arXiv:2603.00983v1 [cs.CV] 1 Mar 2026

面向高效长视频理解的事件锚定帧选择

Wang Chen[1*]、Yongdong Luo[1*]、Yuhui Zeng[1],Luojun Lin[2], Tianyu Xie[1], Fei Chao[1],Rongrong Ji[1],Xiawu Zheng[1†]1多媒体可信感知与高效计算重点实验室，中华人民共和国教育部，厦门大学2计算机与数据科学学院，福州大学

摘要

大量帧冗余与有限的上下文窗口，使得高效帧选择对于基于大型视觉语言模型（*LVLM*）的长视频理解至关重要。然而，主流方法采用平坦采样范式，将视频视为无结构的帧集合。本文提出***Event-Anchored Frame Selection***（***EFS***），一种层次化、事件感知的流程。借助自监督*DINO*嵌入，*EFS*首先将视频流划分为视觉上同质的时间片段，作为语义事件的代理。在每个事件内，再选择与查询最相关的帧作为锚点。这些锚点作为结构先验，引导采用自适应最大边际相关性（*MMR*）方案的全局精炼阶段。该流程确保最终关键帧集同时优化事件覆盖、查询相关性与视觉多样性。作为一个免训练的即插即用模块，*EFS*可无缝集成到现成的大型视觉语言模型中，在具有挑战性的视频理解基准测试上取得显著增益。具体而言，应用于*LLaVA-Video-7B*时，*EFS*在*VideoMME*、*LongVideoBench*和*MLVU*上的准确率分别提升了***4.7%***、***4.9%***和***8.8%*** 。

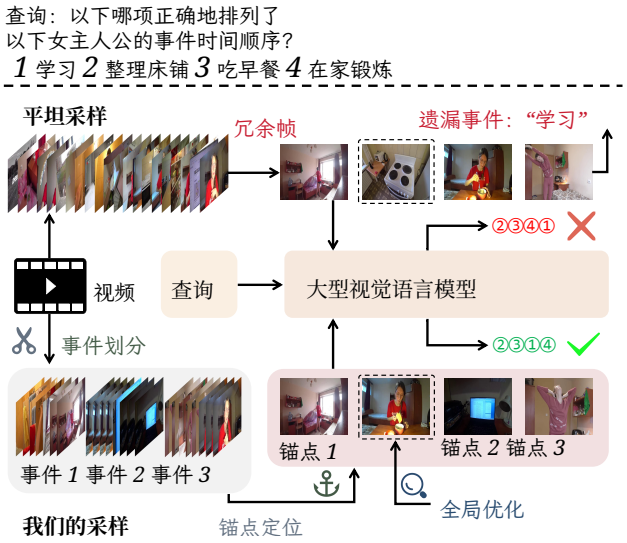


图1. 我们提出的事件锚定帧选择（EFS）与平坦采样基线的概念示意图。平坦采样往往会产生冗余帧并遗漏关键事件，导致大型视觉语言模型无法正确应对；而 EFS 将视频划分为多个事件，选取与查询相关的锚点，并进行锚点引导的全局细化，从而实现正确的时间顺序推理。

1. 引言

大型语言模型（LLM）在自然语言处理领域取得了非凡的成功 [1] ，推动其向多模态领域拓展。通过将强大的视觉编码器与大语言模型先进的推理能力相结合，大型视觉语言模型（LVLM） [15,36]应运而生，推动了机器感知的快速发展。借助在大规模视频文本数据集上的预训练，大型视觉语言模型如今能够执行复杂的视频任务，例如生成细致的视频描述 [5], 、回答细微的问题答 [23],案以及复杂的时间推理 [29]。

尽管取得了这些成功，大型视觉语言模型在处理长视频时仍面临一个根本瓶颈：海量帧数与有限的计算预算和固定的上下文窗口相互冲突。直接处理每一帧根本不可行。因此需要一种机制，将庞大的视觉流压缩为简洁而又具代表性的输入。因此，帧选择已成为将大型视觉语言模型应用于真实世界长视频内容的关键且务实的前提。虽然存在扩展上下文窗口 [16, 42] 或视频转文本摘要 [21]等替代策略，但它们往往带来过高的计算开销或严重的信息损失，这使得高效的帧选择成为一种更直接、更均衡的方法。

*共同贡献。
†通讯作者。

然而，帧选择的效果完全取决于帧的选取方式。许多现有技术 [8, 10, 28, 34, 38–40, 43] 采用平坦采样范式，将视频视为无结构的帧集合，而忽略更高层次的语义关系。该范式本质上不感知时间，忽视了视频固有的叙事与事件结构。因此，这些方法往往无法选出能够同时有效平衡理想选择三大支柱的帧集合：查询相关性、全面的事件覆盖以及足够的视觉多样性。如图1所示，平坦采样策略容易遗漏关键事件，导致大型视觉语言模型得出错误结论。

为克服这一局限，我们提出从不考虑时间因素的视角转向事件感知视角。我们提出事件锚定帧选择（EFS），这是一种基于视频内在事件结构的分层免训练流水线。我们的核心洞见是，最优关键帧集合必须建立在对视频叙事脉络的宏观理解之上。EFS 首先将视频划分为视觉上连贯的时间片段，作为语义事件的可靠代理。随后，它通过从每个事件中选取与查询最相关的帧作为锚点，建立结构性“骨干”。这组初始事件锚点继而引导全局精炼阶段，该阶段利用自适应最大边际相关性（MMR）方案，纳入多样且信息丰富的帧，进一步优化选择结果。这一分层流程确保最终关键帧集合能够同时优化事件覆盖率、查询相关性和视觉多样性。

作为一个即插即用模块，EFS 无需额外训练，即可与现成的大型视觉语言模型无缝集成。我们在三个广泛使用的长视频问答基准测试集上验证了 EFS：VideoMME [9],LongVideoBench [37], 和 MLVU [45]。在所有基准测试集上，EFS 均带来了稳定且显著的性能增益。具体而言，将 EFS 应用于 LLaVA-Video-7B 和 LLaVA-OneVision-7B 时，在 VideoMME、LongVideoBench 和 MLVU 上的准确率分别提升**4.7%、4.9% 和 8.8%**和**3.3%、6.2% 和 8.8%** 。这些结果表明，事件感知帧选择对于释放大型视觉语言模型在长视频理解方面的全部潜力至关重要。

我们的主要贡献总结如下：

•我们提出了事件锚定帧选择（EFS），这是一种无需训练的层次化框架，能够先感知视频的宏观事件结构，再进行细粒度帧选择，突破了扁平且不考虑时间信息的采样方法的局限。

• 我们设计了一种锚点引导的全局精炼策略，采用自适应最大边际相关性（MMR）方案。该方法会根据视频自身的内容统计信息动态校准多样性阈值，从而提升其在不同视频类型中的稳健性和适应性。

• 我们在三个具有挑战性的长视频理解基准上进行了广泛实验，结果表明，EFS 显著提升了现有大型视觉语言模型的推理性能，充分凸显了事件感知帧选择的价值。

2. 相关工作

用于长视频理解的大型视觉语言模型。尽管大型视觉语言模型在短视频片段上表现出色 [14, 22], ，但由于帧冗余严重且上下文窗口有限，它们在处理长视频时面临重大挑战。为缓解这一问题，一类研究致力于改进模型架构，例如扩展模型的上下文窗口 [16, 42]，或通过剪枝和合并来减少视觉词元 [13, 20, 32]。然而，这些方法往往计算开销较高，或有丢弃关键视觉细节的风险。另一种策略是将视频内容转换为文本，以利用大语言模型的语言能力 [7, 19, 21, 28]。尽管这种方法效率较高，但仍会造成显著的信息损失，从而限制对视觉细微差异的理解。

面向视频理解的帧选择。帧选择为上述方法提供了一种实用的替代方案。均匀采样或早期关键帧提取等传统方法 [24, 31] 通常与查询无关，难以胜任复杂的推理任务。近期研究开始探索基于查询的帧选择，根据帧与特定查询的相关性对帧进行排序。例如，BOLT [17]采用逆变换采样，而 Frame-Voyager [40] 则通过基于排序的流程寻找最优帧组合。其他值得关注的方法包括：mDP^3 [34], 将帧选择形式化为马尔可夫决策过程，并结合行列式点过程；AKS [35], 联合优化查询相关性和视频覆盖率；Q-frame [43], 以动态分辨率处理帧；KFC [8], 将该问题建模为子图选择；T* [39], 则采用时序对象中心搜索。尽管这些方法能够提升查询相关性或多样性，但它们大多采用平坦采样范式，将视频视为无结构的帧序列，忽略了事件结构，从而限制了事件覆盖。相比之下，我们的分层事件感知框架联合优化相关性、覆盖率和多样性。

