GranAlign: Granularity-Aware Alignment Framework for Zero-Shot Video Moment Retrieval

Mingyu Jeon[1], **Sunjae Yoon**[2], **Jonghee Kim**[3], **Junyeong Kim**[1]
[1]Department of Artificial Intelligence, Chung-Ang University
[2] Korea Advanced Institute of Science and Technology (KAIST)
[3]Electronics and Telecommunications Research Institute (ETRI)
{smart2557, junyeongkim}@cau.ac.kr, sunjae.yoon@kaist.ac.kr, jhkim27@etri.re.kr

Abstract

Zero-shot video moment retrieval (ZVMR) is the task of localizing a temporal moment within an untrimmed video using a natural language query without relying on task-specific training data. The primary challenge in this setting lies in the mismatch in semantic granularity between textual queries and visual content. Previous studies in ZVMR have attempted to achieve alignment by leveraging high-quality pre-trained knowledge that represents video and language in a joint space. However, these approaches failed to balance the semantic granularity between the pre-trained knowledge provided by each modality for a given scene. As a result, despite the high quality of each modality's representations, the mismatch in granularity led to inaccurate retrieval. In this paper, we propose a training-free framework, called ***Gran**ularity-**A**ware **Align**ment (GranAlign)*, that bridges this gap between coarse and fine semantic representations. Our approach introduces two complementary techniques: granularity-based query rewriting to generate varied semantic granularities, and query-aware caption generation to embed query intent into video content. By pairing multi-level queries with both query-agnostic and query-aware captions, we effectively resolve semantic mismatches. As a result, our method sets a new state-of-the-art across all three major benchmarks (QVHighlights, Charades-STA, ActivityNet-Captions), with a notable 3.23% mAP@avg improvement on the QVHighlights dataset.

1 Introduction

Video Moment Retrieval (VMR)–the task of localizing a segment from a video via a language query–is crucial for efficient video understanding. However, traditional supervised approaches are fundamentally limited by their reliance on large, costly annotated datasets. To overcome this dependency, zero-shot VMR (ZVMR) has emerged as a powerful and successful paradigm, with its feasibility greatly enhanced by recent advances in Vision-Language Models (VLMs) and Large Language Models (LLMs).

Despite its success, the ZVMR paradigm faces a new, fundamental challenge: the 'Granularity Mismatch' between the language query and the visual content. This issue arises because a user might describe the same event at varying levels of detail, from a general phrase like "a cute dog" to a specific one like "a baby Golden Retriever is walking around."

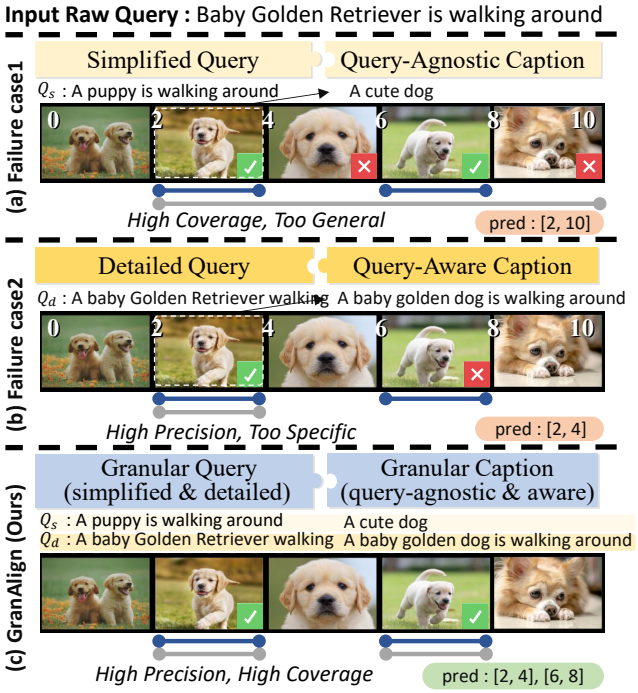

Figure 1: An illustration of how GranAlign resolves the 'Granularity Mismatch'. The (a), (b) fail to localize all moments due to a mismatch in semantic granularity, resulting in low precision and low coverage, respectively. In contrast, our proposed (c) integrates both granular levels, achieving both high precision and high coverage to correctly localize all target moments. **Takeaway:** GranAlign overcomes the core 'Granularity Mismatch' by synergizing the high-recall simple path with the high-precision detailed path.

As illustrated in Figure 1, this mismatch creates an inevitable trade-off: a general query may achieve high coverage but lacks the precision to pinpoint the exact moment, while a specific query offers high precision but often suffers from poor coverage if subtle details do not perfectly align. This limitation is not merely anecdotal. Our quantitative analysis in Figure 2, where performance is broken down by query type (Error, Simple, Detail, and Else), reveals a significant performance gap. This highlights the inability of current ap-

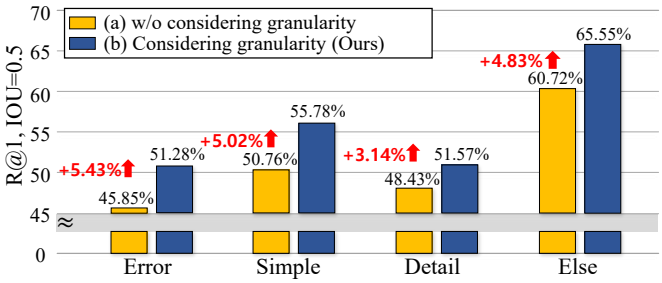

Figure 2: Comparison of our method (b) against a baseline (a) (without considering the granularity of query and video content) on the QVHighlights validation set, categorized by query type (see Section 4.3 for categorization criteria).

