重新思考视频智能体：视频工具框架何时以及为何有效

LuoyuanZhang*,§ Xuanyu Zheng*,†,‡ Yancheng Long*
Haoyang Xu* Bin Wen† Tingting Gao

KlingAI Research

arXiv:2611.31883v1 [cs.CV] 24 Nov 2026

摘要

视频理解不仅取决于模型如何推理视觉证据，也取决于它们观察哪些证据。固定的观察策略会预先确定这些证据；智能体工具框架则允许模型随着理解不断发展而寻求进一步的观察。这一转变为实现固定输入未能发掘的能力创造了机会，但能够请求观察并不意味着模型能够有效利用这些观察。我们研究视频工具框架在何时以及为何能够改善理解，考察任务需求与模型的交互能力如何共同塑造其收益。我们使用 Video-Relook 作为实验框架，让模型从共同的默认配置出发，自适应地调整时间和空间观察，并分析不同模型和任务中的性能与交互轨迹。自适应观察显著提升了细粒度动作理解，其收益也扩展到了 Video-MME-v2。即使使用相同的观测接口，不同模型获得的增益也存在显著差异；交互轨迹显示，它们在证据搜寻以及答案修订可靠性方面也存在相应差异。对于同一模型，不同任务获得的收益同样各不相同，既有显著改进，也有有限增益和回归。总体而言，这些发现支持这样一种解释：工具框架的有效性取决于信息性观测机会与模型将交互转化为可靠修订的能力是否匹配。这些发现推动我们构建能够协调感知与推理的视频工具框架，使观察自主性与任务需求和模型能力相匹配。

日期： 2026年11月18日**关键词：** 视频理解；智能体工具框架；自适应观察

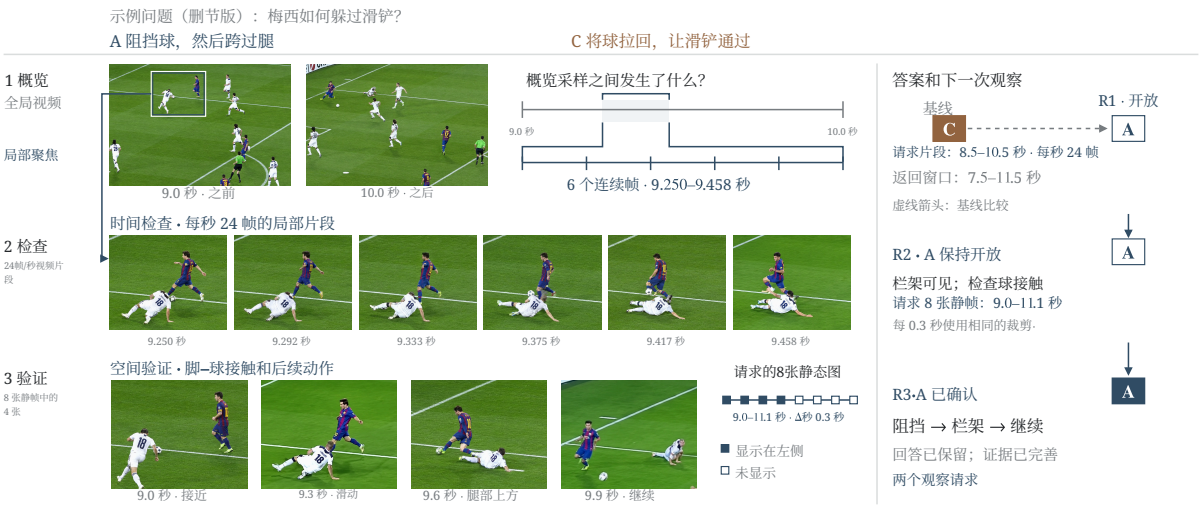

图1。从概览到确认。 该示例说明智能体框架如何帮助模型理解场景细节。

*共同贡献。
†通讯作者。
‡项目负责人。§本项工作由作者在 KlingAI 实习期间完成。

1 引言

智能体范式正在扩展到视觉理解领域，使模型能够参与决定如何获取和处理视觉信息。对于视频任务，这一转变可以通过提取视频片段、裁剪空间区域和选择帧的工具来实现，并配合一个组织这些工具使用的循环。然而，这些组件定义了模型能够做什么，却没有解释为什么这些操作应当改善理解。一个能够调用视觉工具的系统仍然必须决定何时再次观察、哪些观察结果可以消除其不确定性，以及新的证据是否足以促使其修订回答。因此，将智能体范式引入视频领域所提出的问题已经超越了工具是否可用：在什么条件下，对观察的自主控制能够转化为更好的视频理解？

这一问题源于感知与推理之间的相互依赖。模型的判断取决于它能够获得的视觉证据，而推理过程中产生的不确定性则决定了哪些额外证据值得寻找。固定的观察策略会预先限制这一过程：模型必须通过规定的时空采样方案来理解视频，而该方案可能无法足够清晰地呈现相关细节。相较之下，自主观察允许模型根据问题及其不断发展的理解，调整获取证据的方式。我们的出发点是一种常见的默认观察配置，合适的策略可以通过交互逐步形成，从而减少对人工选择的基准测试特定设置的依赖。这为实现固定输入所未能发掘的能力创造了机会，同时也对模型提出了新的要求：它必须识别尚未解决的问题，组织信息性观测，并将返回的信息转化为可靠的判断。

细粒度动作理解提供了一个研究这种关系的具体场景。总体语义相似的动作，可能由于短暂的接触、局部姿势或事件顺序不同而产生不同的回答。识别出一名球员躲开了滑铲，并不能确定他如何控制球，也不能确定他何时跃过防守者的腿。概览可能保留动作的要旨，却无法解决决定回答的细节，这使密集采样的局部片段和空间细节发挥了明确作用。与此同时，更多帧也无法替代对接触关系和动作顺序的正确理解。这一场景同时展现了自主观察的机会及其所需的能力：我们可以考察模型是否会主动寻找相关证据，以及随着观察结果不断累积，其判断是否得到改善。因此，高时间分辨率是使这一问题显现出来的一个条件；而要验证其更广泛的相关性，则需要在证据需求各不相同的视频任务中进行评估。

我们使用 video-relook 作为实验框架来研究这些条件。模型从默认的全局概览出发，根据当前理解请求局部片段、选定的帧或空间细节，接收观察结果，并在有限的多轮对话中修订答案。我们的研究考察两个维度：不同模型使用相同的观测接口和轮次预算，而同一模型则会面对不同类型的视频任务。我们保留首轮预测、答案轨迹、观察请求和停止记录，以追踪增益与损失。

结果首先明确了自适应观察所带来的提升空间。所有五个接受评估的模型在细粒度动作理解方面均取得正向增益，最高达到 14.2 个百分点；在 Video-MME-v2 上，Gemini 3.8 Flash、Gemini 3.6 Flash 和通义千问 Qwen 3.8 Flash 分别提升 3.78、1.91 和 5.25 个百分点，而通义千问 Qwen 3.7 Flash 则没有净增益。然而，较大的基线提升空间并不一定带来更大的增益。通义千问 Qwen 3.7 Flash 留下了最多待纠正的错误，但仅提升 3.0 个百分点，并且在 55.7% 的问题组中于首轮后停止；通义千问 Qwen 3.8 Flash 和 Gemini 3.6 Flash 则分别提升 12.7 和 14.2 个百分点。结合错误恢复和正确答案保持率方面的差异，这些结果表明，必须通过模型的观察和修订行为才能实现现有的提升空间。跨任务结果进一步揭示了这一过程所依赖的条件：在动作问题上获益显著的模型，在其他任务上可能只取得有限增益，甚至表现退步。例如，Gemini 3.6 Flash 在 TempCompass 上下降 1.20 个百分点，但在 Video-MMMU选择题上提升 2.73 个百分点。同一模型—工具框架组合在不同任务中会产生不同结果，这表明观察自主性也取决于进一步观察能否解决任务所呈现的难点。

