See, Rank, and Filter: Important Word-Aware Clip Filtering via Scene Understanding for Moment Retrieval and Highlight Detection

YuEun Lee, Jung Uk Kim*

Kyung Hee University, Yong-in, South Korea
{dbdms8435, ju.kim}@khu.ac.kr

Abstract

Video moment retrieval (MR) and highlight detection (HD) with natural language queries aim to localize relevant moments and key highlights in a video clips. However, existing methods overlook the importance of individual words, treating the entire text query and video clips as a black-box, which hinders contextual understanding. In this paper, we propose a novel approach that enables fine-grained clip filtering by identifying and prioritizing important words in the query. Our method integrates image-text scene understanding through Multimodal Large Language Models (MLLMs) and enhances the semantic understanding of video clips. We introduce a feature enhancement module (FEM) to capture important words from the query and a ranking-based filtering module (RFM) to iteratively refine video clips based on their relevance to these important words. Extensive experiments demonstrate that our approach significantly outperforms existing state-of-the-art methods, achieving superior performance in both MR and HD tasks.

Introduction

The expansion of digital devices and internet platforms has sparked growing interest in video content, resulting in exponential growth in both its volume and diversity (Apostolidis et al. 2021; Foo et al. 2023). While this vast amount of content contains valuable information, reviewing it to extract relevant parts is time-consuming (Apostolidis et al. 2021). To address this, two key tasks have emerged for finding specific clips of interest based on text queries. One is moment retrieval (MR), which aims to locate relevant moments within videos (Gao et al. 2017), and the other is highlight detection (HD), which tries to identify the most important clips of the videos (Sun, Farhadi, and Seitz 2014).

Given the similarity between MR and HD tasks in identifying important video clips, the introduction of Moment-DETR and the QVhighlights dataset (Lei, Berg, and Bansal 2021) has encouraged joint approaches. Moment-DETR (Lei, Berg, and Bansal 2021) first applied the DETR framework (Carion et al. 2020) for this purpose. UMT (Liu et al. 2022) devised a framework by incorporating audio modality, and UVCOM (Xiao et al. 2024) proposed an integration module for the inter- and intra-modality interaction.

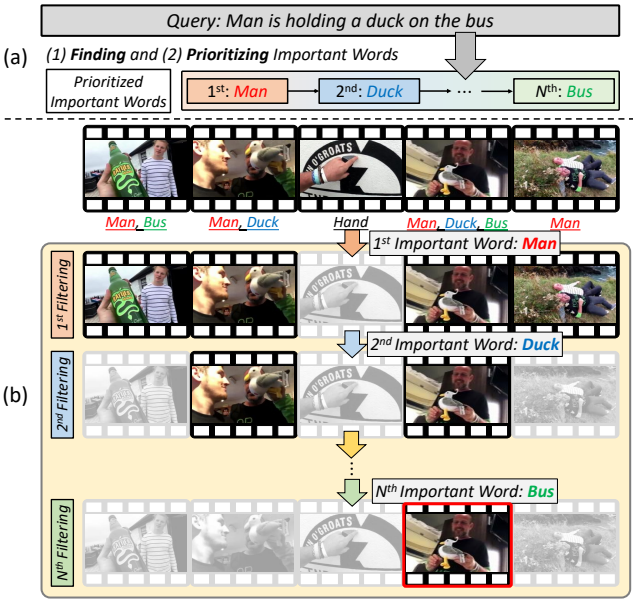

Figure 1: Conceptual illustration of our method. First, we aim to (a) find and prioritize important words in a text query, and (b) filter video clips based on the priority of the words.

TR-DETR (Sun et al. 2024) focused on task interaction during training, while TaskWeave (Yang et al. 2024) proposed a task-oriented framework for more effective MR and HD. Keyword-DETR (Um et al. 2025) enhanced alignment by identifying key words in the query.

While recent studies have achieved significant improvements in MR and HD considering task similarities and differences, their performance is still limited due to underutilization of the unique characteristics of text queries and video clips. To address this limitation, we draw inspiration from how humans utilize both modalities for MR and HD tasks. First, for the aspects of the text query, according to (Just and Carpenter 1980), given a query like '*Man is holding a duck on the bus*', we naturally pre-identify important words (*e.g.,* '*man*,' '*duck*,' '*bus*,' '*hold*') and prioritize these important words before analyzing the video (Figure 1(a)). After that, we filter the video clips related to these prioritized important words to effectively identify the clips that

*Corresponding author.

align with the query (Figure 1(b)). However, existing methods treat the processing of the query text and video clip as a black-box, failing to prioritize the important words and thereby lacking a sufficient understanding of the query.

Second, for the aspect of the video clips, we not only interpret the visual content itself but also leverage scene understanding information (Vo et al. 2022; Buch et al. 2022). Specifically, we analyze spatial layout, objects interactions, actions within the scene, and the temporal changes of the scene to comprehend the context while watching the video. By doing so, we gain a deeper scene understanding, which helps us find the video clip segments that are most relevant to the text query. However, existing methods rely mainly on raw visual content, limiting their ability to fully capture the scene context and properly align it with the query.

In this paper, building on these insights, we propose a novel approach that enables more fine-grained clip filtering related to the text query by fully leveraging image-text scene understanding. To this end, we address two main aspects: (*i*) identifying important words in the text query and understanding video clips effectively, and (*ii*) filtering video clips that are most relevant to these important words.

First, to address the issue (*i*), we introduce a feature enhancement module (FEM) to identify and prioritize the important words given a text query. We also leverage Multimodal Large Language Models (MLLMs) to obtain detailed scene understanding through their rich external knowledge. By integrating MLLMs, ours further enhances the ability to interpret deeper and more complex scene understanding by combining image-text information from video clips.

