arXiv:2603.01104v1 [cs.HC] 1 Mar 2026

Egocentric Co-Pilot：用于辅助性自我中心人工智能的原生网络智能眼镜智能体

Sicheng Yang 清华大学深圳国际研究生院 清华大学 深圳，中国
yangsc25@mails.tsinghua.edu.cn

Yukai Huang 独立研究员 伦敦，英国
u06530032@alum.ccu.edu.tw

Weitong Cai Shitong Sun 伦敦玛丽女王大学 伦敦，英国
weitong.cai@qmul.ac.uk
shitong.sun@qmul.ac.uk

Fengyi Fang You He*清华大学深圳国际研究生院 清华大学 深圳，中国 fangfy22@tsinghua.org.cn
heyou@mail.tsinghua.edu.cn

Yiqiao Xie
Jiankang Deng
伦敦帝国理工学院
英国伦敦
yx2722@ic.ac.uk
j.deng16@imperial.ac.uk

Hang Zhang
独立研究员
伦敦，英国
hz459@cornell.edu.cn

Jifei Song
萨里大学
吉尔福德，英国
j.song@qmul.ac.uk

Zhensong Zhang
独立研究员
伦敦，英国
zhensongzhang@hotmail.com

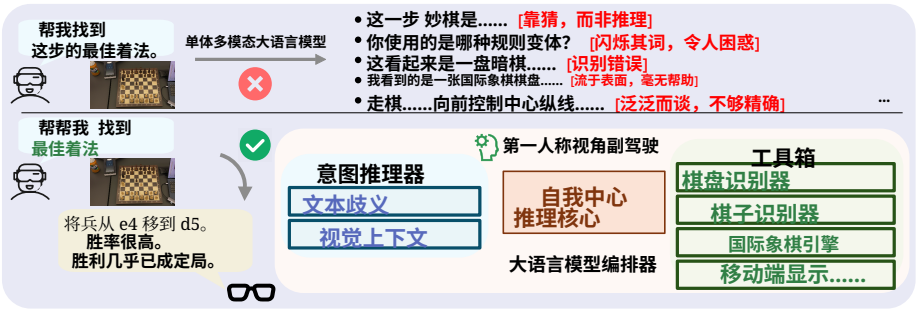

图 1：第一人称视角副驾驶与单体多模态大语言模型的概览。上：单体多模态大语言模型难以处理专门化推理任务（例如策略棋盘游戏），往往给出回避性回答。下：我们的框架使用大语言模型编排器，借助由专门化神经符号模块组成的工具箱。它成功解读用户请求，并调用感知模块生成棋盘状态，再由专用游戏引擎求解，从而得出精确且可执行的解决方案。

摘要

如果访问网络不再需要屏幕、稳定的桌面，甚至空闲的双手，会怎样？对于在拥挤城市中穿行、患有低视力或经历认知过载的人而言，智能眼镜与人工智能代理相结合，可以将网络变成覆盖日常生活的常开辅助层。我们提出第一人称视角副驾驶，一种运行于智能眼镜上的原生于网络的神经符号框架，并使用一个大语言模型（LLM），用于统筹感知、推理和网页工具组成的工具箱。自我中心推理核心将时间链式思维与分层上下文压缩相结合，以支持基于连续第一人称视频的长时程问答和决策支持，能力远超单个模型的上下文窗口。在此基础上，轻量级多模态意图层可将噪声语音和注视转化为结构化的可供工具调用的命令，而无需依赖单一的庞大模型。我们还基于 LiveKit 实现并评估了云原生 WebRTC 流水线，将流式语音、视频和控制消息整合到单一 Web 标准通道中，同时服务智能眼镜客户端和基于浏览器的试验平台。与此同时，我们部署了本地 WebSocket 基线，展现本地推理与云端卸载在延迟、移动性和资源使用方面的具体权衡，

*通讯作者。

本作品采用知识共享署名 4.0 国际许可协议进行许可。*WWW '26*，迪拜，阿拉伯联合酋长国© 2026 版权由所有者/作者持有。ACM ISBN 979-8-4007-2307-0/2026/04 https://doi.org/10.1145/3774904.3792996

以及资源使用方面的权衡。基于 Egolife 和 HD-EPIC 的实验表明，该系统在自我中心问答方面表现出有竞争力或达到最先进水平；针对智能眼镜开展的人机协同研究则显示，其任务完成率和用户满意度均高于领先的商用基线。综合来看，这些结果表明，联网自我中心副驾驶能够成为一条切实可行的路径，为日常生活提供更易获取、具备情境感知能力的辅助。通过以原生网络通信基元和模块化、可审计的工具使用为基础，第一人称视角副驾驶为始终在线的辅助型网络智能体提供了具体蓝图，助力教育、无障碍和社会包容，服务于最能从情境化自我中心人工智能中受益的人群。我们的代码和微调模型可在此处获取。

CCS 概念

• **以人为中心的计算** → **普适与移动计算系统和工具**；• **信息系统** → 万维网；• **计算方法** → 人工智能；• **社会与专业主题** → 残障人士。

关键词

自我中心人工智能、智能眼镜、网络智能体、多模态意图消歧、神经符号系统、WebRTC、可穿戴辅助技术、负责任的人工智能

ACM 参考文献格式：

Sicheng Yang、Yukai Huang、Weitong Cai、Shitong Sun、Fengyi Fang、You He、Yiqiao Xie、Jiankang Deng、Hang Zhang、Jifei Song 和 Zhensong Zhang。2026。Egocentric Co-Pilot：用于辅助性自我中心人工智能的原生网络智能眼镜智能体。载于《*2026*年 *ACM Web* 会议论文集》（*WWW '26*），*2026*年*4*月*13*日至*17*日，阿拉伯联合酋长国迪拜。ACM，纽约州纽约市，美国，14 页。https://doi.org/10.1145/3774904.3792996

1 引言

强大而小巧的硬件与大语言模型（LLM）的结合，催生了新一代由人工智能驱动的个人助理 [61]。尽管自主网络智能体改变了人工智能系统与数字界面的交互方式 [19, 98], ，但将这些能力拓展至物理世界仍面临挑战。在各种潜在平台中，智能眼镜尤为适合：它能提供免提、常开式界面，将数字信息叠加到物理世界中 [2, 25, 33, 45, 86, 89], ，原则上还可让那些难以使用小屏幕、复杂菜单或视觉要求较高的布局的人获取网络内容。实际上，这些助理中的许多都通过基于网络的服务与应用提供，因此网络成为其影响力的主要分发渠道和治理界面。目标不只是信息检索，更是打造一位副驾驶，辅助人类认知：智能体从用户的第一人称（自我中心）视角出发，理解其意图，并在阅读营养标签、遵循多步骤指令或追踪预约等日常任务中主动提供帮助 [22, 51, 65]。

除了技术性能之外，旨在广泛造福人群的助手还必须集成原生网络通信和安全基础机制（标准 Web 协议、云 API 和基于浏览器的客户端），以便能够负责任且高效地大规模部署 [21, 48]。对于我们所面向的用户群体，包括低视力者、面临注意力或记忆障碍的人士，以及行动能力有限者，这类智能体应优先考虑可靠性、隐私和降低认知负荷，而非吸引用户参与或追求新奇。在实际部署中，这意味着行为可预测、系统不确定时能明确提示，并且用户可以明确控制哪些信息会被传输或存储。然而，要实现这一愿景，仍需克服若干根本性挑战。首先，现实世界中的交互本质上具有歧义。例如，一个简单的指示性命令，如“分析这个”，必须结合杂乱的视觉场景来确定其所指对象，这项任务需要强大的多模态推理能力 [100]。其次，没有任何单一人工智能模型能够有效解决所有问题 [85]。许多任务都需要结合鲁棒感知（神经网络在这方面表现出色）与精确的符号推理或工具使用，例如规划游戏中的行动或调用日历 API。当前的端到端模型往往缺乏完成这些专业任务所需的精确度 [72, 84]。最后，自我中心数据流的连续性给上下文窗口有限的模型带来了难题，因为它们难以维持长程依赖和时间上下文 [80]。

在本文中，我们认为，一个有效的自我中心助手应构建于由中央大语言模型编排的模块化神经符号架构之上。我们提出了**第一人称视角副驾驶**，这一框架旨在将人类意图与一组专用工具和可通过网络访问的服务连接起来。我们的框架并不依赖单一模型，而是使用大语言模型作为推理引擎来解释用户的多模态命令。它首先通过交互式对话和视觉定位澄清意图。随后，它通过选择并调用最合适的工具（如神经感知模块、符号推理器或外部网络 API）生成执行计划。这种混合方法将大语言模型的上下文理解能力与专用模块的精确性相结合，同时提供可在资源受限设备和基于浏览器的客户端上于日常环境中运行的原生网络界面。

我们的核心贡献是一个框架，它以与具有社会效益的网络部署相兼容的方式应对上述挑战。具体而言，我们做出以下贡献：

(1) 用于组合专用工具的框架。我们提出了一种神经符号框架，利用大语言模型（LLM）通过轻量、适用于网络的协议（MCP），协调设备端模块与基于网络的服务。这使资源受限设备（如智能眼镜）能够使用丰富的感知、推理和辅助型网络 API 生态系统。

(2) 用于理解模糊指令的模块。为处理不明确的用户请求，我们设计了一个模块，在采取行动前先澄清意图。对于含糊的文本，它会提出后续问题。对于有歧义的视觉输入，它会采用 3D 射线投射方法来确定用户指向的对象，确保正确理解指令。这种审慎、以用户为中心的行为在安全攸关或具有重大社会影响的情境中尤为重要，因为避免造成伤害的误解比追求原始吞吐量更重要。