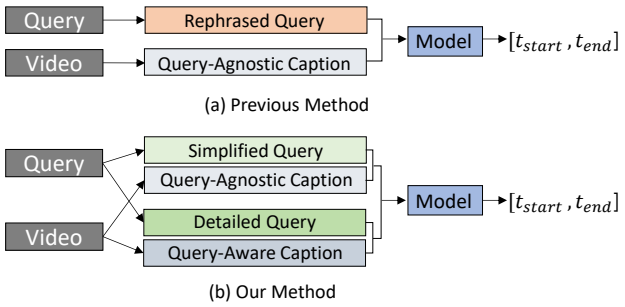

Figure 3: Conceptual Framework Comparison. (a) Previous methods typically adopt a single-path approach, reformulating the query. (b) Our method, GranAlign, employs a dual-path framework, decomposing the query into Simplified and Detailed versions and aligning them with Query-Agnostic and Query-Aware captions, respectively.

proaches to effectively handle varying levels of granularity. Prevailing methods, even when armed with high-quality pretrained knowledge in a joint space, fundamentally fail to resolve this trade-off because they lack an explicit mechanism for granularity-aware alignment.

The root cause of this problem is that existing methods treat queries monolithically. An intuitive solution might be to expand the original query with various rephrases, a strategy employed by several prior works as depicted in Figure 3 (a). However, this "one-size-fits-all" approach is inherently flawed. Even a diverse set of rewritten queries, if confined to a single level of granularity, cannot simultaneously embody the broad scope needed for high recall and the fine-grained detail required for high precision. This approach typically results in a compromise that excels at neither, failing to adapt to the diverse semantic complexity across different query-video pairs. This single-pathway reasoning is the core bottleneck preventing robust and accurate retrieval.

To address these limitations, we introduce Granularity-Aware Alignment (GranAlign), a novel, training-free framework that models and aligns queries and video content at complementary levels of granularity. As conceptually illustrated in Figure 3 (b), GranAlign abandons the single-pathway design. Instead, on the query side, it leverages an LLM (Grattafiori et al. 2024) to reformulate the original query into two distinct paths: a simplified query capturing the core intent for broad coverage, and a detailed query preserving specific nuances for precision. Correspondingly, on the video side, a VLM (Bai et al. 2025) generates a general, query-agnostic caption and a focused, query-aware caption. By aligning these pairs by granularity–simplified query with query-agnostic caption and detailed query with query-aware caption–GranAlign synergizes the high-recall capability of the general path with the high-precision of the specific path, leading to a more robust alignment and superior retrieval accuracy across multiple challenging benchmarks.

2 Related Works

2.1 Zero-Shot Video Moment Retrieval

Zero-shot Video Moment Retrieval (ZVMR) aims to localize relevant moments in a video using pretrained models without task-specific training. Prior work has achieved strong performance on the QVHighlights dataset using off-the-shelf vision-language models and BLIP-2-based (Li et al. 2023a) approaches (Diwan 2023; Wattasseril et al. 2023). However, these methods often rely on coarse frame-level matching or restricted segment-based proposals, limiting their ability to achieve fine-grained temporal alignment and precise semantic grounding. Recent advances such as Moment-GPT (Xu et al. 2025) leverage large language models (LLMs) to rephrase queries and use Video-ChatGPT (Maaz et al. 2023) to score candidate moments, achieving state-of-the-art performance. Despite its advanced architecture, Moment-GPT still struggles with accurate semantic alignment between the rewritten queries and visual content, which hinders precise moment localization. In the next subsection, we analyze this alignment issue through the lens of semantic granularity.

2.2 Granularity in Retrieval

Fine-grained semantic granularity in both *visual* and *linguistic* streams is essential for accurate video grounding. Early approaches relied on coarse frame-level cues or single-sentence queries (Hendricks et al. 2017; Gao et al. 2017a; Yoon et al. 2023), while later works introduced denser visual segments (Zeng et al. 2020; Song et al. 2021) and more elaborate query reasoning (Lin et al. 2019; Yoon et al. 2022; Ryu, Kim, and Kim 2024). However, existing methods still align entire queries with generic captions or raw visual features, without explicitly managing semantic granularity.

