EgoCross：跨域以自我为中心的视频问答多模态大型语言模型的基准测试

Yanjun Li[1*] Yuqian Fu[2*] Tianwen Qian[1†] Qi'ao Xu[1] Silong Dai[1]
Danda Pani Paudel[2] Luc Van Gool[2] Xiaoling Wang[1†]

[1]School of Computer Science and Technology, East China Normal University,
[2]INSAIT, Institute for Computer Science, Artificial Intelligence and Technology

Abstract

多模态大型语言模型（MLLMs）的最新进展显著推动了自我中心视频问答（EgocentricQA）的前沿发展。然而，现有的基准和研究主要局限于烹饪和清洁等常见日常活动。相反，现实世界的部署不可避免地会遇到领域转移，在这些领域中，目标领域在视觉风格和语义内容上存在显著差异。为弥补这一差距，我们引入了EgoCross，这是一个综合性基准，旨在评估 MLLMs 在 EgocentricQA 中跨域泛化的能力。EgoCross 涵盖了四个多样而具有挑战性的领域，包括外科手术、工业、极限运动和动物视角，代表了现实且高影响力的应用场景。它由大约 1,000 对问答对组成，涵盖 798 个视频剪辑，涉及四个关键的问答任务：预测、识别、定位和计数。每个问答对提供了 OpenQA 和 CloseQA 格式，以支持细粒度评估。大量的实验表明，无论是通用的还是专为自我中心设计的现有 MLLMs，大多数在泛化到日常生活之外的领域时存在困难，突显出当前模型的局限性。此外，我们进行了几项初步研究，包括微调和强化学习，以探索潜在的改进。我们希望 EgoCross 及我们的附带分析能够成为推进领域适应性强的鲁棒自我中心视频理解的基础。数据和代码将在以下地址发布：https://github.com/MyUniverse0726/EgoCross.

1. 引言

自我中心视频从第一人称视角捕捉人类如何感知和与物理世界互动，为建模人类行为提供了丰富且独特的数据来源。因此，理解自我中心视觉对于具身 AI、可穿戴助手以及人类对机器人的学习等应用非常有价值。在各种自我中心任务中，视频问答 [1, 2]，这种通过类人问题回答对的方式反映模型理解的问题，已成为一个特别具有挑战性但又影响深远的问题。

相对较早的工作，如 EgoVQA [3]、EgoTaskQA [4] 和 EgoSchema [5]，通过引入首批专用的基准测试，为自我中心视频问答（EgocentricQA）奠定基础。多模态大语言模型（MLLMs）的快速进展进一步显著推进了这一领域的基准构建和模型开发。在基准测试方面，EgoThink [6]、EgoTempo [7] 和 EgoTextVQA [8] 被提出来，针对 QA 任务的不同方面。在建模方面，一些专门设计或改编用于自我中心视频理解的 MLLMs 也已出现。值得注意的例子包括 EgoVLPv2 [9] 和 EgoGPT [10]，这些模型通过在专门的自我中心数据上训练来扩展通用 MLLMs 以用于 EgocentricQA。

尽管最近取得了进展，大多数现有的研究仍然专注于日常生活中的常见活动，如烹饪、吃饭和园艺。然而，现实世界的应用不可避免地超出了这些场景。例如，在手术环境中，一个模型不仅必须识别通用的“切割工具”，还需要精确地区分抓钳、电凝钩和双极镊子等仪器。在这种情况下，无论是视觉外观还是语义背景都与日常活动中发现的有显著不同。这样自然引出了一个基本问题：现有的多模态大模型能够有效地泛化到这些不常见且特定领域的场景吗？

为了解答这个问题，我们引入了 EgoCross，这是一个全面的基准，用于评估多模态大模型在自我中心化问答中的跨领域泛化能力。EgoCross 基于三个核心设计原则构建：(1) 强调跨领域特性，(2) 与实际应用相关，(3) 细粒度的多维度模型评估。遵循这些原则，我们精心挑选了视频来源，并开发了相应的问答对，以反映现实世界具有高影响力的用例。具体来说，我们选择了手术、工业、极限运动和动物视角作为基准的四个基本领域。这些领域在视觉和语义上与典型的日常生活场景有显著偏离，因此对模型的泛化能力提出了独特的挑战。基于这些视频来源，我们设计了一个结构化的数据整理流程，以构建涵盖四种基本问答任务类型：识别、定位、预测和计数的问答对，并进一步扩展到总计 15 个具体的子任务。为了支持判别型和生成型评估协议，每个问答实例都以 CloseQA（多选）和 OpenQA（自由形式回答）格式进行了注释。总体而言，EgoCross 由大约 1000 个问答对组成，涵盖 798 个自我中心视频片段，构成了一个精心构建的数据集，能够系统地评估自我中心化问答中的跨领域泛化。图 1 中提供了视觉概述和代表性示例。

实验表明，大多数通用和自我中心特定的多模态大模型在 EgoCross 上表现不佳，其 CloseQA 的准确率

*Equal Contribution.
†Corresponding Author.

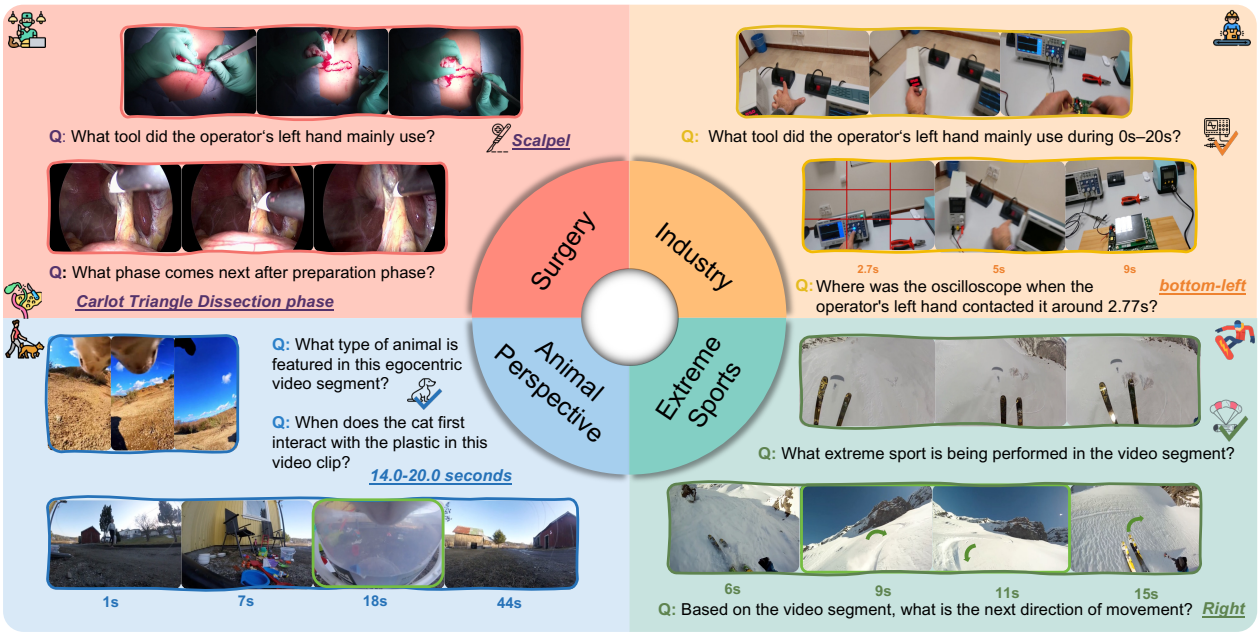


Figure 1. 我们的 EgoCross 基准的例子。我们超越了日常自我中心的场景，涵盖了四个不同的、跨领域的、面向应用的领域：外科手术、工业、极限运动和动物视角。如示例所示，无论是在视觉外观还是语义内容上，都与现有的 EgocentricQA 数据集有显著差异。

低于 55%（随机概率：25%），而 OpenQA 低于 35%，揭示了它们在跨领域设置中的局限性。在相同问题类型从 EgoSchema 到 EgoCross 时出现显著的性能下降（1.6 × ↓），进一步证实了这一挑战。我们还探索了提示学习、微调和强化学习来评估潜在的改进，为未来的研究提供见解。

我们的主要贡献总结如下：

- 我们首次定义并激励跨领域 EgocentricQA 任务，这是一个对现实世界应用至关重要但尚未充分探索的领域。
- 我们发布了 EgoCross，这是第一个面向 EgocentricQA 的跨领域基准，涵盖四个不同的领域（外科手术、工业、极限运动和动物视角），包含 ∼ 1k 高质量的 QA 对。
- 我们对 8 个最先进的 MLLMs 进行全面评估，定量揭示了它们在日常生活领域之外的局限性，并强调了开发更多领域稳健模型的必要性。
- 我们提供了前瞻性试点研究，提出了可操作的见解，并为构建更具泛化性和稳健性的 MLLMs 指明未来方向。

