

# EgoExoBench: 用于多语言大模型中第一人称和第三人称视角视频理解的基准测试

Yuping He<sup>1,2\*</sup>, Yifei Huang<sup>2,3\*</sup>, Guo Chen<sup>1</sup>, Baoqi Pei<sup>2,4</sup>, Jilan Xu<sup>2,5</sup>  
Tong Lu<sup>1†</sup> and Jiangmiao Pang<sup>2†</sup>

<sup>1</sup>Nanjing University, <sup>2</sup>Shanghai AI Laboratory, <sup>3</sup>The University of Tokyo

<sup>4</sup>Zhejiang University, <sup>5</sup>Fudan University

<https://github.com/ayiyayi/EgoExoBench>

## Abstract

在第一人称（自我中心）和第三人称（外部中心）视角之间传递和整合知识是人类智慧的内在特性，使人类能够从他人那里学习并传达自己的经验见解。尽管多模态大型语言模型（MLLMs）迅速发展，但其进行此类跨视角推理的能力仍未得到探索。为了解决这个问题，我们引入了 EgoExoBench，这是首个针对自我中心-外部中心视频理解和推理的基准。EgoExoBench 基于公开的数据集构建，包含超过 7,300 个问题-答案对，涵盖了三个核心挑战中的十一项子任务：语义对齐、视角关联和时间推理。我们评估了 13 个最先进的 MLLMs，发现尽管这些模型在单视角任务中表现出色，但在自我-外部中心的上下文中，它们在跨视角对齐语义、准确关联视图和推断时间动态方面存在困难。我们希望 EgoExoBench 能成为研究具备人类般跨视角智能的具身代理和智能助手的宝贵资源。

## 1 引言

理解和传递第一人称（自我中心）视角与第三人称（他人中心）视角之间的知识是人类智力的基本方面。我们经常通过观察第三人称的演示来学习新技能，并将这些动作映射到我们自己的自我中心体验中。相反，我们可以采用他人中心的视角来表达见解并指导他人。超越简单的传递，人类在跨视角联合推理方面表现出色：在程序化任务中，我们将实时的自我中心观察与他人中心的示例对齐，以预测下一步；在导航中，我们将自我中心的经验与全球地图融合，以规划高效的路线。这种双视角推理能够在动态环境中实现稳健的感知和技能习得。

在人工系统中复制这种自我与外部的推理为更强大的具身智能体提供了一条有前景的途径。能够对齐大规模外中心知识（如教学视频）的机器人可以更有效地执行任务并更好地执行模仿学习 [53, 1, 52]。弥合第一人称和第三人称视角对于无缝人机协作也至关重要 [37, 61]。

最近的基准测试 [40, 14, 78, 34, 8] 在多模态大语言模型（MLLMs）的单视角视频理解方面取得了显著的进展。例如，EgoSchema [40] 评估了以自我为中心流的时间推理，Video-MME [14] 评估了在各种第三人称镜头上的多模态问答，MLVU [78] 则测试了长篇视频理解。虽然每个基准测试从不同的角度推动 MLLMs 在视频理解中的界限，但它们都只在自我中心或外部中心的数据上独立运行。对于模型在整合自我-外部视点信息的能力没有系统的评估，留下了 MLLMs 是否能够像人类一样进行跨视图推理的疑问。

为了填补这一空白，我们引入了 EgoExoBench，这是第一个专门为 MLLM 中的跨视图视频理解而设计的基准。EgoExoBench 基于公开可用的自我中心和外部中心视频源构建，并通过严格的注释协议进行策划。它包括超过 7,300 个问答对，针对三个关键维度：自我-外部关系、自我-外部视图转换和自我-外部时间推理。自我-外部关系通过测试模型是否能够识别出在第一人称和第三人称视图中相同的实体或动作来评估语义对齐。自我-外部视图转换探讨

\*equal contribution. † corresponding authors

空间对应性，衡量模型在自我中心和外部中心视点之间进行转换的能力。自我-外部时间推理通过要求模型对齐并预测跨异步或重叠视频流的事件序列来检查时间序列推理。在这三个维度上的 11 个子任务构成了一个全面的自我-外部视图理解和推理评估套件。

我们对 13 个开源和闭源的 MLLMs 进行了广泛的实验，包括 GPT-4o [25]、GPT-o4-mini [42]、Claude 3.7 Sonnet [2]、Qwen2.5-VL [57] 和 InternVL3 [79]。结果揭示了一个明显的模式：在单视图任务中表现强劲模型在面对跨视图挑战时表现出显著的下降。进一步分析表明，甚至配备了显式推理能力的模型在将基于文本的推理与双视图视频的信息交织在一起时也会遇到困难。这些结果突显了当前能力与化身代理和人-机器人协作所需的人类般跨视图理解之间的显著差距。我们希望 EgoExoBench 将作为一个全面的自我-外部视图推理评估套件，并激发新的架构和训练策略来弥合这一差距。

## 2 相关工作

视频基准测试。随着多模态大型语言模型（MLLMs）在视频理解中的快速应用 [33, 75, 43]，许多工作致力于对 MLLMs 的视频理解能力进行基准测试。像 Next-QA [64]、Next-GQA [65]、MLVU [78]、LV-Bench [59]、Video-MME [14] 和 CG-Bench [8] 等基准测试评估了从一般问答到多模态理解的多个方面。自我中心数据集 [10, 16, 38, 46, 4, 11, 27, 74, 55] 促使 EgoSchema [40] 探索自我中心长形式视频的理解，而 HourVideo [7] 测试空间智能。然而，这些努力各自仅关注单一视角，无论是自我中心还是外部视角，并没有评估模型整合两者的能力。EgoExoBench 填补了这一空白，提供了第一个跨视角视频理解的基准，测量第一人称和第三人称视频流之间的语义、空间和时间推理能力。

自我-外部数据集。之前的工作已经制作了多种配对的自我中心-外部中心的视频集合。Charades-Ego [54] 和 Home Action Genome [47] 捕捉了在家居环境中同步的第一人称和第三人称视频，而 Assembly101 [51] 和 LEMMA [26] 提供了多步骤任务的同步录制。像 Ego-Exo4D [17] 和 EgoExoLearn [23] 这样的视频集合提供了广泛的规模和丰富的注释，但缺乏任何下游的问答或推理基准。专注于机器人技术的数据集如 iGibson [32] 和 H2O [30] 记录了用于姿态估计和导航的固定摄像机阵列，但其范围仅限于这些任务。EgoExoBench 建立在这些数据集的基础上，为评估跨视角视频理解和 MLLM 的推理提供了一个统一的测试平台。

自我外-视频理解。第一人称视频理解方法通常利用大规模的第三人称数据 [41, 18, 56] 来弥补第一人称语料库的有限规模。之前的工作主要分为三大类：联合视角不变学习 [69, 58, 67, 66, 21]、领域适配 [63, 73, 68]，以及知识蒸馏 [35, 45]。虽然这些方法提高了识别性能，但它们主要集中在识别问题上，并没有评估模型跨视角推理的能力，而这正是 EgoExoBench 专门设计用于评测的方面。

## 3 基准测试

### 3.1 任务套件

EgoExoBench 是一个用于评估多模态大语言模型中跨视角视频理解的大规模基准。它涵盖了多样的环境和活动，探索自我和外部推理的三个关键维度：语义对齐、空间对应和时间推理。虽然开放式问答能很好地反映人类对话，但自动评估自由形式的回答可能会很困难且容易出错。为了实现对跨视角视频理解的可靠、可扩展的评估，EgoExoBench 采用了多项选择问题（MCQ）格式。在接下来的部分中，我们详细介绍我们的任务套件和问答生成流水线，旨在产生多样化、高质量的多项选择问题，以严格探测语义对齐、空间对应和时间推理。创建一个跨视角视频理解基准面临独特的挑战：问题不仅要涵盖多个时间段，还需要合成第一人称和第三人称视角之间的信息。EgoExoBench 通过首先建立捕捉自我和外部推理本质的三个核心维度来解决这一问题：语义对齐、空间对应和时间关系。然后，我们利用丰富的、公开可用的自我中心-外部中心数据集，从中挖掘它们的同步和异步多视角录像，空间注释（例如，姿势，边界框）和时间动作标签，以构建高质量的 MCQs。

我们的任务套件包括 11 个子任务，这些任务在 3 个关键维度下组织。自我-外部关系评估跨视角的语义对齐；自我-外部视图转换测量自我中心和外部中心坐标系之间的空间对应；自我-外部时间推理通过要求模型对齐并预测成对视频流中的事件流来检查序列整合。这些子任务共同形成了自我-外部视图视频理解与推理的全面评估套件。

如图 1 所示，我们的问题解答创建遵循一个三阶段的流程，以满足跨视角评估的特定需求。下面，我们描述一般的构建过程。我们将在后续部分提供特定任务的详细信息。

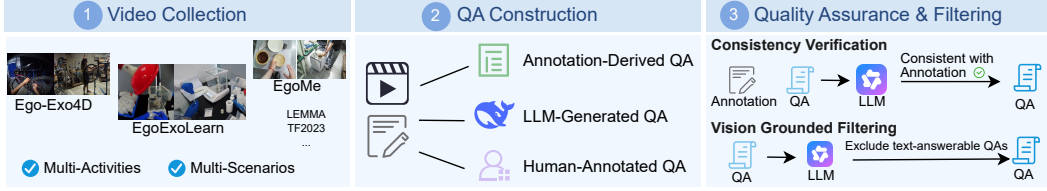


Figure 1: EgoExoBench 的构建流程。

视频合集。我们的基准整合了来自六个自我-外界数据集的视频：Ego-Exo4D [17]、EgoExoLearn [23]、LEMMA [26]、EgoMe [44]、TF2023 和 CVMHAT [19]。这些数据集涵盖了广泛的环境（例如，厨房、实验室、运动场）和活动（例如，烹饪、运动、维修）。Ego-Exo4D [17]、LEMMA [26]、TF2023 [76] 和 CVMHAT [19] 包含同步多视角视频，而 EgoExoLearn [23] 和 EgoMe [44] 则具有异步示范-跟随录制。这种多样化的设置允许全面评估模型在同步和异步跨视角场景中关联和理解能力。

问答构建。为了适应基准中的多样化任务，我们采用了多种 QA 构建策略。(1) 基于注释的 QA。对于具有结构化和确定性注释的任务，我们使用预定义模板直接从注释构建 QA 对。(2) 通过 LLM 生成的 QA。对于在 QA 生成中需要开放式推理的情况，我们利用 LLM [36, 70] 生成具备上下文意识的 QA 对。这些 LLM 被提供任务定义、详细注释、示范 QA 对和任务特定限制。(3) 人工标注的 QA。对于需要细粒度空间理解的任务，目前的 MLLMs 无法产生可靠的结果。因此，我们聘请人类标注员来构建相应的 QA 对。所有 QA 对都被格式化为四选一的多项选择题。

质量保证和筛选。为了确保构建的问答对的可靠性，我们实施了质量保证和筛选过程。(1) 一致性验证。我们通过提示一个大型语言模型 [70] 来执行自动一致性检查，以验证每个问答对相对于原始视频注释的逻辑正确性。如果问题含糊不清、与注释不一致或可有多种合理答案，则弃之。(2) 基于视觉的筛选。为确保每个问题都需要视觉理解，我们将仅文本的问题展示给一个大型语言模型 [70]，并舍弃那些可以仅根据文本输入回答的问答对。

## 3.2 任务特定细节

### 3.2.1 自我-外部关系

自我-外部关系维度评估一个模型在第一人称视角和第三人称视角之间语义对齐内容的能力，这是任何跨视角推理的关键基础。我们设计了四个子任务：任务、动作、人物和对象层次的匹配，旨在进行全面评估。

任务关系 (TR)。我们采用了 Ego-Exo4D [17]、EgoExoLearn [23] 和 LEMMA [26] 中的同步和异步演示。我们将以自我为中心的高层次活动片段（例如，在厨房组装设备）与标注有相同任务 ID 的以他人视角拍摄的视频进行配对。为了防止简单的场景匹配，真实配对来自同一场景内的不同环境，而负面候选则是在视觉上相似的背景中描述其他任务。QA 配对通过基于注释的模板生成，将任务 ID 转换为四选一的多项选择题，并通过自动一致性检查和基于视觉的提示进行过滤，以确保依赖于视觉输入。