(3) 用于对长视频进行推理的模块。 为了处理连续第一人称视角视频，我们开发了一种上下文管理方法，将时间链式思维（*T-CoT*）用于详细的短期推理，并结合分层

上下文压缩（*HCC*）用于长期记忆。这样一来，系统便能利用超出模型标准上下文窗口的时间段中的信息，同时处理由资源受限设备采集的流。

(4) **一套经过真实场景评估的完整系统**。我们将框架构建为一套运行于智能眼镜上的实用系统，采用原生网页后端，并进行了广泛测试，在自我中心问答基准（Egolife、HD-EPIC）上取得了出色成绩。此外，我们通过用户研究展示了其实用价值：在强调建设性日常辅助、而非仅用于提升参与度的真实任务中，该系统显著优于领先的商用系统。

2 相关工作

第一人称视角人工智能。自我中心人工智能从动作识别、手部与物体交互 [14, 50, 59, 99], 和视觉情感分析 [55] 等基础任务，逐步发展到借助多模态大语言模型（MLLM）进行复杂推理[46, 49, 88]。多模态大语言模型支持密集图像描述和问答等高层次、开放式任务，并通过 Ego4D 和 Egolife 等基准进行评估 [20, 27, 46, 64, 76]。然而，处理长视频仍是一项重大挑战，受限于基于Transformer的模型的计算成本和上下文窗口 [70, 80]。常见的解决方案包括将视频转换为文本日志、进行层次建模或生成摘要 [10, 38, 80]。这类研究大多聚焦于离线分析，包括近期反复遍历长篇第一人称视角日志的分层检索智能体。相比之下，我们的系统面向智能眼镜上的实时连续生活记录，这也直接促使我们提出分层上下文压缩（HCC）方法，以实现高效的上下文管理。

大语言模型驱动的智能体与工具使用。 大语言模型如今被广泛用作自主智能体的中央控制器 [29, 69, 96], , 使其能够通过思考—行动循环进行推理 [63], 、规划并与外部工具交互 [56, 95]。虽然 LangChain 和 AutoGPT 等框架简化了开发过程，但在可靠性、长时程规划和泛化方面仍面临重大挑战 [66, 79]。许多相关研究聚焦于数字领域中的智能体，例如网页导航 [19, 28, 81,

83, 98]。最近，研究人员开始将大语言模型应用于物理具身智能体，以执行高层级机器人任务规划，让语言指令能够以物理世界为基础 [1, 3, 9, 53, 60]。我们的工作聚焦于具身智能体研究中的一个特定领域：非机器人式可穿戴智能体，旨在增强而非取代人类行动。这种“第一人称视角副驾驶”是与用户协作的伙伴 [25, 89]。为促进这种人机协作，我们提出了模型上下文协议（MCP），这是一种轻量级协议，专为这类增强人类能力的系统所需的实时边缘端—云端协同而设计。

神经符号系统。 神经符号系统将神经感知与符号推理相结合，兼取两种方法之长。这种综合有助于缓解各自的弱点：Neural Networks 不透明的黑箱本质，以及符号方法在处理噪声数据时的脆弱性 [7, 30, 84]。该领域近期的大量工作聚焦于从非结构化数据中学习、同时又受到约束的系统



图2：我们的 ReasoningCore 流水线。它结合了用于短期分析的时间链式思维（T-CoT）和用于长期记忆的分层上下文压缩（HCC）。该图展示了 T-CoT 路径，其中经过微调的多模态大语言模型处理一个有时间范围限制的查询。

由显式符号知识约束 [8, 13, 35]。我们的工作沿此方向推进：我们采用神经模块，将原始感知输入转换为结构化状态表示，从而为符号推理器提供接地。我们并未构建组件紧密耦合的单体系统，而是将神经符号流水线封装为一个离散、可调用的工具 [67, 72]。随后，该工具由大语言模型（ LLM ）编排，从而形成分层且模块化的神经符号架构 [4, 12, 75]。这一设计与近期的趋势相一致，即构建基于大语言模型的智能体，通过组合专用模块来解决复杂任务 [82]。

多模态意图消歧。 意图澄清研究已从结构化对话 [39, 87, 93]发展到由大语言模型驱动的方法 [52, 91], , 但仍主要局限于非具身的纯文本场景 [16, 68, 92]。这一局限在自我中心视觉（EGV） [37], 中尤为突出，因为用户输入本质上带有噪声且含糊不清 [23, 57], , 尤其是指向手势等关键模态，往往只是被动处理或过度简化 [18, 32, 43]。即使是最先进的视觉语言模型 [26, 34, 74] ，在所需的精确空间推理方面也并不可靠，这一问题更因此加剧，并常常导致幻觉 [24, 31, 44]。将大语言模型推理与专用模块相结合的混合架构 [47, 58, 73] 是一个颇具前景的方向。在我们的工作中，我们迈出了一步，探索主动、迭代式框架，引导用户实时消除多模态歧义，而不是事后被动分析带噪输入 [42, 94]。

3 方法论

3.1 自我中心推理核心

我们框架的核心是推理核心，这是一个基于多模态大语言模型的引擎，旨在处理连续第一人称视角视频

流并回答用户查询。该引擎的基础是统一且按时间顺序排列的事件日志 E，它整合了密集的第一人称视角视频叙述，以及通过自动语音识别（自动语音识别）转录的口语用户查询。每个事件 $e_i = (t_i,\ m_i,\ c_i)$ 记录一个时间戳、模态（视觉或语音）和规范化内容，为系统提供了对第一人称体验紧凑而语义丰富的表征。

收到用户查询 Q后，推理核心会启动一个动态的多阶段流程，为多模态大语言模型构建最优上下文（图2）。我们方法的一项关键创新在于采用双层策略，同时处理短期与长期的时间依赖关系。该流程首先分析查询的意图，以确定其时间范围。（1）对于近期事件或特定时间限定片段的细粒度分析，我们采用时间链式思维（T-CoT）策略 [71, 78]。T-CoT以程序化方式在相关时间戳周围选取狭窄的时间窗口，对相应片段进行叙述或剪辑，并将其编排为连贯的局部故事线。由此可分离出回答当前查询所需的最相关信息。（2）对于超出多模态大语言模型原生上下文窗口的长期推理，我们启用分层上下文压缩（HCC）。历史日志被划分为时间分块，每个分块由一个较小的纯文本模型概括为简短的查询感知描述。随后，HCC选取最相关的摘要，并将其前置到T-CoT上下文中，从而在不超出模型上下文预算的前提下提供长程感知。

最终推理由多模态大语言模型（MLLM） M_{vqa} [5]完成，我们使用第一人称视角的厨房数据集和日常活动数据集（例如 EPIC-KITCHENS）对其进行微调。我们将多项选择选项标准化，把描述不充分的问题改写为明确的、基于视角的提示，并使用基于正则表达式的解析器从自由格式的输出中提取最终选项。为增强鲁棒性，我们通过多数投票集成多个提示变体的预测结果。通过这些步骤，我们将冗长的第一人称视角流转化为适合长时程推理的可处理上下文。

3.2 大语言模型编排的神经符号执行

为了将推理核心的高层理解转化为
为将其转化为具体行动，我们采用模块化、基于工具的模型。与在单次前向传递中处理感知、符号计算和网页交互的单体多模态大语言模型不同，我们的架构将各种能力视为可调用的工具，可按需组合。

这种编排通过 MCP 实现。MCP 是一种轻量级接口，可将工具作为由 JSON Schema 描述的函数公开。执行遵循标准的检索-行动循环：大语言模型发现可用工具、制定计划、通过 MCP 执行工具调用，并综合生成最终响应。该通用循环的形式化伪代码见附录 A。重要的是，MCP 旨在通过标准网络通道运行（例如 WebRTC 数据通道或 HTTPS 后端），因此，同一工具生态系统既可服务于可穿戴设备客户端，也可服务于基于浏览器的用户，并可利用现有网页治理机制进行审计或沙箱隔离。

这一架构的能力通过一个具身策略棋盘助手得以展现，该助手在 MCP 生态系统中封装为单一的神经符号工具。当用户请求

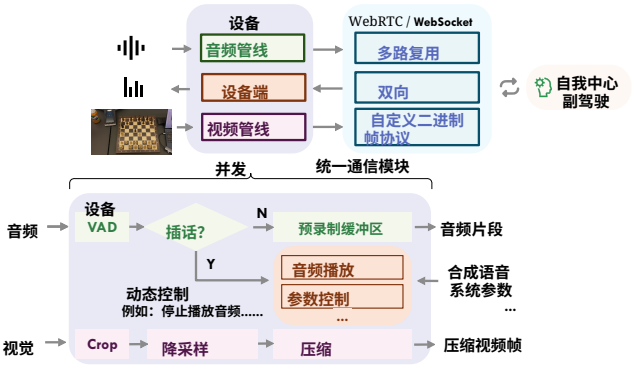

图 3：用于实时多模态交互的设备端架构。音频和视频并行处理，并复用到智能眼镜与云后端之间的双向通道中。我们使用 WebRTC [62]（H.264 视频、Opus 音频、数据通道），并采用定制的 WebSocket 变体作为本地部署基线。

走法建议时，助手会执行一条混合流水线：（i）感知模块将第一人称棋盘视图映射为稳定的符号状态（例如 FEN）；（ii）确定性引擎评估候选走法；（iii）编排器大语言模型将坐标输出转换为非专业棋手也能理解的策略性解说。为确保能够抵御检测噪声，我们实现了一种时间缓冲机制：只有通过多数投票达到稳定性阈值后，才会确认棋盘状态。随后，符号引擎处理稳定后的状态，再由大语言模型综合生成结果。该平滑机制的形式化定义以及完整的执行算法见附录 A。

