实现有效的人在回路中的辅助 AI 代理

Filippos Bellos[1,ꕥ] Yayuan Li[1,ꕥ] Cary Shu[1] Ruey Day[1]
Jeffrey M. Siskind[2] Jason J. Corso[1,3]

[1]University of Michigan [2]Purdue University [3]Voxel51
ꕥEqual Contribution
{ fbellos, yayuanli } @umich.edu

Abstract

有效的人机协作在日常活动和专业领域的物理任务完成中具有重要的潜力。配备有信息指导的 AI 代理可以提升人类表现，但由于人类参与互动的复杂性，评估这种协作仍然具有挑战性。在这项工作中，我们引入了一个评估框架和一个多模态的人机交互数据集，旨在评估 AI 指导如何影响程序性任务表现、错误减少和学习成果。此外，我们开发了一种增强现实（AR）装备的 AI 代理，它在从烹饪到战场医疗的真实任务中提供互动指导。通过人类研究 [1]，我们分享了关于 AI 辅助的人类表现的实证见解，并展示了 AI 辅助协作如何改善任务完成。

1. 介绍

人工智能（AI）的最新进展主要得益于大型语言模型（LLMs）的快速发展，比如 GPT [16]、Claude [1]、Gemini [22]、Qwen [2, 26] 和 LLaMA [23]，这些模型在语言理解、推理和复杂任务执行方面展现了令人印象深刻的能力。这些模型中整合多模态功能进一步拓展了人工智能的应用范围，使得系统能够处理和生成跨越文本、图像和视频的多样化内容。尽管有这些进展，AI 代理在涉及人工参与场景中的实际部署——即 AI 与人共同完成物理任务的场景——仍然是一个未被充分探索的领域，尤其是在动态和互动的环境中。（见图 1）。

辅助自主中的人机协作正成为一种关键范式，其中 AI 代理通过提供上下文感知的指导和互动反馈来增强人类的任务完成能力 [4, 7, 8, 13, 15, 18, 20, 25]。尽管 AI 驱动的代理在受控环境中表现出有效性，但它们在高风险、动态物理任务（如医疗程序）中的应用带来了独特的挑战。通过装备 AI 的增强现实 (AR) 代理增强人类认知能力在专业领域（如战场医疗和手术辅助）以及日常活动（如烹饪和装配）中显示出潜力。然

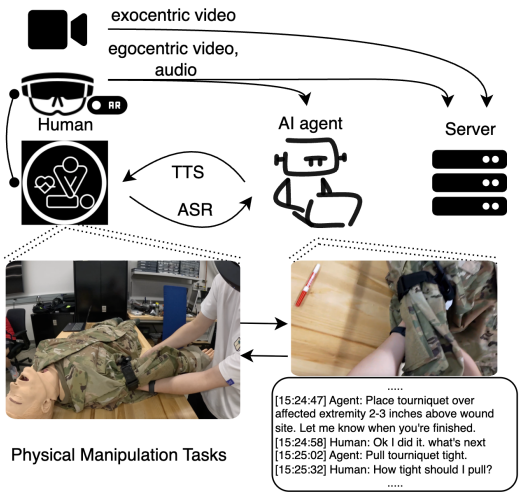

Figure 1. 指导人类完成物理任务的交互式 AI 代理的工作流程。人类在物理世界中执行任务（例如，绑止血带）并与 AI 代理通过 AR 头戴设备进行交流。

而，推进 AI 辅助协作的一个主要限制是缺乏结构化的评估框架，这些框架可以严格评估 AI 代理在有人参与的任务完成中的有效性，而不仅仅是传统自动化基准 [8, 10, 24]。

为填补这一空白，我们引入了一个综合评估框架，专门用于评估在现实世界中人机协作中由 AI 引导的物理任务完成。尽管以前的评估方法主要集中于语言模型或模拟任务，它们未能反映具体现实体协助的复杂性——如交互式指导、执行质量、用户体验和学习成果。我们的框架定义了在这些维度上量化 AI 有效性的重要指标，并基于新收集的人机交互多模态数据集。该数据集包括同步的自我中心和外部视角的视频、音频以及步骤级结果、错误类别和自然语言理由的详细注释——使得对程序性支持和用户适应的细粒度分析成为可能。虽然不是普遍的基准，这一框架代表了朝着在物理环境中对具体现实体 AI 代理进行严格和可重复评估

[1]在研究开始之前，我们机构的伦理审查委员会（IRB）批准了这项关于人类受试者的研究。

的重要第一步。

为了支持该框架，我们开发了一个用于现实世界物理任务协作的 AI 驱动系统。其核心是一个感知能力增强的代理，它将基于 AR 的指导与实时任务监控和反馈相结合。该系统使我们能够执行和评估 AI 引导的协助，任务涵盖日常活动到如战场医疗等高风险领域。在此基础上，我们进行了广泛的人类研究，分析用户如何在物理环境中与 AI 引导系统进行交互。我们的研究考察了对 AI 工作流程的适应性、互动指导的认知影响，以及支持有效协作的设计原则。这些洞察加深了我们对 AI 代理作为协作伙伴而非被动自动化工具的理解，并突出展示了在现实世界部署中的关键挑战和机遇。

