arXiv:2602.01038v1 [cs.CV] 1 Feb 2026

从视频到对话：面向任务辅助的第一人称视角指令

Lavisha Aggarwal[1],Vikas Bahirwani[1],和 Andrea Colaco[1]

Google，美国

摘要。 从家电维修、烹饪到汽车维护，许多日常任务都需要专家知识，尤其是复杂的多步骤程序。尽管人们对用于增强现实（AR）辅助的人工智能代理兴趣日益浓厚，但由于缺乏基于现实世界任务执行的大规模多模态对话数据集，相关进展仍然有限；造成这一问题的部分原因是人工辅助数据收集成本高昂且组织安排复杂。本文提出一种框架，可将单人教学视频自动转换为双人多模态任务指导对话。该框架基于大语言模型构建的全自动流程，为传统数据收集方法提供了一种可扩展且经济高效的替代方案。利用这一框架，我们推出了 HowToDIV，这是一个多模态数据集，包含 507 段对话、6,636 个问答对和 24 小时的视频，涵盖多个领域。每个会话都由多轮专家-新手互动构成。最后，我们报告了 Gemma-3 和 Qwen 2.5 在 HowToDIV 上的基线结果，为多模态程序性任务辅助提供了初步基准。

关键词： 程序性任务辅助 · 增强现实 · 增强现实眼镜 · 多模态对话数据集 · 自我中心视频 · 协作式对话。

1 引言

通用人工智能代理的最新进展，使其能够在各种日常活动中以新的方式为人类提供帮助。这类活动中有一大类涉及程序性任务[1]，例如家具组装、烹饪和机械维修；这些任务需要执行一系列相互依赖的步骤，才能实现明确的目标。由于步骤繁多、细节精细，且对执行顺序和错误十分敏感，这些任务难以记忆，也难以可靠地执行。增强现实（AR）可穿戴设备日益普及，为开发能够实时提供逐步程序性指导的视觉语言代理带来了良好机遇。在此类场景中，人工智能助手可以观察用户的动作，逐步提供指令，并

[1]程序性任务由旨在达成指定目标的结构化步骤序列构成，允许有限的可接受顺序，同时将其他偏离视为执行错误。

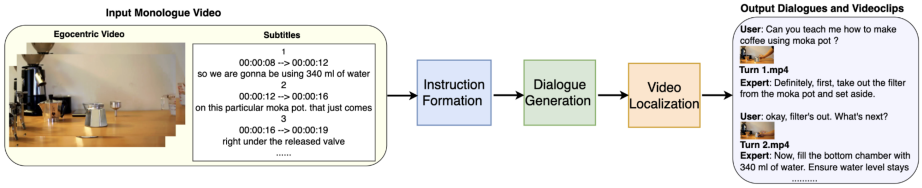


图 1. 所提出的 MDC（独白到对话转换）流程概览。系统处理带有专家字幕的教学独白视频，以推断程序性指令，利用大语言模型生成多轮专家-用户对话，并将视频片段与用户轮次在时间上对齐。

主动检测并纠正执行错误。然而，这一方向的进展受到多模态对话数据集匮乏的限制，因为现有数据集缺乏以现实世界任务执行为基础的内容，这制约了模型训练和代理评估。

捕捉程序性任务中专家-新手交互的数据集对于这一问题设定尤为宝贵。在典型的 AR 任务辅助场景中，配备可穿戴设备的新手用户将视听输入流式传输至视觉语言模型 [14]，该模型必须基于程序性知识进行推理，跟踪任务进度，回答中间问题，并根据观察到的用户行为调整指导。然而，通过人工协助交互收集此类数据集成本高、耗时长且难以扩展，因为这需要在多种任务中进行现场录制、监督和标注。

现有数据集只能部分满足这些要求。例如，HoloAssist [3] 提供了使用 HoloLens 捕捉的、用于协作式物体操作任务的自我中心视频和对话。尽管该数据集具有价值，但其任务种类有限，且需要数百小时的人力投入，因此成本高昂，适用范围也受到限制。更广泛地说，大规模教学视频数据集——例如 NIV [10], EgoPER [11], YouCookII [13], 组装101 [16], CrossTask [17], EPIC-KITCHENS [18], 和 COIN [19]——易于获取，但并不具备对话性。这些数据集通常由旨在离线观看的独白式演示组成，而非用于交互式实时任务辅助，并且缺乏以用户行为为基础的协作式对话。

鉴于人工辅助数据收集成本高昂，而教学视频语料库数量丰富，我们探索了一种自动化替代方案，用于构建程序性任务辅助数据集。我们提出一种基于提示的框架，利用大语言模型将单人教学视频转换为带有轮次级视频对齐的双人专家-新手对话，从而无需人工干预即可大规模构建数据集。

