This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

Context-Enhanced Video Moment Retrieval with Large Language Models

Weijia Liu, Bo Miao, Jiuxin Cao, Xuelin Zhu, Jiawei Ge, Bo Liu, Mehwish Nasim, and Ajmal Mian, *Senior Member, IEEE*

***Abstract*—Current methods for Video Moment Retrieval (VMR) struggle to align complex situations involving specific environmental details, character descriptions, and action narratives. To tackle this issue, we propose a Large Language Model-guided Moment Retrieval (LMR) approach that employs the extensive knowledge of Large Language Models (LLMs) to improve video context representation as well as cross-modal alignment, facilitating accurate localization of target moments. Specifically, LMR introduces a context enhancement technique with LLMs to generate crucial target-related context semantics. These semantics are integrated with visual features for producing discriminative video representations. Finally, a language-conditioned transformer is designed to decode free-form language queries, on the fly, using aligned video representations for moment retrieval. Extensive experiments demonstrate that LMR achieves state-of-the-art results, outperforming the nearest competitor by up to 3.28% and 4.06% on the challenging QVHighlights and Charades-STA benchmarks, respectively. More importantly, the performance gains are significantly higher for localization of complex queries.**

***Index Terms*—Video moment retrieval, multimodal alignment, context enhancement, large language models.**

I. Introduction

Video Moment Retrieval (VMR) [1]–[13] aims to localize the specific moment in an untrimmed video that semantically corresponds to a given language query. The success of VMR enhances video analysis with reduced effort and benefits a wide range of applications, including video semantic segmentation, user-friendly video editing, and mass surveillance. Unlike conventional temporal action localization, constrained by predefined action categories, VMR offers flexibility by localizing complex activities through free-form linguistic expressions.

This work was supported in part by National Natural Science Foundation of China under Grants No.62172089,No.62172090,No.62106045, in part by Natural Science Foundation of Jiangsu province under Grants No.BK20191258, in part by Australian Research Council Future Fellowship Award under Grant FT210100268, in part by Jiangsu Provincial Key Laboratory of Computer Networking Technology, in part by Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No.BM2003201, in part by Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grant No.93K-9, and in part by Nanjing Purple Mountain Laboratory. (Corresponding author: Jiuxin.Cao)

W. Liu, J. Cao, X. Zhu and J. Ge are with School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China. (e-mail: weijia-liu@seu.edu.cn, jx.cao@seu.edu.cn, zhuxuelin@seu.edu.cn, jiawei_ge@seu.edu.cn)

B. Miao, M. Nasim and A. Mian are with School of Computer Science and Software Engineering, University of Western Australia, Perth 6009, Australia. (e-mail: bo.miao@research.uwa.edu.au, mehwish.nasim@uwa.edu.au, ajmal.mian@uwa.edu.au)

B. Liu is with School of Computer Science and Engineering, Southeast University, Nanjing 211189, China. (e-mail: bliu@seu.edu.cn)

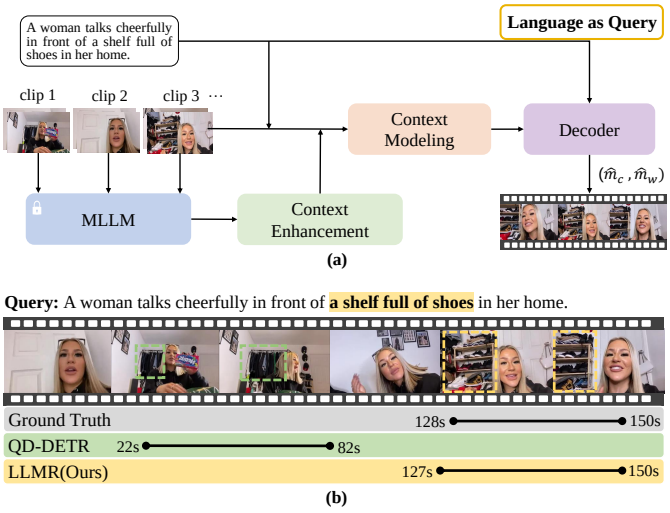


Fig. 1. (a) We introduce LLM-based context enhancement and modeling to improve video representations and directly employ free-form language as queries to decode these representations for accurate moment localization. (b) The benchmark QD-DETR [14] predicts false positives due to insufficient context modeling and decoding (captures major context but misses important details), whereas, our approach accurately localizes the target moment.

Current VMR paradigms primarily include two-stage matching and coordinate regression. Former [15]–[19] pre-segments moment proposals using static sliding windows or proposal-generating networks, followed by repeated query-proposal matching to select the best candidate. The latter [14], [20]–[24] leverages transformers to directly regress the start and end coordinates based on multimodal video and query features. Despite promising results, existing methods struggle to distinguish visually or contextually similar moments, especially when queries describe detailed environments, characters, or actions. Two-stage matching [12], [19], [25] uses segment-level context interactions for proposal representation but fails to capture fine-grained clip-level semantic differences, making it difficult to differentiate similar segments. In coordinate regression, the transformer's self-attention mechanism helps identify event boundaries but often assigns high attention to visually similar clips, treating them as semantically identical, which confuses localization. As illustrated in Fig. 1(b), when multiple moments fit the action "A woman talks cheerfully" and the background "a shelf", the benchmark QD-DETR [14] fails to identify the correct moment for "A woman talks cheerfully in front of a shelf full of shoes".

Recently, Large Language Models (LLMs) [26]–[30] trained with unmatched Internet-scale visual-language data have ex-

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

hibited impressive capabilities in comprehending visual and textual inputs to promote various vision-language tasks [31]–[34]. For instance, contextual object detection models [33] employ the textual output produced by LLMs as prompts to manipulate external vision expert models. Zero-shot image classification models [32] use class descriptions generated by LLMs as guidance for classification. In VMR, the high-level semantic information contained in LLM-generated text can compensate for visual blind spots and enhance the contextual representation of videos, aiding in distinguishing visually similar events from a textual perspective. However, the potential of LLMs in this role has yet to be explored in VMR.

To this end, we propose Large Language Model-guided Moment Retrieval (LMR), an effective VMR approach that repurposes the inherent knowledge of LLMs to enhance video context modeling and adopts a language-conditioned transformer to decode free-form language queries, on the fly. Specifically, the proposed LMR generates multiple textual views for each video, including rich contextual details like environmental information, character descriptions, and action narratives. These crucial views are distilled into target-related context through Video-Query Fusion (VQF). They are then combined with visual features to improve the alignment between video content and complex language queries, resulting in highly discriminative video representations. Finally, we introduce a language-conditioned transformer to directly decode free-form language queries using the highly discriminative video representations for VMR. Furthermore, considering that queries with longer lengths and more clauses may indicates stricter contextual demands for the queried events, we construct a Complex Query Validation (C-QVal) split from the validation split of an existing dataset based on metrics clause and word counts, and then evaluate the effectiveness of LMR in video retrieval for complex queries.

Our main contributions are summarized as follows.

- We propose an LLM-guided Moment Retrieval (LMR) method that integrates visual content with target-related context information derived from LLMs for moment retrieval. We construct C-QVal dataset with complex queries and show the efficacy of LMR in handling complex scenarios.
- We devise a language-conditioned transformer to decode free-form language queries, on the fly, using aligned video representations for moment retrieval.
- Extensive experiments demonstrate that our approach achieves top-ranked performance on popular benchmarks, outperforming the nearest competitor by up to 3.28% and 4.06% on the challenging QVHighlights and Charades-STA, respectively.

