查询到事件分解用于零样本多语言文本到视频检索

Shubhashis Roy Dipta, Francis Ferraro
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, MD 21250 USA
{ sroydip1,ferraro } @umbc.edu

Abstract

最近的方法展示了从大型语言模型（LLMs）和视觉语言模型（VLMs）中提取和利用参数化知识的惊人能力。在这项工作中，我们考虑如何通过自动提取有关复杂现实世界事件的潜在参数知识来改进视频的识别和检索。我们提出 **Q$_2$E**：一种针对零样本多语言文本到视频检索的 Q 问询到 E 事件分解方法，可在数据集、领域、LLMs 或 VLMs 之间进行适应。我们的方法表明，通过使用嵌入在 LLMs 和 VLMs 中的知识来分解查询，我们可以增强对人类查询的理解，避免其过于简化。此外，我们展示了如何将我们的方法应用于视觉和基于语音的输入。为了结合这些多样的多模态知识，我们采用基于熵的融合评分用于零样本融合。通过对两个不同的数据集和多个检索指标的评估，我们证明 **Q$_2$E** 优于多个最先进的基准。我们的评估还表明，整合音频信息可以显著提高文本到视频的检索效果。我们已发布代码和数据以供未来研究。[1]

1 引言

广泛使用的视频检索网站可能会给人一种视频检索问题已经解决的印象：用户输入一个查询，就会得到一份精心整理的相关视频列表。然而，这些方法可以利用丰富的元数据或后续创作者的人工整理（例如标题或经过优化的描述），这些信息超出了视频中直接传达的信息。例如，如 Fig. 1 所示，有许多不同的火灾视频，但如果用户提供一个基本的查询，如“2025 LA 火灾”，以查找提供该特定火灾信息的视频，系统如何满足这一请求？在本文中，我们提出了一种方法，称为 **Q$_2$E**（查询到事件的分解），它利用了 LLM 关于复杂事件的嵌入知识，并使用这些知识自动分解和详细说明用户的查询，同时利用这些知识识别视频中突出的视觉和音频特征。最近的研究 (Liu et al., 2024b) 发现，用户通常输入基本查询，期望系统能够捕捉“细粒度的语义概念”并理解视频与文本之间的复杂关系。我们提出的方法将查询分解为多个细粒度事件，提供了对查询隐含信息请求的关键详解。此外，目前大多数文本到视频的检索系统是在广泛使用的数据集上进行训练和评估的，例如 MSR-VTT (Xu et al., 2016) 和 MSVD (Chen and Dolan, 2011)，这些数据集的特征是通用的、高层次的查询，如“一个人在解释某事”（一个来自 MSR-VTT 的随机查询）。这些通用查询限制了文本到视频检索系统的学习能力，阻碍它们根据精确的用户查询有效检索特定视频，例如“2025 年洛杉矶火灾”。最后，大多数文本到视频检索数据集集中在英语视频上，而视频超越了语言的界限，这需要检索方法能够适应其他语言。

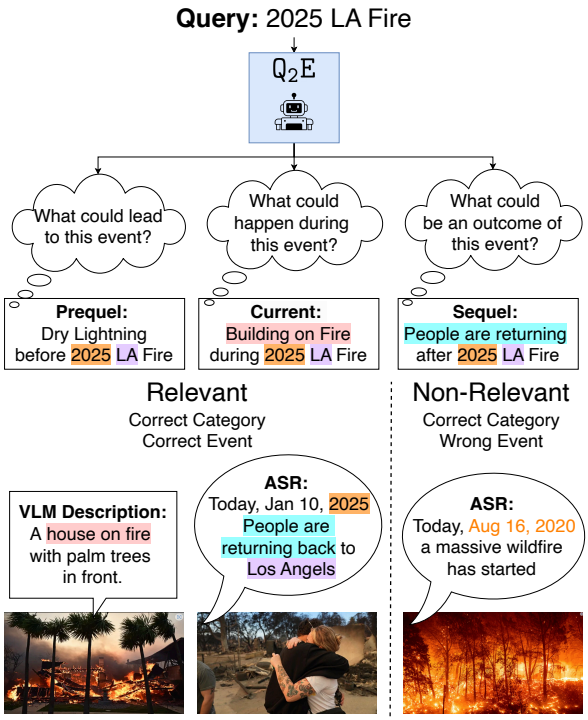

Figure 1: 在 **Q$_2$E** 中，我们提取前传、当前和后续事件，以及音频转录和视频描述，以丰富查询上下文并增强视频检索。该信息与视觉、文本和语音描述相匹配（匹配的短语用相同颜色高亮显示），从而能够检索到正确的视频，同时有效过滤掉视觉上相似但不相关的视频。

[1] https://github.com/dipta007/Q2E

受到这些挑战的启发，我们探索如何使用查询扩展以及视频 + 音频特征来在不同语言中检索高度具体的视频。此外，我们还探讨了如何将大语言模型（LLM）和视觉语言模型（VLM）中嵌入的大量知识有效地转移到文本到视频的检索中。为此，我们提出了一种新颖的零样本文本到视频检索方法，该方法整合了 LLM 和 VLM 的预先存在的知识，以增强文本到视频的检索能力。

我们的方法基于三个关键见解：使用这些见解，Q$_2$E 在两个不同且具有挑战性的数据集上取得了显著的性能提升。

总之，我们的贡献如下：

- 我们提出了一个新颖的框架，利用大型语言模型中的嵌入知识来丰富人类查询，包括前传、当前和续集事件。
- 我们使用大型视频、语音和语言模型来捕捉通过这些多模态传达的关键事件和查询相关信息，如何评估这些信息与查询的相关性，以及如何通过基于熵的排名融合来组合这些评估。
- 通过广泛的消融研究，我们展示了我们方法中每个组件在最终性能中的重要性。
- 我们的零样本方法提高了在广泛使用的文本到视频检索数据集 MSR-VTT (Xu et al., 2016) 上的性能（相比我们的基线增加了 4 个 NDCG 点），以及在更丰富的多语言事件聚焦检索数据集 MultiVENT (Sanders et al., 2023) 上的性能（增加了 8 个 NDCG 点）。

随着 LLMs (Devlin et al., 2019; Radford et al., 2019; Touvron et al., 2023; Dubey et al., 2024; Gunasekar et al., 2023; Liu et al., 2024a; Jiang et al., 2024; Chu et al., 2024) 的快速进步，由于其较低的计算需求和提高的时间效率，诸如零样本学习、小样本学习和上下文学习的方法得到了极大的重视。这些方法已经影响了事件因果推理、多模态语言模型和检索的方法。

1.1 事件因果推理

基于事件因果的推理在多个领域起着至关重要的作用。Sun et al. (2024) 证明，即使是由 LLM 生成的事件因果图，也可以增强故事理解并更好地与人类评估对齐。同样，在 Q$_2$E 中，我们利用人类查询的因果关系构建前传、当前和续集事件，从而更深入地理解用户意图。虽然 Chan et al. (2024) 使用 LLM 推断两事件之间的时间、因果和话语关系，但我们的方法应用相同的时间（前传或续集）和因果（当前）推理来直接从查询中提取相关事件。

1.2 视觉语言模型

视觉语言模型（VLMs）通过利用在图像-文本对上训练的大规模预训练模型，在文本到视频检索任务中得到了显著应用。这些模型展示了令人印象深刻的零样本能力和视频相关任务的迁移学习潜力。最近的进展集中在如何高效地将基于图像的 VLMs 适应到视频领域，以应对时间动态和计算成本的挑战。一些方法提出了新的架构，如稀疏且相关的适配器，以提高检索性能的同时保持效率。此外，研究人员还探索了结合音频信息和使用合成的教学数据进一步改进视频语言模型。

虽然在微调的文本到视频检索领域已经做了很多工作 (Wang et al., 2025; Cicchetti et al., 2024; Chen et al., 2023; Xu et al., 2023)，但在零样本文本到视频检索性能上存在差距。在任务特定的数据集上微调模型可以取得令人印象深刻的性能，但由于高计算成本和时间限制，这并不总是适合的。尽管之前已经研究过零样本文本到视频检索 (Wang et al., 2025; Cicchetti et al., 2024; Chen et al., 2023; Xu et al., 2023)，结果表明与微调检索之间存在显著差距。