基于这一框架，我们推出了 HowToDIV。这是一个多模态数据集，构建于 NIV [10] 和 EgoPER [11] 数据集之上，并可扩展至其他教学视频来源。HowToDIV 由多轮对话、第一人称视角视频片段以及步骤级程序性指令组成。尽管 HowToDIV 是使用大语言模型自动生成的，但对话合成受到真实教学视频、字幕、一组有序的原子化程序步骤以及

轮次级第一人称视角 grounding 的约束。对于 EgoPER 等数据集，用户错误和专家纠正明确以人工标注的错误标签为条件，从而避免凭空生成动作或指令。

我们使用公开可用的视觉语言模型 [14, 15]，对 HowToDIV 进行基准测试，并通过 BLEU、ROUGE 和大语言模型裁判指标评估性能，为程序性任务辅助建立基线结果。

与以往的教学视频数据集或合成对话语料库不同，HowToDIV 是首个在全自动、可扩展的流程中同时满足以下要求的数据集：(i) 原子化程序步骤，(ii) 多轮教学对话，(iii) 轮次级自我中心视频 grounding，以及 (iv) 对用户错误和专家纠正进行明确建模。我们的贡献[2] 总结如下：

– 我们提出一种全自动的教学独白转对话框架，将专家辅助视频中的领域知识与视觉语言模型的语言推理能力相结合，生成结构化对话，并明确建模用户错误和纠正。– 我们推出 HowToDIV，这是一个多模态对话数据集，包含 507 个会话、6,636 个对话轮次，以及涵盖多样任务的 24 小时对齐视频片段；其中包含不同的用户语音风格、动作准确度和步骤顺序。– 我们使用 Gemma-3 和 Qwen-2.5 建立基线性能，为 HowToDIV 上的多模态任务辅助提供首个基准。

2 相关工作

教学视频数据集。 教学视频为学习复杂的多步骤程序提供了丰富的基础，涵盖烹饪、家庭维护和机械维修等领域。HowTo100M [20]等大规模工作通过收集来自超过 122 万个带旁白的教学视频中的 1.36 亿个片段，涵盖 23,000 多种不同任务，极大地推动了这一领域的发展。HowTo100M 利用自动转录的旁白作为弱监督信息，实现了文本-视频表征的可扩展学习，为规模较小的精选数据集（如 MSR-VTT [21]，、YouCook2 [13], 和 EPIC-KITCHENS [18]）所需的高成本人工标注提供了替代方案。

Charades [1] 是另一个大规模数据集，记录了用户在自己家中进行的日常室内活动。虽然它包含各种各样的人类动作，但其重点是动作描述，而非任务执行，也没有交互式对话的概念。其他指令类数据集，例如 COIN [19]和 CrossTask [17]，将视频与 WikiHow 文章 [12]对齐，以支持对步骤级内容的理解。尽管这些数据集对动作识别、时间分割和程序推理研究发挥了重要作用，但其内容主要是独白式指令，无法建模现实世界任务辅助所需的交互式多轮交流。

[2] 我们的数据集和代码已公开发布于 https://github.com/google/howtodiv。

程序性任务第一人称视频数据集。自我中心视频从第一人称视角记录活动，能够直接呈现用户意图、物体操作和动作执行。这一视角催生了 Ego4D [22],等大规模数据集，该数据集涵盖数千小时的日常活动；此外还有面向特定任务的数据集，例如用于烹饪的 EPIC-KITCHENS [18]和用于物体组装的 Assembly 101 [16]。

较新的数据集着重关注第一人称程序性场景中的人际互动。HoloAssist [3] 引入了一种双人协作场景，其中，佩戴增强现实头显的操作者会接受远程指导者的口头指导。该数据集提供丰富的多模态信号——包括 RGB、深度、音频、注视和手部姿态——以及有关动作、错误和干预的标注，凸显了有据可依的指导的重要性。EgoPER [11]进一步推进了这项研究，提出了第一人称烹饪任务中程序性错误的细粒度分类体系，涵盖遗漏、添加、修改和纠正行为。ALFRED [2] 则提供了交互式环境中的长时序组合式家务任务，重点关注序列规划与执行。尽管这些第一人称数据集各有优势，但它们原生并不支持与分步程序执行相对应的结构化专家—新手对话，而是侧重于独白式演示、动作理解，或孤立的错误检测。

3 HowToDIV 数据集

我们介绍 HowToDIV，这是一个为双人程序性任务辅助而设计的多模态数据集，包含专家—用户对话、步骤级指令，以及按对话轮次级对齐的第一人称视角视频片段。第 3.1 节介绍数据来源和整理流程，第 3.2 节介绍数据集标注和统计，第 4 节则详细说明自动数据生成流程。

3.1 数据整理

HowToDIV 由新手用户与一位专家指导者围绕多步骤程序性任务展开的目标导向多轮对话组成。该数据集包含 507 场会话、6,636 个对话轮次，以及约 24 小时的自我中心视频，通过可穿戴摄像机记录用户视角。数据集涵盖烹饪、机械维修和种植三个领域，共包含九项任务。