我们的贡献有三方面：

1. 一个用于衡量 AI 引导物理任务表现的结构化评估框架和数据集；
2. 一个交互式的基于增强现实的人工智能代理，用作现实世界人工智能任务指导的关键参考；
3. 关于人工智能辅助任务完成中表现、经验和工作流程设计的人类研究的实证发现和数据集。

2. 相关工作

最近在大型语言模型（LLMs）方面的进展促进了多模态大型语言模型（MLLMs）的发展，这些模型能够处理超越文本的输入，例如图像、音频和 3D 数据。值得注意的例子包括 OpenAI 的 ChatGPT [16] 、LLaVA [14] 、MiniGPT-4 [30] 、LLaMA-Adapter [27] 以及谷歌的 Gemini [22] 。这些模型扩展了视觉-语言推理功能，对于需要情境理解的任务指导来说是很有前景的。然而，尽管在结构化或虚拟环境中表现出色，但在需要物理适应性和实时互动的现实世界人类参与场景中，它们往往表现不佳。尽管像 “Watch, Talk and Guide” (WTaG) [3] 这样的先前工作探索了基础模型在烹饪任务指导中的零样本能力，但它们在更具挑战性的领域中大多忽视了高效通信的需求。相比之下，我们的工作部署了一个完全实现的配备增强现实（AR）的高效 AI 代理，并具有针对跨领域物理协作的专门 ML 模块——从烹饪到战场医疗——为评估 AI 在职业场合中的人机交互的流畅性、鲁棒性和适应性奠定了基础。

评估 AI 系统对于确保其在真实世界场景中的可靠性能至关重要，特别是在生成式 AI 扩展到任务指导应用中。现有的基准测试如 Big Bench Hard (BBH) [21] 、MMLU-PRO [24] 和 GenAI-Bench [10] 为语言和视觉模型提供了强有力的评估，但主要集中在虚拟任务上。同样，数据集如 IFEval [19] 和 MuSR [29] 测试跨模态和指令跟随能力，但未能捕捉到真实世界中人机交互的互动性和动态性。

任务指导系统需要反映物理交互、实时决策和用户适应性的评估框架。尽管像“观察、对话和指导”(WTaG) [3] 这样的工作收集了有价值的人机互动数据并评估了用户意图的理解，但它们主要集中在评估基础模型的能力，而非衡量 AI 指导系统的实际有效性。我们的框架通过结合任务成功、错误减少、用户满意度和稳健性等指标，填补了这一空白，使得对具身 AI 代理进行更实用和全面的评估成为可能。

3. 人机协作任务完成的评估框架

我们的框架评估了在 AI 辅助任务指导场景中的任务完成性能和用户交互的质量。它由一个多模态数据集支持，该数据集包括同步的自我中心和他人中心录制，以及从包含多项任务的 AI 辅助和非辅助任务执行中获得的详细注释。

3.1. 任务完成质量评估

此组件评估 AI 助手提供准确和及时指导的能力。我们定义了以下指标：

- 成功率指标：我们从两个层面衡量成功：
  - Macro Success Rate (M-SR): 该指标计算所有任务的所有样本的平均成功率。它提供了整个数据集上 AI 有效性的整体视图。
  - Micro Success Rate (μ -SR): 该指标首先分别计算每个任务中的平均成功率，然后平均这些任务特定的成功率。该方法确保了任务的样本大小无论如何都能获得均等的权重，并提供对 AI 在不同任务类型中表现的洞察。

  这两种指标有助于量化由 AI 助手带来的效率增益（或损失），尤其是与传统（非 AI）方法相比。更高的成功率表明 AI 在从头到尾成功帮助用户完成任务。
- 完成时间：我们测量了在有 AI 辅助和无辅助场景下完成任务所需的平均时间。完成时间的减少表明 AI 有效地简化了任务过程。
- 步骤错误率 (S-ER)：我们从用户的角度跟踪错误，包括在有 AI 指导时和没有 AI 指导时。我们将错误分为两种主要类型：
  - Critical Errors: 这些是使任务完成变得不可能的错误。它们代表了显著偏离正确程序的错误，无法不从头开始就进行补救。
  - Step-Specific Errors: 这些是在任务的各个步骤中发生的更细化的错误。这些错误可能是错误的操作（例如，搅拌而不是折叠），错误的对象（例如，用盐代替糖），错误的状态（例如，过度搅打鸡蛋）或其他不属于上述类别的步骤特定错误。
- 误差减少：我们报告了在有 AI 辅助和无 AI 辅助任务完成条件下的步骤误差率 (S-ER)。这些测量使我们能够评估 AI 在不同指导方法中对误差减少的影响。
- 步骤引导一致性：这是衡量 AI 提供的指令与任务当前步骤正确对应的速率。一致性高表明 AI 能保持对用户进度的了解，并提供符合情境的指导。