II. Related Works

Video Moment Retrieval. Two popular VMR techniques include the two-stage matching and semantic decomposition paradigms. The key idea of two-stage matching methods [16]–[19], [35] is to extract a set of moment proposals through multi-scale sliding windows or proposal generating networks, and explicitly model channel-level or sequence-level interactions pairwise for proposal matching and ranking. Despite their promising performance, generating numerous proposals for different durations and locations to ensure high recall greatly increases computational overhead, negatively impacting the efficiency and difficulty of the subsequent matching process.

Semantic decomposition methods [25], [35]–[40] typically decompose video and language into structured hierarchies, such as word-phrase-sentence and event-actions-objects. Thus, the semantics of video and query text can be systematically represented by combining different units, and the semantic correlation between text and vision is then learned through interactions among multimodal semantic units. For instance, [39] uses verb-object pairs and activity concept classifiers to encode multimodal features. Such a decomposition-based paradigm is effective for basic scenarios but struggles to localize moments corresponding to complex situations, where the target moment is conditioned on diverse backgrounds, character appearances, and action sequences. For such intricate queries, thoroughly representing and understanding the spatiotemporal context in videos is imperative. However, existing methods struggle with capturing these detailed semantics. To address this, we properly employ the extensive knowledge within LLMs to construct target-related video context, thereby enhancing video modeling and representation to handle intricate queries.

Vision-Language Tasks with LLMs. External knowledge is commonly employed in VMR tasks to mitigate biases such as temporal location bias [41] and moment length bias [42]. For instance, [43] and [44] distill generalization knowledge from large-scale vision-language pre-trained models and transfer it to VMR networks. Similarly, [45] leverages domain-specific knowledge (either from text or video) to guide neural network training that bridges the two modalities.

While the above methods enhance VMR performance, the advent of LLMs has introduced a new paradigm for incorporating external knowledge into vision-language tasks. In recent years, LLMs have developed rapidly, demonstrating exceptional capabilities in text comprehension and generation. They have been successfully applied to various vision-language tasks, including referring image segmentation [46], [47], image classification [32], contextual object detection [33], and video question answering [48]. These applications leverage the informative textual output produced by LLMs, *e.g.*addpunct., ChatGPT[1], OPT [49] and LLaMA [50], as prompts to guide external vision expert models. For instance, contextual object detection models [33] employ the textual output produced by LLMs as prompts for multimodal context decoding. Zero-shot image classification models [32] use class descriptions generated by LLMs as guidance for classification. [48] utilizes BLIP-2 to generate scores for each frame indicating whether it contains relevant information to answer the question. Despite their success in these fields, the potential of LLMs in VMR has yet to be explored.

Additionally, the progress in LLMs has also spurred the advancement of multimodal large language models (MLLMs) [26]–[30]. These MLLMs, built on pretrained LLMs, are commonly fine-tuned with extra visual components, have expanded the capabilities of LLMs to comprehend both

[1]https://openai.com/blog/chatgpt

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

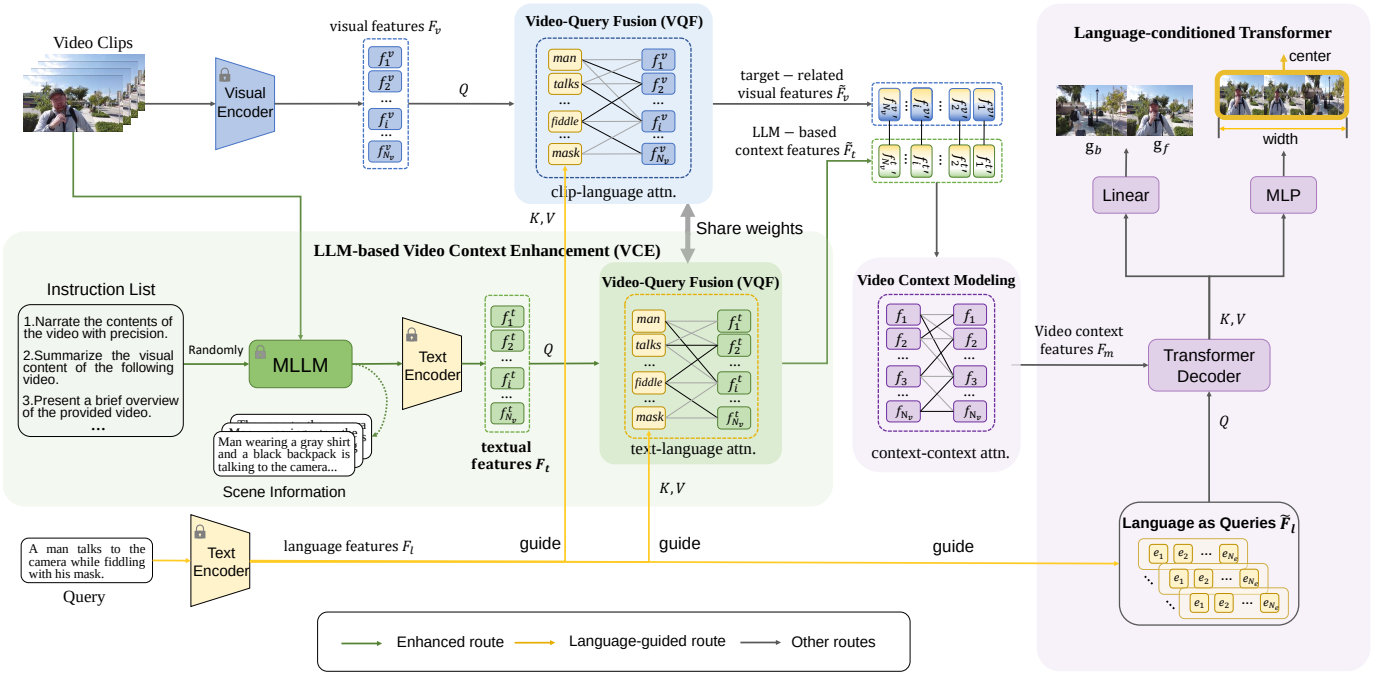


Fig. 2. Architecture of LMR. Given a video sequence and a query language, LLM-based Video Context Enhancement generates target-related context information, Video Context Modeling enhances video representations using target-related visual features and LLM-based video context, and Language-conditioned Transformer directly decodes free-form language queries using aligned video representations for moment retrieval.

language and visual inputs. Inspired by LLM-based prompting techniques and the success of MLLMs in multimodal tasks, we employ MLLMs to generate multi-view description texts as context prompts for the referred moments.

III. LLM-guided Moment Retrieval (LMR)

Given an untrimmed video $V = \{v_i\}_{i=1}^{N^v}$ with N^v clips and a query language $Q = \{q_i\}_{i=1}^{N^l}$ with N^l words, VMR aims to retrieve the best matching temporal moment M corresponding to the query. The target moment M in video V is denoted as $M = (m_c, w_\sigma)$, where m_c represents the center coordinate and w_σ is the temporal width.

The overall architecture of LMR is shown in Fig. 2. It comprises three main components: 1) LLM-based Video Context Enhancement (VCE) employs LLM to create target-related textual views based on videos and randomly sampled instructions; 2) Video Context Modeling encodes aligned target-related visual features and multi-view textual context for comprehensive video representations; 3) Language-conditioned Transformer directly uses free-form language as queries to identify referred moments based on these video representations.