2. 相关工作

自我视角的视频理解因其在模拟人类从第一人称视角进行感知和交互方面的作用而受到了越来越多的关注。除了动作识别 [11]、姿态估计 [12] 和自我-外部对应 [13, 14] 等感知任务外，EgocentricQA [3–8, 15] 已成为一项特别有价值且具有挑战性的任务。广泛使用的 EgocentricQA 基准包括 EgoVQA [3]、EgoTaskQA [4] 和 EgoSchema [5]，最近已引入了几个新的数据集以解决不同的挑战。例如，EgoThink [6] 专注于复杂推理，EgoTempo [7] 专注于时间理解，EgoTextVQA [8] 专注于理解自我视角场景中的文本。数据的增加也推动了专门针对自我视角视频理解的模型的发展，这些模型通常从 MLLMs 改编而来。然而，大多数现有工作仍局限于日常生活场景，对领域转变关注有限。相比之下，我们的工作填补了这一空白，强调了现实世界分布外目标的重要性，并引入了首个全面的跨领域 EgocentricQA 测试平台。

最近在多模态大型语言模型（MLLMs）上的进展展示了其在视频理解中的显著能力。通过广泛的多模态预训练，通用 MLLMs，如 GPT-4.1、Gemini 2.5 Pro、Qwen2.5-VL 和 InternVL，在多个视频任务中都取得了强劲的表现。同时，像 Video-LLaMA3 这样的专用模型通过专门的架构设计进一步提升了时间推理能力。还有一些 MLLMs 专门针对自我中心视频进行了调整，包括 EgoVLPv2 和 EgoGPT。尽管这些模型在第三人称视频和常见日常场景中的自我中心视频中表现良好，它们在面对不熟悉的、领域特定场景时的泛化能力尚未得到广泛检验。在这项工作中，我们系统评估了当前最先进的 MLLMs 在跨领域自我中心目标上的泛化能力，揭示了它们的局限性，并提供了深入的分析，以促进该方向的未来研究。

2.1. 跨领域泛化

跨域泛化是计算机视觉中的一个广泛且长期存在的挑战。之前的工作已经在各种任务上进行过研究，包括图像分类 [16–23]、动作识别 [24–26] 和目标检测 [27–

31]，通过利用领域转移、数据增强和有效的微调技术取得了显著的进展。然而，这些努力主要集中在第三人称视角和低级感知任务。在第一视角视频理解中，由于场景、任务语义和相机运动的显著变化，领域转移尤为显著。一个密切相关的工作 [32] 探讨了第一视角视频中的跨域小样本识别，但其仍然局限于低级感知任务。相反，EgoCross 是第一个专门设计用于评估 EgocentricQA 中跨域泛化的基准，解决了领域差距和高级推理问题。

3. EgoCross 数据集

在本节中，我们为 EgoCross 基准提供了全面的介绍。我们首先讨论领域的选择、视频来源和问答任务的分类，然后解释数据整理流程，并以数据集统计结束。

3.1. 源选择与任务分类

设计原则。 我们建立了领域和数据集选择的关键原则，以及问答任务分类法：(1) 强调跨领域特性。我们需要选择具有独特知识结构、术语和相互作用的领域，这些领域与日常场景有显著不同，以确保模型受到不熟悉概念的挑战。(2) 对实际应用的影响。鼓励与现实世界应用密切相关的数据集，e.g.，如医疗保健和工业操作，因为它们预计会促进 EgocentricQA 在实际应用中的进展。(3) 细粒度的多维模型评估。任务应涵盖广泛的范围，包括多样的考试类型，如复杂推理和时空依赖性，并提供全面的评估指标。

基于上述标准，我们选择了四个专业领域，这些领域具有独特的挑战性并与现实世界高度相关：外科手术、工业、极限运动和动物视角。对于每个领域，我们策划了一到两个高质量的开源数据集，这些数据集附有专家提供的元注释，每个数据集都呈现出独特的感知、认知和推理需求。所选择的领域及其相应的数据集具体如下：

- 手术。外科领域代表了一个高度结构化、知识密集的场景，其中精确性、顺序理解和风险意识至关重要。为了丰富视觉多样性，我们包括了两个数据集：EgoSurgery [33]，该数据集记录了从外科医生视角进行的心脏直视手术的视频，具有详细的手工工具交互和手术阶段注释；以及 CholecTrack20 [34]，该数据集提供了以工具为中心视角的胆囊切除手术的腹腔镜视频。
- 工业。工业场景中的复杂工作流程不仅要求感知精细的物体操作，还需要对步骤顺序和工具使用逻辑进行推理。我们选择了 ENIGMA-51 [35]，这是一个包含实际电路板维修任务的数据集。
- 极限运动。极限运动具有独特的挑战，例如稀有的环境、快速的相机运动和模糊，这些因素可以很好地测试模型的时空感知能力和高速情境推理能力。我们包括了 ExtremeSportFPV [36]，其中包含了各种极限运动的第一人称视频，包括山地自行车、滑雪和跳伞等。
- 动物视角。为了挑战现有模型中的人类中心偏见，我们引入了动物视角，向模型中引入了新的运动模式、摄像机角度和语义重点。因此，包含了如狗、猫、鹰和乌龟等动物的第一视角的数据库 EgoPet [37]。

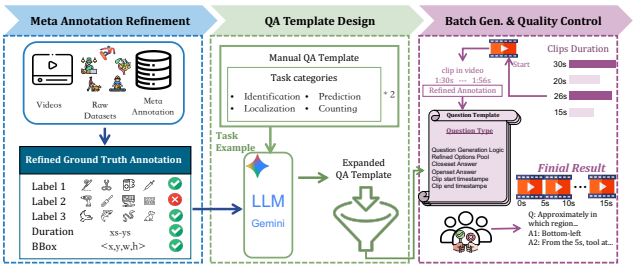


Figure 2. EgoCross 的数据构建管道。

遵循原则 (3)，我们旨在构建多样化的问答对，以全面评估模型能力。如图 3 所示，我们的评估框架围绕四个核心任务类别构建：识别、定位、预测和计数。为应对每个领域的独特挑战，我们进一步将这四个大类别分解为 15 个具体的子任务，共同形成一个全面的评估框架。

- 识别。识别任务评估模型在视频中识别物体、动作和事件的能力。这些任务需要特定领域的知识，并适应对象属性或动作在不同上下文中的细微差异。
- 定位。定位任务评估模型识别物体、动作或交互的精确空间或时间位置的能力。
- 预测。预测任务旨在测试模型基于当前内容预测未来动作或结果的能力。
- 计数。提出计数任务是为了评估模型在时间推移中跟踪和计数不同实例或事件的能力。

3.2. 数据整理管道

基于选择的数据源和问题类别，我们开发了一个多阶段策划管道（图 2），包括三个关键阶段：元注释改进、问答模板设计，以及带有质量控制的批量生成。

虽然所选数据集提供了原始注释，但这些注释通常是为更简单的特定任务目标定制的，例如工具交互的 2D 空间边界框或动作分类的时间段。为了解决这个问题，我们进行了全面的精细化处理，包括统一注释格式和进行人工审核以确保标签的准确性。

根据任务分类学，我们手动设计了 8 个初始 QA 模板，为四个核心任务类别中的每一个创建两个模板。为了增强语言的多样性和复杂性，我们采用了大型语言模型（Gemini 2.5 pro）通过生成特定领域的子任务来扩展初始模板，使用原始模板和精细的注释。所有由 LLM 生成的问题随后均由人工注释员严格验证，以确保其依据视频内容的清晰性、逻辑一致性和可回答性。

批量生成和最终质量控制。 获取问题模板后，我们执行批量实例化以生成最终的问答对。对于每个采样的模板，我们首先根据预定的时间段随机提取相应的视频剪辑，然后通过对剪裁的剪辑执行相关的程序化推理来推导出真实答案。为了进行全面评估，我们采用了传统的固定格式选择题形式（CloseQA）和更灵活的开

放式形式（OpenQA）来回答。

3.3. 数据集统计

我们的 EgoCross 基准覆盖了四个不同的领域：手术、工业、极限运动（XSports）和动物视角（Animal Per.），其来源于五个真实世界的自我中心视频数据集。它包含 798 个视频片段和 957 个问答对，涵盖了 15 种子任务类型，这些类型归纳为四个主要类别。表 1 总结了这五个数据集的关键统计数据，包括片段数量、问答对数量以及视频时长的平均秒数（Dur.(s)）。图 3 进一步展示了 EgoCross 的组成。

