

# UMT: Unified Multi-modal Transformers for Joint Video Moment Retrieval and Highlight Detection

Ye Liu<sup>1</sup> Siyuan Li<sup>2</sup> Yang Wu<sup>2\*</sup> Chang Wen Chen<sup>1,4</sup> Ying Shan<sup>2</sup> Xiaohu Qie<sup>3</sup>

<sup>1</sup> Department of Computing, The Hong Kong Polytechnic University

<sup>2</sup> ARC Lab, Tencent PCG <sup>3</sup> Tencent PCG <sup>4</sup> Peng Cheng Laboratory

csyeliu@comp.polyu.edu.hk, changwen.chen@polyu.edu.hk

{siyuanli, dylanywu, yingsshan, tigerqie}@tencent.com

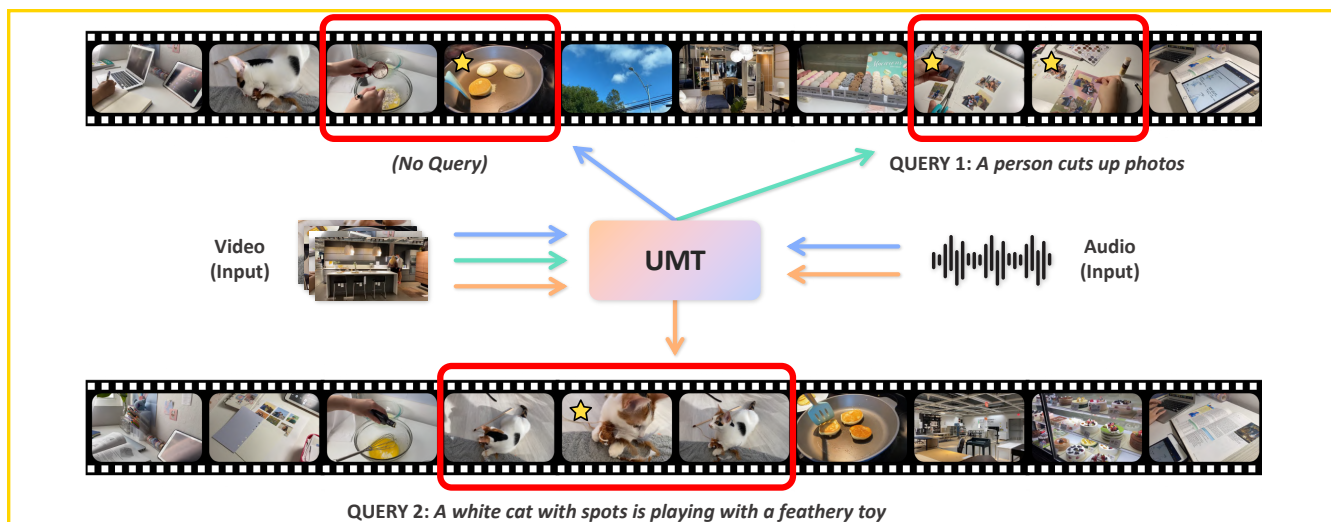


Figure 1. The proposed UMT is a unified and flexible framework which can handle different input modality combinations, and output video moment retrieval and/or highlight detection results (marked by red rectangles and golden stars, respectively). Note that different text queries lead to different outcomes from the same video. Arrows in different colors denote different input-output combinations.

## Abstract

Finding relevant moments and highlights in videos according to natural language queries is a natural and highly valuable common need in the current video content exploration era. Nevertheless, jointly conducting moment retrieval and highlight detection is an emerging research topic, even though its component problems and some related tasks have already been studied for a while. In this paper, we present the first unified framework, named Unified Multi-modal Transformers (UMT), capable of realizing such joint optimization while can also be easily degenerated for solving individual problems. *As far as we are aware, this is the first scheme to integrate multi-modal (visual-audio) learning for either joint optimization or the individual moment retrieval task, and tackles moment retrieval as a keypoint*

detection problem using a novel query generator and query decoder. Extensive comparisons with existing methods and ablation studies on QVHighlights, Charades-STA, YouTube Highlights, and TVSum datasets demonstrate the effectiveness, superiority, and flexibility of the proposed method under various settings. Source code and pre-trained models are available at <https://github.com/TencentARC/UMT>.

## 1. Introduction

Video has already become the major media in content production, distribution, and consumption in our daily lives. It has the unique advantage of being able to include visual, audio, and linguistic information in the same media, *in line with* our natural experiences. Such an advantage on information richness, however, is also a challenging factor limiting its production and consumption, as it *brings about* very

\* Corresponding author.

high costs on satisfying two critical needs. The first one is to find relevant moments in existing videos for producing new content or just getting creation hints from such references. The second one is to glance at large amounts of video content quickly by scanning video highlights rather than watching the entire original videos or video moments at a normal speed, which is needed by both video producers and consumers in such a content explosion era.

These two critical needs lead to two important research topics: video moment retrieval [1, 8] and video highlight detection [32, 37, 44], respectively. Although one may realize that these two tasks are closely related (especially when a text query is given), they have not yet been explicitly jointly studied until a very recent work [18] which builds a novel dataset called QVHighlights for this purpose and presents the first model Moment-DETR optimized for jointly solving both problems. Nevertheless, this seminal work has several limitations. It assumes a text query always exists and it has only applied to the visual modality of each video. Moreover, it is still a very basic model called a strong baseline, although it adopts a transformer framework, the latest and fast-rising neural network architecture type.

This paper goes deeper into designing joint video moment retrieval and highlight detection approaches by mainly exploring two aspects: *multi-modal learning* and *flexibility*, as shown in Figure 1. Apart from text and video (*i.e.* visual information), audio is also treated as an important input. Moreover, a unified yet flexible framework called Unified Multi-modal Transformers (UMT) is proposed to handle different modality reliability situations and combinations. For example, when the text input is unavailable, the task degenerates to highlight detection only. When there is some significant distraction in the text, its reliability will be compromised. Moreover, the audio may also be noisy, which may limit effective exploration. UMT covers all these natural variations which conventionally need to be resolved by different specifically designed models.

