arXiv:2601.00584v1 [cs.CV] 2 Jan 2026

GranAlign：面向零样本视频时刻检索的粒度感知对齐框架

Mingyu Jeon[1], **SunjaeYoon**[2],**JongheeKim**[3],**JunyeoungKim**[1]中央大学人工智能系[2] 韩国科学技术院（KAIST）[3]电子通信研究院（ETRI）{smart2557, junyeongkim}@cau.ac.kr,sunjae.yoon@kaist.ac.kr, jhkim27@etri.re.kr

摘要

零样本视频时刻检索（ZVMR）是一项任务，旨在不依赖特定任务训练数据，利用自然语言查询在未剪辑视频中定位时间片段。这一设定下的主要挑战在于文本查询与视觉内容之间的语义粒度不匹配。以往的 ZVMR 研究尝试利用能够在联合空间中表征视频和语言的高质量预训练知识来实现对齐。然而，对于给定场景，这些方法未能平衡各模态所提供预训练知识的语义粒度。因此，尽管各模态的表征质量很高，粒度不匹配仍会导致检索不准确。本文提出一种名为***Gran****ularity-****A****ware* ***Align****ment(GranAlign)*的免训练框架，以弥合粗粒度和细粒度语义表征之间的差距。我们的方法引入了两种互补技术：基于粒度的查询改写，用于生成不同语义粒度的查询；以及查询感知字幕生成，用于将查询意图融入视频内容。通过将多层级查询与查询无关和查询感知字幕相结合，我们有效解决了语义不匹配问题。因此，我们的方法在三大主流基准测试（QVHighlights、Charades-STA、ActivityNet-Captions）上均创下新的最先进水平，并在具有挑战性的 QVHighlights 数据集上将 mAP@avg 提升了 3.23%。

1 引言

视频时刻检索（VMR）——通过语言查询定位视频中的片段——对于高效的视频理解至关重要。然而，传统的有监督方法从根本上受限于对大量高成本标注数据集的依赖。为克服这一依赖，零样本视频时刻检索（ZVMR）应运而生，成为一种强大且成功的范式；近年来视觉语言模型的进步进一步提升了其可行性（视觉语言模型）和大语言模型。

尽管取得了成功，零样本视频时刻检索范式仍面临一个全新的根本性挑战：语言查询与视觉内容之间的“粒度不匹配”。这一问题源于用户可能会以不同的细节层次描述同一事件，从“一只可爱的小狗”这样笼统的短语，到一个具-

版权所有 © 2026，人工智能促进协会（www.aaai.org）。保留所有权利。

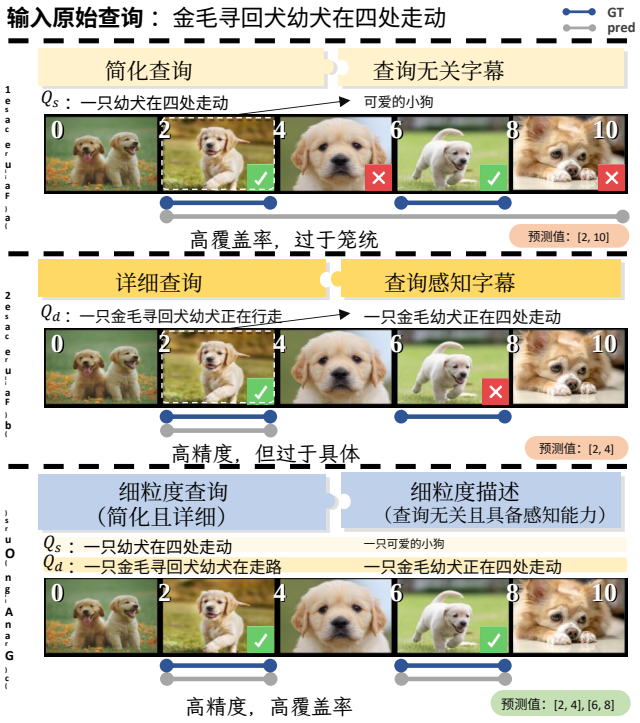

图 1：GranAlign 如何解决“粒度不匹配”的示意图。(a) 和 (b) 由于语义粒度不匹配，无法定位所有时刻，分别导致精确率低和覆盖率低。相比之下，我们提出的 GranAlign (c) 融合了两种粒度层级，实现高精度和高覆盖率，从而正确定位所有目标时刻。**要点：** GranAlign 通过协同高召回率的简单路径与高精确率的详细路径，克服了核心的“粒度不匹配”问题。

例如“金毛寻回犬幼犬正在四处走动”这样具体的查询。如图1所示，这种不匹配造成了不可避免的权衡：一般查询可能具有较高覆盖率，却缺乏精确定位确切时刻所需的精确率；而具体查询虽能提供较高精确率，但如果细微细节无法完美对齐，往往会面临覆盖率不足的问题。这一局限并非仅是个别现象。我们的定量分析—

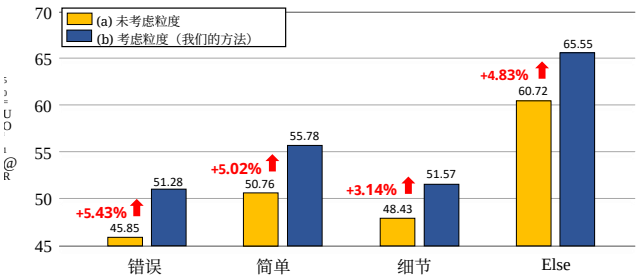

图2：在QVHighlights验证集上，按查询类型分类（分类标准见第4.3节），将我们的方法（b）与基线方法（a）（未考虑查询和视频内容的粒度）进行比较。

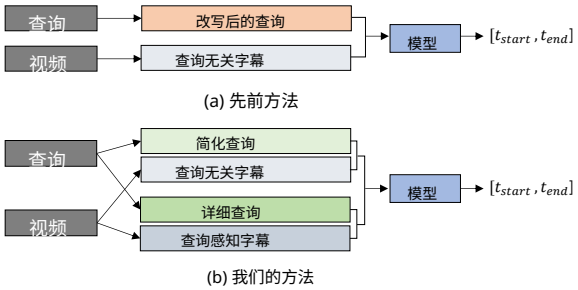

图3：概念框架对比。(a) 先前方法通常采用单路径方法，改写查询。(b) 我们的方法 GranAlign 采用双路径框架，将查询拆分为简化版和详细版，并分别与查询无关字幕和查询感知字幕对齐。

如图2所示，按查询类型（错误、简单、细节和其他）对性能进行细分后，可以发现显著的性能差距。这凸显了当前方法无法有效处理不同粒度层次的问题。现有方法即使具备联合空间中的高质量预训练知识，也从根本上无法解决这一权衡，因为它们缺乏粒度感知对齐的显式机制。

这一问题的根本原因在于，现有方法将查询视为一个整体进行处理。一种直观的解决方案是使用各种改写来扩展原始查询，正如图3（a）所示，这是此前多项工作采用的策略。然而，这种“一刀切”的方法本身存在缺陷。即使改写后的查询集合足够多样，只要局限于单一粒度，就无法同时体现高召回率所需的广泛范围和高精确率所需的细粒度细节。这种方法通常只能采取折中方案，既无法在任何一方面表现出色，也无法适应不同查询—视频对之间多样的语义复杂度。这种单一路径的推理方式正是阻碍稳健、准确检索的核心瓶颈。

