Bridging the Gap: A Unified Video Comprehension Framework for Moment Retrieval and Highlight Detection

Yicheng Xiao[1†], Zhuoyan Luo[1†]
Yong Liu[1], Yue Ma[1], Hengwei Bian[2], Yatai Ji[1], Yujiu Yang[1✉], Xiu Li[1✉]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Carnegie Mellon University

Abstract

*Video Moment Retrieval (MR) and Highlight Detection (HD) have attracted significant attention due to the growing demand for video analysis. Recent approaches treat MR and HD as similar video grounding problems and address them together with transformer-based architecture. However, we observe that the emphasis of MR and HD differs, with one necessitating the perception of local relationships and the other prioritizing the understanding of global contexts. Consequently, the lack of task-specific design will inevitably lead to limitations in associating the intrinsic specialty of two tasks. To tackle the issue, we propose a **U**nified **V**ideo **COM**prehension framework (UVCOM) to bridge the gap and jointly solve MR and HD effectively. By performing progressive integration on intra and inter-modality across multi-granularity, UVCOM achieves the comprehensive understanding in processing a video. Moreover, we present multi-aspect contrastive learning to consolidate the local relation modeling and global knowledge accumulation via well aligned multi-modal space. Extensive experiments on QVHighlights, Charades-STA, TACoS, YouTube Highlights and TVSum datasets demonstrate the effectiveness and rationality of UVCOM which outperforms the state-of-the-art methods by a remarkable margin. Code is available at https://github.com/EasonXiao-888/UVCOM.*

1. Introduction

Video has emerged as a highly favored multi-medium format on the internet with its diverse content. This significant surge in online video encourages users to adjust their strategies for accessing desired video contents. Instead of spending time-consuming efforts inspecting the whole video, they are more inclined to directly obtain particular clips of interest through language descriptions. This shift in user preference gives rise to two significant research topics: Video Moment Retrieval [8, 26, 40, 60, 61], focuses on locating the specific moment, and Highlight Detection [1, 12, 52, 56, 58], is dedicated to identifying segments of high saliency.

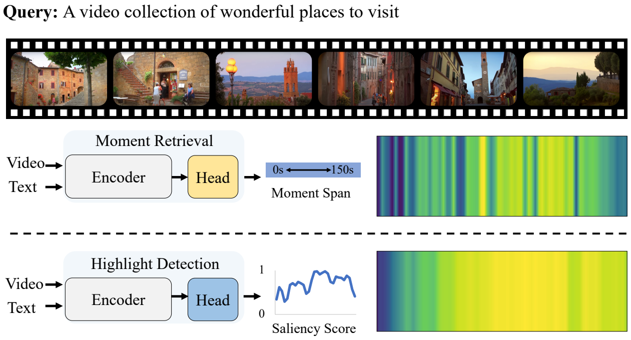

Figure 1. **Illustration of the intrinsic characteristics of Moment Retrieval and Highlight Detection.** We visualize the attention map of the same video under two tasks. The attention map for MR takes on strip-like patterns, indicating the emphasis of local relations. In contrast, it is in band-like format for HD, which signifies the focus on global information.

Actually, it is apparent that two tasks share many common characteristics, *e.g.*, identifying relevant video segments in response to textual expressions. In light of the above, Lei *et.al.* [21] first proposes a novel dataset named QVHighlights and a basic framework called Moment-DETR to jointly solve both tasks. UMT [27] incorporates extra audio modality and QD-DETR [31] produces text-query dependent video representation to achieve better performance. The above-mentioned methods simply model MR and HD as a multi-task problem and mainly concentrate on utilizing non-specific strategy to solve them. In particular, they all adopt a straightforward way to train and optimize both tasks together with general design, *e.g.*, transformer-based models. However, we revisit the characteristics of MR/HD and discover that there exists a gap between them as illustrated in Fig. 1. which leads to the chal-

[†]Equal contribution.
✉ Corresponding author ({yang.yujiu, li.xiu}@sz.tsinghua.edu.cn).

lenge in consistent performance on both tasks, *i.e.*, achieving precise moment localization and accurate highlight-ness estimation simultaneously.

Therefore, we consider that the design of framework should follow two principles to alleviate the above weakness: 1) ***Local Relation Activation:*** MR necessitates the understanding of local relationships within the video to accurately localize specific segments. 2) ***Global Knowledge Accumulation:*** The objective of HD is to fit the saliency distribution of the entire video, emphasizing the importance of global context (in Fig. 1). Based on the principles, we propose a **U**nified **V**ideo **Com**prehension Framework (UVCOM) to seamlessly integrate the emphasis of MR and HD, which effectively bridges the gap and achieves great performance on both tasks consistently. Specifically, we first design a novel Comprehensive Integration Module (CIM) to progressively facilitate the integration on intra and inter-modality across multi-granularity. CIM first efforts to propagate the aggregated semantic phrases from the text into the visual feature to realize local relationship perception. Then, it accumulates global information from video by utilizing the moment-awareness feature as an intermediary. With a comprehensive view of the entire video, CIM facilitates the understanding of particular intervals and highlight contents, which is beneficial to identify the desired moment and non-related ones. Furthermore, we introduce a multi-aspect contrastive learning which incorporates clip-text alignment to consolidate the local relation modeling, and video-linguistic discrimination to enhance the quality of accumulated global information.

We conduct extensive experiments on five popular MR/HD benchmarks to validate the effectiveness of our framework and the results show that UVCOM notably outperforms existing methods for all benchmarks.

Overall, our contributions are summarized as follows:

- Based on our investigation into the emphasis of Moment Retrieval and Highlight Detection, we present two principles for framework design. Guided by them, we propose a Unified Video Comprehension Framework called UVCOM to effectively bridge the gap between two tasks.
- In UVCOM, a Comprehensive Integration Module (CIM) is designed to perform progressive intra and inter-modality interaction across multi-granularity, which achieves locality perception of temporal and multi-modal relationships as well as global knowledge accumulation of the entire video.
- Without bells and whistles, our method outperforms all existing state-of-the-art methods by a remarkable margin, *e.g.*, $+5.97\%$ in R1@0.7 for MR than UniVTG [24] on TACoS [37] and $+3.31\%$ in HIT@1 for HD than QD-DETR [31] on QVHighlights [21].