每场会话均由用户请求针对某项具体任务的指导开始。专家会逐步给出指令，用户则分阶段执行动作、请求后续步骤，并在需要时提出澄清问题。任务完成后，对话即告结束。每个用户轮次都配有一段时间对齐的视频片段，展示所执行的动作；此外，每场会话还提供一份摘要式程序性指令序列。

HowToDIV 构建于两个公开可用的教学视频数据集之上，这些数据集均包含单人演示：Narrated Instruction Videos (NIV) [10] 和 Egocentric Procedural Error (EgoPER) 数据集 [11]。

指令	用户错误	指令	修改
将茶包放在 mug	把茶包掉在地板	将蜂蜜淋入碗中	倒入糖，而不是蜂蜜
将玉米饼卷成 wrap	折叠的 玉米饼 half	in 用勺子搅拌	用刀搅拌
将滤纸折成半圆	撕破的纸滤器	用刀切玉米饼	用手撕玉米饼
		将玉米饼放在砧板上	将玉米饼放在桌上

表1。 HowToDIV 中（左）用户错误和（右）修改及其对应指令步骤示例。

旁白指令视频 (NIV)。NIV 数据集 [10] 包含 150 个 YouTube 视频，涵盖五项任务，每项任务 30 个视频，平均时长约为 2 分钟（或 4,000 帧）。每个视频都展示了一位专家通过旁白讲解并演示一项完整任务，涵盖室内外环境中复杂的人与物体交互。NIV 提供通过自动语音识别获得并经人工标注员校正的英文转录文本。对于 HowToDIV，我们手动筛选具有第一人称视角或近第一人称视角的视频，以更好地模拟用户佩戴式摄像机的拍摄视角。

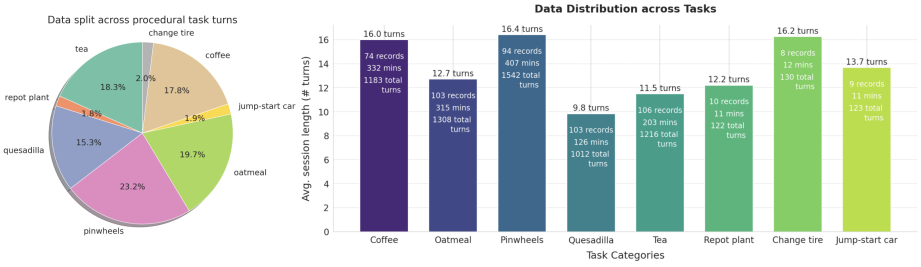


图 2。 HowToDIV 在各种程序性活动中的分布。左：不同任务的用户-专家轮次占比。右：以轮次数量表示的平均对话长度（柱顶）；柱形表示对话数量、视频时长和问答对数量。

第一人称程序性错误（EgoPER）。EgoPER数据集 [11]包含 386 段自我中心程序性任务视频，这些视频由 11 名参与者使用 Microsoft HoloLens 2 录制。数据集涵盖五种烹饪任务的多模态流（RGB、深度、音频、注视和手部跟踪）：风车卷、墨西哥芝士薄饼、燕麦粥、咖啡和茶。与 NIV 不同，EgoPER 不包含口头旁白，而是为每个程序步骤提供带时间戳的细粒度动作标注。数据集包含正确执行的视频（213 段）和错误执行的视频（173 段），总时长约为 28 小时。错误分为步骤遗漏、添加、修改、失误和纠正。对于 HowToDIV，我们保留所有正常视频，并有选择地纳入涉及步骤修改和纠正的错误视频，以符合我们在用户错误后对专家干预进行建模的目标。

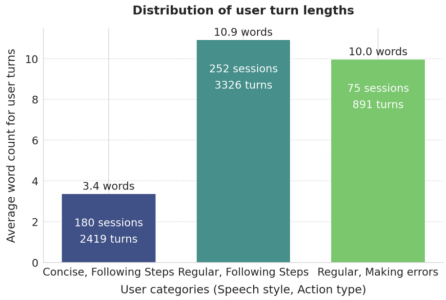


图 3. 不同语音和活动类别下用户对话长度的分布。类别1——简洁型语音与正确动作（平均 3.4 个词）；类别2——常规型（或更冗长）语音与正确动作（平均 10.9 个词）；类别3——常规型语音与用户错误（平均 10 个词）。

