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

1.1 Problem Setting & Motivation

近年来，面向玩家的智能交互快速涌现：从*AI*游戏主播/虚拟角色到*LLM*驱动的*NPC*/插件，社区与产业案例显示“大模型×游戏交互”具备显著关注度与潜在影响（game changer potential）。然而，这些实践多依赖场景定制，缺乏统一动作接口与可复现实验协议；相应研究正在通过统一评测/模块消融与污染控制/协议一致等方法学加以弥合[1], [2]。本文讨论伴随式（*companion-style*）实时助手，并将相关问题置于仅基于*GUI*（*GCC*）与低延迟体验的工作设定下加以界定。

Industry/Community Signals 在社区层面，*Neuro-sama* 的*AI* 游戏主播（*AI streamer*）现象展示了大模型驱动的持续互动与情绪共鸣能力[3]；同时，叙事解谜作品*AI2U: With You 'Til The End* 体现了“对话即操作（*dialogue-as-action*）”与高交互度（LLM-controlled NPCs）的设计潜力[4]。围绕该方向的开源复现与二次开发——如*Open LLM VTuber*、*Airi Project*、*Kimjammer-Neuro* 等——持续出现，反映出应用生态的活跃[5]–[7]。

作为多模态交互智能体（Agent AI）的概念性综述，[8] 将“具身环境中的下一步动作预测（next-embodied action prediction）”与外部知识、人类反馈、多传感输入并置讨论，并提出在*grounded*场景中可缓解大模型“幻觉”问题的观点[8].

作为面向GUI 自动化的综述，[9] 系统梳理了以LLM 为“中枢”的GUI 智能体在框架、训练数据与大动作模型（LAM）、以及评测基准与指标方面的进展与挑战[9].

1.2 Scope & Working Definitions

本文将动作接口（action interface）的工作定义限定为**General Computer Control (GCC)** 的*GUI*范式：*screen-in, keyboard/mouse-out* 的人类同态接口（human-homomorphic interface）。代表性工作*CRADLE* 报告了在不依赖应用API 的前提下完成长链路桌面/游戏任务的可行性与系统结构（规划/技能整理/反思/记忆）[10]。此外，*Model Context Protocol (MCP)* 可作为内部模块/技能编排（registration/orchestration）的通用思路，与具体的输出通道无直接绑定；“统一评测/消融”与*plug-and-play* 思路可见*ORAK*[1]；基于*procedural generation* 的OOD 方法学可见*Benchmarking-VLA-VLM*[11]；将真实游戏“转化为可靠评测”的协议化实践可见*lmgame*[2]。在*GUI* 场景中，*UI-Venus* 强调纯截图（*screenshot-only*）输入与结构化动作（*structured output*）的端到端导航[12]；而*V-MAGE* 聚焦*visual-only/continuous-space* 的视觉中心评测[13]。上述工作将在第二部分的相关研究中系统梳理。

Screenshot-only navigation (GUI). 文献显示，在真实平台上，纯截图输入+ 结构化动作输出亦可实现端到端导航并取得具有竞争力的结果（如AndroidWorld 的*pass@1* 与ScreenSpot 系列的屏幕定位任务）[12]。

作为面向GUI 自动化的综述，[14] 从感知、探索/知识、规划与交互四个组成对(M)LLM-based GUI agents 进行框架化梳理，并总结当前评测在方法学与标准化上的挑战[14].

作为操作系统层面的交互智能体综述，[15] 将OS Agents 定义为在OS 提供的环境与接口（如GUI/CLI）中使用电脑、手机与浏览器完成任务的(M)LLM-based agents，并从“环境/观测/动作”与“理解/规划/落地”两层结构化梳理构建路径与评测要点[15].

作为多模态大模型（MLLM）的通用结构示意，图 1 展示了：文本*query* 经*tokenizer* 输入LLM；非文本模态（image/audio/video）先由*Multimodal Encoder* 提取表示，再经*Multimodal Projector* 对齐到语言嵌入空间，与文本在LLM 中融合并生成响应[16].

