12-2024

Triadic temporal-semantic alignment for weakly-supervised video moment retrieval

Jin LIU

JiaLong XIE

Fengyu ZHOU

Shengfeng HE
Singapore Management University, shengfenghe@smu.edu.sg

Triadic Temporal-Semantic Alignment for Weakly Supervised Video Moment Retrieval

Jin Liu[a], Jialong Xie[a], Fengyu Zhou[a], Shengfeng He[b]

[a] Shandong University, Jinan, China
[b] Singapore Management University, Singapore

Email: {202120638, 202220703, zhoufengyu}@mail.sdu.edu.cn, shengfenghe@smu.edu.sg

Corresponding Author:
Fengyu Zhou
Shandong University
Jinan
Shandong, China
Email: zhoufengyu@mail.sdu.edu.cn

Date: 2024-2-6

Highlights

Triadic Temporal-Semantic Alignment for Weakly Supervised Video Moment Retrieval

Jin Liu, JiaLong Xie, Fengyu Zhou, Shengfeng He

- A weakly supervised video moment retrieval framework to overcome the language bias and relieve data burden.
- Leveraging both fine-grained and coarse-grained features for addressing the multi-modal semantic alignment.
- The method can achieve state-of-the-art performance on out-of-distribution datasets.

Triadic Temporal-Semantic Alignment for Weakly Supervised Video Moment Retrieval

Jin Liu[a], JiaLong Xie[a], Fengyu Zhou[a], Shengfeng He[b]

[a]*Shandong University, JiNan, China*
[b]*Singapore Management University, Singapore*

Abstract

Video Moment Retrieval (VMR) aims to identify specific event moments within untrimmed videos based on natural language queries. Existing VMR methods have been criticized for relying heavily on moment annotation bias rather than true multi-modal alignment reasoning. Weakly supervised VMR approaches inherently overcome this issue by training without precise temporal location information. However, they struggle with fine-grained semantic alignment and often yield multiple speculative predictions with prolonged video spans. In this paper, we take a step forward in the context of weakly supervised VMR by proposing a triadic temporal-semantic alignment model. Our proposed approach augments weak supervision by comprehensively addressing the multi-modal semantic alignment between query sentences and videos from both fine-grained and coarse-grained perspectives. To capture fine-grained cross-modal semantic correlations, we introduce a concept-aspect alignment strategy that leverages nouns to select relevant video clips. Additionally, an action-aspect alignment strategy with verbs is employed to capture temporal information. Furthermore, we propose an event-aspect alignment strategy that focuses on event information within coarse-grained video clips, thus mitigating the tendency towards long video span predictions during coarse-grained cross-modal semantic alignment. Extensive experiments conducted on the Charades-CD and ActivityNet-CD datasets demonstrate the superior performance of our proposed method.

Keywords: Weakly supervised Learning, video moment retrieval, temporal-semantic alignment.

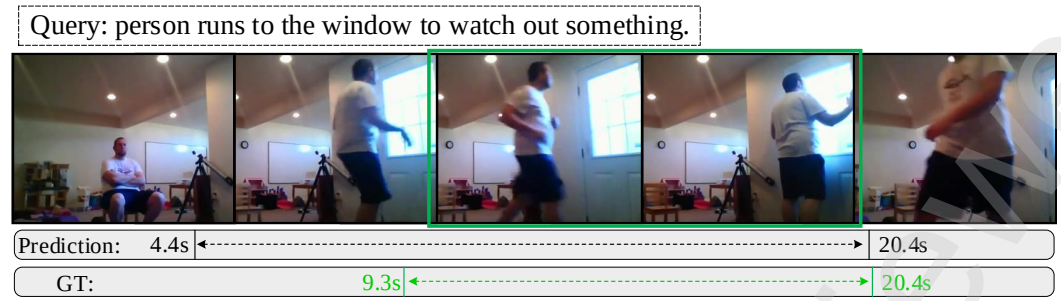
1. Introduction

Video moment retrieval (VMR) aims to detect the temporal location in an untrimmed video based on a given sentence query [1], as depicted in Fig. 1(a), has garnered significant

Email addresses: 202120638@mail.sdu.edu.cn (Jin Liu), 202220703@sdu.edu.cn (JiaLong Xie), zhoufengyu@sdu.edu.cn (Fengyu Zhou), shengfenghe@smu.edu.sg (Shengfeng He)

attention in the field of video understanding [2]. Efficiently identifying the relevant video moment allows users to focus on the most pertinent portion of the video, opening up numerous possibilities for applications in multi-modal tasks such as video summarization [3, 4] and visual question answering [5, 6]. While fully supervised video moment retrieval has achieved impressive results, it is still plagued by temporal bias [7, 8] and the high cost of labor-intensive manual annotations [9], impeding its practicality and progress in real-world applications.

The temporal bias problem refers to the tendency of current VMR models to rely heavily on moment annotation biases present in the training set for retrieving target temporal segments, rather than reasoning based on semantic multi-modal alignment between visual information and textual queries. As illustrated in Fig. 1(a), the VMR model directly infers the inaccurate location of the target video moment based on the memorized biases from the given query "person runs to the window to look out", without fully considering the visual modal information. Recent studies have highlighted that many state-of-the-art models [10] suffer from such temporal bias problems and fail to generalize well on out-of-distribution sets with different moment distributions [11]. Several approaches [12, 13] have attempted to reduce excessive dependence on temporal bias by extra data augmentations. However, they still face the challenge of substantial annotation costs, and these annotations may be accompanied by significant temporal biases. Benefiting from their inherent bias-agnostic nature and low annotation cost, weakly supervised VMR models [14, 15] have garnered significant attention and emerged as optimal solutions. These models operate under weak supervision, where only the video and its corresponding natural language query are provided during training. Compared with unsupervised approaches [16], they have more potential information to utilize and generate more reliable predictions. Weakly supervised VMR models can be broadly classified into two categories: reconstruction-based solutions and multiple instance learning (MIL) based solutions. Reconstruction-based models [17, 18] assume that the video segment



(a) Fully supervised video moment retrieval

(b) Weakly supervised video moment retrieval

Figure 1: (a) An Illustration of fully supervised video moment retrieval methods, where ground-truth moments are utilized to supervise the training process. However, they easily tend to make predictions by using the memorized temporal bias in the training datasets [8]. (b) An example of weakly supervised video moment retrieval approaches, where the ground-truth moments are not available during the training process. Current models often suffer from a temporal ambiguity problem where they tend to predict start or end time points or prolonged segments of the video and fail to take full advantage of the video and query inputs.

that best matches the query can effectively reconstruct the entire query when jointly trained with the words. On the other hand, MIL-based models [19, 20] aim to maximize the matching scores between paired statements and videos while suppressing the scores of unpaired instances, thereby facilitating the learning of visual-text semantic alignment at the video level.

Notwithstanding their demonstrated success, weakly supervised VMR models that lack ground-truth moment supervision often face challenges in achieving fine-grained semantic alignment. These models tend to produce multiple speculative predictions, even prolonged video segments [8] (see Fig. 1(b)). As a consequence, their performance remains lower than that of current fully supervised models and falls short of expectations.