Recent studies continue to exhibit this limitation. Diwan et al. (Diwan 2023) improved ZVMR by 2.5× using off-the-shelf captions, yet these remain query-agnostic. Context-Enhanced VMR (Liu et al. 2024) incorporates BLIP-2 summaries, but lacks alignment with query intent. Moment-GPT (Xu et al. 2025) reformulates queries via LLaMA-3 and scores spans using frozen MLLMs, but still treats each query as a single undifferentiated unit and relies on query-agnostic captions, leading to potential semantic mismatches. Thus, recent methods fall short in achieving robust moment retrieval. Our work, GranAlign, addresses this persistent gap by explicitly modeling and aligning content at complemen-

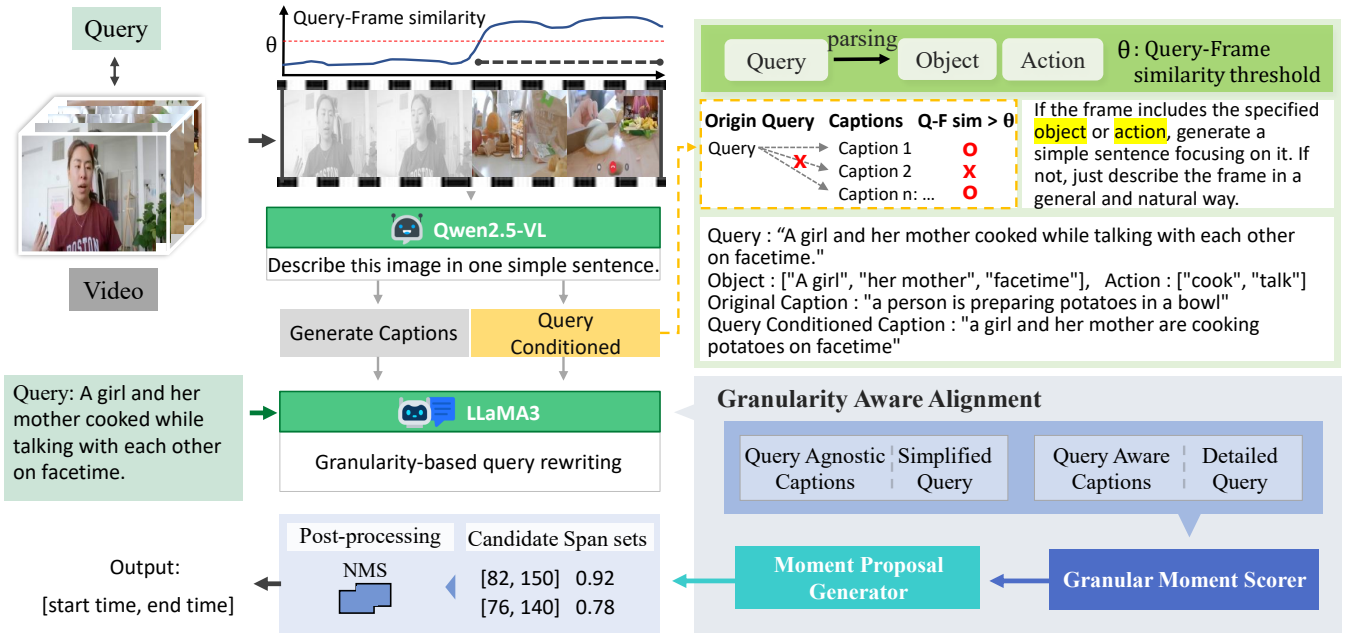

Figure 4: Overview of the GranAlign Framework. In Granularity-Aware Alignment (Sec.3.2), the input query is rewritten at two semantic granularities (simplified/detailed) and matched with either query-agnostic or query-aware captions (captions generated with the query as context) to obtain a *Moment Score* for each video segment. These scores drive the Moment Proposal Generator and NMS stage, after which the post-ranking module produces the final prediction (Sec.3.3). **Takeaway:** GranAlign effectively tackles the ZVMR task by leveraging fine-grained semantic alignment between queries and video content.

tary levels of granularity.

3 Method

3.1 Overview

Given an untrimmed video $V = \{v_i\}_{i=1}^{L_v}$ containing L_v frames, and a textual query $Q = \{q_i\}_{i=1}^{L_q}$ comprising L_q words, the goal of video moment retrieval (VMR) is to predict a set of temporal spans $T = \{t_s, t_e\} \in \mathbb{R}^{N_t \times 2}$. Here, t_s, t_e denote the start and end times of a predicted segment, and N_t is the number of predicted spans.

Our framework consists of three stages. First, to enable our Granularity-Aware Alignment, the input query is reformulated into simplified and detailed versions, and two corresponding caption sets–query-agnostic and query-aware–are generated (Sec.3.2). These are then used to compute a *Granular Moment Score* for each video segment (Sec.3.3). Second, candidate temporal segments (*moment proposals*) are generated based on these scores. Finally, a post-processing step including Non-Maximum Suppression (NMS) selects the most relevant segments.

3.2 Granularity-Aware Alignment

This section introduces the core idea of our approach, *Granularity-Aware Alignment*, which consists of three main components: granularity-based query rewriting, query-aware caption generation, and moment score computation.

