TR-DETR: Task-Reciprocal Transformer for Joint Moment Retrieval and Highlight Detection

Hao Sun[1,2,3*], **Mingyao Zhou**[1,2,3*], **Wenjing Chen**[4†], **Wei Xie**[1,2,3†]

[1]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
Central China Normal University, Wuhan, China
[2]School of Computer Science, Central China Normal University, Wuhan, China
[3]National Language Resources Monitoring and Research Center for Network Media,
Central China Normal University, Wuhan, China
[4]School of Computer Science, Hubei University of Technology, Wuhan, China
haosun@ccnu.edu.cn, zhoumingyao@mails.ccnu.edu.cn, chenwenjing@hbut.edu.cn, XW@mail.ccnu.edu.cn

Abstract

Video moment retrieval (MR) and highlight detection (HD) based on natural language queries are two highly related tasks, which aim to obtain relevant moments within videos and highlight scores of each video clip. Recently, several methods have been devoted to building DETR-based networks to solve both MR and HD jointly. These methods simply add two separate task heads after multi-modal feature extraction and feature interaction, achieving good performance. Nevertheless, these approaches underutilize the reciprocal relationship between two tasks. In this paper, we propose a task-reciprocal transformer based on DETR (TR-DETR) that focuses on exploring the inherent reciprocity between MR and HD. Specifically, a local-global multi-modal alignment module is first built to align features from diverse modalities into a shared latent space. Subsequently, a visual feature refinement is designed to eliminate query-irrelevant information from visual features for modal interaction. Finally, a task cooperation module is constructed to refine the retrieval pipeline and the highlight score prediction process by utilizing the reciprocity between MR and HD. Comprehensive experiments on QVHighlights, Charades-STA and TVSum datasets demonstrate that TR-DETR outperforms existing state-of-the-art methods. Codes are available at https://github.com/mingyao1120/TR-DETR.

Introduction

With the ubiquity of digital devices and the expansion of the Internet, the number and variety of videos are rapidly increasing (Foo et al. 2023). How to quickly search out the desired moments from massive videos (called moment retrieval, MR) (Gao et al. 2017) and efficiently browse videos (called highlight detection, HD) (Molino and Gygli 2018) according to the needs of users has attracted widespread attention. In practical applications, user needs can be expressed in natural language queries (Wang et al. 2022). Due to the complexity of video content as well as the diversity of user needs, MR&HD based on user-provided natural language queries is extremely challenging.

The goal of MR is to precisely search for semantically related moments from whole videos guided by natural language queries (Li et al. 2022). The common pipeline of MR involves several steps. Firstly, pre-trained networks are utilized to extract features from the input video and text. Subsequently, cross-modal interaction is performed based on the extracted features to obtain the query relevance score of the candidate moment or the frame-level start-end probability of the relevant moment (Zhang et al. 2023). HD based on queries strives to assign highlight scores to each video clip based on considering the user needs (Guo et al. 2022). Existing methods (Liu et al. 2022b; Xiong and Wang 2023) utilize transformers (Vaswani et al. 2017) or graph neural networks (Scarselli et al. 2008) to perform single-modal feature encoding or cross-modal interaction.

Due to the task similarity between MR and HD based on queries, and the commonality between their methods involving multi-modal feature extraction, feature interaction, etc., some works (Lei, Berg, and Bansal 2021; Lin et al. 2023) have devoted to designing various multi-task networks for joint MR&HD. For example, Moment-DETR (Lei, Berg, and Bansal 2021) pioneers the application of DETR (Carion et al. 2020) for joint MR&HD. QD-DETR (Moon et al. 2023) introduces a query-dependent video representation module, making moment predictions reliant on user queries. MH-DETR (Xu et al. 2023) introduces a pooling operation into the encoder and incorporates a cross-modality interaction module to fuse visual and query features. In these methods, two isolated task heads are added after the shared multi-modal feature extraction and feature interaction modules for joint MR&HD. These methods generally focus on improving the discrimination of multi-modal feature extraction and feature interaction through a multi-task learning scheme, achieving good performance. However, the reciprocity between MR and HD tasks is ignored.

For MR, the highlight scores from HD based on user-provided queries can be utilized to assist in eliminating query-irrelevant clips, thereby boosting moment retrieval accuracy. In turn, for HD based on queries, the results of mo-

*These authors contributed equally.
†Corresponding authors.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ment retrieval can be used to improve the understanding of videos and user needs. Therefore, MR and HD based on queries are reciprocal.

To fully exploit the reciprocal relationship between the two tasks, we propose a task-reciprocal transformer based on DETR, named TR-DETR, for joint MR&HD. Firstly, visual features and textual features are extracted from user-provided videos and queries through pre-trained networks. Then, we introduce a local-global multi-modal alignment module to perform local and global semantic alignment before modal interaction, respectively. This module encourages the model to distinguish video clips that are semantically similar but irrelevant to the query. Subsequently, we propose a visual feature refinement module for modal interaction, which employs aligned textual features to filter out query-irrelevant information in visual features to avoid it interfering with joint features. Finally, to exploit the complementarities between MR and HD, we propose a task cooperation module consisting of HD2MR and MR2HD. The former explicitly infuses highlight score information into the moment retrieval process, enhancing localization accuracy. The latter exploits localization outcomes to derive clip-level relevant scores, offering visual support for highlight detection. Extensive experiments on QVHighlights (Lei, Berg, and Bansal 2021), Charades-STA (Gao et al. 2017) and TVSum (Song et al. 2015) demonstrate that the proposed TR-DETR outperforms the state-of-the-art methods. The contributions of this paper are summarized as follows:

- We highlight the reciprocity between MR and HD. In addition, we introduce an innovative TR-DETR network that leverages this reciprocity between tasks to optimize performance.
- We introduce the local and global alignment regulators. These regulators are designed to facilitate semantic alignment between video clips and the query, which serves to generate discriminative joint representations.
- To explore the intrinsic complementarity between the two tasks, we construct a task cooperation module. This module explicitly exploits the complementarity between MR and HD by injecting highlight scores into the moment retrieval pipeline and using the retrieved moments to refine the initial highlight distribution.