为解决这些局限性，我们提出粒度感知对齐（Granularity-Aware Alignment，GranAlign），这是一种新颖的免训练框架，可在互补的粒度层面对查询和视频内容进行建模与对齐。如图3（b）所示，GranAlign 摒弃了单路径设计。具体而言，在查询侧，它利用大语言模型（Grattafiori 等，2024）将原始查询改写为两条不同路径：简化查询，用于捕捉核心意图、实现广泛覆盖；详细查询，则保留特定细节以确保精确性。相应地，在视频侧，视觉语言模型（Bai 等，2025）会生成一条通用的查询无关字幕和一条聚焦的查询感知字幕。通过按粒度对齐这些配对——简化查询与查询无关字幕相匹配，详细查询与查询感知字幕相匹配——GranAlign 将通用路径的高召回能力与特定路径的高精确率结合起来，
从而在多个具有挑战性的基准测试中实现更稳健的对齐和更高的检索准确率。

2 相关工作

2.1 零样本视频片段检索

零样本视频时刻检索（ZVMR）旨在利用预训练模型，在无需特定任务训练的情况下定位视频中的相关时刻。此前的研究使用现成的视觉语言模型和基于 BLIP-2 的方法（Li 等人，2023a），在 QVHighlights 数据集上取得了出色的表现（Diwan，2023年；Wattasseril 等人，2023年）。然而，这些方法通常依赖粗粒度的帧级匹配或受限的基于片段的候选方案，因而难以实现细粒度的时间对齐和精确的语义定位。近期的进展，如 Moment-GPT（Xu 等人，2025年），利用大语言模型（LLM）改写查询，并使用 VideoChatGPT（Maaz 等人，2023）为候选时刻打分，从而取得了最先进的性能。尽管架构先进，Moment-GPT 在改写后的查询与视觉内容之间实现准确的语义对齐方面仍面临困难，妨碍了精确的时刻定位。下一小节将从语义粒度的角度分析这一对齐问题。

2.2 检索中的粒度

视觉 和语言流中的细粒度语义粒度对于准确的视频定位至关重要。早期方法依赖粗略的帧级线索或单句查询（Hendricks 等人，2017；Gao 等人，2017a），而后续工作则引入了更密集的视觉片段（Zeng 等人，2020；Song 等人，2021）和更复杂的查询推理（Lin 等人，2019）。然而，现有方法仍将完整查询与通用字幕或原始视觉特征对齐，却没有明确管理语义粒度。

近期研究仍然存在这一局限。Diwan 等人（Diwan（2023年））利用现成的字幕将零样本视频时刻检索提升了 2.5×，但这些字幕仍然与查询无关。上下文增强型视频时刻检索（Liu 等人，2024）引入了 BLIP-2 摘要，但缺乏与查询意图的对齐。Moment-GPT（Xu 等人，2025）通过 LLaMA-3 改写查询，并使用冻结的多模态大语言模型为片段打分，但仍将每个

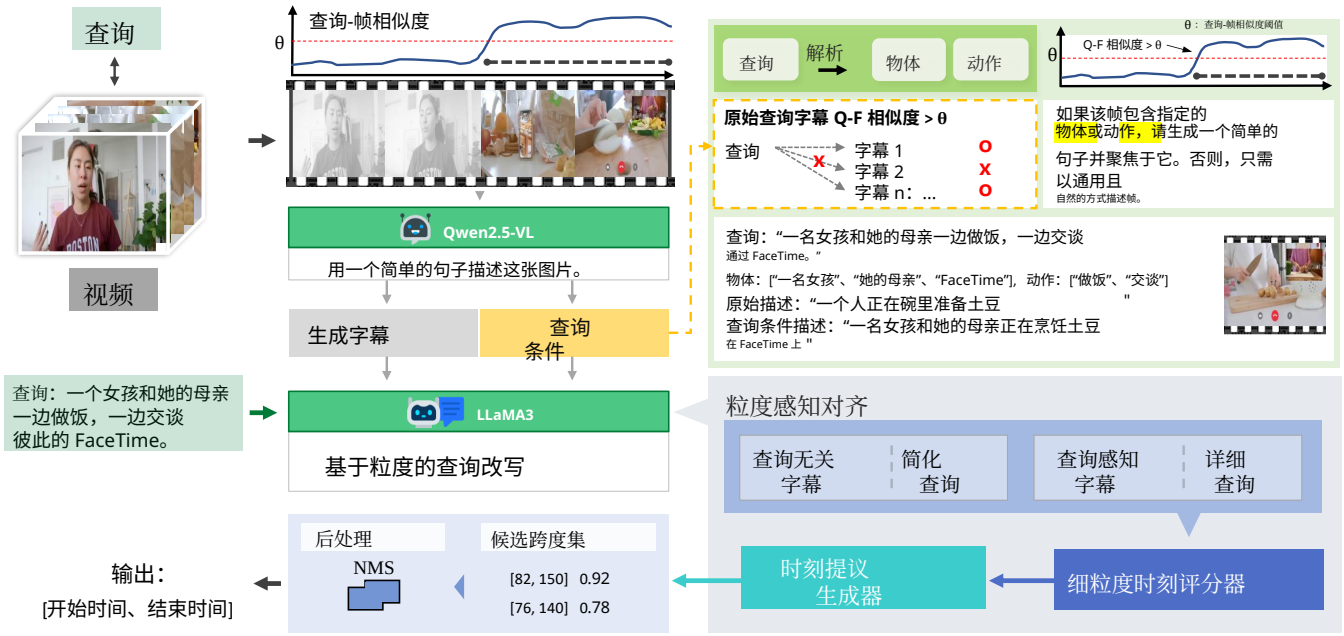

图 4：GranAlign框架概览。在粒度感知对齐（第 3.2 节）中，输入查询会被改写为两种语义粒度（简化/详细），并与查询无关或查询感知的字幕（以查询为上下文生成的字幕）进行匹配，从而为每个视频片段计算一个时刻分数。这些分数用于驱动时刻提议生成器和 NMS 阶段，随后由后排序模块生成最终预测值（第 3.3 节）。**要点：** GranAlign通过利用查询与视频内容之间细粒度的语义对齐，有效解决了零样本视频时刻检索任务。

查询视为一个不可区分的整体，并依赖与查询无关的字幕，从而可能导致语义不匹配。因此，近期方法在实现稳健的时刻检索方面仍有不足。这一持续存在的差距凸显了开发一种能够显式建模并对齐互补粒度内容的框架的必要性。我们的 GranAlign 正是为应对这一挑战而设计的。

第3节 方法

在本文中，我们提出了一种新颖的框架 GranAlign（如图 4 所示），用于在查询与视频内容之间建立细粒度的语义对齐，从而提升零样本视频时刻检索的鲁棒性和精确率。