Granularity-based query rewriting To achieve more robust semantic alignment, we reformulate the input query Q into semantically complementary representations via instruction-guided rewriting with LLaMA-3 (Grattafiori et al. 2024). Defining "simple" and "detail" with a single, rigid mathematical formula is challenging, as such definitions fail to capture the nuances of natural language. Therefore, rather than relying on a single prompt, we utilize multiple, manually crafted instruction pairs to generate the simplified query (Q_s) and the detailed query (Q_d). Figure 5 showcases representative examples from our instruction set.[1] We empirically found that our framework's performance is robust and not overly sensitive to the specific instruction pair used. The simplified query Q_s is derived using a prompt that instructs the model to "replace rare words with common alternatives" while "keeping the core entities and actions." To improve clarity, rare or incidental details are removed, and grammatical errors are corrected. This form provides high generality and effectively retrieves broadly relevant candidate segments. In contrast, the detailed query Q_d preserves fine-grained expressions, temporal context, and specific lexical choices. These elements contribute to more precise alignment and accurate moment localization.

Query-Aware Captioning Conventional video captioning methods generate descriptions from visuals alone, limiting their ability to accurately identify moments semantically aligned with a query. Even a visually correct caption may

[1]The complete set of instruction pairs used in our experiments, along with a sensitivity analysis of their impact on performance, is available in Appendix E.

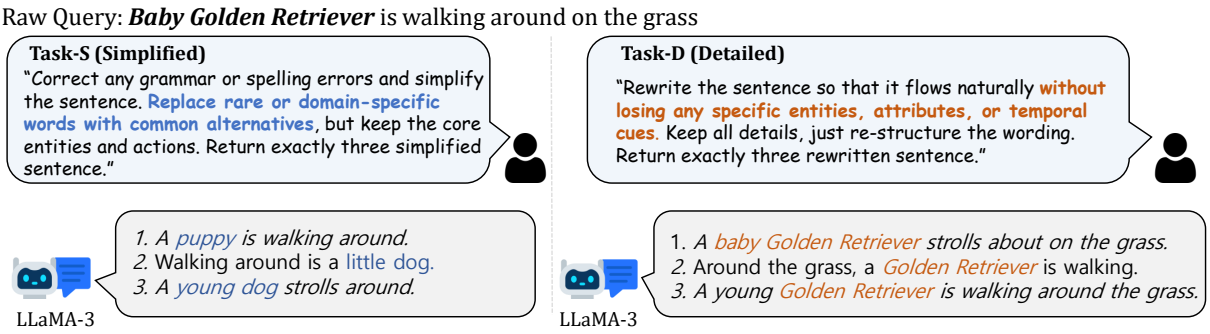

Figure 5: Examples of instruction pairs for granularity-based query rewriting. This figure illustrates sample prompts given to LLaMA-3 to generate simplified (recall-focused) and detailed (precision-focused) queries.

lack the query intent needed for accurate retrieval.

To overcome this limitation, we aim to directly embed the query's intent into captions. However, applying this process to every frame is computationally prohibitive. We therefore propose a hybrid strategy that balances semantic precision with computational feasibility. This strategy first identifies a set of candidate frames by selecting the top-K% with the highest similarity scores between the query Q and each frame. Let L_k denote the number of these candidate frames. Then, using Qwen2.5-VL (Bai et al. 2025), it generates two types of captions. First, a general, *query-agnostic caption* ($C_{agn} \in \mathbb{R}^{L_v \times l}$) is created as a baseline for all L_v frames. Second, a focused, *query-aware caption* ($C_{awr} \in \mathbb{R}^{L_k \times l}$) is generated only for the L_k candidate frames.

To create C_{awr}, we guide the generation process using key semantic elements extracted from the query. For instance, from "A person picking up a pencil from the desk," we extract the entities {*person, pencil*} and the action {*picking up*} as semantic guidance. This hybrid strategy enables our framework to leverage the semantic precision of query-aware captions on the most critical regions while maintaining overall computational efficiency.

This process enables C_{awr} to achieve stronger semantic alignment with the query compared to query-agnostic captions, which in turn improves the accuracy of moment localization. However, this approach can lead to failures, such as hallucinating visual content not present in the video or overly mimicking the query's linguistic structure. Therefore, our final moment scoring process is designed to leverage the advantages of C_{awr} while mitigating these potential failures.

Granular Moment scoring The two query-caption pairs, simplified-agnostic (Q_s, C_{agn}) and detailed-aware (Q_d, C_{awr}), exhibit distinct characteristics with complementary strengths and limitations, as illustrated in Fig 6. The simplified query Q_s and generic caption C_{agn} offer broad coverage of semantically similar scenes, yielding higher recall but often lacking the precision to localize the exact moment. In contrast, the detailed query Q_d and query-aware caption C_{awr} enable more accurate alignment via fine-grained semantics, but are more prone to errors like hallucination or misalignment from over-relying on the query. This highlights the importance of matching granularity levels for successful alignment, which is the core principle of our approach.