To demonstrate the effectiveness and superiority of the proposed framework, we conduct experiments not only on the QVHighlights dataset [18], the only one built for joint video moment retrieval and highlight detection, but also on popular public datasets for moment retrieval (Charades-STA [8]) and highlight detection (TVSum [31], YouTube Highlights [32]), with or without text guidance. For each case, we compare the proposed scheme with several state-of-the-art approaches. Detailed ablation studies are also carried out to evaluate the essential components of the proposed scheme and to reveal meaningful insights.

## 2. Related Works

**Video Moment Retrieval & Highlight Detection** Moment retrieval is a recently studied research topic that focuses on retrieving related moments in a video given a nat-

ural language query. Most existing works [1, 8] assume there is only a single moment in a video corresponding to a given text query, and such queries are usually about activities. The recently proposed QVHighlights dataset [18] goes beyond that by annotating multiple moments in a video for each query and breaks the former moment distribution bias (locating more likely at the beginning of the videos). Video retrieval via text query [24] is a similar task, but it retrieves whole videos rather than video moments. Some works on language grounding [27, 34] align textual phrases to temporal video segments, which operate at a finer scale than moment retrieval and target at different applications. Highlight detection concerns about detecting interesting or salient video segments (*i.e.* highlights) in a video. It has a long history of about two decades with a rich literature, covering various domains of videos, including sports [33, 37], social media [32], and first-person [44]. QVHighlights is the only dataset supporting highlight detection conditioned on text-guided moment retrieval results. Video summarization is a closely related task that targets at summarizing a long video with short video segments. It focuses on representativeness, diversity, and storyline, and thus it tends to be considered as a downstream application of highlight detection [44]. Dynamic video thumbnail generation is another downstream task, which selects attractive video highlights and reforms them into a very short segment [42] to serve as the thumbnail. Among all these tasks, moment retrieval and highlight detection are two fundamental ones and they get highly correlated when a text query is given. This study follows the seminal work of QVHighlights on modeling both tasks together within a single framework. Different from Moment-DETR [18], our model has the flexibility to perform moment retrieval or highlight detection only.

**Text Query Based Models** While text query is a must for moment retrieval, it seldom appears in the studies for video highlight detection, though we believe that providing text queries leads to better results as highlights are usually subjective and interest-dependent. An early work [16] proposes using text to find video highlights, but it is just about using a text ranking algorithm to rank video descriptions in the text domain for providing supervision to video shot ranking, not directly matching text and highlights. The only text-guided highlight detection exists in the very recent work [18]. In the closely related field of video thumbnail generation, text queries were first investigated in [46], where a graph convolutional network is used to model the word-by-clip interactions. Later on, sentence-guided temporal modulation mechanism [28] is proposed to modulate an encoder-decoder based network. All these works assume the reliability of the text query and have to rely on it, while our proposed framework can easily work without text queries or with unreliable text queries.

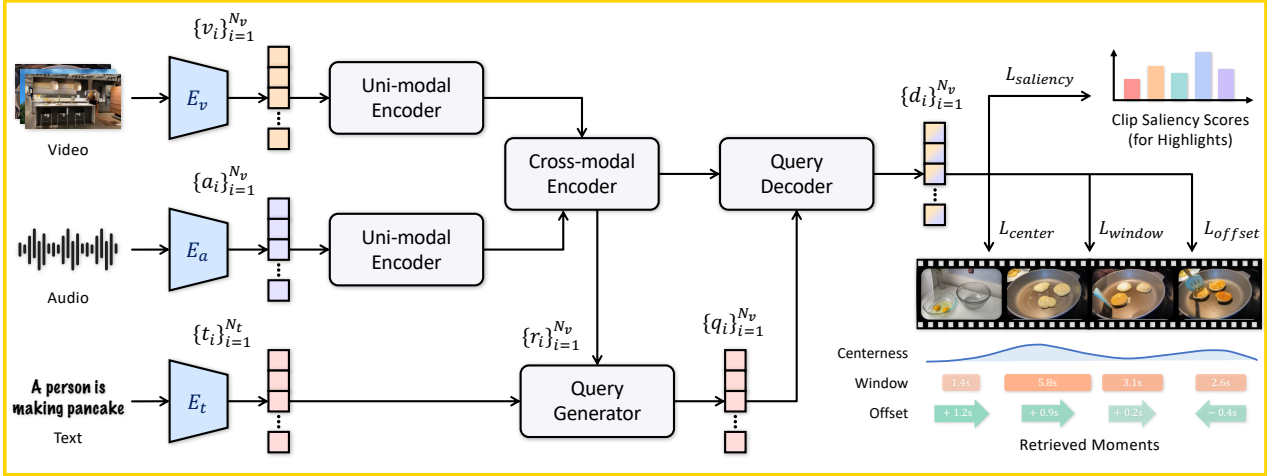


Figure 2. Overall architecture of our framework. When either the video or audio is unavailable, the corresponding uni-modal encoder and cross-modal encoder are deactivated. If text queries are not provided, the model would simply use learnable moment queries instead. Detailed explanations of notations are described in Section 3.1.

