arXiv:2601.10228v1 [cs.CV] 15 Jan 2026

面向第一人称视频理解的多模态大语言模型优化：HD-EPIC VQA 挑战赛解决方案

Sicheng Yang*、Yukai Huang*、孙世通、Weitong Cai、邓建康、宋继飞、Zhensong Zhan g

摘要

多模态大语言模型（*MLLM*）在 *HD-EPIC* 视觉问答等复杂视频问答基准上表现不佳，原因包括查询/选项含糊不清、长程时间推理能力较弱，以及输出格式不统一。我们提出了一种框架，整合了查询/选项预处理、针对特定领域的 *Qwen2.5-VL* 微调、新颖的时间链式思维（*T-CoT*）提示方法以实现多步推理，以及稳健的后处理。该系统在 *HD-EPIC* 视觉问答上取得了 *41.6%* 的准确率，凸显了在要求严苛的视频理解任务中进行全流程优化的必要性。我们的代码和微调模型可在此处获取。

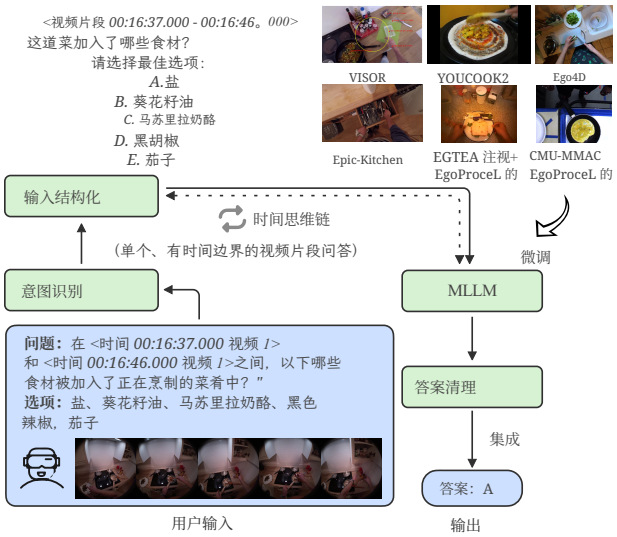

图 1. 所提出的视觉问答系统概览。

1. 引言

视觉问答（VQA）是一项视频理解任务，旨在研究如何根据视频的时间视觉信息回答有关视频内容的问题 [11, 17]。与针对短时、静态或第三人称视角视频的视觉问答相比，为长视频实现自我中心视觉问答 [13, 16–18] 更具挑战性，因为这需要在较长时间跨度内进行稳健的时序推理、推断人类意图、考虑时间信息和历史记忆，并针对细致入微的查询进行复杂的多步推理。

为严格评估模型在这些特定且具有挑战性的自我中心视觉问答任务中的表现，我们聚焦于 HD-EPIC 数据集 [17] 中的视觉问答基准。HD-EPIC 提供了内容丰富、细节高度详尽的长篇第一人称视角视频，记录了复杂的厨房活动，因此其视觉问答基准是检验上述挑战的重要测试平台。该基准覆盖面广，包含 30 种不同的问题原型，可生成 2.6 万个问题，专门用于测试复杂的时序推理、物体交互和细粒度动作理解；这些能力对于评估该领域的模型至关重要。

尽管许多多模态大语言模型（MLLM）在通用视觉问答基准测试中表现出色 [3], 其中包括备受瞩目的闭源模型，如 Gemini 2.5 Pro [12], GPT-4o[14], 和 Grok-3 [22], , 以及 DeepSeek-v3 [9], 、InternVL3 [25], 和 Qwen2.5-VL [1], 等知名开源模型，但它们在一些特定且具有挑战性的任务上往往表现有限，这些任务涉及第一人称视角、复杂的厨房场景、较长的时间上下文以及需要复杂推理的问题。

在这项工作中，我们采用 Qwen-2.5 VL-7B [1]作为基础模型，分析了如图所示、对提升该任务性能至关重要的若干因素1：(1) 仔细进行数据预处理和优化，使用户意图与多模态大语言模型的理解更好地对齐。(2) 使用大量且密集标注的自我中心厨房视频数据进行微调。(3) 针对时间类问题，尤其是长视频和关键帧，实现两阶段推理流程，使多模态大语言模型能够迭代处理并进行推断，而非一次性输出。(4) 采用稳健的后处理，包括处理异常输出、清理结果，以及整合多个提示的回答。这些措施共同促使我们的模型性能超越了以往的基线方法。