3.2. 用户交互质量评估

用户体验对 AI 代理的成功至关重要，它最终将决定其效能和采纳程度。评估用户交互指标对于理解用户如何与代理进行互动，以及识别系统可以改进以更好满足用户需求和期望的领域至关重要。

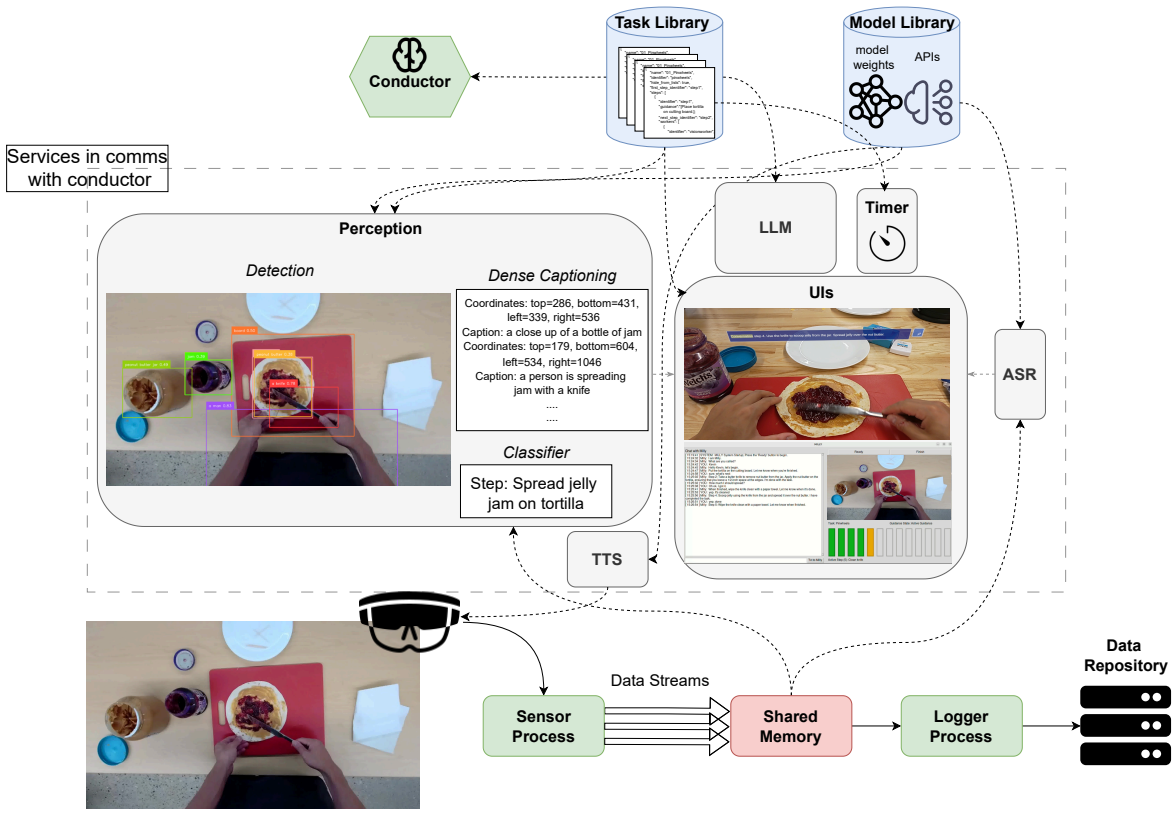

Figure 2. 基线交互代理用于物理任务辅助的概述。

- 清晰度：用户对 AI 生成指令的清晰度和可理解性进行评分。该指标对于理解 AI 传达指导的效果以及用户是否能够轻松解释提供的指令至关重要。
- 主动性：我们评估人工智能提供及时、主动的指导能力。用户根据这些干预措施对当前任务上下文的帮助程度和适用性来进行评估。
- 易用性：用户评估 AI 界面的直观性和用户友好性。该指标有助于评估 AI 助手对不同技术水平用户的可访问性。
- 满意度：这个整体指标捕捉用户对 AI 性能和有用性的一般印象。
- 相关性：我们通过查询响应的适当性来衡量——即 AI 在多大程度上为用户问题提供相关的答案并保持上下文适当的指引。
- 总体评分：结合用户交互体验所有方面的综合指标。