2. Related Work

Moment Retrieval. Moment Retrieval is a task that aims at retrieving the target moment, *i.e.*, one [8] or many [20] continuous intervals in a video given the text description. Generally, the model will focus more on the relationship across adjacent frames for better localization. Previous works retrieve video intervals into two perspectives: proposal-based and proposal-free. The proposal-based methods [8, 9, 15, 44, 50] follow the propose-then-rank pipeline, where they first generate candidate proposals then rank them based on matching scores. Liu *et al.* [25] and Hendricks *et al.* [15] utilize sliding windows to scan the entire video for candidate proposals generation and calculate the similarity with textual embedding for selection. On the other hand, the proposal-free methods [10, 22, 32, 34, 45, 59, 62] directly regress the start and end timestamp via video-text interaction. Yuan *et al.* [59] and Mun *et al.* [32] generate the temporal coordinates of sentence by multi-modal co-attention mechanism. Furthermore, Rodriguez *et al.* [34] utilizes a simple dynamic filter instead of cross attention to match video and text embedding.

Highlight Detection. Highlight Detection aims to identify highlights or important segments with high potential appeal in a video. Compared with Moment Retrieval, It is necessary for the model to associate the whole video content for fitting saliency distribution of each clip. Many prior works [12, 52, 56, 56, 58] adopt ranking formulation where they rank the important segments with higher score. Video2Gif [12] trains a generic highlight predictor to produce GIF from videos. Rochan *et al.* [39] designs a task-specific highlight detectors to automatically create highlights from the user history. Recently, Badamdorj *et al.* [1] elaborates on fusing visual and audio content to generate better video representations.

MR and HD share many similar properties. Moment-DETR [21] puts forward a novel dataset which first includes two tasks and provides a simple DETR-based [2] network. To improve the query quality, UMT [27] proposes to adopt audio, visual and text content for query generation. Furthermore, QD-DETR [31] exploits the textual information by involving video-text pair negative relationship learning, achieving greater performance. However, previous methods simply train and optimize two tasks without considering the different emphasis of each task. To address this issue, we propose a novel unified framework UVCOM that effectively associates the speciality of MR and HD to achieve comprehensive understanding.

3. Method

Given a video of L clips $\{v_1, v_2, \dots, v_l\}$ and a textual expression of N words $\{e_1, e_2, \dots, e_n\}$, the goal of MR is

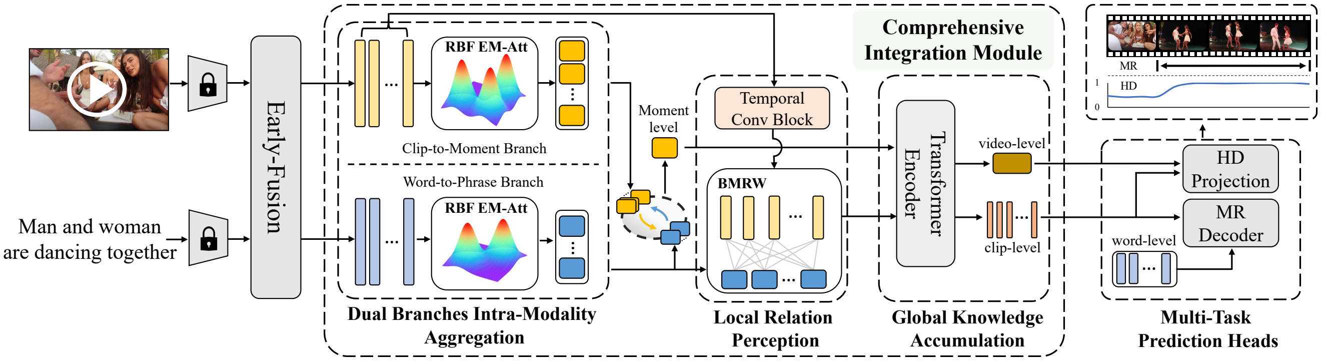

Figure 2. **Overview of UVCOM.** Based on the exploration of MR and HD, we propose a unified video comprehension framework guided by the design principles. Specifically, the model takes a video with language description as input. After encoding and early-fusion process, we design a Comprehensive Integration Module (CIM) to achieve subsequent progressive integration on intra and inter-modality across multi-granularity. Finally, the multi-task heads output the moment spans for MR and saliency scores for HD.

to localize the most relevant moment with the center coordinate and duration, while HL is to generate the saliency score distribution for the whole video.

3.1. Visual-Text Encoding

Visual Encoder. Following previous works [21, 24, 27, 31], we utilize the pretrained backbone, *e.g.*, SlowFast [6], video encoder of CLIP [36] and I3D [3] to extract visual features $\mathcal{F}_v \in \mathbb{R}^{L\times D}$ of the video. Note that D demotes the channel.

Language Encoder. Simultaneously, text encoder of CLIP is adopted to encode the linguistic expression into the textual embedding $\mathcal{F}_t \in \mathbb{R}^{N\times D}$.

With the visual and textual features, we apply a bidirectional transformer-based encoder [13, 29, 55] to perform the early fusion. It coarsely encodes features in different modalities and outputs preliminary aligned visual and textual representations.

3.2. Comprehensive Integration Module

After getting the visual and textual representations, we design a Comprehensive Integration Module (CIM) to perform progressive intra and inter-modality integration across multi-granularity. Specifically, we leverage Expectation-Maximum (EM) Attention [23] on associating inner-modality content to generate the moment-wise visual features and phrase-wise textual features, respectively. Then we propose Local Relation Perception (LRP) module to unify temporal relationship modeling and inter-modality fusion, which reformulates the temporal and modality interconnection to enhance the locality perception. Finally, we utilize a standard encoder to produce the video-wise feature by integrating the correlation between moment and clip-wise visual features.

Dual Branches Intra-Modality Aggregation. A video usually contains more than one event and irrelevant background scenes. The same scenario happens in textual descriptions where insignificant words and unconstrained expressions may cause potential ambiguity. To tackle the problem, we propose to utilize RBF-kernel based EM Attention [17, 23] to aggregate the clip/word-level features. As shown in Fig. 2, it is a dual-branches structure. The clip-to-moment branch aims at incorporating the relationship of each clip to enhance the desired event representations while suppressing the background noise. Meanwhile, the word-to-phrase branch is to emphasize the referred moment description by accumulating contextual information.