Domain	Source	Clips	QA Pairs	
Surgery	CholecTrack20	112	183	29.7
	EgoSurgery	100	100	20.4
Industry	ENIGMA-51	176	245	16.5
XSports	ExtremeSportFPV	242	246	13.7
Animal Per.	EgoPet	168	183	31.5
EgoCross	5 datasets	798	957	22.5

Table 1. EgoCross 基准的关键统计数据。

4. 实验

4.1. 实验装置

在本节中，我们首先在第 4.1 节概述实验设置。然后在第 4.2 节中展示我们 EgoCross 基准测试的主要结果。在第 4.3 节中，我们与领域内的基准进行比较分析。最后，第 4.4 节探讨各种方法，以评估它们在弥合泛化差距方面的潜力和局限性。

为了涵盖主要技术范式，我们选择了一组多样的 MLLMs，分布在三个类别：(1) 为了评估当前最先进的性能，我们包括了领先的专有模型：GPT-4.1 [38] 和 Gemini 2.5 Pro [39]。(2) 对于开源的通用 MLLMs，我们考虑了 Qwen2.5-VL (3B, 7B) [40]、VideoLLaMA3 [41]、和 InternVL3 [42]。(3) 为了评估专为自我中心理解定制的模型，我们还包括了两个自我中心专用模型：EgoVLPv2 [9] 和 EgoGPT [10]。

评价指标。 继之前的工作 [3, 5, 7]，我们使用标准的准确性指标来评估 CloseQA，该指标计算为正确回答问题的百分比。对于 OpenQA，我们采用两阶段的评估过程：(1) 生成答案与真实答案的直接精确匹配，(2) 如果没有找到匹配，我们采用“大型语言模型作为裁判”的方法来评估语义正确性。

所有 MLLM 在零样本设置下进行单轮推理测试。对于视频输入，我们以固定速率 0.5 fps 提取帧。对于提供预采样帧的数据集，我们遵循其原始采样频率。没有设置最大帧数限制，以允许模型处理完整的时间背景。所有实验均在 NVIDIA A6000 GPU 上进行。

4.2. 在 EgoCross 上的结果

评估结果汇总在表 2 中。我们从四个角度分析结果：(1) 任务级挑战，(2) 跨域差异，(3) 模型性能，以及 (4) 指标类型分析。

任务层面的挑战。大多数被评估的多模态大模型在我们的 EgoCross 基准测试中表现不佳，CloseQA 的平均得分低于 55 %，而 OpenQA 的平均得分低于 35 %。考虑到 CloseQA 的随机猜测准确率是 25 %，这些结果表明模型在这个基准测试中确实面临着巨大的挑战。

域间差异。在不同的目标域中，我们观察到不同难度水平的变动，从相对容易的（动物视角）、中等难度的（外科手术）到特别具有挑战性的（极限运动、工业）。为了进一步研究域间差异，我们通过使用 CLIP [43] 作为视觉和文本表示的模态对齐特征提取器，将 EgoSchema 和四个域外目标的 t-SNE 嵌入可视化（图 4）。

模型性能表现。两种专有的多模态大模型（MLLM）达到最高的整体性能，其中 Gemini 2.5 Pro 表现优于 GPT-4.1。接下来是开源模型，包括 Qwen2.5-VL、VideoLLaMA3 和 InternVL3。令人惊讶的是，尽管 EgoVLPv2 和 EgoGPT 是专门为自我中心视频数据设计和训练的模型，但它们的表现最差。

指标类型分析。我们进一步分析了在不同评估指标下的结果，即 CloseQA 和 OpenQA。由于 CloseQA 通过提供明确的候选答案简化了任务，模型在 CloseQA 中自然比在 OpenQA 中取得更高的准确率。此外，我们观察到 CloseQA 的得分在不同的多模态语言模型（MLLMs）中趋于更加稳定，而 OpenQA 对变化更为敏感。例如，GPT-4.1 和 Gemini 2.5 Pro 在 CloseQA 上取得了几乎相同的得分（52.63 对 52.95），但在 OpenQA 上的差异显著（26.65 对 34.40）。

4.3. 跨域差距的更多分析

在第 4.2 节中，我们证明了领域差异显著影响整体的低性能。为了进一步研究这一影响并突出其在 EgoCross 中的独特表现，我们比较了模型在我们的基准和 EgoSchema [5] 上的结果，EgoSchema 是一个典型的日常生活自我中心数据集，包含烹饪、清洁等常见活动。

图 5 的结果显示，所有可比较的 QA 类型的性能均呈现出一致且显著的下降。例如，动作时间定位的性能从域内 EgoSchema 中令人印象深刻的 92.31 % 下降到 EgoCross 中手术、工业和极限运动的新域的仅 34.13 %。同样，下一步动作预测的准确性从 85.71 % 下降到 37.50 %。总体准确率也从 73.58 % 下降到 43.14 %，量化了域迁移带来的重大影响。

4.4. 初步研究

我们积极进行多项试点研究，以探索改善跨领域自我中心 QA 的潜在解决方案。具体来说，我们研究了三种技术：提示学习、监督微调（SFT）和强化学习（RL）。由于 SFT 和 RL 都需要标记的数据，我们随机将初始测试 QA 对分成训练和测试集，比例为 70 % :30 %。我

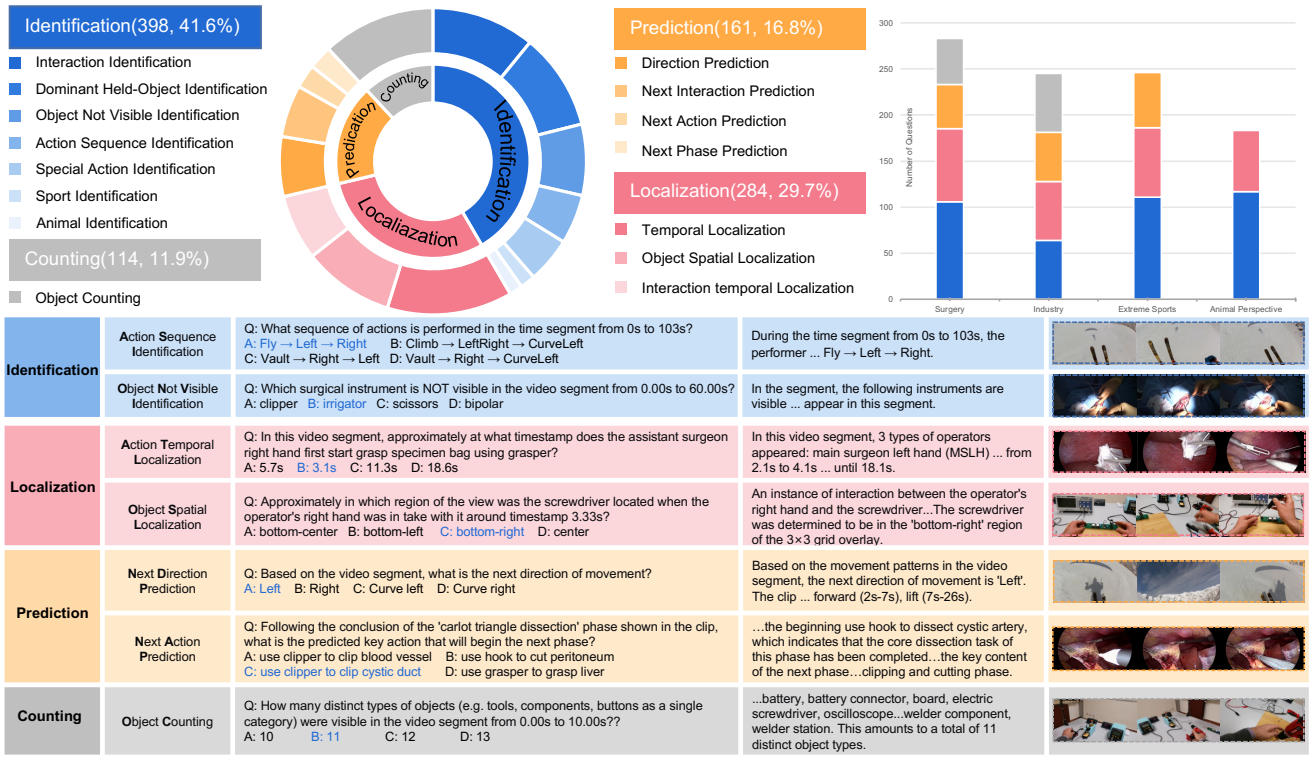


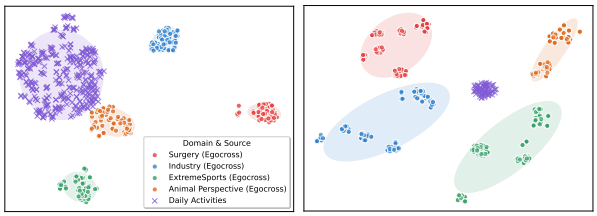
Figure 3. EgoCross 任务分类和统计概览。（左上）四个主要任务类别的总体分布。（右上）跨目标领域的问题数量。（底部）针对每个主要能力的代表性问答示例选择。

