Task-Driven Exploration: Decoupling and Inter-Task Feedback for Joint Moment Retrieval and Highlight Detection

Jin Yang, Ping Wei*, Huan Li, Ziyang Ren
National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
{jin.yang, lh875056558, rzyrzy}@stu.xjtu.edu.cn, pingwei@xjtu.edu.cn

Abstract

Video moment retrieval and highlight detection are two highly valuable tasks in video understanding, but until recently they have been jointly studied. Although existing studies have made impressive advancement recently, they predominantly follow the data-driven bottom-up paradigm. Such paradigm overlooks task-specific and inter-task effects, resulting in poor model performance. In this paper, we propose a novel task-driven top-down framework TaskWeave for joint moment retrieval and highlight detection. The framework introduces a task-decoupled unit to capture task-specific and common representations. To investigate the interplay between the two tasks, we propose an inter-task feedback mechanism, which transforms the results of one task as guiding masks to assist the other task. Different from existing methods, we present a task-dependent joint loss function to optimize the model. Comprehensive experiments and in-depth ablation studies on QVHighlights, TVSum, and Charades-STA datasets corroborate the effectiveness and flexibility of the proposed framework. Codes are available at github.com/EdenGabriel/TaskWeave.

1. Introduction

As videos are prevailing in a wide range of applications, the diversity and massive scales of video content have posed unprecedented challenges in finding relevant moments based on user queries. To this end, the moment retrieval (MR) [60, 63, 65] and highlight detection (HD) [4, 56] tasks have emerged recently. MR aims to retrieve video moments that are relevant to the given query [14, 62]. HD aims to predict the clip-level saliency scores in the video [44].

As MR and HD tasks are closely related, they have been jointly addressed and achieved breakthroughs recently [20, 25, 29, 32, 36, 53]. The existing joint approaches in general utilize a shared backbone to learn the multi-modal features as the common representations for MR and HD. Then a MR prediction head is employed for moments localization and a HD prediction head predicts saliency scores. These methods adhere to the bottom-up, data-driven paradigm, i.e. they capture common features from the input data and then utilize the features for different tasks. The effectiveness of these methods is built upon the premise of the high correlation between MR and HD. As illustrated in Fig. 1, the task MR and HD share the identical query and video inputs, additionally exhibiting a substantial temporal overlap between their respective positive temporal intervals.

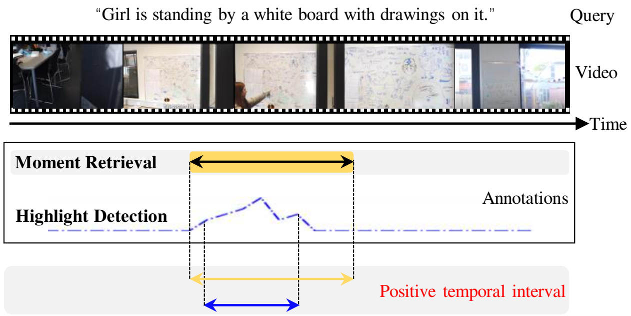
"Girl is standing by a white board with drawings on it."
Query
Video
Time
Moment Retrieval
Highlight Detection
Annotations
Positive temporal interval

Figure 1. Although the positive temporal intervals of moment retrieval and highlight detection exhibit high overlap, they pursue different objectives.

However, the bottom-up, data-driven paradigm tends to excessively rely on the common features, but overlooks the inherent specific characteristics of MR and HD. This tendency might simplify the joint modeling as a problem of feature fusion, without considering the interplay between the two tasks. The distinct objectives pursued by MR and HD rely on distinct task-specific characteristics. Unfortunately, existing methods overlook the specificity.

To address the aforementioned issues, we believe it is essential to leverage the fundamental multi-modal data to

* Corresponding author

mine commonalities across tasks (bottom-up), while also strengthen the awareness of task-specific characteristics (top-down). To this end, we propose a novel paradigm TaskWeave from a task-driven perspective. The key idea is to jointly address the tasks MR and HD by considering the commonality, specificity, and interplay of MR and HD.

To effectively capture the commonality and specificity, we design a task-decoupled unit, i.e. a shared expert to capture common features and two task-specific experts to acquire the distinct characteristics. In order to investigate the interplay between MR and HD in-depth, we design an inter-task feedback mechanism. It converts the predictions of MR/HD into mask information, which are fed back to the input of the HD/MR prediction head. Furthermore, we introduce a principled task-dependent joint loss in which the task-specific weights are dynamically adjusted, rather than manually tuned.

We conduct experiments on the QVHighlights [25] dataset to validate the effectiveness of the proposed method. Moreover, we also conduct experiments for two individual tasks on the datasets Charades-STA [14] (moment retrieval) and TVSUM [43] (highlight detection). The proposed approach outperforms the existing methods.

The key contributions of this paper are three folds.

1. It proposes a novel task-driven, top-down framework for joint moment retrieval and highlight detection.
2. It introduces a task-decoupled unit, an inter-task feedback mechanism, and a principled task-dependent loss.
3. It achieves state-of-the-art performance on three datasets. The ablation study validates the methods.

2. Related Work

We review the related work from four aspects.

Moment Retrieval. Existing approaches for MR mainly include two groups. One group follows a two-stage procedure [1, 14, 60, 65], which involves generating candidate temporal intervals and ranking them based on the correlation with the query. The other group directly regresses the temporal interval based on the aligned visual-text features [8, 37, 58, 59, 61]. Moreover, most datasets provide only one moment annotation for each video-query pair [1, 14], which does not align with real-world scenarios.

Highlight Detection. The saliency score in the highlight detection represents the relevance of a video clip to the given query. Most prior highlight detection benchmark datasets are query-agnostic [15, 44]. The saliency scores for video clips remained constant regardless of the query. As a result, some previous approaches treated HD as a solely visual task [4, 44, 49, 56].