3.3. 成本控制的性能评估

平衡人工智能的性能与计算和财务成本对于可扩展的部署至关重要，特别是在需要人类介入和受限的环境中，如战场医疗和家庭辅助。高性能的人工智能代理，尤其是那些利用大型语言模型（LLMs）的代理，可能会产生显著的推理成本，而这些成本并不总是能带来成比例的收益 [9] 。为了解决这个问题，我们引入了一个成本与性能评估框架，其中包含推理成本，用于量化计算和货币费用，以及成本性能帕累托效率，用于识别资源消耗与引导效果之间的最佳权衡。

3.4. 评估方法

我们的评估过程包括收集任务的视频录制、AI 用户对话日志和任务后用户调查。

评估点是从交互的各个阶段系统地提取的，包括用户输入、AI 响应和任务执行事件。这种方法使我们能够捕捉互动的动态性质，评估 AI 在各种场景下的表现，并识别 AI 指导和用户意识中的潜在盲点。通过将未被注意到的错误作为评估点，我们可以评估 AI 检测和防止错误的能力，以及用户对系统防错的依赖性。

功能指标基于录音和日志的客观测量进行计算，而用户交互指标将客观测量与调查中的用户主观评分结合起来。这种全面的数据收集和分析策略确保了对 AI 助手在任务指导不同方面的表现进行彻底评估。

通过为每个指标提供明确的定义和测量方法，我们确保该框架可以在未来的研究中一致地应用和复制，促进不同 AI 助手实现之间的比较分析。这种标准化的方法将有助于不断改进 AI 辅助任务指导系统，最终提供更有效和用户友好的解决方案。

我们研究了在该问题上应用预训练的大语言和视觉基础模型，而不需要任务特定的训练。我们提出了一种 AI 代理架构和三种不同的配置来提取视觉和对话上下

文：

3.5. 系统设计和开发

我们的任务指导代理（图 2 ）建立在一个模块化的多进程架构之上，旨在实现健壮性、可扩展性和实时性能。代理的核心是指挥过程，它协调整体工作流程。所有主要组件——包括感知、LLM、用户界面、TTS、ASR 和计时器——直接与指挥进行通信，也通过它彼此通信。它接收来自环境和用户的已处理输入信息，根据预定条件进行节点转换。该过程有效地充当状态机，管理任务指导的流程，并确保跨所有系统组件的顺畅任务进展和协调。

在每个新用户会话的开始，指挥器访问任务库，根据用户指明的计划执行的特定任务构建一个任务图。这个任务图由任务库中存储的信息构建而成，作为引导用户完成流程的主干结构。

与控制器协同工作的是数据管理器进程。在我们的架构中，这一组件被表示为传感器进程和共享内存，它不断从传感器和外围设备提取数据，存储在共享内存中，并使其他进程可以方便地访问这些数据。通过集中化的数据处理，我们确保系统的所有组件都能够获得最新的信息，这对于在任务指导中保持一致性至关重要。输出由用户界面（UIs）管理，其中 HL2 输出进程将信息呈现到 Hololens 2 上，为用户提供沉浸式增强现实体验，而 Naive 输出进程则在单独的显示器上显示信息（用于调试和监控）。

视觉理解由主动感知过程处理，该过程基于当前节点的信息持续处理视觉数据。这种自适应方法使系统能够将计算资源集中在视觉输入中最相关的方面，从而提高场景解释的效率和准确性。重要的是，感知服务可以检测用户是否执行了不按顺序的步骤，从而向指导者发出警报，随后提示 AI 代理进入对话模式。在这种模式下，AI 与用户互动以评估情况，并确定是否存在需要解决的问题。我们专门在第 ?? 节对这一过程进行评估。

自然语言处理是我们系统的关键组件，由 LLM Process 管理。这个过程提供了关键的服务，如对自由形式的用户输入进行分类和将标准信息重新措辞为自然语言。通过利用高级语言模型，我们实现了用户与系统之间的直观交互。

音频输入和输出由两个专门的进程处理。ASR 进程不断地从麦克风处理音频，并在必要时广播处理后的文本。这允许进行免提交互，这在许多任务指导场景中至关重要。与此配合的是，TTS 进程根据 Conductor 进程的请求将文本转换为音频，并将其推送到输出进程进行语音化。这使得系统能够提供听觉指导，增强了交互的多模态性。

计时器服务监控任务持续时间，并可以在预定时间阈值超过时中断用户。当被触发时，计时器服务会提示 AI 代理进入对话模式，与用户互动以检查是否有问题或是否需要帮助。此功能确保在时间敏感的任务中或当长时间不活动可能表明问题时的安全性和效率。

日志进程持续访问共享内存，将所有相关数据流记录到数据仓库。这种全面的日志记录有助于系统分析、性能优化和任务指导算法的持续改进。

该架构利用共享内存和 ZeroMQ 进行进程间通信，优化了高吞吐量数据共享和低延迟消息传递。这种混合方法确保了高效的数据传输和系统响应性，对于实时任务指引至关重要。