Specifically, we fit the distribution of $\mathcal{F}_v$ and $\mathcal{F}_t$ by a separated Gaussian Mixture Model [38] to generate the compact moment and phrase-level representations via the centroid of Gaussians. Taking $\mathcal{F}_v$ as an example, we utilize a linear superposition of n_v Gaussians to capture the statistics of $f_v^i \in \mathbb{R}^D$ (the i-th snippet of $\mathcal{F}_v$):

$$p(f_v^i) = \sum_{k=1}^{n_v} z_k^v \mathcal{N}(f_v^i|\mu_k, \Sigma_k), \tag{1}$$

where $z_k^v \in \mathbb{R}$, $\mu_k \in \mathbb{R}^D$ and $\Sigma_k \in \mathbb{R}^{D\times D}$ denote the weight, mean and covariance of k-th Gaussian basis for the clip-to-moment branch. We substitute the covariance with an identity matrix I for simplification and employ the radial basis function (RBF Kernel) $\mathcal{K}(f_v^i, \mu_k)$ to estimate the posterior probability $\mathcal{N}(f_v^i|\mu_k, \mathcal{I})$:

$$\mathcal{K}(f_v^i, \mu_k) = exp(-\lambda \left\| f_v^i - \mu_k \right\|_2^2), \tag{2}$$

where $\lambda > 0$ is an adjustable hyper-parameter to control the distribution. Afterwards, at t-th iteration, we update the weight $Z^{(t)} \in \mathbb{R}^{L\times n_v}$ in the E Step and re-estimate $\mu^{(t)} \in \mathbb{R}^{n_v\times D}$ in the M step, which can be formulated as:

$$\mu^{(t)} = \mathrm{Norm}_1(Z^{(t)})^T F_v, \quad t \in \{1, \ldots, T\}. \tag{3}$$

Furthermore, in contrast to conventional cluster methods that only involve iterative update, the initialized means $\mu^{(0)}$

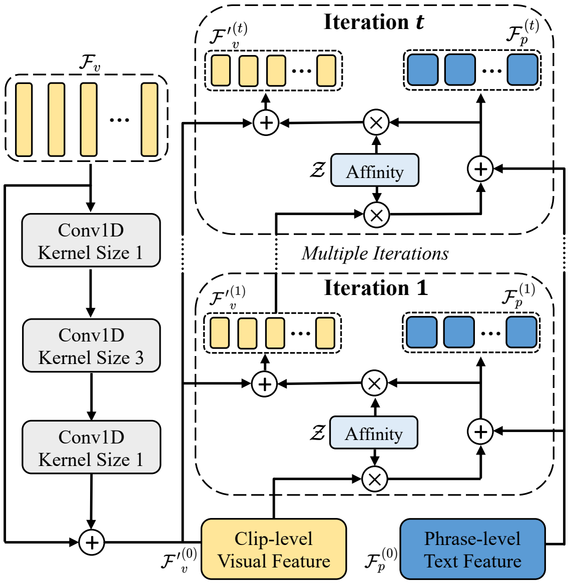

Figure 3. **Illustration of Local Relation Perception (LRP) module.** We first process the visual feature $\mathcal{F}_v$ with a Conv1D Block. Then we develop a Bidirectional Modality Random Walk (BMRW) algorithm to exploit the power of fine-grained multi-modal interaction. The affinity $\mathcal{Z}$ is generates by scaled dot product: $\mathcal{Z} = \lambda_z \mathcal{F'}_v^{(0)} (\mathcal{F}_p^{(0)})^\top$.

we set are learnable. Therefore, they can effectively capture the feature distribution of the dataset through the standard back-propagation.

After t iterations, we obtain the fine-grained moment-wise representation $\mathcal{F}_m$ from $\mu^{(t)}$ which fully aggregates the contextual information. Similarly, we operate the above steps on word-to-phrase branch to generate the phrase-level linguistic feature $\mathcal{F}_p \in \mathbb{R}^{n_t \times D}$, where n_t indicates the number of Gaussian basis in word-to-phrase branch.

Local Relation Perception. Previous methods [21, 27, 31] directly perform cross-modal fusion between clip and word-level features, disregarding the temporal relation and valuable semantic interaction across different granularities. Without the information from adjacent clips, the simple and coarse clip-word fusion will easily deviate the model from focusing on the relevant boundary clips, causing incorrect localization. To address the aforementioned weakness, we design a Local Relation Perception (LRP) module to excavate both temporal and inter-modality relationships. As shown in Fig. 2, we first utilize a temporal convolution block to improve the locality perception of clip-level features, which can be formulated as:

$$\mathcal{F'}_v = \text{Conv}\left(\mathcal{F}_v\right) + \mathcal{F}_v. \tag{4}$$

Since simply incorporate clip-level relation may introduce local redundancy, we leverage fine-grained inter-modal interaction to re-calibrate the attention for activating the relevant moments. Intuitively, a straightforward approach is to utilize cross-attentive mechanism [27, 31] to perform inter-modal interaction. Nevertheless, the complex scenario in an untrimmed video, *e.g.*, footage transitions and irrelevant events, will increase the likelihood of attention drift which leads to the undesirable local activation. Moreover, although phrase-wise linguistic features specify referred moment description and alleviate the impact of noise in contrast to word-wise one, it may potentially contribute to attention drift due to the irrelevant accumulated words. Therefore, inspired by [11, 17, 33], we design a bidirectional modality random walk (BMRW) algorithm to mitigate the mentioned drawbacks and fully exploit the power of the fine-grained multi-modal interaction. It propagates the textual prior into the visual features for highlighting the corresponding local context and suppressing unrelated ones. Simultaneously, linguistic features are refined through the incorporation of updated visual content. As shown in Fig. 3, there are multiple iterations in BMRW where two modalities features learn collaboratively in visual-linguistic shared embedding space until convergence.