A. Visual and Textual Encoders

Following the previous standard [14], [20], [51], we use CLIP [52] and SlowFast [53] as visual backbones to extract visual features $F^v = \{f_i^v\}_{i=1}^{N^v}$ every 2 seconds and CLIP-Text as the textual backbone to extract language features $F^l = \{f_i^l\}_{i=1}^{N^l}$, where f_i^v and f_i^l denote the visual features of the i-th video clip and the language features of the i-th word, respectively.

B. LLM-based Visual Context Enhancement

Although MLLMs can handle both video and text inputs simultaneously, their direct application to the VMR task is unreliable. This is because MLLMs are primarily optimized for knowledge-grounded text generation, without explicit focus on temporal reasoning in their fine-tuning objectives. As a result, we properly leverage the impressive text generation capabilities of MLLMs (with frozen weights) to produce textual descriptions for video clips. By extracting target-related context information from the generated descriptions, *e.g.*addpunct., the background of the current clip and the appearance of individuals, we strengthen the contextual semantic representations of videos, enabling the model to distinguish moments with similar appearances or similar actions.

Video Context Generation. In this work, we use VideoChat [29], which builds on BLIP-2 [28] and Stable-Vicuna[2], as our context generator due to its strong video comprehension capabilities. Specifically, we create an instruction list using ChatGPT, with sample instructions based on LLaVA [54]. The full instruction list can be found in Table I. For each non-overlapping 2s video clip, an instruction from the list is randomly selected for VideoChat to generate diverse descriptions that align with varying query styles. These descriptions range from detailed, such as "A woman in an orange blouse talking to a man in a light blue shirt", to concise, like "person on the phone". We generate descriptions for all clips of a video V simultaneously, then concatenate them chronologically to form the temporal context $T = \{t_i\}_{i=1}^{N^v}$. Finally, we extract target-related information from these de-

[2]https://github.com/Stability-AI/StableLM

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

TABLE I
LIST OF INSTRUCTIONS USED FOR VIDEOCHAT [29] IN VIDEO DESCRIPTION GENERATION.

1.	"Describe the following video concisely."
2.	"Present a brief overview of the provided video."
3.	"Provide a concise description of the given video."
4.	"Convey a short narrative summarizing the provided video."
5.	"Summarize the visual content of the following video."
6.	"Deliver a compact portrayal of the presented video."
7.	"Furnish a concise explanation of the given video."
8.	"Supply a brief account of the provided video."
9.	"Narrate the contents of the video with precision."
10.	"Offer a succinct analysis of the given video."

scriptions to enhance video representations for more precise localization.

Video-Query Fusion (VQF). Considering that LLM-generated description text contains redundant semantic information between video clips, we construct cross-attention layers to highlight referred moment-related context information based on query language. Specifically, initial textual representation $F^t = \{f_i^t\}_{i=1}^{N^t}$ and query representation $F^l = \{f_i^l\}_{i=1}^{N^t}$ are extracted by a frozen text encoder, where f_i^t represents the textual feature of i-th video clip and N^t is the number of description texts. For better feature interaction in cross-attention layers, we keep the language feature F^l and the textual feature F^t in a unified semantic space by extracting both with the same text encoder.

In each cross-attention layer, F^t serve as *query* and language features F^l provide *key* and *value*:

$$\mathrm{Attn}(Q^v, K^l, V^l) = \mathrm{softmax}(\frac{Q^v(K^l)^T}{\sqrt{d_1}})V^l \tag{1}$$

Here, K^l, V^l, Q^v are projected language and textual feature and d_1 representes their feature dimension. Then the textual feature $\tilde{F}^t$ can be represented as weighted sum of F^l. Following the standard transformer architecture, the attention scores are projected through a Multilayer Perceptron (MLP) and incorporated into the original textual representations. We refer the output of this module as LLM-based context feature, denoted as $\tilde{F}^t = \{f^{t'}_i\}_{i=1}^{N^t}$.

C. Video Context Modeling

Video Context Modeling enhances video representations with visual features and LLM-based video context, taking inputs from the LLM-based context feature $\tilde{F}^t$ and the target-related visual feature $\tilde{F}^v$. Here, $\tilde{F}^v$ is also computed with a VQF module, similarly to $\tilde{F}^t$ in Sec. III-B. Due to the temporal consistency between LLM-generated context F_t and visual features F_v, the two VQF modules share weights, facilitating joint learning of query relevance in their respective contexts. The target-related visual features $\tilde{F}^v$ and LLM-based context $\tilde{F}^t$ are then concatenated to form the video contextual sequence. Additionally, a randomly initialized saliency token is prepended to the sequence to predict the relevance score between video and text query. This sequence is encoded through two self-attention layers, yielding LLM-enhanced video representations F_m.

The video tokens $\{x_i\}_{i=1}^{N^v}$ within F_m serve for both decoding and video-query relevance computation. Decoding details are described in Sec. III-D. For the latter, the saliency token x_s and video tokens are projected using a single fully-connected layer, w_s and w_v, respectively. Thus, the predicted relevance score S_i between clip v_i and text query Q is computed as:

$$S_i = \frac{w_s^T x_s \cdot w_v^T x_v^i}{\sqrt{d}} \tag{2}$$

where d is the channel dimension of projected tokens. Subsequently, S_i and $\{x_i\}_{i=1}^{N^v}$ undergo contrastive learning (in Sec. III-E) and decoding, respectively.

D. Language-conditioned Transformer

The DETR-based encoding-decoding architecture is widely adopted in detection and retrieval tasks [14], [51]. It originates from the traditional object detection tasks, aiming to predict a set of bounding boxes and category labels for objects of interest when the exact search target is unknown. However, different from object detection, moment retrieval precisely knows the search target, *i.e.*, the video segment corresponding to the query language. Therefore, using randomly initialized query embeddings as the input for the decoder is not suitable for known-target moment retrieval. However, most existing methods overlook this distinction and apply DETR in a straightforward manner. Randomly initialized embeddings (*query*) struggle to effectively capture referred event related feature information from the visual sequence (*key,value*) in the cross-attention calculation at the decoder layer.

In our work, we address this issue by employing a small set of moment queries, as the input for the decoder. Consequently, all the queries are obligated to find the referred moment only. Specifically, we replicate the query language features F^l k times to create decoder embedding, *i.e.*, moment queries, together with trainable positional parameters as the position encoding, to encode temporal information, where k is a hyperparameter.

Note that the query language has been fused into both target-related visual features and LLM-based context features before being passed into the VCM module. The video context features output by VCM serve as the *K,V* in the decoder. Therefore, during the decoding process, the early-aligned video representations and the query language can better learn their co-relationship using the attention mechanism.

As a result, all moment queries are constrained by the language expressions, ensuring that they target only the referred moment, even when similar events exist in the video. One query will emerge with a significantly higher score, while other query scores will be suppressed. We define the output of the Transformer decoder as $H = \{h_i\}_{i=1}^{N^v+N^t}$, $h_i \in \mathbb{R}^{d_1}$.

E. Localization and Loss Functions

Finally, the temporal bounding box of the target moment is predicted based on the language-conditioned visual features H. As shown in Fig. 2, we feed the language-conditioned

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

visual features into a MLP layer to compute the center coordinate $\hat{m}_c \in [1, N^v]$ and window width $\hat{m}_w$ of the predicted moment $\hat{y}$, and a linear layer to compute the class g either foreground g_f or background g_b.