To leverage the complementary benefits of both pairs, we compute a composite *moment score* by integrating the semantic similarities of (Q_s, C_{agn}) and (Q_d, C_{awr}). Specifically, we evaluate each candidate moment's semantic alignment using both query-caption pairs and take the average as the final score. The combined frame-level similarity score S_f is defined as:

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

The similarity score S_f is computed for each frame f in the video by averaging the semantic similarity between rephrased queries and their corresponding captions. Each rephrased query pair consists of a simplified query $q_s^{(i)}$ and a detailed query $q_d^{(i)}$, both derived from the original input query through controlled rewriting, where m denotes the total number of such pairs. $C_{agn,f}$ and $C_{awr,f}$ represent the caption embeddings for frame f in the query-agnostic and query-aware caption sets, respectively. Here, $\mathrm{g}(\cdot,\cdot)$ denotes the normalized cosine similarity between the sentence embeddings of the query and the caption.

This formulation mitigates potential biases or false positives that may arise from relying on a single query-caption pair, enabling more robust and reliable moment retrieval. We compute the similarity by first embedding queries and captions with a SentenceTransformer model (Reimers and Gurevych 2019), and then calculating their cosine similarity. The resulting frame-level score S_f is then used to generate and rank candidate segments.

3.3 Moment Proposal Generation

In this stage, the computed *Moment Scores* are used to generate candidate segments via a *Moment Proposal Generator*, followed by a *Post Processing* step to refine the final output.

Moment proposal generator The Moment Proposal Generator (MPG) creates semantically coherent candidate spans from frame-level similarity scores S_f. To prevent a single continuous event from being fragmented into multiple short

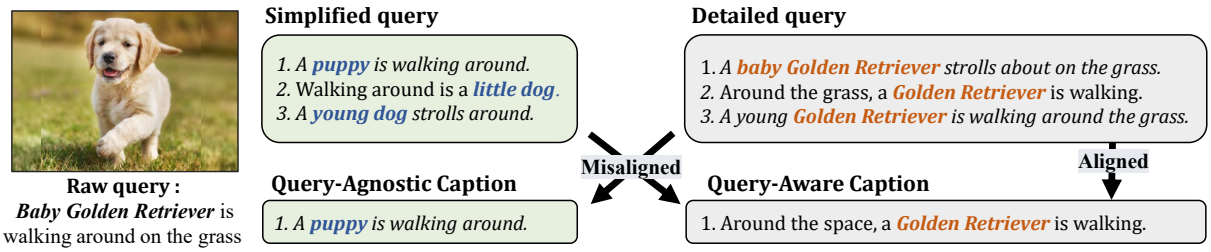

Figure 6: Illustration of the semantic alignment patterns at different granularity levels. A simplified query (e.g., "a puppy") aligns well with a query-agnostic caption, whereas a detailed query (e.g., "Golden Retriever") requires a query-aware caption. Mismatches in granularity can lead to failed alignments.

clips, MPG merges adjacent high-scoring frames into a single proposal if the temporal gap between them is within a threshold of τ frames. While merging frames based on the τ ensures continuity, it can create diluted proposals where high-scoring keyframes are connected by a series of low-relevance frames. To mitigate this issue, spans with an average similarity in the bottom n% are discarded to maintain relevance. This acts as a crucial quality control step, ensuring that only proposals with a consistently high level of semantic relevance proceed to the next stage.

We score each candidate span p based on its semantic relevance and a length regularization. The first term, μ_p, is the average semantic similarity of all frames within the span. The second term, ρ_p, is the length of span p normalized by the total length of all candidate spans. This normalized length ρ_p serves as a crucial regularization factor to penalize overly long proposals that might have low average relevance and to prevent a bias towards trivial, short segments, ensuring that the final candidates are of a meaningful duration. The final score for each candidate span p is computed as:

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

The parameter λ (set to 0.3) balances this trade-off. This regularized scoring provides more robust candidates than simpler threshold- or length-based approaches.

Post processing All candidate spans generated by the *Moment Proposal Generator* are collected into a set, with each span defined by its start and end frame indices. Any two spans p_i and p_j in this set represent distinct candidate intervals. To refine the set of candidate spans, we apply *Non-Maximum Suppression* (NMS), which iteratively discards the lower-scoring of any two spans whose Intersection-over-Union (IoU) exceeds a threshold θ_{NMS}, yielding the final moment prediction $\{t_s, t_e\}$.

4 Experiments

This section details the experiments conducted to validate the effectiveness of the proposed **GranAlign** framework. We present the experimental setup, a performance comparison against state-of-the-art methods, an ablation study analyzing the contributions of our core components, and a further analysis to demonstrate the robustness of our method.

Datasets We evaluate on public benchmarks: *QVHighlights* (Lei, Berg, and Bansal 2021), *Charades-STA* (Gao et al. 2017b), and *ActivityNet-Captions* (Krishna et al. 2017).

Evaluation Metrics Statistics for the datasets and detailed definitions of the metrics are provided in Appendix D.