Second, to address the issue (*ii*), we propose a ranking-based filtering module (RFM) that refines video clips, based on the relevance of the prioritized important word. At this time, important query words are matched with the image-text information in the video clips in an iterative manner, gradually minimizing the effect of the irrelevant clips while highlighting the relevant ones. As a result, more accurate MR and HD are possible. We claim that our use of MLLMs can offer valuable insights for future MR and HD tasks by effectively integrating image-text multimodal information.

The major contributions of our paper are summarized as:

- We propose a feature enhancement module (FEM) that identifies and prioritizes important words in text queries, while enhancing detailed scene understanding through the utilization of MLLMs.
- We introduce a ranking-based filtering module (RFM) that iteratively refines video clips by filtering clips based on the relevance of prioritized important words, improving moment retrieval and highlight detection.

Related work

Moment Retrieval and Highlight Detection

Moment retrieval (MR) aims to find relevant video moments given a natural language query (Gao et al. 2017). There are two main approaches: proposal-based and proposal-free. Proposal-based methods (Gao et al. 2017; Hendricks et al. 2018; Sun et al. 2022) generate candidate segments and rank them based on their match scores with the query. In contrast, proposal-free methods (Li, Guo, and Wang 2021; Mun, Cho, and Han 2020; Rodriguez et al. 2020) directly regress start and end timestamps via video-text interaction.

In addition, the highlights detection (HD) aims to identify the most important video moments, *i.e.,* highlights. Early methods gave high importance scores to important moments regardless of text queries (Sun, Farhadi, and Seitz 2014; Wei et al. 2022; Badamdorj et al. 2022). As user preferences guide content consumption, recent HD approaches incorporate text queries to better personalize highlight selection.

Recently, MR and HD have been studied jointly. Moment-DETR (Lei, Berg, and Bansal 2021) presents QVHighlights dataset and applieds DETR to both tasks. UMT (Liu et al. 2022) incorporates audio alongside visual and textual inputs to enhance query understanding. QD-DETR (Moon et al. 2023) leverages text by modeling negative video-text pairs, while TR-DETR (Sun et al. 2024) and UVCOM (Xiao et al. 2024) emphasize the synergy between MR and HD. TaskWeave (Yang et al. 2024) adopts a task-centric top-down approach, and Keyword-DETR (Um et al. 2025) introduces keyword-aware attention for adaptive focus. However, existing methods fail to capture the overall context of the video and understand the semantic information of each clip, which fails to align properly with the query. To address this issue, we propose important word-aware clip filtering framework that iteratively filters out information irrelevant to the query by integrating and leveraging scene understanding knowledge to better interpret complex scene contexts.

Multimodal Large Language Model

Multimodal Large Language Models (MLLMs) have evolved to meet the growing need for models that can handle multiple modalities, including not only text but also image, video, and audio. For example, CLIP (Radford et al. 2021) aligns visual and language modalities via contrastive learning on a large set of image-text pairs. BLIP-2 (Li et al. 2023) introduces Qformer to efficiently bridge the gap between modalities, while MiniGPT-4 (Zhu et al. 2023) uses a single projection layer to match visual features to textual features. LLaVA (Liu et al. 2023) improves multimodal dialogue ability by tuning instructions on multimodal data generated by GPT-4 (Achiam et al. 2023). QWen-VL (Bai et al. 2023) uses a multi-task training strategy to fine-tune on high-resolution images. InternVL (Chen et al. 2024) scales the vision-based model and gradually aligns it with LLMs. In this paper, we use InternVL2 to obtain rich knowledge for scene understanding of video clips.

Proposed Method

Figure 2 shows the overall framework. The text query with L_q words and the untrimmed video with L_v clips are processed by pre-trained modality-specific encoders (textual and visual) with three-layer feed-forward network to generate query features $F_q \in \mathbb{R}^{L_q \times d}$ and visual features $F_v \in \mathbb{R}^{L_v \times d}$. We also adopt the recent state-of-the-art Multimodal Large Language Models (MLLMs), *i.e.*, InternVL2 (Chen et al. 2024), to generate captions that provide detailed semantic descriptions of each clip with up to L_c words. These

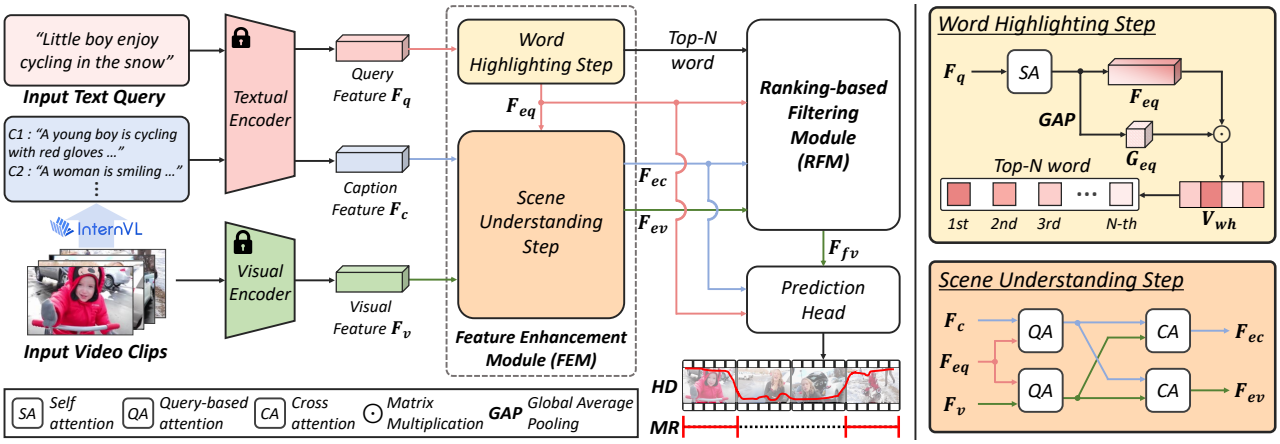

Figure 2: Overall architecture. Query, visual and caption features are prioritize important words and deepen scene understanding via feature enhancement module, then repeatedly filter irrelevant information to the query via ranking-based filtering module.

are encoded by textual encoder to generate caption features $F_c \in \mathbb{R}^{L_v \times d}$.