3.1 概述

给定一个包含 L_v 帧的未剪辑视频 $V = \{v_i\}_{i=1}^{L_v}$，以及一个由 L_q 个词组成的文本查询 $Q = \{q_i\}_{i=1}^{L_q}$，视频时刻检索（VMR）的目标是预测一组时间跨度 $T = \{t_s, t_e\} \in \mathbb{R}^{N_t \times 2}$。其中，$t_s$ 和 t_e 表示预测片段的起始和结束时间，N_t 则表示预测片段的数量。

我们的框架由三个阶段组成。首先，为实现粒度感知对齐，我们将输入查询改写为简化版本和详细版本，并生成两个对应的字幕集合——查询无关和查询感知字幕集合（第 3.2 节）。随后，利用这些集合为每个视频片段计算粒度时刻分数（第 3.3 节）。其次，根据这些分数生成候选时间片段（时刻候选片段）。最后，进行后处理，其中包括采用非极大值抑制（NMS）选出最相关的片段。

3.2 粒度感知对齐

本节介绍我们方法的核心思想——粒度感知对齐，该方法由三个主要组件构成：基于粒度的查询改写、查询感知的描述生成，以及时刻分数计算。

基于粒度的查询改写为实现更稳健的语义对齐，我们通过使用 LLaMA-3（Grattafiori 等，2024）进行指令引导重写，将输入查询Q改写为语义互补的多种表征。要用一个单一、僵化的数学公式来定义"简单"和"详细"颇具挑战，因为这类定义无法捕捉自然语言的细微差别。因此，我们不依赖单一提示，而是使用多个手工编写的指令对，生成简化查询（Q_s）和详细查询（Q_d）。图 5 展示了指令集中的代表性示例。[1] 通过实证我们发现，框架的性能稳健，且不会对所用的特定指令对过度敏感。简化查询 Q_s 通过提示模型"用常见的替代词替换生僻词"，同时"保留核心实体和动作"而生成。为提高清晰度，我们会删除生僻或附带的细节，并

[1]我们实验中使用的完整指令对集合及其对性能影响的敏感性分析，详见附录 E。

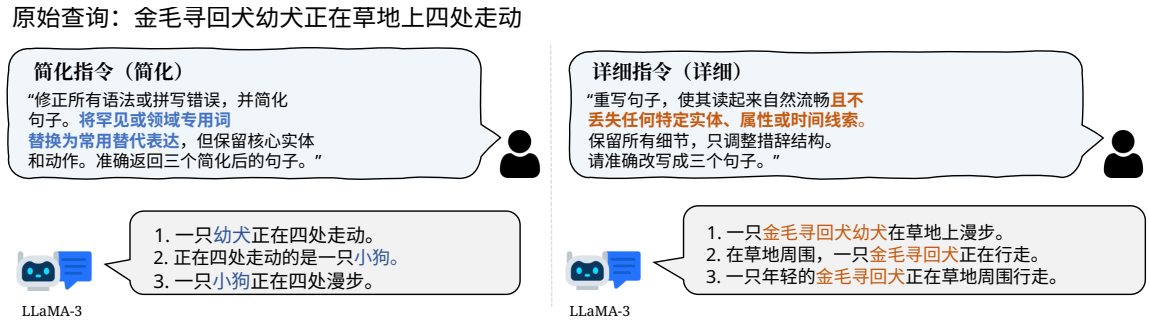

图 5：基于粒度的查询改写指令对示例。该图展示了提供给 LLaMA-3 的示例提示词，用于生成简化（侧重召回率）和详细（侧重精确率）的查询。

纠正语法错误。这种形式具有很强的通用性，能够有效检索广泛相关的候选片段。相比之下，详细查询 Q_d 保留细粒度表达、时间上下文和具体的词汇选择。这些元素有助于实现更精确的对齐和更准确的时刻定位。

查询感知的字幕生成 传统的视频字幕生成方法仅根据视觉内容生成描述，因此难以准确识别在语义上与查询相符的片段。即使字幕在视觉上正确，也可能缺少准确检索所需的查询意图。

为克服这一局限，我们旨在将查询意图直接融入字幕。然而，对每一帧都应用这一过程在计算上难以承受。因此，我们提出一种混合策略，在语义精确性与计算可行性之间取得平衡。该策略首先根据查询 Q 与各帧之间的相似度分数，选出得分最高的前 K%，以确定一组候选帧。令 L_k 表示这些候选帧的数量。随后，利用 Qwen2.5-VL（Bai 等，2025）生成两种字幕。首先，为所有 L_v 帧生成通用的、查询无关字幕 ($C_{agn} \in \mathbb{R}^{L_v \times l}$)，作为基线。其次，仅为 L_k 候选帧生成聚焦的、查询感知字幕 ($C_{awr} \in \mathbb{R}^{L_k \times l}$)。

为了生成 C_{awr}，我们使用从查询中提取的关键语义元素来引导生成过程。例如，对于“某人从桌上拿起一支铅笔”，我们提取实体 {人物、铅笔}以及动作 {拿起} 作为语义引导。这种混合策略使我们的框架能够利用查询感知字幕的语义精确性，重点关注最关键的区域，同时保持整体计算效率。

与查询无关的字幕相比，这一过程使 C_{awr} 能够与查询实现更强的语义对齐，从而提高时刻定位的准确性。然而，这种方法可能导致失败，例如臆造视频中不存在的视觉内容，或过度模仿查询的语言结构。因此，我们最终的时刻评分过程旨在发挥 C_{awr} 的优势，同时减轻这些潜在问题。

细粒度时刻评分两组查询-字幕对，简化无关（Q_s， C_{agn}）和详细感知（Q_d， C_{awr}），呈现出不同的特点，优势与局限互为补充，如图6所示。简化查询 Q_s 和通用字幕 C_{agn} 能够广泛覆盖语义相似的场景，从而提高召回率，但往往缺乏精确定位确切时刻所需的精确率。相比之下，详细查询 Q_d 和查询感知字幕 C_{awr} 借助细粒度语义实现更准确的对齐，但也更容易因幻觉或过度依赖查询而出现错误或错位。这凸显了匹配粒度级别对于成功对齐的重要性，这正是我们方法的核心原则。

为了利用两种查询-字幕对各自的互补优势，我们通过整合 (Q_s, C_{agn}) 和 (Q_d, C_{awr}) 的语义相似度，计算综合时刻分数。具体而言，我们使用两种查询-字幕对评估每个候选时刻的语义对齐程度，并取平均值作为最终分数。组合后的帧级相似度分数 S_f 定义如下：

$$S_f = \frac{1}{2m} \sum_{i=1}^{m} \left[\mathrm{g}(q_s^{(i)}, C_{agn,f}) + \mathrm{g}(q_d^{(i)}, C_{awr,f}) \right] \quad (1)$$

通过对改写后的查询与其对应图注之间的语义相似度取平均值，计算视频中每一帧 f 的相似度分数 S_f。每对改写后的查询都包含一个简化查询 $q_s^{(i)}$ 和一个详细查询 $q_d^{(i)}$，两者均通过受控改写从原始输入查询生成，其中 m 表示此类查询对的总数。$C_{agn,\ f}$ 和 $C_{awr,\ f}$ 分别表示查询无关字幕集合和查询感知字幕集合中第 f 帧对应的字幕嵌入。这里，g(·, ·) 表示查询与图注的句子嵌入之间经过归一化的余弦相似度。