Formally, we first define the $\mathcal{F'}_v$, $\mathcal{F}_p$ as initial features $\mathcal{F'}_v^{(0)}$, $\mathcal{F}_p^{(0)}$ at 0-th iteration and formulate affinity $\mathcal{Z}$ by scaled dot product: $\mathcal{Z} = \lambda_z \mathcal{F'}_v^{(0)} (\mathcal{F}_p^{(0)})^\top$, where λ_z is the scaling factor. At t-th iteration, the phrase-wise linguistic feature $\mathcal{F}_p^{(t)}$ is updated by the original feature $\mathcal{F}_p^{(0)}$ and the visual output $\mathcal{F'}_v^{(t-1)}$ from previous iteration:

$$\mathcal{F}_p^{(t)} = \omega \text{Norm1}(\mathcal{Z})^\top \mathcal{F'}_v^{(t-1)} + (1-\omega)\mathcal{F}_p^{(0)}, \tag{5}$$

Subsequently, it is projected into the temporal-awareness feature $\mathcal{F'}_v^{(t)}$:

$$\mathcal{F'}_v^{(t)} = \omega \mathcal{Z} \mathcal{F}_p^{(t)} + (1-\omega)\mathcal{F'}_v^{(0)}, \tag{6}$$

where $\omega \in (0, 1)$ is the factor which controls the degree of modalities fusion. Then, we substitute $\mathcal{F}_p^{(t)}$ into Eq. (6) to derive the iterative update formula of $\mathcal{F'}_v^{(t)}$:

$$\mathcal{F'}_v^{(t)} = (\omega^2 A)^t \mathcal{F'}_v^{(0)} + (1-\omega) \sum_{i=0}^{t-1} (\omega^2 A)^i (\omega \mathcal{Z} \mathcal{F}_p^{(0)} + \mathcal{F'}_v^{(0)}), \tag{7}$$

where A denotes $\mathcal{Z}\text{Norm1}(\mathcal{Z})^\top$. Intuitively, the moment-specific regions of visual features can be fully activated by the guidance of textual features after multiple iterations. Moreover, to avoid the potential issue of unexpected gradient and high computation cost, we use an approximate inference function based on Neumann Series [30] when $t \rightarrow \infty$:

$$\mathcal{F'}_v^{(\infty)} = (1-\omega)(I - \omega^2 A)^{-1} (\omega \mathcal{Z} \mathcal{F}_p^{(0)} + \mathcal{F'}_v^{(0)}). \tag{8}$$

In this manner, the model realizes a synergistic temporal and inter-modality relation integration and generates a more comprehensive visual representation $\mathcal{F}_v^{new}$, *i.e.*, $\mathcal{F'}_v^{(\infty)}$ in Eq. (8).

Global Knowledge Accumulation. As illustrated in Fig. 1, Highlight Detection prioritizes global information of videos. QD-DETR [31] uses a saliency token to capture general information. However, the input-agonist design might cause the inferior perception of the text-related intervals due to the non-referential search area. To mitigate the concern, we propose to use the moment-aware feature as intermediate guidance to accumulate the global knowledge of a video. Specifically, we derive the most relevant snippet $\mathcal{F'}_m$ by measuring the similarity between the moment-wise $\mathcal{F}_m$ and phrase-wise embeddings $\mathcal{F}_p$. Then, a stack of transformer encoder layers [46] are utilized to excavate the correlation between $\mathcal{F'}_m$ and $\mathcal{F}_v^{new}$. The overall process is:

$$\mathcal{F}_v^g, \mathcal{F}_v^l = Encoder(Concat[\mathcal{F'}_m, \mathcal{F}_v^{new}]). \tag{9}$$

Consequently, the semantic snippet is obliged to focus on the referred moment and suppress the non-target response, which eventually produces the video-wise feature $\mathcal{F}_v^g \in \mathbb{R}^{1\times D}$. In addition, $\mathcal{F}_v^l \in \mathbb{R}^{L\times D}$ is greatly enriched by the supplement of global information.

3.3. Multi-Aspect Contrastive Learning

As discussed in Sec. 3.2, CIM can better accomplish local relation enhancement in temporal and inter-modality as well as global knowledge accumulation of a video. It is anticipated that the explicit supervision of each objective will further consolidate the effectiveness. To this end, we introduce multi-aspect contrastive learning in two folds:

Clip-Text Alignment. This loss bridges the semantic gap between the textual expression and the clip-level features, which further improves the quality of local relation modeling. Specifically, we first average $\mathcal{F}_t$ to get the sentence-level textual embedding $\mathcal{F'}_t \in \mathbb{R}^{1\times D}$ and then measure the relevance with clip-level visual representation $\mathcal{F}_v^{new}$:

$$S_{ct} = \frac{\mathcal{F}_v^{new} \cdot \mathcal{F'}_t^{\top}}{\|\mathcal{F}_v^{new}\| \cdot \|\mathcal{F'}_t\|}. \tag{10}$$

Finally, we compute the contrastive loss by matrix multiplication:

$$\mathcal{L}_{cta} = -\text{LogSoftmax}\left(S_{ct}\right) \cdot G_{ct}, \tag{11}$$

where G_{ct} is annotated to 1 for relevant clips and 0 for others.

Video-Linguist Discrimination. It aims at constructing the fine-grained multi-modal joint space where video-level visual feature $\{\mathcal{F}_{v(i)}^g\}_{i=1}^B$ closens relevant sentence-level textual representation $\{\mathcal{F'}_{t(i)}\}_{i=1}^B$ while distances unrelated ones within a batch B. Similar to [28, 36, 51], the whole process can be formulated as:

$$\mathcal{L}_{vld} = -\sum_{i=1}^{B} \text{Log}\frac{exp\left(\mathcal{F}_{v(i)}^g \cdot \mathcal{F'}_{t(i)}^{\top}\right)}{\sum_{j=1}^{B} exp\left(\mathcal{F}_{v(i)}^g \cdot \mathcal{F'}_{t(j)}^{\top}\right)}. \tag{12}$$

3.4. Prediction Heads and Loss Function

Multi-Task Prediction Heads. As depicted in Fig. 2, there are two simple heads built on top of the Comprehensive Integration Module for Moment Retrieval and Highlight Detection respectively. Similar to [18, 21, 31], Moment Retrieval Head comprises a standard transformer decoder [14, 47, 65] where we leverage $\mathcal{F}_t$ as the query to generate a series of moment spans P_m. Highlight Detection Head consists of two groups of single fully-connected layer for linear projection. Accordingly, we get the prediction saliency scores $P_s \in \mathbb{R}^{L\times 1}$:

$$P_s = \frac{\mathcal{F}_v^g w_g^{\top} \cdot \mathcal{F}_v^l w_l^{\top}}{\sqrt{d}}, \tag{13}$$

where w_g and $w_l \in \mathbb{R}^{d\times D}$ are learnable weights.