Models	Surgery		Industry		XSports		Animal Per.		Overall	
	Closed	Open	Closed	Open	Closed	Open	Closed	Open	Closed	Open
Proprietary MLLMs										
GPT-4.1	57.24	39.58	45.71	12.24	43.09	20.33	64.48	34.43	52.63	26.65
Gemini 2.5 Pro	61.48	42.40	37.55	24.49	43.90	21.54	68.85	49.18	52.95	34.40
Open-source MLLMs										
Qwen2.5-VL-3B	35.69	16.96	36.33	6.94	36.59	6.91	41.53	28.42	37.54	14.81
Qwen2.5-VL-7B	46.29	21.55	37.55	22.04	41.87	6.91	53.55	31.15	44.82	20.41
VideoLLaMA3-7B	39.22	15.90	40.82	13.47	37.80	13.41	50.27	32.24	42.03	18.76
InternVL3-8B	47.00	17.67	33.06	11.84	41.06	11.38	49.18	30.60	42.58	17.87
Egocentric MLLMs										
EgoVLPv2	26.50	-	34.69	-	23.17	-	24.04	-	27.10	-
EgoGPT	31.80	13.07	24.49	10.20	24.80	13.82	41.53	26.78	30.66	15.97

Table 2. MLLMs 在 EgoCross 上的评估结果。所有分数均以百分比报告。最佳结果用粗体标出，第二最佳结果用 underlined 表示。由于模型架构的原因，EgoVLPv2 未在开放集任务上进行评估。

们采用 Qwen2.5-VL-7B 作为基准，并应用 vLLM [44] 进行模型加速。CloseQA 的结果如表 3 所示。

这些结果提供了一些见解：（1）整体趋势。每种方法，无论是提示（不需要标注数据）还是 SFT/RL（需要标注数据），都能在一定程度上提高性能。（2）SFT 的影响。在工业等领域，SFT 显著提高了准确率（几乎提高了 20 % ）。（3）RL 的有效性。在所有领域，RL 显示了最显著的改善（平均提高了 22 % ）。



(a) 视觉特征 t-SNE。 (b) 文本特征 t-SNE。

Figure 4. 文本和视觉特征的 t-SNE 可视化。

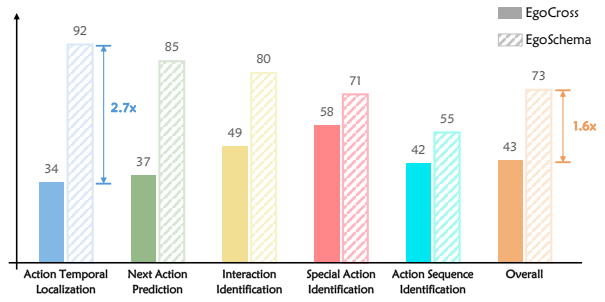


Figure 5. 域内和跨域准确率比较。结果突出了由领域偏移引起的性能差距。

Method	Surgery	Industry	XSports	Animal Per.	Avg.
基线 *	46.29	37.55	41.87	53.55	44.82
Baseline	37.35	35.71	34.72	43.40	37.80
+Prompt	44.58	34.29	52.78	43.40	43.76
+SFT	37.35	52.86	40.28	43.40	43.47
+RL	49.40	61.43	54.17	75.47	60.12

Table 3. CloseQA 在初步研究中的准确性。“+SFT” 和 “+RL” 分别表示监督微调和强化学习。* 表示没有 vLLM 加速的基线。

5. 结论

在这项工作中，我们提出了 EgoCross，这是一个新的基准，用于评估多模态大语言模型（MLLMs）在自我中心视频问答中的跨领域泛化能力。EgoCross 由大约 1k 个问答对组成，这些问答对基于从四个不同且现实的领域精心收集和整理的视频片段：手术、工业、极限运动和动物视角。它涵盖了广泛的问题类型，并支持 CloseQA 和 OpenQA，从而实现细粒度的评估。除了构建基准之外，我们还广泛评估了当前最先进的 MLLMs，发现尽管在现有基准上表现强劲，它们在扩展到不熟悉的、特定领域的场景时仍面临困难。此外，我们进一步探索了几种可能的技术来提高跨领域泛化能力。我们相信，EgoCross 以及我们的实验和分析，为未来关于鲁棒性、领域自适应且实用的自我中心视频理解的研究提供了一个有价值的基础。

References

[1] Y. Zhong, J. Xiao, W. Ji, Y. Li, W. Deng, and T.-S. Chua. Video question answering: Datasets, algorithms and challenges. arXiv preprint arXiv:2203.01225, 2022.

[2] T. Qian, R. Cui, J. Chen, P. Peng, X. Guo, and Y.-G. Jiang. Locate before answering: Answer guided question localization for video question answering. IEEE transactions on multimedia, 2023.

[3] C. Fan. Egovqa: An egocentric video question answering benchmark dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[4] B. Jia, T. Lei, S.-C. Zhu, and S. Huang. Egotaskqa: Understanding human tasks in egocentric videos. Advances in Neural Information Processing Systems, 2022.

[5] K. Mangalam, R. Akshulakov, and J. Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems, 2023.

[6] S. Cheng, Z. Guo, J. Wu, K. Fang, P. Li, H. Liu, and Y. Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In CVPR, 2024.

[7] C. Plizzari, A. Tonioni, Y. Xian, A. Kulshrestha, and F. Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In CVPR, 2025.

[8] S. Zhou, J. Xiao, Q. Li, Y. Li, X. Yang, D. Guo, M. Wang, T.-S. Chua, and A. Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In CVPR, 2025.

[9] S. Pramanick, Y. Song, S. Nag, K. Q. Lin, H. Shah, M. Z. Shou, R. Chellappa, and P. Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[10] J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang, et al. Egolife: Towards egocentric life assistant. In CVPR, 2025.

[11] X. Wang, L. Zhu, H. Wang, and Y. Yang. Interactive prototype learning for egocentric action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[12] J. Wang, D. Luvizon, W. Xu, L. Liu, K. Sarkar, and C. Theobalt. Scene-aware egocentric 3d human pose estimation. In CVPR, 2023.

[13] Y. Fu, R. Wang, Y. Fu, D. P. Paudel, X. Huang, and L. Van Gool. Objectrelator: Enabling cross-view object relation understanding in ego-centric and exo-centric videos. ICCV, 2025.

[14] Y. Fu, R. Wang, Y. Fu, D. P. Paudel, and L. Van Gool. Cross-view multi-modal segmentation@ ego-exo4d challenges 2025. arXiv preprint arXiv:2506.05856, 2025.

[15] K. Li, Q. Xu, T. Qian, Y. Fu, Y. Jiao, and X. Wang. Clivis: Unleashing cognitive map through linguistic-visual synergy for embodied visual reasoning. arXiv preprint arXiv:2506.17629, 2025.

[16] Y. Zhu, F. Zhuang, and D. Wang. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In Proceedings of the AAAI conference on artificial intelligence, 2019.

[17] Y. Fu, Y. Fu, and Y.-G. Jiang. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In Proceedings of the 29th ACM international conference on multimedia, 2021.

[18] J. Zhang, J. Song, L. Gao, and H. Shen. Free-lunch for cross-domain few-shot learning: Style-aware episodic training with robust contrastive learning. In ACM Multimedia, 2022.

[19] Y. Fu, Y. Xie, Y. Fu, J. Chen, and Y.-G. Jiang. Wave-san: Wavelet based style augmentation network for cross-domain few-shot learning. arXiv preprint arXiv:2203.07656, 2022.

[20] L. Zhuo, Y. Fu, J. Chen, Y. Cao, and Y.-G. Jiang. Tgdm: Target guided dynamic mixup for cross-domain few-shot learning. In Proceedings of the 30th ACM International Conference on Multimedia, 2022.

[21] Y. Fu, Y. Xie, Y. Fu, and Y.-G. Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In CVPR, 2023.

[22] J. Zhang, J. Song, L. Gao, N. Sebe, and H. T. Shen. Reliable few-shot learning under dual noises. IEEE TPAMI, 2025.