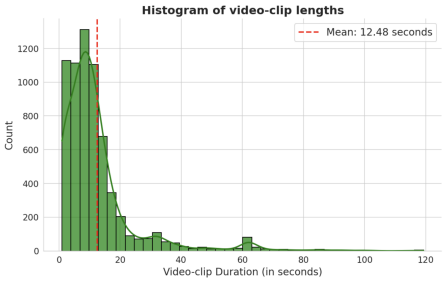


图4. 用户轮次视频长度直方图，平均时长为 12.48 秒。

数据集	领域	对话	时间步骤	叙述	指令	错误	第一人称视角	视频		
NIV [10]	多样	✗	✗	✓	✗	✗	✗	6.66
HowTo100M [20]	多样	✗	✗	✓	✗	✗	✗	134.4k
HoloAssist [3]	组装	✓	✓	✗	✓	✓	✓	2221
EpicTent [4]	帐篷制作	✗	✓	✗	✓	✓	✓	7
EgoPER [11]	烹饪	✗	✓	✗	✓	✓	✓	28
EGTEA [5]	烹饪	✗	✓	✗	✗	✗	✓	28
组装101 [16]	玩具组装	✗	✓	✗	✓	✓	✓	167
COIN [19]	多样	✗	✓	✓	✗	✗	✗	476
CrossTask [17]	多样	✗	✓	✓	✓	✗	✗	376
YouCook2 [13]	烹饪	✗	✓	✓	✓	✗	✗	176
EPIC-KITCHENS [18]	烹饪	✗	✓	✓	✗	✗	✓	100
HowToDIV（我们的）	多样	✓	✓	✓	✓	✓	✓	576

表2。 HowToDIV 与现有程序性任务教学视频和对话数据集的比较。

3.2 标注和统计

各任务类别的会话分布如图 2 所示。每个会话的平均轮次数从较简单任务的 9.8 轮到较复杂任务的 16.4 轮不等，整体平均为 13 轮。

HowToDIV 按用户言语风格和动作类型两个维度进行标注（图 3）。用户对话分为简洁型和常规型两种言语风格。该数据集包含 180 个简洁型会话（用户每轮平均 3.4 个词语）和 327 个常规型会话（用户每轮平均 10.7 个词语）。常规型会话进一步分为按步骤执行（252 个会话）和出现错误（75 个会话）：前者中用户正确执行指令，后者中用户偏离规定流程，随后由专家纠正。表 1 总结了具有代表性的错误类型和步骤修改。

专家回答平均每轮包含 19.3 个词语。由于每轮用户发言都配有一段第一人称视角视频片段，片段时长差异很大，从 1 秒到 2 分钟以上不等，平均时长为 12.5 秒（图 4）。表 2 对比了现有的指导类数据集和自我中心数据集。据我们所知，HowToDIV 是首个同时提供程序性指令、专家—用户对话和第一人称视角视频片段的数据集；它以完全自动化且经济高效的方式构建，适用于基于增强现实的程序性任务辅助领域的可扩展研究。

3.3 数据集质量控制

由于 HowToDIV 是自动生成的，我们同时进行了人工和自动质量检查。我们随机抽取 175 个对话轮次，由两名经过培训的标注员使用二元标签（可用/不可用）评估指令正确性、对话自然度和视频与步骤对齐情况。如果两名标注员都认为某个轮次至少符合两项标准，则将其视为可用；在此标准下，抽样轮次中 **93.2%** 被判定为可用。此外，我们还自动筛除重复项、退化响应、长度异常值、亵渎性语言以及时间定位失败的轮次，剔除的轮次不到最初生成总数的**4%**。尽管 HowToDIV 的规模小于网络规模的教学语料库，但其每个样本的结构密度显著更高，因为每段对话都以第一人称视角视频和程序状态为依据。

4 方法

我们提出了一种基于提示工程的框架，用于将第一人称视角教学视频转换为多轮对话。我们将这一过程形式化为一种算法转换，将单人讲述式教学视频转换为结构化的辅助对话，并为每个对话轮次提供视频依据。我们将这一转换称为独白到对话转换（*MDC*）流程。我们的方法采用一个冻结权重的预训练多模态大语言模型（LLM），并基于两个关键观察：（i）第一人称视角教学视频包含足够的视觉与语言线索，可用于还原完整、逐步展开的任务规范；（ii）结构化的指令集可以系统地转换为交互式教学对话，从而模拟现实世界中的程序性辅助。如图 1 和图 5 所示，我们的流程包含三个阶段：指令构建、对话生成和视频定位。我们首先简要介绍底层语言模型（第 4.1 节），形式化问题设定（第 4.2 节），然后详细说明流程的各个阶段（第 4.3–4.5 节）。

4.1 Gemma 3

Gemma 3 [14] 是一个开放权重多模态模型系列，借助基于 SigLIP 的视觉编码器支持文本和图像输入。它提供较长的上下文窗口（最多 128K 个词元），能够对冗长的转录文本、指令