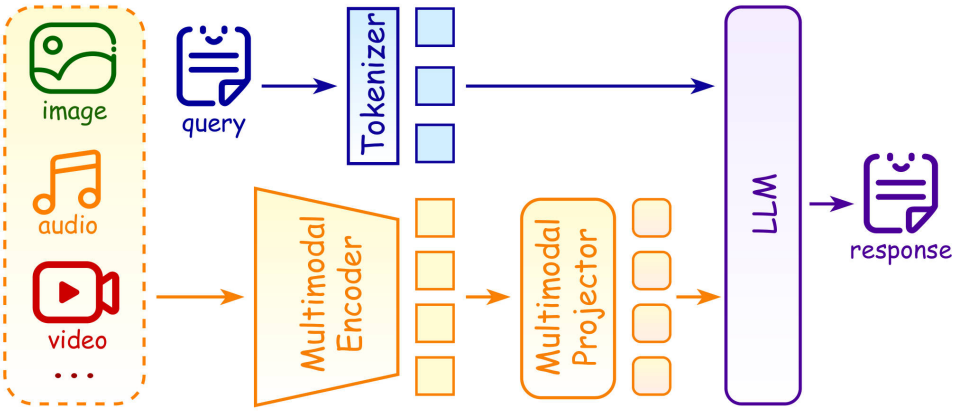

Figure 1: The overall architecture of MLLMs.

1.3 Key Challenges

（i）长链路稳定性（**long-horizon stability**）：在*GUI*（*GCC*）通道上，错误累积与状态漂移更易放大；文献通过*skills*/反思/记忆等管线缓解，但挑战仍存[10]；（ii）视觉中心定位与记忆（**vision-centric grounding & memory**）：*visual-only/continuous-space* 设定对定位/时机/视觉记忆/高层推理提出更高要求[13]；（iii）**OOD**与协议一致性（**OOD & protocol consistency**）：评测需在过程生成与变量可控的条件下比较架构/数据/后处理，减少不可比性[11]；（iv）提示方差与污染（**prompt variance & contamination**）：将“游戏→评测”落到可复现协议需稳定交互回路并记录后处理[2]；（v）无效动作与幻觉（**invalid actions & hallucination**）：结构化输出/约束解码可降低*invalid action*，但*think–action mismatch* 等现象仍被报告[12]；（vi）时延与交互体验（**latency & UX**）：实时伴随式场景强调语音往返（voice RTT）与帧到提示的响应时间，需与稳定性指标共同考量[1], [2]。

1.4 Our Positioning & Contributions

TODO: (1) 研究定位：GUI（GCC）下的伴随式助手设定；(2) 概念性模块化：以MCP-style 作为内部“技能/工具总线”进行注册/编排（与输出通道无关）；(3) 评测要

素：统一的任务脚本与指标（advice adoption, voice RTT, macro success 等）；(4) 预期交付物：原型/评测脚本/文档（按C.1中的Expected Results表述）。

1.5 Design Principles & System Preview

TODO: 系统流一句话：*screen/audio* → *VLM* → *LLM/agentic*（planning/memory/reflection）→ *MCP-style*（技能注册/路由）→ *GUI*执行（kb/mouse）→ *safety*（permissions, rollback, kill-switch）。设计原则：结构化输出（structured output）、可审计（auditability）、可复现（reproducibility）。

2 Literature Review

2.1 Perception: Modalities & Grounding

该方向通常以屏幕帧（screen frames）或短视频（video clips）作为主要输入，辅以窗口/坐标等轻量上下文；可选接入音频（audio）以形成语音闭环（ASR/TTS）。多模态模型在此承担检测/描述/定位（detection/description/grounding）、UI元素识别与状态读出等能力；与直接产出动作的*VLA* 相比，*VLM+tool* 将视觉理解与动作执行解耦，通过结构化调用（*structured tool calls*）或技能库（*skills/macros*）完成闭环。文献中的代表实践包括：*UI-Venus* 在*screenshot-only* 条件下通过结构化动作实现端到端导航[12]；*V-MAGE* 强调*visual-only/continuous-space* 设定下的定位、时机与视觉记忆压力[13]；*lmgame* 提供感知/记忆脚手架（scaffolds）以稳定交互与提示方差[2]。