To address these challenges, our aim is to disambiguate the weak supervision by incorporating a triadic temporal-semantic alignment approach that considers both fine-grained and coarse-grained multi-modal semantic alignment perspectives. For the former, we propose aligning the content through concept-aspect and action-aspect alignment to mine fine-grained cross-modal semantic correlations. The concept-aspect alignment involves reconstructing masked noun-concepts using original video clip proposals and minimizing correlations between masked noun-concepts and irrelevant video clip proposals from other videos. This allows the model to leverage clip-level context effectively. Conversely, the action-aspect alignment focuses on reconstructing masked verb-concepts using original video clip proposals while suppressing correlations with shuffled video clip proposals from the same video. This enables the model to capture reliable temporal information from motion-level context. Furthermore, we introduce event-aspect alignment to emphasize coarse-grained cross-modal semantic alignment. We use three different combinations of video clips from the original and irrelevant videos in event-aspect alignment, allowing the model to determine the correct time segment of events rather than predicting specific start or end timestamps. This reduces reliance on the entire video and enhances the model's focus on event information within video clips at the video-level context.

Extensive experiments conducted on the challenging Charades-CD and ActivityNet-CD datasets demonstrate that our proposed strategy achieves superior performance compared to state-of-the-art approaches. Specifically, our model surpasses the current state-of-the-art methods by 2.08% and 2.89% on the Charades-CD and ActivityNet-CD datasets, respectively. These results serve as strong evidence of the effectiveness and generalization capability of our model on both datasets. Moreover, the concept of triadic-grained alignment introduced in our work can serve as valuable insights for addressing various weakly-supervised multi-modal analysis tasks.

2. Related Work

2.1. Video Moment Retrieval

Given an untrimmed video clip and its corresponding natural language query, the task of video moment retrieval (VMR) requires the model to accurately predict the start and end time points that best match the paired video and query. This entails identifying the specific temporal location within the video that corresponds to the query [21].

Existing fully supervised VMR models can be categorized into three groups: two-stage methods, single-stage methods, and reinforcement learning-based methods. Two-stage methods involve pre-segmented moment candidates and the query as inputs, generating proposals offline. Gao et al. [22] propose a multi-scale sliding window sampling approach for candidate proposal generation. In contrast, single-stage methods perform one-pass retrieval of the best matching moment for the given query. Chen et al. [23] propose a novel semantic-aware graph calibration network from global and local multi-modal fusion perspective to make accurate predictions in an end-to-end manner. Sun et al. [24] first treat video moment retrieval as one video reading comprehension task and propose a comprehensive relation-aware network to perceive comprehensive relations and conduct coarse-and-fine cross-modal interaction. Sun et al. [25] propose an end-to-end framework that efficiently retrieves video at segment granularity with two branches, where one branch is for stand-alone ranking and another one for text-video modal alignment.

RL-based methods view proposal generation as a sequential decision-making process. Wu et al. [26] present an iterative coarse-to-fine decision-making process for regulating the temporal boundary. However, these methods are limited by the high annotation costs and temporal biases associated with moments, restricting their practicality in real-world applications.

Different from fully supervised VMR, weakly supervised VMR aims to predict accurate moment timestamps without access to precise temporal locations during training, thereby being free from temporal biases and saving annotation costs. As discussed in Sec. 1, existing weakly supervised methods can be categorized into multiple instance learning (MIL) based models and reconstruction-based models. For the former, Wang et al. [27] propose a context-aware multiple instance learning module that considers adjacent contexts and enhances multi-modal context alignment. For the latter, Lin et al. [28] introduce a semantic completion module that measures semantic similarity between proposals and masked queries, selecting the best matching predicted timestamps with low reconstruction loss. However, these models often align multi-modal context at a coarse granularity and tend to predict specific time points, resulting in their performance being inferior to most state-of-the-art fully supervised models.

2.2. Temporal Bias in Video Moment Retrieval

The problem of temporal moment annotation bias in VMR models has been investigated in recent works [8, 7]. This bias refers to the models' tendency to rely on temporal biases in the training dataset rather than reasoning based on semantic alignment between visual and textual information to determine the target moment timestamp. For example, Yuan et al. [8] address this issue by re-splitting the original Charades-STA [22] and ActivityNet Captions [29] datasets to create debiasing datasets, Charades-CD, and ActivityNet-CD, to evaluate the models' reasoning ability. Their findings indicate that weakly supervised VMR models, benefiting from the bias-agnostic nature of weak supervision, outperform some fully supervised methods. However, these models lack fine-grained multi-modal alignments, leading to multiple speculative predictions on out-of-distribution datasets. In this paper, we propose a triadic temporal-semantic alignment model that addresses these limitations by considering

both fine-grained and coarse-grained semantic alignments in weakly supervised VMR.

3. Triadic Temporal-Semantic Alignment

In this section, we begin by providing an overview of the weakly supervised video moment retrieval task in Sec. 3.1. Then, we describe the procedure for extracting visual and textual features in Sec. 3.2. Next, we present our proposed triadic temporal-semantic alignment model, which comprises a main prediction branch and three alignment modules at both fine-grained and coarse-grained levels, namely, concept-aspect alignment and action-aspect alignment at the fine-grained level, and event-alignment at the coarse-grained level. Finally, we discuss the process of obtaining accurate temporal moments using a well-trained model during the inference stage.

3.1. Task Overview

Given an unmodified video V and a natural language query Q, the weakly supervised video moment retrieval task aims to predict the accurate time locations $\tau = \{t_s, t_e\}$ of the specific video moment corresponding to the query, where t_s and t_e represent the start and end time points, respectively. The video is denoted as $V = \{v_i\}_{i=1}^{n_f}$, where v_i is the i-th video frame and n_f is the number of frames. The natural language query is represented as $Q = \{q_i\}_{i=1}^{n_w}$, where q_i is the i-th word in the query and n_w is the number of words. Importantly, the temporal moment annotation is unavailable during the training process, and it is only utilized for evaluation purposes. In the context of weakly supervised setting, we further propose a triadic temporal-semantic alignment model to enhance the weak supervision and evaluate the generalization ability on out-of-distribution datasets [8].

3.2. Visual and Textual Feature Representations

Following previous works [17], we first pre-extract the video and query representations for the sequential modeling process.

Visual Representation. Given an unmodified video $V = \{v_i\}_{i=1}^{n_f}$, we sample it as images and utilize a pre-trained 3D convolutional network (e.g., I3D [30]) to encode the frame features. We then use a linear projection network to map these features into the multi-modal joint modeling space, obtaining the final representations $F^v = \{f_i^v\}_{i=1}^{n_f} \in \mathbb{R}^{n_f \times D_f}$, where f_i^v denotes the representation of the i-th frame and D_f is the dimension of the video frame features.

Textual Representation. The natural language query $Q = \{q_i\}_{i=1}^{n_w}$ is initialized with pre-trained GloVe embeddings [31]. Additionally, we incorporate positional embeddings [32] into these embeddings to capture the sequential context of the query. Finally, we project the word representations into the multi-modal joint modeling space using a linear projection network, resulting in the final representations $F^q = \{f_i^q\}_{i=1}^{n_w} \in \mathbb{R}^{n_w \times D_w}$, where f_i^q represents the representation of the i-th word and D_w is the word feature dimension (which is the same as D_f).