动作关系 (AR)。给定一个视频片段，此任务旨在从另一个视角中识别出捕捉到相同动作的相应片段。它需要弥合由于视角引起的变化，如摄像机运动和遮挡，以建立细粒度的动作对应关系。我们使用来自 LEMMA [26] 和 EgoExoLearn [23] 的视频构造问答对。对于 LEMMA [26]，真实视频片段在时间上与查询对齐，但从不同视角录制。对于 EgoExoLearn [23]，正确匹配的视频片段来自描述相同动作的不同视频。

对象关系 (OR)。给定一个查询视频，此任务旨在识别来自另一个视角的候选视频，该视频涉及与同一对象的交互。我们从 LEMMA [26] 中整理视频并从注释中导出对象交互。为了增加难度，未提供关于对象的文本提示。模型必须首先推断每个视频中交互的对象，然后在不同视角之间进行匹配。正确的视频在时间上与查询对齐，但来自另一个视角，而负样本候选则是在相同环境中与不同对象的交互视频。

人物关系 (PR)。给定一个视频片段，这项任务旨在从不同的视角识别同一个人。我们从 CVMHAT [19] 中整理数据，该数据在户外场景中提供了同步的第一人称视角和顶视图录制。在每个查询中，目标个体用边界框突出显示。正确答案对应于从另一个视角观察到的相同

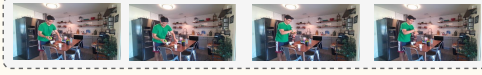
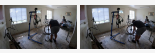
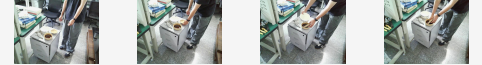
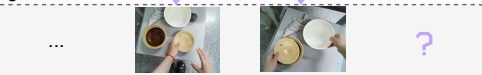
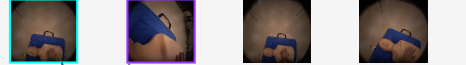
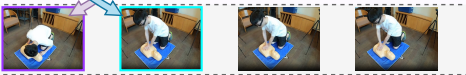
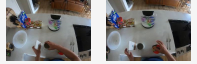
| Ego-Exo Relation                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Task Relation</b><br>Query Video:<br><br>Q: Which candidate video most accurately matches the activities in the query video?<br>A.  B.  C.  D.                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | <b>Action Relation</b><br>Query Video:<br><br>Q: Which candidate video most accurately matches the actions in the query video?<br>A.  B.  C.  D.                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <b>Object Relation</b><br>Query Video:<br><br>Q: Which candidate video most accurately matches the interacted object in the query video?<br>A.  B.  C.  D.                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | <b>Person Relation</b><br>Query Video:<br><br>Q: Which candidate video most accurately matches the person (within the bbox) in the query video?<br>A.  B.  C.  D.                                                             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Ego-Exo View Transition                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <b>Egocentric Wearer Identification</b><br>Egocentric Video:<br><br>Q: Which boxed person in the third-person view is the camera wearer who captured the egocentric video?<br>A.  B. <br>C.  D.   | <b>Body Part Action Understanding</b><br>Video1:<br><br>Video2:<br><br>Q: What is the difference in the actions involving the person's feet in the two videos?<br>A. In Video 1, the person lifts left foot...<br><b>B. In Video 1, the person steps on the solid base with left foot; in Video 2, the person does it with right foot.</b><br>C. In Video 1, the person places right foot on the pedal...<br>D. In Video 1, the person steps on the solid base ... |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <b>Direction Prediction</b><br>Egocentric Video:<br><br>Q: In which view does the person move to the foreground of the scene?<br>A.  B. <br>C.  D.  |
| Ego-Exo Temporal Reasoning                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <b>Action Prediction</b><br>Reference Video:<br><br>Egocentric Video:<br><br>Q: Based on the reference video, what is the next action in the egocentric video?<br>A. The right hand picks up the pot and pours the water into the bowl.<br><b>B. The left hand holds the lid and cleans it with water in the pot.</b><br>C. The right hand pours some water from the bowl into the pot.<br>D. The left hand holds the lid and places it on a yellow pot. |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | <b>Sequence Alignment</b><br>Video1:<br><br>Video2:<br><br>Q: What is the key difference in the action sequences between Video 1 and Video 2?<br>A. Video 1 confirms patient consciousness, while Video 2 does not.<br>B. Video 1 includes chest compressions, while Video 2 does not.<br><b>C. Video 1 places the hand on top of the first before confirming consciousness, while Video 2 does it after.</b><br>D. Video 2 includes a conversation before confirming patient... |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <b>Action Ordering</b><br>Video1:<br><br>Video2:<br><br>Q: What is the correct temporal order between the actions in these two videos?<br>A. Actions in Video 2 occur before actions in Video 1.<br>B. Actions in Video 1 are part of the sequence of actions in Video 2.<br>C. Actions in Video 2 are part of the sequence of actions in Video 1.<br><b>D. Actions in Video 1 occur before actions in Video 2.</b>                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | <b>Skill Evaluation</b><br>Reference Video:<br><br>Q: Which candidate video performs the task best according to the reference video?<br>A.  B.  C.  D.                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |

Figure 2: 自我拓展基准中的示例选择题。正确答案以粗体突出显示。

时间戳的同一个体。为了防止基于场景或时间线索的捷径，负候选项从相同时间戳中选取，但描绘的是不同的个体。

### 3.2.2 自我视角与外部视角转换

自我-外部视角转换中的子任务评估模型在第一人称和第三人称视角之间转译空间信息的能力。我们定义了三个子任务：自我中心的佩戴者识别、方向预测和身体部位动作理解，每个子任务都由定制的数据源和问答策略构建。

第一人称视角佩戴者识别 (EWI)。给定一个第一人称视角视频，该任务旨在从第三人称视角识别摄像机佩戴者。与人物关系任务相比，这个设置更加具有挑战性，因为第一人称视角佩戴者通常在自己的视角中缺席。模型必须通过分析周围人和物体的相对位置来推断空间关系，然后将这些关系从第一人称视角映射到第三人称视角。在 TF2023 数据集 [76] 的基础上构建，该数据集提供同步的第一人称-第三人称图像对以及标注的人的边界框。我们仅保留包含四个或更多个体的样本，以增加空间复杂性。为了构建 MCQ 候选项，我们根据同一第三人称视角框架选择不同的人。

方向预测 (DP)。给定一个自我中心的视频，此任务考察模型将自我视角中的运动投射到全局场景的能力。利用 Ego-Exo4D 的 [17] 同步多视角录像，我们首先通过 Qwen2.5 32B 过滤的解说识别出具有明显实体移动的片段。专业注释员随后标记穿戴者或物体在自我中心视角中的移动方向。每个问题提供多个异中心片段，其中只有一个符合指定的方向条件。手动验证确保仅保留模糊性小且视觉上有据可循的示例。

身体部位动作理解 (BPA)。该任务探测两个视角下肢体运动的细粒度空间映射。我们从 Ego-Exo4D [17] 和 EgoExoLearn [23] 中提取片段，这些片段在旁白中明确提到身体部位，如左手或右手。Qwen2.5-32B [70] 识别相关的动词和对象，而 DeepSeek-V3 [36] 生成至少共享一个动作或对象提示的干扰项。我们生成的问题是要求将表现相同身体部位活动的自我视角和外部视角片段配对。最后，我们应用一个基于 Qwen2.5-32B 的过滤器，去除任何仅通过文本就可以解决的实例。

### 3.2.3 自我-外部时间推理

自我-外部时间推理评估模型对齐和推断自我中心和外部中心视频流中事件流的能力。我们定义了四个子任务——下一个动作预测、动作排序、序列对齐和技能评估。

动作预测 (AP)。我们将部分自我中心视角的剪辑与更长的非自我中心演示配对，并要求模型预测后续的自我中心动作。为了构建数据，我们从 LEMMA [26] 和 EgoMe [44] 收集视频。对于 LEMMA [26]，我们利用动作注释提取真实后续动作和负面候选项。具体来说，我们在自我中心序列中选择当前动作周围的十个动作，排除真正的后续动作，并提示 Qwen2.5-32B [70] 从这组动作中识别出三个最可能的干扰项。在来自 EgoMe [44] 的异步设置中，我们依靠自我中心视角的带注释动作描述，并使用 DeepSeek-V3 [36] 相应地生成三个干扰项。最后，我们使用 Qwen2.5-32B [70] 筛选出可以仅基于文本输入回答的问题。

动作排序 (AO)。在这个任务中，模型被给予两个短视频片段，一个是第一视角的，一个是外视角的，需要判断它们的时间关系。为了构建数据，我们从 LEMMA [26] 中获取原始动作注释。每个问题实例由选择两个时间上连续的动作片段组成。为了确保时间排序的有效性，我们使用动作的文本描述提示 Qwen2.5-32B [70] 以滤除模糊或无序的动作对。

序列比对 (SA)。序列比对通过询问以自我为中心和以他人为中心的视频是否共享相同的动作顺序、是否顺序相反或在关键步骤上存在差异，来扩展多步骤活动的排序。为了构建数据，我们首先从 Ego-Exo4D [17] 中策划包括关键步骤注释的视频。从同一活动类别中选择视频对，每对视频至少包含两个共享步骤。对于每对视频，我们组织步骤描述，并使用 DeepSeek-V3 [36] 生成强调动作序列关键差异的问题。为了确保问题质量，我们使用 Qwen2.5-32B [70] 来验证正确答案是否与注释步骤一致。最后，我们利用 Qwen2.5-32B [70] 筛选出只能通过文本输入回答的问题。

技能评估 (SE)。给定一个专家演示视频作为参考，此任务评估从不同视角拍摄的其他视频的技能水平。它被形式化为一个多项选择问题，其中模型在候选中选择最佳或最差的表现。与文本指南相比，专家演示提供了更细致的行为线索，这常常被基于规则的方法忽略。我们从 EgoExoLearn [23] 和 Ego-Exo4D [17] 中构建数据。对于 EgoExoLearn [23]，我们使用外部视角的演示视频作为参考，并从成对的技能注释中构建传递性排名链，以在候选视频之间建立明确的层次结构。对于 Ego-Exo4D [17]，我们选择“迟到专家”视频作为参考，并从相

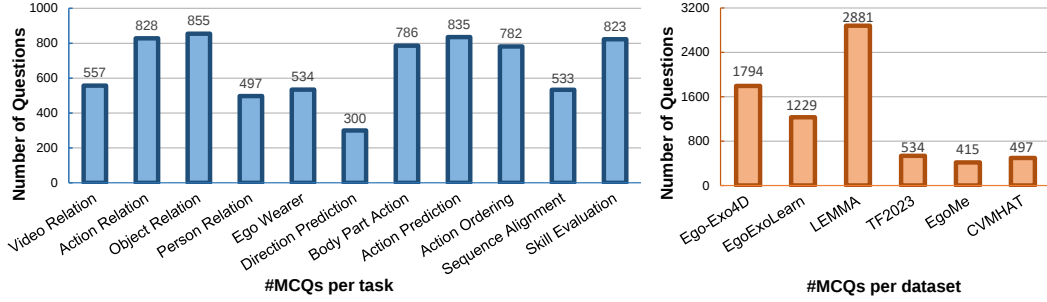


Figure 3: MCQs 在 EgoExoBench 中的分布。左：每个子任务的问题数量。右：每个数据集的问题数量。

同活动中策划具有不同熟练度评分的候选人。视角不佳且无法清楚看到动作的自我视角视频被手动排除。

### 3.3 基准统计

EgoExoBench 包括总计 7,330 个选择题，每个题目均采用四选一格式设计。这些问题是从六个自我-外部配对数据集中构建的，涵盖了广泛的场景和活动。如图 2 所示，基准包括三个任务类别，总共包括十一项子任务。图 3 展示了选择题在子任务和数据集之间的分布。