这种模块化的高层抽象使编排语言模型能够利用成熟的符号引擎，而无需管理其内部机制。视觉模型、状态稳定化启发式方法和第一人称提示的实现细节概述于附录 A。与系统中的其他工具一样，最终的文本响应可以合成为富有表现力的语音，整个端到端流程则由异步事件循环管理，以确保实时交互期间的响应速度。

3.3 设备端感知与基于 WebRTC 的交互

我们的前端专为资源受限的智能眼镜设计，可与云后端进行实时双向多模态通信。在设备端，音频和视频流水线并发运行于单个事件循环中。轻量级的基于能量的语音活动检测搭配较短的预录制缓冲区，可捕获完整语句，同时支持全双工插话；视频管线则会裁剪并缩小第一人称视角流，再对其编码以供传输。

我们没有采用自定义的 WebSocket 分帧协议，而是使用标准的 WebRTC 协议栈（通过 LiveKit [41]），在统一信道上传输音频、视频和低速率 JSON 控制消息。音频以 Opus 格式流式传输，视频以 H.264 格式传输，数据通道则承载对齐元数据和工具调用信号。在服务器端，单个语音流水线智能体整合了神经网络语音活动检测、流式自动语音识别、

多模态大语言模型和文本转语音，构成如下形式：

$$\mathcal{I}_{\text{webrtc}} = \mathcal{F}_{\text{vad}} \circ \mathcal{F}_{\text{asr}} \circ \mathcal{F}_{\text{llm}} \circ \mathcal{F}_{\text{tts}}, \quad (1)$$

并配置为支持全双工、可感知中断的运行模式。智能眼镜和基于浏览器的试验平台复用同一操作端 $\mathrm{I}_{\text{webrtc}}$，实现亚秒级、原生网络的多模态交互。作为对比，我们还部署了本地 WebSocket 变体，以牺牲部分部署简便性和移动性为代价，换取略低的延迟。

3.4 主动式多模态意图消歧

即使拥有强大的多模态骨干网络，用户在第一人称视角场景中仍经常发出信息不充分或含糊不清的指令（例如，指着一块板或一台电器问“能再给我展示一下吗？”）。为减少误解——这在辅助或教育场景中可能造成尤为严重的影响——我们在交互流水线末端集成了一个轻量、即插即用的澄清器。当大语言模型检测到较高的语义不确定性或相互冲突的解释时，澄清器会将当前情形重新表述为一个受约束的决策问题。给定一个输入 $(x_{1:t},\ v_{1:t},\ s_t)$ 和一小组候选解释 $\{\phi_k\}$，它会选择直接回答，或提出一个简短的澄清问题：

$$\phi^\star = \arg\max_{\phi_k \in \Phi \cup \{\text{ask}\}} U\big(\phi_k \mid x_{1:t}, v_{1:t}, s_t\big), \quad (2)$$

其中，$U(\cdot)$权衡信息量与交互成本。如果 $\phi^\star = \text{ask}$，系统会发出简短的追问（例如：“你指的是左边那件，还是角落附近的那件？”），并在执行操作前根据回复更新上下文。我们通过改造近期提出的即插即用多模态澄清器 [77], 来实现该模块，并在此重点介绍其在基于 WebRTC 的自我中心助手中的集成。该澄清器作为底层大语言模型和感知算子的黑盒封装器，无需重新训练基础模型，还可针对导航辅助工具或日常协助等敏感领域选择性启用。

运行时护栏与模式管理。为便于说明，算法1展示了一个线性的“先规划、后执行”循环；而在实际实现中，我们加入了实用的防护措施，以避免不安全或非预期的操作。首先，通过明确的允许列表向大语言模型开放工具；在所有智能眼镜实验中，我们将工具集限制为非破坏性能力（例如回答查询），不允许直接控制执行器或外部账户。其次，在每次调用工具之前，MCP 都会根据函数签名和结构化文档字符串所定义的带类型注释的模式验证参数；如果不匹配，系统会记录日志并中止调用，而不是强行转换参数。第三，对于可能产生副作用的命令（例如编辑日历条目），编排器必须发出自然语言确认提示，只有用户明确确认后才会执行工具。多工具计划采用尽力执行的方式：如果任何中间调用失败，就会跳过其余调用，并指示大语言模型总结部分结果，而不是尝试自动回滚。本文重点关注推理和编排方面；工业部署还需要更强的机制，例如为每个应用设置能力白名单、模式版本控制，以及事务性提交/中止语义。

数据集	模型	准确率（%）
Egolife [76]	LLaVA-OV [36]	30.8*
	GPT-4o [34]	36.2*
	Gemini-1.5-Pro [54]	36.9*
	Qwen2.5 VL [5]	38.1
	Ours	**40.9**
HD-EPIC [50]	VideoLlama 2 [11]	27.4*
	LongVA [90]	29.3*
	LLaVA-Video [40]	32.4*
	Qwen2.5 VL [5]	33.5
	Gemini-1.5-Pro [54]	37.6*
	Ours	**46.2**

表1：在 Egolife 和 HD-EPIC 基准测试上与最先进的方法进行比较。带星号（*）的结果引自原论文；其他结果均为我们使用官方代码复现所得。

4 实验

4.1 在自我中心问答基准测试中的应用

我们将推理核心应用于 Egolife 和 HD-EPIC 基准测试，这两者各自带来了独特的挑战。对于 Egolife 中需要基于大量历史信息进行推理的长视频，我们采用分层上下文压缩（HCC）。具体而言，历史日志被划分为时间分块（例如按小时划分）。随后，纯文本大语言模型会评估每个分块与用户查询的相关性，并仅为相关分块生成简洁的摘要。这个过程会生成一份紧凑且针对查询的历史信息表示，并将其置于近期事件的详细日志之前，从而构成供多模态大语言模型进行推理的最终上下文。相比之下，对于 HD-EPIC 中以动作为重点的片段，我们采用专门的时间链式思维（T-CoT）策略。为对多个片段进行推理，我们会以编程方式将相关视频片段拼接成一条时间线，并重新规范化时间戳。对于超出上下文窗口长度的单个视频，我们会通过描述连续片段生成文本摘要，并将其用作上下文。对于所有基准测试，我们采用两阶段流程来生成可靠答案：首先由基于正则表达式的解析器提取主要答案选项，然后对五种语法形式不同的提示所产生的输出进行多数投票。

表1展示了我们在 Egolife 和 HD-EPIC 问答基准上的主要结果，并比较了我们的 ReasoningCore 与最先进方法的性能。我们的方法取得了出色的结果，尤其是在 HD-EPIC 上，这凸显了动态 T-CoT 策略在以动作 为中心的推理中的效用。

4.2 消融实验与敏感性分析

为明确各组件的贡献，我们进行了消融实验，结果汇总于表2，关键发现如下。在 Ego- life 上，移除 HCC 会使准确率降低 2.0 个百分点，移除 T-CoT 则会降低 1.4 个百分点，这证实了长时程摘要和局部时间结构化均有实质性贡献。在第一人称视角数据上进行微调可进一步提升 1.7 个百分点，而去除自动语音识别转录文本会导致准确率下降 0.8 个百分点，这表明

数据集	变体	准确率（%）
Egolife [76]	Ours	**40.9**
	无微调	39.2 (-1.7)
	无 T-CoT	39.5 (-1.4)
	无 HCC	38.9 (-2.0)
	不含转录	40.1 (-0.8)
HD-EPIC [50]	Ours	**46.23**
	不含微调	40.61 (-5.62)
	不含 T-CoT	42.68 (-3.55)
	不含 HCC	41.55 (-4.68)
	不含预处理	43.56 (-2.67)

表2：我们在 Egolife 和 HD-EPIC 上对推理核心进行的消融研究。我们报告完整模型（“我们的方法”）以及移除单个组件后的变体；括号中显示性能下降幅度。

语音线索能提供有用但次要的上下文。在 HD-EPIC 上，由于其侧重于短小但复杂的动作片段，领域特定微调尤为关键：省略微调会导致准确率下降 5.62 个百分点。移除 HCC 和 T-CoT 分别会使准确率降低 4.68 和 3.55 个百分点，这表明即使在片段尺度上，时间组织仍然重要。最后，禁用我们的提示与输出规范层（预处理、正则表达式答案提取和提示集成）会导致准确率下降 2.67 个百分点，因此，看似只是“工程实现”的细节在实证上对稳健性也很重要。我们还通过改变分块大小和摘要长度来分析 HCC 的敏感性：将摘要长度减半或将分块大小加倍，Egolife 准确率最多下降 1.0–1.6 个百分点；而将用于选择的 LLM 换成更小的模型，则会导致准确率下降 2.6 个百分点。这些趋势表明，该框架对超参数变化具有较好的稳健性，而且 HCC 在 T-CoT 单独使用的基础上仍能带来额外价值。

4.3 真实场景中的工具使用

为了验证我们的大语言模型编排的神经符号框架——第一人称视角副驾驶，我们在运行于智能眼镜上的端到端系统中开展了真实世界的第一人称视角实验。实验通过英语或普通话下达用户指令，评估系统理解多模态意图、组合工具并执行复杂任务的能力。我们采用任务完成率（TCR）衡量性能，其定义为无需用户干预即可成功执行任务；配置详情见附录 A。我们将任务划分为三个类别，以体现系统复杂度和真实世界风险的逐步提升。类别 1（基础工具使用）包含常见、低风险的请求，例如事实查询、简单提醒和记笔记；这些任务用于检验能否将自然语言可靠地映射到网络 API 和本地实用工具。类别 2（具身与时空任务）聚焦于以感知为基础的活动，例如棋盘游戏辅助和物体跟踪，这些任务需要稳定的第一人称视角感知和短期记忆。类别 3（复杂神经符号推理）涵盖专家级任务，这些任务将含噪视觉输入与确定性符号求解器相结合。在总共，我们定义了数十种任务模板，在三个类别中的分布大致均衡；评估过程中，每种模板都会实例化为多个具体试验。这一层级划分使我们能够分别对核心 API 落地、具身感知和完整的神经符号推理进行压力测试，同时涵盖常见的日常需求和更具挑战性的长尾场景，例如策略辅导。

