DTS311TC FINAL YEAR PROJECT

Player-Aware Intelligent Monitoring and Operations Navigator

Proposal Report

In Partial Fulfillment
of the Requirements for the Degree of
Bachelor of Engineering

Student Name	:	Taimingwang Liu
Student ID	:	2037690
Supervisor	:	Xihan Bian

School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University

November 2025

Abstract

Apply the font of Times New Roman to the paragraphs of the abstract using font size of 12. An abstract is usually one to three paragraphs long with a length of 150 to 350 words.

Contents

1 Introduction

The demand for **interactive and companion experiences** in real-time entertainment is on the rise. Players are no longer satisfied with simple automation or static overlays; they seek dynamic, engaging partners that offer meaningful interaction. As gaming experiences evolve, the focus is shifting towards assistants that not only provide practical help but also enrich the player's journey through companionship and engagement.

Games like AI2U: "With You 'Til The End" are already capitalizing on the demand for interactive experiences, offering players the novelty of engaging with generative AI characters [1]. At the same time, major companies such as NVIDIA, with its Avatar Cloud Engine (ACE), and Ubisoft, with its "NEO NPCs", are advancing foundational technologies to create autonomous agents that enhance gameplay by offering more than just conversation [2], [3]. These developments demonstrate the growing commercial viability of AI-driven experiences, supported by increased player engagement and media attention [4].

The rise of AI-driven virtual streamers, especially the Neuro-sama phenomenon, highlights a significant shift in both technology and community-driven commercialization. Neuro-sama, an AI-powered VTuber, engages in real-time conversations and dynamic gameplay, capturing the attention of a wide audience [5]–[7]. Although Neuro-sama remains closed-source, its success has sparked a vibrant open-source ecosystem, with developers working to replicate or expand upon its capabilities [8]–[10]. This technological shift has been met with strong commercial traction on platforms like Twitch, underscoring the growing market demand for AI that combines both utility and companionship [11].

The success of these AI-driven personalities validates the market's desire for this combination of utility and companionship; this project builds upon this principle, seeking to translate this proven engagement from a broadcast-entertainment model into a player-centric, companion-style assistant. This assistant is designed to augment, not automate, the player's agency. Its role is distinct from broadcast-focused AI-Vtubers and player-replacing automation bots; it is scoped as a persistent, in-game partner that uses a **full voice-loop** to provide support that is both relational and functional. The system's core capabilities are threefold: offering **proactive, contextual companionship** by using in-character awareness to identify opportunities (e.g., spotting missed items); engaging in **collaborative problem-solving** to understand player goals and access external knowledge (e.g., researching online strategies); and supporting **on-demand task delegation**, allowing the player to hand over control for specific, well-defined tasks (e.g., 'explore this area while I'm away').

The technical scope to achieve these capabilities involves three key techniques. First, the system will use **a Unified Action Interface via GUI (GCC)**, a human-homomorphic interface with a screen-in, keyboard/mouse-out paradigm. This approach eliminates the need for game-specific APIs and ensures cross-platform adaptability [12], [13]. Second, actions will be managed through **Constrained Action Generation with Structured Output**. This method, which

formats commands from a predefined set of valid actions (e.g., 'move forward') into structured output (e.g., JSON), reduces errors like hallucinations and ensures that delegated tasks are legal and reproducible [14]. Third, the system will employ **a Low-Coupling Orchestration** (i.e., modularization by MCP-style). This plug-and-play approach is crucial, ensuring both scalability for integrating new skills like problem-solving and task-delegation, and testability which enables systematic ablation studies to evaluate each component's contribution [15].

With a clear understanding of market demand and technological feasibility, this project aims to develop a companion-style assistant that seamlessly integrates into gameplay. By demonstrating a system capable of this deeper, functional partnership—blending both companionship and shared agency—this work will enhance the overall player experience through dynamic, interactive assistance.