Given F_v, F_q, and F_c, the feature enhancement module (FEM) first identifies important words from the text query in a self-supervised manner in the word highlighting step to generate enhanced query features F_{eq} and generates a word-highlighted vector V_{wh} to rank important words. This module also enriches scene understanding by associating video-query and caption-query pairs in the scene understanding step, yielding enhanced visual features F_{ev} and enhanced caption features F_{ec}. Then, based on the identified important words, the ranking-based filtering module (RFM) filters clips by emphasizing those most relevant to the query while reducing the effect of unrelated ones. This process is repeated N times to obtain filtered visual features F_{fv}. Finally, F_{eq}, F_{fv}, and F_{ec} pass through a transformer encoder-decoder to perform moment retrieval (MR) and highlight detection (HD). More details are in the following subsections.

Feature Enhancement Module

Since the text query specifies the exact moments the user is looking for, accurately recognizing the words in the text query is essential for effective MR and HD. The video offers visual cues, while the generated clip captions provide the detailed meaning of each clip. Therefore, to effectively capture the specific moments that user is searching for based on the text query, it is important to establish strong cross-modal associations to interpret complex scene contexts.

To this end, we propose a feature enhancement module (FEM) which consists of two steps: (*i*) word highlighting step and (*ii*) scene understanding step. First, in the word highlighting step, as text queries contain important words and contextual information that convey the meaning of the sentence, but direct label supervision is not available, we generate enhanced query features $F_{eq} \in \mathbb{R}^{L_q \times d}$ by identifying the context through self-attention mechanism. Afterwards, the word-highlighted vector $V_{wh} \in \mathbb{R}^{L_q}$, which represents the relationship between words and between sentence-word, is generated by calculating the similarity with $G_{eq} \in \mathbb{R}^d$ generated through global average pooling (GAP). This process can be represented as:

$$F_{eq} = \text{Attention}(F_q, F_q, F_q), \tag{1}$$

$$V_{wh} = Sim(F_{eq}, G_{eq}), \tag{2}$$

$$Sim(X, Y) = \frac{XY^\top}{\|X\|\|Y\|}, \tag{3}$$

where $\text{Attention}(Q, K, V) = \text{Softmax}\left(QK^\top/\sqrt{d}\right)V$. By doing so, V_{wh} captures the importance of words by modeling relationships between words and between words and the sentence. After learning the contextual importance by utilizing self-attention, we fine-tune the relative importance of each word in the sentence by comparing it with the global meaning. That is, each element of V_{wh} can be interpreted as a score reflecting how important a specific word is in the query. If the score is high, the word is considered important, otherwise it is considered less important. After that, we rank the words to find the most important N words.

Second, in the scene understanding step, similarities between F_v and F_{eq}, F_c and F_{eq} are calculated to generate video-query similarity scores $A_{vq} \in \mathbb{R}^{L_v \times L_q}$ and caption-query similarity scores $A_{cq} \in \mathbb{R}^{L_v \times L_q}$, computed as:

$$A_{vq} = \frac{P(F_v)\,P(F_{eq})^\top}{\sqrt{d}}, \; A_{cq} = \frac{P(F_c)\,P(F_{eq})^\top}{\sqrt{d}}, \tag{4}$$

where $P(\cdot)$ is a linear projection layer. Then, row-wise softmax is applied to A_{vq} and A_{cq} to obtain A^r_{vq} and A^r_{cq}, capturing the correlation between each clip or caption and all words in the text query. Column-wise softmax is also applied to obtain A^c_{vq}, A^c_{cq}, representing the correlation between a specific word in the text query and all clips or all captions.

Then, the video-to-query features F_{v2q}, caption-to-query features F_{c2q}, and the query-to-video features F_{q2v} and query-to-caption features F_{q2c} are calculated as follows:

$$F_{v2q} = A^r_{vq}F_{eq}, \; F_{c2q} = A^r_{cq}F_{eq}, \tag{5}$$

$$F_{q2v} = A^r_{vq}A^{c\top}_{vq}F_v, \; F_{q2c} = A^r_{cq}A^{c\top}_{cq}F_c. \tag{6}$$

Finally, to maximize interaction between the query and F_v/F_c, we compute the query-related visual features F_{qv} and query-related caption features F_{qc} as follows:

$$\widehat{F}_v = P(F_v \parallel F_{v2q} \parallel F_v \odot F_{v2q} \parallel F_v \odot F_{q2v}), \quad (7)$$

$$\widehat{F}_c = P(F_c \parallel F_{c2q} \parallel F_c \odot F_{c2q} \parallel F_c \odot F_{q2c}), \quad (8)$$

$$F_{qv} = \text{ReLU}(\text{Conv1D}(\widehat{F}_v \parallel F'_{eq})), \quad (9)$$

$$F_{qc} = \text{ReLU}(\text{Conv1D}(\widehat{F}_c \parallel F'_{eq})), \quad (10)$$

where F'_{eq} is sentence-level enhanced query features via a weighted sum of words (Huang et al. 2022). $(\cdot||\cdot)$ indicates concatenation and $\odot$ is Hadamard Product.