3. 方法

本节将详细介绍我们的事件锚定帧选择框架。如图2所示，EFS 首先将视频划分为连贯的视觉事件，从而获得宏观理解。随后，这种事件结构将引导全局细化过程，生成一组关键帧

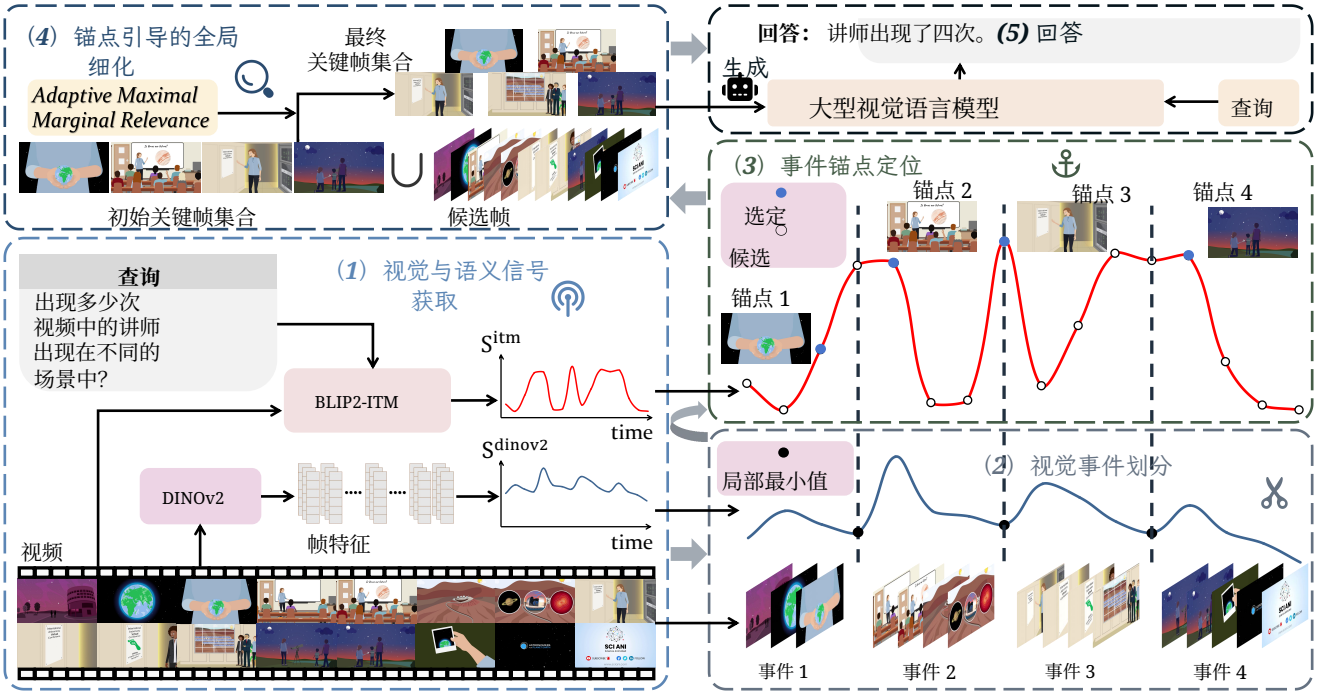


图2. 我们的事件锚定帧选择（EFS）流程示意图。给定一个用户查询，例如“讲师在不同场景中出现了多少次？”，EFS 分为多个阶段运行：(1) 首先获取视觉信号（来自 DINOv2）和语义相关性分数（来自 BLIP2-ITM）。(2) 利用视觉信号，在时间相似度的局部最小值处将视频划分为连贯的事件。(3) 在每个事件中，定位与查询最相关的帧作为锚点。(4) 随后通过自适应 MMR对这组初始锚点进行全局优化，形成最终关键帧集。(5) 最后，这组经过筛选的帧使大型视觉语言模型能够准确回答查询，并正确识别出四次出现。

，并针对相关性、覆盖率和多样性进行优化。

3.1. 问题定义

给定一个包含 T帧的长视频， $V = \{I_1, I_2, ..., I_T\}$，以及用户查询 Q，关键帧选择的任务是提取一个简洁且具有代表性的帧子集 $K = \{I_{t_1}, I_{t_2}, ..., I_{t_k}\}$，其中索引按时间顺序排列，且 $k << T$。 k 的值通常根据大型视觉语言模型的上下文窗口限制以及下游任务的需求或用户偏好预先设定。目标是选择 K，使得将其与查询 Q 一同作为视觉上下文提供后，大型视觉语言模型能够生成尽可能准确且全面的答案。这需要平衡三个核心目标：确保充分覆盖关键事件、与查询保持高度相关，并提升多样性以尽量减少冗余。

然而，现有基于大型视觉语言模型的长视频理解流程中，最常用的策略是均匀采样。尽管简单高效，它在很大程度上忽视了视频内在的时间与语义结构，常常错过与查询相关的关键事件，并引入冗余或缺乏信息的帧。因此，更智能、基于结构的——因此，需要具备感知能力的关键帧选择框架。

3.2. 视觉与语义信号获取

我们首先以每秒 1 帧的速度对视频进行采样，得到长度为 N 的候选帧序列 $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$。在该序列中，我们为每一帧提取两种基本信号。

图文匹配。为了量化每一帧与查询的语义相关性，我们使用 Blip2 [12]的图文匹配（ITM）头（称为 BLIP2-ITM），该模块旨在输出一个分数，用以衡量给定图像与文本描述之间的对齐程度。此分数是衡量查询相关性的出色指标。对于每一帧 I_i 和文本查询 Q，相关性分数的计算方式如下：

$$s_i^{\text{itm}} = \text{BLIP2-ITM}(I_i, Q), \tag{1}$$

这些分数构成相关性信号 $\{s_1^{\text{itm}}, \ldots, s_N^{\text{itm}}\}$，进而使我们能够优先选择与用户特定信息需求相关的帧。

时间相似度计算。为了理解视频的时间结构，我们使用 DINOv2 [27], a 这一强大的自监督模型来衡量帧之间的视觉相似度。该模型以稳健的视觉嵌入而闻名，

这些嵌入能够捕捉高层语义内容和底层结构细节，且无需显式标签。其特征尤其擅长辨别视觉内容中的有意义变化。对于每一帧 I_i，我们提取其 DINOv2 特征向量 $\mathbf{f}_i \in \mathbb{R}^d$ 并进行 L2 归一化。随后，利用这些嵌入，通过将每一帧 i 与加权滑动窗口内的相邻帧进行比较，计算其帧间时间相似度分数 s_i^{dinov2}。窗口大小为 l。这能够捕捉局部视觉连贯性：

$$s_i^{\text{dinov2}} = \frac{\sum_{j \in \mathcal{N}(i)} w_{|i-j|} \cdot \cos(\mathbf{f}_i, \mathbf{f}_j)}{\sum_{j \in \mathcal{N}(i)} w_{|i-j|}}, \tag{2}$$

其中，$\mathcal{N}(i)$表示帧 i 的邻域，权重w_d 会随着距离 $d = |i-j|$ 的增加而线性递减。根据经验性设计，我们将 l 设置为 3；详细的消融研究请参见附录。由此得到的帧间时间相似度曲线 $\{s_1^{\text{dinov2}}, \ldots, s_N^{\text{dinov2}}\}$揭示了整个视频中视觉变化的节奏。

3.3. 事件锚定帧选择

我们的 EFS 方法遵循一个由粗到细的三阶段流程：将视频划分为事件，在每个事件中定位一个锚点，并执行全局感知的细化。

视觉事件划分。 我们将“事件”概念化为视觉上同质的时间片段。在视频制作中，视觉内容的显著变化——例如镜头切换或场景的大幅变化——通常标志着不同场景或事件之间的边界。这些时刻直接对应于我们的时间相似度曲线 $\{s_1^{\text{dinov2}}, \ldots, s_N^{\text{dinov2}}\}$中的局部最小值，因为它们表示视觉变化达到最大。与依赖聚类或其他计算密集型分割方法的方法不同，直接将这些最小值选作初始事件边界非常高效，且不会产生额外开销。

然而，长视频通常包含数百个事件，超出大型视觉语言模型的实际词元预算。为此，我们设定目标事件数量 M。如果初始划分得到的事件数量超过 M，我们便迭代合并最相似的相邻事件对，依据是它们的 DINOv2 平均特征之间的余弦相似度。该过程持续进行，直到恰好剩下 M 个视觉上彼此不同且时间上连贯的事件 $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_M\}$，从而形成用于选择的宏观结构。