MSR-VTT (Xu et al., 2016) 是一个著名的基准测试。然而，数据集中的查询往往更为通用，例如，“背景中有一场龙卷风和雷暴”，而不是针对特定事件的丰富信息，如 Fig. 1 中的 “2025 LA 火灾”。相比之下，MultiVENT (Sanders et al., 2023) 定义了一项多语种视频检索任务，其中查询指向具体事件。

2 Q$_2$E: 查询到事件的分解

现代密集检索方法通常计算查询和每个候选之间的基于嵌入的相似性（例如，余弦相似性）；这些分数用于对候选进行排名，并返回得分最高的候选。例如，在视频检索中，可以使用适当的 LLM 对用户的查询进行嵌入，使用适当的视频嵌入器对视频进行嵌入，然后计算这些嵌入之间的余弦相似性 (Sanders et al., 2023)。这种 “查询与视频” 的相似性评分被用作视频与查询相关程度的代理。

我们的方法（如 Fig. 2 所示）基于这种通用的嵌入相似性计算，不过为了更好地捕捉视频和人类查询中所传达的多样化和细微的多模态信息——包括由 Q$_2$E 产生的事件分解——我们必须扩展这些相似性计算。为了捕捉我们从分解中获得的丰富事件信息（**??**），单一的表示视频的嵌入是不够的。我们还需要获取视频的多模态描述（**??** and §2.1）；这些是基于文本的视频和音频描述，可以更容易地与查询进行评估。利用这些组件，Q$_2$E 为每个视频 v 计算五种类型的分数 $S_{i,v}$：

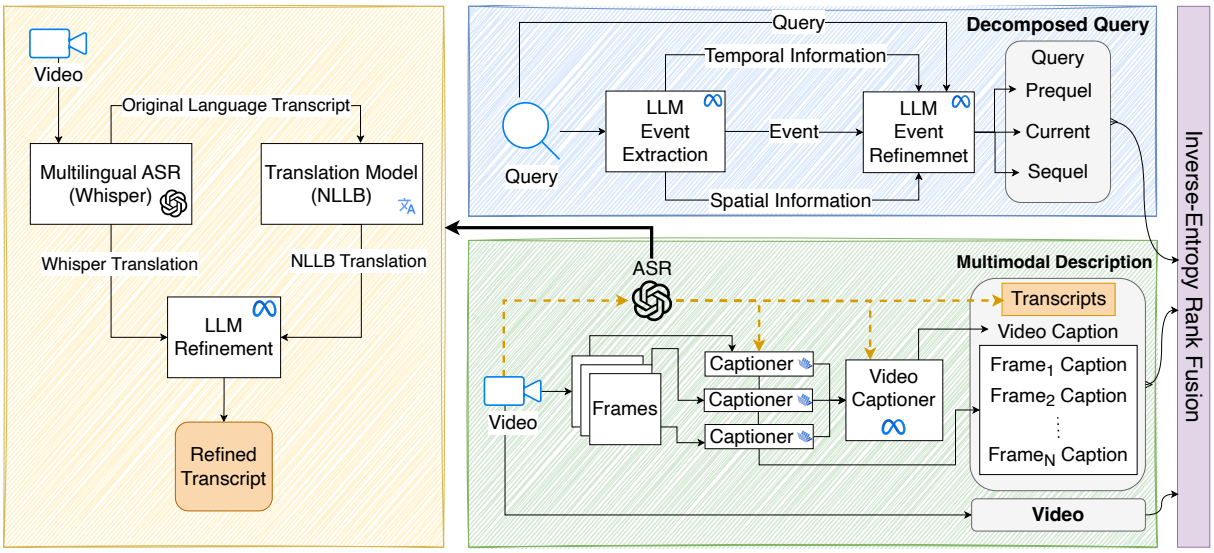

Figure 2: 我们文本到视频检索系统的完整方法。顶部的蓝色框代表我们方法的查询分解模块，而底部的绿色框展示了视频字幕和音频转录模块。左边的橙色框表示我们的多层 ASR 模块。右边的紫色框融合了从查询-视频、查询-描述和事件描述（事件 = 前传 + 当前 + 续集，描述 = 多模态描述）计算得出的排名。文本到文本的相似度分数是使用 ColBERT (Khattab and Zaharia, 2020) 计算的，文本到视频的相似度分数是使用 MultiCLIP (Delitzas et al., 2023) 计算的。在我们方法的非 ASR 变体中，橙色框和橙色虚线中的组件被排除。

[(a)]正如上面所述，查询与视频的评分；每个前传、当前和续集分解的事件分解对比多模态描述评分；以及 一个查询与多模态描述评分。

。虽然 (a) 有助于识别诸如“火灾”或“洪水”之类的粗粒度特征，但 (b) 和 (c) 都提供了细粒度的上下文细节。通过使用基于熵的融合（§2.2）聚合这些组件，它们相互补充以改善特定视频事件的检索。

类似于任何传统的文本视频检索系统，用户提供一个自然语言查询，这个查询可以是几个词到 10 多个词。在这种方法中，我们使用了一个大型语言模型（LLM）将其分解为三个子事件：（1）序幕，（2）当前，（3）续集。基于初步实验，为了管理计算，我们仅考虑了五个序幕、当前和续集事件。我们在 App. A.3 中提供了每个事件的模板提示。

前传 “前传”主要指的是那些可能导致查询中所描述的当前事件的事件。一个事件可以有多个前传；例如，野火可能有强风、高温警报和干雷电作为前传事件。前传事件可能会使查询中的事件能够发生，但不一定会导致它。

当前 “当前”主要指发生在查询中描述的事件期间的简单分解事件。这种方法将查询分解为可能被观察到的子事件。例如，“2025 年洛杉矶大火”可以分解为“建筑物着火”或“到处都是烟”。

续集 “续集”事件可能由查询中描述的事件导致。就像前传一样，一个事件可以导致多个续集事件。

分解细化 我们使用时间、空间和主要事件数据来细化这些事件分解。为此，我们使用相同的大型语言模型 (LLM) 提取时间 (2025 年)、空间 (LA) 和主要事件 (火灾) 信息。虽然这些丰富的信息可以模板化地结合，例如 “P/S/C (2025, LA, Fire)”，但在初步开发实验中，我们发现对其进行细化以使其读作更自然的、人工提供的查询显著提高了性能。我们使用 LLM 自动生成了这种细化的查询；我们在 App. A.3.3 和 **提供所有响应作为我们匿名数据发布的一部分** 中展示了这一点。[2]

在实践中，我们发现通过基本的相似性计算，视频/视频帧的嵌入不足以捕捉由我们的事件分解所描述的各种动作。虽然更高级的相似性方法（如交叉编码器）可能是可行的，但它们带来了计算方面的挑战（需要对每个视频/查询对运行交叉编码器）。为了解决这个问题，我们考虑如何直接从视频中提取文本描述（字幕），目标是让这些描述与我们的查询事件分解足够相似。我们发现，两种方法在捕获视频中的互补局部和全局元素方面效果良好：(1) 为选定的帧生成情境化的描述，以及 (2) 生成整体视频描述。

[2] https://github.com/dipta007/Q2E

情景化框架描述 我们从视频中抽取了 K=16 帧（通过权衡性能与计算得出的结论——详见 App. A.1.2 中的消融实验）。我们尝试了均匀采样和基于场景变化的采样方法 [3] 进行帧采样，发现尽管均匀采样很简单，但效果优于基于场景变化的采样（详见 App. A.1.3 中的消融实验）。由于视频包含时间信息，并且设计上视频帧在语义上应该是连贯的，因此我们没有孤立地为每一帧生成字幕，而是使用了一种滑动窗口（窗口大小 =1）的方法，并提供给 VLM 模型之前帧的字幕（如果可用）。基于之前帧的字幕进行条件处理有助于后续字幕的上下文化。每个视频，我们都有 K 个上下文化的帧字幕。

我们发现当前的 VLM 模型能够进行密集的文字描述，但却忽略了“更大的图景”。例如，一个 VLM 可能会描述一个有野火的视频，其中有人在奔跑，火焰四处蔓延，但未能识别出更高层次的“野火”结论。为了应对这个限制，我们提供了一个 LLM 模型，并给出了上面所有 16 帧的描述，然后要求它将这些描述总结成一个单一的视频密集文字描述，同时保留帧描述的顺序。