MR and HD tasks are traditionally studied separately. They have been jointly addressed recently as the introduction of QVHighlights dataset [25]. QVHighlights provides multiple moment-annotations for each query and ensures these moments are uniformly distributed throughout the video. It also provides query-dependent highlightness annotations. With QVHighlights, the model Moment-DETR [25] is proposed for joint MR and HD. Following Moment-DETR, a growing number of approaches have been proposed to accomplish the joint task [20, 29, 32, 36]. However, these methods adopt the data-driven and bottom-up paradigm. Different from them, we propose a novel task-driven and top-down paradigm.

Vision Transformers. Transformer-based models [2] have brought huge achievement in both the image and video related domains [5, 11, 27, 28, 54, 55]. One of the most well-known methods is DETR [5], which regards object detection as a set prediction problem. Its end-to-end prediction procedure eliminates the intermediate or post-processing steps. On the other hand, some studies have employed cross-attention mechanism [35, 36, 50] to inject multi-modal data into the Transformer architecture. In this paper, we also adopt a DETR-like architecture [26, 31]. However, different from those methods, we incorporate diverse network architectures for feature extraction. Our focus lies in the collaboration of the network architectures.

Multi-task Learning. Multi-task learning (MTL) aims to train a single model for multiple tasks [7]. The most straightforward approach is to utilize a shared backbone to extract common features, which relies on the data-driven manner. However, it often leads to suboptimal performance for unrelated tasks and lacks flexibility. In response, some studies introduce MOE [19] and MMOE [33], which utilize a set of the shared experts in place of the shared bottom layer. To mitigate the seesaw phenomenon of multi-task learning, PLE [45] employs separate experts for each task while still retaining the shared experts. From the perspective of MTL, existing approaches for joint MR and HD follow the shared bottom paradigm. Moreover, directly applying MTL methods to joint MR and HD does not yield favorable results. Existing MTL methods are complex, therefore direct usage of them leads to a sharp increase in model complexity. This might decrease the performance. To this end, we aim to design an effective task-driven framework to jointly address MR and HD.

3. Methodology

3.1. Overview

Given a text query with W words and an untrimmed video composed of N clips, the objective of the joint moment retrieval (MR) and highlight detection (HD) is to localize the center coordinate q_c and width q_w of temporal intervals that are relevant to the text query, in addition to ranking clip-wise saliency scores.

Architecture Overview. Given the high correlation between moment retrieval and highlight detection tasks, the

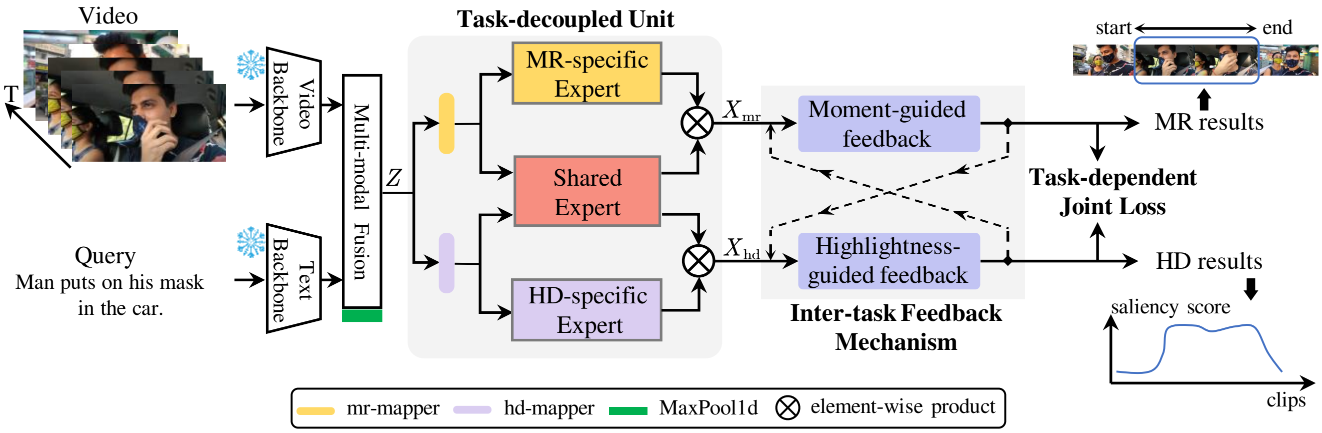

Figure 2. The overall pipeline of the proposed task-driven model TaskWeave. We propose the task-decoupled unit to capture task-specific and common features. Various experts can adopt different network implementations, showcasing the flexibility of the model. Inter-task feedback mechanism is designed to investigate the influence between both tasks. There are two feedback manners: Moment-guided and Highlightness-guided feedback. The principled task-dependent joint loss is introduced for jointly optimize the model.

most intuitive approach is using a shared backbone in conjunction with two task-specific prediction heads. It is a data-driven paradigm that is commonly employed by previous methods, for its simplicity and ease of implementation. However, the task-specific characteristics are inherently present, since MR and HD have distinct objectives. Additionally, the interplay between MR and HD should also be considered, which can enhance the model's performance.

The overall pipeline of our approach is illustrated in Fig. 2. We employed the frozen video/text-encoder backbones to extract the video/text features, while ensuring their dimensions remained consistent at D by the projection. These features are utilized for multi-modal fusion through methods such as cross-attention or concatenation, resulting in query-related video representations $Z \in \mathbb{R}^{N \times D}$. In our paper, we extract these representations through the cross-attention. After cross-attention, inspired by [41, 46], a 1D Max Pooling (MaxPool1d) with the kernel size 5, stride 1 and padding 2 is utilized to eliminate the rank loss problem in the attention mechanism. These representations are fed into the task-decoupled unit to capture the task-related features $X_{\mathrm{mr}} \in \mathbb{R}^{N \times D}$, $X_{\mathrm{hd}} \in \mathbb{R}^{N \times D}$. Then, we employ the task-specific decoders with inter-task feedback mechanism to make predictions for moments localization and clip-wise saliency scores. We introduce the principled task-dependent loss to jointly optimize the model.

