



Xi'an Jiaotong-Liverpool University

西交利物浦大学

# 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

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Problem Setting & Motivation . . . . .	1
1.2	Scope & Working Definitions . . . . .	1
1.3	Design Principles & System Preview . . . . .	2
1.4	Key Challenges . . . . .	2
1.5	Project Objectives & Expected Deliverables . . . . .	3
1.6	Assumptions & Out-of-Scope . . . . .	3
<b>2</b>	<b>Literature Review</b>	<b>4</b>
2.1	Perception: Modalities & Grounding . . . . .	4
2.2	Action Interfaces: GUI (GCC) & MCP-style Orchestration . . . . .	4
2.3	Agentic Modules: Planning, Memory, Reflection, Skills . . . . .	5
2.4	Learning Paradigms: Zero-shot, RAG, Finetune, IL/RL, Distillation . . . . .	6
2.5	Benchmarks & Datasets (OS-like, Games, Desktop) . . . . .	7
2.6	Evaluation Protocols & Metrics . . . . .	9
2.7	Deployment & Real-time Considerations . . . . .	9
2.8	Safety, Permissions & Robustness . . . . .	10
2.9	Synthesis: Trends, Gaps & Our Niche . . . . .	11
<b>3</b>	<b>Project Plan</b>	<b>12</b>
3.1	Proposed Solution / Methodology . . . . .	12
3.2	Experimental Design . . . . .	12
3.3	Expected Results . . . . .	12
3.4	Progress Analysis and Gantt Chart . . . . .	12
3.4.1	Risk & Ethics . . . . .	12
<b>4</b>	<b>Conclusion</b>	<b>13</b>
<b>References</b>		<b>14</b>
<b>Appendix A.</b>	<b>Title of Appendix A</b>	<b>I</b>
A.1	Appendix Heading 1 . . . . .	I
A.2	Appendix Heading 2 . . . . .	I
A.3	Appendix Table and Figure Captions . . . . .	I
<b>Appendix B.</b>	<b>Title of Appendix B</b>	<b>I</b>

# 1 Introduction

## 1.1 Problem Setting & Motivation

本项目拟构建一个面向玩家的伴随式 (*companion-style*) 实时游戏助手：围绕事件提示、策略建议与语音交互闭环提供低时延 (low-latency) 体验与可复现评测 (reproducible evaluation)。当前生态从AI游戏主播/虚拟角色到LLM 驱动的NPC/插件快速涌现，但大量实践依赖场景定制，统一的动作接口与可复现实验协议仍不充分；研究侧正通过统一评测/模块消融与污染控制/协议一致等方法学收敛评测与对比[1], [2]。本提案据此明确项目目标、范围与评测要素。

## 1.2 Scope & Working Definitions

本项目以**GUI (GCC)** 通道为主要执行路径 (*screen-in, keyboard/mouse-out* 的人类同态接口 *human-homomorphic interface*)，强调跨应用/游戏的通用性与可迁移性 (portability)。代表性工作 *CRADLE* 报告了在不依赖应用 API 的前提下，通过规划/技能整理/反思/记忆的管线完成长链路任务的可行性[3]。**MCP-style** 编排在本项目中被用作内部模块/技能的注册与路由 (registration/orchestration) 的通用思路；不涵盖针对具体应用/游戏的专用 API 适配工作。关于“统一评测/消融、plug-and-play”的组织可参考 *ORAK*[1]；基于 *procedural generation* 的 OOD 方法学可参考 *Benchmarking-VLA-VLM*[4]；把真实游戏“转化为可靠评测”的协议化实践可参考 *lmgame*[2]。在 GUI 场景中，*UI-Venus* 展示了截图输入 (*screenshot*) + 结构化动作 (*structured output*) 的端到端导航路径[5]；*V-MAGE* 聚焦 *visual-only/continuous-space* 的视觉中心评测[6]。相关工作将在第二部分系统梳理。