2. 方法与实验

2.1. 问题意图识别与澄清

为了提升多模态大语言模型在 HD-EPIC 视觉问答等专业基准上的表现，对查询和选项进行细致的预处理至关重要。针对这套包含 30 道题的题型中固有的歧义和多样的格式，我们采用多方面策略来识别和澄清意图。首先，我们对每个查询进行严格的输入模态分析，将其视觉上下文划分为以下 4 种类型：(1) 单张静态图像（通常是关键帧，例如用于三维感知和器具定位）。(2) 多张静态图像（例如用于营养图像营养估算等比较任务）。(3) 单个有明确时间范围的视频片段（例如用于细粒度动作识别）。(4) 多个不同的视频片段（例如用于食谱准备定位）。这种分类对于指导后续针对多模态大语言模型的处理至关重要。

其次，我们实施了针对任务的提示优化。我们有策略地将原始问题转换为更清晰、结构化的形式，以便多模态大语言模型更好地理解。通过基于正则表达式的解析，我们提取核心实体、时间标记和关系约束。随后，我们将这些信息重新整合为改进后的提示，例如，将简洁的查询（如“X 在哪里？”）改写为更明确、以视角为依据的问题（如“根据这张展示我当前视角的图像，确定 X 的方向。”）。

第三，基于多模态大语言模型对选项呈现方式高度敏感这一观察，[17], 我们从以下几个方面规范并优化多项选择选项的结构。将数字编号改为字母编号（如 A.、B.），使准确率初步 +1提升了 6%。进一步微调格式——例如调整选项间距或使用分号——分别使准确率提高了 +1.8% 和+2.0%。尤为重要的是，为每个选项清晰地换行分隔（\n）带来了显著的 +2.4% 提升。对于包含多个时间片段的复杂时间选项，我们也会重新排版以提高清晰度（例如，A. [V1] HH:MM:SS.sss - HH:MM:SS.sss）。这些改进配合统一的分隔符，降低了解析歧义 [23]。

模型	食谱	食材	营养	动作		3D	运动	注视	平均值
VideoLlama 2 [17]	30.8	25.7	32.7	27.2	25.7	28.5	21.2	27.4	
LongVA [17]	29.6	30.8	33.7	30.7	32.9	22.7	24.5	29.3	
LLaVA-Video [17]	36.3	33.5	**38.7**	**43.0**	27.3	18.9	29.3	32.4	
Gemini Pro [17]	60.5	**46.2**	34.7	39.6	32.5	20.8	28.7	37.6	
Qwen2.5 VL 7B In. [1]	40.6	35.8	32.0	37.3	35.0	23.9	29.7	33.5	
Qwen2.5 VL 32B In. [1]	59.0	37.0	33.0	40.3	35.6	19.8	**33.6**	36.9	
Ours	**64.8**	43.3	37.0	42.0	**40.9**	**29.9**	33.0	**41.6**	
未进行预处理	62.0	40.3	35.7	39.0	35.4	25.0	29.4	38.1	
未进行微调 / T-CoT	62.6	41.5	32.0	38.3	35.2	23.5	29.1	37.5	
未进行后处理	65.0	43.0	35.0	40.3	38.4	26.5	31.6	40.0	

表 1. 各类别的视觉问答结果（准确率 %）。‘w/o’ 是消融研究中 ‘without’（不含）的缩写。

这些预处理提升了多模态大语言模型对任务的理解——通过降低认知负荷并使输入结构与模型的处理优势相匹配，进一步提升了模型对任务的理解。如表 1所示，仅这些有针对性的改进就带来了显著的准确率提升（+3.5%），凸显了稳健的输入工程是发挥模型高级能力的重要前提。