2 Literature Review

This section reviews the foundational research underpinning the development of AI-driven game companions. It first establishes the technical feasibility of using human-homomorphic interfaces for game-agnostic control. Following this, it examines the evaluation protocols and performance metrics critical for benchmarking such systems. The review then explores the core agentic modules—planning, memory, and reflection—necessary for maintaining stability in long-horizon tasks. Subsequently, it details the modern learning paradigms required for robust agent training, before concluding with a discussion of persistent challenges, such as latency and safety, and a synthesis of the key literature gaps this project aims to address.

2.1 Feasibility of Human-Homomorphic Interfaces

Achieving the proposed unified action interface requires grounding the agent in **human-homomorphic interfaces**, a paradigm often referred to as General Computer Control (GCC). This approach uses screen-in, keyboard/mouse-out interaction to bypass the need for game-specific APIs, which is foundational for a truly game-agnostic assistant [12]. The feasibility of this paradigm, however, rests on solving key challenges in stability, representation, and perception. A primary challenge is ensuring long-term stability, which the CRADLE framework addresses by enforcing **legal move constraints**. By restricting the agent to a valid-action set and enforcing keyboard/mouse action validity, it significantly reduces error accumulation and demonstrates improved generalization across unseen games [12].

Beyond just constraining moves, the representation of both actions and states is a non-trivial factor in cross-game performance. This is corroborated by the ORAK benchmark, which emphasizes that the **move representation format**—specifically, the methods used for action mapping—directly affects an agent's ability to generalize across diverse game environments [15]. This requirement for clear representation extends to the agent's understanding of the game state. The UI-Venus framework highlights that **state representation**—specifically, structured output at multiple granularities such as format, type, and coordinates—is crucial for improving reliability when controlling UI elements, leading to more dependable agent behaviour [13].

While these works provide a clear path forward for API-free interaction, this "vision-in, action-out" model is not without its perceptual challenges. The V-MAGE benchmark, which is designed to assess vision-centric capabilities, demonstrates that **vision-only inputs** pose significant generalization challenges for game-playing agents, particularly in handling novel scenes or tracking objects [16].

2.2 Evaluation Protocols & Performance Metrics

To rigorously measure the performance of AI assistants in complex, dynamic environments, a new generation of evaluation frameworks and specific metrics has been established.

Proper benchmarking is essential for systematically evaluating AI performance across diverse game types and tasks, with systems like ORAK and lmgame-bench providing comprehensive frameworks for ensuring consistent and reliable results [15], [17]. These frameworks guide the design of scaffolding and feedback mechanisms that enhance AI stability.

Within these evaluations, researchers rely on a suite of metrics to assess an agent's effectiveness, efficiency, and error rate. General metrics such as `pass@k`, for instance, are commonly used to measure how many attempts an AI needs to successfully complete a task over repeated trials. More specific to GUI agents, metrics like `Invalid%` track errors by measuring the proportion of invalid actions taken by the AI. This is a critical diagnostic, as a high `Invalid%` indicates the model is struggling to produce valid outputs within the specified action space [14].

Beyond simple success or error rates, recent work has introduced more sophisticated systems for comparative and granular assessment. To address the difficulty of comparing performance across different games, V-Mage introduces a **Dynamic ELO system** to standardize agent performance [16]. The same framework also moves beyond a single success score by using **"Unit Tests for Core Visual Abilities"**, a method for providing a granular assessment of an agent's specific skills, such as positioning, trajectory tracking, and visual memory [16].

2.3 Agentic Modules and Stability in Long-Horizon Tasks

For an assistant to be a reliable, long-term partner, it must employ a set of **agentic modules**—such as planning, memory, and reflection—to manage complex tasks and mitigate error accumulation. While single-step actions are relatively simple, maintaining stability over long-horizon tasks requires a robust architecture where these modules can handle complex, sequential decision-making.

The most critical of these modules are **planning and reflection**, which work in a continuous loop. Planning, often achieved through **task segmentation**, involves decomposing complex goals into manageable sub-tasks, while reflection allows the agent to self-correct and adapt to evolving situations [12], [18]. Frameworks like CRADLE demonstrate how an agent can adjust its plan iteratively during a task based on self-reflection. This is supported by surveys like OS-Agents, which identify the ability to adapt plans based on **environmental feedback** as a cornerstone of modern agent design [12], [18].