**Screenshot-first GUI feasibility.** 在真实平台上，“截图输入+ 结构化动作输出”路线已有具体成绩：*UI-Venus* 在 *AndroidWorld* 报告 **65.9% pass@1**，在 *ScreenSpot-V2/Pro* 分别为 **95.3%/61.9%** [5]，表明在不依赖 A11y/DOM 的条件下亦可实现可比导航能力。

**MLLM architecture at a glance.** 图 1 给出多模态大模型 (MLLM) 的通用结构：文本经 *tokenizer* 输入 LLM，非文本模态由 *Multimodal Encoder/Projector* 对齐后与文本融合[7]。

**Survey cues.** 更广义的“Agent AI”综述从“下一步具身动作预测 (next-embody action prediction)”出发，讨论外部知识、人类反馈与多传感输入在 *grounded* 场景中的作用[8]；面向 GUI 自动化的综述系统梳理了以 LLM 为中枢的框架、*LAM* (large action models)、基准与指标[9]；另有针对 (M)LLM-based GUI agents 的综述从感知—探索/知识—规划—交互四组成对齐术语与挑战[10]；在 OS 视角，*OS Agents* 综述提出“环境/观测/动作—理解/规划/落地”的要素图谱[11]；端侧推理综述为“低时延”语境提供工程参考 **ondevice-llm**；“LLM×游戏智能体”的方法与基准的更广泛总览可参见近期 arXiv 工作 **game-agents-large-models**。

**Industry/Community signals.** 作为动机与生态线索（非技术有效性论证），*Neuro-*

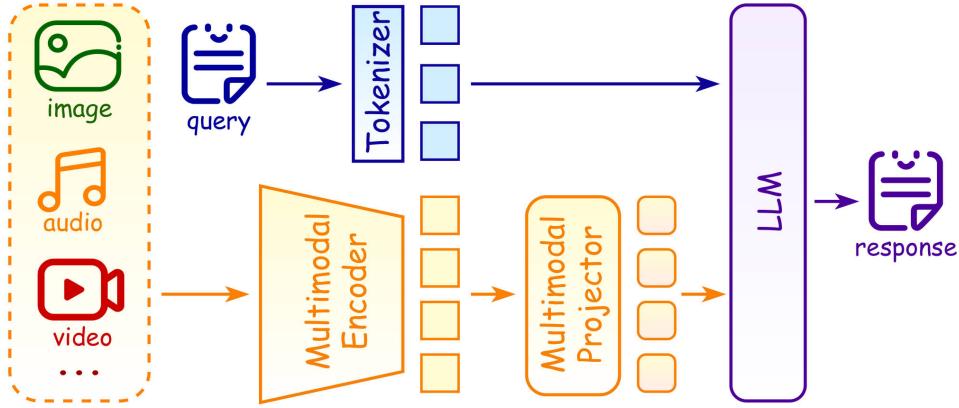


Figure 1: The overall architecture of MLLMs [7].

sama 的 AI streamer 现象与相关开源生态（如 *Open LLM VTuber*、*Airi Project*、*Kimjammer-Neuro*）反映出该方向的关注与活跃[12]–[15]；平台侧量化数据亦显示较大受众体量：Neuro-sama 于 2025/01 创下 **Hype Train Level 111**（约 85K 付费订阅、约 1.2M bits），Twitch 2024 年总观看时长约 **18.5–20.8B** 小时、2024Q4 全行业约 **21B** 小时[16]–[21]。

### 1.3 Design Principles & System Preview

**Design principles:** 结构化输出 (structured output)、可审计 (auditability)、可复现 (reproducibility)、低耦合 (MCP-style 编排)。 **System preview** (一句话流程)：*screen/audio* → 轻量 *VLM* → *LLM/agentic* (planning/memory/reflection) → *MCP-style* 技能注册/路由 → *GUI* 执行 (kb/mouse) → *safety* (permissions, rollback, kill-switch)。默认工程姿态 (不作实现承诺)：single-flight (不并发大模型请求)、event-triggered (事件触发)、frame-window (3–5 帧堆叠)、text-first (可解析为结构化文本则优先文本路径)，以降低延迟与方差。