Next, we apply cross-attention to F_{qv} and F_{qc} to obtain enhanced visual features $F_{ev} \in \mathbb{R}^{L_v \times d}$ and enhanced caption features $F_{ec} \in \mathbb{R}^{L_v \times d}$, which can be represented as:

$$F_{ev} = \text{Attention}(F_{qv}, F_{qc}, F_{qc}), \quad (11)$$

$$F_{ec} = \text{Attention}(F_{qc}, F_{qv}, F_{qv}). \quad (12)$$

This effectively integrates the two features, allowing complementary visual-textual information to enhance query-based scene understanding.

Ranking-based Filtering Module

We propose a ranking-based filtering module (RFM) to emphasize video clips related to important words in the text query while suppressing unrelated ones. At this time, since the important words in the text query are diverse, our goal is to repeat this process N times based on the priority of the important N words to find query-relevant clips.

As shown in Figure 3, we calculate query-video and query-caption similarity matrix $S_{qv}, S_{qc} \in \mathbb{R}^{L_v \times L_q}$ to measure query relevance. Then, they are combined to obtain fusion similarity matrix S_{qvc}, defined as follows:

$$S_{qvc} = WS_{qv} + (1 - W)S_{qc}, \quad (13)$$

where W is a learnable weight matrix that balances S_{qv} and S_{qc}, dynamically adjusting the relative importance between video clips and captions based on the situation.

After that, the word-highlighted vector V_{wh} in Eq. (4) acts as explicit prior knowledge for iterative clip filtering. The top-N important words in V_{wh} are used to iteratively filter the enhanced visual features F_{ev}. Specifically, in the first iteration ($N = 1$), let i be the position of the word with the highest value in V_{wh}. Then, we extract the i-th column of S_{qvc} as the important word vector $V_s^i \in \mathbb{R}^{L_v}$, where each element indicates the similarity score between each clip and the i-th word in the text query. V_s^i is applied to F_{ev} with residual connection. This process is repeated N times as:

$$x_0 = F_{ev}, \quad F_{fv} = \sum_{j=1}^{N} x_{j-1}(1 + V_s^i). \quad (14)$$

By repeating the process N times, the influence of unnecessary clips is minimized, emphasizing only important clips associated with the similarity of words. The output of the iteration process is the filtered visual features F_{fv}. Finally, F_{eq} from text query, F_{fv} from the video, and F_{ec} from the caption are considered to the prediction head. Note that, we follow the prediction head as (Sun et al. 2024).

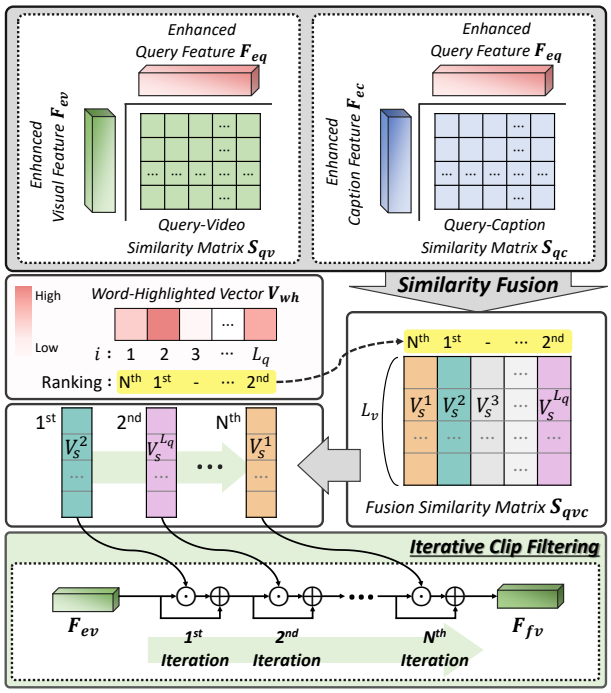

Figure 3: The detailed process of the ranking-based filtering module (RFM). Video clips are iteratively filtered based on the ranking of the most important query tokens.

Modal Alignment Loss

To bridge the inherent gap between text and video arising from their different modalities, we introduce a modal alignment loss inspired by (Sun et al. 2024) that maps them to a shared semantic space. It consists of three losses: (*i*) query-video alignment loss between sentence-level query features and video-level visual features, (*ii*) query-clip alignment loss between sentence-level query features and clip-level visual features within a query-video pair, and (*iii*) caption-clip alignment loss between sentence-level caption features and clip-level visual features within a caption-clip pair. First, the query-video alignment loss $\mathcal{L}_{q\text{-}v}$ is calculated as:

$$\mathcal{L}_{q\text{-}v} = -\frac{1}{B}\sum_{j=1}^{B} \log \frac{\exp(Sim(G_{v_j}, G_{q_j}))}{\sum_{i=1}^{B} \exp(Sim(G_{v_i}, G_{q_j}))}, \quad (15)$$

where B is the batch size, $G_{v_i}, G_{q_i} \in \mathbb{R}^d$ are the i-th global visual and query features, obtained via GAP of F_{v_i} and F_{q_i} respectively. $\mathcal{L}_{q\text{-}v}$ enhances global correlation of similar query-video pairs by separating them from dissimilar pairs.