3.3. Models

As shown in Fig. 2, our proposed triadic temporal-semantic alignment model consists of two main components: the main prediction branch and the triadic multi-modal alignment module. The main prediction branch takes the video and a natural language query as inputs and predicts candidate moment intervals. The multi-modal alignment module, on the other hand, comprises three different-aspect alignment components. This module aims to enhance weak supervision by addressing the multi-modal grounding process from both fine-grained and coarse-grained semantic alignment perspectives.

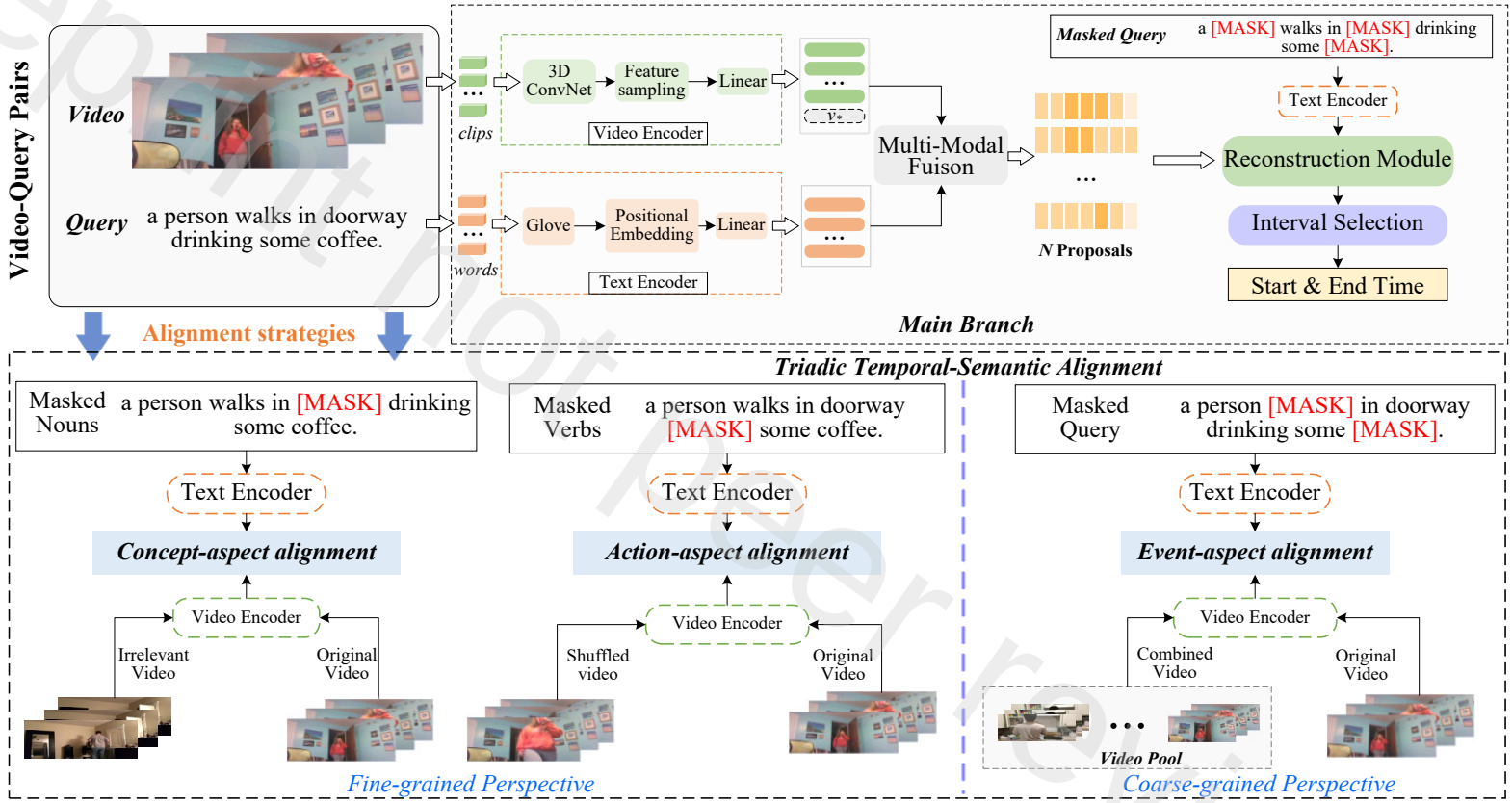


Figure 2: Overview of the proposed model. The model consists of a main prediction branch responsible for generating candidate proposals, and a reconstruction-based multi-modal alignment module for moment prediction. Additionally, the model incorporates triadic temporal-semantic alignment, encompassing both fine-grained (concept-aspect alignment and action-aspect alignment) and coarse-grained (event-aspect alignment) perspectives.

3.3.1. Main Prediction Branch

The main prediction branch of our model follows the reconstruction-based approach commonly used in weakly supervised video moment retrieval methods [14, 17]. This branch conditions moment prediction on the assumption that the video segment that best matches the query should effectively reconstruct the masked parts of the query. It consists of two key modules: the proposal generation module and moment prediction module:

Proposal Generation Module. In this module, inspired by CPL [14], we introduce an additional learnable token v_* with the same feature dimension as the video representations F^v. This token is appended to the video representations, resulting in a modified video representation $\hat{F^v} = \{v_1, v_2, \cdots, v_{nf}, v_*\}$. We then perform multi-modal fusion between the query representation F^q and the modified video representation $\hat{F^v}$ using a transformer-based approach [14, 32]. This fusion process produces the hidden interaction representations H, as shown in Eq. (1).

$$H = \mathrm{MD}(\mathrm{ME}(F^q, \hat{F^v})), \qquad H \in \mathcal{R}^{K \times D_f}, K = n_f + n_w + 1, \tag{1}$$

where ME denotes the transformer encoder. MD denotes the transformer decoder.

After obtaining the hidden representations, we adopt a fully-connected layer to project the learnable token representation h_* to obtain predicted time points $T = \{(\hat{t}_{s1}, \hat{t}_{e1}), (\hat{t}_{s2}, \hat{t}_{e2}), \cdots, (\hat{t}_{sN}, \hat{t}_{eN})\}$, where N is the number of proposals. Given the predicted time intervals, we generated the temporal mask $M = \{m_i\}_{i=1}^{N} \in \mathcal{R}^{N \times n_f}$ to promote the calculation of multi-modal fusion, which is generated from time points. Considering that the predicted time points are not differentiable, we employ a Gaussian shape to approximate the mask [14, 33]. After that, we obtain the candidate proposal representations $P^f = \{p_i^f\}_{i=1}^{N}$ by fusing the temporal mask M and video representations F^v, i.e., $P^f = M \odot F^v$.

Moment Prediction Module. We first randomly mask a part of words (e.g., 1/3) with [MASK] tokens, and require the model to predict the masked word given the proposal representation P^f and the masked query representation F^{qm} by the reconstruction module. The module is transformer-based and formed in Eq. (2).

$$\mathrm{W} = \mathrm{FC}(\mathrm{MD}(F^{qm}, P^f)), \quad \mathrm{W} \in \mathcal{R}^{n_w \times D_w}, \tag{2}$$

where FC denotes a fully-connected layer. W denotes the final predicted word logits.