## 4 实验

我们在零样本多项选择问答 (MCQ) 格式中使用多样化的 MLLMs 对 EgoExoBench 进行评估。闭源模型包括 Claude 3.7 Sonnet [2]、GPT-4o [25] 和 GPT-o4-mini [42]。对于开源模型，我们评估了 Qwen2.5-VL [57]、InternVL3 [79]、LLaVA-OV [31]、LLaVA-Video [75]、NVILA [39] 和 EgoGPT [72]，涵盖了多样化的架构和参数规模。每个模型接收到相同的标准化提示，其中提出问题并附有标记选项，指示模型仅返回其选择答案的字母。我们采用基于规则的方法来提取预测答案。所有评估均使用准确率作为主要指标，且不进行微调或梯度更新以确保公平性。对于闭源模型，我们通过官方 API 进行评估。对于开源模型，所有实验均在四个 A100 GPUs 上进行，每个 GPU 拥有 80GB 的内存。

### 4.1 主要结果

在主要实验之前，我们首先确认 MLLMs 能够可靠地区分多个视频流作为独立的输入，从而验证我们的多视频实验设置（细节在补充材料中）。

表 1 总结了开源和闭源模型在 EgoExoBench 上的性能表现。在开源模型中，Qwen2.5-VL-72B [57] 在整体准确率上达到 47.0%，而 GPT-o4-mini [42] 在闭源组中以 48.0% 领先。在 Ego-Exo 关系类别中，GPT-o4-mini [42] 平均比 Qwen2.5-VL-72B [57] 高出 5%，这表明开源架构在跨视角的显式语义对齐方面仍然落后于最先进的闭源系统。在 Ego-Exo 视角转换和 Ego-Exo 时间推理中，Qwen2.5-VL-72B [57] 分别达到 47.3% 和 37.4% 的平均准确率。其在空间和视频定位上的相对优势可能对这些空间对应和序列整合任务的更好表现有所助益。值得注意的是，尽管 EgoGPT [72] 在大型以自我为中心的视频语料库 (EgoIT-99K) [72] 上进行了微调，但相比于同样大小的模型仅提供了有限的性能提升，这表明仅靠专门的预训练不足以在没有特定任务目标的情况下掌握跨视角推理。

人类表现。为了将模型性能置于背景中，我们从每个子任务中随机抽取 30 个问题，总计 330 个问题，并测量人类在相同选择题上的准确率。给评估者的指示是分别在 1 分钟（快速回答）和 3 分钟（深思熟虑回答）内得出答案。如表 1 所示，在深思熟虑条件下（每题 3 分钟）进行的人员达到了 90.1% 的平均准确率，超过了最佳模型 42%。这一差距表明，跨视角视频理解中，MLLMs 仍然远远低于人类水平。最大差距出现在“自我中心佩戴识别” (EWI) 子任务中，在该任务中，所有人类评估者都能成功地基于人与物体之间的空间关系，在第三人称视角中推断自我中心摄像机佩戴者的身份。相比之下，当前的 MLLMs 难以在不同视角间推理此类空间配置。最小的差异出现在“技能评估”中，即使是人类也发现领域特定的评估（例如，判断篮球水平）具有挑战性。这表明，有效的跨视角技能评估要求 MLLMs 整合领域知识和多视角推理。

Table 1: 各种开源和闭源 MLLMs 在 EgoExoBench 上的性能表现。所有模型中最好的结果用粗体显示，第二好的结果用下划线标出。

|                                   | Avg.        | Relation    |             |             |             | View Transition |             |             | Temporal Reasoning |             |             |             |
|-----------------------------------|-------------|-------------|-------------|-------------|-------------|-----------------|-------------|-------------|--------------------|-------------|-------------|-------------|
|                                   |             | TR          | AR          | OR          | PR          | EWI             | DP          | BPA         | AP                 | AO          | SA          | SE          |
| Human Performance (330 Questions) |             |             |             |             |             |                 |             |             |                    |             |             |             |
| Human (1min)                      | 64.6        | 72.2        | 58.9        | 62.2        | 70.0        | 91.1            | 78.9        | 53.3        | 68.9               | 61.1        | 60.0        | 34.4        |
| Human (3min)                      | 90.1        | 95.5        | 90.0        | 92.2        | 94.4        | 100.0           | 96.7        | 85.5        | 91.1               | 87.8        | 88.9        | 68.9        |
| Qwen2.5-VL-72B                    | 48.5        | 56.7        | 46.7        | 66.7        | 56.7        | 56.7            | 43.3        | 46.7        | 46.7               | 43.3        | 43.3        | 26.7        |
| Claude-3.7-Sonnet                 | 32.8        | 47.1        | 30.0        | 36.7        | 46.7        | 33.3            | 30.0        | 26.7        | 20.0               | 30.0        | 30.0        | 30.0        |
| GPT-4o                            | 38.5        | 53.8        | 30.0        | 60.0        | 56.7        | 33.3            | 26.7        | 30.0        | 26.7               | 33.3        | 43.3        | 30.0        |
| Open-source MLLMs                 |             |             |             |             |             |                 |             |             |                    |             |             |             |
| Qwen2.5-VL-7B                     | 32.8        | 40.2        | 34.4        | 45.5        | 36.0        | 26.4            | 30.7        | 34.5        | 19.2               | 31.0        | 37.1        | 26.1        |
| Qwen2.5-VL-32B                    | 39.7        | 43.3        | 40.1        | 50.4        | 42.3        | 41.7            | <u>34.0</u> | <u>42.4</u> | 38.6               | 31.1        | 45.8        | 27.5        |
| Qwen2.5-VL-72B                    | <u>44.7</u> | 51.0        | 43.5        | 56.6        | <u>49.7</u> | <u>56.7</u>     | <u>37.0</u> | <u>48.1</u> | 39.9               | <u>33.6</u> | 46.1        | <u>29.9</u> |
| InternVL3-8B                      | 31.3        | 36.8        | 30.9        | 37.0        | 27.2        | <u>16.5</u>     | 33.7        | 37.2        | 27.4               | <u>29.8</u> | 47.2        | 20.9        |
| InternVL3-14B                     | 35.1        | 38.8        | 31.9        | 43.5        | 31.8        | 30.5            | 29.0        | 41.1        | 33.9               | 32.9        | 48.1        | 24.2        |
| InternVL3-78B                     | 41.4        | 50.6        | 37.3        | 48.9        | 39.0        | 46.9            | 32.3        | 38.0        | 51.0               | 31.5        | <u>50.6</u> | <u>29.5</u> |
| LLaVA-OV-7B                       | 29.5        | 30.7        | 28.7        | 34.0        | 28.0        | 22.7            | 27.0        | 31.4        | 21.7               | 29.3        | 44.1        | 26.8        |
| LLaVA-Video-7B                    | 31.2        | 33.9        | 29.1        | 35.4        | 27.0        | 29.0            | 27.0        | 36.4        | 23.5               | 28.9        | 43.8        | 28.9        |
| NVILA-8B                          | 29.6        | 30.0        | 25.0        | 30.8        | 31.6        | 23.6            | 27.0        | 37.7        | 24.4               | 26.2        | 44.5        | 24.8        |
| EgoGPT-7B                         | 29.6        | 29.8        | 29.5        | 35.6        | 30.4        | 22.3            | 26.3        | 32.1        | 22.3               | 29.4        | 40.1        | 27.6        |
| Closed-source MLLMs               |             |             |             |             |             |                 |             |             |                    |             |             |             |
| Claude-3.7-Sonnet                 | 31.3        | 33.4        | 33.0        | 35.1        | 30.6        | 37.8            | 30.0        | 33.3        | 26.0               | 28.5        | 30.2        | 26.4        |
| GPT-4o                            | 38.5        | <u>52.9</u> | <u>44.6</u> | 57.4        | 48.0        | 41.6            | 24.1        | 37.1        | 27.7               | 26.1        | 37.4        | 26.3        |
| GPT-o4-mini                       | 48.0        | <u>65.3</u> | 45.2        | <u>56.8</u> | 53.7        | 73.8            | 30.0        | 40.0        | <u>46.4</u>        | 35.9        | 51.5        | 30.1        |

#### 4.2 推理能提高 MLLM 的性能吗？

提示技术 [28, 62] 已被证明可以增强多模态大型语言模型在各种推理任务上的表现 [62, 77, 15, 50]。为了研究相似的推理提示是否可以改善跨视角理解，我们从每个子任务中抽取 100 个问题，并将基线零样本选择题表现与 CoT 增强的提示进行比较。根据 [28, 60]，我们在每个提示后附加“让我们一步一步地思考”，并评估四个代表不同架构和规模的模型：LLaVA-OV-7B [31]、Qwen2.5-VL-32B [70]、Qwen2.5-VL-72B [57] 和 GPT-4o [25]。

图 4 显示了所有子任务中的准确性差异 ( $Acc_{CoT} - Acc_{baseline}$ )。CoT 提示在大多任务上降低了性能。特别是对于“人物关系”和“动作关系”子任务，Qwen2.5-VL-72B [57]，本来在没有 CoT 的情况下表现良好，却遭遇了显著的准确性下降 (20 % 和 19 %)。性能下降的一个原因是 EgoExoBench 中的任务要求模型在解释一个视频流、将理解转化为语言并应用于第二个流之间交替。标准的 CoT 提示仅关注文本分解，打乱了这种交错的视觉-语言 workflow，导致性能下降。

图 5 提供了一个来自与 GPT-4o [25] 的自我中心穿戴者识别的具体示例，使用 CoT 提示。模型从第一视角的片段中正确推断出穿戴者站在一个穿蓝色夹克的人旁边 (用绿色标出)。然而，在随后的推理步骤中，它错误地将蓝色夹克的人识别为摄像头穿戴者，而不是旁边的人 (用红色标出)。这一失败说明了纯文本导向的 CoT 推理如何可以打破跨视角推断的连续性，强调了需要新的提示或架构技术来结合视觉和语言推理的必要性。

#### 4.3 MLLMs 能否利用跨视角指导？

为了确定 MLLMs 是否可以利用额外的跨视角上下文，我们通过增加来自不同视角的额外视频来增强动作预测和技能评估任务作为参考。对于每个问题，模型同时接收主要片段 (例如，用于动作预测的自我中心视角) 和参考片段 (例如，他视视角)，然后选择答案。同时，我们通过移除参考片段并测量准确率的变化进行消融实验。结果显示于表 2。在动作预测任务中，提供参考视频稳定地提升了性能。例如，Qwen2.5-VL [57] 实现了 8.2 % 的提升。这一改进证实了跨视角线索在更好理解动作时的实际价值。相反，在技能评估任务中的结果则不太一致。虽然 GPT-4o [25] 和 InternVL3 [79] 显示出小的性能差异，但 Qwen2.5-VL [57]

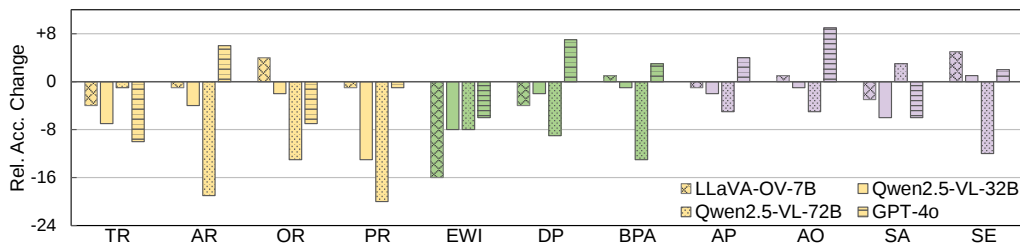


Figure 4: 应用 CoT 后，相对性能变化。平均而言，CoT 导致各项任务的性能下降。这表明仅靠语言推理无法有效解决 EgoExoBench 问题。

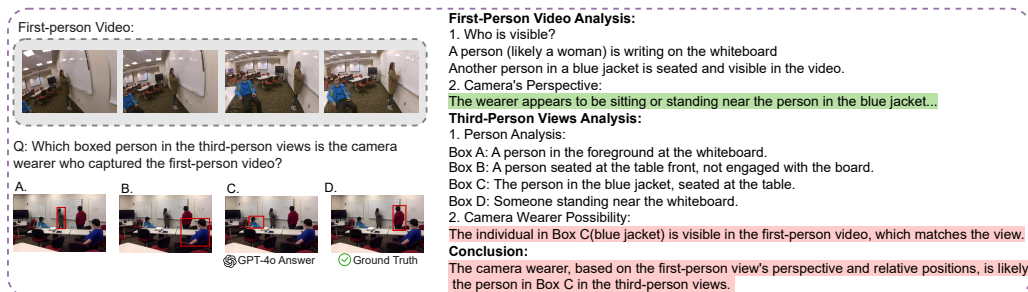


Figure 5: 使用 CoT 提示时 GPT-4o 的一个失败案例。模型未能保持跨视角视觉-语言推理的一致性。

在包含参考视频时表现出 1.5 % 的下降。我们假设该任务对特定领域专业知识和微妙质量判断的依赖，例如评估运动技能熟练度，超过了额外视觉上下文的益处。在当前模型能力下，单靠跨视角输入可能不足以在没有更深入的领域知识整合的情况下增强细粒度的技能评估。