类别 *1*：基础工具使用。此类别旨在检验将自然语言映射为特定 API 调用的核心能力。任务包括查询知识库（例如“查一下这个苹果的卡路里”）、管理个人信息（例如“提醒我下午 3 点的会议”）或创建笔记。如图 4(a–b)所示，这些任务要求大语言模型解析意图、提取实体（有时需要从视觉上下文中提取），并使用恰当的参数调用正确的工具（例如 NutritionAPI、CalendarAPI、MemoTool）。这些任务的 TCR 高达 98.5%，展现了我们基础执行循环的可靠性。

类别 *2*：具身策略与时空任务。此类别以棋盘策略助手为例，该助手用于实体棋盘上的国际象棋类游戏。该助手在 MCP 生态系统中封装为单一神经符号工具。当用户请求走法建议时，该工具会运行算法 2 中概述的混合流水线：视觉模块观察当前棋盘局面，并将原始画面转换为稳定的符号状态（公式 3）；随后，确定性引擎针对合法走法执行符号搜索；最后，编排大语言模型将坐标式输出转化为适合玩家技能水平的策略性自然语言指导。这种设计体现了具身时空推理：任务以持续变化的物理场景为基础，但助手通过语音和文本进行交流。通过仅向 MCP 编排器提供简洁的工具接口，我们使语言模型能够调用复杂的游戏引擎，而不必纠缠于其内部逻辑，同时仍可通过合成语音向用户提供实时多模态反馈。

类别 *3*：复杂神经符号推理。 这最后一个类别评估整个神经符号流水线在需要将现实世界感知与形式化符号推理紧密结合的任务上的表现。此类任务的特点是需要：(i) 将含噪视觉输入转换为结构化的符号表示，(ii) 对该表示应用确定性或启发式规则引擎，以及 (iii) 将符号输出转换回具有上下文感知的自然语言指导。我们使用在若干国际象棋类游戏上实例化的棋盘游戏副驾驶作为该类别的代表性基准测试。在 50 局游戏中，该系统在生成策略合理且与上下文相关的走法建议方面实现了 98% 的端到端成功率，表明感知模块、符号引擎以及由大语言模型驱动的语义解释器能够良好协同工作，正如算法 2 所形式化的那样。

失败分析。 为了更好地理解自我中心副驾驶的局限性，我们人工检查了全部三个任务类别中一组有代表性的失败案例。大多数失败可分为四类：(i) 感知错误，例如误检的棋盘状态



图 4：大语言模型编排框架的核心能力。我们的系统解读多模态用户意图，并通过 MCP 协议动态组合神经符号工具。(a) 基础工具使用：一个简单查询会触发视觉语言模型进行物体识别，并发起一次外部 API 调用。(b) 结构化任务管理：自然语言被转换为面向原生设备应用的结构化 API 调用。(c) 复杂神经符号推理：棋盘游戏副驾驶整合视觉工具（神经）、确定性游戏引擎（符号）以及用于语义解释的大语言模型。(d) 时空记忆：系统通过在遮挡过程中对物体进行视觉跟踪，并从记忆中召回该物体，从而解析指示性指称（"这个"）。

或在杂乱场景中错误定位目标物体；(ii) 意图误解，即大语言模型根据上下文过度泛化，选择了错误的工具或误解了指示性指称；(iii) 工具层面问题，包括缺失参数或意外的 API 响应；以及 (iv) 长时程记忆错误，即相关过往事件被从压缩上下文中遗漏。感知和意图错误最为常见，尤其是在光线不佳或头部快速运动的情况下。这些类别表明了具体的缓解措施，例如更强的自我中心视角主干网络、明确

在情境模糊时增加确认轮次，并加强参数验证。我们将对这些方向进行系统探索，留待未来工作开展。在各类别中，我们有意聚焦于体现建设性日常辅助的任务，例如阅读标签、管理简单日程以及接受棋盘旁辅导，而非以娱乐或单纯提升参与度为导向的场景。此选择旨在更好地体现此类智能体如何帮助用户完成影响自主性与福祉的实际日常活动。



图 5：对第一人称视角副驾驶、商用智能眼镜设备及人工基准进行的主观评估。柱形表示 5 点李克特量表评分的均值（越高越好）；星号（*）表示默认交互模式偏离持续对话式人工智能的设备。

4.4 人在回路评估

为评估实际应用效果，我们开展了一项人在回路研究，将我们的第一人称视角副驾驶与若干商用智能眼镜设备及人类基线进行比较。我们没有开展实时交互测试——因为这会使人工智能质量与硬件、网络和连接性差异相互混淆——而是采用受控的离线方案：针对每个系统和每项任务，我们使用相同的提示和环境记录交互日志（音频、视频和转录），随后请独立评估者对这些日志进行评估。

参与者观看了经过匿名处理且随机排序的片段，并且不知道底层系统的身份。对于每个片段，他们使用5分李克特量表进行评分（分数越高越好），判断助手是否（i）正确理解了多模态意图，以及（ii）成功执行了相应任务。我们还记录了客观的任务完成率（TCR），其定义为按照预先设定的成功检查清单，在没有人工干预的情况下完成任务。默认交互模式与持续对话式人工智能差异较大的设备（例如仅通知模式）在图5中以星号标记；我们仍将其作为基线纳入比较，但在解读其评分时会考虑这一点。

如图 5 所示，我们的模型部署在标准的现成硬件上，平均评分达到 4.70，显著超过所有商业竞争产品，并接近 4.92 的人类基线。平均 TCR 也呈现出类似趋势：自我中心副驾驶端到端完成的任务数超过了任何单一设备基线。这些提升与我们的设计目标一致：增强意图消歧能力并实现稳健的工具组合，从而减少用户纠正，提供更令人满意的协助。由于这项研究涉及对低风险日常任务（例如天气查询、简单规划和棋盘游戏建议）的预先录制且完全匿名化的交互日志进行评分，因此在本机构属于“最小风险”类别，无需接受正式的机构审查委员会审查；所有参与者均在参与前提供了知情同意。更多实验方案细节汇总于附录 D，示例请见图 6。这些结果表明，即使在普通硬件上，经过精心编排的 Web 原生助手也能为用户提供更可靠、少些挫败感的支持，帮助他们完成日常任务，其表现优于当前的商业智能眼镜软件；这为部署自我中心网页智能体提供了一条切实可行的路径，使其真正提升日常自主性，而非仅仅增加通知。

5 局限性与未来工作

第一人称视角副驾驶是一个研究原型，存在若干局限性。首先，其行为最终取决于底层的大语言模型/视觉语言模型骨干模型以及人工设计的工具模式。感知、推理或工具选择中的错误仍可能沿流水线逐步传播，而我们目前的防护措施（允许列表中的工具、基于模式的参数检查和显式确认）尚不及形式化安全保证。进一步扩展该框架，加入更强的能力管理、应用级策略以及事务式提交/中止语义，是一个重要的发展方向。

其次，我们依赖领域适配的自我中心模型和云后端。虽然使用第一人称数据进行微调能够提升性能，但未必能完美迁移到新的领域或相机外形规格，而且连续流传输会带来延迟和能耗开销。我们计划探索参数高效适配、能够减少流式处理的端侧模型，以及对不同部署选项下计算与能耗开销的明确核算。

最后，我们的评估聚焦于在受控环境中为健康成年人提供短期辅助和策略指导。我们未考察长期影响、高风险场景，也未考虑残障人士、老年人或其他可能最能从中受益群体的需求。对于始终开启的自我中心式采集，隐私和旁观者同意仍是尚待解决的问题。未来工作包括面向多元人群开展纵向研究，并针对网络规模部署加强端侧过滤和隐私保护训练。

6 结论

我们推出了第一人称视角副驾驶，这是一款模块化神经符号框架，将第一人称视角感知、长时程上下文管理和大语言模型编排的工具使用整合到一款智能眼镜助手中。它结合了时间链式思维、分层上下文压缩、原生于网络的工具生态系统和云原生 WebRTC 后端，在 Egolife 和 HD-EPIC 上取得了有竞争力的准确率，并在真实世界的人在回路研究中胜过了多款商业助手。

除了原始性能之外，这项设计还强调日常辅助应用场景，例如情境式辅导、情境感知提醒和阅读辅助，旨在提升用户的独立性和数字福祉，而非一味优化参与度。我们希望，第一人称视角副驾驶能为未来原生于网络的自我中心智能体提供一个具体范本，使其不仅具备技术能力，也能作为负责任、包容性的技术部署，为最能从情境化、始终在线辅助中受益的人群提供支持。更广泛地说，我们的研究结果表明，围绕有原则的感知与上下文管理技术栈，精心编排各类专用工具，可能比单纯扩展单体模型更能切实推动网络上值得信赖的辅助型人工智能发展。

7 致谢

本研究由深圳市科技计划（项目编号：ZDSYS20220323112000001）资助。

参考文献

[1] Michael Ahn、Anthony Brohan、Noah Brown 等。2022。《照我能做的做，而不是照我说的做：将语言 grounding 于机器人可供性》。CoRR abs/2204.01691（2022）。arXiv:2204.01691 doi:10.48550/ARXIV.2204.01691[2] 匿名作者。2025。《WearVox：面向可穿戴设备的第一人称视角多通道语音助手基准测试》。载于提交至第十四届国际学习表征会议的论文。https://openreview.net/forum?id=QpaNErg7ug，审稿中。[3] Raghav Arora、Shivam Singh、Karthik Swaminathan 等。2024。《预判与行动：整合大语言模型与经典规划，以高效执行家庭环境中的任务》。载于IEEE国际机器人与自动化会议（ICRA）。IEEE，14038 14045 10 1109 57147 2024 10611164 –。doi: ./ICRA。2025

