UNIVERSITY OF LEEDS

# Improving Task Difficulty Modelling for Robot Teaming in Multi-Dimensional Contexts

## with Applications in Performance Prediction and LLM-Driven Multi-Robot Planning

**Yuhui Wan**

**Submitted in accordance with the requirements for the degree of PhD in Mechanical Engineering**

**The University of Leeds**

**School of Mechanical Engineering**

**June 2025**

# Intellectual Property and Publication Statements

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and the other authors to this work has been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 3 is based on work from jointly authored publication: "3D Fitts' Law for Performance Prediction of Human-Machine Teaming", Yuhui Wan and Chengxu Zhou, Published by IEEE Transactions on Industrial Informatics [1].

Chapter 4 is based on work from jointly authored publication: "Predicting Human-Robot Team Performance Based on Cognitive Fatigue", Yuhui Wan and Chengxu Zhou, Published by International Conference on Automation and Computing [2].

Chapter 5 is based on work from jointly authored publication: "Performance and usability evaluation scheme for mobile manipulator teleoperation", Yuhui Wan, Jingcheng Sun, Christopher Peers, Joseph Humphreys, Dimitrios Kanoulas, and Chengxu Zhou, Published by IEEE Transactions on Human-Machine Systems [3].

In publication "3D Fitts' Law for Performance Prediction of Human-Machine

Teaming" Yuhui Wan developed the task modelling method, designed and performed the experiment, collected the data, and wrote the paper. Dr. Chengxu Zhou directed me and provided guidance during the research.

In publication "Predicting Human-Robot Team Performance Based on Cognitive Fatigue" Yuhui Wan developed the task modelling with fatigue, designed and performed the experiment, collected the data, and wrote the paper. Dr. Chengxu Zhou directed me and provided guidance during the research.

In publication "Performance and usability evaluation scheme for mobile manipulator teleoperation" Yuhui Wan developed the evaluation scheme, designed and performed the experiment, collected the data, and wrote the paper. Dr. Chengxu Zhou directed me and provided guidance during the research. Jingcheng Sun, Christopher Peers, and Joseph Humphreys helped with the experiment and data collecting.



I acknowledge the use of the closed-source model Chatgpt-4o (Open AI, link: https://chat.openai.com/) to improve grammar and proofread my final draft. Also, I acknowledge the use of open-source models: codegeex4-all-9b, deepseek-r1-distill-qwen-32b, internlm3-8b-instruct, internvl2.5-26b, internvl2.5-38b, internvl2.5-78b, llama-3.1-70b-instruct, llama-3.1-8b-instruct, llama-3.1-nemotron-70b-instruct, llama-3.3-70b-instruct, llama3.1-70b, qwen2-72b-32k, qwen2-vl-72b-instruct, qwen2.5-14b-instruct, qwen2.5-32b-instruct, qwen2.5-32b-instruct-vllm, qwen2.5-72b-32k, qwen2.5-72b-instruct, qwen2.5-72b-instruct-lmdeploy, qwen2.5-7b-instruct, qwen2.5-coder-7b-instruct, qwq-32b-preview, qwq-32b-preview-awq as part of the research, as detailed in Chapter 6.

# Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Chengxu Zhou, Professor Robert Richardson, Dr. Andrew R. Barber, and Dr. George Jackson-Mills. Their guidance, support, and encouragement throughout the course of my research have been invaluable. I am sincerely grateful for their insightful advice and constructive feedback, which have shaped both this thesis and my development as a researcher.

I would also like to thank Jingcheng Sun, Christopher Peers, and Joseph Humphreys for their assistance with the experiments and data collection. Their help was crucial to the success of this work.

I would also like to thank my family for their unwavering support and encouragement throughout my academic journey. Their belief in me and generous sponsorship have been instrumental in enabling me to pursue my studies.

# Abstract

This thesis introduces a novel framework for modelling task difficulty in Human-Machine Teaming (HMT), inspired by Fitts' Law and extended into a six-degrees-of-freedom spatial domain. The proposed method accounts for both translational and rotational constraints between machine agents and their targets, enabling precise HMT performance prediction in complex, real-world tasks. By integrating cognitive fatigue into the model using the SAFTE (Sleep, Activity, Fatigue, and Task Effectiveness) framework, the approach holistically captures both long-term skill levels and short-term cognitive effectiveness of human operators. This enables realistic and adaptive forecasting of team performance under varying operational demands. The framework supports multiple applications. First, it provides a robust predictive model for HMT performance, useful for mission planning, workload estimation, and system adaptation. Second, it enables quantitative evaluation through a hybrid scheme combining objective measures, predictive curves, and subjective assessments (e.g., NASA-TLX, SUS). The model has been validated through a comprehensive human study, encompassing a simulation and real-world experiments involving the teleoperation of a quadruped mobile manipulator with different interfaces. Finally, this task difficulty model is adapted to support decision-making in large language model (LLM)-driven multi-robot task allocation. We introduce FittsPrompt, a pre-processing scheme that abstracts

spatial complexity into structured difficulty descriptors. This abstraction allows LLMs to make more efficient and scalable task allocation decisions compared to raw observation inputs. Evaluations across 42 open- and closed-source LLMs demonstrate that the proposed approach not only surpasses traditional baseline methods but also outperforms expert human planners in real-world robotic task allocation.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Modelling task and predicting performance

Human-Machine Teaming (HMT), especially Human-Robot Teaming (HRT), represents a pivotal advancement in intelligent systems design, enabling coordinated and synergistic interaction between human operators and autonomous or semi-autonomous machines. This paradigm facilitates the execution of complex tasks in highly dynamic and unstructured environments by leveraging the complementary strengths of human cognitive flexibility and machine precision and endurance. The importance of HMT has been increasingly recognised across a spectrum of critical domains, including high-throughput manufacturing environments [4], next-generation aerial mobility and autonomous flight operations [5], advanced aerospace missions requiring real-time adaptability [6], and disaster and emergency response scenarios demanding robust situational awareness and rapid intervention [7]. In these settings, human-machine collaboration is not merely a convenience but a necessity for achieving mission success, particularly where full autonomy remains infeasible or undesirable due to ethical, legal, or technical

constraints.

A central challenge in designing effective HMT systems is the accurate prediction of team performance under varying operational conditions. Such prediction is essential not only for real-time decision support and adaptive task allocation but also for the pre-deployment planning phase, where understanding potential bottlenecks and operator workload is crucial. Accurate performance forecasting allows system designers and operators to anticipate risks, optimise the allocation of human and machine resources, and refine training and control interfaces. Moreover, predictive models serve as foundational tools for safety assurance, helping to prevent human errors and machine failures by proactively identifying mismatches between task demands and system capabilities [8]. As HMT systems become increasingly embedded in safety-critical applications, the demand for robust performance modelling frameworks that account for human variability and environmental uncertainty continues to grow.

Despite rapid advancements in robotic automation, sensor technologies, and artificial intelligence, accurately forecasting the time required for HMT to complete specific tasks in real-world conditions remains an enduring and complex challenge. While significant progress has been made in modelling operator behaviour and task execution in constrained environments, the variability and uncertainty inherent in open-world scenarios continue to hinder precise performance prediction [2], [4]. Factors such as unpredictable terrain, inconsistent sensor data, dynamic goals, and variations in human operator proficiency all contribute to the difficulty of generating reliable models. This complexity is further exacerbated in heterogeneous HMT systems, where task requirements often demand nuanced coordination between different types of robots and human roles, each possessing distinct motion capabilities and functional limitations [9], [10].

This study seeks to address these multifaceted challenges by enhancing models of task difficulty and human skill, with the goal of enabling more robust, accurate, and context-sensitive performance forecasting. Central to the approach is the extension of Fitts' Law—a well-established predictive model of human motor behaviour—into a three-dimensional context suitable for embodied interaction between agents and their physical environment. Traditional Fitts' Law formulations focus on 2D target acquisition tasks, limiting their applicability in robotic settings where movement occurs across spatial planes and includes orientation constraints. By adapting the law to a 3D framework and incorporating motion capability identification for both human and robotic agents, the method captures not only the spatial distribution of targets but also the dynamic relationship between agent positioning, reachable workspace, and environmental constraints.

One of the persistent and pressing challenges in HMT applications is the accurate prediction of task execution time in real-world scenarios. Unlike controlled laboratory environments, real-world operational contexts are often characterised by uncertainty, variability in human behaviour, and unpredictable environmental dynamics, all of which complicate performance estimation. The difficulty lies in modelling the intricate interactions between human cognitive processes, such as attention, decision-making speed, and fatigue, and machine dynamics, which may include latency, mechanical limitations, or perception errors.

Accurately forecasting how long a human-machine team will take to complete a given task is not just a matter of academic interest; it has profound implications for mission success, resource allocation, and operator safety. Performance prediction serves as a cornerstone for numerous downstream applications, including adaptive scheduling, dynamic replanning, interface personalization, and supervisory control. For instance, in time-critical domains such as emergency response or

autonomous aerial surveillance, delays in execution can lead to mission failure or even endanger human lives. Furthermore, inaccurate predictions can undermine trust in automation, especially in settings where humans rely on systems to operate reliably under variable conditions [11]. Thus, developing robust models that can predict HMT performance with high fidelity remains a fundamental research challenge with high practical relevance.

Given the inherent complexity of human-machine systems, comprehensively modelling performance requires accounting for a wide range of variables, including machine dynamics, human cognitive and physical capabilities, and potential sources of system-level error. Explicitly modelling each of these elements is not only computationally intensive but also often infeasible due to the unpredictability and variability introduced by human behaviour and environmental factors. To address this challenge, the approach adopts Fitts' Law [12] as a unifying framework, treating the human-machine team as an integrated entity. In this formulation, task-relevant parameters—such as latency, dexterity, precision, and control variability—are implicitly incorporated into the predictive model as hidden variables, allowing for a more abstract yet effective representation of overall system performance.

To further adapt this framework to the spatially rich and physically grounded context of real-world HMT, this study extends Fitts' Law into a 3D formulation encompassing six degrees of freedom (DoF)—three for translation and three for orientation. This extension enables the model to more accurately capture the spatial and kinematic relationship between the machine agent and its target, reflecting real-world challenges such as approach angles, grasp alignment, and workspace constraints. By doing so, this study not only improve the precision of task difficulty estimation but also enable the identification of motion capa-

bilities that are essential for generating standardised benchmark tasks for HMT performance assessment.

Crucially, the method builds upon and advances prior work in the human-computer interaction (HCI) community [13], [14], where extensions of Fitts' Law have been applied to two-dimensional interfaces and limited 3D motion contexts. However, these approaches often overlook the complexities introduced by full spatial orientation, which are fundamental in robotics and HMT scenarios. By incorporating both translational and rotational components, the framework represents the first known extension of Fitts' Law tailored specifically for real-world HMT applications. This novel contribution fills a critical gap in existing models and provides a robust foundation for predicting performance in complex, unstructured, and multidimensional task environments.

Predictive models for HMT performance span a wide array of methodologies, each contributing unique insights into various dimensions of human-machine interaction [15]. A significant portion of the literature has focused on human-centric factors, acknowledging the critical influence of human cognition, attention, and trust in shaping collaborative outcomes. For example, cognitive models have been developed to represent mental workload and decision-making under varying task loads [16], while other studies have explored the impact of operator trust, situational awareness, and automation transparency on performance and safety [17]. Further extending the human-centric lens, operator models have been proposed to capture the semantic structure of working memory, helping to explain how humans perceive, store, and recall task-relevant information during interaction with autonomous systems [18].

In parallel, efforts have been made to quantify the scalability of HMT systems, particularly in multi-robot scenarios. The concept of neglect tolerance—which

estimates how long a robot can operate autonomously without human intervention—has been used to determine the maximum number of robots an operator can supervise effectively, providing valuable guidelines for workload balancing and autonomy allocation [19]. Beyond static models, dynamic approaches such as higher-order Markov chains have been employed to learn and predict sequential human behaviour, such as assembly steps in collaborative manufacturing tasks, enabling anticipatory planning for the robotic partner based on historical action sequences [20].

Physiological sensing has also emerged as a promising avenue for performance prediction. Studies have demonstrated the feasibility of using real-time neurophysiological signals, including brain activity, heart rate variability, and eye movement, to assess situational awareness and cognitive engagement during teleoperated exploration tasks with drones [21]. These biometrics-driven approaches offer a direct window into human mental states, potentially enabling adaptive systems that respond dynamically to fluctuations in operator capacity.

While these diverse models provide important building blocks for understanding HMT performance, they often suffer from key limitations. Most notably, they tend to prioritise human state modelling in isolation, frequently overlooking the mechanical, spatial, and task-specific characteristics of robotic agents and mission contexts. Additionally, many of these approaches lack standardised evaluation frameworks or real-world benchmarks, making it difficult to generalise findings across different platforms and operational settings. These limitations highlight the need for an integrated, task-centred modelling approach—such as the one proposed in this thesis—that captures the coupled dynamics of both human and machine contributions to performance.

Fitts' Law [12], originally developed within the field of human motor control and

later widely adopted in HCI, provides a robust mathematical model for predicting the time required to move a pointer to a target based on the distance to the target and its size. As a foundational tool in interface design, Fitts' Law enables system developers to estimate user effort and efficiency in selecting on-screen elements, making it a key component of usability studies. Its strength lies in its simplicity and predictive power for 1D and 2D planar movements, typically modelled as logarithmic functions of the ratio between movement amplitude and target width. However, this simplicity also introduces limitations: the model's traditional formulation assumes point-to-point linear movements in a one-dimensional space and treats the motor task as a purely human-centric activity, without accounting for physical embodiment or environmental complexity.

While extensions of Fitts' Law have been proposed—incorporating variables such as cursor diameter, angular displacement, and probabilistic target areas—most of these remain grounded in virtual or simplified 2D user interface contexts [22]–[26]. These adaptations are primarily geared toward improving interface design for traditional computing systems rather than addressing the full complexity of physical interactions in robotics or HMT applications. As a result, the applicability of these models to real-world, embodied tasks remains limited.

A growing number of studies have attempted to bridge this gap by applying Fitts' Law and its derivatives to real-world tasks involving physical systems and robots [27], [28]. However, even in these efforts, the role of machines is often conceptualised as passive extensions of human operators—i.e., tools or actuators controlled entirely by human input. Consequently, the predictive frameworks continue to emphasise human movement characteristics while mainly neglecting the mechanical constraints, actuation delays, spatial configurations, and control fidelity of the robotic systems involved. This perspective severely limits the utility

of Fitts' Law in evaluating or forecasting performance in fully integrated HMT systems, where both the human and machine contribute actively and jointly to task execution.

Therefore, there remains a critical need to develop a formulation of Fitts' Law that not only generalises to higher-dimensional, embodied contexts but also recognises the autonomy and capability of robotic agents as active participants in task execution. The work presented in this thesis addresses this gap by proposing a novel extension of Fitts' Law to six degrees of freedom, tailored specifically for spatial HMT scenarios, and modelling performance as a function of both human and machine factors within a unified predictive framework.

As a result, there remains a substantial gap in existing task difficulty models, which often struggle to offer a comprehensive framework capable of capturing the diverse and intricate nature of HMT systems and the varied tasks they must perform. Current approaches frequently fall short in addressing key challenges such as accurately characterising machine motion capabilities, modelling multi-dimensional task constraints, and generalising to real-world operational environments. These limitations underscore the urgent need for predictive frameworks that are both holistic—capturing the interaction dynamics of human and machine agents—and scalable—applicable across a broad range of task scenarios and robotic platforms.

In response to these challenges, this thesis introduces a novel and unified task difficulty modelling framework for predicting the performance of HMT systems in realistic operational settings. The method bridges insights from human-computer interaction, embodied cognition, and robotics, offering a multidimensional modelling approach grounded in an extended formulation of Fitts' Law. The key contributions of the work are as follows:

1. **This study proposes a systematic method to model HMT task difficulty and predict real-world HMT performance.** The approach comprises three key components:

    - The creation of an expansive mathematical model of HMT tasks by extending Shannon's index of difficulty formulation into 3D space with 6 DoFs, considering both translation and orientation between the target and the agent.

    - The development of principles for designing standardised benchmark tasks, which ascertain the motion capabilities of the targeted machine agent, laying the foundation for a comprehensive model for all subsequent tasks.

    - The development of a formulated performance prediction model based on Fitts' Law, employing standard tasks to understand the characteristics of the HMT and predict future system performance.

    - Implementation of a method to identify the minimum set of standard tasks required for accurate prediction based on the machine agent's motion capabilities.

2. **Following the theoretical contributions, this study conducted two empirical validations to demonstrate the model's efficacy and applicability in diverse HMT systems:**

    - A web-based simulation involving 16 human participants using a remote control excavator, providing insights into virtual HMT interactions.

    - A real-world experiment with a quadruped manipulator robot, involving 7 human users and 2 distinct human-machine interfaces, to assess

the model in a tangible, real-life scenario.

Together, these contributions advance the state of the art in HMT performance modelling by offering a robust, interpretable, and generalizable framework that captures the multidimensional nature of collaborative human-robot tasks. This framework not only improves predictive accuracy but also supports practical applications such as task allocation, human-robot interface design, and adaptive autonomy in real-world deployments.

## 1.2 Employing Quadruped Manipulator

This study chooses the mobile manipulators as an example of robot agents (RAs). Nowadays, RAs are increasingly deployed by public safety and emergency response agencies to support high-risk missions [29], including explosive ordnance disposal (EOD) [30]. In these contexts, mobile manipulators—particularly those based on quadruped locomotion platforms—are gaining traction due to their versatility and operational advantages over human agents (HAs) in hazardous environments.

In detail, a quadruped manipulator was used in the experiment. In certain mission scenarios, quadruped manipulators offer clear benefits when compared to human first responders. For example, a human operator wearing a level-A hazardous materials (HAZMAT) suit and a self-contained breathing apparatus (SCBA) faces severe operational constraints. The duration of the mission is directly limited by the capacity of the oxygen tank in the SCBA and further constrained by the physical workload, body heat retention, and the additional weight of the protective gear [31]. These physiological and ergonomic limitations can significantly reduce the efficiency and safety of human responders in prolonged or physically demanding tasks.

By contrast, a quadruped robot's operational limits are primarily defined by its battery life—typically ranging from 2.5 to 4.5 hours for commercially available platforms such as the Unitree AlienGo—and can be extended with the integration of external power sources. In addition to their operational endurance, quadruped manipulators offer logistical and economic advantages. The long-term deployment and maintenance cost of robotic systems is often lower than that of training, equipping, and sustaining human teams for comparable missions. Furthermore, the physical footprint of these robots is generally smaller than that of a human responder in full protective gear, allowing for greater manoeuvrability in tight or cluttered environments such as collapsed structures or chemically contaminated areas—conditions frequently encountered in HAZMAT operations.

Given these benefits, quadruped manipulators can, in specific mission profiles, outperform human responders in terms of safety, endurance, and task accessibility. However, to fully exploit the potential of these robotic platforms, it is essential to integrate them into a broader framework that allows seamless collaboration with human intelligence. In this regard, a well-designed Human-Machine Teleoperation Interface (HMTI) plays a critical role. Such interfaces serve as the conduit through which human expertise and real-time decision-making can be effectively coupled with robotic precision and durability, enabling optimal performance of the integrated system across a wide range of mission scenarios.

## 1.3 Considering Fatigue in Prediction

The preceding section introduced a methodology for evaluating HMT performance through predictive modelling and empirical validation. While these approaches effectively capture task complexity, interface usability, and robotic capabilities, human cognitive factors—particularly fatigue—remain an essential yet

challenging element to model precisely. Building upon the extended Fitts' Law framework introduced in the modelling task and predicting performance section, and the evaluation methodologies from the Evaluating Human-Machine Teaming Performance section, this section further enhances the predictive model by explicitly incorporating cognitive fatigue. By integrating physiological models of human cognitive effectiveness with task difficulty estimations, the refined approach offers a more holistic and realistic prediction of HMT system performance under varying operational conditions.

Predicting the performance of a robotic system for a given task, especially when parameters such as distance, orientation, and mechanical constraints are well-defined, is typically a deterministic and tractable problem. Robotic motion models and control algorithms offer high degrees of predictability and repeatability under such conditions. However, in the context of MHT, performance prediction becomes significantly more complex and less intuitive. This complexity arises from the incorporation of human agents whose cognitive and physical states vary dynamically throughout task execution.

Among the various human factors influencing HMT performance, *cognitive fatigue* plays a particularly critical role. Fatigue can degrade human decision-making accuracy, reaction time, motor coordination, and situational awareness—factors that directly impact task effectiveness and team safety. Unlike mechanical degradation in robots, which can often be quantified through sensor feedback and predictive maintenance models, human fatigue manifests non-linearly and can be influenced by a variety of interacting variables, including task complexity, workload, duration of operation, environmental stressors, and individual differences.

Modelling HMT performance with cognitive fatigue considerations introduces unique challenges due to the stochastic and individualised nature of human be-

haviour. Traditional task modeling frameworks typically assume a constant or ideal level of human performance, thereby failing to capture the performance degradation associated with fatigue over time. Nonetheless, incorporating cognitive fatigue into predictive models could offer valuable insights for mission planning, interface adaptation, workload balancing, and the design of more robust shared autonomy systems.

Fatigue is a well-documented risk factor in the operation of complex machinery, including vehicles and robotic systems. In human-in-the-loop systems, fatigue has been shown to significantly impair operator judgment, reaction time, and overall situational awareness—factors that can lead to critical errors. The aviation sector, for example, has extensively studied fatigue-related risks. According to the Federal Aviation Administration (FAA), approximately 21% of reports submitted to the Aviation Safety Reporting System (ASRS) mention pilot or crew fatigue, and 3.8% of these reports directly attribute incidents to fatigue-related causes [32]. Similarly, data from the U.S. National Highway Traffic Safety Administration (NHTSA) estimate that drowsy driving contributes to over 100,000 crashes annually, resulting in more than 1,500 fatalities and 71,000 injuries [32].

While the transportation industry has developed robust fatigue models and mitigation protocols, corresponding research in the field of robotics—particularly teleoperated and human-supervised systems—remains limited. Nevertheless, emerging evidence suggests that fatigue poses a comparable threat in robotic domains. For instance, a study analysing 237 reported robotics incidents revealed that 63.27% were directly attributed to excessive workload and operational fatigue, exacerbated by factors such as insufficient staffing, night shifts, and lack of rest [33]. Moreover, 31.39% of the root causes for unsafe behaviour in robot operation cases were linked explicitly to night-time operations and fatigue accumulation. These

findings underscore the pressing need to account for cognitive fatigue in predictive models for Human-Robot Teaming, particularly in mission-critical or round-the-clock deployments.

Human-robot teams are increasingly deployed in mission-critical domains such as search and rescue, firefighting, and EOD, where operational reliability, rapid decision-making, and sustained performance are essential. In such high-stakes scenarios, assessing whether a human operator is cognitively fit for the mission—and predicting how their performance may evolve over time—becomes a vital component of the overall system's effectiveness and safety. Traditional models often overlook the impact of operator fatigue, despite its critical influence on human reliability and decision-making.

To address this gap, I introduces a performance prediction model that explicitly incorporates the cognitive effectiveness of human operators by modelling fatigue dynamics over time. The approach quantifies cognitive fatigue by jointly considering the demands of the task and the operator's physiological sleep condition, based on a homeostatic sleep regulation framework. This enables the model to dynamically estimate an operator's *cognitive effectiveness*—a continuous measure reflecting their mental readiness and performance capacity at any given time. By integrating this fatigue-aware component into task performance prediction, the model offers a more realistic and adaptive foundation for mission planning, human-robot task allocation, and real-time supervisory control.

The operator's cognitive effectiveness level can also be leveraged to enhance the accuracy of human-robot team performance prediction. Traditional evaluation frameworks often separate system performance—measured through objective metrics such as task completion time or success rate—from user experience, which is typically assessed through subjective instruments such as workload or

usability scales [34], [35]. While this bifurcation allows for modular analysis, it fails to capture the intertwined nature of human cognitive state and system-level outcomes during dynamic operations.

In particular, operator fatigue—reflected in diminished cognitive effectiveness—can have a profound impact on real-time decision-making, control precision, and error rates, all of which directly influence the overall performance of the human-robot team. As missions grow in complexity and duration, overlooking this factor can lead to overly optimistic or inaccurate performance estimates. Therefore, this study argues that integrating the operator's cognitive effectiveness with their skill level is essential for a more comprehensive and realistic performance prediction model. By accounting for both long-term proficiency and short-term cognitive readiness, the approach provides a unified framework that better reflects the dynamic capabilities of human-robot teams in real-world scenarios.

The proposed model integrates the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model [36] with an extended formulation of Fitts' Law [35], as illustrated in Fig. 4.1. The SAFTE model is a validated biomathematical framework that simulates human cognitive effectiveness based on three key physiological components: the homeostatic sleep drive, circadian rhythm, and sleep inertia. These elements collectively determine fluctuations in an operator's mental readiness over time, accounting for sleep history, time of day, and recovery cycles.

This study enhances the SAFTE model by coupling it with a task demand model derived from the extended Fitts' Law. This formulation, which characterises task difficulty in (6-DoF), enables a more precise estimation of the cognitive resources required to perform specific motion-level tasks. This study uses the difficulty score to modulate the depletion rate of attention capacity in the SAFTE model, effectively amplifying fatigue effects during cognitively demanding tasks.

The resulting cognitive effectiveness value reflects an operator's real-time mental readiness, which this study uses to assess their suitability for engaging in robotic teleoperation. In parallel, operator skill level is independently quantified through performance on standardised tasks defined by the extended Fitts' Law, representing long-term proficiency. By integrating these two components—*cognitive effectiveness* (short-term readiness) and *skill level* (long-term capability)—alongside specific mission requirements, the model enables accurate prediction of human-robot team performance, particularly in terms of task execution time and anticipated variability under fatigue.

The contribution to the field lies in the development of a novel performance prediction model that simultaneously accounts for the cognitive fatigue and skill level of human operators, offering a more accurate and realistic representation of human-robot team performance. By integrating these two critical human factors—short-term cognitive effectiveness and long-term operational proficiency—into a unified predictive framework, the model provides actionable insights into how team performance may fluctuate under varying task demands and fatigue conditions.

The adoption of this model enables stakeholders to make more informed and adaptive decisions regarding mission assignments, operator scheduling, and autonomy allocation. In high-stakes and time-sensitive domains, such as emergency response, this can significantly enhance both the efficiency and safety of operations. Moreover, this research lays the foundation for future investigations into the temporal dynamics of fatigue and its broader implications for human-robot collaboration. By incorporating physiological and cognitive considerations into robotic system design and planning, this work contributes to the ongoing effort to increase the robustness, adaptability, and human-centeredness of robotic de-

ployments across a wide range of real-world applications.

## 1.4 Evaluating Human-Machine Teaming

As outlined in the previous section, the task difficulty modelling and performance prediction framework developed in this thesis provides a robust approach for anticipating HMT performance across a diverse range of operational scenarios. It can serve as a foundational component for comprehensively evaluating HMT systems. By quantifying task difficulty and forecasting performance, the proposed predictive model is integrated with additional statistical evaluation tools and subjective assessment measures, thus providing a robust methodology for assessing the effectiveness of various human-robot interaction strategies. In the following subsections, this study outlines the systematic evaluation approach, demonstrating how these predictive insights are empirically validated through rigorous experimental testing with real-world teleoperation interfaces and robotic platforms.

As modern control methodologies continue to advance, RAs have become increasingly capable and intelligent, particularly within structured environments. With the integration of artificial intelligence (AI), current-generation robotic systems can now operate with near-full autonomy in domains such as manufacturing and warehouse logistics, executing repetitive and well-defined tasks with high precision and reliability. Despite these achievements, significant challenges remain when robots are deployed in unstructured, open-world environments, where unpredictability, variability, and high-stakes decision-making prevail.

In such contexts—especially in high-risk, mission-critical scenarios—fully autonomous operation is often infeasible. Tasks such as HAZMAT rescue, chemical decontamination, and EOD exemplify the types of operations that require not only complex real-time interaction with dynamic environments but also deep do-

main knowledge, situational awareness, and adaptive judgment—traits that are inherently human. HAs bring valuable expertise and flexibility to these missions, but their direct physical involvement also exposes them to considerable risks, including injury or death.

Motion-level teleoperation has emerged as a viable solution to reconcile the complementary strengths of both human and robotic agents while mitigating human exposure to danger. In this paradigm, RAs serve as the physical executors of tasks in hazardous zones, while HAs provide high-level cognitive guidance and control from a remote location. This division of labour enables both agents to operate within their respective domains of strength: RAs handle physical manipulation and traversal under challenging conditions, while HAs contribute domain-specific reasoning and strategic decision-making [37].

Therefore, evaluating such HMT in real-world conditions necessitates not only an assessment of autonomous capabilities but also a rigorous examination of teleoperation interfaces, shared autonomy strategies, and performance prediction models. The subsequent sections present the evaluation methodology, including both simulation-based and real-world experiments, designed to validate the proposed predictive framework across diverse HMT applications.

While mobile manipulators are becoming increasingly prevalent in field operations, research explicitly targeting teleoperation strategies and interface design for these systems remains limited [38]. This is especially notable considering the rapid evolution of HMTs, including HMTIs, which now incorporate a wide array of emerging technologies. Recent developments have introduced novel modalities such as inertial measurement units (IMUs), vision-based recognition, wearable motion capture systems, and even haptic feedback mechanisms. Each of these innovations contributes uniquely to control precision, user experience, and situa-

tional awareness in remote operation.

Yet, the growing diversity of HMTI designs presents a major obstacle to systematic comparison. Due to their varying input modalities, degrees of embodiment, and interaction metaphors, it is currently infeasible to evaluate these interfaces side by side using a consistent standard. This heterogeneity creates a barrier to empirical validation and prevents the community from drawing generalizable conclusions about interface performance. As such, the establishment of a standardised HMTI evaluation scheme is critical. A robust and repeatable evaluation framework would not only facilitate the benchmarking of teleoperation systems but also guide future development by identifying best practices and quantifiable metrics for performance across diverse robotic platforms.

This study presents a standardised HMT evaluation scheme specifically designed for mobile manipulators. This scheme enables a comprehensive assessment of HMTs and HMTIs by integrating a suite of robot motion tests, both objective and subjective evaluation metrics, and a quantified performance prediction model. The evaluation includes statistical side-by-side comparisons of motion execution times across different interface types, alongside first-hand user feedback on system usability and workload. To ensure predictive utility, the performance model incorporates both human and robot system characteristics, leveraging data from standardised tasks to forecast performance in future real-world missions.

To validate and refine the proposed evaluation scheme, this study conducted an experimental study comparing two HMTIs—namely, a conventional gamepad and a wearable motion capture system (WMCS)—for teleoperating a quadruped mobile manipulator. The main contributions of this work are summarised as follows:

1. **A standardised HMTI evaluation scheme for mobile manipulators**,

composed of three key components:

- A suite of standard motion tests, evaluating locomotion, manipulation, and combined control performance.
- An objective evaluation protocol based on statistical analysis of operator motion time across interface types.
- A generalisable performance prediction model from the previous section, extended from Fitts' Law, enabling performance forecasting in unseen tasks using data from standard tests.

2. **Standardised subjective evaluation metrics**, including the NASA Task Load Index (NASA-TLX) and the System Usability Scale (SUS), to assess perceived workload and system usability.
3. **Empirical comparison of two HMTIs for quadruped manipulator teleoperation**, demonstrating the scheme's utility in evaluating both performance and usability. The study benchmarks the gamepad and WMCS interfaces through real-world experimentation, enabling a data-driven understanding of interface trade-offs.

## 1.5 Apply Task Modelling with LLM

The preceding sections introduced a task difficulty modelling framework that captures the spatial and kinematic intricacies of HMT systems. Building upon this foundation, this section explores the integration of task modelling with Large Language Models (LLMs) to enhance autonomous decision-making for task allocation in multi-robot systems. By abstracting complex spatial and environmental data into structured task difficulty representations, the proposed approach leverages the predictive power of LLMs to achieve more efficient, scalable, and

context-aware planning and coordination among heterogeneous robotic agents in dynamic and unstructured environments.

In modern manufacturing environments, multi-robot systems have demonstrated significant success, primarily owing to the structured nature of these settings and the use of task-specific programming developed by skilled engineers [39]. Within these controlled environments, each robot is typically assigned a predefined role, executing repetitive motions and trajectories that have been meticulously designed and optimised offline. This high degree of determinism facilitates seamless coordination, minimises conflict or redundancy among RAs, and leads to high operational efficiency.

However, the deployment of multi-robot systems in unstructured, open-world, or dynamic environments remains a formidable challenge. Outside the confines of factories and warehouses, such systems encounter unpredictable terrain, variable task demands, incomplete information, and rapidly changing mission objectives. In these contexts, robots must be capable of real-time sensing, reasoning, and decision-making without relying on pre-scripted control policies. The absence of fixed infrastructure and human supervision further complicates coordination. As a result, traditional multi-robot paradigms often fail to achieve the same level of reliability or efficiency in these settings [40].

This disparity highlights the need for advanced methods in autonomous task allocation, environmental understanding, and adaptive behaviour for multi-robot teams operating in complex environments. In particular, there is growing interest in leveraging learning-based, distributed, and human-in-the-loop strategies to improve the resilience and autonomy of such systems beyond the factory floor.

Using AI systems, including LLMs, has emerged as a powerful tool for generalisation and decision-making [41]–[45], offering the ability to process diverse inputs

and adapt to various tasks. Others have already recognised these problems, and efforts to find appropriate solutions for them have been underway in the field of robot learning [46], [47]. Many distinct categories of approaches have emerged.

The first category follows a one-model-fits-all paradigm, where a single neural network is trained to map raw sensory inputs (e.g., camera images) directly to robot actions (e.g., end-effector velocities) [48]–[50]. While promising in its generality, this approach encounters several critical limitations. A major challenge lies in the substantial data requirements for training such models. Unlike in simulation, where data can be generated in large quantities and at low cost, real-world robotics data collection is time-consuming, expensive, and constrained by hardware wear, safety concerns, and environmental variability. Scaling data collection to encompass the diversity of tasks, objects, and environments required for generalisation remains a significant bottleneck.

Another inherent drawback is the risk of catastrophic forgetting—a phenomenon in which pre-trained models lose previously acquired knowledge when fine-tuned on new tasks or domains. This issue is particularly problematic in robotics, where continual learning and task adaptation are essential. Without mechanisms for retaining and integrating prior knowledge, one-model systems may underperform or fail outright when deployed in scenarios that differ from their fine-tuning data.

Additionally, this approach often assumes a monolithic model without external system dependencies, which conflicts with the modular and heterogeneous nature of practical robotic systems. In real-world applications, frameworks such as the Robot Operating System (ROS) are widely adopted due to their flexibility, modularity, and support for diverse hardware and software components. A single end-to-end neural model may lack the structural modularity required for seamless integration with such frameworks, limiting its adaptability to changes

in hardware platforms, sensing modalities, or task configurations.

In summary, while the one-model-fits-all strategy holds promise for general-purpose learning, it faces considerable hurdles regarding scalability, data efficiency, lifelong learning, and practical deployment. These challenges underscore the need for more robust, modular, and context-aware approaches that can support real-world multi-robot collaboration in complex and evolving environments.

Noticing the limitations of monolithic, data-intensive learning systems, researchers have increasingly turned to LLMs as a promising alternative for robotic decision-making and task planning [51]–[54]. LLMs offer strong symbolic reasoning capabilities and broad generalisation across diverse tasks, making them well-suited to serve as high-level priors in robotic frameworks. Several recent systems exemplify this trend. For instance, LLM-Planner [55] demonstrates few-shot grounded planning for embodied agents using LLMs, while LaMI [56] enables enhanced multi-modal human-robot interaction through language-conditioned policies. Other works translate natural language commands into structured representations such as behaviour trees or state machines [57], [58], facilitating interpretable and modular task execution.

These approaches aim to leverage the general-purpose capabilities of pre-trained language models without requiring task-specific retraining from scratch. However, they introduce a distinct set of challenges. A fundamental limitation is the assumption that a pre-defined and finite set of actions is available to the robot, along with sufficient task-specific knowledge to ensure successful execution [55], [59]. In practice, robotic actions are often highly contextual and may need to be dynamically adjusted or even discovered through environmental interaction. Rigidly pre-defining action spaces restricts a system's adaptability, particularly in real-world environments where task variability and unforeseen events are com-

mon. This rigidity hampers the system's ability to generalize, reducing robustness in unfamiliar settings that deviate from training data distributions.

Given their capacity to understand abstract instructions and reason over general logic, LLMs hold considerable promise for enabling autonomous decision-making in dynamic and unstructured environments. In single-robot scenarios, task allocation is relatively straightforward, since only one agent is available, the question of who should perform the task does not arise. Most focus instead lies in interpreting, planning, and executing the task effectively, for example, Imitation Learning. However, extending this paradigm to multi-robot systems introduces a new layer of complexity. The presence of multiple heterogeneous RAs necessitates an explicit decision-making process to determine which robot is best suited for each task, based on capabilities, proximity, workload, and environmental constraints. Moreover, this decision must be made dynamically, often with incomplete information and under time pressure, as conditions evolve in real-time. Incorporating LLMs into such a framework raises fundamental questions: How should task-relevant context be encoded and communicated to the model? Can LLMs reason over robot capabilities and constraints? And to what extent can language models support decentralised coordination among autonomous agents? These challenges highlight the need for new architectures that tightly couple the reasoning power of LLMs with real-world action grounding and coordination strategies suitable for multi-robot systems operating in open-ended environments.

One of the most significant challenges in multi-robot task allocation in open-world settings lies in the nature and structure of the input information. Unlike structured environments where the state space is well-defined and relatively static, open-world scenarios demand that robots interpret, reason, and act based on complex, high-dimensional, and often unstructured raw sensory data. To enable

effective decision-making, models typically rely on large-scale raw input from visual, spatial, and semantic observations to form a coherent understanding of the task environment. This includes detailed spatial configurations of robots and targets, dynamic obstacles, terrain variability, and inter-agent relationships—all of which are crucial for determining which agent is best suited for which task, and how tasks should be sequenced or distributed across the team.

While LLMs excel at processing symbolic and textual data, they are not inherently designed to parse or reason over dense spatial information. The intricate and high-dimensional nature of real-world spatial relationships, such as three-dimensional (3D) translation positions, orientations, visibility constraints, reachability, and occlusions, is often underrepresented or abstracted in natural language. Consequently, when tasked with multi-robot planning, LLMs frequently struggle to maintain a precise mental model of the scene, leading to reasoning errors or poor alignment between the plan and the physical environment [60].

A further complication arises from the sheer volume of information that must be encoded into the input prompt. As the number of robots and objects increases, so too does the complexity of the environment and the combinatorial space of possible interactions. In order to provide sufficient context for decision-making, prompts must grow longer to include spatial layouts, robot states, object attributes, and mission constraints. However, LLMs are subject to memory limitations and token constraints that restrict the total volume of input they can effectively process. This often leads to "prompt saturation," where critical details are either omitted or diluted amidst less relevant information, reducing the model's ability to attend to key variables and increasing the risk of suboptimal or erroneous task allocation decisions [61].

These limitations underscore the urgent need for innovative strategies to enhance

LLMs' capacity for spatial reasoning and input abstraction in multi-robot systems. Approaches may include the use of compact scene representations, structured prompt engineering, hybrid architectures that combine language models with geometric or graph-based encoders, or pre-processing layers that translate raw spatial data into semantically meaningful language tokens. Such methods aim to streamline input data, reduce cognitive overload within the model, and maintain the clarity and precision required for effective task assignment and coordination in real-world robotic teams.

To address the aforementioned challenges in spatial reasoning and prompt scalability for multi-robot task planning, this study proposes a novel approach grounded in task difficulty modelling [3], which is itself an extension of the well-established Fitts' Law [12]. Originally developed to model human motor behaviour, Fitts' Law has been extended in this work to quantify the difficulty of robotic tasks by incorporating spatial relationships, target/tool size, and dynamic constraints. By leveraging this extended model, it transforms complex environmental and agent-specific data into a concise, structured representation of task difficulty. This abstraction significantly reduces the need for LLMs to process raw, high-dimensional spatial inputs, instead allowing them to reason over more meaningful and compact descriptors of task complexity.

The method enables LLMs to focus on the most relevant decision-making variables while mitigating the risk of prompt saturation and information dilution. In doing so, it facilitates more scalable and interpretable planning for heterogeneous multi-robot systems operating in dynamic environments.

The major contributions of the work are as follows:

- **FittsPrompt:** This study introduces *FittsPrompt*, a novel preprocessing framework that encodes spatial task complexity into structured, low-

dimensional representations. This preprocessing layer streamlines input for LLMs, enhancing both decision-making quality and robustness in multi-robot task allocation.

- **Integration of Extended Fitts' Law:** This study adapts and operationalize the extended formulation of Fitts' Law into a practical algorithm suited for LLM-driven planning. This allows task difficulty to be numerically quantified and compared across agents and tasks, enabling grounded and interpretable action selection.
- **Comprehensive Evaluation:** The methodology is evaluated through a combination of standardised benchmarks, user studies, and real-world robotic experiments. These evaluations demonstrate the effectiveness and generalizability of the approach across different robotic platforms and environmental settings, with significant improvements observed in planning efficiency and task success rates.
- **Cross-Model Benchmarking:** This study conducts a large-scale comparative analysis of LLM performance in multi-robot planning tasks, evaluating a total of 42 models—38 open-source and 4 closed-source. This benchmark provides side-by-side comparisons of task allocation accuracy, error profiles, and decision consistency across diverse model architectures.

# Chapter 2

# Related work

## 2.1 Fitts' Law and extensions

P. M. Fitts proposed a widely recognised method, known as Fitts' Law, for predicting human-machine interface performance [12]. This foundational model, rooted in information theory, quantifies the trade-off between speed and accuracy in human computer control tasks. It has since become a cornerstone in the evaluation of user interface efficiency and human psychomotor behaviour. Fitts' Law predicts the Movement Time (MT) a user takes to move a pointing device, such as a mouse cursor, to a designated target location through the optimal route, as a function of the Index of Difficulty (ID) of the movement task:

$$\mathrm{MT} = a + b \cdot \mathrm{ID}, \tag{2.1}$$

where the constants $a$ and $b$ are empirically determined coefficients that define a linear regression line, often referred to as the prediction line, linking MT with ID. This relationship reflects a consistent, measurable pattern in human-machine actions under varied task demands. The foundation of the model's predictive

power lies in the accurate representation of task difficulty, encapsulated by the ID metric. In its original formulation, ID considers only two spatial parameters of the task—the distance ($d$) from the starting point to the center of the target and the width ($w$) of the target along the axis of motion:

$$\mathrm{ID} = \log_2\left(\frac{2d}{w}\right). \tag{2.2}$$

This logarithmic formulation mirrors the concept of information transmission, where greater distance and smaller target width imply higher difficulty, thereby requiring more time to complete the task. Notably, in Fitts' original formulation, the ratio $\frac{2d}{w}$ retains physical units (distance over distance), giving ID the dimension of bits per movement. However, many later extensions of Fitts' Law reformulated ID as a dimensionless quantity for generalisability across contexts. In line with these extensions, the ID defined in this study is treated as unitless. It is important to emphasise that the Index of Difficulty, as defined in this study, characterises only the intrinsic properties of the task itself. In other words, ID captures the spatial and geometric constraints of the task, and does not vary according to whether the task is performed by a human operator or a robotic system. This separation ensures that ID remains a purely task-centric metric, while performance differences between agents are reflected in the parameters of the predictive model (e.g., $a$ and $b$), rather than in the formulation of ID. Furthermore, ID should be interpreted as providing a theoretical lower bound on task difficulty, since the model assumes an optimal solution path; in practice, suboptimal strategies or execution variability may result in greater effective difficulty.

Building upon the foundational principles introduced by Fitts, subsequent research explored more nuanced representations of the factors contributing to movement difficulty. A particularly influential refinement is Welford's model [23],

which reexamines the composition of the ID by separating the effects of distance and target width. Unlike Fitts' original formulation, which encapsulates both variables into a single term, Welford proposed a bifactorial model that allows the individual contributions of distance ($d$) and width ($w$) to be explicitly and independently modelled. This approach acknowledges that the two variables may not symmetrically influence user performance, particularly under varying task conditions or cognitive constraints. The movement time is then predicted as a linear function of the logarithm of each variable:

$$\mathrm{MT} = a + b_1 \cdot \log_2(d) + b_2 \cdot \log_2(w), \tag{2.3}$$

where $a$ denotes a constant intercept, and $b_1$ and $b_2$ are empirical coefficients reflecting the relative impact of distance and width on the movement duration. By decoupling the influence of $d$ and $w$, Welford's formulation provides a more flexible and potentially more accurate representation of user behaviour in target acquisition tasks, particularly in scenarios where the spatial properties of the task are complex or asymmetric. This model has been instrumental in deepening the understanding of human-machine performance and has served as the basis for numerous subsequent studies aiming to improve interface design and task performance analysis.

An alternative formulation to Fitts' Law was proposed by Kvålseth, known as the power model formulation [62]. Unlike Fitts' Law, which models movement time as a linear function of the logarithmic index of difficulty and involves two empirical constants, the power model introduces a power-law relationship with three empirically determined constants. This added flexibility allows the model to achieve higher multiple correlations when fitted to experimental data, offering the potential for improved predictive accuracy in certain contexts. The power

model can be particularly useful in scenarios where the linearity assumption of Fitts' Law does not hold or where more complex behaviour must be captured.

Despite these advantages, the power model has not achieved widespread adoption in the human-computer interaction and robotics communities. One of the primary reasons is its increased mathematical complexity, which makes it less intuitive and more computationally intensive compared to the simplicity and interpretability of Fitts' original formulation. Moreover, the added parameter introduces a risk of overfitting in empirical studies with limited sample sizes. As a result, the power model remains a lesser-used alternative, often referenced in theoretical discussions but infrequently employed in practical performance modelling applications.

Over the years, numerous modifications and enhancements have been introduced to improve the original formulation of Fitts' Law (2.2), with the goal of increasing its empirical validity and robustness across a wider range of task conditions. Among these, one of the most widely adopted variants is the Shannon formulation [24], which revises the computation of the ID to address a key limitation in the original model. Specifically, Fitts' original logarithmic expression may yield negative ID values when the target distance ($d$) is less than half the target width ($w$), which can occur in small-scale or precision-oriented tasks. To circumvent this issue and ensure non-negativity, the Shannon formulation introduces an additive constant within the logarithmic term, yielding the following expression:

$$\mathrm{ID} = \log_2\left(\frac{d}{w} + 1\right). \tag{2.4}$$

This adjustment not only prevents undefined or negative ID values but also aligns more closely with observed user behaviour in empirical studies, particularly under high-precision or short-distance conditions. As a result, the Shannon formulation has gained widespread acceptance in both experimental psychology and human-

computer interaction research, where it is often preferred for its stability and consistency in modelling a broad spectrum of human-machine tasks. Its improved numerical behaviour makes it especially suitable for computer implementations in user interface evaluation and design.

However, both the Shannon formulation (2.4) and the original Fitts' Law are inherently constrained to one-dimensional movements and primarily focus on linear translations along a single axis. These limitations restrict their applicability in more complex, real-world scenarios where movement typically occurs in two or more dimensions and involves additional factors beyond linear displacement. To address these shortcomings, subsequent research has sought to generalise the model to accommodate 2D spatial configurations and compound movement patterns. Most of these adaptations retain the core structure of the Shannon formulation due to its improved empirical alignment and numerical stability.

One such extension incorporates not only the translational distance to the target position but also the rotational displacement required to align with the target orientation. This model, proposed by Stølen and Akin [25], augments the traditional difficulty index by adding a rotational term that captures the angular effort needed to reach and orient toward the target. Specifically, the model introduces rotational distance ($\theta$) and rotational tolerance ($\delta$) to quantify the difficulty associated with orientation alignment:

$$\mathrm{ID} = \log_2\left(\frac{d}{w} + 1\right) + \log_2\left(\frac{\theta}{\delta} + 1\right). \tag{2.5}$$

This formulation allows for a composite evaluation of task difficulty by summing the contributions of both translational and rotational components. It is particularly relevant in applications involving complex spatial manipulation, such as robotic arm control, virtual object manipulation, and user interface interac-

tions in 3D environments. By recognising and modelling the added cognitive and machinery demands of rotational adjustments, this approach provides a more comprehensive framework for analysing human performance in multidimensional tasks.

Since the majority of Fitts' Law-based models are designed for human control of a virtual cursor—typically idealized as a dimensionless point with no physical properties such as size or mass—they present inherent limitations when applied to modeling embodied systems. These simplifications are particularly problematic in the context of HMT, where both the human operator and the machinery system possess spatial extent and physical constraints. Representing a complex, embodied effector as a point target fails to account for interactions influenced by geometry, surface area, and physical clearance.

To address this issue, an innovative study by Cha and Myung [13] employed a more physically grounded representation by using a human finger as the cursor, thereby introducing a tangible spatial dimension into the model. This approach integrated the size of the finger ($f$) into the calculation of the Index of Difficulty, resulting in a revised formulation:

$$\mathrm{ID} = \log_2\left(\frac{2d}{w+f}\right). \tag{2.6}$$

By explicitly incorporating the end effector's (in this case, the human finger's) physical dimensions, this method offers a more realistic estimation of task difficulty in settings where spatial constraints and surface interaction are non-negligible. Such considerations are particularly pertinent in HMT tasks, where robot end-effectors or manipulators must operate in constrained or cluttered environments. This formulation serves as a conceptual bridge toward more comprehensive models that take into account not only movement distances and target

sizes but also the embodied characteristics of the agents involved.