2.1 ASR 增强视频字幕

语音是许多视频的重要组成部分，尤其是在仅视觉元素可能不足以传达完整语境的时候。尽管使用了多语言 ASR 模型 (Radford et al., 2023)，多语言视频仍然面临着获得一致且准确的 ASR 转录的独特挑战。为了解决这个问题，我们提出了一种多层翻译流程 (Fig. 2，左侧橙色框)。该流程利用了三个模型：(1) 多语言 ASR 模型，`whisper-v3` (Radford et al., 2023)，将音频转录为原始语言和其英文翻译；(2) 翻译模型，`NLLB` (Costa-jussà et al., 2022)，通过翻译原始转录本生成英文转录本；以及 (3) LLM，`Llama-70B` (Dubey et al., 2024)，改进前两个模型生成的英文翻译。早期实验表明，所有三个模型对于获得足够的转录和翻译质量都是至关重要的。

2.2 评分和评分融合

密集视频嵌入 我们利用原始查询和视频来计算相似度分数。选择 MultiCLIP (Delitzas et al., 2023) 模型是因为与 CLIP (Radford et al., 2021) 模型相比，它在多语言视频上的表现更为优越。尽管 MultiCLIP 传统上是为图像设计的，我们通过遵循 Sanders et al. (2023) 的方法将其适配用于视频：均匀采样 16 帧并平均它们的嵌入以生成最终的视频嵌入。采用余弦相似度来计算相似度分数，然后将其缩放到 0-100 范围内，以确保与其他分数的可比性：$S_{i,v}|Q = 100.0 \times \text{multiCLIP}(Q, V)$，其中 $S_{i,v}|Q$ 表示给定查询 Q 的分数向量，$\text{multiCLIP}(x, y)$ 返回 x 和 y 之间的相似度分数。

查询与多模态描述 在该方法中，我们使用了原始查询，但将视频内容替换为多模态描述集，从而获得文本对文本的相似性分数，而不是文本对视频的分数。具体来说，我们计算查询与所有多模态描述之间的最大相似性。形式上：

$$S_{i,v}|Q = \max_{c=1:C} Sim(Q, Cap(v)_c), \quad (1)$$

，其中 C 是与视频相关联的描述的数量，$Cap(v)$ 返回视频 v 的多模态描述。$Sim(x, y)$ 计算 x 和 y 之间的 ColBERT [4] 相似度分数。

我们使用查询中的精炼前篇/当前/后篇事件和多模态描述集来计算多对多相似度分数。具体来说，我们计算每个事件对于所有描述的最大相似度分数，然后取所有事件中的最大相似度分数作为最终分数。这种方法通过利用全局最大值而不是平均值来减少单个噪声分解或字幕的负面影响，平均值更容易受到噪声数据的影响。形式上：

$$S_{i,v}|Q = \max_{l=1:L} \max_{c=1:C} Sim(E_i(Q)_l, Cap(v)_c), \quad (2)$$

，其中 L 是查询中事件（前篇/当前/后篇）的数量，而 $E_i()$ 返回相应的事件分解（前篇/当前/后篇）。

由于 **Q₂E** 是零样本的，将所有分数结合起来是一个挑战。简单地对所有分数取平均会假设每个分数的重要性相等，这并不符合真实场景。不同的分数根据查询对最终结果的贡献不同，例如，对于与地震相关的视频，续集与描述的分数比前传与描述的分数有显著更大的权重，因为地震发生前没有视觉上的指示。Yin and Jiang (2024) 证明了基于熵的方法在零样本排名融合任务中表现优于其他方法。在这种熵排名中，个别视频的不同相似度分数被归一化为 softmax 分布，低熵表示高置信度，而高逆熵对应于高置信度。基于这一见解，我们使用逆熵排名融合来结合这些分数。

具体来说，我们将之前计算的五个相似度得分 S_i 转换为加权分布 $P_i = \text{softmax}(S_i)$，并

[3] https://www.scenedetect.com/

[4] 我们对基于余弦的 SBERT (Reimers and Gurevych, 2019) 进行了实验，发现 ColBERT 在准确性和速度方面都优于 SBERT。尤其是在我们的案例中，LLM 生成的字幕可能会有一些噪声。ColBERT 通过在标记级别进行最大聚合，而不是将信息压缩成单个向量来减轻这种噪声的影响。这种方法有效地减少或忽略了多模态描述中的噪声元素。此外，之前的研究 (Warner et al., 2024) 表明，ColBERT 的方法能够最小化 SBERT 在平均过程中引入的信息损失，从而进一步提升其性能。

Model	R@1 ↑	R@5 ↑	R@10 ↑	P@1 ↑	P@5 ↑	P@10 ↑	MRR ↑	NDCG ↑	MAP ↑	MnR ↓	MdR ↓
	MultiVENT										
MultiCLIP	9.83	44.32	70.82	88.80	80.77	65.25	0.92	75.34	86.33	22.12	6
MultiCLIP + ASR	10.08	46.64	74.70	90.73	84.32	68.57	0.93	78.76	88.87	22.66	6
MultiCLIP + **Q$_2$E**	10.24	46.71	75.76	92.28	85.25	69.73	0.95	80.04	89.42	21.13	6
MultiCLIP + **Q$_2$E** + ASR	10.32	49.00	79.60	93.05	88.73	73.09	0.95	83.24	91.20	16.44	6
InternVideo2-1B	5.60	28.92	49.12	51.35	53.28	45.44	0.68	50.43	63.77	235.49	11
InternVideo2-1B + ASR	9.68	40.48	62.43	87.64	73.59	57.34	0.91	68.30	83.84	109.70	7
InternVideo2-1B + **Q$_2$E**	9.54	40.88	63.40	87.26	75.21	58.73	0.92	69.15	83.96	53.84	7
InternVideo2-1B + **Q$_2$E** + ASR	10.24	44.94	70.79	92.28	81.85	65.14	0.95	76.10	88.09	42.81	6
	MSR-VTT-1kA										
MultiCLIP	43.52	69.05	76.88	43.62	13.85	7.71	0.54	59.72	54.27	20.29	2
MultiCLIP + ASR	45.43	70.75	77.19	45.53	14.19	7.74	0.56	61.22	56.11	31.31	2
MultiCLIP + **Q$_2$E**	44.52	71.26	79.40	44.62	14.29	7.96	0.56	61.51	55.84	17.91	2
MultiCLIP + **Q$_2$E** + ASR	46.23	73.37	81.71	46.33	14.71	8.19	0.58	63.59	57.83	18.73	2
InternVideo2-1B	52.56	73.07	80.10	52.66	14.65	8.03	0.62	66.07	61.62	25.12	1
InternVideo2-1B + ASR	53.17	73.27	81.61	53.27	14.69	8.18	0.62	67.08	62.49	23.41	1
InternVideo2-1B + **Q$_2$E**	53.47	73.57	82.11	53.57	14.75	8.23	0.62	67.16	62.47	16.73	1
InternVideo2-1B + **Q$_2$E** + ASR	56.28	76.58	83.72	56.38	15.36	8.39	0.65	69.53	65.06	15.85	1

Table 1: MultiVENT 和 MSRVTT "1k-A'' 分割上的性能比较。我们采用 MultiCLIP（每个 Sanders et al. (2023)）和 InternVideo2（在 排行榜 上表现最好的方法）作为我们的基线编码器。我们的方法（**Q$_2$E**），即使没有 ASR，也在 MultiVENT 数据集上提高了 5-19 点的 NDCG，在 MSRVTT 数据集上提高了 1-2 点。添加 ASR 进一步增强了性能，在 MultiVENT 数据集上增加了 3-7 点，在 MSRVTT 数据集上增加了 2 点，总体在两种不同的编码器上对 MultiVENT 提高约 8-26 点，对 MSRVTT 提高约 3-4 点。

计算该分布的熵 $H(P_i)$。这个 P_i 根据 i 评分标准，提供了对 V 视频的初始可能性权重。我们通过其逆熵重新调整 P_i，最终得分 $\hat{S}$ 通过对所有重新调整的 P_i 求和来计算：

$$\hat{S} = \sum_{i}^{5} \frac{1}{H(P_i)} * P_i. \tag{3}$$

通过应用 Eq. (3)，我们有效地为低熵得分分配更大的权重，强调更加确信的预测。这个 $\hat{S}$ 包含了每个视频的汇总得分；$\hat{S}$ 可以被排序以返回查询的视频排名列表。