[4] Ali Baheri 和 Cecilia O. Alm。《分层神经符号决策变换器》。CoRR abs/2503.07148（2025）。arXiv:2503.07148 doi:10.48550/ARXIV. 2503.07148[5] Shuai Bai、Keqin Chen、Xuejing Liu 等。2025。《Qwen2.5-VL技术报告》。CoRR abs/ 2502 13923 2025 2502 13923 10 48550 2502 13923 。（）。arXiv:。doi: ./ARXIV。Siddhant Bansal、Chetan Arora 和 C. V. Jawahar。2022。《我的视角最佳：从第一人称视角视频中进行程序学习》。载于《计算机视觉——ECCV 2022——第17届欧洲会议》，以色列特拉维夫（《计算机科学讲义》，第 *13673* 657 675 10 1007 978 3 031 19778 9_38 卷）。Springer，–。2024

[6] doi: ./ - - - - Emily Barnes 和 James Hutson。《自然语言处理与神经符号人工智能：神经网络在知识引导的符号方法中的作用》。Journal of Artificial Intelligence and Robotics 2, 1（2024）。[8] Bikram Pratim Bhuyan、Amar Ramdane-Cherif、Ravi Tomar 和 T. P. Singh。2024 36, 21 2024 12809 12844 10 1007 00521 024 09960。《神经符号人工智能：综述》。Neural Comput. Appl.（），–。doi: ./S - - -Z[9] Yongchao Chen、Jacob Arkin、Charles Dawson 等。2024。《AutoTAMP：以大语言模型作为翻译器和检查器的自回归任务与运动规划》。载于IEEE国际机器人与自动化会议（ICRA），日本横滨。IEEE， 6695 6702 10 1109 57147 2024 10611163 –。doi: ./ICRA。2024

[10] Dingxin Cheng、Mingda Li、Jingyu Liu 等。《借助基于分层事件的记忆增强长视频理解》。CoRR abs/2409.06299（ 2024 2409 06299 10 48550 2409 06299 ）。arXiv:。doi: ./ARXIV。Zesen Cheng、Sicong Leng、Hang Zhang 等。2024。《VideoLLaMA 2：推进视频大语言模型中的时空建模与音频理解》。CoRR abs/ 2406 07476 2024 2406 07476 10 48550 2406 07476 。（）。arXiv:。doi: ./ARXIV。Yoonyoung Cho、Junhyek Han、Jisu Han 和 Beomjoon Kim。2025。《面向通用环境中非抓取式操作的分层模块化网络》。CoRR abs/2502.20843（2025）。arXiv:2502.20843 doi:10.48550/ARXIV.2502. 20843[13] Brandon Curtis Colelough 和 William Regli。2025。《2024年神经符号人工智能：系统综述》。CoRR abs/2501.05435（2025）。arXiv:2501.05435 doi:10.48550/ARXIV.2501.05435[14] Dima Damen、Hazel Doughty、Giovanni Maria Farinella 等。2018。《扩展第一人称视角视觉：EPIC-KITCHENS数据集》。CoRR abs/1804.02748（2018）。arXiv: 1804 02748 1804 02748 。http://arxiv.org/abs/。Dima Damen、Hazel Doughty、Giovanni Maria Farinella 等。2022。《重新扩展第一人称视角视觉：EPIC-KITCHENS-100的采集、流水线与挑战》。Int. J. Comput. Vis. 130, 1 2022 33 55 10 1007 11263 021 01531 2 （），–。doi: ./S - - -Preetam Prabhu Srikar Dammu 、 Omar Alonso 和 Barbara Poblete。2025。《用于满足主观产品需求的购物智能体》。载于第十八届ACM国际网络搜索与数据挖掘会议（WSDM）论文集，德国汉诺威。ACM，1032 1035 10 1145 3701551 3704124 –。doi: ./。2025

[17] Ahmad Darkhalil、Dandan Shan、Bin Zhu 等。《EPIC-KITCHENS VISOR基准：视频分割与物体关系》。载于《神经信息处理系统进展35：NeurIPS会议》。[18] Shome S. Das。2021。《一种基于深度图像估计指向方向的方法及数据集，适用于人机交互和虚拟现实应用》。载于IEEE国际机器人与自动化会议（ICRA），中国西安。IEEE，11485–11491。doi:10.1109/ICRA48506.2021.9561143[19] Xiang Deng、Yu Gu、Boyuan Zheng 等。2023。《Mind2Web：迈向网络通用型智能体》。载于《神经信息处理系统进展36：2023年神经信息处理系统年度会议》，NeurIPS 2023。[20] Shangzhe Di 和 Weidi Xie。2024。《长篇第一人称视角视频中的基于依据的问答》。载于IEEE/CVF计算机视觉与模式识别会议（CVPR） *2024* 12934 12943 10 1109 52733 2024 01229 。IEEE，–。2026

[21] doi: ./CVPR。Yicheng Di、Hongjian Shi、Ruhui Ma 等。《FedRL：一种利用强化学习选择器和超网络生成器实现高效通信的强化学习联邦推荐系统》。Trans. Recomm. Syst. 4, 1（ 2026 7 1 7 31 10 1145 3682076 ），：–：。doi: ./ Yicheng Di、Xiaoming Wang、Hongjian Shi 等。2025。《利用细粒度变换和混合信息共享的个性化消费者联邦推荐系统》。IEEE消费者电子学汇刊 71, 2（2025），7254–7268。doi:10.1109/TCE.2025.3526427[23] Chenyou Fan。2019。《EgoVQA——第一人称视角视频问答基准数据集》。载于IEEE/CVF国际计算机视觉会议研讨会，ICCV研讨会。IEEE，4359 4366 10 1109 2019 00536 –。doi: 。/ICCVW. 。Tony Haoran Feng、Paul Denny、Burkhard C. Wünsche 等。2024。《为人工智能慧眼识物：利用计算机图形学问题评估 GPT-4o 的视觉感知能力与几何推理能力》。载于SIGGRAPH Asia 教育者论坛，SA，日本东京。ACM，5 1 5 8 10 1145 3680533 3697064 : – : 。doi: 。/ 。Pascale Fung、Yoram Bachrach、Asli Celikyilmaz 等。2025。具身人工智能体：建模世界。CoRR abs/2506.22355（2025）。arXiv:2506.22355 doi:10.48550/ARXIV.2506.22355[26] Google DeepMind。2025。Gemini 2.5 Pro Preview 模型卡。技术报告。谷歌。https://storage.googleapis.com/model-cards/documents/gemini-2.5- pro-preview.pdf 技术报告（预览版发布）。[27] Kristen Grauman、Andrew Westbury、Eugene Byrne 等。2022。《Ego4D：三千小时自我中心视频，纵览世界》。载于 IEEE/CVF计算机视觉与模式识别会议，CVPR，美国路易斯安那州新奥尔良。IEEE，18973–18990。doi:10.1109/CVPR52688.2022.01842[28] Hongliang He、Wenlin Yao、Kaixin Ma 等。2024。《WebVoyager：使用大型多模态模型构建端到端网页智能体》。载于计算语言学协会年会论文集，ACL 2024。计算语言学协会，6864–6890。doi:10.18653/V1/2024.ACL-LONG. 371[29] You He、Shulan Ruan、Dong Wang 等。2025。《大型人工智能模型驱动的智能决策：进展、挑战与展望》。CAAI智能技术汇刊 10，6（2025），1573–1592。[30] Pascal Hitzler、Aaron Eberhart、Monireh Ebrahimi 等。2022。人工智能中的神经符号方法。《国家科学评论》9，6（2022年03月），nwac035。doi:10.1093/nsr/nwac035[31] Kung-Hsiang Huang、Can Qin 等。2025。《为何视觉语言模型难以应对视觉算术？迈向更强的图表与几何理解能力》。载于计算语言学协会会议论文集，ACL。ACL，4830–4843。[32] Yichao Huang、Xiaorui Liu、Xin Zhang 和 Lianwen Jin。2016。《基于指向手势的第一人称视角交互系统：数据集、方法与应用》。载于2016年IEEE计算机视觉与模式识别会议研讨会，CVPR研讨会。IEEE计算机学会， 370 377 10 1109 2016 53 – 。2025

[33] doi: 。/CVPRW. 。Yifei Huang、Jilan Xu 等。。Vinci：基于自我中心视觉语言模型、面向便携设备的实时智能助手。《ACM交互、移动、可穿戴与普适技术汇刊》 9, 3 2025 88 1 88 33 10 1145 3749513 ( )，：– : 。2024 4

[34] doi: 。/ Aaron Hurst、Adam Lerer、Adam P. Goucher 等。。GPT- o 系统卡。CoRR abs/ 2410 21276 2024 2410 21276 10 48550 2410 21276 。（）。2024

[35] arXiv: 。doi: 。/ARXIV. 。Minseo Kwon、Yaesol Kim 等。。使用神经符号语言模型和多层次目标分解进行快速、准确的任务规划。CoRR abs/ 2409 19250 2024 2409 19250 10 48550 2409 19250 。（）。arXiv: 。2025