[23] L. Zhuo, Z. Wang, Y. Fu, and T. Qian. Prompt as free lunch: Enhancing diversity in source-free cross-domain few-shot learning through semantic-guided prompting. arXiv preprint arXiv:2412.00767, 2024.

[24] B. Pan, Z. Cao, E. Adeli, and J. C. Niebles. Adversarial cross-domain action recognition with co-attention. In Proceedings of the AAAI conference on artificial intelligence, 2020.

[25] Y. Xu, H. Cao, K. Mao, Z. Chen, L. Xie, and J. Yang. Aligning correlation information for domain adaptation in action recognition. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[26] K.-Y. Lin, J. Zhou, and W.-S. Zheng. Human-centric transformer for domain adaptive action recognition. IEEE TPAMI, 2024.

[27] Y. Fu, Y. Wang, Y. Pan, L. Huai, X. Qiu, Z. Shangguan, T. Liu, Y. Fu, L. Van Gool, and X. Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In European Conference on Computer Vision, 2024.

[28] Y. Zheng, D. Huang, S. Liu, and Y. Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In CVPR, 2020.

[29] Y. Li, X. Qiu, Y. Fu, J. Chen, T. Qian, X. Zheng, D. P. Paudel, Y. Fu, X. Huang, L. Van Gool, et al. Domain-rag: Retrieval-guided compositional image generation for cross-domain few-shot object detection. arXiv preprint arXiv:2506.05872, 2025.

[30] J. Zhang, L. Gao, B. Hao, H. Huang, J. Song, and H. Shen. From global to local: Multi-scale out-of-distribution detection. IEEE Transactions on Image Processing, 2023.

[31] Y. Fu, X. Qiu, B. Ren, Y. Fu, R. Timofte, N. Sebe, M.-H. Yang, L. Van Gool, K. Zhang, Q. Nong, et al. Ntire 2025 challenge on cross-domain few-shot object detection: Methods and results. In CVPR Workshop, 2025.

[32] M. Hatano, R. Hachiuma, R. Fujii, and H. Saito. Multimodal cross-domain few-shot learning for egocentric action recognition. In European Conference on Computer Vision, 2024.

[33] R. Fujii, M. Hatano, H. Saito, and H. Kajita. Egosurgery-phase: a dataset of surgical phase recognition from egocentric open surgery videos. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2024.

[34] C. I. Nwoye, K. Elgohary, A. Srinivas, F. Zaid, J. L. Lavanchy, and N. Padoy. Cholectrack20: A multi-perspective tracking dataset for surgical tools. In CVPR (CVPR), 2025.

[35] F. Ragusa, R. Leonardi, M. Mazzamuto, C. Bonanno, R. Scavo, A. Furnari, and G. M. Farinella. Enigma-51: Towards a fine-grained understanding of human behavior in industrial scenarios. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024.

[36] S. Singh, C. Arora, and C. Jawahar. Trajectory aligned features for first person action recognition. Pattern Recognition, 2017.

[37] A. Bar, A. Bakhtiar, D. Tran, A. Loquercio, J. Rajasegaran, Y. LeCun, A. Globerson, and T. Darrell. Egopet: Egomotion and interaction data from an animal's perspective. In European Conference on Computer Vision, 2024.

[38] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[39] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261, 2025.

[40] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[41] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106, 2025.

[42] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, et al. Internvl3: Exploring advanced training and test-time

recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025.

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In ICML, 2021.

[44] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.

EgoCross：跨域以自我为中心的视频问答多模态大型语言模型的基准测试

Supplementary Material

6. 与现有基准的比较

如表 4 所示，EgoCross 通过独特地整合跨域挑战、时间任务以及双开放/封闭问答格式，将自己与之前的基准测试区分开来，以便对自我中心视频理解中的模型泛化进行更严格和全面的评估。

7. 更多实现细节

本节提供了我们实验设置的更详细描述，以支持可重复性并实现公平比较。

本节补充了第 4.1 节的设置，以确保在第 4.2 节中报告的结果能够完全重现，并在所有模型之间进行公平比较。

我们的评估方法基于零样本、单轮推理范式，在 NVIDIA A6000 GPU 上执行。如前所述，视频帧以数据集特定的速率（0.5-1 fps）采样，没有最大帧数限制。为了保证确定性和可重现的输出，我们将关键推理参数 do_sample=False 设置为所有实验中的贪婪解码。为确保公平和标准化的评估，下面详细介绍了每个任务使用的确切提示。

封闭式问答（CloseQA）。 对于选择题，图 6 中所示的提示指示模型以结构化的 JSON 格式返回其答案和推理。随后将“预测”后面的字母提取为最终答案。

CloseQA Inferenece Prompt

Please carefully read the question and its options, then select the most appropriate answer. Question: {question}{options_str}
The original FPS of the video is {original_video_fps}. This image set is obtained by sampling at {sampling_fps} fps.
Respond in JSON format with two fields: 'prediction' (the correct option letter: A, B, C, or D) and 'reason' (a brief explanation of your choice). Do not include any other content.

Example response:
{
"prediction": "B",
"reason": "Paris is the capital city of France."
}

Figure 6. CloseQA 推理的提示

对于自由形式的问题，图 7 中显示的提示指导模型生成文本答案。“prediction”之后的字符串用作模型响应以进行评估。

鉴于开放式答案的主观性，我们采用 Qwen-Max 作为自动评判，以确保评估的一致性和可扩展性。LLM 评判根据提示评估模型预测与真实答案在语义上的正确性，如图 8 所示

7.1. 领域差距实验实施细节

如同在第 4.3 节中提到的，为了能够直接比较 EgoSchema（域内）和 EgoCross（跨域）上的性能，需

OpenQA Inferenece Prompt

Please carefully read the question and the provided context, then provide a clear and concise answer based on your understanding. Question: {question}
The original FPS of the video is {original_video_fps}. This image set is obtained by sampling at {sampling_fps} fps.
Your answer should be reasoned and directly address the question.
Respond in JSON format with two fields: 'prediction' (your answer as text) and 'reason' (a brief explanation of your reasoning). Do not include any other content.

Example response:
{
"prediction": "The time duration of the video is 10 seconds.",
"reason": "The total number of frames is 240 and the frame rate is 24 FPS, so 240 / 24 = 10 seconds."
}

Figure 7. OpenQA 推理提示

OpenQA LLM-as-Judge Prompt

You are an intelligent chatbot designed for evaluating the correctness of AI assistant predictions for question-answer pairs. Your task is to compare the predicted answer with the ground-truth and determine if the predicted answer is correct or not. Here's how you can accomplish the task:
##INSTRUCTIONS:
- Focus on the factual accuracy and semantic equivalence of the predicted answer with the ground-truth.
- Consider uncertain predictions, such as 'it is impossible to answer the question from the video', as incorrect, unless the ground truth answer also says that.
Please evaluate the following video-based question-answer pair:
Question: {question}
Correct ground truth: {answer}
Detailed ground truth: {detailed_answer}
Predicted Answer: {pred}
Provide your evaluation as a correct/incorrect prediction along with the score where the score is an integer value between 0 (fully wrong) and 5 (fully correct). The middle score provides the percentage of correctness. Please generate the response in the form of a json object with the following fields:
{
"mark": 4, # The score of the prediction
"reason": "", # The reason for the score
}

Figure 8. 大语言模型作为裁判的提示

要将 EgoSchema 中的问答对与我们预定义的任务类别对齐。由于 EgoSchema 不提供官方的任务类型标签，我们开发了一个半自动的、迭代的分类过程。本节详细描述了该方法、分类规则以及对齐问答对的最终分布。

7.1.1. 半自动分类的方法

为了使 EgoSchema 中的问题与我们预定义的任务类别对齐，我们采用了半自动化工作流程。我们提示大型语言模型（LLM）为 EgoCross 中的每种任务类型模板生成关键词。这些关键词被提取出来以捕捉每种任务类型的核心概念和特征，例如识别、定位、预测和计数。然后我们使用这些关键词集通过将 EgoSchema 中的问题与相关关键词匹配来对问题进行分类。对于问题数量有限的类别，我们避免进一步细分以防止统计不稳定。

关键是，这一初步分类通过严格的人在回路过程进行了细化。人工专家审查了自动化结果，分析了被错误

Dataset	Cross Domain	Video Length	# Test	# Categories	Temporal	QA Types
EgoVQA	✗	(25s, 100s)	250	3	✗	OpenQA
EgoTaskQA	✗	25s	8k	4	✗	OpenQA
EgoSchema	✗	3 min	500	-	✗	CloseQA
EgoThink	✗	-	750	12	✗	OpenQA
EgoTempo	✗	45s	500	10	✓	OpenQA
EgoCross	✓	22.5s	957	15	✓	CloseQA & OpenQA