Table 2: 关于使用参考视频的消融研究。

|                | Action Prediction |             | Skill Evaluation |             |
|----------------|-------------------|-------------|------------------|-------------|
|                | w.o./ref          | w./ref      | w.o./ref         | w./ref      |
| Qwen2.5-VL-72B | 31.7              | 39.9 (+8.2) | 31.4             | 29.9 (-1.5) |
| InternVL3-78B  | 41.9              | 51.0 (+9.1) | 30.4             | 29.5 (-0.9) |
| GPT-4o         | 26.1              | 27.7 (+1.1) | 26.5             | 26.3 (-0.2) |

## 5 结论

EgoExoBench 是第一个专门设计用于评估多模态大型语言模型跨视角视频理解的大规模基准。通过整合配对的自中心与外中心记录，并在十一个子任务中设计了超过 7,300 道高质量的多项选择题，EgoExoBench 探索了自中心与外中心推理的三个基本维度：语义对齐、空间对应和序列整合。我们对开源和闭源模型的广泛评估显示，尽管在单视角任务中表现良好，但当前的多模态大型语言模型难以跨越视角之间的鸿沟。思想链提示 (Chain-of-thought prompting) 和额外的跨视角指导仅带来有限的改进，这凸显了需要新颖的架构和训练模式，以便能够在多视角之间交织视觉和语言推理。尽管该基准涵盖多种任务，但它可能无法完全反映真实世界自中心与外中心场景的广度，我们将此作为未来工作的方向。我们希望 EgoExoBench 能够成为一个有价值的资源，以促进对具有人类般跨视角智能的具身代理和协作系统的研究。

## References

- [1] Cihan Acar, Kuluhan Binici, Alp Tekirdağ, and Yan Wu. Visual-policy learning through multi-camera view to single-camera view knowledge distillation for robot manipulation tasks. *IEEE Robotics and Automation Letters*, 9(1):691–698, 2023.

- [2] Anthropic. Claude 3.7 sonnet and claude code, 2025. URL <https://www.anthropic.com/news/claude-3-7-sonnet>.
- [3] Albert Bandura. Observational learning. The international encyclopedia of communication , 2008.
- [4] Siddhant Bansal, Chetan Arora, and C.V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In Eur. Conf. Comput. Vis. , 2022.
- [5] Cristina Becchio, Marco Del Giudice, Olga Dal Monte, Luca Latini-Corazzini, and Lorenzo Pia. In your place: neuropsychological evidence for altercentric remapping in embodied perspective taking. Social cognitive and affective neuroscience , 8(2):165–170, 2013.
- [6] Neil Burgess. Spatial memory: how egocentric and allocentric combine. Trends in cognitive sciences , 10(12):551–557, 2006.
- [7] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyza-guirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language under-standing. Advances in Neural Information Processing Systems , 37:53168–53197, 2024.
- [8] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. arXiv preprint arXiv:2412.12075 , 2024.
- [9] Massimiliano Conson, Elisabetta Mazzarella, Dalila Esposito, Dario Grossi, Nicoletta Marino, Angelo Massagli, and Alessandro Frolli. “put myself into your place” : Embodied simulation and perspective taking in autism spectrum disorders. Autism Research , 8(4):454–466, 2015.
- [10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kaza-kos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In Proceedings of the European conference on computer vision (ECCV) , pages 720–736, 2018.
- [11] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks , 2022.
- [12] DeepMind. Gemini 2.5 pro, 2025. URL <https://deepmind.google/models/gemini/pro/>.
- [13] Arne D Ekstrom, Derek J Huffman, and Michael Starrett. Interacting networks of brain regions underlie human spatial navigation: a review and novel synthesis of the literature. Journal of neurophysiology , 118(6):3328–3344, 2017.
- [14] Chaoyou Fu, Yuhao Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075 , 2024.
- [15] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. arXiv preprint arXiv:2210.00720 , 2022.
- [16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 18995–19012, 2022.
- [17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Under-standing skilled human activity from first-and third-person perspectives. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 19383–19400, 2024.
- [18] Chunhui Gu et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In IEEE Conf. Comput. Vis. Pattern Recog. , pages 6047–6056, 2018. doi: 10.1109/CVPR.2018.00633.
- [19] Ruize Han, Wei Feng, Yujun Zhang, Jiwen Zhao, and Song Wang. Multiple human association and tracking from egocentric and complementary top views. IEEE Transactions on Pattern Analysis and Machine Intelligence , 44(9):5225–5242, 2022. doi: 10.1109/TPAMI.2021.3070562.
- [20] Nicola J Hodges, A Mark Williams, Spencer J Hayes, and Gavin Breslin. What is modelled during observational learning? Journal of sports sciences , 25(5):531–545, 2007.

- [21] Yifei Huang, Minjie Cai, and Yoichi Sato. An ego-vision system for discovering human joint attention. *IEEE Transactions on Human-Machine Systems* , 50(4):306–316, 2020.
- [22] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* , pages 14024–14034, 2020.
- [23] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 22072–22086, 2024.
- [24] Yifei Huang, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Lijin Yang, Xinyuan Chen, Yaohui Wang, Zheng Nie, Jinyao Liu, et al. Vinci: A real-time embodied smart assistant based on egocentric vision-language model. *arXiv preprint arXiv:2412.21080* , 2024.
- [25] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* , 2024.
- [26] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *European Conference on Computer Vision* , pages 767–786. Springer, 2020.
- [27] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems* , 35:3343–3360, 2022.
- [28] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems* , 35:22199–22213, 2022.
- [29] Lukas Kunz, Armin Brandt, Peter C Reinacher, Bernhard P Staresina, Eric T Reifenshtein, Christoph T Weidemann, Nora A Herweg, Ansh Patel, Melina Tsitsiklis, Richard Kempter, et al. A neural code for egocentric spatial maps in the human medial temporal lobe. *Neuron* , 109(17):2781–2796, 2021.
- [30] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Int. Conf. Comput. Vis.* , 2021.
- [31] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* , 2024.
- [32] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272* , 2021.
- [33] KunChang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* , 2023.
- [34] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 22195–22206, 2024.
- [35] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 6943–6953, 2021.
- [36] Aixiu Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* , 2024.
- [37] Yueyue Liu, Zhijun Li, Huaping Liu, and Zhen Kan. Skill transfer learning for autonomous robots and human-robot cooperation: A survey. *Robotics and Autonomous Systems* , 128:103515, 2020.
- [38] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pages 21013–21022, 2022.

- [39] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024.
- [40] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.
- [41] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.
- [42] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 07 2024. URL <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- [43] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023.
- [44] Heqian Qiu, Zhaofeng Shi, Lanxiao Wang, Huiyu Xiong, Xiang Li, and Hongliang Li. EgoME: Follow me via egocentric view in real world. *arXiv preprint arXiv:2501.19061*, 2025.
- [45] Camillo Quattrocchi, Antonino Furnari, Daniele Di Mauro, Mario Valerio Giuffrida, and Giovanni Maria Farinella. Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs, 2023. *arXiv:2312.02638*.
- [46] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021.
- [47] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.
- [48] Richard Ramsey, David M Kaplan, and Emily S Cross. Watch and learn: the cognitive neuroscience of learning from others’ actions. *Trends in neurosciences*, 44(6):478–491, 2021.
- [49] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <https://arxiv.org/abs/1908.10084>.
- [50] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.
- [51] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhanian, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.
- [52] Jinghuan Shang and Michael S Ryoo. Self-supervised disentangled representation learning for third-person imitation learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 214–221. IEEE, 2021.
- [53] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. *Advances in Neural Information Processing Systems*, 32, 2019.
- [54] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018.
- [55] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Adv. Neural Inform. Process. Syst.*, 2024.
- [56] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. *arXiv:1212.0402*.

- [57] Qwen Team. Qwen2.5-v1, January 2025. URL <https://qwenlm.github.io/blog/qwen2.5-v1/>.
- [58] Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Int. Conf. Comput. Vis.*, 2023.
- [59] Weihai Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.
- [60] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- [61] Yeping Wang, Gopika Ajaykumar, and Chien-Ming Huang. See what i see: Enabling user-centric robotic assistance using first-person demonstrations. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 639–648, 2020.
- [62] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [63] Haifeng Xia, Pu Wang, and Zhengming Ding. Incomplete multi-view domain adaptation via channel enhancement and knowledge transfer. In *Eur. Conf. Comput. Vis.*, 2022.
- [64] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.
- [65] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.
- [66] Boshen Xu, Sipeng Zheng, and Qin Jin. Pov: Prompt-oriented view-agnostic learning for egocentric hand-object interaction in the multi-view world. In *ACM Int. Conf. Multimedia*, pages 2807–2816, 2023.
- [67] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024.
- [68] Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. *arXiv preprint arXiv:2504.11732*, 2025.
- [69] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Bain-ing Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022.
- [70] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [71] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024.
- [72] Jingkan Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. EgoLife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*, 2025.
- [73] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.
- [74] Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutso, and Yoichi Sato. Fine-grained affordance annotation for egocentric hand-object interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.
- [75] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

- [76] Ziwei Zhao, Yuchen Wang, and Chuhua Wang. Fusing personal and environmental cues for identification and segmentation of first-person camera wearers in third-person views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 16477–16487, 2024.
- [77] Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. Progressive-hint prompting improves reasoning in large language models. arXiv preprint arXiv:2304.09797 , 2023.
- [78] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv preprint arXiv:2406.04264 , 2024.
- [79] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 , 2025.

## A 附加细节和基准统计

我们首先展示 EgoExoBench 的额外统计数据。接下来，我们详细描述每个任务的 QA 构建过程。最后，我们提供关于人类评估表现的详细信息。

### A.1 基准统计

EgoExoBench 由六个公开可用的自我-外部数据集构建，包括三个任务类别，总共包含 11 个子任务。图 6 总结了每个数据集为每个子任务贡献的 QA 对的数量。

| Ego-Exo Relation                                 |                                |                      |                      |
|--------------------------------------------------|--------------------------------|----------------------|----------------------|
| Task Relation(557)                               | Action Relation(828)           | Object Relation(855) | Person Relation(497) |
| LEMMA: 212<br>EgoExoLearn: 108<br>Ego-Exo4D: 237 | LEMMA: 612<br>EgoExoLearn: 216 | LEMMA: 855           | CVMHAT: 497          |

| Ego-Exo View Transition   |                                       |                                      |
|---------------------------|---------------------------------------|--------------------------------------|
| Direction Prediction(300) | Egocentric Weaer Identification (534) | Body Part Action Understanding (786) |
| Ego-Exo4D: 497            | TF2023: 534                           | Ego-Exo4D: 325<br>EgoExoLearn: 461   |

| Ego-Exo Temporal Reasoning |                   |                          |                                    |
|----------------------------|-------------------|--------------------------|------------------------------------|
| Action Prediction(835)     | Action Order(782) | Sequence Alignment (533) | Skill Evaluation(823)              |
| LEMMA: 420<br>EgoMe: 415   | LEMMA: 782        | Ego-Exo4D: 533           | EgoExoLearn: 444<br>Ego-Exo4D: 379 |

Figure 6: EgoExoBench 中 QA 分布概览。

### A.2 自我-外部关系

自我-外部关系评估模型的能力，以在第一人称和第三人称视角之间关联语义上相似的视觉内容。以下是任务、动作、人物和对象级关系子任务的 QA 构建过程的详细信息。每个子任务的 QA 示例如图 7 和 8 所示。