To ensure better alignment between visual proposals and sentence queries and high-quality predictions, we employ a cross-entropy loss for the reconstructed query and the original query, which is shown in Eq. (3).

$$L_{rec}^k = -\sum_{i=1}^{n_w-1} \log\, p(q_{i+1}|w_i^k), \quad k \in 1, v_2, \cdots, N, \tag{3}$$

where L_{rec}^k represents the reconstruction loss. w_i^k denotes the $i-$th predicted logit of k-th proposal.

After obtaining the loss, we select the optimal matching moment with the following Eq. 4.

$$\begin{aligned} k^* &= \arg\min(L_{rec}), \qquad k^* \in 1, 2, \cdots, N \\ (\hat{t_s}, \hat{t_e}) &= T[k^*], \quad T = \{(\hat{t}_{s1}, \hat{t}_{e1}), \cdots, (\hat{t}_{sN}, \hat{t}_{eN})\}, \end{aligned} \tag{4}$$

where k^* is the index of the best proposal with the smallest reconstruction loss.

Nevertheless, the above main prediction branch suffers from coarse-grained multi-modal semantic alignment in the moment prediction module. As a result, the models are unable to comprehensively understand the paired video context and query information, resulting in their performance still being lower than that of current fully supervised models. To address

this challenge, we align visual and textual information through concept-aspect alignment manner, action-aspect alignment manner, and event-aspect alignment manner from fine-grained and coarse-grained perspectives.

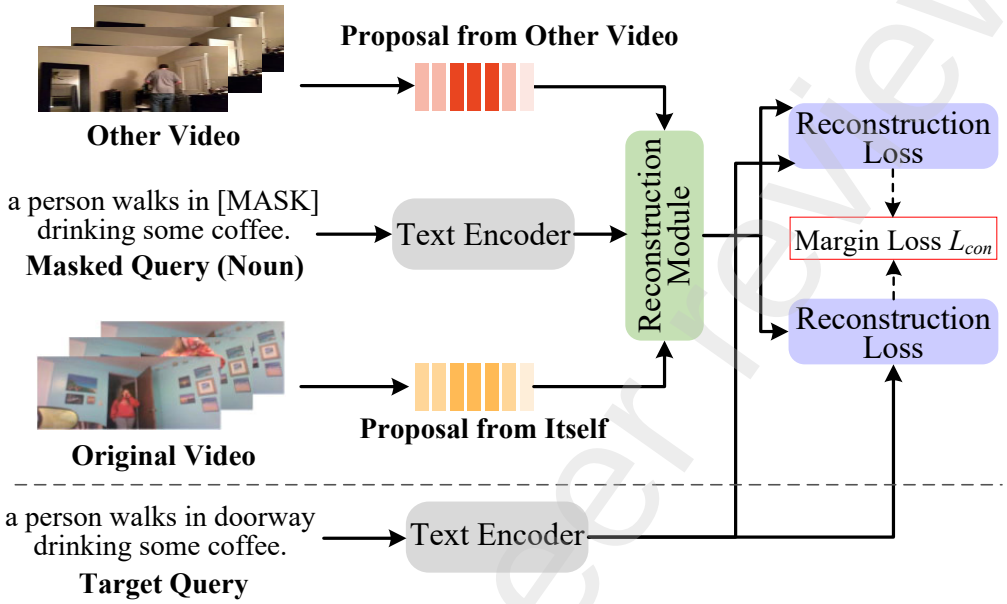


Figure 3: Architecture of the concept-aspect alignment procedure.

3.3.2. Concept-Aspect Alignment

To enhance the alignment between the concept information from a query and proposal context (i.e., clip-level) from the video, and encourage the model to capture more salient concept context from concept-specific proposals, we propose concept-aspect alignment in addition to the main prediction branch, shown in Fig. 3. This module facilitates the reconstruction of masked noun-concepts using original video clip proposals, while simultaneously minimizing correlations between masked noun-concepts and irrelevant video clip proposals from other videos.

Specifically, we first randomly mask a noun (i.e., concept) in the sentence query and obtain the masked query representations. Subsequently, an irrelevant proposal context from

other videos is selected and paired with the masked query to form the negative sample, while the original proposal representation and the masked query form the positive sample. Next, we feed both positive and negative samples into the reconstruction module and utilize Eq. (3) to obtain the positive reconstruction loss L_p^c and negative reconstruction loss L_n^c. To ensure that the better reconstruction of the concept is achieved only through positive proposal information, we introduce one margin loss, shown in Eq. (5), to supervise the training process.

$$L_{con} = max(L_p^c - L_n^c + \alpha_1, 0), \tag{5}$$

where α_1 denotes the margin parameter.

3.3.3. Action-Aspect Alignment

Current weakly supervised video moment retrieval approaches typically focus on continual video clips and sentence queries to implement the multi-modal alignment [14, 34]. However, these approaches often overlook the rich temporal information available in the video from a fine-grained alignment perspective. To this end, we propose the action-aspect alignment shown in Fig. 4, which facilitates the capturing of more temporal context from sequential clips (i.e., motion-level). This approach involves the reconstruction of masked verb-concepts using original video clip proposals, while suppressing correlations with shuffled video clip proposals from the same video.

Particularly, we first randomly mask a verb (i.e., action) in the sentence query and obtain the masked query representations. Subsequently, a shuffled proposal context from the original videos is selected and paired with the masked query to form the negative sample, while the original proposal representation and the masked query form the positive sample. Next, we feed both positive and negative samples into the reconstruction module and utilize Eq. (3) to obtain the positive reconstruction loss L_p^a and negative reconstruction loss L_n^a. To ensure that

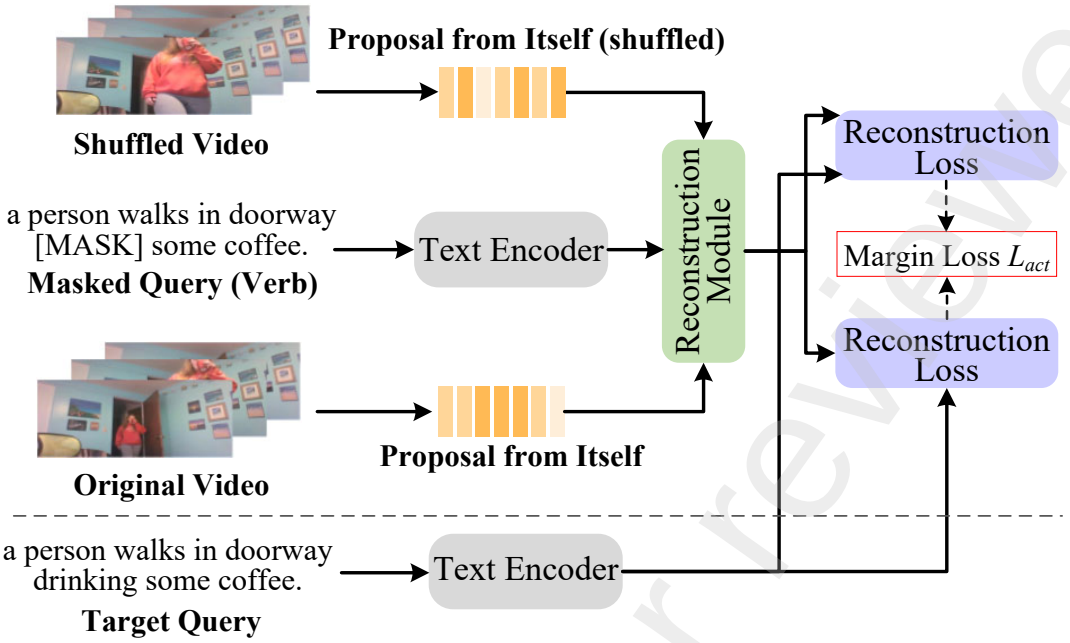


Figure 4: Architecture of the action-aspect alignment procedure.

the better reconstruction of the action is achieved only through positive proposal information, we introduce one margin loss to supervise the training process, as shown in Eq. (6).

$$L_{act} = max(L_p^a - L_n^a + \alpha_2, 0), \tag{6}$$

where α_2 denotes the margin parameter.