**Multi-modal Learning** Recently, multi-modal learning approaches have been explored for highlight detection by jointly modeling visual and audio modalities. The earliest work seems to be MINI-Net [12], which simply concatenates the feature vectors of both modalities. Very recently, two more sophisticated modality fusion models [3, 45] have been proposed and significantly boosted the performances. One of them invents a visual-audio tensor fusion mechanism [45] to learn cross-modality relationships with **tensor decomposition and low-rank constraints**. The other does the fusion via cross-modal bidirectional attention layers [3] which extract audio-attended visual features and visually-attended audio features. Though all the three approaches share the same idea with us on learning multi-modal fused representations for highlight detection, only the attention-based work build their model under the same supervised-learning setting as ours (the other two are weakly-supervised and thus not fairly comparable). Moreover, to the best of our knowledge, our approach is the first one to solve joint moment retrieval and highlight detection with multi-modal (visual-audio) learning.

## 3. Method

### 3.1. Overview

Given an untrimmed video  $V$  containing  $N_v$  clips and a natural language query  $T$  with  $N_t$  tokens, the goal of joint video moment retrieval and highlight detection is to localize all the moments (represented by temporal boundaries  $b \in \mathbb{R}^2$ ) in  $V$ , in which the visual and/or audio contents are highly relevant to  $T$ , while predicting clip-level saliency scores  $\{s_i\}_{i=1}^{N_v}$  for each moment simultaneously.

As shown in Figure 2, the overall architecture of our framework derives from the transformer encoder-decoder

structure, and can be divided into five parts, *i.e.* **uni-modal encoder**, **cross-modal encoder**, **query generator**, **query decoder**, and **prediction heads**. The input video and text are firstly processed by pre-trained feature extractors. Specifically, we use three different models ( $E_v$ ,  $E_a$ , and  $E_t$ ) to extract visual, audio, and textual features, respectively. Each video-text pair can be therefore represented by three collections of feature vectors, namely visual features  $\{v_i\}_{i=1}^{N_v}$ , audio features  $\{a_i\}_{i=1}^{N_v}$ , and textual features  $\{t_i\}_{i=1}^{N_t}$ . The visual and audio features are fed into separate uni-modal encoders to be contextualized under global receptive field, then be fused by the cross-modal encoder for visual-audio joint representations  $\{r_i\}_{i=1}^{N_v}$ . These representations, together with textual features, are used to generate clip-level moment queries  $\{q_i\}_{i=1}^{N_v}$  that can be utilized to retrieve moments and highlights from joint representations in the query decoder. After decoding query-guided video features  $\{d_i\}_{i=1}^{N_v}$ , we use two prediction heads to obtain the final moment retrieval and highlight detection results.

### 3.2. Uni-modal Encoder

Most existing feature extractors for videos and audios [4, 15, 35] are under the sliding window scheme, thus these methods only consider local temporal correlations, without being aware of the global context information, which is of great essence for video understanding tasks. Detecting queried moments and highlights in a video also requires an overall view of the global content. Therefore, to augment the features with global context within each modality, we adopt uni-modal encoder to process the input visual and audio features. This module is constructed by stacking standard transformer encoder layers [36], each consisting of a multi-head self-attention block and a feed-forward network. In each attention head, self-attention for either of the visual

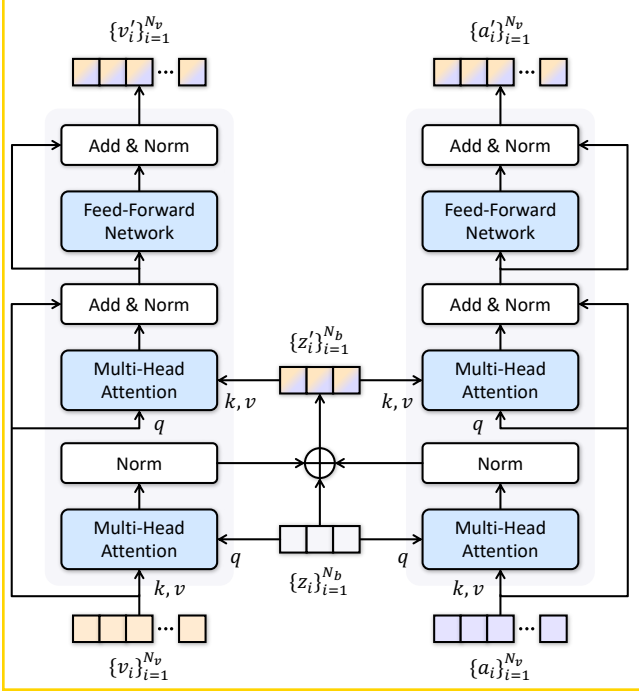


Figure 3. The architecture of bottleneck transformer module. We introduce bottleneck tokens for cross-modal feature compression and expansion, largely reducing the computational cost.

or audio modality  $x \in \{v, a\}$  can be computed as

$$x'_i = x_i + w_z \sum_{j=1}^{N_v} \frac{\exp(w_q x_i \times w_k x_j)}{\sum_{m=1}^{N_v} \exp(w_q x_i \times w_k x_m)} w_v x_j \quad (1)$$

where  $x_i$  and  $x'_i$  are the input and output features of clip  $i$ , and  $w_{\{q,k,v,z\}}$  indicates linear transform weights for the query, key, value, and output matrices. More details about self-attention computation are referred in [36]. The above formula computes the embedded gaussian correlations among clips, and aggregates the global context information into each clip. After aggregating the features, subsequently, a two-layer feed-forward network formed by Linear  $\rightarrow$  ReLU  $\rightarrow$  Dropout  $\rightarrow$  Linear is used to further project the features.