Total Loss. We supervise our framework by four groups of training objective functions. For MR, $L1$ loss and $GIoU$ loss are adopted to measure the disparity between GT moment G_m and prediction spans P_m :

$$\mathcal{L}_{MR} = \lambda_{gIoU}\mathcal{L}_{gIoU}(P_m, G_m) + \lambda_{L1}\mathcal{L}_{L1}(P_m, G_m). \tag{14}$$

Moreover, the loss functions for HD consist of margin ranking loss $\mathcal{L}_{margin}$ and rank-aware loss $\mathcal{L}_{rank}$ following [31]. Both losses work in tandem to ensure the predicted saliency scores P_s conform to the ground truth scores G_s :

$$\mathcal{L}_{HD} = \lambda_{HD}\left[\mathcal{L}_{margin}(P_s, G_s) + \mathcal{L}_{rank}(P_s, G_s)\right] \tag{15}$$

Inspired by [5, 31], we involve hard samples into training process for diversifying the formulations of local and global relationships of different video-text pairs. Briefly, we categorize the lowest relevance between video and text as hard samples and suppress their saliency scores P_s^{hard} during training:

$$\mathcal{L}_{hard} = -\lambda_{hard}\text{Log}(1 - P_s^{hard}) \tag{16}$$

In addition, the objective of multi-aspect contrastive learning promotes semantic associations between text descriptions and visual contents of multi-granularity:

$$\mathcal{L}_{con} = \lambda_{cta}\mathcal{L}_{cta} + \lambda_{vld}\mathcal{L}_{vld}. \tag{17}$$

Generally, the total loss is expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{HD} + \mathcal{L}_{MR} + \mathcal{L}_{hard} + \mathcal{L}_{con}. \tag{18}$$

The λ above are hyper-parameters for balancing the losses.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate our model on five prevalent MR/HD benchmarks: QVHighlights [21], Charades-STA [7], TaCoS [37], TVSum [42] and YouTube Highlights [43]. Due to the space limitation, the details of each datasets are included in the supplementary material.

Method	MR					HD	
	R1		mAP			≥ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
M-DETR [21]	52.89	33.02	54.82	29.40	30.73	35.69	55.60
UMT† [27]	56.23	41.18	53.83	37.01	36.12	38.18	59.99
UniVTG [24]	58.86	40.86	57.60	35.59	35.47	38.20	60.96
MH-DETR [54]	60.05	42.28	60.75	38.13	38.38	38.22	60.51
QD-DETR† [31]	63.06	45.10	63.04	40.1	40.19	39.04	62.87
EaTR [18]	61.36	45.79	61.86	41.91	41.74	37.15	58.65
UVCOM	**63.55**	**47.47**	**63.37**	**42.67**	**43.18**	**39.74**	**64.20**
UVCOM †	**63.81**	**48.70**	**64.47**	**44.01**	**43.27**	**39.79**	**64.79**
With ASR Captions Pretrain							
M-DETR [21]	59.78	40.33	60.51	35.36	36.14	37.43	60.17
UMT [27]	60.83	43.26	57.33	39.12	38.08	39.12	62.39
QD-DETR [31]	64.10	46.10	64.30	40.50	40.62	38.52	62.27
UVCOM	**64.53**	**48.31**	**64.78**	**43.65**	**43.80**	**39.98**	**65.58**

Table 1. **Jointly MR and HD results on QVHighlights test split.** † indicates training with audio modality. *With ASR Caption Pretrain* denotes models pretrained on ASR captions [21].

Method	Charades-STA		TACoS	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7
2D TAN [64]	46.02	27.50	27.99	12.92
VSLNet [62]	42.69	24.14	23.54	13.15
M-DETR [21]	53.63	31.37	24.67	11.97
QD-DETR [31]	57.31	32.55	–	–
UniVTG [24]	58.01	35.65	34.97	17.35
UVCOM	**59.25**	**36.64**	**36.39**	**23.32**

Table 2. **MR results on Charades-STA test split and TACoS test split**. The pre-extracted features are from SlowFast [6] and CLIP [36].

Metrics. Following [2, 21, 27], we measure the performance of our model by the same criteria for QVhighlights, Charades-STA, TACoS, YouTube Highlights and TVSum. For descriptions of the metrics corresponding to datasets, please see the supplementary material.

4.2. Implementation Details

Pre-extracted Features. For a fair comparison, we take the same features of video, text and audio from corresponding pretrained feature extractors, *e.g.*, SlowFast [6], CLIP [36], PANN [19]. For more details please refer to supplementary material.

Training Settings. Our model is trained with AdamW optimizer where the learning rate is 1×10^{-4} and weight decay is 1×10^{-4} by default. The encoder of Global Knowledge Accumulation and the decoder of Moment Retrieval Head compose of three layers of transformer blocks. The coefficients for losses are set to $\lambda_{cta} = 0.5, \lambda_{hard} = 1, \lambda_{vld} = 0.5, \lambda_{HD} = 1, \lambda_{gIoU} = 1, \lambda_{L1} = 10$ in default. Due to space limitations, please see the supplementary material for more training details.

4.3. Main Result

QVHighlights. We compare our method to previous methods on QVHighlights in Tab. 1. Benefiting from the comprehensive understanding of the video, our UVCOM achieves new state-of-the-art performance on different settings and shows a significant margin across all metrics. Specifically, our approach outperforms EaTR [18] by 2.25% on the average of all metrics. Incorporating with video and audio modality, UVCOM yields a clear improvement of 3.6% in R1@0.7, 4% in mAP@0.75 for MR and 2% in HID@1 for HD compared to QD-DETR [31]. Furthermore, with ASR caption pretraining, UVCOM achieves the greatest performance on more stringent metrics, *e.g.*, 43.8% in Avg. mAP for MR and 39.98% in Avg. mAP for HD, demonstrating the effectiveness of our method.

Charades-STA & TACoS. In order to evaluate the performance of our method in precise moment localization, we report the results on Charades-STA and TACoS benchmarks. As depicted in Tab. 2, UVCOM outperforms QD-DETR [31] by about 4% R1@0.7 using SlowFast and CLIP features in Charades-STA dataset while boosts 6% R1@0.7 than UniVTG [24] in TaCoS. It is worth noting that we also validate our model surpasses the existing SOTA methods using VGG features (see in supplementary material).

YouTube Highlights & TVSum. For Video Highlight Detection, we conduct experiments on TVSum and YouTube Highlights. Considering the fact that the scale and scoring criteria of TVSum is small and inconsistent, our method gains incoherently among domains. However, in Tab. 4, it still boost an improvement of 1.3% in Avg. mAP compared with the SOTA methods. As shown in Tab. 3, our method achieves 76.4% and 77.4% in Avg. mAP without audio source under different settings. Note that the features used in UniVTG [24] and UMT [27] on YouTube Highlights are different. Therefore, we follow the same protocol of each for a fair comparison.