我们的贡献主要有三点：

1. 我们提出假设：当信息丰富的观察机会与模型获取证据并可靠修正其判断的能力相匹配时，视频工具框架才会发挥作用。2. 我们引入 video-relook，这是一种基于共享默认设置进行自适应时空观察的框架。3. 我们刻画了不同模型和任务中自主观察的优势与局限，增益最高可达

14.2 个百分点。

2 相关工作

视频理解及其证据要求。 动作数据集 [10, 23] 用于区分细粒度运动；NExT-QA [29] 和 MVBench [17] 用于探究时间推理；Video-MME [8] 涵盖不同持续时长和任务。Video-MME-v2 [9]增加了推理层级；Video-MMMU [12]研究从专业视频中学习。Moment- Video [18] 发现，即使短暂事件被覆盖，也会出现错误。

对视频的自适应访问。 主动感知 [3, 4] 和循环注意力 [22] 将观察视为一种决策。VideoAgent [7, 26] 和 VideoTree [27] 根据问题调整访问方式。EVA [37] 通过分阶段微调和强化学习学习时间和空间选择。VideoNexus [20] 通过紧凑的证据交换，将视频感知与规划分离。我们在共享接口下比较固定模型。

用于推理的视觉工具。 VisProg [11] 和 ViperGPT [25] 通过程序组合工具； V^* [28] 引导视觉搜索，而 Visual Sketchpad [13] 支持视觉中间推理。video-relook 在推断过程中提供视频片段、裁剪图和静帧。

评估交互式视频系统。 对于语言和多模态模型而言， [2, 6, 14, 16, 30, 32, 33, 36],ReAct [31] 交替进行推理与行动；Reflexion [24] 和 Self-Refine [21] 利用反馈进行修订。没有外部反馈的自我纠正也可能削弱推理能力 [15]。AgentVidBench [1], 、 Video-Arena [35], 和 VideoGAIA [34] 用于评估交互式视频理解。我们通过观察请求、停止和修订，追踪增益与损失，并区分持续交互与可靠使用新获取的视觉证据。

3 研究框架与设计

video-relook 提供了一个通用的观测接口，用于比较模型和任务。轮次级回答、请求和停止决策，将最终准确率与模型的交互历史联系起来。

3.1 从固定观察到自主观察

单次传递基线根据预先规定的全局概览生成回答。测试框架从同一概览请求开始，并允许模型获取局部的时间和空间观测结果，在同一对话中解读这些结果，然后修订其回答。模型和评分规则保持不变。

这项比较评估的是完整的测试框架——包括提示、观察决策和多轮推理——而未匹配计算量或单独考察视觉输入。我们将新观察之前的修正与之后的修正区分开来
它们的返回结果。



图2。 请求用于选择观察结果；返回循环会将它们附加到同一对话中。参考标签仅交由评分器使用。

3.2 有界观察循环

模型接收概览和问题，生成带有证据注释的初步答案，并可请求观察结果。每个请求都会指定一个问题、工具和参数。执行器将源视频观察结果返回到同一对话中。所有会话评估最多允许三个模型轮次，并且每个生成请求的轮次最多允许六个请求；第三轮返回最终答案。

概览与请求生成。 概览请求以每秒 1 帧进行全局采样；提供方预处理会单独记录，因为 API 不会公开其内部视觉表示。第一轮提示要求先提供证据再回答，标记未解决的问题，并提供根据问题文本推断出的检查清单。参考答案、基准标签和单独的证据标注均不会提供；问题中的时间范围仍然可见。不同的提示可能产生不同的基线答案和第一轮答案。

观察接口。 clip 以最高每秒 24 帧的速率重新查看一个窗口，并可选择空间裁剪；frames 返回带标签的静帧以及可选的放大裁剪图；montage 将候选事件与边界上下文分组，以便进行计数；scan 对更长的时间范围进行粗略搜索；overview 则重新查询全局输入。选定的运动问题还会获得运动能量摘要。实现细节见补充材料。

修订与停止。 后续轮次会追加观察结果，并允许提供回答、提出进一步请求或停止。对于没有获得自身观察结果的问题，已确认答案会予以保留。当无需进一步请求即可获得答案，或达到轮次上限时，循环结束。如果未回答的问题没有获得初始请求，执行器可以检查默认窗口。

3.3 衡量增益并追踪修订

配对回答揭示了准确率提升背后的纠正与回归；首轮预测则能确定这些变化发生在新观察之前还是之后。

令 B_i 和 H_i 分别表示单次传递基线和测试框架对问题 i 的回答是否正确。对于相同的 N个问题，将纠正 定义为 $(B_i, H_i) = (0,1)$，将退步 定义为 $(1,0)$。准确率差恰好为

$$\Delta A = \frac{N_{\text{fix}} - N_{\text{break}}}{N}. \tag{1}$$

为考虑不同的基线准确率，将恢复率 r 定义为基线错误中的恢复比例，并将回退率 d 定义为正确回答中的回退比例：

$$r = \Pr(H = 1 \mid B = 0), \quad d = \Pr(H = 0 \mid B = 1). \tag{2}$$

将基线准确率记为 a，可得

$$\Delta A = (1 - a)r - ad. \tag{3}$$

对于较强的基线，即使回退率较低，也可能抵消高得多的恢复率。

区分修订的各个阶段。令 R_i 表示在首次测试框架轮次之后、返回任何请求的观察结果之前的正确性。于是

$$\Delta A = \underbrace{(A_R - A_B)}_{\text{initial-round change}} + \underbrace{(A_H - A_R)}_{\text{later-round change}}. \tag{4}$$

第一项包含提示和推理变异性；第二项包含新观察、进一步推理和累积上下文。两者都无法单独确定视觉证据的因果效应。我们枚举 (B_i, R_i, H_i)，以区分持续性早期错误与后续反转。未作答的预测均计为错误。

4 VideoHarnesses 何时有效?

我们比较不同任务和模型中的增益，然后考察时间要求和问题类别。

4.1 实验设置

任务。 我们评估 Video-MME-v2 [9], TempCompass [19],Video-Holmes [5], 以及 Video-MMMU 的选择题部分 [12]。为了更深入地研究动作细节，我们使用 FastAct，这是一个基于高质量视频数据构建、经过人工标注和审核的内部基准测试。它测试短暂接触、局部运动和动作顺序。

FastAct 中的动作细节。 FastAct 包含 1,000 道四选项问题，涉及 100 段时长约为 36–60 秒的无声视频片段，涵盖体育运动、格斗、舞蹈和其他动作。问题涉及接触、物体操控、动作顺序和参与者反应。例如，踢桶问题需要区分抬起的脚、支撑腿和踢击脚。问题会指定相关时间范围；标注证据区间的中位时长为 0.90 秒。视频以每秒 24 帧提供，其中 693 道问题被标注为需要每秒 8 帧，307 道被标注为需要每秒 24 帧。给定的时间范围减少了广泛搜索的必要性，使我们能够考察针对性观察是否可以厘清概览中仍然模糊的细节。