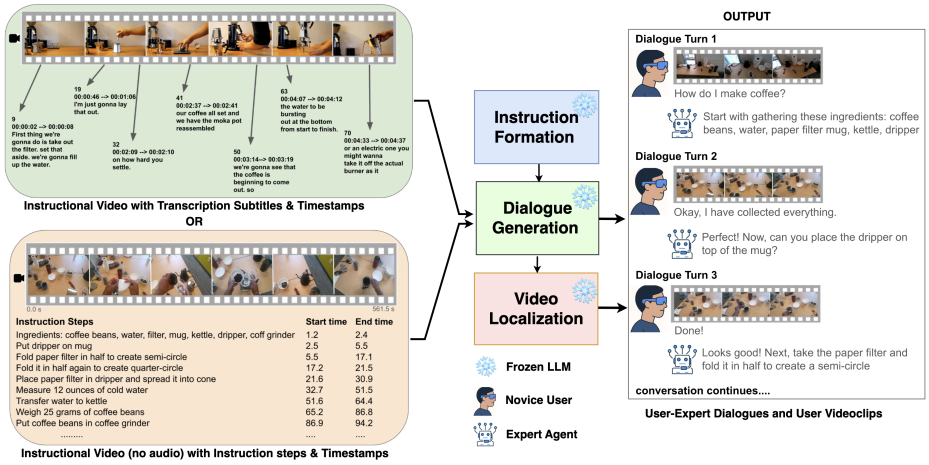


图 5. 我们方法的输入可以是左上：带有旁白的教学视频，或左下：不带旁白的教学视频及步骤详情。它们经三步法处理后，生成用户与专家助手之间的对话（包含第一人称视角视频）。

列表和多图像输入进行推理。指令微调变体具备出色的多模态和多语言性能，同时计算开销较低，因此适合在不进行微调的情况下生成可扩展的数据集。

4.2 问题设定

我们考虑这样一种辅助场景：用户佩戴一台配备摄像头、麦克风和扬声器等传感器的增强现实（AR）设备。通过这一界面，用户可以与人工智能代理交互；该代理能够感知第一人称视角视觉输入，并实时响应口头查询。代理的目标是根据用户的进度，通过逐步发布可执行指令，引导用户完成一项程序性任务。

每一轮中，代理都会收到一条自然语言查询（例如：“我接下来该做什么？”），以及一段记录用户当前状态的短第一人称视角视频片段。代理会回复下一条指令（例如：“往锅里加两杯水”）。构建支持这一设定的数据集，需要包含多轮对话，并确保每一轮都以语言和视频为依据。

我们提出从第一人称视角教学视频中生成此类数据，这些视频通常属于以下类别之一：

1. **带旁白的教学视频**，其中专家在演示任务的同时口头讲解每个步骤。这些视频由一系列 N帧 $V=(I_1,\ I_2,\ \ldots\ldots,\ I_N)$ 和字幕片段 $S=(S_1,\ S_2,\ \ldots\ldots,\ S_M)$ 表示，其中每个 $S_j=(s_j,\ t_{js},\ t_{je})$ 都包含旁白文本 s_j、时间边界 t_{js} 和 t_{je}。

2. **无声演示视频**，其中专家在没有旁白的情况下执行任务，并配有描述动作 q_j 及其时间范围 t_{js} 和 t_{je} 的步骤级标注 $Q_j = (q_j, t_{js}, t_{je})$。

我们的目标是生成一个对话序列 $C = (c_1, c_2, ..., c_P)$，其中每个轮次 $c_i = (u_i, e_i, v_i, t_{is}, t_{ie})$ 都由用户提问 u_i、专家回答e_i、视频 v_i 以及对应视频片段的时间跨度构成。这种表示方式与现代多模态大语言模型在训练和推理期间所需的输入—输出结构相匹配。

4.3 指令构建

第一阶段提取一套完整且有序的任务指令，其中每条指令对应一个原子动作。为确保每个步骤都能在对话轮次中独立执行和提问，我们要求指令具备原子性。

鉴于不同数据集所提供的监督信息各不相同，我们采用针对数据集的提取策略。对于带旁白的教学视频（例如 NIV [10]），字幕转录文本取自现有标注，或通过自动语音识别获取。随后，将这些转录文本交由大语言模型处理，并提示其推断出简洁型、分步骤的指令列表；必要时还会补充隐含的程序性知识。

对于具有细粒度动作标注的数据集（例如 EgoPER [11]），我们会进行后处理，合并重复动作、规范描述并提升语言清晰度。对于其他尚未涵盖的数据集，例如 EGTEA [5], ，可以将语义相似的动作归类，并滤除非必要动作（例如“打开抽屉”“合上盖子”），从而生成精简而完整的指令集。最终得到一组规范的原子指令序列，适用于多种任务领域。

4.4 对话生成

给定有序指令列表后，我们会再次进行大语言模型推理，合成新手与专家之间的多轮对话。假设新手无法预先查看指令列表，而专家则充分了解任务。每条原子指令对应一个对话轮次，包含用户提问和相应的专家回答。