4.1 Comparison with the State-of-the-Art

Table 1 shows that GranAlign consistently surpasses the previous zero-shot SOTA (Xu et al. 2025) on QVHighlights. On the validation split, it improves all metrics by **+3.04% to +3.93%**, with the largest gain in mAP@0.5; on the hidden test set, it still leads by **+1.6% to +3.84%**. These gains confirm that granularity-based query rewriting and query-aware captioning enhance semantic alignment and, consequently, retrieval accuracy. Table 2 shows strong transfer to Charades-STA and ActivityNet-Captions. On Charades-STA, GranAlign exceeds the previous SOTA by +1.2% (R1@0.5) and +1.5% (mIoU); on ActivityNet-Captions, it gains +2.9% (R1@0.5) and +2.3% (mIoU). Thus, the proposed alignment strategy scales from short to long, diverse videos, providing robust zero-shot moment retrieval across benchmarks.

4.2 Ablation Studies

Notation In the following ablation experiments, we revisit the notation for clarity: Q_r denotes the raw query, Q_s, the simplified query; Q_d, the detailed query; C_{agn}, the query-agnostic caption; and C_{awr}, the query-aware caption.

Impact of Query-Aware Caption As shown in Table 3, using a query-aware caption (C_{awr}) with the raw query yields consistent improvements over a query-agnostic one (C_{agn}) across all metrics. This suggests that tailoring captions to the query content enhances semantic alignment.

Effectiveness of Granularity-Aligned Pairing We analyze the effects of granularity alignment in Table 3. Performance declines when query and caption granularities are mismatched (e.g., pairing Q_s with C_{awr}), whereas aligned pairs (e.g., pairing Q_d with C_{awr}) perform better.

However, these single aligned pairs present a trade-off: the simplified pair (Q_s, C_{agn}) offers high recall but low precision, while the detailed pair (Q_d, C_{awr}), despite achieving the best performance among single pairs, provides high precision but can fail on subtle mismatches. Our GranAlign framework resolves this by combining both pairs, leveraging their complementary strengths. As a result, GranAlign achieves the best overall performance with consistent gains

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

Table 1: Performance comparison on QVHighlights test and val sets. "MLLM" indicates the use of a multimodal large language model, and "FS/WS/US/ZS" denote fully-supervised, weakly-supervised, unsupervised, and zero-shot settings.

Method	MLLM	Setting	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 et al. 2024)	✓	FS	–	29.6	11.9	–	–	–	–	–
VTimeLLM (Huang et al. 2024)	✓	FS	55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4
TimeChat (Ren et al. 2024)	✓	FS	–	43.8	22.7	–	–	–	–	–
Moment-DETR (Lei, Berg, and Bansal 2021)	–	FS	62.1	48.2	25.3	42.3	52.6	32.5	15.3	37.8
CPL (He et al. 2022)	–	WS	56.0	38.1	20.3	37.8	52.4	30.9	12.0	32.6
Huang et al. (Huang, Yang, and Sato 2023)	–	WS	59.2	44.2	22.1	39.4	54.8	32.9	–	36.4
PSVL (Nam et al. 2021)	–	US	45.2	30.9	14.2	30.9	45.1	29.8	15.73	30.2
Liu et al. (Zheng et al. 2022)	–	US	44.2	28.7	14.7	–	47.3	28.2	–	–
TimeChat (Ren et al. 2024)	✓	ZS	–	32.2	13.4	–	–	–	–	–
Luo et al. (Luo et al. 2024)	✓	ZS	53.4	36.0	19.3	34.1	45.6	27.4	12.3	28.4
Moment-GPT (Xu et al. 2025)	✓	ZS	58.2	38.4	21.6	36.5	48.1	31.1	14.9	30.8
GranAlign (Ours)	✓	ZS	**59.1**	**39.6**	**22.7**	**38.0**	**50.3**	**34.0**	**16.5**	**33.1**

Table 2: Performance comparison on Charades-STA and ActivityNet-Captions under various supervision settings.

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

Table 3: Consolidated ablation study results. We denote queries as Q and captions as C, with subscripts for raw (r), simplified (s), detailed (d), query-agnostic (agn), and query-aware (awr). The (Q_s, Q_d) row represents our full model.

across all metrics. We also confirm this approach's robustness with a BLIP-2 backbone (please see Appendix A).

Implementation Details We implemntd our framework using LLaMA3-8B (Grattafiori et al. 2024) for query rewriting and Qwen2.5-VL-7B (Bai et al. 2025) for caption generation, with initial frames filtered by CLIP ViT-B/32 (Radford et al. 2021). All experiments were run on four NVIDIA A6000 GPUs, and a comprehensive list of implementation details is available in Appendix B.

4.3 Further Analysis

Robustness to Query Variation As shown in Fig 7, we evaluated GranAlign's robustness across four linguistically diverse query types (Error, Simplified, Detailed, and Else).[2] The baseline combinations, (a) (Q_s, C_{agn}) and (b) (Q_d, C_{awr}), exhibit complementary but limited behaviors. The former provides broad recall but lacks precision, while the latter captures fine-grained semantics yet suffers from overfitting to the query, leading to hallucinations or misalignment. In contrast, our proposed GranAlign integrates both simplified and detailed query reformulations with query-aware captioning, achieving robust performance across all categories by effectively balancing recall and precision.