This feedback loop is heavily reliant on **memory**. Both short-term memory (e.g., recent actions) and long-term memory (e.g., learned strategies) are essential for stabilizing performance and preventing task drift. The importance of this module is highlighted by benchmarks like ORAK, which is designed to test long-term memory in adventure games, and lmgame-bench, which explicitly provides **memory scaffolding** to even allow LLMs to attempt long-horizon tasks [15], [17].

To make long-term planning tractable, agents also rely on a curated set of **atomic skills**.

The CRADLE framework, for example, introduces **"Skill Curation"** as a core module, where the agent dynamically generates and updates skills that can be recombined to solve complex tasks [12]. This modularity is complemented by advanced refinement techniques. The UI-Venus report introduces a **"Self-Evolving Trajectory History Alignment"** framework, which acts as a form of automated reflection. By re-evaluating and filtering its own past "thought-action pairs," the agent refines its historical context, leading to "more coherent planning" and improved performance over time [13].

Collectively, these modules of planning, memory, and skill curation are not just for task completion but are essential for **error mitigation**. The feedback loops described in the OS-Agents survey and the memory scaffolds in lmgame-bench are designed to prevent the agent from losing focus or drifting off-task, ensuring it can adjust its approach in real-time to maintain stability during long sessions [17], [18].

2.4 Learning Paradigms for Robust Agent Training

A stable agent architecture must be supported by a robust training paradigm that bridges the gap between single-step imitation and sustained, interactive assistance. While many foundational models are trained via supervised imitation (e.g., behavior cloning), this paradigm struggles with the reward sparsity and long-horizon, multi-step nature of interactive tasks. Consequently, modern agent systems are increasingly applying **Reinforcement Fine-Tuning (RFT)** on top of pre-trained models. This approach, as demonstrated in works like UI-Venus, AgentGym-RL, and research on autonomous driving, has proven essential for adapting agents to the complex, sequential decision-making required in dynamic environments [13], [19], [20].

This shift to RFT is enabled by two key mechanisms: sophisticated reward shaping and iterative refinement loops. To overcome sparse rewards, systems employ a combination of **dense rewards** (feedback at each timestep) and **auxiliary rewards** (feedback for sub-goals or correct formatting). This reward structure is shown to accelerate learning and improve task performance, especially in complex, long-horizon scenarios [19], [20]. This is then combined with iterative **feedback loops** that allow the agent to refine its strategy over time. Examples include the "progressive interaction scaling" proposed by AgentGym-RL, or the "self-evolving trajectories" of UI-Venus, both of which are methods for the agent to correct mistakes and adapt its policy [13], [19].

Finally, a critical component for reliability, especially in GUI environments, is the use of **structured action generation**. To reduce errors and constrain the vast output space of a large language model, agents are trained to generate actions in a highly structured form, such as JSON or discrete macro actions, rather than as free-form text. The UI-Venus framework, for example, relies on this approach to ensure that generated actions are valid and reliable, which is a crucial step in preventing model-induced errors during execution [13].

2.5 Challenges and Future Directions

Despite the rapid progress in agent capabilities, several persistent challenges must be addressed to develop truly robust, real-time AI companions. The most immediate of these is **latency**, which is critical for a seamless user experience in interactive entertainment. Research into on-device processing, such as the UI-Venus framework, aims to solve the "unacceptably long inference latency" of large models [13]. This is a non-trivial requirement; studies in real-time human-AI coordination have empirically demonstrated that latency beyond approximately 100ms is perceived as lag and significantly degrades the sense of fluid cooperation [21].

Beyond performance, significant questions of **safety and robustness** remain. For an agent to be trusted with any level of control, it must be able to handle dangerous actions and critical errors. The OS-Agents survey identifies "Error Recovery Mechanisms" (such as rollbacks) and "Human-in-the-Loop Control" (such as confirmation mechanisms) as major unsolved challenges for the field [18]. This need for alignment is a key research area, focusing on robustifying agents to ensure safe "transfer" to new situations and prevent unintended consequences [22].