事件锚点定位与初始化。 为同时确保完整的事件覆盖与较强的查询对齐，我们通过从每个事件中选取一个代表性帧来初始化关键帧集合。鉴于用户查询在视频理解任务中的重要性，我们在每个事件 $\mathcal{G}_j$ 中选择查询相关性分数 s^{itm} 最高的帧作为其“锚点”。这可以表述为

$$k_j^{\text{anchor}} = \arg\max_{i \in \mathcal{G}_j} s_i^{\text{itm}}, \tag{3}$$

这组锚点$\mathcal{K}_{\text{init}} = \{k_1^{\text{anchor}}, \ldots, k_M^{\text{anchor}}\}$构成了一个初始关键帧集合，它牢固立足于视频的事件结构，同时以查询的关注点为中心。

算法1 锚点引导的全局细化

输入： 候选帧 $\mathcal{C}$、锚点集 $\mathcal{K}_{\text{init}}$、DINOv2 帧特征 $\mathbf{f}$、相关性分数 s^{itm}、目标大小 k，阈值松弛因子 α，增量 δ。
输出： 最终关键帧集 $\mathcal{K}$。
$\mathcal{K} \leftarrow \mathcal{K}_{\text{init}}$；按相关性分数 s^{itm}降序排列 $\mathcal{C}$。
$\mu, \sigma \leftarrow$ 均值,标准差({最大值$_{I_j \in \mathcal{K}_{\text{init}}}$ 余弦相似度($\mathbf{f}_i$,$\mathbf{f}_j$) $\mid I_i \in \mathcal{C}$})
// 估计 $\mathcal{C}$ 到 $\mathcal{K}_{\text{init}}$的相似度分布。
$\theta_{\text{strict}}, \theta_{\text{loose}} \leftarrow$ 截断($\mu - \alpha\sigma, 0, 1$), 截断($\mu + \alpha\sigma, 0, 1$)
// 设置自适应严格阈值和宽松阈值。
$\theta_{\text{current}} \leftarrow \theta_{\text{strict}}$
当 $|\mathcal{K}| < k$ **且** $\theta_{\text{current}} \leq \theta_{\text{loose}}$ **时执行**
  对于 I_c 在 $\mathcal{C} \setminus \mathcal{K}$ **中执行**
    若 $|\mathcal{K}| \geq k$ **则中断**
    结束若
    若 $\max^{I_j \in \mathcal{K}} \cos(\mathbf{f}_c, \mathbf{f}_j) < \theta_{\text{current}}$ **则**
      $\mathcal{K} \leftarrow \mathcal{K} \cup \{I_c\}$
      // 若低于当前阈值，则添加帧。
    结束 if
  结束 for
  $\theta_{\text{current}} \leftarrow \min(\theta_{\text{current}} + \delta, \theta_{\text{loose}})$
  // 放宽下一轮的阈值。
结束 while
返回 $\mathcal{K}$

锚点引导的全局细化。 锚点集提供了强但稀疏的表示。最终步骤是通过添加更多帧来细化该集合，以增强细节和多样性。一种经典方法是最大边际相关性（MMR） [3], 其目标是选择既与查询相关、又与已选定项不相似的项：

$$\arg\max_{I_i \in \mathcal{C}} \left[\lambda \cdot \text{sim}(I_i, Q) - (1-\lambda) \max_{I_j \in \mathcal{K}} \text{sim}(I_i, I_j) \right]. \tag{4}$$

此处，$\mathcal{K}$ 表示选定帧的集合，$\mathcal{C}$ 指剩余帧，超参数 λ 用于平衡相关性与多样性。该方法在每一步对所有候选重新评分，导致计算复杂度约为$\mathcal{O}(|\mathcal{C}||\mathcal{K}|^2)$，可能较为耗时。为提高效率， [6] 提出一种近似贪心算法，按相关性对候选排序，并应用固定的多样性阈值。然而，该方法不够稳健：单一阈值无法适应不同的视觉节奏——适用于动作片的阈值可能不适用于节奏缓慢的纪录片。

我们的核心动机是让多样性阈值能够自适应并由数据驱动，依据每段视频内在的视觉结构进行调整。通过将先前生成的事件锚点用作统计先验，我们的锚点引导细化策略会根据内容

表 1. 以往大型视觉语言模型在 Video-MME [9], LongVideoBench [37] 和 MLVU [45]上的基于视频的问答准确率（%）。我们的 EFS 应用于三种现成的大型视觉语言模型。Frames 和 LLM 分别表示输入大型视觉语言模型的视频帧数量，以及 LLM 部分的参数数量。

模型	参数量	帧	VideoMME（不含子集）				LongVideo 基准测试集 (验证集)	MLVU (平均值)
			短	中	Long	总体		
专有大型视觉语言模型								
GPT-4o mini [25]	-	250	72.5	63.1	58.6	64.8	-	-
GPT-4o [26]	-	384/256/0.5fps	80.0	70.3	65.3	71.9	66.7	64.6
开源大型视觉语言模型								
Video-LLaVA [14]	7B	8	45.3	38.0	36.2	39.9	39.1	47.3
LongVA [42]	7B	128/256	61.1	50.4	46.2	52.6	-	56.3
Video-XL [33]	7B	128/256	64.0	53.2	49.2	55.5	50.7	64.9
Frame-Voyager [40]	7B	8	67.3	56.3	48.9	57.5	-	65.6
LongVU [32]	7B	1fps	-	-	59.5	60.6	-	65.4
NVILA [18]	8B	256	75.7	62.2	54.9	64.2	57.7	70.1
LLaVA-OneVision [11]	7B	8	65.2	52.0	45.0	54.1	54.1	58.6
LLaVA-OneVision + EFS	**7B**	**8**	**70.4**(+5.2)				**60.3** (+6.2)	**67.4** (+8.8)
Qwen2.5-VL [2]	7B	16	68.0	57.0	47.8	57.6	57.0	56.5
Qwen2.5-VL + EFS	**7B**	**16**	**70.2**(+2.2)				**60.5** (+3.5)	**66.0** (+9.5)
LLaVA-Video [44]	7B	64	76.4	63.1	54.2	64.6	58.8	68.1
LLaVA-Video + EFS	**7B**	**64**	**77.3**(+0.9)				**62.1** (+3.3)	**70.9** (+2.8)

冗余度量身定制源视频的处理方式。这种方法会动态调整多样性阈值，使视觉内容密集的片段接受更严格的去重处理，而较稀疏的区域则采用更宽松的处理方式。这种内容感知的自适应机制，使关键帧选择能够适用于各种类型的视频，并兼顾稳健性与公平性。有关实现细节，请参阅算法1。由于外层 while 循环在实际应用中通常只执行少数几次，因此总体复杂度约为 $\mathcal{O}(\max(|\mathcal{K}||\mathcal{C}|, |\mathcal{C}| \log |\mathcal{C}|))$，因此该方法适用于实际应用。由我们的 EFS 确定最终关键帧集 $\mathcal{K}$ 后，将其与查询 Q 拼接，并作为输入提供给预训练的大型视觉语言模型。随后，该模型会生成全面的回答： $\mathcal{R} =$ 大型视觉语言模型$(\mathcal{K}, Q)$。

4. 实验

4.1. 实验设置

评测基准与模型。我们在三个长视频问答基准测试上评估所提出的方法：1）VideoMME [9], 包含 2,700 组人工标注的问答对，视频平均时长为 17 分钟；2）LongVideoBench [37], 采用包含 1,337 组问答对的验证集，视频平均时长为 12 分钟；3）MLVU [45], 是一个多任务基准测试集，我们重点关注其中的多项选择（M-avg）任务，该任务涵盖 7 个类别，共包含 2,174 个问题，每段视频的平均时长为 11 分钟。为公平评估关键帧选择的贡献，我们在所有实验中均不使用视频字幕。我们选用三个大型视觉语言模型（LVLM）作为基线：LLaVA-OneVision [11],、Qwen2.5-VL [2], 和 LLaVA-Video [44], , 均为 7B 版本。为确保公平比较，我们在统一的实验设置下复现了所有模型的结果。

实现细节。 为降低计算成本，我们从每个视频中均匀采样候选帧，采样频率为每秒 1 帧。使用 BLIP2-ITM-ViT-g 模型计算查询与帧的匹配分数 [12], , 同时使用 DINOv2 [27]进行视觉特征提取。对于所有任务，我们使用 LMMs-Eval 工具包 [41]报告测得的准确率。所有实验均在配备 80 GB 内存的 NVIDIA A800 GPU 上完成。