3 结果

我们在事件为中心的 MultiVENT 数据集 (Sanders et al., 2023) 和广泛使用的 MSR-VTT 数据集 (Xu et al., 2016) 上进行了实验。MultiVENT 提供了丰富的、事件特定的人类查询，而 MSR-VTT 包含 10,000 个视频片段，分布在 20 个类别中，每个片段都有 20 个英语句子的注释。我们使用了标准的 "1k-A" 测试集划分，包含 1,000 个视频，以适应计算约束，与之前的工作中使用的一致 (Jiang et al., 2022; Guan et al., 2023; Jiang et al., 2023a)。数据集的详细信息在 Table 11 中。完整的实现细节可以在 **??** 中找到，但简单来说，我们使用了 Llama-3 作为 LLM，InternVL-2.5 作为 VLM，和 Whisper-large-v3 作为多语言 ASR 模型。此外，我们使用 ColBERT 和 MultiCLIP 用于相似度得分的计算。

3.1 零样本文本到视频检索

我们的方法与基线在 MultiVENT 和 MSR-VTT-1kA 上的对比显示在 Table 1 上。我们使用标准检索指标：在排名为 K 的召回率/精确率，平均倒数排名（MRR），平均平均精度（MAP），归一化折扣累计增益（NDCG），以及平均和中位数排名（MnR，MdR）。

总体而言，与 ASR 结合使用，在两个不同的数据集和两个不同的文本到视频编码器中，**Q$_2$E 在所有指标上都得到了最佳分数**。但即使没有音频信息，我们在所有指标上都优于基线。**Q$_2$E** 也可以在不同的文本-视频编码器之间进行泛化。为了证明这一点，我们将其应用于零次测试 MSR-VTT 排行榜 [5] 中的 Rank 1 编码器 (Wang et al., 2025) 以及 MultiCLIP 编码器（遵循 Sanders et al. (2023)）。

Table 1 展示了我们方法的通用性。当应用于 InternVideo2 时，我们的方法在没有音频的情况下将 MSR-VTT 的 NDCG 得分提高了 1 分，加入音频后提升了 4 分。此外，InternVideo2 在 MultiVENT 数据集上观察到显著的性能下降，强调出该数据集的丰富性与复杂性。尽管如此，加入我们的方法后，在没有音频的情况下将 NDCG 得分提高了 19 分，加入音频后提升了 26 分，进一步验证了其有效性。

相反，MultiCLIP 在 MultiVENT 数据集 (Sanders et al., 2023) 上取得了最佳基线得分。

[5] https://paperswithcode.com/sota/zero-shot-video-retrieval-on-msr-vtt

Model	R@10	P@10	MRR	NDCG
		Arabic		
MC	72.31	62.55	0.94	76.10
MC + $\mathbf{Q_2E}$	74.46	63.92	0.94	78.09
MC + $\mathbf{Q_2E}$ + ASR	79.41	68.24	0.94	82.07
		Chinese		
MC	74.22	68.65	0.91	77.29
MC + $\mathbf{Q_2E}$	82.13	76.35	0.99	86.32
MC + $\mathbf{Q_2E}$ + ASR	83.02	76.92	0.97	86.61
		English		
MC	85.79	81.92	1.00	89.98
MC + $\mathbf{Q_2E}$	86.45	82.50	1.00	90.57
MC + $\mathbf{Q_2E}$ + ASR	87.80	83.85	1.00	91.43
		Korean		
MC	65.33	62.12	0.89	70.38
MC + $\mathbf{Q_2E}$	72.65	69.04	0.93	76.92
MC + $\mathbf{Q_2E}$ + ASR	76.90	73.27	0.92	80.58
		Russian		
MC	79.09	71.15	0.96	82.84
MC + $\mathbf{Q_2E}$	81.61	73.46	0.96	84.62
MC + $\mathbf{Q_2E}$ + ASR	85.40	76.54	0.99	88.70

Table 2: 不同语言的比较。结果在 MultiVENT 数据集上报告。MC = MultiCLIP 编码器。结果在 MultiVENT 数据集上报告。完整表格见 Table 7 中的 App. A.1.4 。

将我们的方法应用于 MultiCLIP 使得 NDCG 提升了 8 点，创造了 MultiVENT 数据集的最新记录。然而，与 InternVideo2 不同的是，MultiCLIP 没有使用交叉编码器重新排序器，这解释了其在 MSR-VTT 数据集上的相对较低表现。我们的方法仍然使 MSR-VTT 的 NDCG 得分提高了 4 点，显示了在不同编码器上的有效性。接下来，我们将研究 $\mathbf{Q_2E}$ 在不同语言（§3.2 ）、不同 LLM 大小（**??** ）、不同排序融合方法（§3.3 ）以及我们方法的各种组件（§3.4 ）下的表现。在附录中，我们研究 VLM 大小（App. A.1.1 ）、帧数（App. A.1.2 ）和帧选择方法（App. A.1.3 ）的影响；$\mathbf{Q_2E}$ 在编码决策中相当稳定。

3.2 跨语言性能

MultiVENT 包括五种语言——阿拉伯语、中文、英语、韩语和俄语。Table 2 显示，在 MultiCLIP 基线的基础上构建，结合我们的方法在所有指标上持续提高了检索性能。此外，添加音频特征在所有语言中带来了额外的提升，展示了我们方法的不论语言如何的有效性。

MultiCLIP 在阿拉伯语、中文和韩语视频检索方面遇到的困难最大。然而，当我们的方法被添加后，这些语言的检索性能分别提高了大约 6、9 和 10 个 NDCG 点。

Method	R@10	P@10	MRR	NDCG
MultiCLIP	70.82	65.25	0.92	75.34
		Without Audio		
Neg. Exp. Ent.	60.05	55.60	0.90	66.61
RRF	62.91	58.30	0.92	69.74
Max	71.98	66.41	0.93	76.60
Mean	73.54	67.88	0.94	78.37
Inv. Ent. ($\mathbf{Q_2E}$)	75.76	69.73	0.95	80.04
		With Audio		
Neg. Exp. Ent.	67.23	61.97	0.93	73.20
RRF	70.91	65.29	0.93	76.29
Max	76.10	70.04	0.93	80.04
Mean	78.64	72.47	0.95	82.44
Inv. Ent. ($\mathbf{Q_2E}$)	79.60	73.09	0.95	83.24

Table 3: 零样本排序融合中聚合方法的比较：负指数熵（Neg. Exp. Ent.）、倒数排序融合（RRF）、逆熵（Inv. Ent.）、平均和最大值。结果在 MultiVENT 数据集上报告。完整结果见 Table 9 ， App. A.1.6 。

虽然这个模型具有很高的鲁棒性和可推广性，但它的大尺寸需要显著的计算能力和更长的推理时间。为了展示我们方法的鲁棒性，**??** 提出了不同 LLM 大小的结果，范围从 1B 到 70B 参数。尽管通常是通过 70B 模型实现最佳性能，但即使是 1B 模型在没有音频的情况下也始终比基线 NDCG 高至少 4 个 NDCG 点，带有音频时高出 8 个点。这些结果突出了我们方法的鲁棒性。

3.3 排序融合方法的效果

在零样本场景中融合不同的排名是一项不小的任务。传统方法通常涉及训练密集网络来结合分数 (Yu et al., 2024) 或利用 LLMs 来集成多种专家输出 (Lu et al., 2024b,a; Jiang et al., 2023b) 。然而，对于 LLMs 来说，排名或分数融合是一个重要的挑战，因为它们必须处理和比较大量的数值数据，包括浮点值（用于分数）和离散值（用于排名）。在这两种情况下，LLMs 都难以产生合理的融合排名。我们考虑了几种聚合方法，从简单的均值技巧到更复杂的方法，如倒数排名融合（RRF）和逆熵；由于篇幅，我们在 App. A.1.6 中描述这些公式。Table 3 中的结果表明，即使是简单的均值函数与 RRF 和负指数熵相比也表现良好。逆熵显示出更好的表现，这使得我们采用它作为聚合方法。

3.4 不同成分的影响

Table 4 对查询事件分解和 $\mathbf{Q_2E}$ 的多模态描述组件进行了一项消融研究。结果强调了视频是最关键的组件。移除查询导致两种变体的性能下

Components	R@10	P@10	MRR	NDCG
MultiCLIP	70.82	65.25	0.92	75.34
		Without Audio		
Q$_2$E	75.76	69.73	0.95	80.04
Q$_2$E – Video	58.00	53.90	0.89	64.83
Q$_2$E – Query	74.74	68.80	0.93	78.78
Q$_2$E – Events	74.77	68.88	0.93	79.02
		With Audio		
Q$_2$E	79.60	73.09	0.95	83.24
Q$_2$E – Video	67.74	62.43	0.93	73.96
Q$_2$E – Query	78.12	71.74	0.93	81.54
Q$_2$E – Events	77.77	71.47	0.94	81.75