提示词旨在生成简洁型和较长的对话回复。为模拟真实的失败模式，我们引入了 EgoPER 等数据集中的错误标注 [11]，这些标注指出错误步骤及其纠正动作。与错误对应的指令步骤会用特殊词元 <user error> 标记，使大语言模型能够生成识别错误并提供纠正性指导的专家回答。此外，从旁白中提取的程序性约束和细节会转化为澄清问题，从而以领域知识和自然的互动模式丰富对话。

4.5 视频定位

最后阶段会将每条原子指令在时间上与原教学视频中的一个片段对齐，并将该片段重新用作用户执行这一步骤时的第一人称视角视觉输入。定位策略取决于源数据集。

对于具有步骤级时序标注的数据集（例如 EgoPER、EGTEA），我们可以直接使用提供的时间戳；当多个动作合并为一条指令时，再将相应片段聚合起来。对于缺少明确步骤边界的带旁白的教学视频，我们依赖转录过程中生成的字幕时间戳。在提取指令的同一次大语言模型调用中推断时间边界。尽管更先进的时序动作定位方法可能改善对齐效果，但我们将其整合留待未来工作。在实践中，基于字幕的定位已足以支持我们实验中的轮次级定位。

确定性与随机性。尽管该流程依赖大语言模型推理，其结构仍受固定约束支配。在给定相同转录文本或标注的情况下，指令构建会确定性地产生有序、原子化的步骤序列。对话生成确保步骤与轮次一一对应，随机性仅限于表层语言表达的变化。存在时序标注时，视频定位具有确定性；否则则使用推断出的字幕时间戳。因此，相同的提示会产生一致的流程结构和视频对齐结果，变化仅限于词汇表达。

总体而言，MDC 流水线执行指令规范化、按轮次组织的对话合成和时序定位，从而实现了从教学视频到多模态对话数据的可复现映射。

5 实验

5.1 实现细节

我们使用 Gemma-3 [14] 作为指令提取、对话合成和视频步骤标注的基础语言模型，因为它具有出色的指令遵循能力和长上下文处理能力。除非另有说明，数据集生成期间均采用经过指令微调的 27B 变体。所有实验均在配备 8 个计算节点和 4 个 Dragonfish 芯片的 TPU v3 硬件上，使用全分片数据并行（FSDP）进行训练。

我们按照 70% / 10% / 20% 的比例将数据集划分为训练集、验证集和测试集，并根据任务类型和用户类别进行分层。由此得到 355 个训练会话、44 个验证会话和 108 个测试会话。所有定量结果均在留出的测试集上报告。

对话多样性。为鼓励生成的交互具有多样性，我们采用温度为 1.5 的随机采样，以及 top-$p = 0.9$ 核采样。正如图 5 所示，这种选择带来了多样的用户话语和专家回应风格

我们使用自回归解码为每个对话轮次生成一个回应，不进行重排序。对于 EgoPER 会话，我们为每段对话生成两个版本，分别对应常规型（通常较冗长）和简洁型的用户语音，从而提升语言覆盖度。

推理设置。 为在推理时生成专家回答，我们对 Gemma-3-4B [14] 和 Qwen2.5-VL-7B [15] 进行了基准测试，并评估了两种提示配置：

– **仅历史记录**：模型接收先前的对话，并必须利用其内部知识推断后续步骤。– **历史记录 + 步骤**：将完整的指令列表附加到系统提示中，以便显式推理任务结构。

后者作为上限参考，而前者则反映了更切合实际的部署场景。

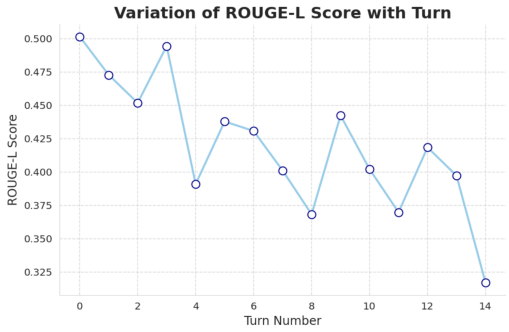


图 6. HowTo-DIV 会话中，随着对话轮次增加，ROUGE-L 平均分数的变化，展示了长时程程序性对话中的累积误差

5.2 评估指标

我们使用三项互补指标评估专家响应生成。

LLM-as-a-Judge。 遵循先前的研究 [8, 9], ，我们采用基于大语言模型的评估方案来评估语义响应质量。向 Gemma-3 12B 模型提供任务名称、用户提问、参考响应和生成输出作为提示；模型会根据正确性、相关性、完整性和有用性给出简短的理由说明，并按照 1-5 分制给出标量分数。为减轻顺序偏差，参考响应和生成输出会以随机顺序呈现。我们对所有测试轮次的分数取平均。虽然基于大语言模型的评估可能会继承模型偏差，但它能够大规模开展开放式程序指导的语义评估，而这类评估难以仅通过词汇指标实现。