Building on these insights, recent research has explored the integration of multiple such modifications to assess the performance of human-robot systems more holistically [3]. These efforts aim to develop a unified modelling framework capable of accurately reflecting the complex dynamics of shared autonomy tasks, where physical embodiment and cooperative interaction are essential dimensions of task difficulty.

Nevertheless, despite the advancements introduced by various formulations and extensions of Fitts' Law, the majority of these methodologies remain confined to a two-dimensional spatial framework. This constraint poses significant challenges when attempting to accurately model real-world HMT operations, which are inherently three-dimensional and often require precise coordination across both translational and rotational degrees of freedom. To date, only limited research has explored the extension of Fitts' Law into 3D environments [63] and its specific application within HMT contexts [27], where the nature of interaction extends beyond planar motion.

Recent studies have attempted to bridge this gap by introducing formulations that combine both translational and rotational distances in a unified difficulty index [64], and by incorporating target depth as an additional variable within the traditional Fitts' Law structure [65]. These efforts represent important steps toward capturing the spatial richness of natural human movement. Parallel developments in the field of human-computer interaction have applied Fitts' Law to evaluate performance in virtual reality (VR) environments, particularly focusing on 3D cursor control tasks [14], [66]. While these VR-based models offer valuable insights into human spatial behavior, they remain constrained by the nature of the cursor metaphor, which typically only involves translational control in 3D

space without accounting for orientation or torque—critical factors in robotic or HMT systems.

Thus, although these adaptations are effective in characterising HMT performance in immersive virtual interfaces, they fall short in representing the full complexity of HMT scenarios, where manipulators must not only reach a target but also align with specific orientations and constraints dictated by the environment or task. As a result, such models may underestimate the cognitive and physical demands inherent in collaborative manipulation and control tasks involving embodied robotic systems.

To address these shortcomings, the proposed model leverages the diversity of approaches derived from Fitts' Law and synthesizes them into a unified framework that accounts for both translational and rotational demands in a three-dimensional space. Specifically, this study introduces a 3D prediction model with six degrees of freedom (6-DoFs), purpose-built to represent the operational realities of HMT. This formulation is designed to more accurately reflect the multidimensional nature of task difficulty in practical human-robot collaboration scenarios, thereby offering a more comprehensive and predictive understanding of joint performance in real-world settings.

Despite the extensive body of work exploring and extending Fitts' Law in various contexts, there remains a conspicuous lack of research dedicated to developing a fully 3D predictive model tailored to HMT performance. Most existing models are either constrained to planar motions or exclude rotational dimensions, thereby limiting their applicability to embodied systems operating in complex spatial environments. This gap is particularly critical given the increasing prevalence of mobile manipulation systems that must navigate and interact within dynamic, unstructured settings.

A predictive framework grounded in a 3D formulation of Fitts' Law would offer significant benefits across both system development and operational deployment phases. From a design perspective, it could provide quantitative insights into task difficulty, enabling the optimisation of control strategies, interface design, and motion planning for shared autonomy. In real-world operations, such a model would be instrumental in forecasting performance constraints, informing task allocation decisions, and adapting robotic behaviour in response to human input and environmental variability.

This need is especially pronounced in mobile manipulation missions, where the robot must coordinate movement across 6-DoFs to accomplish tasks such as grasping, insertion, or collaborative object transport. In these scenarios, accurately modelling the interplay between translational reach, rotational alignment, and environmental constraints is essential to achieving seamless and effective HMT. Consequently, the development of a robust 3D Fitts' Law model tailored to these requirements constitutes a necessary advancement toward more intelligent, adaptive, and efficient human-robot collaboration in real-world applications.

## 2.2 Evaluation for Human-Machine Teaming Interfaces

Numerous scholars have undertaken empirical investigations into integrating human and robotic intelligence within human-machine collaboration, mainly focusing on the design and evaluation of Human-Machine Teaming Interfaces (HMTIs) [37]. Over the years, a wide range of control modalities—ranging from traditional joysticks and teleoperation systems to immersive virtual and motion-capture-based interfaces—have been developed and applied to robotic platforms in both research

and industry. These diverse HMTIs aim to enhance the fluency, adaptability, and effectiveness of human-robot cooperation.

However, identifying the most effective control strategy remains an ongoing challenge, particularly given the variability in task requirements, user expertise, and environmental complexity. This challenge is compounded by the lack of a unified framework for objectively quantifying interface performance. In response, a growing body of literature has highlighted the need for a standardised methodology to evaluate and compare the efficacy of different HMTIs [67]. Such a framework would not only facilitate the benchmarking of interface technologies but also guide interface selection and design based on empirical performance predictions.

Motivated by these observations, this research addresses this critical gap by proposing a coherent and generalizable approach to HMTI performance evaluation. Specifically, we employ the extended 3D formulation of Fitts' Law that captures the multidimensional nature of embodied interaction in HMT [1], along with subjective evaluation.

Evaluating the effectiveness of teleoperation interfaces remains a critical component in the development and refinement of human-robot teaming systems. A task-based evaluation framework for teleoperation is introduced in [68], where performance is quantified using task completion metrics. This framework incorporates measurable parameters based on the number of successful and unsuccessful task executions, offering a practical approach to assessing objective performance. In addition to quantitative data, user feedback is collected through structured questionnaires, enabling a parallel evaluation of subjective user experience and interface usability.

Building upon this foundation, more recent studies have presented broader and more integrated frameworks for assessing operator performance in robotic sce-

narios [34], [67]. These works emphasise the importance of combining objective performance indicators—such as task efficiency, error rates, and accuracy—with subjective measures, including perceived workload, ease of use, and user satisfaction. Such comprehensive evaluation methods are essential for capturing the multifaceted nature of teleoperation performance, particularly in tasks involving prolonged cognitive engagement or physical coordination.

Despite the progress made in establishing standardised evaluation methodologies, the majority of existing frameworks rely on movement models constrained to lower-dimensional task representations, often limited to one or two spatial dimensions. While such models are helpful for simplified experimental setups or interface benchmarking, they may prove inadequate for capturing the full complexity of real-world robotic missions, which typically involve 3D motion, orientation alignment, and dynamic interaction with the environment.

As robotic systems continue to expand into more complex domains—ranging from mobile manipulation in unstructured environments to collaborative industrial tasks—there is a growing need for evaluation frameworks that reflect these operational realities. Accurate and scalable evaluation models are essential not only for quantifying system performance but also for guiding interface design, task allocation, and operator training in practical applications of HMT.

#### Usability and Workload Evaluation

In the evaluation of HMTIs, subjective response measurements are commonly categorised into two principal dimensions: the system's perceived workload—both mental and physical—on the human operator, and the overall usability of the interface [69], [70]. These assessments provide valuable insights into the cognitive and ergonomic impact of different interface modalities, complementing objective

performance metrics.

Among the most widely adopted tools for measuring cognitive demand is the NASA Task Load Index (NASA-TLX) [71], which captures an operator's perceived workload across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. The NASA-TLX is typically implemented via a structured questionnaire and employs weighted averages to reflect the relative importance of each factor. Its comprehensive yet intuitive structure has made it particularly popular in applied engineering domains. Empirical studies have shown that NASA-TLX remains the preferred choice for evaluating cognitive load in real-world systems due to its reliability, ease of administration, and strong correlation with task complexity [72].

In addition to workload assessment, situational awareness is a critical aspect of teleoperation, especially in scenarios involving remote perception or video feedback. The NASA Situation Awareness Rating Technique (SART) [73] is specifically designed to measure an operator's perceived awareness of their environment. Unlike NASA-TLX, which focuses on task-related cognitive effort, SART emphasises the human agent's understanding of the external environment. It is particularly relevant in supervisory control tasks and operations involving indirect visual input. However, SART is less concerned with machinery manipulation itself and is therefore typically used in conjunction with other evaluation tools in the context of HMTIs.

Usability testing plays a crucial role in interface evaluation to further capture the subjective quality of interaction. The System Usability Scale (SUS) [74] has emerged as a standardised and widely accepted method for usability assessment. Comprising a ten-item Likert-scale questionnaire, the SUS is designed to be easily understood by users from diverse backgrounds and is applicable across a broad

range of domains. Its simplicity, quick deployment, and industry-wide acceptance have made it a favoured tool in HMTI research as well [75], [76], where it serves to gauge user satisfaction, ease of learning, and perceived system effectiveness.

Together, these subjective evaluation methods provide a multi-faceted understanding of HMTI performance, bridging the gap between technical capabilities and user experience.

## 2.3 Human-machine Interface Hardwares

In the context of HMT, two primary categories of Human-Agent (HA) to machine, in this case, Robotic-Agent (RA), interfaces have emerged. The first category involves the use of remote control devices, such as gamepads, joysticks, or keyboards, which enable HAs to issue commands to RAs through discrete or analog inputs [77]. These interfaces are typically characterised by their low latency, portability, and intuitive mapping of control schemes, making them widely adopted in various robotic applications, from aerial drone navigation to ground-based manipulation tasks.

The second category encompasses more immersive interaction modalities, wherein HAs control RAs through natural body movements captured using motion capture (MoCap) systems [78]. These interfaces translate the physical gestures or postures of the human operator into robotic actions, offering a more direct and embodied form of teleoperation. MoCap-based systems are particularly advantageous in complex manipulation or collaborative tasks, where the nuances of human motion—such as hand orientation, joint articulation, and spatial awareness—can be leveraged to achieve more precise and intuitive control.

In recent years, there has been a marked increase in research and development

efforts focused on the application of motion capture technologies within robotic teleoperation. Advances in wearable sensors, real-time kinematic tracking, and machine learning-based motion interpretation drive this trend. As a result, MoCap-enabled interfaces have gained traction for their potential to enhance task fluency, reduce operator workload, and improve the overall transparency of human-robot interaction. However, systematic methods for evaluating the performance benefits of such interfaces—particularly in comparison to traditional control modalities—remain underexplored, motivating the need for unified evaluation frameworks such as the one proposed in this work.

### 2.3.1 Simulation Technologies

Simulation environments play a central role in developing and evaluating HMT systems. They provide a safe and cost-effective platform for rapid prototyping, task benchmarking, and operator studies before real-world deployment. Modern simulators combine physics engines with robot middleware to support realistic interaction, data generation, and human-in-the-loop testing.

Physics-based simulators such as Gazebo/Ignition, Webots, MuJoCo [79], and Isaac Sim (PhysX) model robot dynamics, collisions, and sensing with sufficient fidelity for both manipulation and navigation. Platforms such as AI2-THOR [80], Habitat, RLBench, and BEHAVIOR-1K [81] further provide standardised 3D tasks and reproducible evaluation scenarios. These environments are often integrated with ROS, enabling consistent interfaces across simulation and hardware.

Simulation in this study was conducted through an interactive web-based platform built using the WebGL Application Programming Interface (API). This approach allows the entire simulation to run locally in a web browser, without the need for plug-ins or additional software installation. After the initial load-

ing, all computation is executed on the participant's machine, ensuring that task execution is unaffected by internet latency; connectivity is only required at the beginning and for submitting final results.

A key strength of WebGL is its direct access to the user's GPU through HTML5, enabling efficient use of available hardware for rendering and control. This design supports high-fidelity visualisation and responsive interaction while adapting dynamically to the computational capacity of different devices. As a result, the platform offers a consistent and scalable environment across a diverse participant pool.

Participants controlled the excavator using a standard keyboard interface, with specific keys mapped to both locomotion and manipulation functions. This scheme balances accessibility with sufficient complexity to capture realistic task demands. Overall, the WebGL-based system provides a robust and reproducible environment for evaluating teleoperation performance, supporting the investigation of task difficulty and user variability under controlled yet realistic conditions.

### 2.3.2 Gamepad Technologies

Gamepad controllers have become one of the most prevalent and standardised input devices for remotely operating RAs, mainly due to their ergonomic design, widespread availability, and ease of integration with robotic systems. Their intuitive layout—typically composed of analogue sticks, directional pads, and buttons—provides a versatile platform for issuing both discrete and continuous commands, making them especially suitable for tasks that require simultaneous control of multiple degrees of freedom.

As a result, gamepads are extensively used across a wide array of robotic applications. In the field of healthcare robotics, for instance, researchers have employed

gamepad-based control for nursing and assistive robots, enabling operators to perform tasks such as guiding patients, manoeuvring mobile robots in constrained environments, and assisting with activities of daily living [82], [83]. Their compactness and responsiveness make them particularly advantageous in clinical settings, where space and time efficiency are critical.

Beyond practical applications, gamepads have also served as a baseline control interface in numerous empirical studies examining human-robot interaction. These studies often explore how gamepad-based teleoperation compares with alternative modalities in terms of performance, usability, and user preference. For example, Zhao et al. [84] conducted a comparative analysis between gamepad control and hand gesture-based interfaces, evaluating factors such as response time, accuracy, and user cognitive load. Similarly, Oshita [85] examined the trade-offs between gamepad and touchscreen controls, offering insight into interface selection for different types of robotic tasks.

Commercial quadruped robots frequently incorporate gamepads as their primary control method, given the need for simultaneous navigation and posture adjustment. The dual analog sticks, particularly, are well-suited for controlling locomotion while allocating other buttons for limb or gripper articulation. This compatibility with complex locomotion and manipulation tasks has cemented the gamepad's role as a standard interface in many commercial robotic platforms.

### 2.3.3 Motion Capture Technologies

In addition to traditional gamepad controllers, MoCap systems have become an increasingly prominent teleoperation modality within HMTIs, offering an immersive and intuitive means of control. These systems utilise real-time human motion tracking to generate corresponding control signals for robotic agents, enabling a

more embodied and natural interaction paradigm. Depending on the specific implementation, motion capture input can be derived from visual sensors such as RGB and RGB-D cameras [86], or wearable devices equipped with IMUs [87] that track body segment orientation and acceleration.

Contemporary MoCap technologies incorporate a diverse array of sensing approaches, including optical, inertial, mechanical, magnetic, and acoustic systems. Optical systems often employ multiple high-speed cameras and reflective markers to reconstruct full-body kinematics with high spatial accuracy. In contrast, inertial-based systems rely on IMUs attached to key body segments to estimate joint angles and motion trajectories without needing external infrastructure. Mechanical and magnetic systems provide alternative tracking solutions using linkages or field sensors, while acoustic tracking uses time-of-flight measurements to infer position.

Many MoCap systems are integrated with programming by demonstration frameworks to enhance their applicability and adaptability in robotics. These include techniques such as keyframing—where representative postures are selected and interpolated over time—and clustering, which groups motion trajectories to identify reusable motion primitives [88]. Such methods allow for the efficient encoding of complex human-machine behaviours, reducing the need for low-level programming and enabling operators to specify tasks through natural movement patterns.

These technologies have been widely explored for their ability to replicate complex manipulative and navigational actions, often with greater fluidity and responsiveness than traditional input devices. Their ability to track full-body motion makes them particularly effective in applications that require whole-arm or whole-body teleoperation, such as mobile manipulation, object handover, and collaborative motions.

### Vision-based Motion Capture Technologies

Recent advancements in vision-based technologies have prompted growing interest in utilising camera images as input for MoCap systems, particularly due to their non-intrusive nature and ease of deployment. Vision-based MoCap systems rely on visual sensors—often in the form of RGB cameras, depth sensors, or RGB-D devices—to estimate human body pose and movement by analysing image sequences. These systems typically employ computer vision algorithms and deep learning models to detect key body landmarks, reconstruct skeletal structures, and track dynamic motions in real time.

A representative example of this approach is presented in [89], where a camera-based motion capture method is implemented using the Microsoft Kinect V2 sensor. This system leverages the Kinect's built-in depth sensing capabilities to perform human-body motion analysis, enabling full-body pose estimation without the need for physical markers or wearable devices. The study outlines the application of the Kinect sensor in teleoperation scenarios and demonstrates its ability to capture joint positions and movements with reasonable fidelity and responsiveness.

Overall, vision-based MoCap systems like the one described in [89] demonstrate the feasibility of using camera input for intuitive human-robot interaction. They offer promising avenues for achieving hands-free, naturalistic control of robotic agents, especially in applications where freedom of movement and ease of use are prioritised.

### Wearable Motion Capture Technologies

Wearable motion capture technologies offer enhanced stability and robustness compared to vision-based motion capture systems, particularly in dynamic or

unstructured environments. These systems are less susceptible to external visual disturbances such as varying lighting conditions, background clutter, or occlusion by other objects or users. As a result, wearable systems are often favoured in scenarios that require reliable tracking performance across diverse operating contexts.

Wearable motion capture systems typically consist of suits or modules embedded with IMUs, which track the orientation, velocity, and acceleration of individual body segments. By aggregating data from multiple IMUs placed across the body, these systems can reconstruct full-body skeletal motion in real time. Such configurations have enabled the accurate and responsive mapping of human motion to RAs, facilitating intuitive and high-fidelity teleoperation [90], [91]. These systems are particularly effective in tasks involving whole-body manipulation or coordinated locomotion, where joint-level correspondence between human and robot is crucial.

In addition to motion tracking, wearable systems also support higher-level interaction paradigms by integrating workspace mapping and path planning features. For instance, virtual obstacles can be configured within the robot's operating environment to impose motion constraints, thereby enhancing safety and improving user experience [92]. These constraints help guide the RA's actions within acceptable boundaries, reducing the risk of collisions and allowing the operator to focus on task-level objectives rather than low-level motion details.

Through the combination of stable motion capture and intelligent interaction design, wearable systems offer a compelling solution for intuitive human-robot collaboration, particularly in applications requiring fine-grained control and situational adaptability.

## 2.4 Prediction and Evaluation with Human Fatigue

HMT has become a central focus in robotics research, particularly for applications where full autonomy is currently infeasible or undesirable. In many real-world missions, the complexity, unpredictability, and dynamic nature of the environment necessitate effective cooperation between HAs and RAs. As robotic systems and the missions become increasingly complex, the workload on the HAs significantly increases and impacts the HMT performance, especially in longer-period missions. Therefore, the need for systematic evaluation and modelling of HMT performance becomes critical when considering the fatigue of HAs.

Such evaluation frameworks must consider not only the capabilities of the robotic system and the quality of the control interface but also the cognitive and physical states of the human operator. In particular, factors such as task difficulty, human workload, and fatigue can substantially influence overall system performance.

This section reviews existing research related to the evaluation of human-robot collaboration, with a particular focus on performance prediction with fatigue modelling. This study first examines evaluation methods used to measure HMT performance, emphasising task-based frameworks and workload assessment techniques. This study then reviews approaches to modelling human fatigue, distinguishing between measurement-based methods and bio-mathematical models, and discusses their relevance to predicting human performance in teleoperation scenarios.

To this end, research has diverged into several streams focusing on different aspects of the collaborative system [93], [94]. Some studies emphasise the evaluation of the robot itself, analysing aspects such as mechanical design, autonomy capa-

bilities, and task execution efficiency [95]. Others focus on the development and assessment of HRIs, aiming to improve the quality of interaction, reduce operator workload, and enhance task fluency [96], [97]. In parallel, an important line of inquiry has emerged that seeks to incorporate human factors into the overall system performance evaluation. Recognising that human operators are integral components of collaborative systems, these studies examine cognitive load, situational awareness, mental workload, and fatigue as critical determinants of overall HMT performance [34], [35], [68], [98].

By integrating these perspectives, researchers aim to develop comprehensive frameworks that not only assess the technical proficiency of the robotic systems but also account for the human-centred dimensions that fundamentally shape collaborative effectiveness in complex environments.

Beyond task outcome analysis, other evaluation methods place a stronger emphasis on understanding the cognitive and physical workload experienced by participants during robotic teleoperation. Tools such as the NASA-TLX and other customised workload assessment scales have been employed for this purpose [34], [35]. These instruments measure multiple dimensions of perceived workload—including mental demand, physical demand, temporal demand, effort, and frustration—providing a comprehensive and multi-faceted view of the operator's experience. Insights gained from workload evaluations are instrumental in informing system design improvements, with the goal of minimising operator strain and enhancing overall system usability and performance.

While workload is widely recognised as a key factor influencing system performance, particularly through its contribution to human error and reduced operational efficiency, relatively few studies have extended their analysis to examine the impact of workload-induced cognitive fatigue over time. Cognitive fatigue

can significantly impair an operator's ability to sustain high levels of attention, decision-making quality, and teleoperation control, especially during prolonged or demanding missions. Despite its importance, the relationship between accumulated workload, mental fatigue, and performance degradation in robot teleoperation contexts remains underexplored, highlighting a critical area for future investigation.

Accurate modelling of task difficulty is a fundamental prerequisite for predicting system performance during the execution of specific tasks. Task difficulty directly influences the level of cognitive and physical demand placed on a human operator, and in turn, significantly affects the rate of fatigue accumulation over the course of task execution [99]–[103]. As cognitive and physical resources are progressively depleted under high-demand conditions, an operator's ability to maintain optimal performance deteriorates, underscoring the critical need for predictive models that can effectively quantify task difficulty.

Consequently, the development of precise and reliable task difficulty models is essential for two interconnected purposes: task performance modelling and fatigue modelling. From a performance standpoint, task difficulty serves as a predictor of execution time, error likelihood, and control fluency. From a fatigue perspective, task difficulty modulates the cognitive load that contributes to fatigue onset and accumulation, thereby impacting long-term operational safety and efficiency.

Within the field of psychology, Fitts' Law [12] stands as one of the most influential tools for modelling human-computer system performance. As stated in the previous section, it describes the relationship between task difficulty and the time required for targeted movements. Building upon this body of work, the previous study proposed a refined task difficulty model specifically designed to better capture the unique demands of HMT in 3D. This refinement considers the additional

degrees of freedom, environmental complexity, and embodiment characteristics typical of teleoperated robotic systems. Nevertheless, to the best of our knowledge, no existing research has yet leveraged such extended task difficulty models to systematically generate task demand profiles and predict cognitive fatigue accumulation in human operators during robot teleoperation. Addressing this gap could offer a powerful framework for anticipating operator state degradation and optimising human-robot collaboration in complex, real-world missions.

### 2.4.1 Modelling Human Fatigue

Fatigue is a critical factor that substantially influences human performance in HMT, impacting both task efficiency and operational safety. In collaborative robotics, fatigue can manifest in multiple forms, each exerting distinct effects depending on the nature of the interaction between the HA and the RA.

Physical fatigue primarily affects scenarios involving human-robot co-manipulation, where the human operator is required to exert sustained physical effort to coordinate or assist the robot's movements [104]. Prolonged physical exertion can lead to muscular fatigue, reduced precision, slower reaction times, and increased risk of injury, ultimately degrading the effectiveness and safety of collaborative operations.

Conversely, cognitive fatigue plays a predominant role in human-robot teleoperation tasks, where the human operator must maintain high levels of concentration, decision-making, and perceptual processing over extended periods [105], [106]. Cognitive fatigue can impair attention, situational awareness, and problem-solving abilities, leading to delayed responses, increased error rates, and diminished overall system performance. Given that many teleoperation tasks require continuous monitoring of robot status, dynamic environments, and task objec-

tives, understanding and mitigating cognitive fatigue becomes essential for maintaining mission success.

Accurately modelling both physical and cognitive fatigue is thus crucial for developing predictive frameworks that can enhance the design of HRC systems, inform workload management strategies, and support real-time adaptation to operator state fluctuations. The following subsections review existing approaches for measuring and modelling human fatigue, distinguishing between measurement-based methods and bio-mathematical models.

### Measurement-Based Approaches

Cognitive fatigue can be influenced by a multitude of interacting factors, including task demands, environmental conditions, and individual physiological states. To capture the dynamic and multifaceted nature of fatigue, some research efforts have focused on continuous monitoring approaches that estimate fatigue levels by observing human behaviour and environmental variables in real time [107]–[111].

One notable study [107] proposed a comprehensive model of human fatigue that integrates a wide array of measured elements, such as ambient temperature, environmental noise, circadian rhythm variations, and other contextual factors. This model not only considers instantaneous indicators of operator state but also captures the accumulative property of fatigue as it builds up over time during prolonged task execution. To model this temporal evolution of fatigue, the study introduced a dynamic fatigue detection framework based on Dynamic Bayesian Networks, which can probabilistically infer the fatigue state of an individual by integrating multiple streams of sensory data over time.

Building on this foundational work, subsequent studies have developed additional

dynamic fatigue detection models tailored for applications such as driver monitoring. These models leverage various machine learning techniques to track and predict changes in cognitive state. For example, fatigue detection approaches based on Hidden Markov Models [108] and Dynamic Bayesian Networks [109] have been proposed, capturing sequential dependencies in behavioural data. More recently, deep learning-based methods have been explored [110], utilising neural networks to model complex, nonlinear patterns in physiological and behavioural signals associated with fatigue.

While these models offer comprehensive and sophisticated mechanisms for fatigue estimation, a common limitation is their heavy reliance on large volumes of multimodal sensor data, including physiological measurements (e.g., heart rate variability, skin conductance), environmental readings, and behavioural observations (e.g., gaze patterns, body posture). The need for continuous, high-fidelity data collection can impose significant practical challenges, particularly in real-world mission scenarios where extensive sensor deployment may be impractical, costly, or intrusive. This constraint highlights the necessity of exploring more scalable and lightweight approaches to fatigue modelling, especially for operational environments involving human-robot collaboration.

#### Bio-mathematical Model

Bio-mathematical modelling approaches offer an alternative pathway for estimating human cognitive performance and fatigue levels based on structured behavioural and environmental schedules [112]. Rather than attempting to model the complex neurophysiological mechanisms underlying brain function—which remains a significant scientific challenge due to individual variability and the intricate nature of cognitive processes—these models are typically derived from large-scale experimental data collected across diverse volunteer populations. By

abstracting individual differences, bio-mathematical models provide generalised, population-level predictions of performance fluctuations over time.

One of the most prominent examples of this approach is the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model and its operational implementation in the Fatigue Avoidance Scheduling Tool (FAST) [36], [113]. The SAFTE model integrates sleep/wake histories, circadian rhythms, and homeostatic sleep drive mechanisms to predict variations in human cognitive performance. It captures the effects of sleep deprivation, time of day, and accumulated fatigue to generate a continuous estimate of operator effectiveness. The FAST software, widely used in military and aerospace applications, leverages these predictions to design schedules that minimise fatigue-related risks.

Further refinements to the SAFTE model have been proposed to enhance its realism and applicability across different operational contexts. Studies have incorporated additional factors such as task demand intensity, exposure to light countermeasures (to modulate circadian rhythms), pharmaceutical interventions, and the disruptive effects of night shifts and jet lag [114]–[116]. These enhancements recognise that external interventions and environmental conditions can significantly alter the trajectory of cognitive performance, and thus must be considered when modelling fatigue for mission-critical operations.

As robotic systems continue to expand into sectors such as manufacturing, logistics, healthcare, and defence, the issue of human operators managing robots while under the influence of fatigue has emerged as a growing concern. Fatigue can impair situational awareness, decision-making, and control precision, undermining the safety and effectiveness of human-robot teams. Despite its critical importance, relatively little research has incorporated fatigue modelling into the broader context of human-robot collaboration [106]. Combining human-robot

teleoperation models with well-established bio-mathematical fatigue models, such as SAFTE, holds considerable promise for predicting human performance under varying states of cognitive fatigue. Such integration could enable dynamic workload management, mission planning adjustments, and real-time adaptations to ensure sustained operational effectiveness in fatigue-prone environments.

## 2.5 Multi-Robot Control with LLMs

Multi-robot task allocation (MRTA) is a critical problem in the field of robotics, concerned with the efficient and reliable assignment of tasks to a team of robotic agents. The primary objective of MRTA is to optimise mission outcomes, ensuring that tasks are completed within specified constraints of time, cost, and resource utilisation. Traditional MRTA approaches typically formulate the problem as an optimisation or matching problem, relying on predefined cost metrics, capability matrices, and heuristic or algorithmic methods to assign tasks to robots based on their suitability [117], [118]. These classical frameworks are often grounded in mathematical programming, graph theory, or auction-based mechanisms and have demonstrated considerable success in structured and predictable environments.

However, classical approaches can face substantial limitations when applied to more dynamic, uncertain, and unstructured operational contexts. In real-world missions, the task environment may undergo frequent and unpredictable changes, task descriptions may be provided at a high level or expressed in natural language, and complex interdependencies among tasks may require flexible reasoning rather than rigid optimisation. Traditional models, which depend heavily on predefined cost functions and static assumptions about the environment and agent capabilities, can struggle to adapt to these evolving and complex mission requirements.

In response to these challenges, Large Language Models (LLMs) have recently

emerged as promising decision-making agents in robotic systems. Leveraging their advanced language understanding, reasoning capabilities, and generative flexibility, LLMs offer a fundamentally different approach to MRTA. Researchers are increasingly exploring how LLMs can enhance task allocation by interpreting high-level mission objectives, reasoning about the relationships between tasks and robot capabilities, and dynamically adapting allocation strategies to optimise performance metrics such as task completion time, success rate, and resource usage [119].

By integrating LLMs into the task allocation process, robotic teams can achieve a higher degree of autonomy and adaptability. LLMs are capable of parsing complex instructions, inferring implicit task requirements, and proposing allocation plans that are contextually aware and sensitive to changing operational conditions. This new paradigm opens exciting opportunities for more intelligent, flexible, and human-aligned multi-robot collaboration, especially in domains where mission specifications cannot be fully predefined or where human-robot communication needs to occur in naturalistic, high-level language forms.

### 2.5.1 Traditional Task Allocation Methods

A foundational approach in MRTA involves matching each task's specific requirements with the skills, capabilities, or resource profiles of the available robotic agents. This capability-task matching ensures that the selected robots are appropriately suited to accomplish their assigned missions efficiently and reliably. In many systems, humans can play an active role in the allocation process. Tasks may be assigned by a human operator acting as a mission supervisor [120], or humans may themselves be treated as agents within the team, receiving and executing tasks alongside robotic counterparts [121]. Incorporating humans into the

agent pool introduces additional complexity, such as modelling human cognitive load, task preferences, and dynamic availability.

Classical automatic task allocation algorithms have been widely developed to automate the assignment process without requiring manual intervention. These methods typically frame the task allocation problem as an optimisation problem, seeking to optimise a global objective function such as minimising the total mission execution time, minimising energy consumption, or maximising the number of tasks successfully completed. Among the most commonly used techniques are the Hungarian method for solving assignment problems with polynomial-time complexity and linear programming formulations that allow the incorporation of complex constraints and cost structures [122].

These traditional approaches have demonstrated strong effectiveness in structured environments where task definitions, agent capabilities, and environmental conditions are known a priori and remain relatively stable over time. They provide mathematically rigorous solutions with guarantees of optimality or bounded sub-optimality under well-defined assumptions. However, their reliance on static cost models and limited reasoning about unstructured or dynamic elements can restrict their applicability in open, real-world mission scenarios where tasks may evolve, interdependencies emerge, or new information becomes available during execution.

### 2.5.2 LLMs in Robot Task Planning

Recent advances in LLMs have sparked growing interest in their integration into robotic systems, particularly for enhancing HMT. Kim et al. [123] conducted a user study to systematically investigate the capabilities and limitations of LLM-powered robots in real-world HMT scenarios. Their findings highlight that LLM-

enabled robots exhibit notable strengths in tasks that involve building social connections, engaging in deliberative dialogues, and providing empathetic or contextually sensitive responses. These qualities suggest that LLMs can significantly enhance the relational and communicative dimensions of human-robot collaboration, fostering trust and rapport between users and robotic agents.

However, the study also identified several critical challenges associated with deploying LLMs in robotic systems. Specifically, LLM-powered robots demonstrated difficulties in maintaining logically coherent communication over extended interactions, occasionally producing inconsistent or factually incorrect statements. Furthermore, the unpredictability and lack of full transparency in LLM-generated responses were found to induce anxiety and uncertainty in users, particularly in safety-critical or decision-making contexts. These findings underscore the importance of developing methods to improve the logical reasoning capabilities, reliability, and explainability of LLM-driven robotic systems.

### Action Planning from Language

The integration of LLMs with robotic systems for task planning through action composition has been an active area of exploration in recent studies. Action composition refers to the process of translating high-level language instructions into structured sequences of executable actions that a robotic system can perform, bridging the gap between natural language understanding and low-level robotic control.

Zeng et al. [51] introduced Socratic Models, a framework designed to enhance multimodal capabilities by enabling different specialised models (e.g., vision models, language models) to communicate and reason collaboratively without the need for fine-tuning. This modular interaction between models allows LLMs to contribute

to complex action planning tasks while leveraging the strengths of domain-specific perception modules.

Li et al. [52] proposed the Chain of Code method, which extends the Chain of Thought prompting paradigm by incorporating executable code generation. This approach improves reasoning reliability by enabling LLMs to reason through intermediate code-based steps. It supports complex task decomposition, making it particularly suitable for robotic applications requiring precise logical sequencing.

In another study, Kwon et al. [53] demonstrated the direct application of LLMs to robotic control by predicting robot end-effector poses from visual inputs and task descriptions. Their results showcase the potential of LLMs to serve as high-level policy generators, translating multimodal observations into actionable control outputs without extensive task-specific retraining.

Furthermore, Silver et al. [54] explored the use of closed-source models such as GPT-4 for task planning within Planning Domain Definition Language domains. By generating Python programs that map to formal PDDL specifications, LLMs are positioned to automate classical planning tasks—a domain with a long-standing history in artificial intelligence for solving complex, multi-step problems.

Collectively, these studies highlight the expanding role of LLMs in enabling robots to autonomously interpret, plan, and execute complex tasks derived from natural language instructions, moving toward greater autonomy and adaptability in real-world environments.

Song et al. [55] introduced the LLM-Planner, a framework designed for embodied agents focusing on few-shot grounded planning. Their approach leverages the capabilities of Large Language Models to perform high-level task decomposition and action sequencing with minimal task-specific fine-tuning. A key emphasis of

their work is the importance of dynamic planning adaptability, enabling agents to respond flexibly to complex and evolving environments. LLM-Planner builds upon the foundations of traditional symbolic planning methodologies [59], yet it introduces a novel integration of natural language processing techniques, effectively bridging the gap between symbolic task representations and the dynamic, ambiguous nature of real-world interactions.

In parallel, Wang et al. [56] developed LaMI, a system designed to enhance multi-modal human-robot interaction by incorporating LLMs into the planning and control loop. LaMI integrates high-level linguistic guidance with atomic action primitives and multi-modal sensory inputs, enabling robots to interpret rich human instructions and regulate their behavior accordingly. By combining language understanding, visual perception, and action execution within a unified framework, LaMI demonstrates the potential of LLMs to serve as central reasoning engines for more natural, intuitive, and adaptive human-robot collaboration.

Together, these studies highlight the expanding frontier of LLM-integrated robotic systems, illustrating how the fusion of natural language processing, symbolic reasoning, and multi-modal interaction can enable more intelligent, flexible, and human-centric robot behaviors.

Izzo et al. [57] and Yang et al. [58] adopt a structured approach to integrating natural language processing with robotic systems by translating natural language instructions into formalized control architectures, such as behavior trees and state machines. In these frameworks, natural language inputs are parsed and mapped onto predefined structural templates that guide robotic behavior in a fixed, sequential manner. While such approaches benefit from the formal verifiability and robustness of symbolic systems, they also inherently impose rigidity on the generated plans, limiting flexibility and adaptability to unforeseen changes in the

environment.

Most existing robotic systems operate primarily in an open-loop configuration, executing planned actions without the capability for autonomous error detection or recovery. This limitation constrains their robustness and adaptability in real-world, dynamic environments. To address this, several recent studies have introduced frameworks that incorporate human feedback into the robot learning and task execution process.

Shi et al. [124] proposed a proactive framework in which robots anticipate potential failures during task execution and autonomously request human assistance to refine their strategies. By incorporating anticipatory mechanisms, the system reduces failure rates and enhances mission success by leveraging human expertise at critical decision points. Similarly, Han et al. [125] developed a system that allows robots to explain their planned actions to users and receive corrective feedback. This interpretability improves transparency, fosters greater trust in the robot's decision-making process, and empowers users to intervene effectively when necessary.

Singh et al. [126] introduced an interactive prompting system that utilizes structured feedback to correct robot behaviors, particularly targeting educational and technical training environments. Their system demonstrates how structured user input can efficiently guide robot behavior refinement without requiring deep technical knowledge from users. Additionally, Liu et al. [127] presented Operation-relabeled Learning with Language Feedback (OLAF), a system that enables robots to update their visuomotor neural policies based on natural language corrections. OLAF allows verbal feedback to relabel failed experiences, thereby helping robots learn to avoid repeating mistakes through interactive human guidance.

Building upon these ideas, the framework advances the state of the art by enabling real-time, closed-loop interaction between the user and the robotic system during task execution. The approach allows users to naturally interact with robots between task steps, offering corrections, adjustments, or new instructions without requiring prior coding expertise. This seamless integration of real-time feedback significantly enhances the flexibility, accessibility, and effectiveness of human-robot collaboration, particularly in unstructured or rapidly changing environments.

**Utilizing open-source models**

The use of open-source models in robotics research plays a crucial role in promoting accessibility, transparency, and reproducibility—factors that are essential for the collective advancement of the field. Open-source frameworks enable researchers and developers to replicate experimental studies, verify reported results, and build upon prior work without prohibitive barriers to entry. This openness not only enhances the credibility of technological innovations but also fosters faster dissemination and collaborative development across the global research community.

Several recent studies have effectively leveraged open-source language and multimodal models to drive advancements in robotic capabilities. For instance, Mu et al. [128] and Huang [129] utilise a range of open-source LLMs to power their embodied AI systems, demonstrating that competitive performance in robotic task planning and execution can be achieved without reliance on closed, proprietary models. Their approaches underscore the potential for democratizing access to powerful AI tools and enabling broader participation in cutting-edge robotics research.

Similarly, Pueyo et al. [130] explore using the multimodal CLIP model, originally developed for vision-language tasks, to unlock new functionalities in robotic control. By repurposing the visual-semantic embedding capabilities of CLIP, their work showcases how multimodal models can serve as bridges between perception and action, enabling more flexible and intuitive robotic behaviours.

Collectively, these efforts illustrate the growing impact of open-source models in advancing embodied intelligence, highlighting how publicly available AI resources can catalyse innovation, enhance reproducibility, and lower the barrier to entry for robotics research and development.

The approach builds on these foundational developments by exclusively employing open-source models, thereby ensuring that the methodologies remain transparent, accessible, and reproducible within the broader robotics research community. This commitment to open-source solutions not only promotes stable development environments but also facilitates the verification and extension of the results by other researchers, contributing to the collaborative advancement of the field.

In addition to leveraging the accessibility benefits of open-source software, the framework integrates real-world feedback mechanisms into the task execution loop. This integration significantly enhances the system's dynamic control capabilities, enabling robots to adapt their behaviour in response to environmental changes and user interactions in real time. By addressing critical limitations observed in prior studies—such as the lack of empirical validation on actual robotic platforms, as noted in the work of Cao and Lee [131]—the approach bridges the gap between theoretical planning frameworks and practical deployment challenges.

Thus, the work not only capitalises on the foundational strengths of open-source models to ensure reproducibility and foster community engagement but also ex-

tends their application to dynamic, feedback-driven control scenarios. In doing so, this study pushes the boundaries of what open-source, LLM-integrated robotic systems can achieve in complex, real-world environments.

#### Experiment on real robots wth LLM

Experimentation on real robotic platforms represents a crucial step in robotics research, as it enables the validation, refinement, and practical assessment of theoretical models and control strategies. Testing in real-world environments exposes robotic systems to unpredictable dynamics, sensor noise, hardware limitations, and environmental uncertainties that cannot be fully captured in simulation. As a result, real-robot experiments provide indispensable insights into system robustness, adaptability, and practical feasibility.

Several recent studies have demonstrated the implementation of LLM-based methodologies on real robotic systems. For instance, Tanneberg et al. [132], Kwon et al. [53], and Chu et al. [133] have showcased the successful application of LLMs to real-world tasks, highlighting the potential of language models to facilitate robot planning and action generation.

Moreover, in contrast to simulation-only studies such as Pueyo et al. [130], the framework emphasises extensive and diverse real-world testing. This study validates system performance to ensure not only theoretical soundness but also operational reliability, scalability, and practical readiness.

### 2.5.3 Multi-Robot Task Allocation with LLMs

In the context of MRTA, researchers have begun to explore the potential of LLMs as high-level decision-makers capable of interpreting complex mission objectives and facilitating task distribution among robotic teams. One promising line of

research leverages LLMs to parse complex, often unstructured tasks described in natural language into structured sub-tasks that can be systematically assigned to individual robots [134]. This capability is particularly valuable in dynamic or uncertain environments where explicit pre-programming of tasks is impractical or where human operators issue high-level, goal-oriented commands that require further decomposition.

LLMs have demonstrated remarkable proficiency in breaking down abstract or high-level mission descriptions into actionable steps and generating structured plans that align with robotic capabilities. Beyond language understanding, recent studies have shown that LLMs can also reason directly over raw observational data, such as positional or state information, to support task allocation and decision-making [135]. By combining linguistic reasoning with perception-based situational awareness, LLMs offer a novel approach to bridging the gap between human intent and robotic execution.

However, despite their impressive generalisation and reasoning abilities, LLMs face notable challenges when applied to MRTA scenarios involving complex spatial relationships. Understanding and accurately interpreting relative positions, orientations, and kinematic constraints between robots and targets remains a difficult problem. Traditional language models, primarily trained on textual data, may lack the necessary inductive biases to fully capture geometric or spatial dependencies intrinsic to multi-robot coordination. As a result, while LLMs can generate reasonably high-level plans, their effectiveness in precise spatial reasoning and fine-grained task optimisation often requires supplementary mechanisms or integration with specialised planning modules.

One of the recent works [136] developed a prompt-based system to assign tasks across a heterogeneous team of robots, including a robotic manipulator, a mobile

manipulator, and an aerial drone. Their system, built upon the BEHAVIOR-1K benchmark [81], utilises feedback mechanisms to iteratively refine task planning and correct errors, ultimately achieving higher success rates compared to baseline methods. Performance is evaluated in terms of the number of steps required to complete assigned tasks, providing a quantitative measure of planning efficiency and effectiveness.

In parallel, recent studies have introduced formal mathematical problem formulations and heuristic prompt optimisation techniques specifically targeting multi-step task planning scenarios [137]. These approaches aim to improve the coherence and reliability of LLM-generated plans over sequences of actions, addressing the inherent difficulty of maintaining consistency across multi-stage missions.

Another line of work proposes the use of conformal prediction techniques to enhance completion guarantees in language-guided robotic planning. Within simulated environments such as AI2-THOR [80], conformal prediction frameworks have been applied to quantify uncertainty and ensure safer execution of planned behaviors [138]. These efforts contribute to building more trustworthy LLM-integrated systems, particularly in settings where task execution risks must be carefully managed.

Beyond safety and step-level optimisation, further research explores different aspects of task planning and execution for multi-robot systems. Topics include few-shot prompting strategies for adapting LLMs to new tasks with minimal additional data [134], [139], as well as architectural considerations regarding centralised versus decentralised planning frameworks. Another significant contribution discusses task-and-motion planning, emphasising the potential of few-shot natural language translation into structured task representations that facilitate seamless integration with motion planning pipelines [140].

However, it is important to note that these recent methodologies predominantly focus on heterogeneous multi-robot task allocation, optimising the sequencing and assignment of distinct capabilities across robots with different hardware configurations. They emphasise step order optimisation and inter-task reasoning but pay relatively less attention to scenarios involving homogeneous robot teams, where multiple robots possess identical or highly similar skill sets.

In contrast, the proposed method addresses this gap by targeting homogeneous multi-robot task allocation within each planning step. This approach is designed to complement and extend existing LLM-based frameworks by providing a mechanism for efficiently distributing tasks among identical or similar robots, thereby broadening the applicability, scalability, and efficiency of task allocation strategies in larger, more uniform multi-robot teams.

# Chapter 3

# Modelling task difficulty

## 3.1 Process Overview

The section below provides a comprehensive, step-by-step overview of the algorithmic process involved in implementing the proposed 3D Fitts' Law model for predicting Human-Machine Teamwork (HMT) performance, based on our publication [1]. The process is designed to systematically guide the development and application of this difficulty model, ensuring that all relevant factors influencing task performance are considered and appropriately accounted for. This process can be divided into several key stages, each of which contributes to the creation of a robust predictive model capable of accurately forecasting the performance of HMT systems across various tasks and configurations. The following stages outline the procedure in a logical sequence:

1. **Motion Capability Identification:**

    - The first step in the process involves identifying the motion capabilities of the system. This is a crucial task as it establishes the fundamental building blocks for the model. The motion capabilities refer to the

various ways in which the HMT system can manipulate or interact with the environment, such as locomotion movements from the robot base and manipulation movements from the robot arms and tools. These capabilities need to be clearly defined to ensure that the model is representative of the system's actual abilities and constraints.

- Once the motion capabilities are identified, the next task is to determine the minimum set of standard tasks required. Standard tasks act as the foundation of the model, allowing for the decomposition of complex tasks into simpler, more manageable components. These tasks are defined in terms of their relevant parameters, such as distance, angle, and target size, which are then used to evaluate the difficulty of a task in terms of the 3D Fitts' Law. The selection of the minimum set of standard tasks is guided by the principle of ensuring model comprehensiveness, as detailed in Section 3.1.1. These tasks are chosen to cover the range of potential motion capabilities of the system, enabling the model to predict performance across a broad spectrum of possible tasks.

2. **Standard Task Modelling:**

   - In this stage, the focus shifts to the modelling of the standard tasks identified in the previous step. To build a robust predictive model, it is essential to clearly define the parameters for each standard task. These parameters include, but are not limited to, the linear distances the system must travel, the angles through which the system must rotate, and the sizes of the targets the system must interact with. Each of these parameters plays a vital role in determining the task's difficulty and directly impacts the calculation of the task's Index of

Difficulty (ID). A thorough and precise definition of these parameters ensures that the model will be able to accurately represent real-world task requirements and variations. The standard tasks are treated as exemplar subtasks executed in an optimal order, thereby establishing a theoretical lower bound on overall task difficulty.

- After the parameters for the standard tasks are defined, the next step is to calculate the Index of Difficulty for each task. This is done using the extended 3D formulation of Fitts' Law, which incorporates both translational and rotational motion, as discussed in Section 3.1.2. The calculation of the ID for each standard task is a critical step, as it provides the foundational data that will be used in subsequent stages of the model. The ID is used to quantify the difficulty of each task in terms of the time required to complete it, and thus forms the core component of the predictive model. By calculating the ID for each standard task, the model can generate predictions for a variety of practical tasks by aggregating the difficulties of the subtasks that make up those tasks.

### 3.1.1 Motion Capability Identification

Standard tasks play an integral role in understanding the system's capabilities and constructing an effective predictive model. These tasks serve as fundamental components in accurately characterising the performance of HMT systems, as they allow for the systematic decomposition of complex real-world tasks into more straightforward, more manageable elements. The process of identifying the motion capabilities of the targeted machine agent forms the cornerstone of the predictive model. Through this identification, the full range of the agent's capa-

bilities is understood, ensuring that the model is realistic and versatile enough to handle diverse tasks across various configurations.

The identification process determines the minimum standard tasks ($p$) necessary for constructing a comprehensive model. These standard tasks must cover a wide variety of motion capabilities, enabling the model to predict the performance of the machine agent across a broad spectrum of scenarios. By carefully selecting and defining these standard tasks, this study ensures that the model is equipped to handle different types of movements and interactions that the agent may encounter. The central goal here is to configure every potential real-world task as a combination of these pre-defined subtasks, thereby simplifying the task space and making the model adaptable to a variety of practical applications.

The calculation of the minimum set of standard tasks, $p$, is inherently linked to the total number of distinct motion capabilities ($m$) of the machine agent. These motion capabilities encompass both the individual capabilities and their possible combinations, as certain tasks may require multiple capabilities working together. To account for these combinations, this study employs the binomial coefficient $C_m^r$, which represents the number of ways in which $r$ capabilities can be selected from a total of $m$ distinct capabilities. The relationship between the number of distinct motion capabilities and the minimum number of standard tasks is mathematically described by the following formula:

$$p = \sum_{r=1}^{m} C_m^r = \sum_{r=1}^{m} \frac{m!}{r!(m-r)!} = 2^m - 1. \tag{3.1}$$

This equation captures the combinatorial nature of the problem, where the number of standard tasks grows exponentially with the total number of distinct motion capabilities. To illustrate this, consider a machine agent with multiple capabili-

ties, such as locomotion and manipulation movements. Each combination of these capabilities requires a corresponding standard task, leading to a rapidly increasing number of tasks as the system's motion capabilities expand. The value $2^m - 1$ provides the minimum number of tasks needed to represent all combinations of these capabilities, excluding the trivial case where no capabilities are used at all. Thus, this formulation serves as the foundation for ensuring that the model remains both comprehensive and efficient in predicting HMT performance.

To illustrate the application of the proposed model, consider the case of a wheeled-legged humanoid robot, which possesses multiple motion capabilities that can be classified into distinct categories. In this example, the robot has two types of locomotion capabilities: one through its legs and the other through its wheels. Additionally, the robot is equipped with two manipulation capabilities, one for each arm. These capabilities represent the fundamental actions the robot can perform, including both the translational and rotational motions required for tasks such as walking, wheeling, and manipulating objects.