Agent AI（multimodal interaction survey） 该综述给出"*Agent AI*"的工作定义：能感知多模态输入（视觉/语言/环境信号）并在具身或虚拟环境中产生动作（embodied actions）的交互系统；作者从"下一步具身动作预测（next-embodied action prediction）"出发，讨论外部知识（external knowledge）、多传感输入（multi-sensory inputs）与人类反馈（human feedback）在*grounded* 场景下提升稳健性的作用，并主张通过虚拟/模拟环境加速研究进展[8]。与本文关系：作为2.1小节的术语锚点与范围图谱，用于对齐"感知—定位—交互"相关表述。

2.2 Action Interfaces: GUI (GCC) & MCP-style Orchestration

在动作接口上，**GUI** 路线以**General Computer Control (GCC)** 为统一通道（*screen-in, keyboard/mouse-out*），强调跨应用/跨游戏的可迁移性（portability）与人类同态交互（human-homomorphic interface）。代表性工作*CRADLE* 展示了在不依赖应用专用接口的前提下，通过规划—技能整理（*skill curation/registry*）—反思—记忆的管线完成长链路任务（desktop/games），为*GUI* 可行性提供了实证支持[10]。与此同时，**MCP**（Model Context Protocol）提供了模块注册/编排（*module registration/orchestration*）的协议化思路：在不改变输出仍为*GUI* 的条件下，*skills/macros*、*planning*、*memory*、*reflectio*
等可在统一接口下组织，便于可复现实验与消融比较[1]。

Takeaway 文献显示：*GUI*（*GCC*）提供统一接口与较低移植门槛；协议化编排（如*MCP*）有助于模块化与复现性。与本文关系：本节作为动作接口背景与术语界定。

如Figure 2 所示，CRADLE 以统一的GUI（GCC）通道展示了从"屏幕输入→内在推理→键鼠控制"的闭环。

UI-Venus（screenshot-only） 端到端GUI 导航，无需planner/A11y；强调截图→结构化动作的通道在真实平台可达到具有竞争力的结果（如AndroidWorld 的pass@1

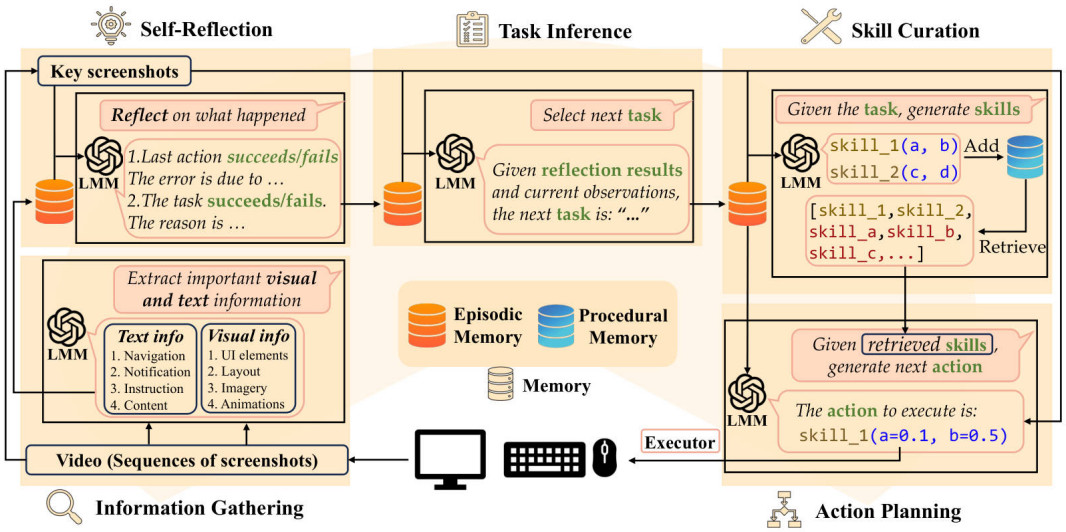

Figure 2: An overview of the CRADLE framework: CRADLE takes video from the computer screen as input and outputs computer keyboard and mouse control determined through inner reasoning (planning, skill curation, reflection, memory) [10].