4.4. Ablation Study

In this section, we conduct a series of analysis experiments on the val split of QVHighlights benchmark and train the model from scratch without audio modality.

Component Analysis. We first verify the effectiveness of the proposed Comprehensive Integration Module (CIM) and Multi-Aspect Contrastive Learning (MCL). As illustrated in Tab. 5, both of them brings improvement and their combination contributes to better performance ,*i.e.*, +5.71% in Avg. mAP, which demonstrates the effectiveness of the comprehensive understanding. To further investigate the validity of three modules involved in CIM, we provide additional experiments on Dual Branches Intra-Modality Aggregation (DBIA), Local Relation Perception

Method	Dog	Gym.	Par.	Ska.	Ski.	Sur.	**Avg.**
GIFs [12]	30.8	33.5	54.0	55.4	32.8	54.1	46.4
LSVM [43]	60.0	41.0	61.0	62.0	36.0	61.0	53.6
LIM-S [52]	57.9	41.7	67.0	57.8	48.6	65.1	56.4
SL-Module [53]	70.8	53.2	77.2	72.5	66.1	76.2	69.3
MINI-Net† [16]	58.2	61.7	70.2	72.2	58.7	65.1	64.4
TCG† [57]	55.4	62.7	70.9	69.1	60.1	59.8	63.0
Joint-VA† [1]	64.5	71.9	80.8	62.0	73.2	78.3	71.8
UMT†[27]	65.9	75.2	81.6	71.8	72.3	82.7	74.9
UniVTG [24]	71.8	76.5	73.9	73.3	73.2	82.2	75.2
UVCOM[1]	**73.8**	**77.1**	75.7	**75.3**	**74.0**	**82.7**	**76.4**
UVCOM[2]	**66.5**	**77.4**	**82.8**	**78.7**	**74.2**	**84.6**	**77.4**

Table 3. **HD results of mAP on YouTube HL.** † denotes using audio modality. 1 and 2 indicate using the same visual and textual features of UniVTG and UMT.

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	**Avg.**
sLSTM [63]	41.1	46.2	46.3	47.7	44.8	46.1	45.2	40.6	47.1	45.5	45.1
LIM-S [52]	55.9	42.9	61.2	54.0	60.4	47.5	43.2	66.3	69.1	62.6	56.3
Trailer [48]	61.3	54.6	65.7	60.8	59.1	70.1	58.2	64.7	65.6	68.1	62.8
SL-Module [53]	86.5	68.7	74.9	86.2	79.0	63.2	58.9	72.6	78.9	64.0	73.3
MINI-Net† [16]	80.6	68.3	78.2	81.8	78.1	65.8	57.8	75.0	80.2	65.5	73.2
TCG† [57]	85.0	71.4	81.9	78.6	80.2	75.5	71.6	77.3	78.6	68.1	76.8
Joint-VA† [1]	83.7	57.3	78.5	86.1	80.1	69.2	70.0	73.0	**97.4**	67.5	76.3
UniVTG [24]	83.9	85.1	89.0	80.1	84.6	81.4	70.9	91.7	73.5	69.3	81.0
UMT†[27]	87.5	81.5	88.2	78.8	81.5	**87.0**	76.0	86.9	84.4	**79.6**	83.1
QD-DETR [31]	**88.2**	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7	85.0
UVCOM	87.6	**91.6**	**91.4**	**86.7**	**86.9**	86.9	**76.9**	**92.3**	87.4	75.6	**86.3**

Table 4. **HD results of Top-5 mAP on TVSum.** † denotes using audio modality. The 2-nd performance values are highlighted by underline.

CIM	**MCL**	**MR**			**HD**	
		R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
		61.55	44.84	40.08	37.10	62.0
✓		62.84	48.77	43.6	39.33	62.97
	✓	60.77	44.06	40.48	38.81	62.06
✓	✓	**65.10**	**51.81**	**45.79**	**40.03**	**63.29**

Table 5. **Effectiveness of the proposed modules.**

DBIA	**LRP**	**GKA**	**MR**			**HD**	
			R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
			60.77	44.06	40.48	38.81	62.06
		✓	62.32	46.71	41.03	38.73	62.58
	✓		62.06	46.45	41.42	38.57	62.45
✓		✓	63.74	49.16	43.45	39.54	**64.26**
✓	✓		64.71	50.0	43.69	39.69	63.16
	✓	✓	64.84	50.0	44.02	39.58	64.13
✓	✓	✓	**65.10**	**51.81**	**45.79**	**40.03**	63.29

Table 6. **Effects of the components designed of proposed CIM module.**

Method	**MR**			**HD**	
	R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
Average	63.48	49.87	44.10	39.81	63.16
K-Means	62.58	48.39	43.13	39.47	62.26
EM-Att	64.32	50.26	44.49	39.82	**64.0**
EM-Att†	**65.10**	**51.81**	**45.79**	**40.03**	63.29

Table 7. **Impact of various aggregation methods.** † indicates the EM Attention module with RBF kernel.

Method	**MR**			**HD**	
	R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
Cross Attention	63.03	49.87	43.79	39.63	**63.94**
BMRW	**65.10**	**51.81**	**45.79**	**40.03**	63.29

Table 8. **Comparison of different modality interaction strategies.**

(LRP) and Global Knowledge Accumulation (GKA). As shown in Tab. 6, since GKA facilitate the understanding of global context, the ablation of it leads to inferior performance on HD, *i.e.*, -1.1% in HIT@1. Moreover, LRP brings a clear improvement of $+2.34\%$ in Avg. mAP on MR, proving the enhancement on locality perception.

Aggregation Method. We study the impacts on various aggregation methods utilized in DBIA module. As illustrated in Tab. 7, we believe the superiority of our RBF kernel based EM-Attention derives from two aspects: 1) Compared with "Average" and K-Means, our method enhances the desired moment representation while suppresses noises. 2) RBF kernel maps features into a high-dimensional latent space while modeling the relationship within it, which is beneficial for the subsequent aggregation.

Modality Interaction Strategy. We investigate the effects of different modality interaction strategies in Local Relation Perception. As shown in Tab. 8, replacing BMRW by cross attention mechanism results in 2% performance degradation, which demonstrates the effectiveness of BMRW. Furthermore, we provide visualization of features to prove the rationality of LRP. It can be seen in Fig. 5 that the utilization of cross-attentive mechanism leads to the emergence of attention drift. In contrast, through iterative multi-modal learning in shared space, BMRW mitigates the issue, thereby facilitating more precise localization. Moreover, LRP achieves the local relation perception evidenced by clearer strip-like attention patterns in Fig. 5.