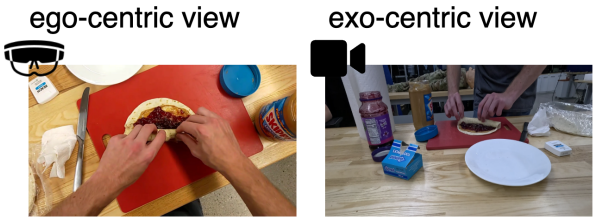

Figure 3. 我们数据集中的示例。左：微软 Hololens 2 视图。右：GoPro Hero 12 视图。

4. 数据集收集

我们为 12 名参与者执行的 4 个任务捕捉了同步的自我中心和外部中心视角的多模态录制。除了步骤时间界限外，我们还提供了任务和步骤级别的错误检测注释，并附有自然语言描述。这些媒体和注释不仅对人类动作理解训练有用，还对人机协作有帮助。

参与者我们招募了 12 名参与者（6 名男性，6 名女性；年龄 19-29 岁），来自一家位于美国的大学。这个群体在种族和文化上具有多样性：5 名参与者自认是中国人（包括 3 名国际学生和 2 名在美国长大的学生），3 名自认为是印度/南亚人，3 名是白人美国人，1 名是中欧混血。学术水平从一年级本科生到博士阶段不等，包括 7 名本科生和 5 名研究生（3 名硕士和 1 名博士），其专业涵盖机器人学、计算机与电气工程、航空航天工程、神经科学和运动学，几位参与者有双专业方向。所有参与者的视力均为正常或矫正至正常。这种文化背景和专业知识的广度为数据收集提供了一个丰富多样的群体。

任务每位参与者完成了四项任务，这些任务从日常烹饪到战场医疗：泡一杯茶、准备卷饼三明治、准备甜点薄饼和应用止血带。对于每项任务，我们评估了三种辅助条件。在未协助的条件下，参与者仅被提供任务名称和简要的目标描述，仅依靠之前的知识。在纸质说明条件下，用户获得了任务的详细、逐步说明。这种方法代表传统的菜谱风格说明（?? ）。在 AI 代理条件下，参与者通过我们的基于 AR 的 AI 系统（?? ）获得交互式、情境感知的指导。

录音实验中共有 144 次会话，每次会话记录一位参与者执行一项带有一种说明类型的任务。对于每个会话，我们使用 GoPro Hero 12 Black 相机录制了参与者动作的第三人称视角视频。对于 AI 说明的会话，我们还使用微软 Hololens 2 记录了参与者的第一人称视角和对话，并收集了所有传感器数据，包括前置摄像头、4 个侧面摄像头、深度传感器、IMU 和音频。会话的平

均时长是 4.73 ± 2.01 分钟，总共得到 15.15 小时的有效实验时长（所有会话的外中心视角录制时长）。我们在同一个房间里录制所有会话。“Pinwheel” 任务的自我中心-外中心视角（前置 RGB 相机）的示例如 Fig. 3 所示。

标注两位具有本科学历工程背景的训练标注者按照预定义的协议对视频录制进行了标注。标注过程使用 VGG 图像标注工具（VIA）[5, 6] 进行，该工具允许高效的视频标注并支持时间分割。我们的标注捕获了五个互补的数据字段，每个字段存储在一个机器可读的 JSON 文件中，以便于后续分析：(1) 任务成功（布尔值），如果失败则提供一个自由格式的 comment 字符串。(2) 任务持续时间记录为两个浮点时间戳（start_sec，end_sec），标记从参与者最初阅读指令到他们明确确认任务完成的时间间隔。(3) 步骤边界是条目，每个条目包含 start_sec 和一个 end_sec 。当存在明确的视觉证据显示某个隐性动作（例如，手向纸巾移动以开始“清洁刀子”）时，一个新步骤开始。(4) 通过将标注步骤的时间顺序与标准食谱顺序进行比较，标记出顺序错误。(5) 记录步骤中细粒度的错误——例如使用错误的工具或错误测量配料——为 { ”step # ”: ”free-form description” } 对象列表。(6) 同步——对于包含自我中心和他我中心录制的会议，我们通过标注时间偏移手动同步这两个流，以确保在不同视角之间的精确时间对齐。

5. 人工智能辅助任务协作的人类研究

我们进行了一项结构化的用户研究，以评估不同的指导方法如何影响实际任务表现、学习结果和用户体验。参与者在使用第 4 节中介绍的相同四种场景和指导条件的情况下完成实际任务。

5.1. 实验设计

每位参与者完成相同的任务三次，每次在三种指导条件中的一种下进行：无辅助（UA）、纸质说明（PI）和人工智能代理（AI）。为了控制顺序效应，我们在六种可能的排列（例如，UA → PI → AI、UA → AI → PI 等）中进行了完全平衡，参与者被随机分配到其中一种顺序。