4.2. 与最先进方法的比较。

与不同大型视觉语言模型的比较。我们首先将本方法的视频问答准确率与领先的大型视觉语言模型进行比较，其中包括 GPT-4o 等专有模型和 NVILA 等开源模型。这些模型在参数规模、输入帧容量和架构方面各不相同，为评估本方法的泛化能力提供了广泛的测试。如表1所示，我们的事件锚定帧选择（EFS）在所有模型和基准测试集上都稳定且显著地提升了准确率。例如，将 EFS 应用于 LLaVA-OneVision（8 个输入帧）时，VideoMME、LongVideoBench 和 MLVU 上的准确率分别提升了 3.3%、6.2% 和 8.8%。对于 Qwen2.5-VL

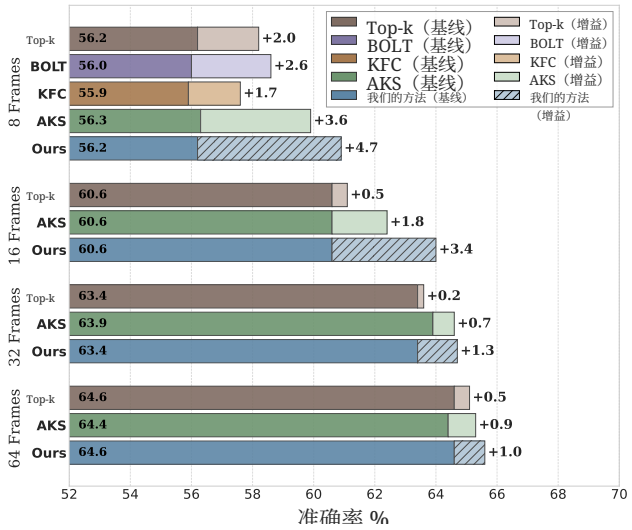


图3. 使用 LLaVA-Video-7B 在 VideoMME 上进行的定量比较。与不同采样策略相比，在 8/16/32/64 帧预算下，EFS 始终带来最大的准确率提升。每个柱形左侧的文本表示该方法的基线性能，右侧的文本则表示相应的增益。

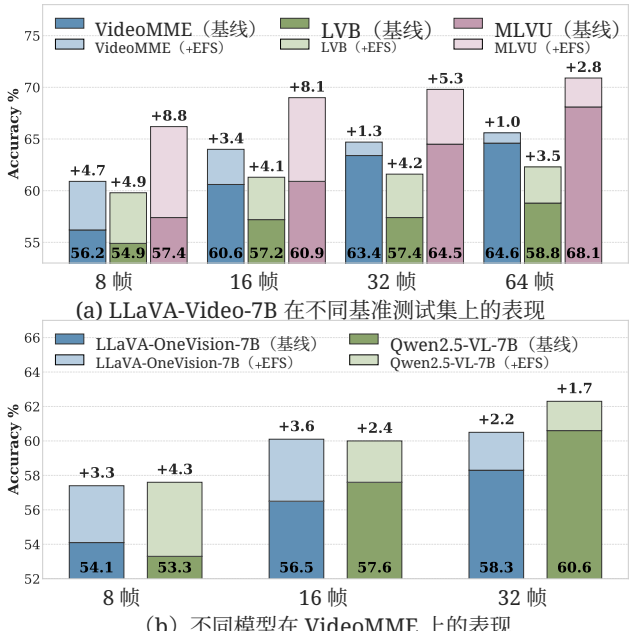


图4. 不同输入帧数量下基于视频的问答准确率（%）。

（16 帧），增益分别为 2.4%、3.5% 和 9.5%。即使是面对最强的基线 LLaVA-Video（64 帧），EFS 也分别带来了 1.0%、3.3% 和 2.8% 的提升。这些改进使较小的 70 亿参数模型能够在相近的计算预算下超越竞争方法，在某些情况下甚至可以媲美大得多的模型。例如，采用 EFS 的 LLaVA-Video-7B 在 VideoMME 上的准确率达到 65.6%，比 GPT-4o-mini 高出 0.8%。同样，采用 EFS 时，仅使用 8 帧的 LLaVA-OneVision 在 MLVU 上的表现也比 GPT-4o 高出 2.8%，而均匀帧选择无法取得这一结果。这些发现凸显了 EFS 出色的泛化能力，以及其在各种模型和输入配置中的简便易用、即插即用优势。

不同帧采样策略的比较。我们进一步与近期基于查询的帧采样方法进行直接比较，包括 Top-k、BOLT [17], KFC [8], 和 AKS [35]。所有方法均以 LLaVA-Video-7B 为骨干网络，在 VideoMME 上进行评估，并报告相较于均匀采样基线的提升，涵盖不同输入帧数。如图 3所示，EFS 在所有帧数下均带来了最大的准确率提升，在 8、16、32 和 64 帧时分别比基线高出 4.7%、3.4%、1.3% 和 1.0%。Top-k 方法由于存在冗余选择，带来的提升微乎其微。尽管 BOLT、KFC 和 AKS 等策略旨在提高查询相关性和帧的多样性，但其平坦采样范式往往忽略视频内在的事件结构。总而言之，EFS 始终实现最稳健的提升，彰显了其在信息帧选择方面的有效性，以及其对准确理解长视频的优势。

4.3. 消融与分析

不同数量的关键帧。为评估 EFS 的泛化能力，我们在不同帧预算和多个大型视觉语言模型上评估其性能，如图 4所示。在 LLaVA-Video-7B 上（子图 a），EFS 在所有三个基准测试集上始终优于均匀采样，在帧预算紧张时优势尤其显著：仅使用 8 帧时，VideoMME 上的准确率提高了 +4.7%，LongVideoBench 上提高了 +4.9%，MLVU 上提高了 +8.8%；即使在 64 帧时——此时均匀采样基线表现更强——EFS 仍保持明显优势。EFS 展现出模型无关性（子图 b），同样能提升 LLaVA-OneVision-7B 和 Qwen2.5-VL-7B 在 VideoMME 上的表现；使用 8 帧时，二者的准确率分别提高了 +3.3% 和 +4.3%。总体而言，这些结果表明，EFS 是一个稳健的即插即用模块，无论大型视觉语言模型架构或帧预算如何，都能带来显著的准确率提升，彰显了其在各种场景中的实用价值。

EFS 策略消融实验。为验证 EFS 流水线，我们使用 16 个输入帧在 LLaVA- Video-7B 上进行了组件级消融实验（见表2和表3）。

1）事件划分。在表2中，将事件边界定义在DINOv2相似度曲线的局部最小值处取得了最佳结果，优于随机划分和

表2. EFS中不同策略的基于视频的问答准确率（%）。我们在VideoMME和LVB基准测试上，将最终方法的性能与采用不同事件划分、锚点初始化和精炼过程策略的变体进行比较。

方法		VideoMME	LVB
均匀采样		60.6	57.2
随机划分		62.9	59.8
聚类划分		63.3	60.7
随机初始化		62.1	60.3
集中式初始化		61.9	60.1
不使用自适应的最大边际相关性	τ=0.4	63.7	59.9
	τ=0.6	63.4	61.2
	τ=0.8	63.6	60.9
我们的方法		**64.0**	**61.3**

表 3。在 LLaVA-Video-7B 上使用不同镜头边界检测器的基于视频的问答准确率（%）。

方法	VideoMME	LVB
均匀采样	60.0	57.2
自适应检测器 [4]	63.6	60.3
内容检测器 [4]	63.6	60.4
哈希检测器 [4]	62.8	60.4
直方图检测器 [4]	62.7	60.3
阈值检测器 [4]	63.2	59.8
我们的方法（基于 DINOv2）	**64.0**	**61.3**

聚类划分。这表明，检测自然时间边界比任意划分或基于聚类的划分更有效，且计算效率更高。此外，为阐明我们的方法与传统镜头边界检测之间的关系，我们进行了一项额外实验（表3）。我们将基于DINOv2的划分模块替换为PySceneDetect中若干成熟的检测器 [4], ，这些检测器基于较低层次的视觉线索运行。我们的方法再次全面优于它们，从而支持了我们的假设：尽管在概念上与镜头检测相关，但利用DINOv2等强大的自监督特征能够对视频事件进行语义上更忠实、更准确的划分——这些能力对下游推理任务至关重要。

2）锚点初始化。接下来，我们考察如何在每个事件中选定初始锚点帧。为每个事件选择查询相关性最高的帧，效果显然优于随机选择或基于视觉质心的选择，这证实了对于视频问答而言，语义相关性比视觉代表性更为重要。优先考虑查询相关性，能够使锚点集合与用户意图高度一致。