模型与比较。 FastAct涵盖 Qwen 3.7 Flash、Qwen 3.8 Flash、Gemini 3.6 Flash、Gemini 3.8 Flash 和 Gemini 4 Argon；公开基准涵盖四个 Flash 模型。每个模型都与其单次传递基线进行比较，同时固定权重、选项和评分方式。两组均请求获取 1 帧/秒的概览；测试框架额外加入局部观察。Gemini 使用 HIGH 思考模式；通义千问使用网关默认设置。

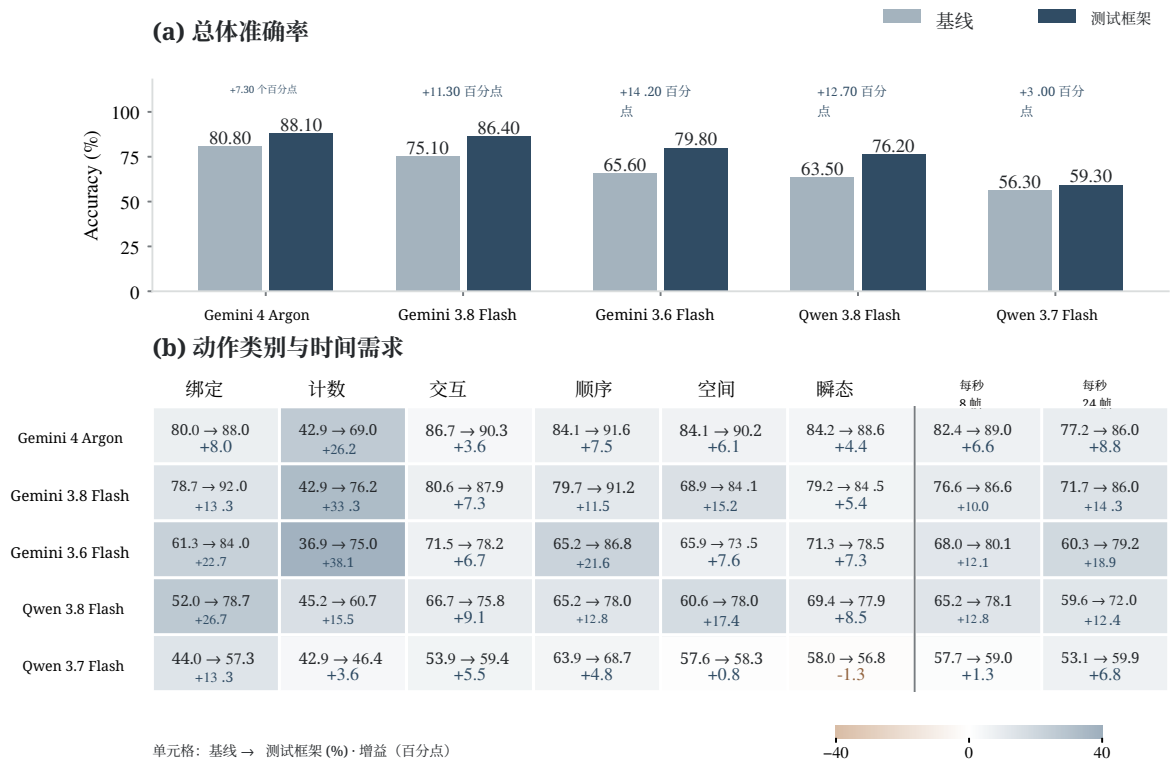

图 3。 FastAct 的总体结果和子组结果。配对柱状图比较基线与测试框架的准确率；单元格报告配对准确率和增益，阴影表示增益大小。

4.2 一种揭示自适应观察的设置

概览可能会遗漏决定性细节，或者模型可能会误解可用证据。细粒度动作会暴露这两种困难：短暂的接触或肢体位置的变化，就能在一项易于识别的活动中区分不同回答。

FastAct 提供相关时间范围，但将动作细节留待进一步解析。密集的视频片段和空间裁剪允许进行检查；模型必须决定请求什么，以及证据是否足以支持修订。这既提供了找回被忽略细节的机会，也要求模型对观察进行引导和解读。

公开基准显示了这一收益的局限性（图4）：Gemini 3.6 Flash 在 Video-MMMU 上提升了 2.73 分，但在 TempCompass 上下降了 1.20 分；Qwen 3.8 Flash 在 Video-MME-v2 上提升了 5.25 分，但在 Video-Holmes 上下降了 0.33 分。

W我们考察相关的交互行为 low.

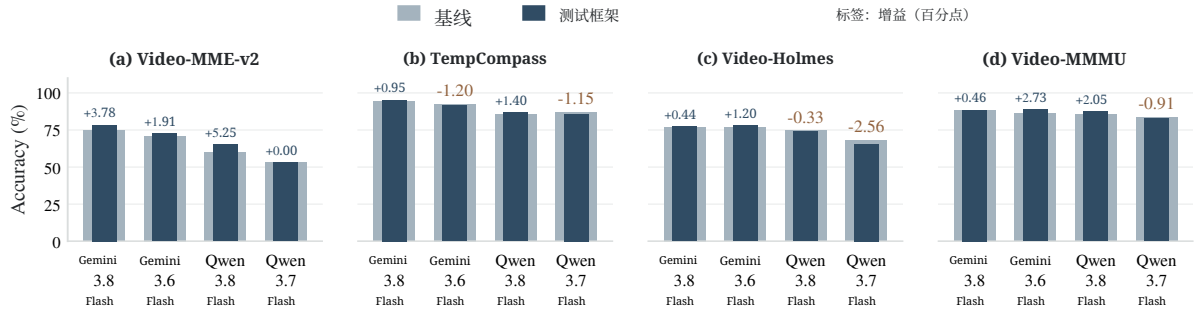

图4。A 四个公开基准上的准确率。Harness柱状条叠加在基线上；标签显示以百分点计的增益 ts.