与ScreenSpot 系列定位）[12]。与本文关系：作为screenshot-only 路线可行性的文献实例。

LLM-brained GUI agents（survey） 该综述将"以LLM 为中枢的GUI 智能体"作为统一对象，围绕GUI 自动化的框架、数据与训练、面向动作的专门化模型（*large action models, LAM*）与评测基准/指标展开，总结跨Web/移动/桌面平台的通用交互能力与挑战，并给出未来路线图[9]。

(M)LLM-based GUI agents（survey） 该综述将GUI 智能体拆解为四大组件：*perception*（文本解析与多模态理解）、*exploration/knowledge*（内部模型、历史回放与外部检索）、*planning*（推理与任务分解）与*interaction*（动作生成与安全控制），并回顾跨桌面/移动/Web 的研究进展；作者据此指出元素定位、知识检索、长时规划与安全执行控制仍是前沿挑战，同时强调现有评测在指标与协议上亟需标准化**mllm-gui-survey**。

OS Agents（OS-level scope survey） 该综述在OS-level 上界定了(M)LLM 驱动的计算设备智能体：以操作系统提供的接口（如GUI/CLI）为通道，跨电脑/手机/浏览器执行任务；作者提出"环境—观测空间—动作空间"的要素划分，并将"理解/规划/动作落地（grounding）"作为核心能力，系统回顾领域化基础模型、代理框架、评测指标与基准，以及产业产品版图与安全威胁[15]。与本文关系：提供2.2 的术语与范围锚点，可与*CRADLE/UI-Venus/ORAK* 等文献在"GUI（GCC）/评测与安全"维度对齐。

2.3 Agentic Modules: Planning, Memory, Reflection, Skills

TODO: 规划（planning）、记忆（memory, 用户偏好/历史）、反思（self-reflection, 纠错/风格一致）、技能库（skills/macros, 原子→复合）。与本文关系：仅作机制分类与代表性做法的回顾。

文献中的agentic modules 常见于四类：*planning*（分解与策略选择）、*memory*（短长时/用户偏好）、*self-reflection*（纠错与风格一致）与*skills/macros*（原子→复合操作）。例如*CRADLE* 组合了*planning/skills/reflection/memory* 的管线以缓解长链路误差累积，*ORAK* 在统一评测中对上述模块进行消融比较[1], [10]；*UI-Venus* 则在训练与数据层面探索轨迹历史对齐与稀疏动作增强**uivenus_rft**。与本文关系：作为模块分类与代表做法的回顾。

历史对齐与稀疏动作增强　提出*Self-Evolving Trajectory History Alignment & Sparse Action Enhancement*：用当前模型重写历史"思维—动作"轨迹以对齐风格/细节，并上采样稀疏但关键动作（如LongPress），以改善长链路一致性与泛化**uivenus_rft**。与本文关系：作为处理长链路与长尾动作的文献做法。

补充而言，[8] 将"下一步具身动作预测（next-embodied action prediction）"与人类反馈（human feedback）并置为提升*agentic* 能力的关键因素，强调在*grounded* 环境中校准策略与记忆的重要性。与本文关系：用于模块分类讨论的语义背景与术语对齐。

2.4 Learning Paradigms: Zero-shot, RAG, Finetune, IL/RL, Distillation

TODO: 零样本/提示工程、检索增强（RAG for UI schema/FAQ）、轻量微调（LoRA）、模仿/强化（IL/RL）、蒸馏到小模型。与本文关系：范式综述，不含实现承诺。

以"指令化（instructionalization）"增强RL 代理的上下文理解是一条代表性路线。*R2-Play* 将多模态游戏指令（MGI）并入*Decision Transformer*（*DTGI*），并通过超网络（*SHyperGenerator*）在训练任务与未见任务间共享知识；作者报告多模态指令较文本/轨迹单模态在多任务与泛化上更优（动机见Figure 3），MGI 的三段式结构——*game description*、*game trajectory*、*game guidance*（含动作、语言引导及关键元素位置）——给出了指令模板（见Figure 4）[17]。与本文关系：作为"指令化/Decision Transformer"方向的文献背景。