Related Works

MR and HD

Video moment retrieval is originally introduced by the literature (Gao et al. 2017), with the objective of retrieving moments from a video based on a given natural language query. Moment retrieval typically includes two types of methods: proposal-based and proposal-free methods. In the proposal-based methods, candidate moments are initially generated through techniques such as sliding windows (Gao et al. 2017), proposal generation networks (Xu et al. 2019), or 2D-Maps (Zhang et al. 2020). These candidates are subsequently ranked based on the similarity scores to the query, where the candidate with the highest score is used as the result. Although these methods have high accuracy, they necessitate additional pre- and post-processing steps, introducing computational redundancy. Moreover, their performance heavily relies on the quality of candidate moments. On the other hand, proposal-free methods (Ghosh et al. 2019; Zhang et al. 2021; Mun, Cho, and Han 2020) directly predict start-end probabilities for target moments within a video, which eliminates the need to rank a large number of candidate moments, thereby improving training efficiency.

In contrast, highlight detection concentrates on measuring the significance of each clip within a given video. Slightly different from moment retrieval, highlight detection initially is proposed as a single-modal task and does not rely on text queries. However, highlight determination is often a subjective matter and users' preferences should be taken into account. Therefore, the literature (Kudi and Namboodiri 2017) proposes to integrate text queries as supplementary information for highlight detection. Nonetheless, this work relies solely on text ranking algorithms to rank video descriptions in the text domain to guide video clip ranking. It does not entail a direct alignment of text and highlights. Subsequently, in video thumbnail generation, which closely parallels highlight detection, Yuan *et al.* (Yuan, Ma, and Zhu 2019) delves into text queries and uses graph convolutional networks to model the interaction between each clip and text.

Conventionally, moment retrieval and highlight detection are addressed in isolation, lacking an integrated framework for joint learning. Recent research (Lei, Berg, and Bansal 2021) constructs the QVHighlights dataset to facilitate joint learning of MR&HD and proposes a baseline model based on DETR. Building upon this, Liu *et al.* (Liu et al. 2022b) incorporates audio modality into the model, catering to scenarios for missing queries. Additionally, Moon *et al.* (Moon et al. 2023) prioritizes full integration of provided query information into the joint representation, enabling the text to guide both moment retrieval and highlight detection. Different from previous methods, this paper focuses on exploiting the natural reciprocity between two tasks.

Multi-Modal Alignment

Recently, researchers in the multimodal field have focused on constructing contrastive losses to fit the interactions and correspondences between different modalities (Luo et al. 2020; Sun et al. 2020; Miech et al. 2020; Yan et al. 2023). For example, the literature (Ging et al. 2020) introduces a cycle consistency loss to align video clip-level features and query word-level features. Similarly, the literature (Zhang et al. 2022) introduces a multi-level contrast loss to capture multi-granular interactive alignment details within queries and videos, enhancing the performance of moment retrieval. Although these methods share similarities with the multi-modal alignment in our approach, they do not explicitly align the semantic information of different modalities before modality interaction, resulting in insufficient discrimination of joint features.

Method

The overview of TR-DETR is shown in Figure 1. TR-DETR comprises four core modules: feature extraction,

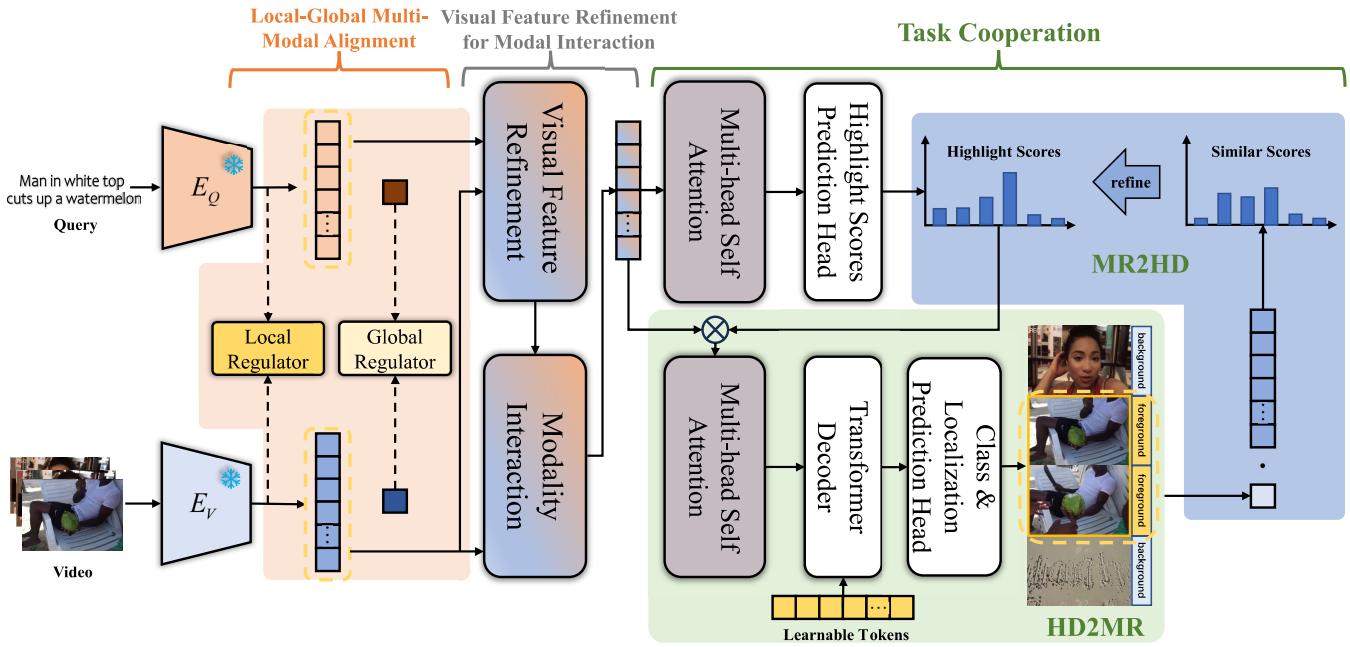

Figure 1: The proposed TR-DETR involves several key steps. Initially, two frozen pre-trained networks are employed to extract visual and textual features from videos and queries. Subsequently, a local-global multi-modal alignment module is constructed to effectively align the extracted visual and textual features. Then, the visual features are refined under the guidance of textual features for obtaining discriminative joint features. Finally, a task cooperation module is implemented to enhance prediction outcomes based on task reciprocity. Additionally, two multi-head self-attention components share weights.

local-global multi-modal alignment, visual feature refinement for modal interaction, and task cooperation. Details are introduced as follows.