[36] doi: 。/ARXIV. 。Bo Li、Yuanhan Zhang、Dong Guo 等。。LLaVA-OneVision：轻松实现视觉任务迁移。《机器学习研究汇刊》2025（2025）。[37] Xiang Li、Heqian Qiu、Lanxiao Wang 等。2025。《第一人称视角视觉的挑战与趋势：综述》。CoRR abs/2503.15275（2025）。arXiv:2503.15275 doi:10.48550/ARXIV.2503.15275[38] Xinhao Li、Yi Wang、Jiashuo Yu 等。2025。《VideoChat-Flash：面向长上下文视频建模的分层压缩》。CoRR abs/2501.00574（2025）。arXiv:2501.00574 doi:10.48550/ARXIV.2501.00574[39] Xi Li、Xiaoxu Wu、Lijuan Xiao 等。2025。《MDSD：用于理解领域特定不完整请求的多轮多样化合成对话生成》。《SSRN电子期刊》（2025），1–9。[40] Bin Lin、Yang Ye、Bin Zhu 等。2024。《Video-LLaVA：通过先对齐再投影学习统一视觉表征》。载于自然语言处理实证方法会议论文集，EMNLP。计算语言学协会， 5971 5984 10 18653 1 2024 342 – 。2025

[41] doi: 。/V / 。EMNLP-MAIN。LiveKit 贡献者。。LiveKit：开源 WebRTC 与实时人工智能基础设施。https://github.com/livekit/livekit。[42] Yaxi Lu、Shenzhi Yang、Cheng Qian 等。2025。主动智能体：推动大语言模型智能体从被动响应转向主动辅助。载于第十三届国际学习表征会议，ICLR，新加坡。OpenReview.net。[43] Atharv Mahesh Mane、Dulanga Weerakoon 等。2025。Ges3ViG：将指向手势融入基于语言的三维视觉定位，以理解具身指代。载于 IEEE/CVF计算机视觉与模式识别会议，CVPR。计算机视觉基金会 / IEEE，9017–9026。[44] Spyridon Mouselinos、Henryk Michalewski 和 Mateusz Tomasz Malinowski。2024。《超越线条与圆形：揭示大语言模型中的几何推理差距》。载于计算语言学协会会议论文集：EMNLP *2024* 6192 6222 10 18653 1 2024 360 。ACL，– 。2024 2

[45] doi: 。/V / 。FINDINGS-EMNLP。Wiktor Mucha、Florin Cuconasu、Naome A. Etori 等。。TEXT TASTE：一种利用大语言模型提供智能阅读辅助的通用第一人称视角视觉系统。载于《帮助特殊需求人群的计算机》——第19届国际会议，ICCHP 2024（计算机科学讲义，第 *14751* 285 291 10 1007 978 3 031 62849 8_35 卷）。Springer，– 。2025

[46] doi: 。/ - - - - Alkesh Patel、Vibhav Chitalia 和 Yinfei Yang。。利用多模态大语言模型推进自我中心视频问答。CoRR abs/2504.04550（2025）。arXiv:2504.04550 doi:10.48550/ARXIV.2504.04550

[47] Avinash Patil。2025。推进大语言模型推理：有前景的方法与途径。CoRR abs/2502.03671（2025）。arXiv:2502.03671 doi:10. 48550/ARXIV.2502.03671[48] Shishir G. Patil、Tianjun Zhang、Xin Wang 等。2024。Gorilla：连接海量 API 的大语言模型。载于《神经信息处理系统进展》第38卷：2024年神经信息处理系统年会，NeurIPS 2024，加拿大不列颠哥伦比亚省温哥华，2024年12月10日至15日。[49] Taiying Peng、Jiacheng Hua、Miao Liu 等。2025。《在多模态大语言模型的视野中：使用注视引导提示对第一人称视角视频意图理解进行基准测试》。CoRR abs/ 2509 07447 2025 2509 07447 10 48550 2509 07447 。（）。[50] 2025 arXiv: 。doi: 。/ARXIV. 。Toby Perrett、Ahmad Darkhalil、Saptarshi Sinha 等。。HD-EPIC：一个高精度第一人称视角视频数据集。载于 IEEE/CVF 计算机视觉与模式识别会议，CVPR。计算机视觉基金会 / IEEE，23901–23913。[51] Zhangyang Qi、Zhixiong Zhang、Ye Fang 等。2025。GPT4Scene：使用视觉语言模型从视频理解三维场景。CoRR abs/2501.01428（ 2025 2501 01428 10 48550 2501 01428 ）。arXiv: 。doi: 。/ [52] 2024 ARXIV. 。Cheng Qian、Bingxiang He、Zhong Zhuang 等。。告诉我更多！迈向对语言模型驱动智能体的隐式用户意图理解。载于计算语言学协会第62届年会论文集。ACL， 1088 1113 10 18653 1 2024 61 – 。doi: 。/ V / 。[53] 2025 ACL-LONG。Ram Ramrakhya、Matthew Chang、Xavier Puig 等。。通过强化学习，将多模态大语言模型与会主动求助的具身智能体对接。CoRR abs/2504.00907（2025）。arXiv:2504.00907 doi:10.48550/ARXIV.2504. 00907[54] Machel Reid、Nikolay Savinov、Denis Teplyashin 等。2024。Gemini 1.5：解锁跨数百万上下文词元的多模态理解能力。CoRR abs/ 2403 05530 2024 2403 05530 10 48550 2403 05530 。（）。[55] 2024 arXiv: 。doi: 。/ARXIV. 。Shulan Ruan、Kun Zhang、Le Wu 等。。用于图像情感分析的颜色增强交叉相关网络。《IEEE多媒体汇刊》26（2024），4097–4109。doi:10.1109/TMM.2021.3118208[56] Timo Schick、Jane Dwivedi-Yu、Roberto Dessì 等。2023。Toolformer：语言模型可以自学使用工具。载于《神经信息处理系统进展》第36卷：神经信息处理系统年会，NeurIPS，美国路易斯安那州新奥尔良。[57] Ashish Seth、Utkarsh Tyagi 等。2025。EGOILLUSION：对第一人称视角视频理解中的幻觉进行基准测试。载于2025年自然语言处理实证方法会议论文集。28449–28468。[58] Aditya Sharma、Aman Dalmia、Mehran Kazemi 等。2025。GeoCoder：通过视觉语言模型生成模块化代码来解决几何问题。载于计算语言学协会会议论文集：北美计算语言学协会，美国新墨西哥州阿尔伯克基。ACL，7340–7356。doi:10.18653/V1/2025.FINDINGS-NAACL.410[59] Tsukasa Shiota、Motohiro Takagi、Kaori Kumagai 等。2024。通过捕捉手物接触和物体状态进行第一人称视角动作识别。载于 IEEE/CVF冬季计算机视觉应用会议，WACV，美国夏威夷州威可洛亚。IEEE， 6527 6537 10 1109 57701 2024 00641 – 。doi: 。/WACV . 。[60] 2023 Ishika Singh、Valts Blukis 等。。ProgPrompt：使用大语言模型生成情境化机器人任务规划。载于国际机器人与自动化会议，ICRA。IEEE，11523–11530。doi:10.1109/ICRA48891.2023.10161317[61] Kiran K. Somasundaram、Jing Dong、Huixuan Tang 等。2023。Project Aria：第一人称视角多模态人工智能研究的新工具。CoRR abs/2308.13561（2023）。arXiv:2308.13561 doi:10.48550/ARXIV.2308.13561[62] Branislav Sredojev、Dragan Samardzija 等。2015。WebRTC技术概述及信令解决方案的设计与实现。载于国际信息与通信技术、电子与微电子大会，MIPRO *2015* 1006 1009 10 1109 2015 7160422 。IEEE，–。[63] 2025 1 doi: 。/MIPRO. 。Shulin Tian 、 Ruiqi Wang 、 Hongming Guo 等。。Ego-R：面向超长第一人称视角视频推理的工具链式思维。CoRR abs/2506.13654（ 2025 2506 13654 10 48550 2506 13654 ）。arXiv: 。doi: 。/ [64] 2025 ARXIV. 。Shulin Tian、Ziniu Zhang、Liangyu Chen 等。。MMInA：对多跳多模态互联网智能体进行基准测试。载于计算语言学协会会议论文集，ACL 2025。ACL，13682–13697。[65] Nenad Tomašev、Julien Cornebise 等。2020。公益人工智能：释放创造积极影响的机遇。《自然·通讯》11，1（2020），2468。[66] Oguzhan Topsakal 和 Tahir Cetin Akinci。2023。利用 LangChain 创建大语言模型应用：快速开发大语言模型应用入门。载于国际应用工程与自然科学会议，第1卷。1050–1056。[67] Zishen Wan、Che-Kai Liu、Hanchen Yang 等。2024。《迈向认知人工智能系统：神经符号人工智能综述与展望》。CoRR abs/2401.01040（ 2024 2401 01040 10 48550 2401 01040 ）。arXiv: 。doi: 。/ [68] ARXIV. 。Jinqiang Wang、Huansheng Ning、Tao Zhu 和 Jianguo Ding。2025。一种由大语言模型驱动的数据合成方法，用于主动挖掘旅游领域中的隐式用户意图。CoRR abs/2505.11533（2025）。arXiv:2505.11533 doi:10.48550/ARXIV.2505.11533