Next, the query-video similarity matrix S_{qv} is passed through the sigmoid and average pooling to generate G_{qc_i}. Then, the query-clip alignment loss $\mathcal{L}_{q\text{-}c}$ is defined as:

$$\mathcal{L}_{q\text{-}c} = -\sum_{i=1}^{L_v} \left(M_i \log(G_{qc_i}) + (1 - M_i) \log(1 - G_{qc_i})\right), \quad (16)$$

where M_i is a ground-truth mask, which means 1 if the i-th video clip is relevant to the query, and 0 otherwise. G_{qc_i}

Method	Source	MR					HD	
		R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	mAP(Avg.)	mAP	HIT@1
M-DETR (NeurIPS'21) (Lei, Berg, and Bansal 2021)	$\mathcal{V}$	52.89	33.02	54.82	29.40	30.73	35.69	55.60
QD-DETR (CVPR'23) (Moon et al. 2023)	$\mathcal{V}$	62.40	44.98	62.52	39.88	39.86	38.94	62.40
UniVTG (ICCV'23) (Lin et al. 2023)	$\mathcal{V}$	58.86	40.86	57.60	35.59	35.47	38.20	60.96
TR-DETR (AAAI'24) (Sun et al. 2024)	$\mathcal{V}$	64.66	48.96	63.98	43.73	42.62	39.91	63.42
UVCOM (CVPR'24) (Xiao et al. 2024)	$\mathcal{V}$	63.55	47.47	63.37	42.67	43.18	39.74	64.20
Keyword-DETR (AAAI'25) (Um et al. 2025)	$\mathcal{V}$	<u>66.86</u>	<u>51.23</u>	<u>67.73</u>	<u>46.24</u>	<u>45.69</u>	<u>40.94</u>	<u>64.79</u>
Proposed Method	$\mathcal{V}$	**68.09**	**52.20**	**67.81**	**46.74**	**46.54**	**42.24**	**68.22**
UMT (CVPR'22) (Liu et al. 2022)	$\mathcal{V}+\mathcal{A}$	56.23	41.18	53.38	37.01	36.12	38.18	59.99
QD-DETR (CVPR'23) (Moon et al. 2023)	$\mathcal{V}+\mathcal{A}$	63.06	45.10	63.04	40.10	40.19	39.04	62.87
TR-DETR (AAAI'24) (Sun et al. 2024)	$\mathcal{V}+\mathcal{A}$	65.05	47.67	64.87	42.98	43.10	39.90	63.88
UVCOM (CVPR'24) (Xiao et al. 2024)	$\mathcal{V}+\mathcal{A}$	63.81	48.70	64.47	44.01	43.27	39.79	64.79
Keyword-DETR (AAAI'25) (Um et al. 2025)	$\mathcal{V}+\mathcal{A}$	<u>67.77</u>	<u>50.52</u>	**68.30**	<u>45.88</u>	<u>45.52</u>	<u>41.15</u>	<u>65.82</u>
Proposed Method	$\mathcal{V}+\mathcal{A}$	**68.87**	**52.27**	<u>68.09</u>	**46.55**	**46.23**	**42.36**	**69.78**

Table 1: Results of moment retrieval and highlight detection experiments on the QVHighlights *test* set using video only ($\mathcal{V}$) and video and audio together ($\mathcal{V}+\mathcal{A}$). Best/second-best results are marked in **Bold**/<u>underlined</u>.

is the similarity score between the global query features and the i-th clip-level visual features. Through $\mathcal{L}_{q\text{-}c}$, relevant and irrelevant video clips are differentiated based on the query.

Finally, the caption-clip alignment loss $\mathcal{L}_{c\text{-}c}$ is defined as:

$$\mathcal{L}_{c\text{-}c} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\sum_{j=1}^{L_v}\exp(Sim(F_{v_{kj}}, F_{c_{kj}}))}{\sum_{i=1}^{B}\sum_{j=1}^{L_v}\exp(Sim(F_{v_{ij}}, F_{c_{kj}}))}, \tag{17}$$

where $F_{v_{kj}}, F_{c_{kj}}$ denotes the visual and caption features for the j-th clip of the k-th video. This improves the correlation of similar caption-clip pairs in a video and better separates dissimilar pairs or similar pairs with different meanings.

Finally, the modal alignment loss $\mathcal{L}_{ma}$ is formulated as:

$$\mathcal{L}_{ma} = \lambda_{q\text{-}v}\mathcal{L}_{q\text{-}v} + \lambda_{q\text{-}c}\mathcal{L}_{q\text{-}c} + \lambda_{c\text{-}c}\mathcal{L}_{c\text{-}c}, \tag{18}$$

where $\lambda_{q\text{-}v}$, $\lambda_{q\text{-}c}$ and $\lambda_{c\text{-}c}$ are balancing weights. Finally, the total training loss function is formulated as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_{mr} + \mathcal{L}_{hd} + \mathcal{L}_{ma}, \tag{19}$$

where $\mathcal{L}_{mr}$ and $\mathcal{L}_{hd}$ denote the MR/HD losses from (Moon et al. 2023; Sun et al. 2024).

Experiments

Datasets and Evaluation Metrics

Dataset. We use three benchmark datasets. **QVHighlights** dataset (Lei, Berg, and Bansal 2021) contains 10,148 content-rich YouTube videos, paired with text queries that identifies a specific highlight moment. It includes annotations for both moment retrieval (MR) and highlight detection (HD). Test annotations are hidden, and results are evaluated via the CodaLab server. **TVSum** dataset (Song et al. 2015) includes 50 videos across 10 categories for HD. Following (Moon et al. 2023), 80% of the dataset is used for training, and 20% is used for testing. **Charades-STA** dataset (Gao et al. 2017) contains 9,848 videos of indoor daily activities and 16,128 human-annotated text queries. Following (Moon et al. 2023), 12,408 samples are used for training and 3,720 samples for testing.