Table 4: 我们方法中不同组件的分析。每个区块的第一行对应的是我们完整方法，其中包含所有组件。在随后的几行中，我们逐次移除一个主要组件（视频、查询、事件），以展示其对整体方法的贡献。事件 = （前情 + 当前 + 后续）。结果是在 MultiVENT 数据集上报告的。完整结果在 Table 10 中报告于 App. A.1.7 。

降 1 分，而移除事件导致非音频变体的性能下降 1 分，音频变体的性能下降 2 分，突出了分解事件在检索相关视频中的重要性。

4 结论 & 未来工作

我们利用了嵌入在 LLMs 和 VLMs 中的世界知识来分解查询和视频以用于视频检索任务，展示了如何将这些知识与视觉和音频信号相匹配。我们采用逆熵作为一种排名融合方法，而无需微调任何额外的融合网络。不同特征的两个数据集的实验结果表明了我们方法的鲁棒性和广泛适用性。虽然我们展示了 LLMs 和 VLMs 的现有知识如何有效地转移到文本到视频的检索中，但我们的工作为未来的研究奠定了基础，以探索如何利用事实和反事实信息来积极或消极地与查询对齐。

尽管我们的方法在多个数据集上表现出色，但其计算成本高且耗时。未来的研究应该集中于优化效率，以减少时间和成本。

由于 LLM 和 VLM 的固有特性，某种程度的信息误传或幻觉可能会发生。虽然我们在评估中没有发现任何信息误传，但在我们提示事件发生前、发生时或发生后可能导致幻觉。同样，依靠这些模型的参数化知识，可能会传播模型或其输出中固有的偏见。我们没有明确探索或缓解这些模型中固有的偏见，但我们也没有注意到反感或刻板的输出。

5 致谢

本材料部分基于由国家科学基金会在资助计划编号 IIS-2024878 下支持的工作。一些实验是在 UMBC HPCF 上进行的，该平台由国家科学基金会在资助计划编号 CNS-1920079 下支持。此材料也基于部分由陆军研究实验室资助的，资助计划编号为 W911NF2120076 以及在协议编号 HR00112520301 下由 DARPA 为 SciFy 项目提供支持的研究。尽管上面注明了版权，允许美国政府为政府目的再现和分发。文中包含的观点和结论为作者所持观点，不应被解释为必然代表 ARL、DARPA 或美国政府的正式政策或背书，无论是明示的还是默示的。

References

Meng Cao, Haoran Tang, Jinfa Huang, Peng Jin, Can Zhang, Ruyang Liu, Long Chen, Xiaodan Liang, Li Yuan, and Ge Li. 2024. Rap: Efficient text-video retrieval with sparse-and-correlated adapter. arXiv preprint arXiv:2405.19465 .

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In Findings of the Association for Computational Linguistics: EACL 2024 , pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies , pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset. arXiv preprint . ArXiv:2305.18500.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. arXiv preprint arXiv:2407.10759 .

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. 2024. Gramian multimodal representation learning and alignment. arXiv preprint arXiv:2412.11959 .

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 .

Alexandros Delitzas, Maria Parelli, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2023. Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes. arXiv preprint arXiv:2306.02329 .

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT , volume 1, page 2.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783 .

Peiyan Guan, Renjing Pei, Bin Shao, Jianzhuang Liu, Weimian Li, Jiaxi Gu, Hang Xu, Songcen Xu, Youliang Yan, and Edmund Y Lam. 2023. Pidro: Parallel isomeric attention with dynamic routing for text-video retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision , pages 11164–11173.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. arXiv preprint arXiv:2306.11644 .

Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. 2023. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision , pages 12054–12064.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. arXiv preprint arXiv:2401.04088 .

Chen Jiang, Hong Liu, Xuzheng Yu, Qing Wang, Yuan Cheng, Jia Xu, Zhongyi Liu, Qingpei Guo, Wei Chu, Ming Yang, et al. 2023a. Dual-modal attention-enhanced text-video retrieval with triplet partial margin contrastive learning. In Proceedings of the 31st ACM International Conference on Multimedia , pages 4626–4636.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023b. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. arXiv preprint arXiv:2306.02561 .

Jie Jiang, Shaobo Min, Weijie Kong, Hongfa Wang, Zhifeng Li, and Wei Liu. 2022. Tencent text-video retrieval: Hierarchical cross-modal interactions with multi-level representations. IEEE Access .

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval , pages 39–48.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434 .

Haowei Liu, Yaya Shi, Haiyang Xu, Chunfeng Yuan, Qinghao Ye, Chenliang Li, Ming Yan, Ji Zhang, Fei Huang, Bing Li, et al. 2024b. Unifying latent and lexicon representations for effective video-text retrieval. arXiv preprint arXiv:2402.16769 .

Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024a. Routing to the expert: Efficient reward-guided ensemble of large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) , pages 1964–1974, Mexico City, Mexico. Association for Computational Linguistics.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. 2024b. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 6159–6172, Bangkok, Thailand. Association for Computational Linguistics.

Taichi Nishimura, Shota Nakada, and Masayoshi Kondo. 2024. Vision-language models learn super images for efficient partially relevant video retrieval. ACM Trans. Multimedia Comput. Commun. Appl. Just Accepted.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning , pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In International conference on machine learning , pages 28492–28518. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog , 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kate Sanders, David Etter, Reno Kriz, and Benjamin Van Durme. 2023. Multivent: Multilingual videos of events and aligned natural text. Advances in Neural Information Processing Systems , 36:51065–51079.

Yidan Sun, Qin Chao, and Boyang Li. 2024. Event causality is key to computational story understanding. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) , pages 3493–3511, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 .

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. 2025. Internvideo2: Scaling foundation models for multimodal video understanding. In European Conference on Computer Vision , pages 396–416. Springer.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv preprint arXiv:2412.13663 .

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. In International Conference on Machine Learning , pages 38728–38748. PMLR.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE conference on computer vision and pattern recognition , pages 5288–5296.

Siqi Yin and Lifan Jiang. 2024. Multi-method integration with confidence-based weighting for zero-shot image classification. arXiv preprint arXiv:2405.02155 .

Shi Yu, Chenghao Fan, Chenyan Xiong, David Jin, Zhiyuan Liu, and Zhenghao Liu. 2024. Fusion-in-t5: Unifying variant signals for simple and effective document ranking with attention fusion. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) , pages 7556–7561, Torino, Italia. ELRA and ICCL.

Yue Zhao, Long Zhao, Xingyi Zhou, Jialin Wu, Chun-Te Chu, Hui Miao, Florian Schroff, Hartwig Adam, Ting Liu, Boqing Gong, et al. 2024. Distilling vision-language models on millions of videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 13106–13116.

A 附录

A.1 消融研究（续）

A.1.1 VLM 尺寸的影响

在 Q_2E 中，我们使用了 `InternVL-2.5-78B` 模型。类似于 `Llama-3.3-70B` ，该模型具有很高的鲁棒性和广泛的适用性，但也伴随着显著的计算和时间限制。为了评估我们方法的效率和灵活性，Table 5 报告了在 InternVL-2.5 系列中的各种 VLM 规模的结果。

结果表明，从 4B 到 38B 参数的模型实现了几乎相同的性能，分数上的差异可以忽略不计。这表明即使使用较小的视觉语言模型（VLM），我们的方法也能获得可比较的结果，从而降低计算需求。此外，加入音频后，性能差距进一步缩小；例如，1B 的 VLM 仅在 NDCG 得分上下降了 1 分。这些发现强调了我们方法的鲁棒性和普适性，无论 VLM 的大小如何。

A.1.2 帧数的影响

Fig. 3 展示了帧数对不同方法的性能影响。结果表明，增加采样帧数通常会提高性能。然而，随着每秒比较次数的增加（即，采样帧更少），在所有方法中 NDCG 均有所下降，这突出显示了计算效率与检索性能之间的权衡。

尽管上升趋势在超过 64 帧后继续，我们假设进一步增加帧数将带来额外的性能提升。然而，收益递减和计算成本增加将减少模型作为检索系统的实用性。因此，我们将评估限制在 64 帧，并选择 16 帧采样作为性能和速度之间的最佳平衡。