任务是从用于数据集收集的任务中选择出来的：三个以食谱为风格的程序和一个医疗场景。我们战略性地选择任务以涵盖一系列复杂性水平。食谱任务——泡茶、制作旋转饼和制作甜点玉米饼——常用于评估指导系统，因为它们具有清晰的时间结构和客观的成功标准。为了验证我们系统在高风险领域的适用性，我们包括了止血带应用，这是一项来自战场医疗的显著更复杂且与安全性相关的任务。

参与者。本研究涉及 12 名参与者（不同于标注者），每位参与者在不同的指导条件下完成多个任务变体。

曝光考虑。由于每位参与者在不同的指导方法下多次执行相同的任务，表现不仅可能受到指导条件本身的影响，还可能受到先前任务经验的影响。为了解释这一点，我们的分析考虑了第一次和重复尝试，从而让我们可以研究学习效果以及先前对一种方法的经验如何影响另一种方法下的表现。

5.2. 评估结果

我们报告不同任务指导方法在功能性能、用户交互质量和技能获取方面的表现。我们根据提出的评估框架展示了物理任务完成的结果。结果根据我们提出的评估框架进行了解释，并针对参与者的任务接触进行了分析。

任务完成质量评估。表 1 中展示的任务绩效评估结果清楚地证明了 AI 辅助指导的有效性。

当用户在没有任何预训练（训练 = 无）的情况下第一次尝试任务时，那些受到 AI 系统引导的用户达到了显著更高的宏观成功率（M-SR）70 %，而未受辅助引导（UA）的只有 20 %，使用纸质说明（PI）的则是 28.57 %。步骤错误率（S-ER）也遵循了类似的趋势，更有利于 AI，分别为 16.43 %，而 PI 为 18.37 %，UA 则为 38.75 %。虽然 AI 引导的任务完成时间较长（186.54 秒），但这种权衡带来了更高的成功率和更低的错误率。

我们还研究了最初接触 AI 指导如何影响使用其他方法的后续表现。

Training	Guidance	M-SR ↑	S-ER ↓	Time(s) ↓
None	UA	20.00 %	38.75 %	106.26
	PI	28.57 %	18.37 %	144.29
	AI	70.00 %	16.43 %	186.54
AI	UA	66.67 %	18.45 %	104.77
	PI	75.00 %	6.70 %	132.29
PI	UA	50.00 %	20.09 %	99.80
	AI	80.00 %	5.00 %	186.69
UA	AI	100.00 %	0.00 %	217.49
	PI	60.00 %	25.00 %	97.82

Table 1. 任务绩效评估结果。我们展示宏观成功率（M-SR）、步骤错误率（S-ER）和任务持续时间（Time）。

具体而言，表格 1 展示了在不同指导方法下的技能获得数据。

关键发现是在 “AI” 条件下，参与者在第一次试验中使用了 AI 辅助：

- 随后的无人协助（UA）性能显著提高到 66.67 % 的成功率，错误率降低到 18.45 %。
- 在接触 AI 后，通过纸质说明，表现进一步提高至 75 % 成功且仅有 6.70 % 错误。

仅仅是重复任务可能就足以提升表现，因为参与者会对任务结构和物理动作变得更熟悉。然而，我们观察到，在接触 AI 后的改善明显大于 UA 或 PI 后的改善。AI 系统提供的互动和情境感知支持似乎能促进更有效的学习，使得用户在后续任务中表现更佳，无论使用何种指导方法。这种增强的技能获得强调了 AI 代理不仅

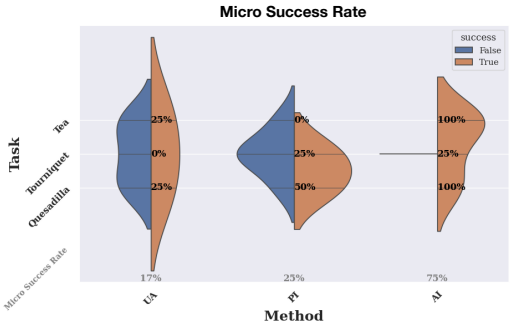

Figure 4. 微任务表现评估结果。对于所有任务，即使是像战场医学中使用止血带这样具有挑战性的专业任务，AI 指导也能帮助用户取得更好的表现。

仅是任务助手，更是物理操作任务的有效训练工具的潜力。

如图 4 所示的微任务性能结果在各个单独任务中强化了这一趋势，包括最复杂的任务（止血带应用），在这些任务中，人工智能始终能在各种任务领域中实现更好的表现。

用户交互质量评价。我们报告了用户对以 AI 助手方法开始完成后续任务的帮助程度的看法。大多数参与者（77.8 %）表示，认为 AI 助手方法在过渡到其他方法时更有帮助。这表明 AI 助手可能为用户提供了强有力的基础理解或方法，从而在后续任务中受益。另一方面，22.2 % 的参与者认为，以 AI 助手方法开始并没有对他们完成其他方法的能力产生影响，表明对某些用户而言，AI 的影响是中性的。值得注意的是，没有参与者认为 AI 助手方法较少有帮助。