Evaluation Metrics. We follow the evaluation metrics used in previous works (Sun et al. 2024; Xiao et al. 2024) for fair comparison. For QVHighlights, we measure Recall@1 (R1) at IoU thresholds of 0.5 and 0.7 for MR, and compute the average mAP (mAP@Avg) for IoU thresholds sampled at 0.05 intervals from 0.5 to 0.95. We also evaluate mAP at specific thresholds of 0.5 and 0.75 for more detailed performance comparison. For HD, we use the average precision (mAP) and HIT@1, which represents the hit rate of the highest-scoring clip. For TVSum, we evaluate HD using the top-5 mAP values. For Charades-STA, we measure R1 at IoU thresholds of 0.5 and 0.7 for MR.

Implementation Details

Following (Sun et al. 2024), we extracted video, query, caption, and audio features using pre-trained models. For video, SlowFast and CLIP (Radford et al. 2021) were used for QVHighlights; VGG (Simonyan and Zisserman 2014) and SlowFast+CLIP for Charades-STA; and I3D for TVSum. Query and caption features were extracted using CLIP for QVHighlights and TVSum, and GLoVe for Charades-STA. All audio features were obtained using PANN (Kong et al. 2020) trained on the AudioSet (Gemmeke et al. 2017).

All experiments were performed on an NVIDIA RTX 3090, with $\lambda_{q\text{-}v} = 0.3$, $\lambda_{q\text{-}v} = 0.5$, $\lambda_{c\text{-}c} = 1.5$ in Eq. (18). Other training settings followed TR-DETR (Sun et al. 2024)

Experimental Results

Results on the QVHighlights. Table 1 shows the experimental results for MR and HD on the QVHighlights *test* set. We compared state-of-the-art methods (Lei, Berg, and Bansal 2021; Moon et al. 2023; Lin et al. 2023; Sun et al. 2024; Xiao et al. 2024; Um et al. 2025; Liu et al. 2022). Our method shows superior performance across all metrics when using only video ($\mathcal{V}$). With the concatenated video and audio setting ($\mathcal{V}+\mathcal{A}$), it still outperforms other methods on most metrics. These results

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Avg.
LIM-S (CVPR'19) (Xiong et al. 2019)	55.9	42.9	61.2	54.0	60.3	47.5	43.2	66.3	69.1	62.6	56.3
Trailer (ECCV'20) (Wang et al. 2020)	61.3	54.6	65.7	60.8	59.1	70.1	58.2	64.7	65.6	68.1	62.8
SL-Module (ICCV'21) (Xu et al. 2021)	86.5	68.7	74.9	86.2	79.0	63.2	58.9	72.6	78.9	64.0	73.3
UMT† (CVPR'22) (Liu et al. 2022)	87.5	81.5	88.2	78.8	81.5	87.0	76.0	86.9	84.4	79.6	83.1
QD-DETR (CVPR'23) (Moon et al. 2023)	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7	85.0
UniVTG (ICCV'23) (Lin et al. 2023)	83.9	85.1	89.0	80.1	84.6	87.0	70.9	91.7	73.5	69.3	81.0
TR-DETR (AAAI'24) (Sun et al. 2024)	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	**81.6**	88.1
UVCOM (CVPR'24) (Xiao et al. 2024)	87.6	91.6	91.4	86.7	86.9	86.9	76.9	92.3	87.4	75.6	86.3
TaskWeave (CVPR'24) (Yang et al. 2024)	88.2	90.8	93.3	**87.5**	87.0	82.0	80.9	92.9	89.5	81.2	87.3
Keyword-DETR (AAAI'25) (Um et al. 2025)	**89.9**	93.8	94.4	85.9	89.2	89.4	**81.5**	92.6	**90.1**	80.6	88.7
Proposed Method	**89.9**	**94.1**	**95.0**	**87.5**	**89.7**	**90.4**	80.6	**93.3**	89.9	**81.6**	**89.2**

Table 2: Results on highlight detection experiments on the TVSum. † means training with audio modality. Best/second-best results are marked in **Bold**/underlined.

Method	Feat	R1@0.5	R1@0.7
UMT† (CVPR'22) (Liu et al. 2022)	VGG	48.31	29.25
QD-DETR (CVPR'23) (Moon et al. 2023)	VGG	52.77	31.13
TR-DETR (AAAI'24) (Sun et al. 2024)	VGG	53.47	30.81
TaskWeave (CVPR'24) (Yang et al. 2024)	VGG	56.51	33.66
Keyword-DETR (AAAI'25) (Um et al. 2025)	VGG	54.89	31.97
Proposed Method	VGG	**61.51**	**37.58**
QD-DETR (CVPR'23) (Moon et al. 2023)	SF+C	57.31	32.55
UniVTG (ICCV'23) (Lin et al. 2023)	SF+C	58.01	35.65
TR-DETR (AAAI'24) (Sun et al. 2024)	SF+C	57.61	33.52
UVCOM (CVPR'24) (Xiao et al. 2024)	SF+C	59.25	36.64
Keyword-DETR (AAAI'25) (Um et al. 2025)	SF+C	**61.08**	37.89
Proposed Method	SF+C	60.97	**38.52**

Table 3: Results on moment retrieval experiments on the Charades-STA. † indicates training with audio modality. Best/second-best results are marked in **Bold**/underlined.

demonstrate the effectiveness of our method in identifying important query words and understanding video content.

Results on the TVSum. We evaluated the HD performance on the TVSum dataset. As shown in Table 2, the overall average performance (Avg.) of our method still outperforms the existing methods in almost all categories. This highlights that our method is still effective approach for HD task.

Results on the Charades-STA. As shown in Table 3, when evaluating the MR task on the Charades-STA dataset, our method outperforms state-of-the-art models in most cases, regardless of the video feature type (i.e., VGG or Slow-Fast+CLIP(SF+C)). These experimental results demonstrate that proposed method, which effectively understands both query text and video clips, remains effective in MR task.