### 3.3. Cross-modal Encoder

Previous works [3, 23] have claimed that jointly modeling multi-modal features can better obtain the overall representations. Hence after the uni-modal encoders, an extra cross-modal encoder is utilized to jointly capture the global correlations across modalities. Here, the exact form of the cross-modal encoder is not crucial. A straightforward approach is to apply cross-modal attention [3]. However, such a strategy has two weaknesses. First, as typical natural signals, both visual and audio features have heavy spatial-temporal redundancy and noisy information that are use-

less for other modalities. Second, the computation of cross-modal attention is costly, with square complexity when calculating clip-to-clip correlations. A recent work [23] tried to tackle the first problem by introducing **attention bottlenecks** that can be regarded as the information bridge across modalities. Although promising results have been achieved, this module still suffers from high computational cost since inter- and cross-modal correlations are jointly modeled. In this work, we extend this idea and propose to disentangle these strategies, thus the resulting bottleneck transformer module can be divided into two stages, *i.e.* feature compression and expansion, shown in Figure 3.

**Feature Compression** Following [23], we introduce bottleneck tokens  $\{z_i\}_{i=1}^{N_b}$  to capture the compressed features from all modalities. Here  $N_b$  is a number much smaller than the number of video clips  $N_v$ . The feature compression is realized by several multi-head attentions between bottleneck tokens and the features from different modalities. Since there are only visual and audio modalities in this case, the compression process can be represented by

$$z'_i = z_i + w_z \sum_{j=1}^{N_v} \frac{\exp(w_q z_i \times w_k x_j)}{\sum_{m=1}^{N_v} \exp(w_q z_i \times w_k x_m)} w_v x_j \quad (2)$$

where  $z_i$  and  $z'_i$  are input and output features of bottleneck tokens. Other notations are consistent with Eq. 1. The only difference between Eq. 1 and Eq. 2 is that the query matrix is replaced by  $z_i$ , aiming to aggregate features into bottleneck tokens. We apply this operation for both visual and audio features, so that multi-modality information is refined and compressed into bottleneck tokens.

**Feature Expansion** After compressing the multi-modal information, we expand the features and propagate them into each modality using another multi-head attention. Formally, the computation is as follows.

$$x'_i = x_i + w_z \sum_{j=1}^{N_v} \frac{\exp(w_q x_i \times w_k z_j)}{\sum_{m=1}^{N_v} \exp(w_q x_i \times w_k z_m)} w_v z_j \quad (3)$$

Here,  $x'_i$  represents the cross-modality enhanced features of clip  $i$ . These features are then fed into feed-forward networks for further projection. Leveraging such a two-stage feature propagation across modalities, **visual and audio features are augmented under linear computational complexities, without incorporating noisy information.**

### 3.4. Query Generator

As transformers are firstly introduced for language translation tasks, the lengths of the input and output sequences may not be the same, **where the length of the output sequence is determined by the query embeddings** fed into the



Table 1. Experimental results on QVHighlights *test* split. MR and HD represent moment retrieval and highlight detection, respectively. w/ PT means pre-training with ASR captions.

Method	MR					HD	
	R1		mAP			$\geq$ Very Good	mAP HIT@1
	@0.5	@0.7	@0.5	@0.75	Avg.		
BeautyThumb [30]	—	—	—	—	—	14.36	20.88
DVSE [21]	—	—	—	—	—	18.75	21.79
MCN [1]	11.41	2.72	24.94	8.22	10.67	—	—
CAL [6]	25.49	11.54	23.40	7.65	9.89	—	—
XML [19]	41.83	30.35	44.63	31.73	32.14	34.49	55.25
XML+ [18]	46.69	33.46	47.89	34.67	34.90	35.38	55.06
Moment-DETR [18]	52.89	33.02	54.82	29.40	30.73	35.69	55.60
Moment-DETR w/ PT	59.78	40.33	<b>60.51</b>	35.36	36.14	37.43	60.17
UMT (Ours)	56.23	41.18	53.83	37.01	36.12	38.18	59.99
UMT (Ours) w/ PT	<b>60.83</b>	<b>43.26</b>	57.33	<b>39.12</b>	<b>38.08</b>	<b>39.12</b>	<b>62.39</b>

decoder. When generalized to vision tasks, query embeddings are randomly initialized and learned during training. Such a scheme may not be suitable for video highlight detection, since the outputs ought to be strictly aligned with the input tokens. Besides, query embeddings shall naturally guide the process of representation decoding. Therefore, we introduce a query generator to adaptively generate temporally aligned moment queries depending on the natural language input. This module is also constructed by a multi-head attention layer, in which visual-audio joint representations  $\{r_i\}_{i=1}^{N_v}$  act as query, textual features are key and value. Our hypothesis is that by computing the attention weights between video clips and text queries, each clip can learn whether it contains which of the concepts described in the text, and predict a query embedding that can be used to decode the learned information for different needs. Note that when text queries are not available, joint representations and learnable positional encodings are summed up to serve as **moment queries** instead.

### 3.5. Query Decoder and Prediction Heads

Query decoder takes visual-audio joint representations  $\{r_i\}_{i=1}^{N_v}$  and text-guided moment queries  $\{q_i\}_{i=1}^{N_v}$  as inputs, and decodes the video features for joint moment retrieval and highlight detection. The output sequence of the query decoder has the same length as the encoder input. Such a peculiarity has two advantages: 1) We may obtain the clip-level saliency (highlight) scores as simple as adding a linear projection layer with sigmoid activation. 2) The dynamic length of output sequence also enables us to define moment retrieval as a **keypoint detection problem** [17, 50]. That is, each moment can be represented by its temporal center and duration (window), where the center point can be estimated by predicting a temporal heatmap and extracting local maxima. The window can be further regressed from features of

Table 2. Comparison with representative moment retrieval methods on Charades-STA *test* split. All the models use the officially released VGG and/or optical flow features of Charades.