任务关系。我们从 LEMMA [26]、EgoExoLearn [23] 和 EgoExo4D [17] 数据集中策划了视频。由于 LEMMA [26] 包含涉及多个并发任务的多代理场景，我们仅包括描绘单一任务的视频，以避免在任务识别中产生歧义。对于每个 QA 对，真实的候选视频与查询视频展示相同的任务，但视角不同（例如，第一人称视角与第三人称视角）。为了构建负面候选，我们避免仅基于环境差异的简单区分。特别地，干扰视频从与查询视频相同类别的场景中选择（例如，

均来自厨房环境)。这种设计确保模型必须依赖与任务相关的视觉线索而不是背景差异。总共,我们从 LEMMA [26]、EgoExoLearn [23] 和 Ego-Exo4D [17] 中分别构建了 212、108 和 237 个 QA 对。

动作关系。我们使用来自 LEMMA [26] 和 EgoExoLearn [23] 数据集的视频构建 QA 对。对于 LEMMA [26], 我们利用提供的动作注释。真实标签 (GT) 视频来源于与查询相同来源的视频, 捕捉同一时间段但从不同视角拍摄。负样本候选从同一视频的不同时间段中选择, 分别对应不同的动作。对于 EgoExoLearn [23], 我们使用来自关联基准的 QA 对。GT 视频是直接提供的。为了构建负样本候选, 我们从基准中包含的 20 个候选视频开始。我们首先筛选出与查询共享至少一个动词或名词的候选, 然后使用 SentenceTransformer [49] 计算查询叙述与每个候选叙述之间的相似性。选择三个最相似的候选 (不包括 GT) 作为负选项。这确保了干扰项是相关的, 并保持任务难度。总共, 我们分别从 LEMMA [26] 和 EgoExoLearn [23] 构建了 612 和 216 个 QA 对。

物体关系。我们从 LEMMA [26] 数据集中整理出视频。根据动作注释提取与人物相关的交互物体。与动作关系子任务类似, 对于每个查询视频 (例如, 从第一人称视角), 其真实的候选视频是在对应的第三人称视角下时间对齐的片段。负候选视频是从同一第三人称视角的视频中采样的, 但对应于人物与不同物体交互的不同时间段。

人物关系。我们从 CVMHAT [19] 数据集中构建数据, 该数据集提供了五个户外场景中的同步自我中心和顶视录像。为了指示每个视频中的目标人物, 我们叠加了来自数据集注释的边框。查询视频是由特定个体拍摄自我中心视频片段, 而所有候选视频则来自相应时间戳的顶视帧。候选者之间的区别在于由边框突出显示的人的身份。

### A.3 自我-外部视图转换

自我-外观视图转换中的子任务评估模型在第一人称和第三人称视角之间转换空间信息的能力。图 9 展示了每个子任务的问答示例。以下是我们提供的关于每个子任务问答对构建的详细信息。

自我中心佩戴者识别。我们采用注释派生的策略来构建问答对。我们从 TF2023 [76] 数据集中策划数据, 该数据集提供同步的自我中心-他人视角图像对。给定一个自我中心视频, 目标是在第三人称视角中识别相应的佩戴者。候选者是按照人与人关系子任务中的相同方法构建的。对于每个选项, 将用边界框标记目标人物。为了增加空间复杂性, 我们仅选择包含场景中四个或更多人的样本。

方向预测。我们采用人工标注的策略来构建数据。这个过程包括五个主要步骤: (1) 视频选择。我们从 Ego-Exo4D [17] 数据集中选择同步的多视角视频。(2) 动作片段过滤。我们使用 Ego-Exo4D [17] 中提供的基本描述注释, 并仅保留在第一人称视角中标记为可见的动作片段。然后我们应用 Qwen2.5-32B [70] 来判断动作描述是否包含方向信息 (例如, 向前移动)。只有具有明确方向性内容的片段才会被保留。(3) QA 标注。我们聘请三位学生标注者根据过滤后的片段创建 QA 对。每位标注者可以访问同步的第一人称和第三人称视频, 并被指示关注演员或正在互动的物体的移动方向。如果在第一人称视角中动作不清楚, 或者无法从任一第三人称视角判断移动方向, 则样本被丢弃。要构建问题, 标注者选择在一个第三人称视角中观察到的移动方向, 并撰写一个涉及该方向的问题。动作描述以自由形式的自然语言书写。每个问题设计为在候选答案中有且仅有一个正确答案。(4) 问题润色。我们使用 Qwen2.5-32B [70] 来完善问题, 以确保其清晰性和一致性。(5) 质量保证。我们聘请了一位不参与标注过程的额外学生来审查每个 QA 对。审查者确保每个问题只有一个正确答案, 并且所描述的方向与正确的视频准确对应。

身体部位动作理解。我们采用基于 LLM 的方法来构建问答对。每个问题都来自一对视频, 这些视频至少共享一个动词或对象。我们将两个视频的文本描述输入 DeepSeek-V3 [36], 并提示它生成一个比较动作的问题, 明确关注涉及的身体部位。提示策略如图 10 所示。为了确保这些问题需要视觉信息, 我们进一步使用 Qwen2.5-32B [70] 来筛选那些不需要视觉输入即可回答的问题。筛选提示如图 11 所示。

### A.4 自我-外部时间推理

自我-外部时序推理评估模型在对齐和推断自我中心和外中心视频流事件序列方面的能力。图 16 和图 17 展示了每个子任务的示例问答对。下面, 我们提供了关于每个子任务问答构造过程的进一步细节。

**Ego-Exo Relation**

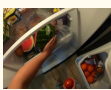
**Task Relation**

**Query Video:**



**Question:**  
Which candidate video most accurately reflects the activities observed in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

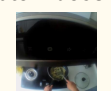
---

**Query Video:**

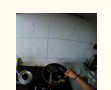


**Question:**  
Which candidate video most accurately reflects the activities observed in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

---

**Query Video:**



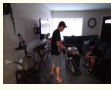
**Question:**  
Which candidate video most accurately reflects the activities observed in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

**Action Relation**

**Query Video:**



**Question:**  
Which candidate video most accurately reflects the actions observed in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

---

**Query Video:**



**Question:**  
Which candidate video most accurately reflects the actions observed in the query video?

**Candidate Videos:**

Video1: 

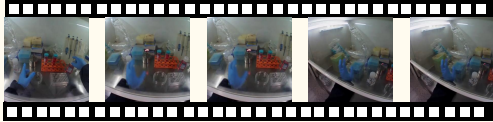
Video2: 

Video3: 

Video4: 

---

**Query Video:**



**Question:**  
Which candidate video most accurately reflects the actions observed in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

Figure 7: 自我-外部关系的例子。正确答案用绿色标出。

**Ego-Exo Relation**

**Object Relation**

**Query Video:**



**Question:**  
Which candidate video most accurately matches the interacted object seen in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

---

**Query Video:**



**Question:**  
Which candidate video most accurately matches the interacted object seen in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

---

**Query Video:**

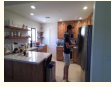


**Question:**  
Which candidate video most accurately matches the interacted object seen in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

**Person Relation**

**Query Video:**



**Question:**  
Which candidate video most accurately matches the person (within the bbox) in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

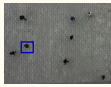
---

**Query Video:**



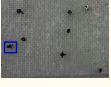
**Question:**  
Which candidate video most accurately matches the person (within the bbox) in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

---

**Query Video:**

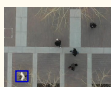


**Question:**  
Which candidate video most accurately matches the person (within the bbox) in the query video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

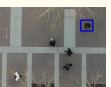
Video4: 

Figure 8: 自我-外部关系的示例。正确答案以绿色突出显示。

**Ego-Exo View Transition**  
**Egocentric Wearer Identification**

**First-person Video:**



**Question:**  
Which boxed person in the third-person views is the camera wearer who captured the first-person video?

**Third-person views:**

A. 

B. 

C. 

D. 

---

**First-person Video:**



**Question:**  
Which boxed person in the third-person views is the camera wearer who captured the first-person video?

**Third-person views:**

A. 

B. 

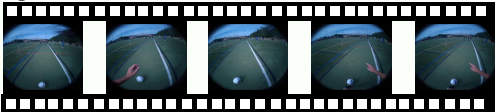
C. 

D. 

---

**Direction Prediction**

**EgoVideo:**



**Question:**  
In which view does the man run diagonally forward towards the left side of the scene?

**Options:**

A. 

B. 

C. 

D. 

---

**EgoVideo:**



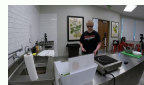
**Question:**  
In which view does the person approach the foreground of the scene?

**Options:**

A. 

B. 

C. 

D. 

---

**Body Part Action Understanding**

**Video1:**



**Video2:**



**Options:**

A. In Video 1, the right hand transfers an object, while in Video 2, it interacts with the object.

B. In both videos, the right hand remains passive and does not interact with the object.

C. In both videos, the right hand is responsible for cutting an object.

D. In Video 1, the right hand stabilizes the object, while in Video 2, it manipulates the object.

---

**Video1:**



**Video2:**



**Options:**

A. In Video 1, the left hand picks up items and the right hand holds them, while in Video 2, both hands perform actions simultaneously.

B. In Video 1, both hands are used for picking up items, while in Video 2, the right hand holds an item.

C. In Video 1, both hands perform actions simultaneously, while in Video 2, the left hand holds an item and the right hand picks it up.

D. In Video 1, the left hand holds items and the right hand picks them up, while in Video 2, the left hand holds an item and the right hand performs an action.

Figure 9: 自我视角到外视角转换的示例。正确答案以绿色突出显示。

You will be provided with two text descriptions of different video scenes. Each description involves a person performing actions that engage specific body parts. Based on these descriptions, you are required to generate a set of question-and-answer (QA) pairs.

Requirements:

1. Each QA pair should include one question, four answer options, and one correct answer.
2. The question must clearly address details from both descriptions, particularly focusing on the roles of the body parts, without directly referencing specific actions or objects to prevent giving away hints.
3. Ensure there is one correct answer, and the other three options should be plausible but incorrect to increase the difficulty.
4. Use 'Video 1' to refer Video 1, use 'Video 2' to refer Video 2.
5. The output must be in JSON format as follows:

```
```json
{
  "Question": "The question text goes here",
  "Options": {
    "A": "Option A text",
    "B": "Option B text",
    "C": "Option C text",
    "D": "Option D text"
  },
  "Answer": "Correct option letter (e.g., 'A')"
}
...
```
```

Example Input:

Video 1: I hold the knife with my left hand and press the garlic with my right hand to slice it.

Video 2: Hold the chili pepper with your left hand and cut it into pieces with a knife in your right hand.

Output Example Based on the Above Input:

```
```json
{
  "Question": "How do the tasks performed by the left and right hands differ in the two videos?",
  "Options": {
    "A": "In Video 1, the left hand holds the knife and the right hand slices; in Video 2, the left hand holds the chili and the right hand cuts.",
    "B": "In Video 1, the left hand presses the garlic and the right hand holds the knife; in Video 2, the left hand cuts the chili and the right hand holds the knife.",
    "C": "In Video 1, the left hand holds the knife and the right hand presses the garlic; in Video 2, the left hand holds the knife and the right hand cuts the chili.",
    "D": "In Video 1, the left hand cuts the garlic and the right hand holds the knife; in Video 2, the left hand holds the chili and the right hand cuts."
  },
  "Answer": "A"
}
...
```
```

Video1: [description]

Video2: [description]

Figure 10: 身体部位动作理解子任务中的问答生成提示。

Carefully read the question and all the options, and based on the information provided, select and provide the most likely correct answer.