Grounding Consistency. Benefiting from the task-specific design, our method yields greater consistency in the joint solution of MR and HD. To quantify the performance coherence, on one hand, we count the videos with accurate hightlight-ness estimation ($mAP_{\text{HD}} > 0.8$) and calculate MR mAP for those videos as shown in Fig. 6 (a). On the other hand, we measure the HD mAP and quantities of videos with precise moment spans ($mAP_{\text{MR}} > 0.8$) as

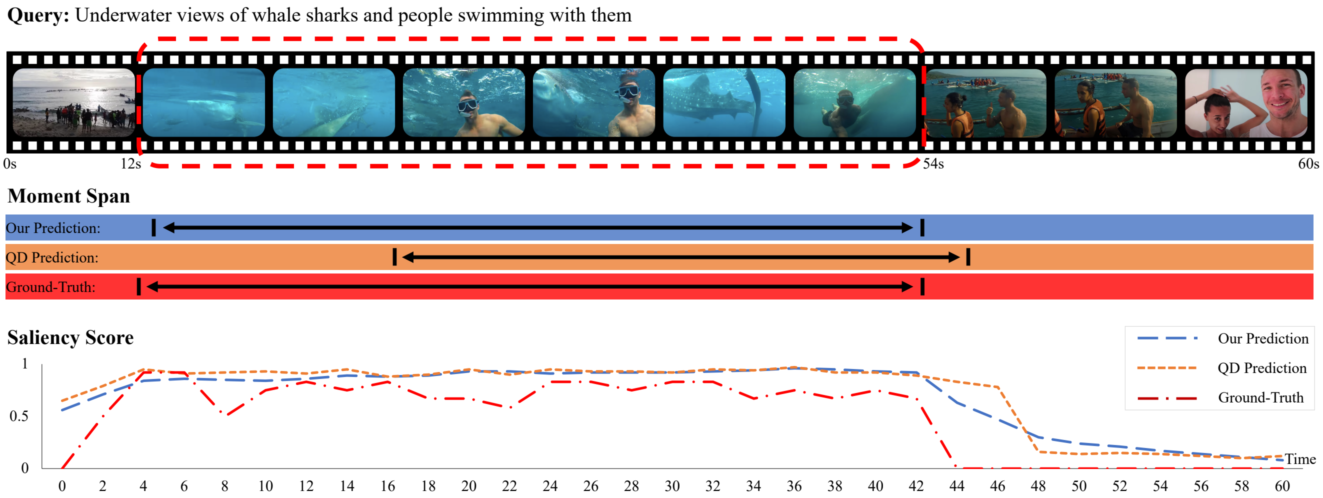

Figure 4. **Visullization comparison on MR and HD.** QD indicates previous state-of-the-art method QD-DETR [31]

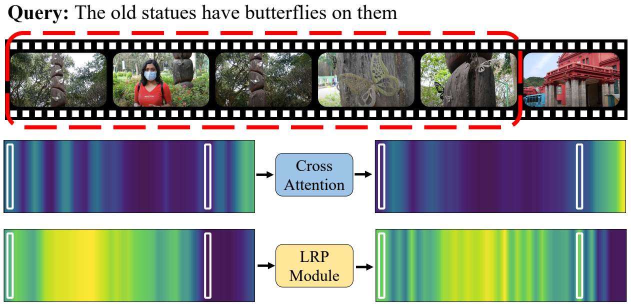

Figure 5. **Illustration of different modality interaction strategies.** The red bounding box indicates the relevant interval and the white bounding box denotes the start and end clips.

shown in Fig. 6 (b). The results demonstrate that UVCOM effectively bridges the gap between two tasks for which our method is superior on all statistics, *i.e.*, MR and HD precision as well as quantity.

4.5. Qualitative Results

As shown in Fig. 4, The local-global enhancement and comprehensive understanding allows our method to accurately model the saliency distribution and localize timestamps of the moment precisely. Comparatively, without the explicit association of characteristics of two tasks, QD-DETR [31] struggles to handle simultaneously in complex scenarios.

5. Conclusion

In light of the different emphasis on MR and HD, we propose a unified video comprehension framework called UVCOM under the guidance of design principles to effectively bridge the gap between two tasks. By performing progressive intra and inter-modality interaction across multi-granularity, UVCOM achieves locality perception of temporal and multi-modal relationship as well as global knowledge accumulation of the entire video. Moreover, we introduce multi-aspect contrastive learning to provide the explicit supervision of above two objectives. Extensive studies validate our model's comprehensive understanding of videos and show our UVCOM remarkably outperforms the existing state-of-the-art methods.

Limitations. Since we just use a simple way to handle audio features instead of specific design, we think that the explicit design for audio features is an interesting future direction.

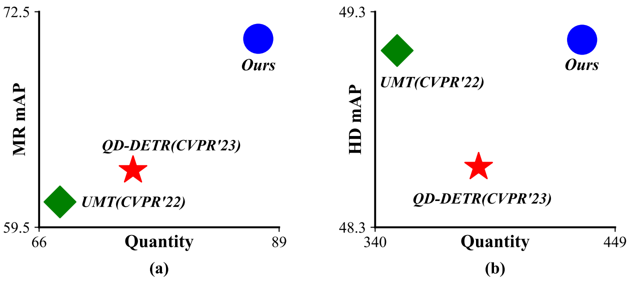

Figure 6. **Illustration of grounding consistency of MR and HD.** (a) indicates the videos collected by $mAP_{\text{HD}} > 0.8$. (b) indicates the videos collected by $mAP_{\text{MR}} > 0.8$.

Acknowledgements

This work is supported by National Key R&D Program of China (Grant No. 2020AAA0108303), Shenzhen Science and Technology Project (Grant No. JCYJ20200109143041798), Shenzhen Stable Supporting Program (WDZC20200820200655001) and partly supported by the National Natural Science Foundation of China (Grant No. U1903213) and the Shenzhen Science and Technology Program (JSGG20220831093004008).