Method	R@1		R@5	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
SAP [5]	27.42	13.36	66.37	38.15
SM-RL [39]	24.36	11.17	61.25	32.08
MAN [47]	41.24	20.54	83.21	51.85
2D-TAN [49]	40.94	22.85	83.84	50.35
FVMR [9]	42.36	24.14	83.97	50.15
UMT <sup>†</sup> (Ours)	48.31	<b>29.25</b>	88.79	<b>56.08</b>
UMT <sup>‡</sup> (Ours)	<b>49.35</b>	26.16	<b>89.41</b>	54.95

<sup>†</sup> video + audio, <sup>‡</sup> video + optical flow

the center. Note that the points in the heatmap are **discrete**, which may be misaligned with the real temporal center and inevitably damage the retrieval performances. An extra offset term used to adjust the center should also be predicted. We adopt four linear projection layers to predict the saliencies, centers, windows, and offsets, respectively.

During training, the clip-level saliency score prediction is optimized using a binary cross-entropy loss  $\mathcal{L}_s$ . For each ground truth moment with center  $p \in [1, N_v]$  and window  $d$ , we **quantize** the center point to  $\tilde{p}$  and fill the heatmap  $H \in [0, 1]^{N_v}$  using a 1D gaussian kernel  $H_x = \exp(-\frac{(x-\tilde{p})^2}{2\sigma_p^2})$ , where  $x$  is the temporal coordinate and  $\sigma_p$  is the window-adaptive standard deviation. We optimize the center point prediction using the gaussian **focal** loss [20] as

$$\mathcal{L}_c = -\frac{1}{N} \sum_x \begin{cases} (1 - \hat{H}_x)^\alpha \log(\hat{H}_x) & \text{if } H_x = 1 \\ (1 - H_x)^\gamma \hat{H}_x^\alpha \log(1 - \hat{H}_x) & \text{otherwise} \end{cases} \quad (4)$$

Here,  $N$  is the number of moments,  $\alpha$  and  $\gamma$  denote the **weighting** and the exponent of the modulating factors in the **focal loss**, which are set to 2.0 and 4.0 in practice. For window and offset regression, we simply adopt L1 losses to optimize the actual values for all ground truth centers as

$$\mathcal{L}_w = -\frac{1}{N} \sum_p |w_p - \hat{w}_p| \quad (5)$$

$$\mathcal{L}_o = -\frac{1}{N} \sum_p |(o_p - \tilde{p}) - \hat{o}_p| \quad (6)$$

where  $w_p$ ,  $\hat{w}_p$ ,  $o_p$ , and  $\hat{o}_p$  are the ground truth and predicted windows and offsets respectively. The overall training loss would be the weighted sum of all the losses above as

$$\mathcal{L} = \lambda_s \mathcal{L}_s + \lambda_c \mathcal{L}_c + \lambda_w \mathcal{L}_w + \lambda_o \mathcal{L}_o \quad (7)$$

where  $\lambda_{\{s,c,w,o\}}$  are the weights for saliency, center, window, and offset losses, respectively. When testing, the moment boundaries are obtained by combining the center, window, and offset terms as introduced in [17, 50].

Table 3. Experimental results on YouTube Highlights (metric: mAP). Above are the methods using visual features only, the others are using visual-audio features.

Method	Dog	Gym.	Par.	Ska.	Ski.	Sur.	Avg.
RRAE [43]	49.0	35.0	50.0	25.0	22.0	49.0	38.3
GIFs [11]	30.8	33.5	54.0	55.4	32.8	54.1	46.4
LSVM [32]	60.0	41.0	61.0	62.0	36.0	61.0	53.6
LIM-S [40]	57.9	41.7	67.0	57.8	48.6	65.1	56.4
SL-Module [41]	<b>70.8</b>	53.2	77.2	<b>72.5</b>	66.1	76.2	69.3
MINI-Net [12]	58.2	61.7	70.2	72.2	58.7	65.1	64.4
TCG [45]	55.4	62.7	70.9	69.1	60.1	59.8	63.0
Joint-VA [3]	64.5	71.9	80.8	62.0	<b>73.2</b>	78.3	71.8
UMT (Ours)	65.9	<b>75.2</b>	<b>81.6</b>	71.8	72.3	<b>82.7</b>	<b>74.9</b>

Table 4. Comparison with representative highlight detection methods on TVSum (metric: Top-5 mAP). Above are the methods using visual features only, the others are using visual-audio features.

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Avg.
sLSTM [48]	41.1	46.2	46.3	47.7	44.8	46.1	45.2	40.6	47.1	45.5	45.1
SG [22]	42.3	47.2	47.5	48.9	45.6	47.3	46.4	41.7	48.3	46.6	46.2
LIM-S [40]	55.9	42.9	61.2	54.0	60.4	47.5	43.2	66.3	69.1	62.6	56.3
Trailer [38]	61.3	54.6	65.7	60.8	59.1	70.1	58.2	64.7	65.6	68.1	62.8
SL-Module [41]	86.5	68.7	74.9	<b>86.2</b>	79.0	63.2	58.9	72.6	78.9	64.0	73.3
MINI-Net [12]	80.6	68.3	78.2	81.8	78.1	65.8	57.8	75.0	80.2	65.5	73.2
TCG [45]	85.0	71.4	81.9	78.6	80.2	75.5	71.6	77.3	78.6	68.1	76.8
Joint-VA [3]	83.7	57.3	78.5	86.1	80.1	69.2	70.0	73.0	<b>97.4</b>	67.5	76.3
UMT (Ours)	<b>87.5</b>	<b>81.5</b>	<b>88.2</b>	78.8	<b>81.4</b>	<b>87.0</b>	<b>76.0</b>	<b>86.9</b>	84.4	<b>79.6</b>	<b>83.1</b>

## 4. Experiments

### 4.1. Datasets and Experimental Settings

**Datasets** QVHighlights [18] is the only existing public dataset that has ground-truth annotations for both moment retrieval and highlight detection, thus being suitable for evaluating the full version of our proposed model. This dataset contains videos cropped into 10,148 short (150s-long) segments, and had each segment annotated with at least one text query depicting its relevant moments. There are averagely about 1.8 disjoint moments per query, annotated on non-overlapping 2s-long clips. In total, there are 10,310 queries with 18,367 annotated moments. We follow the original QVHighlights data splits in all experiments.