4.3 自主观察带来的增益

FastAct 在五个模型上的增益范围为 3.0 至 14.2 分（图 3）。Qwen 3.7 Flash 的 95% 区间跨越零点（–0。4 至6。4分），因此其改进的确定性较低。

多样化视频理解。在 Video-MME-v2 上，Gemini 3.8 Flash、Gemini 3.6 Flash 和 Qwen 3.8 Flash 分别提升 3.78、1.91 和 5.25 个百分点，而 Qwen 3.7 Flash 仍保持在 53.16%（图4）。按照相同的模型顺序，官方组别得分分别从 62.21 变为 65.95、从 56.30 变为 58.45、从 43.80 变为 49.78，以及从 36.32 变为 36.46。这一收益并不局限于动作理解。

4.4 提升空间必须通过交互来实现

在 FastAct 上，通义千问 3.7 Flash 的基线最弱，增益也最小。通义千问 3.8 Flash 和 Gemini 3.6 Flash 的起点更高，但提升幅度也更大；即使是基线最强的 Gemini 4 Argon，也获得了 7.3 个百分点的增益。仅凭基线提升空间无法解释增益的排名。

我们将这些差异与恢复、回归和停止联系起来（图 5 和第 5 节）。较低的基线意味着有更多错误可供纠正，但这并不能说明进一步观察是否会帮助模型解决这些错误，同时不推翻原本正确的回答。

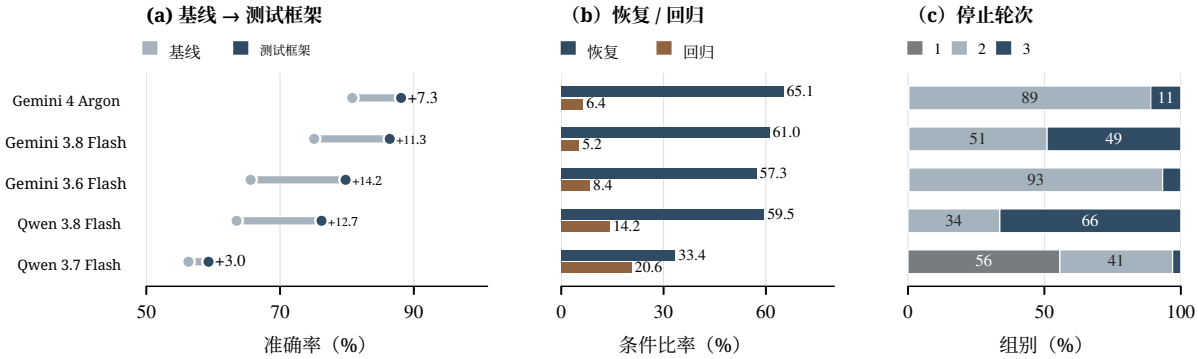

图 5。**在 FastAct 上的准确率、修订和停止**。 恢复取决于基线错误；回归取决于基线正确。轮次计数以问题组为单位。

4.5 时间需求制约机会

图6 对比了两个时间需求分层。Gemini 3.6 Flash 在每秒 8 帧问题上提升了 12.1 个百分点，在每秒 24 帧问题上提升了 18.9 个百分点；Gemini 3.8 Flash 分别提升了 10.0 和 14.3 个百分点。通义千问 3.7 Flash 在每秒 24 帧子集上的提升也更大，为 6.8 个百分点，而在每秒 8 帧子集上为 1.3 个百分点。然而，通义千问 3.8 Flash 在两个子集上的提升相近（分别为 12.8 和 12.4 个百分点）。因此，对于多个模型而言，较高的时间需求与更大的增益相关，但时间需求本身并不能决定增益大小。

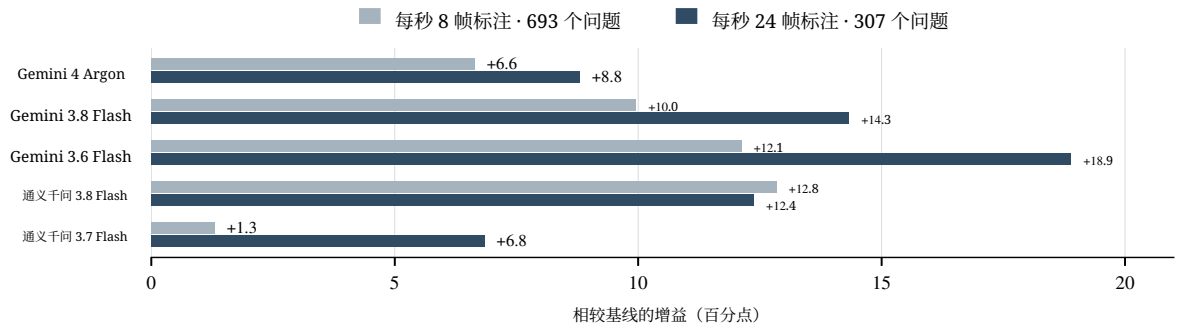

图6. 按标注的时间需求划分的增益。条形图比较了各分层中的五个模型；这些标注并非受控采样干预。

4.6 任务中的观察机会

表1. 按类别、认知水平和组类型划分的 Video-MME-v2 准确率（%）。每个单元格报告基线 → 测试框架。

子集	N	Gemini 3.8 Flash	Gemini 3.6 Flash	Qwen 3.8 Flash	Qwen 3.7 Flash
总体	3200	74.56 → 78.34	70。84 → 72。75	59.88 → 65.13	53.16 → 53.16
动作与运动	282	54.96 → 59.22	49。65 → 55。32	41。49 → 50。71	34.40 → 31.91
变化	242	76.03 → 76.86	67.77 → 70.66	61.98 → 71.90	56.20 → 59.92
复杂图表理解	249	78。71 → 79。52	74.30 → 75.10	59.84 → 64.26	49.80 → 52.61
仅帧	575	78.78 → 85.91	78.78 → 81.39	76。87 → 82。61	73.04 → 72.70
帧与音频	478	88.28 → 90.17	89。54 → 88。08	62.97 → 65.90	52。93 → 52。30
顺序	343	84.26 → 86.88	82.22 → 82.22	72.89 → 80.47	72.59 → 70.55
物理世界推理	298	51.34 → 65.10	45。97 → 45。64	35.23 → 41.28	29.53 → 26.85
社会行为分析	269	67.29 → 66.17	61.71 → 60.97	46。47 → 47。58	39.03 → 39.78
时间推理	182	79.12 → 80.77	70.88 → 74.73	68.13 → 67.58	47.80 → 53.30
基于视频的知识获取	282	74.11 → 75.89	64.89 → 73.40	54。26 → 59。22	50。35 → 50。00
第1级	1053	83。10 → 87。84	83.67 → 84.43	70。56 → 75。02	63.91 → 63.44
第2级	1049	73.59 → 76.07	68。16 → 71。02	61。11 → 68。26	54。24 → 54。72
第3级	1098	67.30 → 71.40	61.11 → 63.21	48。45 → 52。64	41。80 → 41。80
逻辑	1124	82.12 → 84.16	76.60 → 78.65	67.62 → 71.62	59。96 → 60。14
相关性	2076	70.47 → 75.19	67。73 → 69。56	55.68 → 61.61	49.47 → 49.37

MME 类别和认知层级 表1 报告了 800 个视频中各类别的结果，并将预测失败计为错误。类别和认知层级标签仅用于分析，不会提供给测试框架。

不同模型的类别增益各不相同（表1）。Qwen 3.8 Flash 在动作与运动上获得 9.22 个百分点的增益，但在时间推理上下降了 0.55 个百分点。Gemini 3.8 Flash 在物理世界推理上获得 13.76 个百分点的增益，但在社会行为分析上下降了 1.12 个百分点；Gemini 3.6 Flash 在基于视频的知识获取上获得 8.51 个百分点的增益，但在帧与音频上下降了 1.46 个百分点。Qwen 3.7 Flash 的总体表现没有变化，但在时间推理上提升了 5.49 个百分点，在动作与运动上下降了 2.48 个百分点。因此，总体增益可能掩盖相互对立的类别变化。

仅帧增益带来的好处并不局限于时间整合。静帧和空间裁剪图同样可用，但类别得分无法单独分离出它们各自的贡献。

细粒度动作概况 FastAct 也呈现出类似的变化（图 3）：Qwen 3.7 Flash 在绑定方面有所提升，但在瞬态状态方面出现退步。增益较大的模型在提升最显著的类别上也各不相同。

5 为什么视频工具框架有效?