Sensitivity Analysis As shown in Figure 8, our novel framework, GranAlign, is robust to variant hyperparameter settings. For example, performance peaks when using

[2]Queries were classified by length: Simple ($\leq$6 tokens), Detail ($\geq$20 tokens or containing a proper noun), or Else (all others). Independently, queries with grammatical errors were flagged as Error.

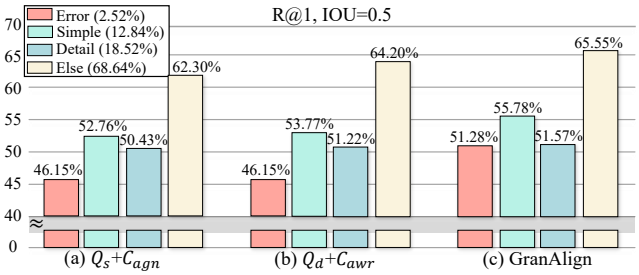

Figure 7: Analysis of performance across simplified, detailed, error, and other query types on QVHighlights val. The Y-axis is truncated below 40 for clarity.

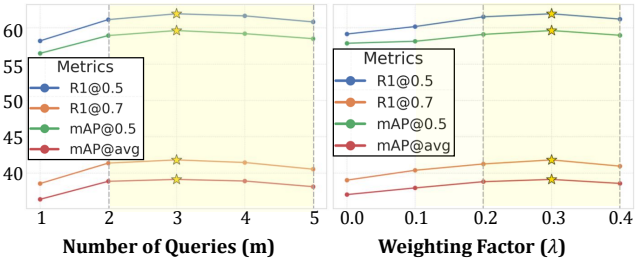

Figure 8: Analysis of performance across variant hyperparameters. The yellow shaded regions highlight the parameter range where our framework demonstrates robust and stable performance around the optimal settings (marked by stars).

3 rewritten queries (m=3) and a weighting factor of 0.3 (λ=0.3), while remaining stable across a range of nearby values. Our full experiments on hyperparameter selection are detailed in Appendix C.

Efficiency Analysis In Table 4, we compare the training cost, GPU memory usage, and inference time against fully supervised and zero-shot methods. As a zero-shot method, GranAlign incurs no training cost and, for efficiency, pre-generates query-agnostic captions offline.

Subsequently, during the online process, it selectively generates additional captions only for key frames that exceed a certain threshold. This two-stage approach contributes to reducing inference time and helps the model maintain a competitive memory footprint. As a result, GranAlign achieves the best performance, outperforming other configurations on most key metrics such as R1@0.5 and mAP while also recording the shortest inference time.

Qualitative Results Fig 9 demonstrates how GranAlign utilizes semantic granularity for accurate zero-shot localization. Given the query "An Asian girl wearing a white face mask with a heart...", the ground truth (GT) spans 90–142s. The Simple path, relying on coarse alignment, fails to distinguish specific attributes and over-extends the prediction to start at 78s, including irrelevant content. Conversely, the Detail path identifies fine-grained cues like "heart" for a precise onset but under-segments the end (missing frames after 138s). By integrating both scores, GranAlign successfully retrieves the exact GT (90–142s). This validates that combining complementary semantics is essential to resolve the trade-off between coverage and precision.

Setup	Methods	R1@0.5	mAP	TrC	InT (s)	GMU (G)
ZS	VideoChatGPT	22.4	10.3	0	9.8	11
FS	VTimeLLM	48.8	26.8	4090 40h	11.2	18
ZS	Wattersseril	53.1	30.2	0	12.7	14
ZS	Moment-GPT	58.9	35.9	0	16.1	32
ZS	Ours	**61.94**	**39.12**	**0**	**6.2**	**22**

Table 4: Comparison of training cost (TrC), GPU memory usage (GMU), and inference time (InT) on the QVHighlights dataset.

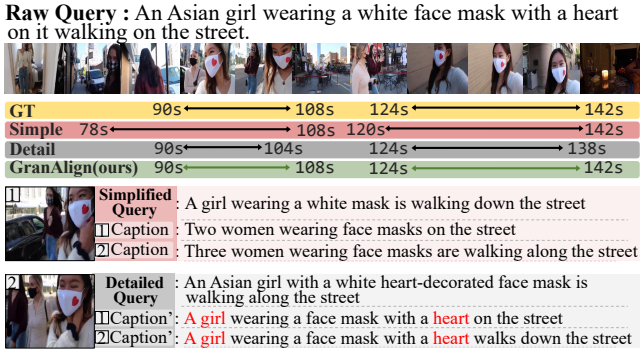

Figure 9: Qualitative results on QVHighlights. Caption' in Detailed Query denotes query-aware caption.

5 Discussions and Limitations

While GranAlign achieves strong performance, its limitations stem from the generative nature of its core components, presenting clear avenues for future research. For instance, the risk of hallucination in query-aware captioning could be mitigated by a fact-checking mechanism that verifies generated details against the visual evidence. Similarly, a semantic verification step could ensure that LLM-based query rewriting preserves the original intent. Pursuing these enhancements would further improve the reliability and generalizability of granularity-aware moment retrieval systems.