Table 4. 各数据集特征概述，包括平均视频长度、测试样本数量、类别数量、视频拍摄场景数量及问题类型（OpenQA 或 CloseQA）。

分类实例的问题和选项，并迭代地更新了关键词列表。该细化循环重复进行了五轮，直到分类稳定并达到高准确度，最终由人工验证确认。这一方法论产生了一套经过细化和一致的问题，完全与我们的任务定义保持一致。

7.1.2. 分类规则和关键词

这种细化过程产生了一组关键词和一个用于解决歧义的层次结构。

分类层次 针对重叠案例建立了严格的优先顺序：
1) 主要类别：预测 → 计数 → 定位 → 识别。
2) 识别子类型：主要持有物 → 动作序列 → 交互 → 特殊动作。

最终关键词集合。 最终的关键词集和问题分布详见表格 6 。为了处理 EgoSchema 独有的问题，我们指定了 Inference Prediction（在 EgoCross 中不存在）并创建了 Action State Identification 作为通用动作识别查询的收集类别。剩余六个计数不为零的子任务构成了我们用来直接跨数据集性能对比的核心任务类型集。

定量跨领域比较。 两个领域之间显著的性能差距如图 5 所示，在表格 5 中有详细的量化说明。为了测量这一差距，我们对 Qwen2.5-VL 在完整数据集和我们精心匹配的子集上进行了评估。在匹配的问答子集上，该模型的准确率从领域内 EgoSchema 任务的 73.58% 暴跌至跨领域 EgoCross 对应任务的仅 43.14%。值得注意的是，这一子集上的性能趋势与模型在完整数据集上的总体结果（EgoSchema 为 69.60%，EgoCross 为 44.31%）非常相似，这验证了我们的匹配子集作为评估领域差距的可靠指标。表格进一步分解了各类类似任务中的这种降级，突显出领域转移对模型能力构成了显著且持续的挑战。

7.2. 实施试点研究的细节

为了确保实验的一致性和效率，我们采用 vLLM 框架 [44] 对 Qwen2.5-VL-7B 模型进行训练和推理。虽

	In-Domain		Cross-Domain	
Question Type	Num.	Acc. (%)	Num.	Acc. (%)
Special Action Id.	56	71.43	46	58.70
Interaction Id.	50	80.00	104	49.04
Temporal Loc.	26	92.31	126	34.13
Action Sequence Id.	47	55.32	50	42.00
Next Action Pred.	14	85.71	24	37.50
Overall Above	193	73.58	350	43.14
Overall Dataset	500	69.60	957	44.31

Table 5. 对 Qwen2.5-VL 在域内 EgoSchema 基准测试和我们的跨域 EgoCross 基准测试相似任务类型的表现进行直接比较。这突出显示了即使在结构相似的任务上，转移到新领域时性能的显著下降。

然 vLLM 显著加速了这些过程，但我们注意到，如表 3 显示的那样，存在轻微的性能取舍，这在 RL 训练所需的大量采样中是可以接受的折衷。所有实验都在配备 8 个 NVIDIA H100 GPU 的服务器上进行。SFT 和 RL 的训练设置受到 Video-R1 项目的启发，我们计划发布代码以便于复现。在训练期间的视频输入方面，我们在 4 到 16 帧之间进行采样；超过 16 帧的序列会被截断至 16 帧。对初步研究的推理设置保持与主要结果相一致。

提示学习。 这个方法的提示由两部分组成。第一部分提供领域特定的背景和例子，而第二部分则直接提出关于输入视频的问题。这种结构，如图 9 所示，旨在引导模型先了解每个领域的具体特征，然后才尝试回答问题。

我们对基础模型执行全参数 SFT，更新其所有权重。训练使用了 4 个 H100 GPU，每个设备的批处理大小为 1，进行 2 次梯度累积，因此有效批处理大小为 8。我们将学习率设置为 1e-6，并训练 12 个轮次。为了管理内存和加速训练，我们利用了一系列优化技术，包括 DeepSpeed ZeRO-2、BF16 混合精度、梯度检查点和 Flash Attention 2。SFT 使用的数据格式如图 10 所

Category	Sub-tasks	Count	Keywords
Prediction	Next Action Prediction	14	what will happen, future action, next phase, next direction, prepare for, ready for
	Inference Prediction	26	taking into account, analyze, evaluate, compare, discuss, deduce, overall focus
Counting	Object Counting	0	how many, number of, count how, quantity, total number
Localization	Action Temporal Localization	26	at what time, when did start/end/occur, key moments when, before, after, during
	Object Spatial Localization	0	where is/are, in which region/location, where located
Identification	Dominant Held-Object Identification	10	primary/main tool, tool used, effectiveness tools, how tools contribute
	Action Sequence Identification	47	sequence of actions, from start to finish, key steps, main stages, overarching process
	Interaction Identification	50	interaction between, two characters, both characters, collaborate, relationship
	Special Action Identification	56	most significant/important/critical, key turning points, pivotal, vital steps
	Sport/Animal/Not Visible Identification	0	sport, game play, animal, pet, not visible/shown/present
	Action State Identification	239	primary objective/goal, describe, summarize, explain, infer, deduce, what is/was primary

Table 6. 自我模式任务分类的最终关键词集和问题分布。

示。

我们的强化学习方法基于生成性奖励优化策略 (GRPO)，从零开始训练模型，没有进行 SFT 预训练。训练过程分布在全部 8 个 H100 GPU 上。我们使用 1e-6 的学习率，采用余弦调度器，并训练 16 个周期。通过 DeepSpeed ZeRO-3 来加强优化策略，以适应强化学习过程，同时使用 BF16、梯度检查点和 Flash Attention 2。关键的强化学习特定超参数包括奖励塑性参数 beta 为 0.04，以及在训练期间为策略更新每个提示生成 8 个响应。用于强化学习的提示结构如图 11 所示。

8. 更多实验结果

本节提供了 CloseQA 和 OpenQA 评估的详细结果，并按不同领域的任务能力进行分类。结果展示在表格 ?? 和表格 ?? 中。

为了剖析这种性能下降的来源，表 ?? 和表 ?? 分别提供了对 CloseQA 和 OpenQA 功能的细化分析。这个详细的分析揭示了一般化差距在不同技能之间并不一致。在这两种评估设置中，专有的 SOTA 模型在需要复杂推理的任务中表现出明显的优势。这在 OpenQA 预测 (P) 功能（表 ?? ）中最为明显，其中 Gemini 2.5 Pro 在外科预测任务中取得了显著的 62.50 %，大多数其他模型在此任务中的得分都只有个位数。这突出了

Domain Prompt Learning

```
"{Question}\n"
"Please think about this question as if you were a human pondering deeply,"
"then give your final answer between the <answer> </answer> tags."
{DOMAIN_CONTEXTS}
Please provide only the single option letter (e.g., A, B, C, D, etc.) within the
<answer> </answer> tags.
```

Domain Prompt

```
DOMAIN_CONTEXTS = {
    'Surgery': {
        'context': "Surgical procedure analysis from first-person perspective.
        Key focus: Precise tool identification, surgical phases, hand-specific
operations." },
    'Industry': {
        'context': "Technical repair/assembly tasks from first-person perspective.
        Key focus: Component identification, systematic workflows,
troubleshooting."},
    'Extreme Sports': {
        'context': "High-speed extreme sports from first-person perspective.
        Key focus: Motion patterns, terrain navigation, handling motion blur."
    },
    'Animal Perspective': {
        'context': "Animal perspective videos with alternative movement patterns.
        Key focus: Species identification, behavioral patterns, scale differences."
    }
}
```

Figure 9. 用于领域特定提示学习的两部分提示。第一部分（上部）提供领域背景，第二部分（下部）提出具体问题。

SFT Learning Prompt

```
"{Question}\n"
"Please think about this question as if you were a human pondering deeply,"
"then give your final answer between the <answer> </answer> tags."
Please provide only the single option letter (e.g., A, B, C, D, etc.) within the
<answer> </answer> tags.
```

Figure 10. SFT 的数据格式。每个样本由一个视频和一个对话组成，其中模型被训练根据用户的查询生成助手的回应。

RL Learning Prompt

```
DOMAIN_CONTEXTS = {
    'Surgery': {
        'context': "Surgical procedure analysis from first-person perspective.
        Key focus: Precise tool identification, surgical phases, hand-specific
operations." },
    'Industry': {
        'context': "Technical repair/assembly tasks from first-person perspective.
        Key focus: Component identification, systematic workflows,
troubleshooting."},
    'Extreme Sports': {
        'context': "High-speed extreme sports from first-person perspective.
        Key focus: Motion patterns, terrain navigation, handling motion blur."
    },
    'Animal Perspective': {
        'context': "Animal perspective videos with alternative movement patterns.
        Key focus: Species identification, behavioral patterns, scale differences."
    }
}
```