这种表述能够减轻依赖单个查询-图注对可能产生的偏差或误报，从而实现更稳健、可靠的时刻检索。我们首先使用 SentenceTransformer 模型（Reimers 和 Gurevych，2019年）对查询和图注进行嵌入，然后计算它们的余弦相似度。随后，利用所得的帧级分数 S_f 生成候选片段并对其排序。

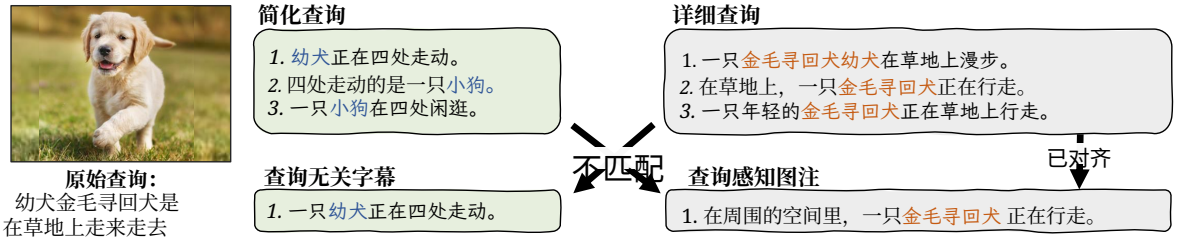

图 6：不同粒度级别下语义对齐模式的示意图。简化查询（例如“幼犬”）与查询无关字幕对齐良好，而详细查询（例如“金毛寻回犬”）则需要查询感知图注。粒度不匹配可能导致对齐失败。

3.3 时刻候选生成

在这一阶段，利用计算得到的片段分数 通过时刻提议生成器生成候选片段，然后通过后处理 步骤优化最终结果。

时刻候选生成器时刻候选生成器（MPG）根据帧级相似度分数 S_f生成语义连贯的候选片段。为防止单个连续事件被切分成多个短片段，如果相邻高分帧之间的时间间隔在 τ 帧的阈值范围内，MPG 就会将它们合并为一个候选片段。虽然根据τ 合并帧能够确保连续性，但这可能生成被稀释的候选片段，即高分关键帧之间连接着一系列相关性较低的帧。为缓解这一问题，我们会丢弃平均相似度处于最低 n% 的片段，以保持相关性。这是至关重要的质量控制步骤，可确保只有语义相关性始终较高的候选片段才能进入下一阶段。

我们根据语义相关性和长度正则化对每个候选片段 p 进行评分。第一项 μ_p是片段内所有帧的平均语义相似度。第二项 ρ_p是片段 p 的长度占所有候选片段总长度的比例。该归一化长度 ρ_p 是至关重要的正则化因子，可惩罚平均相关性较低的过长候选片段，并防止偏向无意义的短片段，从而确保最终候选片段具有合理的时长。每个候选片段 p的最终得分计算如下：

$$\text{Score}(p) = (1-\lambda)\mu_p + \lambda\rho_p, \quad 0 \le \lambda \le 1. \qquad (2)$$

参数 λ（设为 0.3）用于平衡这一权衡。这种正则化评分比简单的基于阈值或长度的方法能提供更稳健的候选项。

后处理 将片段提议生成器 生成的所有候选片段汇集到一个集合中，每个片段由其起始帧和结束帧的索引定义。该集合中的任意两个片段 p_i 和 p_j 代表不同的候选区间。为进一步筛选候选片段集合，我们采用非极大值抑制（NMS），迭代丢弃任意两个交并比（IoU）超过阈值 θ_{NMS}的片段中得分较低者，从而得到最终的片段预测值 $\{t_s, t_e\}$。

4 实验

本节详细介绍了为验证所提出的**GranAlign**框架有效性而开展的实验。我们将介绍实验设置、与最先进方法的性能比较、分析核心组件贡献的消融研究，以及进一步的分析，以展示我们方法的稳健性。

数据集 我们在公开基准测试上进行评估：*QVHigh-lights* （Lei、Berg 和 Bansal，2021年）、*Charades-STA* （Gao 等人，2017b）和 *ActivityNet-Captions*（Krishna 等人，2017年）。**评估指标** 遵循先前工作，我们报告标准的视频时刻检索分数，包括召回率@1（R1@n），即在给定 IoU 阈值 n 下，排名第一的预测正确的查询所占百分比；以及平均精度均值（mAP@n），用于评估整体定位精确率。数据集的完整统计信息以及指标的详细定义见附录 D。

4.1 与最先进方法的比较

表1显示，GranAlign 在 QVHighlights 上始终优于此前的零样本最先进水平（Xu 等人，2025）。在验证集上，各项指标均提升了 **+3.04% 至 +3.93%**，其中 mAP@0.5 的增幅最大；在隐藏测试集上，仍领先 **+1.6% 至 +3.84%**。这些提升证实，基于粒度的查询改写和查询感知图注生成增强了语义对齐，进而提高了检索准确率。表2显示，该方法在迁移到 Charades-STA 和 ActivityNet-Captions 时表现出色。在 Charades-STA 上，GranAlign 的表现比此前的最先进方法高出 +1.2%（R1@0.5）和 +1.5%（平均交并比）；在 ActivityNet-Captions 上，则分别提升 +2.9%（R1@0.5）和 +2.3%（平均交并比）。因此，所提出的对齐策略可从较短视频扩展至较长且内容多样的视频，在各项基准测试中实现稳健的零样本时刻检索。

4.2 消融研究

我们在 QVHighlights 验证集上进行消融研究，以评估框架中各组件的作用，所有结果汇总于表3。

符号说明 在以下消融实验中，为清晰起见，我们重新说明符号： Q_r 表示原始查询， Q_s表示简化查询； Q_d表示详细查询； C_{agn}表示查询无关字幕； C_{awr}表示查询感知字幕。

方法	MLLM	设置	QVHighlights 测试集				QVHighlights 验证集			
			R1@0.5	R1@0.7	mAP@0.5	mAP@avg	R1@0.5	R1@0.7	mAP@0.5	mAP@avg
VTimeLLM（Huang 等，2024）	✓	FS	47.2	29.3	47.3	27.4	48.8	29.5	49.3	26.8
LLaViLo（Ma 等，2023年）	✓	FS	48.6	29.7	48.7	27.9	49.0	30.4	49.4	28.9
Moment-DETR（Lei、Berg 和 Bansal，2021年）	–	FS	52.9	33.0	54.8	30.7	54.2	33.4	55.4	31.1
MomentDiff（Li 等，2023b）	–	FS	-	-	-	-	57.8	39.2	54.6	35.3
CPL（He 等，2022年）	–	WS	30.8	10.8	22.8	–	–	–	–	–
CPI（Kong 等人，2023年）	–	WS	32.3	11.8	23.7	–	–	–	–	–
Liu 等人（Zheng 等人（2022））	–	US	–	–	–	–	12.3	3.5	10.4	2.7
PZVMR（Wang 等人（2022））	–	US	14.2	4.9	15.7	4.6	12.6	5.1	16.2	5.3
VideoLLaMa（Zhang、Li 和 Bing（2023））	✓	ZS	17.1	6.7	18.2	6.2	18.5	6.9	17.8	7.1
VideoChatGPT（Maaz 等人，2023）	✓	ZS	21.1	10.2	22.8	9.5	22.4	10.8	21.9	10.3
UniVTG（Lin等人（2023年））		ZS	25.2	9.0	27.4	10.9	–	–	–	–
Diwan（Diwan（2023年））	✓	ZS	–	–	–	–	48.3	31.0	47.3	28.0
Moment-GPT（Xu 等人，2025）	✓	ZS	58.3	37.7	55.1	35.0	58.9	38.6	55.7	35.9
GranAlign（我们的方法）	✓	ZS	**59.92**	**39.3**	**58.94**	**38.23**	**61.94**	**41.81**	**59.63**	**39.12**