3）自适应细化。最后，我们验证了

表4. 信号采集消融。“+”左侧表示用于事件检测的视觉特征提取器，右侧表示用于查询帧相关性评分的模型。

方法	帧	VideoMME	LVB
均匀采样	8/16	56.2/60.6	54.9/57.2
CLIP + CLIP	8/16	60.2/63.1	57.6/60.2
DINOv2 +CLIP	8/16	60.5/63.1	58.4/60.3
CLIP + BLIP2-ITM	8/16	60.4/63.3	58.8/60.4
DINOv2 +BLIP2-ITM	8/16	**60.9/64.0**	**59.8/61.3**

表5. 对超参数 M（最大事件数）和 α（阈值松弛因子）的消融实验。所有实验均使用 LLaVA-Video-7B 在 VideoMME 上进行。

M \ α	0.1	0.3	0.4	0.5	0.6	0.7	0.9
1	61.4	61.3	61.0	61.5	61.7	61.9	61.6
4	62.5	62.4	62.4	62.5	62.2	61.9	61.4
8	62.6	62.9	62.8	62.7	62.4	62.6	62.2
10	63.0	63.4	63.8	**64.0**	63.6	63.4	63.5
12	63.0	63.1	63.1	63.3	62.9	63.4	63.4
14	63.4	63.2	63.3	63.4	62.9	63.2	62.7
16	63.1	63.0	63.1	63.2	63.0	63.2	63.0

自适应细化机制。如表2所示，我们的锚点引导的自适应细化始终优于采用固定多样性阈值 τ的标准最大边际相关性实现，因为没有任何静态阈值能够在两个基准测试集上都具有良好的泛化能力。根据每个视频的独特内容动态调整多样性标准，可获得最佳结果（64.0/61.3%），证明了内容自适应细化的稳健性和有效性。

视觉与语义信号消融。 我们进一步对框架中的信号获取组件进行消融（表 4）。在事件边界检测方面，DINOv2 的表现始终优于 CLIP [30], ，这可能是因为它更擅长捕捉细粒度视觉变化，因此比 CLIP 的高层语义特征更准确地表征视觉事件。在相关性评分方面，BLIP2-ITM 明显优于 CLIP，凸显了专门进行图文匹配的优势。将 DINOv2 用于事件结构分析、BLIP2-ITM 用于相关性评分，能够取得最佳性能，这也印证了我们的设计选择。

超参数消融。 我们通过全面的网格搜索，研究了两个关键超参数：最大事件数 M 和松弛因子 α，搜索结果汇总于表5中。 M 控制视频划分的粒度。性能在 M=10时达到峰值，因为较小的值会导致事件划分较粗、出现欠分割，而较大的值则会造成过度划分和锚点噪声。与此同时， α 决定了自适应MMR中的多样性阈值。

结果表明，α=0.5 是最佳的中间值，因为它既能捕捉内在的视觉方差，又不会过于严格地过滤掉相关帧。这一最优配置（M=10, α=0.5）在 VideoMME 上取得了 64.0% 的准确率，达到当前最先进水平。

表 6. 使用 LLaVA-Video-7B 在 VideoMME 上的计算成本与准确率比较。我们报告各组件的单视频平均耗时（秒）。所有实验均在单张 NVIDIA A800 GPU 和两颗 Intel Xeon Silver 4314 CPU（16 核）上进行，采样率为 1 fps，输入帧数为 16。

组件	短	中	Long	总体
预处理时间（秒）：				
视频读取 1.45		9.40	48.94	19.93
视觉信号 1.16		7.45	33.75	14.12
语义信号 1.34		8.48	40.75	16.85
EFS 选择 0.03		0.16	0.95	0.38
预处理总耗时（秒）	4.0	25.5	124.4	51.3
大型视觉语言模型推理（秒）	2.39	2.99	3.51	2.97
准确率（%）	**76.33**	**62.67**	**53.00**	**64.00**
对比均匀采样	+4.22	+4.34	+1.78	+3.44

效率与成本效益分析。 为说明我们方法的效率，我们在表 6 中报告了其计算开销。尽管 EFS 会产生均匀采样所没有的预处理成本，但显著的 +3.44% 总体准确率增益足以证明这项开销是合理的。分析表明，主要的计算瓶颈在于视觉处理和语义信号提取步骤，而 EFS 选择阶段本身则非常轻量，耗时不到预处理总时间的 1%。这表明我们方法的核心逻辑具有很高的效率。通过投入适度的计算资源换取显著的准确率提升，这种权衡使 EFS 成为一种切实可行且经济高效的选择，适用于性能至关重要且能够进行离线处理的应用场景，例如深度内容分析。此外，这种开销并非固定不变，还可以大幅降低。未来可通过模型量化和硬件专用优化等方式，加速特征提取这一瓶颈环节。此外，对于成本最高的长视频，适度降低输入帧率即可在维持性能的同时显著缩短处理时间。

4.4. 可视化

图5 展示了来自 VideoMME 的两个定性示例，说明与均匀采样相比，我们的事件锚点帧选择（EFS）具有优势。在两个视频问答案例——瓶子 DIY 和手工教程中，均匀采样基线不考虑时间信息，因此会漏掉关键事件并给出错误回答。相比之下，EFS 利用图文匹配（ITM）分数和帧间相似度，揭示视频内容背后的事件结构，然后有策略地选择帧，以确保全面覆盖事件。因此，大型视觉语言模型（LVLM）能够获得准确推理并给出正确回答所需的证据。

5. 结论

本文提出了事件锚定帧选择（EFS），这是一种免训练帧选择方法，旨在提升大型视觉语言模型对长视频的理解能力。EFS 利用视频内在的事件结构，构建了一种针对事件覆盖、查询相关性和视觉多样性进行优化的关键帧表示。这种事件感知方法在多个主要基准测试上取得了显著增益，证实了其在有效长视频推理方面优于扁平且不考虑时间信息的采样方法。更多讨论和补充实验见附录。

参考文献

[1] Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ah- mad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。Gpt-4 08774，2023 2303 1技术报告。arXiv 预印本 arXiv: . . [2] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2. 5-vl 技术报告。arXiv 预印本 arXiv:2502.13923，2025。5 [3] Jaime Carbonell 和 Jade Goldstein。使用 MMR（基于多样性的重排序）重新排列文档并生成摘要。载于 ACM SIGIR，第 335–336 页，1998。4[4] Brandon Castellano。PySceneDetect。https://www. scenedetect.com/，2025。0.6.7 版本；在线；访问日期：2025-08-24。7[5] Lin Chen、Xilin Wei、Jinsong Li、Xiaoyi Dong、Pan Zhang、Yuhang Zang、Zehui Chen、Haodong Duan、Zhenyu Tang、Li Yuan 等。ShareGPT4Video：通过更优质的字幕提升视频理解与生成。载于 NeurIPS，第 19472–19495 页，2024。1[6] Chuanqi Cheng、Jian Guan、Wei Wu 和 Rui Yan。通过分层差分蒸馏将视频语言模型扩展至 1 万帧。arXiv 预印本 arXiv:2504.02438，2025。4[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。载于 ECCV，第 75–92 页。Springer，2024。2[8] Bo Fang、Wenhao Wu、Qiangqiang Wu、Yuxin Song 和 Antoni B Chan。将关键帧与叙事串联：多模态大语言模型是强大的长视频理解模型。arXiv 预印本 arXiv:2505.24158，2025。2，6