The loss $\mathcal{L}_{mr}$ is used to measure the discrepancy between the prediction and ground truth moments. Following [14], [51], [55], it consists of three losses: a $\mathcal{L}1$ loss for m_c, a generalized IoU loss $\mathcal{L}_{IoU}$ for m_c, m_w and a cross-entropy loss $\mathcal{L}_{CE}$ for g. Therefore, $\mathcal{L}mr$ is calculated as follows:

$$\mathcal{L}_{mr} = \lambda_{\mathcal{L}1}||m_c - \hat{m}_c|| + \lambda_{IoU}\mathcal{L}_{IoU}(m_w, \hat{m}_w) + \lambda_{CE}\mathcal{L}_{CE}, \quad (3)$$

$$\mathcal{L}_{CE} = -\sum_{y \in \mathcal{G}} ylog(\hat{y}), \mathcal{G} = [g_f, g_b], \quad (4)$$

where (m_c, m_w) are the center coordinates and window width of the ground truth moments y, and $\lambda_* \in \mathbb{R}$ are hyperparameters balancing the terms. Moreover, we treat sentences from other samples within batches as negative samples, which should have lower relevancy scores S to the videos, and introduce inter-video contrastive learning for cross-sample supervision:

$$\mathcal{L}_{cont} = -log(1 - S_{neg}) \quad (5)$$

Finally, our overall training objective is defined as:

$$\mathcal{L} = \lambda_{mr}\mathcal{L}_{cont} + \lambda_{cont}\mathcal{L}_{mr} \quad (6)$$

IV. Experiments

A. Datasets and Evaluation Metrics

In order to validate the effectiveness of the proposed LMR, we conducted extensive experiments on two publicly available datasets: Charades-STA [56] and QVHighlights [51].

Charades-STA [56] is a dataset for temporal sentence grounding. Derived from Charades, it provides 12,408 training and 3,720 testing moment-sentence pairs. The videos capture human actions and queries, offering an average video length of 29.8 seconds.

QVHighlights [51] is a dataset for moment retrieval and highlight detection, offering over 10,000 videos of 150-seconds each, along with human-written text queries. These queries describe relevant video segments averaging 24.6 seconds. The dataset contains 10,310 queries and 18,367 moments. We adopt the original data splits, using the train split for model training and the test split for evaluation.

TACoS [57] contains 127 videos collected from cooking scenarios. We also follow the public split, which includes 10146, 4589, 4083 query segment pairs for training, validation and testing.

Evaluation Metrics. We use evaluation metrics 'R@n, IoU=m' [58]–[61], which calculate the percentage of language queries having at least one correct retrieval (temporal IoU with ground truth moment is larger than m) in the top-n retrieved moments. Following standard practice, we use $n \in 1, 5$ and $m \in 0.5, 0.7$. For the Charades-STA dataset, we also adopt the mean average precision (mAP) at various thresholds, similar to prior works [20], [51].

B. Implementation Details

For a fair comparison, we simply leverage the pre-extracted SlowFast [53] and CLIP [52] features on QVHighlights and official VGG [67] features on Charades-STA, provided by other baselines: Moment-DETR [51] and UMT [20], respectively. For query text and generated description text, we use only the frozen embedding layer of Llama [50]. On QVHighlights, we use the query embeddings from CLIP instead before performing cross-attention with visual features, bringing them into the same feature spaces. The encoder is composed of 4 layers of transformer block (2 cross-attention layers and 2 self-attention layers) while the decoder has two layers, following [14]. For LLM, we employ the VideoChat [29], which is built on BLIP-2 [28] and StableVicuna[3] and remains frozen throughout its usage. The weights of losses are set as $\lambda_{mr} = 1$, $\lambda_{cont} = 1$, $\lambda_{\mathcal{L}_1} = 10$, $\lambda_{\mathcal{L}_{IoU}} = 1$ and $\lambda_{\mathcal{L}_{CE}} = 4$. The hidden dimensions of transformers are set to 256, with no expansions in feed-forward networks. In all experiments, we use Adam [68] optimizer with 1e-4 learning rate and 1e-4 weight decay. The model is trained with batch size 32 for 200 epochs on QVHighlights, and batch size 8 for 100 epochs on Charades-STA. All experiments are conducted on a single RTX 3090 GPU. LLM is utilized offline only once for video preprocessing, and our model's training process requires only 2-5 hours. Furthermore, inference on a video takes only 0.45 seconds after training.

C. Comparisons with State-of-the-art Methods

Comparison on QVHighlights. In Table II, we compare our performance with other state-of-the-art (SOTA) methods on the QVHighlights dataset. As shown, LMR achieves substantial improvements across all evaluation metrics. Among methods using video as the source, LMR outperforms previous SOTA methods by significant margins, with gains of 2.01% on both R1@0.7 and mAP@0.75. Notably, even when QD_DETR utilizes both video and audio, LMR still surpasses it on all metrics while relying solely on video. Specifically, LMR (V) outperforms QD_DETR (V+A) by 1.34%, 2.11%, 1.61%, and 3.06% on R1@0.5, R1@0.7, mAP@0.5, and mAP@0.75, respectively. Incorporating both video and audio further boosts LMR's performance, with increases of 2.21%, 2.28%, and 2.1% on R1@0.7, mAP@0.75, and mAP@Avg, respectively. For audio, we use the same query alignment structure as the video to extract target-related audio features for decoding. These results highlight the effectiveness of leveraging LLMs to model multimodal contexts in video moment retrieval.

Comparison on Charades-STA. Table III presents a detailed performance comparison with other state-of-the-art (SOTA) methods on the Charades-STA dataset. The results clearly demonstrate the proposed LMR's significant enhancements on R@1 metrics. LMR outperforms previous SOTA methods by substantial margins, achieving impressive improvements of 3.14% and 4.06% on the R1@0.5 and R1@0.7 metrics, respectively. Remarkably, even when QD_DETR employs both video and audio as data sources, LMR maintains its superior performance across R1@* evaluation metrics when relying solely on video source. More specifically, when assessed with VGG features, LMR surpasses QD_DETR by

[3]https://github.com/stability-AI/

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

TABLE II
PERFORMANCE COMPARISON ON THE QVHIGHLIGHTS *test* SPLIT. 'V' AND 'A' IN THE 'SRC' COLUMN REPRESENT THE VIDEO AND AUDIO MODALITIES, RESPECTIVELY. ALL METHODS USE SLOWFAST AND CLIP AS VISUAL BACKBONES. RED VALUES INDICATE PERFORMANCE IMPROVEMENTS ACHIEVED BY INCORPORATING THE AUDIO SOURCE.