3.2. Task-decoupled Unit

Since moment retrieval and highlight detection have distinct objectives, the specificity of each task should be considered, rather than solely focusing on their commonalities. For this purpose, we propose a task-decoupled unit from a task-driven perspective to capture the task-related features, which involves task-specific features and common features.

The task-decoupled unit is depicted in Fig. 2. Inspired by the attention mechanism [2], the query-related video representation Z is initially fed into two task-specific mappers, mr-mapper and hd-mapper. Each mapper is implemented using one-layer feed-forward network. The output of each task-specific mapper is directed into both the respective task-specific expert network (MR-specific Expert or HD-specific Expert) and the Shared Expert network.

Thanks to the design of our task-decoupled framework, each expert can employ various networks, such as convolutional network, Transformer, and feed-forward network. It offers more flexibility in addressing the joint MR and HD tasks. In our study, we also investigate the influence of different expert networks, referring to Sec. 4 for more details. Consequently, we can capture the task-related features X_{mr} and X_{hd} via element-wise product between the output of the task-specific expert and that of the shared expert.

The MR-specific feature X_{mr} can be computed as:

$$X_{\mathrm{mr}} = \mathcal{P}_{\mathrm{mr}}\left(\mathcal{M}_{\mathrm{mr}}\left(Z\right)\right) \otimes \mathcal{S}\left(\mathcal{M}_{\mathrm{mr}}\left(Z\right)\right), \qquad (1)$$

where $\mathcal{M}_{\mathrm{mr}}\left(\cdot\right)$ refers to the mr-mapper operation. $\mathcal{S}\left(\cdot\right)$ is the shared expert calculation. $\otimes$ denotes the element-wise product. $\mathcal{P}_{\mathrm{mr}}\left(\cdot\right)$ is the mr-specific expert calculation. The calculation method for the HD-specific feature X_{hd} follows the similar approach.

3.3. Inter-task Feedback Mechanism

It is imperative to investigate the interplay between the moment retrieval and highlight detection tasks for their significant correlation. However, existing methods overlook this aspect and employ two decoders for direct prediction.

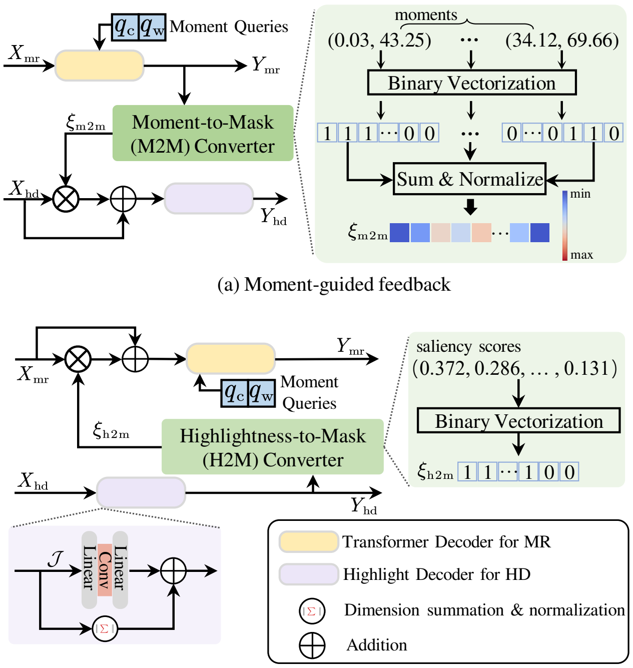

Figure 3. Illustration of the inter-task feedback mechanism. (a) moment-guided feedback manner. (b) highlightness-guided feedback manner.

To resolve this issue, we propose an inter-task feedback mechanism, which contains two task-specific decoders (Transformer decoder for MR and lightweight decoder for HD) and two feedback manners (moment-guided feedback and highlightness-guided feedback). The output of one task is transformed into the mask to assist another task in the feedback mechanism. All components of the inter-task feedback mechanism is detailed as follows.

Transformer Decoder for MR. Previous studies have demonstrated that it is effective to employ dynamic anchors as queries for the decoder within the DETR-like structure [26, 31]. They iteratively update anchor boxes, thereby transforming the procedure for updating queries into the cascaded soft ROI-pooling. In our work, the Transformer decoder for MR follows the work [31]. Each moment query is represented by its temporal center coordinate q_c and width q_w. The output of the Transformer decoder for MR is denoted as $Y_{\text{mr}} \in \mathbb{R}^{N_q \times 2}$. N_q is the number of moment queries and set to 10 in our work.

Lightweight Decoder for HD. The most straightforward method for the HD decoder would be utilizing one or more fully-connected layers, as seen in prior studies such as Moment-DETR [25] and UMT [32]. However, such design overlooks the diversity of video-query pairs [36], which provides identical criterias for the saliency score prediction of each video-query pair. Although QD-DETR [36] proposes a global saliency token to predict the saliency scores, it remains highly coupled with the encoder.

Different from those methods, we introduce CNN structures in the decoder. CNNs can capture local details, ensuring accurate saliency predictions across varying queries within the same video. The structure of the lightweight decoder for HD prediction is illustrated in Fig. 3, which alternates between *Linear* and *Conv1d* layers. We denote the input of the HD decoder as $\mathcal{J} \in \mathbb{R}^{N \times d}$. The saliency prediction $Y_{\text{hd}} \in \mathbb{R}^{N \times 1}$ can be computed as follows:

$$Y_{\text{hd}} = D_{\text{HD}}\left(\mathcal{J}\right) + \frac{\mathbb{I}_{i=1}^{N}\left(\sum_{k=1}^{d} \mathcal{J}^{ik}\right)}{\sqrt{d}}, \qquad (2)$$

where $\mathbb{I}_{i=1}^{N}\left(\cdot\right)$ represents the traversal operation. $D_{\text{HD}}\left(\cdot\right)$ refers to the decoder network operation.