表 1：QVHighlights 测试集和验证集上的性能比较。“MLLM”表示使用多模态大语言模型，“FS/WS/US/ZS”分别表示全监督/弱监督/无监督/零样本设置。

方法	MLLM	设置	Charades-STA				ActivityNet-Captions			
			R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
GroundingGPT（Li等人，2024）	✓	FS	–	29.6	11.9	–	–	–	–	–
VTimeLLM（Huang 等，2024）	✓	FS	55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4
TimeChat（Ren 等人，2024 年）	✓	FS	–	43.8	22.7	–	–	–	–	–
Moment-DETR（Lei、Berg 和 Bansal，2021年）	–	FS	62.1	48.2	25.3	42.3	52.6	32.5	15.3	37.8
CPL（He 等，2022年）	–	WS	56.0	38.1	20.3	37.8	52.4	30.9	12.0	32.6
Huang等人（Huang、Yang 和 Sato，2023年）	–	WS	59.2	44.2	22.1	39.4	54.8	32.9	–	36.4
PSVL（Nam等人，2021年）	–	US	45.2	30.9	14.2	30.9	45.1	29.8	15.73	30.2
Liu 等人（Zheng 等人（2022））	–	US	44.2	28.7	14.7	–	47.3	28.2	–	–
TimeChat（Ren 等人，2024 年）	✓	ZS	–	32.2	13.4	–	–	–	–	–
Luo 等人（Luo 等人 2024）	✓	ZS	53.4	36.0	19.3	34.1	45.6	27.4	12.3	28.4
Moment-GPT（Xu 等人，2025 年）	✓	ZS	58.2	38.4	21.6	36.5	48.1	31.1	14.9	30.8
GranAlign（我们的方法）	✓	ZS	**59.1**	**39.6**	**22.7**	**38.0**	**50.3**	**34.0**	**16.5**	**33.1**

表2：在不同监督设置下，Charades-STA 和 ActivityNet-Captions 上的性能比较。

C_{agn}	C_{awr}	R1@0.5	R1@0.7	mAP@0.5	mAP@avg
Q_r	-	57.94	38.84	51.64	31.8
-	Q_r	58.19	39.03	52.12	32.13
Q_s	-	58.97	40.71	56.88	37.13
-	Q_s	57.55	38.39	56.02	36.22
Q_d	-	58.19	39.42	56.4	36.82
-	Q_d	59.48	40.52	57.01	37.65
Q_s	Q_s	61.03	**42.06**	59.12	38.78
Q_d	Q_s	60.90	41.26	58.83	38.63
Q_s	Q_d	**61.94**	41.81	**59.63**	**39.12**
Q_d	Q_d	61.19	41.35	59.52	39.01

表3：汇总的消融实验结果。我们用 Q 表示查询，用 C 表示字幕，下标分别表示原始（r）、简化（s）、详细（d）、查询无关（agn）和查询感知（awr）。(Q_s, Q_d) 这一行表示我们的完整模型。

查询感知字幕的影响 如表3所示，在原始查询中使用查询感知字幕（C_{awr}），相比使用查询无关字幕（C_{agn}），在所有指标上均取得了稳定提升。这表明，根据查询内容定制字幕能够增强语义对齐。

粒度对齐配对的有效性我们在表3中分析了粒度对齐的影响。性能会在查询与图注的粒度不匹配时下降（例如，将 Q_s 与 C_{awr} 配对），而粒度对齐的配对（例如，将 Q_d 与 C_{awr}配对）则表现更好。

然而，这些单一对齐对存在权衡：简化对（Q_s, C_{agn}）召回率高，但精确率低；而详细对（Q_d, C_{awr}）尽管在单对方案中表现最佳，精确率高，却可能无法捕捉细微的不匹配。我们的 GranAlign 框架通过结合这两种配对方案，利用它们互补的优势，解决了这一问题。因此，GranAlign 的整体表现最佳，并且在所有指标上均取得了稳定提升。我们还通过采用 BLIP-2 主干网络，证实了该方法的稳健性（请参见附录A）。

实现细节我们使用 LLaMA3-8B（Grattafiori 等，2024年）进行查询改写，并使用 Qwen2.5-VL-7B（Bai 等，2025年）生成图注；初始帧则通过 CLIP ViT-B/32（Radford 等，2021年）筛选。所有实验均在四块 NVIDIA A6000 GPU 上运行，实现细节的完整列表见附录B。

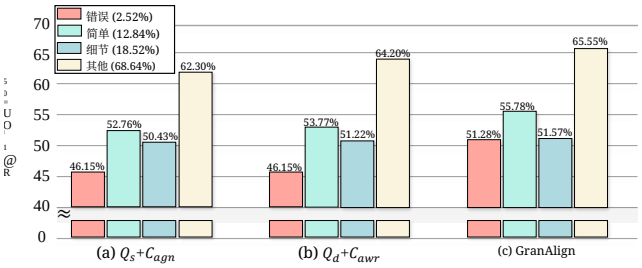

图7：QVHighlights 验证集上，简化、详细、错误和其他类型查询的性能分析。为清晰起见，Y 轴在 40 以下截断。

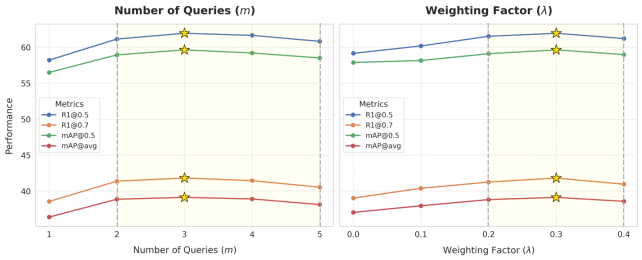

图8：不同超参数设置下的性能分析。黄色阴影区域突出显示了我们的框架在最优设置（以星号标记）附近表现稳健且稳定的参数范围。

4.3 进一步分析

对查询变化的鲁棒性如图7所示，我们评估了 GranAlign 在四种语言特征各异的查询类型（错误、简化、详细和其他）下的鲁棒性。[2]基线组合 (a)（Q_s、C_{agn}）和 (b)（Q_d、C_{awr}）表现互补，但各有局限。前者广泛召回率较高，但精确率不足；后者能够捕捉细粒度语义，却容易过度拟合查询，从而导致幻觉或对齐偏差。相比之下，我们提出的 GranAlign 将简化和详细的查询改写与查询感知图注生成相结合，通过有效平衡召回率和精确率，在所有类别中都实现了稳健的性能。