### 1.4 Key Challenges

- (i) 长链路稳定性 (**long-horizon stability**)：在 GUI (GCC) 通道下，错误累积与状态漂移更易放大；文献中的 *skills*/反思/记忆管线可缓解，但挑战仍存[3]；
- (ii) 视觉中心定位与记忆 (**vision-centric grounding & memory**)：*visual-only/continuous-space* 对定位、时机、视觉记忆与高层推理提出更高要求[6]；
- (iii) **OOD** 与协议一致 (**OOD & protocol consistency**)：需要在过程生成与变量可控条件下比较架构/数据/后处理，减少不可比性[4]；
- (iv) 提示方差与污染 (**prompt variance & contamination**)：把“游戏→评测”落到可复现协议需稳定交互通路并记录后处理[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[5]；
- (vi) 时延与交互体验 (**latency & UX**)：实时伴随式场景强调 *voice RTT* 与帧到提示响应时间，需要与稳定性指标共同考量[1], [2]。

## 1.5 Project Objectives & Expected Deliverables

### Objectives.

- 实现一个以*GUI (GCC)* 为主的伴随式实时游戏助手原型，覆盖事件提示、策略建议与语音回路，面向低时延与可复现评测。
- 采用*MCP-style* 的内部编排思想组织*skills/macros*、*planning*、*memory*、*reflection*，在不依赖应用专用API的前提下实现可插拔与可审计（auditability）。
- 明确一套小而可复现的评测要素（任务脚本与指标族），关注*advice adoption*、*voice RTT*、*macro success* 等体验相关量。

### Expected Deliverables.

- 系统原型：屏幕采集与轻量感知、*agentic* 模块、*MCP-style* 技能总线、*GUI* 执行器、基础安全护栏（权限/回滚/急停）。
- 评测脚本与配置：可复现实验的任务脚本、指标计算与日志审计工具（含模块开关用于对照）。
- 使用文档与演示：安装/运行说明、配置模板与演示视频；范围外：针对具体应用/游戏的专用API适配与大规模模型微调。

## 1.6 Assumptions & Out-of-Scope

**Working assumptions.** 默认采用“*single-flight + event-triggered + frame-window* (3–5 帧) + *text-first*”的工程姿态以降低时延与方差；评测记录*post-processing* 与交互版本以保持协议一致。

**Out-of-scope.** 不开展每个游戏/应用的专用API适配；不在本报告中承诺大规模端到端训练与数据采集；不涉及平台级增强权限（如A11y/私有DOM钩子）的依赖。

## 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）为主输入，辅以窗口/坐标等轻量上下文；必要时接入音频（audio）形成语音闭环（ASR/TTS）。多模态模型承担检测/描述/定位（detection/description/grounding）、UI 元素识别与状态读出等能力；相较直接产出动作的VLA，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）闭环。代表实践包括：*UI-Venus* 在*screenshot-only* 条件下以结构化动作实现端到端导航[5]；*V-MAGE* 在*visual-only/continuous-space* 设定下强调定位、时机与视觉记忆压力[6]；*lmgame* 以感知/记忆脚手架（scaffolds）稳定交互并控制提示方差[2]。另外，也有直接从Let's Play 视频侧推断玩家体验（PX）的做法，为交互或评测中的“机会对齐（opportunity alignment）”提供情绪/投入度代理[22]。

**Agent AI (multimodal interaction survey)** 该综述将“Agent AI”界定为：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；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* 场景中的作用，并建议以虚拟/模拟环境加速研究进展[8]。该视角为“感知—定位—交互”的表述提供了概念锚点。

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

动作接口层面，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）。CRADLE 显示：在不依赖应用专用接口的前提下，结合规划—技能整理（skill curation/registry）—反思—记忆的管线亦可完成长链路任务（desktop/games）[3]。与此同时，**MCP**（Model Context Protocol）提供模块注册/编排（module registration/orchestration）的协议化思路，使*skills/macros*、*planning*、*memory*、*reflection* 能在统一接口下组织与对比[1]。由此形成“以GUI 为执行通道、以协议化编排（MCP-style）组织模块”的常见范式。