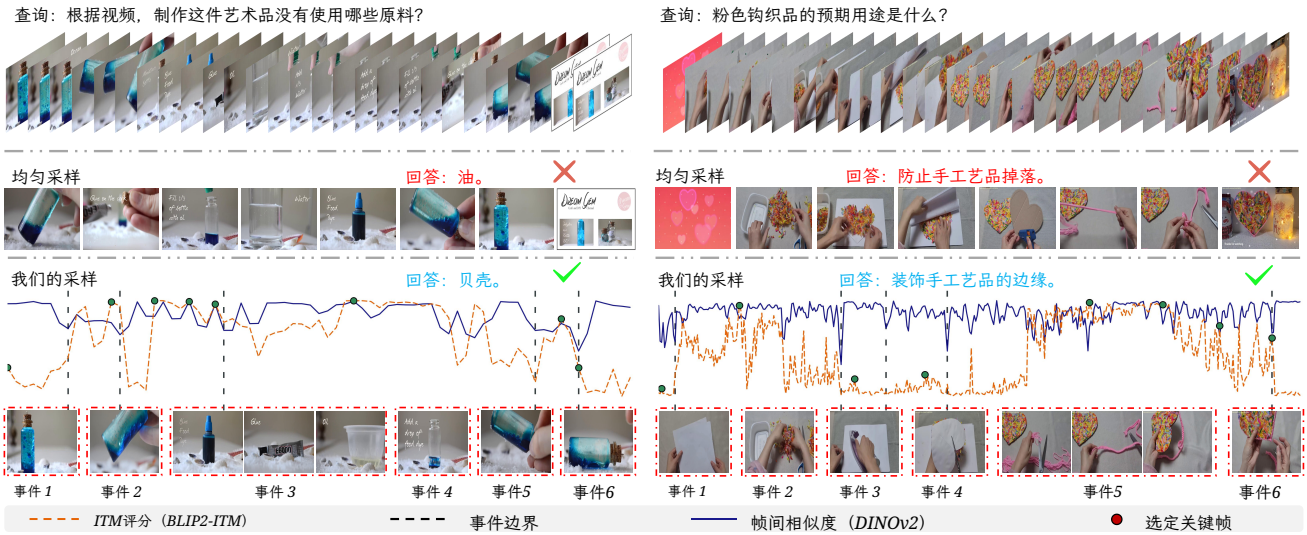


图5. 我们的 EFS 与均匀采样的定性比较。均匀基线因遗漏关键事件而失败，而 EFS 明确识别视频底层的事件结构，从而实现准确推理并给出正确答案。请放大查看细节。

[9] Chaoyou Fu、Yuhan Dai、Yongdong Luo、Lei Li、Shuhuai Ren、Renrui Zhang、Zihan Wang、Chenyu Zhou、Yunhang Shen、Mengdan Zhang 等。Video-MME：首个全面评测多模态大语言模型视频分析能力的评测基准。载于 CVPR，第 24108–24118 页，2025。2，5[10] De-An Huang 、Subhashree Radhakrishnan、Zhiding Yu 和 Jan Kautz。Frag：面向长视频与长文档理解的帧选择增强生成。arXiv 预印本 arXiv:2504.17447，2025。2[11] Bo Li、Yuanhan Zhang、Dong Guo、Renrui Zhang、Feng Li、Hao Zhang、Kaichen Zhang、Peiyuan Zhang、Yanwei Li、Zi-wei Liu 等。LLaVA-OneVision：轻松实现视觉任务迁移。arXiv 预印本 arXiv:2408.03326，2024。5[12] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。Blip-2：利用冻结图像编码器和大语言模型引导语言-图像预训练。载于 ICML，第 19730 19742 2023 3, 5, 1 – 页。PMLR，。[13] Yanwei Li、Chengyao Wang 和 Jiaya Jia。Llama-vid：在大语言模型中，一张图像胜过 2 个词元。载于 ECCV，第 323–340 页。Springer，2024。2[14] Bin Lin、Yang Ye、Bin Zhu、Jiaxi Cui、Munan Ning、Peng Jin 和 Li Yuan。Video-LLaVA：通过先对齐再投影学习统一的视觉表征。载于 EMNLP，第 5971–5984 页，2024。2，5[15] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令微调。载于 NeurIPS，第 34892 34916 页，2023 1 –。[16] Hao Liu、Wilson Yan、Matei Zaharia 和 Pieter Abbeel。利用分块环形注意力实现百万长度视频与语言的世界模型。arXiv 预印本 arXiv:2402 08268，2024 1，2 。[17] Shu ming Liu、Chen Zhao、Tianqi Xu 和 Bernard Ghanem。BOLT：无需训练即可增强长视频理解能力的大型视觉语言模型。载于 CVPR，第 3318 3327 页，2025 2，6 –。

[18] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yux- ian Gu, Dacheng Li 等。NVILA：高效的前沿视觉语言模型。载于 CVPR，第 4122–4134 页，2025年。5[19]Yongdong Luo, Wang Chen, Xiawu Zheng, Weizhong Huang, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Jiebo Luo 和 Rongrong Ji。Video-RAG：视觉对齐的检索增强长视频理解。arXiv 预印本 arXiv:2411.13093，2024年。2[20] Yongdong Luo, Wang Chen, Xiawu Zheng, Weizhong Huang, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Jiebo Luo 等。Quota：通过思维链查询解耦实现面向查询的长视频理解令牌分配。arXiv 预印本 arXiv:2503.08689，2025年。2[21] Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi 和 Jianfei Cai。DrVideo：基于文档检索的长视频理解。载于 CVPR，第 18936–18946 页，2025年。1, 2[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan 和 Fahad Khan。Video-ChatGPT：通过大型视觉语言模型实现细致的视频理解。载于 ACL，第 12585–12602 页，2024年。2[23] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho 和 Cordelia Schmid。MoreVQA：探索用于视频问答的模块化推理模型。载于 CVPR，第 13235–13245 页，2024年。1[24] Azra Nasreen 和 G Dr Shobha。使用边缘变化率进行镜头分割的关键帧提取。IJARCCE，2(11)：4421–4423，2013年。2[25] OpenAI。GPT-4o mini：推进高性价比智能。https : / / openai . com / index / gpt - 4o - mini - advancing - cost - efficient - intelligence/，2024年。访问日期：2024年6月18日。5[26] OpenAI。Hello GPT-4o。https : / / openai . com /

index/hello-gpt-4o/，2024年。访问日期：2024年5月13日。5[27] Maxime Oquab, Timoth´ee Darcet, Th´eo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby 等。DINOv2：无需监督地学习鲁棒视觉特征。arXiv 预印本 arXiv:2304.07193，2023年。3, 5, 1[28] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim 和 Michael S Ryoo。帧数过多并不意味着都有用：长视频问答的高效策略。arXiv 预印本 arXiv:2406.09396，2024年。2[29] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat- Seng Chua, Yueting Zhuang 和 Siliang Tang。Momentor：以细粒度时间推理增强视频大语言模型。载于 ICML，第 41340–41356 页，2024年。1[30]Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark 等。从自然语言监督中学习可迁移视觉模型。载于 ICML，第 8748–8763 页。PMLR，2021年。7[31] C Va Sheena 和 NK Narayanan。通过统计方法分析视频帧直方图来提取关键帧。Procedia Computer Science，70：36–40，2015年。2[32] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Bal- akrishnan Varadarajan, Florian Bordes 等。LongVU：面向长视频语言理解的时空自适应压缩。arXiv 预印本 arXiv:2410.17434，2024年。2, 5[33] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Jun- jie Zhou, Zhengyang Liang, Tiejun Huang 和 Bo Zhao。Video-XL：用于小时级视频理解的超长视觉语言模型。载于 CVPR，第 26160–26169 页，2025年。5[34] Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang 和 Ming Li。MDP3：一种无需训练的视频大语言模型列表式帧选择方法。arXiv 预印本 arXiv:2501.02885，2025年。2[35] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao 和 Qixiang Ye。用于长视频理解的自适应关键帧采样。载于 CVPR，第 29118–29128 页，2025年。2, 6[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge 等。Qwen2-VL：增强视觉语言模型对任意分辨率下世界的感知能力。arXiv 预印本 arXiv:2409.12191，2024年。1[37]Haoning Wu, Dongxu Li, Bei Chen 和 Junnan Li。LongVideoBench：长上下文交错视频语言理解基准测试集。载于 NeurIPS，第 28828–28857 页，2024年。2, 5[38] Zeyu Xu, Junkang Zhang, Qiang Wang 和 Yi Liu。E-vrag：通过资源高效的检索增强生成提升长视频理解。arXiv 预印本 arXiv:2508.01546，2025年。2[39] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chan- drasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei 等。重新思考长视频理解中的时序搜索。载于 CVPR，第 8579–8591 页，2025年。2[40] Sicheng Yu、Chengkai Jin、Huanyu Wang、Zhenghao Chen、Sheng Jin、Zhongrong Zuo、Xiaolei Xu、Zhenbang Sun、Bingni Zhang、Jiawei Wu 等。Frame-voyager：学习为视频大语言模型查询帧。*arXiv* 预印本 *arXiv:2410.03226*，2024年。2，5[41]Kaichen Zhang、Bo Li、Peiyuan Zhang、Fanyi Pu、Joshua Adrian Cahyono、Kairui Hu、Shuai Liu、Yuanhan Zhang、Jingkang Yang、Chunyuan Li 等。LMMs-Eval：大型多模态模型评估的现实检验。发表于*NAACL*，第 881–916 页，2025年。5[42] Peiyuan Zhang、Kaichen Zhang、Bo Li、Guangtao Zeng、Jingkang Yang、Yuanhan Zhang、Ziyue Wang、Haoran Tan、Chunyuan Li 和 Ziwei Liu。将长上下文从语言迁移到视觉。*arXiv* 预印本 *arXiv:2406.16852*，2024年。1，2，5[43] Shaojie Zhang、Jiahui Yang、Jianqin Yin、Zhenbo Luo 和 Jian Luan。Q-frame：面向视频大语言模型的查询感知帧选择与多分辨率自适应。*arXiv* 预印本 *arXiv:2506.22139*，2025年。2[44] Yuanhan Zhang、Jinming Wu、Wei Li、Bo Li、Zejun Ma、Ziwei Liu 和 Chunyuan Li。利用合成数据进行视频指令微调。*arXiv* 预印本 *arXiv:2410.02713*，2024年。5[45] Junjie Zhou、Yan Shu、Bo Zhao、Boya Wu、Shitao Xiao、Xi Yang、Yongping Xiong、Bo Zhang、Tiejun Huang 和 Zheng Liu。MLVU：多任务长视频理解综合基准测试集。*arXiv* 预印本 *arXiv:2406 04264*，2024年2，5。。