FastAct 轨迹显示模型是否会寻求更多证据，以及修正带来的正确答案是否多于退步的答案。我们还比较了 Video-MME-v2 上最终答案的变化。

5.1 寻找证据并继续交互

在 FastAct 上，通义千问 3.7 Flash 在 55.7% 的组别中仅进行一轮便停止，而其他四个模型的这一比例为 0–0.3%。它平均每组使用 1.47 轮，而通义千问 3.8 Flash 为 2.66 轮，Gemini 3.6 Flash 为 2.06 轮。因此，通义千问 3.7 Flash 经常会在收到请求的观察结果之前就停止。它的停止行为反映了对这一测试框架的有限使用，而不是对其跨工具和任务的一般智能体能力的独立衡量。

更长的交互并不能按准确率对模型进行排名。通义千问 3.8 Flash 有 66.3% 的组别达到第三轮，而 Gemini 3.6 Flash 仅有 6.7% 达到第三轮，却取得了更高的最终准确率。轮数衡量的是后续动作，而不是返回的证据被成功利用的程度。

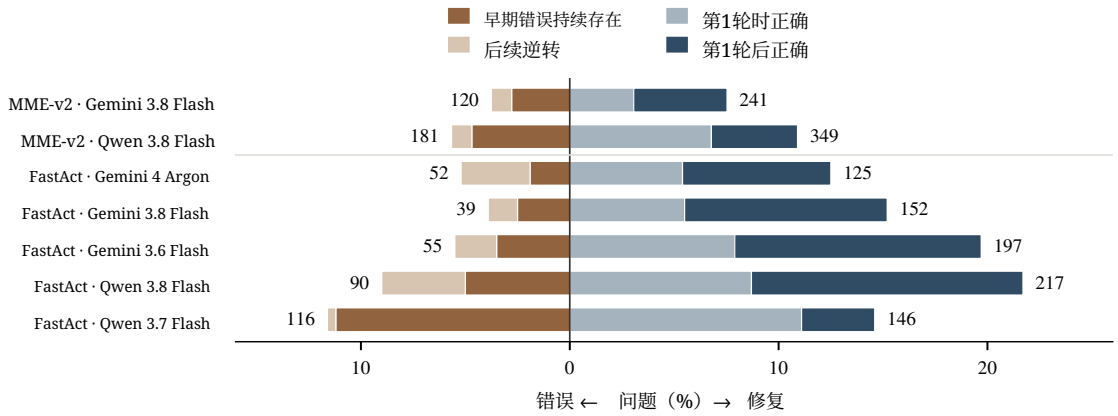

图7。回答被修复或出错的情况。条形宽度表示所有问题所占的比例；端点标签表示数量。阴影区分第一轮和后续变化。

5.2 可靠的修订决定增益

在 FastAct 上，通义千问 3.7 Flash 纠正了 33.4% 的基线错误，但推翻了 20.6% 的正确基线答案（图5）。通义千问 3.8 Flash 纠正了 59.5%，推翻了 14.2%；Gemini 3.8 Flash 纠正了 61.0%，推翻了 5.2%。二者都改善了错误恢复和正确答案保持率。

在 Video-MME-v2 上，Qwen 3.7 Flash 修正了 250 个基线错误，却又推翻了 250 个正确答案，因此总体准确率没有变化。Gemini 3.6 Flash 修正了 245 个错误，推翻了 184 个正确答案，净增益为 61 个答案。因此，当正确答案保持率不同时，修正数量相近也可能带来不同的净增益。

图7 将五个 FastAct 模型以及 Qwen 3.8 Flash 和 Gemini 3.8 Flash 在 Video-MME-v2 上第一轮与后续轮次的变化区分开来。在 FastAct 上，Qwen 3.7 Flash 的 116 个从基线到最终结果的退步中，有 112 个在第一轮后就已答错；只有 4 个原本在第一轮答对的答案后来被推翻。因此，它的大部分回归都发生在补充观察结果返回之前。

对其他模型而言，损失更常发生在后续轮次。Qwen 3.8 Flash 的 90 次退步中，有 50 次是第一轮就出现且持续存在的错误，另有 40 次是后续反转。Gemini 4 Argon 的 52 次退步中，后续反转占 33 次。因此，即使总体增益较大的模型，后续轮次仍然是回归的重要来源。

同样的区别也体现在修正上。在通义千问 Qwen 3.8 Flash 的 217 个基线错误中，最终回答纠正了这些错误，其中 87 个在第一轮后就已答对，另外 130 个则在之后变得正确。Gemini 3.8 Flash 则有 55 个较早的修正和 97 个较晚的修正。这些计数仅对基线到最终回答之间的变化进行分类：它们不包括那些基线时答对、第一轮时暂时答错、随后又恢复正确的情况。

同一个问题，相反的修订。在图 8中，通义千问 Qwen 3.7 Flash 在第一轮后以错误回答停止，且没有请求观察。在观看每秒 8 帧的视频片段后，通义千问 Qwen 3.8 Flash 将 C 改为参考答案 B，并在观看每秒 24 帧的裁剪片段后仍保留该答案。Gemini 3.6 Flash 则在观看每秒 8 帧的视频片段后将 B 改为 C。Gemini 3.8 Flash 和 Gemini 4 Argon 始终保留 B。