Model	Src	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	mAP Avg.
MCN [62]	V	11.41	2.72	24.94	8.22	10.67
CAL [63]	V	25.49	11.54	23.40	7.65	9.89
XML [64]	V	41.83	30.35	44.63	31.73	32.14
XML+ [64]	V	46.69	33.46	47.89	34.67	34.90
M-DETR [51]	V	52.89	33.02	54.82	29.40	30.73
SEVILA [48]	V	54.5	36.5	-	-	-32.3
MomentDiff [65]	V	57.42	39.66	54.02	35.73	35.95
UniVTG [61]	V	58.86	40.86	57.60	35.59	35.47
QD-DETR [14]	V	62.40	44.98	62.52	39.88	39.86
MESM [66]	V	62.78	45.2	62.64	41.45	40.68
Ours	V	**64.40**	**47.21**	**64.65**	**43.16**	**42.56**
UMT [20]	V+A	56.23	41.18	53.38	37.01	36.12
QD-DETR [14]	V+A	63.06 (↑ 0.66)	45.1 (↑ 0.12)	63.04 (↑ 0.52)	40.1 (↑ 0.22)	40.19 (↑ 0.33)
Ours	V+A	**65.82** (↑ **1.42**)	**49.42** (↑ **2.21**)	**65.71** (↑ **1.06**)	**45.44** (↑ **2.28**)	**44.66** (↑ **2.1**)

TABLE III
PERFORMANCE COMPARISONS ON CHARADES-STA *test* SPLIT. ALL METHODS USE VIDEOS AS DATA SOURCES AND VGG FEATURES.

Model	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
MCN [62]	27.42	13.36	66.37	38.15
SAP [40]	27.42	13.36	66.37	38.15
MAN [38]	41.21	20.54	83.21	51.85
2D-TAN [12]	40.94	22.85	83.84	50.35
FVMR [69]	42.36	24.14	83.97	50.15
CBLN [70]	43.67	24.44	88.39	56.49
MMRG [71]	44.25	-	60.22	-
MMN [72]	47.31	27.28	83.74	58.41
G2L [58]	47.91	28.42	84.80	**59.33**
UMT [20]	48.31	29.25	**88.79**	56.08
QD-DETR [14]	52.77	31.13	-	
Ours	**55.91**	**35.19**	83.79	50.48

0.4% and 1.02% on the R1@0.5 and R1@0.7 metrics, respectively. These results underscore the effectiveness of integrating knowledge of LLMs for multimodal contextual semantic modeling. Furthermore, it is important to note that R1@* metrics are more challenging than R5@*, requiring models to achieve high levels of semantic alignment accuracy by satisfying IoU requirements with only one predicted moment. These results verify the significance of augmented query text in guiding the decoding process to focus on the most semantically relevant moments.

Furthermore, we investigate the performance improvements of our LMR compared to the previous state-of-the-art (QD-DETR [14]) across dense IoU values on mAP and R1 metrics. In Fig. 3, the horizontal axis represents IoU values in increments of 0.05, while the vertical axis indicates the percentage

TABLE IV
PERFORMANCE COMPARISONS ON THE TACOS [57] BENCHMARK. ALL BASELINES USE THE SAME VIDEO FEATURES (CLIP AND SLOWFAST) AND TEXT FEATURES (CLIP TEXT ENC.).

Method	R1@0.3	R1@0.5	R1@0.7	mIoU
2D-TAN [12]	40.01	27.99	12.92	27.22
VSLNet [24]	35.54	23.54	13.15	24.99
M-DETR [51]	37.97	24.67	11.97	25.49
UniVTG [61]	51.44	34.97	17.35	33.6
CG-DETR [73]	52.23	-	22.23	36.48
Ours	**53.24**	**37.16**	**22.81**	**37.25**

of performance improvement of LMR over the state-of-the-art. It can be observed that as the IoU requirements with ground truth increase, the relative improvement of LMR over state-of-the-art also gradually increase. This indicates that LMR's superiority becomes more pronounced when stricter requirements for moment localization are imposed. These results verify the significant enhancement of the proposed LMR in semantic alignment between language and vision as well.

Comparison on TACoS. Results in Table IV also compare VMR performances against the models using Slowfast and CLIP features [51] on TACoS dataset. As reported, we validate that our model surpasses the existing SOTA methods across all metrics.

D. Comparison on Complex Query split

Construction of *Complex Query* (C-QVal) split. To evaluate the effectiveness of our proposed LMR in video retrieval for complex queries, we extract splits of complex queries from the val split of the QVHighlights dataset, referred to as "*C-QVal*" split. Considering that queries with longer lengths and more clauses may have a higher chance of including

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

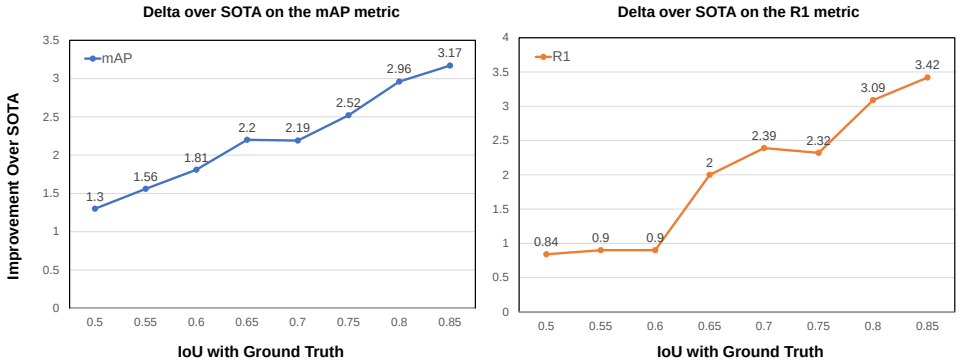


Fig. 3. Improvements over the state-of-the-art (QD-DETR [14]) on mAP and R1 metrics on the Charades-STA dataset.

strict contextual requirements, such as video background requirements, character appearance specifications, and action constraints, etc., we determine whether a query belongs to complex queries based on the presence of clauses and word numbers. Moreover, the complexity of a query increases with the simultaneous inclusion of more clauses and words, indicating higher contextual demands for the queried activities and greater localization challenges. Specifically, we utilize the syntactic analysis library spaCy[4] from the NLP field to calculate the number of clauses in a query. We set different thresholds for the number of clauses and words to obtain *C-QVal* split with varying levels of complexity. For complex query extraction under different parameter settings (clauses and word count), we obtained multiple sets of *C-QVal* splits with varying numbers of queries, ranging from 48 to 388. After training the model on the complete *train* split of the QVHighlights dataset, we perform inference on *C-QVal* split and compute metrics to validate the model's alignment capability for complex queries.

Comparison with SOTA Method. Table VI presents detailed comparison results with the SOTA method QD-DETR [14] on multiple *C-QVal* splits. Compared with QD-DETR, our method demonstrates a significant performance improvement across all *C-QVal* splits, achieving an average boost of 7.99% and 7.52% on the most challenging evaluation metrics R@1 IoU=0.7 and mAP IoU=0.7, respectively. This validates the effectiveness of LMR in complex query localization. Moreover, the boosts for both metrics generally exhibit an increasing trend with the growing difficulty of *C-QVal* splits, reaching their peaks at (least clause count, least word count)=(1,16) with boosts of up to 13.79% and 10.30%, respectively. This suggests the tremendous potential of LMR in localizing complex contextualized queries. The experimental results above confirm that fusing visual content with complementary context information derived from LLMs has a profound impact on enhancing the semantic alignment between videos and complex contextualized queries.