In this scenario, the total number of distinct motion capabilities ($m$) of the robot is 4, as it encompasses the two locomotion modes (legs and wheels) and the two arms for manipulation. Given this total of $m = 4$ motion capabilities, this study can apply the formula derived earlier to calculate the minimum number of standard tasks ($p$) required to fully represent the system's motion capabilities in the predictive model.

According to the model, the required number of standard tasks ($p$) for accurate task prediction is computed as follows:

$$p = 2^m - 1 = 2^4 - 1 = 15. \tag{3.2}$$

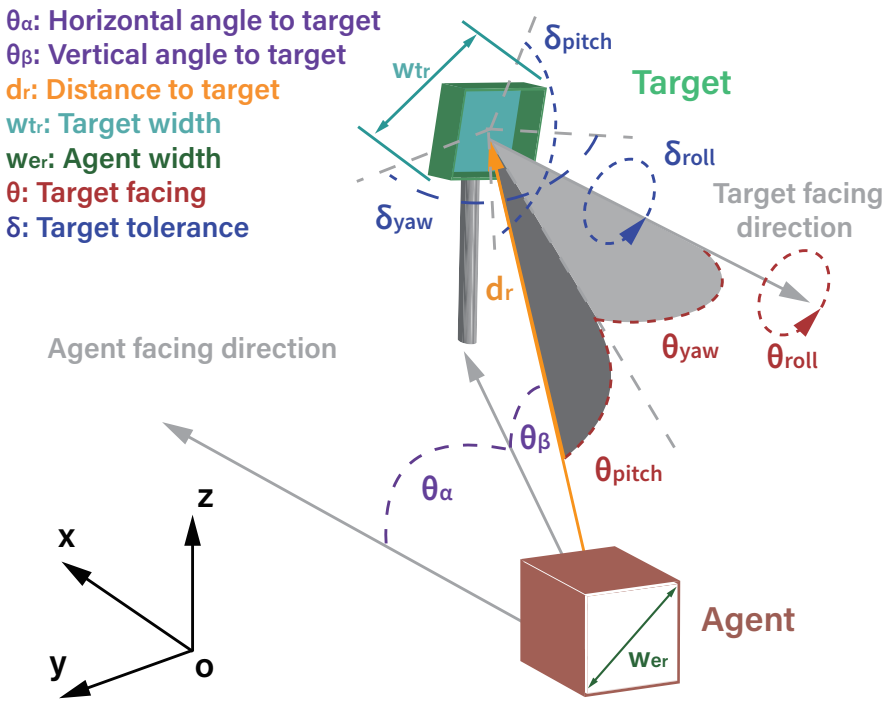



Figure 3.1: Giving the position of the machine agent and the target, the definition of each parameter used in calculating the 3D index of difficulty in a polar coordinate system.

This result demonstrates that, despite the robot having only four distinct motion capabilities, the model predicts that 15 standard tasks are necessary to capture all possible combinations of these capabilities. These tasks would serve as the basis for building a comprehensive predictive model, allowing the robot's performance to be accurately forecasted across a wide range of complex tasks involving various combinations of locomotion and manipulation capabilities. Thus, the number of required standard tasks grows rapidly.

### 3.1.2 Extending Fitts' Law to 3D

The proposed model extends the classical formulation of Fitts' Law by adapting it to predict task performance in the context of HMT. The central objective of the model is to estimate the motion time required for a machine agent to complete a

given task by systematically quantifying its associated difficulty. This difficulty is captured using an extended notion of the ID, tailored to the characteristics and capabilities of the agent in a 3D operational space.

The key advancement introduced by the model lies in its comprehensive treatment of task difficulty. Traditional applications of Fitts' Law in robotics and human-computer interaction have predominantly modelled task execution in simplified settings, often treating the agent as a dimensionless point and the target as a static volume in space. These models are typically limited to considering only translational motion and largely ignore the role of orientation, thereby restricting their applicability in complex, real-world scenarios. In contrast, the model overcomes these limitations by incorporating both translational and rotational aspects of motion, offering a complete 6-DoF formulation. Specifically, it accounts for three degrees of translational freedom (movement along the $x$, $y$, and $z$ axes) and three degrees of rotational freedom (yaw, pitch, and roll), thereby enabling predictions in truly three-dimensional task spaces.

This enhancement is particularly crucial for robotic systems engaged in real-world operations, where orientation plays a vital role. For example, tasks such as pressing a button, aligning a connector, or inserting a component during assembly require not just precise positioning but also specific approach angles. Neglecting orientation in such contexts leads to an incomplete characterisation of the task and, consequently, inaccurate performance predictions. By explicitly modelling orientation alongside position, the extended 3D Fitts' Law captures the full spatial complexity of these interactions.

To operationalize this, the total index of difficulty in the 3D formulation is decomposed into two distinct but complementary components: the translation index of difficulty ($\mathrm{ID}_{\mathrm{trans}}$), which quantifies the effort required to move the agent to the

target location, and the orientation index of difficulty ($\mathrm{ID}_{\mathrm{ori}}$), which reflects the effort needed to align the agent's end-effector or tool with the desired target orientation. These two components are additive and together represent the overall difficulty of the task. The formulation is expressed as:

$$\mathrm{ID} = \mathrm{ID}_{\mathrm{trans}} + \mathrm{ID}_{\mathrm{ori}}. \tag{3.3}$$

**Modelling distances**

Inspired by the foundational formulations of Fitts' Law, such as those expressed in Equations (2.5) and (2.6), the model distinguishes between two fundamental types of spatial difficulty: linear distance and rotational distance. These two components correspond to the physical translation and angular alignment challenges that a robotic agent must overcome to successfully interact with a target in 3D space. Together, they serve as the building blocks for the extended Index of Difficulty, allowing us to systematically quantify motion effort across all six degrees of freedom.

The first component, linear distance, refers to the straight-line displacement between the agent and the target. It is influenced not only by the spatial separation $d$ but also by the physical dimensions of both the target and the agent. These dimensions determine how much leeway the agent has in positioning its end-effector near or on the target. A larger target or a wider end-effector generally makes the task easier, while smaller or more precise targets require finer control and hence incur greater difficulty.

This study define the index of difficulty associated with linear distance as follows:

$$\mathrm{ID}_{\mathrm{linear}} = \log_2\left(\frac{d}{w_t \pm w_e} + 1\right), \tag{3.4}$$

where $d$ denotes the Euclidean distance between the agent's operational point (e.g., end-effector or base) and the target center. The terms $w_t$ and $w_e$ represent the effective widths (or tolerances) of the target and agent, respectively, measured along the relevant contact or interaction surface. The $\pm$ sign in the denominator is chosen based on the task's spatial configuration: in some cases, the target and agent widths may be combined (e.g., when both contribute to interaction range), whereas in others, a difference may be more appropriate (e.g., when the agent must fit inside a narrow cavity).

The second component, rotational distance, captures the angular deviation between the agent's current orientation and the desired orientation required to interact effectively with the target. This is especially important in tasks where alignment—such as inserting a peg into a hole, aligning a tool with a surface, or orienting a sensor toward a field of interest—is critical to success. Rotational distance is measured by the angle $\theta$ between the current and target orientations, while $\delta$ represents the allowable tolerance or precision limit of the system.

The rotational index of difficulty is defined analogously to its linear counterpart:

$$\mathrm{ID}_{\mathrm{rot}} = \log_2\left(\frac{\theta}{\delta} + 1\right), \tag{3.5}$$

where $\theta$ is typically expressed in degrees or radians, and $\delta$ defines the maximum permissible error in alignment. Smaller values of $\delta$ indicate tighter tolerances and therefore increase the difficulty of the task, while larger tolerances make the alignment easier to achieve.

An additional consideration arises when the target possesses rotational symmetry. For example, when the target is circular or when its geometry is invariant under certain rotations, the required angular alignment is effectively relaxed. In

such cases, the relevant angular displacement $\theta_{\text{eff}}$ is reduced by subtracting the symmetry angle $\phi$, corresponding to the smallest rotation that leaves the target unchanged:

$$\theta_{\text{eff}} = \max\big(0,\, \theta - \phi\big).$$

Accordingly, the rotational index of difficulty is evaluated as

$$\text{ID}_{\text{rot}}^{\text{sym}} = \log_2\left(\frac{\theta_{\text{eff}}}{\delta} + 1\right). \tag{3.6}$$

This formulation ensures that for fully symmetric targets (e.g., a circle, $\phi = 360°$), the effective displacement is $\theta_{\text{eff}} = 0$, yielding $\text{ID}_{\text{rot}}^{\text{sym}} = 0$ regardless of $\delta$. For partially symmetric objects (e.g., hexagons with $\phi = 60°$), the alignment tolerance is widened by the symmetry factor, thereby reducing the effective difficulty while still accounting for task precision.

By combining both linear and rotational indices, the model provides a holistic measure of task difficulty that encompasses both translation and orientation challenges. This is essential for accurately modelling robotic interactions in 3D environments, where most real-world tasks involve coordinated control over both position and attitude. The flexibility to incorporate both types of spatial relationships enables the framework to generalise across a wide variety of applications, from object manipulation and tool usage to locomotion and inspection tasks.

### Translation index of difficulty

For the translation index of difficulty ($\text{ID}_{\text{trans}}$), the calculation begins with the establishment of a consistent coordinate system centred at the initial position of the machine agent. This spatial reference frame provides a standardised basis for quantifying movement and ensures that translational distances are measured

uniformly across different task scenarios. In this coordinate system, the agent is oriented such that its default forward-facing direction aligns with the positive $x$-axis. The $+y$ direction is defined to point laterally to the agent's left side, while the $+z$ direction points vertically upward, following the conventional right-hand rule used in robotics and mechanical systems. This configuration provides an intuitive and consistent spatial mapping that facilitates the measurement of target locations and movement trajectories in 3D space.

To describe the relative position of a given target with respect to the agent, two types of coordinate systems can be employed: Cartesian and polar. The choice of a coordinate system depends on the mechanical structure and kinematic behaviour of the robot or machine agent in question. For instance, Cartesian robots—such as those that operate along orthogonal linear rails—naturally align with the Cartesian coordinate system, as their motion is inherently constrained along independent $x$, $y$, and $z$ axes. In this case, the distance to the target can be directly calculated using linear displacement components along each axis, simplifying the formulation of $\mathrm{ID}_{\mathrm{trans}}$.

In contrast, robots with articulated or rotational joints, such as robotic arms or humanoid manipulators, often benefit from a polar (or spherical) coordinate system, which more naturally reflects their kinematic structure. These robots typically perform tasks by rotating joints to achieve certain angles, making angular measurements—such as azimuth and elevation angles—more meaningful than Cartesian distances. For such systems, the position of the target relative to the agent is expressed using radial distance and angular displacements, allowing for a more accurate and intuitive representation of motion effort.

By supporting both Cartesian and polar coordinate systems, the model ensures compatibility with a wide range of robotic architectures, thereby enhancing its

generality and applicability. This flexibility allows the model to be seamlessly integrated into diverse robotic platforms without requiring major reconfiguration of the coordinate systems or motion planning frameworks. Ultimately, the coordinate system serves as the foundational structure for computing the translation index of difficulty, enabling precise quantification of the spatial challenge involved in reaching a target location.

In a Cartesian coordinate system, the translation is described by linear distances along the $x$, $y$, and $z$ axes. Thus, the $\mathrm{ID}_{\mathrm{trans}}$ can be expressed as the sum of three $\mathrm{ID}_{\mathrm{linear}}$ components:

$$\mathrm{ID}_{\mathrm{trans}} = \sum_{i \in \{x,y,z\}} \mathrm{ID}_{\mathrm{linear}_i} = \sum_{i \in \{x,y,z\}} \log_2 \left( \frac{d_i}{w_{t_i} \pm w_{e_i}} + 1 \right). \tag{3.7}$$

In a polar coordinate system, the position of a target relative to the agent is described using two angular components and one radial component, making it well-suited for systems with rotational kinematics, such as articulated manipulators or mobile manipulators. Specifically, the location is characterized by two rotational distances—denoted as $\alpha$ and $\beta$—and a linear distance, denoted as $r$. These three parameters form a complete representation of a point in 3D space from the perspective of a rotationally capable agent.

The first rotational component, $\alpha$, corresponds to the azimuthal angle ($\theta_\alpha$), which is defined as the angle between the projection of the target point onto the $xy$-plane and the positive $x$-axis. This angle captures the horizontal deviation from the forward-facing direction of the agent, essentially describing how far left or right the target is located relative to the agent's heading. The second rotational component, $\beta$, corresponds to the polar or elevation angle ($\theta_\beta$), which is defined as the angle between the vector pointing from the agent to the target and the

$xy$-plane. This angle quantifies the vertical displacement of the target, indicating whether the agent must reach upwards or downwards to access the target.

The linear component, $r$, represents the Euclidean distance from the agent to the target, measuring the straight-line displacement required to physically reach the target's location. Together, $(r, \alpha, \beta)$ form a spherical coordinate representation of the target's position, which is particularly effective for capturing the full spatial relationship in systems where angular movement is more relevant than linear translation.

Each angular component—$\theta_\alpha$ and $\theta_\beta$—is constrained by the mechanical configuration of the agent. These constraints are captured by their respective maximum movement ranges, denoted as $\delta_\alpha$ and $\delta_\beta$, which define the bounds within which the agent can rotate to reach a target. Typically, these values lie within the range $[0°, 180°]$, although they may vary depending on the joint limits and design of the agent.

In cases where the agent possesses a full rotational capability of 360°, the effective maximum range is still modelled as 180° in either direction. This simplification is adopted based on the assumption that the agent will always choose the shortest angular path to the target. Thus, regardless of the full rotation capability, the model assumes a maximum effective angle of 180°, allowing the agent to approach the target from the nearer side. This modelling choice ensures consistency in difficulty calculations and reflects the intelligent behaviour of most robotic systems, which naturally optimise for minimal effort in reaching a goal.

As depicted in Fig. 3.1, this polar representation provides an intuitive and spatially rich framework for defining target positions in relation to the agent, making it a critical component in the calculation of the translation index of difficulty for robots with articulated or rotational degrees of freedom.

Thus, the translation index of difficulty ($\mathrm{ID}_{\mathrm{trans}}$) in a polar coordinate system can be formulated as the cumulative difficulty arising from both rotational and linear movements required to reach the target. Specifically, this index is composed of two rotational components, corresponding to angular displacements in azimuth ($\alpha$) and elevation ($\beta$), and one linear component, corresponding to the radial distance ($r$) between the agent and the target. These three components collectively describe the spatial effort involved in translating the agent's end-effector to the desired position in 3D space.

The rotational components are each quantified using a logarithmic formulation consistent with Fitts' Law, which models the trade-off between movement amplitude and precision. For each angle $\theta_i \in \{\alpha, \beta\}$, the corresponding index of difficulty is computed as a function of the ratio between the angular displacement to be covered and the allowable angular tolerance $\delta_i$, which reflects the agent's rotational precision limits. Similarly, the radial or linear component is evaluated using a standard $\mathrm{ID}_{\mathrm{linear}}$ formulation, involving the ratio between the radial distance $d_r$ and the combined widths of the target and agent along the contacting surface.

The complete expression for $\mathrm{ID}_{\mathrm{trans}}$ in a polar coordinate system is given by:

$$\begin{aligned} \mathrm{ID}_{\mathrm{trans}} &= \sum_{i \in \{\alpha,\beta\}} \mathrm{ID}_{\mathrm{rot}_i} + \mathrm{ID}_{\mathrm{linear}_r} \\ &= \sum_{i \in \{\alpha,\beta\}} \log_2\left(\frac{\theta_i}{\delta_i} + 1\right) + \log_2\left(\frac{d_r}{w_{t_r} \pm w_{e_r}} + 1\right), \end{aligned} \tag{3.8}$$

In this equation, $w_{t_r}$ and $w_{e_r}$ denote the effective widths of the target and the agent, respectively, along the relevant contact surface. These widths are crucial for defining the spatial precision required in reaching the target, as they set the bounds within which successful contact or alignment is considered acceptable.

The angular displacements are defined using the two-argument arctangent function atan2, which avoids quadrant ambiguity:

$$\theta_\alpha = \text{atan2}(y, x), \qquad \theta_\beta = \text{atan2}\Big(z, \sqrt{x^2 + y^2}\Big),$$

where $(x, y, z)$ is the position of the target relative to the agent.

Since atan2 returns values in the range $[-180°, 180°]$, the effective angular displacement used in (3.8) is mapped to the minimal equivalent within $[0°, 180°]$:

$$\theta_i = \min\big(|\theta_i|,\, 360° - |\theta_i|\big), \quad i \in \{\alpha, \beta\}.$$

This modelling choice reflects the assumption that the agent always follows the shortest angular path to the target. Consequently, even when an agent possesses full 360° rotation capability, the effective difficulty is bounded by 180°, ensuring consistency with the intelligent behaviour of robotic systems that naturally optimise for minimal effort.

The interpretation of these widths depends on the nature of the task. In locomotion tasks, where the agent is navigating toward a physical location on a surface (e.g., walking to a waypoint), the contacting surface is typically the ground plane. In this case, $w_t$ and $w_e$ are computed as the diagonal lengths of the vertical projection areas of the target and the agent, respectively, providing an accurate representation of their effective ground footprint. In manipulation tasks, by contrast, the interaction usually occurs through the front face of the end-effector. Here, the relevant contact surface is the end-effector's working plane, and the sizes $w_t$ and $w_e$ are measured along the projected frontal areas involved in the interaction.

By incorporating both angular and linear movement challenges, this formulation enables the model to quantify the full complexity of spatial interaction in 3D environments, whether for positioning during locomotion or alignment during fine manipulation. This holistic view ensures that the predicted performance accurately reflects the real-world difficulty of the movement, grounded in the physical characteristics and precision limitations of the HMT system.

### Orientation index of difficulty

The orientation index of difficulty ($ID_{ori}$) is a critical component of the extended 3D Fitts' Law model, capturing the rotational effort required for the agent to align itself with the desired orientation of the target. In many real-world scenarios—particularly those involving fine manipulation, assembly, or interaction with oriented surfaces—reaching the correct spatial location alone is insufficient. The agent must also adopt an appropriate orientation to effectively engage with the target, which may involve aligning tools, sensors, or grippers with specific approach directions. As such, the orientation component is essential for accurately modelling the true complexity of these tasks.

The formulation of $ID_{ori}$ accounts for two primary variables: the angular deviation from the optimal approach angle, denoted by $\theta$, and the allowed angular tolerance or acceptance window, denoted by $\delta$. The optimal approach angle represents the desired orientation of the agent's end-effector relative to the target's facing direction, which is typically dictated by the geometry or functional requirements of the task. The tolerance $\delta$ defines how precisely this orientation must be achieved in order for the task to be completed successfully, encapsulating the precision constraints inherent to the task.

To systematically quantify the orientation difficulty, this study decomposes both

the deviation $\theta$ and the tolerance $\delta$ into three principal components corresponding to the standard Euler angles: yaw, pitch, and roll. These components describe the rotation of the agent around the vertical, lateral, and longitudinal axes, respectively. This decomposition is visualised in Fig. 3.1, which shows how the agent's orientation must be adjusted across all three rotational degrees of freedom to match the target's pose.

With this decomposition, the total orientation index of difficulty is expressed as the sum of three $\mathrm{ID}_{\mathrm{rot}}$ components, each of which is calculated using a logarithmic formulation that mirrors classical Fitts' Law:

$$\mathrm{ID}_{\mathrm{ori}} = \sum_{i \in \{\mathrm{yaw,pitch,roll}\}} \mathrm{ID}_{\mathrm{rot}_i} = \sum_{i \in \{\mathrm{yaw,pitch,roll}\}} \log_2 \left( \frac{\theta_i}{\delta_i} + 1 \right). \tag{3.9}$$

Each term in this summation reflects the difficulty associated with rotating the agent about one of the three axes, depending on how far it must rotate ($\theta_i$) and how precise the alignment must be ($\delta_i$). Tasks requiring exact orientation, such as inserting a key into a lock or aligning a connector with a socket, will result in larger values of $\mathrm{ID}_{\mathrm{ori}}$, indicating a higher overall difficulty due to tight tolerances.

The inclusion of orientation in the computation of task difficulty is one of the main contributions of the work. While prior formulations of Fitts' Law and its adaptations to robotics have primarily focused on translational movement, often treating the robot as a point mass moving in space, the approach explicitly models orientation by accounting for all six degrees of freedom (three translational and three rotational). This extension provides a more comprehensive and realistic representation of tasks encountered in robotic applications, where orientation often plays a crucial role. By doing so, the model not only improves predictive accuracy but also offers a principled framework for planning, control,

and evaluation of complex HMT tasks in fully 3D environments.

### Generalization of the model

HMTs may exhibit diverse motion capabilities across different DoFs, and the level of difficulty associated with executing a movement can vary substantially depending not only on the specific DoF but also on the direction, type, and dynamics of the motion. For example, a robot might perform forward translations more efficiently than lateral ones, or it may exhibit greater precision in yaw rotations compared to pitch or roll due to mechanical constraints or sensor placement. These disparities arise from variations in actuation strength, mechanical structure, control algorithms, joint limits, payload distribution, and compliance characteristics, among other factors.

To faithfully capture such variability and ensure that the model is capable of generalising across a wide range of HMT systems and task settings, we introduce a set of weighting factors, denoted by $k$, for each component of the ID. These weighting factors serve as scaling parameters that adjust the contribution of individual DoFs to the total task difficulty based on the HMT system's unique motion characteristics and execution proficiency. Essentially, the weights encode domain-specific knowledge about the system's capabilities and limitations, allowing the model to be customised or calibrated for different robotic platforms or use-case scenarios.

By incorporating these weights, the model gains the flexibility to account for anisotropic performance—that is, unequal ease or difficulty of movement across different directions or modes. This is particularly important in heterogeneous robot teams or adaptive systems, where capabilities may differ significantly between agents or evolve due to wear, reconfiguration, or learning. The weighting

scheme thus enhances the adaptability and realism of the model, enabling it to remain robust and predictive across a broad spectrum of real-world applications.

In practice, these weights can be empirically determined from experimental performance data or derived analytically from the system's kinematic and dynamic specifications. Once defined, they are applied multiplicatively to the corresponding translational and orientational components of the ID, as elaborated in the following sections. This weighted formulation forms the basis for a generalised and system-aware extension of Fitts' Law to 6-DoF task performance modelling.

These weighting factors reflect the intrinsic motion capabilities of the HMT system across different degrees of freedom, enabling the model to more accurately represent the varying levels of difficulty associated with specific types of motion. In real-world robotic systems, it is common for certain motions to be inherently more efficient or precise than others due to differences in hardware design, actuation mechanisms, control fidelity, or physical constraints. For example, an HMT system might exhibit high speed and accuracy in translational motion along the $x$-axis due to optimised linear actuators, while rotational movements around the roll axis may be less stable or slower due to mechanical limitations or controller sensitivity. These disparities in performance must be acknowledged in the difficulty model to ensure realistic and system-aware predictions.

To accommodate such differences, the original unweighted formulation of the total index of difficulty given in Equation(3.3) is extended by incorporating two sets of weight vectors: $\boldsymbol{k}_{\text{trans}}$ for translational difficulty and $\boldsymbol{k}_{\text{ori}}$ for orientational difficulty. These vectors encode the relative contribution or influence of each DoF to the overall task difficulty. Depending on the structure of the robot, different weights may be assigned to different axes or rotation types. For example, if a system exhibits low precision in yaw but high accuracy in pitch and roll, a larger

value of $k_{\text{yaw}}$ can be applied to reflect the additional challenge. The modified formulation of the total difficulty is expressed as:

$$\text{ID} = \boldsymbol{k}_{\text{trans}} \cdot \text{ID}_{\text{trans}} + \boldsymbol{k}_{\text{ori}} \cdot \text{ID}_{\text{ori}}, \tag{3.10}$$

where $\boldsymbol{k}_{\text{trans}} = [k_x, k_y, k_z]$ and $\boldsymbol{k}_{\text{ori}} = [k_{\text{yaw}}, k_{\text{pitch}}, k_{\text{roll}}]$ are the weight vectors applied to each corresponding DoF. These weights can either be scalars applied uniformly across each subcomponent or vectors allowing fine-grained tuning per axis, depending on the specificity and granularity required by the application.

These weights can either be scalars applied uniformly across each subcomponent or vectors allowing fine-grained tuning per axis, depending on the specificity and granularity required by the application. Moreover, the $k$-weightings need not be fixed a priori; they can be empirically learned through experimental data, in a manner analogous to how the coefficients $a$ and $b$ in Fitts' Law are determined via linear regression.

By integrating these weights, the model gains a valuable mechanism for tailoring itself to a wide range of heterogeneous robotic systems. It ensures that the predictive outputs remain faithful to the system's actual operational capabilities, regardless of whether the robot is an aerial drone, a ground vehicle, a manipulator arm, or a hybrid configuration. This flexibility greatly enhances the generalisability of the model and allows it to be deployed across diverse application domains, including industrial assembly, search and rescue, teleoperation, and assistive robotics.

For clarity and focus in this paper, this study assumes a simplified case where the HMT is evenly capable across all DoFs. This assumption allows us to set all weights to unity, i.e., $k = 1$ for all DoFs. This baseline configuration serves

as a default setting for evaluation and comparison, while still leaving the door open for more specialised tuning in future implementations or platform-specific deployments.

### Practical task modelling

In real-world applications, practical tasks are rarely monolithic; rather, they are typically composed of a sequence or combination of multiple subtasks, each representing a distinct movement, alignment, or interaction event. These subtasks may vary in their spatial and temporal characteristics, and they often engage different degrees of freedom to varying extents. For instance, a pick-and-place operation may involve navigating toward an object, grasping it with an appropriate end-effector orientation, transporting it across a workspace, and then precisely aligning and releasing it at a target location. Each of these steps constitutes a subtask with its own motion requirements and associated difficulty.

To model such complex task sequences within the framework, this study adopts a modular approach that decomposes a practical task into its constituent subtasks. This decomposition not only simplifies the analysis and planning process but also aligns with the capabilities-based structure outlined in Section 3.1.1. As long as each subtask falls within the identified motion capability set of the HMT system—i.e., it can be executed using one or more of the standard task primitives—the entire task remains representable within the framework.

Once the task has been broken down into subtasks, each one can be individually evaluated in terms of difficulty using the weighted index of difficulty defined in Equation (3.10). This approach allows for fine-grained modelling of performance cost, taking into account both the spatial configuration and the specific movement dynamics required for each subcomponent of the task. It also facilitates

comparative analysis and optimisation, where individual subtasks may be reallocated, reordered, or adapted based on their respective difficulty scores and the agent's performance profile.

By enabling the difficulty estimation of practical tasks through the aggregation of weighted subtask difficulties, the model provides a principled and scalable way to analyse complex operations. This decomposition strategy is essential for capturing the hierarchical and sequential nature of tasks in real-world HMT applications, from industrial automation to service robotics and human-robot collaboration scenarios.

Thus, for any feasible, practical task composed of multiple subtasks, the overall index of difficulty for the proposed 3D Fitts' Law model can be expressed as the weighted sum of the individual subtask difficulties. Formally, this is written as:

$$\mathrm{ID}_{\mathrm{prac}} = \sum_{i=1}^{n} k_i \cdot \mathrm{ID}_i, \tag{3.11}$$

where $n$ denotes the number of subtasks comprising the complete task, $\mathrm{ID}_i$ represents the difficulty of the $i$-th subtask as computed using the 3D model, and $k_i$ is the corresponding weight that modulates the contribution of that subtask to the overall task difficulty. The weight $k_i$ may reflect a variety of factors, including the intrinsic complexity of the motion, the dynamic properties of the system in that configuration, or the strategic importance of that subtask within the broader task context.

This formulation serves as a generalisation of Fitts' Law from simple, single-target movements to realistic, composite tasks that require coordinated sequences of actions across multiple degrees of freedom. It allows us to capture the heterogeneity of effort involved in practical HMT tasks by appropriately scaling the

difficulty contributions of each subtask. This is particularly important in scenarios where some subtasks are disproportionately demanding—such as requiring high-precision manipulation or long-range navigation—while others may be comparatively trivial.

The decomposition into weighted subtasks also provides significant practical advantages. It enables modular analysis, making it possible to evaluate and optimise individual components of a task independently or in parallel. Moreover, this structure supports the implementation of learning-based or adaptive planning algorithms that iteratively refine execution strategies based on the observed or predicted difficulty of subtasks. By assigning task-specific weights, the model can dynamically prioritise or reallocate resources to optimise performance outcomes.

Overall, Equation (3.11) forms the final step in the generalised difficulty modelling pipeline, bringing together the geometric, kinematic, and system-specific factors into a unified and interpretable measure of task complexity. The resulting framework is capable of predicting HMT performance in a broad spectrum of real-world scenarios, including those involving heterogeneous agents, intricate spatial configurations, and diverse motion requirements. Through this approach, the proposed 3D Fitts' Law transcends its classical roots and becomes a powerful tool for quantifying, comparing, and optimising robot behaviour in complex environments.

## 3.2 Validation

This section presents a comprehensive validation process for the proposed prediction model, encompassing both simulation-based evaluations and real-world experimental trials. The primary objective of this validation is to rigorously assess the model's versatility, robustness, and predictive accuracy across different

domains and agent types. By applying the model to both simulated and physical robotic systems, this study aims to demonstrate its generalisability and practical utility in diverse application scenarios.

The validation process is structured around two complementary experimental platforms. The first involves a high-fidelity simulation environment featuring a construction machinery agent, designed to test the model's performance under controlled and parameterised conditions. This simulation allows for repeatable trials and systematic exploration of motion difficulty across a wide range of spatial configurations and subtasks. The second validation platform consists of real-world experiments conducted with a legged robotic agent, providing a physically grounded assessment of the model's applicability in more dynamic and unpredictable environments. These two platforms together offer a comprehensive testbed for evaluating both the theoretical soundness and the empirical viability of the proposed framework.

In both cases, the model is used to predict the performance of the agent on a complex multistep task, representative of practical operations encountered in construction, logistics, or field robotics. The prediction is grounded in the performance parameters obtained from a predefined set of standard tasks, as outlined in earlier sections. These standard task results serve as the foundation for estimating the difficulty—and hence the expected execution time—of a more advanced composite task involving multiple degrees of freedom and sequential actions.

As mentioned, the advanced practical task is decomposed into subtasks, each of which is mapped to its corresponding standard task for difficulty estimation. The predicted total execution time is then obtained by aggregating the individual subtask predictions using the extended and weighted formulation of the 3D Fitts' Law. This predicted value is subsequently compared against the actual execution

time measured during the experiment to assess the predictive accuracy of the model.

By comparing predicted performance times with actual measurements, this study can quantify the model's effectiveness in both real-world and simulated contexts. Strong alignment between the predicted and observed results provides evidence for the validity of the model, confirming its capacity to represent and anticipate HMT behaviour in practical scenarios. Discrepancies, where present, are analysed to understand their origin—whether due to modelling assumptions, environmental noise, or execution uncertainties—offering further insight into areas for future refinement.

### 3.2.1 General Validation

#### Design principle

This study selects articulated machines and robotic platforms as the focus of the validation studies due to their broad applicability and increasing prevalence in modern industrial and field robotics. These systems, which combine multiple joints and linkages to provide flexible motion across a range of degrees of freedom, are well-suited for complex manipulation and interaction tasks. Their widespread adoption in domains such as construction, logistics, manufacturing, and service robotics underscores the importance of ensuring that the prediction model is well-adapted to their unique motion characteristics.

To reflect the kinematic structure of these systems, this study employs a polar coordinate system for task modelling and difficulty estimation, as formalised in Equation (3.8). This coordinate system offers a natural and intuitive way to describe spatial relationships and target positions from the perspective of articulated agents, especially those with rotational bases or serial manipulators. It

allows us to effectively capture both the radial and angular components of motion, which are particularly relevant for agents that rely on rotational joints for reaching and alignment.

In the experimental setup, this study uses a mobile base equipped with an articulated robotic arm as the representative agent platform. This class of platform offers the combined advantages of locomotion and manipulation: it can traverse large or cluttered environments using the mobile base, and it can perform fine-grained interaction tasks using the articulated arm. Such hybrid capabilities make these platforms ideal candidates for deployment in unstructured environments where flexibility and adaptability are paramount.

For the simulation-based validation, this study selected a hydraulically actuated construction excavator as the testbed. Excavators are quintessential articulated machines widely used in heavy-duty industrial tasks such as digging, lifting, grading, and material handling. Their motion involves coordinated control of multiple rotational and prismatic joints, making them an ideal candidate for evaluating the 3D Fitts' Law model under realistic control constraints.

For the real-world experimental validation, this study employed a legged mobile manipulator robot—an advanced robotic platform that integrates legged locomotion with upper-body manipulation. These robots represent the frontier of mobile manipulation systems and are especially suited for environments where wheeled mobility is inadequate. Their articulation and dynamic balance capabilities enable them to navigate irregular terrain, climb stairs, or reposition their base to gain better manipulation leverage, all while performing fine-motor tasks with their arm(s).

Although the excavator and the legged manipulator share a common high-level structural architecture—namely, the integration of base mobility with articulated

manipulation—their control strategies and motion capabilities differ significantly. Excavators typically rely on hydraulic actuation and are operated using joint-level teleoperation or programmable control sequences, often exhibiting non-trivial dynamics and limited feedback loops. In contrast, legged mobile manipulators usually involve sophisticated onboard control algorithms, sensor fusion for perception, and real-time motion planning to maintain balance and coordination across limbs.

By selecting these two representative platforms, this study ensures that the validation process covers a broad spectrum of control paradigms and operational contexts, thus reinforcing the generalisability and practical relevance of the proposed prediction model.

The selected mobile manipulator platform possesses two primary motion capabilities ($m = 2$): a locomotion capability that enables it to navigate through its environment, and a manipulation capability provided by its articulated arm for object interaction and task execution. According to Equation (3.1), the total number of distinct standard tasks ($p$) required to fully characterize a system with $m$ motion capabilities is computed as $p = 2^m - 1 = 3$. This result implies that three standard tasks are sufficient to capture all necessary combinations of motion capabilities for this platform.

These three standard tasks are strategically designed to independently and jointly evaluate the key motion modalities of the HMT system. The first standard task isolates and evaluates the agent's locomotion capability, requiring the robot to reach a spatial goal using its mobile base while excluding any manipulation activity. This task enables us to measure baseline navigation performance, such as travel time, spatial accuracy, and responsiveness under varying spatial conditions.

The second standard task focuses solely on the robot's manipulation capability.

In this task, the base remains stationary while the manipulator performs an object interaction task, such as reaching, aligning, or grasping. This isolated evaluation allows us to characterise the dexterity, precision, and control accuracy of the arm across different spatial orientations and interaction conditions.

The third standard task integrates both capabilities, requiring coordinated use of locomotion and manipulation. For example, the robot may need to reposition itself to gain access to a workspace and then use its arm to complete a manipulation subtask. This integrated task assesses the system's ability to plan and execute compound actions, including base-arm coordination and task sequencing. Importantly, it also reflects the kinds of challenges typically encountered in real-world applications, where mobility and dexterity must be used in tandem.

Collectively, these three tasks form a minimal yet complete set of training examples that span the full motion capability space of the platform. As a result, any future practical task—no matter how complex—can be decomposed into a combination of these three standard task types. This decomposition is central to the modelling process: the performance data collected from each standard task serve as reference points for estimating the difficulty and expected completion time of more advanced, composite tasks encountered in deployment.

Although the model supports generalisation via variable weighting, as introduced in Section 3.1.2, this study adopts a conservative assumption for the current study to isolate and evaluate the structural effectiveness of the framework. Specifically, this study simplifies the system by assigning uniform weights ($k = 1$) to all motion components and task types. This simplification allows us to assess the model's predictive capability without introducing additional tuning parameters, thereby emphasising the foundational utility of the standard task formulation and the additive index structure.

### Modelling tasks

The three standard tasks in the selected mobile manipulation platform are locomotion task ($T_{mob}$), manipulation task ($T_{mani}$), and combined task ($T_{comb}$), as shown in Fig. 3.2. The practical task ($T_{prac}$) is a complex Explosive Ordnance Disposal (EOD) mission formed of a series of different subtasks.

**Locomotion task** This task necessitates only the locomotion capability. It contains only locomotion movement, as shown in Fig. 3.2a. It has index of difficulty ($ID_{mob}$) calculated as (3.3).

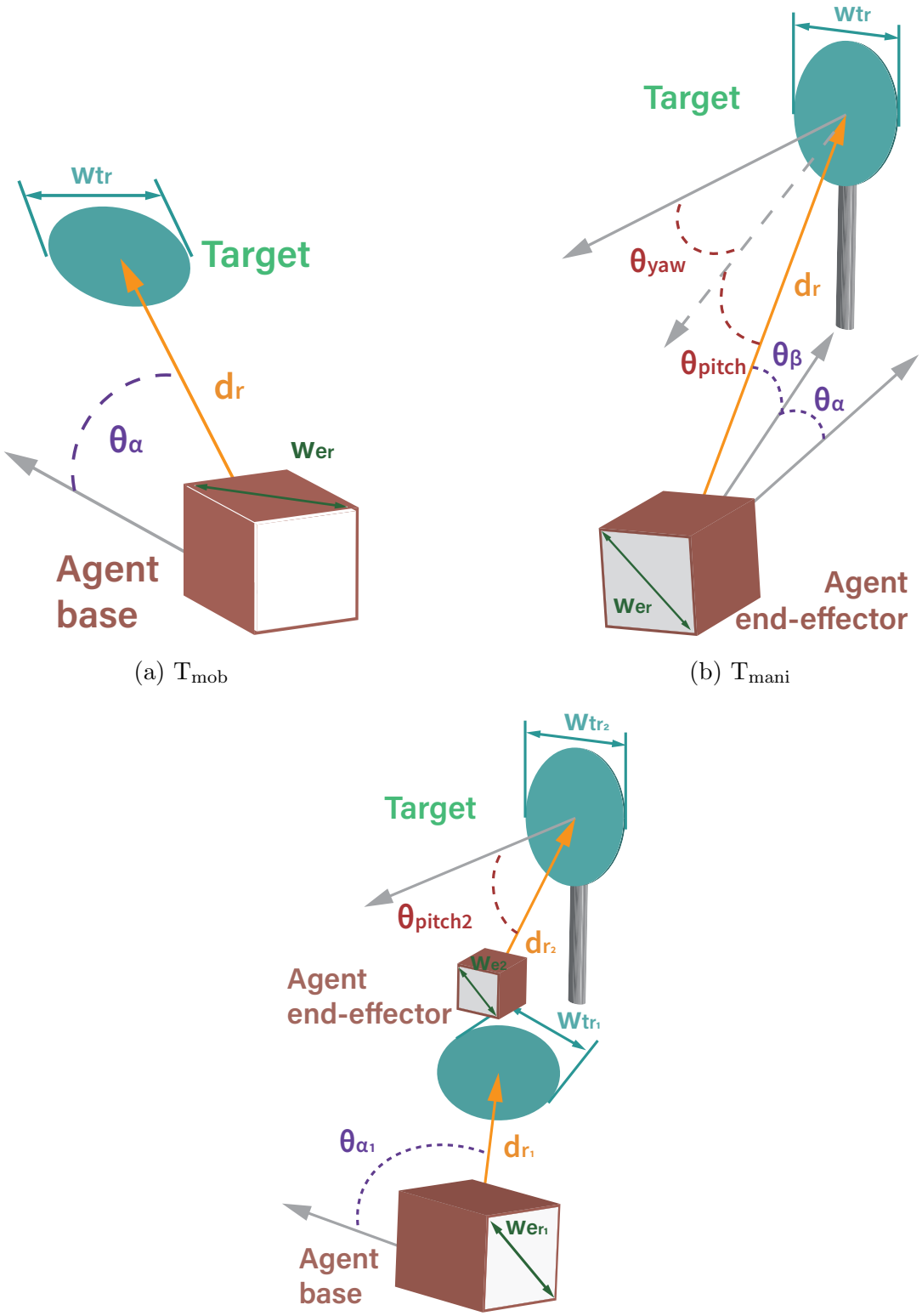
**Manipulation task** This task utilises only the manipulation capability and involves a single task, as shown in Fig. 3.2b. Thus, its index of difficulty ($ID_{mani}$) can be calculated as (3.3) as well.

**Combined task** The last standard task requires both locomotion and manipulation capabilities, as shown in Fig. 3.2c. As such, the total index of difficulty ($ID_{comb}$) is modelled as a sum of two subtasks ($n = 2$), as per (3.11):

$$ID_{comb} = ID_{comb_1} + ID_{comb_2}. \tag{3.12}$$

In the first step, the machine travels to the target location in subtask 1 ($T_{comb_1}$). In the second step, the machine employs the arm to touch the target in subtask 2 ($T_{comb_2}$), where the agent is expected to already stand directly in front of the target, with the direction angle and the target's best approach angle both set to zero ($\theta_\alpha = 0$ and $\theta_{yaw} = 0$).

**Practical task** As stated in Section 3.1.2, all practical tasks that are feasible to the HMT can be decomposed into a combination of one or more standard tasks




(a) $T_{mob}$ (b) $T_{mani}$

(c) $T_{comb}$

Figure 3.2: Example of parameters for the selected platform in three required standard tasks: (a) locomotion task, (b) manipulation task, and (c) combined task.

with (3.11). Therefore, a practical task ($T_{prac}$) has its index of difficulty ($ID_{prac}$) to be calculated by the summation of these subtasks' difficulty.

### 3.2.2 Simulation with Excavator

As depicted in Fig. 3.3, this study developed a web-based physical simulation of a teleoperated excavator to evaluate the proposed task performance prediction framework. The excavator, a widely deployed HMT system in industrial and construction contexts, offers a rich testing ground due to its multiple degrees of freedom and combined manipulation and locomotion capabilities. This simulation-based platform allows for controlled experimentation across a variety of scenarios while facilitating scalable data collection.

The excavator used in the simulation supports two primary motion capabilities: locomotion and manipulation. According to the task decomposition strategy discussed in Section 3.2.1, three standard tasks are defined. The first task isolates the locomotion capability, evaluating the agent's ability to reposition itself within the environment. The second task focuses solely on the manipulator arm, measuring the operator's performance in handling objects without base movement. The third task combines both locomotion and manipulation to assess compound skill coordination under more complex conditions.

#### Simulation design

Participants engaged with the simulation through an interactive web-based platform developed using the WebGL Application Programming Interface (API). This platform allows the simulation to run entirely within a web browser, without requiring any additional plug-ins or software installation. Once fully loaded, the WebGL API executes the simulation locally on the user's machine, leveraging the

Table 3.1: Parameters of tasks and their subtasks with calculated indexes of difficulty, where $d$ and $w$ are in meters, $\theta$ and $\delta$ are in degrees. Also, $\delta_\beta = 90°$ for simulation, $\delta_\beta = 180°$ for experiment, $\delta_\alpha = 180°$ for both.

| | | Distance | Target size | Agent size | Direction horizontal | Direction vertical | Tolerance yaw | Best angle yaw | Tolerance pitch | Best angle pitch | Task difficulty |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $d$ | $w_t$ | $w_e$ | $\theta_\alpha$ | $\theta_\beta$ | $\delta_{\text{yaw}}$ | $\theta_{\text{yaw}}$ | $\delta_{\text{pitch}}$ | $\theta_{\text{pitch}}$ | **ID** |
| Simulation | $T^{\text{sim}}_{\text{mob}}$ | 32.56 | 2 | 10.44 | 10.62 | 0 | 360 | 0 | - | 0 | 1.94 |
| | $T^{\text{sim}}_{\text{mani}}$ | 6.63 | 1 | 1.41 | 71.57 | 45 | 180 | 18.43 | 180 | 45 | 3.44 |
| | $T^{\text{sim}}_{\text{comb}_{\text{mob}}}$ | 25.18 | 3 | 10.44 | 80.84 | 0 | 180 | 83.16 | - | 0 | 4.90 |
| | $T^{\text{sim}}_{\text{comb}_{\text{mani}}}$ | 5 | 3 | 1.41 | 0 | 63.43 | 180 | 0 | 180 | 63.43 | |
| | $T^{\text{sim}}_{\text{prac}_{\text{taeget1.mob}}}$ | 24.41 | 2 | 10.44 | 55.01 | 0 | 180 | 124.99 | - | 0 | avg:13.42 min:12.29 max:16.05 |
| | $T^{\text{sim}}_{\text{prac}_{\text{taeget1.mani}}}$ | 2 | 2 | 1.41 | 0 | 63.43 | 180 | 0 | 180 | 63.43 | |
| | $T^{\text{sim}}_{\text{prac}_{\text{taeget2.mob}}}$ | 28.42 | 1 | 10.44 | 39.29 | 0 | 360 | 0 | - | 0 | |
| | $T^{\text{sim}}_{\text{prac}_{\text{taeget2.mani}}}$ | 0 | 1 | 1.41 | 0 | 0 | 360 | 0 | 360 | 0 | |
| | $T^{\text{sim}}_{\text{prac}_{\text{taeget3.mob}}}$ | 45.71 | 3 | 10.44 | 10.08 | 0 | 120 | 79.92 | - | 0 | |
| | $T^{\text{sim}}_{\text{prac}_{\text{taeget3.mani}}}$ | 2.86 | 3 | 1.41 | 0 | 70.73 | 120 | 0 | 120 | 70.73 | |
| Experiment | $T^{\text{exp}}_{\text{mob}}$ | 2 | 0.1 | 0.69 | 45 | 0 | 360 | 0 | - | 0 | 2.14 |
| | $T^{\text{exp}}_{\text{mani}}$ | 0.75 | 0.20 | 0.13 | 45 | 75 | 180 | 45 | 180 | 75 | 3.37 |
| | $T^{\text{exp}}_{\text{comb}_{\text{mob}}}$ | 1.12 | 0.20 | 0 | 70 | 0 | 360 | 20 | - | 0 | 6.00 |
| | $T^{\text{exp}}_{\text{comb}_{\text{mani}}}$ | 0.75 | 0.20 | 0.13 | 0 | 75 | 180 | 0 | 180 | 75 | |
| | $T^{\text{exp}}_{\text{prac}_{\text{step1}}}$ | 1.45 | 0.14 | 0 | 5 | 0 | 90 | 45 | - | 0 | 13.63 |
| | $T^{\text{exp}}_{\text{prac}_{\text{step2}}}$ | 0.75 | 0.14 | 0.06 | 0 | 10 | 90 | 0 | 90 | 10 | |
| | $T^{\text{exp}}_{\text{prac}_{\text{step3}}}$ | 0.75 | 0.08 | 0.06 | 0 | 10 | 60 | 0 | 60 | 10 | |

device's hardware to provide a consistent and high-performance experience. This design ensures that participant interactions are not affected by internet latency or network instability during the trial itself; internet connectivity is only required for loading the simulation initially and for submitting the final results upon task completion.

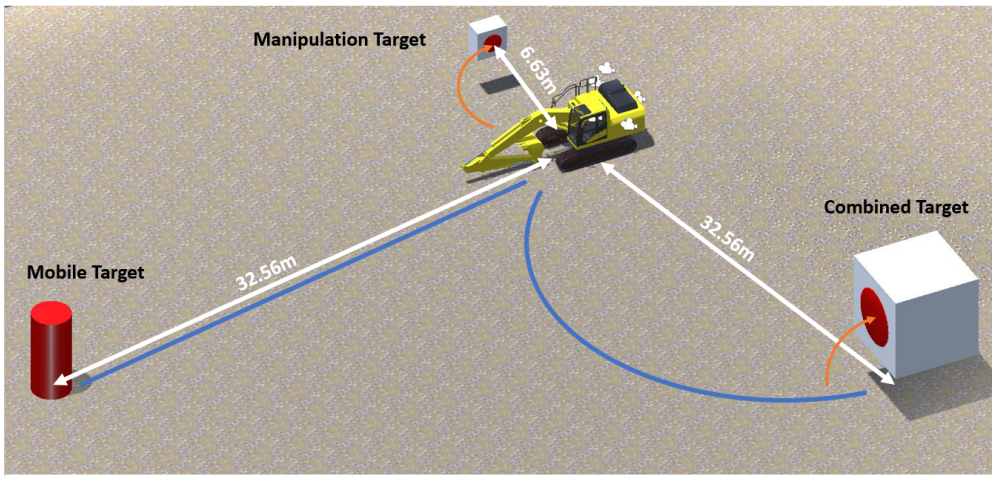
A significant advantage of WebGL is its ability to access the user's GPU directly from the browser via HTML5 elements. This hardware-level integration allows the simulation to utilise available graphical processing power efficiently, supporting high-fidelity visualisations and responsive controls. As a result, the system dynamically adapts to the capabilities of different user-end devices, with visual performance scaling according to the available computational resources.

Participants control the excavator using a standard keyboard interface. Specific keys are mapped to locomotion and manipulation functions, enabling users to intuitively perform the required tasks. This control scheme ensures accessibility while preserving the complexity necessary to evaluate skill and task difficulty effectively.