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/AI1y；“截图→结构化动作”在真实平台报告了具有竞争力的结果（如AndroidWorld 的pass@1 与 ScreenSpot 系列定位）[5]，强化了“GUI 统一通道”的可达性认识。

**LLM-brained GUI agents (survey)** 该综述以“LLM 为中枢的GUI 智能体”为统一

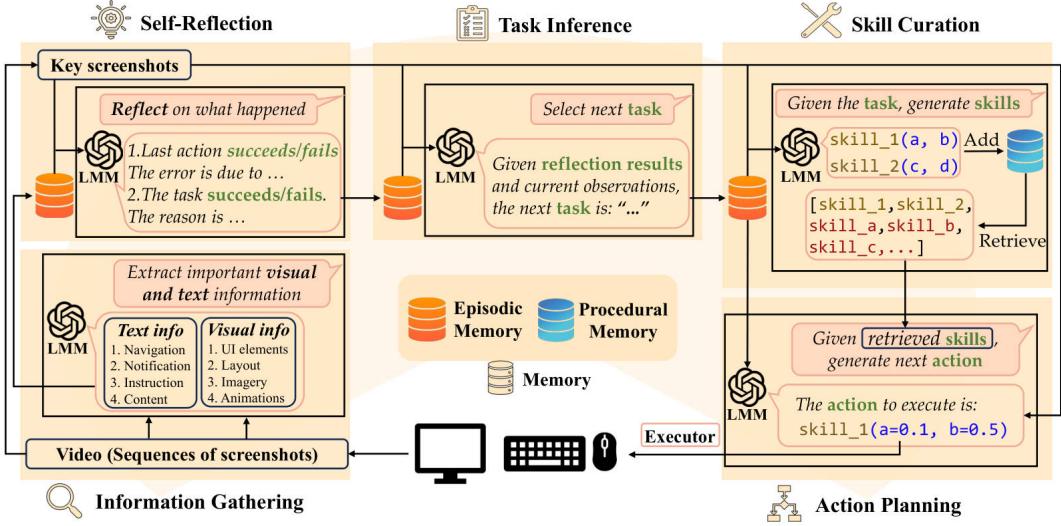


Figure 2: An overview of the CRADLE framework: CRADLE takes video from the computer screen as input and outputs computer keyboard and mouse control determined through inner reasoning (planning, skill curation, reflection, memory) [3].

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，并提出若干路线图[9]。

**(M)LLM-based GUI agents (survey)** 该综述将GUI智能体拆解为四大组件：*perception*（多模态理解）、*exploration/knowledge*（内部模型、历史回放与外部检索）、*planning*（任务分解与推理）与*interaction*（动作生成与安全控制），回顾跨桌面/移动/Web的研究进展，并指出元素定位、知识检索、长时规划与安全执行控制仍是挑战，同时强调评测在指标与协议上的标准化需求[10]。

**OS Agents (OS-level scope survey)** 在OS视角下，(M)LLM驱动的计算设备智能体通过操作系统提供的接口（如GUI/CLI）跨电脑/手机/浏览器执行任务；综述提出“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，系统回顾基础模型、代理框架、评测与安全威胁[11]。该范围化视角有助于统一“设备—接口—能力”的讨论语言。

## 2.3 Agentic Modules: Planning, Memory, Reflection, Skills

常见*agentic modules*包括：*planning*（分解与策略选择）、*memory*（短长时与用户偏好）、*self-reflection*（纠错与风格一致）与*skills/macros*（原子→复合）。例如，CRADLE组合*planning/skills/reflection/memory*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], [3]；UI-Venus在训练与数据层面探索轨迹历史对齐与稀疏动作增强ivenus\_rft。这类机制为长链路稳定性与一致性提供了可分析的结构要素。

历史对齐与稀疏动作增强 *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），报告了对长链路一致性与泛化的改善uivenus\_rft。另有观点将“下一步具身动作预测（next-embody action prediction）”与人类反馈并置为提升*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[8]。