敏感性分析 如图8所示，我们提出的新框架 GranAlign 对不同的超参数设置具有鲁棒性。例如，使用 3 个改写后的查询（m=3）和 0.3 的加权因子（λ=0.3）时，性能达到峰值，同时在附近的一系列取值范围内保持稳定。关于超参数选择的完整实验详见附录C。

效率分析 在表4中，我们将训练成本、GPU内存使用量和推理时间与全监督方法及零样本方法进行了比较。作为一种零样本方法，GranAlign 无需训练成本，并为提高效率，预先离线生成查询无关字幕。随后，

[2]根据长度将查询分类为：简单（≤6 个词元）、细节（≥20 个词元或包含专有名词）或其他（所有其他查询）。此外，语法错误的查询会被单独标记为错误。

设置	方法	R1@0.5	mAP	TrC	InT（秒）	GMU（G）
ZS	VideoChatGPT（Maaz 等人，2023）	22.4	10.3	0	9.8	11
FS	VTimeLLM（Huang等人，2024年）	48.8	26.8	4090，40小时	11.2	18
ZS	Wattersseril（Wattasseril 等人，2023年）	53.1	30.2	0	12.7	14
ZS	Moment-GPT（Xu 等人，2025年）	58.9	35.9	0	16.1	32
ZS	Ours	**61.94**	**39.12**	**0**	**6.2**	**22**

表4：QVHighlights 数据集上的训练成本（TrC）、GPU内存使用量（GMU）和推理时间（InT）对比。



图9：QVHighlights上的定性结果。详细查询中的字幕'表示查询感知字幕。

在在线处理过程中，它仅针对超过特定阈值的关键帧选择性地生成额外字幕。这种两阶段方法有助于缩短推理时间，并使模型保持较低的内存占用。因此，GranAlign取得了最佳性能，在R1@0.5和平均精度均值等大多数关键指标上优于其他配置，同时推理时间也是最短的。

定性结果图 9 定性展示了 GranAlign 如何利用语义粒度，在零样本设置下准确定位时刻。对于查询“一个戴着印有爱心的白色口罩的亚洲女孩走在街道上”，真实值（GT）片段的时间范围为 90 至 142 秒。简单路径将简化查询与查询无关字幕进行对齐，无法捕捉具体属性，而是依赖一般信息，因此预测范围过长，从 78 秒开始，包含了无关内容。相比之下，详细路径将详细查询与查询感知字幕进行对齐，能够识别“爱心”等细粒度线索，并从更准确的位置开始，但片段末尾略有欠分割（漏掉了 138 秒之后的帧）。GranAlign 融合两条路径的对齐分数，成功预测了准确的真实值片段（90–142 秒），兼具两种方法的优势。这一定性结果清楚地表明，我们的粒度化设计是克服覆盖率与精确率之间核心权衡的关键，从而实现更准确、稳健的时刻定位。

5 讨论与局限性

尽管 GranAlign 性能出色，但其局限性源于核心组件的生成特性，也为未来研究指明了明确方向。例如，可以通过一种事实核查机制来降低查询感知字幕生成中的幻觉风险，核实生成的细节是否与视觉证据一致。同样，

可以加入语义验证步骤，确保基于大语言模型的查询重写能够保留原始意图。推进这些改进，将进一步提升粒度感知时刻检索系统的可靠性和泛化能力。

6 结论

我们提出了粒度感知对齐（GranAlign），这是一种用于零样本视频时刻检索（ZVMR）的新型免训练框架。通过在多个语义粒度层面对齐查询与视频内容，GranAlign 化解了覆盖率与精确率之间的权衡，兼顾了通用表征的高召回率与细粒度属性的高精确率。因此，GranAlign 无需特定任务训练成本，便在多个主要基准上达到了最先进水平，为零样本视频时刻检索开辟了新的可能性。我们认为，我们提出的粒度感知对齐策略为未来的多模态理解系统提供了一个前景广阔且有效的发展方向。

附录

A 进一步分析

A.1 对下游任务的泛化能力

为进一步验证粒度感知设计的有效性，我们在视频高光检测（VHD）任务上进行了消融实验，如表5所示。我们首先评估了 GranAlign†，这是一个仅使用详细查询和查询感知字幕生成（Q_d, C_{awr}）的变体。尽管这一设置已取得出色表现，但完整的 GranAlign 模型——结合简化原始配对（Q_s, C_{agn}）和详细感知配对（Q_d, C_{awr}）——进一步提升了结果，达到 39.35% 的平均精度均值（mAP）和 66.34% 的命中率@1（HIT@1）。值得注意的是，GranAlign 的表现甚至优于全监督的 QD-DETR（mAP 为 39.04%，HIT@1 为 62.87%），这表明利用互补粒度对能够显著提升检索准确率。这些结果证实，我们的双粒度对齐策略不仅对时刻检索至关重要，也能广泛迁移至相关下游任务。

A.2 粒度感知配对

为了验证在使用 BLIP2 生成的图注时，消融实验中观察到的相同粒度模式是否依然成立，我们使用 BLIP2 图注重新进行了消融实验。如表 6 所示，混合粒度设置 (b) 和 (c) 再次优于同质设置，其中配置 (c) 的平均精度均值最高，从而证实了 GranAlign 在不同图注生成骨干网络上的稳健性。

B 实现细节

本节详细描述了我们的方法所使用的实现细节和超参数。(i) 帧级字幕由 Qwen2.5-VL-7B（Bai 等，2025）以 0.5 帧/秒的采样率生成。随后，我们使用 CLIP ViT-B/32 模型（Radford 等，2021）从帧和查询中提取特征，并仅选择帧级相似度排名位于前 10%（K）的字幕进行后续处理。

模型	设置	mAP	命中率@1
TimeChat（Ren 等人，2024 年）	FS	14.5	23.9
Moment-DETR（Lei、Berg 和 Bansal，2021年）	FS	35.7	55.7
QD-DETR（Moon等，2023年）	FS	39.04	62.87
Moment-GPT（Xu 等人，2025）	ZS	36.7	62.7
GranAlign†	ZS	37.41	64.98
GranAlign（我们的方法）	ZS	**39.35**	**66.34**

表 5：完全监督/零样本设置下的 VHD 性能比较。

配置	C_{agn}	C_{awr}	平均交并比	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	mAP@avg
(a)	Q_s	Q_s	54.53	58.00	38.71	55.05	36.33	36.20
(b)	Q_d	Q_s	54.50	57.74	38.45	55.16	36.36	36.30
(c)	Q_s	Q_d	56.00	59.32	40.68	56.71	38.05	37.80
(d)	Q_d	Q_d	55.05	58.45	39.81	55.71	37.01	36.73

表6：不同 C_{agn}/C_{awr}设置下的 BLIP2 消融实验结果。

（ii）使用 LLaMA3-8B（Grattafiori 等，2024）执行查询改写。通过实验我们发现，生成 3 个改写后的查询（m）可获得最佳性能。改写后的查询 Q' 旨在保留原始查询 Q 的语义含义。