RFT（GRPO）与动作粒度奖励　将奖励拆分为格式/动作类型/坐标/内容四部分并加权，以同时度量结构化输出合规性与细粒度定位/文本输入正确性，作为GUI 导航中RL-finetune 的代表做法之一**uivenus_rft**。与本文关系：作为RL-finetune 在GUI 导航中的奖励设计范例。

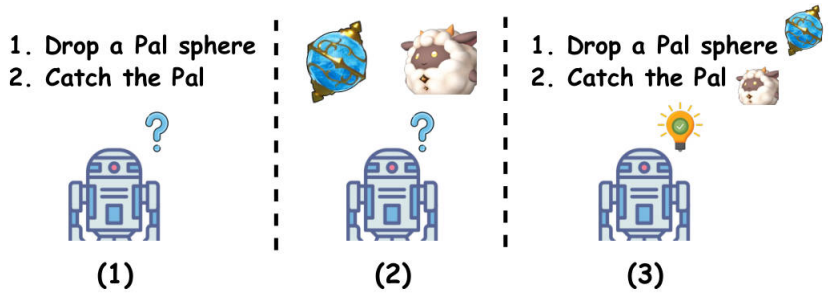

Figure 3: Imagine an agent learning to play Palworld. (1) The agent exhibits confusion when only relying on textual guidance. (2) The agent is confused when presented with images of a Pal sphere and a Pal. (3) The agent understands how to catch a pet through *multimodal guidance*, which combines textual guidance with images of the Pal sphere and Pal [17].

Figure 4: An illustrative example of *multimodal game instructions (MGI)*. Each instruction consists of three sections: *game description*, *game trajectory*, and *game guidance* (including action, language guidance, and the position of key elements) [17].

Tool-augmented MLLMs（survey） 该综述在MLLM 框架下扩展“工具”概念，覆盖API、专家模型与知识库等外部手段，并从“数据—任务—评测”三条主线梳理近年的方法学：在任务层面，总结了外部工具在六类挑战场景（多模态RAG、推理、幻觉、安全、代理、视频感知）中的典型用法；在流程层面，以MRAG 的“检索—重排—整合”三段式为例展示工具化管线；在评测层面，指出既有指标难以全面刻画多模态生成与对齐，主张引入工具协助的更系统评测[16]。

2.5 Benchmarks & Datasets (OS-like, Games, Desktop)

Orak（统一评测/消融/MCP思想） *ORAK* 通过*MCP* 实现*plug-and-play* 的代理—环境解耦，并在统一配置下检验*planning / reflection / memory / skills* 等*agentic modules* 的边际贡献（ablation），配套*Leaderboard/Battle Arena* 与训练轨迹数据（fine-tuning trajectories），将机制—性能—配置一体化呈现[1]。与本文关系：作为“统一评测与消融”的代表性基准。

Procedural-generation（OOD方法学） 基于*procedural generation* 的开放式评测在可控生成下构造*OOD* 与多步任务压力，比较*VLA/VLM* 在架构/训练数据/输出后处理等变量下的泛化与稳健性，并配套工具链以保证*reproducibility*[11]。与本文关系：作为OOD/变量可控的评测方法学背景。

lmgame-Bench（脚手架与污染控制） *lmgame* 将“游戏→评测”系统化：用*Gym-style* 接口与*perception/memory scaffolds* 稳定*prompt* 并剔除污染，在多模型下获得良好分离度，并通过相关性分析展示“各游戏探测的能力混合不相同”；另报告单一游戏的*RL* 训练对未见游戏/外部规划任务存在迁移[2]。与本文关系：作为“脚手架/污染控制/迁移观察”的评测文献。

...为减少提示方差并抑制污染，lmgame 以模块化脚手架稳定“感知—记忆—推理”的交互回路（见Figure 5）。

V-MAGE（vision-centric, visual-only, continuous-space） 该框架以仅视觉输入与连续空间的游戏环境，评测多模态模型的视觉中心能力，覆盖定位、轨迹追踪、时机、视觉记忆及更高层时序推理；其评测管线支持分离“模型/策略”，并采用Elo 风格排名进行相对强度比较；作者报告现有模型与人类表现存在差距、常见感知错误与锚定偏差，且有限历史上下文会限制长时规划[13]。与本文关系：作为视觉中心评测的代表性基准。