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

学习与推理范式并行发展：零样本/提示工程（prompting）、检索增强（RAG）、轻量微调（LoRA）、模仿/强化（IL/RL）与蒸馏（distillation）。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），且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）——给出清晰的指令模板（见Figure 4）[23]。

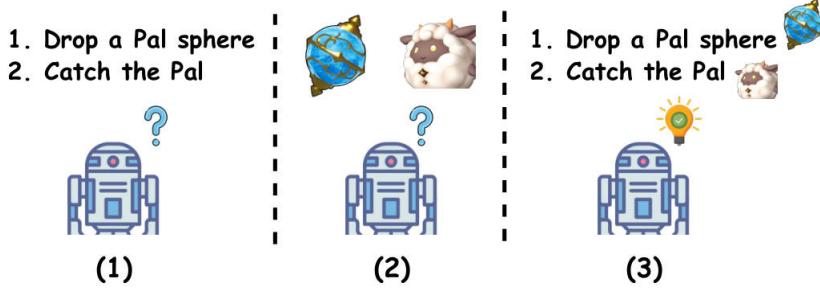


Figure 3: Imagine an agent learning to play Palworld. (1) The agent exhibits confusion when only relying on textual guidance. (2) The agent is confused when presented with images of a Pal sphere and a Pal. (3) The agent understands how to catch a pet through *multimodal guidance*, which combines textual guidance with images of the Pal sphere and Pal [23].

**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。

“工具增强型MLLM（tool-augmented MLLM）”的综述从“数据—任务—评测”三条主线梳理外部工具（API、专家模型、知识库等）的作用边界：任务侧涵盖多模态RAG、推理、幻觉、安全、代理与视频感知，流程侧以MRAG的“检索—重排—整合”三段式为例，评测侧指出既有指标难以全面刻画多模态生成与对齐[7]。更广泛的总览从方法、应用与挑战三方面回顾LLM×游戏的研究脉络，串联学习范式、*planning/memory/reflection* 与工具调用（tool-use/MCP-style）的组合，并联通环境类型、输入模态与动作接口以对齐评测与基准的讨论**game-agents-large-models**。



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) [23].

## 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]。更广泛的总览亦梳理了多样基准与评测设置，强调跨游戏可比性与协议一致的重要性**game-agents-large-models**。

**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不同游戏探测的能力混合；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]。整体上，评测组织正从“单点游戏”走向“脚手架化、协议一致”的可复现比较。…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”的交互通路（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[6]。这一路线补充了跨类型真实游戏基准的视角。

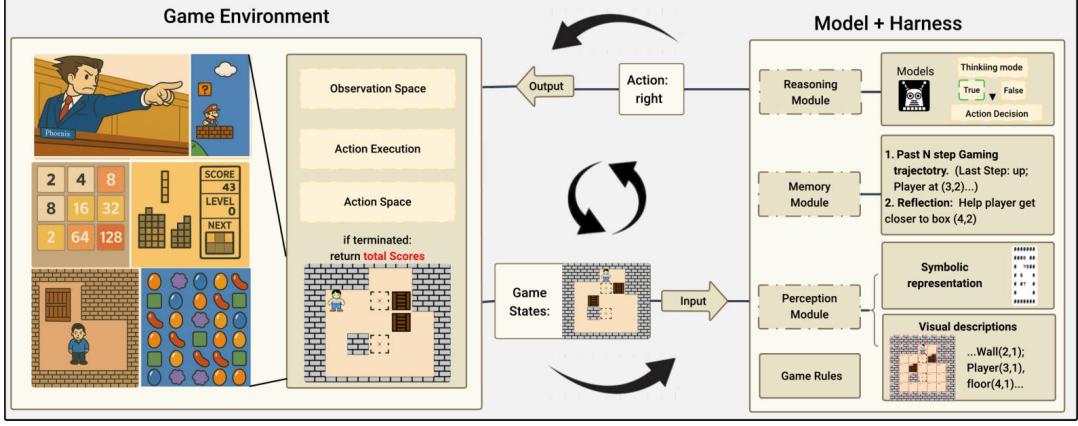


Figure 5: Imgame-Bench uses modular harnesses—such as perception, memory, and reasoning modules—to systematically extend a model’s game-playing capabilities, enabling iterative interaction loops with a simulated game environment [2].