A.1.3 帧选择方法的影响

在 Q_2E 中，我们采用了一种均匀采样方法进行帧选择。然而，为了比较其与更复杂方法的有效性，我们使用来自 `PySceneDetector` [6] 的自适应检测器进行了一个消融研究。如 **??** 所示，

[6] https://www.scenedetect.com/

VLM Size	R@1 ↑	R@5 ↑	R@10 ↑	P@1 ↑	P@5 ↑	P@10 ↑	MRR ↑	NDCG ↑	MAP ↑	MnR ↓	MdR ↓
MultiCLIP	9.83	44.32	70.82	88.80	80.77	65.25	0.92	75.34	86.33	22.12	6
						Without Audio					
1B	9.80	43.76	71.73	89.19	79.85	66.02	0.93	75.86	86.05	24.96	6
2B	9.94	45.23	73.86	89.96	82.55	68.03	0.93	77.90	87.49	22.45	6
4B	9.97	46.75	75.85	90.73	85.17	69.92	0.93	79.81	89.05	19.68	6
8B	10.03	46.40	75.73	90.35	84.71	69.81	0.93	79.72	88.98	21.24	6
26B	10.24	47.03	76.16	92.66	85.64	70.19	0.95	80.47	89.94	19.87	6
38B (**Q$_2$E**)	10.24	46.71	75.76	92.28	85.25	69.73	0.95	80.04	89.42	21.13	6
						With Audio					
1B	10.18	47.87	78.15	91.89	86.56	71.74	0.94	81.77	90.33	18.79	6
2B	10.14	48.27	78.46	91.51	87.26	72.05	0.94	82.14	90.58	17.21	6
4B	10.19	48.90	79.72	92.28	88.57	73.17	0.94	83.13	91.05	16.77	6
8B	10.40	48.87	79.88	93.82	88.34	73.36	0.95	83.52	91.53	17.27	6
26B	10.40	48.64	80.12	93.82	88.03	73.51	0.95	83.61	91.13	15.97	6
38B (**Q$_2$E**)	10.32	49.00	79.60	93.05	88.73	73.09	0.95	83.24	91.20	16.44	6

Table 5: 全面比较 InternVL-2.5 家族中不同 VLM 大小（从 1B 到 38B 参数），专门用于帧字幕生成。在 MultiVENT 数据集上报告结果。

# frames	R@1 ↑	R@5 ↑	R@10 ↑	P@1 ↑	P@5 ↑	P@10 ↑	MRR ↑	NDCG ↑	MAP ↑	MnR ↓	MdR ↓
MultiCLIP	9.83	44.32	70.82	88.80	80.77	65.25	0.92	75.34	86.33	22.12	6
						Without Audio					
2	9.31	41.71	64.46	83.78	76.37	59.54	0.90	69.64	83.75	36.17	7
4	9.75	44.00	71.05	88.03	80.39	65.64	0.92	75.43	86.54	25.32	6
8	10.10	46.56	74.84	90.73	84.79	68.88	0.94	79.05	89.05	23.01	6
16	10.24	46.71	75.76	92.28	85.25	69.73	0.95	80.04	89.42	21.13	6
32 (**Q$_2$E**)	10.15	47.22	75.95	91.89	85.95	70.00	0.94	80.29	90.15	20.05	6
64	10.09	47.15	76.77	91.12	85.95	70.66	0.94	80.66	89.74	19.98	6
						With Audio					
2	10.26	46.63	74.93	92.28	84.48	68.88	0.94	79.42	89.33	21.66	6
4	10.16	47.37	77.54	91.51	85.79	71.27	0.94	81.26	90.05	19.20	6
8	10.26	48.74	79.39	92.66	88.11	72.86	0.95	82.93	90.99	16.97	6
16	10.32	49.00	79.60	93.05	88.73	73.09	0.95	83.24	91.20	16.44	6
32 (**Q$_2$E**)	10.41	48.83	80.09	93.82	88.26	73.55	0.95	83.68	91.32	16.38	6
64	10.40	48.90	80.42	93.82	88.49	73.82	0.95	83.95	91.58	15.90	6

Table 6: 不同帧数（从 2 到 64）的全面对比。结果在 MultiVENT 数据集上进行了报告。

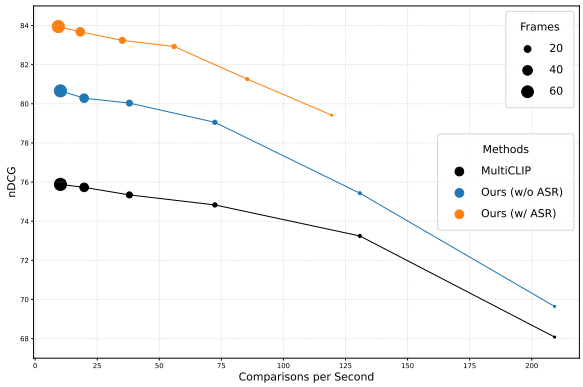

Figure 3: 帧数对检索性能的影响。该图显示了三个方法每秒比较的 nDCG 性能：MultiCLIP（黑色）、**Q₂E**（不含 ASR）（蓝色）和 **Q₂E**（含 ASR）（橙色）。点的大小与采样帧数成正比（即，2、4、8、16、32、64）。每秒比较数 = 运行时间 /（查询 × 视频数量）。结果在 MultiVENT 数据集上报告。完整表格报告在 Table 6 。

均匀采样在这两种变体中都实现了最佳性能，显示了其在音频可用性无论如何的情况下的鲁棒性。此外，场景检测算法产生可变数量的帧（从 2 到 100 不等），这使得检索时间相比 16 均匀采样显著增加了大约 2.8 倍。

A.1.4 不同语言的影响

在 §3.2 中描述的实验的详细结果已在 Table 7 上报告。

A.1.5 LLM 规模的影响

在 **??** 中描述的实验的详细结果在 Table 8 中有报道。

A.1.6 不同排名融合的效果

在 §3.3 中描述的实验的详细结果在 Table 9 中报告。这些实验考虑了五种不同的方法来聚合分数并执行零样本排序融合。在所有方法中，我们计算了一个分布 P_i ，它跨越所有视频，对于每个 i 分数组件，如 §3.3 中讨论的那样。我们考虑的具体排序融合方法在于我们如何组合五种 P_i 中的每一种以获得跨越所有视频的单一排名 $\hat{S}$ 。这些方法是：

平均聚合 我们简单地对 P_i 取平均，计算 $\hat{S} = \frac{1}{5}\sum_i P_i$ 。

最大聚合 对于每个视频 v ，我们对该视频的 P_i 分数取最大值：$\hat{S}_v = \max_i P_{i,v}$ 。

互反排名融合（RRF） 对于每个视频 v ，我们计算该视频在 P_i 中的排名，然后累加这些互反排名。v 的最终排名是 $\sum_i (1/Rank(P_{i,v}))$ 。

负指数熵 我们计算 $\hat{S} = \sum_i \exp(-H(P_i)) * P_i$ ，其中 $H(P_i)$ 是 P_i 的熵。

逆熵 这是主论文中描述的方法，在 §3.3 中。

A.1.7 不同成分的影响

在 §3.4 中描述的实验的详细结果在 Table 10 中报告。

A.2 数据集统计

详细的数据集统计信息展示在 Table 11 。

我们使用 Llama-3 [7] 作为 LLM，InternVL-2.5 [8] 作为 VLM，以及 Whisper-large-v3 [9] 作为用于查询解构、帧字幕和音频转录的多语言 ASR 模型。在嵌入计算中，我们采用 ColBERT [10] 进行文本到文本相似度计算，并使用 MultiCLIP [11] 进行文本到视频相似度评分的计算。查询解构和视频字幕编写在温度为 0.8 和 top-p 采样为 0.95 的情况下进行。对于帧提取，我们应用了 16 帧的均匀采样。我们的评估框架基于 torchmetrics [12] 。数据提取提示使用了两个 A100 GPU 执行，而文本到视频检索则在 RTX 6000 GPU 上进行。

A.3 提示

A.3.1 提取空间、时间和事件数据

以下三个提示分别用于从查询中提取主要事件、地点和时间。这些提示的唯一输入是原始查询。在查询未提及具体时间或地点的情况下，指示 LLM 输出 “NOT AVAILABLE”。

Extract Primary Event

You are given a video search query. Your task is to extract the primary event(s) being referenced or discussed in the query.