6 Conclusion

We proposed Granularity-Aware Alignment (GranAlign), a novel, training-free framework for ZVMR. By aligning queries and video content at multiple levels of semantic granularity, GranAlign resolves the trade-off between coverage and precision, balancing the high recall of general representations with the high precision of detailed attributes. As a result, GranAlign establishes a SOTA across major benchmarks without task-specific training cost, presenting new possibilities for zero-shot VMR. We believe our granularity-aware alignment strategy offers a promising and effective direction for future multimodal understanding systems.

Acknowledgements

This work was supported by IITP grant funded by the Korea government(MSIT) (No. RS-2020-II200004, Develop-

ment of Previsional Intelligence based on Long-term Visual Memory Network) and supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program(Chung-Ang University)].

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Diwan, A. 2023. Zero-shot video moment retrieval with offtheshelf models. In *Transfer Learning for Natural Language Processing Workshop*, 10–21.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017a. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5267–5275. IEEE.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017b. TALL: Temporal Activity Localization via Language Query. *2017 IEEE International Conference on Computer Vision (ICCV)*, 5277–5285.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

He, X.; Yang, D.; Feng, W.; Fu, T.-J.; Akula, A. R.; Jampani, V.; Narayana, P.; Basu, S.; Wang, W. Y.; and Wang, X. 2022. CPL: Counterfactual Prompt Learning for Vision and Language Models. In *Conference on Empirical Methods in Natural Language Processing*.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing Moments in Video with Natural Language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5803–5812. IEEE.

Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14271–14280.

Huang, Y.; Yang, L.; and Sato, Y. 2023. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18908–18918.

Kong, S.; Li, L.; Zhang, B.; Wang, W.; Jiang, B.; Yan, C.; and Xu, C. 2023. Dynamic contrastive learning with pseudo-samples intervention for weakly supervised joint video mr and hd. In *Proceedings of the 31st ACM International Conference on Multimedia*, 538–546.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023b. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36: 65948–65966.

Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Vu, V. T.; et al. 2024. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*.

Lin, K.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A.; Yan, R.; and Shou, M. Z. 2023. UniVTG: Towards Unified Video-Language Temporal Grounding. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2782–2792.

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2019. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8803–8810. AAAI Press.

Liu, Z.; Qu, X.; Zhou, P.; Xu, Z.; and Cheng, Y. 2024. Context-Enhanced Video Moment Retrieval with Large Language Models. *arXiv preprint arXiv:2405.12540*.

Luo, D.; Huang, J.; Gong, S.; Jin, H.; and Liu, Y. 2024. Zero-shot video moment retrieval from frozen vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5464–5473.

Ma, K.; Zang, X.; Feng, Z.; Fang, H.; Ban, C.; Wei, Y.; He, Z.; Li, Y.; and Sun, H. 2023. LLaViLo: Boosting Video Moment Retrieval via Adapter-Based Multimodal Modeling. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2790–2795.

Maaz, M.; Rasheed, H.; Khan, S. H.; and Khan, F. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Annual Meeting of the Association for Computational Linguistics*.

Nam, J.; Ahn, D.; Kang, D.; Ha, S. J.; and Choi, J. 2021. Zero-shot natural language video localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 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. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14313–14323.

Ryu, J.; Kim, J.; and Kim, J. 2024. A Study on the Representativeness Heuristics Problem in Large Language Models. *IEEE Access*, 12: 147958–147966.

Song, Z.; Liu, D.; Qu, X.; Cheng, Y.; and Zhou, P. 2021. Fine-Grained Video Temporal Retrieval With Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15804–15813. IEEE.

Wang, G.; Wu, X.; Liu, Z.; and Yan, J. 2022. Prompt-based Zero-shot Video Moment Retrieval. In *ACM Multimedia*.

Wattasseril, J. I.; Shekhar, S.; Döllner, J.; and Trapp, M. 2023. Zero-Shot Video Moment Retrieval Using BLIP-Based Models. In *International Symposium on Visual Computing*, 160–171. Springer.

Xu, Y.; Sun, Y.; Zhai, B.; Li, M.; Liang, W.; Li, Y.; and Du, S. 2025. Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8978–8986.

Yoon, S.; Hong, J. W.; Yoon, E.; Kim, D.; Kim, J.; Yoon, H. S.; and Yoo, C. D. 2022. *Selective Query-Guided Debiasing for Video Corpus Moment Retrieval*, 185–200. Springer Nature Switzerland. ISBN 9783031200595.

Yoon, S.; Koo, G.; Kim, D.; and Yoo, C. D. 2023. SCANet: Scene Complexity Aware Network for Weakly-Supervised Video Moment Retrieval. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13530–13540. IEEE.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense Regression Network for Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10287–10296. IEEE.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Conference on Empirical Methods in Natural Language Processing*.

Zheng, M.; Huang, Y.; Chen, Q.; and Liu, Y. 2022. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3517–3525.