Figure 11. 用于强化学习训练的提示结构。模型生成一个响应，然后由奖励模型进行评估，以提供优化策略的反馈。

高级时序和因果推理是一个关键的区别因素。相反，对于更直接的感知任务如计数 (C)，性能差距缩小，几个开源模型表现出竞争力。这种详细视角也揭示了引人注目的失败模式。例如，在动物视角定位任务中，Gemini 2.5 Pro 取得了 42.42 % 的高分，而 GPT-4.1 却意外地得了零分。质性审查表明，尽管 Gemini 正确提供了时间戳，GPT-4.1 却默默认指帧索引（例如，“猫首次与塑料交互在第四幅图像中”），未能遵循提示中明确指示使用提供的视频 FPS 的指令。这种在相同模型的其他领域中未被观察到的特定故障表明其指令跟随能力可能易碎且依赖于上下文。最终，这种细致的分析确认了主要瓶颈：挑战不在于基本的感知能力，而更多在于在新的、专业化的环境中强有力地应用高级技能——如时序推理、知识整合以及一致的指令遵循。

9. 数据集的更多细节

本附录为我们提出的 QA 任务分类法中的 15 个子任务提供了详细的说明性示例。如正文中所述，这些任务被分为四个核心类别：识别、定位、预测和计数。图 12 展示了每个子任务的一个代表性问题、封闭答案、开放答案和三个对应的视觉框架，突出了在不同自我中心视频领域中存在的多样挑战。

这些例子涵盖了从手术录像中的细粒度对象识别到动态户外场景中高层次动作预测的情形，反映了所需推理能力的广度。每个子任务都是精心设计的，以针对自我视角理解的特定维度，例如检测细微的手-物体交互以及从部分观察中预测即将发生的活动。总体而言，它们构成了一个严格而全面的基准，用于评估在真实自我视角环境中多模态模型的表现，其中成功依赖于时间上下文、视觉细节和任务知识的无缝整合。

在下文中，我们为外科手术（图 13）、工业（图 14）、极限运动（图 15）和动物视角（图 16）这四个不同领域的所有子任务展示了具有代表性的问答示例。

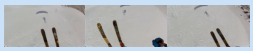
Categories	Sub-tasks	Open-set Questions, Options and Answer	Close-set Answer	Video Frames
Identification	Interaction Identification	Q: What object is the cat interacting with in this video segment? A: Plastic B: Bench C: Vehicle D: Filament	The cat is interacting with a filament.	
	Dominant Held-Object Ident.	Q: Which tool was the operator's left hand predominantly interacting with in the video segment from 0.00s to 30.00s? A: Battery B: battery connector C: Board D: Button	During the segment from 0.00s to 21.01s, the operator's right hand ... with most frequently (10 instance(s)).	
	Object Not Visible Identification	Q: Which surgical instrument is NOT visible in the video segment from 0.00s to 60.00s? A: clipper B: irrigator C: scissors D: bipolar	In the segment, the following instruments are visible ... appear in this segment.	
	Action Sequence Identification	Q: What sequence of actions is performed in the time segment from 0s to 103s? A: Fly → Left → Right B: Climb → LeftRight → CurveLeft C: Vault → Right → Left D: Vault → Right → CurveLeft	During the time segment from 0s to 103s, the performer ... Fly → Left → Right.	
	Special Action Identification	Q: What action is being performed in the video segment from 6s to 8s? A: Spin B: Flip C: Jump D: Fly	The special action 'jump' is performed in the video ... a distinctive athletic maneuver.	
	Sport Identification	Q: What extreme sport is being performed in the video segment? A: Jetski B: Skiing C: Parkour D: Speedflying	The video shows skiing activity based on the movement patterns and environment.	
	Animal Identification	Q: What type of animal is featured in this egocentric video segment? A: Cheetah B: Cat C: Turtle D: Alligator	The video shows a cat from an egocentric perspective.	
Localization	Action Temporal Localization	Q: In this video segment, approximately at what timestamp does the assistant surgeon right hand first start grasp specimen bag using grasper? A: 5.7s B: 3.1s C: 11.3s D: 18.6s	In this video segment, 3 types of operators appeared: main surgeon left hand (MSLH) ... from 2.1s to 4.1s ... until 18.1s.	
	Object Spatial Localization	Q: Approximately in which region of the view was the screwdriver located when the operator's right hand was in take with it around timestamp 3.33s? A: bottom-center B: bottom-left C: bottom-right D: center	An instance of interaction between the operator's right hand and the screwdriver...The screwdriver was determined to be in the 'bottom-right' region of the 3×3 grid overlay.	
	Interaction Temporal Local.	Q: When does the cat first interact with the plastic in this video clip? A: 13.4-15.4 seconds B: 10.1-12.1 seconds C: 2.2-4.2 seconds D: 6.5-8.5 seconds	The cat first interacts with the plastic from 6.0 to 9.0 seconds in this video clip.	
Prediction	Next Direction Prediction	Q: Based on the video segment, what is the next direction of movement? A: Left B: Right C: Curve left D: Curve right	Based on the movement patterns in the video segment, the next direction of movement is 'Left'. The clip ... forward (2s-7s), lift (7s-26s).	
	Next Interaction Prediction	Q: Based on the activity observed up to 30.00s, what is the predicted next type of interaction for the operator's left hand? A: contact the board B: contact the oscilloscope C: contact the oscilloscope component D: contact the pliers	At 2.47s, the right hand performed 'take' with 'board' ... at 25.87s approximately 1.07s. The predicted next interaction by the operator's left hand is 'contact the board'.	
	Next Action Prediction	Q: Following the conclusion of the 'carlot triangle dissection' phase shown in the clip, what is the predicted key action that will begin the next phase? A: use clipper to clip blood vessel B: use hook to cut peritoneum C: use clipper to clip cystic duct D: use grasper to grasp liver	…the beginning use hook to dissect cystic artery, which indicates that the core dissection task of this phase has been completed…the key content of the next phase…clipping and cutting phase.	
	Next Phase Prediction	Q: Based on the surgical activity shown in this clip, which surgical phase immediately follows the 'carlot triangle dissection' phase? A: gallbladder extraction phase B: gallbladder packaging phase C: clipping and cutting phase D: preparation phase	The current phase is the calot triangle dissection phase. From the captured video clip, we can observe at the beginning ... next phase will be the clipping and cutting phase.	
Counting	Object Counting	Q: How many distinct types of objects (e.g. tools, components, buttons as a single category) were visible in the video segment from 0.00s to 10.00s?? A: 10 B: 11 C: 12 D: 13	...battery, battery connector, board, electric screwdriver, oscilloscope...welder component, welder station. This amounts to a total of 11 distinct object types.	

Figure 12. 我们四个核心类别中十五个子任务的详细示例：识别、定位、预测和计数。对于每个子任务，我们展示一个代表性问题、对应的答案以及视频中的一个视觉帧。这些示例说明了我们基准的多样性，涵盖从低级感知（例如，对象识别、空间定位）到高级推理（例如，下一步动作预测、动态事件计数）。

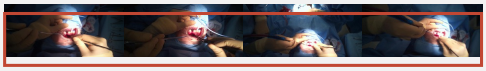
1. Dominant Held-object Identification

Question: Which tool was the main surgeon left hand predominantly interacting with in the video segment from 0.00s to 10.00s?

Options:
A: hook
B: clipper
C: grasper
D: scissors

Detailed Answer: In the segment from 0.00s to 10.00s, the main surgeon left hand used: grasper: 10 times. 'grasper' was clearly the most frequent with 10 occurrences.

2. Object Not Visible Identification



Question: Which of the following objects was NOT visible in the video segment from 0.0s to 18s?

Options:
A: Skewer
B: Raspatory
C: Mouth Gag
D: Bipolar Forceps

Detailed Answer: In the video segment from 0.0s to 18s, the following (refined) object types were observed: [Bipolar Forceps, Mouth Gag, Raspatory, Suction Cannula]. Considering the options, 'Skewer' was not observed in this segment.

3. Object Spatial Localization

Question: Approximately in which region of the view was the 'grasper' (operated by main surgeon left hand) located around timestamp 3.33s?