ROUGE。 ROUGE [7]用于衡量生成回复与参考回复之间以召回率为导向的词汇重叠。我们报告 ROUGE-1、ROUGE-2 和 ROUGE-L，分别衡量一元词召回率、二元词召回率和最长公共子序列重叠。

BLEU。 BLEU [6] 采用长度惩罚几何平均数，衡量生成回复与参考回复之间的 n 元精确率，得分归一化至 $[0, 1]$。

对于基于参考的指标（BLEU 和 ROUGE），我们在对话轮次级计算分数，并在测试集上取平均值。由于 LLM-as-a-Judge 能够衡量高层次的语义充分性，ROUGE 强调内容覆盖度，而 BLEU 反映局部精确率和流畅度，因此我们报告这三项指标，以全面评估专家回复质量的互补方面。

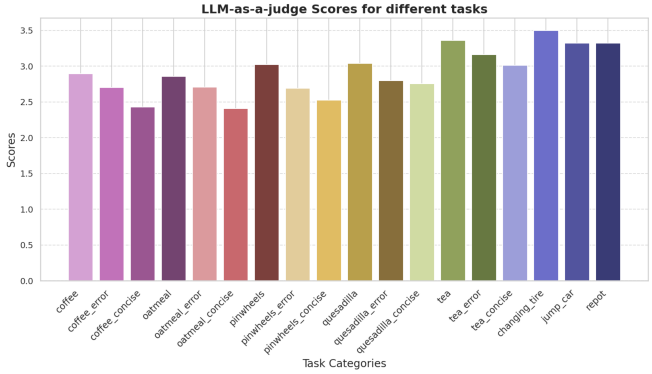


图7。 Gemma-3-4B 专家回答在不同任务、语音风格和动作类别下的 LLM-as-a-Judge 评分变化。

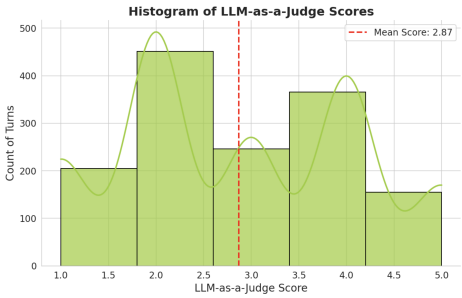


图8。 Gemma-3-4B 回答在 HowToDIV 上使用仅历史提示时的 LLM-as-a-Judge 评分直方图。

5.3 结果与观察

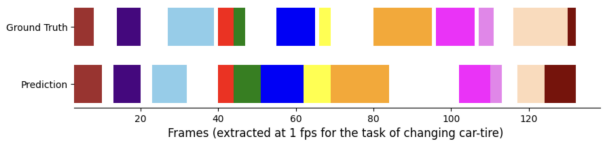


图9。 更换汽车轮胎的时序动作定位；每种颜色表示一个步骤

方法	BLEU	大语言模型评审	ROUGE-1	ROUGE-2	ROUGE-L
Gemma-3-4B（仅历史记录）	0.321	2.870	0.325	0.125	0.270
Gemma-3-4B（历史记录 +步）	**0.457**	4.101	**0.489**	**0.268**	**0.429**
Qwen2.5-VL-7B（仅历史记录）	0.281	2.995	0.310	0.105	0.220
Qwen2.5-VL-7B（历史 + 步骤）	0.441	**4.232**	0.466	0.199	0.356

表 3. 在 HowToDIV 上对 Gemma-3 和 Qwen2.5 配置进行评估。LLM-as-a-Judge 指标使用 12B Gemma-3 变体。

用户类别	BLEU	大语言模型评审	ROUGE-1	ROUGE-2	ROUGE-L
简洁跟随	0.307	2.616	0.309	0.116	0.258
常规跟随	**0.335**	**3.066**	**0.342**	**0.135**	**0.282**
常规-错误	0.304	2.816	0.304	0.109	0.257

表4。 使用仅历史提示配置的 Gemma-3 在 HowToDIV 上跨用户语音风格和动作类型的性能。

表3比较了不同的提示策略和模型。在所有指标上，对于 Gemma-3 和 Qwen-2.5，History +Steps 配置的表现都显著优于仅使用 History 的设置，这凸显了明确任务结构的重要性。由于无法获取步骤级指令，模型只能依赖先验知识，这往往会导致幻觉和领域特定错误。Gemma-3-4B 在 BLEU 和 ROUGE 指标上表现更好，而 Qwen2.5-VL-7B 的 LLM-as-a-Judge 得分略高。

图8显示，使用 Gemma-3 的仅历史设置下，LLM-as-a-Judge 的中位得分为 3.0。如图6所示，随着对话长度增加，性能持续下降，表明累积误差加剧，长时程推理能力降低。