3.3.4. Event-Aspect Alignment

Video-level content typically involves event-level context [35, 36], existing approaches that solely focus on aligning a completely irrelevant video with a sentence query tend to predict prolonged video segments. To encourage the model to focus on coarse-grained cross-modal semantic alignment at the video event level and mitigate the temporal bias problem, we introduce an event-aspect alignment module shown in Fig. 5.

In Specific, we first randomly mask several words in the sentence query and obtain the

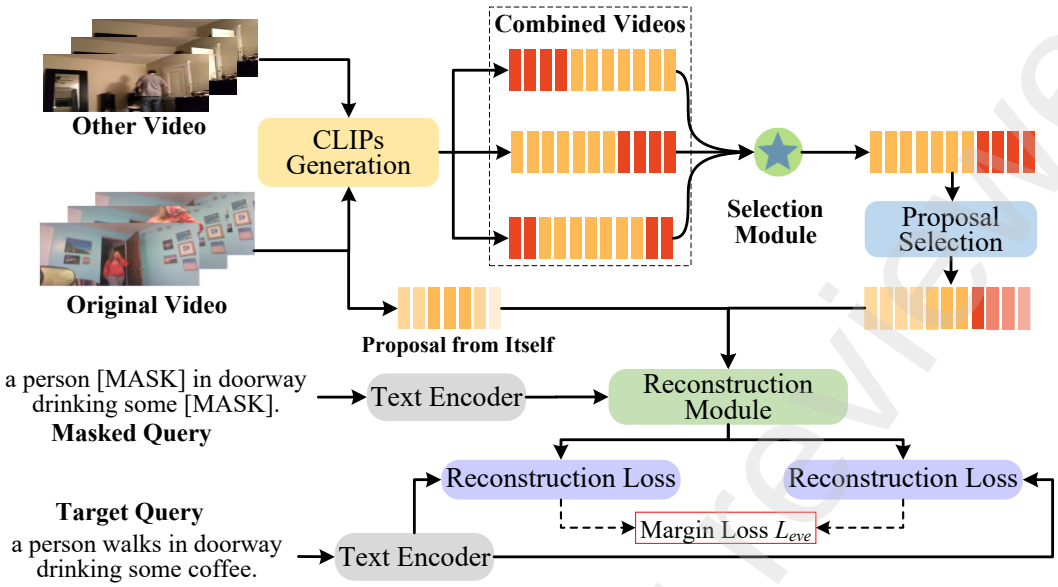


Figure 5: Architecture of the event-aspect alignment procedure.

masked query representations. Then, we randomly select a certain length of proposal representation (e.g., 10 clips) from other irrelevant videos and concatenate it with the original video proposal. Here, we propose three video combination strategies towards the temporal bias conditioned on the relative position of irrelevant video clips from other videos and the original video: head-first, tail-first, and side-first. Subsequently, the combined videos form a new proposal representation through a selection gate [1], and pair with a masked query to generate negative samples. In contrast, the original proposal representation and the masked query form the positive samples. Next, we feed both positive and negative samples into the reconstruction module and utilize Eq. (3) to obtain the positive reconstruction loss L_p^e and negative reconstruction loss L_n^e. To ensure that the better reconstruction of the action is achieved only through positive proposal information, we introduce one margin loss to

[1]We employ the random module in Python to select the combined videos. Note that in our preliminary experiments, we observed that the model achieves superior performance when employing three strategies. Therefore, in this paper, we continue to adhere to this paradigm.

supervise the training process, as shown in Eq. (7).

$$L_{eve} = max(L_p^e - L_n^e + \alpha_3, 0), \tag{7}$$

where α_3 denotes the margin parameter.

3.4. Training and Inference

Our model comprises four loss components, namely the reconstruction loss L_{rec} in Eq. (3) of the main prediction branch for masked query completion, concept-aspect alignment loss L_{rec} in Eq. (5) for enhancing the alignment between the concept information from a query and proposal context from the video clip, action-aspect alignment loss L_{act} in Eq. (6) for facilitating the model to capture more temporal context from sequential motion clips, and event-aspect alignment loss L_{eve} in Eq. (7) for encouraging the model to focus on coarse-grained cross-modal semantic alignment at the video event level and mitigate the temporal bias problem. After that, we utilize a multi-task loss L in the following Eq. (8) to optimize the whole model.

$$L = L_{rec} + L_{con} + L_{act} + L_{eve}, \tag{8}$$

During Inference, we only utilize the main prediction branch and select the optimal matching time proposal by the Eq. (4).

4. Experiments

4.1. Datasets

Charades-CD [8] is a re-splitted dataset based on Charades-STA [22], primarily focused on indoor activities and their corresponding language descriptions. It comprises a total of

Table 1: The detailed statistics of the number of videos and video-query pairs in Charades-CD and ActivityNet-CD datasets and different splits.

Split	Charades-CD		ActivityNet-CD	
	Videos	Pairs	Videos	Pairs
training	4564	11071	10984	51415
validation	333	859	746	3521
test-iid	333	823	746	3443
test-ood	1442	3375	2450	13578

16,128 video-query pairs, with 11,071 pairs used for training. The remaining pairs are divided into a validation set (859 pairs), a test-iid set (823 pairs), and a test-ood set (3,375 pairs).

ActivityNet-CD [8] is another re-splitted dataset, built upon largest VMR dataset ActivityNet Captions [29], which consists of human activities extracted from YouTube videos. The dataset contains a total of 71,957 video-query pairs, with 51,415 pairs used for training. The remaining pairs are divided into a validation set (3,521 pairs), a test-iid set (3,443 pairs), and a test-ood set (13,578 pairs).

The detailed statistics of Charades-CD and ActivityNet-CD are reported in Table 1.

4.2. Evaluation Metrics

Following previous works [8], we adopt the calibrated evaluation metric dR@n, IoU@m to assess the performance of our model. This metric takes into account the "temporal distance" between the predicted moments and the ground truths. The metric is defined as:

$$\text{dR@}n, \text{IoU@}m = \frac{1}{N_q} \sum_i r(n, m, q_i) \cdot \alpha_i^s \cdot \alpha_i^e, \tag{9}$$

where $\alpha_i^* = 1 - abs(p_i^*\text{-}g_i^*)$, and $abs(p_i^* - g_i^*)$ denotes absolute distance between the predicted starting timestamp p_i^s (or end timestamp p_i^e) and ground-truth starting timestamp g_i^s (or ground-truth end timestamp g_i^e).