E. Comparison without temporal bias

We conducted additional experiments on the Charades-CD dataset [74] to evaluate the robustness of our model. In Charades-CD, the temporal locations of samples in the training, validation, and in-distribution (iid) test splits follow the original distribution of the Charades-STA [56] dataset, while the ood test split contains out-of-distribution samples. For fair comparison, we re-implemented state-of-the-art (SOTA) models under the same training conditions. As shown in Table VII, our method consistently outperforms SOTA methods, exhibiting notably smaller performance drops in the ood setting compared to competitors. Specifically, compared to M-DETR [52], our method reduces the performance drop by 3.33% on average, and by 2.13% compared to QD-DETR [14], indicating stronger robustness to ood data. Furthermore, our method achieves 79.38% in R5@0.5 in the ood setting, with only a 3.46% drop, highlighting its ability to localize query-related moment candidates without relying on temporal bias.

TABLE V
ABLATION ON LMR COMPONENTS.VCE, VFE, AND LMQ DENOTE LLM-BASED VIDEO CONTEXT ENHANCEMENT, VISUAL FEATURE EXTRACTION, AND LANGUAGE-CONDITIONED MOMENT QUERIES.

Model	R1 @0.5	R1 @0.7	mAP @0.5	mAP @0.75	mAP Avg.
w/o VCE	60.55	44.39	60.18	39.77	38.68
w/o VFE	55.13	37.81	56.51	34.83	33.14
w/o LMQ	63.87	47.39	63.36	42.49	42.12
Full Model	65.48	48.58	65.06	44.45	43.46

F. Ablation Study

Here, we investigate the effectiveness of the individual components within the proposed LMR as well as the effectiveness of the internal design in LLM-based Video Context Enhancement module (VCE). These experiments are conducted on the QVHighlights validation set.

Effectiveness of Individual Components. we conduct a thorough ablation study on the three main components in LMR to verify their effectiveness. As depicted in Table V, the removal of the entire LLM-based Video Context Enhancement (w/o VCE) leads to a performance degradation of up to 4.54%, underscoring the significance of incorporating contextual scene information from LLMs. Additionally, from w/o VFE, we observe that the model can still function normally without the

[4]https://spacy.io/

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

TABLE VI
COMPARISON TO THE SOTA METHOD QD-DETR [14] ON THE LOCALIZATION OF COMPLEX QUERIES ON MULTIPLE *C-QVal* SPLIT.

Metrics	Model	(Least Clause Count, Least Word Count)												
		(1,8)	(1,9)	(1,10)	(1,11)	(1,12)	(1,13)	(1,14)	(1,15)	(1,16)	(2,8)	(2,9)	(2,10)	**Avg**
R1@0.7	QD_DETR	46.39	45.89	44.59	43.61	44.69	42.55	44.44	44.00	37.93	50.00	51.02	50.00	45.43
	LMR	54.12	52.97	52.46	53.01	53.98	53.72	52.94	52.8	51.72	54.00	55.10	54.17	53.42
	Perf. Gain	**+7.73**	**+7.08**	**+7.87**	**+9.40**	**+9.29**	**+11.17**	**+8.50**	**+8.80**	**+13.79**	**+4.00**	**+4.08**	**+4.17**	**+7.99**
mAP@0.7	QD_DETR	43.96	43.54	41.94	39.92	41.06	40.72	42.68	40.95	34.86	45.24	46.17	46.09	42.26
	LMR	50.15	49.01	48.58	47.25	48.50	49.08	49.58	48.91	45.16	52.95	54.03	54.12	49.78
	Perf. Gain	**+6.19**	**+5.47**	**+6.64**	**+7.33**	**+7.44**	**+8.36**	**+6.90**	**+7.96**	**+10.30**	**+7.71**	**+7.86**	**+8.03**	**+7.52**

(clause count, word count) represents the minimum number of clauses and words simultaneously included in a query. Higher values indicate greater complexity of the queries, implying more stringent contextual requirements for locating the activity moments.

TABLE VII
COMPARISON ON CHARADES-CD SPLIT USING VGG FEATURES. RED VALUES (↓) INDICATE THE PERFORMANCE DROP OF METHODS IN THE OUT-OF-DISTRIBUTION (OOD) SETTING RELATIVE TO THE IN-DISTRIBUTION (IID) SETTING.

Model	Split	R1@0.3	R1@0.5	R1@0.7	R5@0.5	R5@0.75
M-DETR [52]	iid	62.04	49.27	27.62	76.03	41.61
	ood	53.51 (↓ 8.53)	35.8 (↓ 13.47)	12.98 (↓ 14.64)	67.53 (↓ 8.5)	25.04 (↓ 16.57)
QD-DETR [14]	iid	63.15	53.31	30.69	75.93	47.22
	ood	55.03 (↓ 8.12)	40.25 (↓ 13.06)	16.72 (↓ 13.97)	67.79 (↓ 8.14)	32.83 (↓ 14.39)
Ours	iid	**65.69**	**55.72**	**35.77**	**82.85**	**51.09**
	ood	**59.15** (↓ 6.54)	**44.50** (↓ 11.22)	**23.63** (↓ 12.14)	**79.38 (↓ 3.46)**	**39.38 (↓ 11.71)**

TABLE VIII
ABLATION OF COMPONENTS WITHIN VCE.

Model	R1 @0.5	R1 @0.7	mAP @0.5	mAP @0.75	mAP Avg.
repl. LLM_{LLaVA}	61.61	46.95	62.60	43.41	42.13
repl. LLM_{ImgCap}	60.82	44.98	61.76	41.03	39.91
w/o WS	63.35	48.13	63.59	44.36	43.12
w/o VQF	63.22	46.50	62.87	42.48	41.30
Full Model	65.48	48.58	65.06	44.45	43.46

repl. LLM (LLaVA): replace the video-based LLM VideoChat with a image-based LLM LLaVA [54]; repl. LLM (ImgCap): replace LLM with a general image captioning model [75]; w/o WS: no weight sharing between the two VQFs; w/o VQF: remove VQF.

original video-based visual features, although there is obvious performance drop. The combined results from w/o VCE and w/o VFE demonstrate that visual features play an important role in VMR. However, LLM-generated temporal scene information effectively provides complementary contextual details, enhancing localization performance. In experiments without LMQ, we replace Language-conditioned positional embedding with randomly initialized positional embedding as the decoder input. This leads to an average performance drop of 1.56 points, emphasizing the importance of using language as moment queries for improved decoder guidance in identifying the target moment.

Effectiveness of LLM-based Video Context Enhancement. In Table VIII, we conduct a detailed ablation study on the internal design of VCE. For the LLM replacements (LLaVA and ImgCap), we replace the video-based LLM VideoChat [29] with an image-based LLM LLaVA-1.6 [54] and a general image captioning model UTD [75] to generate image captions. The images are sampled every 2 seconds in videos, and we apply the same instructions for LLaVA as those used in VideoChat. As illustrated in Table VIII, the performance of LLaVA replacement exhibits a decrease of 1.04% to 3.87%, while ImgCap experiences a more significant decline, ranging from 3.3% to 4.66%. This is because, even though image-based LLMs lack temporal and audio information like video-based LLMs, their generated captions can still offer more valid and detailed scene information than general image captioning models, given that LLMs are pretrained on web-scale data. In w/o WS, the model shows a significant 2.13% drop in R1@0.5, verifying the advantages of weight sharing between VQFs. This is due to the temporal correspondence between LLM-generated scene information and visual features, which enables joint learning of query relevance in their respective contexts. Directly removing VQF in VCE results in a performance degradation of 2.13% on average, highlighting the effectiveness of using query language to emphasize related context information derive from LLM.