Options:
A: top-right
B: top-left
C: top-center
D: center-right

Detailed Answer: Around 3.33s, the visible surgical instruments in the frame include: grasper, scissors. The 'grasper' operated by main surgeon left hand has a bounding box at coordinates [x=0.041, y=0.000, width=0.281, height=0.256] (normalized coordinates). Based on the tool's center point, it is located in the 'top-left' region of the 3×3 grid overlay (top-left, top-center, top-right, center-left, center, center-right, bottom-left, bottom-center, bottom-right).

4. Action Temporal Localization

Question: In this video segment, approximately at what timestamp does the assistant surgeon right hand first start grasp specimen bag using grasper?
Options:
A: 5.7s
B: 3.1s
C: 11.3s
D: 18.6s

Detailed Answer: In this video segment, 3 types of operators appeared: main surgeon left hand (MSLH), assistant surgeon right hand (ASRH), main surgeon right hand (MSRH)...The MSRH performed use grasper to grasp specimen bag from 3.1s to 5.1s; performed use irrigator to retract liver from 5.1s to 9.1s; performed use irrigator to retract liver from 10.1s to 18.1s. The assistant surgeon right hand first started to grasp specimen bag using grasper at 2.1s and continued until 18.1s.

5. Next Phase Prediction

Question: Based on the surgical activity shown in this clip, which surgical phase immediately follows the 'clipping and cutting' phase?

Options:
A: gallbladder extraction phase
B: cleaning and coagulation phase
C: gallbladder packaging phase
D: gallbladder dissection phase

Detailed Answer: The current phase is the clipping and cutting phase. From the captured video clip, we can observe at the beginning use scissors to cut cystic duct, which indicates that the gallbladder's "lifeline" has been severed; while at the end of the clip, we can see use hook to dissect gallbladder, which is the key content of the next phase, we can predict that the next phase will be the gallbladder dissection phase.

6. Next Action Prediction

Question: Following the conclusion of the 'cleaning and coagulation' phase shown in the clip, what is the predicted key action that will begin the next phase?

Options:
A: use grasper to retract liver
B: use grasper to grasp specimen bag
C: use bipolar to dissect cystic plate
D: use irrigator to retract omentum

Detailed Answer: The current phase is the cleaning and coagulation phase. From the captured video clip, we can observe at the beginning use irrigator to aspirate fluid, which indicates that surgical field cleaning and final hemostasis have been completed; while at the end of the clip, we can see use grasper to grasp specimen bag, which is the key content of the next phase, we can predict that the next phase will be the gallbladder extraction phase.

7. Object Counting

Question: How many distinct types of surgical instruments were fully or partially visible in the video segment from 0.00s to 10.00s?

Options:
A: 4
B: 0
C: 1
D: 2

Detailed Answer: In the segment from 0.00s to 10.00s, the distinct instruments that were fully or partially visible include: grasper, hook. Total count: 2 distinct types of surgical instruments.

Figure 13. 外科领域的代表性 QA 示例。

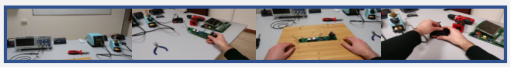
1. Object Not Visible Identification

Question: Which of the following objects (using refined categories like 'button' for all buttons) was NOT visible in the video segment from 0.00s to 10.00s?

Options:
A: battery
B: battery connector
C: board
D: oscilloscope

Detailed Answer: In the video segment from 0.00s to 10.00s, the following (refined) object types were observed: [battery, board, oscilloscope, screwdriver, socket, welder component, welder station]. Considering the options, 'battery connector' was not observed in this segment.

2. Dominant Held-object Identification

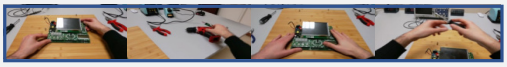


Question: Which tool was the operator's left hand predominantly interacting with in the video segment from 0.00s to 30.00s?

Options:
A: battery
B: battery connector
C: board
D: button

Detailed Answer: During the segment from 0.00s to 30.00s, the operator's left hand had the following interaction dynamics: 'battery': 1 instance(s)… 'board': 3 instance(s), observed in 3 labelled frames. The tool 'board' was interacted with most frequently and/or for the longest based on timestamp analysis.

3. Object Spatial Localization



Question: Approximately in which region of the view was the electric screwdriver located when the operator's right hand was in contact with it around timestamp 2.37s (within segment 0.00s-10.00s)?

Options:
A: bottom-center
B: bottom-left
C: bottom-right
D: center

Detailed Answer: An instance of interaction (contact) between the operator's right hand and the electric screwdriver was noted around 2.37s. At this moment, the electric screwdriver was determined to be in the 'center' region of the view. This is one example within the segment 0.00s to 10.00s.

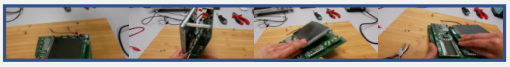
4. Action Temporal Localization

Question: At what approximate timestamp did the operator's right hand perform the second 'contact' with the button within the video segment from 0.00s to 15.00s?

Options:
A: 1.50s
B: 12.50s
C: 2.27s
D: 7.50s

Detailed Answer: Within the segment from 0.00s to 15.00s, the operator's right hand performed 'contact' with the button multiple times: the first at 1.20s, the second at 2.27s. The second occurrence was at approximately 2.27s.

5. Next Interaction Prediction



Question: Based on the activity observed up to 30.00s, what is the predicted next type of interaction for the operator's left hand?

Options:
A: contact the board
B: contact the oscilloscope
C: contact the oscilloscope component
D: contact the screen

Detailed Answer: Observations up to 30.00s suggest a forthcoming interaction. During the observed context: at 2.93s, the right hand performed 'contact' with 'screen'…'screen' was visible in the context for approximately 1.00s…

6. Object Counting

Question: How many distinct types of objects (e.g. tools, components, buttons as a single category) were visible in the video segment from 0.00s to 10.00s?

Options:
A: 10
B: 11
C: 12
D: 13

Detailed Answer: The special action 'jump' is performed in the video segment from 4s to 6s, demonstrating a distinctive athletic maneuver.

Figure 14. 来自工业领域的代表性问答示例。

1. Special Action Identification

Question: What action is being performed in the video segment from 4s to 6s?

Options:
A: Climb
B: Spin
C: Jump
D: Fly

Detailed Answer: The special action 'jump' is performed in the video segment from 4s to 6s, demonstrating a distinctive athletic maneuver.

2. Sport Identification

Question: What extreme sport is being performed in the video segment?

Options:
A: Mountain Bike
B: Jetski
C: Skiing
D: ParkourDetailed

Detailed Answer: The video shows jetski activity based on the movement patterns and environment.

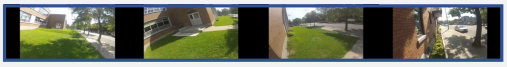
3. Action Sequence Identification

Question: What sequence of actions is performed in the time segment from 0s to 29s?

Options:
A: Curveleft → Jump → Curveleft
B: Leftright → Vault → Left
C: Right → Flip → Left
D: Right → Fly → Left

Detailed Answer: During the time segment from 0s to 29s, the performer executes the following sequence: Curveleft → Jump → Curveleft.

4. Action Temporal Localization



Question: At what approximate time does the 'walk' action begin in this video segment?

Options:
A: 13.3s
B: 6.7s
C: 10.2s
D: 2.3s

Detailed Answer: The 'walk' action begins at 10.2s in this video segment.

5. Next Direction Prediction

Question: Based on the video segment, what is the next direction of movement?

Options:
A: Forward
B: Left
C: Left then right
D: Curve right

Detailed Answer: Based on the movement patterns in the video segment, the next direction of movement is 'Forward'. The clip includes the following actions: right (3s-4s), left (4s-5s), right (5s-6s), curveleft (6s-11s), left (11s-15s).

Figure 15. 极限运动领域的代表性问答示例。

1. Animal Identification

Question: What type of animal is featured in this egocentric video segment?

Options:
A: Alligator
B: Cheetah
C: Turtle
D: Lizard

Detailed Answer: The video shows a alligator from an egocentric perspective.

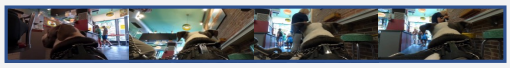
2. Interaction Identification

Question: What object is the dog interacting with in this video segment?

Options:
A: Ball
B: Person
C: Bench
D: Bird

Detailed Answer: The dog is interacting with a bench.

3. Interaction Temporal Localization



Question: When does the dog first interact with the person in this video clip?

Options:
A: 9.0-13.0 seconds
B: 2.5-6.5 seconds
C: 30.8-34.8 seconds
D: 66.2-70.2 seconds

Detailed Answer: The dog first interacts with the person from 7.0 to 15.0 seconds in this video clip.

Figure 16. 动物视角领域的代表性问答示例。