Feature Extraction

Visual Features. Following the literature (Lei, Berg, and Bansal 2021), the video is first divided into non-overlapping clips according to a certain time interval, such as 2s. Then the pre-trained ViT-B/32 in CLIP (Radford et al. 2021) and SlowFast (Feichtenhofer et al. 2019) are utilized to extract clip-level visual features $F_v = \left[f_v^1, f_v^2, \ldots, f_v^L\right] \in \mathbb{R}^{L \times d_v}$, where L and d_v are the number of clips and the visual feature dimension, respectively. Following the way that UMT (Liu et al. 2022b) uses audio information, we use the pre-trained audio feature extractor to extract the audio features $F_a \in \mathbb{R}^{L \times d_a}$, and then splice them behind the visual features F_v. See the experimental settings for details.

Textual Features. For a natural language query, we use the textual encoder in the pre-trained CLIP to extract textual features $F_t = \left[f_t^1, f_t^2, \ldots, f_t^N\right] \in \mathbb{R}^{N \times d_t}$, where N and d_t are the number of words and the textual feature dimension, respectively.

Local-Global Multi-Modal Alignment

Existing methods (Moon et al. 2023; Lei, Berg, and Bansal 2021; Liu et al. 2022b) for joint MR&HD directly input the extracted visual and textual features into the modal interaction module to obtain joint features. However, there is a natural information mismatch between visual features and textual features, resulting in insufficient discrimination of joint features (Xu, Zhu, and Clifton 2022). In this study, to reduce the modal gap, we propose a local-global multi-modal alignment module, comprising local and global regularization components. The local regulator helps the model distinguish semantically similar but undesired clips, while the global regulator ensures that both modalities share a unified semantic space. Integrating these alignment regulators can significantly promote multimodal associations and facilitate subsequent modal interactions.

Given the clip-level visual features F_v of the video and the word-level textual features F_t of the query, we first map them into the same dimension d by using three-layer multi-layer perceptions (MLP).

$$\widehat{F}_v = \mathrm{MLP}_v(F_v), \tag{1}$$

$$\widehat{F}_t = \mathrm{MLP}_t(F_t). \tag{2}$$

For the local regulator, we calculate the cosine similarity between each clip and each word by using the following formula, obtaining a similarity matrix $S_{loc} \in \mathbb{R}^{L \times N}$.

$$S_{loc} = \sigma\left(\frac{\widehat{F}_v \, \widehat{F}_t^{\mathsf{T}}}{\|\widehat{F}_v\|_2 \|\widehat{F}_t\|_2}\right), \tag{3}$$

where σ is the sigmoid function. We employ mean-pooling to get $\widehat{S}_{loc} = \mathrm{MeanPooling}(S_{loc}) \in \mathbb{R}^L$, which measures the similarity between each video clip and the global textual

features. Then, a local regular loss $\mathcal{L}_{loc}$ is used to encourage distinguishing video clips that are irrelevant to the query.

$$\mathcal{L}_{local} = -\sum_{i=1}^{L} \left(C^i \log(\widehat{S}^i_{loc}) + (1 - C^i) \log(1 - \widehat{S}^i_{loc}) \right), \tag{4}$$

where $\widehat{S}^i_{loc}$ is the similarity score between the i-th video clip and the global textual features, and C^i indicates whether the i-th video clip and the query are actually relevant. Specifically, according to ground truth in MR, if the i-th clip is relevant to the query, C^i is 1, otherwise 0. For the global regulator, a multi-modal contrastive loss (Li et al. 2021) is employed to promote the similarity of global representations of paired videos and queries.

$$\mathcal{L}_{global} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp((G^i_v)\,(G^i_t)^\mathsf{T})}{\sum_{i=1}^{B} \sum_{j=1}^{B} \exp((G^i_v)\,(G^j_t)^\mathsf{T})}, \tag{5}$$

where B is the batch size, $G^i_v \in \mathbb{R}^d$ and $G^i_t \in \mathbb{R}^d$ are the global feature of the i-th video and the i-th query in a training batch, respectively. Specifically, G^i_v is obtained by averaging all clip features $\widehat{F}_v$ within the i-th video, and G^i_t is derived by averaging word-level features $\widehat{F}_t$ in the i-th query.

Visual Feature Refinement for Modal Interaction

The goal of modal interaction is to generate discriminative joint features from visual and textual features (Lei, Berg, and Bansal 2021), which play a key role in joint MR&HD. In the literature (Lei, Berg, and Bansal 2021), visual and textual features are simply concatenated for modal interaction. However, videos generally contain a large number of clips irrelevant to the textual query, which may cause the model to pay too much attention to these irrelevant contents, resulting in ignoring the really important clips.

To suppress the interference of query-irrelevant information in visual features, we introduce a query-guided visual feature refinement module inspired by the literature (Xiong, Zhong, and Socher 2017) for modal interaction. This module employs the textual query as a guide to refine clip-level visual features to effectively suppress irrelevant information present in the video and retain temporal cues. The similarity matrix between aligned clip-level visual features and word-level textual features is calculated as:

$$A = \frac{\text{Linear}(\widehat{F}_v)\,\text{Linear}(\widehat{F}_t)^\mathsf{T}}{\sqrt{d}}, \tag{6}$$

where $A \in \mathbb{R}^{L \times N}$ is the similarity matrix and Linear$(\cdot)$ represents the linear projection layer. Then the similarity matrix is used to weigh and sum the query and video features respectively to obtain preliminary refinement features.

$$F_{v2q} = A_r\, \widehat{F}_t, \tag{7}$$

$$F_{q2v} = A_r\, A_c^\mathsf{T}\, \widehat{F}_v, \tag{8}$$

where A_r and A_c represent the results after row softmax normalization and column softmax normalization of A, F_{v2q} and F_{q2v} are the clip-level textual features and word-level visual features, respectively. Finally, to further use text queries to optimize clip-level visual features $\widehat{F}_v$, we perform the following feature concatenation and obtain the final refined clip features $\overline{F}_v$ through linear projection.

$$F_v^{Cat} = \left[\widehat{F}_v \| F_{v2q} \| \widehat{F}_v \odot F_{v2q} \| \widehat{F}_v \odot F_{q2v} \| F_t^G \right], \tag{9}$$

$$\overline{F}_v = \text{Linear}(F_v^{Cat}), \tag{10}$$

where $F_t^G \in \mathbb{R}^{L \times d}$ is a matrix formed by copying and splicing the text global features obtained through the pooling operation, $[\cdot\|\cdot]$ means concatenation, and $\odot$ is the Hadamard product. Then, modality fusion is performed using a cross-attention layer to further incorporate query features into the joint features, where textual features are from the refined clip feature $Q_v = \text{Linear}_q(\overline{F}_v)$, key and value features are from the textual features $K_t = \text{Linear}_k(\widehat{F}_t)$ and $V_t = \text{Linear}_v(\widehat{F}_t)$.

$$Z = \text{Attention}(Q_v, K_t, V_t) = \text{Softmax}\left(\frac{Q_v K_t^\mathsf{T}}{\sqrt{d}} \right) V_t, \tag{11}$$

where $Z \in \mathbb{R}^{L \times d}$ represents joint features through modal interaction between refined visual features and textual features.