2.2. 模型微调与基于时间域的思考

鉴于 HD-EPIC [17]是一个聚焦于厨房场景的第一人称视频数据集，领域特定微调对于获得最佳性能至关重要 [16]。我们在多种自我中心视角厨房视频数据集上对 Qwen2.5-VL-7B-Instruct [1] 模型进行了微调，这些数据集包括 EPIC-KITCHENS [4–6], CMU-MMAC [8]、EgoProceL [2] 的 EGTEA Gaze+[15] 子集、YOUCOOK 2 [24], VISOR[7], ， 以及 Ego4D [13] 中与物体交互和程序理解相关的部分。微调配置为学习率 2×10^{-7}、批量大小 2、1 步梯度累积和 1 个训练轮次。仅对大语言模型组件进行了调优，同时冻结视觉塔和 MLP 投影器。我们采用 bfloat16 精度、AdamW 优化器，并将最大序列长度设为 131072 个词元。视频处理时，每个样本最多使用 768 帧（最少 4 帧），并根据每个视频动态调整总像素数（3136 至 846720）。

尽管经过微调，Qwen2.5-VL 与许多多模态大语言模型一样，在理解长期时间关系方面仍面临挑战。例如，在多步定位任务中，我们微调后的模型准确率为 26%（微调前为 22%），远低于 Gemini Pro 的 88%。在步骤定位（25% vs. 70%）和粗略步骤定位（28% vs. 74%）等任务中也出现了类似差距，食材顺序任务亦是如此。我们推测，这在一定程度上源于模型的训练序列长度（8192/32768）以及其视频处理策略：分析的帧数上限为 768（视频词元总数 ≤ 24576）。对于时长为 >12 分钟的视频（以 1 FPS 处理时为 >720 帧），动态分辨率缩放或词元限制可能导致有效输入帧不足，无法进行细粒度的时间分析。

为解决这些时间方面的局限，我们设计了一种时间链式思维（T-CoT）提示策略[20, 21], ，引导多模态大语言模型通过中间推理步骤，筛选并理解相关的时间背景，而非直接提出复杂的时间查询。该策略包括：（1）明确利用时间线索：对于包含局部视觉信息或指定时间点/片段的任务（例如，带边界框（BBOX）的三维感知、带时间片段的注视），我们会先处理这些线索。对于 BBOX 信息，我们通过提示多模态大语言模型为 BBOX 内的对象生成一个名词短语来解析，随后

| | | |

详尽短视频片段的效果不如数量较少、时长较长的片段）。针对这些更大的 32B 或 72B 模型进行微调，被认为是更有前景的途径。尽管基线性能显著提升，我们的多阶段架构仍会增加延迟，造成性能与效率之间的关键权衡，尤其是在实时应用中。未来工作必须优化这种平衡。此外，在实现人类水平认知方面仍存在显著差距 [19]，，尤其是在需要对长视频进行深度推理和稳健长期记忆的任务中，这也凸显了一条重要的研究方向。

总之，我们提出了一套全面的方法，证明其能够提升多模态大语言模型在复杂第一人称视频理解任务中的效能。我们的结果凸显了结构化输入/输出处理、领域自适应和引导式时间推理的重要性。尽管计算效率和实现人类水平认知方面仍面临持续挑战，我们的工作仍为未来开发面向第一人称视频分析的先进实用人工智能系统提供了可靠的基线和关键见解。

参考文献

[1] Shuai Bai、Keqin Chen、Xuejing Liu 等。Qwen2.5-vl 技术报告。CoRR，abs/2502.13923，2025。1、2、3[2] Siddhant Bansal、Chetan Arora 和 C. V. Jawahar。我的视角就是最佳视角：从第一人称视角视频中学习流程。载于《计算机视觉—ECCV 2022》第十七届欧洲计算机视觉会议，以色列特拉维夫，2022年10月23日至27日，论文集，第XIII部分，第657–675页。施普林格，2022年。2[3] Wei-Lin Chiang、Lianmin Zheng、Ying Sheng 等。Chatbot arena：通过人类偏好评估大语言模型的开放平台。载于第四十一届国际机器学习大会，ICML 2024。OpenReview.net，2024年。1[4] Dima Damen、Hazel Doughty、Giovanni Maria Farinella 等。扩展第一人称视觉：EPIC-KITCHENS 数据集。CoRR，abs/1804.02748，2018年。2[5] Dima Damen、Hazel Doughty、Giovanni Maria Farinella 等。EPIC-KITCHENS 数据集：收集、挑战与基线。IEEE《模式分析与机器智能汇刊》，43（11）：4125–4141，2021年。[6] Dima Damen、Hazel Doughty、Giovanni Maria Farinella 等。重新扩展第一人称视觉：EPIC-KITCHENS-100 的收集、流程与挑战。《国际计算机视觉期刊》，130(1)：33–55，2022年。2[7] Ahmad Darkhalil、Dandan Shan、Bin Zhu 等。EPIC-KITCHENS VISOR 评测基准：视频分割与物体关系。载于《神经信息处理系统进展35：NeurIPS 会议》，2022年。2[8] Fernando De la Torre、Jessica Hodgins、Adam Bargteil、Xavier Martin、Justin Macey、Alex Collado 和 Pep Beltran。卡内基梅隆大学多模态活动（CMU-MMAC）数据库指南。机器人研究所，2009年。2[9] DeepSeek-AI、Aixin Liu、Bei Feng 等。DeepSeek-V3 技术报告。CoRR，abs/2412.19437，2024年。1