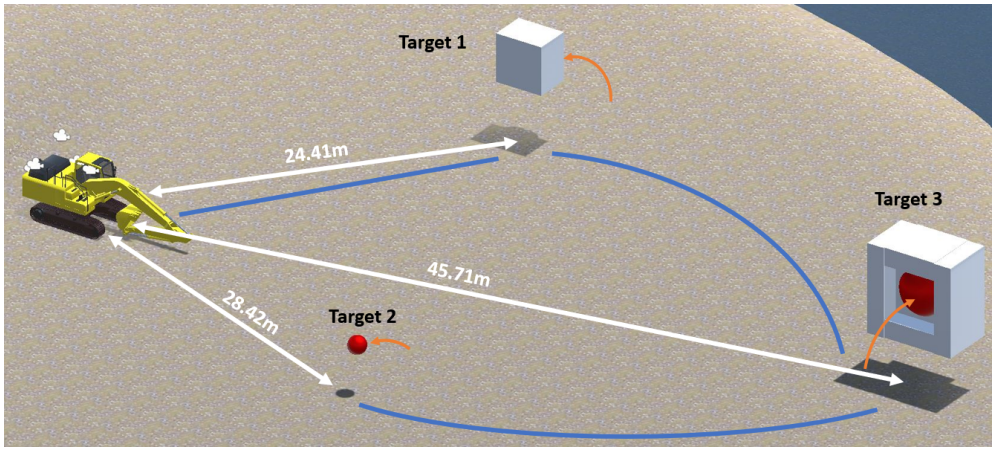
Overall, the simulation provides a robust and scalable environment for testing and validating the proposed performance prediction methodology under realistic teleoperation constraints and user variability.

The simulation environment consists of three standard tasks—$T^{sim}_{mob}$ (mobility task), $T^{sim}_{mani}$ (manipulation task), and $T^{sim}_{comb}$ (combined task)—as well as one practical task denoted as $T^{sim}_{prac}$, which is composed of multiple subtasks that emulate real-world operational sequences. The key parameters for each task are detailed in Table 3.1. This structure adheres to the standard-task modelling approach introduced in the framework, where isolated and compound capabilities are evaluated in a systematic and controlled manner.

In designing these tasks, this study accounted for the kinematic limitations of the simulated excavator. Specifically, the excavator's bucket is not actuated in the roll direction, rendering it incapable of performing rolling movements. As a result, the orientation difficulty associated with roll is inherently zero for all tasks, i.e., $\mathrm{ID}_{\mathrm{ori}}^{\mathrm{roll}} = 0$. This simplification is essential for maintaining realism in simulation fidelity, as it reflects the mechanical constraints of actual excavator hardware.




(a) Standard task targets with markers indicating trajectories




(b) Practical task targets with markers representing trajectory of target order 1-3-2

Figure 3.3: Simulation environment setup. The white lines show the relative distance to the targets. The blue and orange lines show the potential locomotion and manipulation paths.

Each standard task involves interacting with a designated spatial target. As shown in Fig. 3.3a, the system requires the excavator to make physical contact with the target to mark successful task completion. In the mobility-focused task $\mathrm{T}^{\mathrm{sim}}_{\mathrm{mob}}$, the goal is to navigate the excavator's body to reach a specified location. Here, the distance metric is calculated from the centre of the excavator's base to the centre of the target. Given that this task tests only locomotion, the approach direction is not constrained—thus, the yaw angle difference is set to $\theta_{\mathrm{yaw}} = 0$ with a full rotational tolerance of $\delta_{\mathrm{yaw}} = 360°$, indicating that the excavator can approach the target from any heading.

Additionally, since the terrain in the simulation environment is a perfectly flat plane, there is no pitch-based orientation requirement in any locomotion-related task. This results in $\theta_{\mathrm{pitch}} = 0$ for all locomotion tasks and subtasks, as there is no inclination that would otherwise introduce pitch-related difficulty. This setup isolates translational performance during locomotion and simplifies the interpretation of resulting Index of Difficulty values.

In the manipulation-focused task $\mathrm{T}^{\mathrm{sim}}_{\mathrm{mani}}$, the task involves positioning the excavator's arm to bring its end-effector into contact with the target object. The distance metric in this case is measured from the end-effector to the centre of the target, capturing the full spatial complexity of the manipulative action. Since this task emphasises precise articulation without base movement, it effectively isolates the challenges associated with the manipulator's kinematic reach and orientation.

To ensure experimental control and isolate capability assessment, the control interface is intentionally constrained in each standard task. In $\mathrm{T}^{\mathrm{sim}}_{\mathrm{mob}}$, only base movement controls are active, while arm controls are disabled. Conversely, in $\mathrm{T}^{\mathrm{sim}}_{\mathrm{mani}}$, only manipulator controls are enabled, with base movement locked. This

separation ensures that each task evaluates only the intended capability, without interference or compensatory behaviour from other subsystems.

This structured task design allows for consistent difficulty quantification and serves as a foundation for evaluating how well the proposed framework predicts performance in the subsequent practical task $T_{prac}^{sim}$, where both locomotion and manipulation capabilities must be employed in sequence.

The practical task $T_{prac}^{sim}$ is composed of three sequential subtasks, each requiring the participant to navigate the excavator to a different target location and interact with the designated object using both locomotion and manipulation. These targets are spatially distributed across the environment to necessitate full-body coordination, thus providing a realistic approximation of operational tasks encountered in real-world teleoperated excavation scenarios. A visual illustration of the target arrangement and execution sequence is presented in Fig. 3.3b.

To ensure experimental rigour and enhance the generalisability of the results, participants are required to complete the three subtasks in a randomised order, determined by the system at runtime. This randomisation serves multiple purposes. First, it mitigates potential learning effects or memorisation biases that could occur if the task order were fixed across participants. Second, it maintains participant engagement by introducing variation in task flow. Third, it allows for broader coverage of possible trajectory sequences, enabling us to evaluate the robustness of the proposed performance prediction model under diverse motion patterns and difficulty combinations.

Due to the randomised execution order, the total difficulty of $T_{prac}^{sim}$ varies across trials. The computed Index of Difficulty values for the full task range from a minimum of 12.29—when the execution follows the order Target $2 \Rightarrow 3 \Rightarrow 1$—to a maximum of 16.05, corresponding to the order Target $3 \Rightarrow 2 \Rightarrow 1$. These values

reflect accumulated difficulty from the start point through all subtasks, taking into account both translational and orientational demands for each transition.

The detailed parameters, including distances, angular deviations, and tolerances from the initial position to each target, are summarised in Table 3.1. These parameters serve as input for calculating the predicted task performance using the extended Fitts' Law framework described in earlier sections. The varying levels of task difficulty across different sequences also allow us to examine how well the model generalises across complex multi-step tasks that require sustained coordination between motion subsystems.

The simulation system is designed to automatically record the duration of each task performed by the participants. This includes the execution time for each standard task, as well as the total time taken to complete the practical task $T_{prac}^{sim}$. Upon completing all tasks, participants are prompted to submit their performance records, along with demographic and background information (e.g., prior experience with robotics or teleoperation), to a cloud-based data collection server. This cloud infrastructure enables centralised storage and streamlined analysis of user data for subsequent validation of the task performance prediction model.

To increase user engagement and encourage broader participation, the simulation interface also features a real-time leaderboard. The leaderboard displays top-performing participants based on task completion times, allowing users to benchmark their performance against others. This competitive element has proven effective in motivating participants to engage more thoroughly with the simulation and strive for improved performance.

For the purpose of analysis, this study selected data from 9 participants whose submitted records met the inclusion criteria for completeness and execution in-

tegrity. The recorded data revealed that, across these participants, the randomly generated practical task sequences in $T_{prac}^{sim}$ yielded an average computed ID of 11.53. This average reflects the diverse nature of the randomly assigned target sequences and serves as a representative benchmark for evaluating the predictive power of the proposed FittsPrompt-based performance model under real-world teleoperation constraints.

### Participants

A total of 16 participants from various global locations took part in the simulation-based study. These participants were recruited online and accessed the simulation remotely through the web-based platform. Of the 16 individuals who initially engaged with the experiment, 9 participants successfully completed all required simulation tasks within the designated time constraints and without any critical execution failures. Data from these 9 participants were selected for final analysis.

The selected cohort consisted of 2 female participants, 5 male participants, and 2 individuals who chose not to disclose their gender. The participants' ages ranged from 21 to 29 years, with a mean age of 25.8 years and a standard deviation of 2.4 years, indicating a relatively young and demographically consistent group. Additionally, participants were asked to self-report their prior experience with computer games on a scale from 0 (no experience) to 9 (expert level). The average reported gaming experience across this cohort was 7.6, suggesting a generally high level of familiarity with interactive digital environments and user interfaces.

This participant's background information helps contextualise the results by providing insight into the user population's technical familiarity and motor coordination, both of which are relevant factors in assessing performance in teleoperation-based simulation tasks.

### 3.2.3 Experiment with Quadruped

To further validate the proposed task modelling and prediction framework under real-world conditions, this study conducted an additional set of experiments involving human participants operating a physical quadruped manipulator robot. This experiment focused on assessing human-in-the-loop performance in executing teleoperated robotic tasks via two distinct control interfaces: a conventional gamepad and a wearable motion capture suit (WMCS). The goal was to compare operator performance across different input modalities while evaluating the effectiveness of the task design and difficulty modelling described in Section 3.2.1.

In alignment with the methodological framework, the experiment comprised three standard tasks—$T^{exp}_{mob}$ (locomotion-only), $T^{exp}_{mani}$ (manipulation-only), and $T^{exp}_{comb}$ (combined locomotion and manipulation)—along with one practical task, $T^{exp}_{prac}$, which involved a realistic multi-step objective. All tasks were conducted with reference to the parameters detailed in Table 3.1, ensuring consistency with the difficulty modelling approach used throughout this study.

Due to mechanical limitations in the robot's hardware, its end-effector lacks roll actuation. As a result, the system inherently incurs no difficulty associated with roll movement, and the roll component of orientation difficulty is set to zero, i.e., $ID^{roll}_{ori} = 0$, for all tasks. This constraint was factored into the difficulty calculations for the standard and practical tasks.

In terms of spatial configuration, the robot's starting positions were standardised across tasks to ensure consistency in trajectory computation and difficulty modelling. For tasks involving locomotion—namely, $T^{exp}_{mob}$, $T^{exp}_{comb}$, and $T^{exp}_{prac}$—the robot began each trial from the location designated as "Start 1." For the manipulation-only task $T^{exp}_{mani}$, the robot started from a separate position labelled "Start 2," allowing for isolation of manipulation performance without the influence of pre-

ceding locomotion. Each task required the participant-controlled robot to reach and interact with predefined targets as illustrated in Fig. 3.4.

The practical task $T^{exp}_{prac}$ was designed to simulate a time-sensitive and high-stakes bomb disposal operation. The scenario involved a target labelled "EOD" (Explosive Ordnance Disposal) and consisted of three sequential subtasks. First, the operator directed the robot to walk toward the EOD site using its quadrupedal locomotion system. Upon arrival, the robot was required to use its manipulator arm to open a container placed at the site. In the final stage, the robot had to reach inside the box and carefully extract a red wire, symbolising the act of disabling a bomb.

This structured experimental design allowed us to test the task modelling framework in a complex real-world scenario with multiple execution phases. It also facilitated performance comparisons across input modalities and provided data for validating the ID model in realistic human-robot interaction settings.

Similar to the simulation setup, in the locomotion-only task $T^{exp}_{mob}$, the robot is permitted to approach the target from any direction. As such, the yaw alignment requirement is relaxed, with the angular deviation set to $\theta_{yaw} = 0°$ and a full tolerance of $\delta_{yaw} = 360°$, indicating that the robot may navigate to the target from any heading without penalty. Additionally, the experimental environment features a flat and uniform ground surface. As a result, there is no pitch-related challenge associated with approaching the targets in any locomotion task or subtask, and thus $\theta_{pitch} = 0°$ is assumed throughout all relevant stages.

However, in contrast to the simulation, the real-world experimental setup introduces physical interaction constraints that require more cautious execution. Specifically, several of the targets used in the experiment are rigid and sensitive to accidental collision. To prevent mechanical damage or unintended contact,

the operator is instructed to bring the robot to a halt at a predefined safe distance from the target, rather than making direct contact. This constraint applies particularly to the locomotion phases of combined and practical tasks, namely $\mathrm{T}^{\mathrm{exp}}_{\mathrm{comb}_{\mathrm{mob}}}$ and $\mathrm{T}^{\mathrm{exp}}_{\mathrm{prac}_{\mathrm{step1}}}$.

To model this constraint within the task difficulty framework, the robot is treated as a point mass during these subtasks, i.e., the robot's effective width is set to $w_e = 0$. This modelling choice eliminates the interaction margin normally afforded by the robot's physical footprint, thereby enforcing stricter spatial precision. Instead of requiring contact with the target object, the task is redefined as requiring the robot to reach a designated safe location in proximity to the target, thereby satisfying the objective while avoiding physical impact.

This adaptation ensures that the ID remains realistic and reflective of actual operational constraints encountered during teleoperated robot missions, particularly in scenarios involving sensitive equipment or restricted manoeuvring zones.

In this experiment, human operators are required to control a quadruped manipulator robot (RA) using two distinct teleoperation interfaces: a gamepad and a WMCS. However, due to the fundamental kinematic dissimilarity between the human body and the robotic agent—particularly in terms of joint types, motion constraints, and degrees of freedom—directly mapping human joint positions or velocities to the robot's actuators is not feasible. This mismatch necessitates the design of a structured teleoperation strategy that translates intuitive human motions into robot-executable commands.

To address this challenge, this study developed a two-layered teleoperation control framework comprising a set of high-level robot strategies. These strategies are divided into two functional groups: *trigger strategies* and *argument strategies*, as summarised in Table 3.2. The trigger strategies serve to switch between different

Table 3.2: Teleoperation strategies and teleoperation interfaces

| Robot strategies | | Interfacess | |
|---|---|---|---|
| | | Gamepad | WMCS |
| Trigger strategies | Walking trigger | LT | Right hand down |
| | Arm trigger | LB | Left hand up |
| | Gripper trigger | RT | Right hand up |
| | Homing trigger | RB | Left hand down |
| Argument strategies | Walking arguments | LS + RS | Relative foot position |
| | Arm arguments | LS + RS | Relative hand position |

control modes, such as enabling locomotion or manipulation, while the argument strategies determine the magnitude or directional parameters of the commanded motions within the active mode.

This modular control logic allows operators to fluidly switch between robot functionalities (e.g., walking, arm movement, or gripper control) and continuously adjust motion parameters without requiring complex kinematic mappings. The design prioritises intuitiveness, enabling operators to focus on task-level decision-making rather than low-level control complexities.

To ensure experimental consistency and fairness in interface comparison, both teleoperation interfaces—the gamepad and WMCS—employ the same underlying control strategy. This shared strategy architecture minimises variability between interface modalities and isolates the effect of interface type on task performance. As a result, observed differences in execution efficiency or user experience can be more confidently attributed to the interface itself rather than underlying differences in control logic.

**Trigger Strategies**

The trigger strategies are responsible for switching between high-level operational modes of the robotic system. Each trigger is mapped to a distinct control command that enables or disables a specific category of motion or behaviour. These

triggers serve as mode selectors, ensuring that only the desired subsystem is active at a given time, thus preventing conflicting commands and improving operator focus and control accuracy. The defined trigger strategies are as follows:

- *Walking Trigger:* Activates the robot's locomotion mode. When this trigger is engaged, the robot is permitted to perform walking or repositioning actions. Locomotion commands are ignored unless this mode is explicitly enabled.
- *Arm Trigger:* Enables manipulation mode. Once activated, this trigger allows the operator to control the movements of the robot's arm. Arm commands are inactive unless the system is in this designated mode, ensuring that manipulation and locomotion are mutually exclusive when necessary.
- *Gripper Trigger:* Controls the gripper mechanism located at the end of the robotic arm. When activated, the gripper initiates a closing motion and maintains its closed state until the trigger is released. This toggle-like behaviour enables discrete grasping actions during manipulation tasks.
- *Homing Trigger:* Issues a homing command to the robotic arm, returning it to a predefined home position. This function is used to reset the arm configuration, either at the beginning of a task or to recover from undesired configurations during operation.

#### Argument Strategies

Argument strategies define the continuous control parameters sent to the robot once a specific operational mode is activated via trigger strategies. These arguments determine the direction and magnitude of movement based on user input. While the WMCS collects full-body 3D motion data, allowing for six degrees of freedom (DoF) control, the gamepad interface is limited to 2D motion inputs. To

compensate for this limitation, both the left stick (LS) and right stick (RS) on the gamepad are used in tandem to provide sufficient input for multi-axis control, as illustrated in Fig. 3.5.

- *Walking Arguments:* In walking mode, the operator controls the robot's trunk velocity across three components: forward/backwards translation, left/right strafing, and rotation about the vertical axis (yaw). The argument values are linearly proportional to the joystick displacement or body motion captured by the WMCS. These commands are continuously updated and directly control the robot's base velocity in real time.

- *Arm Arguments:* When manipulation mode is activated, the operator is given control over the arm's end-effector. The arm argument inputs include linear displacement forward/backwards, vertical movement up/down, and rotational movement of the arm's base joint (yaw). These controls enable intuitive Cartesian motion of the end-effector, allowing the operator to position the tool with precision. On the gamepad, this is achieved by assigning motion axes to joystick components, whereas in WMCS mode, body posture and limb movement are used to derive the same arguments.

This dual-interface design ensures that both control modalities—despite their different input mechanisms—deliver equivalent motion commands to the robot. This consistency in control logic allows for fair comparison in performance analysis between the WMCS and gamepad interfaces.

**Experiment hardware**

The robotic platform used in this experiment comprises a legged-base mobile manipulator designed for high-mobility teleoperation tasks. The mobile base is a Unitree AlienGo quadruped robot, which weighs approximately 21.5 kg and

provides stable, agile locomotion over uneven terrain. Mounted on the base is a modified ViperX 300 robotic arm, weighing 2.5 kg, which serves as the manipulator component for executing precision tasks [141].

Participants interacted with the robot using two distinct control interfaces: a Logitech F710 wireless gamepad and a Noitom Perception Neuron-based WMCS. These interfaces were selected to evaluate the usability and effectiveness of traditional input devices versus immersive, full-body teleoperation systems. To ensure fair comparison and consistency, both interfaces employed the same unified set of teleoperation strategies. This strategy set defines both trigger and argument control layers, ensuring that differences in performance can be attributed to the interface modality rather than variations in control logic.

The robot's base locomotion is controlled using velocity commands, allowing for continuous adjustment of walking speed and direction in real time. In contrast, the robot arm operates under position control, enabling precise placement of the end-effector during manipulation tasks. This hybrid control architecture supports fine-grained manipulation while maintaining responsive and stable mobility.

For a comprehensive overview of the teleoperation strategy architecture and control logic, readers are referred to our prior work in [3].

### Participants

To minimise bias introduced by individual differences in experience, skill level, or familiarity with robotic systems, the experiment involved multiple participants. This approach enables a more representative evaluation of HMT performance within the intended user demographic and strengthens the generalisability of the findings.

A total of 7 participants (comprising 3 females and 4 males) voluntarily partici-

pated in the experiment. Their ages ranged from 21 to 28 years, with a mean age of 26.3 years and a standard deviation of 4.2 years. Participants were also asked to self-report their prior experience with robots on a scale from 0 (no experience) to 5 (expert level), resulting in an average reported experience of 3.57, indicating moderate familiarity with robotic systems. The detailed background information is in Appendix.B.1.

To mitigate learning effects and reduce potential bias in the evaluation, participants were provided with only a brief instruction session prior to task execution and were not given opportunities for extended practice. This design choice ensured that performance reflected intuitive control and natural interaction rather than rehearsed proficiency. Furthermore, to eliminate order effects between interface conditions, the sequence in which participants used the gamepad and WMCS was randomised.

This experimental setup allows for robust comparative analysis of user performance across control interfaces, while maintaining ecological validity and fairness in participant evaluation.

### 3.2.4 Basic Training

To ensure a baseline level of familiarity and minimise initial user confusion, each participant underwent a structured training process prior to commencing the experimental tasks. The training was designed to introduce participants to the control interfaces and mission objectives while avoiding overfitting through extended practice, thus preserving the integrity of the comparative analysis.

The training session began with a brief demonstration video, which showcased real-world teleoperation exercises performed by an experienced user operating the robot through the WMCS. This visual overview provided participants with a clear

understanding of the robot's capabilities, movement dynamics, and the type of tasks they would be expected to perform. The video emphasised both locomotion and manipulation actions, illustrating the fluid switching between control modes and the appropriate physical motions used to command the robot via the WMCS.

Following the demonstration, participants received step-by-step verbal instructions on how to operate both teleoperation interfaces—namely, the gamepad and the WMCS. For the gamepad, the training covered joystick mappings, trigger functions, and the procedures for activating walking, arm control, and gripper modes. Similarly, for the WMCS, participants were instructed on how body movements and gestures correspond to robot actions, including walking directions, arm positioning, and interaction commands. This training aimed to establish a mental model of the control logic across both interfaces.

To reinforce learning and provide ongoing reference, each participant was also given a printed control diagram that outlined the command structure for both interfaces. These physical instruction sheets, illustrated in Fig. 3.5, served as quick-access visual aids during the missions. By offering participants tangible materials, the design supported memory retention and reduced cognitive load during high-demand phases of the task.

Importantly, no hands-on practice trials were provided before the official data collection began. This decision was made to maintain fairness across participants with differing levels of prior experience and to prevent learning bias from influencing the standard task performance. Instead, the training focused on cognitive familiarisation, ensuring that all participants had sufficient theoretical understanding of the system and its operation without gaining a motor advantage through repetition.

Overall, the basic training process was structured to balance clarity and neutral-

ity. It provided enough information to enable informed interaction with the robot system while preserving the authenticity of participant performance during the subsequent experimental trials.

### 3.2.5 Experiment Performing

After completing the training phase and confirming that users had acquired a sufficient understanding of the teleoperation strategies and mission requirements, participants proceeded to the experimental phase of the study. To minimise potential bias introduced by the learning curve associated with repeated task execution, a counterbalanced experimental design was employed to evaluate the two human-robot teleoperation interfaces: the gamepad and the WMCS.

Specifically, the ten participants were divided into two groups. Five randomly selected participants were instructed to perform all experimental procedures—including the standardised tasks and the full real-world mission—using the gamepad interface first, followed by the same sequence using the WMCS. The remaining five participants executed the experimental sequence in the reverse order, starting with the WMCS and then switching to the gamepad. This design ensures that any improvements in user performance due to task repetition or growing familiarity with the robot and environment are evenly distributed across both interface conditions. The procedural flow for both groups is illustrated in Fig. 3.5, which highlights the consistent application of teleoperation strategies across interface types.

During task execution, users stood in close proximity to the robot, allowing them to observe its actions and maintain spatial awareness. Participants were permitted to walk around the robot to gain a better viewpoint or adjust their perspective during control, provided they did not interfere with the robot's operational

space or trajectory. This arrangement allowed for naturalistic user behaviour, replicating how operators might reposition themselves in real-world teleoperation scenarios.

To ensure participant safety and task integrity, experiment facilitators monitored user positioning throughout each trial. Participants were advised to remain outside of the robot's projected motion path, particularly during dynamic locomotion and arm manipulation phases. Any encroachment into the robot's movement zone was gently corrected to avoid unintentional interference.

Notably, no time constraints were imposed on task execution. Each trial was conducted without a strict time limit to reduce psychological pressure on the participants and to allow them to perform each task with a focus on precision rather than speed. This approach aligns with the study's emphasis on naturalistic task performance and ensures that the motion times recorded reflect true user-system interaction rather than time-pressured behaviour.

This controlled yet flexible procedure allowed for a fair and comprehensive evaluation of both teleoperation interfaces, while preserving the authenticity of user interaction and minimising order-related bias in the experimental results.

## 3.3 Results

The core principle of the proposed framework is based on the extended 3D Fitts' Law, which models system performance by quantifying the relationship between motion time and the inherent complexity of the motion. This study presents a comprehensive evaluation of the model's predictive capability using data collected from both the simulation environment and real-world experiments.

Task difficulty values—expressed as ID—were computed for all standard and

practical tasks according to the formulations described in earlier sections. These computed values are summarised in Table 3.1. Correspondingly, the task execution time for each trial was recorded and analysed using MATLAB and Microsoft Excel. In the simulation environment, motion times were automatically logged by the system, whereas in the physical experiment, task durations were extracted from video recordings and manually annotated. An overview of the experiment flow and data collection pipeline is depicted in Fig. 3.6.

For the simulation study, data from 9 participants were included in the final analysis. These participants successfully completed all assigned tasks within a reasonable time frame and without system or procedural errors. The collected motion time data and predicted task difficulties for these participants are visualised in Fig. 3.6a.

In the real-world experiment, 7 volunteers who completed all four defined tasks using both control interfaces (gamepad and wearable motion capture suit) were selected for performance analysis. The interface-specific results are presented in Fig. 3.6b and Fig. 3.6c, allowing for direct comparison between input modalities.

To validate the accuracy of the proposed prediction model, the measured performance on standard tasks (i.e., tasks isolating locomotion, manipulation, or combined capabilities) was used to forecast the motion time of the practical task for each participant. The predicted times were then compared to the actual measured durations to evaluate the reliability and generalisability of the extended Fitts' Law formulation in both simulated and real-world teleoperation contexts.

### 3.3.1 Prediction

To evaluate the predictive capability of the extended 3D Fitts' Law model, this study analysed the relationship between the computed ID and the corresponding

Table 3.3: Parameters in the prediction lines

| Constant | Simulation | Experiment | |
|---|---|---|---|
| | | GP | WMCS |
| a | 3.57 | -0.12 | 13.18 |
| b | 2.60 | 7.07 | 3.96 |
| RMSE | 0.06 | 2.32 | 16.61 |

average motion time for each task. The IoD values were derived from Table 3.1, and the associated task durations were averaged across participants for both the simulation and real-world experiments.

The results are visualised in Fig. 3.7, where the average motion time for each task is plotted against its corresponding task difficulty. For both datasets, a linear regression model was fitted to capture the relationship between task complexity and execution time. This relationship is described using the predictive expression previously defined in Equation 2.1, where task time is modelled as a linear function of difficulty.

The fitted line parameters, including the regression constants and the root-mean-square deviation (RMSE) for each dataset, are summarised in Table 3.3. The RMSE serves as a quantitative measure of prediction accuracy, indicating the deviation between the predicted and actual motion times. Lower RMSE values reflect a stronger alignment with observed data and thus better predictive performance.

The linear trends observed in both simulation and experimental results affirm the effectiveness of the proposed model in capturing the relationship between task difficulty and system performance. This confirms the utility of the 3D Fitts' Law formulation in teleoperated robot task analysis.

### 3.3.2 Verification

To assess the practical utility and accuracy of the proposed 3D Fitts' Law-based prediction model, this study used the fitted prediction line to estimate the execution time for the practical tasks in both the simulation and real-world experiment settings. The predicted times were then compared against the actual measured average times to quantify the prediction error.

In the simulation setting, the average ID for the practical task was calculated to be 13.42. Using the fitted prediction line (as defined in Equation 2.1), the predicted motion time was 38.42 seconds. The measured average time taken by participants to complete the practical task was 43.7 seconds. This yields a prediction error of approximately 12%, indicating a strong alignment between predicted and observed performance in the simulated environment.

In the real-world experiment, task completion times were predicted separately for each control interface. For the gamepad, the task was assigned an IoD of 13.63, resulting in a predicted time of 96.19 seconds compared to an actual average of 128.6 seconds, corresponding to a prediction error of 25%. For the WMCS interface, the same task difficulty yielded a predicted time of 67.16 seconds, while the measured average was 81.3 seconds, resulting in a prediction error of 17%. The results are summarised in Table 3.4.

Table 3.4: Predicted vs. actual task completion times for different interfaces.

| **Interface** | **IoD** | **Predicted (s)** | **Actual (s)** | **Error (%)** |
|---|---|---|---|---|
| Gamepad | 13.63 | 96.19 | 128.6 | 25 |
| WMCS | 13.63 | 67.16 | 81.3 | 17 |

These results demonstrate that while the prediction model maintains high accuracy in the simulation environment, slightly larger deviations occur in real-world settings, likely due to increased variability in human input, environmental dis-

turbances, and mechanical latency. Nevertheless, the prediction errors remain within an acceptable range, confirming the effectiveness and applicability of the proposed difficulty-based performance prediction framework across both virtual and physical teleoperation contexts.

To further evaluate the distinctiveness and effectiveness of the proposed model, this study conducted a comparative analysis against several well-established Fitts-based prediction methods under identical task conditions in both the simulation and real-world experiments. These prior models were originally designed for human-computer interaction scenarios, particularly cursor-based pointing tasks, and include the formulations by Ware and Mikaelian [66], Stølen and Akin [25], and Kulik et al. [64].

The results of this comparison revealed that the traditional models exhibited substantially higher prediction errors when applied to mobile manipulator scenarios. Specifically, the model by Ware and Mikaelian produced an average error of 168%, the Stølen and Akin model 82%, and the Kulik et al. model 72%. These significant deviations highlight the inadequacy of conventional Fitts-based models in handling the complexities of mobile robotic systems.

The primary reason for their poor performance lies in their foundational assumptions, which are deeply rooted in 2D cursor-based control paradigms. As discussed in Section 2.4, these models do not account for the coupled translational and rotational dynamics inherent to mobile manipulators. Consequently, they fail to capture the true task complexity and spatial constraints that arise during real-world robot control and teleoperation.

These findings underscore the necessity of the proposed 3D Fitts' Law extension, which explicitly models both translation and orientation across six degrees of freedom. The model provides a more accurate and robust framework for predict-

ing task performance in robotic and machinery applications, particularly those involving mobile manipulators operating in unstructured or task-constrained environments.

## 3.4 Discussion

The results of the study affirm the efficacy of the proposed 3D Fitts' Law-based method in predicting HMT performance, with an error margin consistently maintained within 25%. This level of accuracy holds across both heavy machinery simulations and real-world robotic experiments, underscoring the robustness and generalisability of the model in diverse teleoperation contexts.

One key observation that illustrates the limitations of traditional models is the overestimation of rotational difficulty. For instance, in the method proposed by Stølen and Akin [25], the cursor rotation is treated as a single discrete action with a disproportionately high impact on total task difficulty. However, in real-world robotic teleoperation, operators often adjust the robot's orientation continuously during locomotion. As a result, orientation corrections are inherently integrated into the movement process and do not contribute as significantly to the overall motion time as predicted by cursor-based models.

A significant advantage of the proposed method lies in its ability to consider the HMT system as an integrated whole. Instead of isolating factors such as interface type, mechanical design, or operator variability, the framework encapsulates all related influences within the performance of standard tasks. These standard tasks serve as implicit representations of system-level characteristics, capturing latent variables that are often difficult to model explicitly.

For example, in the real-world experiments, this study observed greater variabil-

ity in task execution time during manipulation tasks when participants used the gamepad, as visualised in Fig. 3.6. Participant feedback corroborated this observation, with several users reporting difficulty achieving fine-grained control using the gamepad interface. Interestingly, although none of the participants had prior experience with the WMCS, it consistently outperformed the gamepad in both prediction and validation phases—reflected by a lower skill coefficient ($b$ value) in Table 3.3 and a smaller prediction error in Fig. 3.7.

This suggests that certain interface-related factors, such as usability, intuitiveness, and learning curve, which are inherently difficult to quantify or model directly, are nevertheless embedded in the empirical performance data captured by the standard tasks. The model effectively absorbs these hidden variables through its calibration process, providing a more holistic and adaptable framework for performance prediction in teleoperated robotic systems.

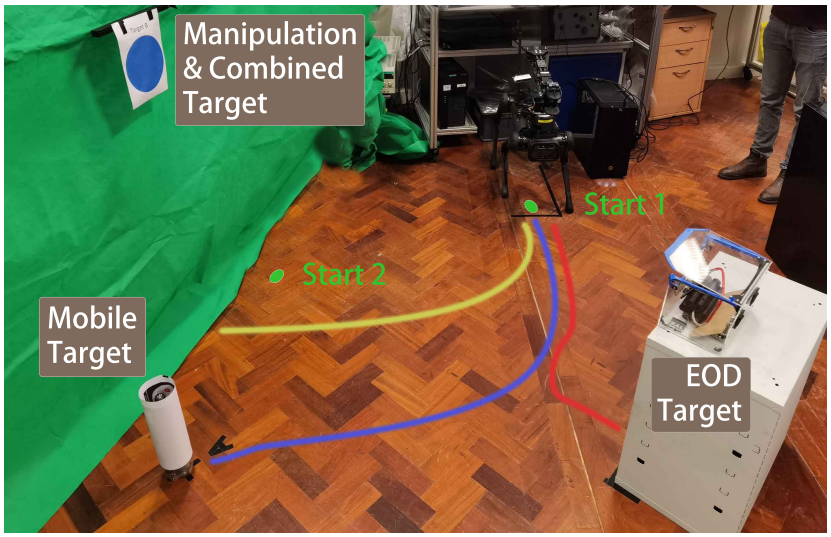
The RMSE values presented in Table 3.3 reveal that the prediction model exhibits a stronger correlation with the simulation data than with the experimental data. This discrepancy is primarily attributed to the increased disturbances and uncontrolled variables present in real-world environments, which are inherently absent in simulation. For example, in experimental locomotion tasks, the robot's stopping distances were observed to vary depending on subtle shifts in its center of mass, which were influenced by the arm's position at the time of movement. These dynamic changes introduce variability into motion execution that cannot be easily predicted or accounted for in a simulation environment where the physical interactions are idealised. This observation reinforces the sensitivity of the model to the specific mechanical configuration of the robot and its environment, and suggests that factors such as the type of mobility platform—legged versus wheeled—can significantly influence prediction accuracy. Legged robots,

for instance, often exhibit greater oscillations and micro-adjustments during locomotion, which may affect timing outcomes in ways that differ from smoother wheeled motion.

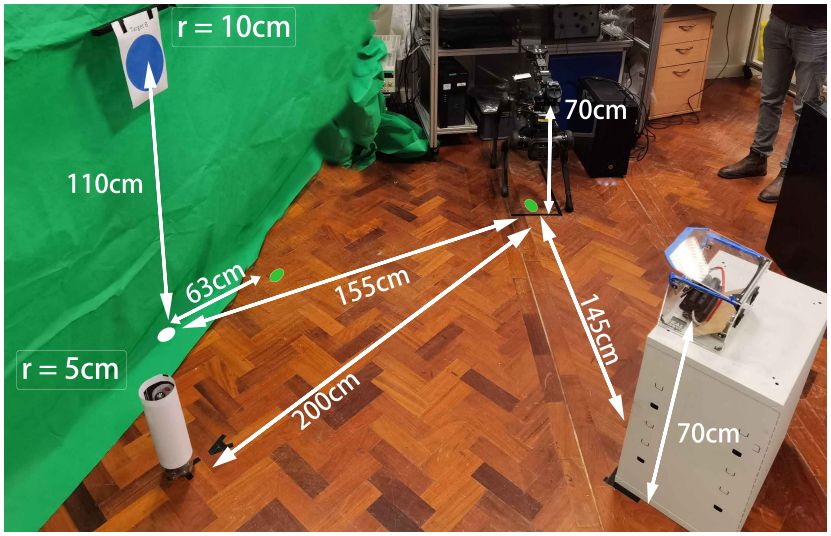
Additionally, participant variability in interface familiarity played a significant role in the observed results. Participants who had limited or no prior experience using gamepads tended to perform more poorly, particularly during fine manipulation tasks. This disparity in user proficiency impacted overall task execution time and contributed to increased variance in the experimental data. Such findings imply that in the context of HMT system design, interface selection should consider the target user demographic. Interfaces such as WMCS, which may offer more intuitive control through natural body movement, could be more suitable for novice users or operators without prior gaming or joystick experience. On the other hand, experienced users may benefit more from conventional input devices like gamepads due to their familiarity with joystick-based control paradigms.

An additional noteworthy trend was that the predicted motion time for the practical task was consistently shorter than the actual motion time recorded from users in both the simulation and experimental settings. This discrepancy likely results from a learning effect, whereby participants gradually became more adept at controlling the robot and understanding the task structure through repeated exposure to the standard tasks. As a result, by the time participants reached the practical task, their skill levels had improved beyond what was initially measured in standard tasks. This performance gain, while beneficial for real-world execution, introduces a divergence between predicted and actual times, since the model uses initial task data as a baseline for performance prediction. It underscores the importance of factoring learning curves into predictive models, especially in scenarios involving novice users or tasks that require sequential skill acquisition.

Furthermore, qualitative feedback from participants indicated certain limitations in the simulation interface that may have affected task performance. Specifically, several users reported difficulty in estimating distances to target objects due to constraints in the camera view provided within the simulation environment. This visual limitation impaired depth perception and spatial awareness, potentially causing errors in manipulation and locomotion alignment. Such feedback draws attention to an often-overlooked aspect of HMT systems—the design and fidelity of visual feedback. Enhancing camera systems, integrating dynamic view controls, or incorporating augmented visual aids could substantially improve user perception and task accuracy. Overall, these insights emphasise the importance of holistic interface and feedback system design in the broader context of HMT performance and model validation.



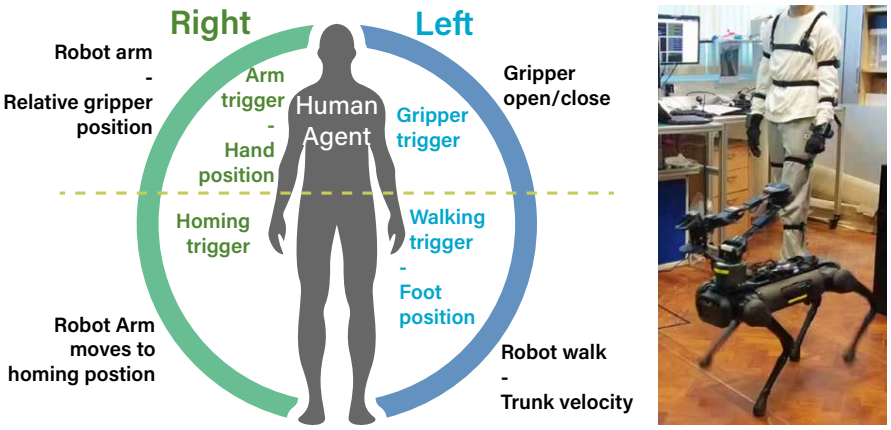

(a) Target and potential paths




(b) Parameter of the setup

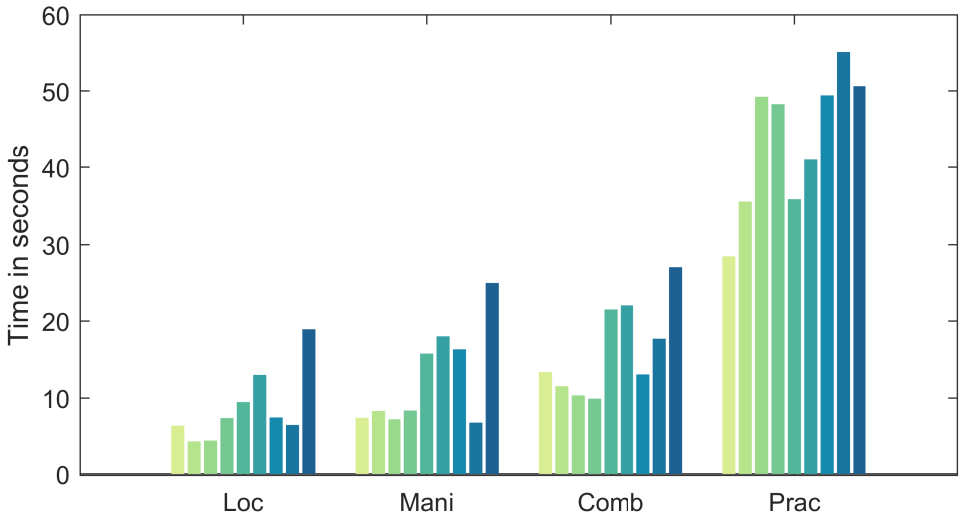
Figure 3.4: Experiment setup: In (a), blue, yellow, and red trajectories represent the locomotion path of $T^{exp}_{mob}$, $T^{exp}_{comb}$, and $T^{exp}_{prac}$. In (b), the white lines show the relative positions.


(a) Gamepad

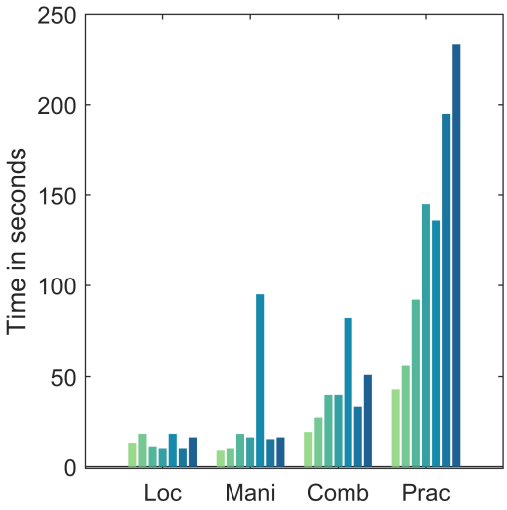



(b) WMCS

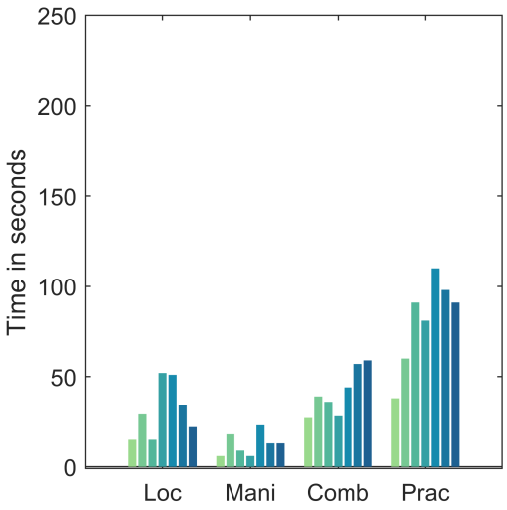
Figure 3.5: Details of mapping from interfaces to trigger and argument strategies, and experiment operation example: (a) gamepad, (b) WMCS. For the WMCS, each trigger is active by the user closing his/her hand.




(a) Simulation

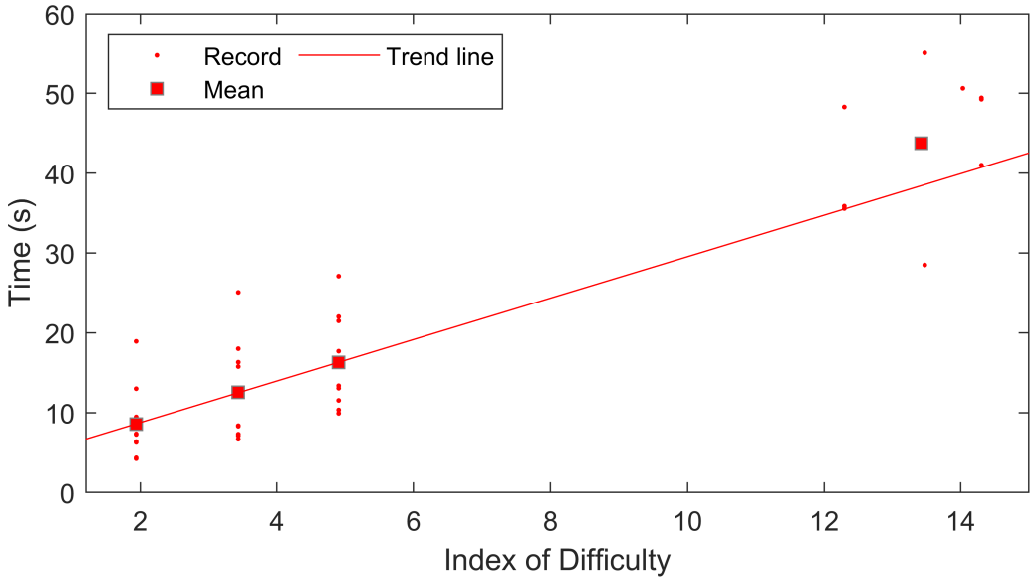



(b) Gamepad

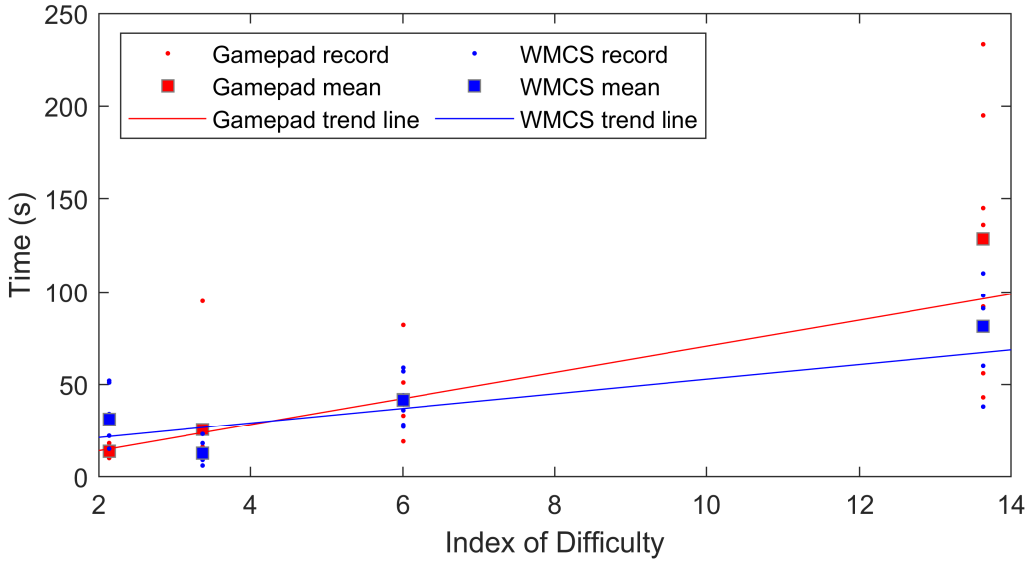



(c) Wearable motion capture suit

Figure 3.6: Motion time the users took to complete the standard tasks and practical tasks in (a) simulation and in the experiment with the (b) gamepad and (c) WMCS, where each bar represents time for a participant.




(a) Simulation




(b) Experiment

Figure 3.7: The motion time users took to complete each task versus different ID values from Table 3.1. The fitted prediction line to the average motion time of standard tasks, with $a$ and $b$ values in Table3.3.

# Chapter 4

# Predicting Performance Based on Cognitive Fatigue

## 4.1 Methodology

This chapter presents the integrated modelling framework developed to predict operator performance in complex missions, where both cognitive and physical factors play critical roles. The methodology combines cognitive state estimation based on the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model [36] and task demand modelling [116] with an extended 3D Fitts' Law [1], [35] to assess task difficulty and operator skill level. These two components are then fused to capture the dynamic interaction between an operator's cognitive effectiveness and their capacity to execute physically or mentally demanding tasks, as shown in Fig. 4.1.

The core idea of the framework is to provide a continuous and adaptive prediction of system performance by accounting for both the human operator's physiological condition and the task's complexity. The model allows for mission-time predic-

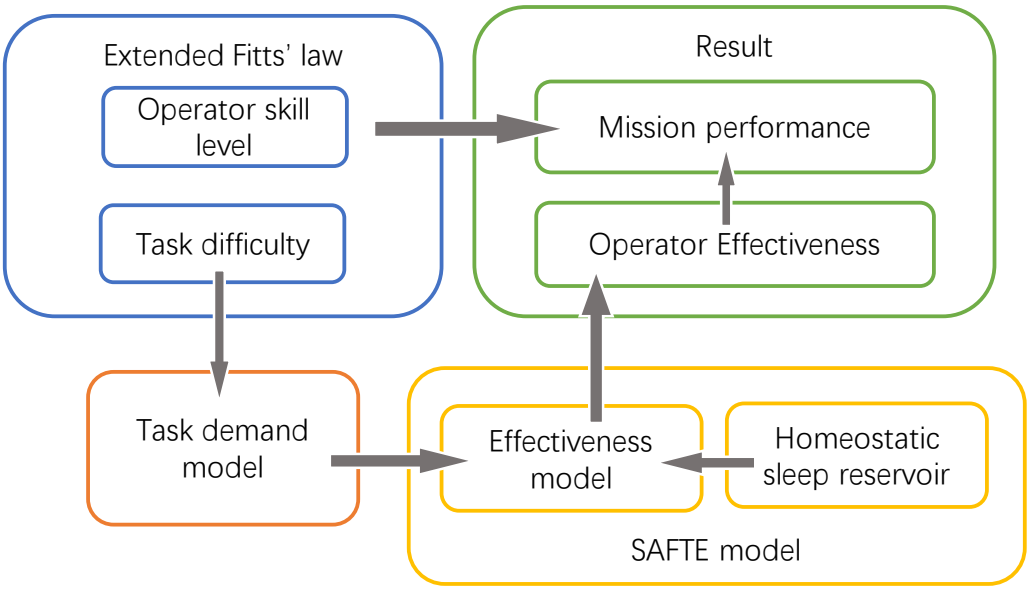



Figure 4.1: Structure of the prediction model.

tion, as detailed in Equation. 2.1, with operator fatigue, circadian rhythms, and task demands affecting mission execution capability. The integration of cognitive effectiveness into the task performance model enables more accurate forecasting of outcomes, informed operator scheduling, and human-machine decision-making strategies in high-risk or extended missions.