Task Cooperation

Although previous methods (Lei, Berg, and Bansal 2021; Liu et al. 2022b; Moon et al. 2023) have attempted to jointly solve MR and HD, these methods usually focus on optimizing the shared multi-modal feature extraction and feature interaction modules to improve the discrimination of joint features using a multi-task learning framework. However, the inherent complementarity between MR and HD tasks is underutilized.

In essence, video clips with high highlight scores are often strong candidates for MR. Because highlight-worthy clips tend to possess enhanced visual significance and attraction. Additionally, clips within the moment relevant to the current query probably cover the highlights, too. This is because query-relevant moments also contain visual expressions of user needs, which helps to refine the highlight score distribution from the visual perspective. Given these insights, we propose a task cooperation module consisting of HD2MR and MR2HD components.

HD2MR MR can leverage the highlight scores obtained by HD to empower the exclusion of irrelevant or less attractive video clips. We first use the multi-head attention mechanism and a linear layer to obtain clip-level highlight scores from the joint features Z.

$$H = \text{Linear}(\text{MHA}(Z)), \tag{12}$$

where MHA$(\cdot)$ represents multi-head attention that is employed to model video temporal information and $H \in \mathbb{R}^L$ is the predicted highlight scores.

To filter out non-highlight information in Z and explicitly inject highlight scores information into the MR pipeline, we

multiply the clip-level highlight scores H with the joint features Z to obtain the enhanced joint features $\overline{Z} \in \mathbb{R}^{L\times d}$. Then, $\overline{Z}$ is input into the MHA again for joint features encoding.

$$\begin{aligned} \overline{Z} &= \text{Softmax}(H) \odot Z, \\ \widehat{Z} &= \text{MHA}\left(Z + \overline{Z}\right), \end{aligned} \tag{13}$$

where $\widehat{Z}$ is the joint features of the perceived highlight scores. Finally, these enhanced features $\widehat{Z}$ are fed into the transformer decoder and prediction head from the literature (Liu et al. 2022a) to obtain the ultimate retrieved moments.

MR2HD HD, in turn, gains a deeper understanding of video content and user needs by leveraging the text query and retrieved moments from MR. We employ the gated recurrent unit (GRU) (Chung et al. 2014) to effectively capture global information from the retrieved moments.

$$F_v^M = \text{GRU}(m), \tag{14}$$

where m represents the clip feature vectors in $\widehat{F}_v$ of the retrieved moments from HD2MR and $F_v^M \in \mathbb{R}^d$ is the global feature vector of these retrieved moments. To use the visual information of the retrieved moments to refine highlight scores prediction, we calculate similarity scores between F_v^M and visual features $\widehat{F}_v$.

$$S_{ref} = \frac{F_v^M \, \widehat{F}_v^{\mathsf{T}}}{\|F_v^M\|_2 \, \|\widehat{F}_v\|_2}, \tag{15}$$

where $S_{ref} \in \mathbb{R}^L$ is the correlation between clips and F_v^M. The highlight score refinement process involves multiplying the clip-level correlation scores by $\widehat{Z}$, then adding them to Z, and finally obtaining refined scores by linear projection. The formulation is as follows:

$$\overline{H} = \text{Linear}(Z + \text{Softmax}(S_{ref}) \odot \widehat{Z}), \tag{16}$$

where $\overline{H} \in \mathbb{R}^L$ is the refined highlight scores.

Objective Losses

The objective losses of TR-DETR include three parts: MR loss $\mathcal{L}_{mom}$, HD loss $\mathcal{L}_{high}$, regulators losses $\mathcal{L}_{local}$ and $\mathcal{L}_{global}$.

$$\mathcal{L}_{total} = \mathcal{L}_{mom} + \mathcal{L}_{high} + \lambda_{lg}(\mathcal{L}_{local} + \mathcal{L}_{global}), \tag{17}$$

where λ_{lg} is the coefficient of local-global regulators losses. $\mathcal{L}_{mom}$ and $\mathcal{L}_{high}$ are consistent with QD-DETR (Moon et al. 2023).

Experiment

Datasets

QVHighlights dataset (Lei, Berg, and Bansal 2021) comprises 10,148 content-rich videos from YouTube. Each video is accompanied by at least one manually annotated text query, where the highlight clips are located within the corresponding moment. The evaluation process of this dataset is particularly fair as the annotations of the test set are inaccessible. The prediction results of the model need to be uploaded to the QVHighlights server's CodaLab competition platform[1] for impartial performance assessment.

Charades-STA dataset (Gao et al. 2017) contains 9,848 videos capturing daily indoor activities and 16,128 human-tagged query texts. Following QD-DETR(Moon et al. 2023), we allocate 12,408 samples for training while the remaining 3,720 samples are for testing.

TVSum dataset (Song et al. 2015) is a benchmark dataset for HD. It contains 10 different categories of videos, and each category comprises 5 videos. To ensure consistency with QD-DETR (Moon et al. 2023), 80% of the dataset is utilized for training and the remaining for testing.

Metrics and Experimental Settings

We use common metrics from recent studies like Moment-DETR, UMT, QD-DETR, and MH-DETR. For QVHighlights, we calculate Recall@1 with IoU $\in \{0.5, 0.7\}$ and mean average precision (mAP) with IoU $\in \{0.5, 0.75\}$. Following Lei *et al.* (Lei, Berg, and Bansal 2021), we also uniformly sample 10 IoU thresholds from $\{0.5, 0.95\}$ to calculate mAP, and take the average as the average mAP metric. For highlight detection, we use mAP and HIT@1. Charades-STA involves Recall@1 with IoU $\in \{0.5, 0.7\}$, while for TVSum, top-5 mAP is the main metric.