[69] Lei Wang、Chen Ma、Xueyang Feng等。2024。基于大语言模型的自主智能体综述。Frontiers Comput. Sci. 18, 6（2024），186345。doi:10.1007/S11704-024-40231-1[70] Ying Wang、Yanlai Yang和Mengye Ren。2023。LifelongMemory：利用大语言模型回答自我中心视频中的查询。CoRR abs/2312.05269（2023）。arXiv:2312.05269 doi:10.48550/ARXIV.2312.05269[71] Jason Wei、Xuezhi Wang、Dale Schuurmans等。2022。思维链提示激发大语言模型中的推理。在《神经信息处理系统进展》第35卷：NeurIPS年度会议。[72] Lilian Weng。2023。大语言模型驱动的自主智能体。lilianweng.github.io（2023年6月）。https://lilianweng.github.io/posts/2023-06-23-agent/ [73] Zhuofeng Wu、Richard He Bai、Aonan Zhang等。2024。拆分还是征服？应该蒸馏大语言模型的哪一部分？载于《计算语言学协会会议论文集：EMNLP》。ACL，2572–2585。doi:10.18653/V1/2024. FINDINGS-EMNLP.145[74] xAI。2025。Grok 3 Beta——推理智能体时代。https://x.ai/blog/grok-3。[75] Haoyi Xiong、Zhiyuan Wang等。2024。范式融合：符号人工智能与连接主义人工智能在大语言模型赋能的自主智能体中的协同作用。CoRR abs/ 2407 08516 2024 2407 08516 10 48550 2407 08516 。（）。[76] 2025 arXiv: 。doi: 。/ARXIV. 。Jingkang Yang、Shuai Liu、Hongming Guo等。Egolife：迈向自我中心生活助理。载于《计算机视觉与模式识别会议论文集》。28885–28900。[77] Sicheng Yang、Yukai Huang、Weitong Cai等。2025。即插即用澄清器：用于第一人称视角意图消歧的零样本多模态框架。arXiv预印本 arXiv:2511.08971（2025）。[78] Sicheng Yang、Yukai Huang、Shitong Sun等。2026。面向第一人称视角视频理解优化多模态大语言模型：HD-EPIC视觉问答挑战赛解决方案。arXiv: 2601 10228 [ ] 2601 10228 。cs.CV [79] 2023 https://arxiv.org/abs/ 。Shunyu Yao、Jeffrey Zhao、Dian Yu等。ReAct：语言模型中推理与行动的协同。在第十一届国际学习表征会议（ICLR 2023），基加利，卢旺达，2023年5月1日至5日。OpenReview.net。[80] Hanrong Ye、Haotian Zhang、Erik A. Daxberger等。2025。MMEgo：迈向面向视频问答的自我中心多模态大语言模型。载于第十三届国际学习表征会议（ICLR），新加坡。OpenReview.net。[81] Asaf Yehudai、Lilach Eden等。2025。基于大语言模型的智能体评估综述。CoRR abs/2503.16416（2025）。arXiv:2503.16416 doi:10.48550/ARXIV.2503. 16416[82] Kexin Yi、Jiajun Wu、Chuang Gan等。2018。神经符号视觉问答：将推理与视觉和语言理解分离。在《神经信息处理系统进展》第31卷：神经信息处理系统年度会议（NeurIPS），加拿大蒙特利尔。1039–1050。[83] Ori Yoran、Samuel Joseph Amouyal、Chaitanya Malaviya等。2024。AssistantBench：网络智能体能否完成真实且耗时的任务？载于《2024年自然语言处理经验方法会议论文集》（EMNLP） *2024* 8938 8968 10 18653 1 2024 505 。ACL，–。[84] 2023 doi: 。/V / .EMNLP-MAIN。Dongran Yu、Bo Yang等。神经符号学习系统综述。Neural Networks 166 2023 105 126 10 1016 2023 06 028 （），–。[85] 2025 doi: 。/J.NEUNET. 。Zihao Yue、Zhenru Lin、Yifan Song等。MiMo-VL技术报告。CoRR abs/ 2506 03569 2025 2506 03569 10 48550 2506 03569 。（）。[86] 2024 arXiv: 。doi: 。/ARXIV. 。Dell Zhang、Yongxiang Li等。借助大语言模型赋能智能眼镜：迈向无处不在的通用人工智能。载于2024年ACM普适与泛在计算国际联合会议伴随论文集。631–633。[87] Guanhua Zhang、Mohamed Adel Naguib Ahmed、Zhiming Hu等。2025。SummAct：通过交互行为摘要揭示用户意图。载于CHI计算系统人因会议论文集，CHI，横滨日本。ACM， 265 1 265 17 10 1145 3706598 3713190 : –: 。doi: 。/。[88] 2025 2 Haoyu Zhang、Qiaohui Chu、Meng Liu等。Exo Ego：由外部视角知识引导、用于第一人称视角视频理解的多模态大语言模型。CoRR abs/2503.09143（ 2025 2503 09143 10 48550 2503 09143 ）。arXiv: 。doi: 。/ARXIV. 。[89] 2024 Hangtao Zhang、Chenyu Zhu、Xianlong Wang等。BadRobot：在物理世界中越狱基于大语言模型的具身人工智能。CoRR abs/2407.20242（2024）。arXiv:2407.20242 doi:10.48550/ARXIV.2407.20242[90] Peiyuan Zhang、Kaichen Zhang、Bo Li等。2024。从语言到视觉的长上下文迁移。CoRR abs/2406.16852（2024）。arXiv:2406.16852 doi:10.48550/ ARXIV.2406.16852[91] Tong Zhang、Peixin Qin、Yang Deng等。2024。CLAMBER：用于识别和澄清大语言模型中模糊信息需求的基准测试。载于计算语言学协会第62届年会论文集。ACL， 10746 10766 10 18653 1 2024 578 –。doi: 。/V / . [92] 2024 ACL-LONG。Xuan Zhang、Yang Deng等。先询问后规划：面向现实世界规划的主动式语言智能体。载于计算语言学协会会议：EMNLP。ACL， 10836 10863 10 18653 1 2024 636 –。doi: 。/V / . [93] 2025 FINDINGS-EMNLP。Xuan Zhang、Yongliang Shen、Zhe Zheng等。AskToAct：通过自我纠正式澄清增强大语言模型的工具使用能力。CoRR abs/2503.01940（2025）。arXiv:2503.01940 doi:10.48550/ARXIV.2503.01940[94] Yichi Zhang、Xin Luna Dong、Zhaojiang Lin等。2025。基于流式第一人称视角视频的主动式助理对话生成。CoRR abs/2506.05904（2025）。arXiv:2506.05904 doi:10.48550/ARXIV.2506.05904

[95] Boyuan Zheng、Boyu Gou、Jihyung Kil、Huan Sun和Yu Su。2024。如果有依据，GPT-4V(ision)就是通用型网络智能体。载于第四十一届国际机器学习会议，*ICML 2024*，奥地利维也纳，*2024*年*7*月*21*日至*27*日。OpenReview.net。[96] Zhuo Zhi、Qiangqiang Wu等。2025。VideoAgent2：通过不确定性感知思维链增强用于长视频理解的基于大语言模型的智能体系统。*CoRR* abs/ 2504 04471 2025 2504 04471 10 48550 2504 04471 。（）。[97] arXiv: 。doi: 。/ARXIV. 。Luowei Zhou、Chenliang Xu和Jason J. Corso。从网络教学视频中自动学习程序。载于第三十二届*AAAI*人工智能会议论文集（*AAAI-18*）。AAAI出版社，7590–7598。doi:10.1609/AAAI.V32I1.12342[98] Shuyan Zhou、Frank F. Xu、Hao Zhu等。2024。WebArena：用于构建自主智能体的真实网络环境。载于第十二届国际学习表征会议，*ICLR 2024*。OpenReview.net。[99] Zhifan Zhu和Dima Damen。2023。抓稳：在第一人称视角视频中重建稳定的手物抓握姿势。*CoRR* abs/2312.15719（2023）。arXiv:2312.15719 doi:10.48550/ARXIV.2312.15719[100] Jie Zou、Jimmy Xiangji Huang、Zhaochun Ren等。2024。学习如何提问：通过表征学习实现对话式商品搜索。*CoRR* abs/2411.14466（2024）。arXiv:2411.14466 doi:10.48550/ARXIV.2411.14466

A 扩展方法

为求完整，我们简要总结了正文中未详述的若干 Egocentric Co-Pilot 实现细节。统一事件日志 E 通过以 1 FPS 对自我中心视频进行采样，并运行经过微调的 Qwen2.5-VL-7B-Instruct 模型 [5] 来构建，以生成关于动作、物体状态变化和场景上下文的密集第一人称描述。这些视觉条目与自动语音识别转录文本合并，形成按时间排序的事件序列。我们会预处理查询，分析其模态要求（图像、视频或混合模态），将描述不充分的问题改写为明确的、基于视角的提示，并将多项选择选项重新格式化为统一模板；这能减少解析歧义并提升稳健性。

自我中心骨干模型通过在多种第一人称视频数据集上微调 Qwen2.5-VL-7B-Instruct [5] 获得，其中包括 EPIC-KITCHENS [14, 15], EgoProceL [6], YOUCOOK2 [97], VISOR [17], EgoIT [76], 以及 Ego4D [27]的相关部分。我们冻结视觉塔和投影器，使用 AdamW 更新大语言模型层（学习率 2×10^{-7}、批量大小为 2、训练一个轮次、采用 bfloat16 精度），并限制帧数和序列长度（最多 131,072 个词元）。时间链式思维（T-CoT）通过简单的提示模板来实现，这些模板鼓励模型进行中间推理步骤；同时，我们按第 3.1 节所述，以编程方式裁剪或拼接时间窗口。后处理包括使用正则表达式提取答案字母，并对每个问题的五个语法形式各异的提示进行多数投票。

广义执行循环

第一人称视角副驾驶使用的广义编排循环详见算法1。

棋盘游戏工具实现细节

棋盘游戏副驾驶是我们神经符号工具使用的典型示例。为了使感知到的棋盘状态不受逐帧检测噪声的影响，我们会在棋盘的每个位置维护一个预测时间缓冲区 $V_{(i)r,\ c,\ k}$ 其中（r，c）处，k 表示棋子类型索引，i 表示帧索引。最终确定的状态 $P_{r,\ c}$ 通过以下方式获得：