Moment-guided Feedback. We propose a moment-guided feedback manner to investigate the influence of MR on HD. Our key focus is on how to effectively utilize the results of moment retrieval for highlight detection. The implementation procedure of moment-guided feedback is illustrated in Fig. 3 (a).

We transform the tentatively predicted moment results from the "(center, width)" format to the "(start, end)" format. These moment results are fed into the Moment-to-Mask (M2M) converter to generate moment-aware masks ξ_{m2m}. We initialize a clip-wise vector with a length of N for each moment prediction. The indices in these vectors correspond to the clip indexes within the video. We binarize these vectors based on the moment results. We can obtain the clip indexes that are included in the moments through the prior information and set the values at those positions in the vector to 1, while the rest to 0. The obtained N_q binary vectors are then summed and normalized with L2 norm.

The obtained moment-aware masks ξ_{m2m} integrate the prediction results from N_q queries and provide guidance for the highlight detection. In the moment-guided feedback manner, the input $\mathcal{J}$ of the HD decoder is updated by $X_{\text{hd}} + X_{\text{hd}} * \xi_{\text{m2m}}$.

Highlightness-guided Feedback. Similar to the moment-guided feedback, we introduce a highlightness-guided feedback manner with the Highlightness-to-Mask (H2M) converter, to explore the influence of HD on MR. As illustrated in Fig. 3 (b), the saliency score vector is binarized in the H2M to obtain the highlightness-aware mask ξ_{h2m}. Values in the saliency score vector that are below the mean of the vector are set to 0, while those above are set to 1. Then ξ_{h2m} would provide guidance for moment retrieval. In this feedback manner, the input of the MR decoder is updated by $X_{\text{mr}} + X_{\text{mr}} * \xi_{\text{h2m}}$.

3.4. Task-dependent Joint Loss

A direct way to optimize the joint task is to sum the respective task-specific loss functions. However, it neglects the variations in magnitudes of different task losses, leading to dominance of one task. Existing methods [20, 25, 29, 32, 36] utilize manually-set weights for the weighted sum of task losses, which limit the task learning as tasks evolve at different rates. To address this problem, inspired by the studies using dynamic weights [10, 16, 22, 23, 30] based on task learning stages, we introduce the task-dependent joint loss as a more effective and flexible solution.

MR Loss. For moment retrieval, we define the MR likelihood using a Gaussian distribution as shown in Eq. (3), with its mean determined by the output $f^{\theta_{\text{mr}}}(x)$ of the neural network with θ_{mr}. x is the input of the neural network.

$$p\left(Y_{\text{mr}} | f^{\theta_{\text{mr}}}(x)\right) = \mathcal{N}\left(f^{\theta_{\text{mr}}}(x), \delta_{\text{mr}}\right), \tag{3}$$

where δ_{mr} is a learnable parameter that quantifies the uncertainty of moment retrieval. The negative log-likelihood can be derived as follows:

$$\begin{aligned} &-\log p\left(Y_{\text{mr}} | f^{\theta_{\text{mr}}}(x)\right) \\ &\qquad \propto \frac{1}{2{\delta_{\text{mr}}}^2}\left\|Y_{\text{mr}} - f^{\theta_{\text{mr}}}(x)\right\|^2 + \log \delta_{\text{mr}}. \end{aligned} \tag{4}$$

$\left\|Y_{\text{mr}} - f^{\theta_{\text{mr}}}(x)\right\|^2$ measures the offset between the model's predicted value and the ground-truth value. Therefore, inspired by Eq. (4), the MR loss function $\mathcal{L}_{\text{mr}}\left(\theta_{\text{mr}}, \delta_{\text{mr}}\right)$ can be defined as:

$$\mathcal{L}_{\text{mr}}\left(\theta_{\text{mr}}, \delta_{\text{mr}}\right) = \frac{1}{2{\delta_{\text{mr}}}^2}\mathcal{L}\left(\theta_{\text{mr}}\right) + \log \delta_{\text{mr}}. \tag{5}$$

Following existing approaches [25, 32, 36], $\mathcal{L}\left(\theta_{\text{mr}}\right)$ consists of three components: the L1 loss L_{L1}, the generalized IoU loss [40] L_{gIoU}, and the cross-entropy loss L_{BCE}. L_{L1} and L_{gIoU} are employed to calculate the mean absolute error and gIoU deviation between ground-truth moments and predicted moments, respectively. L_{BCE} is used to classify whether the predicted moments belong to the foreground or background. In summary, $\mathcal{L}\left(\theta_{\text{mr}}\right) = L_{\text{L1}} + L_{\text{gIoU}} + L_{\text{BCE}}$.

HD Loss. For highlight detection, inspired by the Boltzmann distribution, the HD likelihood is modeled using a softmax function applied to the scaled model output similar to [23], as Eq. (6). The learnable scaling factor is denoted as δ_{hd}. $f^{\theta_{\text{hd}}}(x)$ is the output of the network with θ_{hd} on input x.

$$p\left(Y_{\text{hd}} | f^{\theta_{\text{hd}}}(x), \delta_{\text{hd}}\right) = \operatorname{softmax}\left(\frac{1}{{\delta_{\text{hd}}}^2} f^{\theta_{\text{hd}}}(x)\right). \tag{6}$$