Table 2: Comparison on the test-ood set of Charades-CD and ActivityNet-CD. We use $n = 1$ and $m \in 0.1, 0.3, 0.5, 0.7$ for dR@n, IoU@m.

Methods	Supervision	**Charades-CD**				**ActivityNet-CD**			
		m=0.1	m=0.3	m=0.5	m=0.7	m=0.1	m=0.3	m=0.5	m=0.7
Biased-based [8]	-	14.75	9.30	5.04	2.21	21.89	9.21	0.26	0.11
PredictAll [8]	-	37.43	27.13	0.06	0.00	21.87	9.01	0.00	0.00
CTRL [22]	Full	52.80	44.97	30.73	11.97	26.23	15.68	7.89	2.53
2D-TAN [10]		50.87	43.45	30.77	11.75	44.37	30.86	18.38	9.11
SCDM [38]		59.08	52.38	41.60	22.22	45.08	31.56	19.14	9.31
WSSL [39]	Weak	49.92	35.86	23.67	8.27	30.71	17.00	7.17	1.82
CNM [17]		52.66	38.50	23.56	8.96	32.08	16.50	5.05	0.76
CPL [14]		60.41	53.88	36.36	18.41	33.28	17.68	6.01	1.01
Our		62.49	55.31	37.01	19.10	36.17	22.23	11.42	4.57

4.3. Implementation Details

For the video features, we employ a pre-trained I3D model [30] to extract representations for Charades-CD, and a C3D model [37] for ActivityNet-CD. As for the natural language, we utilize 300d GloVe embeddings [31] as the initial word representations. Regarding the model architecture, the multi-modal fusion Transformer consists of 3 layers with 4 attention heads for both the encoder and decoder modules. To ensure a fair comparison with existing methods [14], we set the number of proposals N to 8 for Charades-CD and 7 for ActivityNet-CD. Additionally, the margin parameters are determined through grid search: for Charades-CD, $\alpha_1 = 0.15$ in Eq. (5), $\alpha_2 = 0.12$ in Eq. (6), and $\alpha_3 = 0.11$ in Eq. (7); for ActivityNet-CD, $\alpha_1 = 0.11$, $\alpha_2 = 0.1$, and $\alpha_3 = 0.11$.

During training, the model is optimized on a single RTX 3090 GPU using PyTorch. We employ the Adam optimizer with a learning rate of 4e-4, 15 epochs, and a batch size of 32 for all datasets.

4.4. Comparison with State-of-the-arts

We compare our model with various state-of-the-art (SOTA) on the out-of-distribution Charades-CD and ActivityNet-CD [8] datasets. In Tab. 2, we present the results of our model

alongside fully supervised VMR methods (CTRL [22], 2D-TAN [10], SCDM [38]), weakly supervised VMR methods (WSSL [39], CNM [17], CPL [14]), and two biased prediction methods [8] (Biased-based and PredictAll).

Our model consistently achieves the best performance on both datasets compared to the SOTA weakly supervised VMR methods, indicating its ability to alleviate temporal bias. Even on the challenging ActivityNet-CD dataset, which features longer videos, longer queries, and complex content, our model outperforms the current SOTA weakly supervised VMR model CPL [14] by a significant margin. For example, our model exhibits improvements of 4.55% in m=0.3 and 5.41% in m=0.5.

Furthermore, our model consistently outperforms the heuristic algorithm model in the first column, demonstrating its effectiveness in learning and leveraging modal alignment features for accurate moment predictions. Notably, our model achieves promising performance on the Charades-CD dataset even when compared to fully supervised VMR models. It even outperforms the SOTA model SCDM [38] with significant improvements in m=0.1 (62.49% vs. 59.08%) and m=0.3 (55.31% vs. 52.38%). However, our model still exhibits a performance gap when compared to fully supervised VMR models on the ActivityNet-CD dataset. Nonetheless, our model offers significant advantages in terms of practical usability and low annotation costs.

4.5. Ablation Study

To demonstrate the effectiveness of different model variants, we conducted ablation studies on the Charades-CD dataset and obtained the results reported in Tab. 3 above the dashed line. From the results, we observed that compared to the baseline model that only preserves the main prediction branch, our model exhibited significant improvements in performance with the continuous introduction of different aspect alignment modules, including concept-

Table 3: Ablation study of different aspect multi-modal alignments on the Charades-CD dataset. We use $n = 1$ and $m \in 0.1, 0.3, 0.5, 0.7$ for dR@n, IoU@m.

Alignment Terms			m=0.1	m=0.3	m=0.5	m=0.7
concept	action	event				
-	-	-	46.65	38.34	24.21	9.29
✓	-	-	59.94	52.25	36.17	17.49
✓	✓	-	61.13	53.99	37.45	17.30
✓	✓	✓	62.49	55.31	37.01	19.10

Table 4: Ablation study of different combinations for different aspect multi-modal alignments on the Charades-CD dataset. We use $n = 1$ and $m \in 0.1, 0.3, 0.5, 0.7$ for dR@n, IoU@m.

Alignment Terms			m=0.1	m=0.3	m=0.5	m=0.7
concept	action	event				
✓	✓	-	61.13	53.99	37.45	17.30
✓	-	✓	61.83	54.39	36.79	18.83
-	✓	✓	61.47	54.47	36.50	18.88

aspect alignment, action-aspect alignment, and event-aspect alignment. The model with all these alignment modules achieved the best performance, indicating the significance of each designed module.

As discussed in the model design section, our model can align multi-modal information at both fine-grained (concept-aspect alignment and action-aspect alignment) and coarse-grained levels (event-aspect alignment). To further explore the impact of combining modules with different granularities on model performance, we performed additional ablation experiments with the following combinations: (1) only fine-grained combination: concept-aspect alignment and action-aspect alignment, and (2) mixed-grained combination: concept-aspect alignment and event-aspect alignment, action-aspect alignment and event-aspect alignment. The results reported in Table 4 revealed that solely using the fine-grained combination was not sufficient for achieving effective performance and yielded inferior results compared to

Table 5: Ablation study of different selection methods in the event-aspect alignment module on the Charades-CD dataset. We use $n = 1$ and $m \in 0.1, 0.3, 0.5, 0.7$ for dR@n, IoU@m.

Method	m=0.1	m=0.3	0.5	0.7
Full Model	62.49	55.31	37.01	19.10
only head-first	61.67	54.57	37.83	18.42
only tail-first	61.12	54.28	37.53	18.89
only mid-first	60.98	53.86	37.39	19.22

the mixed-grained combination. Importantly, the two different mixed-grained combinations achieved comparable performance, but there was still a notable performance gap compared to the full model. We speculate that this is mainly because each aspect alignment can contribute to the multi-modal grounding process, and these combinations do not fully account for the diverse distribution of temporal bias cases, thereby limiting the model's generalization capacity. These findings further underscore the importance of the different aspect alignment modules in our model.

Notably, there exist three video combination strategies in event-aspect alignment module. To investigate the effectiveness of these strategies, we implement extra ablation experiments and obtain the results reported in Table 5. The results shown that solely using the fine-grained combination was not sufficient for achieving effective performance and yielded inferior results compared to the mixed-grained combination. Importantly, the two different mixed-grained combinations achieved comparable performance, but there was still a notable performance gap compared to the full model. We speculate that this is mainly because each aspect alignment can contribute to the multimodal grounding process, and these combinations do not fully account for the diverse distribution of temporal bias cases, thereby limiting the model's generalization capacity.