We also utilize three more datasets: Charades-STA [8], YouTube Highlights [32], and TVSum [31] for further evaluation on the moment retrieval or highlight detection task only, as our model has the flexibility for tasks. Charades-STA contains 16,128 query-moment pairs annotating different actions. YouTube Highlights has 6 domains with 433 videos currently available. TVSum includes 10 domains with 5 videos each. We follow the tradition to work on a random 0.8/0.2 training/test split. Note that the annotators for TVSum were aware of the video titles, so we believe that these titles can serve as noisy text queries. Our model’s flexibility to handle this situation is studied.

**Evaluation Metrics** We use the same evaluation metrics used in existing works. Specifically, for QVHighlights, Recall@1 with IoU thresholds 0.5 and 0.7, mean average precision (mAP) with IoU thresholds 0.5 and 0.75, and the average mAP over a series of IoU thresholds [0.5:0.05:0.95] are used for moment retrieval. For highlight detection, mAP and HIT@1 are utilized, where a clip prediction is treated as a true positive if it has the saliency score of *Very Good*. For Charades-STA, Recall@1 and Recall@5 with IoU thresholds 0.5 and 0.7 are used. For YouTube Highlights and TVSum, mAP and Top-5 mAP are adopted, respectively.

**Implementation Details** On QVHighlights, we simply leverage the pre-extracted features using SlowFast [7] and CLIP [26]. Official VGG [29] and optical flow features, as well as GloVe [25] embeddings, are used for Charades-STA. On YouTube Highlights and TVSum, we obtain clip-level visual features using an I3D [4] pre-trained on Kinetics 400 [13]. Since each feature vector captures 32 consecutive frames, we follow [3] and consider the feature vector belonging to a clip if their overlap is more than 50%. We also use CLIP to extract the title features in TVSum. Audio features of all the datasets are extracted by a PANN [15] model pre-trained on AudioSet [10]. Visual and audio features are temporally aligned at clip level.

All the models in our experiments contain one uni-modal and cross-modal encoder layer each. The number of decoder layers is set to 3 for QVHighlights and Charades-STA, and 1 for YouTube Highlights and TVSum since they have smaller scales. The number of bottleneck tokens  $N_b$  is insensitive thus be set to 4. The weights of losses are set as  $\lambda_s = 3.0$ ,  $\lambda_c = 1.0$ ,  $\lambda_w = 0.1$ , and  $\lambda_o = 1.0$ , while  $\lambda_w$  and  $\lambda_o$  are reduced to 0.05 and 0.5 specially for Charades-STA. Following [18], we set the hidden dimensions to 256, with  $4 \times$  dimension expansions in feed forward networks. Learnable positional encodings, pre-norm style layer normalizations [2], 8 attention heads, and 0.1 dropout rates are used in all transformer layers. We also adopt extra pre-dropouts with rate 0.5 for visual and audio inputs, and 0.3 for text inputs. In all experiments, we use Adam [14] optimizer with  $1e-3$  learning rate and  $1e-4$  weight decay. The model is trained with batch size 32 for 200 epochs on QVHighlights, batch size 8 for 100 epochs on Charades-STA, batch size 4 for 100 epochs on YouTube Highlights, and batch size 1 for 500 epochs on TVSum, respectively.

### 4.2. Results on Joint Video Moment Retrieval and Highlight Detection

We first evaluate our proposed UMT on QVHighlights test split. The results are shown in Table 1, in comparison with all the other performances ever reported. On both mo-