A primary cause of such errors is the agent's difficulty with **out-of-distribution (OOD)** generalization. Both V-Mage and the Benchmarking-VLA-VLM paper identify OOD generalization as a critical failure point [14], [16]. V-Mage, for example, highlights that vision-only models struggle significantly to adapt to new visual scenarios [16]. This finding is not limited to vision; the Benchmarking-VLA-VLM study found that all evaluated models (VLAs and VLMs) had "significant limitations in zero-shot generalization to OOD tasks," noting that their performance was heavily influenced by factors like action representation and prompt engineering [14].

These challenges collectively define the **future directions** for the field. As outlined in the CRADLE framework, research must prioritize enhancing an agent's ability to handle complex, multi-step tasks while ensuring real-time reliability. This requires enhancing **multi-modal capabilities**, improving accuracy in fine-grained control, and solving the "prohibitive inference latency" of current systems, all while ensuring agents remain resilient in unseen environments [12].

2.6 Synthesis and Gaps in the Literature

This review reveals that while significant progress has been made in functional agent control, a clear and actionable gap exists in the literature regarding unified, companion-style assistants. Current research is largely bifurcated, prioritizing either high-performance task completion or the fundamentals of relational AI, with a significant lack of academic frameworks that effectively merge the two. For instance, while AI systems show growing competence in task execution, recent work such as the **Emotion AWARE** framework highlights that integrating "multi-granular and explainable" **emotional intelligence** remains a complex, unsolved challenge—a feature largely absent in current agent architectures [23]. This gap is evident in state-

of-the-art functional systems; frameworks like UI-Venus have made strides in performance via RFT and structured actions, but their architecture is exclusively focused on functional task completion and does not address this relational component [13].

This project is therefore positioned to fill this gap by creating an assistant that architecturally blends both companionship and agency. In doing so, it will also address two other persistent challenges in the field: **real-time adaptability** and **cross-game transferability**. The former remains a recognized challenge, with benchmarks such as lmgame-bench being specifically designed to test the limitations of current models in spatiotemporal and long-context reasoning [17]. The latter, the difficulty of creating **game-agnostic agents**, has been pointed out by numerous studies. Both the ORAK benchmark, which highlights the difficulty of evaluating agents across diverse games, and the CRADLE framework, which was built to solve the reliance on proprietary game APIs, confirm that cross-game transferability is a key hurdle for the field [12], [15].

By focusing on a human-homomorphic interface and feedback-driven learning, this project will directly contribute to these unsolved areas, demonstrating a system capable of the deeper, functional partnership that current literature lacks.

References

[1] AlterStaff, *Ai2u: With you 'til the end*, https://store.steampowered.com/app/2880730/AI2U_With_You_Til_The_End/, Accessed: 2025-10-10, 2025.

[2] NVIDIA. "Nvidia ace for games - autonomous game characters." Accessed: 2025-11-05, NVIDIA Developer. [Online]. Available: https://developer.nvidia.com/ace-for-games.

[3] Ubisoft. "How ubisoft's new generative ai prototype 'neo npcs' changes the narrative." Accessed: 2025-11-05, Ubisoft News. [Online]. Available: https://news.ubisoft.com/en-gb/article/5qXdxhshJBXoanFZApdG3L/how-ubisofts-new-generative-ai-prototype-changes-the-narrative-for-npcs.

[4] J. Kim. "Bringing personality to pixels, inworld levels up game characters using generative ai," NVIDIA Blog. [Online]. Available: https://blogs.nvidia.com/blog/generative-ai-npcs/.

[5] Vedal and Neuro-sama, *Neuro-sama official youtube channel*, https://www.youtube.com/@Neurosama, Accessed: 2025-10-10, 2022.