5.3 有效交互的条件

轨迹能够区分提前停止与后续逆转。鼓励检查可以解决前者；后者则需要可靠解释。FastAct 将这一区别具体呈现出来：相关时间段已知，但

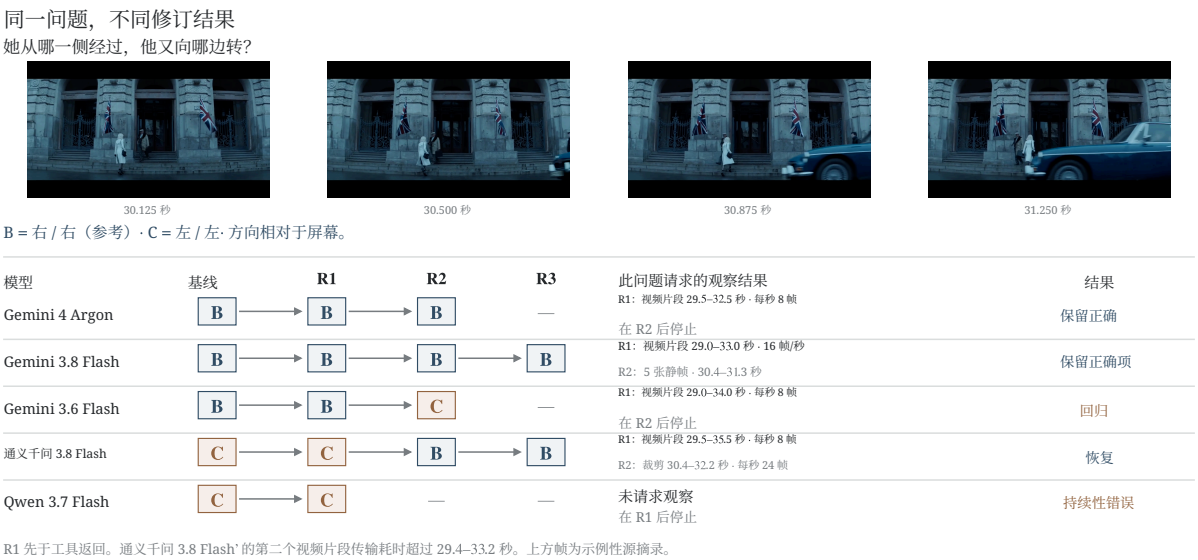

图 8。同一问题上的相反修正。 所有五个动作模型的源帧摘录和完整回答路径。请求按轮次显示；短横线表示会话已结束。

模型在检查哪些内容以及如何解释返回的证据方面仍然各不相同。任务比较能够定位增益；轨迹则展示模型如何获得这些增益，但无法分离观察与推理各自的因果贡献。

6 讨论与结论

视频工具框架使观察成为推断的一部分：模型所看到的内容可能会受到其尚未理解的内容影响。这种灵活性可以实现固定观察所未能发掘的能力。细粒度动作展现了这一机会：识别一个事件可能仍无法确定关键细节。不同模型和任务所获得的增益各不相同：进一步检查必须应对任务的难点，而模型必须组织观察结果，并可靠地利用其反馈。

因此，观测策略应随着理解的深入而发展。从一个通用的默认策略出发，推理决定哪些时间和空间细节值得关注。设计上的挑战在于维持这种相互作用：不确定性引导证据获取，而获取到的证据则决定是否应改变判断。这种协调必须符合任务对证据的需求，以及模型维持可靠交互的能力。因此，有效的反馈必须同时指导下一步观察，并帮助决定是否保留现有答案。

局限性与未来方向。 我们的评估涵盖视频问答。本文报告的相对增益是在统一的观测接口和相同的三轮上限下测得的；采用其他工具或停止规则，可能会改变哪些模型获益最多。通过受控的证据干预，可以将视觉访问与提示及额外推理区分开来。交互式三维环境、自我中心世界建模和具身任务则可用于测试那些会改变环境的行动。使用交互数据训练较小的模型，可以直接检验其证据搜寻、工具使用和答案修订能力。

大语言模型使用声明

本工作仅在以下有限范围内使用了大语言模型。在写作方面，大语言模型仅起辅助作用：作者先起草文本，再使用大语言模型进行编辑和润色。在代码和实现方面，大语言模型同样仅用于协助编辑和完善作者编写的代码。在数据准备方面，大语言模型被用于生成部分数据（合成数据、数据集创建或标注），作者则审核、编辑并批准了所有生成内容。大语言模型未用于相关工作与引用、实验设计、执行或分析、图表及视觉材料，也未用于研究思路、假设、论断和结论。作者未向任何大语言模型提供商分享完整手稿或其中的大部分内容，以获取类似审稿意见的反馈。

参考

[1] Seoyeon An、Hyeonseo Jang、Minsu Kim、Chanho Lee、Younghan Park 和 Kang wook Lee。《AgentVid-Bench：用于评估多模态大语言模型智能体的多跳视频问答基准》。arXiv预印本 arXiv:2609.21386，2026年。[2] Shuai Bai、Yuxuan Cai、Ruizhe Chen、Keqin Chen、Xionghui Chen、Zesen Cheng、Lianghao Deng、Wei Ding、Chang Gao、Chunjiang Ge 等。《通义千问3-VL技术报告》。arXiv预印本 arXiv:2511.21631，2025年。[3] Ruzena Bajcsy。《主动感知》。IEEE论文集，76(8):996–1005，1988年。doi: 10.1109/5.5968。[4] Dana H. Ballard。《动画视觉》。人工智能，48(1):57–86，1991年。doi: 10.1016/0004-3702(91) 90080-4。[5] Junhao Cheng、Yuying Ge、Teng Wang、Yixiao Ge、Jing Liao 和 Ying Shan。《Video-Holmes：多模态大语言模型能否像福尔摩斯一样进行复杂视频推理？》arXiv预印本 arXiv:2505.21374，2025年。https://arxiv.org/abs/2505.21374。[6] Junbo Cui、Bokai Xu、Chongyi Wang、Tianyu Yu、Weiyue Sun、Yingjing Xu、Tianran Wang、Zhihui He、Wenshuo Ma、Tianchi Cai 等。《MiniCPM-o 4.5：迈向实时全双工全模态交互》。arXiv预印本 arXiv:2604.27393，2026年。[7] Yue Fan、Xiaojian Ma、Rujie Wu、Yuntao Du、Jiaqi Li、Zhi Gao 和 Qing Li。《VideoAgent：用于视频理解的记忆增强型多模态智能体》。arXiv预印本 arXiv:2403.11481，2024年。[8] Chaoyou Fu、Yuhan Dai、Yongdong Luo、Lei Li、Shuhuai Ren、Renrui Zhang、Zihan Wang、Chenyu Zhou、Yunhang Shen、Mengdan Zhang 等。《Video-MME：首个用于视频分析的多模态大语言模型综合评估基准》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2025年。[9] Chaoyou Fu、Haozhi Yuan、Yuhao Dong、Yi-Fan Zhang、Yunhang Shen、Xiaoxing Hu、Xueying Li、Jinsen Su、Chengwu Long、Xiaoyao Xie 等。《Video-MME-v2：迈向综合视频理解基准的下一阶段》。arXiv预印本 arXiv:2604.05015，2026年。[10] Raghav Goyal、Samira Ebrahi mi Kahou、Vincent Michalski、Joanna Materzynska、Susanne Westphal、Heuna Kim、Valentin Haenel、Ingo Fruend、Peter Yianilos、Moritz Mueller-Freitag、Florian Hoppe、Christian Thurau、Ingo Bax 和 Roland Memisevic。《用于学习和评估视觉常识的“Something Something”视频数据库》。载于《IEEE国际计算机视觉会议论文集》（ICCV），第5842–5850页，2017年。[11] Tanmay Gupta 和 Aniruddha Kembhavi。《视觉编程：无需训练的组合式视觉推理》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），第14953–14962页，2023年。[12] Kairui Hu、Penghao Wu、Fanyi Pu、Wang Xiao、Yuanhan Zhang、Xiang Yue、Bo Li 和 Ziwei Liu。《Video-MMMU：评估从多学科专业视频中获取知识的能力》。arXiv预印本 arXiv:2501.13826，2025年。[13] Yushi Hu、Weijia Shi、Xingyu Fu、Dan Roth、Mari Ostendorf、Luke Zettlemoyer、Noah A Smith 和 Ranjay Krishna。《Visual Sketchpad：将绘图作为多模态语言模型的视觉思维链》。载于《神经信息处理系统进展》（NeurIPS），2024年。[14] Ailin Huang、Chengyuan Yao、Chunrui Han、Fanqi Wan、Hangyu Guo、Haoran Lv、Hongyu Zhou、Jia Wang、Jian Zhou、Jianjian Sun 等。《STEP3-VL-10B技术报告》。arXiv预印本 arXiv:2601.09668，2026年。[15] Jie Huang、Xinyun Chen、Swaroop Mishra、Huaixiu Steven Zheng、Adams Yu、Xinying Song 和 Denny Zhou。《大语言模型尚不能进行自我纠正推理》。载于《国际学习表征会议》（ICLR），2024年。[16] 快手可灵团队、Bin Wen、Changyi Liu、Chengru Song、Chongling Rao、Guowang Zhang、Han Li、Haonan Fan、Hengrui Ju、Jiankang Chen、Jiapeng Chen、Jiawei Yuan、Kaixuan Yang、Kaiyu Jiang、Kun Gai、Lingzhi Zhou、Na Nie、Sen Na、Tianke Zhang、Tingting Gao、Xuanyu Zheng、Yulong Chen、Fan Yang、Haixuan Gao、Lele Yang、Mingqiao Liu、Muxi Diao、Qi Zhang、Qile Su、Wei Chen、Wentao Hong、Xingyu Lu、Yancheng Long、