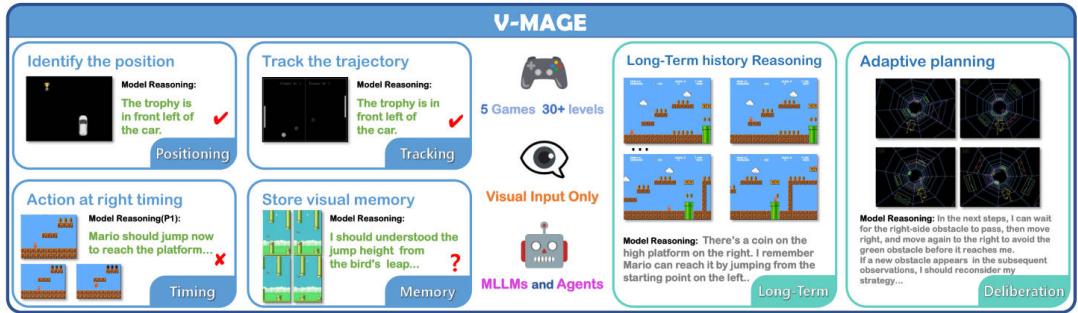


Figure 6: The overview of the V-MAGE benchmark, designed to evaluate vision-centric capabilities and higher-level reasoning of MLLMs across 5 free-form games with 30+ levels [6].

## 2.6 Evaluation Protocols & Metrics

**Methodological note (Benchmarking-VLA-VLM)** 该工作以*procedural generation* 的多子任务为载体，总结了动作映射与评测原则：将模型输出显式映射到合法离散动作空间，并联合使用*Invalid%*、*Micro/Macro Precision/Recall/F1* 与*Brier/MAE* 等度量；实验还显示输出约束/动作空间显式化能显著降低无效输出、提升稳定性，同时指出视觉增益依模型/任务而异、*OOD*/时序敏感动作是难点[4]。这些方法学结论为结构化输出/约束解码 (*structured output & constrained decoding*) 与无效动作率 (*invalid actions*) 的纳入提供依据。

常见客观指标：任务成功率 (*success/pass@k*)、完成时间 (*time-to-completion*)、误操作/回滚 (*misclick/rollback rate*) 与延迟 (*latency*, 如语音往返*voice RTT*)；主观/行为指标：采纳程度 (*advice adoption*) 与满意度等[1], [2], [4]。在类别分布不均 (*class imbalance*) 情形，*macro-averaged* 与 *micro* 指标并举以减轻多数类偏置[4]。协议层面，文献强调结构化输出与约束解码以降低无效动作与坐标偏差，也有将无效动作率纳入评估与训练信号设计（动作拆分为格式/类型/坐标/内容四粒度）的做法[5]；“游戏→可靠评测”路线强调记录后处理/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]。跨任务比较上，*Elo* 风格排名 (*Elo-style ranking*) 可缓解关卡难度不均与多任务汇总问题[6]；而*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* 稳健性与可复现性[4]。以上做法逐步形成了通用的评测组织与指标图谱。

**Label-free PX proxies from Let's Play videos** 除任务成功与时延外，亦有工作直接从*Let's Play* 视频构建无标签 (*label-free*) 玩家体验 (PX) 的代理信号：以画面、解说音频等多模态特征弱/自监督提炼*arousal/valence/engagement*，并与小规模人工评分或高激励事件对齐，用于机会窗口 (*opportunity windows*) 与提示时机 (*hint timing*) 的度量[22]。