To gain further insights into our model, we conducted additional ablation experiments to address the following questions: (1) Can our model maintain its performance on the

Table 6: Comparison on the test-iid set of Charades-CD and ActivityNet-CD. We use $n = 1$ and $m \in 0.1, 0.3, 0.5, 0.7$ for dR@n, IoU@m.

Methods	**Charades-CD**				**ActivityNet-CD**			
	m=0.1	m=0.3	m=0.5	m=0.7	m=0.1	m=0.3	m=0.5	m=0.7
WSSL [39]	45.90	34.99	14.06	4.27	36.67	26.06	17.20	6.16
CNM [17]	48.44	35.28	17.37	7.58	36.28	35.62	24.86	13.51
CPL [14]	66.53	61.39	50.49	22.91	43.93	34.68	24.80	13.62
Our	67.57	62.29	51.30	23.02	44.73	35.27	25.39	14.11

independent and identically distributed (IID) split? (2) Does the choice of n in the top-n retrieval have an impact on model performance? (3) How does the number of proposals affect our model's performance?

Q1: *Can our model maintain its performance on the independent and identically distributed (IID) split?* To assess the performance of our model on the IID split, we conducted additional comparison experiments with existing weakly supervised video moment retrieval methods on the test-iid sets of Charades-CD and ActivityNet-CD datasets. The detailed results are presented in Tab. 6. Our model consistently outperformed other weakly supervised video moment retrieval methods on both datasets, indicating its ability to maintain strong generalization performance on test-iid datasets.

Q2: *Does the choice of n in the top-n retrieval have an impact on model performance?* To investigate the impact of different retrieval numbers n in Eq. (9) on our model's performance, we conducted an extensive analysis comparing $n = 3$ and $n = 5$ with the baseline model (utilizing only the main prediction branch) on the test-ood set of Charades-CD. The results are presented in Fig. 6. Our model consistently outperformed the baseline models with increasing margins as the evaluation metrics became more challenging. For example, our model achieved relative improvements of 22.45% for R@3, IoU=0.7, and 13.51% for R@3, IoU=0.1. These findings indicate that our model can achieve superior performance as the

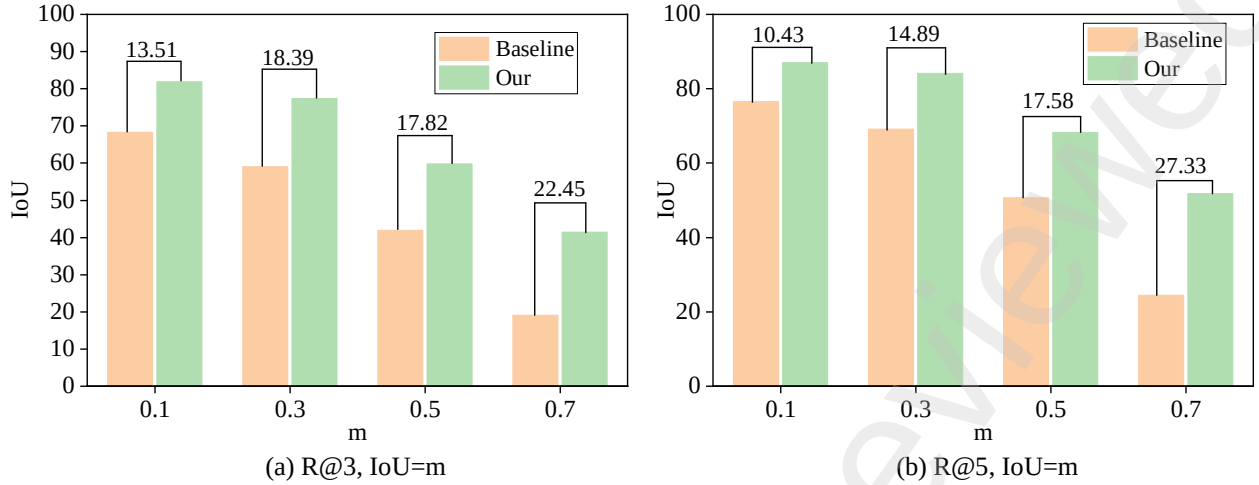


Figure 6: Comparison results for different top-n retrieval predictions with the baseline model (only utilize the main prediction branch) on the test-ood set of Charades-CD.

value of n increases, benefiting from the designed alignment modules.

Q3: *How does the number of proposals affect our model's performance?* To further explore the impact of the number of proposals on our model's performance, we conducted an additional ablation study on the test-ood set of Charades-CD by varying the number of proposals. The results are presented in Fig. 7. It can be observed that our proposed model achieves optimal performance on Charades-CD and ActivityNet-CD when the number of proposals is set to 8 and 7, respectively, for both R@1, IoU=0.1 and R@1, IoU=0.3 cases. In all our experiments, we consistently set the number of proposals to 8 and 7 for Charades-CD and ActivityNet-CD, respectively.

4.6. Qualitative Results

To qualitatively demonstrate the effectiveness of our proposed model, we present two examples along with an additional failure case from the Charades-CD dataset in Fig. 8. Each example consists of a query and the corresponding ground-truth temporal locations denoted as 'GT'. We compare our model with the state-of-the-art weakly supervised video

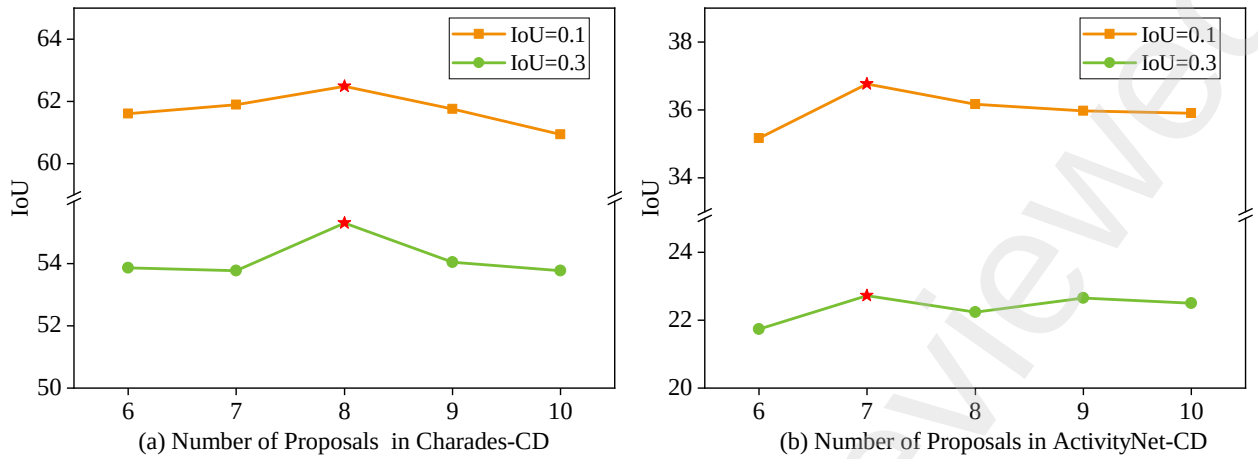
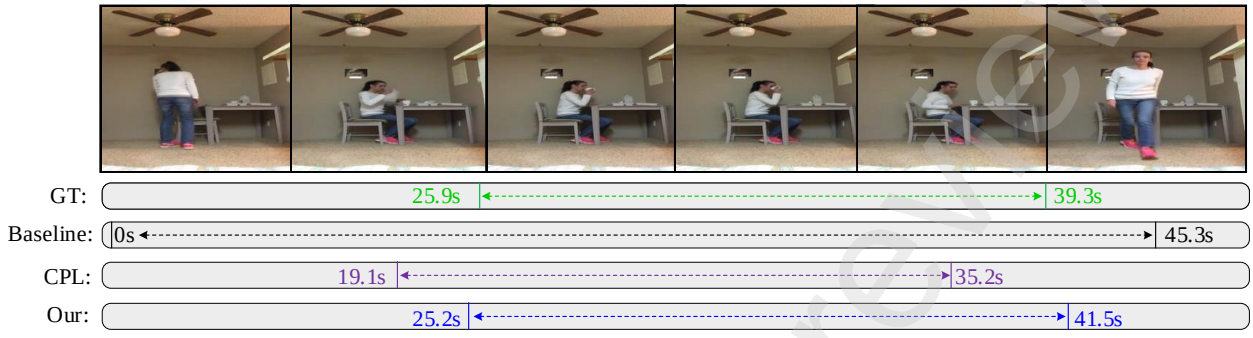