事件锚 定帧选择，以有效理解长视频 g

补充材料

6. 局限性与未来工作

6.1. 局限性

尽管已证明其有效性，我们提出的事件锚定帧选择（EFS）框架仍存在若干固有局限，值得进一步讨论。

• **预处理开销**。 预处理开销。我们的流程需要一个基于预训练 DINOv2 [27]和 BLIP2-ITM [12], 的离线信号提取阶段，会产生均匀采样所避免的计算成本。对于长视频，生成这些信号会造成沉重的预处理负担，使 EFS 在对延迟敏感或实时场景中难以实用。

• **启发式与超参数敏感性**。 EFS 是一个多阶段的启发式驱动流程，其性能取决于若干关键超参数，尤其是事件数量 M 以及松弛因子（α）。尽管我们通过网格搜索确定了可行配置，但这些设置可能无法跨视频领域或下游任务迁移。当前框架缺乏一种有原则的数据驱动方法来实现超参数的自动自适应，从而使性能对人工选择较为敏感。

• **对上游模型的依赖**。 EFS 的性能本质上与基础模型 DINOv2 和 BLIP2-ITM 的表征能力紧密耦合。因此，EFS 不仅继承了它们的能力，也继承了它们的局限。这些上游模型中的任何固有偏差、领域差异或灾难性故障都可能向下游传播，从而对最终帧选择的质量产生不利影响。

6.2. 未来工作

我们当前方法的局限性凸显了若干有前景的未来研究方向。

• **系统级效率提升**。 首要目标是降低所提出框架的端到端延迟。未来工作将聚焦于整个流程的整体系统级优化。这包括加速数据摄取与处理、设计更高效的计算路径，以及探索硬件感知执行策略，以在严格的资源约束下确保稳定、低延迟的性能，同时不牺牲选择质量。

•**全模态事件理解**。 为了实现更稳健且能感知上下文的分析，我们计划突破纯视觉线索的局限，开发一个全模态框架。这包括整合多种数据流，例如高层语义、运动动态、音频信号和文本转录，并将其纳入统一表征。通过融合这些多模态信号，模型能够学习识别复杂的事件边界，并消除仅凭视觉信息无法察觉的歧义，从而更全面地理解视频内容。

• **端到端可训练选择**。一项重要目标是用完全端到端可训练的模型替代当前基于启发式的流程。为此，需要设计一个可微选择模块，通过直接优化下游任务目标来学习识别关键内容。这种数据驱动的方法不仅可以免去手动超参数调优，还可能发掘出比手工设计流程更有效、更具自适应性的选择策略，最终提供更强大、泛化能力更强的解决方案。

7. 更多实验结果

7.1. 在VideoMME的不同子任务上进行比较

为细粒度评估我们的事件锚定帧选择（EFS）方法，我们在VideoMME基准上将其与均匀采样基线进行了全面比较分析。两种方法均使用LLaVA-Video-7B模型进行评估，输入帧数限制为八帧。该设置能够详细评估模型在全部12个子任务上的表现，并揭示EFS在哪些特定领域具有最显著的优势。

如图6所示，EFS 在每项子任务上的表现均优于基线，这凸显了其广泛的效用，而非仅对某一特定任务有益。在需要较高感知敏锐度和时间意识的任务中，增益最为显著，包括空间感知（+11.11%）、时间感知（+9.09%）和 OCR 问题（+7.91%）。这凸显了 EFS 在捕捉并保留关键场景语境、时序顺序和细粒度视觉细节方面的优势。此外，在空间推理（+5.35%）、动作推理（+4.92%）和时间推理（+4.52%）等复杂高层次任务上的显著提升，进一步印证了我们方法的泛化能力。从低层次感知到高级推理，在这类多样化任务中始终如一的性能提升，验证了 EFS 是一种稳健有效的长视频理解框架，尤其适用于上下文窗口有限这一实际限制。

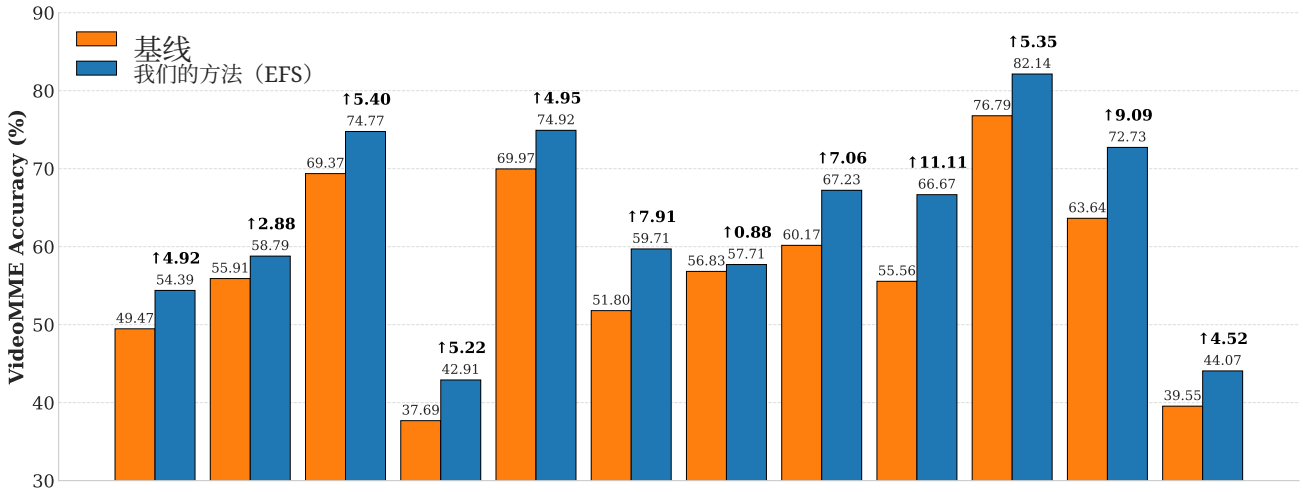


图6. 我们的方法（EFS）与基线在 Video-MME 基准测试集全部 12 个子任务上的详细性能对比 [9]。结果表明，我们的 EFS 方法在每个类别中均持续优于基线，凸显了该方法在提升多样化视频理解能力方面的广泛有效性。

表7. 窗口大小消融 l。我们报告了 VideoMME、LVB 和 MLVU 基准测试上的基于视频的问答准确率（%）。最佳结果以**粗体**标出。

方法	VideoMME	LVB（验证集）	MLVU
均匀采样	60.6	57.2	60.9
EFS(l=3)	**64.0**	61.3	**69.0**
EFS(l=5)	63.3	**61.7**	68.0
EFS(l=7)	63.4	60.9	68.6

7.2. 窗口大小消融研究

窗口大小（记作 l）是我们提出的事件锚定帧选择（EFS）框架中的关键超参数，因为它决定了计算帧间相似度的时间范围。选择 l 需要权衡：较小的窗口对快速、细粒度的场景变化更敏感，而较大的窗口则能更稳健地衡量局部时间连贯性。为系统地研究这一权衡并确定最优配置，我们开展了消融实验，将 l 分别设为 3、5 和 7。我们在三个不同的视频理解基准测试上，将各个变体的表现与均匀采样基线进行比较：VideoMME [9], LongVideoBench（LVB）[37], 和 MLVU[45]。