[Question]

[Options]

The answer is:

Figure 11: 用于筛选出可以仅使用文本输入正确回答的问答对的提示。

动作预测。我们采用一种基于注释的策略从 LEMMA [26] 和 EgoMe [44] 数据集中构建 QA 数据。对于 LEMMA [26]，我们利用所提供的动作注释。为了生成候选答案，从自我中心序列中选出当前动作周围的十个动作，不包括真实的下一个动作。然后将这些输入到 Qwen2.5-32B [70] 中，选择三个最可能的干扰项。提示策略如图 12 所示。对于 EgoMe [44]，我们使用细粒度的步骤注释。将自我中心序列中的前一个、当前和下一个动作提供给 DeepSeek-V3 [36]，生成三个干扰项。相应的提示如图 13 所示。为了确保数据质量，我们仅保留当前动作在序列中唯一出现的样本，从而消除下一步预测中的歧义。此外，还使用 Qwen2.5-32B [70] 来确认干扰项不包含正确答案。根据身体部位动作理解任务中的方法，我们也应用 Qwen2.5-32B [70] 来移除任何仅使用文本就可以正确回答的 QA 对。

For the current action, which of the following actions is most likely to be the next action?  
Current action: [action name]  
Actions: [action names], [action names], ...  
The answer is:

Figure 12: 从 LEMMA 数据集构建的动作预测问题中生成否定选项的提示。

**Task Description:**  
You are tasked with generating multiple-choice distractor options (negative options) for a next-action prediction task. The goal is to create plausible and reasonable options that could be mistaken for the correct next action. Each distractor must adhere to the following criteria:  
**Logical Consistency:** All options must align with the sequence of actions described in the "previous" and "current" actions. They should follow naturally from the context provided.  
**Reasonableness:** Each option must represent a potential next action that could reasonably occur in the given scenario. Avoid introducing actions that are unrelated, overly complex, or implausible.  
**Non-redundancy:** Distractors must not duplicate the correct answer or each other. Each option should introduce a distinct possibility for the next action.  
**No Obvious Errors:** Ensure that none of the distractors contain logical flaws, contradictions, or actions that clearly violate the context.

**Input Format:**  
You will receive the following inputs:  
**Previous Actions:** A description of the actions that occurred before the current moment.  
**Current Action:** A description of the action currently being performed.  
**Correct Next Action:** The correct next action in the sequence (this will not be included in your output).

**Output Format:**  
Generate 3 plausible distractor options that meet the criteria above. Each option should be concise and clearly describe a potential next action.

**Example Input:**  
**Previous Actions:** "On the white table in front, there are two silver mice and a screwdriver. The left hand picks up the mouse on the left, and the right hand picks up the screwdriver on the right."  
**Current Action:** "The right hand turns the screwdriver to loosen the screws on the mouse."  
**Correct Next Action:** "The right hand rotates the screwdriver to tighten the screws on the mouse."

**Example Output:**  
"The left hand steadies the mouse while the right hand inspects the loosened screws."  
"The person pauses to examine the screwdriver for any signs of damage before continuing."  
"The right hand uses the screwdriver to start adjusting another screw on the mouse."

**Previous Actions:** [action name], [action name], ...  
**Current Action:** [action name]  
**Correct Next Action:** [action name]

Figure 13: 生成用于从 EgoMe 数据集构建的动作预测问题中的负面选项的提示。

动作顺序。我们采用基于 LEMMA [26] 数据集的注释衍生策略来构建 QA 数据。每个问题实例由一对时间上相邻的动作片段组成：一个来自外部视角，另一个来自对应的自我视角。为了确保有效的时间关系，我们用这两个动作的描述来提示 Qwen2.5-32B [70]，以过滤掉不具合理时间顺序的对。提示策略如图 14 所示。

You are an assistant that understands the logical sequence of everyday actions. Please determine whether there is a clear cause-effect or sequential dependency between the following two actions:  
Action1: [action name]  
Action2: [action name]  
Rules:  
Respond with "True" if one action must precede the other.  
Respond with "False" if the actions can be performed in any order or independently.  
The answer is:

Figure 14: 在动作顺序子任务中提示过滤动作对以确保合理的时间关系。

序列比对。我们采用基于 LLM 的策略，基于 Ego-Exo4D [17] 数据集构建 QA 数据。对于每个视频对，我们使用 DeepSeek-V3 [36] 与相应的关键步骤注释提示，生成强调动作序列关键差异的问答。提示策略如图 15 所示。为了质量控制，我们使用 Qwen2.5-32B [70] 根据关键步骤注释验证每个选项，并舍弃有多个有效答案的问答。最后，如同在动作预测子任务中，我们过滤掉仅依靠文本输入即可正确回答的问答。

技能评估。我们采用一种基于注释的策略，使用 EgoExoLearn [23] 和 Ego-Exo4D [17] 数据集构建问答数据。对于 EgoExoLearn [23]，我们使用外部中心演示视频作为参考。然后，我们从成对技能注释中构建传递式排序链以推导候选视频。对于 Ego-Exo4D [17]，参与者的熟练程度被分类为初学者、早期专家、中期专家和晚期专家。我们选择晚期专家视频作为参考，并从同一活动中采样不同熟练程度的候选人。总共，我们从 EgoExoLearn [23] 和 Ego-Exo4D [17] 中分别精选了 444 和 379 对问答对。

为了估计 EgoExoBench 上的人类水平表现，我们从每个子任务中随机抽取 30 个问答，总计 330 个问题。邀请了两名研究生来完成这个子集。为了避免标注偏差，评估者不参与数据构建过程。每个评估者与 MLLMs 接收到相同的输入，包括问题、选项和对应的视频。评估者被指导尽其所能回答所有问题。他们可以暂停、重播，并在没有时间限制的情况下多次观看视频。我们报告评估者的平均准确率作为该评估子集的人类表现基准。

## B 实验

我们首先设计了一个实验，以评估当前 MLLMs 是否能够准确识别多个串联视频输入之间的边界。然后，我们展示了在 EgoExoBench 基准测试上的额外结果，并将模型在 EgoExoBench 上的性能与其他视频理解基准测试上的性能进行比较。最后，我们提供了与思维链 (CoT) 提示策略相关的实验细节和定性示例，以及关于使用参考视频影响的消融研究。

在进行主要实验之前，我们验证了 MLLMs 是否能够将多个视频输入区分为不同的流。这项能力至关重要，因为 EgoExoBench 任务需要同时比较或关联多个视频中的信息。为此，我们设计了一个相同视频对识别任务。每个问题提供五个视频片段，其中两个是完全相同的副本。模型的目标是识别出匹配的对。本任务的一个例子在图 18 中展示。

我们从 Ego-Exo4D [17] 中整理了视频片段，包括以自我为中心和他我为中心的视角。我们构建了共 100 个问题实例用于评估。对于每个视频，我们均匀采样 8 帧作为输入。准确率被用作评估指标，只有当模型准确识别出匹配的视频对时，响应才被视为正确。在随机猜测下，期望的准确率是 10 %。

如表 3 所示，所有评估模型的准确率均超过 77 %，远高于随机猜测的性能。该结果证实了当前的 MLLMs 一般能够区分多个视频输入。这一发现支持了我们主要实验中结果的可行性。

我们还评估了 Gemini 2.5 Pro 在 EgoExoBench 上的表现。如表 4 所示，Gemini 2.5 Pro [12] 实现了最高的整体性能，平均准确率为 51.7 %，超过第二好的模型 GPT-4o-mini [42]，领先 3.7 %。我们还报告了 MLLM 在几个广泛使用的视频理解基准上的表现。如表 5 所示，开

You are a professional test designer tasked with creating a multiple-choice question based on the following two video action sequences.

Design Requirements:

Question: Clearly focus on the differences in the order of actions or the presence/absence of specific actions between Video 1 and Video 2.

Options: Provide four logically clear options (A, B, C, D) directly sourced from the given action sequences. Each option must closely address the action sequences or specific actions within the two videos.

Correct Option: This option should precisely identify one specific difference between the two videos, whether it's about the sequence of actions or the presence/absence of an action.

Incorrect Options: These options must contradict the video content by:

1. Providing an incorrect sequence of actions;
2. Making false statements about the presence or absence of specific actions;

Accuracy: Avoid vague or overly subjective questions. Do not introduce actions or details that are not present in the provided sequences.

Naming Convention: Refer to the first video as "Video 1" and the second video as "Video 2".