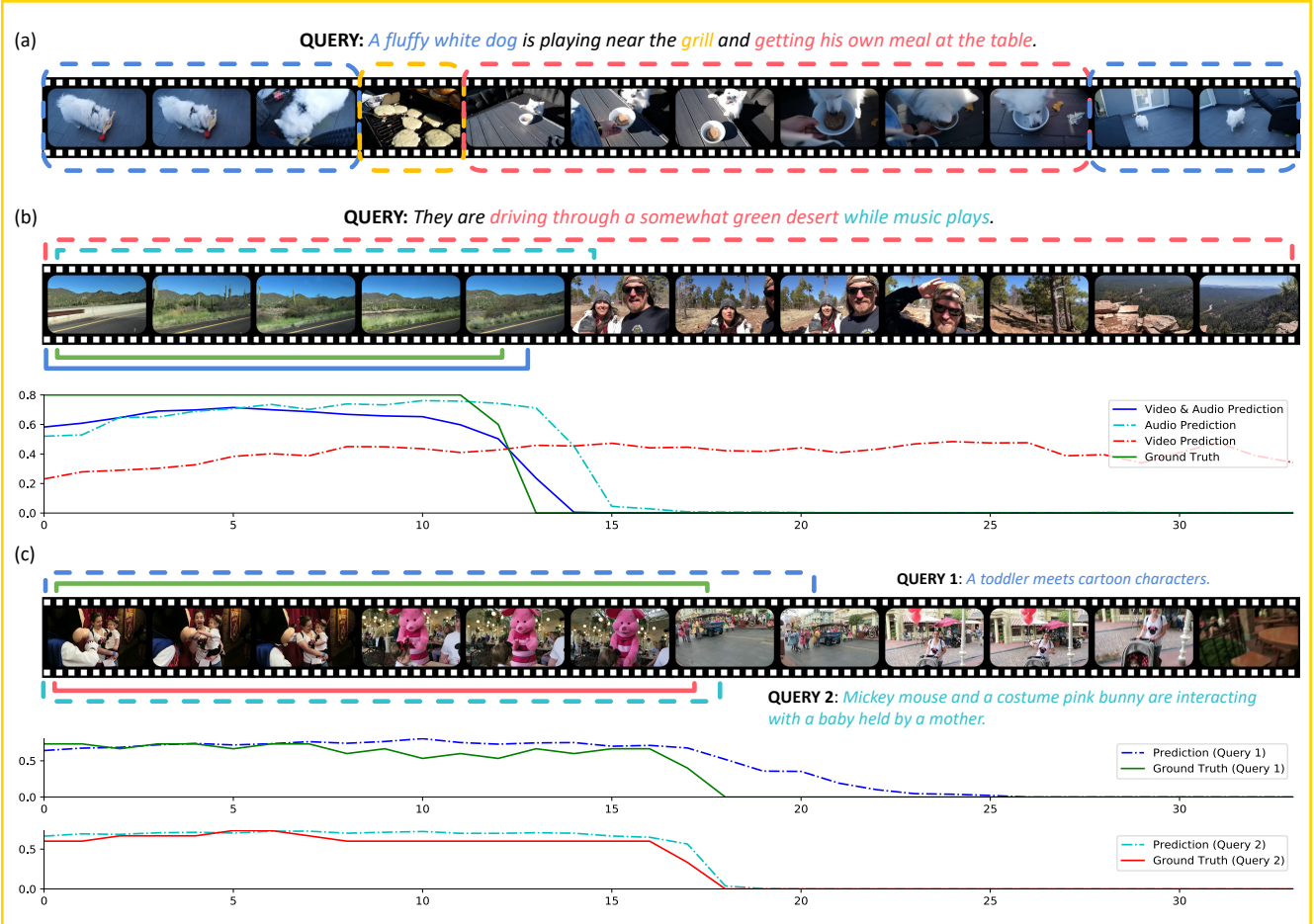


Figure 4. Qualitative results on QVHighlights. The predicted moments and saliency scores are shown by brackets and lines. a) All the highlight clips are presented, indicating that UMT can learn implicit correlations between video content and query semantics. b) Different modality combinations guide the model pay attention to different moments. c) Our model can handle multiple queries in a single video.

ment retrieval and highlight detection tasks, our proposed model outperforms all the existing approaches, including the previous state-of-the-art method Moment-DETR [18] under both settings (with or without pre-training with automatic speech recognition captions). Figure 4 presents some qualitative results of our method on QVHighlights.

### 4.3. Results on Moment Retrieval

Table 2 shows the comparison of UMT with some representative methods on Charades-STA test split. Our approach performs better than previous methods under different metrics. We also tried adopting optical flow instead of audio and obtained similar performances.

### 4.4. Results on Highlight Detection

Highlight detection results on YouTube Highlights and TVSum are presented in Table 2 and Table 3, respectively. On both datasets, UMT performs better than not only representative methods which only use video, but also existing multi-modal ones that utilize both video and audio.

### 4.5. Ablation Studies

**Multi-modality (Visual-Audio)** Table 5 shows the performances of all the multi-modal methods when different modalities are used. Note that Moment-DETR<sup>+</sup> is a multi-modal extension of original Moment-DETR [18] by implementing a similar bottleneck structure as UMT. Clearly, multi-modal learning can significantly boost most methods' performance on all the datasets and tasks in comparison with using a single modality, since it can capture more useful information. Compared with the most similar competitor Moment-DETR<sup>+</sup>, UMT can better explore the complementary information from different modalities and suppress the possible noise during information transfer.

**Multi-task Co-optimization** Given a text query for a video, retrieving the related moments and detecting salient highlights in such moments seem to be highly correlated tasks. Therefore, it is interesting to see how the multi-task co-optimization performs in comparison with training for

Table 5. The effectiveness of multi-modal learning on YouTube Highlights, TVSum, and QVHighlights val split. MR and HD denote moment retrieval and highlight detection, respectively.

Method	YouTube	TVSum	QVHighlights	
	mAP	Top-5 mAP	MR (mAP)	HD (mAP)
MINI-Net <sup>†</sup> [12]	61.38	69.79	–	–
Joint-VA <sup>†</sup> [3]	70.50	74.80	–	–
Moment-DETR <sup>‡</sup> [18]	–	–	32.20	36.52
UMT <sup>†</sup> (Ours)	73.48	81.89	37.79	38.97
MINI-Net <sup>‡</sup> [12]	52.23	59.72	–	–
Joint-VA <sup>‡</sup> [3]	67.00	68.70	–	–
Moment-DETR <sup>‡</sup> [18]	–	–	16.69	26.00
UMT <sup>‡</sup> (Ours)	65.61	76.51	13.73	23.91
MINI-Net [12]	64.36	73.24	–	–
Joint-VA [3]	71.80	76.30	–	–
Moment-DETR <sup>+</sup> [18]	–	–	34.05	37.67
UMT (Ours)	<b>74.93</b>	<b>83.14</b>	<b>38.59</b>	<b>39.85</b>

<sup>†</sup> video only, <sup>‡</sup> audio only, <sup>+</sup> w/ bottleneck transformer

Table 6. Comparison with Moment-DETR using different training task combinations on QVHighlights val split. MR and HD denote moment retrieval and highlight detection, respectively.

Method	Tr. Task(s)		MR			HD	
	MR	HD	R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
Moment-DETR [18]	✓	–	44.84	25.87	25.05	–	–
	–	✓	–	–	–	36.52	56.45
	✓	✓	53.94	34.84	32.20	35.65	55.55
UMT (Ours)	✓	–	54.14	33.82	34.02	–	–
	✓	✓	<b>60.26</b>	<b>44.26</b>	<b>38.59</b>	39.85	64.19

each individual task when the same framework and backbone are used. We conduct single-task experiments by **turning off** the losses corresponding to each task and training the rest of the model. Note that moment retrieval is considered to be a harder task than highlight detection as explained in Section 1. The results in Table 6 show that the co-optimization not only generates results for both tasks simultaneously, but also significantly boosts the performance on moment retrieval. This is clear for both our UMT and Moment-DETR [18]. Note that when training for moment retrieval only, UMT performs much better than Moment-DETR, indicating its superiority on the model design. **Our UMT better models the moment retrieval task as a keypoint detection problem [17, 50] rather than set prediction or clip classification.** Moreover, the inputs to the UMT decoder are clip-aligned text-guided queries instead of positional encodings, which enables more flexible output sequence lengths and may provide stronger query information for each clip. Therefore, we believe that UMT can model the relationship between the two tasks better than Moment-DETR does.