如表7所示，EFS 的所有配置均显著优于均匀采样基线，证实了内容感知选择的根本优势。

在测试的变体中，窗口大小为 l=3 时总体表现最佳，在 VideoMME 和 MLVU 上分别取得了 64.0% 和 69.0% 的最先进结果。尽管 l=5 设置在 LVB 验证集上略占优势，但其在其他基准测试集上的表现有所下降。将窗口增大至 l=7 会导致准确率更明显地下降。这些结果充分表明，紧凑的窗口（l=3）最能有效捕捉对下游理解任务至关重要的显著短期视觉动态。相比之下，较大的窗口可能会过度平滑时间信号，从而掩盖重要的转折时刻。因此，在后续所有实验中，我们均采用 l=3 作为 EFS 框架的默认窗口大小。

7.3. 超参数消融研究

为剖析本框架关键组件的贡献，我们对两个关键超参数进行了全面的消融实验：自适应 MMR 算法中使用的最大视频事件数 M，以及阈值松弛因子 α。我们对这些参数进行网格搜索，并报告在 VideoMME、LongVideoBench（LVB）和 MLVU 基准测试集上的准确率。

事件粒度的影响（M）。超参数 M 决定初始视频划分的粒度。M 的最优取值必须取得平衡：若 M 过小，不同的语义事件可能被合并，从而导致粗粒度感知以及较差的初始锚点集合。

相反，若 M 过大，单个事件可能被过度划分，从而引入噪声或冗余锚点，削弱后续的精炼过程。

如表5所示，我们在VideoMME上的实证结果证实了这一权衡。随着 M 从1增加到10，性能稳步提升，并在准确率为**64.0%**时达到峰值。然而，进一步将 M 增加到12及以上会导致性能出现轻微但持续的下降。在LongVideoBench和MLVU数据集（表8）上也观察到类似趋势，当 M 处于12到14的范围内时，性能达到峰值。这一跨基准测试的一致模式验证了我们的假设：中等事件粒度对于构建具有代表性且有效的帧选择结构先验至关重要。

阈值松弛因子（α）的影响。 松弛因子 α控制我们自适应MMR框架中多样性阈值的灵活性。较小的 α 会强制形成狭窄而僵化的多样性范围，可能无法适应视频固有的视觉方差。相反，较大的 α 会使相似度阈值过于宽松（或过于严格，取决于具体形式），从而可能滤除相关帧并降低覆盖率。

有趣的是， α 的最优值因数据集而异。对于VideoMME（表5），适中的α=0.5 能实现最佳平衡，使模型能够根据视频的统计特征动态调整。然而，对于LongVideoBench 和 MLVU（表8），较小的 α=0.1 则始终能取得最佳结果。这表明，LVB 和 MLVU 中的视频可能具有更均匀的事件结构或统计特性，因此更适合采用严格且适应性较弱的多样性准则。LVB 的最优配置为（M=14， α=0.1），准确率达到**61.3%**；MLVU 的最优配置为（M=14， α=0.1），准确率达到**69.0%**。

这些发现表明，尽管我们的超参数总体趋势一致，但其最优值会受到各数据集独特的统计特征和事件结构的影响。这凸显了利用数据感知先验有效引导帧选择过程的重要性。

7.4. 对信息概要查询的泛化能力

我们的 EFS 方法本质上是基于查询的，因此引出了一个问题：它对信息概要查询（例如“这个视频描述了什么？”）是否有效。直观而言，对于这类缺乏特定语义信号的宽泛问题，EFS 以查询为中心的锚点选择可能并无助益，而更简单的均匀采样方法或许就足够了。

为探究这一问题，我们开展了一项实验，以确定大型视觉语言模型能否动态选择合适的

表8. M 和 α的消融实验。结果以 LongVideoBench/MLVU 上的准确率（%）表示。

M \ α	0.1	0.3	0.5	0.7
1	59.5/68.2	58.7/68.0	58.3/67.6	58.8/67.6
4	60.0/67.9	59.6/67.6	59.8/68.3	59.6/67.8
8	59.7/68.0	60.0/68.0	59.7/67.5	59.0/67.9
10	60.0/68.6	60.0/68.5	60.3/67.7	60.4/67.7
12	60.6/**69.0**	59.8/68.0	60.4/68.0	59.9/68.1
14	**61.3/69.0**	60.4/68.6	61.1/68.5	60.3/68.7
16	60.1/67.8	59.9/67.8	59.9/67.8	60.1/67.8

采样策略。我们实现了一个基于思维链的路由器：对于每个查询，首先提示 LLaVA-Video 模型将其分类为具体且有针对性的查询，或一般且侧重概要的查询。用于此分类的确切提示详见7.5。如果模型将查询识别为具体查询，我们就采用 EFS 方法；如果模型认为查询较为一般，系统则会改用标准均匀采样策略。我们将这种混合方法称为“部分 EFS”。所有实验均采用最多输入 16 帧的限制。

结果见表9。与直觉相反，自适应“部分 EFS”策略在三个基准测试集上的表现始终不如完整 EFS 方法。例如，在 VideoMME 上，对一般查询改用均匀采样，准确率比始终使用 EFS 低了 1.5%。这一性能下降表明，对一般查询强制回退到均匀采样是一种有害的选择。结果有力地说明，EFS 提供的结构先验和更全面的事件覆盖，即使对于信息摘要任务也有益，其表现优于均匀采样在时间维度上不加区分的特性。图6所示的信息摘要子任务上的性能增益，进一步支持了这一结论。

表 9. 基于 CoT 的自适应采样策略的性能比较。“部分 EFS”性能下降凸显了 EFS 的普遍有效性，即使对于系统回退到均匀采样的一般查询也是如此。

方法	VideoMME	LVB	MLVU
LLaVA-Video（均匀采样）	60.6	57.2	60.9
LLaVA-Video +EFS	64.0	61.3	69.0
LLaVA-Video +部分 EFS	62.5(-1.5)	60.7(-0.6)	68.6(-0.4)

7.5. 更多定性示例

为进一步验证我们事件划分的合理性，图7展示了从采样视频帧中提取的 DINOv2 特征的 t-SNE 可视化结果。清晰的聚类

以及点之间的分离表明，我们的方法能够将视觉上一致的事件归为一组，并识别有意义的事件边界。这为后续关键帧选择奠定了可靠的基础。

系统提示：视频查询分类器

你是一名精通视频分析的专家助手。你的任务是判断用户的问题是否需要采用基于查询的关键帧选择方法。此方法旨在查找视频中的特定时刻、对象或动作。

分析用户的问题。请逐步思考，判断这是一个具体查询还是一般性问题。完成推理后，请在新的一行给出最终答案。最终答案必须仅为“Yes”或“No”。“Yes”表示该问题需要进行具体、有针对性的搜索。“No”表示该问题较为一般，旨在询问摘要或总体概况。

示例 1：

问题：“总结一下这个视频的内容。”推理：这是一个要求提供总体概述的高层次请求，不需要定位某个特定时刻。

最终答案：否

示例 2：

问题：“找到猫撞倒花瓶的场景。”推理：这是要定位一个非常具体的动作（“撞倒花瓶”），涉及特定对象（“猫”“花瓶”），因此需要定向搜索。最终答案：是

示例 3：

问题：“说话者的总体情绪基调如何？”推理：这个问题关注的是视频全程的整体基调，而不是某个瞬间。这是一种整体性分析。最终答案：否

示例 4：

问题：“首席执行官何时出现在屏幕上？”推理：这需要识别特定人物（“首席执行官”）可见的所有特定时刻。这是一项查询特定的任务。最终答案：是

你的任务：

问题：{你的问题}推理：

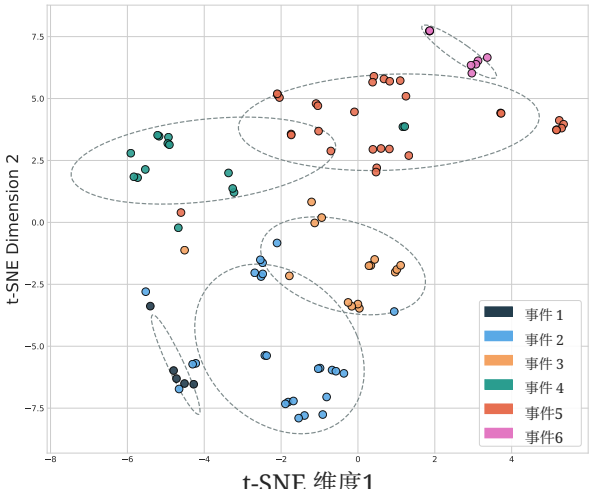


图7. 采样视频帧的 DINOv2 特征的 t-SNE 可视化。每个点代表一帧，颜色表示我们的 EFS 流水线预测的事件。