In addition, we introduce implementation details and hyperparameters as follows. The hidden layer dimension d is 256, and λ_{lg} is set to 0.3. We use PANN (Kong et al. 2020) trained on the AudioSet dataset (Gemmeke et al. 2017) to extract audio features. For QVHighlights, we use Slow-Fast (Feichtenhofer et al. 2019) and CLIP to extract visual features and the text encoder in CLIP to extract textual features. The training phase involves 200 epochs, a batch size of 32, and a learning rate of 1e-4. For TVSum, we use the I3D pre-trained on Kinetics-400 for visual features and CLIP for textual features. Training spans 2000 epochs with a batch size of 4 and a learning rate of 1e-3. In Charades-STA, we extract visual features with VGG (Simonyan and Zisserman 2015), I3D (Carreira and Zisserman 2017), SlowFast, and CLIP, and use GLoVe (Pennington, Socher, and Manning 2014) for textual features. The training phase includes 100 epochs, a batch size of 8, and a learning rate of 1e-4. Moreover, all our experiments are conducted on Nvidia RTX 4090 and Gen Intel(R) Core(TM) i7-12700 CPU.

Comparison with Other Methods

Table 1 reports the TR-DETR's performance on joint moment retrieval and highlight detection tasks. Meanwhile, Tables 2 and 3 list the results of different methods on moment retrieval and highlight detection, respectively.

In Table 1, we evaluate the performance of moment retrieval and highlight detection simultaneously based on the QVHighlights dataset. For a fair comparison, we compare the performance with UniVTG (Lin et al. 2023) without pre-training. As shown in Table 1, our TR-DETR method

[1] https://codalab.lisn.upsaclay.fr/competitions/6937

Method	Src	Moment Retrieval					HD	
		R1		mAP			≥Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
BeautyThumb (Song et al. 2016)	V	-	-	-	-	-	14.36	20.88
DVSE (Liu et al. 2015)	V	-	-	-	-	-	18.75	21.79
MCN (Hendricks et al. 2018)	V	11.41	2.72	24.94	8.22	10.67	-	-
CAL (Escorcia et al. 2019)	V	25.49	11.54	23.40	7.65	9.89	-	-
XML (Lei et al. 2020)	V	41.83	30.35	44.63	31.73	32.14	34.49	55.25
XML+ (Lei, Berg, and Bansal 2021)	V	46.69	33.46	47.89	34.67	34.90	35.38	55.06
MDETR (Lei, Berg, and Bansal 2021)	V	52.89	33.02	54.82	29.40	30.73	35.69	55.60
QD-DETR (Moon et al. 2023)	V	62.40	44.98	62.62	39.88	39.86	38.64	62.40
UniVTG (Lin et al. 2023)	V	58.86	40.86	57.60	35.59	35.47	38.20	60.96
TR-DETR	V	**64.66**	**48.96**	**63.98**	**43.73**	**42.62**	**39.91**	**63.42**
UMT (Liu et al. 2022b)	V+A	56.23	41.18	53.38	37.01	36.12	38.18	59.99
QD-DETR (Moon et al. 2023)	V+A	63.06	45.10	63.04	40.10	40.19	39.04	62.87
TR-DETR	V+A	**65.05**	**47.67**	**64.87**	**42.98**	**43.10**	**39.90**	**63.88**

Table 1: Experimental results on the QVHighlights *test* set. HD represents the results of highlight detection. 'V' and 'A' represent using video and audio features, respectively. Bold letters indicate the best results, while underlined results are suboptimal.

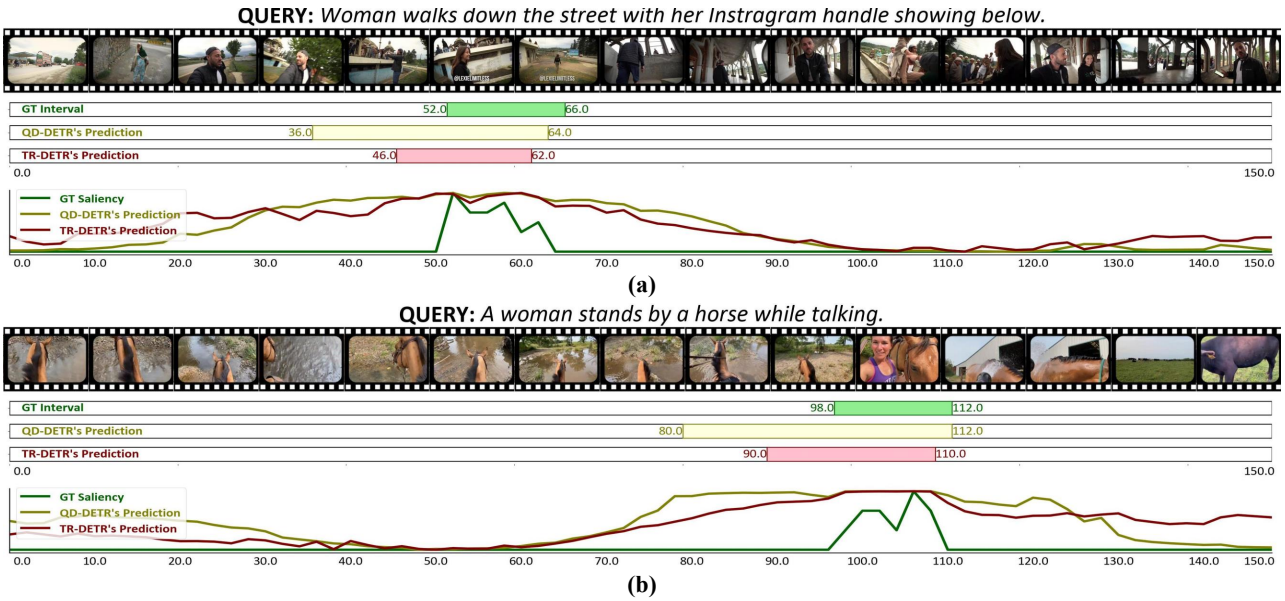

Figure 2: Qualitative results of TR-DETR on QVHighlights *val* set.

outperforms the current best approach on all metrics. Especially with visual features only, TR-DETR exhibits a significant increase in performance under more stringent metrics and high IOU thresholds. Compared with previous methods, TR-DETR improves R1@0.7 and mAP@0.75 by 3.98% and 3.75%, respectively. In addition, after introducing audio information, the performance of a few indicators decreases. This may be because the audio features are spliced directly behind the video features, causing misaligned multi-modal features to be combined and thus impairing modal interactions.