The overall structure of the model is illustrated in Figure 4.2, which maps out the dependencies and interactions between the main components. Each component is linked to a corresponding mathematical formulation described in subsequent sections. The model outputs—cognitive effectiveness, operator performance index, and task completion estimates—are highlighted in green, indicating their role in influencing operational decisions or system-level planning.

### 4.1.1 Modelling Cognitive Effectiveness

Cognitive effectiveness, denoted by $E(t)$, serves as a dynamic indicator of an operator's readiness and mental capability to perform tasks at any given time

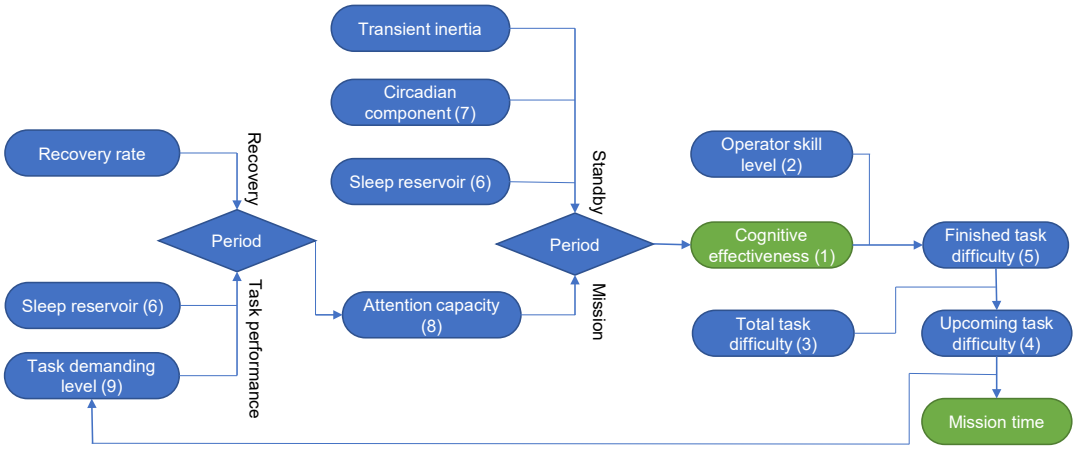



Figure 4.2: Relationship of components with corresponding equation numbers in the model. Output Components are marked in green.

$t$. The proposed model is based on the SAFTE model [36], which incorporates circadian rhythms, sleep pressure, and sleep inertia to estimate human fatigue and cognitive performance. This study adapts the SAFTE model to focus solely on waking periods, as mission-critical operations occur during these intervals. This study further divides waking time into two distinct phases: the *standby* period, which begins immediately after awakening, and the *mission* period, during which the operator is actively engaged in performing tasks.

During the standby period, cognitive effectiveness is primarily a function of circadian effects, residual sleep inertia, and sleep reservoir status. During the mission period, cognitive effectiveness progressively declines as a function of accumulated workload and attentional demand. The overall formulation for $E(t)$ is given by:

$$E(t) = \begin{cases} R(t)/R_{\max} + C_t - I_t, & \text{during standby} \\ E(t-1) - \frac{W_{\max} - W(t)}{W_{\max}}, & \text{during mission} \end{cases} . \tag{4.1}$$

In this equation:

- $R(t)/R_{\max}$ represents the normalised sleep reservoir, which quantifies how

rested the operator is. The reservoir model is described in detail in Equation (4.6).

- $C_t$ captures the circadian rhythm component, which reflects the biological clock's modulation of alertness throughout the day, as discussed in Equation (4.7).
- $I_t$ denotes the transient sleep inertia immediately following awakening, reducing cognitive performance temporarily (see Section 4.1.3 for details).
- $W(t)$ denotes the current attentional capacity during mission engagement, and $W_{\max} = 75$ is the maximum theoretical attention level (explained further in Section 4.1.4).

This dual-phase structure captures how cognitive effectiveness initially rebounds following sleep but then gradually depletes under task-induced strain. The model provides a principled way to assess whether an operator is cognitively fit to undertake a mission at a specific time. For mission planning and safety-critical applications, the value of $E(t)$ can serve as a threshold parameter to accept, delay, or reassign tasks based on real-time assessments of operator state.

Moreover, this measure of effectiveness can be directly integrated with physical task performance models, as described in the following sections, to create a comprehensive human-machine teaming model that adapts to both internal (cognitive) and external (task) constraints.

### 4.1.2 Task-Based Performance Prediction

To model and predict the operator's task performance, this study adopts and extends the classical formulation of Fitts' Law [12], a well-established empirical model that relates the time required to complete a task to its inherent difficulty.

Originally developed in the context of human motor control and aimed at characterising pointing or reaching movements, Fitts' Law has since been widely applied in human factors, ergonomics, and human-robot interaction research.

In its canonical form, Fitts' Law expresses the predicted task completion time ($PT$) as a linear function of the task's Index of Difficulty ($ID$):

$$PT = a + b \cdot ID, \tag{4.2}$$

where:

- $PT$ denotes the predicted time required to complete the task,
- $ID$ represents the quantified difficulty of the task,
- $a$ is a constant offset representing baseline or reaction time, and
- $b$ is the slope of the linear relationship, reflecting how sensitive the operator's performance time is to increasing task difficulty.

The parameter $b$ plays a crucial role in the modelling framework, as it encapsulates the rate at which task time increases with difficulty. Conceptually, this study interprets $b$ as the operator skill coefficient, which serves as a proxy for the operator's proficiency or efficiency. A lower value of $b$ indicates that the operator can handle increasingly difficult tasks with only modest increases in time—suggesting higher skill—whereas a higher $b$ implies that performance deteriorates more rapidly as task difficulty grows, signaling lower skill or increased cognitive or physical strain.

This linear formulation allows for straightforward interpretation and regression-based parameter identification from empirical data. When integrated with the extended version of Fitts' Law (discussed in previous chapters), which includes

both translational and rotational motion components, the model enables nuanced performance prediction across a range of 3D tasks involving spatial positioning and orientation.

This study further extends this framework by allowing the skill coefficient $b$ to be dynamically modulated by the operator's cognitive state, as detailed in the next section. This integration bridges the gap between physical task difficulty and mental readiness, allowing us to model how factors such as fatigue, attention, and circadian rhythm influence the operator's effective skill level during mission execution.

However, the original formulation of Fitts' Law is inherently one-dimensional, designed to model simple pointing or reaching tasks along a linear axis, typically in the context of two-dimensional interfaces such as computer screens or planar workspaces. While highly effective in such constrained environments, this one-dimensional approach proves inadequate for modelling complex interactions in real-world spatial settings, particularly in robotics or teleoperation scenarios where the operator must manage multiple degrees of freedom in three-dimensional space.

To address the limitation of original Fitt's Law, the privious work in [1] extends the concept of task difficulty beyond one-dimensional pointing to encompass the full complexity of three-dimensional motion tasks. This extension enables Fitts' Law to be used in applications involving physical robot platforms, where operators control agents to perform navigation, manipulation, or interaction in unstructured and dynamic environments. In this extended formulation, task difficulty is decomposed into multiple components that reflect the multifaceted challenges faced during 3D motion.

Again, the total index of difficulty ($ID_{\text{total}}$) for a practical task consisting of

multiple motion steps can be extract from Equation.3.10 as:

$$ID_{\text{total}} = \sum_{1}^{n} ID_{\text{trans}_i} + ID_{\text{ori}_i} + ID_{\text{dir}_i}, \tag{4.3}$$

where:

- $n$ is the total number of motion steps that comprise the full task,
- $ID_{\text{trans}_i}$ is the translational index of difficulty for the $i$-th motion step, representing the spatial distance the agent must traverse,
- $ID_{\text{ori}_i}$ quantifies the difficulty of achieving the desired orientation, accounting for rotational displacements and tolerances,
- $ID_{\text{dir}_i}$ represents the directional difficulty associated with executing motion in a specific direction relative to the agent's configuration or environmental constraints.

Each of these components captures a distinct aspect of 3D movement, enabling the model to reflect not only the physical distance to the target but also the complexity of aligning orientation and dealing with constrained or awkward motion directions. This decomposition is especially valuable for evaluating multi-step tasks, where the difficulty may accumulate or vary significantly from one step to another.

For detailed formulations and examples of how each $ID$ component is computed, the reader is referred to [35], where the extended 3D Fitts' Law is applied and validated in robot teleoperation scenarios. This study adopts this multidimensional difficulty model as the foundation for coupling physical task requirements with dynamic cognitive effectiveness, enabling more realistic and adaptive performance prediction for human-machine systems operating in complex environments.

In practical task execution scenarios—especially those involving sequential operations or multi-stage missions—the cumulative difficulty of the remaining task naturally diminishes as the operator completes more motion steps. That is, the overall task becomes progressively easier to complete as time progresses and more subtasks are executed. To model this dynamic, this study defines a time-dependent metric: the remaining task difficulty or residual Index of Difficulty, denoted as $ID_{\text{left}}(t)$, which captures the portion of the total task that remains to be completed at a given time $t$.

The remaining difficulty is calculated as the difference between the total planned task difficulty ($ID_{\text{total}}$) and the cumulative difficulty of all completed motion steps up to time $t-1$, which I denote as $ID_{\text{done}}(t-1)$. Formally, this is expressed as:

$$ID_{\text{left}}(t) = ID_{\text{total}} - ID_{\text{done}}(t-1). \tag{4.4}$$

This formulation reflects the intuitive notion that each completed subtask incrementally reduces the burden of the task as a whole. It also introduces a dynamic element into the performance prediction framework, allowing us to monitor progress and forecast task completion in real-time.

To compute $ID_{\text{done}}(t)$, this study leverages the earlier definition of operator skill and cognitive effectiveness. Specifically, this study assumes that the amount of difficulty that can be completed within a given time interval $\delta t$ depends on two main factors: the operator's skill coefficient ($b$), which represents how efficiently the operator can convert effort into completed work, and the operator's current cognitive effectiveness ($E(t)$), which modulates that efficiency based on fatigue, alertness, or workload.

Accordingly, the cumulative difficulty completed up to time $t$ is given by:

$$ID_{\text{done}}(t) = \frac{1}{b} \cdot E(t) \cdot \delta t, \tag{4.5}$$

In this expression:

- $b$ is the operator skill coefficient introduced in Equation (4.2), with smaller values corresponding to higher skill,
- $E(t)$ is the operator's cognitive effectiveness at time $t$, as defined in Equation (4.1),
- $\delta t$ is the time interval under consideration (e.g., between system update ticks or control cycles).

This formulation bridges the cognitive and physical models by dynamically adjusting the operator's performance potential over time. When cognitive effectiveness is high (e.g., immediately after rest or during peak circadian phases), the operator completes more difficult units per unit time. Conversely, when the operator is fatigued or distracted, $E(t)$ decreases, thereby reducing $ID_{\text{done}}(t)$ and leaving more residual difficulty for future intervals.

Together, Equations (4.4) and (4.5) provide a powerful mechanism for real-time monitoring and forecasting of task completion, grounded in both cognitive modelling and empirical performance theory. This integration enables a deeper understanding of how human state and task dynamics co-evolve during mission execution.

Building upon this, this study implements an iterative simulation process to project forward in time and predict whether the operator will be able to complete the mission within the bounds of acceptable cognitive effectiveness. Specifically, the model continuously evaluates whether the operator's cognitive effectiveness $E(t)$ remains above a defined safety threshold during the execution of each re-

maining task segment. At each iteration, $ID_{\text{done}}(t)$ is recomputed based on the updated $E(t)$, which itself evolves according to the operator's fatigue, circadian rhythm, and attentional workload. The remaining difficulty $ID_{\text{left}}(t)$ is updated accordingly.

This iterative process continues until one of two conditions is met. The first condition occurs when $ID_{\text{left}}(t) \leq 0$, indicating that the full task has been completed. The second condition occurs when the operator's cognitive effectiveness $E(t)$ drops below a predefined safety threshold $E_{\min}$, indicating that the operator has become cognitively unfit to continue the mission safely.

Through this process, the model produces two key outputs: (1) a predicted mission duration based on the operator's current and projected cognitive state, and (2) a safety evaluation indicating whether the task is feasible given the operator's cognitive trajectory. These outputs are essential for proactive decision-making in human-machine teaming contexts, allowing systems to trigger adaptive strategies such as workload redistribution, task reassignment, or operator rest recommendations before performance degradation or mission failure occurs.

By tightly coupling cognitive modelling with a task difficulty framework, this predictive loop enables mission-level performance forecasting that accounts for both the complexity of the work and the readiness of the human operator. As such, it represents a step forward in enabling cognitively aware autonomy and resilient human-machine collaboration in extended or high-risk operations.

### 4.1.3 Awaking and Standby Period

In the context of human performance modelling, the awake and standby period is a critical phase in which the operator transitions from a sleep state to a state of readiness for mission execution. During this period, cognitive effectiveness

is primarily governed by homeostatic and circadian factors, as described in the SAFTE model [36]. One of the key mechanisms in this phase is the homeostatic sleep drive, which reflects both the accumulated need for sleep and the restorative effects of prior rest. This process is represented through the concept of a sleep reservoir, denoted as $R(t)$, which serves as a proxy for the operator's available cognitive capacity upon awakening.

When the operator awakens from a sufficient period of sleep, the sleep reservoir is assumed to be fully replenished, attaining its maximum value: $R(t_0) = R_{\max} = 2880$. This corresponds to an optimal physiological state, where the operator is well-rested and cognitively capable of performing demanding tasks. However, empirical findings in sleep science and neuroscience indicate that performance does not peak immediately after awakening. This is due to the presence of sleep inertia—a temporary decline in cognitive functioning caused by a mismatch in the reactivation rates of different brain regions. Specifically, while subcortical areas such as the brainstem activate rapidly upon waking, the prefrontal cortex, responsible for executive functions and decision-making, lags behind [142].

This short-term impairment is captured in the model by a transient inertia term, $I_t$, which can reduce performance immediately following awakening. The maximum value of this inertia is set at 0.05 and is assumed to diminish after approximately two hours. During this initial post-sleep period, even if the sleep reservoir is full, the inertia term lowers the effective cognitive readiness of the operator.

Following the dissipation of inertia, the operator enters the standby period, during which they are not yet engaged in active mission tasks but are awake and maintaining alertness. In this phase, the cognitive reservoir begins to deplete gradually due to the effects of continued wakefulness. This depletion is modelled as a linear decline, given by the following equation:

$$R(t) = R(t-1) - V_k \delta t, \tag{4.6}$$

where $V_k = 30$ hour$^{-1}$ represents the rate at which the sleep reservoir is consumed during wakefulness, and $\delta t$ is the time increment over which the model is evaluated [116].

This formulation captures the natural decline in cognitive capacity over time, even in the absence of active workload. The standby period thus serves as a preparatory interval between rest and mission execution, during which the operator's available resources begin to diminish. The values of $R(t)$ and $I_t$ during this phase are essential inputs for the subsequent prediction of task readiness and performance during the mission period, ensuring that both physiological recovery and short-term impairments are taken into account.

The circadian component, denoted as $C_t$, represents the influence of the body's internal biological clock on cognitive performance. The circadian rhythm is a fundamental physiological mechanism that regulates sleep-wake cycles, hormone release, core body temperature, and overall alertness. It plays a significant role in modulating human cognitive effectiveness over the course of a 24-hour period, independently of sleep pressure or task demand.

To capture the effect of circadian variation in the model, $C_t$ is expressed as a function of both the intrinsic circadian phase and the current state of the sleep reservoir. The formulation used in this study follows the approach introduced in [143], with additional modulation based on homeostatic sleep depletion. The circadian contribution to cognitive effectiveness at time $t$ is given by:

$$C_t = c_t \cdot (a_1 + a_2 \frac{R_{\max} - R(t)}{R_{\max}}), \tag{4.7}$$

where:

- $c_t$ is the normalized circadian arousal value at time $t$, which oscillates between $+1$ and $-1$ over a 24-hour period, capturing daily fluctuations in alertness levels [143];
- $R(t)$ is the current sleep reservoir level at time $t$;
- $R_{\max}$ is the maximum sleep reservoir capacity;
- $a_1 = 0.07$ and $a_2 = 0.05$ are empirical coefficients that define the base circadian influence and its sensitivity to homeostatic sleep loss, which are determined from user experiments without cognitive fatigue.

This equation models the idea that the impact of circadian rhythms on cognitive performance is not static but is modulated by the current level of sleep deprivation. When the sleep reservoir is near its maximum (i.e., the operator is well-rested), the term $(R_{\max} - R(t))/R_{\max}$ approaches zero, and the circadian component simplifies to $C_t \approx c_t \cdot a_1$. This corresponds to a relatively stable modulation of cognitive effectiveness due to circadian phase alone.

However, as the reservoir depletes with time awake, the term $(R_{\max} - R(t))/R_{\max}$ increases, enhancing the influence of $c_t$ on performance. This models the interaction between circadian and homeostatic effects, acknowledging that circadian low points (e.g., early morning hours) have a stronger negative effect when the operator is also fatigued.

The result is a dynamic, biologically plausible modulation of performance that captures real-world phenomena such as the "post-lunch dip" and early morning performance decline, even in the absence of task execution. Incorporating $C_t$ into the overall cognitive effectiveness model enables more accurate prediction of performance trajectories in extended operations, shift work, or around-the-clock

mission scenarios.

### 4.1.4 Mission Period

During the mission period, the operator is actively engaged in performing tasks, and cognitive resources are consumed at a variable rate depending on both the task's demand level and the operator's physiological condition. Unlike the standby period, where reservoir depletion is driven primarily by passive wakefulness, the mission period introduces task-induced cognitive load, which directly impacts attention capacity. The key variable representing the operator's cognitive readiness during this phase is the attention capacity $W(t)$, which serves as a proxy for the amount of available mental energy or bandwidth at time $t$.

Extensive research has demonstrated that cognitive workload and sleep quality jointly influence the rate at which attentional resources are depleted [99]–[103]. High-demand tasks impose greater strain on cognitive systems, resulting in faster depletion of attention capacity. In parallel, poor sleep quality or insufficient rest reduces the system's resilience to workload, thereby accelerating resource consumption even further [103], [144], [145].

To capture these interacting effects, this work models the attention capacity $W(t)$ as a time-varying function updated through a rate-based formulation. The change in attention capacity over a small time interval $\delta t$ is given by:

$$W(t) = W(t-1) + \dot{W}_t \, \delta t, \tag{4.8}$$

follows a discretised rate-based formulation. In this expression, $W(t)$ denotes the operator's attention capacity at the current time step, while $W(t-1)$ is the value from the previous step. The term $\dot{W}_t$ represents the instantaneous rate of change

of attention capacity at time $t$, which may be negative during task performance (depletion) or positive during recovery (replenishment).

The increment $\delta t$ defines the temporal resolution of the model, i.e., the size of the discrete time step over which changes are computed. In continuous time, the relationship would be written as

$$\frac{dW}{dt} = \dot{W}_t, \tag{4.9}$$

where $\dot{W}_t$ represents the instantaneous rate of change in attention capacity at time $t$, but in practice this derivative is approximated by discrete updates over intervals of length $\delta t$. For example, if $\delta t = 1$ minute, the update describes how much attention capacity changes per minute, given the current depletion or recovery rate. A sufficiently small $\delta t$ ensures that the model provides a close approximation to continuous cognitive dynamics while remaining computationally efficient. This rate is defined piecewise depending on whether the operator is actively engaged in task performance or is in a recovery phase:

$$\dot{W}_t = \begin{cases} -\frac{R_{\max}}{100+R(t)} \cdot L(t) \cdot V_d, & \text{during task performance} \\ V_r, & \text{during recovery} \end{cases}.$$

In this equation:

- $V_d = 1.14$ hour$^{-1}$ is the base depletion rate of attention during task execution, as suggested by empirical fatigue models [146],
- $L(t)$ denotes the task demand level at time $t$, representing how cognitively taxing the current task is (to be further discussed below),
- $R(t)$ is the current level of the sleep reservoir, as defined in Equation (4.6),

- $R_{\max}$ is the maximum sleep reservoir capacity,
- $V_r = 11$ hour$^{-1}$ is the recovery rate of attention capacity during resting or non-task periods [146].

The formulation captures two important dependencies. First, the depletion of attention capacity is directly proportional to task demand, $L(t)$, such that higher workload leads to faster exhaustion of cognitive resources. Second, the depletion rate is inversely related to current sleep reservoir level $R(t)$: when the operator is well-rested ($R(t)$ near $R_{\max}$), the system can better tolerate high workload, and depletion proceeds more slowly. Conversely, when the sleep reservoir is low, even moderate task demands result in steep declines in attention capacity.

This dual dependency on physiological state and task characteristics makes the model sensitive to real-world performance variability. It reflects the empirical observation that the same task may feel manageable or overwhelming depending on the operator's rest status and the time of day. During recovery intervals—such as breaks or standby between tasks—the model switches to a fixed positive recovery rate, $V_r$, allowing attention capacity to replenish toward its baseline.

This dynamic attention model enables the prediction of not just performance degradation but also potential points of failure or recovery throughout the mission, making it a valuable tool for adaptive planning and operator state management.

The task demand level at time $t$, denoted as $L(t)$, refers to the amount of cognitive attention required from the operator to effectively perform the task assigned at that time. It acts as a dynamic multiplier in the attention depletion model, modulating the rate at which cognitive resources are consumed based on the complexity of the task being executed.

In this framework, $L(t)$ is directly influenced by the difficulty of the current task segment. More specifically, tasks with higher spatial, temporal, or cognitive complexity demand greater levels of sustained attention, decision-making, and motor coordination from the operator. These demands translate into increased cognitive workload, which in turn accelerates the depletion rate of the operator's attention capacity as modelled in Equation (4.10).

The level of task demand is therefore a function of task difficulty metrics derived from the extended Fitts' Law model, such as the translation, orientation, and directional indices of difficulty. As task difficulty increases, so does $L(t)$. For instance, a high-precision manipulation task that requires the operator to align an end-effector with a narrow or moving target would be assigned a higher $L(t)$ compared to a simple forward navigation task.

While $L(t)$ may be predefined based on mission design or task type, it can also be dynamically estimated in real-time by monitoring task parameters or using performance metrics (e.g., time-on-task, error rates, or control effort). This allows the model to remain responsive to variations in task complexity and operator behaviour throughout the mission.

In summary, $L(t)$ operationalises the cognitive demand imposed by the task and serves as a critical input to the attention capacity model. By linking it to task difficulty, this study enables a unified treatment of cognitive and physical workload within the overall performance prediction framework.

Although a task may theoretically possess an unbounded level of difficulty—particularly when it becomes physically or cognitively infeasible to complete—it is neither realistic nor cognitively meaningful to assume that such tasks impose infinite mental demand on a human operator. In practice, even when a task is perceived as overwhelmingly difficult or impossible, the operator's cognitive system does

not continuously escalate its internal workload. Instead, mental effort tends to saturate or plateau, especially when the operator consciously acknowledges the limits of their capabilities or the impossibility of the task at hand.

To reflect this bounded nature of cognitive demand, this study introduces a non-linear mapping function that transforms the raw task difficulty—represented by the residual Index of Difficulty $ID_{\text{left}}(t)$—into a normalized mental workload level, denoted as $L(t)$. This mapping ensures that task demand remains within a physiologically and behaviorally plausible range, avoiding unrealistic spikes in attention depletion during infeasible or prolonged operations.

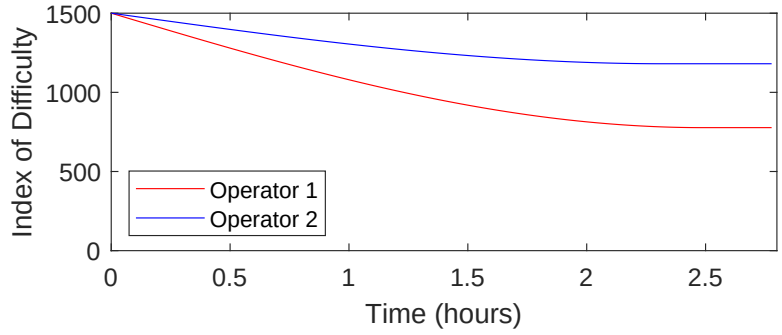
The transformation is performed using the following arctangent-based function:

$$L(t) = \frac{\tan^{-1}(ID_{\text{left}}(t)/3600)}{\pi/6}, \tag{4.10}$$

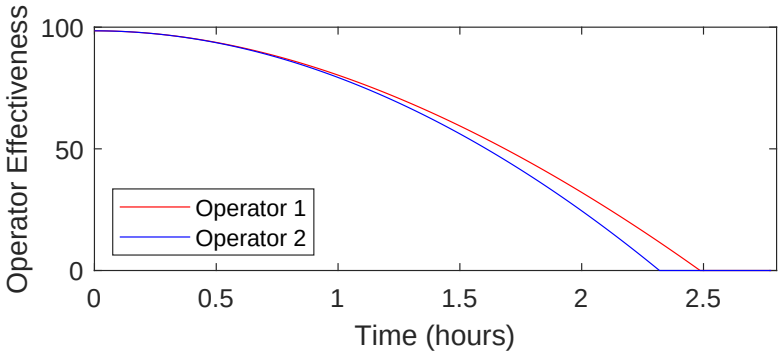
where $L(t)$ is the task demand level at time $t$, scaled to lie within the range $[0, 3]$ [116]. The use of the arctangent function ensures asymptotic behaviour as $ID_{\text{left}}(t)$ increases, which prevents the task demand from growing without bound. This reflects the cognitive reality that, beyond a certain point, increases in task difficulty no longer translate into proportionally higher perceived demand.

The division by 3600 in the input argument serves as a unit conversion factor. Since $ID_{\text{left}}(t)$ is measured in units equivalent to predicted task time in seconds—according to the extended Fitts' Law model—this normalization scales the input to a unitless form that fits consistently within the range of the arctangent function. The denominator $\pi/6$ in the overall expression scales the maximum output of the arctangent to reach an upper bound of approximately 3, aligning with established conventions in mental workload modelling literature [116].

This approach allows for a smooth, continuous mapping from objective task diffi-




(a) Index of Difficulty




(b) Operator Effectivenes

Figure 4.3: In the situation of both operators performing the whole mission continuously, the upcoming task difficulty and current operator effectiveness in task performance at the current time are shown in the graph. 0 upcoming task difficulty means all tasks have been completed, and 0 effectiveness means the operator is no longer suitable for a mission.

culty to subjective cognitive demand. It is suitable for integration with real-time models of attention capacity, fatigue, and operator state monitoring. It also provides a robust foundation for simulating operator responses to dynamic changes in task complexity throughout a mission.

## 4.2 Case Study and Results

To evaluate the feasibility and applicability of the proposed integrated cognitive-task performance prediction model, this study conducted a case study based on a hypothetical mission scenario. This scenario is constructed using parameters

derived from the prior human-subject study on mobile manipulator teleoperation, as detailed in [35]. By leveraging empirical data from this previous study, this study aims to simulate realistic operator performance and assess how the model captures differences in skill, cognitive state, and task progression.

The case study focuses on two representative human operators with contrasting levels of teleoperation experience, corresponding to the user groups identified in the prior experiment. Specifically, Group A comprised users with more extensive prior experience in teleoperating robotic systems, while Group B consisted of relatively inexperienced users who had limited exposure to such systems. From the performance data collected, this study calculated average skill coefficients ($b$ values from Fitts' Law) for each group.

Based on these group characteristics, this study defines two hypothetical operators for the simulation:

- **Operator 1** represents an experienced teleoperator, modelled using the average skill coefficient of Group A. This operator has a skill level of $b = 7.9$, indicating greater efficiency in completing tasks of varying difficulty. A lower $b$ value implies that task performance time increases more slowly with difficulty, reflecting higher operational skill.

- **Operator 2** represents a less experienced user, modelled using the average skill coefficient of Group B. This operator is characterised by a skill level of $b = 17.0$, signifying reduced proficiency and a steeper increase in task time with difficulty. This higher $b$ value reflects the performance limitations associated with inexperience, such as slower movement, higher error rates, and less effective task strategies.

These two operator profiles allow us to explore how the model differentiates be-

tween users with varying skill levels and how these differences interact with fatigue, task complexity, and time-dependent cognitive changes. By simulating the same mission for both operators under identical task conditions, this study can compare predicted mission durations, cognitive effectiveness trajectories, and attention depletion trends. The case study thereby provides insight into the model's ability to support individualised prediction and decision support in human-robot teaming scenarios.

Both operators are required to complete the same mobile manipulator teleoperation mission, which is designed to reflect a realistic multi-step operation commonly encountered in remote manipulation scenarios. The mission consists of three sequential sub-tasks that involve a combination of locomotion, perception-guided alignment, and manipulation actions. The total difficulty of the mission is set to $ID_{\text{total}} = 1500$, which is decomposed evenly across three stages, with each subtask assigned a difficulty of $ID_i = 500$. These values are determined based on the extended 3D Fitts' Law formulation, incorporating translation, orientation, and direction indices of difficulty [35].

The goal of this case study is to evaluate, under identical mission conditions, whether each operator can successfully complete the mission given their respective skill levels and initial cognitive states. Furthermore, this study aims to estimate the time required for each operator to complete the mission using the integrated model that combines cognitive effectiveness and task-based performance prediction.

The initial cognitive and physiological conditions are assumed to be ideal for both operators. Specifically, each operator begins the mission fully rested, with a sleep reservoir $R(t) = R_{\max} = 2880$, and at the start of their circadian peak ($c_t = 0.5$). The inertia term $I_t$ is set to 0.05 to reflect the transient post-awakening decline in

Table 4.1: Parameters used in the case study, with calculated values in bold. All parameters are dimensionless.

| Parameters | $ID_{\text{total}}$ | $ID_i$ | $R(t)$ | $I_t$ | $c_t$ | $\mathbf{C_t}$ | $\mathbf{E(t_0)}$ |
|---|---|---|---|---|---|---|---|
| Value | 1500 | 500 | 2880 | 0.05 | 0.5 | **0.035** | **98.5%** |

effectiveness. The circadian component $C_t$ and the initial cognitive effectiveness $E(t_0)$ are calculated using Equations (4.7) and (4.1), respectively. These values are summarised in Table 4.1.

With these initial conditions and task parameters defined, the model simulates the mission execution for both Operator 1 ($b = 7.9$) and Operator 2 ($b = 17.0$). For each operator, the simulation iteratively computes the amount of task difficulty completed per time step based on current cognitive effectiveness and skill level, updating $E(t)$ and $ID_{\text{left}}(t)$ accordingly. The simulation continues until either the mission is completed ($ID_{\text{left}}(t) \leq 0$) or the operator's cognitive effectiveness drops below a critical threshold, indicating potential failure.

This setup allows for directly comparing estimated completion times, attention capacity dynamics, and fatigue-induced limitations across operators of differing experience. The results provide insight into how skill and cognitive state jointly influence mission success, which is crucial for field commanders seeking to assign tasks based on operator readiness and mission criticality.

### 4.2.1 Standby Period

In the simulated scenario, this study assumes that both operators arrive at the workplace under favourable physiological conditions, particularly in terms of homeostatic sleep status. Specifically, both individuals are presumed to have had a full night of rest, resulting in a fully replenished sleep reservoir at the time of awakening: $R(t) = R_{\text{max}} = 2880$. This assumption reflects a best-case scenario where the operators are well-prepared to engage in cognitively demanding tasks

from a sleep and recovery perspective.

However, in realistic work environments, operators do not typically begin performing mission-critical tasks immediately upon waking. Instead, there is usually a delay between waking up and task initiation, due to activities such as commuting, briefing, or preparation. This study assumes a delay of two hours between the time of awakening and the start of the mission. This interval corresponds to the standby period, during which the operators are awake but not yet engaged in active task performance.

During this standby phase, two important physiological factors influence the operator's cognitive effectiveness. First, there is a transient inertia effect associated with the early post-wake period. As discussed in Section 4.1.3, this inertia arises from a temporary lag in the reactivation of cognitive control networks and is modelled as an additive reduction in cognitive effectiveness. For both operators, this inertia is set to $I_t = 0.05$, consistent with the expected magnitude within the first two hours post-awakening.

Second, this study accounts for the influence of circadian rhythms, which fluctuate over the course of the day. At the assumed mission start time—two hours after awakening—this study sets the circadian phase to $c_t = 0.5$, representing a moderately elevated level of circadian arousal. Using Equation (4.7), the circadian contribution to cognitive effectiveness at this time is computed as:

$$C_t = c_t \cdot \left( a_1 + a_2 \cdot \frac{R_{\max} - R(t)}{R_{\max}} \right) = 0.5 \cdot \left( 0.07 + 0.05 \cdot \frac{2880 - 2880}{2880} \right) = 0.035. \tag{4.11}$$

With all the components in place, the initial cognitive effectiveness $E(t_0)$ at the start of the mission is calculated using Equation (4.1) for the standby condition:

$$E(t_0) = \frac{R(t)}{R_{\max}} + C_t - I_t = \frac{2880}{2880} + 0.035 - 0.05 = 0.985 \text{ or } 98.5\%. \tag{4.12}$$

This value indicates that both operators begin the mission in an optimal cognitive state, with only a slight decrement from full effectiveness due to transient sleep inertia. These initial conditions are summarised in Table 4.1 and provide a strong baseline for comparing the impact of operator skill and task load during the subsequent mission period.

### 4.2.2 Executing Mission as a Whole

To evaluate how the two operators would perform under continuous working conditions, this study first applies the proposed integrated performance model to simulate a scenario in which each operator attempts to complete the entire mission in a single uninterrupted session. That is, no breaks, pauses, or task reallocations are permitted throughout the execution. This test case serves to demonstrate the effect of sustained workload and cognitive fatigue accumulation on overall task feasibility, particularly in high-demand teleoperation scenarios.

Under these conditions, the model continuously tracks both task progress—via $ID_{\text{left}}(t)$—and the operator's cognitive effectiveness—via $E(t)$—at each time step. The mission consists of a total task difficulty of $ID_{\text{total}} = 1500$, as detailed in Section 4.2. Each operator's performance rate is modulated by their skill level (via parameter $b$) and their current cognitive state (via $E(t)$). As task time progresses, attention capacity depletes and cognitive effectiveness declines accordingly, reducing the operator's ability to complete further difficulties.

The simulation results show that neither operator is able to complete the full mission when executed continuously. Both operators eventually reach a state of

complete cognitive exhaustion, defined by $E(t) = 0$, rendering them unfit for safe or effective operation. This highlights the limitations imposed by sustained workload on operator performance, even when initial conditions are optimal.

As illustrated in Figure 4.3, the total progress made by each operator plateaus before reaching the full task difficulty threshold. Figure 4.3b provides a detailed view of the cognitive effectiveness trajectory over time. For Operator 1 (the more experienced user), effectiveness reaches zero at $t = 2.49$ hours. For Operator 2 (the less experienced user), cognitive effectiveness falls to zero slightly earlier, at $t = 2.32$ hours.

These results clearly demonstrate the impact of individual skill and cognitive fatigue on mission feasibility. While Operator 1 is able to maintain effective control for a longer duration due to a lower skill coefficient ($b = 7.9$), even this advantage is insufficient to complete the mission without breaks. In contrast, Operator 2, with a higher skill coefficient ($b = 17.0$), experiences more rapid fatigue and reaches the critical threshold sooner.

This simulation validates the importance of accounting for cognitive state in mission planning and suggests that uninterrupted execution of high-difficulty missions may not be realistic or safe. The next sections explore more adaptive strategies, such as task segmentation or scheduled breaks, to improve mission success and operator sustainability.

### 4.2.3 Dividing Mission without Resting

In the third simulation approach, the mission is divided into three discrete subtasks, each with a difficulty of $ID_i = 500$, and a short rest period of 5 minutes (0.083 hours) is inserted between each sub-task. This setup simulates a practical intervention strategy whereby task planners incorporate short recovery windows

into the mission timeline to allow operators to partially restore attention capacity before resuming work. Such segmentation is often employed in operational settings to mitigate fatigue and enhance sustained performance during prolonged or high-demand activities.

The inclusion of rest breaks allows the attention capacity model to temporarily switch from depletion mode to recovery mode, governed by the recovery rate $V_r$ introduced in Section 4.1.4. This enables the operator's cognitive effectiveness $E(t)$ to rebound during the rest periods, thereby improving the likelihood of successful task completion compared to a continuous execution scenario.

As shown in Figure 4.4, this strategy significantly benefits the more experienced Operator 1. With a lower skill coefficient ($b = 7.9$), Operator 1 successfully completes all three sub-tasks with the assistance of recovery periods, reaching mission completion at $t = 3.86$ hours. More importantly, Operator 1 retains a high level of residual cognitive effectiveness at the end of the mission, with $E_{\text{end}} = 76.91\%$. This indicates that not only was the mission completed successfully, but the operator finished the task in a relatively safe and sustainable state, preserving reserve capacity for future operations or contingencies.

In contrast, Operator 2, who possesses a higher skill coefficient ($b = 17.0$) and therefore requires more time per unit of task difficulty, is unable to complete the full mission despite the added recovery periods. As illustrated in Figure 4.4, Operator 2 experiences gradual cognitive decline and eventually reaches zero effectiveness at $t = 6.45$ hours, just before completing the final sub-task. This outcome highlights the limitations of rest-based strategies for operators with lower baseline performance or in cases where task load remains excessively high over extended periods.

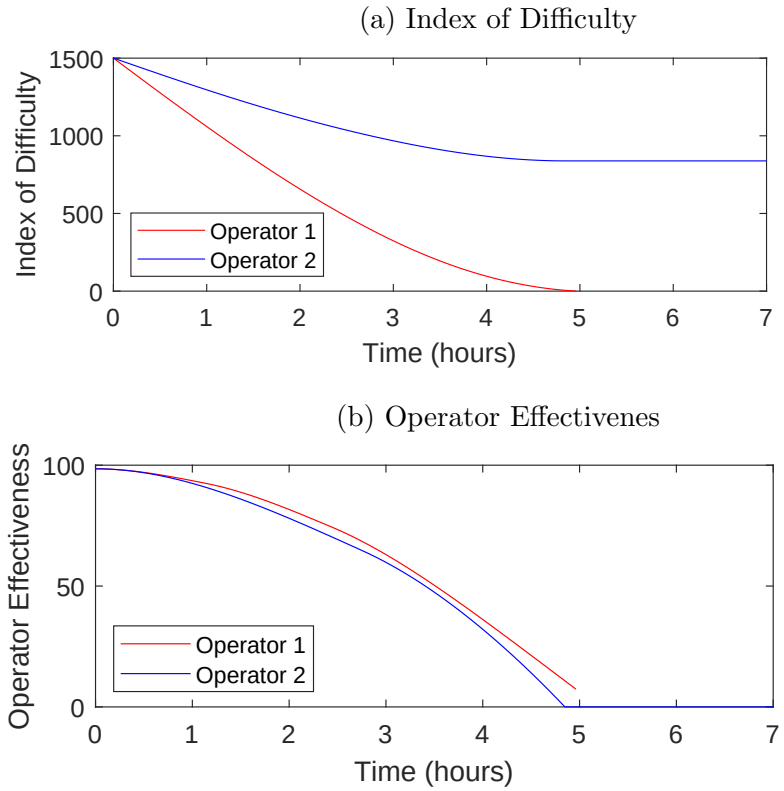



Figure 4.4: In the situation of the mission being split into 3 sub-tasks without resting time in between, the upcoming task difficulty and current operator effectiveness in task performance at the current time are shown in the graph.

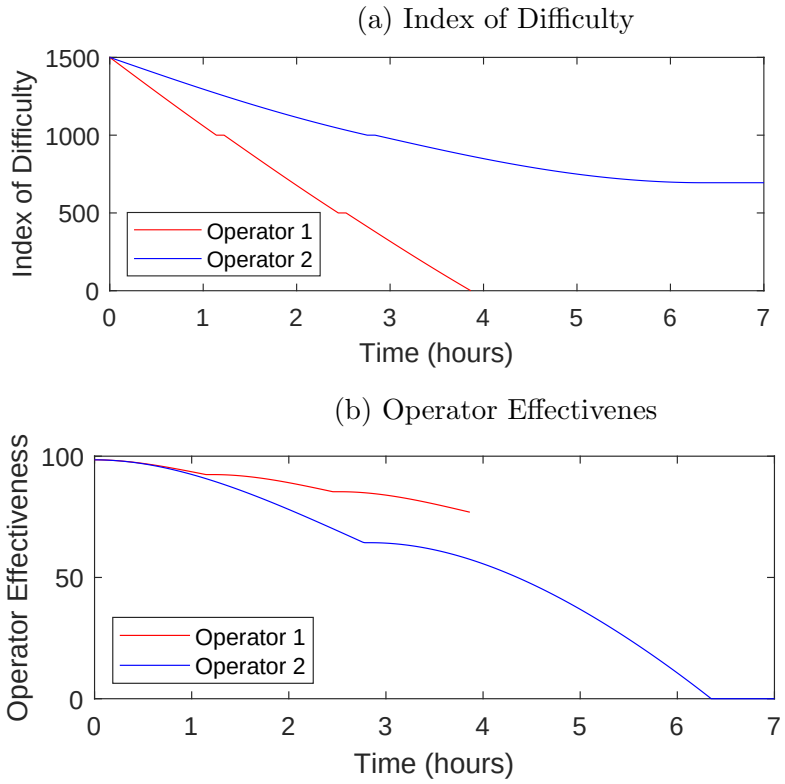



Figure 4.5: In the situation of the mission being split into 3 sub-tasks and having resting time in between, the upcoming task difficulty and current operator effectiveness in task performance at the current time are shown in the graph.

### 4.2.4 Division of Mission with Rest Periods

In the third approach, the agency splits the mission into three sub-tasks. It also gives the operators 5 minutes of rest between each subtask to recover their attention capacity, leading to a better result, as shown in Fig. 4.5. Operator 1 completes the mission at $t = 3.86$ hours, with the remaining effectiveness at $E_{\text{end}} = 76.91\%$. However, Operator 2 is still unable to finish the mission and reaches 0 effectiveness at $t = 6.45$ hours.

These results underscore the utility of integrating rest planning into mission execution for enhancing operator sustainability, but also reveal that rest alone may not fully compensate for lower skill levels or sustained high difficulty. The findings motivate the need for more adaptive strategies, such as dynamic task redistribu-

tion or predictive operator-task matching based on real-time monitoring.

## 4.3 Discussion

The comparative analysis of the three mission execution strategies—illustrated in Figures 4.3, 4.4, and 4.5—demonstrates the critical role of task structuring and rest integration in enabling successful human-machine teaming under cognitively demanding conditions. The results clearly show that partitioning a complex mission into smaller sub-tasks significantly improves mission feasibility and that supplementing this structure with scheduled recovery periods yields further gains in performance and sustainability.

The proposed model provides a mechanistic explanation for these observed outcomes. Specifically, the act of dividing a high-difficulty mission into multiple segments reduces the instantaneous task demand level, $L(t)$, associated with each segment. This leads to a corresponding decrease in the rate at which attention capacity, $W(t)$, is depleted. Because cognitive effectiveness, $E(t)$, is closely tied to attention capacity, a slower rate of depletion results in a more gradual decline in effectiveness over time. This relationship underscores the importance of managing task difficulty dynamically to preserve human operator capability throughout a mission.

This finding aligns with established principles in cognitive ergonomics and workload management. In real-world operational settings, substituting a single complex task with a sequence of smaller, more manageable subtasks reduces cognitive load. By reducing the volume of information to be processed and narrowing the scope of planning and decision-making required at any given moment, operators can better maintain situational awareness, avoid cognitive overload, and sustain higher levels of task performance.

Moreover, incorporating short rest periods between sub-tasks leverages the natural recovery dynamics modelled by the attention recovery rate $V_r$. These breaks allow attention capacity to partially regenerate, which in turn elevates or stabilises cognitive effectiveness. This intervention is particularly effective for experienced operators, as demonstrated by Operator 1's ability to complete the mission with high residual effectiveness. However, as seen with Operator 2, rest intervals alone may not be sufficient when baseline skill levels are low or when the mission difficulty remains excessive relative to the operator's capability.

These results indicate the value of intelligent mission design and operator-aware task scheduling in complex human-machine systems. Rather than treating human operators as fixed resources with static performance capabilities, the model supports a more adaptive and predictive approach—one that accounts for fatigue, skill variability, and task complexity in real time. Such insights directly apply to domains including teleoperation, remote inspection, space robotics, defence, and other time-critical or safety-critical operations where human performance limits must be proactively managed.

Figure 4.4 further highlights that, when the mission is divided into sub-tasks, Operator 1 is not only able to complete the mission but does so with a relatively high residual cognitive effectiveness of $E_{\text{end}} = 76.91\%$. This outcome demonstrates the benefit of aligning mission complexity with operator skill level, as the more experienced operator (with a lower skill coefficient $b = 7.9$) can sustain high performance under segmented execution. In contrast, Operator 2, despite benefiting from the same task structure and recovery opportunities, remains unable to complete the mission due to cumulative fatigue. Their effectiveness eventually drops to zero before the final task segment is completed. This result reflects real-world trends where operators with higher proficiency are more capable of

managing complex or extended workloads. At the same time, less experienced users may require additional support or task simplification to succeed.

The comparison between Figures 4.4 and 4.5 reinforces the additional benefit of incorporating short rest intervals between sub-tasks. These recovery periods allow attention capacity to partially regenerate before the onset of the next task, leading to improved endurance across the full mission. As shown in Figure 4.4b, the operators regain part of their cognitive reserves during the 5-minute breaks, and their effectiveness trajectories resume with a slower rate of decline after each rest period. This observed recovery dynamic supports the theoretical prediction in Section 4.1.4 that recovery, modelled through the rate $V_r$, can significantly buffer against cognitive depletion over time.

Taken together, these findings underscore a key insight: partitioning demanding missions into manageable segments, coupled with strategically placed rest periods, is an effective strategy for sustaining operator performance, especially in high-risk or long-duration scenarios. The model not only captures these effects quantitatively but also provides a planning tool that can be used to anticipate mission feasibility based on individual operator profiles. From an operational standpoint, this suggests that mission design should consider both cognitive workload distribution and personalised recovery scheduling to optimise human-machine system effectiveness.

# Chapter 5

# Evaluation Scheme for Human-Machine Team

## 5.1 Methods of Evaluation Scheme

The task difficulty modelling and performance prediction framework developed in Chapter 3 provides a robust approach for anticipating HMT performance across a diverse range of operational scenarios, as detailed in Equation.3.10. It can serve as a component for comprehensively evaluating Human-Machine Team (HMT) systems.

The proposed HMT evaluation scheme has four major components: standard tests, objective measures, a prediction model based on extended Fitts' Law [1], and subjective measures, as illustrated in Fig. 5.1. This comprehensive structure enables both quantitative and qualitative assessment of human-robot teaming interfaces (HMTIs), taking into account performance, predictability, and user experience.

As stated in Chapter. 3, the standard tests consist of four mission types: locomo-

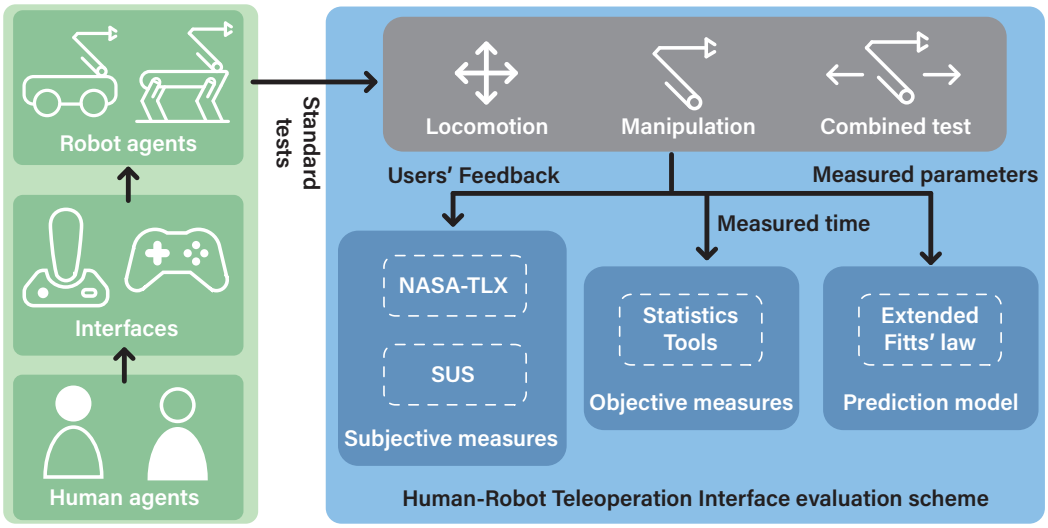



Figure 5.1: Structure of the HMTI evaluation scheme for mobile manipulator applications.

tion, manipulation, combined operations, and an Explosive Ordnance Disposal (EOD) task. These tests are designed to reflect essential capabilities in real-world teleoperation scenarios and provide a consistent benchmark for comparing different HMTIs.

The objective measures rely on time-based metrics and statistical analysis, including T-tests and ANOVA, to evaluate system efficiency and user performance variability. These measures offer a data-driven foundation to compare performance across interfaces and user groups.

The prediction model extends Fitts' Law to estimate task difficulty based on spatial characteristics of the environment and target. By fitting a linear polynomial to measured data, the model allows for motion time prediction in unseen tasks and provides insight into how different interfaces scale with difficulty.