Ablation Studies on Language-conditional Moment Query Number. In Table IX, we evaluate the performance of our model with different numbers of Moment Queries. Previous methods [14], [51] need to carefully set the number k of Moment Queries to achieve good performance. For

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797



Fig. 4. Qualitative results. (a)(b): queries with specific backgrounds, (c)(d): queries in specific contextual events, (e): queries with particular appearances, (f): queries specifying certain actions. The above constraints are highlighted in yellow in the queries. The yellow rectangular boxes represent moments semantically aligning with the query without yellow-highlighted scene constraints. We observed that QD-DETR [14] tends to localize the moments missing the background and action details.

example, reducing k from 20 to 10 in [51] leads to substantial drops in R1@0.5, R1@0.7, and average mAP metrics by 6%, 5.74%, and 7.39%, respectively. In contrast, our model consistently achieves good performance for different values of k, even with $k = 1$. This benefits from the design of language-conditioned moment queries focusing on localizing referred activities. Moreover, our model shows potential for further improvement since its accuracy increases with the number of moment queries. For instance, increasing k from 10 (adopted in our main experiments to align with prior works) to 20 boosts the R1@0.7 score of our model by 1.5%. We conjecture that since language-conditional moment queries provide explicit retrieval targets during decoding, additional moment queries enable the model to generate a larger number of quality moment candidates, enhancing its ability to handle complex video scenes with multiple similar activities.

TABLE IX
ABLATION STUDY FOR THE NUMBER (k) OF LANGUAGE-CONDITIONAL MOMENT QUERIES ON QVHIGHLIGHTS VALIDATION SPLIT.

k	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	mAP Avg.
1	62.84	48.58	59.10	41.54	40.00
3	62.84	48.58	59.10	41.54	40.08
5	63.03	48.19	60.91	41.15	40.60
8	63.87	48.13	63.79	42.8	42.17
10	63.74	49.16	63.96	44.03	43.45
20	64.26	50.65	64.73	45.37	44.52
30	64.65	50.06	64.64	45.26	44.00

G. Qualitative Analysis

In Fig. 4, we present a qualitative comparison with QD-DETR [14]. Without the constraints highlighted in yellow in the queries, each example may encompass multiple matched moments, such as "a couple is traveling" (a), "a woman standing before a mirror" (e), and "a cat on the floor"(f). However, when specific constraints are applied, such as backgrounds, contextual events, appearances, and actions, the existing method struggles to distinguish the correct moment from similar-looking and similar-action moments. In contrast, our proposed LMR successfully localizes the target moment. This is attributed to the extensive scene information obtained from Large Language Models (LLMs) for each video clip. When modulated by the query, this information enables the model to dynamically perceive query-relevant detailed changes in scene semantics, such as the color change of character's outfit in (e), the background alteration in (a) and (b), and the subtle movement changes in (d). Consequently, this improves the alignment between complex queries and videos. Additional qualitative examples are provided in the supplementary material.

The Fig. 5 compares the attention values of full-text query and partial-text query by removing "taking a photo shoot" for example (c) in Fig. 4. They are calculated between video context features and text queries. The lower attention value in "first walking" segment in the full-text branch shows that the model truly focuses on the key information in the query text.

V. CONCLUSION

We proposed a Large Language Model-based Moment Retrieval (LMR) approach to tackle the localization challenges posed by complex queries with spatiotemporal context constraints. We presented a context enhancement technique with the inherent knowledge of LLMs and video-query fusion to generate crucial target-related video context. We also developed a language-conditioned transformer to directly decode free-form language queries for moment retrieval. Finally,

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

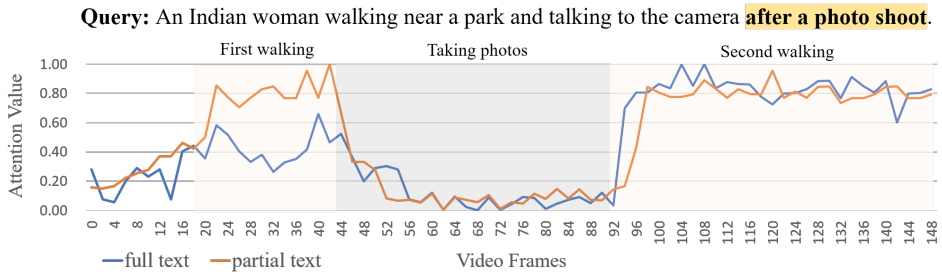


Fig. 5. Comparison of attention values with/without specific information.

we evaluated our models on popular VMR benchmarks and our constructed C-QVal, and demonstrated that our approach achieves top-ranked performance. While our model demonstrates strong performance, it may struggle with ambiguous queries containing polysemous words. Future work will explore leveraging LLMs to enrich query semantics and enhance cross-modal alignment in such cases.

REFERENCES

[1] D. Liu, J. Zhu, X. Fang, Z. Xiong, H. Wang, R. Li, and P. Zhou, "Conditional video diffusion network for fine-grained temporal sentence grounding," *IEEE Transactions on Multimedia*, 2023.

[2] D. Liu, X. Fang, W. Hu, and P. Zhou, "Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding," *IEEE Transactions on Multimedia*, 2023.

[3] X. Fang, D. Liu, P. Zhou, Z. Xu, and R. Li, "Hierarchical local-global transformer for temporal sentence grounding," *IEEE Transactions on Multimedia*, 2023.

[4] G. Gong, L. Zhu, and Y. Mu, "Language-guided multi-granularity context aggregation for temporal sentence grounding," *IEEE Transactions on Multimedia*, 2022.

[5] T. Chen, W. Wang, Z. Jiang, R. Li, and B. Wang, "Cross-modality knowledge calibration network for video corpus moment retrieval," *IEEE Transactions on Multimedia*, 2023.

[6] X. Fang, D. Liu, P. Zhou, and Y. Hu, "Multi-modal cross-domain alignment network for video moment retrieval," *IEEE Transactions on Multimedia*, 2022.

[7] H. Tang, J. Zhu, M. Liu, Z. Gao, and Z. Cheng, "Frame-wise cross-modal matching for video moment retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 1338–1349, 2021.

[8] G. Wang, X. Xu, F. Shen, H. Lu, Y. Ji, and H. T. Shen, "Cross-modal dynamic networks for video moment retrieval with text query," *IEEE Transactions on Multimedia*, vol. 24, pp. 1221–1232, 2022.

[9] R. Chen, J. Tan, Z. Yang, X. Yang, Q. Dai, Y. Cheng, and L. Lin, "Dphanet: Discriminative parallel and hierarchical attention network for natural language video localization," *IEEE Transactions on Multimedia*, 2024.

[10] Y. Zeng, N. Han, K. Pan, and Q. Jin, "Temporally language grounding with multi-modal multi-prompt tuning," *IEEE Transactions on Multimedia*, 2023.

[11] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury, "Weakly supervised video moment retrieval from text queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 592–11 601.

[12] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 870–12 877.

[13] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7190–7198.

[14] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-dependent video representation for moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 023–23 033.

[15] S. Huo, Y. Zhou, R. Wang, W. Xiang, and S.-Y. Kung, "Semantic relevance learning for video-query based video moment retrieval," *IEEE Transactions on Multimedia*, 2023.

[16] M. Zhang, Y. Yang, X. Chen, Y. Ji, X. Xu, J. Li, and H. T. Shen, "Multi-stage aggregated transformer network for temporal language localization in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 669–12 678.