2.6 Evaluation Protocols & Metrics

文献在客观指标上常使用任务成功率（*success/pass@k*）、完成时间（*time-to-completion*）、误操作/回滚相关比率（*misclick/rollback rate*）以及延迟（*latency*，如语

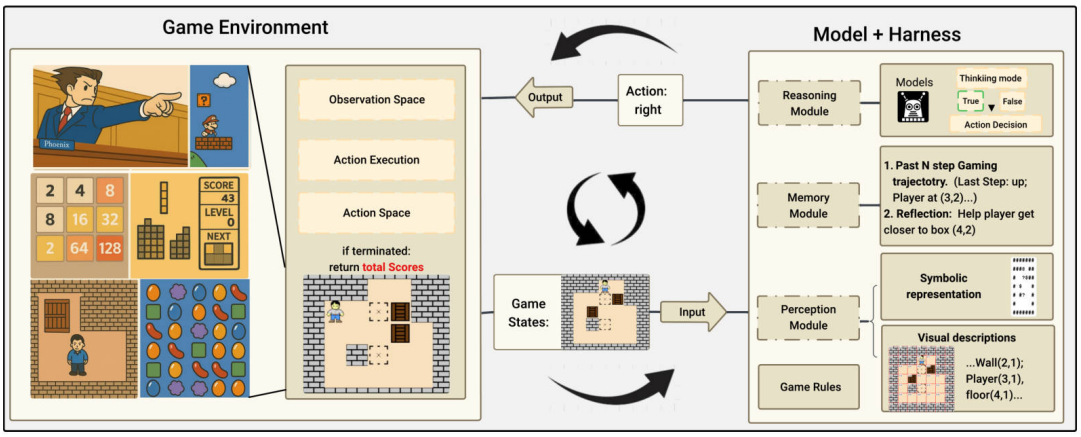

Figure 5: lmgame-Bench uses modular harnesses—such as perception, memory, and reasoning modules—to systematically extend a model's game-playing capabilities, enabling iterative interaction loops with a simulated game environment [2].

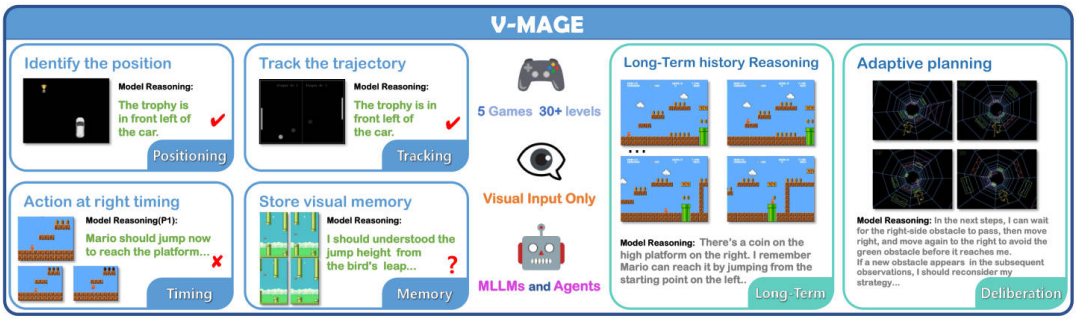

Figure 6: The overview of the V-MAGE benchmark, designed to evaluate vision-centric capabilities and higher-level reasoning of MLLMs across 5 free-form games with 30+ levels [13].

音往返*voice RTT*）等；在主观或行为性指标上，亦见用户采纳程度（*advice adoption*）与满意度等表述[1], [2], [11]。

此外，针对类别分布不均（class imbalance）的场景，亦有工作同时报告*macro-averaged* 与*micro* 指标，以减轻多数类偏置对总体判断的影响[11]。

在协议层面，部分工作强调结构化输出与约束解码（*structured output & constrained decoding*）以降低无效动作和坐标偏差；也有将无效动作率纳入评估与惩罚设计的做法（例如将动作拆分为格式/类型/坐标/内容等粒度进行评测与训练信号的组合）[12]。与“游戏→可靠评测”相关的工作则强调记录后处理/解码策略（*post-processing*）、提示方差与污染控制，以减少实现细节对可比性的影响[2]。