Here are some guidelines:
- Identify the main event(s) that the query is about, whether explicitly stated or implied
- If multiple events are mentioned, list all of them
- If no clear event can be identified from the query, output "NOT AVAILABLE" in the EVENTS section
- First, explain your reasoning in 1-2 sentences about how you identified the event(s) from the query
- Structure your response into two sections, "EXPLANATION:" and then "EVENTS:"
- The events section should contain either the identified event(s) in a numbered list, or "NOT AVAILABLE" if no event can be determined
- Don't include temporal or spatial information in this event, just focus on the primary event(s) being discussed or referenced

[7] meta-llama/Llama-3.3-70B-Instruct
[8] OpenGVLab/InternVL2_5-38B
[9] openai/whisper-large-v3
[10] hltcoe/plaidx-large
[11] laion/Clip-VIT-H14
[12] https://lightning.ai/docs/torchmetrics

Model	R@1 ↑	R@5 ↑	R@10 ↑	P@1 ↑	P@5 ↑	P@10 ↑	MRR ↑	NDCG ↑	MAP ↑	MnR ↓	MdR ↓
						Arabic					
MultiCLIP	10.91	44.32	72.31	92.16	76.47	62.55	0.94	76.10	84.59	14.44	6.0
MultiCLIP + Q_2E	10.82	45.86	74.46	90.20	78.43	63.92	0.94	78.09	86.08	14.91	6.0
MultiCLIP + Q_2E + ASR	10.79	49.66	79.41	90.20	83.53	68.24	0.94	82.07	88.32	11.59	6.0
						Chinese					
MultiCLIP	9.12	45.59	74.22	84.62	83.85	68.65	0.91	77.29	86.52	12.60	6.0
MultiCLIP + Q_2E	10.66	48.81	82.13	98.08	90.00	76.35	0.99	86.32	94.64	9.59	6.0
MultiCLIP + Q_2E + ASR	10.46	50.55	83.02	96.15	93.08	76.92	0.97	86.61	93.96	8.73	5.5
						English					
MultiCLIP	10.61	50.82	85.79	100.00	96.15	81.92	1.00	89.98	97.47	7.64	5.0
MultiCLIP + Q_2E	10.61	51.66	86.45	100.00	97.69	82.50	1.00	90.57	98.04	8.07	5.0
MultiCLIP + Q_2E + ASR	10.61	51.66	87.80	100.00	97.69	83.85	1.00	91.43	98.12	7.74	5.0
						Korean					
MultiCLIP	9.30	39.98	65.33	86.54	75.00	62.12	0.89	70.38	82.86	20.78	7.0
MultiCLIP + Q_2E	9.68	43.65	72.65	90.38	82.31	69.04	0.93	76.92	87.04	15.87	6.0
MultiCLIP + Q_2E + ASR	9.68	45.93	76.90	90.38	86.15	73.27	0.92	80.58	89.02	13.85	6.0
						Russian					
MultiCLIP	10.57	49.57	79.09	92.31	87.69	71.15	0.96	82.84	91.20	10.65	6.0
MultiCLIP + Q_2E	10.39	50.55	81.61	92.31	89.62	73.46	0.96	84.62	91.46	10.88	6.0
MultiCLIP + Q_2E + ASR	11.21	53.49	85.40	98.08	94.23	76.54	0.99	88.70	95.02	9.13	5.0

Table 7: 不同语言的综合比较。结果基于 MultiVENT 数据集报告。

LLM Size	R@1 ↑	R@5 ↑	R@10 ↑	P@1 ↑	P@5 ↑	P@10 ↑	MRR ↑	NDCG ↑	MAP ↑	MnR ↓	MdR ↓
MultiCLIP	9.83	44.32	70.82	88.80	80.77	65.25	0.92	75.34	86.33	22.12	6
						Without Audio					
1B	10.05	46.07	75.45	90.35	83.94	69.54	0.93	79.34	88.15	22.34	6
3B	10.18	46.70	75.50	91.51	85.25	69.54	0.94	79.78	89.40	21.66	6
8B	10.21	46.60	75.02	92.28	84.94	69.11	0.94	79.41	89.37	19.74	6
70B (Q_2E)	10.24	46.71	75.76	92.28	85.25	69.73	0.95	80.04	89.42	21.13	6
						With Audio					
1B	10.26	48.31	78.71	92.66	87.49	72.28	0.95	82.50	91.23	17.64	6
3B	10.32	48.88	79.17	93.05	88.57	72.74	0.95	83.03	91.55	16.51	6
8B	10.41	48.62	79.04	93.82	88.03	72.59	0.95	82.91	91.28	16.79	6
70B (Q_2E)	10.32	49.00	79.60	93.05	88.73	73.09	0.95	83.24	91.20	16.44	6

Table 8: 对 Llama-3 家族中不同 LLM 大小（从 1B 到 70B 参数）的全面比较，特别用于查询分解和将帧字幕改进为视频字幕。结果在 MultiVENT 数据集上报告。

Method	R@1 ↑	R@5 ↑	R@10 ↑	P@1 ↑	P@5 ↑	P@10 ↑	MRR ↑	NDCG ↑	MAP ↑	MnR ↓	MdR ↓
MultiCLIP	9.83	44.32	70.82	88.80	80.77	65.25	0.92	75.34	86.33	22.12	6
						Without Audio					
Neg. Exp. Ent.	9.66	39.97	60.05	87.26	73.36	55.60	0.90	66.61	83.60	68.44	7
RRF	9.92	42.10	62.91	89.58	77.37	58.30	0.92	69.74	86.94	33.80	7
Max	9.98	45.15	71.98	90.35	82.32	66.41	0.93	76.60	87.77	29.31	6
Mean	10.19	45.99	73.54	92.28	84.09	67.88	0.94	78.37	89.50	26.32	6
Inv. Ent. (Q_2E)	10.24	46.71	75.76	92.28	85.25	69.73	0.95	80.04	89.42	21.13	6
						With Audio					
Neg. Exp. Ent.	10.11	43.98	67.23	90.73	80.00	61.97	0.93	73.20	87.91	59.09	7
RRF	10.01	45.75	70.78	90.73	83.17	65.17	0.93	76.14	88.35	26.77	6
Max	9.94	46.58	76.20	89.96	84.79	70.12	0.93	79.86	88.82	22.89	6
Mean	10.23	48.06	78.30	92.66	87.41	72.16	0.95	82.20	91.04	20.32	6
Inv. Ent. (Q_2E)	10.37	48.42	79.06	93.44	87.80	72.66	0.95	82.89	91.10	17.14	6

Table 9: 对零样本排名融合的不同聚合方法进行全面比较：负数指数熵（Neg. Exp. Ent.）、互惠排名融合（RRF）、逆熵（Inv. Ent.）、平均值和最大值。结果在 MultiVENT 数据集上报告。

Components	R@1 ↑	R@5 ↑	R@10 ↑	P@1 ↑	P@5 ↑	P@10 ↑	MRR ↑	NDCG ↑	MAP ↑	MnR ↓	MdR ↓
MultiCLIP	9.83	44.32	70.82	88.80	80.77	65.25	0.92	75.34	86.33	22.12	6
						Without Audio					
Q$_2$E	10.24	46.71	75.76	92.28	85.25	69.73	0.95	80.04	89.42	21.13	6
Q$_2$E - Video	9.28	39.56	58.00	84.17	72.97	53.90	0.89	64.83	82.66	65.97	8
Q$_2$E - Query	10.08	45.97	74.74	91.12	83.78	68.80	0.93	78.78	87.92	21.00	6
Q$_2$E - Events	9.95	46.38	74.77	89.96	84.48	68.88	0.93	79.02	88.98	18.35	6
						With Audio					
Q$_2$E	10.32	49.00	79.60	93.05	88.73	73.09	0.95	83.24	91.20	16.44	6
Q$_2$E - Video	10.16	44.61	67.74	91.51	81.16	62.43	0.93	73.96	88.97	50.96	6
Q$_2$E - Query	10.08	47.39	78.12	91.12	85.95	71.74	0.93	81.54	89.18	16.32	6
Q$_2$E - Events	10.22	48.09	77.77	91.89	87.18	71.47	0.94	81.75	90.38	16.64	6

Table 10: 我们方法中不同组件的全面分析。顶部的模块展示了我们方法在有音频文本时的性能，而底部的模块则代表了在无音频文本时的性能。每个模块的第一行对应的是包含所有组件的完整方法。在后续行中，每次移除一个主要组件（查询、视频、事件），以展示其对整个方法的贡献。事件 =（前序 + 当前 + 后序）。结果报告基于 MultiVENT 数据集。

Dataset	# of query	Avg. words	# of video	Avg. length (sec.)
MSR-VTT-1kA	995	9	1000	15
MultiVENT	259	27	2394	83

Table 11: 本研究中使用的数据集概览。表格总结了 MSR-VTT-1kA 和 MultiVENT 数据集的关键统计信息。

Here is the query I'd like you to extract events from: {{query}}

Extract Spatial Information

You are given a video search query. Assume the video search query is an ongoing event. Your task is to extract any locations mentioned in the query.