In Table 2, we use VGG, C3D, and SF+C features to comprehensively evaluate the performance of TR-DETR on the Charades-STA dataset. For each feature of VGG, C3D and SF+C, we follow the data preparation settings of UMT (Liu et al. 2022b), VSLNet (Zhang et al. 2021), and Moment-DETR (Lei, Berg, and Bansal 2021), respectively. As shown in Table 2, our TR-DETR shows comparable performance on VGG and SF+C features. Also, performance on some metrics degrades with the introduction of audio, possibly due to insufficient modal interaction. Compared with using only VGG features, the performance of the proposed method is slightly different from UniVTG when using SF+C features. We believe the reasons are as follows: semantic in-

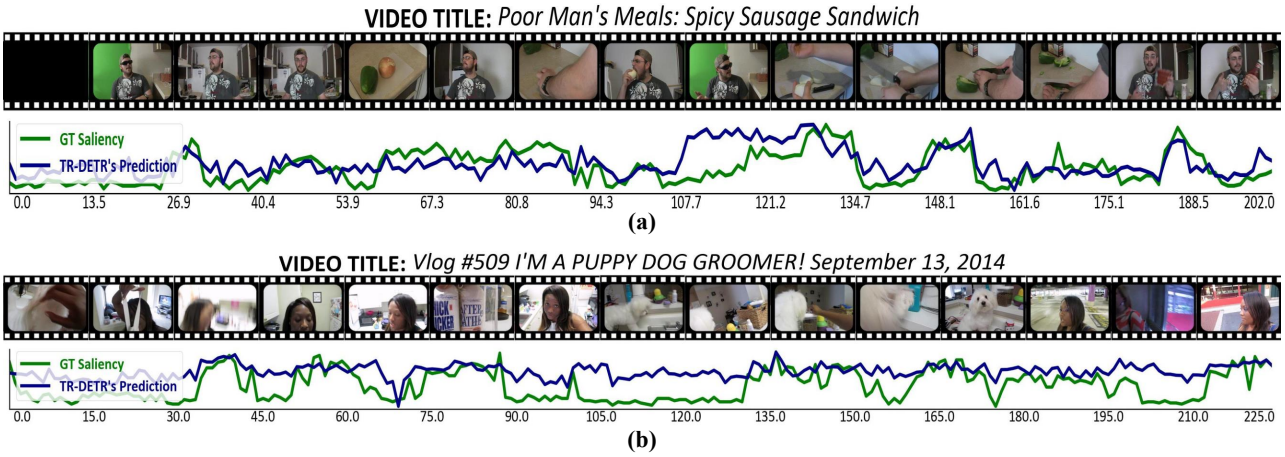

Figure 3: Qualitative results of TR-DETR on TVSum *val* set.

Method	Feat	R1@0.5	R1@0.7
SAP (Chen and Jiang 2019)	VGG	27.42	13.36
TripNet (Hahn et al. 2020)	VGG	36.61	14.50
MAN (Zhang et al. 2019)	VGG	41.24	20.54
2D-TAN (Zhang et al. 2020)	VGG	40.94	22.85
FVMR (Li et al. 2021)	VGG	42.36	24.14
UMT† (Liu et al. 2022b)	VGG	48.31	29.25
QD-DETR (Moon et al. 2023)	VGG	52.77	31.13
QD-DETR† (Moon et al. 2023)	VGG	**55.51**	**34.17**
TR-DETR	VGG	53.47	30.81
TR-DETR†	VGG	54.49	32.37
CTRL (Gao et al. 2017)	C3D	23.63	8.89
ACL (Ge et al. 2019)	C3D	30.48	12.20
MAN (Zhang et al. 2019)	C3D	46.53	22.72
DEBUG (Lu et al. 2019)	C3D	37.39	17.69
VSLNet (Zhang et al. 2021)	I3D	47.31	30.19
QD-DETR (Moon et al. 2023)	I3D	50.67	31.02
TR-DETR	I3D	**55.51**	**33.66**
QD-DETR (Moon et al. 2023)	SF+C	57.31	32.55
UniVTG (Lin et al. 2023)	SF+C	**58.01**	**35.65**
TR-DETR	SF+C	57.61	33.52

Table 2: Experimental results on the Charades-STA *test* set. '†' represents using audio features.

formation of features extracted by different-scale feature extractors (*e.g.* VGG and PANN) varies greatly. In our method, the local-global multi-modal alignment module is used to force the alignment of the visual features of VGG, the audio features of PANN, and the text features of GLoVe, which is challenging and results in relatively weak performance. However, when text, visual and audio features are all derived from large models, such as CLIP, our method shows excellent performance on the QVHighlights dataset.

Consistent with previous work on highlight detection, we evaluate the performance of the proposed TR-DETR on each video category and calculate the top-5 mAP scores. The results are shown in Table 3. In addition, to comprehensively evaluate the overall performance of TR-DETR, we calculate

Figure 4: The impact of local-global alignment loss and λ_{lg} based on QVHighlights *val* set, introducing audio features.

the average value of top-5 mAP on 10 categories. The proposed TR-DETR exceeds the previous method by approximately 3.1% when using only video features, which demonstrates the powerful performance of TR-DETR in solving HD alone.

Visualization

In Figures 2 and 3, we visualize the qualitative analysis results of TR-DETR on the QVHighlights and TVSum datasets, respectively. In Figure 2, compared with QD-DETR, TR-DETR shows more reasonable and accurate results in terms of retrieved accuracy and highlight score distribution. In Figure 3 a), the proposed TR-DETR can accurately fit the highlight score distribution. We believe that these performance improvements are due to the combination of the proposed modules. In addition, in Figure 3 b), it may be that the model only noticed the concept of "puppy dog", resulting in unreasonable high highlight scores in the middle of the result.

Ablation

To verify the effect of each module in the proposed TR-DETR, we conduct a comprehensive ablation experiment,