References

[1] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *ICCV*, pages 8107–8117, 2021. 1, 2, 7

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 2, 6

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 3, 1

[4] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, pages 8199–8206, 2019. 1

[5] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *ICCV*, pages 1851–1860, 2017. 5

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210, 2019. 3, 6, 1

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 5, 2

[8] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *ICCV*, pages 5277–5285, 2017. 1, 2

[9] Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, and Shu-Tao Xia. Backdoor defense via adaptively splitting poisoned dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4005–4014, 2023. 2

[10] Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. In *ICLR*, 2024. 2

[11] Leo Grady. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(11):1768–1783, 2006. 4

[12] Michael Gygli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. In *CVPR*, pages 1001–1009, 2016. 1, 2, 7

[13] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *NeurIPS*, 36, 2024. 3

[14] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. In *ICLR*, 2024. 5

[15] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with temporal language. In *EMNLP*, pages 1380–1390, 2018. 2

[16] Fa-Ting Hong, Xuanteng Huang, Weihong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. In *ECCV*, pages 345–360, 2020. 7

[17] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *CVPR*, pages 3262–3271, 2022. 3, 4

[18] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. *arxiv preprint arXiv: 2308.06947*, 2023. 5, 6

[19] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28: 2880–2894, 2020. 6, 1

[20] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. TVR: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, pages 447–463, 2020. 2

[21] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *arXiv preprint arXiv: 2107.09609*, 2021. 1, 2, 3, 4, 5, 6

[22] Kun Li, Dan Guo, and Meng Wang. Proposal-free video grounding with contextual pyramid network. In *AAAI*, pages 1902–1910, 2021. 2

[23] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, pages 9166–9175, 2019. 3

[24] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. *arXiv preprint arXiv: 2307.16715*, 2023. 2, 3, 6, 7

[25] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, pages 15–24, 2018. 2

[26] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *ACM MM*, pages 843–851, 2018. 1

[27] Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. UMT: unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, pages 3032–3041, 2022. 1, 2, 3, 4, 6, 7

[28] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *NeurIPS*, 36, 2024. 5

[29] Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. Visual knowledge graph for human action reasoning in videos. In *ACM MM*, pages 4132–4141, 2022. 3

[30] Carl D Meyer and Ian Stewart. *Matrix analysis and applied linear algebra*. SIAM, 2023. 4, 1

[31] WonJun Moon, Sangeek Hyun, Sanguk Park, Dongchan Park, and Jae-Pil Heo. Query - dependent video representa-

tion for moment retrieval and highlight detection. In *CVPR*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[32] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, pages 10807–10816, 2020. 2

[33] Giannis Nikolentzos and Michalis Vazirgiannis. Random walk graph neural networks. In *NeurIPS*, 2020. 4

[34] Cristian Rodriguez Opazo, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*, pages 2453–2462, 2020. 2

[35] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 1

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3, 5, 6, 1

[37] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Trans. Assoc. Comput. Linguistics*, 1:25–36, 2013. 2, 5

[38] Sylvia Richardson and Peter J Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4):731–792, 1997. 3

[39] Mrigank Rochan, Mahesh Kumar Krishna Reddy, Linwei Ye, and Yang Wang. Adaptive video highlight detection by learning from user history. In *ECCV*, pages 261–278, 2020. 2

[40] Erica K Shimomoto, Edison Marrese-Taylor, Hiroya Takamura, Ichiro Kobayashi, Hideki Nakayama, and Yusuke Miyao. Towards parameter-efficient integration of pre-trained language models in temporal video grounding. *arXiv preprint arXiv:2209.13359*, 2022. 1

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[42] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, pages 5179–5187, 2015. 5

[43] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *ECCV*, pages 787–802, 2014. 5, 7

[44] Xin Sun, Xuan Wang, Jialin Gao, Qiong Liu, and Xi Zhou. You need to read again: Multi-granularity perception network for moment retrieval in videos. In *SIGIR*, pages 1022–1032, 2022. 2

[45] Haoyu Tang, Jihua Zhu, Meng Liu, Zan Gao, and Zhiyong Cheng. Frame-wise cross-modal matching for video moment retrieval. *TMM*, 24:1338–1349, 2022. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 5

[47] Jiangshan Wang, Yifan Pu, Yizeng Han, Jiayi Guo, Yiru Wang, Xiu Li, and Gao Huang. Gra: Detecting oriented objects through group-wise rotating and attention. *arXiv preprint arXiv:2403.11127*, 2024. 5

[48] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N Metaxas. Learning trailer moments in full-length movies with co-contrastive attention. In *ECCV*, pages 300–316, 2020. 7

[49] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, pages 334–343, 2019. 1

[50] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *AAAI*, pages 2986–2994, 2021. 2

[51] Yicheng Xiao, Yue Ma, Shuyan Li, Hantao Zhou, Ran Liao, and Xiu Li. Semanticac: semantics-assisted framework for audio classification. In *ICASSP*, pages 1–5. IEEE, 2023. 5

[52] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *CVPR*, pages 1258–1267, 2019. 1, 2, 7

[53] Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. Cross-category video highlight detection via set-based learning. In *ICCV*, pages 7950–7959, 2021. 7

[54] Yifang Xu, Yunzhuo Sun, Yang Li, Yilei Shi, Xiaoxiang Zhu, and Sidan Du. Mh-detr: Video moment and highlight detection with cross-modal transformer. *arXiv preprint arXiv:2305.00355*, 2023. 6

[55] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *ICCV*, pages 17503–17512, 2023. 3

[56] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*, pages 982–990, 2016. 1, 2

[57] Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *ICCV*, pages 7950–7959, 2021. 7

[58] Youngjae Yu, Sangho Lee, Joonil Na, Jaeyun Kang, and Gunhee Kim. A deep ranking model for spatio-temporal highlight detection from a 360 video. In *AAAI*, pages 7525–7533, 2018. 1, 2

[59] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, pages 9159–9166, 2019. 2

[60] Yawen Zeng, Ning Han, Keyu Pan, and Qin Jin. Temporally language grounding with multi-modal multi-prompt tuning. *TMM*, 2023. 1

[61] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, pages 1247–1257, 2019. 1

[62] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, pages 6543–6554, 2020. 2, 6

[63] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, pages 766–782, 2016. 7

[64] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, pages 12870–12877, 2020. 6, 1

[65] Hantao Zhou, Rui Yang, Yachao Zhang, Haoran Duan, Yawen Huang, Runze Hu, Xiu Li, and Yefeng Zheng. Unihead: unifying multi-perception for detection heads. *arXiv preprint arXiv:2309.13242*, 2023. 5