The output must be in JSON format as follows:

```
```json
{
  "Question": "The question text goes here",
  "Options": {
    "A": "Option A text",
    "B": "Option B text",
    "C": "Option C text",
    "D": "Option D text"
  },
  "Answer": "Correct option letter (e.g., 'A')"
}
```

Video 1 Action Sequence: []

Video 2 Action Sequence: []

Figure 15: 序列比对子任务中用于生成问答的提示。

Table 3: 模型在初步实验中的表现。

| Model             | Acc. |
|-------------------|------|
| Qwen2.5-VL-7B     | 97   |
| InternVL3-8B      | 85   |
| LLaVA-Video-7B    | 77   |
| NVILA-8B          | 78   |
| EgoGPT            | 79   |
| Claude-3.7-Sonnet | 89   |

源模型如 Qwen2.5VL-72B [57] 和 InternVL3-78B [79] 在某些基准上超过了闭源模型 GPT-4o [25]。

如在主论文中所述，我们评估了链式思维（CoT）提示在基准上的有效性。根据 [71]，我们在每个问题后面附加短语“让我们一步一步地思考。”解码参数设置为温度为 0，top-p 为 1，top-k 为 1。为了从模型的输出中提取最终答案，我们使用 Qwen2.5-32B [70] 作为解析器来明确提取答案。提示设计如图 19 所示。

案例分析。我们展示了链式思维（CoT）推理示例，以说明模型如何解释视频并推断答案。图 20 显示了 GPT-4o [25] 在任务关系子任务中成功的一个案例。模型通过整合环境上下文

**Ego-Exo Temporal Reasoning**  
**Action Prediction**

**Exocentric Video:**



**Egocentric Video:**



**Question:**

Which option most accurately matches the next action expected to occur in the egocentric video?

**Options:**

- A. get controller from sofa with hand
- B. sit on sofa
- C. put controller to sofa with hand
- D. play game-console with controller

---

**Exocentric Video:**



**Egocentric Video:**



**Question:**

Which option most accurately matches the next action expected to occur in the egocentric video?

**Options:**

- A. Lift the top cover of the bag with the left hand
- B. Put the book into the open bag with the right hand
- C. Pick up a book from the shelf with both hands
- D. Use the left hand to shut the upper cover of the bag

---

**Skill Evaluation**

**Reference Video:**



**Question:**

Which candidate video performs the task best according to the reference video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

---

**Reference Video:**



**Question:**

Which candidate video performs the task worst according to the reference video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

---

**Reference Video:**



**Question:**

Which candidate video performs the task best according to the reference video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

Figure 16: 自我与外部时间推理的示例。正确答案用绿色标记。

和个体的关键动作进行推理。它识别出视频中的关键行为，推断出潜在任务，并通过与查询视频比较候选视频来选择答案。图 21 展示了序列对齐子任务中成功的一个示例。模型首先提供每个视频动作序列的逐步描述。值得注意的是，它关注参与者与关键对象（例如试

**Ego-Exo Temporal Reasoning**  
**Action Order**

**Video1:**



**Video2:**



**Question:**  
What is the correct temporal relationship between the actions in these two videos?

- A. Actions in Video 1 occur before actions in Video 2
- B. Actions in Video 2 are part of the sequence of actions in Video 1
- C. Actions in Video 2 occur before actions in Video 1
- D. Actions in Video 1 are part of the sequence of actions in Video 2

---

**Video1:**



**Video2:**



**Question:**  
What is the correct temporal relationship between the actions in these two videos?

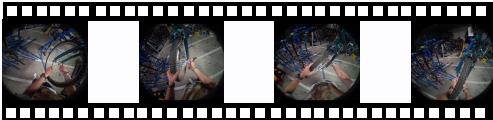
- A. Actions in Video 2 occur before actions in Video 1
- B. Actions in Video 1 are part of the sequence of actions in Video 2
- C. Actions in Video 2 are part of the sequence of actions in Video 1
- D. Actions in Video 1 occur before actions in Video 2

**Sequence Alignment**

**Video1:**



**Video2:**



**Question:**  
Which of the following statements accurately describes the difference in action sequences between Video 1 and Video 2?

- A. Both videos show the same sequence of actions without any differences
- B. Video 1 starts with rolling the wheel, while Video 2 starts with tightening both axle nuts
- C. Video 1 includes tightening the brake cable to the rear axle, while Video 2 does not
- D. Video 2 includes pushing the level inward and turning the axle clockwise, while Video 1 does not

---

**Video1:**



**Video2:**



**Question:**  
Which of the following statements accurately describes the difference in action sequences between Video 1 and Video 2?

- A. Video 1 includes reading the instructions, while Video 2 does not
- B. Video 2 includes arranging test material, while Video 1 does not
- C. Video 1 includes rotating and swirling the swab, while Video 2 does not
- D. Video 2 includes carefully opening the test tube seal paper, while Video 1 does not

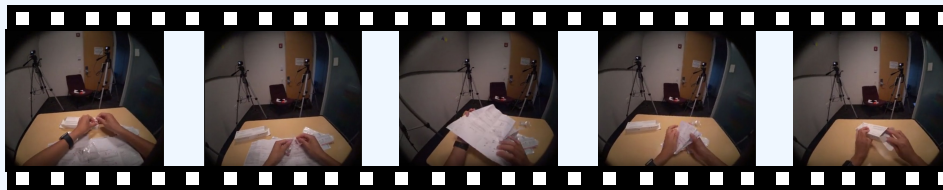
Figure 17: 自我-外部时间推理的示例。正确答案以绿色突出显示。

管、测试板) 互动的顺序。此外, 它能区分细微动作, 例如遮盖和打开。这种对动作语义和时间顺序的详细理解使模型能够有效比较视频之间的序列并推断正确的答案。图 22 展示了 GPT-4o 在身体部位动作理解子任务中的失败案例。虽然 GPT-4o 成功识别出两个视频中的动作, 但它错误地将动作归因于自我中心视频中的右手。相比之下, 它在外中心视频中正

**[Task Instruction]**

You will be provided with 5 videos. Please carefully analyze the content of each video and identify the two videos that are most likely to be identical.

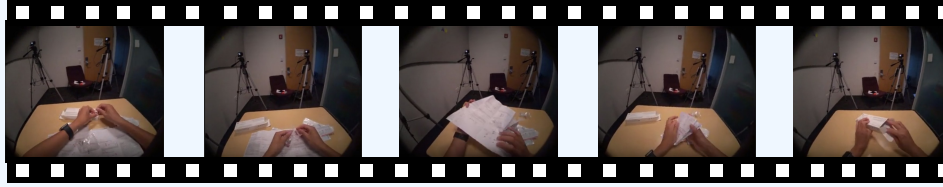
**Video1:**



**Video2:**



**Video3:**



**Video4:**



**Video5:**



**[Response Format]**

Output your answer in the format VideoX-VideoY, indicating the two videos you believe are most similar. The answer is:

**GT: Video1-Video3**

Figure 18: 初步实验中提出的相同视频对识别任务示例。

Given a prediction for a multiple-choice question, directly extract the selected answer while skipping the reasoning process. If the prediction explicitly chooses option (e.g., A, B, C, D), return the corresponding letter. If the prediction does not specify a choice or indicates that none of the options are correct, return None.

[Prediction]

Figure 19: 用于从 CoT 预测中提取最终答案的提示。

Table 4: 各种 MLLM 在 EgoExoBench 上的表现。所有模型中最好的结果用粗体突出显示，第二好的用下划线标出。

|                   | Avg. | Relation    |             |             |             | View Transition |             |             | Temporal Reasoning |             |             |             |
|-------------------|------|-------------|-------------|-------------|-------------|-----------------|-------------|-------------|--------------------|-------------|-------------|-------------|
|                   |      | TR          | AR          | OR          | PR          | EWI             | DP          | BPA         | AP                 | AO          | SA          | SE          |
| Qwen2.5-VL-72B    | 44.7 | 51.0        | 43.5        | 56.6        | 49.7        | 56.7            | <u>37.0</u> | 48.1        | 39.9               | 33.6        | 46.1        | 29.9        |
| InternVL3-78B     | 41.4 | 50.6        | 37.3        | 48.9        | 39.0        | 46.9            | 32.3        | 38.0        | 51.0               | 31.5        | <u>50.6</u> | 29.5        |
| Claude-3.7-Sonnet | 31.3 | 33.4        | 33.0        | 35.1        | 30.6        | 37.8            | 30.0        | 33.3        | 26.0               | 28.5        | 30.2        | 26.4        |
| GPT-4o            | 38.5 | <u>52.9</u> | 44.6        | 57.4        | 48.0        | 41.6            | 24.1        | 37.1        | 27.7               | 26.1        | 37.4        | 26.3        |
| GPT-o4-mini       | 48.0 | <u>65.3</u> | <u>45.2</u> | <u>56.8</u> | <u>53.7</u> | <u>73.8</u>     | 30.0        | 40.0        | 46.4               | 35.9        | 51.5        | <u>30.1</u> |
| Gemini 2.5 Pro    | 51.7 | 63.1        | 52.1        | 67.3        | 56.8        | 76.0            | 38.0        | <u>47.7</u> | <u>50.8</u>        | <u>33.9</u> | 49.3        | <u>33.4</u> |

Table 5: 跨 EgoExoBench 和现有视频理解基准的 MLLMs 性能比较。

|                | Video-MME | MLVU | LongVideoBench | CG-Bench  | Egoschema | EgoExoBench |
|----------------|-----------|------|----------------|-----------|-----------|-------------|
| Qwen2.5-VL-7B  | 65.1/71.6 | 70.2 | 56.0           | -         | 65.0      | 28.8        |
| LLaVA-OV-7B    | 58.2/-    | 64.7 | 56.4           | 31.1/43.2 | 60.1      | 29.9        |
| LLaVA-Video-7B | 46.5/-    | -    | 43.5           | -         | 57.3      | 30.4        |
| InternVL3-8B   | 66.3/68.9 | 71.4 | 58.8           | 38.6/55.2 | -         | 31.9        |
| InternVL3-14B  | 70.4/73.0 | 73.3 | 63.9           | 44.1/60.6 | -         | 35.0        |
| Qwen2.5-VL-72B | 73.3/79.1 | 74.6 | 60.7           | -         | 76.2      | 45.3        |
| InternVL3-78B  | 72.7/75.7 | 79.5 | 65.7           | 48.4/65.3 | -         | 40.5        |
| GPT-4o         | 71.9/77.2 | 64.6 | 66.7           | 41.8/58.3 | 72.2      | 38.5        |

确识别了对应的身体部位。这表明 GPT-4o 在不同视角下仍难以准确定位左右身体部位的动作。图 23 展示了 GPT-4o 在人物关系子任务中的失败案例。模型在推理过程中表现出不一致性。在查询视频分析期间，它推断目标人物与其他人保持一定距离。但随后它选择了一名与其他人位置较近的个体，与其早期观察相矛盾。这种不一致性表明模型在跨视图输入时难以保持推理连贯性。

在主论文中，我们评估模型在动作预测和技能评估子任务中是否能有效利用跨视图信息。为了评估参考视频的贡献，我们通过移除参考视频但保留原始输入视频进行消融研究。为了保证公平比较，我们尽量保持输入提示一致，仅修改必要组件以反映缺少参考视频的情况。其他实验设置保持不变。图 24 和图 25 分别展示了动作预测和技能评估子任务的任务提示。图 26 展示了动作预测子任务的示例，比较了使用和不使用参考视频时 GPT-4o 的预测。图 27 展示了技能评估子任务的类似比较。

**Task Relation**

**Query Video:**



**Question:**  
Which candidate video most accurately reflects the activities observed in the query video? Let's think step by step.

---

 **GPT-4o Prediction:**

To determine which video closely matches the query video, let's analyze the query video step by step:

**Query Video Analysis:**  
 Setting: The video takes place in a kitchen.  
 Actions:  
 One person initially walks across the kitchen.  
 The second person is seen standing, then goes to the refrigerator.  
 The second person opens the refrigerator to take something out.  
 This person then proceeds to the counter and engages in food preparation near or at the stove.

**Comparing Candidate Videos:**  
 Let's examine each candidate video:

**Video 1:**  
 Setting: Different kitchen setup with a dark theme.  
 Actions: Involves actions like taking items from the refrigerator and using the microwave, which is not aligned with the query.

**Video 2:**  
 Setting: Similar kitchen setup to the query.  
 Actions:  
 A person goes to the refrigerator.  