这些发现表明，AI 助手方法通常提供了积极的学习体验或提供了用户可以应用到其他任务完成方法中的见解。高比例的用户发现它更有帮助，这表明 AI 助手可能在训练或熟悉任务中是有效的，即使当 AI 没有直接指导他们时，也能提高他们的表现，这一观点在技能获取的定量分析中得到了验证。

Question	Score Logit (5) ↑	Score Percentage ↑
Clarity	3.42	68.33 %
Proactivity	3.17	63.33 %
Ease of use	3.08	61.67 %
Satisfaction	3.00	60.00 %
Relevance	2.67	53.33 %
Overall	3.07	61.33 %

Table 2. 用户交互质量指标。

表 2 中提出的用户交互质量指标提供了关于用户感知的 AI 助手性能的各个方面的见解。这些指标涵盖了清晰度、主动性、易用性、满意度和相关性，每项均在 5 分制上进行评分。结果表明，AI 助手在提供清晰指示方面表现最佳，得分为 3.42 分（68.33 %）。在引导用户完成任务的主动性方面也获得好评，得分为 3.17 分（63.33 %）。易用性和整体满意度的评分略高于平均水平，分别得分约为 3 分（61.67 % 和 60.00 %）。AI 对用户查询的回应的相关性得分最低，为 2.67 分（53.33 %），显示了潜在改进的空间。综合绩效，即所有问题的平均得分，计算为 3.07 分（61.33 %），这表明尽管总体上 AI 助手满足了用户的期望，但仍有改进的空间。

成本。我们的 AI 任务指导系统有效平衡了成本和性能，运行在配备 NVIDIA RTX 5000 GPU 的联想 ThinkPad P16 Gen 2 和微软 HoloLens 2 上，总费用为 $ 9,019。每个任务会话使用 OpenAI 的 ChatGPT [17]，平均推理成本为 $ 0.002，推理成本与成功率比率为 0.000029 $ / %，显示出在 AI 辅助任务完成中的强大成本效益。

为了充分了解 AI 代理在任务指导中的潜力，关键是要单独评估其感知能力。这种隔离使我们能够独立评估它在多大程度上能够解释和响应来自环境的视觉输入，这是实时指导中的关键因素。为此，我们在系统中集成了两种方法，一种采用零样本方法，另一种采用监督方法。

5.2.1. 场景描述 (Sce)

该方法通过结合对象检测器、字幕模块和大型语言模型 (LLM) 生成自由文本场景描述，从而增强情境感知。

如图 5 (a) 所示，我们应用了一个目标检测器，这里我们使用 DINO [11]，来识别最近捕获帧中的感兴趣区域 (ROI)。这些 ROI 然后被输入到一个生成描述的模型中，我们尝试使用 BLIP-2 [12] 和 LaViLa [28]，并结合提示为每个区域生成描述。与 BLIP-2 不同，LaViLa 接收一系列帧作为输入，使其能够更好地捕捉任务完成的时间动态。

来自任一字幕生成器的描述结果及其对应的区域坐标被用作提示输入给大型语言模型（LLM）(GPT-3.5-turbo)，并附上该任务步骤的完整列表。然后，LLM 将单个描述映射到整体场景状态识别，起到任务步骤分类器的作用，同时还提供带有位置信息的全面场景描述。

这种零样本方法使系统能够推广到任务特定数据可能不可用的多样或不熟悉的环境。该方法的分类准确性受限于流水线中不同组件施加的限制。例如，对象检测和描述阶段虽然在生成丰富和详细的场景描述方面很有效，但会引入噪声和变化，从而影响 LLM 准确分类任务步骤的能力。

5.2.2. ResNet 步分类器（Cls）

为了更准确地跟踪任务进度，我们实现了一个基于 ResNet 的步骤分类器。尽管它需要特定任务的训练数据，但它特别适合于计算资源有限的场景，如远程环境，从而在战场医学等专业领域的任务指导应用中具有价值。

正如预期，由于其任务特定的训练和分布内测试环境数据（图 5 (b)），Cls 在各项任务上表现优于 Sce。然而，Sce 准确地检测出显著区域（这里没有进行定量评估），这在我们系统未来的迭代中可能会非常有价值。