The subjective measures include user feedback on workload and usability, collected through the NASA Task Load Index (NASA-TLX) and the System Usability Scale (SUS). These tools evaluate mental and physical demand, as well as overall user satisfaction, offering critical insights into user acceptance and prac-

tical usability of each interface.

Together, these four components form an integrated evaluation framework that captures both system-level performance and user-centred perspectives, supporting the development and selection of effective HMTIs.

### 5.1.1 Objective Measure

The quantified model provides essential data on the system's operational performance by focusing on measurable outcomes from each user trial. The primary metric used is the motion time—defined as the time taken by the user to successfully complete a given mission segment—which serves as the response variable across the evaluation. This motion time is measured with high temporal precision and verified by multiple observers using video recordings to ensure reliability and consistency.

Once the response data is collected, it is cross-compared across a set of predictor variables. These include the type of HMTI employed—such as the gamepad or the wearable motion-capture system (WMCS)—and the user's background, particularly their prior experience with gamepad devices. The comparative analysis is conducted using statistical measures such as mean, standard deviation, and inferential tests including T-tests and ANOVA. These tests reveal the presence of statistically significant performance differences between groups and conditions, thereby uncovering trends in system usability and adaptability.

Beyond motion time, additional performance indicators such as the number of attempts required to complete a task, the frequency of errors, and mission-specific metrics (e.g., successful manipulation or navigation accuracy) are also recorded. These secondary variables enrich the understanding of user interaction with each HMTI under different mission circumstances. For instance, a lower number of

attempts paired with reduced motion time may indicate both system efficiency and user proficiency.

Together, these objective measures establish a robust quantitative foundation for evaluating HMTIs, allowing the identification of both strengths and limitations in interface design and user interaction across varied task complexities.

### 5.1.2 Subjective Measures

The subjective measures analyse the usability and perceived workload of the system from the user's perspective, complementing the objective performance metrics. These measures are critical for understanding the cognitive and physical demands imposed by each HMTI, as well as the user's overall comfort, effort, and satisfaction during operation. In this evaluation, two widely accepted instruments are employed: the NASA-TLX and the SUS.

NASA-TLX is a multidimensional assessment tool developed by the National Aeronautics and Space Administration to evaluate perceived workload. It has been extensively validated across various high-demand operational settings, including aviation, medical systems, and more recently, robotics. The full NASA-TLX assesses six workload dimensions: mental demand, physical demand, temporal demand, effort, frustration, and perceived performance. However, given the short-term and mission-oriented nature of robot teleoperation tasks in this study, only the most directly relevant dimensions—mental and physical demands—are selected for analysis. These two dimensions are sufficient to capture the primary cognitive and bodily strain imposed during interaction without overburdening the participant with unnecessary complexity.

In the questionnaire, participants rate the intensity of each selected workload component on a scale from 0 (very low demand) to 100 (very high demand). A

lower score corresponds to lower perceived workload, indicating that the interface is more intuitive, less mentally taxing, or physically easier to operate. As shown in Table 5.3, average workload ratings are reported separately for locomotion, manipulation, and combined tasks, allowing for fine-grained comparison between interfaces across different mission types. This distinction is particularly important in robotic systems, where some interfaces may excel in 3D manipulation but underperform in basic navigation or locomotion due to limitations in control mapping or feedback.

The System Usability Scale (SUS) is applied to measure users' overall impressions of system usability. It consists of ten fixed statements rated on a five-point Likert scale, ranging from 1 ("Strongly disagree") to 5 ("Strongly agree"), as summarised in Table 5.4. The statements alternate between positive and negative phrasing to control for acquiescence bias and ensure a balanced assessment. While this design increases reliability, it also complicates direct interpretation. Therefore, the responses are converted into a normalised usability score, where a higher value reflects better usability.

Together, NASA-TLX and SUS offer a comprehensive perspective on the subjective experience of users, revealing not only how well an interface performs but how well it aligns with human preferences, cognitive ergonomics, and physical comfort. These insights are essential for designing intuitive and effective HMTIs that can be reliably deployed in high-stakes field scenarios.

### 5.1.3 Experiment Participation

The experiment involved voluntary participation from a group of users, each of whom contributed to the evaluation of the proposed HMTIs. Prior to the start of the experiment, each participant was required to complete a pre-test

questionnaire. This questionnaire was designed to capture baseline information about the users' demographic background, prior experience with gamepads or motion capture systems, and any previous exposure to robot teleoperation tasks. This information serves as a key factor in interpreting the variability in user performance and preference.

At the beginning of each experimental session, participants received standardised basic training. This training included a demonstration video and hands-on guidance to ensure users were familiar with the operation principles and control strategies of both HMTIs. The goal was to minimise confusion and establish a consistent knowledge base across all participants, thereby reducing bias due to uneven prior experience, as detailed in Appendix.B.1.

The experimental trials were conducted in a controlled environment where users performed a series of standard tasks using each interface. Two cameras were set up to record the entire process: one focusing on the user's interaction with the HMTI and the other on the robot's execution of tasks. This dual-angle recording enabled thorough post-hoc analysis of both human input and robotic response, and facilitated accurate measurement of motion time, task success, and other performance metrics.

Upon completion of the experimental tasks, users were asked to fill out a post-test questionnaire. This questionnaire included both the NASA-TLX and SUS forms, as discussed in the previous section, and provided a structured way to capture subjective feedback on usability, workload, and interface preferences. Informal interviews and open-ended feedback were also encouraged to gain deeper insights into the participants' experiences.

Overall, this structured participation protocol ensured a consistent experimental procedure while capturing both quantitative and qualitative data from diverse

user backgrounds, thereby supporting a comprehensive evaluation of the proposed HMTIs.

## 5.2 Result

The proposed HMTI evaluation scheme was applied to systematically assess the performance and usability of two distinct interfaces across users with different backgrounds. The evaluation aimed to capture both quantitative and qualitative differences between the interfaces, focusing on task execution efficiency, predicted performance trends, and user experience.

The results demonstrate that the choice of HMTI, as well as the user's prior experience, particularly with gamepad control, significantly influence task performance across various mission types. Notable differences were observed in motion time, task completion attempts, and perceived workload. These findings are presented through three complementary analyses: objective performance measures, prediction using extended Fitts' Law, and subjective evaluations based on standardised questionnaires. Together, they provide a comprehensive understanding of the relative strengths and limitations of each interface under different task demands.

### 5.2.1 Objective Measure

The objective evaluation relies on motion time as the primary response variable to assess system performance. This metric represents the time required for each of the ten participants to complete the designated experiment missions. As shown in Fig. 3.6, motion time data was systematically recorded and analysed using MATLAB and Excel. To ensure accuracy and consistency, three independent evaluators reviewed video recordings of each trial and manually annotated the

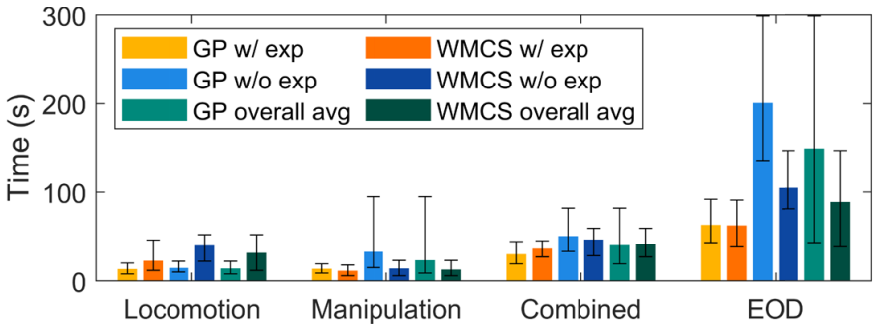



Figure 5.2: The side-by-side comparison of the motion time to complete each mission between user group A with past gamepad experience, user group B without past gamepad experience, and the total average of all the users.

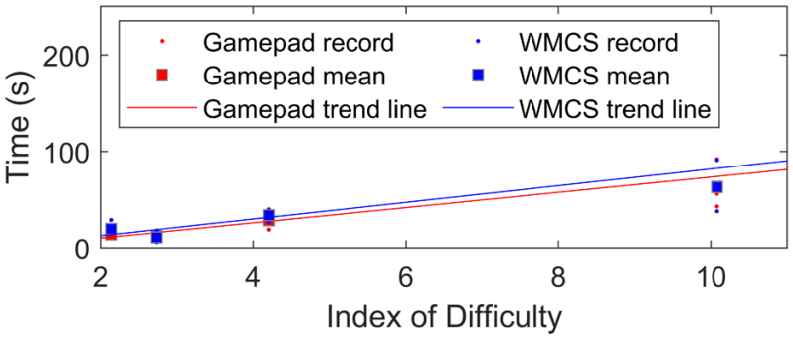
time, which was then averaged across evaluators to reduce individual bias.

The experimental design includes two key predictor variables: the type of HMTI used—either the gamepad or the wearable WMCS—and the user's prior experience with gamepads. These predictors allow for a comparative analysis across both interface types and user backgrounds. Fig. 3.6 presents the aggregated motion times for all users, offering a side-by-side comparison of performance outcomes across different mission types and interface conditions.

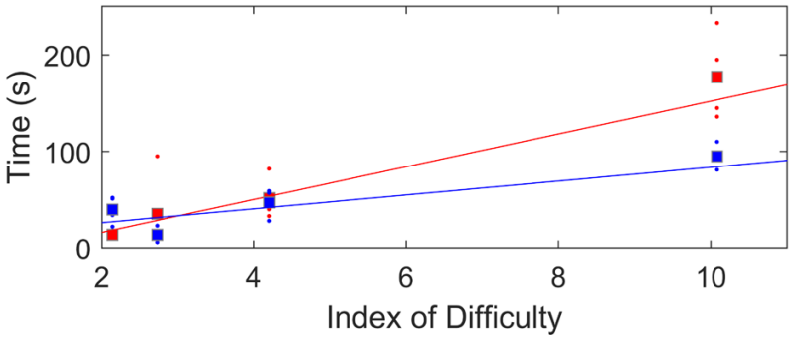
Fig. 5.2 presents the mean and range of motion time results for user groups A and B, enabling a direct comparison of performance between users with and without prior gamepad experience. This comparison highlights how user background influences task efficiency across different HMTIs. To support this analysis, statistical methods—including the calculation of mean values, standard deviations, and P-values from independent T-tests—are applied, as summarised in Table 5.1. These statistical comparisons reveal the extent and significance of performance differences between interfaces and user groups.

In addition to motion time, the number of attempts required to complete each mission is also examined as an indicator of system usability. Specifically, for the complex Explosive Ordnance Disposal (EOD) task, users required an average

(a) User group A




(b) User group B




(c) All user groups

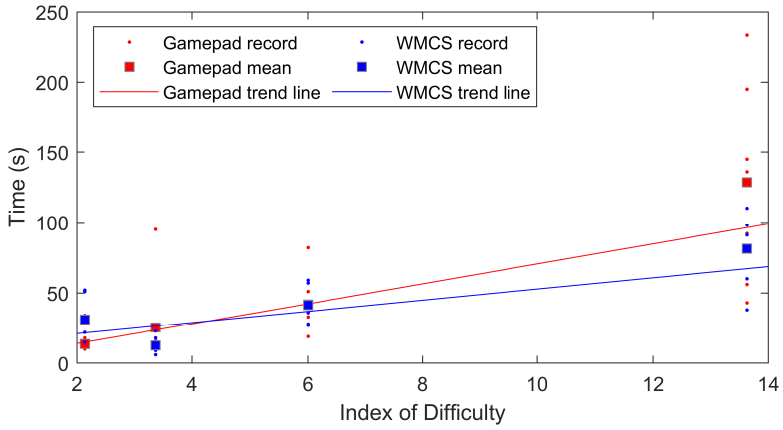



Figure 5.3: The motion time of users took to complete each mission, which is represented by different IDs, and the fitted linear polynomial line for average motion time: (a) user group A, (b) user group B, and (c) all 7 selected user. (The lower the motion time, the better performance)

Table 5.1: Statistical analysis shows the mean, standard deviation (Std), and p-value comparing two HMTIs and two groups of users. The most representative results (p-value <0.1) appear in bold, and significant results (p-value <0.05) appear in italic.

| Users | HMTI | Result | Locomotion | Manipulation | Combined | EOD |
|---|---|---|---|---|---|---|
| All users | GP | Mean | 14.60 | 23.80 | 40.70 | 149.88 |
| | | Std | 4.59 | 24.07 | 17.20 | 82.90 |
| | WMCS | Mean | 31.90 | 13.40 | 41.88 | 89.50 |
| | | Std | 14.73 | 5.22 | 11.14 | 30.41 |
| | | P-value | ***0.002*** | **0.086** | 0.364 | ***0.016*** |
| Group A (w/ exp) | GP | Mean | 14.00 | 14.40 | 30.40 | 63.67 |
| | | Std | 4.43 | 4.13 | 9.89 | 20.73 |
| | WMCS | Mean | 23.40 | 12.20 | 36.75 | 63.00 |
| | | Std | 12.75 | 4.40 | 6.50 | 21.74 |
| | | P-value | **0.050** | 0.237 | **0.089** | 0.411 |
| Group B (w/o exp) | GP | Mean | 15.20 | 33.20 | 51.00 | 201.60 |
| | | Std | 4.66 | 31.07 | 16.79 | 60.04 |
| | WMCS | Mean | 40.40 | 14.60 | 47.00 | 105.40 |
| | | Std | 11.25 | 5.68 | 12.39 | 22.84 |
| | | P-value | ***0.007*** | 0.119 | 0.379 | ***0.008*** |
| Group A VS. Group B | GP | P-value | 0.360 | 0.141 | **0.057** | ***0.005*** |
| | WMCS | P-value | **0.062** | 0.206 | **0.094** | **0.093** |
| | ALL | P-value | **0.054** | 0.102 | ***0.016*** | ***0.012*** |

of 3.4 attempts when using the gamepad, compared to only 1.8 attempts with the WMCS. This suggests that the WMCS may offer more intuitive control in high-difficulty, manipulation-heavy tasks, particularly for users with limited prior experience.

The study adopts widely accepted statistical thresholds in the field of human-robot interaction, setting a p-value of less than 0.1 as indicative of representative results and a p-value below 0.05 as statistically significant, in line with prior work [147]. As shown in Table 5.1, when considering all users, the comparison between the two HMTIs in the combined test does not yield statistically significant results. However, in all other mission categories—locomotion, manipulation, and the EOD task—representative or significant differences are observed, highlighting meaningful variations in interface performance.

Notably, the locomotion and EOD tasks exhibit statistically significant performance differences between the two HMTIs, suggesting that interface design plays a more critical role in these mission types. Further analysis of results within individual user groups reveals interesting contrasts. For example, in the locomotion test, both user groups show a performance advantage with the gamepad, but the effect is more pronounced in group B (p-value = 0.007) than in group A (p-value = 0.05). Similarly, the WMCS shows a stronger advantage in the manipulation task for group B (p-value = 0.119) compared to group A (p-value = 0.237). In the combined task, the results are reversed between the groups, and in the EOD task, the WMCS yields a statistically significant advantage for group B (p-value = 0.008), whereas no significant difference is observed for group A (p-value = 0.411).

These findings suggest that users without prior gamepad experience may benefit more from the intuitive mapping offered by the WMCS, particularly in complex manipulation or integrated tasks. Conversely, experienced users are better able to leverage the precision of the gamepad in simpler locomotion tasks, leading to nuanced performance dynamics across mission types and user backgrounds.

The standard deviation (std) provides insight into the variability of user performance across trials. As shown in Table 5.1, the gamepad generally results in more consistent performance during the locomotion test, with a notably smaller standard deviation compared to the WMCS (std = 4.59 vs. 14.73). In contrast, the WMCS demonstrates more stable performance in tasks involving manipulation (std = 5.22 vs. 24.07) and in the EOD task (std = 30.41 vs. 82.90), indicating that users were able to achieve greater consistency with the wearable system in scenarios requiring complex 3D arm movement.

When comparing across user groups, group A (users with prior gamepad experi-

ence) exhibits more minor standard deviations in most missions, reflecting a more uniform level of performance. The only exception is the locomotion task using the WMCS, where group A shows more significant variability. This suggests that while prior experience may enhance performance consistency overall, it does not necessarily translate into more stable control when using unfamiliar interfaces like the WMCS in locomotion tasks.

The performance differences between user groups operating the same interface also present notable trends. To investigate this, the study conducts independent T-tests comparing group A (users with prior gamepad experience) and group B (users without prior experience) across each standard test and the EOD task, as reported in the final three columns of Table 5.1. Results indicate a more pronounced performance gap between the two groups when using the WMCS in the locomotion test (p-value = 0.062) compared to the gamepad (p-value = 0.36). This suggests that while the WMCS offers intuitive control, its effective use in locomotion tasks may still be influenced by individual adaptability, particularly for inexperienced users.

In the more complex EOD task, a significant advantage is observed for gamepad users when using the gamepad interface (p-value = 0.005), whereas the performance gap is less pronounced with the WMCS (p-value = 0.093). This result aligns with the expectation that prior familiarity with discrete controls benefits users in tasks that require both navigation and manipulation under time pressure.

Furthermore, when all users are considered collectively, statistically significant differences between the user groups emerge in both the combined test and the EOD task. These findings indicate that task complexity amplifies performance disparities, reinforcing the importance of evaluating HMTIs across a range of mission types to uncover how user experience interacts with interface design.

### 5.2.2 Prediction Model using the Extended Fitts' Law

As part of the proposed HMTI evaluation scheme, predictive modelling is incorporated to assess how well system performance can be anticipated based on task characteristics. This component serves not only to validate existing performance data but also to enable forward-looking evaluations of HMTI suitability for tasks that have not been tested. In this study, an extended version of Fitts' Law is applied to model and predict motion time based on task difficulty, quantified through an Index of Difficulty (ID) tailored to real-world robot teleoperation.

Fitts' Law posits a logarithmic relationship between movement time and the difficulty of a task, where higher IDs are associated with longer completion times. By extending this framework with mission-specific spatial and operational parameters, the model provides a scalable and interpretable method for forecasting task performance across different interfaces and user groups.

To test the validity of this predictive model, the analysis focuses on a subset of seven users who completed all four missions using both HMTIs. These users—A1, A3, A4, and B1 through B4—were selected due to the availability of complete data, as shown in Fig. 3.6. MATLAB was used to construct linear polynomial fits between the extended ID values and the users' measured motion times for standard tests, forming the basis for predicting performance on the EOD task. This predictive analysis complements the objective measurements by offering an analytical tool to evaluate the expected behaviour of users across tasks of varying complexity.

Due to technical constraints, three participants were unable to complete all missions in the experiment. As a result, the predictive model was evaluated using data from the remaining seven users who successfully completed all four mission types—locomotion, manipulation, combined, and EOD—using both HMTIs.

These users include A1, A3, A4, and B1 through B4, as referenced in Fig. 3.6.

To construct the prediction model, MATLAB was used to fit motion time data against mission difficulty, as defined by the extended difficulty modelling. The ID values for each mission had been previously calculated based on environmental layout and target parameters, as formalised in Chapter 3, and are summarised in Table 3.1. These values formed the basis for evaluating the relationship between task complexity and user performance, enabling the prediction of motion time for future missions.

The extended ID values were then plotted against the average motion time for each mission, aggregated by user group, as shown in Fig. 5.3. The four missions—locomotion, manipulation, combined, and EOD—are presented from left to right in each graph. A linear polynomial fit was applied to the average motion times from the first three standard tests to model the relationship between ID and performance. The resulting fit was then used to predict the motion time for the EOD task, which serves as an unseen validation case. The fitted models follow the form of extended Fitts' Law, with coefficients $a$ and $b$ summarised in Table 5.2.

To begin, the model was applied to data from the three users in group A. MATLAB was used to generate linear regression curves for both the gamepad and WMCS conditions. The root-mean-square deviation (RMSE) of the fitted lines was 4.65 for the gamepad and 10.12 for the WMCS, indicating a closer fit in the gamepad condition. The predicted EOD performance also aligned with real-world observations: the model correctly projected that group A users—who had prior gamepad experience—would perform better with the gamepad than with the WMCS. These findings, illustrated in Fig. 5.3a, support the validity of the extended Fitts' Law as a predictive tool within this evaluation framework.

Table 5.2: Constant $a$ and $b$ in Fitts' Law and the difference between predicted motion time and average measured time.

| Users: | HMTI | Measured MT | a | b | Predicted MT | Difference |
|---|---|---|---|---|---|---|
| Group A | Gamepad | 63.67s | -5.504 | 7.884 | 73.98s | -16.18% |
| | WMCS | 63s | -4.513 | 8.622 | 82.40s | -30.79% |
| Group B | Gamepad | 177.25s | -17.89 | 17.00 | 153.45s | 13.43% |
| | WMCS | 95s | 11.80 | 7.178 | 84.16s | 11.42% |
| All users | Gamepad | 128.57s | -12.58 | 13.09 | 119.39s | 7.14% |
| | WMCS | 81.29s | 4.807 | 7.797 | 83.40s | -2.6% |

Next, the model was applied to the four users from group B, who lacked prior gamepad experience. Linear polynomial curves were fitted to their average motion time data for both HMTIs using MATLAB. The resulting root-mean-square deviation (RMSE) values were 8.81 for the gamepad and 22.26 for the WMCS, indicating greater variability and a less precise fit compared to group A. Notably, as shown in Fig. 5.3b, the fitted lines for the gamepad and WMCS intersect at approximately an ID value of 3.1. This crossover point suggests that for simpler tasks (lower ID), the gamepad yields better performance, while the WMCS becomes more effective as task complexity increases.

When comparing across groups, group A demonstrates consistently lower RMSE values than group B (4.65 vs. 8.81 for the gamepad, and 10.12 vs. 22.26 for the WMCS), indicating that the model fits more accurately when users have greater interface familiarity. This higher accuracy translates into more reliable predictions. For instance, the difference between the predicted and actual motion time for the EOD task is notably smaller in group B's WMCS model (11.42%) than in group A's WMCS model ($-30.79\%$), as detailed in Table 5.2. This finding highlights that while model accuracy is influenced by user consistency, prediction precision may also depend on interface adaptability to novice users.

Although users in groups A and B differed in their prior gamepad experience, they shared several other relevant characteristics, including similar levels of tech-

nical proficiency and task exposure during the study. Therefore, to evaluate the generalisability of the prediction model, all seven selected users were aggregated into a single group, and the extended Fitts' Law model was reapplied. As shown in Fig. 5.3c, linear polynomial curves were fitted to the combined dataset using MATLAB. The resulting RMSE values were 3.04 for the gamepad and 17.06 for the WMCS, suggesting improved model accuracy when aggregating across users.

Interestingly, the intersection point of the fitted lines occurs around an ID value of 3.5, indicating a performance crossover between the two HMTIs. Specifically, the gamepad outperforms the WMCS in lower-difficulty missions, whereas the WMCS demonstrates superior performance as task complexity increases. This trend aligns with earlier observations from individual groups and supports the notion that interface effectiveness is task-dependent.

Furthermore, as the sample size increases, the model's predictive reliability improves. The differences between predicted and measured motion times for both HMTIs fall below 10%, as shown in Table 5.2. This outcome reinforces the value of incorporating prediction as a component of the evaluation framework, demonstrating that the extended Fitts' Law can effectively model and anticipate performance trends across diverse user groups and mission types.

The proposed extended Fitts' Law successfully predicted the performance outcomes of different user groups across varying mission complexities. The accuracy of the prediction was found to be influenced by both the sample size and the users' prior experience with the control interfaces. Specifically, users with consistent backgrounds and familiarity yielded tighter model fits and smaller prediction errors. Importantly, a power analysis confirmed that the selected group of seven users provides sufficient statistical power (0.8) at a type I error rate of 0.1, affirming the adequacy of the sample for distinguishing performance differences

Table 5.3: Mean scores for NASA Task Load Index, on a scale of 0 to 100. (The lower score, the lower workload, marked in bold.)

| NASA-TLX | | Gamepad | WMCS |
|---|---|---|---|
| Mentally demanding: | Locomotion | 52 | **51** |
| | Manipulation | 67 | **35** |
| | Combined and EOD task | 69 | **49** |
| | Average workload | 63 | **45** |
| Physically demanding: | Locomotion | **27** | 55 |
| | Manipulation | 39 | **32** |
| | Combined and EOD task | **43** | 45 |
| | Average workload | **36** | 44 |

between the two HMTIs in the context of quadruped manipulator teleoperation.

These results underscore the utility of integrating a predictive modelling component within the evaluation framework, offering not only retrospective analysis but also forward-looking insights into system scalability and task adaptability.

### 5.2.3 Subjective Measure

Subjective measures were gathered through post-experiment questionnaires to capture user perceptions of workload and usability, complementing the objective performance data. Two standardised assessment tools were employed: the NASA-TLX, as shown in Appendix. B.2, and the SUS, as demonstrated in Appendix. B.3. These instruments provided insight into how users experienced each HMTI during different tasks, particularly in terms of cognitive effort, physical demand, and interface intuitiveness.

Overall, user preferences varied across task types. Specifically, 58% of participants preferred the gamepad for the locomotion test, while 44% favoured the WMCS for the manipulation task. Interestingly, only 28% preferred the WMCS during the complex EOD task, suggesting that interface preference may shift with task complexity and required precision.

Table 5.4: Users' average scores for System Usability Scale (on a scale of 1 (Strongly disagree) to 5 (Strongly agree)).

| No. | System Usability Scale statements | Response type | User response | | Converted score | |
|---|---|---|---|---|---|---|
| | | | Gamepad | WMCS | Gamepad | WMCS |
| 1 | I think that I would like to use this system frequently. | Positive | 3.2 | 3.4 | 2.2 | 2.4 |
| 2 | I found the system unnecessarily complex. | Negative | 2.5 | 2.6 | 2.5 | 2.4 |
| 3 | I thought the system was easy to use. | Positive | 4.3 | 2.7 | 3.3 | 1.7 |
| 4 | I think that I would need the support of a technical person to be able to use this system. | Negative | 2.1 | 4 | 2.9 | 1 |
| 5 | I found the various functions in this system were well integrated. | Positive | 3 | 3.4 | 2 | 2.4 |
| 6 | I thought there was too much inconsistency in this system. | Negative | 2.2 | 3.3 | 2.8 | 1.7 |
| 7 | I would imagine that most people would learn to use this system very quickly. | Positive | 3.3 | 3.1 | 2.3 | 2.1 |
| 8 | I found the system very cumbersome to use. | Negative | 2.7 | 2.9 | 2.3 | 2.1 |
| 9 | I felt very confident using the system. | Positive | 3.3 | 3.6 | 2.3 | 2.6 |
| 10 | I needed to learn a lot of things before I could get going with this system. | Negative | 1.7 | 2.5 | 3.3 | 2.5 |
| | (For odd-numbered questions, subtract 1 from the score. For even-numbered questions, subtract the score from 5.) | | | Total score: | 64.75 | 52.25 |

To evaluate workload, two key dimensions—mental and physical demand—were extracted from the NASA-TLX, as shown in Apeendix.B.2, and applied to each mission type. As summarised in Table 5.3, the WMCS consistently resulted in lower reported mental workload across all task types. This suggests that the spatially intuitive control mapping of the WMCS reduced users' cognitive burden. However, physical demand scores were more balanced: while the gamepad imposed less physical strain in locomotion tasks, the WMCS was perceived as physically easier during manipulation, likely due to its natural alignment with arm movements.

Usability scores derived from the SUS, as shown in Table 5.4, reflect a nuanced perspective. While many users found the WMCS more complex to operate—especially at first contact—they reported greater confidence when using it, indicating that the wearable system may offer a steeper learning curve but higher perceived control once mastered. These results highlight the importance of considering both usability and learnability when evaluating HMTIs for field deployment.

## 5.3 Discussion

This section interprets the findings from the objective, predictive, and subjective evaluations of the two HMTIs: the gamepad and the wearable WMCS. By integrating performance metrics, model predictions, and user feedback, the discussion aims to highlight key insights into the practical implications of interface design, user experience, and task complexity in teleoperated robot systems.

The results underscore the importance of selecting an appropriate HMTI based on mission demands and user profiles. Significant differences were observed in user performance and perception depending on the interface used and the type of

task performed. These differences were especially prominent in locomotion and manipulation scenarios—two fundamental operations in real-world teleoperation tasks such as search and rescue, remote maintenance, and explosive ordnance disposal. In light of these findings, the discussion explores the trade-offs between intuitive control, usability, adaptability, and learnability of each interface, and how they influence task success under varying operational conditions.

From the results presented in Table 5.1, it is evident that the most representative (p-value ¡ 0.1) and statistically significant (p-value ¡ 0.05) findings appear across the majority of mission types. These outcomes highlight notable performance differences between the two evaluated HMTIs—the gamepad and the wearable WMCS—particularly in tasks involving locomotion and manipulation. Such results are consistent with the distinct operational modalities of the two interfaces: while the gamepad offers more structured and discrete control inputs, the WMCS enables more natural and continuous movement mapping.

This distinction is highly relevant to real-world scenarios, where crisis management missions such as hazardous material handling, search and rescue, or explosive ordnance disposal typically require seamless integration of both locomotion and manipulation within a single task [29]. For instance, navigating a robot through rubble or a confined space often precedes or coincides with the manipulation of tools, objects, or hazardous components. In such high-stakes environments, the interface must not only support accurate and efficient control but also reduce operator cognitive load and facilitate intuitive interaction.

Therefore, the presence of measurable differences between HMTIs in these core operational domains underscores the importance of employing a comprehensive and multi-faceted evaluation framework—one that accounts for both performance outcomes and user experience—to inform the selection and deployment of suitable

interfaces. In particular, hybrid performance advantages suggest that combining the intuitive, spatial mapping of the WMCS with the ergonomic familiarity and accessibility of gamepads could offer a more robust solution. An HMTI that leverages the best of both systems has the potential to improve operator effectiveness and reduce the likelihood of failure during critical missions.

The control mechanisms of the two HMTIs differ fundamentally in how they translate human input into robotic motion, which directly impacts user performance across different task types. The gamepad interface, with its dual joystick configuration, primarily provides linear commands in a two-dimensional control space. This modality is well-suited for planar navigation tasks, such as forward/backwards and lateral movement during locomotion. As a result, users generally found it easier to perform locomotion tasks using the gamepad, as it aligns closely with conventional input paradigms found in video games and remote-controlled vehicles.

In contrast, the WMCS allows for position input in a three-dimensional space by directly mapping the human operator's arm and body movements to the robot's manipulator. This naturalistic mapping enables users to execute complex manipulation tasks more intuitively, particularly when the robot arm must move in all three spatial dimensions simultaneously. For example, actions such as reaching, rotating, and grasping objects could be performed more fluidly and with greater spatial awareness using the WMCS.

However, in practical trials, it was observed that users made more errors when attempting to control the manipulator with the gamepad. These errors often stemmed from the cognitive disconnect between the 2D joystick inputs and the 3D motion required for effective arm manipulation. Users had to mentally translate joystick movements into Cartesian space operations, which increased their

cognitive load and led to more frequent mistakes, particularly in tasks involving precise spatial coordination or rotational alignment. In contrast, the WMCS reduced this cognitive translation effort, allowing for more direct and embodied control over the manipulator's motion. These findings highlight the impact of control dimensionality on task performance and suggest that interface suitability is closely linked to the nature of the motion being executed.

The extended Fitts' Law model used in this study reinforces a foundational principle in human motor control: motion time increases as task difficulty rises. This observation is consistent with the original formulation of Fitts' Law [12], as well as with several of its later modifications designed for digital and embodied interaction contexts [13], [24], [25]. By introducing an extended Index of Difficulty tailored to real-world robot teleoperation scenarios, the model enabled nuanced prediction of user performance across both standardised and complex tasks.

Interestingly, empirical results show that both HMTIs—despite their differences—consistently outperformed the predicted motion times generated by the extended Fitts' Law, as shown in Table 5.2. This suggests that participants may have leveraged strategic or adaptive behaviours during the experiment, particularly in tasks requiring higher-order reasoning or environmental improvisation. For example, the fastest user shortened their task duration in the EOD scenario by pushing out the wire connector instead of pulling it, as originally instructed. Such improvisation likely reflects a combination of user familiarity with digital input devices and prior gaming experience, which may have facilitated more efficient motion planning and execution.

Another contributing factor lies in the order of task exposure. All users interacted with both interfaces in the same progression—from tasks with the lowest ID to those with the highest. Although none of the participants had prior experience

with the WMCS, they progressively improved their performance over the course of the experiment. This learning effect is reflected in the slope differences between the linear fits shown in Fig. 5.3c, where motion times using the WMCS decreased more sharply than those of the gamepad. Early-stage WMCS use was associated with more frequent mistakes, but these errors diminished in later trials, indicating a rapid increase in user proficiency through short-term adaptation.

Interviews and post-experiment feedback further support this observation. Several users reported initial difficulty coordinating trigger-based activation with gesture-driven input (referred to as "trigger-argument strategy coordination") when using the WMCS. However, as they gained more exposure, they became more confident and efficient, reinforcing the hypothesis that the WMCS may have a steeper learning curve but ultimately enables more expressive and efficient control in high-difficulty tasks. These findings validate the usefulness of the extended Fitts' Law not only as a predictive model but also as a diagnostic tool for capturing learning dynamics and adaptive behaviour across different interface modalities.

The usability results reveal a trade-off between cognitive and physical demands across the two interfaces. Specifically, the gamepad was associated with a higher mental workload but lower physical exertion, as indicated by NASA-TLX ratings. In practical terms, users occasionally encountered difficulties recalling the functions mapped to various buttons and joysticks, particularly when transitioning between locomotion and manipulation tasks. This cognitive overhead was more pronounced during high-pressure missions, where rapid response and multi-step coordination were required. As a result, the mental load associated with interface memorisation and discrete input control contributed to elevated subjective workload scores.

Conversely, the WMCS offered a more direct and embodied form of control, reducing the need for abstract button mapping. Users were able to operate the manipulator through natural arm and body movements, which made the interface more intuitive, especially for spatially complex tasks. However, this advantage came at the cost of increased physical effort. Unlike the gamepad, which relies on finger movement, the WMCS demands full-body engagement—including arm extension, rotation, and balance coordination—which can lead to fatigue during prolonged use. These physical demands were reflected in higher physical workload scores for locomotion tasks using the WMCS.

The SUS results suggest that users generally rated the gamepad as more usable overall, particularly due to its familiarity and plug-and-play simplicity. Many users had prior exposure to gamepads through gaming or other remote control applications, giving them a significant advantage in initial trials. In contrast, the WMCS required additional setup effort, including sensor calibration, wireless connection, and battery management, which introduced logistical complexity and potential points of failure. While users acknowledged the intuitive nature of motion-based control, these operational barriers contributed to lower overall usability scores in SUS evaluations.

These findings highlight an important distinction between perceived and practical usability: while the WMCS offered a more natural control scheme, its setup and maintenance requirements, combined with greater physical strain, made it less accessible for users with limited technical experience. In contrast, the gamepad—despite its steeper cognitive demands—benefited from existing user familiarity and a lightweight, portable form factor that made it more convenient for rapid deployment in field operations.

In addition to the structured questionnaires, direct message feedback and in-depth

interviews provided valuable qualitative insights into user perceptions of the two HMTIs. Several users shared specific reflections on the strengths and limitations of each interface. One participant described the gamepad as "more sensitive and user-friendly," highlighting its responsiveness and ease of use, particularly in tasks requiring quick directional input. Another user expressed a preference for the WMCS, noting that "the motion-capture suit had more straightforward controls," referring to the intuitive spatial mapping between human movement and robot action.

Some concerns were also raised regarding the reliability of the WMCS. One user pointed out an initial issue with control accuracy, which was later traced back to miscalibrated IMU sensors. After recalibration, the issue was resolved and did not affect the overall experimental results. This highlights the importance of system readiness and sensor integrity in ensuring consistent performance, particularly in wearable interface setups.

Importantly, a recurring theme emerged from multiple interviews: the potential benefit of a hybrid control system that combines the strengths of both interfaces. As one user articulated, "In an ideal world, I'd have a hybrid system with a joystick for locomotion and hand controls for the arm." This sentiment reflects the general consensus that while the gamepad excels in navigation due to its precision and simplicity, the WMCS provides superior control for manipulation through its natural motion correspondence.

This user-driven insight suggests a promising direction for future interface development. A modular system that leverages the gamepad for locomotion tasks and the WMCS for manipulator control could offer an optimal balance between usability, intuitiveness, and task-specific efficiency. Such a hybrid HMTI would not only enhance operational flexibility but also reduce training time by allowing

users to engage with each subsystem in the context where it performs best.

The comparison between the gamepad and the WMCS reveals an important discrepancy between system performance and perceived usability. While one interface may yield faster or more consistent task execution, it does not necessarily translate to higher user satisfaction or lower cognitive and physical workload. This phenomenon highlights a potential dissociation between objective performance metrics and subjective user experience, emphasising that high performance does not inherently imply high usability, and vice versa.

Such a separation suggests that evaluating only one dimension—either quantitative task efficiency or qualitative user preference—would provide an incomplete picture of system effectiveness. For instance, the gamepad may appear superior in terms of usability due to its simplicity and widespread familiarity, yet the WMCS may offer superior manipulation precision and lower mental demand in complex tasks. These contrasting outcomes underscore the necessity of incorporating both performance and usability evaluations within any comprehensive human-robot interaction assessment.

Therefore, a dual-faceted evaluation approach is essential for informed decision-making in HMTI design and deployment. By considering both how well a system performs and how intuitively it can be used and maintained, researchers and practitioners can make more nuanced judgments about interface suitability across different users, tasks, and environments. This comprehensive perspective is particularly critical in high-stakes applications, such as search and rescue or hazardous material handling, where both efficiency and operator well-being are paramount.

# Chapter 6

# Multi-Robot Task Planning Application

The task modelling proposed in Section 3 can also be used to help large language models (LLMs) in decision making for multi-robot task allocation based on mission spatial information.

## 6.1 Methodology

The cornerstone of the methodology is the introduction of a prompt-layer framework tailored for LLM-based decision-making systems. This framework extends the conventional multi-robot task allocation paradigm by incorporating not only the static capabilities of each robot but also dynamic performance factors, such as expected completion time and probability of success. These additional dimensions are derived from a formal task difficulty model, which allows the system to reason about task allocation in a manner that prioritises both efficiency and robustness. As illustrated in Figure 6.1, this enables the system to make more

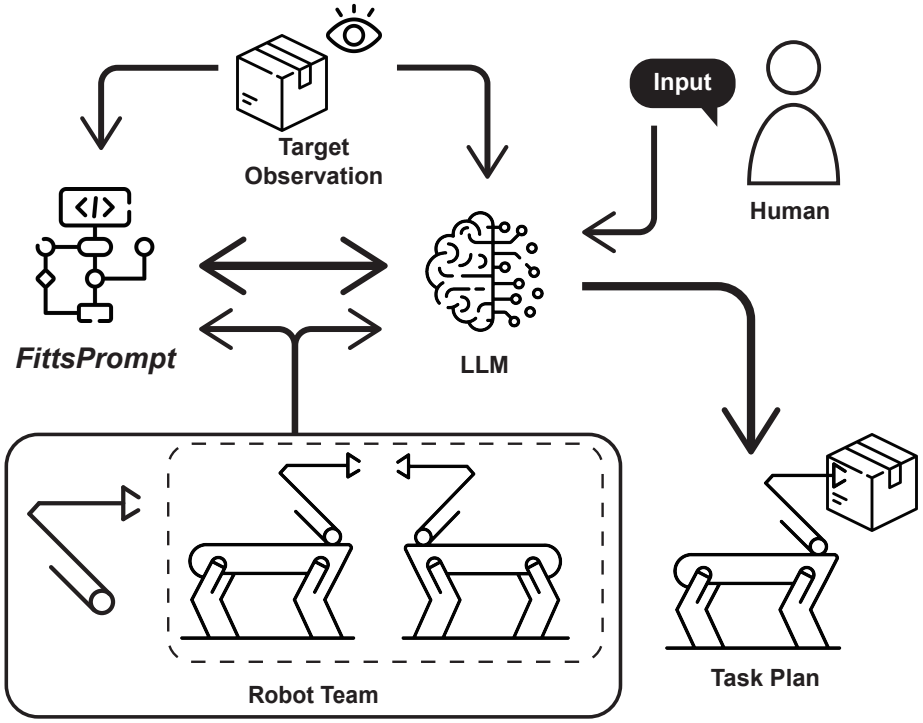



Figure 6.1: Illustrative representation of the proposed FittsPrompt for multi-robot task allocation.

nuanced allocation decisions by selecting the most appropriate robot for a given task—not merely based on its ability to perform the task, but by evaluating how challenging the task is for that robot and optimizing for reduced execution time and higher likelihood of success.

This extension moves beyond traditional methods that often rely solely on rule-based or cost-function-driven allocation schemes. Instead, this study embeds cognitive and spatiotemporal reasoning into the LLM's prompt design, allowing it to account for task difficulty in a generalizable and interpretable way. To address the intrinsic limitations of LLMs in processing complex, high-dimensional spatial observations, such as robot-to-target distances, orientations, or joint constraints, this study introduces a specialised preprocessing mechanism called *FittsPrompt*. This module transforms raw spatial state data into a scalar task difficulty representation using the extended Index of Difficulty (ID) model. By encoding relevant

environmental complexity into this compact form, FittsPrompt prevents raw, verbose state descriptions from overloading the LLM prompt, thereby improving the model's ability to generalise across environments while reducing token consumption.

**Algorithm 1** FittsPrompt pseudocode

**Require:** Set of robots $R$, set of tasks $T$, and set of targets $G$
**Ensure:** Optimal robot selection for each task

**for** each task $t_j \in T$ **do**
    Identify candidate robots $R_c \subseteq R$ based on capability using LLM
    **if** $|R_c| = 1$ **then**
        Select the only robot $r_j^* \in R_c$
    **else**
        **for** each robot $r_i \in R_c$ **do**
            Compute $\mathrm{ID}(r_i, g_k)$ with observation
        **end for**
        Select $r_j^* = \arg\min_{r_i \in R_c} \mathrm{ID}(r_i, g_k)$ optional using LLM
    **end if**
    Assign $r_j^*$ to task $t_j$
    Execute task $t_j$ with $r_j^*$
**end for**

This preprocessing step is crucial in ensuring the effectiveness of the LLM as a reasoning engine for robotics tasks. Without this abstraction, the model's input space would be cluttered with detailed metric data, making it prone to confusion or hallucination. By contrast, with FittsPrompt, the LLM receives a clear, semantically meaningful representation of the task environment, which allows it to focus on optimising assignment logic. The full task allocation pipeline, which combines the task difficulty scoring mechanism with LLM-driven planning and selection, is formally described in Algorithm 1.

The design of the framework is rooted in two central assumptions, both consistent with the empirical findings of Fitts' Law [12] as applied to robotic contexts. First, tasks that present higher inherent difficulty—as captured by geometric complexity, spatial alignment, or orientation constraints—require more time for robots

to execute. This relationship aligns with the original principle that movement time increases logarithmically with task complexity. Second, this study assumes that higher task difficulty also correlates with lower success rates, due to greater sensitivity to execution errors, delays, or environmental disturbances. These assumptions are supported by observations in robotic manipulation, navigation, and coordination literature, where increased task difficulty often leads to higher variance in performance outcomes.

By integrating the ID as a core component of the LLM decision process, the method introduces a quantifiable, interpretable, and scalable measure of task complexity that enhances the performance of multi-robot systems. This allows the LLM to make informed decisions about not just whether a robot can perform a task, but how well it is likely to do so, optimising allocation strategies with respect to both time and reliability. As a result, the proposed framework supports more adaptive, efficient, and resilient task assignment in scenarios ranging from warehouse automation to disaster response, where robot heterogeneity and mission unpredictability are significant concerns.

### 6.1.1 Task Definition

This study considers a shared environment where multiple autonomous robots operate concurrently to complete a predefined set of tasks. Each task involves interacting with one or more physical targets and varies in complexity due to differences in spatial configuration, required skills, and surrounding contextual constraints. The objective is to assign these tasks to the most appropriate robots in a manner that maximises mission efficiency, minimises execution time, and increases the likelihood of success.

Let the set of robots in the environment be defined as $R = \{r_1, r_2, \ldots, r_n\}$. Each

robot $r_i$ is associated with a specific state that includes spatial parameters such as position and orientation, a defined set of capabilities or skills (also referred to as a kill set), and possibly dynamic attributes like energy levels or current workload. These state descriptors influence the robot's suitability and effectiveness in performing a given task.

Similarly, this study defines the set of targets as $G = \{g_1, g_2, \ldots, g_m\}$, where each target $g_k$ represents a physical objective within the environment that must be approached, manipulated, or otherwise acted upon. Targets may impose spatial constraints—such as occlusion, limited access angles, or confined workspace—and can vary in how accessible or demanding they are depending on the robot's configuration.

The overall task set is denoted as $T = \{t_1, t_2, \ldots, t_l\}$, where each task $t_j \in T$ corresponds to an allocation instance defined by three core elements: a designated target $g_k \in G$, the skill requirements needed to accomplish the task, and any relevant environmental or operational constraints. These constraints may include spatial difficulty, temporal urgency, or sequencing dependencies with other tasks.

In existing LLM-based task allocation frameworks, large language models are typically used to parse task specifications and identify a subset of robots, $R_c \subseteq R$, that possess the necessary skills to accomplish a given task $t_j$. While this step ensures basic feasibility, it does not account for spatial efficiency or potential variation in performance across robot candidates.

To address this limitation, the proposed FittsPrompt framework introduces task difficulty modelling into the LLM-driven decision process. Once the capable subset $R_c$ is identified, FittsPrompt enables the LLM to further evaluate each robot $r_i \in R_c$ in relation to the assigned target $g_k$, using spatial parameters to estimate the difficulty of completing the task. These parameters—such as Euclidean

distance, alignment angles, and task-specific interaction constraints—are used to compute a scalar task difficulty value based on the extended 3D Index of Difficulty formulation.

The resulting difficulty scores provide the LLM with a compact yet informative abstraction of the task environment. Rather than processing raw geometric data, the LLM receives a semantically meaningful prompt encoding the estimated effort required for each robot-target pair. This allows the LLM to select the most efficient robot $r_i \in R_c$ for executing task $t_j$, ensuring that both capability and spatial feasibility are taken into account during task assignment.

By embedding task difficulty directly into the decision pipeline, the framework supports more context-aware and resource-efficient task allocations in multi-robot systems operating in dynamic and spatially complex environments.

The task difficulty, quantified by the ID, is used to evaluate how challenging it is for a given robot to complete a specific task based on its spatial relationship to the target. For a task $t_j$, the ID values are computed for each robot $r_i$ within the subset of candidate robots $R_c$, in relation to a designated target $g_k$. This is expressed as:

$$\{\mathrm{ID}(r_i, g_k) \mid r_i \in R_c\} = \{f(p_{r_i}, p_{g_k}) \mid r_i \in R_c\} \tag{6.1}$$

Here, $R_c \subseteq R$ represents the subset of robots that possess the required capabilities to perform task $t_j$. Each robot $r_i$ is characterised by a set of spatial parameters, denoted as $p_{r_i}$, which includes its current position, orientation, and the physical dimensions of its end-effector or tool. Similarly, the target $g_k$ is defined by a corresponding parameter set $p_{g_k}$, which includes its location, orientation in the environment, physical size, and any relevant interaction constraints such as

approach angles or workspace limitations.

The function $f$ serves as a mapping between the robot-target parameter pairs and a scalar difficulty score. This function integrates geometric and kinematic information to estimate how demanding the task would be for each robot given its current state. The formulation of $f$ is based on an extended version of Fitts' Law, adapted for 3D space and robot-specific execution contexts. It accounts for both translational and rotational displacements, as well as task-specific tolerances. The result is a set of ID values that reflect the relative ease or difficulty with which each candidate robot can approach and interact with the assigned target.

This representation provides the LLM with a compact and interpretable abstraction of task complexity, reducing the burden of processing raw spatial inputs and enabling more informed, context-aware decision-making during the task allocation process.

### 6.1.2 Optimization Layer

The optimisation process proceeds in two stages. First, the LLM performs a capability filtering step to ensure that only feasible robots are considered for allocation. For each task $t_j$, the required skill set $S(t_j)$ is compared against the capability set $C(r_i)$ of each robot $r_i \in R$. The feasible candidate subset is then defined as:

$$R_c = \{r_i \in R \mid S(t_j) \subseteq C(r_i)\}. \tag{6.2}$$

This ensures that only robots possessing all of the required capabilities are eligible for assignment. For example, if $S(t_j)$ requires grasping and cutting, then any robot lacking either a gripper or a cutter will be excluded from $R_c$, regardless of

its proximity to the target. The prompt is provided in the Appendix.C.1.

Second, once the feasible subset $R_c$ has been determined, task difficulty modelling is applied to discriminate among candidates. For each $r_i \in R_c$ and the target $g_k$ of task $t_j$, a scalar difficulty score $\mathrm{ID}(r_i, g_k)$ is computed using the FittsPrompt formulation introduced earlier. The final optimisation step then selects:

$$r_j^* = \arg\min_{r_i \in R_c} \mathrm{ID}(r_i, g_k). \tag{6.3}$$

Here, $r_j^*$ denotes the optimal robot selected by the LLM to perform task $t_j$. The function $\mathrm{ID}(r_i, g_k)$ quantifies the spatial and operational difficulty associated with assigning robot $r_i$ to target $g_k$. The set $R_c$ includes all robots capable of executing $t_j$, as determined during the capability filtering phase. The prompt is provided in the Appendix.C.2. It corresponds to the robot expected to complete the task with the greatest efficiency and reliability.