 The person then uses the stove with utensils similar to those in the query, aligned with the sequence of opening the fridge, taking items, and preparing something on the stove.

**Video 3:**  
 Setting: Kitchen setup similar to the query.  
 Actions: Focuses on cutting a watermelon, which is not present in the query video.

**Video 4:**  
 Setting: Wooden kitchen cabinets, different style from query.  
 Actions: Involves using the refrigerator and pouring a drink. Does not involve using stove or food preparation activities.

**Conclusion:**  
 Video 2 fits the actions and setting closest to the query video, with the refrigerator and stove being used in a sequence consistent with the query. Thus, Video 2 most accurately reflects the activities observed in the query video. ✓

**Candidate Videos:**

Video1:



Video2:



Video3:



Video4:



Figure 20: GPT-4o 在任务关系子任务中成功使用链式思维推理的一个案例。模型比较视频中的关键动作以识别匹配的候选者。关键动作作用绿色突出显示。

## Sequence Alignment

### Video1:



### Video2:



### Question:

What is the key difference in the action sequences between Video 1 and Video 2? Let's think step by step.

### Options:

- A. In Video 1, the test tube is covered after unwrapping the testing plate, whereas in Video 2, the test tube is covered before unwrapping the test tube cap.
- B. In Video 1, the test tube is covered before unwrapping the testing plate, whereas in Video 2, the test tube is covered after unwrapping the test tube cap.
- C. In Video 1, the swab is swirled after covering the test tube, whereas in Video 2, the swab is swirled before removing it from the testing tube.
- D. In Video 1, the testing plate is unwrapped before swirling the swab, whereas in Video 2, the testing plate is unwrapped after covering the test tube.

### GPT-4o Prediction:

To determine the key difference in the action sequences, let's analyze each video step-by-step based on the provided videos:

#### Video 1:

1. The person appears to be preparing materials on the table.
2. They seem to be handling a test tube, possibly **doing a covering or sealing action**.
3. There's a focus on opening or handling some instruction or paper.
4. **The testing plate is being unwrapped** and possibly manipulated.

#### Video 2:

1. The person is seen handling materials on a table.
2. They **unwrap the testing plate** and handle the test tube.
3. After unwrapping, they perform actions with the test tube, such as **covering it**.

#### Conclusion:

After reviewing the sequences, Option B correctly describes the actions. ✓

Figure 21: 一个关于序列比对子任务的 GPT-4o 的 CoT 推理成功案例。该模型抓住了与关键对象交互的顺序，并区分了细粒度的动作，从而实现了准确的跨视频序列比较。关键的推理步骤用绿色标出。

### Body Part Action Understanding

#### Video1:



#### Video2:



#### Question:

How do the roles of the left and right hands differ between Video 1 and Video 2? Let's think step by step.

#### Options:

- A. In Video 1, the left hand handles seasoning; in Video 2, the left hand supports the pot and the right hand handles scooping.
- B. In Video 1, the right hand handles seasoning; in Video 2, the right hand handles scooping and the left hand supports the pot.
- C. In Video 1, the right hand handles seasoning and the right hand assists; in Video 2, the right hand supports the pot and the left hand handles scooping,
- D. In Video 1, the left hand handles seasoning; in Video 2, the left hand handles scooping and the right hand supports the pot.



#### GPT-4o Prediction:

To determine the roles of the hands in each video, we analyze their actions separately.

#### Video 1:

The images show someone cooking over a stove. The right hand is adding seasoning from a packet to the pan. There is no clear indication that the left hand is supporting the pot, but it appears to assist during the process.

#### Video 2:

The cook is preparing a dish and serving it onto a plate. The left hand is holding and supporting the pan. The right hand is using a utensil to scoop food onto the plate.

#### Comparison of hand roles:

In Video 1, the right hand handles seasoning, and the left hand assists.

In Video 2, the right hand handles scooping, and the left hand supports the pot.

Based on this analysis, the correct option is: B. ❌

Figure 22: 这是 GPT-4o 在身体部位动作理解子任务中链式推理的一个失败案例。模型在自我中心视角中误识别了活动的手，但在外部视角中正确识别，表明在不同视角中保持左右位置一致性存在挑战。正确预测用绿色标出；错误预测用红色标出。

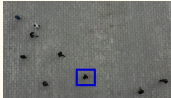
Person Relation

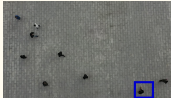
**Query Video:**



**Question:**  
Which candidate video most accurately matches the person (within the bbox) in the query video? Let's think step by step.

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

---

**GPT-4o Prediction:**

To determine which candidate video most accurately matches the person in the query video, let's follow these steps:

- Query Video Analysis:**  
The query video features a person annotated with a bounding box who is dressed in blue and is standing in a more distant position relative to others.
- Candidate Videos Analysis:**

Video 1:  
The person within the bounding box is located in the middle area, positioned closer to others, and is facing towards the lower part of the image.

Video 2: The person within the bounding box is positioned at the bottom of the scene, facing upwards, and is in a different area relative to others.

Video 3: The person within the bounding box is located towards the top left of the image, facing right.

Video 4: The person within the bounding box is near the bottom, facing upwards.
- Matching Characteristics:**  
Considering the horizontal perspective of the query video, the person stands out due to their attire (blue) and position. Video 1's individual matches in positioning, general stance, and isolation relative to other figures more closely than the others. Therefore, Video 1 most accurately matches the person in the query video.

Figure 23: GPT-4o 在人物关系子任务中的一个失败案例。模型推断目标人物在查询视频中与其他人距离较远，但选择了一个与其他人位置较近的候选者，暴露了跨视图推理中的不一致性。正确的推理步骤用绿色突出显示。不一致的推理步骤用红色突出显示。

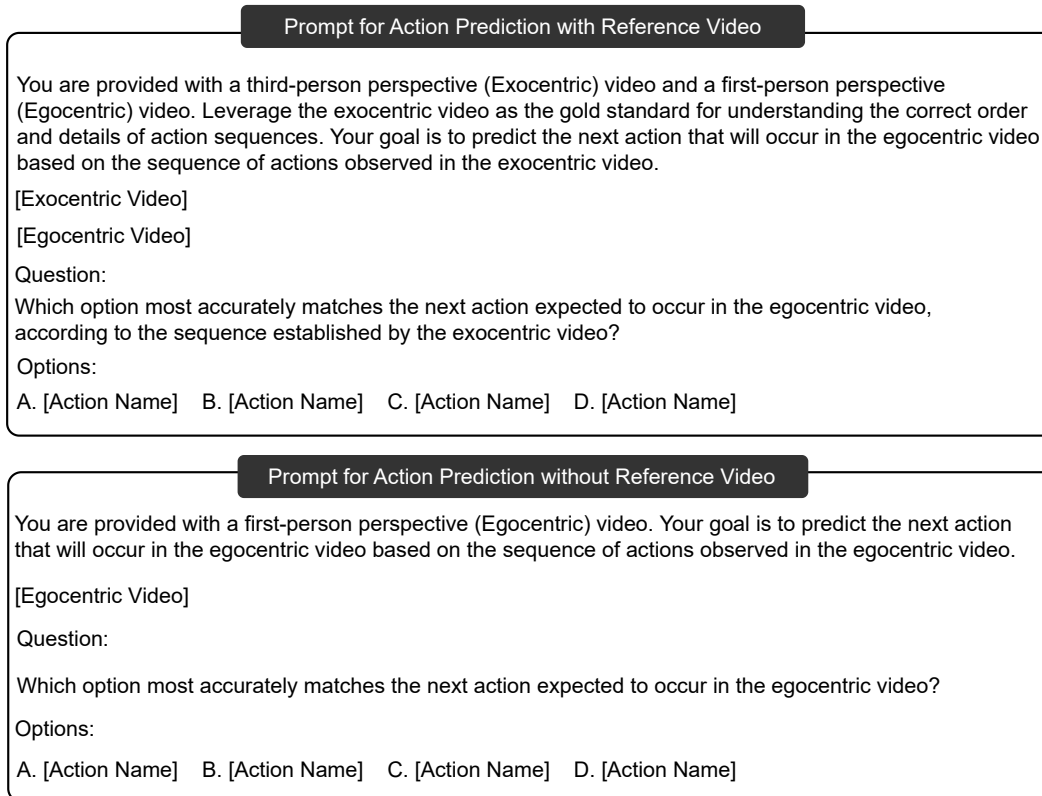


Figure 24: 动作预测子任务的提示，包含参考视频和不包含参考视频的情况。

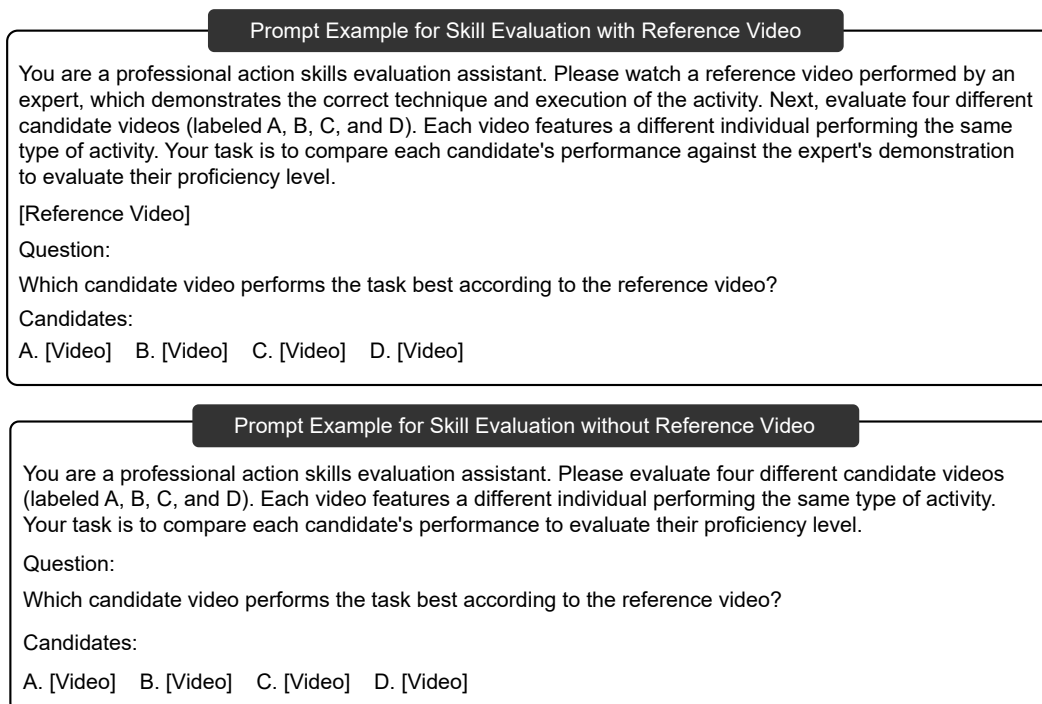


Figure 25: 技能评估子任务的提示，有参考视频和没有参考视频两种情况。

| Action Prediction                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Refrence Video:</b></p>  <p><b>Egocentric Video:</b></p>      | <p><b>Question:</b></p> <p>Which option most accurately matches the next action expected to occur in the egocentric video?</p> <p><b>Options:</b></p> <p>A. close tank <span>🌀 w./ref</span> <span>✓</span></p> <p>B. put tank to table with hand <span>🌀 w.o./ref</span> <span>✗</span></p> <p>C. fill tank with sink <span>🌀 w.o./ref</span> <span>✗</span></p> <p>D. get basin from table with hand <span>🌀 w.o./ref</span> <span>✗</span></p>                              |
| <p><b>Refrence Video:</b></p>  <p><b>Egocentric Video:</b></p>      | <p><b>Question:</b></p> <p>Which option most accurately matches the next action expected to occur in the egocentric video?</p> <p><b>Options:</b></p> <p>A. switch remote <span>🌀 w./ref</span> <span>✓</span></p> <p>B. stand-up <span>🌀 w./ref</span> <span>✓</span></p> <p>C. get remote from table with hand <span>🌀 w.o./ref</span> <span>✗</span></p> <p>D. put remote to table with hand <span>🌀 w.o./ref</span> <span>✗</span></p>                                     |
| <p><b>Refrence Video:</b></p>  <p><b>Egocentric Video:</b></p>   | <p><b>Question:</b></p> <p>Which option most accurately matches the next action expected to occur in the egocentric video?</p> <p><b>Options:</b></p> <p>A. get bottle-water from table with hand <span>🌀 w./ref</span> <span>✓</span></p> <p>B. close juicer <span>🌀 w./ref</span> <span>✓</span></p> <p>C. close bottle-water <span>🌀 w.o./ref</span> <span>✗</span></p> <p>D. open bottle-water <span>🌀 w.o./ref</span> <span>✗</span></p>                                  |
| <p><b>Refrence Video:</b></p>  <p><b>Egocentric Video:</b></p>  | <p><b>Question:</b></p> <p>Which option most accurately matches the next action expected to occur in the egocentric video?</p> <p><b>Options:</b></p> <p>A. fill cup with water-dispenser <span>🌀 w./ref</span> <span>✓</span></p> <p>B. put remote to table with hand <span>🌀 w./ref</span> <span>✓</span></p> <p>C. put cup to table with hand <span>🌀 w.o./ref</span> <span>✗</span></p> <p>D. drink water with cup <span>🌀 w.o./ref</span> <span>✗</span></p>              |
| <p><b>Refrence Video:</b></p>  <p><b>Egocentric Video:</b></p>  | <p><b>Question:</b></p> <p>Which option most accurately matches the next action expected to occur in the egocentric video?</p> <p><b>Options:</b></p> <p>A. point to game-console <span>🌀 w./ref</span> <span>✓</span></p> <p>B. get remote from table with hand <span>🌀 w./ref</span> <span>✓</span></p> <p>C. put controller to person with hand <span>🌀 w./ref</span> <span>✓</span></p> <p>D. play game-console with controller <span>🌀 w.o./ref</span> <span>✗</span></p> |

Figure 26: GPT-4o 在有和没有参考视频情况下对动作预测子任务的预测。正确的预测用绿色标出；错误的则用红色标出。

## Skill Evaluation

**Reference Video:**



**Question:**

which candidate video performs the task best?

**Candidates:**



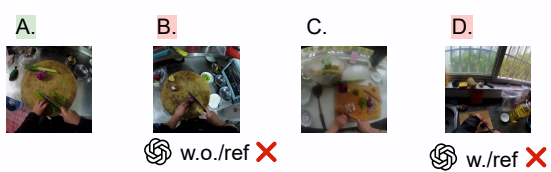
**Reference Video:**



**Question:**

which candidate video performs the task best?

**Candidates:**



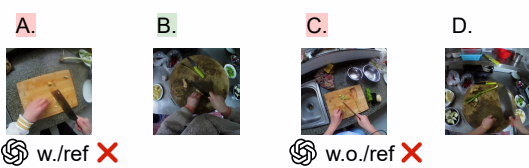
**Reference Video:**



**Question:**

which candidate video performs the task best?

**Candidates:**



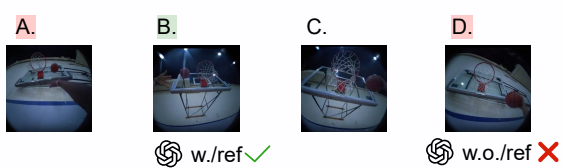
**Reference Video:**



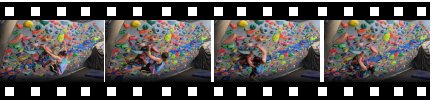
**Question:**

which candidate video performs the task worst?

**Candidates:**



**Reference Video:**



**Question:**

which candidate video performs the task worst?

**Candidates:**

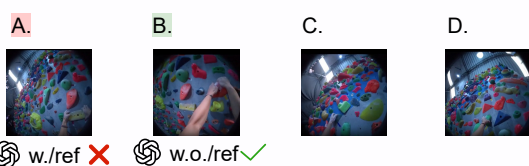


Figure 27: GPT-4o 在有和没有参考视频的情况下对技能评估子任务的预测。正确的预测用绿色标出；错误的用红色标出。