Table 7. Effectiveness justification of the offset loss for moment retrieval on QVHighlights val split. Both models are trained using the co-optimization recipe.

Losses	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
$\mathcal{L}_c + \mathcal{L}_w$	<b>62.32</b>	43.23	<b>57.78</b>	38.61	37.36
$\mathcal{L}_c + \mathcal{L}_w + \mathcal{L}_o$	60.26	<b>44.26</b>	56.70	<b>39.90</b>	<b>38.59</b>

Table 8. The influence of weakly relevant (TVSum) or highly relevant (QVHighlights) text queries on highlight detection.

Text Query	TVSum	QVHighlights	
		mAP	HIT@1
✓	81.42	25.14	33.42
	<b>83.14</b>	<b>39.85</b>	<b>64.19</b>

**Moment Retrieval Losses** Table 7 presents the performances of UMT when different combinations of moment retrieval losses are used. Since the center loss  $\mathcal{L}_c$  and the window loss  $\mathcal{L}_w$  are **mandatory** for representing a moment, only the necessity of the offset loss  $\mathcal{L}_o$  is justified. As the comparison shows, modeling the temporal offset does make the moment boundary prediction more accurate.

**Justification of Text Queries** We believe that highlight detection based on text queries is an important setting for highlight detection, as different interests can favor quite different highlights from the same video. Table 8 reports the results of our model, with or without the text queries. It can be seen that when the query is relevant, it does improve the highlight detection performance, and such an improvement is more significant when the relevance is greater.

## 5. Conclusion

This paper introduces a novel and also the first framework for solving joint moment retrieval and highlight detection as well as its individual component problems in a unified way. It is also the first to integrate multi-modal learning into its model for such a purpose. The effectiveness and superiority of the proposal have been demonstrated on diverse and representative public datasets, in comparison with relevant methods under various settings. The framework is robust to modality quality variations and also flexible enough to work under different text query conditions.

## Acknowledgements

This research is supported in part by Key-Area Research and Development Program of Guangdong Province, China with Grant 2019B010155002 and financial support from ARC Lab, Tencent PCG.



## References

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5803–5812, 2017. 2, 5
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 6
- [3] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8127–8137, 2021. 3, 4, 6, 8
- [4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 3, 6
- [5] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8199–8206, 2019. 5
- [6] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. Technical Report arXiv:1907.12763, 2019. 5
- [7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. 6
- [8] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5267–5275, 2017. 2, 6
- [9] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1523–1532, 2021. 5
- [10] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 6
- [11] Michael Gygli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1001–1009, 2016. 6
- [12] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. Mini-net: Multiple instance rank- ing network for video highlight detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 345–360, 2020. 3, 6, 8
- [13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. Technical Report arXiv:1705.06950, 2017. 6
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 6
- [15] 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 Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 3, 6
- [16] Sukanya Kudi and Anoop M Namboodiri. Words speak for actions: Using text to find video highlights. In *Proceedings of the IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 322–327, 2017. 2
- [17] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 5, 8
- [18] Jie Lei, Tamara L Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 5, 6, 7, 8
- [19] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2020. 5
- [20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 5
- [21] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3707–3715, 2015. 5
- [22] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 202–211, 2017. 6
- [23] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4
- [24] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 651–667, 2016. 2
- [25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 6
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language super-

- vision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 6
- [27] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzl, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 2
- [28] Mrigank Rochan, Mahesh Kumar Krishna Reddy, and Yang Wang. Sentence guided temporal modulation for dynamic video thumbnail generation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 2
- [29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–14, 2015. 6
- [30] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 659–668, 2016. 5
- [31] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187, 2015. 2, 6
- [32] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 787–802, 2014. 2, 6
- [33] Hao Tang, Vivek Kwatra, Mehmet Emre Sargin, and Ullas Gargi. Detecting highlights in sports videos: Cricket as a test case. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2011. 2
- [34] Stefanie Tellex and Deb Roy. Towards surveillance video search by natural language query. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, pages 1–8, 2009. 2
- [35] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 3
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 3, 4
- [37] Jinjun Wang, Changsheng Xu, Engsiong Chng, and Qi Tian. Sports highlight detection from keyword sequences using hmm. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 599–602, 2004. 2
- [38] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N. Metaxas. Learning trailer moments in full-length movies. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–316, 2020. 6
- [39] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 334–343, 2019. 5
- [40] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1258–1267, 2019. 6
- [41] Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. Cross-category video highlight detection via set-based learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7970–7979, 2021. 6
- [42] Yi Xu, Fan Bai, Yingxuan Shi, Qiuyu Chen, Longwen Gao, Kai Tian, Shuigeng Zhou, and Huyang Sun. Gif thumbnails: Attract more clicks to your videos. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3074–3082, 2021. 2
- [43] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4633–4641, 2015. 6
- [44] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 982–990, 2016. 2
- [45] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7950–7959, 2021. 3, 6
- [46] Yitian Yuan, Lin Ma, and Wenwu Zhu. Sentence specified dynamic video thumbnail generation. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 2332–2340, 2019. 2
- [47] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1247–1257, 2019. 5
- [48] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016. 6
- [49] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12870–12877, 2020. 5
- [50] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 8