In this two-stage structure, the LLM serves as the high-level planner that integrates both logical feasibility checks (via capability matching) and spatial reasoning (via difficulty-aware optimisation). By first pruning infeasible options and then reasoning over structured difficulty scores, the system ensures that each selected robot is not only able to perform the task but also likely to do so in the most efficient manner.

**Implement Task Difficulty Modelling**

Unlike conventional optimisation approaches that require manually defined cost functions or rule-based heuristics, this approach enables the LLM to internalise these decision patterns from data and natural language instructions. This not only simplifies the control logic but also enables more flexible generalisation to

novel scenarios, provided that the prompt structure includes all relevant task difficulty values and constraints.

Since task difficulty is influenced not only by the distance between the robot and the target but also by their respective spatial tolerances, this study adopts an extended version of Fitts' Law [3] to model task difficulty more comprehensively. This extended formulation accounts for both translational and orientational complexity, allowing for a more accurate prediction of execution effort in robotic systems. Specifically, this study begins by quantifying the translational component of the ID, which reflects the physical separation between the robot and the target along a straight-line path in 3D space.

The translational task difficulty between a robot $r_i$ and a target $g_k$ is defined as:

$$\mathrm{ID}_{\mathrm{trans}}(r_i, g_k) = \log_2 \left( \frac{d(r_i, g_k)}{w_g \pm w_r} + 1 \right), \tag{6.4}$$

In this expression, $d(r_i, g_k)$ denotes the Euclidean distance between the robot and the target, which captures the magnitude of spatial displacement required to reach the target. The denominator incorporates both the physical size of the target ($w_g$) and the size of the robot's tool or end-effector ($w_r$). The operator $\pm$ accounts for task-specific interactions: in some tasks, such as docking or contact-based manipulation, the effective interaction area increases with the combined sizes of the robot and target, while in others—such as insertion or precision tasks—the difference between these sizes determines the spatial tolerance.

This formulation is grounded in the original principles of Fitts' Law, which posits a logarithmic relationship between movement time and task difficulty. By applying this relationship in the context of robot-target interactions, this study creates a scalable and interpretable measure of task complexity that can be computed

for any robot-target pair. This metric serves as a foundational component of the ID used in the optimisation layer described previously, allowing both LLMs and traditional planners to reason over spatial task difficulty in a consistent and data-efficient manner.

In addition to translational distance, task difficulty is also influenced by orientational alignment between the robot and the target. Many real-world tasks—such as insertion, assembly, docking, or precise manipulation—require the robot to approach the target not only from a specific position but also with an appropriate orientation. Misalignment in orientation can significantly increase the likelihood of task failure, even if the robot is spatially close to the target. Therefore, this study incorporates an orientation-based task difficulty component into the extended ID formulation.

The orientation-based task difficulty between a robot $r_i$ and a target $g_k$ is given by:

$$\mathrm{ID}_{\mathrm{ori}}(r_i, g_k) = \log_2 \left( \frac{\theta(r_i, g_k)}{\delta} + 1 \right). \tag{6.5}$$

In this expression, $\theta(r_i, g_k)$ represents the angular deviation between the robot's current orientation and the desired orientation required for successful task execution at the target. This angular deviation may be computed as the absolute rotation difference in yaw, pitch, and roll, or via more compact representations such as quaternion angular distance, depending on the application and platform.

The denominator $\delta$ denotes the allowable angular tolerance for the task. It defines the maximum permissible orientation error within which the robot can still successfully interact with the target. Tasks that require high precision, such as aligning a connector to a socket, will have smaller $\delta$ values. In contrast, more

forgiving tasks, such as general object pushing or rough placement, may allow for higher orientation variability.

This orientation component mirrors the structure of the translational difficulty formulation and maintains consistency with the logarithmic scaling of Fitts' Law. By introducing $\mathrm{ID}_{\mathrm{ori}}$ alongside $\mathrm{ID}_{\mathrm{trans}}$, this study can capture both the positional and directional challenges involved in robot-target interactions. This enriched representation enables more nuanced and accurate assessments of task difficulty in both planning and decision-making contexts.

To capture the full complexity of task execution from both positional and orientational perspectives, the overall task difficulty is computed as a weighted sum of the translational and orientational Index of Difficulty components. This composite metric, denoted as $\mathrm{ID}(r_i, g_k)$, provides a unified scalar representation of task complexity that accounts for the spatial configuration and alignment required between a robot $r_i$ and a target $g_k$, as shown below:

$$\mathrm{ID}(r_i, g_k) = \alpha \cdot \mathrm{ID}_{\mathrm{trans}}(r_i, g_k) + \beta \cdot \mathrm{ID}_{\mathrm{ori}}(r_i, g_k). \tag{6.6}$$

In this formulation, $\mathrm{ID}_{\mathrm{trans}}(r_i, g_k)$ represents the difficulty arising from translational displacement, while $\mathrm{ID}_{\mathrm{ori}}(r_i, g_k)$ quantifies the difficulty due to orientation misalignment. The parameters $\alpha$ and $\beta$ are tunable weight coefficients that determine the relative influence of each component in the overall score. These weights can be set empirically based on task requirements, robot capabilities, or prior experimental calibration. For instance, if a task is highly sensitive to alignment precision, a higher value of $\beta$ may be assigned to emphasise the importance of orientation. Conversely, if task completion depends predominantly on reaching a specific location, $\alpha$ would be weighted more heavily.

The weighted combination enables flexible adaptation to diverse robotic applications. For example, in assembly tasks involving fine motor control, both translational reach and precise orientation are critical, whereas in transport tasks, translation may dominate. This adaptability allows the ID framework to generalise across multiple domains, robot types, and mission scenarios.

By unifying both spatial dimensions into a single metric, $\mathrm{ID}(r_i, g_k)$ serves as an effective cost function for robot-task allocation decisions within the LLM-based framework. It provides the basis for optimising robot selection in a way that is both interpretable and computationally efficient, enabling the LLM to compare robot-target pairs and select the most suitable option for each task. This integration supports robust, skill-aware, and spatially-informed multi-robot coordination in real-world environments.

Once the translational and orientational components of task difficulty have been computed for each candidate robot-target pair, the LLM processes these precomputed ID values within a structured prompt. This prompt presents the LLM with a concise representation of the task complexity associated with each robot in the candidate set $R_c$ for a given task $t_j$. Rather than analysing raw spatial or sensory input, the LLM reasons over these scalar difficulty values to identify the most suitable robot for execution.

In this context, the original optimisation objective from Equation 6.3 can be refined to incorporate the weighted difficulty model, yielding the following formulation:

$$r_j^* = \arg\min_{r_i \in R_c} \left( \alpha \cdot \mathrm{ID}_{\mathrm{trans}}(r_i, g_k) + \beta \cdot \mathrm{ID}_{\mathrm{ori}}(r_i, g_k) \right). \tag{6.7}$$

Here, the LLM selects the optimal robot $r_j^*$ for task $t_j$ by minimising the total task

difficulty, defined as the weighted sum of the translational and orientational ID components. The parameters $\alpha$ and $\beta$ balance the influence of positional distance and angular alignment based on the task's spatial and control requirements. This formulation enables a more nuanced and physically grounded form of reasoning, where the LLM selects robots not solely on binary skill compatibility but on predicted performance efficiency.

It is worth noting that once scalar Index of Difficulty values have been computed, a simple deterministic arg min could be used to select the robot with the lowest task difficulty. In fact, a practical hybrid system might combine both approaches, where the LLM provides an initial reasoning-based decision and a deterministic controller subsequently verifies or overrides the choice based on efficiency. In this work, however, we deliberately allow the LLM to perform the comparison step as a proof of concept. This enables us to study the extent to which large language models can reason over structured numeric abstractions, with the understanding that additional decision factors (e.g., mission rules, safety constraints, or operator preferences) could later be integrated into the same framework.

By leveraging the LLM's language-based reasoning capabilities in conjunction with structured, interpretable ID metrics, the system achieves context-aware task allocation that reflects real-world spatial constraints. This approach significantly enhances the decision quality over traditional rule-based or purely symbolic planners, particularly in environments where geometry, orientation, and physical tolerances play critical roles in task feasibility.

By incorporating this optimisation step, the LLM enhances the task allocation process beyond traditional capability-based matching. Rather than merely selecting a robot that can perform a task, the LLM leverages precomputed spatial difficulty metrics to reason about which robot is best suited to execute the task

with minimal effort and maximum reliability. This shift from binary feasibility checks to quantitative difficulty-aware optimisation enables a more refined and context-sensitive decision-making process.

As a result, the selected robot for each task is not only capable but also spatially and kinematically optimal relative to the target, thereby reducing the execution burden. This significantly improves overall system performance by minimising execution time, lowering energy consumption, and increasing task success rates, especially in scenarios involving complex geometries, limited workspace, or heterogeneous robot platforms.

Incorporating task difficulty into the LLM's reasoning process contributes to a more intelligent and adaptive multi-robot system, capable of making efficient allocation decisions in dynamic and spatially constrained environments. This integration forms the foundation for robust, geometry-aware planning in large-scale autonomous systems. It supports scalable deployment across various domains such as warehouse automation, field robotics, and collaborative human-robot operations.

## 6.2 Validation

To demonstrate the effectiveness and generalizability of the proposed FittsPrompt framework, this study performs a two-stage validation process that combines large-scale simulation-based benchmarking with targeted real-world robot experiments. This dual approach ensures that the framework is evaluated not only in controlled, reproducible simulation environments, but also in physical deployments that expose it to sensor noise, actuation imperfections, and environmental uncertainties.

It is important to note that, in this validation, the LLM itself is used to perform the difficulty-aware optimisation step. While a deterministic arg min over precomputed Index of Difficulty values would suffice for selecting the numerically optimal candidate, the goal in this work is to treat the LLM as the decision engine as a proof of concept. This design allows us to directly evaluate the reasoning capabilities of large language models when confronted with structured numeric abstractions, while recognising that in future hybrid systems, the LLM could provide initial reasoning and constraint handling, with a deterministic optimiser ensuring efficiency.

The validation strategy is designed with three objectives:

1. **Evaluate efficiency and correctness at scale:** Using the benchmark, this study tests FittsPrompt across diverse simulated household and task environments. This provides systematic evidence of whether difficulty-aware prompting consistently improves allocation quality compared to raw state prompting.

2. **Benchmark against human baselines:** To contextualise the results, this study compares the performance of LLM-based task planning with human planners operating under the same information constraints. This establishes whether the framework can match or exceed human decision-making in time-critical allocation scenarios.

3. **Assess transferability to physical platforms:** By conducting experiments with quadruped robots in a controlled laboratory setting, this study evaluates whether plans generated via FittsPrompt can be reliably executed in the real world under noisy and dynamic conditions.

Together, these validation steps provide both breadth (through high-volume, di-

verse simulations) and depth (through focused physical trials), ensuring that the proposed approach is rigorously assessed in terms of scalability, robustness, and practical applicability. The following subsections detail the benchmark and real-robot evaluations.

### 6.2.1 Benchmark Evaluation

To systematically evaluate the task allocation efficiency and overall planning quality of FittsPrompt, this study develops a structured benchmark evaluation procedure based on the BEHAVIOR-1K benchmark suite [81]. BEHAVIOR-1K is implemented on the OmniGibson simulation environment, which is built atop NVIDIA Omniverse and PhysX 5, enabling realistic, interactive simulation of complex object dynamics, deformables, fluids, and enriched object states. Representative scenes are shown in Fig. 6.2, illustrating the diversity of challenges that the models encountered during testing. Fig. 6.3, illustrating the robot family that can be used in the benchmark. Within this validation framework, the LLM is presented with a structured, spatially detailed environmental observation comprising:

- **Robot specifications:** Descriptions of all available robots, including identities, skill/capability sets, and relevant operational parameters (e.g., mobility type, tool availability, kinematic constraints).
- **Spatial states:** Explicit spatial data specifying the current poses of each robot, as well as the precise locations, orientations, and dimensions of all target objects and other salient items in the environment.
- **Task instruction:** A clearly defined instruction stating the objective (target object/location) and the required action (e.g., manipulation, transportation, assembly).

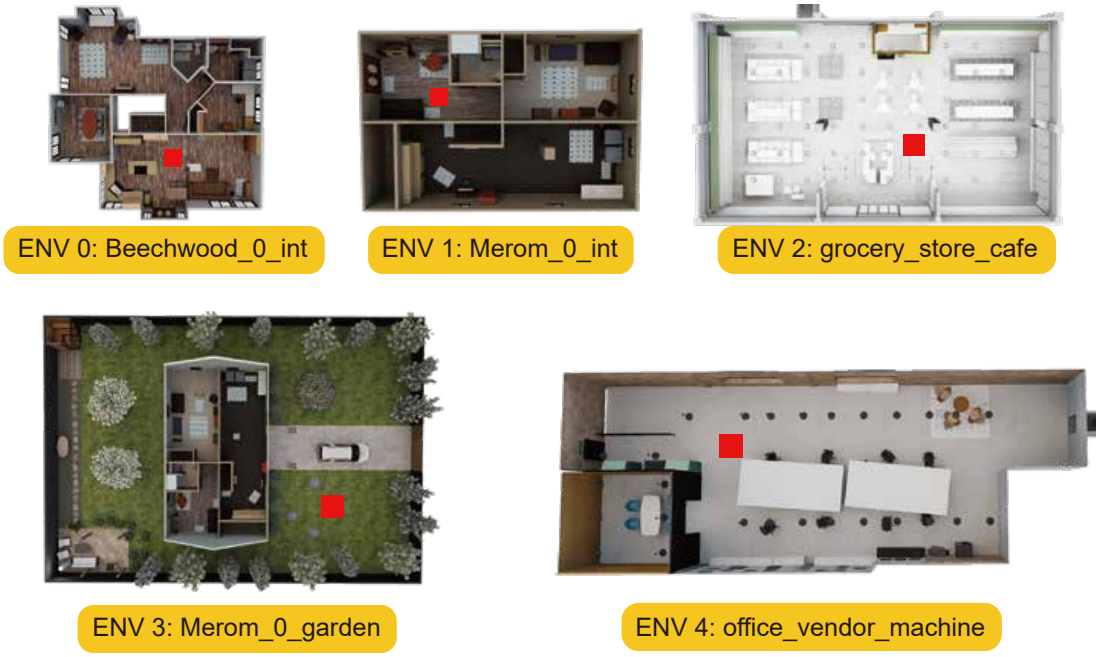



Figure 6.2: Visualisation of scenes used in the benchmark, where the robots' initial positions are shown in red squares.

The observation is instantiated from an environment JSON file that generates the simulation world; useful fields are extracted to inform the decision-making process and are fed into the enquiry LLM. This comprehensive information ensures the LLM has sufficient context to generate optimised and executable task-allocation plans, allowing rigorous assessment of performance in complex and dynamic multi-robot scenarios.

The environmental observations used in FittsPrompt are instantiated from environment JSON files, which specify both robot agents and target objects. Relevant attributes such as positions, orientations, capabilities, and interaction properties are extracted from these files and encoded into the LLM prompt. Tables 6.1 and 6.2 show simplified examples drawn from one such observation.

These structured entries demonstrate how raw simulation definitions are translated into concise, tabular inputs that inform the LLM's decision-making. In

Figure 6.3: Different types of robot used in the benchmark simulation.

practice, FittsPrompt further compresses this information into scalar difficulty values, reducing token usage while preserving the spatial and functional semantics required for effective task allocation.

### Top model selection

In addition to providing a rigorous evaluation protocol, the benchmark also serves as the basis for model selection. The performance of each LLM is quantified using two primary metrics: *success rate* and *optimisation rate*.

The success rate measures the proportion of trials in which the LLM produces an allocation plan that is both valid and achieves the specified task objective. For a given model $M$ evaluated over $N$ benchmark trials, this is computed as:

$$\text{SuccessRate}(M) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\big[\text{plan}(M, t_j) \text{ succeeds}\big], \tag{6.8}$$

where $\mathbb{I}[\cdot]$ is the indicator function, returning 1 if the generated plan for trial $t_j$ successfully achieves the objective, and 0 otherwise.

Table 6.1: Example robot specifications extracted from environment JSON (with randomised locations and vertical formatting).

| **ID** | **Class** | **Movable** | **Actions** | **Location** | **Orientation** | Speed(m/s)<br>Mobile<br>Arm |
|---|---|---|---|---|---|---|
| 23 | Stretch3 | True | REACH_TO<br>PICK_UP | (3.42<br>17.89<br>0.00) | (0.00<br>1.00<br>0.00) | 0.35<br>0.50 |
| 24 | UR5 | False | PICK_UP<br>OPEN<br>CLOSE | (12.76<br>8.54<br>0.00) | (0.00<br>0.00<br>0.00) | –<br>1.00 |
| 25 | Fetch | True | REACH_TO<br>PICK_UP<br>OPEN<br>CLOSE | (19.34<br>2.71<br>0.00) | (0.00<br>0.00<br>0.00) | 1.00<br>0.75 |
| 26 | Fetch | True | REACH_TO<br>PICK_UP<br>OPEN<br>CLOSE | (7.82<br>14.29<br>0.00) | (0.00<br>0.00<br>0.00) | 1.00<br>0.75 |

The optimisation rate measures the proportion of trials in which the LLM not only succeeds but also selects the robot (or target) corresponding to the lowest predicted task difficulty value according to FittsPrompt. Formally:

$$\text{OptRate}(M) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\Big[r_j^*(M) = \arg\min_{r_i \in R_c} \text{ID}(r_i, g_k)\Big], \tag{6.9}$$

where $r_j^*(M)$ denotes the robot chosen by model $M$ for task $t_j$, and $\text{ID}(r_i, g_k)$ is the difficulty score defined in Section 6.

Since the benchmark covers both *multi-robot task allocation* and *multi-target selection*, the overall optimisation performance of each model is obtained by aggregating across both task types:

$$\text{OverallOptRate}(M) = \frac{N_{\text{allo}} \cdot \text{OptRate}_{\text{allo}}(M) + N_{\text{exec}} \cdot \text{OptRate}_{\text{exec}}(M)}{N_{\text{allo}} + N_{\text{exec}}}, \tag{6.10}$$

Table 6.2: Example target specifications extracted from environment JSON.

| **ID** | **Category** | **Class** | **Properties** | **Location** (x,y,z) | **States** |
|---|---|---|---|---|---|
| 1 | Furniture | coffee table | SURFACES<br>LOW_HEIGHT | 4.27<br>15.68<br>0.00 | N/A |
| 2 | Facility | grill | HEATABLE<br>SURFACES<br>ON_HIGH_SURFACE | 10.53<br>7.89<br>0.00 | HEATED |
| 3 | Appliances | car | N/A | 18.42<br>2.75<br>0.00 | N/A |
| 4 | Appliances | mailbox | CONTAINERS<br>LOW_HEIGHT | 12.34<br>3.29<br>0.00 | OPEN |
| 5 | Ball | soccer ball | GRABABLE<br>MOVABLE | 6.77<br>19.21<br>0.00 | N/A |
| 6 | Drinks | water bottle | GRABABLE<br>MOVABLE<br>ON_HIGH_SURFACE | 2.89<br>11.46<br>0.00 | N/A |

where $N_{\text{allo}}$ and $N_{\text{exec}}$ denote the number of trials in allocation and execution (target selection) benchmarks, respectively.

In this study, the top-performing model is selected as the LLM that achieves the highest OverallOptRate($M$) across all benchmark trials, subject to maintaining a non-trivial success rate. This ensures that the chosen model is not only capable of generating valid, executable plans but also demonstrates consistent difficulty-aware optimisation across both allocation and selection scenarios, in line with the design goals of FittsPrompt.

### 6.2.2 Real-Robot Evaluation

To thoroughly assess the practical applicability and transferability of FittsPrompt beyond simulated environments, this study conducted real-world experiments us-

ing physical legged robot platforms. These experiments were designed to evaluate whether the task allocation plans generated by the LLM-driven FittsPrompt method could be effectively deployed under realistic conditions, characterised by uncertainty, sensory noise, and dynamic changes commonly encountered in physical environments.

The experimental setup, as depicted in Fig. 6.4, involved multiple legged-robots operating concurrently in a carefully controlled yet dynamically complex workspace. The environment was structured to present challenges representative of real-world operational scenarios, including spatial obstacles, varied terrain textures, and precisely placed target objects requiring accurate navigation and manipulation. This arrangement provided conditions that demanded spatial reasoning and precise task difficulty modelling, which were essential for validating the effectiveness.

**Experiment Setup**

To validate the real-world applicability of the proposed FittsPrompt-based task allocation framework, this study deployed a multi-robot system integrated with a high-precision motion capture (MOCAP) system in a controlled indoor environment. The setup enabled continuous tracking of both robot agents and physical targets, ensuring accurate spatial data acquisition for real-time task reasoning.

The OptiTrack MOCAP system, equipped with multiple ceiling-mounted cameras and reflective markers affixed to the robots and target objects, provided precise 3D positional and orientational data. This information was used to construct a structured observation of the environment, including both robot states and object configurations. The structured observation was then translated into a prompt format that encoded all relevant task information, such as robot capabilities, tool sizes, and target parameters, along with a natural language instruction specifying

the intended task.

This prompt was fed into the large language model `gpt-o1-preview`, chosen for its top-tier performance in previous benchmark evaluations. Based on the spatial reasoning embedded via the FittsPrompt difficulty model, the LLM selected the robot best suited to execute the given task under real-world conditions. The output of the LLM—specifying the chosen robot—was parsed and forwarded to the robotic control stack, prompting the selected agent to initiate and complete the assigned task autonomously.

The experimental environment consisted of the following core components:

- **Robot agents:** Two legged-robots from Unitree Robotics—an Aliengo and an A1 quadruped platform [148]. These robots were selected for their differing physical specifications, particularly in terms of tool size and payload capacity, thereby introducing variation in execution difficulty that FittsPrompt could leverage in its decision-making process.
- **Targets:** Two physical objects were used as representative task goals: a 24-can soft drink box and a model explosive device. These items were chosen to simulate common robotic search, rescue, and transport tasks with varied interaction demands and size constraints.
- **MOCAP system:** The tracking infrastructure consisted of an OptiTrack motion capture and 3D tracking system. Cameras were mounted on the ceiling to provide complete coverage of the task space, while active markers were placed on all mobile agents and target items to ensure precise real-time localisation and orientation estimation.

Fig. 6.4 presents visual documentation of the setup, showing the spatial arrangement of the robots and targets during task execution. This setup allowed for

a faithful validation of the FittsPrompt framework's ability to transition from simulation to physical execution and confirmed the effectiveness of LLM-driven task allocation in dynamic, sensor-instrumented environments.

### Experiment Planning

To validate the real-world applicability and decision-making effectiveness of FittsPrompt in robotic task allocation, this study designed a set of three controlled experiments. Each experiment presents a distinct scenario that challenges different aspects of task difficulty modelling, including spatial reasoning, tool utility, and response efficiency. These scenarios were selected to demonstrate FittsPrompt's ability to optimise task planning by considering multiple contextual parameters beyond mere capability matching.

**Experiment 1: Obstructed Path Removal.** In the first experiment, an unexpected obstacle—a 24-can soft drink box—is placed on a factory floor pathway, simulating an urgent scenario where access must be restored quickly. The system is instructed to select the most efficient robot to push the box away as rapidly as possible. As shown in Fig. 6.4a, the LLM receives spatial observations of the environment and must reason about both the relative position and orientation between each robot and the obstacle. In this scenario, the optimal solution requires FittsPrompt to consider not only Euclidean distance but also angular alignment, ensuring the robot approaches the object from a direction that minimises time-consuming repositioning or reorientation.

**Experiment 2: Tool-Based Precision Placement.** The second experiment focuses on task specificity involving hardware constraints. A robot is instructed to push a box into a designated goal region. The available robots differ in their end-effector or tool size, with one having a larger, more suitable pushing sur-

face. As illustrated in Fig. 6.4b, the LLM must select the robot that is best equipped to accomplish the task, prioritising tool appropriateness over raw proximity. This scenario challenges FittsPrompt to reason beyond spatial distance and incorporate physical parameters—such as tool dimensions—into its task allocation process. An optimised plan, in this case, would prefer a robot with a larger tool that enables more reliable contact and faster task completion, even if it is farther from the object.

**Experiment 3: Explosive Threat Inspection.** In the third experiment, the scenario simulates a high-priority inspection task involving a suspicious object (the bomb used in the EOD practical task experiment in Chapter.3), representing a model explosive device located in a semi-obstructed area. The robot must reach the site and inspect it as quickly as possible. Fig. 6.4c depicts this complex navigation setting where direct access may be impeded by nearby obstacles. Here, FittsPrompt must prioritise time efficiency and feasible access routes rather than the shortest distance alone. The optimised task plan requires the LLM to infer which robot can reach the inspection site fastest, taking into account potential navigational constraints that may not be captured by distance metrics alone.

Together, these three experiments comprehensively test FittsPrompt's capacity to reason under real-world physical constraints, adapt to diverse contextual cues, and generate high-quality task plans that are both executable and optimised for real-time deployment.

**Human baseline cohort.**

To contextualise the benchmarked LLM performance, this study additionally collected a human baseline using a cohort of 8 participants (6 male, 2 female), all with higher-education backgrounds (ages 23–30). Under the same observation

format and timing constraints as the LLMs (one minute per trial), participants produced task-allocation (and target-selection) plans from the benchmark scenes. These human results serve as a comparative baseline for the analyses reported later in this chapter and the Human study subsection).

## 6.3 Results

This section presents the results of the evaluation of the proposed FittsPrompt framework. The focus of this study is limited to assessing the effectiveness of FittsPrompt in generating valid and optimised task plans based on the proposed difficulty-aware prompting strategy. The study explicitly exclude the downstream execution of these plans by real robots, as execution dynamics—such as physical actuation errors, delays, or environmental uncertainties—fall outside the scope of this work. Instead, the evaluation centres on the quality, correctness, and optimality of the plans generated by LLMs in response to FittsPrompt-structured inputs.

To thoroughly evaluate the generalizability and robustness of the approach, this study conducted extensive testing across a wide range of simulated environments. These environments include diverse object arrangements and spatial configurations to reflect varying levels of task complexity.

The evaluation pipeline involved a large-scale comparison of LLMs, comprising 42 models in total. This included 38 open-source LLMs and 4 proprietary models, spanning a broad spectrum of parameter sizes and architectures. Notable models evaluated include `deepseek-r1-distill-qwen-32b`, `llama-3.3-70b-instruct`, and `gpt-o1-preview`, which represent the state-of-the-art in large-scale generative reasoning.

Among the tested models, 34 successfully produced task plans that were both syntactically valid and executable according to the simulation criteria. This success rate demonstrates that the FittsPrompt framework is compatible with various model backbones and can guide even distilled or lower-parameter models to generate high-quality planning outputs.

From this pool, this study identified the top-performing model based on a combination of plan correctness, spatial efficiency, and alignment with the task difficulty minimisation objective. This model was then used for a comparative analysis against human-generated task plans, as detailed in Fig. 6.5. The results clearly illustrate the superior planning capabilities enabled by FittsPrompt, highlighting its ability to outperform human baselines in scenarios that require spatial reasoning and optimal robot-task pairing. This comparative evaluation further validates the utility of the proposed approach in supporting high-performance, difficulty-aware planning in multi-robot systems.

To validate the effectiveness and generalizability of the proposed FittsPrompt framework, this study conducted a two-pronged evaluation consisting of both simulation-based benchmarking and real-world testing. The evaluation strategy is summarised as follows:

- **Benchmark Evaluation:** this study performed multi-robot task allocation across 10 distinct trials, each conducted in a different environment inspired by the BEHAVIOR-1K benchmark suite [81]. These environments were designed to present diverse spatial layouts, object distributions, and interaction requirements, thereby assessing the LLM's ability to generalise task planning across varied task contexts. In addition to the allocation tasks, this study also conducted 10 target selection trials involving different household items, focusing on the LLM's capacity to choose the most

suitable object or destination under spatial constraints and ambiguity.

- **Real-World Robot Experiments:** To evaluate the deployability of FittsPrompt beyond simulation, this study conducted real-world trials using a quadruped robot equipped with onboard computation and sensing. These experiments aimed to validate whether the plans generated via FittsPrompt were not only spatially optimised but also practically executable under real-world conditions involving terrain variation, actuation noise, and perception uncertainty. The robot successfully performed tasks based on FittsPrompt-generated instructions, demonstrating the framework's utility in bridging high-level decision-making and low-level execution in physical settings.

### 6.3.1 Benchmark Results

To systematically evaluate the task allocation efficiency and overall planning quality of FittsPrompt, this study designed a structured benchmark evaluation procedure inspired by the BEHAVIOR-1K benchmark suite [81]. Within this validation framework, the LLM is presented with a structured, spatially detailed environmental observation. Each environmental observation provided to the LLM includes:

- **Robot Specifications:** Detailed descriptions of all available robots, encompassing their respective identities, skill sets or capabilities, and relevant operational parameters such as mobility type, tool availability, and kinematic constraints.
- **Spatial States:** Explicit spatial data specifying the current positions and orientations of each robot, as well as the precise locations, orientations, and dimensions of all target objects and other relevant items within the environment.

- **Task Instruction:** A clearly defined task instruction that explicitly states the objective, identifying the target object or location and specifying the required action, such as manipulation, transportation, or assembly.

This comprehensive information ensures the LLM has sufficient context to generate optimised and executable task allocation plans, allowing us to rigorously assess its performance in complex and dynamic multi-robot scenarios.

To comprehensively assess the advantages of the proposed FittsPrompt framework compared to traditional raw data prompts, this study evaluated its performance across two essential metrics:

- **Success Rate:** This metric measures the proportion of trials in which the generated task allocation plan successfully addresses and fulfils the specified task objectives. A successful plan must correctly assign capable robots to relevant tasks and produce executable sequences of actions resulting in task completion.
- **Optimization Rate:** This metric assesses not only whether the task was completed successfully but also if the allocation was optimized. Specifically, this study measured the proportion of trials where the generated plans assigned robots in a manner that minimised overall task execution time, demonstrating effective utilisation of the difficulty-aware approach inherent to FittsPrompt.

To maintain high reliability and objectivity in the evaluations, two independent human evaluators performed manual verification of all generated plans. These evaluators confirmed task completion status, assessed the correctness of robot assignments, and verified the optimality of each allocation. Discrepancies between evaluators were resolved through discussion, ensuring robust and accurate

assessment of the FittsPrompt framework's efficacy.

### Multi-Robot Task Allocation

In this benchmark evaluation, this study conducted a comprehensive set of 10 trials involving multi-robot task allocation scenarios. Each trial involved 4 distinct robots operating within 5 varied environmental scenes designed to challenge the spatial reasoning and allocation capabilities of the tested LLM models. Fig. 6.2 provides visual examples of the diverse scenes used in these trials, illustrating the complexity and variability of spatial arrangements encountered by the robots.

Out of the 34 LLMs that successfully produced syntactically valid and executable task plans, 27 models were able to generate at least one fully correct task allocation plan, thereby demonstrating their baseline capability for accurate robot-task matching, as shown in Appendix D.1. Among these models, a direct comparative analysis was performed between the proposed FittsPrompt method and conventional raw data prompting approaches.

As shown clearly in Fig. 6.6, FittsPrompt significantly outperformed traditional raw data prompts across both evaluated metrics. Specifically, this study observed an approximate 9% increase in overall task success rates, indicating that FittsPrompt leads to more reliably executable and contextually appropriate robot assignments. Even more pronounced was the improvement in optimisation rates, where FittsPrompt achieved an impressive 30% enhancement. This metric underscores FittsPrompt's ability to consistently produce not just feasible but optimally efficient task allocations, minimising execution time and operational complexity.

Collectively, these quantitative results confirm the substantial advantages provided by FittsPrompt over conventional prompt-based methods. They highlight the transformative potential of integrating explicit task difficulty modelling with

LLM-driven planning to enhance the effectiveness and reliability of multi-robot task allocation systems.

### Multi-Target Selection

To further evaluate the capabilities of FittsPrompt, this study conducted an additional benchmark focusing on multi-target selection tasks. In this scenario, the large language model was required to choose the optimal target among several available options within the environment. Unlike the multi-robot task allocation scenario, where the target object is explicitly predefined, the multi-target selection benchmark granted the LLM autonomy to select the most appropriate object based on task instructions, using a single predetermined robot. This selection process required the LLM to exhibit spatial reasoning and optimality-driven decision-making, choosing the object that allowed for the most efficient task execution.

The multi-target selection benchmark comprised a total of 10 trials, each featuring various types of objects placed in diverse spatial configurations. Some of these objects appeared multiple times within the same scene, thus compelling the LLM to differentiate and select the best instance of a given target type. For instance, Fig. 6.7 visually exemplifies one such scenario, depicting a situation where the robot is tasked with picking up a banana, necessitating the selection of the most spatially suitable banana from several options available in the environment.

Out of the 34 LLM models that successfully generated executable plans, 30 were capable of producing at least one correct multi-target selection, as shown in Appendix D.2. The performance comparison between FittsPrompt and conventional raw data prompts, as clearly depicted in Fig. 6.8, demonstrates a significant advantage for the FittsPrompt approach. Specifically, FittsPrompt achieved a sub-

stantial 24% increase in overall task success rates, highlighting its improved accuracy in correctly selecting the optimal targets. Even more notably, FittsPrompt demonstrated a remarkable 45% improvement in optimisation efficiency, reflecting its superior capability in consistently identifying the targets that minimised overall execution complexity and time.

These results further underline the effectiveness and robustness of FittsPrompt, confirming its ability to significantly enhance the decision-making quality of LLM-driven robotic systems, particularly in scenarios requiring precise target selection and spatial reasoning.

**Human study**

To further contextualize the performance of the proposed FittsPrompt framework, this study conducted a comparative human study aimed at evaluating task-planning capabilities between human planners and the proposed LLM-driven approach. For this comparison, this study selected the `GPT-o1-preview` model due to its outstanding performance observed in prior benchmarking tests. A total of 8 independent trials were conducted with the model, each configured with a generation temperature parameter of 1 to encourage diverse responses and reflect realistic decision-making variability.

The human study involved 8 participants, each possessing prior experience in robotic teleoperation and task planning. Participants were presented with the same environmental observations and robot capability information provided to the LLM, ensuring consistency in task conditions. Each participant was tasked with manually assigning robots to perform the tasks defined in the Multi-Robot Task Allocation benchmark scenarios, replicating the conditions under which the LLM was evaluated. To establish equitable conditions, each human participant

was allocated exactly 1 minute per task, matching the timeout threshold given to the LLM. This design ensured fairness in comparisons by imposing identical planning constraints for both human and model-based planners.

The comparative results of this human versus LLM evaluation are illustrated in Fig. 6.5, clearly highlighting the performance differences between human planners and the FittsPrompt-driven `GPT-o1-preview` model. The outcomes of this experiment indicate that the FittsPrompt framework significantly surpassed human planners in terms of optimisation efficiency. Specifically, the proposed approach achieved a 35% improvement over human performance in multi-robot task allocation scenarios and a 30% improvement in target selection tasks. These marked improvements underscore the capability of FittsPrompt to consistently generate optimized and highly effective plans under tight time constraints—conditions under which human planners commonly exhibit limitations due to cognitive load and time pressures.

These findings not only validate the practical superiority of the proposed method but also highlight the broader implications of leveraging structured difficulty-aware prompting in robotic task planning. FittsPrompt enables more reliable and efficient robotic coordination than human planning, particularly in complex, real-time operational environments.

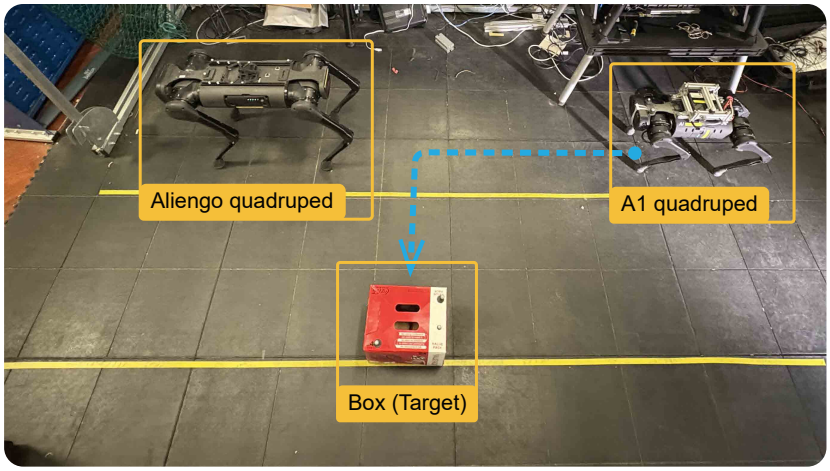
### 6.3.2 Experiment Results

As a result of the real-world experimental evaluation, FittsPrompt—powered by the `gpt-o1-preview` model—demonstrated exceptional precision in robotic task allocation, achieving a perfect 100% alignment with the ground-truth optimal task plans across all three tested scenarios. This outcome underscores the framework's effectiveness in translating difficulty-aware spatial reasoning into practical, real-

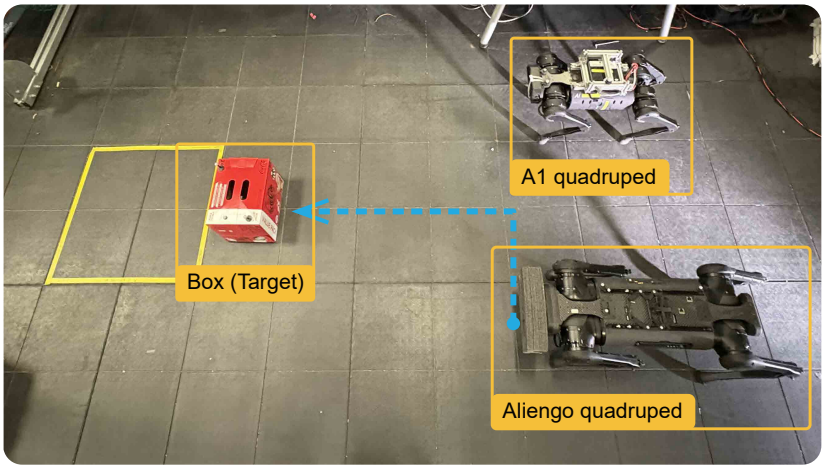
world decision-making.

The system consistently succeeded in the following key aspects:

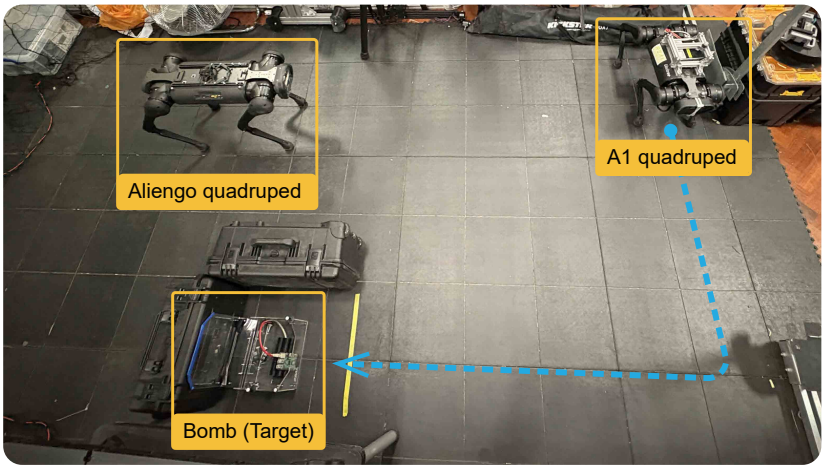
- **Optimising task allocation based on spatial state:** FittsPrompt accurately evaluated the spatial configurations of the environment, enabling the selection of the most time-efficient robot for each task. This directly contributed to reduced task execution durations and more responsive system behaviour.
- **Selecting functionally appropriate robots:** Beyond mere spatial proximity, the framework effectively considered robot-specific functional attributes, such as tool size, to determine the best-suited agent for the task. This improved the overall quality and reliability of task execution.
- **Handling real-world uncertainty and sensor noise:** Despite the presence of environmental variability, actuation imperfections, and sensory noise, the system maintained robust decision-making capabilities. This robustness illustrates FittsPrompt's potential for deployment in operational multi-robot systems beyond controlled simulation settings.




(a) Experiment 1

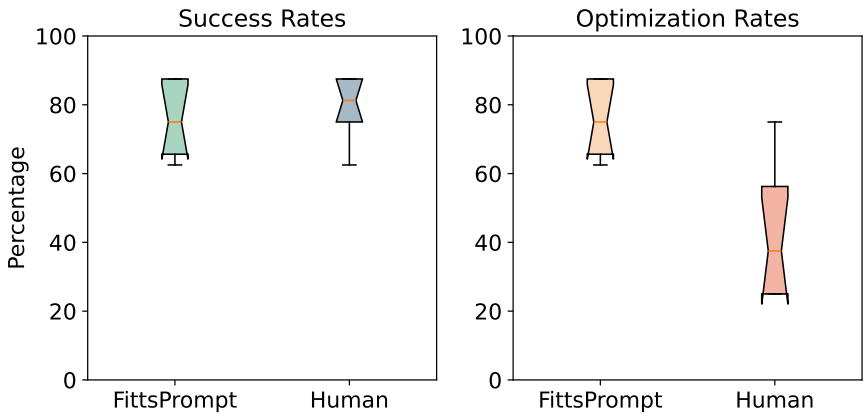



(b) Experiment 2

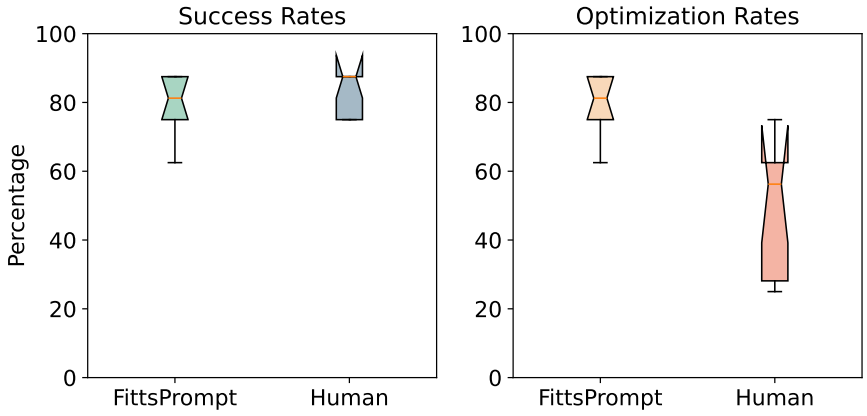



(c) Experiment 3

Figure 6.4: Setup and process of three real-world robot experiments, where robots and targets are marked in yellow boxes and planned paths are marked in blue dashed lines.




(a) Task allocation results.




(b) Target selection results.

Figure 6.5: The boxplots of the success rate and optimisation rate results from FittsPrompt and Human trials.

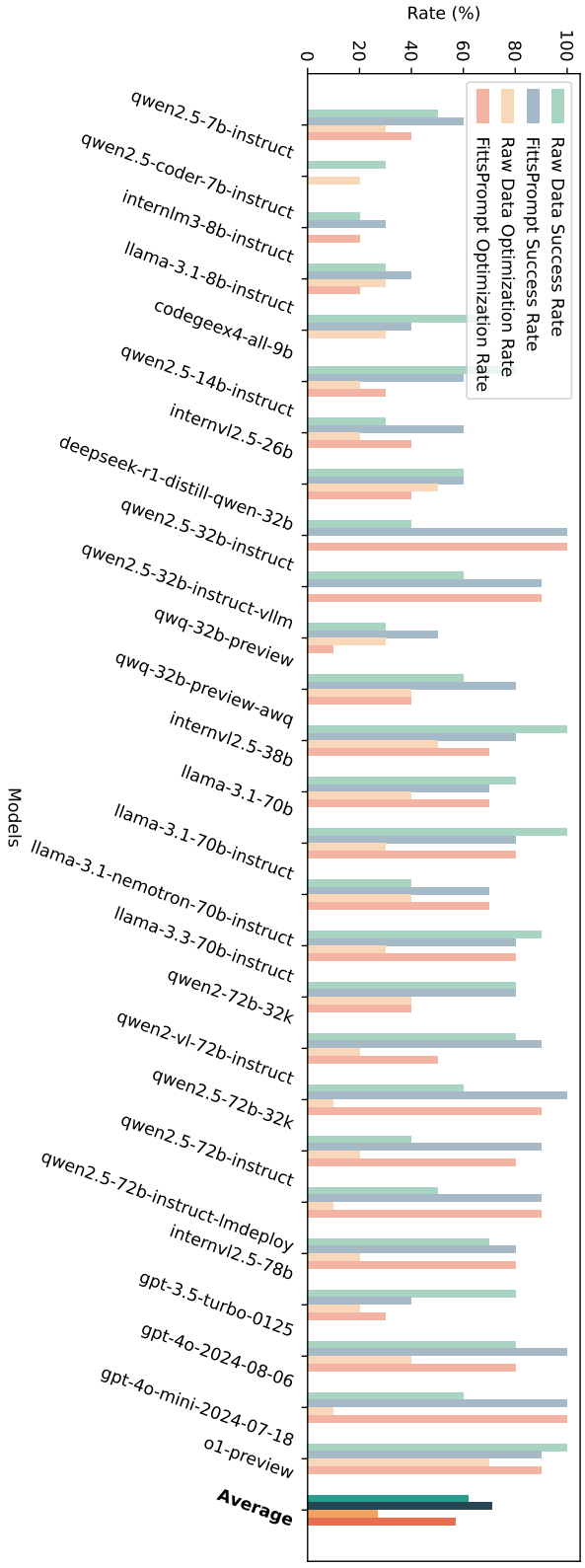



Figure 6.6: Comparison of Success and Optimisation Rate between raw data and FittsPrompt for different LLMs in task allocation. Each plot uses a notched boxplot representation: the central horizontal line shows the median, the box indicates the interquartile range (IQR, 25th–75th percentile), the whiskers extend to capture variability in the data, and the notch reflects an approximate 95% confidence interval around the median. Narrower boxes and shorter whiskers denote more consistent performance, while taller boxes indicate greater variability.


Figure 6.7: Visualization of sense where robot encounters multiple target objects. In this case, the robot needs to pick up a banana, but there is more than one in the observation.

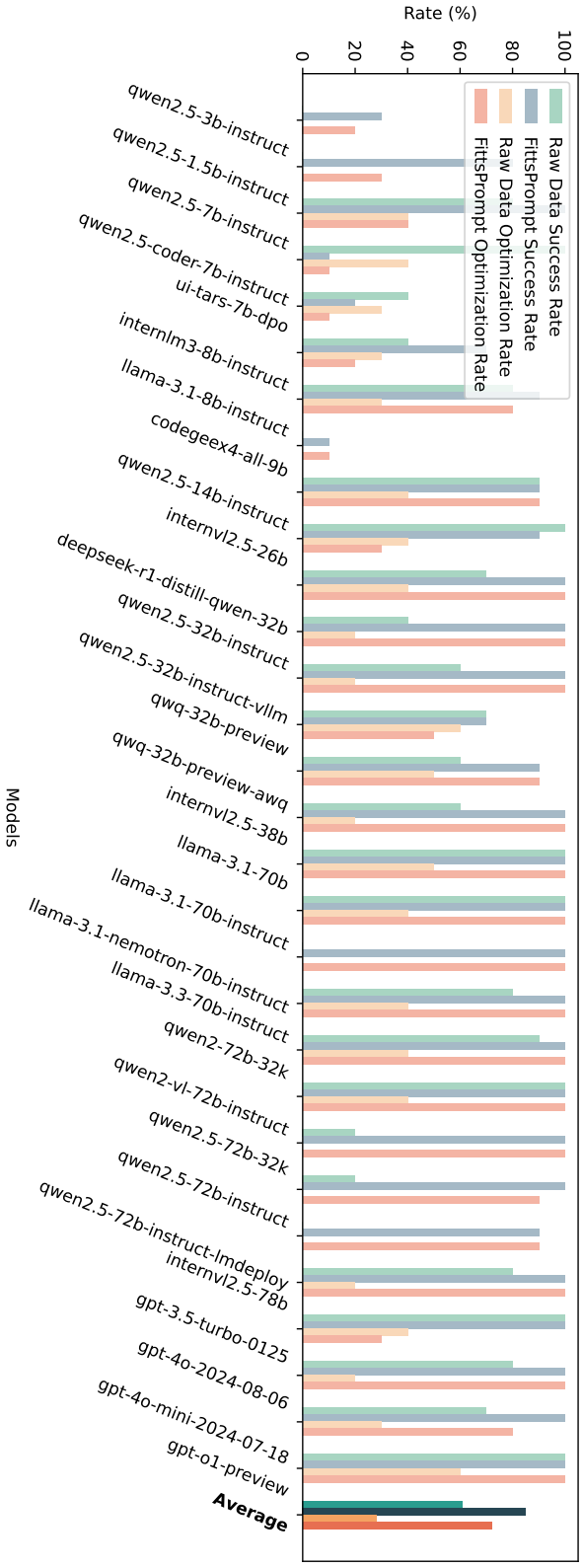



Figure 6.8: Comparison of Success and Optimisation Rate between raw data and FittsPrompt for different LLMs in target selection.