算法1大语言模型编排的广义执行循环

要求： 用户查询 Q、多模态上下文 C_{mm}
$T_{available} \leftarrow$ MCP.ListTools() ▷ 发现可用工具
$tool_{plan} \leftarrow$ 大语言模型.GeneratePlan($Q, C_{mm}, T_{available}$) ▷ 大语言模型 制定工具使用计划
对于所有 $call \in tool_{plan}$ **执行**
$Tool_{selected}, Args \leftarrow call.name, call.arguments$
$result \leftarrow$ MCP.CallTool($Tool_{selected}$， $Args$)
使用 $result$更新执行上下文
结束循环
$R_{final} \leftarrow$ LLM.SynthesizeResponse(Q，执行上下文) ▷ 生成最终输出
返回 R_{final}

通过多数投票并设置稳定性阈值 τ：

$$P_{r,c} = \begin{cases} \arg\max_k \sum_{i=1}^{N} V_{r,c,k}^{(i)} & \text{if } \frac{1}{N} \max_k \sum_{i=1}^{N} V_{r,c,k}^{(i)} \geq \tau, \\ P_{r,c}^{\text{prev}} & \text{otherwise,} \end{cases} \tag{3}$$

其中，$P_{r\ c}^{\text{prev}}$，表示在（r，c）处先前提交的状态，N 表示缓冲区大小，而 τ 用于控制响应速度与稳定性之间的权衡。

执行逻辑封装在算法 2 中，展示了如何将视觉流转换为自然语言建议。

算法 2 混合神经符号国际象棋风格工具执行

要求：视觉流 V、大语言模型 $\mathbf{M}_{\text{LLM}}$、符号引擎 $\mathbf{S}_{\text{eng}}$ 确保：自然语言战略建议 A 1：函数 ExecuteBoardTool(V) 2：感知： $S_{\text{FEN}} \leftarrow$ **PerceiveStableState(V)** ▷ **使用式 (3) 3：符号搜索：** $M_{\text{sym}} \leftarrow \mathbf{S}_{\text{eng}}$**.GetBestMove(**$S_{\text{FEN}}$**) 4：语义解释：5：** $P \leftarrow$ **“作为一名棋盘游戏教练，请根据当前位置解释走法** M_{sym} **背后的思路。” 6：A** $\leftarrow \mathbf{M}_{\text{LLM}}$**.Generate(**$P$**) 7：返回 A 8：结束函数**

在工具端，每项能力都通过装饰器注册到 MCP 中，该装饰器会公开带类型注释的签名和文档字符串。算法 2 所述的棋盘游戏副驾驶使用紧凑型卷积网络进行逐格分类，使用式 3 中的时间平滑规则来稳定感知到的位置，并以标准国际象棋引擎作为符号核心。在我们的智能眼镜原型中，同一 MCP 注册表还公开了网络 API（例如天气或营养相关 API）、本地实用工具（例如笔记和提醒事项），以及设备桥接工具，用于向附近的手机或计算机发送结构化 JSON 消息。

B 客户端设备上的实时音频处理

算法 3 详细介绍了我们的智能眼镜原型所采用的设备端音频流水线。该流水线实现了轻量级语音活动检测、预录制缓冲区和插话检测，以便将语音片段

算法 3 客户端设备上的实时音频处理

```
初始化环形缓冲区 B_ring，状态 S ← 空闲
定义阈值 θ_start、θ_barge-in 以及时长 T_silence、T_min
当真时执行
  获取音频块 b，设置 b′ ← g · b 为 g = 5.0
    更新 B_ring 为 b′，令 A ← 取最大值(|b′|)
    如果系统正在播放音频 且 A > θ_barge-in 则
      停止播放() ▷ 检测到插话
    结束 if
    如果 S =空闲 则
      如果 A > θ_start 那么
        S ←录制
        开始新的片段，使用 B_ring
      结束如果
    否则，如果 S =录制 那么
      追加 b′ 到当前片段
      如果 A < θ_start 则
        启动或继续时长为 T_silence的静音计时器
        如果计时器已到期则
          完成片段 S_audio
          如果时长(S_audio) > T_min 则
            调度(S_audio)
          结束条件判断
          S ←空闲
        结束条件判断
      else
        重置静音计时器
      结束条件判断
    结束条件判断
</style}结束
while
```

以低延迟传输至云端，同时允许用户在需要时打断正在进行的播放。第 3.3 节所述的所有基于 WebRTC 的部署均采用了这一客户端流水线。

C 关于Egolife和EgoGPT的其他结果

为补充上下文，我们总结了 [76] 中报告的 EgoGPT 在 Egolife 基准测试上的离线结果。EgoGPT 采用多阶段检索增强生成（RAG+）流水线，并进行大量离线处理：首先为长时段自我中心视频日志建立索引，然后在整个数据集上执行多轮大语言模型推理。在这一设置下，EgoGPT（EgoIT–Egolife）的准确率为 38.5%，EgoGPT（EgoIT）在 Ego-life [76] 上的准确率为 42.6%。然而，在我们的设置中复现这种方案，需要在整个数据集上多次运行离线推理，这与我们对智能眼镜低延迟、始终在线辅助的关注以及能量感知部署的考量不相容。因此，我们仅将这些 EgoGPT 数值作为上下文参考，而不将其视为可直接比较的基线；表1中的主要比较则集中于单次处理、适合流式处理的系统。

D 更多 Human 评估详情

第 4.4 节的人在回路研究邀请了四名此前有使用 AI助手经验的参与者。对于每种任务场景，我们收集了九个系统的记录：我们的第一人称视角副驾驶、若干商用设备以及人类助手基线。记录包括音频转录文本、关键视频帧以及对系统操作的简短文本摘要（匿名化片段示例见图 6，这些片段展示给参与者观看）。所有记录均经过匿名化并随机打乱，因此评分者无法得知系统身份。参与者使用 5分李克特量表对每条记录的两个方面进行评分：（1）助手对多模态意图的理解程度；（2）助手成功执行任务的程度。图 5 中报告的分数是这两个维度的平均值。采用非标准交互模式的设备（例如以通知为主的设备或单一功能应用）以星号标注。由于研究使用了预先录制且完全匿名化的记录，任务为低风险的日常任务，且不涉及敏感人群，因此根据当地指南，该研究被归类为最小风险；参与者均提供了知情同意，并可随时退出且不受任何惩罚。

E 关于系统部署与局限性的补充讨论

本节将进一步介绍系统的部署约束、安全考量和算法适应性，并回应针对第一人称视角流式智能体实际应用所提出的具体关切。

E.1 连接性与离线回退策略

关于系统在网络状况不佳时的行为（离线回退）：鉴于可穿戴增强现实设备受到严格的硬件限制，我们当前的架构优先采用云端方案。

- 硬件限制：部署设备（RayNeo X2 Pro）在尺寸、重量和功耗方面受到显著限制。尽管我们尝试在眼镜本地部署轻量级模型（例如 0.5B 参数模型），但推理延迟过高，无法满足实时交互的需求，且模型容量不足以支持复杂推理。
- 设计选择：因此，我们目前尚未为复杂查询实现完整的离线回退机制。系统面向高带宽环境（WiFi/4G）设计，利用云端进行高强度计算，以维持可穿戴设备形态。未来的迭代可能会探索混合卸载，但目前稳定连接是先决条件。

E.2 数据隐私与安全

关于数据保护，我们承认，这项工作主要聚焦于自我中心智能体的架构可行性。传输采用标准 Web 协议。端到端加密和严格的数据保留政策（例如，推理完成后立即删除数据）计划在生产阶段实施，但尚未在此原型中实现。我们为未来工作提出一种“隐私优先的混合架构”：在边缘设备本地处理或屏蔽敏感视觉数据（例如人脸、文本），并且仅将非敏感抽象特征传输至云端。

E.3 算法适应性：HCC 和 T-CoT

我们希望阐明，与标准的检索增强生成或提示工程相比，我们的历史上下文控制（HCC）和时间思维链（T-CoT）有何独特之处。

- 在线离线检索增强生成：传统检索增强生成需要对离线数据进行索引，这会增加延迟，也不适用于自我中心视频连续、流式处理的特性。我们的 HCC 机制会在流中动态执行由粗到细检索，与从静态向量数据库中检索相比，大幅缩短了耗时。
- 处理长上下文：标准多模态大语言模型在输入较长的视频历史记录时，难以应对“迷失在中间”现象。我们通过动态压缩，模仿人类记忆模式（近因偏差）。HCC 与 T-CoT 的结合专为第一人称视频的时间依赖关系优化，因为理解刚刚发生的事情往往比了解久远的历史更重要。

E.4 工具箱方法的可扩展性

“工具箱”机制旨在构建可扩展的混合代理系统，而非僵化的规则集合。该系统遵循标准的代理范式：为高精度任务定义特定工具（应用程序接口）（例如日历、天气）。然而，当用户的意图与预定义工具不匹配（或属于日常生活中的“长尾需求”）时，系统会退回到底层多模态大语言模型的通用能力（零样本视觉问答）。

E.5 硬件性能：电池与热约束

摄像头、显示屏与高频网络传输的同时运行极为耗电。在无外部供电的连续流传输模式下，设备电池大约可维持运行20分钟。我们观察到，随着电池电量下降，设备固件会触发省电模式，显著限制性能（例如中央处理器），从而导致系统卡顿。热耗散仍处于用户舒适度可接受的范围内，但设备在长时间使用时会产生明显热量。这些发现进一步说明，我们的云卸载架构对于尽量降低设备端计算负载是必要的，尽管电池技术仍是所有增强现实硬件的瓶颈。



图 6：向参与者展示的交互日志示例。每一列对应于不同系统对同一用户查询的响应。通过评估预先录制的日志，而不是实时交互，我们避免了硬件、网络或用户界面差异对人工智能质量造成的混淆。