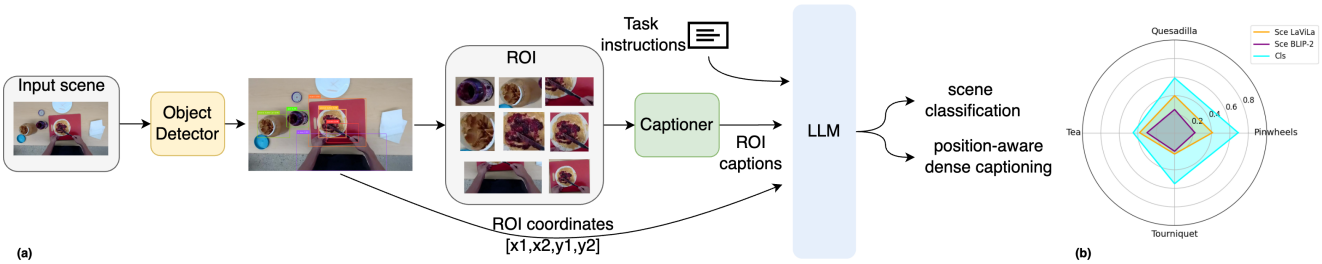

Figure 5. (a)：零样本场景分类和密集描述。(b)：感知方法的准确性报告。

此外，对描述生成器进行微调可能会带来显著的性能提升，这是我们计划在未来工作中探索的内容。

对于人类与 AI 协作完成任务，我们引入了一个综合评估框架并开发了一个配备 AR 功能的 AI 代理用于互动引导。我们的人工研究验证了该框架的有效性，从技术设计和人类学习的角度为多样化任务情境下的 AI 辅助协作提供了宝贵见解。首先，结果表明我们的 AI 代理显著提高了任务成功率并减少了错误。此外，我们的研究提供了关于 AI 辅助环境下人类技能发展的更深入见解，揭示了 AI 引导如何影响学习曲线、任务适应性和用户信心。这些发现不仅强调了 AI 在提升任务表现方面的潜力，还在促进结构化技能获得上扮演重要角色。此外，我们通过分享人类研究中的匿名多模态数据及专家标注的任务评估，来为研究社区做出贡献，从而支持进一步的分析和基准测试。基于我们的研究成果，未来的工作可以探索直观的人类查询界面、先进的感知模型以及主动干预策略，以增强人类与 AI 协作完成任务的适应性和用户体验。

References

[1] Antropic. The claude 3 model family: Opus, sonnet, haiku. 1

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. ArXiv , abs/2309.16609, 2023. 1

[3] Yuwei Bao, Keunwoo Yu, Yichi Zhang, Shane Storks, Itamar Bar-Yossef, Alex de la Iglesia, Megan Su, Xiao Zheng, and Joyce Chai. Can foundation models watch, talk and guide you step by step to make a cake? In Findings of the Association for Computational Linguistics: EMNLP 2023 , pages 12325–12341, Singapore, 2023. Association for Computational Linguistics. 2

[4] Filippos Bellos, Yayuan Li, Wuao Liu, and Jason Corso. Can large language models reason about goal-oriented tasks? In Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024) , pages 24–34, 2024. 1

[5] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In Proceedings of the 27th ACM International Conference on Multimedia , New York, NY, USA, 2019. ACM. 5

[6] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/, 2016. 5

[7] Qiuyuan Huang, Naoki Wake, Bidipta Sarkar,

Zane Durante, Ran Gong, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Noboru Kuno, Ade Famoti, et al. Position paper: Agent ai towards a holistic intelligence. arXiv preprint arXiv:2403.00833 , 2024. 1

[8] Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. Ai agents that matter. arXiv preprint arXiv:2407.01502 , 2024. 1

[9] Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. Ai agents that matter, 2024. 3

[10] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. arXiv preprint arXiv:2406.13743 , 2024. 1, 2

[11] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 13619–13627, 2022. 6

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning . JMLR.org, 2023. 6

[13] Yayuan Li, Zhi Cao, and Jason J Corso. Instructional video generation. arXiv preprint arXiv:2412.04189 , 2024. 1

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Advances in Neural Information Processing Systems , pages 34892–34916. Curran Associates, Inc., 2023. 2

[15] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292 , 2024. 1

[16] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. Gpt-4 technical report, 2024. 1, 2

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems , 35:27730–27744, 2022. 6

[18] Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents, 2024. 1

[19] Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. arXiv preprint arXiv:2310.16049 , 2023. 2

[20] Shane Storks, Itamar Bar-Yossef, Yayuan Li, Zheyuan Zhang, Jason J Corso, and Joyce Chai. Explainable procedural mistake detection. arXiv preprint arXiv:2412.11927 , 2024. 1

[21] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. arXiv preprint arXiv:2210.09261 , 2022. 2

[22] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, et al. Gemini: A family of highly capable multimodal models, 2025. 1, 2

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 , 2023. 1

[24] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574 , 2024. 1, 2

[25] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864 , 2023. 1

[26] Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-

Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. ArXiv , abs/2412.15115, 2024. 1

[27] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In The Twelfth International Conference on Learning Representations , 2024. 2

[28] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In CVPR , 2023. 6

[29] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911 , 2023. 2

[30] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In The Twelfth International Conference on Learning Representations , 2024. 2