（iii）对于候选片段评分，我们使用基于相似度分布的逆累积直方图。该分布被划分为 10 个区间，并选取位于前 8 个区间内的候选片段。根据分数分布动态调整片段选择，以纳入处于边界的情况。此外，候选片段之间最多 τ=6 个相似度较低的连续帧间隔将被合并为一个片段。此处，相似度是在去偏查询与经过条件处理的查询以及原始查询之间计算的；如果片段的相似度位于分布底部 20%（n）范围内，则不参与合并。(iv) 初始阶段和第二阶段的片段选择阈值根据验证集性能通过经验确定。片段评分器中的加权因子 λ 设为 0.3。

(v) 最终片段选择采用基于 IoU 的非极大值抑制（NMS），阈值为 0.9（θ_{NMS}）。所有实验均使用 4 块 NVIDIA A6000 48GB GPU 进行。

C 超参数分析

本节介绍了消融研究，旨在确定我们框架中超参数的最优取值。

m	R1@0.5	R1@0.7	mAP@0.5	mAP@avg
1	58.21	38.55	56.49	36.38
2	61.14	41.37	58.94	38.86
3	**61.94**	**41.81**	**59.63**	**39.12**
4	61.65	41.45	59.20	38.90
5	60.82	40.53	58.51	38.12

表 7：QVHighlights 验证集上查询数量（m）的消融实验。

λ	R1@0.5	R1@0.7	mAP@0.5	mAP@avg
0.0	59.15	39.02	57.88	37.03
0.1	60.18	40.38	58.15	37.95
0.2	61.52	41.25	59.10	38.81
0.3	**61.94**	**41.81**	**59.63**	**39.12**
0.4	61.21	40.95	58.98	38.57

表8：在 QVHighlights 验证集上对加权因子（λ）进行的消融实验。

K（%）	R1@0.3	R1@0.5	mIoU
5%	58.2	38.5	37.1
10%	**59.1**	**39.6**	**38.0**
20%	58.8	39.1	37.7
30%	57.9	38.2	36.9

表 9：在 Charades-STA 数据集上对字幕选择比例（Top-K%）进行的消融实验。

τ	R1@0.3	R1@0.5	mIoU
0	57.5	37.9	36.5
2	58.6	39.0	37.4
4	58.9	39.4	37.8
6	**59.1**	**39.6**	**38.0**
8	58.3	38.8	37.2

表 10：在 Charades-STA 数据集上对片段合并间隔（τ）进行的消融实验。

n（%）	R1@0.3	R1@0.5	mIoU
后 10%	58.7	39.2	37.6
后 20%	**59.1**	**39.6**	**38.0**
后 30%	58.2	38.6	37.1
底部 40%	57.6	38.0	36.4

表11：在 Charades-STA 数据集上对合并排除阈值进行的消融实验。

θ_{NMS}	R1@0.3	R1@0.5	mIoU
0.8	58.1	38.8	37.0
0.85	58.8	39.3	37.7
0.9	**59.1**	**39.6**	**38.0**
0.95	58.9	39.5	37.9

表12：在 Charades-STA 数据集上对 NMS IoU阈值（θ_{NMS}）进行的消融实验。

D 数据集、评估指标

数据集 QVHighlights（Lei、Berg 和 Bansal，2021年）是一个基准数据集，包含 10,148 个 YouTube 视频，每个视频时长约 150 秒，并包含 7,218/1,550/1,542 个训练/验证/测试查询。每个查询对应一个或多个标注片段。测试标签未公开；评估通过公共服务器进行。

Charades-STA（Gao 等人，2017b）包含 9,848 个室内视频，平均时长为 31 秒，以及 16,128 个查询-片段对。我们采用标准的 12,408/3,720 训练/测试划分，并抽取 10% 的训练数据用于验证。

ActivityNet-Captions（Krishna等人，2017）是一个基准数据集，包含 20,000 段总时长约 849 小时的未剪辑视频，并配有约 100,000 条字幕（每段视频平均包含 3.6 个片段），广泛用于密集事件图注和时刻检索。

评估指标 为确保公平比较，我们遵循先前工作的评估协议（Xu 等人，2025；Huang 等，2024），采用标准视频时刻检索指标：R1@n、mAP@m、mAP@avg 和平均交并比。具体而言，R1@n 表示排名第一的预测与真实标注之间的 IoU 大于 n 的查询所占比例。mAP@m 指 IoU 阈值为 m 时的平均精度均值，而 mAP@avg 则计算 IoU 阈值从 0.5 到 0.95、步长为 0.05 时的平均精度均值。平均交并比衡量所有查询中预测片段与真实标注片段之间的最大 IoU 平均值，用以反映定位准确度。对于视频精彩片段检测（VHD），我们报告命中率@1，用于衡量排名最高的片段是否与真实标注匹配；同时报告平均精度均值，用于评估排序质量。

E 指令对与敏感性分析

本节给出了第3.2节所述基于粒度的查询改写所使用的指令对集合（如图10所示）以及每个指令对对应的性能。我们进行了敏感性分析，以通过实验证实我们的观点：该框架的性能具有稳健性，不会过度依赖提示词的具体措辞。

为通过实验证实我们方法的稳健性，我们在QVHighlights验证集上评估了五组指令对中的每一组的性能。表13中的结果表明，不同措辞下的性能具有高度一致性。例如，R1@0.5得分均集中在一个很窄的0.7%范围内，最低为61.26%，最高为61.94%。这种较低的方差为以下结论提供了有力的实证证据：我们的框架学习并泛化了“简化”和“细节保留”这两个底层概念，而不是对单一的精调提示词过拟合。该分析验证了正文中的论断，即我们的基于粒度的改写方法对于指令提示的变化具有稳健性。对于我们的主要

Pair	**R1@0.5**	**R1@0.7**	**mAP@0.5**	**mAP@avg**
1	**61.94**	**41.81**	59.63	**39.12**
2	61.81	41.60	59.27	39.01
3	61.63	41.10	58.99	38.65
4	61.52	41.35	**59.78**	38.93
5	61.26	41.42	59.45	38.53

表 13：指令对在 QVHighlights 验证集上的敏感性分析。所有五组提示对均取得了稳定的高性能，这表明我们的查询改写方法具有鲁棒性，且不会过拟合特定的提示措辞。