Ablation Study

We perform an ablation study on the QVHighlights *val* set to see the effectiveness of each module in our method. As shown in Table 4, both FEM and RFM contribute to improved performance compared to the baseline, and their combination achieves the best results. Adding captions further enhances scene understanding and overall performance.

Cap.	FEM	RFM	MR			HD	
			R1@0.5	R1@0.7	mAP(Avg.)	mAP	HIT@1
-	-	-	66.13	49.74	43.24	40.11	64.77
-	✓	-	66.32	49.81	43.53	40.80	65.10
-	-	✓	67.55	49.61	44.38	41.29	65.81
-	✓	✓	69.42	51.81	45.56	41.32	66.06
✓	-	-	69.03	52.71	46.34	42.23	67.94
✓	✓	-	69.16	53.81	46.81	42.86	69.61
✓	-	✓	69.42	53.87	47.43	42.29	68.00
✓	✓	✓	**70.00**	**55.68**	**48.14**	**43.24**	**71.23**

Table 4: Effect of our components (caption usage (Cap.), feature enhancement module (FEM), and ranking-based filtering module (RFM)) on the QVHighlights *val* set.

Visualization Results

Qualitative Comparisons. We compared our method with state-of-the-art approaches (Sun et al. 2024; Xiao et al. 2024; Um et al. 2025) on the QVHighlights *val* set. As shown in Figure 4, ours outperformed existing ones, demonstrating its superiority in achieving more accurate MR and HD predictions.

Results of the Ranking Important Words. Figure 5 shows the top-N words from the word-highlighted vector V_{wh}, obtained during the word highlighting step of RFM. It demonstrates that the model prioritizes words strongly connected to other components or central to actions or objects, which effectively helps retrieve relevant clips.

Discussions

Effect of the Caption Information. While leveraging captions with MLLMs is a promising recent trend, previous MR/HD studies have not yet explored integrating visual-text multimodal information to enhance video understanding. In contrast, we are the first to incorporate multimodal information in MR/HD tasks. For fair comparison, Table 5 reports results for adding MLLM-generated captions to SOTA models. These results demonstrate that the superiority of this study lies not in the captions, but in effectively identifying and prioritizing important query words.

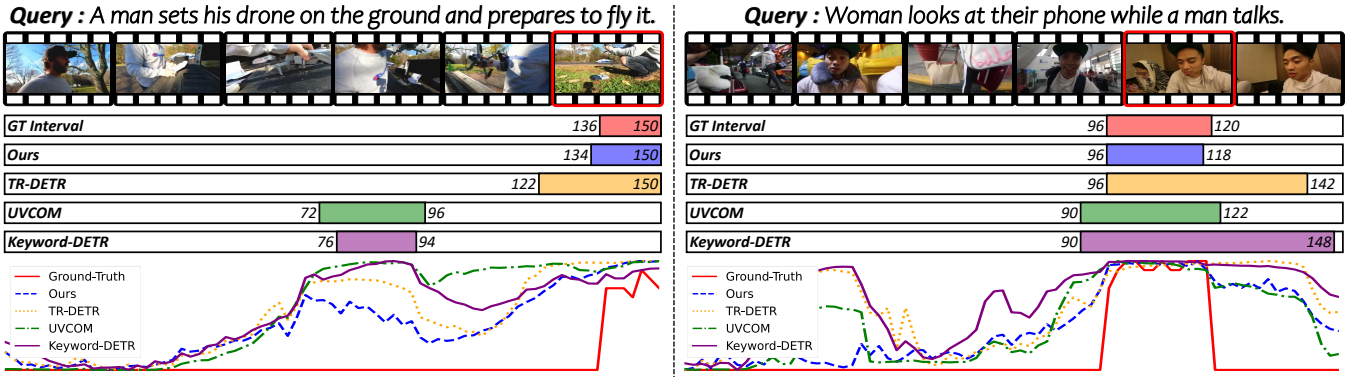

Figure 4: Visualization comparison of moment retrieval (MR) and highlight detection (HD) for the QVHighlights *val* set.

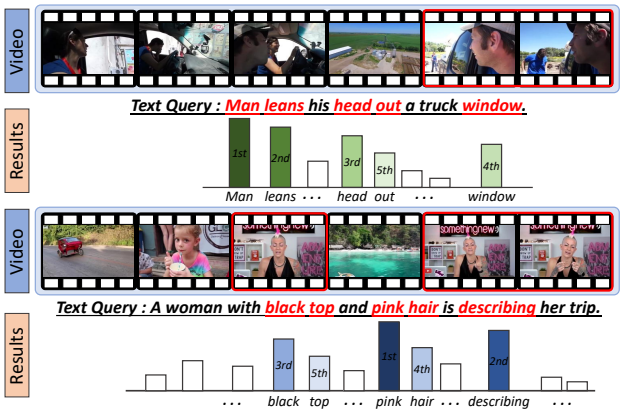

Figure 5: Visualization results of the prioritized N important words on the QVHighlights *val* set ($N = 5$).

Method	MR			HD	
	R1 @0.5	R1 @0.7	mAP (Avg.)	mAP	HIT @1
TR-DETR (AAAI'24)	68.90	51.94	46.60	42.15	68.13
UVCOM (CVPR'24)	68.71	53.23	47.68	41.95	68.32
Keyword-DETR (AAAI'25)	68.32	53.35	47.73	41.97	69.55
Proposed Method	**70.00**	**55.68**	**48.14**	**43.24**	**71.23**

Table 5: Effect of the caption information on the QVHighlights *val* set. InternVL2 is used for caption extraction.

Effect of Important Word-based Iterative Filtering. We evaluate the effect of the number of iterations in RFM on the QVHighlights *val* set, as shown in Table 6. Increasing the number of iterations leads to performance gains by refining clip selection, with the best performance at $N = 5$. At $N = 7$, it slightly drops due to over-filtering but still outperforms the no-iteration baseline.

MLLM Variations. Table 7 shows results using two recent MLLMs: LLaVA (Liu et al. 2023) and InternVL2 (Chen et al. 2024). InternVL2 performs best overall, while LLaVA remains competitive and exceeds InternVL2 on R1@0.5.

# of Iter	MR			HD	
	R1@0.5	R1@0.7	mAP(Avg.)	mAP	HIT@1
0	69.16	53.81	46.81	42.86	69.61
1	69.48	54.13	47.10	42.91	70.06
3	69.74	54.45	47.30	42.98	70.90
5	**70.00**	**55.68**	**48.14**	**43.24**	**71.23**
7	**70.00**	53.87	47.09	42.76	69.29

Table 6: Effect of the number of filtering iterations in the ranking-based filtering module on the QVHighlights *val* set.

MLLMs	MR			HD	
	R1@0.5	R1@0.7	mAP(Avg.)	mAP	HIT@1
LLaVA	**70.71**	54.90	47.69	42.80	69.68
InternVL2	70.00	**55.68**	**48.14**	**43.24**	**71.23**

Table 7: Effect of MLLM variants on QVHighlights *val* set.

These results highlight the robustness of our method and its effectiveness in leveraging scene understanding for query-aware filtering across different MLLMs.

Limitations. We incorporated MLLMs into MR/HD tasks to enhance scene understanding with richer external knowledge. However, this increases inference time and the number of parameters. Our future work will focus on reducing reliance on MLLMs during inference and finding ways to utilize caption knowledge without directly using captions.

Conclusion

We propose an important word-aware clip filtering framework to imporve MR and HD tasks by focusing on the most relevant information in video content. We adopt MLLM to fully leverage the knowledge of the captions from each video clip to further understand the video. Our approach includes a feature enhancement module to identify/prioritize important words and enhance the semantic understanding of video clips, while a ranking-based filtering module iteratively refines video clips based on their relevance to the query. This results in improved performance on MR and HD tasks.

Acknowledgments

This work was supported by the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00252391), and by IITP grant funded by the Korea government (MSIT) (No. RS-2022-00155911: Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University), No. RS-2025-25442384, IITP-2023-RS-2023-00266615: Convergence Security Core Talent Training Business Support Program, No. RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Apostolidis, E.; Adamantidou, E.; Metsai, A. I.; Mezaris, V.; and Patras, I. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11): 1838–1863.

Badamdorj, T.; Rochan, M.; Wang, Y.; and Cheng, L. 2022. Contrastive learning for unsupervised video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14042–14052.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Buch, S.; Eyzaguirre, C.; Gaidon, A.; Wu, J.; Fei-Fei, L.; and Niebles, J. C. 2022. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2917–2927.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Foo, L. G.; Gong, J.; Fan, Z.; and Liu, J. 2023. System-status-aware adaptive network for online streaming video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10514–10523.

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

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 776–780. IEEE.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2018. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337*.

Huang, J.; Jin, H.; Gong, S.; and Liu, Y. 2022. Video activity localisation with uncertainties in temporal boundary. In *European Conference on Computer Vision*, 724–740. Springer.

Just, M. A.; and Carpenter, P. A. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4): 329.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2880–2894.

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. 2023. 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, K.; Guo, D.; and Wang, M. 2021. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1902–1910.

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

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022. 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*, 3042–3051.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23023–23033.

Mun, J.; Cho, M.; and Han, B. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10810–10819.

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

Rodriguez, C.; Marrese-Taylor, E.; Saleh, F. S.; Li, H.; and Gould, S. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2464–2473.

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

Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5179–5187.

Sun, H.; Zhou, M.; Chen, W.; and Xie, W. 2024. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4998–5007.

Sun, M.; Farhadi, A.; and Seitz, S. 2014. Ranking domain-specific highlights by analyzing edited videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 787–802. Springer.

Sun, X.; Wang, X.; Gao, J.; Liu, Q.; and Zhou, X. 2022. You need to read again: Multi-granularity perception network for moment retrieval in videos. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1022–1032.

Um, S. J.; Kim, D.; Lee, S.; and Kim, J. U. 2025. Watch Video, Catch Keyword: Context-aware Keyword Attention for Moment Retrieval and Highlight Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7473–7481.

Vo, K.; Yamazaki, K.; Nguyen, P. X.; Nguyen, P.; Luu, K.; and Le, N. 2022. Contextual explainable video representation: Human perception-based understanding. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, 1326–1333. IEEE.

Wang, L.; Liu, D.; Puri, R.; and Metaxas, D. N. 2020. Learning trailer moments in full-length movies with co-contrastive attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, 300–316. Springer.

Wei, F.; Wang, B.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022. Learning pixel-level distinctions for video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3073–3082.

Xiao, Y.; Luo, Z.; Liu, Y.; Ma, Y.; Bian, H.; Ji, Y.; Yang, Y.; and Li, X. 2024. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18709–18719.

Xiong, B.; Kalantidis, Y.; Ghadiyaram, D.; and Grauman, K. 2019. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1258–1267.

Xu, M.; Wang, H.; Ni, B.; Zhu, R.; Sun, Z.; and Wang, C. 2021. Cross-category video highlight detection via set-based learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7970–7979.

Yang, J.; Wei, P.; Li, H.; and Ren, Z. 2024. Task-Driven Exploration: Decoupling and Inter-Task Feedback for Joint Moment Retrieval and Highlight Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18308–18318.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.