Yankai Yang、Yingxin Li、Yiyang Fan、Yu Xia、Yuzhe Chen、Ziliang Lai、Chuan Yi、Haonan Jia、Tianming Liang、Weixin Xu、Xiaoxiao Ma、Yang Tian、Yufei Han、Feng Han、Hang Li、Jing Wang、Jinghui Jia、Junmin Chen、Junyu Shi 和 Ruilin Zhang。Kwai Keye-VL-2.0 技术报告，2026年。https://arxiv.org/abs/2606.10651。[17] Kunchang Li、Yali Wang、Yinan He、Yizhuo Li、Yi Wang、Yi Liu、Zun Wang、Jilan Xu、Guo Chen、Ping Luo、Limin Wang 和 Yu Qiao。MVBench：一项全面的多模态视频理解基准测试。发表于 IEEE/CVF计算机视觉与模式识别会议论文集（CVPR），第 22195–22206 页，2024年。[18] Xiaolin Liu、Yilun Zhu、Xiangyu Zhao、Xuehui Wang、Yan Li、Xin Li、Haoyu Cao、Xing Sun、Shaofeng Zhang、Xu Yang、Zhihang Zhong 和 Xue Yang。Moment-Video：诊断视频多模态大语言模型对瞬时视觉事件的时间保真度。arXiv 预印本 arXiv:2606.02522，2026年。[19] Yuanxin Liu、Shicheng Li、Yi Liu、Yuxiang Wang、Shuhuai Ren、Lei Li、Sishuo Chen、Xu Sun 和 Lu Hou。Tempcompass：视频大语言模型真的理解视频吗？arXiv 预印本 arXiv:2403.00476，2024年。https://arxiv.org/abs/2403.00476。[20] Yancheng Long、Xuanyu Zheng、Wei Chen、Luoyuan Zhang、Haoyang Xu、Bin Wen 和 Tingting Gao。VideoNexus：超越推理上下文的视频感知，2026年。https://arxiv.org/abs/2611.21392。[21] Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhu moye、Yiming Yang、Shashank Gupta、Bodhisattwa Prasad Majumder、Katherine Her mann、Sean Welleck、Amir Yazdanbakhsh 和 Peter Clark。Self-Refine：基于自我反馈的迭代优化。发表于神经信息处理系统进展（NeurIPS），第 36 卷，2023年。[22] Volodymyr Mnih、Nicolas Heess、Alex Graves 和 Koray Kavukcuoglu。视觉注意力循环模型。发表于神经信息处理系统进展（NeurIPS），第 27 卷，2014年。[23] Dian Shao、Yue Zhao、Bo Dai 和 Dahua Lin。FineGym：用于细粒度动作理解的分层视频数据集。发表于 IEEE/CVF计算机视觉与模式识别会议论文集（CVPR），第 2616–2625 页，2020年。[24] Noah Shinn、Federico Cassano、Ashwin Gopinath、Karthik Narasimhan 和 Shunyu Yao。Reflexion：采用语言强化学习的语言智能体。发表于神经信息处理系统进展（NeurIPS），第 36 卷，2023年。[25] Dídac Surís、Sachit Menon 和 Carl Vondrick。ViperGPT：通过 Python 执行进行视觉推理以实现推理。发表于 IEEE/CVF国际计算机视觉会议（ICCV），第 11888–11898 页，2023年。[26] Xiaohan Wang、Yuhui Zhang、Orr Zohar 和 Serena Yeung-Levy。VideoAgent：以大语言模型为智能体理解长篇视频。发表于欧洲计算机视觉会议（ECCV），2024年。[27] Ziyang Wang、Shoubin Yu、Elias Stengel-Eskin、Jaehong Yoon、Feng Cheng、Gedas Bertasius 和 Mohit Bansal。VideoTree：用于大语言模型推理长视频的自适应树状视频表示。发表于 IEEE/CVF计算机视觉与模式识别会议论文集（CVPR），2025年。[28] Penghao Wu 和 Saining Xie。 *V**：引导式视觉搜索作为多模态大语言模型的核心机制。发表于 IEEE/CVF计算机视觉与模式识别会议论文集（CVPR），2024年。[29] Junbin Xiao、Xindi Shang、Angela Yao 和 Tat-Seng Chua。NExT-QA：问答的下一阶段——解释时间动作。发表于 IEEE/CVF计算机视觉与模式识别会议论文集（CVPR），第 9777–9786 页，2021年。[30] 小米大语言模型核心团队、Zihao Yue、Zhenru Lin、Yifan Song、Weikun Wang、Shuhuai Ren、Shuhao Gu、Shicheng Li、Peidian Li、Liang Zhao 等。MiMo-VL 技术报告。arXiv 预印本 arXiv:2506.03569，2025年。[31] Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。ReAct：让语言模型中的推理与行动相互促进。发表于国际学习表征会议（ICLR），2023年。

[32] Yuan Yao、Tianyu Yu、Chongyi Wang、Junbo Cui、Bokai Xu、Hongji Zhu、Tianchi Cai、Fuwei Huang、Tianran Wang、Wenshuo Ma 等。MiniCPM-o 2.6：可在手机上实现视觉、语音和多模态实时流媒体的 GPT-4o 级多模态大语言模型，2025年。[33] Tianyu Yu、Zefan Wang、Chongyi Wang、Fuwei Huang、Wenshuo Ma、Zhihui He、Tianchi Cai、Weize Chen、Yuxiang Huang、Ranchi Zhao 等。MiniCPM-V 4.5：通过架构、数据和训练方案打造高效的多模态大语言模型。发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 11704–11715 页，2026年。[34] Fan Zhang、Guangming Yao、Jinyang Wu、Hao Wu、Zheng Lian、Xinyu Geng、Jingdong Chen、Yi Yuan 和 Pheng-Ann Heng。VideoGAIA：面向通用人工智能助手的智能体视频理解基准测试。*arXiv* 预印本 *arXiv:2608.14718*，2026年。[35] Kuan Zhang、Yukun Chen、Zhihao Yang、Yue Su、Xiangnan Wu、Zirong Chen、Run Luo、Tao Tan、Yinhe Zheng 和 Yiming Li。Video-Arena：一项可扩展的智能体视频基准测试，用于评估时空能力。*arXiv* 预印本 *arXiv:2610.14158*，2026年。[36] Luoyuan Zhang、Bokai Xu、Junbo Cui、Weiyue Sun、Yingjing Xu、Hanyu Liu 和 Yuan Yao。释放全双工语音模型中的大语言模型能力。*arXiv* 预印本 *arXiv:2606.07547*，2026年。[37] Yaolun Zhang、Ruohui Wang、Jiahao Wang、Yepeng Tang、Xuanyu Zheng、Haonan Duan、Hao Lu、Hanming Deng 和 Lewei Lu。EVA：面向端到端视频智能体的高效强化学习。*arXiv* 预印本 *arXiv:2603.22918*，2026年。