Similar to Eq. (4), the negative log-likelihood of Eq. (6) can be calculated as follows:

$$\begin{aligned} &-\log p\left(Y_{\text{hd}} = y | f^{\theta_{\text{hd}}}(x), \delta_{\text{hd}}\right) \\ &\quad = \frac{1}{{\delta_{\text{hd}}}^2} \log \frac{\sum_{y'} \exp\left(f_{y'}^{\theta_{\text{hd}}}(x)\right)}{\exp\left(f_{y}^{\theta_{\text{hd}}}(x)\right)} \\ &\quad + \log \frac{\sum_{y'} \exp\left(\frac{1}{{\delta_{\text{hd}}}^2} f_{y'}^{\theta_{\text{hd}}}(x)\right)}{\left(\sum_{y'} \exp\left(f_{y'}^{\theta_{\text{hd}}}(x)\right)\right)^{\frac{1}{{\delta_{\text{hd}}}^2}}} \\ &\quad \approx -\frac{1}{{\delta_{\text{hd}}}^2} \log \operatorname{softmax}\left(Y_{\text{hd}}, f^{\theta_{\text{hd}}}(x)\right) + \log \delta_{\text{hd}}, \end{aligned} \tag{7}$$

where $f_{y}^{\theta_{\text{hd}}}(x)$ refers to the y-th value of the $f^{\theta_{\text{hd}}}(x)$. We introduce a simplify assumption $\sum_{y'} \exp\left(\frac{1}{{\delta_{\text{hd}}}^2} f_{y'}^{\theta_{\text{hd}}}(x)\right) \approx \left(\sum_{y'} \exp\left(f_{y'}^{\theta_{\text{hd}}}(x)\right)\right)^{\frac{1}{{\delta_{\text{hd}}}^2}}$ when $\delta_{\text{hd}} \rightarrow 1$.

$-\log \operatorname{softmax}\left(Y_{\text{hd}}, f^{\theta_{\text{hd}}}(x)\right)$ of Eq. (7) represents the cross-entropy classification loss of Y_{hd}. Inspired on this term, we generalize this expression to other classification losses. The HD loss function $\mathcal{L}_{\text{hd}}\left(\theta_{\text{hd}}, \delta_{\text{hd}}\right)$ can be formulated as:

$$\mathcal{L}_{\text{hd}}\left(\theta_{\text{hd}}, \delta_{\text{hd}}\right) = \frac{1}{{\delta_{\text{hd}}}^2}\mathcal{L}\left(\theta_{\text{hd}}\right) + \log \delta_{\text{hd}}. \tag{8}$$

Consistent with prior approaches [25, 36], $\mathcal{L}\left(\theta_{\text{hd}}\right)$ includes the hinge loss L_{hinge}, negative video-query pairs loss L_{neg}, and rank-aware contrastive loss [17] L_{cont}, i.e. $\mathcal{L}\left(\theta_{\text{hd}}\right) = L_{\text{hinge}} + L_{\text{neg}} + L_{\text{cont}}$. L_{hinge} is computed between two pairs of positive and negative clips, with its margin 0.2 to maintain consistency with [25] for fairness. L_{neg} and L_{cont} from [36] are used to reduce the saliency of negative pairs.

We can obtain final task-dependent joint loss $\mathcal{L}_{\text{joint}}$ through integrating Eq. (5) and Eq. (8):

$$\mathcal{L}_{joint} = \mathcal{L}_{\text{mr}}\left(\theta_{\text{mr}}, \delta_{\text{mr}}\right) + \mathcal{L}_{\text{hd}}\left(\theta_{\text{hd}}, \delta_{\text{hd}}\right). \tag{9}$$

4. Experiments

4.1. Experimental Setup

Datasets. Extensive experiments are conducted on three benchmark datasets: QVHighlights [25], TVSum [43], and Charades-STA [14]. QVHighlights is currently the sole publicly dataset for joint moment retrieval and highlight detection tasks. It provides 10,310 queries associated with 18,367 moments, with an average of 1.8 disjoint moments per query. In contrast to other MR datasets with one-to-one query-moment mappings, QVHighlights aligns more closely with real-world scenarios. Each video in the dataset comprises 75 clips, each of which is 2s-long.

We also utilize two task-specific datasets, Charades-STA [14] for MR and TVSum [43] for HD, to evaluate the model.

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	**Avg.**
sLSTM [64]*ECCV'16*	41.1	46.2	46.3	47.7	44.8	46.1	45.2	40.6	47.1	45.5	45.1
SG [34]*CVPR'17*	42.3	47.2	47.5	48.9	45.6	47.3	46.4	41.7	48.3	46.6	46.2
LIM-S [51]*CVPR'19*	55.9	42.9	61.2	54.0	60.4	47.5	43.2	66.3	69.1	62.6	56.3
Trailer [47]*ECCV'20*	61.3	54.6	65.7	60.8	59.1	70.1	58.2	64.7	65.6	68.1	62.8
SL-Module [52]*ICCV'21*	86.5	68.7	74.9	86.2	79.0	63.2	58.9	72.6	78.9	64.0	73.3
MINI-Net† [18]*ECCV'20*	80.6	68.3	78.2	81.8	78.1	65.8	57.8	75.0	80.2	65.5	73.2
TCG† [57]*ICCV'21*	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]*ICCV'21*	83.7	57.3	78.5	86.1	80.1	69.2	70.0	73.0	97.4	67.5	76.3
UMT† [32]*CVPR'22*	87.5	81.5	88.2	78.8	81.4	**87.0**	76.0	86.9	84.4	79.6	83.1
QD-DETR [36]*CVPR'23*	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7	85.0
UniVTG‡ [29]*ICCV'23*	83.9	85.1	89.0	80.1	84.6	81.4	70.9	91.7	73.5	69.3	81.0
TaskWeave(Ours)	**88.2**	**90.8**	**93.3**	**87.5**	**87.0**	82.0	**80.9**	**92.9**	**89.5**	**81.2**	**87.3**

Table 1. Experimental results (%) on TVSum. † means including audio modality. ‡ means following the pretrain-finetune paradigm.

Backbone	Method	R1@0.5	R1@0.7
VGG	SAP [9]*AAAI'19*	27.42	13.36
	SM-RL [48]*CVPR'19*	24.36	11.17
	2D-TAN [65]*AAAI'20*	40.94	22.85
	FVMR [13]*CVPR'21*	24.36	11.17
	UMT† [32]*CVPR'22*	48.31	29.25
	QD-DETR [36]*CVPR'23*	52.77	31.13
	TaskWeave(Ours)	**56.51**	**33.66**
I3D	CTRL [14]*ICCV'17*	23.63	8.89
	MAN [60]*CVPR'19*	46.53	22.72
	VSLNet [61]*ACL'20*	47.31	30.19
	QD-DETR [36]*CVPR'23*	50.67	31.02
	TaskWeave(Ours)	**53.36**	**31.4**

Table 2. Experimental results (%) on Charades-STA test split. † means including audio modality. ‡ means following the pretrain-finetune paradigm.