Method	**Src**	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Avg
sLSTM (Zhang et al. 2016)	V	41.1	46.2	46.3	47.7	44.8	46.1	45.2	40.6	47.1	45.5	45.1
SG (Yuan et al. 2020)	V	42.3	47.2	47.5	48.9	45.6	47.3	46.4	41.7	48.3	46.6	46.2
LIM-S (Xiong et al. 2019)	V	55.9	42.9	61.2	54.0	60.3	47.5	43.2	66.3	69.1	62.6	56.3
Trailer (Wang et al. 2020)	V	61.3	54.6	65.7	60.8	59.1	70.1	58.2	64.7	65.6	68.1	62.8
SL-Module (Xu et al. 2021)	V	86.5	68.7	74.9	**86.2**	79.0	63.2	58.9	72.6	78.9	64.0	73.3
QD-DETR (Moon et al. 2023)	V	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7	85.0
UniVTG (Lin et al. 2023)	V	83.9	85.1	89.0	80.1	84.6	87.0	70.9	**91.7**	73.5	69.3	81.0
TR-DETR	V	**89.3**	**93.0**	**94.3**	85.1	**88.0**	**88.6**	**80.4**	91.3	**89.5**	**81.6**	**88.1**
MINI-Net (Hong et al. 2020)	V+A	80.6	68.3	78.2	81.8	78.1	65.8	75.8	75.0	80.2	65.5	73.2
TCG (Ye et al. 2021)	V+A	85.0	71.4	81.9	78.6	80.2	75.5	71.6	77.3	78.6	68.1	76.8
Joint-VA (Badamdorj et al. 2021)	V+A	83.7	57.3	78.5	86.1	80.1	69.2	70.0	73.0	**97.4**	67.5	76.3
UMT (Liu et al. 2022b)	V+A	87.5	81.5	88.2	78.8	81.4	87.0	76.0	86.9	84.4	79.6	83.1
QD-DETR (Moon et al. 2023)	V+A	87.6	91.7	90.2	**88.3**	84.1	**88.3**	78.7	91.2	87.8	77.7	86.6
TR-DETR	V+A	**90.6**	**92.4**	**91.7**	81.3	**86.9**	85.5	**79.8**	**93.4**	88.3	**81.0**	**87.1**

Table 3: Experimental results on the TVSum *val* set. 'V' and 'A' represent using video and audio features, respectively.

Setting	LGAM	VFR	MR2HD	HD2MR	**Moment Retrieval**					**HD**	
					R1		mAP			≥Very Good	
					@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
(a)					57.72	42.35	59.10	38.16	38.03	36.76	57.44
(b)	✓				63.10	44.97	63.13	40.22	40.47	39.92	63.87
(c)		✓			64.19	47.61	63.50	42.90	41.74	39.71	64.13
(d)			✓		58.39	42.71	59.28	39.19	38.76	37.80	58.8
(e)				✓	59.61	42.26	60.91	39.28	39.26	37.67	58.45
(f)			✓	✓	59.81	44.71	60.25	39.33	39.80	37.86	57.94
(g)	✓		✓	✓	62.13	47.16	62.00	42.79	41.21	39.76	62.65
(h)		✓	✓	✓	63.23	46.90	63.30	42.47	41.64	38.12	59.55
(i)	✓	✓			66.32	50.71	65.71	44.82	43.95	40.35	64.90
(j)	✓	✓	✓	✓	67.10	51.48	66.27	46.42	45.09	40.55	64.77

Table 4: Comparison with the baseline (Moment-DETR with cross-attention module and DAB-DETR's decoder (Liu et al. 2022a)) with different module combinations on QVHighlights *val* set. LGAM represents the local-global alignment module, and VFR is the visual feature refinement module.

and the results are listed in Table 4. Settings (b) to (e) show the performance of each component on the baseline model compared to setting (a). Setting (f) demonstrates the existence of task reciprocity. Compared with setting (c), the reason for the performance degradation in setting (h) may be the semantic mismatch between modalities, resulting in mutual degradation of tasks. Setting (i) shows the huge performance improvement of the proposed local-global alignment loss combined with visual feature refinement.

To further verify the effect of the proposed local-global alignment loss, we also conduct ablation experiments on its coefficients. As shown in Figure 4, after adding the local global regularization term, the model's performance has been significantly improved by about 5%. In addition, as the value of the hyperparameter λ_{lg} gradually increases, the performance improvement becomes more significant. When λ_{lg} is set to 0.3, the model performance reaches its peak and then begins to decline slowly. Comparing the hyperparameter values of 0 and 0.3, the model performance has been improved by about 7% in total, confirming the significant role of the local-global alignment regulators.

Conclusion

This paper proposes a TR-DETR to explore the reciprocity between HD and MR tasks. First, local-global alignment regulators are designed to align visual and textual features. Then, a visual feature refinement module is constructed to obtain discriminative joint features. Finally, a task-reciprocal module is proposed to inject highlight score information into the moment retrieval pipeline and optimize highlight score prediction by utilizing retrieved moments. Extensive experiments on several datasets demonstrate the effectiveness of TR-DETR. However, TR-DETR cannot efficiently utilize data from the audio modality. In the future, we will study novel multi-modal feature interaction networks to coordinate information from multiple modalities.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 62201222 and 62377026, in part by Hubei Provincial Natural Science Foundation of China under Grant 2022CFB954, in part by Knowledge Innovation Program of Wuhan-Shuguang Project under Grant 2023010201020377 and 2023010201020382, in part by self-determined research funds of CCNU from the colleges' basic research and operation of MOE under Grant CCNU22QN014, CCNU22JC007 and CCNU22XJ034, and in part by Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning NO. 2023AISL003 and 2023AISL010.

References

Badamdorj, T.; Rochan, M.; Wang, Y.; and Cheng, L. 2021. Joint Visual and Audio Learning for Video Highlight Detection. In *IEEE ICCV*, 8107–8117.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *ECCV*, 213–229. Springer.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE CVPR*, 4724–4733.

Chen, S.; and Jiang, Y. 2019. Semantic Proposal for Activity Localization in Videos via Sentence Query. In *AAAI*, 8199–8206.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeurIPS 2014 Workshop on Deep Learning*.

Escorcia, V.; Soldan, M.; Sivic, J.; Ghanem, B.; and Russell, B. C. 2019. Temporal Localization of Moments in Video Collections with Natural Language. *CoRR*, abs/1907.12763.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *IEEE ICCV*, 6201–6210.

Foo, L. G.; Gong, J.; Fan, Z.; and Liu, J. 2023. System-Status-Aware Adaptive Network for Online Streaming Video Understanding. In *IEEE CVPR*, 10514–10523.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. TALL: Temporal Activity Localization via Language Query. In *IEEE ICCV*, 5277–5285.

Ge, R.; Gao, J.; Chen, K.; and Nevatia, R. 2019. MAC: Mining Activity Concepts for Language-Based Temporal Localization. In *IEEE WACV*, 245–253.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE ICASSP*, 776–780.

Ghosh, S.; Agarwal, A.; Parekh, Z.; and Hauptmann, A. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *NAACL*, 1984–1990. Minneapolis, Minnesota: ACL.

Ging, S.; Zolfaghari, M.; Pirsiavash, H.; and Brox, T. 2020. COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning. In *NeurIPS*.