补充材料

A 实现细节

对话与观察预算。 会话以全局视频部分、结构化的回答/规划提示以及问题组开场。后续轮次会附加新执行的观察结果和后续提示，同时保留对话历史。每个请求生成轮次最多允许三轮模型交互和六次工具请求。请求按问题排序后再执行。冗余请求以及超出服务提供方限制的观察结果可能会被跳过；执行日志会区分已请求与已交付的观察结果。Gemini 路径还会限制对话中的视频观察数量。这些操作限制是评估框架的一部分。

工具。 片段执行器会将请求的局部帧率限制在每秒 24 帧，并可降低帧率以满足帧预算；较短的窗口会进行边界填充。帧工具会返回选定的静帧，并附带带标签网格、可选裁剪图以及计数标记。蒙太奇会按顺序排列候选事件，并添加边界上下文。扫描请求以粗时间分辨率覆盖更长的时间区间。概览工具会对全局视频进行聚焦的文字重述。运动能量摘要会自动附加到符合条件的运动问题中，而不是由模型请求的操作。这些工具会呈现证据，但不提供参考标签。

FastAct 数据来源与标注。 FastAct 基于高质量内部视频数据构建，并经过人工标注和审核。其 100 个源视频片段包含 1,000 个四选项问题，分为 300 个问题组进行评估。问题明确指出相关时间范围，并区分局部动作细节；标注证据区间的中位时长为 0.90 秒。时间需求标注包括 693 个每秒 8 帧的问题和 307 个每秒 24 帧的问题。这些分层描述的是问题，而非对模型输入进行的受控干预。

B 提示模板

以下模板摘自受评估的提示构建器：build_public_test_prompt、build_prompt 和 followup_text。尖括号占位符表示变量字段或重复块。问题块会针对每个 N 问题重复出现；类型和检查清单根据问题文本推断得出。基线使用基准测试中可用的选项字母。未包含参考答案。

单次传递基线。 模型根据默认概览回答问题组。

目标视频文件： <VIDEO_PATH>

回答以下关于此视频的 <N> 个问题：

问题 1：<
QUESTION_TEXT>
选项：<OPTIONS>

对于每个问题，将单个选项字母（从 “A” 到 “<LAST_OPTION>”）设为 “"answer"“。请按以下确切结构，以 JSON 块形式输出最终回答：“‘json { "questions": [{ "question": 1, "answer": "A", "evidence": "简短的带时间戳说明" }]} “‘

初始测试框架轮次。 第一轮加入了证据评估和观察规划。

你正在回答关于一段视频的 <N> 道多项选择题。视频的第一部分是整段视频，每秒一帧，并包含音频。这是第 1 轮（共 3 轮）：根据当前观察结果尽可能作答；对于仍需更高帧率或空间缩放才能回答的问题，请请求近距离查看（使用 MM:SS 格式）。

问题 1（类型： <QUESTION_TYPE>）<
QUESTION_TEXT>选项：<OPTIONS>

判断问题 1 的方法：<
QUESTION_CHECKLIST>

<TOOL_DEFINITIONS>

说明：
- [ ] 每个问题都要先填写“证据”，并附上 MM:SS.s 格式的精确时间戳，然后填写“状态”（“已确认”或“开放”）和“回答”。- [ ] 如果你看到的视觉或音频证据足以判定答案，就将“状态”标记为“已确认”；否则标记为“开放”，在“尚不明确”中说明需要核查的内容，并请求适当的近距离查看。- [ ] 如果某个问题附有新的近距离观察结果，请根据这些观察结果核对每个选项的具体细节；如果近距离查看支持不同的选项，则更新“回答”。- [ ] 仅以 JSON 格式回复：{"questions": [{"question": 1, "evidence": "...", "unclear": "...", "status": "confirmed or open", "answer": "A"}], "requests": [...]}

可用的观察请求。 以下阻挡会替换初始提示和第二轮提示中的 <TOOL_DEFINITIONS> 。

可在“requests”中请求更细致的查看（每轮最多 6 项；时间请写成“MM:SS”或“MM:SS.s”，例如“01:54”或“04:43.5”）： - {"tool": "clip", "question": k, "start": "MM:SS", "end": "MM:SS", "fps": 4|8|24, "box": [x0, y0, x1, y1]or null, "purpose": "..."}: 高分辨率视频片段（最高每秒 24 帧）。可灵活选择“fps”和时间窗口长度（每个片段保持在 ~60-120 帧以内：15 秒时使用 ~4 帧/秒，10 秒时使用 ~8 帧/秒；对于快速或亚秒级动作，5 秒时最高可用每秒 24 帧；也接受“detail”和“motion”）。 - {"tool": "frames", "question": k, "times": ["MM:SS.s", ...],"cells": ["B2", "B3"] or null, "crop": [x0, y0, x1, y1] or null, "points": [[x, y],...] or null, "purpose": "..."}: 最多 8 张全分辨率静帧。完整帧包含一个 4×4 网格（A1..D4）和 0.1-0.9 坐标标尺；传入“cells”（例如["B2", "C2"]）或“crop”（[x0, y0, x1, y1]）可将该区域放大至 1536 px；传入“points”（[[x, y], ...]）可绘制编号标记 #1..#N，以核验密集的单帧计数。 - {"tool": "montage", "question": k, "times": ["MM:SS.s", ...], "cells": ["B2"] or null, "purpose": "..."}: 根据候选事件时间戳 [t1, t2, ...] 以及边界探测帧 [PRE] 和 [POST] 生成编号事件记录拼图，以核验时间计数，避免重复计数或漏掉边界处的重复事件。

- {"tool": "scan", "question": k, "start": "MM:SS", "end": "MM:SS", "purpose": "..."}: 针对较长时间范围（30-120 秒），以默认分辨率每秒 2 帧进行扫描。 - {"tool": "overview", "question": k, "purpose": "what to search for across the whole video"}: 如果未在局部片段中找到目标事件，则以每秒 1 帧（含音频）重新扫描整个视频，并将结果附加到概览文件中。

后续轮次。 新观察会追加到保留的对话中。初始轮次中的回答说明和 JSON 格式会重复提供。如果无法执行某项请求，后续内容也会说明该请求以及跳过的原因。最后一轮会省略工具定义。

你在上一条回复中请求的更仔细观察已按请求的顺序附在上方。这是第 2 轮，共 3 轮。请结合目前为止在本次对话中看到的所有内容，重新检查每个问题；对于仍需仔细观察的问题，请求进一步仔细观察。

<TOOL_DEFINITIONS><
ANSWER_INSTRUCTIONS_AND_JSON_FORMAT>

你在上一条回复中请求的更仔细观察已按请求的顺序附在上方。这是第 3 轮，共 3 轮。这是最后一轮：请根据你在本次对话中看到的所有内容给出最终答案。

<回答说明和 JSON 格式>
对于 Video-MME-v2 中的逻辑问题组，回答说明还要求按顺序解答问题，因为组别评分会在第一个错误答案处停止。

C 图像中的观测接口



图9。Ac 实际的工具输出。每个面板都将请求与返回的观测配对；运动摘要会自动附加 ly.