Method	**MR**					**HD**	
	R1		mAP			≥ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
Moment-DETR [25]*NIPS'21*	53.94	34.84	-	-	32.20	35.65	55.55
UMT† [32]*CVPR'22*	60.26	44.26	-	-	38.59	**39.85**	**64.19**
QD-DETR [36]*CVPR'23*	62.68	46.66	62.23	41.82	41.22	39.13	63.03
EaTR [20]*ICCV'23*	61.36	45.79	61.86	41.91	41.74	37.15	58.65
UniVTG‡ [29]*ICCV'23*	59.74	-	-	-	36.13	38.83	61.81
TaskWeave(Ours)	**64.26**	**50.06**	**65.39**	**46.47**	**45.38**	39.28	63.68

Table 3. Experimental results (%) on QVHighlights val split. † means including audio modality. ‡ means following the pretrain-finetune paradigm.

Task-decoupled Unit	Inter-task feedback	Joint loss	**MR**		**HD**	
			R1 @0.7	Avg. mAP	mAP	HIT@1
			46.26	41.0	38.94	62.84
✓			47.87	43.24	38.58	61.81
✓		✓	49.29	45.12	38.96	62.0
✓	✓	✓	50.06	45.38	39.28	63.68

Table 4. The ablation results (%) of the components of our proposed method.

Charades-STA provides 16,128 query-moment pairs. TVSum comprises videos from 10 domains, with each domain containing 5 videos.

Evaluation Metrics. We utilize the same evaluation metrics used in prior approaches [25, 32, 36]. For QVHighlights, mean average precision (mAP) with different tIoU thresholds 0.5, 0.75, the average mAP over [0.5:0.05:0.95], and Recall@1 with tIoU 0.5, 0.7 are utilized for MR evaluation. mAP and HIT@1 are used for HD evaluation, where a clip is considered as a true positive when it achieves a "Very Good" [25] saliency score. For Charades-STA, we utilize Recall@1 with tIoU thresholds 0.5, 0.7. For TVSum, Top-5 mAP is adopted.

4.2. Implementation Details

Feature representations. For OVHighlights, we employ the pre-trained SlowFast [12] and CLIP [39] backbones to extract video features, following [25, 32, 36] for fairness. For Charedes-STA, we leverage the official release of VGG [42] and I3D [6] features as video embeddings. For TVSum, we extract video features by the I3D [6] pre-trained on Kinetics 400 [21]. For QVHighlights and TVSum, we use CLIP [39] to extract text features, while using GloVe [38] text embeddings for Charades-STA.

Training settings. We leverage AdamW [24] optimizer with 1e-4 learning rate and 1e-4 weight decay. We train 200, 100, and 2000 epochs with batch size 32, 32 and 2 for QVHighlights, Charades-STA and TVSum, respectively. All Transformer layers follow the consistent configuration, including sinusoidal positional encodings, 8 at-

Index	MR-expert	HD-expert	MR R1 @0.7	MR Avg. mAP	HD mAP	HD HIT@1
(a)	Iden.	Iden.	44.77	40.23	38.38	60.19
(b)	Line.	Line.	47.55	43.65	38.32	60.77
(c)	CNN	CNN	49.23	43.73	38.82	60.77
(d)	Trans.	Trans.	48.32	42.85	38.61	61.10
(e)	Trans.	Iden.	47.1	42.1	38.7	61.68
(f)	CNN	Iden.	49.29	45.12	38.96	62.0

Table 5. Flexibility validation (%) of the task-decoupled unit. "Iden.": identity mapping; "Trans.": Transformer; "Line.": linear layer.

Type	MR R1 @0.7	MR Avg. mAP	HD mAP	HD HIT@1
Sum	47.68	44.27	38.7	60.32
Weighted Sum	48.32	43.21	38.31	61.35
Ours	50.06	45.38	39.28	63.68

Table 6. Effectiveness justification (%) of the task-dependent joint loss.

Type	MR R1 @0.7	MR Avg. mAP	HD mAP	HD HIT@1
MR2HD	50.06	45.38	38.73	62.84
HD2MR	48.65	44.78	39.28	63.68
Bi-MRHD	48.84	45.1	38.36	61.81
MR-HD	50.0	45.3	38.96	62.45
HD-MR	49.1	44.87	39.03	59.94

Table 7. Comparing various inter-task feedback combinations (%) on QVHighlights val split.

tention heads, and a dropout rate of 0.1. Multi-modal fusion employs a 2-layer Transformer with cross attention, where video features serve as the *query* and text features serve as *key* and *value*. For Eq. (9), we define $\gamma_{\rm mr}, \gamma_{\rm hd}$ as $\log{\delta_{\rm mr}}^2, \log{\delta_{\rm hd}}^2$, respectively. Therefore, Eq. (9) can be rewritten as $\mathcal{L}_{joint} = \exp\left(-\gamma_{\rm mr}\right)\mathcal{L}\left(\theta_{\rm mr}\right) + 2\exp\left(-\gamma_{\rm hd}\right)\mathcal{L}\left(\theta_{\rm hd}\right) + \gamma_{\rm mr} + \gamma_{\rm hd}$. $\gamma_{\rm mr}$ and $\gamma_{\rm hd}$ are learnable parameters. They are initialized to 0. To further stabilize the training, the model EMA strategy has also undertaken. All experiments are conducted with Pytorch v1.13.1 on a single NVIDIA RTX 3090.