[17] S. Chen, W. Jiang, W. Liu, and Y.-G. Jiang, "Learning modality interaction for temporal sentence localization and event captioning in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 333–351.

[18] X. Qu, P. Tang, Z. Zou, Y. Cheng, J. Dong, P. Zhou, and Z. Xu, "Fine-grained iterative attention network for temporal language localization in videos," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4280–4288.

[19] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] Y. Liu, S. Li, Y. Wu, C.-W. Chen, Y. Shan, and X. Qie, "Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3042–3051.

[21] J. Hao, H. Sun, P. Ren, J. Wang, Q. Qi, and J. Liao, "Query-aware video encoder for video moment retrieval," *Neurocomputing*, vol. 483, pp. 72–86, 2022.

[22] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *European Conference on Computer Vision*. Springer, 2022, pp. 105–124.

[23] Y.-W. Chen, Y.-H. Tsai, and M.-H. Yang, "End-to-end multi-modal video temporal grounding," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28 442–28 453, 2021.

[24] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," *arXiv preprint arXiv:2004.13931*, 2020.

[25] Z. Zhang, Z. Lin, Z. Zhao, and Z. Xiao, "Cross-modal interaction networks for query-based moment retrieval in videos," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 655–664.

[26] O. OpenAI, "Gpt-4 technical report," Mar 2023.

[27] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[28] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[29] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[30] L. Li, Y. Yin, S. Li, L. Chen, P. Wang, S. Ren, M. Li, Y. Yang, J. Xu, X. Sun *et al.*, "M3it: A large-scale dataset towards multi-modal multilingual instruction tuning," *arXiv preprint arXiv:2306.04387*, 2023.

[31] J. Pan, Z. Lin, Y. Ge, X. Zhu, R. Zhang, Y. Wang, Y. Qiao, and H. Li, "Retrieving-to-answer: Zero-shot video question answering with frozen large language models," *arXiv preprint arXiv:2306.11732*, 2023.

[32] M. F. Naeem, M. G. Z. A. Khan, Y. Xian, M. Z. Afzal, D. Stricker, L. Van Gool, and F. Tombari, "I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 169–15 179.

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3581797

[33] Y. Zang, W. Li, J. Han, K. Zhou, and C. C. Loy, "Contextual object detection with multimodal large language models," *arXiv preprint arXiv:2305.18279*, 2023.

[34] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, "Bliva: A simple multimodal llm for better handling of text-rich visual questions," *arXiv preprint arXiv:2308.09936*, 2023.

[35] H. Wang, Z.-J. Zha, L. Li, D. Liu, and J. Luo, "Structured multi-level interaction network for video moment localization via language query," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7026–7035.

[36] J. Li, J. Xie, L. Qian, L. Zhu, S. Tang, F. Wu, Y. Yang, Y. Zhuang, and X. E. Wang, "Compositional temporal grounding with structured variational cross-graph correspondence learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3032–3041.

[37] D. Liu, X. Qu, X.-Y. Liu, J. Dong, P. Zhou, and Z. Xu, "Jointly cross-and self-modal graph attention network for query-based moment localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4070–4078.

[38] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1247–1257.

[39] R. Ge, J. Gao, K. Chen, and R. Nevatia, "Mac: Mining activity concepts for language-based temporal localization," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 245–253.

[40] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8199–8206.

[41] X. Yang, F. Feng, W. Ji, M. Wang, and T.-S. Chua, "Deconfounded video moment retrieval with causal intervention," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 1–10.

[42] X. Yang, S. Wang, J. Dong, J. Dong, M. Wang, and T.-S. Chua, "Video moment retrieval with cross-modal neural architecture search," *IEEE Transactions on Image Processing*, vol. 31, pp. 1204–1216, 2022.

[43] J. Dong, M. Zhang, Z. Zhang, X. Chen, D. Liu, X. Qu, X. Wang, and B. Liu, "Dual learning with dynamic knowledge distillation for partially relevant video retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 302–11 312.

[44] J. Gao and C. Xu, "Learning video moment retrieval without a single annotated video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1646–1657, 2021.

[45] V. Tran, N. Balasubramanian, and M. Hoai, "From within to between: Knowledge distillation for cross modality retrieval," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3223–3240.

[46] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," *arXiv preprint arXiv:2308.00692*, 2023.

[47] W. Ji, L. Li, H. Fei, X. Liu, X. Yang, J. Li, and R. Zimmermann, "Towards complex-query referring image segmentation: A novel benchmark," *arXiv preprint arXiv:2309.17205*, 2023.

[48] S. Yu, J. Cho, P. Yadav, and M. Bansal, "Self-chained image-language model for video localization and question answering," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[49] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[50] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[51] J. Lei, T. L. Berg, and M. Bansal, "Detecting moments and highlights in videos via natural language queries," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 846–11 858, 2021.

[52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[53] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[54] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[55] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[56] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 510–526.

[57] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013.

[58] H. Li, M. Cao, X. Cheng, Y. Li, Z. Zhu, and Y. Zou, "G2l: Semantically aligned and uniform video grounding via geodesic and game theory," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 032–12 042.

[59] H. Li, X. Shu, S. He, R. Qiao, W. Wen, T. Guo, B. Gan, and X. Sun, "D3g: Exploring gaussian prior for temporal sentence grounding with glance annotation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 734–13 746.

[60] S. Yan, X. Xiong, A. Nagrani, A. Arnab, Z. Wang, W. Ge, D. Ross, and C. Schmid, "Unloc: A unified framework for video localization tasks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 623–13 633.

[61] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou, "Univtg: Towards unified video-language temporal grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2794–2804.

[62] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5803–5812.

[63] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5267–5275.

[64] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvr: A large-scale dataset for video-subtitle moment retrieval," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 447–463.

[65] P. Li, C.-W. Xie, H. Xie, L. Zhao, L. Zhang, Y. Zheng, D. Zhao, and Y. Zhang, "Momentdiff: Generative video moment retrieval from random to real," *Advances in neural information processing systems*, vol. 36, 2023.

[66] Z. Liu, J. Li, H. Xie, P. Li, J. Ge, S.-A. Liu, and G. Jin, "Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3855–3863.

[67] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[69] J. Gao and C. Xu, "Fast video moment retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1523–1532.

[70] D. Liu, X. Qu, J. Dong, P. Zhou, Y. Cheng, W. Wei, Z. Xu, and Y. Xie, "Context-aware biaffine localizing network for temporal sentence grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 235–11 244.

[71] Y. Zeng, D. Cao, X. Wei, M. Liu, Z. Zhao, and Z. Qin, "Multi-modal relational graph for cross-modal video moment retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2215–2224.

[72] Z. Wang, L. Wang, T. Wu, T. Li, and G. Wu, "Negative sample matters: A renaissance of metric learning for temporal grounding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2613–2623.

[73] W. Moon, S. Hyun, S. Lee, and J.-P. Heo, "Correlation-guided query-dependency calibration in video representation learning for temporal grounding," *arXiv preprint arXiv:2311.08835*, 2023.

[74] J. Hao, H. Sun, P. Ren, J. Wang, Q. Qi, and J. Liao, "Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding," in *European Conference on Computer Vision*. Springer, 2022, pp. 130–147.

[75] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

Authorized licensed use limited to: University of Western Australia. Downloaded on September 11,2025 at 00:45:09 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.