[6] StreamElements. "State of the stream: Twitch 2024 year in review." Accessed: 2025-10-11. [Online]. Available: https://blog.streamelements.com/state-of-the-stream-twitch-2024-year-in-review-ef4d739e9be9.

[7] "Q4 2024 global live streaming landscape." Accessed: 2025-10-11, Streams Charts. [Online]. Available: https://streamscharts.com/news/q4-2024-global-livestreaming-landscape.

[8] O.-L.-V. contributors, *Open-llm-vtuber: An open-source ai vtuber framework*, https://github.com/Open-LLM-VTuber/Open-LLM-VTuber, Accessed: 2025-10-10, 2025.

[9] moeru-ai, *Airi: Ai waifu / virtual character container inspired by neuro-sama*, https://github.com/moeru-ai/airi, Accessed: 2025-10-10, 2025.

[10] kimjammer, *Neuro: A local-model recreation of neuro-sama*, https://github.com/kimjammer/Neuro, Accessed: 2025-10-10, 2025.

[11] "Vedal's ai vtuber neuro-sama sets new twitch hype train world record." Accessed: 2025-10-11, Streams Charts. [Online]. Available: https://streamscharts.com/news/vedals-ai-vtuber-neuro-twitch-hype-train-record.

[12] W. Tan *et al.*, "Cradle: Empowering foundation agents towards general computer control," *arXiv preprint arXiv:2403.03186*, 2024.

[13] Z. Gu *et al.*, "Ui-venus technical report: Building high-performance ui agents with rft," *arXiv preprint arXiv:2508.10833*, 2025.

[14] P. Guruprasad, Y. Wang, S. Chowdhury, H. Sikka, and P. P. Liang, "Benchmarking vision, language, & action models in procedurally generated, open ended action environments," *arXiv preprint arXiv:2505.05540*, 2025.

[15] D. Park *et al.*, "Orak: A foundational benchmark for training and evaluating llm agents on diverse video games," 2025, arXiv:2506.03610. arXiv: 2506.03610.

[16] X. Zheng *et al.*, "V-mage: A game evaluation framework for assessing vision-centric capabilities in multimodal large language models," *arXiv preprint arXiv:2504.06148*, 2025.

[17] L. Hu *et al.*, "Lmgame-bench: How good are llms at playing games?, 2025a," *URL https://arxiv. org/abs/2505.15146*,

[18] X. Hu *et al.*, *Os agents: A survey on mllm-based agents for computer, phone and browser use*, 2024.

[19] Z. Xi *et al.*, "Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning," *arXiv preprint arXiv:2509.08755*, 2025.

[20] Z. Peng *et al.*, "Improving agent behaviors with rl fine-tuning for autonomous driving," in *European Conference on Computer Vision*, Springer, 2024, pp. 165–181.

[21] J. Liu *et al.*, "Llm-powered hierarchical language agent for real-time human-ai coordination," *arXiv preprint arXiv:2312.15224*, 2023.

[22] M. Zubia, T. D. Simão, and N. Jansen, "Robustifying RL agents for safe transfer through action disturbances," in *Proceedings of the BNL Conference on Artificial Intelligence and Machine Learning (BNAIC/BeNeLearn 2024)*, 2024. [Online]. Available: https://bnaic2024.sites.uu.nl/wp-content/uploads/sites/986/2024/10/Robustifying-RL-Agents-for-Safe-Transfer-through-Action-Disturbances.pdf.

[23] G. Gamage, D. De Silva, N. Mills, D. Alahakoon, and M. Manic, "Emotion aware: An artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis," *Journal of Big Data*, vol. 11, no. 1, p. 93, 2024.

Appendix A. Title of Appendix A

A.1 Appendix Heading 1

Text of the appendix goes here

A.2 Appendix Heading 2

Text of the appendix goes here

A.3 Appendix Table and Figure Captions

In appendices, table and figure caption labels and numbers are typed in manually (e.g., Table A1, Table A2, etc.). These do not get generated into the lists that appear after the Table of Contents.

Appendix B. Title of Appendix B

Text of the appendix goes here if there is only a single heading.