[10] Yicheng Di、Hongjian Shi、Xiaoming Wang 等。基于扩散增强和引导式去噪的联邦推荐系统。ACM《信息系统汇刊》，43：36，2025年，2、31、1、31、3（）：：–：。[11] Chaoyou Fu、Yuhan Dai 等。Video-MME：首个全面评估视频中多模态大语言模型的评测基准。CoRR，abs/2405 21075, 2024 1。[12] 2 5

[12] 2405 21075, 2024 1 Google DeepMind。Gemini Pro 预览版模型卡。技术报告，Google，2025年。技术报告（预览版）。1[13] Kristen Grauman、Andrew Westbury、Eugene Byrne 等。Ego4D：全球3,000小时的第一人称视角视频。载于IEEE/CVF计算机视觉与模式识别会议，CVPR 2022，美国路易斯安那州新奥尔良，*18 24, 2022*，18973 18990 2022 1, 2 —第–页。[14] IEEE，。Aaron Hurst、Adam Lerer、Adam P. Goucher 等。GPT-4o 系统卡。CoRR，abs/2410 21276, 2024 1。Yin Li、Miao Liu 和 James M. Rehg。观察者眼中的世界：在第一人称视角视频中联合学习注视与动作。载于《计算机视觉—ECCV 2018》第十五届欧洲计算机视觉会议，德国慕尼黑，2018年9月8日至14日，论文集，第V部分，第639 655 2018 2–页。施普林格，。Baoqi Pei、Guo Chen、Jilan Xu 等。EgoVideo：探索第一人称视角基础模型与下游适配。CoRR，abs/2406.18070，2024年。1、2[17] Toby Perrett、Ahmad Darkhalil、Saptarshi Sinha 等。HD-EPIC：高度详细的第一人称视角视频数据集。CoRR，abs/2502.04144，2025年。1、2、3[18] Heqian Qiu、Zhaofeng Shi、Lanxiao Wang 等。EgoMe：在真实世界中通过第一人称视角跟随我。CoRR，abs/2501.19061，2025年。1[19] Agnese Taluzzi、Davide Gesualdi、Riccardo Santambrogio 等。从像素到图：使用场景图和知识图解决 HD-EPIC VQA 挑战赛。arXiv 预印本 arXiv:2506.08553，2025年。4[20]Chen Wang、Fei Xia、Wenhao Yu 等。模态链：利用视觉语言模型从多模态人类视频中学习操作程序。arXiv 预印本 arXiv:2504.13351，2025年。2[21] Jason Wei、Xuezhi Wang、Dale Schuurmans 等。思维链提示能够激发大语言模型的推理能力。载于《神经信息处理系统进展》*35* 2022 2：NeurIPS 年度会议，[22]。xAI。Grok 3 测试版——推理智能体的时代。https://x.ai/blog/grok-3，2025年。1[23] Sicheng Yang、Yukai Huang、Weitong Cai 等。即插即用澄清器：用于第一人称视角意图消歧的零样本多模态框架。CoRR，abs/2511.08971，2025年。2[24] Luowei Zhou、Chenliang Xu 和 Jason J. Corso。迈向从网络教学视频中自动学习流程。载于第三十二届 AAAI人工智能会议（AAAI）论文集，第7590–7598页。AAAI Press，2018年。2[25] Jinguo Zhu、Weiyun Wang、Zhe Chen 等。InternVL3：探索开源多模态模型的先进训练与测试时方案。arXiv 预印本 arXiv:2504.10479，2025年。1