在实验中，我们选择配对1作为最佳组合，因为它在大多数关键指标上表现最佳，包括R1@0.5、R1@0.7和mAP@avg。

任务-S（简化）	任务-D（详细）
## 配对1 更正语法或拼写错误，并简化句子。 将生僻词或领域专用词替换为常用替代表达，但 保留核心实体和动作。请仅返回三个简化后的 句子。	改写句子，使其读起来自然流畅，同时不遗漏任何 具体实体、属性或时间线索。保留所有细节，只需重新 组织措辞。请恰好返回三个改写后的句子。
## 配对2 校对给定句子，检查是否存在语法或拼写 纠正错误。通过替换任何 术语或生僻词来简化句子，换成更常见的表达。确保 主要主体及其动作保持不变。请提供 恰好三句简化后的句子。	改写以下句子，使其读起来更自然。你必须 保留所有原始细节，包括所有具体实体及其 属性以及任何与时间相关的信息。只改变结构 和措辞。请准确提供三个改写版本。
## Pair3 请改正以下句子中的拼写或语法错误。 请用更简单的语言改写，但务必保留原文 的含义，包括关键实体及其行为。请提供 这个句子的三个简化版本。	重构这个句子，让它听起来更自然，不那么 生硬。务必保留所有原有成分： 具体的人物或物品、所有描述性细节，以及任何时间 参考信息。提供三种不同的措辞。
## Pair4 检查这句话是否正确、清晰，并修正其中的错误 。修正语法或拼写错误后，用日常用语重新表述， 同时保留核心主题和动作。恰好返回三个 简化后的句子。	请重新调整这句话的结构，使其更易读、 更流畅。请勿删除或更改任何具体实体及其 特征，或任何对时间的提及。请准确返回三个改写后的 句子，并包含所有原始信息。
## Pair5 请根据以下说明编辑给定的句子： 纠正：消除所有拼写和语法错误。 简化：用更常见的词替换专业或生僻词。 保留：保持主要实体及其核心动作不变。 提供：给出恰好三个简化后的句子版本。	请按照以下规则修改句子： 目标：通过调整句子结构，使表达更加自然流畅。 限制：必须毫无例外地保留所有原始细节 （例如姓名、地点、描述、日期和时间）。 动作：仅修改措辞和句子结构。 输出：提供恰好三个改写后的句子。

图10：指令提示。

参考文献

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; 等。2025年。Qwen2.5-vl技术报告。*arXiv*预印本 *arXiv:2502.13923*。Diwan, A. 2023年。使用现成模型进行零样本视频片段检索。载于自然语言处理迁移学习研讨会，第10–21页。Gao, J.; Sun, C.; Yang, Z.; 和 Nevatia, R. 2017a。TALL：基于语言查询的时间活动定位。载于*IEEE*国际计算机视觉会议（*ICCV*）论文集，第5267–5275页。IEEE。Gao, J.; Sun, C.; Yang, Z.; 和 Nevatia, R. 2017b。TALL：基于语言查询的时间活动定位。*2017*年 *IEEE* 国际计算机视觉会议（*ICCV*），第5277–5285页。Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; 等。2024年。llama 3模型系列。*arXiv*预印本 *arXiv:2407.21783*。He, X.; Yang, D.; Feng, W.; Fu, T.-J.; Akula, A. R.; Ja m- pani, V.; Narayana, P.; Basu, S.; Wang, W. Y.; and Wang, X. 2022年。CPL：面向视觉语言模型的反事实提示学习。载于自然语言处理实证方法会议。Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Dar- rell, T.; and Russell, B. 2017. 使用自然语言定位视频时刻。载于*IEEE*国际计算机视觉会议（*ICCV*）论文集，5803–5812。IEEE。Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2024年。Vtimellm：让大语言模型理解视频时刻。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，14271–14280。Huang, Y.; Yang, L.; and Sato, Y. 2023年。通过不确定性引导的自训练进行弱监督时间句子定位。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，18908–18918。Kong, S.; Li, L.; Zhang, B.; Wang, W.; Jiang, B.; Yan, C.; and Xu, C. 2023年。采用伪

样本干预的弱监督联合视频时刻检索与高光检测动态对比学习。载于第*31*届*ACM*国际多媒体会议论文集，538–546。Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Car- los Niebles, J. 2017. 视频中的事件密集描述。载于*IEEE*国际计算机视觉会议论文集，706–715。Lei, J.; Berg, T. L.; and Bansal, M. 2021. 通过自然语言查询检测视频中的时刻和高光。神经信息处理系统进展，34：11846–11858。Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2：利用冻结图像编码器和大语言模型实现语言-图像预训练引导。载于国际机器学习会议，19730–19742。PMLR。Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023b。MomentDiff：从随机到真实的生成式视频时刻检索。神经信息处理系统进展，36：65948–65966。Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Vu, V. T.; 等。2024年。GroundingGPT：语言增强的多模态定位模型。*arXiv*预印本 *arXiv:2401.06071*。Lin, K.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A.; Yan, R.; and Shou, M. Z. 2023年。UniVTG：迈向统一的视频-语言时序定位。*2023 IEEE/CVF*国际计算机视觉会议（*ICCV*），2782–2792。Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2019年。通过语义补全网络进行弱监督视频时刻检索。载于*AAAI*人工智能会议论文集，8803–8810。AAAI Press。Liu, Z.; Qu, X.; Zhou, P.; Xu, Z.; and Cheng, Y. 2024年。利用大语言模型进行上下文增强视频时刻检索。*arXiv*预印本 *arXiv:2405.12540*。Luo, D.; Huang, J.; Gong, S.; Jin, H.; and Liu, Y. 2024年。基于冻结视觉-语言模型的零样本视频时刻检索。载于*IEEE/CVF*冬季应用计算机视觉会议论文集，5464–5473。Ma, K.; Zang, X.; Feng, Z.; Fang, H.; Ban, C.; Wei, Y.; He, Z.; Li, Y.; and Sun, H. 2023年。LLaViLo：通过基于适配器的多模态建模增强视频时刻检索。*2023 IEEE/CVF*国际计算机视觉会议研讨会（*ICCVW*），2790–2795。Maaz, M.; Rasheed, H.; Khan, S. H.; and Khan, F. 2023年。Video-ChatGPT：借助大型视觉语言模型实现详细视频理解。载于计算语言学协会年会。Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023年。查询依赖的视频表示：用于时刻检索和高光检测。*2023 IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），23023–23033。Nam, J.; Ahn, D.; Kang, D.; Ha, S. J.; and Choi, J. 2021年。零样本自然语言视频定位。载于*IEEE/CVF*国际计算机视觉会议论文集，1470–1479。

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021年. 从自然语言监督中学习可迁移视觉模型. 收录于机器学习国际会议. Reimers, N.; and Gurevych, I. 2019年. Sentence-BERT：使用孪生BERT网络生成句子嵌入. 收录于自然语言处理经验方法会议. Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2024年. Timechat：面向长视频理解的时间敏感多模态大语言模型. 收录于*IEEE/CVF*计算机视觉与模式识别会议论文集，14313–14323. Song, Z.; Liu, D.; Qu, X.; Cheng, Y.; and Zhou, P. 2021年. 通过对比学习实现细粒度视频时序检索. 收录于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，15804–15813. IEEE. Wang, G.; Wu, X.; Liu, Z.; and Yan, J. 2022年. 基于提示的零样本视频片段检索. 收录于*ACM*多媒体会议. Wattasseril, J. I.; Shekhar, S.; D¨ollner, J.; and Trapp, M. 2023年. 使用基于BLIP的模型进行零样本视频片段检索. 收录于国际视觉计算研讨会，160–171. Springer. Xu, Y.; Sun, Y.; Zhai, B.; Li, M.; Liang, W.; Li, Y.; and Du, S. 2025年. 通过现成的多模态大语言模型进行零样本视频片段检索. 收录于*AAAI*人工智能会议论文集，第39卷，8978–8986. Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020年. 用于视频定位的密集回归网络. 收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，10287–10296. IEEE. Zhang, H.; Li, X.; and Bing, L. 2023年. Video-LLaMA：用于视频理解的指令调优音视语言模型. 收录于自然语言处理经验方法会议. Zheng, M.; Huang, Y.; Chen, Q.; and Liu, Y. 2022年. 通过对比式负样本挖掘进行弱监督视频片段定位. 收录于*AAAI*人工智能会议论文集，第36卷，3517–3525。