4.3. Comparison with State-of-the-arts

Results on TVSum. In Tab. 1, we compare our TaskWeave with existing state-of-the-arts (SOTA) methods, where the methods with † are incorporate with audio features. UniVTG [29] follows pretrain-then-finetune paradigm, therefore we use its non-pretrained results for a fair comparison. We observe that: i) our methods outperforms SOTA methods in 9 out of the 10 categories; ii) TaskWeave significantly outperforms all methods in terms of Avg. mAP, with a remarkable 2.71% improvement over the previous SOTA method.

Results on Charades-STA. In Tab. 2, we utilize different backbones to demonstrate the effectiveness of our TaskWeave. We compare our results with other methods for MR only (white background) and methods for joint MR and HD (gray background). It results in significant improvements, *i.e.*, +7.09% in R1@0.5 and +8.13% in R1@0.7.

Results on QVHighlights. As shown in Tab. 3, we present the performance comparisons with existing methods. We report the MR results with the MR-guided feedback, while HD results with the HD-guided feedback. We observe that: i) our TaskWeave surpasses SOTA methods by a large margin in all moment retrieval metrics, with a remarkable 8.72% improvement in Avg. mAP; ii) our method achieves the second-best performance on the HD. However, It's worth noting that UMT [32] incorporates audio features, while our method does not. When compared to other methods without audio features, our method outperforms them all. In summary, these results validate the effectiveness of our task-driven framework.

We provide visualization examples in Fig. 4 for the qualitative analysis among our TaskWeave, Moment-DETR [25], and QD-DETR [36]. Given diverse queries for a video, TaskWeave precisely localizes moments for queries and presents high IoU with the ground-truth. Moment-DETR misses some instances, and QD-DETR exhibits lower retrieval accuracy. Moreover, our method obtains higher saliency scores for relevant clips in response to the query.

4.4. Ablation Studies and Discussions

We present some ablation studies and discussions about TaskWeave, with all experiments are conducted on the QVHighlights *val* split [25].

Ablation of components. We conduct ablation studies on each component of TaskWeave, as shown in Tab. 4. Sequentially employing the task-decoupled unit and our proposed joint loss contributes a 5.5% improvement in average mAP and a 6.5% increase in R1@0.7 for MR. Utilizing all components, we observe that a significant 10.7% improvement in average mAP of MR and a 1.3% enhancement in HIT@1 of HD. These results indicate the effectiveness of the proposed components in out TaskWeave.

Flexibility of the task-decoupled unit. To demonstrate the flexibility of the task-decoupled unit, we investigate the performance of applying different networks within various experts, as shown in Tab. 5. The shared expert utilize a fixed configuration with 2 Transformer layers. We provide 4 different methods for each task-specific expert, including the feed-forward network (one layer Linear), the identity mapping, CNN (composed of one depthwise convolution layer with the kernel/stride/padding of 5/1/2 and 1D convolution), and the Transformer. Due to space constraints, we present

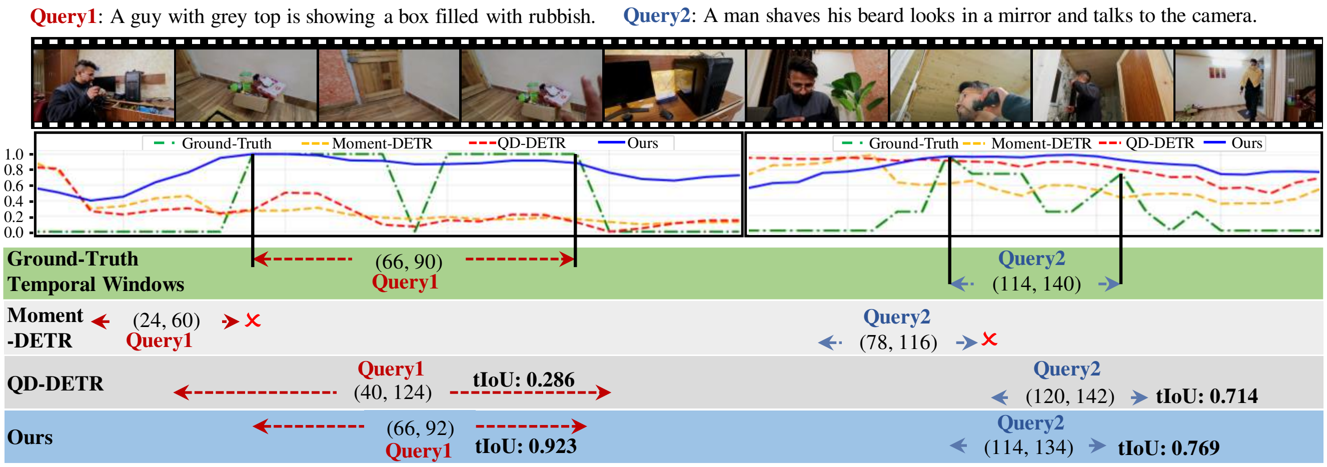

Figure 4. Qualitative results on the QVHighlights for Ground-Truth, Moment-DETR, QD-DETR and our method. The predicted moments and saliency scores are illustrated through intervals and lines.

results of 6 out of 16 combinations. The results in Tab. 5 are obtained by TaskWeave without inter-task feedback.

From this table, we can observe several interesting facts. First, different task-specific experts have different performance. The need for experts is reflected in (a). Second, although Transformer has made significant progress in visual, it's not a panacea. Results in (d) are lower compared to (b) and (c), we believe that the high computational complexity of Transformer decrease the performance. Third, the task-specific expert should be designed based on the task objective. For instance, (c) and (f) perform well on the MR because they focus on local features, which is important for localizing moments. Finally, we also find evidence that MR and HD are highly related. Improvement in one task enhances the other (compare (c) and (f)), while a decline in one limits the other's performance (compare (d) and (e)). In this paper, our mr-specific expert utilizes CNN and hd-specific expert is implemented by identity mapping.

Task-dependent joint loss. Tab. 6 shows the performance of TaskWeave with different losses. "Sum" and "Weighted Sum" refer to the fusion manner of loss for different tasks, respectively. Weights in "Weighted Sum" are consistent with the existing methods [25, 36] for fairness. Comparing "Sum" and "Weighted Sum", we find that "Weighted Sum" can better balance the performance of the two tasks. However, it is obvious that the proposed task-dependent joint loss is optimal.

Inter-task interactions. We believe that each task requires learning before providing effective feedback. Therefore, in our inter-task feedback mechanism, the feedback process starts when the model is trained to half of the max epoch. For brevity, we write Moment/Highlightness-guided feedback as MgF/HgF. In Tab. 7, we explore five feedback manners as follows: "MR2HD" (with MgF only), "HD2MR" (with HgF only), "Bi-MRHD" (with MgF and HgF simultaneously), "MR-HD" (MgF first, then HgF), and "HD-MR" (HgF first, then MgF).

We find that the inter-task feedback makes both tasks gain simultaneously. The performance of the model with "HD2MR" is slightly worse, this is because HD focuses on more refined moments than MR. Ground-truth annotations and the results of "MR-HD" and "HD-MR" also illustrate this fact. The gain brought by "Bi-MRHD" is limited, which we believe that it's a natural result of feedback not always being effective. In general, the inter-task feedback not only contributes to bring gains for both tasks but also helps to understand the characteristics of each task.

5. Conclusion

This paper proposes a novel task-driven paradigm for addressing joint moment retrieval and highlight detection. Different from existing data-driven methods, we utilize the task-decoupled unit to capture the task-specific and common features, respectively. We also explore different network architecture for moment retrieval and highlight detection. We design the inter-task feedback mechanism to in-depth investigate the interplay between both tasks. Different from prior methods, we introduce the principled joint loss to optimize the model. The effectiveness, flexibility, and superiority of the proposed method have been demonstrated on three benchmark datasets.

Acknowledgement

This research was supported by the National Natural Science Foundation of China (No. U23B2060, 62088102), the Youth Innovation Team of Shaanxi Universities, and the Fundamental Research Funds for the Central Universities.

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 international conference on computer vision*, pages 5803–5812, 2017. 2

[2] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Łukasz Kaiser Ashish Vaswani, Noam Shazeer and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5999–6009, 2017. 2, 3

[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*, pages 8127–8137, 2021. 6

[4] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Contrastive learning for unsupervised video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14042–14052, 2022. 1, 2

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

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[7] Rich Caruana. Multitask learning. *Machine learning*, 28: 41–75, 1997. 2

[8] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8175–8182, 2019. 2

[9] 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*, pages 8199–8206, 2019. 6

[10] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018. 5

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[12] 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*, pages 6202–6211, 2019. 6

[13] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1523–1532, 2021. 6

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1, 2, 5, 6

[15] Ana Garcia del Molino and Michael Gygli. Phd-gifs: personalized highlight detection for automatic gif creation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 600–608, 2018. 2

[16] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–287, 2018. 5

[17] David T Hoffmann, Nadine Behrmann, Juergen Gall, Thomas Brox, and Mehdi Noroozi. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 897–905, 2022. 5

[18] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 345–360. Springer, 2020. 6

[19] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 2

[20] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13846–13856, 2023. 1, 2, 5, 6

[21] 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. *arXiv preprint arXiv:1705.06950*, 2017. 6

[22] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 5

[23] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 5

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 1, 2, 4, 5, 6, 7, 8

[26] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 2, 4

[27] Huan Li, Ping Wei, Jiapeng Li, Zeyu Ma, Jiahui Shang, and Nanning Zheng. Asymmetric relation consistency reasoning

for video relation grounding. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022. 2

[28] Huan Li, Ping Wei, Zeyu Ma, and Nanning Zheng. Inverse compositional learning for weakly-supervised relation grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15477–15487, 2023. 2

[29] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804, 2023. 1, 2, 5, 6, 7

[30] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019. 5

[31] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 2, 4

[32] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022. 1, 2, 4, 5, 6, 7

[33] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939, 2018. 2

[34] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211, 2017. 6

[35] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10553–10563, 2022. 2

[36] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023. 1, 2, 4, 5, 6, 7, 8

[37] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 2

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 6

[39] 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 supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[40] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[41] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 3

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[43] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 2, 5

[44] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 787–802. Springer, 2014. 1, 2

[45] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 269–278, 2020. 2

[46] Tuan N Tang, Kwonyoung Kim, and Kwanghoon Sohn. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*, 2023. 3

[47] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N Metaxas. Learning trailer moments in full-length movies with co-contrastive attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 300–316. Springer, 2020. 6

[48] 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*, pages 334–343, 2019. 6

[49] Fanyue Wei, Biao Wang, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Learning pixel-level distinctions for video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3073–3082, 2022. 2

[50] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF con-*

ference on computer vision and pattern recognition, pages 10941–10950, 2020. 2

[51] 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*, pages 1258–1267, 2019. 6

[52] 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*, pages 7970–7979, 2021. 6

[53] Yifang Xu, Yunzhuo Sun, Yang Li, Yilei Shi, Xiaoxiang Zhu, and Sidan Du. Mh-detr: Video moment and highlight detection with cross-modal transformer. *arXiv preprint arXiv:2305.00355*, 2023. 1

[54] Jin Yang, Ping Wei, Ziyang Ren, and Nanning Zheng. Gated multi-scale transformer for temporal action localization. *IEEE Transactions on Multimedia*, 2023. 2

[55] Jin Yang, Ping Wei, and Nanning Zheng. Cross time-frequency transformer for temporal action localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[56] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990, 2016. 1, 2

[57] Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7950–7959, 2021. 6

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

[59] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 2

[60] 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*, pages 1247–1257, 2019. 1, 2, 6

[61] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, 2020. 2, 6

[62] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Temporal sentence grounding in videos: A survey and future directions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10443–10465, 2023. 1

[63] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Temporal sentence grounding in videos: a survey and future directions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[64] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 766–782. Springer, 2016. 6

[65] 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*, pages 12870–12877, 2020. 1, 2, 6