Figure 7: Results of different numbers of proposals on the test-ood set of Charades-CD and ActivityNet-CD. We use $n = 1$ and $m \in 0.1, 0.3$ for dR@n, IoU@m. Red stars denote the best performance.

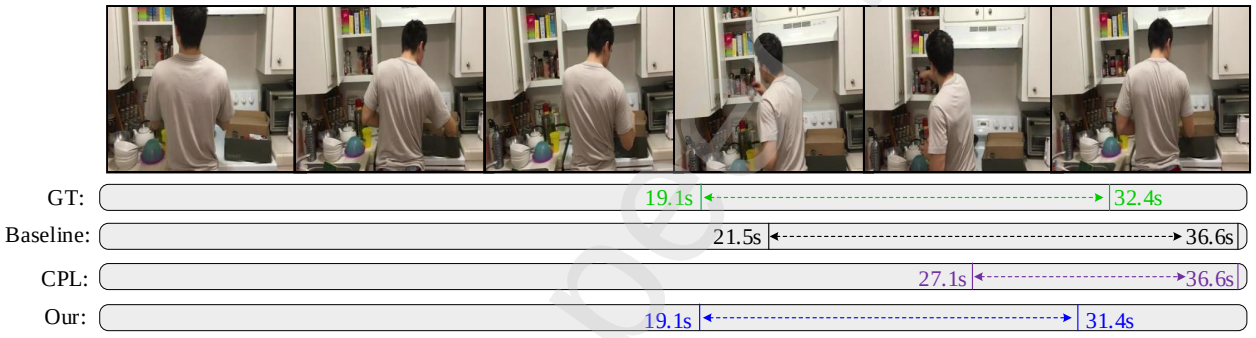
moment retrieval models and the baseline method that only preserves the main prediction branch.

From Fig. 8(a) and (b), it can be observed that the baseline model tends to predict either the start timestamp (e.g., 0.0s in Fig. 8(a)) or the end timestamp (e.g., 36.6s in Fig. 8). The SOTA model CPL [14] fails to capture more fine-grained details about the video context and thus makes inaccurate predictions (e.g., CPL fails to capture the state change between the action "putting"). However, our model successfully predicts precise temporal moments, encompassing essential concepts, actions, and events. This demonstrates the effectiveness of our three alignment modules: concept-aspect alignment, action-aspect alignment, and event-aspect alignment.

Nevertheless, our model also exhibits failure, as shown in Fig. 8(c). We observe that although our model captures a reasonable temporal moment, it fails to cover the overall query-specific moment. This failure can be attributed to the ambiguity between the visual content in the video and the textual query. Specifically, the query contains the concept "chair", which is not present in the video, leading the model to struggle in inferring the



(a) *Query*: the person drinks the glass of water.



(b) *Query*: person putting things on a shelf.

(c) A failure case. *Query*: a person is sitting in a chair.

Figure 8: Three qualitative examples from the test-ood split of the Charades-CD dataset, including one failure case. Our model is compared with the state-of-the-art method CPL [14] and the baseline method that only preserves the main prediction branch. The ground-truth moment is denoted as GT and plotted in green.

human posture and subsequently producing incorrect predictions.

5. Conclusion

In this paper, we propose a novel triadic temporal-semantic alignment model for weakly supervised video moment retrieval. Our model enhances weak supervision by incorporating both fine-grained and coarse-grained multi-modal semantic alignments. We introduce concept-aspect alignment and action-aspect alignment to facilitate fine-grained cross-modal grounding, aligning visual context with nouns and verbs in the query. Additionally, we incorporate event-aspect alignment, encompassing three video combinations (head-first, tail-first, and side-first), to promote coarse-grained multi-modal semantic alignment at the video event level and mitigate temporal bias. Our extensive quantitative and qualitative experiments on Charades-CD and ActivityNet-CD datasets demonstrate that our model achieves new state-of-the-art performance.

In the future, we plan to extend our model to handle compositional video moment retrieval cases, enabling the model to effectively generalize a combination of novel concepts.

References

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

Y. Liu, X. Zhang, F. Huang, S. Shen, P. Tian, L. Li, Z. Li, Dynamic self-attention with vision synchronization networks for video question answering, Pattern Recognition 132 (2022) 108959.

Y. Zhu, W. Zhao, R. Hua, X. Wu, Topic-aware video summarization using multimodal transformer, Pattern Recognition 140 (2023) 109578.

P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, L. Shao, Exploring global diverse attention via pairwise temporal relation for video summarization, Pattern Recognition 111 (2021) 107677.

J. Liu, C. Fan, F. Zhou, H. Xu, Be flexible! learn to debias by sampling and prompting for robust visual question answering, Information Processing and Management (2023) 103296–103310.

S. A. M. Mohamud, A. Jalali, M. Lee, Encoder–decoder cycle for visual question answering based on perception-action cycle, Pattern Recognition 144 (2023) 109848.

X. Yang, F. Feng, W. Ji, M. Wang, T.-S. Chua, Deconfounded video moment retrieval with causal intervention, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1–10.

Y. Yuan, X. Lan, X. Wang, L. Chen, Z. Wang, W. Zhu, A closer look at temporal sentence grounding in videos: Dataset and metric, in: Proceedings of the 2nd international workshop on human-centric multimedia analysis, 2021, pp. 13–21.

R. Tan, H. Xu, K. Saenko, B. A. Plummer, Logan: Latent graph co-attention network for weakly-supervised video moment retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2083–2092.

S. Zhang, H. Peng, J. Fu, J. Luo, Learning 2d temporal adjacent networks for moment localization with natural language, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 12870–12877.

J. Hao, H. Sun, P. Ren, J. Wang, Q. Qi, J. