Watch Video, Catch Keyword: Context-aware Keyword Attention for Moment Retrieval and Highlight Detection

Sung Jin Um[1], Dongjin Kim[1], Sangmin Lee[2,†], Jung Uk Kim[1,†]

[1]Kyung Hee University, Yong-in, South Korea
[2]Sungkyunkwan University, Seoul, South Korea
{sungzin1, rlaehdwls310, ju.kim}@khu.ac.kr, sangmin.lee@skku.edu

Abstract

The goal of video moment retrieval and highlight detection is to identify specific segments and highlights based on a given text query. With the rapid growth of video content and the overlap between these tasks, recent works have addressed both simultaneously. However, they still struggle to fully capture the overall video context, making it challenging to determine which words are most relevant. In this paper, we present a novel Video Context-aware Keyword Attention module that overcomes this limitation by capturing keyword variation within the context of the entire video. To achieve this, we introduce a video context clustering module that provides concise representations of the overall video context, thereby enhancing the understanding of keyword dynamics. Furthermore, we propose a keyword weight detection module with keyword-aware contrastive learning that incorporates keyword information to enhance fine-grained alignment between visual and textual features. Extensive experiments on the QVHighlights, TVSum, and Charades-STA benchmarks demonstrate that our proposed method significantly improves performance in moment retrieval and highlight detection tasks compared to existing approaches.

Introduction

With the exponential growth of video content, precise video moment retrieval and highlight detection have become crucial (Snoek, Worring et al. 2009; Apostolidis et al. 2021). Video moment retrieval enables users to find specific segments within videos based on natural language queries (Anne Hendricks et al. 2017), while highlight detection helps extract the most engaging parts from long-form videos (Gygli et al. 2014). These technologies enhance user experience and productivity across various applications, including video searching, video editing, social media, and e-learning, by enabling quick and accurate access to relevant content.

Extensive research has been conducted on moment retrieval (Gao et al. 2017; Hendricks et al. 2018; Xiao et al. 2021; Sun et al. 2022) and highlight detection (Sun, Farhadi, and Seitz 2014; Xu et al. 2021; Wei et al. 2022; Badamdorj et al. 2022) as separate tasks. However, with the introduction of Moment-DETR and the QVHighlights dataset (Lei, Berg, and Bansal 2021), which allows for the simultaneous execution of these tasks, new studies (Liu et al. 2022; Moon et al. 2023; Xiao et al. 2024; Sun et al. 2024) have emerged that aim to address moment retrieval and highlight detection based on text queries concurrently. Following the Moment-DETR (Lei, Berg, and Bansal 2021), UMT (Liu et al. 2022) utilizes audio-visual multi-modal learning. TR-DETR (Sun et al. 2024) leverages the reciprocal relationship between two tasks, refining visual features through textual guidance to enhance both tasks simultaneously. UVCOM (Xiao et al. 2024) introduces integration module for progressive intra- and intermodality interaction across multi-granularity.

Despite these advancements, existing methods often fail to capture the dynamic importance of keywords within the context of the video, which is crucial for accurate moment retrieval and highlight detection. As illustrated in Figure 1, the importance of each word in a text query can vary de-

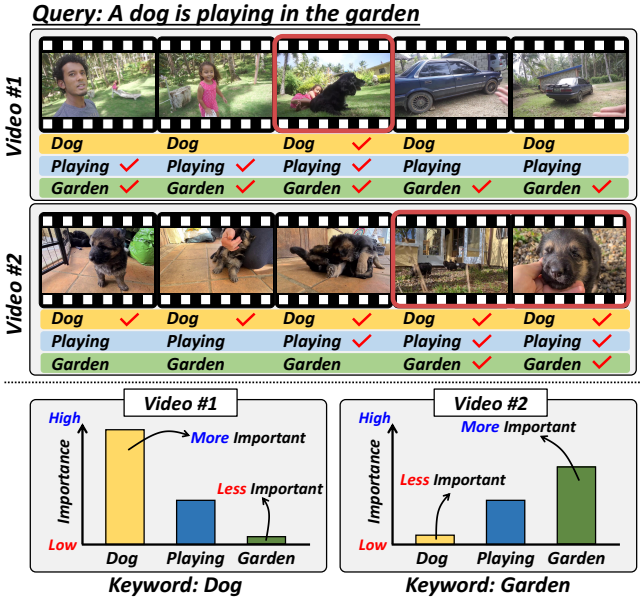

Figure 1: Text keywords can vary by video context. The less frequently a word appears in the video clip, the more important it becomes within the text query. In Video #1, '*dog*' is important, while in Video #2, '*garden*' is important.

[†]Corresponding author.

pending on the video content. For instance, when we consider a text query *'A dog is playing in the garden'*, the importance of words such as *'dog'* or *'garden'* can shift based on the predominant scenes in the video. In Video #1, where most scenes contain *'garden'*, the word *'dog'* is more critical than *'garden'* for specifying the desired video segment. Conversely, in Video #2, *'dog'* is predominantly *'playing'* indoors, making *'garden'* more essential than *'dog'* for identifying the relevant video segment. This indicates that the importance of words in a text query can vary significantly depending on the video context. Therefore, it is necessary to consider such keyword variations for effective moment retrieval and highlight detection. However, existing methods fall short in addressing this keyword variation, as they rely on text features extracted independently of the video context, failing to capture the dynamic importance of words relative to the visual content.

In this paper, we propose a Video Context-aware Keyword Attention module that effectively captures keyword variations by considering the overall video context. Our approach addresses two main challenges: (*i*) how to effectively encode the overall context of a video to capture keyword variation, and (*ii*) how to capture and utilize desired text keywords within their relevant video contexts.

First, effective keyword extraction requires a comprehensive understanding of the overall context of the video. To address this, we tackle the challenge (*i*) by introducing a video context clustering module that leverages temporally-weighted clustering to group similar video scenes. This approach allows our model to grasp the high-level flow and structure of the video. The resulting cluster assignments provide a concise representation of the overall video context and are leveraged to understand keyword dynamics. Furthermore, since these clustered features contain information about scene changes, they are further used as additional hints for moment retrieval and highlight detection.

To address the challenge (*ii*), we propose a keyword weight detection module. This module recognizes less frequently occurring but important words in the text query and calculates the similarity between clustered video features and text features to generate a keyword weight vector. This vector captures information about the important words in the text query within the video context, allowing our framework to adjust the keywords based on the overall video context dynamically. Based on this, we introduce keyword-aware contrastive learning to incorporate keyword weights and facilitate a fine-grained alignment between visual and text features. As a result, our method allows for accurate moment retrieval and highlight detection.

The major contributions of our paper are as follows:

- We propose a video context-aware keyword attention module to capture keyword variations by considering overall context of the video for effective moment retrieval and highlight detection. To the best of our knowledge, this is the first work to address this aspect in video moment retrieval and highlight detection tasks.
- We introduce keyword-aware contrastive learning to integrate keyword weight information, enhancing the fine-grained alignment between visual and text features. This approach improves the ability of model to understand the relationship between textual queries and video content.
- Experimental results on QVHighlights, TVSum, and Charades-STA demonstrate the effectiveness of our method for moment retrieval and highlight detection.

Related Works

Moment Retrieval

The connection between visual and language cues has become important in machine learning (Lee et al. 2022b; Park et al. 2024b; Lee et al. 2024). Moment retrieval aims to locate relevant moments in a video based on a natural language query (Gao et al. 2017). This task is typically approached using either proposal-based or proposal-free methods. The proposal-based methods (Gao et al. 2017; Hendricks et al. 2018; Xiao et al. 2021; Sun et al. 2022) generate candidate proposals and rank them by matching scores. On the other hand, the proposal-free methods (Yuan, Mei, and Zhu 2019; Mun, Cho, and Han 2020; Rodriguez et al. 2020; Li, Guo, and Wang 2021) directly regress the start and end timestamps through video-text interaction.

Highlight Detection

Highlight detection identifies the most significant parts of a video, which might not necessarily be tied to specific textual queries. Early highlight detection approaches are unimodal, assessing the salience score of video clips without external textual data (Sun, Farhadi, and Seitz 2014; Xu et al. 2021; Wei et al. 2022; Badamdorj et al. 2022). However, as user preferences have increasingly influenced content consumption, integrating text queries has become common to tailor the detection process to individual user needs (Dagtas and Abdel-Mottaleb 2004; Kudi and Namboodiri 2017; Lei, Berg, and Bansal 2021). It has been demonstrated that audio cues provide complementary information to visual features (Lee et al. 2022a; Park et al. 2024a; Kim et al. 2024).

Traditionally, moment retrieval and highlight detection are addressed separately, but recent work has explored their joint learning. MomentDETR (Lei, Berg, and Bansal 2021) introduces the QVHighlights dataset to facilitate joint learning of moment retrieval and highlight detection, proposing a DETR-based model. UMT (Liu et al. 2022) proposes adopting audio, visual, and text content for query generation to improve query quality. QD-DETR (Moon et al. 2023) further leverages textual information by incorporating negative relationship learning between video-text pairs. UVCOM (Xiao et al. 2024) and TR-DETR (Sun et al. 2024) integrate the specialties of moment retrieval and highlight detection to achieve a comprehensive understanding.

Despite these advancements, many existing methods do not capture the overall context of the video. Capturing the overall context is essential for accurate moment retrieval and highlight detection as it provides a comprehensive understanding of the content and narrative flow. To this end, we present a video context-aware keyword attention module that understands the video context and identifies keywords between the video and text query.

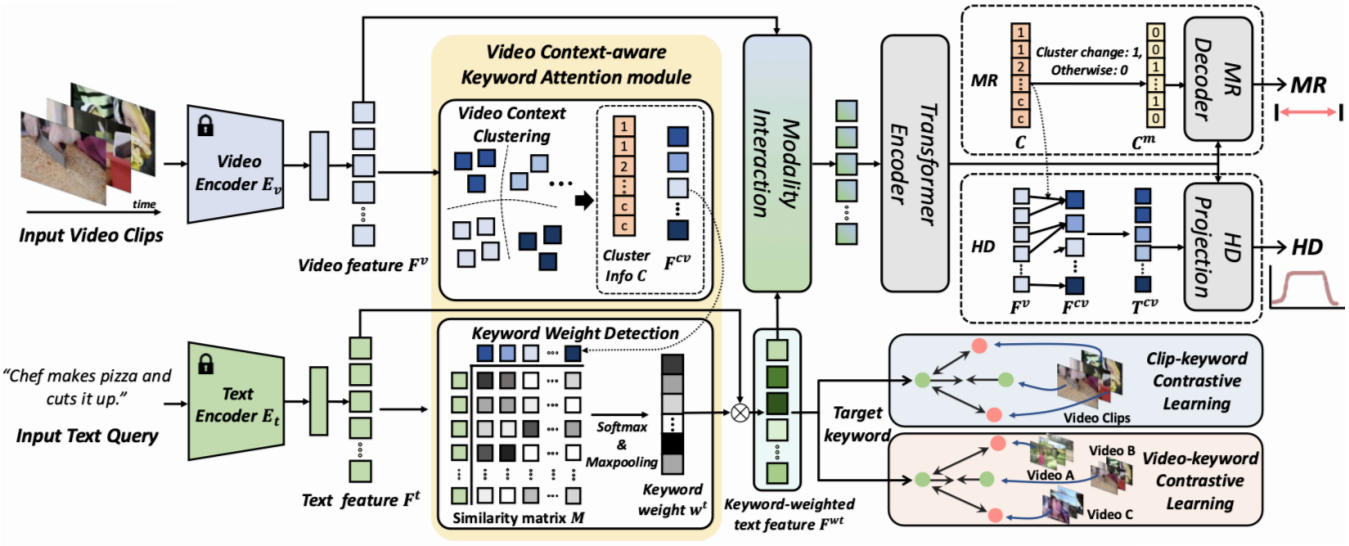

Figure 2: Overall configuration of our moment retrieval and highlight detection. $\otimes$ indicates element-wise multiplication.

Proposed Method

Figure 2 shows the overall architecture of our framework. Similar to previous works (Moon et al. 2023; Sun et al. 2024; Xiao et al. 2024), we employ a two-stream network to extract video and text features. Input video of L clips and a text query with N words pass through each modal encoder (*i.e.,* video encoder and text encoder) and three-layer feed-forward network to generate video features $\mathbf{F}^v \in \mathbb{R}^{L\times d}$ and text features $\mathbf{F}^t \in \mathbb{R}^{N\times d}$, respectively. To capture the overall flow of the video, $\mathbf{F}^v$ is passed through a video context clustering module to generate c clustered video features $\mathbf{F}^{cv} \in \mathbb{R}^{c\times d}$. Subsequently, to identify video-related keywords in the text query, a keyword weight detection module calculates the similarity between clustered video features and text features, resulting in a keyword-weighted text feature $\mathbf{F}^{wt}$. We then perform modality interaction between the video features and keyword-weighted text features, which is processed through a transformer encoder. Finally, we utilize the video-contextual information from $\mathbf{F}^{cv}$ and cluster information $\mathbf{C}$, along with the transformer-encoded feature, to conduct moment retrieval and highlight detection. More details are in the following subsections.

Video Context-aware Keyword Attention Module

To understand the entire video sequence and extract relevant text keywords corresponding to the video content, we propose a Video Context-aware Keyword Attention Module. As shown in Figure 2, our video context-aware keyword attention module consists of two steps: (1) video context clustering and (2) keyword weight detection.

(1) Video Context Clustering. We propose a video context clustering module to cluster video clips, capturing the overall context of the video and identifying each scene. To account for the temporal order of the video, we adopt a temporally-weighted FINCH algorithm (Sarfraz et al. 2021). With video features $\mathbf{F}^v \in \mathbb{R}^{L\times d}$ as input, the algorithm clusters video clips based on their adjacency relation and merges them hierarchically. Consequently, the most similar clips are grouped into c clusters $(c_1, ...c_c)$ based on their relations. The output is a cluster information vector $\mathbf{C} = \{C_i\}_{i=1}^{L} \in \mathbb{R}^L$, where each element $C_i \in \{c_1, ..., c_c\}$ indicates the cluster assignment value for each video clip. Then, the clustered video features $\mathbf{F}^{cv}$ are generated by averaging the information within each cluster.

(2) Keyword Weight Detection. Aligning video and text features is essential for moment retrieval and highlight detection tasks. However, direct interaction between original video and text features can lead to information loss due to the misalignment (Xu, Zhu, and Clifton 2023). Particularly in video moment retrieval and highlight detection tasks based on specific text queries, the emphasis on certain text keywords varies depending on the overall video context. Consider the text query *"Chef makes pizza and cuts it up"* in Figure 2. From the perspective of moment retrieval and highlight detection, the word *'chef'* might be less important if it appears consistently throughout the video. Instead, a sudden appearance of *'pizza'* or the action of *'cuts'* could be more significant. Conversely, if a machine is making pizza and a chef suddenly appears and makes pizza, *'chef'* becomes a crucial keyword. As keyword importance changes with video content, we conduct keyword weight detection to identify the most important keyword related to the video.

To this end, we calculate cosine similarity matrix $M \in \mathbb{R}^{N\times c}$ between the text feature $\mathbf{F}^t$ and the clustered video feature $\mathbf{F}^{cv}$ (see Figure 2). We then apply each column-wise softmax and max-pooling to obtain a keyword weight vector $w^t \in \mathbb{R}^N$. Higher values in the keyword weight vector indicate words that are strongly associated with only specific clusters in the video, while lower values are associated with most clusters similarly. Finally, we multiply w^t with

the original text feature $\mathbf{F}^t$ to generate the keyword-weighted text feature $\mathbf{F}^{wt} \in \mathbb{R}^{N\times d}$, which is be represented as:

$$M = \frac{\mathbf{F}^t\, \mathbf{F}^{cv\top}}{||\mathbf{F}^t||\,||\mathbf{F}^{cv}||}, \tag{1}$$

$$w^t = \text{MaxPooling}(\text{Softmax}(M/\tau)), \tag{2}$$

$$\mathbf{F}^{wt} = w^t \mathbf{F}^t, \tag{3}$$

where τ is a temperature hyper-parameter. This keyword-weighted text feature $\mathbf{F}^{wt}$ emphasizes contextually important words in the query, enhancing video-text alignment and improving performance in moment retrieval and highlight detection tasks.

Video Contextual MR/HD Prediction

Moment retrieval (MR) and highlight detection (HD) are two crucial tasks in video understanding. Moment retrieval aims to localize the center coordinates and duration of moments related to a given text query, while highlight detection generates a saliency score distribution across the entire video. To improve the effectiveness of these tasks, we utilize the video-contextual information $\mathbf{C}$ and $\mathbf{F}^{cv}$ generated by the video context clustering module in each prediction head.

In the context of moment retrieval, transition points between clusters are crucial. These points often correspond to scene changes in the video, providing valuable information for the moment retrieval task. By leveraging the cluster information vector $\mathbf{C}$, we create a binary context change vector $\mathbf{C}^m \in \mathbb{R}^L$ that encapsulates information about these transitions. In this vector, we assign a value of 1 if the cluster number changes between the i-th and (i+1)-th frame (*e.g.,* $c_j \rightarrow c_{j+1}$), and 0 otherwise. Then, following previous works (Moon et al. 2023), we employ a standard transformer decoder structure for moment retrieval head. However, instead of the traditional approach, we use $\mathbf{C}^m$ as the initial embedding for the learnable anchors, which helps the MR decoder better focus on scene transition points in the video, potentially leading to more accurate moment identification.

For highlight detection, we focus on the representative values of each cluster obtained through our clustering approach. These values provide information about the best representation of each scene context. To obtain this information, we compute a cosine similarity between each video clip feature $\mathbf{F}^v$ and the average information of the cluster it belongs to, which we extract from $\mathbf{F}^{cv}$ using $\mathbf{C}$. This similarity computation results in $\mathbf{C}^h \in \mathbb{R}^L$, which indicates how well each clip represents the information of its cluster. To generate saliency score distribution for highlight detection, we use two groups of single fully connected layers for linear projection to calculate the saliency score, following (Moon et al. 2023). As input to this process, we use a context-aware video token $T^{cv} \in \mathbb{R}^{L\times(d+1)}$, which is created by concatenating $\mathbf{C}^h$ with the video token $T^v \in \mathbb{R}^{L\times d}$ obtained from passing through a transformer encoder. The predicted saliency scores $\mathbf{S}$ are then computed using the following equation:

$$\mathbf{S} = \frac{T^s w^{s\top} \cdot T^{cv} w^{cv\top}}{\sqrt{p}}, \tag{4}$$

where $T^s \in \mathbb{R}^d$ is the randomly initialized input-adaptive saliency token, $w^s \in \mathbb{R}^{p\times d}$ and $w^{cv} \in \mathbb{R}^{p\times(d+1)}$ are learnable parameters, and p is the projection dimension.

Keyword-aware Contrastive Loss

To enhance the alignment between text query features and video features by leveraging the overall flow of the video, we introduce keyword-aware contrastive loss. This loss is composed of two components: a clip-keyword contrastive loss and a video-keyword contrastive loss. The clip-keyword contrastive loss focuses on intra-video relationships between text queries and visual features of each clip, while the video-keyword contrastive loss addresses inter-video relationships across the dataset.

Clip-keyword Contrastive Loss. Existing methods (Xiao et al. 2024; Sun et al. 2024) typically construct loss functions that bring the clip features of ground-truth moments closer to the text query features while pushing background clip features away. However, as illustrated in Figure 1, even background clips considered irrelevant to the text query may still have high relevance to specific words in the text query. In such cases, the feature of background clips can be misrepresented through contrastive loss from existing methods. To address this, we utilize the keyword weight vector w^t to emphasize keywords in advance. This approach enables more robust alignment between the clip features $\mathbf{F}^v$ and the keyword-weighted text features $\mathbf{F}^{wt} = w^t\mathbf{F}^t$ of Eq.(3). We formulate the clip-keyword contrastive loss $\mathcal{L}_{ck}$ as follows:

$$\mathbf{G}^{wt} = \text{MeanPooling}(\mathbf{F}_i^{wt}), \tag{5}$$

$$\mathcal{L}_{ck} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{j\in R_i} exp(Sim(\mathbf{F}_i^{v_{(j)}}, \mathbf{G}_i^{wt}))}{\sum_{j=1}^{L} exp(Sim(\mathbf{F}_i^{v_{(j)}}, \mathbf{G}_i^{wt}))}, \tag{6}$$

$$Sim(A,B) = \frac{AB^\top}{||A||\,||B||}, \tag{7}$$

where $\mathbf{G}_i^{wt} \in \mathbb{R}^d$ is average of word-level text feature, R_i denotes relevant ground-truth clips in the *i*-th video and B indicates the batch number. The $\mathcal{L}_{ck}$ maximizes the learning effect of essential central information, thereby enabling more accurate moment retrieval and highlight detection.

Video-keyword Contrastive Loss. Extending beyond single video contexts, we propose a global contrastive loss, called video-keyword contrastive loss, to operate across the entire dataset. Unlike existing methods (Xiao et al. 2024; Sun et al. 2024) that use unweighted global information from video-text pairs, we incorporate keyword-weighted text features $\mathbf{F}^{wt}$ to obtain a better global representation by utilizing the global information of relevant videos and keyword-weighted text queries. We define the video-keyword contrastive loss $\mathcal{L}_{vk}$ as:

$$\mathbf{G}_i^v = \text{MeanPooling}(r_i^b \mathbf{F}_i^v), \tag{8}$$

$$\mathcal{L}_{vk} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{exp(Sim(\mathbf{G}_i^v, \mathbf{G}_i^{wt}))}{\sum_{j=1}^{B} exp(Sim(\mathbf{G}_j^v, \mathbf{G}_i^{wt}))}, \tag{9}$$

Method	Src.	MR					HD	
		R1		mAP			≥Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
M-DETR (Lei, Berg, and Bansal 2021)	$\mathcal{V}$	52.89	33.02	54.82	29.40	30.73	35.69	55.60
QD-DETR (Moon et al. 2023)	$\mathcal{V}$	62.40	44.98	62.52	39.88	39.86	38.94	62.40
UniVTG (Lin et al. 2023)	$\mathcal{V}$	58.86	40.86	57.60	35.59	35.47	38.20	60.96
TR-DETR (Sun et al. 2024)	$\mathcal{V}$	64.66	48.96	63.98	43.73	42.62	39.91	63.42
UVCOM (Xiao et al. 2024)	$\mathcal{V}$	63.55	47.47	63.37	42.67	43.18	39.74	64.20
Ours	$\mathcal{V}$	**66.86**	**51.23**	**67.73**	**46.24**	**45.69**	**40.94**	**64.79**
UMT (Liu et al. 2022)	$\mathcal{V}+\mathcal{A}$	56.23	41.18	53.38	37.01	36.12	38.18	59.99
QD-DETR (Moon et al. 2023)	$\mathcal{V}+\mathcal{A}$	63.06	45.10	63.04	40.10	40.19	39.04	62.87
TR-DETR (Sun et al. 2024)	$\mathcal{V}+\mathcal{A}$	65.05	47.67	64.87	42.98	43.10	39.90	63.88
UVCOM (Xiao et al. 2024)	$\mathcal{V}+\mathcal{A}$	63.81	48.70	64.47	44.01	43.27	39.79	64.79
Ours	$\mathcal{V}+\mathcal{A}$	**67.77**	**50.52**	**68.30**	**45.88**	**45.52**	**41.15**	**65.82**

Table 1: Experimental results on the QVHighlights *test* set for moment retrieval and highlight detection when using either video only ($\mathcal{V}$) or video and audio ($\mathcal{V}+\mathcal{A}$). **Bold**/underlined fonts indicate the best/second-best results.

where $\mathbf{G}^v \in \mathbb{R}^d$ is average of clip-level visual feature, r^b is a binary value (1 for ground-truth clips, 0 otherwise). The $\mathcal{L}_{vk}$ strengthens the global representation based on keywords, facilitating more effective cross-video learning.

Finally, we devise a keyword-aware contrastive loss $\mathcal{L}_{kw}$ that combines $\mathcal{L}_{ck}$ and $\mathcal{L}_{vk}$, which can be formulated as:

$$\mathcal{L}_{kw} = \mathcal{L}_{ck} + \mathcal{L}_{vk}. \tag{10}$$

The $\mathcal{L}_{vk}$ enables our model to optimize both temporal relevance within videos and global semantic coherence across the dataset, achieving a comprehensive alignment between text queries and video contents.

Training Objective

To train our proposed method, we construct the total training loss function as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_{mr} + \mathcal{L}_{hd} + \lambda_{kw}\mathcal{L}_{kw}, \tag{11}$$

where $\mathcal{L}_{mr}$ and $\mathcal{L}_{hd}$ denote the loss functions for moment retrieval and highlight detection as outlined in (Moon et al. 2023). λ_{kw} is a balancing parameter. The $\mathcal{L}_{Total}$ enables effective moment retrieval and highlight detection.

Experiments

Datasets and Evaluation Metrics

QVHighlights. The QVHighlights dataset (Lei, Berg, and Bansal 2021) includes 10,148 YouTube videos with rich content, each paired with an annotated text query that indicates highlight moments. This is the only dataset that includes both annotations for moment retrieval and highlight detection. Following (Lei, Berg, and Bansal 2021), to ensure a fair evaluation, we submitted our model predictions to the QVHighlights server CodaLab competition platform, with test set annotations remaining confidential.

TVSum. The TVSum dataset (Song et al. 2015) is also a standard benchmark for highlight detection, comprising videos from 10 different categories, with each category containing 5 videos. For a fair comparison, we use the same train/test split as utilized in QD-DETR(Moon et al. 2023).

Charades-STA. The Charades-STA dataset (Gao et al. 2017) contains 9,848 videos depicting indoor activities with 16,128 human-annotated query texts. Following QD-DETR (Moon et al. 2023), we use 12,408 samples for training, with the remaining 3,720 samples allocated for testing.

Evaluation Metric. For the evaluation, we follow the metrics of prior works (Lei, Berg, and Bansal 2021; Xiao et al. 2024; Sun et al. 2024) for a fair and comprehensive comparison. In the QVHighlights dataset, we evaluate Recall@1 (R1) at IoU thresholds of 0.5 and 0.7, and mean average precision (mAP) at thresholds from 0.5 to 0.95 in steps of 0.05 (mAP@Avg). We compare performance at IoU thresholds of 0.5 and 0.75, referred to as mAP@0.5 and mAP@0.75. For highlight detection, we use mAP and HIT@1 (hit ratio of the highest-scored clip). In the Charades-STA dataset, we evaluate Recall@1 at IoU thresholds of 0.5 and 0.7. For the TVSum dataset, the primary evaluation metric is top-5 mAP.

Implementation Details

Pre-extracted Features. Following (Moon et al. 2023), we use the pre-extracted video, text, and audio features from the various models. For video features, we use the pre-trained SlowFast (Feichtenhofer et al. 2019) and CLIP (Radford et al. 2021) models for QVHighlights, VGG (Simonyan and Zisserman 2014) and SlowFast+CLIP (SF+C) for Charades-STA, and I3D pre-trained on Kinetics 400 (Carreira and Zisserman 2017) for TVSum. For text features, we use CLIP (Radford et al. 2021) for QVHighlights and TVSum, and GloVe (Pennington, Socher, and Manning 2014) for Charades-STA. We use audio features from all datasets using a PANN (Kong et al. 2020) model pre-trained on AudioSet.

Training Settings. We set the loss weights to $\lambda_{kw} = 0.3$ and use the Adam optimizer (Kingma and Ba 2014) with

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Avg.
sLSTM (Zhang et al. 2016)	41.1	46.2	46.3	47.7	44.8	46.1	45.2	40.6	47.1	45.5	45.1
LIM-S (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 (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 (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† (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 (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 (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 (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 (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
Ours	**89.9**	**93.8**	**94.4**	85.9	**89.2**	**89.4**	**81.5**	**92.6**	**90.1**	80.6	**88.7**

Table 2: Experimental results on the TVSum for highlight detection. † indicates training with audio modality. **Bold**/underlined fonts indicate the best/second-best results.

Method	Feat	R1@0.5	R1@0.7
2D-TAN (Zhang et al. 2020)	VGG	40.94	22.85
UMT† (Liu et al. 2022)	VGG	48.31	29.25
QD-DETR (Moon et al. 2023)	VGG	52.77	31.13
TR-DETR (Sun et al. 2024)	VGG	53.47	30.81
Ours	VGG	**54.89**	**31.97**
QD-DETR (Moon et al. 2023)	SF+C	57.31	32.55
UniVTG (Lin et al. 2023)	SF+C	58.01	35.65
TR-DETR (Sun et al. 2024)	SF+C	57.61	33.52
UVCOM (Xiao et al. 2024)	SF+C	59.25	36.64
Ours	SF+C	**61.08**	**37.89**

Table 3: Experimental results on the Charades-STA for moment retrieval. † indicates training with audio modality. **Bold**/underlined fonts indicate the best/second-best results.

VCKA	VCP	MR			HD	
		R1		mAP Avg.	mAP	HIT@1
		@0.5	@0.7			
-	-	66.39	49.03	45.55	40.78	65.42
✓	-	68.90	52.32	46.98	41.38	66.19
-	✓	66.90	51.03	46.42	41.30	66.71
✓	✓	**68.97**	**53.35**	**47.69**	**41.67**	**67.03**

Table 4: Effect of the proposed component (video context-aware keyword attention (VCKA) module, and video contextual MR/HD prediction (VCP)) on QVHighlights *val* set.

a learning rate of 1e-3 and a weight decay of 1e-4. We train QVHighlights with a batch size of 32 for 200 epochs, Charades-STA with a batch size of 8 for 100 epochs, and TVSum with a batch size of 4 for 2000 epochs, using a single RTX 4090 GPU. For detailed network configurations, please refer to the supplementary material.

Comparison to Prior Works

Results on the QVHighlights. Table 1 compares the performance of our method with existing state-of-the-art methods on the QVHighlights dataset for moment retrieval (MR) and highlight detection (HD). When using only the video ($\mathcal{V}$), our method shows superior performance, achieving an average of 2.12% higher performance across all metrics. A similar trend is observed when combining video with audio ($\mathcal{V} + \mathcal{A}$). This indicates that the keyword-aware contrastive loss $\mathcal{L}_{kw}$ effectively enables our framework to understand video content and capture relevant keywords.

Results on the TVSum. We use the TVSum dataset to evaluate video highlight detection performance, following protocols of previous works (Moon et al. 2023; Sun et al. 2024; Xiao et al. 2024). As shown in Table 2, our method shows superior performance in 8 out of 10 categories, with an overall average performance (Avg.) of 88.7%, indicating an improvement over existing methods.

Results on the Charades-STA. We also evaluate the performance of our method on moment retrieval using the Charades-STA dataset. As shown in Table 3, our approach consistently shows improved performance, achieving a 0.84% improvement over VGG feature and a 1.25% improvement over SF+C (SlowFast+CLIP) feature in R1@0.7. This demonstrates that our method is robust across various pre-extracted video features for the moment retrieval.

Ablation Study

In this section, we conduct various ablation studies to investigate (1) the effect of our proposed components and (2) the effect of our proposed losses. All ablation studies are conducted using QVHighlights *val* set to evaluate both moment retrieval and highlight detection.

Effect of the Proposed Components. Table 4 shows the effect of the two components Video Context-aware Keyword Attention module (VCKA) and Video Contextual MR/HD Prediction (VCP) heads. The results clearly demonstrate that both of proposed components bring better performance, highlighting the effectiveness of our approach.

Proposed Losses. We evaluate our method with respect to the two keyword-aware contrastive losses, $\mathcal{L}_{ck}$ and $\mathcal{L}_{vk}$. The results in Table 5 show that considering each loss individually consistently surpasses the baseline, which does not use these contrastive losses. Combining all proposed losses achieves the highest performance, significantly enhancing

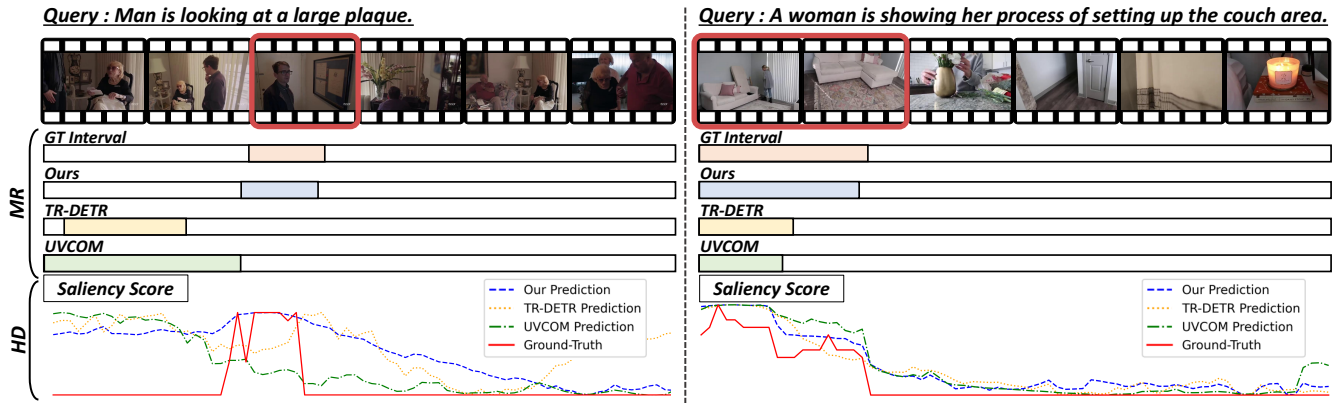

Figure 3: Visualization examples for moment retrieval (MR) and highlight detection (HD) on the QVHighlights *val* set.

$\mathcal{L}_{ck}$	$\mathcal{L}_{vk}$	MR			HD	
		R1		mAP	mAP	HIT@1
		@0.5	@0.7	Avg.		
-	-	66.90	49.94	45.02	40.53	65.48
✓	-	67.74	51.16	46.37	41.32	66.45
-	✓	67.16	51.42	46.76	40.84	65.10
✓	✓	**68.97**	**53.35**	**47.69**	**41.67**	**67.03**

Table 5: Effect of the proposed keyword-aware contrastive losses on QVHighlights *val* set.

the method by improving its capacity to learn robust and discriminative video features.

Visualization Results

As shown in Figure 3, We provide examples of moment retrieval and highlight detection on QVHighlights *val* set by comparing our method with TR-DETR and UVCOM. The visualizations demonstrate the efficacy of our method in both moment retrieval and highlight detection.

Discussion

Effectiveness of Our Keyword Weight. We visualize the keyword weight w^t of each query word on QVHighlights *val* set. As shown in Figure 4, in the first sample, *'woman wearing glasses'* appears in most scenes, resulting in low keyword weights. Conversely, *'microphone'* receives a high weight due to its specificity in the video context. Similarly, in the second sample, *'woman'* has a low weight due to frequent appearance, while *'fruit'*, *'stick'*, and *'eats'* get higher weights. These visualizations demonstrate the effectiveness of our keyword weight w^t which can capture contextual importance within video.

Limitation

Our method has demonstrated robust performance across diverse datasets. However, when processing both visual and audio inputs, we use a simplified audio representation without a detailed audio-specific framework. Therefore, developing a more sophisticated audio-visual integration approach is an important direction for future work.

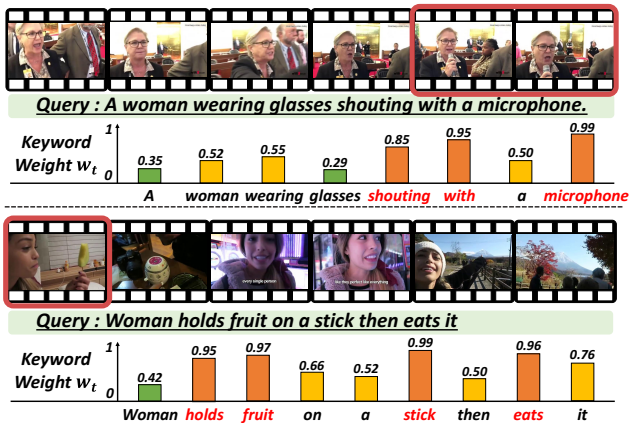

Figure 4: Visualization results of keyword weight effectiveness on QVHighlights *val* set.

Conclusion

In this paper, we propose an innovative approach to moment retrieval and highlight detection that improves understanding of the video context and utilizes contextually appropriate keywords. The core of our method is the video context-aware keyword attention module, which effectively comprehends the entire video sequence and extracts relevant text keywords corresponding to the video content. Our keyword-aware contrastive loss functions successfully enhance the alignment between text query features and video clip features by leveraging the overall flow of the video. We believe that our method not only enhances accuracy but also holds diverse practical applications.

Acknowledgments

This work was supported by the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00252391), and by the IITP grant funded by the Korea government (MSIT) (No. RS-2022-00155911: Artificial Intelligence

Convergence Innovation Human Resources Development (Kyung Hee University), No. 2022-0-00124: Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities, IITP-2023-RS-2023-00266615: Convergence Security Core Talent Training Business Support Program), and conducted by CARAI grant funded by DAPA and ADD (UD230017TD).

References

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

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.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Dagtas, S.; and Abdel-Mottaleb, M. 2004. Multimodal detection of highlights for multimedia content. *Multimedia Systems*, 9: 586–593.

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

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.

Gygli, M.; Grabner, H.; Riemenschneider, H.; and Van Gool, L. 2014. Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, 505–520. Springer.

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*.

Kim, D.; Um, S. J.; Lee, S.; and Kim, J. U. 2024. Learning to Visually Localize Sound Sources from Mixtures without Prior Source Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26467–26476.

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

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.

Kudi, S.; and Namboodiri, A. M. 2017. Words speak for actions: Using text to find video highlights. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, 322–327. IEEE.

Lee, S.; Kim, H.-I.; Ro, Y. M.; and . 2022a. Weakly paired associative learning for sound and image representations via bimodal associative memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10534–10543.

Lee, S.; Kim, H.-I.; Ro, Y. M.; and . 2024. Text-guided distillation learning to diversify video embeddings for text-video retrieval. *Pattern Recognition*, 156: 110754.

Lee, S.; Park, S.; Ro, Y. M.; and . 2022b. Audio-Visual Mismatch-Aware Video Retrieval via Association and Adjustment. In *European Conference on Computer Vision*, 497–514. Springer.

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, 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(3), 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, 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.

Park, K. R.; Lee, H. J.; Kim, J. U.; and . 2024a. Learning Trimodal Relation for Audio-Visual Question Answering with Missing Modality. In *European Conference on Computer Vision*.

Park, K. R.; Oh, Y.; Kim, J. U.; and . 2024b. Enhancing Audio-Visual Question Answering with Missing Modality via Trans-Modal Associative Learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 5755–5759. IEEE.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

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.

Sarfraz, S.; Murray, N.; Sharma, V.; Diba, A.; Van Gool, L.; and Stiefelhagen, R. 2021. Temporally-weighted hierarchical clustering for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11225–11234.

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

Snoek, C. G.; Worring, M.; et al. 2009. Concept-based video retrieval. *Foundations and Trends® in Information Retrieval*, 2(4): 215–322.

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(5), 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.

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, S.; Chen, L.; Zhang, S.; Ji, W.; Shao, J.; Ye, L.; and Xiao, J. 2021. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(4), 2986–2994.

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.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12113–12132.

Yuan, Y.; Mei, T.; and Zhu, W. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(01), 9159–9166.

Zhang, K.; Chao, W.-L.; Sha, F.; and Grauman, K. 2016. Video summarization with long short-term memory. In *ECCV*. Springer.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(07), 12870–12877.