表4和图7分析了不同用户语音风格和动作类型下的性能。在各项指标上，常规语音交互始终优于简洁语音变体。此外，用户遵循指令的对话得分显著高于包含用户错误的对话，反映出纠正性推理的复杂性更高。这一趋势在大多数任务类别中都成立。最后，我们还在图9中展示了时间步定位的质量：针对一段具有代表性的 NIV录音，将预测的动作片段与人工标注进行了比较。

6 结论

我们提出了一种基于提示的框架，可将第一人称独白式教学视频转换为多轮双人对话，以提供程序性任务辅助。通过利用大语言模型的推理与语言生成能力，我们的方法能够合成自然的互动，并以

时间定位的自我中心视频为依据，为成本高昂的现场数据收集和人工标注提供了一种可扩展的替代方案。基于这一框架，我们推出了 HowToDIV，这是一个以自我中心视频为依据的专家—新手程序性对话多模态数据集。它填补了现有资源在面向任务的多模态对话方面的关键空白，并支持对话式教学智能体的基准测试。我们使用 Gemma-3 和 Qwen-2.5 模型，在多项评估指标上建立了基线结果，为未来研究提供参考。我们依赖基于字幕的定位进行轮次级定位，并将学习式动作定位留待未来工作。我们相信，该数据集将推动交互式 AI 智能体的发展，使其能够在实时场景中协助用户完成复杂流程。

参考文献

1. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes：用于活动理解的众包数据收集。在：欧洲计算机视觉会议（ECCV），第510–526页。Springer，Cham（2016） 2. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettle- moyer, L., Fox, D.: Alfred：用于理解日常任务中基于情境的指令的基准。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第10740–10749页（2020） 3. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., Joshi, N., Pollefeys, M.: HoloAssist：面向现实世界交互式人工智能助手的自我中心人类交互数据集。在：IEEE/CVF国际计算机视觉会议（ICCV）论文集，第20270–20281页（2023） 4. Jang, Y., Sullivan, B., Ludwig, C., Gilchrist, I., Damen, D., Mayol-Cuevas, W.: Epic-Tent：用于露营帐篷组装的自我中心视频数据集。在：IEEE/CVF国际计算机视觉会议研讨会（ICCVW）论文集（2019） 5. Li, Y., Liu, M., Rehg, J.M.: 旁观者眼中的世界：第一人称视频中注视与动作的联合学习。在：欧洲计算机视觉会议（ECCV），第619–635页。Springer，Cham（2018） 6. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU：一种机器翻译自动评估方法。在：计算语言学协会（ACL）第40届年会论文集，第311–318页（2002） 7. Lin, C.-Y.: ROUGE：用于摘要自动评估的工具包。在：Text Summarization Branches Out，第74–81页（2004） 8. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.: 使用MT-Bench和Chatbot Arena评判LLM-as-a-Judge。神经信息处理系统进展**36**，46595–46623（2023） 9. Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., Sun, L.: MLLM-as-a-judge：使用视觉—语言基准评估多模态LLM-as-a-judge。在：第41届国际机器学习会议（ICML）论文集（2024） 10. Alayrac, J.-B., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., Lacoste-Julien, S.: 从旁白指令视频中进行无监督学习。在：IEEE计算机视觉与模式识别会议（CVPR）论文集，第4575–4583页（2016）

11. Lee, S.-P., Lu, Z., Zhang, Z., Hoai, M., Elhamifar, E.: 自我中心程序性任务视频中的错误检测。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第18655–18666页（2024） 12. Koupaee, M., Wang, W.Y.: WikiHow：大型文本摘要数据集。arXiv:1810.09305（2018） 13. Zhou, L., Louis, N., Corso, J.J.: 通过损失加权和物体交互从文本中进行弱监督视频物体定位。arXiv:1805.02834（2018） 14. Gemma Tea m, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M.: Gemma 3技术报告。arXiv:2503.19786（2025） 15. Qwen Team: Qwen2.5：基础模型系列。https://qwenlm.github.io/blog/qwen2.5/（2024） 16. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101：用于理解程序性活动的大规模多视角视频数据集。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（2022） 17. Zhukov, D., Alayrac, J.-B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: 从教学视频中进行跨任务弱监督学习。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第3537–3545页（2019） 18. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: 扩展自我中心视觉：EPIC-KITCHENS数据集。在：欧洲计算机视觉会议（ECCV）。Springer，Cham（2018） 19. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: COIN：用于全面教学视频分析的大规模数据集。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第1207–1216页（2019） 20. Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M：通过观看一亿个旁白视频片段学习文本—视频嵌入。在：IEEE/CVF国际计算机视觉会议（ICCV）论文集，第2630–2640页（2019） 21. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT：用于连接视频与语言的大型视频描述数据集。在：IEEE计算机视觉与模式识别会议（CVPR）论文集，第5288–5296页（2016） 22. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X.: Ego4D：全球范围内的3,000小时自我中心视频。在：IEEE计算机视觉与模式识别会议（CVPR）论文集，第18995–19012页（2022）