Guo, Z.; Zhao, Z.; Jin, W.; Wang, D.; Liu, R.; and Yu, J. 2022. TaoHighlight: Commodity-Aware Multi-Modal Video Highlight Detection in E-Commerce. *IEEE TMM*, 24: 2606–2616.

Hahn, M.; Kadav, A.; Rehg, J. M.; and Graf, H. P. 2020. Tripping through time: Efficient Localization of Activities in Videos. In *BMVC*.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. C. 2018. Localizing Moments in Video with Temporal Language. In *EMNLP*, 1380–1390. ACL.

Hong, F.; Huang, X.; Li, W.; and Zheng, W. 2020. MINI-Net: Multiple Instance Ranking Network for Video Highlight Detection. In *ECCV*, 345–360. Springer.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE TASLP*, 28: 2880–2894.

Kudi, S.; and Namboodiri, A. M. 2017. Words speak for actions: Using text to find video highlights. In *ACPR*, 322–327. IEEE.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting Moments and Highlights in Videos via Natural Language Queries. In *NeurIPS*, 11846–11858.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval. In *ECCV*, 447–463. Springer.

Li, J.; Selvaraju, R. R.; Gotmare, A.; Joty, S. R.; Xiong, C.; and Hoi, S. C. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*, 9694–9705.

Li, J.; Xie, J.; Qian, L.; Zhu, L.; Tang, S.; Wu, F.; Yang, Y.; Zhuang, Y.; and Wang, X. E. 2022. Compositional Temporal Grounding with Structured Variational Cross-Graph Correspondence Learning. In *IEEE CVPR*, 3022–3031.

Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. UniVTG: Towards Unified Video-Language Temporal Grounding. *CoRR*, abs/2307.16715.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022a. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *ICLR*.

Liu, W.; Mei, T.; Zhang, Y.; Che, C.; and Luo, J. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *IEEE CVPR*, 3707–3715.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022b. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *IEEE CVPR*, 3042–3051.

Lu, C.; Chen, L.; Tan, C.; Li, X.; and Xiao, J. 2019. DEBUG: A Dense Bottom-Up Grounding Approach for Natural Language Video Localization. In *EMNLP-IJCNLP*, 5143–5152. ACL.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *ArXiv preprint ArXiv:2002.06353*.

Miech, A.; Alayrac, J.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. In *IEEE CVPR*, 9876–9886.

Molino, A. G. D.; and Gygli, M. 2018. PHD-GIFs: Personalized Highlight Detection for Automatic GIF Creation. In *MM*, 600–608. ACM.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *IEEE CVPR*, 23023–23033.

Mun, J.; Cho, M.; and Han, B. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *IEEE CVPR*, 10807–10816. IEEE.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, 1532–1543. ACL.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *TNNLS*, 20(1): 61–80.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Song, Y.; Redi, M.; Vallmitjana, J.; and Jaimes, A. 2016. To Click or Not To Click: Automatic Selection of Beautiful Thumbnails from Videos. In *ACM CIKM*, 659–668.

Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. TVSum: Summarizing web videos using titles. In *IEEE CVPR*, 5179–5187.

Sun, C.; Baradel, F.; Murphy, K.; and Schmid, C. 2020. Learning Video Representations using Contrastive Bidirectional Transformer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Wang, L.; Liu, D.; Puri, R.; and Metaxas, D. N. 2020. Learning Trailer Moments in Full-Length Movies with Co-Contrastive Attention. In *ECCV*, 300–316. Springer.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. In *AAAI*, 2613–2623. AAAI Press.

Xiong, B.; Kalantidis, Y.; Ghadiyaram, D.; and Grauman, K. 2019. Less Is More: Learning Highlight Detection From Video Duration. In *IEEE CVPR*, 1258–1267.

Xiong, C.; Zhong, V.; and Socher, R. 2017. Dynamic Coattention Networks For Question Answering. In *ICLR*.

Xiong, Z.; and Wang, H. 2023. Dual-Stream Multimodal Learning for Topic-Adaptive Video Highlight Detection. In *ICMR*, 272–279. ACM.

Xu, H.; He, K.; Plummer, B. A.; Sigal, L.; Sclaroff, S.; and Saenko, K. 2019. Multilevel Language and Vision Integration for Text-to-Clip Retrieval. In *AAAI*, 9062–9069.

Xu, M.; Wang, H.; Ni, B.; Zhu, R.; Sun, Z.; and Wang, C. 2021. Cross-category Video Highlight Detection via Set-based Learning. In *IEEE ICCV*, 7950–7959.

Xu, P.; Zhu, X.; and Clifton, D. A. 2022. Multimodal Learning With Transformers: A Survey. *IEEE TPAMI*, 45: 12113–12132.

Xu, Y.; Sun, Y.; Li, Y.; Shi, Y.; Zhu, X.; and Du, S. 2023. MH-DETR: Video Moment and Highlight Detection with Cross-modal Transformer. *ArXiv preprint ArXiv:2305.00355*.

Yan, Z.; Chen, Y.; Song, J.; and Zhu, J. 2023. Multimodal feature fusion based on object relation for video captioning. *CAAI TRIT*, 8(1): 247–259.

Ye, Q.; Shen, X.; Gao, Y.; Wang, Z.; Bi, Q.; Li, P.; and Yang, G. 2021. Temporal Cue Guided Video Highlight Detection with Low-Rank Audio-Visual Fusion. In *IEEE ICCV*, 7930–7939.

Yuan, L.; Tay, F. E. H.; Li, P.; and Feng, J. 2020. Unsupervised Video Summarization With Cycle-Consistent Adversarial LSTM Networks. *IEEE TMM*, 22(10): 2711–2722.

Yuan, Y.; Ma, L.; and Zhu, W. 2019. Sentence Specified Dynamic Video Thumbnail Generation. In *MM*, 2332–2340. ACM.

Zhang, B.; Yang, C.; Jiang, B.; and Zhou, X. 2022. Video Moment Retrieval with Hierarchical Contrastive Learning. In *MM*, 346–355. ACM.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.; and Davis, L. S. 2019. MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment. In *IEEE CVPR*, 1247–1257.

Zhang, H.; Sun, A.; Jing, W.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*, 44(8): 4252–4266.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2023. Temporal Sentence Grounding in Videos: A Survey and Future Directions. *IEEE TPAMI*, 45(8): 10443–10465.

Zhang, K.; Chao, W.; Sha, F.; and Grauman, K. 2016. Video Summarization with Long Short-Term Memory. In *ECCV*, 766–782. Springer.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI*, 12870–12877.