在跨任务比较上，亦有采用*Elo* 风格排名（*Elo-style ranking*）报告相对强度的实践，便于处理关卡难度不均与多任务汇总[13]。同时，基于*procedural generation* 的方法学将架构/数据/后处理视作可控变量进行统一对比，强调*OOD* 情形下的稳健性与可复现性[11]。

与本文关系：本小节仅提供指标与评测协议的文献背景与术语对齐，不涉及实现承诺。

2.7 Deployment & Real-time Considerations

部署相关文献关注资源与实时性约束：本地—云混合（*local–cloud hybrid*）与推理量化（*inference quantization,* 如*INT4/FP8*）用于降低时延与成本；流式解码（*streaming decoding*）与语音中断（*barge-in*）用于缩短交互回路；并评估对帧率与CPU/GPU 占用的影响。为保证可比性，协议层面强调固定提示、记录*post-processing* 与环境版本；在真实设备/平台的在线评测也逐渐出现（如AndroidWorld 场景与脚手架化交互）[2], [12]。与本文关系：作为工程现实与实时交互的背景回顾。

2.8 Safety, Permissions & Robustness

文献强调权限模型（*permission scoping/whitelisting*）与操作确认（*confirmation*）以约束高风险动作；影子模式（*shadow execution*）先预测后执行以降低副作用，并配套回滚（*rollback*）与急停（*kill-switch*）保障可逆性与故障恢复。在GUI 场景下，*think–action mismatch* 揭示了多模态模型可能产生的“幻觉（hallucination）”与不一致风险，提示需要日志与审计（*auditability*）以支持溯源与复查[12]。与本文关系：作为权限/鲁棒性与审计机制的文献背景。

如Figure 7 所示，think–action 不一致（mismatch）揭示了MLLM 的“幻觉”（hallucination）风险[12]。

2.9 Synthesis: Trends, Gaps & Our Niche

当前趋势是在*GUI*（*GCC*）通道上引入协议化/模块化编排以支撑可复现实验与消融；真实多类型游戏的统一评测（如*Orak*）与*visual-only/continuous-space* 的视觉中心

Figure 7: One trace of UI-Venus on the task named MarkorDeleteAllNotes in AndroidWorld. We can observe that UI-Venus successfully achieves the goal and has the reflection ability in Step 3. However, there also exists the conflict between think and action in Step 5, remaining as a future work about how to solve MLLM's hallucination.[12]

评测（如*V-MAGE*）并行发展；把脚手架/污染控制引入评测协议（如*lmgame-Bench*）成为常见做法。与本文关系：本文研究定位于伴随式（*companion-style*）场景的文献回顾与术语/评测背景梳理。

与基于GUI（GCC）的实践并行，*grounded* 的感知—行动—人类反馈闭环正成为多模态交互智能体的共识性趋势[8].

3 Project Plan

3.1 Proposed Solution / Methodology

3.2 Experimental Design

3.3 Expected Results

3.4 Progress Analysis and Gantt Chart

3.4.1 Risk & Ethics

4 Conclusion

References

[1] 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.

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

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

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

[5] 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.

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

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

[8] Z. Durante *et al.*, "Agent ai: Surveying the horizons of multimodal interaction," *arXiv preprint arXiv:2401.03568*, 2024.

[9] C. Zhang *et al.*, "Large language model-brained gui agents: A survey," *arXiv preprint arXiv:2411.18279*, 2024.

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

[11] 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.

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

[13] 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.

[14] F. Tang *et al.*, "A survey on (m) llm-based gui agents," *arXiv preprint arXiv:2504.13865*, 2025.

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

[16] W. An, J. Nie, Y. Wu, F. Tian, S. Lu, and Q. Zheng, "Empowering multimodal llms with external tools: A comprehensive survey," *arXiv preprint arXiv:2508.10955*, 2025.

[17] Y. Jin *et al.*, "Read to play (r2-play): Decision transformer with multimodal game instruction," *arXiv preprint arXiv:2402.04154*, 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.