## 2.7 Deployment & Real-time Considerations

部署相关研究聚焦资源与实时性约束：本地—云混合 (*local–cloud hybrid*) 与推理量化 (*inference quantization, INT4/FP8*) 降低时延与成本；流式解码 (*streaming decoding*) 与语音中断 (*barge-in*) 缩短交互动路；并评估对帧率与CPU/GPU 占用的影响。为保证可比性，协议侧采用固定提示、记录*post-processing* 与环境版本；在真实设备/平台的在线评测亦逐步增多（如AndroidWorld 与脚手架化交互）[2], [5]。同时，端侧部署 (*on-device*) 综述强调量化 (*INT4/FP8*)、蒸馏/适配、*KV* 缓存与分页注意力以及*speculative/streaming* 解码等路径以降低时延与能耗，并讨论资源约束下的稳定性折中与隐私优势**ondevice-llm**。

## 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场景中，*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*) 支持溯源与复查[5]。总体上，权限边界、影子执行与可审计性构成了“安全—鲁棒”的基本支架。

如Figure 7所示，*think-action* 不一致 (*mismatch*) 提示在复杂交互中需关注模型的幻觉与不一致风险[5]。



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. [5]

## 2.9 Synthesis: Trends, Gaps & Our Niche

总体趋势是：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；跨类型真实游戏的统一评测（如*ORAK*）与*visual-only/continuous-space* 的视觉中心评测（如*VMAGE*）并行发展，将脚手架/污染控制纳入协议（如*lmgame-Bench*）逐步形成共识[1], [2], [6]。同时，沿着*grounded* 的“感知—行动—人类反馈”闭环，研究正从单点Demo 转向过程变量与稳健性，进一步强调跨任务与OOD 条件下的可比性与一致性[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](https://arxiv.org/abs/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] W. Tan *et al.*, “Cradle: Empowering foundation agents towards general computer control,” *arXiv preprint arXiv:2403.03186*, 2024.
- [4] 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.
- [5] Z. Gu *et al.*, “Ui-venus technical report: Building high-performance ui agents with rft,” *arXiv preprint arXiv:2508.10833*, 2025.
- [6] 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.
- [7] 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.
- [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] F. Tang *et al.*, “A survey on (m) llm-based gui agents,” *arXiv preprint arXiv:2504.13865*, 2025.
- [11] X. Hu *et al.*, *Os agents: A survey on mllm-based agents for computer, phone and browser use*, 2024.
- [12] Vedral and Neuro-sama, *Neuro-sama official youtube channel*, <https://www.youtube.com/@Neurosama>, Accessed: 2025-10-10, 2022.
- [13] 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.
- [14] moeru-ai, *Airi: Ai waifu / virtual character container inspired by neuro-sama*, <https://github.com/moeru-ai/airi>, Accessed: 2025-10-10, 2025.
- [15] kimjammer, *Neuro: A local-model recreation of neuro-sama*, <https://github.com/kimjammer/Neuro>, Accessed: 2025-10-10, 2025.

- [16] “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>.
- [17] P. Staff. “With valorant’s help, this ai vtuber just beat a massive twitch record.” Accessed: 2025-10-11, PCGamesN. [Online]. Available: <https://www.pcgamesn.com/valorant/neuro-sama-twitch-record>.
- [18] “Vedal987 — streamer overview & stats.” Accessed: 2025-10-11, TwitchTracker. [Online]. Available: <https://twitchtracker.com/vedal987>.
- [19] C. Xiang. “Meet neuro-sama, the ai twitch streamer who plays minecraft, sings karaoke, loves art.” Accessed: 2025-10-11, Bloomberg. [Online]. Available: <https://www.bloomberg.com/news/newsletters/2023-06-16/neuro-sama-an-ai-twitch-influencer-plays-minecraft-sings-karaoke-loves-art>.
- [20] 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>.
- [21] “Q4 2024 global live streaming landscape.” Accessed: 2025-10-11, Streams Charts. [Online]. Available: <https://streamscharts.com/news/q4-2024-global-livestreaming-landscape>.
- [22] D. Goel, A. Mahmoudi-Nejad, and M. Guzdial, “Label-free subjective player experience modelling via let’s play videos,” in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 20, 2024, pp. 46–53.
- [23] 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.