# Chapter 7

# Conclusion

This thesis presented a novel and comprehensive framework for task difficulty modelling in HMT, grounded in an extended formulation of Fitts' Law. By generalising the traditional 2D movement model to a six-degrees-of-freedom spatial domain, the proposed method effectively captures both translational and rotational constraints inherent in real-world robotic interactions. This advancement enables more accurate prediction of HMT performance in dynamic, unstructured environments.

Recognising the pivotal role of human operators in teleoperated and shared autonomy scenarios, the model further integrates cognitive fatigue through the SAFTE framework. This integration allows for realistic and adaptive performance forecasting, taking into account not only the long-term skill of operators but also their short-term cognitive readiness under operational stress. The framework thus bridges the gap between physical task complexity and human variability, offering a holistic tool for mission planning, operator scheduling, and system adaptation.

The proposed methodology was validated through rigorous empirical studies, in-

cluding a simulation-based evaluation and two real-world experiments involving a quadruped mobile manipulator. These studies demonstrated the model's practical utility in quantifying task difficulty and evaluating Human-Machine Interfaces through both objective metrics and subjective assessments such as NASA-TLX and SUS scores.

In extending this framework to modern AI-driven robotics, the thesis introduced FittsPrompt, a novel pre-processing mechanism for enabling effective multi-robot task allocation using LLMs. By abstracting spatial and motion-related complexity into structured difficulty descriptors, FittsPrompt significantly improved LLM-based decision-making performance, surpassing both traditional rule-based methods and expert human planners in benchmark and real-world settings.

Overall, this work provides a generalisable and interpretable foundation for task difficulty modelling, with implications for performance prediction, system evaluation, and intelligent decision-making across a broad spectrum of robotic applications. Future research can build upon this foundation by exploring adaptive learning of difficulty models in real-time, integrating physiological sensing for fatigue estimation, and scaling FittsPrompt to broader domains of collaborative autonomy.

## 7.1 Future Works

While this thesis has introduced a robust framework for task difficulty modelling and demonstrated its utility across diverse HMT and multi-robot scenarios, several promising directions remain for future exploration. These avenues aim to further enhance the adaptability, scalability, and real-world applicability of the proposed methods.

The future research will extend the current linear framework to a non-linear model, which is crucial in scenarios where human physical and cognitive fatigue influences performance during high-stress missions. Besides manually modelling human fatigue, artificial intelligence can be used to consider more complex human factors to provide a more accurate prediction in the real world. On the other hand, the quantified results of the proposed modelling can be leveraged by AI algorithms to tune parameters in human-machine interface design, as well as to refine task setups based on predictive feedback.

Moreover, the scope of the research will expand to encompass fine manipulation manoeuvres post-target contact. These include detailed tasks like opening a box, rotating a handle, or relocating a bottle. Incorporating these advanced manoeuvres into the model will enhance its predictive accuracy for complex tasks and offer insights into how human factors like fatigue can impact task execution in HMT systems.

Given that the influence of fatigue on human-robot collaboration remains an emerging area of research, many of the model's elements and coefficients are derived from studies in related fields, such as transportation and mechanical operation. Future investigations will concentrate on measuring these elements within human-robot interactions to refine the model's accuracy. Additionally, in the case study, the operator is considered unsuitable for executing missions when their effectiveness falls to zero. In real-world applications, a more stringent evaluation involving a higher cut-off for operator effectiveness would be imposed as a safety precaution, particularly in high-risk missions.

The human study experiment provided an example of applying the presented MHT evaluation scheme to real RA and HMTIs. From the results of the experiment with the gamepad and WMCS interfaces, the proposed model can predict

system performance in future missions. However, the HAs in the experiment had limited experience in robot teleoperation compared to professional operators in actual missions. Therefore, the results from the experiment only represent the user group with limited robot teleoperation experience, and a professional user group may produce different results.

Future work will concentrate on advancing the automation of data processing from observations, specifically by enhancing the generation of task difficulty metrics from detailed 3D models of the environment. Additionally, there is a significant opportunity to refine the modelling of the proposed tasks. This will involve developing multipurpose algorithms capable of handling complex tasks with higher precision.

# References


[1] Y. Wan and C. Zhou, "3-d fitts' law for performance prediction of human–machine teaming," *IEEE Transactions on Industrial Informatics*, pp. 1–10, 2025. DOI: 10.1109/TII.2025.3563553.

[2] Y. Wan and C. Zhou, "Predicting human-robot team performance based on cognitive fatigue," in *International Conference on Automation and Computing*, 2023, pp. 575–580.

[3] Y. Wan, J. Sun, C. Peers, J. Humphreys, D. Kanoulas, and C. Zhou, "Performance and usability evaluation scheme for mobile manipulator teleoperation," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 5, pp. 844–854, 2023. DOI: 10.1109/THMS.2023.3289628.

[4] Y. Ren and G.-P. Li, "An interactive and adaptive learning cyber physical human system for manufacturing with a case study in worker machine interactions," *IEEE Trans. Ind. Inform.*, vol. 18, no. 10, pp. 6723–6732, 2022.

[5] M. Xu, A. Hu, and H. Wang, "Visual-impedance-based human–robot cotransportation with a tethered aerial vehicle," *IEEE Trans. Ind. Inform.*, vol. 19, no. 10, pp. 10 356–10 365, 2023.

[6] C.-Y. Weng, Q. Yuan, F. Suárez-Ruiz, and I.-M. Chen, "A telemanipulation-based human–robot collaboration method to teach aerospace masking skills," *IEEE Trans. Ind. Inform.*, vol. 16, no. 5, pp. 3076–3084, 2020.

[7] C. D. Bellicoso, M. Bjelonic, L. Wellhausen, *et al.*, "Advances in real-world applications for legged robots," *Journal of Field Robotics*, vol. 35, no. 8, pp. 1311–1326, 2018.

[8] P. Damacharla, A. Y. Javaid, J. J. Gallimore, and V. K. Devabhaktuni, "Common metrics to benchmark human-machine teams: A review," *IEEE Access*, vol. 6, pp. 38 637–38 655, 2018.

[9] C. R. Kube, "Task modelling in collective robotics," *Auton. Robots*, vol. 4, pp. 53–72, 1997.

[10] D. Y. Y. Sim and C. K. Loo, "Extensive assessment and evaluation methodologies on assistive social robots for modelling human–robot interaction–a review," *Inf. Sci.*, vol. 301, pp. 305–344, 2015.

[11] J. B. Lyons, K. T. Wynne, S. Mahoney, and M. A. Roebke, "Trust and human-machine teaming: A qualitative study," in *Artificial Intelligence for the Internet of Everything*, Elsevier, 2019, pp. 101–116.

[12] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement.," *J. Exp. Psychol.*, vol. 47, no. 6, p. 381, 1954.

[13] Y. Cha and R. Myung, "Extended Fitts' law for 3d pointing tasks using 3d target arrangements," *International Journal of Industrial Ergonomics*, vol. 43, no. 4, pp. 350–355, 2013.

[14] L. D. Clark, A. B. Bhagat, and S. L. Riggs, "Extending fitts' law in three-dimensional virtual environments with current low-cost virtual reality technology," *International Journal of Human-Computer Studies*, vol. 139, p. 102 413, 2020.

[15] B. Rohrer, "A developmental agent for learning features, environment models, and general robotics tasks," *IEEE ICDL/Eprirob*, 2011.

[16] A. Hasselberg and D. Söffker, "Petri-net-based modeling of human operator's planning for the evaluation of task performance using the example of air traffic control," *IEEE Trans Hum Mach Syst*, vol. 45, no. 6, pp. 676–685, 2015.

[17] J. Y. C. Chen and M. J. Barnes, "Human–agent teaming for multirobot control: A review of human factors issues," *IEEE Trans Hum Mach Syst*, vol. 44, no. 1, pp. 13–29, 2014. DOI: 10.1109/THMS.2013.2293535.

[18] M. Freed and R. Remington, "A conceptual framework for predicting error in complex human-machine environments," in *CogSci*, 2022, pp. 356–361.

[19] J. Crandall, M. Goodrich, D. Olsen, and C. Nielsen, "Validating human-robot interaction schemes in multitasking environments," *IEEE Trans. Syst. Man Cybern. A Syst. Hum.*, vol. 35, no. 4, pp. 438–449, 2005. DOI: 10.1109/TSMCA.2005.850587.

[20] A. M. Zanchettin, A. Casalino, L. Piroddi, and P. Rocco, "Prediction of human activity patterns for human–robot collaborative assembly tasks," *IEEE Trans. Ind. Inform.*, vol. 15, no. 7, pp. 3934–3942, 2019.

[21] A. H. Memar and E. T. Esfahani, "Physiological measures for human performance analysis in human-robot teamwork: Case of tele-exploration," *IEEE Access*, vol. 6, pp. 3694–3705, 2018.

[22] H. Drewes, "Only one Fitts' law formula please!" *Proc. SIGCHI Conf. Hum. Factor Comput. Syst.*, pp. 2813–2822, 2010.

[23] A. T. Welford, *Fundamentals of skill*. Methuen, 1968.

[24] I. S. MacKenzie, "Fitts' law as a research and design tool in human-computer interaction," *Hum.-Comput. Interact.*, vol. 7, pp. 91–139, 1992.

[25] M. F. Stoelen and D. L. Akin, "Assessment of Fitts' law for quantifying combined rotational and translational movements," *Human Factors*, vol. 52, no. 1, pp. 63–77, 2010.

[26] C. Radix, P. Robinson, and P. Nurse, "Extension of Fitts' law to modeling motion performance in man-machine interfaces," *IEEE Trans. Syst. Man Cybern. A Syst. Hum.*, vol. 29, no. 2, pp. 205–209, 1999.

[27] T. Petrič, R. Goljat, and J. Babič, "Cooperative human-robot control based on Fitts' law," *IEEE Int. Conf. on Humanoid Robots*, pp. 345–350, 2016.

[28] S. Sutjipto, Y. Lai, M. G. Carmichael, and G. Paul, "Fitts' law in the presence of interface inertia," *Conf Proc IEEE Eng Med Biol Soc*, pp. 4749–4752, 2020.

[29] J. C. Jurmain, A. J. Blancero, J. A. Geiling, *et al.*, "HazBot: Development of a telemanipulator robot with haptics for emergency response," *American Journal of Disaster Medicine*, vol. 3, no. 2, pp. 87–97, 2008.

[30] J. Carpenter, *Culture and human-robot interaction in militarized spaces: A war story*. Routledge, 2016.

[31] S. L. Murray, Y. L. Simon, and H. Sheng, "The effects of chemical protective suits on human performance," *Journal of Loss Prevention in the Process Industries*, vol. 24, no. 6, pp. 774–779, 2011.

[32] S. Board, "Evaluation of us department of transportation efforts in the 1990s to address operator fatigue," *Safety Report*, 1999.

[33] K. Lee, J. Shin, and J.-Y. Lim, "Critical hazard factors in the risk assessments of industrial robots: Causal analysis and case studies," *Safety and Health at Work*, vol. 12, no. 4, pp. 496–504, 2021.

[34] E. Rosen, D. Whitney, E. Phillips, *et al.*, "Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays," *The International Journal of Robotics Research*, vol. 38, no. 12-13, pp. 1513–1526, 2019.

[35] Y. Wan, J. Sun, C. Peers, J. Humphreys, D. Kanoulas, and C. Zhou, "Performance and usability evaluation scheme for mobile manipulator teleportation," *IEEE Transactions on Human-Machine Systems*, 2023.

[36] S. R. Hursh, D. P. Redmond, M. L. Johnson, *et al.*, "Fatigue models for applied research in warfighting," *Aviation, Space, and Environmental Medicine*, vol. 75, no. 3, A44–A53, 2004.

[37] C. Yang, Y. Zhu, and Y. Chen, "A review of human–machine cooperation in the robotics domain," *IEEE Trans Hum Mach Syst*, vol. 52, no. 1, pp. 12–25, 2022.

[38] R. Chipalkatty, H. Daepp, M. Egerstedt, and W. Book, "Human-in-the-loop: MPC for shared control of a quadruped rescue robot," in *IEEE/RSJ IROS*, 2011, pp. 4556–4561.

[39] J. A. Marvel, R. Bostelman, and J. Falco, "Multi-robot assembly strategies and metrics," *ACM Computing Surveys (CSUR)*, vol. 51, no. 1, pp. 1–32, 2018.

[40] H. Chakraa, F. Guérin, E. Leclercq, and D. Lefebvre, "Optimization techniques for multi-robot task allocation problems: Review on the state-of-the-art," *Robotics and Autonomous Systems*, p. 104 492, 2023.

[41] OpenAI, J. Achiam, S. Adler, *et al.*, *Gpt-4 technical report*, 2024. arXiv: 2303.08774 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2303.08774.

[42] DeepSeek-AI, D. Guo, D. Yang, *et al.*, *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*, 2025. arXiv: 2501.12948 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2501.12948.

[43] Qwen, : A. Yang, *et al.*, *Qwen2.5 technical report*, 2025. arXiv: 2412.15115 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.15115.

[44] A. Grattafiori, A. Dubey, A. Jauhri, *et al.*, *The llama 3 herd of models*, 2024. arXiv: 2407.21783 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2407.21783.

[45] Z. Cai, M. Cao, H. Chen, *et al.*, *Internlm2 technical report*, 2024. arXiv: 2403.17297 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.17297.

[46] M. Ahn, A. Brohan, N. Brown, *et al.*, *Do as i can, not as i say: Grounding language in robotic affordances*, 2022. arXiv: 2204.01691 [cs.RO].

[47] M. Kim, K. Pertsch, S. Karamcheti, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[48] A. Brohan, N. Brown, J. Carbajal, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *arXiv preprint arXiv:2307.15818*, 2023.

[49] S. Belkhale, T. Ding, T. Xiao, *et al.*, *Rt-h: Action hierarchies using language*, 2024. arXiv: 2403.01823 [cs.RO].

[50] S. Reed, K. Zolna, E. Parisotto, *et al.*, *A generalist agent*, 2022. arXiv: 2205.06175 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2205.06175.

[51] A. Zeng, M. Attarian, B. Ichter, *et al.*, *Socratic models: Composing zero-shot multimodal reasoning with language*, 2022. arXiv: 2204.00598 [cs.CV].

[52] C. Li, J. Liang, A. Zeng, *et al.*, *Chain of code: Reasoning with a language model-augmented code emulator*, 2023. arXiv: 2312.04474 [cs.CL].

[53] T. Kwon, N. D. Palo, and E. Johns, *Language models as zero-shot trajectory generators*, 2023. arXiv: 2310.11604 [cs.RO].

[54] T. Silver, S. Dan, K. Srinivas, J. B. Tenenbaum, L. P. Kaelbling, and M. Katz, *Generalized planning in pddl domains with pretrained large language models*, 2023. arXiv: 2305.11014 [cs.AI].

[55] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with

large language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2998–3009.

[56] C. Wang, S. Hasler, D. Tanneberg, *et al.*, "Lami: Large language models for multi-modal human-robot interaction," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ACM, May 2024, pp. 1–10. DOI: 10.1145/3613905.3651029. [Online]. Available: http://dx.doi.org/10.1145/3613905.3651029.

[57] R. A. Izzo, G. Bardaro, and M. Matteucci, *Btgenbot: Behavior tree generation for robotic tasks with lightweight llms*, 2024. arXiv: 2403.12761 [cs.RO].

[58] Y. Yang, J.-R. Gaglione, C. Neary, *et al.*, "Large language models for verifiable sequential decision-making in autonomous systems," in *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

[59] C. Belta, A. Bicchi, M. Egerstedt, E. Frazzoli, E. Klavins, and G. J. Pappas, "Symbolic planning and control of robot motion [grand challenges of robotics]," *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 61–70, 2007. DOI: 10.1109/MRA.2007.339624.

[60] R. Fu, J. Liu, X. Chen, Y. Nie, and W. Xiong, *Scene-llm: Extending language model for 3d visual understanding and reasoning*, 2024. arXiv: 2403.11401 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2403.11401.

[61] Z. Fountas, M. A. Benfeghoul, A. Oomerjee, *et al.*, *Human-like episodic memory for infinite context llms*, 2024. arXiv: 2407.09450 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2407.09450.

[62] T. O. Kvålseth, "An alternative to Fitts' law," *Bulletin of the psychonomic Society*, vol. 16, no. 5, pp. 371–373, 1980.

[63] E. Triantafyllidis and Z. Li, "The challenges in modeling human performance in 3D space with Fitts' law," *Proc. SIGCHI Conf. Hum. Factor Comput. Syst.*, pp. 1–9, 2021.

[64] A. Kulik, A. Kunert, and B. Froehlich, "On motor performance in virtual 3D object manipulation," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 5, pp. 2041–2050, 2020.

[65] M. D. Barrera Machuca and W. Stuerzlinger, "The effect of stereo display deficiencies on virtual hand pointing," *Proc. SIGCHI Conf. Hum. Factor Comput. Syst.*, pp. 1–14, 2019.

[66] C. Ware and R. Balakrishnan, "Reaching for objects in VR displays: Lag and frame rate," *ACM Trans Comput Hum Interact*, vol. 1, no. 4, pp. 331–356, 1994.

[67] E. Triantafyllidis and Z. Li, "Considerations and challenges of measuring operator performance in telepresence and teleoperation entailing mixed reality technologies," in *ACM CHI*, 2021, pp. 1–5.

[68] N. Mavridis, N. Giakoumidis, and E. L. Machado, "A novel evaluation framework for teleoperation and a case study on natural human-arm-imitation through motion capture," *International Journal of Social Robotics*, vol. 4, no. 1, pp. 5–18, 2012.

[69] J. Zhang, Z. Yin, and R. Wang, "Recognition of mental workload levels under complex human–machine collaboration by using physiological

features and adaptive support vector machines," *IEEE Trans Hum Mach Syst*, vol. 45, no. 2, pp. 200–214, 2014.

[70] K. Stowers, J. Oglesby, S. Sonesh, K. Leyva, C. Iwig, and E. Salas, "A framework to guide the assessment of human–machine systems," *Human factors*, vol. 59, no. 2, pp. 172–188, 2017.

[71] S. Hart and L. E. Staveland, "Development of NASA-TLX (task load index)," in *Advances in psychology*, vol. 52, 1988, pp. 139–183.

[72] R. A. Grier, "How high is high? A meta-analysis of NASA-TLX global workload scores," in *The Human Factors and Ergonomics Society Annual Meeting*, vol. 59, 2015, pp. 1727–1731.

[73] M. R. Endsley, S. J. Selcon, T. D. Hardiman, and D. G. Croft, "A comparative analysis of SAGAT and SART for evaluations of situation awareness," in *Proc Hum Factors Ergon Soc Annu Meet*, vol. 42, 1998, pp. 82–86.

[74] J. Brooke, "Sus: A 'quick and dirty' usability scale," in *Usability Evaluation In Industry*. CRC Press, 1996, ch. 21.

[75] A. Huber and A. Weiss, "Developing human-robot interaction for an industry 4.0 robot: How industry workers helped to improve remote-HRI to physical-HRI," in *ACM/IEEE HRI*, 2017, pp. 137–138.

[76] R. Bevilacqua, E. Felici, F. Marcellini, *et al.*, "Robot-Era project: Preliminary results on the system usability," in *International conference of design, user experience, and usability*, 2015, pp. 553–561.

[77] V. R. Garate, S. Gholami, and A. Ajoudani, "A scalable framework for multi-robot tele-impedance control," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 2052–2066, 2021.

[78] C. Stanton, A. Bogdanovych, and E. Ratanasena, "Teleoperation of a humanoid robot using full-body motion capture, example movements, and machine learning," in *Australasian Conf. on Robotics and Automation*, vol. 8, 2012, p. 51.

[79] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*, IEEE, 2012, pp. 5026–5033.

[80] E. Kolve, R. Mottaghi, W. Han, *et al.*, *Ai2-thor: An interactive 3d environment for visual ai*, 2022. arXiv: 1712.05474 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1712.05474.

[81] C. Li, R. Zhang, J. Wong, *et al.*, *Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation*, 2024. arXiv: 2403.09227 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2403.09227.

[82] T. L. Chen and C. C. Kemp, "Lead me by the hand: Evaluation of a direct physical interface for nursing assistant robots," in *ACM/IEEE HRI*, 2010, pp. 367–374.

[83] T. I. Chowdhury, S. M. S. Ferdous, and J. Quarles, "VR disability simulation reduces implicit bias towards persons with disabilities," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 6, pp. 3079–3090, 2019.

[84] J. Zhao, R. An, R. Xu, and B. Lin, "Comparing hand gestures and a gamepad interface for locomotion in virtual environments," *International Journal of Human-Computer Studies*, vol. 166, p. 102 868, 2022.

[85] M. Oshita and H. Ishikawa, "Gamepad vs. touchscreen: A comparison of action selection interfaces in computer games," in *Special Interest Group on Computer Graphics and Interactive Techniques Asia*, 2012, pp. 27–31.

[86] S. Gaglio, G. L. Re, and M. Morana, "Human activity recognition process using 3-d posture data," *IEEE Trans Hum Mach Syst*, vol. 45, no. 5, pp. 586–597, 2015.

[87] W. Takano, "Annotation generation from IMU-based human whole-body motions in daily life behavior," *IEEE Trans Hum Mach Syst*, vol. 50, no. 1, pp. 13–21, 2020.

[88] M. Field, Z. Pan, D. Stirling, and F. Naghdy, "Human motion capture sensors and analysis in robotics," *Industrial Robot*, vol. 38, no. 2, pp. 163–171, 2011.

[89] F. Schlagenhauf, P. P. Sahoo, and W. Singhose, "Comparison of single-kinect and dual-kinect motion capture of upper-body joint tracking," in *Asian Control Conference*, 2017, pp. 256–261.

[90] H. Zhou, G. Yang, H. Lv, X. Huang, H. Yang, and Z. Pang, "IoT-enabled dual-arm motion capture and mapping for telerobotics in home care," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1541–1549, 2019.

[91] J. Humphreys, C. Peers, J. Li, Y. Wan, and C. Zhou, "High utility teleoperation framework for legged manipulators through leveraging whole-body control," *Journal of Intelligent & Robotic Systems*, 2023.

[92] L. Penco, B. Clément, V. Modugno, *et al.*, "Robust real-time whole-body motion retargeting from human to humanoid," in *IEEE-RAS Humanoids*, 2018, pp. 425–432.

[93] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, "Progress and prospects of the human-robot collaboration," *Autonomous Robots*, vol. 42, pp. 957–975, 2018.

[94] A. Bauer, D. Wollherr, and M. Buss, "Human-robot collaboration: A survey," *International Journal of Humanoid Robotics*, vol. 5, no. 01, pp. 47–66, 2008.

[95] N. G. Tsagarakis, D. G. Caldwell, F. Negrello, *et al.*, "Walk-man: A high-performance humanoid platform for realistic environments," *Journal of Field Robotics*, vol. 34, no. 7, pp. 1225–1259, 2017.

[96] L. Peternel, N. Tsagarakis, and A. Ajoudani, "Towards multi-modal intention interfaces for human-robot co-manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016, pp. 2663–2669.

[97] C. Zhou, C. Peers, Y. Wan, R. Richardson, and D. Kanoulas, *TeLeMan: Teleoperation for legged robot loco-manipulation using wearable IMU-based motion capture*, arXiv:2209.10314, 2022.

[98] C. P. Chanel, R. N. Roy, F. Dehais, and N. Drougard, "Towards mixed-initiative human–robot interaction: Assessment of discriminative physiological and behavioral features for performance prediction," *Sensors*, vol. 20, no. 1, p. 296, 2020.

[99] M. S. Young and N. A. Stanton, "Malleable attentional resources theory: A new explanation for the effects of mental underload on performance," *Human Factors*, vol. 44, no. 3, pp. 365–375, 2002.

[100] S. D. Baulk, K. J. Kandelaars, N. Lamond, G. D. Roach, D. Dawson, and A. Fletcher, "Does variation in workload affect fatigue in a regular 12-hour shift system?" *Sleep and Biological Rhythms*, vol. 5, pp. 74–77, 2007.

[101] K. S. Gould, K. Hirvonen, V. F. Koefoed, *et al.*, "Effects of 60 hours of total sleep deprivation on two methods of high-speed ship navigation," *Ergonomics*, vol. 52, no. 12, pp. 1469–1486, 2009.

[102] P. A. Hancock and G. Matthews, "Workload and performance: Associations, insensitivities, and dissociations," *Human Factors*, vol. 61, no. 3, pp. 374–392, 2019.

[103] C. D. Smith, A. D. Cooper, D. J. Merullo, *et al.*, "Sleep restriction and cognitive load affect performance on a simulated marksmanship task," *Journal of Sleep Research*, vol. 28, no. 3, e12637, 2019.

[104] L. Peternel, N. Tsagarakis, D. Caldwell, and A. Ajoudani, "Robot adaptation to human physical fatigue in human–robot co-manipulation," *Autonomous Robots*, vol. 42, pp. 1011–1021, 2018.

[105] S. Hopko, R. Khurana, R. K. Mehta, and P. R. Pagilla, "Effect of cognitive fatigue, operator sex, and robot assistance on task performance metrics, workload, and situation awareness in human-robot collaboration," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3049–3056, 2021.

[106] M. Selvaggio, M. Cognetti, S. Nikolaidis, S. Ivaldi, and B. Siciliano, "Autonomy in physical human-robot interaction: A brief survey," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7989–7996, 2021.

[107] Q. Ji, P. Lan, and C. Looney, "A probabilistic framework for modeling and real-time monitoring human fatigue," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Syst. Hum*, vol. 36, no. 5, pp. 862–875, 2006.

[108] R. Fu, H. Wang, and W. Zhao, "Dynamic driver fatigue detection using hidden markov model in real driving condition," *Expert Systems with Applications*, vol. 63, pp. 397–411, 2016.

[109] G. Yang, Y. Lin, and P. Bhattacharya, "A driver fatigue recognition model based on information fusion and dynamic bayesian network," *Information Sciences*, vol. 180, no. 10, pp. 1942–1954, 2010.

[110] D. Tran, H. Manh Do, W. Sheng, H. Bai, and G. Chowdhary, "Real-time detection of distracted driving based on deep learning," *IET Intelligent Transport Systems*, vol. 12, no. 10, pp. 1210–1219, 2018.

[111] C. Ahlström, W. van Leeuwen, S. Krupenia, *et al.*, "Real-time adaptation of driving time and rest periods in automated long-haul trucking: Development of a system based on biomathematical modelling, fatigue and

relaxation monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4758–4766, 2022.

[112] A. Gundel, K. Marsalek, and C. Thoren, "A critical review of existing mathematical models for alertness," *Somnologie*, vol. 3, no. 11, pp. 148–156, 2007.

[113] H. Van Dongen, "Comparison of mathematical model predictions to experimental data of fatigue and performance," *Aviation, Space, and Environmental Medicine*, vol. 75, no. 3, A15–A36, 2004.

[114] D. F. Dinges, "Critical research issues in development of biomathematical models of fatigue and performance," *Aviation, Space, and Environmental Medicine*, vol. 75, no. 3, A181–A191, 2004.

[115] K. E. Friedl, M. M. Mallis, S. T. Ahlers, S. M. Popkin, and W. Larkin, "Research requirements for operational decision-making using models of fatigue and performance," *Aviation, Space, and Environmental Medicine*, vol. 75, no. 3, A192–A199, 2004.

[116] H. T. Peng, F. Bouak, W. Wang, R. Chow, and O. Vartanian, "An improved model to predict performance under mental fatigue," *Ergonomics*, vol. 61, no. 7, pp. 988–1003, 2018.

[117] S. Choudhury, J. K. Gupta, M. J. Kochenderfer, D. Sadigh, and J. Bohg, "Dynamic multi-robot task allocation under uncertainty and temporal constraints," *Autonomous Robots*, vol. 46, no. 1, pp. 231–247, 2022.

[118] M. Ahn, A. Brohan, N. Brown, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[119] K. Obata, T. Aoki, T. Horii, T. Taniguchi, and T. Nagai, "Lip-llm: Integrating linear programming and dependency graph with large language models for multi-robot task planning," *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 1122–1129, 2025. DOI: 10.1109/LRA.2024.3518105.

[120] R. Sakagami, S. G. Brunner, A. Dömel, A. Wedler, and F. Stulp, "Rosmc: A high-level mission operation framework for heterogeneous robotic teams," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5473–5479. DOI: 10.1109/ICRA48891.2023.10161133.

[121] M. Lippi, P. Di Lillo, and A. Marino, "A task allocation framework for human multi-robot collaborative settings," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 7614–7620. DOI: 10.1109/ICRA48891.2023.10161458.

[122] A. Monguzzi, M. Badawi, A. M. Zanchettin, and P. Rocco, "A mixed capability-based and optimization methodology for human-robot task allocation and scheduling," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022, pp. 1271–1276. DOI: 10.1109/RO-MAN53752.2022.9900823.

[123] C. Y. Kim, C. P. Lee, and B. Mutlu, "Understanding large-language model (llm)-powered human-robot interaction," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 371–380.

[124] L. X. Shi, Z. Hu, T. Z. Zhao, *et al.*, *Yell at your robot: Improving on-the-fly from language corrections*, 2024. arXiv: 2403.12910 [cs.RO].

[125] M. Han, Y. Zhu, S.-C. Zhu, Y. N. Wu, and Y. Zhu, *Interpret: Interactive predicate learning from language feedback for generalizable task planning*, 2024. arXiv: 2405.19758.

[126] I. Singh, V. Blukis, A. Mousavian, *et al.*, *Progprompt: Generating situated robot task plans using large language models*, 2022. arXiv: 2209.11302 [cs.RO].

[127] H. Liu, A. Chen, Y. Zhu, A. Swaminathan, A. Kolobov, and C.-A. Cheng, *Interactive robot learning from verbal correction*, 2023. arXiv: 2310.17555 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2310.17555.

[128] Y. Mu, Q. Zhang, M. Hu, *et al.*, *Embodiedgpt: Vision-language pre-training via embodied chain of thought*, 2023. arXiv: 2305.15021 [cs.RO].

[129] W. Huang, F. Xia, D. Shah, *et al.*, *Grounded decoding: Guiding text generation with grounded models for embodied agents*, 2023. arXiv: 2303.00855 [cs.RO].

[130] P. Pueyo, E. Montijano, A. C. Murillo, and M. Schwager, *Clipswarm: Generating drone shows from text prompts with vision-language models*, 2024. arXiv: 2403.13467 [cs.RO].

[131] Y. Cao and C. S. G. Lee, *Robot behavior-tree-based task generation with large language models*, 2023. arXiv: 2302.12927 [cs.RO].

[132] D. Tanneberg, F. Ocker, S. Hasler, *et al.*, *To help or not to help: Llm-based attentive support for human-robot group interactions*, 2024. arXiv: 2403.12533 [cs.RO].

[133] K. Chu, X. Zhao, C. Weber, M. Li, W. Lu, and S. Wermter, *Large language models for orchestrating bimanual robots*, 2024. arXiv: 2404.02018 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2404.02018.

[134] S. S. Kannan, V. L. N. Venkatesh, and B.-C. Min, "Smart-llm: Smart multi-agent robot task planning using large language models," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 12 140–12 147. DOI: 10.1109/IROS58592.2024.10802322.

[135] C. E. Mower, Y. Wan, H. Yu, *et al.*, *Ros-llm: A ros framework for embodied ai with task feedback and structured reasoning*, 2024. arXiv: 2406.19741 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2406.19741.

[136] K. Liu, Z. Tang, D. Wang, Z. Wang, B. Zhao, and X. Li, *Coherent: Collaboration of heterogeneous multi-robot system with large language models*, 2024. arXiv: 2409.15146 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2409.15146.

[137] Y. Chen, J. Arkin, Y. Hao, Y. Zhang, N. Roy, and C. Fan, *Prompt optimization in multi-step tasks (promst): Integrating human feedback and heuristic-based sampling*, 2024. arXiv: 2402.08702 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2402.08702.

[138] J. Wang, G. He, and Y. Kantaros, *Safe task planning for language-instructed multi-robot systems using conformal prediction*, 2024. arXiv: 2402.15368 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2402.15368.

[139] Z. Mandi, S. Jain, and S. Song, "Roco: Dialectic multi-robot collaboration with large language models," in *2024 IEEE International Conference on*

*Robotics and Automation (ICRA)*, 2024, pp. 286–299. DOI: 10.1109/ICRA57147.2024.10610855.

[140] Y. Chen, J. Arkin, C. Dawson, Y. Zhang, N. Roy, and C. Fan, "Autotamp: Autoregressive task and motion planning with llms as translators and checkers," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6695–6702. DOI: 10.1109/ICRA57147.2024.10611163.

[141] C. Zhou, Y. Wan, C. Peers, A. M. Delfaki, and D. Kanoulas, "Advancing teleoperation for legged manipulation with wearable motion capture," *Frontiers in Robotics and AI*, vol. 11, p. 1 430 842, 2024. DOI: 10.3389/frobt.2024.1430842.

[142] T. J. Balkin, A. R. Braun, N. J. Wesensten, *et al.*, "The process of awakening: A pet study of regional brain activity patterns mediating the reestablishment of alertness and consciousness," *Brain*, vol. 125, no. 10, pp. 2308–2319, 2002.

[143] P. Lavie, "The 24-hour sleep propensity function (spf): Practical and theoretical implications.," *Sleep, Sleepiness and Performance*, pp. 65–93, 1991.

[144] G. Belenky, N. J. Wesensten, D. R. Thorne, *et al.*, "Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study," *Journal of Sleep Research*, vol. 12, no. 1, pp. 1–12, 2003.

[145] C. M. Jung, J. M. Ronda, C. A. Czeisler, and K. P. Wright Jr, "Comparison of sustained attention assessed by auditory and visual psychomotor

vigilance tasks prior to and during sleep deprivation," *Journal of Sleep Research*, vol. 20, no. 2, pp. 348–355, 2011.

[146] S. Folkard, K. A. Robertson, and M. B. Spencer, "A fatigue/risk index to assess work schedules," *Somnologie-Schlafforschung und Schlafmedizin*, vol. 11, no. 3, pp. 177–185, 2007.

[147] C. Pinheiro, J. Figueiredo, N. Magalhães, and C. P. Santos, "Wearable biofeedback improves human-robot compliance during ankle-foot exoskeleton-assisted gait training: A pre-post controlled study in healthy participants," *Sensors*, vol. 20, no. 20, p. 5876, 2020.

[148] J. Humphreys and C. Zhou, *Learning to adapt: Bio-inspired gait strategies for versatile quadruped locomotion*, 2024. arXiv: 2412.09440 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2412.09440.

# Appendix A

# Ethics

This appendix includes the documentation related to the ethical approval for the studies conducted as part of this research. All experiments involving human participants were reviewed and approved by the appropriate ethics committee to ensure compliance with institutional guidelines and ethical standards. The approval covered participant recruitment, informed consent procedures, data collection, and data handling protocols.

# A.1 Ethics approval

The related human study has been approved by the Engineering and Physical Science Ethics (EPS/FREC) Committee.

Dear Yuhui

**MEEC 21-027 - Robot shared autonomy human study**

***NB: All approvals/comments are subject to compliance with current University of Leeds and UK Government advice regarding the Covid-19 pandemic.***

I am pleased to inform you that the above research ethics application has been reviewed by the Engineering and Physical Science Ethics (EPS/FREC) Committee and on behalf of the Chair, I can confirm a conditional favourable ethical opinion based on the documentation received at date of this email and *subject to the following condition/s which must be fulfilled prior to the study commencing:*

1. ***Please complete Section C21 i.e. box ticked for storage in University approved cloud storage.***

The study documentation must be amended where required to meet the above conditions and submitted for file and possible future audit.

*Once you have addressed the conditions and submitted for file/future audit, you may commence the study and further confirmation of approval is not provided.*

*Please note*, failure to comply with the above conditions will be considered a breach of ethics approval and may result in disciplinary action.

***Please retain this email as evidence of conditional approval in your study file.***

Please notify the committee if you intend to make any amendments to the original research as submitted and approved to date. This includes recruitment methodology; all changes must receive ethical approval prior to implementation. Please see https://ris.leeds.ac.uk/research-ethics-and-integrity/applying-for-an-amendment/ or contact the Research Ethics & Governance Administrator for further information on researchethics@leeds.ac.uk if required.

Ethics approval does not infer you have the right of access to any member of staff or student or documents and the premises of the University of Leeds. Nor does it imply any right of access to the premises of any other organisation, including clinical areas. The committee takes no responsibility for you gaining access to staff, students and/or premises prior to, during or following your research activities.

*Please note:* You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, risk assessments and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited.

It is our policy to remind everyone that it is your responsibility to comply with Health and Safety, Data Protection and any other legal and/or professional guidelines there may be.

I hope the study goes well.

Best wishes
Rachel P
***On behalf Dr Jim Young, CHAIR, EPS***

# Appendix B

# Human Study Forms

This appendix includes the questionnaires used in the user study, specifically the background questionnaire, the NASA Task Load Index (NASA-TLX), and the System Usability Scale (SUS).

# B.1 Background questionnaire

Table B.1: Questions used in the participant background questionnaire.

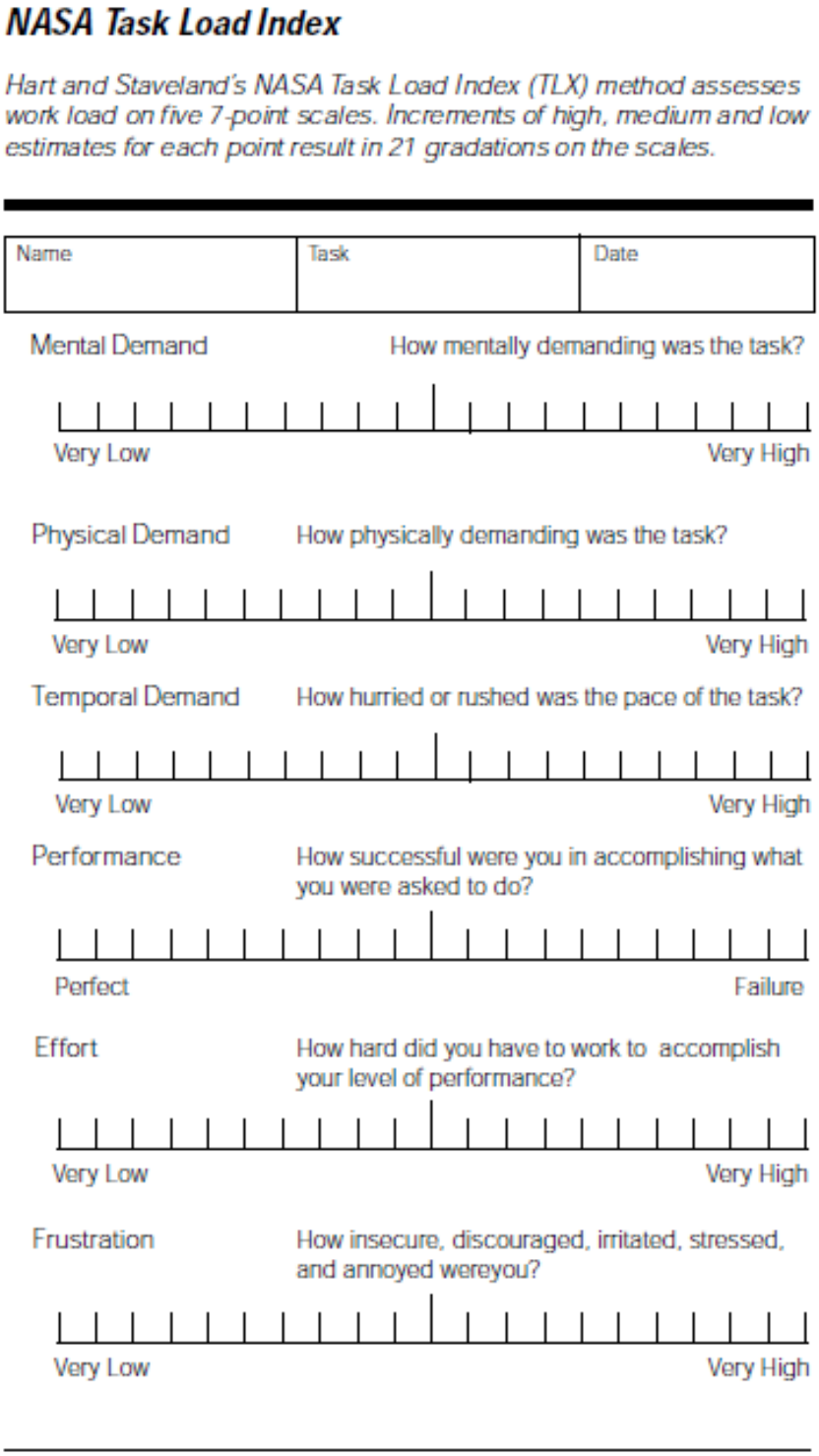
| ID | Question |
|---|---|
| Q1 | How much experience do you have with gaming joysticks? |
| Q2 | How much experience do you have with the motion-capture suit? |
| Q3 | How much experience do you have with the VR display? (If your answer is 1 please skip next question) |
| Q4 | How do you feel during your experience using VR? |
| Q5 | Do you have robot remote control experience? |

## B.1.1 Participant Responses

Table B.2: Participant responses to the questionnaire (Likert scale: 1 = very low / none, 5 = very high / extensive).

| ID | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| A1 | 5 | 1 | 4 | 4 | 5 |
| A2 | 4 | 1 | 3 | 3 | 5 |
| A3 | 5 | 1 | 2 | 4 | 5 |
| B1 | 1 | 3 | 3 | 5 | 3 |
| B2 | 1 | 1 | 2 | 2 | 1 |
| B3 | 5 | 1 | 2 | 3 | 3 |
| B4 | 2 | 2 | 3 | 3 | 3 |

## B.2 NASA-TLX



***NASA Task Load Index***

*Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.*

| Name | Task | Date |
|---|---|---|
| | | |

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Figure B.1: NASA Task Load Index (NASA-TLX) form used for subjective workload assessment used in the study.

## B.3 System Usability Scale

| No. | SUS Questions |
|---|---|
| 1 | I think that I would like to use this system frequently. |
| 2 | I found the system unnecessarily complex. |
| 3 | I thought the system was easy to use. |
| 4 | I think that I would need the support of a technical person to be able to use this system. |
| 5 | I found the various functions in this system were well integrated. |
| 6 | I thought there was too much inconsistency in this system. |
| 7 | I would imagine that most people would learn to use this system very quickly. |
| 8 | I found the system very cumbersome to use. |
| 9 | I felt very confident using the system. |
| 10 | I needed to learn a lot of things before I could get going with this system. |

Table B.3: System Usability Scale (SUS) Questionnaire

# Appendix C

# Prompt used in FittsPrompt

For reproducibility, this appendix provides the exact prompts used in the two steps of FittsPrompt. The prompts were structured to ensure consistency in model behaviour, encourage reasoning over task difficulty values, and enforce output in a machine-readable format.

## C.1 Step 1: Capability Filtering

```
You are an expert robot planning assistant.
Your task is to create a plan for the robot to achieve its goal
efficiently and effectively.
Use principles from robotics planning instruction above to ensure
a structured and logical approach.

If there are multiple agent of the available to achieve the task,
do not out put your plan instead indicating there are more than
one agent.
And output the object's id and agent's id as following
{
    "object_id":object_id,
    "agent_id":agent_id,
    "agent_id":agent_id,
    "agent_id":agent_id
}

Else please output your plan in the form of:

[
    {"agent_id": agent_id, "action": action, "object_id": object_id},
    ......
    {"agent_id": agent_id, "action": action, "object_id": object_id},
    {"agent_id": agent_id, "action": action, "object_id": object_id}
]
```

## C.2 Step 2: Difficulty-Aware Selection

```
You are an expert robot planning assistant.
Your task is to create a plan for the robot to achieve its goal
efficiently and effectively. Use principles from robotics planning
instruction above to ensure a structured and logical approach.

Think about your plan, make sure you select object with the
Smaller Difficulty when multiple item available.
Please output your plan in the form of:

[
    {"agent_id": agent_id, "action": action, "object_id": object_id},
    ......
    {"agent_id": agent_id, "action": action, "object_id": object_id},
    {"agent_id": agent_id, "action": action, "object_id": object_id}
]
```

# Appendix D

# LLM Benchmark Results

# D.1 Multi-robot Task Allocation Results

| Model | Raw data Success Rate | FittsPrompt Success Rate | Raw data Optimal Rate | FittsPrompt Optimal Rate |
|---|---|---|---|---|
| qwen2.5-7b-instruct | 50% | 60% | 30% | 40% |
| qwen2.5-coder-7b-instruct | 30% | 0% | 20% | 0% |
| internlm3-8b-instruct | 20% | 30% | 0% | 20% |
| llama-3.1-8b-instruct | 30% | 40% | 30% | 20% |
| codegeex4-all-9b | 70% | 40% | 30% | 0% |
| qwen2.5-14b-instruct | 80% | 60% | 20% | 30% |
| internvl2.5-26b | 30% | 60% | 20% | 40% |
| deepseek-r1-distill-qwen-32b | 60% | 60% | 50% | 40% |
| qwen2.5-32b-instruct | 40% | 100% | 0% | 100% |
| qwen2.5-32b-instruct-vllm | 60% | 90% | 0% | 90% |
| qwq-32b-preview | 30% | 50% | 30% | 10% |
| qwq-32b-preview-awq | 60% | 80% | 40% | 40% |
| internvl2.5-38b | 100% | 80% | 50% | 70% |
| llama-3.1-70b | 80% | 70% | 40% | 70% |
| llama-3.1-70b-instruct | 100% | 80% | 30% | 80% |
| llama-3.1-nemotron-70b-instruct | 40% | 70% | 40% | 70% |
| llama-3.3-70b-instruct | 90% | 80% | 30% | 80% |
| qwen2-72b-32k | 80% | 80% | 40% | 40% |
| qwen2-vl-72b-instruct | 80% | 90% | 20% | 50% |
| qwen2.5-72b-32k | 60% | 100% | 10% | 90% |
| qwen2.5-72b-instruct | 40% | 90% | 20% | 80% |
| qwen2.5-72b-instruct-lmdeploy | 50% | 90% | 10% | 90% |
| internvl2.5-78b | 70% | 80% | 20% | 80% |
| gpt-3.5-turbo-0125 | 80% | 40% | 20% | 30% |
| gpt-4o-2024-08-06 | 80% | 100% | 40% | 80% |
| gpt-4o-mini-2024-07-18 | 60% | 100% | 10% | 100% |
| o1-preview | 100% | 90% | 70% | 90% |
| Average | 62% | 71% | 27% | 57% |

Table D.1: Task allocation Success Rates and Optimal Rate of Models under Raw and FittsPrompt Conditions

## D.2 Multi-robot Task Execution Results

| Model | Raw data Success Rate | FittsPrompt Success Rate | Raw data Optimal Rate | FittsPrompt Optimal Rate |
|---|---|---|---|---|
| qwen2.5-3b-instruct | 0% | 30% | 0% | 20% |
| qwen2.5-1.5b-instruct | 0% | 80% | 0% | 30% |
| qwen2.5-7b-instruct | 80% | 100% | 40% | 40% |
| qwen2.5-coder-7b-instruct | 100% | 10% | 40% | 10% |
| ui-tars-7b-dpo | 40% | 20% | 30% | 10% |
| internlm3-8b-instruct | 40% | 70% | 30% | 20% |
| llama-3.1-8b-instruct | 80% | 90% | 30% | 80% |
| codegeex4-all-9b | 0% | 10% | 0% | 10% |
| qwen2.5-14b-instruct | 90% | 90% | 40% | 90% |
| internvl2.5-26b | 100% | 90% | 40% | 30% |
| deepseek-r1-distill-qwen-32b | 70% | 100% | 40% | 100% |
| qwen2.5-32b-instruct | 40% | 100% | 20% | 100% |
| qwen2.5-32b-instruct-vllm | 60% | 100% | 20% | 100% |
| qwq-32b-preview | 70% | 70% | 60% | 50% |
| qwq-32b-preview-awq | 60% | 90% | 50% | 90% |
| internvl2.5-38b | 60% | 100% | 20% | 100% |
| llama-3.1-70b | 100% | 100% | 50% | 100% |
| llama-3.1-70b-instruct | 100% | 100% | 40% | 100% |
| llama-3.1-nemotron-70b-instruct | 0% | 100% | 0% | 100% |
| llama-3.3-70b-instruct | 80% | 100% | 40% | 100% |
| qwen2-72b-32k | 90% | 100% | 40% | 100% |
| qwen2-vl-72b-instruct | 100% | 100% | 40% | 100% |
| qwen2.5-72b-32k | 20% | 100% | 0% | 100% |
| qwen2.5-72b-instruct | 20% | 100% | 0% | 90% |
| qwen2.5-72b-instruct-lmdeploy | 0% | 90% | 0% | 90% |
| internvl2.5-78b | 80% | 100% | 20% | 100% |
| gpt-3.5-turbo-0125 | 100% | 100% | 40% | 30% |
| gpt-4o-2024-08-06 | 80% | 100% | 20% | 100% |
| gpt-4o-mini-2024-07-18 | 70% | 100% | 30% | 80% |
| gpt-o1-preview | 100% | 100% | 60% | 100% |
| Average | 61% | 85% | 28% | 72% |

Table D.2: Task execution Success Rates and Optimal Rate of Models under Raw and FittsPrompt Conditions