Here are some guidelines:
- Extract any location information mentioned in the query, including specific cities, towns, countries, regions, areas, neighborhoods, or other geographical references
- If no location information is found in the query, output "NOT AVAILABLE" in the LOCATION INFORMATION section
- First, explain your reasoning in 1-2 sentences about what location information appears in the query (or lack thereof) and why you're extracting it
- Structure your response into two sections, "EXPLANATION:" and then "LOCATION INFORMATION:"
- The location information section should contain either the extracted location information in one line, or "NOT AVAILABLE" if no location information is provided.
- Output in the following format: CITY, COUNTRY, REGION, AREA, NEIGHBORHOOD, etc. (e.g. New York, USA, Europe, Downtown). If one of the fields is not available, you can leave it out.

Here is the query I'd like you to extract location information from: {{query}}

Extract Temporal Informaton

You are given a video search query. Assume the video search query is an ongoing event. Your task is to extract any dates, times, or years mentioned in the query.

Here are some guidelines:
- Extract any temporal information mentioned in the query, including specific dates, times of day, years, seasons, or other time-related references
- If no temporal information is found in the query, output "NOT AVAILABLE" in the TEMPORAL INFORMATION section
- First, explain your reasoning in 1-2 sentences about what temporal information appears in the query (or lack thereof) and why you're extracting it
- Structure your response into two sections, "EXPLANATION:" and then "TEMPORAL INFORMATION:"
- The temporal information section should contain either the extracted date/time information in one line, or "NOT AVAILABLE" if no temporal information is provided.
- Output in the following format: DAY MONTH YEAR, TIME, SEASON, etc. (e.g. 1 January 2022, 3:00 PM, Summer). If one of the fields is not available, you can leave it out.

Here is the query I'd like you to extract temporal information from: {{query}}

A.3.2 将查询分解为前传、当前和续集事件

以下提示被用于从查询中提取前传、当前和续集事件，尽量少做修改以保持在所有情况下的普遍适用性。

Prequel / Current / Sequel

You are given a video search query. Assume the video search query is an ongoing event. You have to find out what previous events could lead to this given event / decompose the ongoing event in multiple simple events / find out what events can be the outcome of this ongoing event.The extracted events should be concrete enough to be visualized in a video.

Here are some guidelines:
- Extract events that could be visualized in a video.

- Before extracting an event, consider whether the event text could serve as a reasonable caption for a frame of a video associated with the query.
- Give specific events related to the query, use your previous WORLD KNOWLEDGE to extract events that are likely to be seen in a video. For example, if the query is "2019 Christmas Eve Protests in Hong Kong", you might return "Protesters marching in the streets", but not "Protestors marching in snow" because it is not likely to have snow in Hong Kong, though its pretty common in other places.
- First, explain your reasoning in 1-2 sentences about which events it makes sense to extract. Then, provide a numbered list with the events (anywhere from one to five event; use your best judgement to determine how many to extract)
- Structure your response into two sections, ""EXPLANATION:"" and then ""EVENTS:"". The events sections should contain nothing but the events.

Here is the query I'd like you to extract events from: {{query}}

A.3.3 优化查询

从先前的提示中提取信息后，使用以下提示通过结合提取的事件、时间和空间细节来完善前序、当前和后续事件。目标是在保留所有相关信息的同时，构建自然的、一致的、类似人类的搜索查询。

Refine Query

You will receive a query along with optional time, location, and main event information. Your task is to integrate this information naturally into a refined, human-like search query.

Here are some guidelines:
- Take the original query and incorporate any provided time, location, and event details seamlessly
- If any of these elements (time, location, event) are not provided, work with just the available information
- The refined query should sound natural, as if someone is searching with that query in Google, YouTube, etc.
- The base query information must always be preserved in the refined version
- First, explain your reasoning in 1-2 sentences about which events it makes sense to extract. Then, provide the refine query in one line
- Structure your response into two sections, ""EXPLANATION:"" and then ""REFINED QUERY:"". The Refined Query section should contain nothing but the refined query in one line.

Base Query: {{prequel/Current/sequel}}
Event: {{event}}
Place: {{place}}
Time: {{time}}

A.3.4 帧说明器

使用以下提示为每个视频帧生成字幕。与其他提示不同的是，利用 VLM 来有效地处理多模态输入。如 Fig. 2 所示，我们实现了两种变体：(1) ASR 和 (2) 非 ASR。在非 ASR 变体的提示中排除了 橙色 文本。

Frame Caption

You are provided with the description of the previous frames, the current frame (image) of a single video, and the original transcript of the audio (ASR) from the entire video. Your task is to describe the main event or activity depicted in the current frame.

Instructions:
- Use the original ASR to enhance your understanding of the video context and inform your description of the current frame.
- Focus on identifying and detailing the key actions, participants, and context of the event depicted in the current frame.
- Maintain continuity with the previous frame' s description, connecting the current frame to the overall flow of the video.
- Ignore unrelated visual elements like colors or shapes unless they are critical to understanding the event.
- Use any visible text in the image to enhance your interpretation of the frame.

Only output the current frame description, integrating insights from the original ASR and previous frame description. Don't output any additional commentary or content.

Original ASR Transcript (Entire Video):
{ { ASR } }

Previous Frame Description:
{{prev_caption}}

Frame:
{{frame}}

Current Frame Description:

A.3.5 完整视频字幕生成器

以下提示为整个视频生成了一个总结性的字幕。鉴于处理视频所需的高计算要求以及 VLM 在处理视频方面的有限能力，我们利用 LLM 来总结前一步中提取的所有字幕。完整的精炼音频转录也被包含在 ASR 变体中，以提供额外的背景信息。与之前一样， 橙色 文本元素在非 ASR 变体中被排除在提示之外。

如

Video Caption using Frame Captions

You will receive individual descriptions of consecutive frames from a single video, along with the original transcript of the audio (ASR) from the entire video. Your task is to provide a detailed description of the whole video, summarizing the key actions, participants, and context of the event depicted in the video.

Instructions:
- Use the original ASR to enhance your understanding of the video context and inform your description of the whole video.
- Focus only on relevant aspects, such as what is happening, who is involved, and the overall context or setting.
- Ignore unrelated visual details like colors, shapes, or other non-essential information.
- Your summary should concisely and accurately capture the essence of the event, integrating insights from the frame descriptions, the Original ASR and the ASR sum-

mary.

Only output the video summary, without any additional commentary or content.

Frame Descriptions:
Frame [i] Description
{{frame_i_description}}

Original ASR Transcript (Entire Video):
{ { ASR } }

Summary:

和

A.3.6 精炼音频转录

中所示，我们利用了一个 LLM 来改进由多语言-ASR 和特定翻译模型生成的翻译。随后使用了以下提示语来利用三个输入细化音频抄本：(1) 来自 ASR 的原语言抄本，(2) 来自多语言-ASR 的翻译抄本，以及 (3) 来自翻译器的翻译抄本。

Refine Audio Transcript

You are an expert in language translation and text refinement. Your task is to produce a highly accurate and natural-sounding translation by analyzing the original transcript and two provided translations.

Instructions:
- Combine the strengths of both translations and use the original transcript as a reference to improve accuracy and contextual meaning.
- Prioritize clarity, fluency, and maintaining the original intent and tone of the text.
- Where translations differ significantly, cross-check with the original transcript for fidelity.
- Ensure the final output is free from any awkward phrasing or mistranslation.
- If the transcripts can't be refined with reasonable accuracy, just output "Not Available".
Transcripts:
Original Transcript:
{{original_language_asr}}

Translation 1:
{{whisper_translated_asr}}

Translation 2:
{{nllb_translated_asr}}

Output the refined translation only, without additional notes or explanations.

Refined Translation:

作者在开发过程中使用了 GitHub copilot [13] ，并在校对和润色最终写作时使用了 ChatGPT [14] 。提供给这些工具的内容是作者原创的。

[13] https://github.com/features/copilot
[14] https://chatgpt.com/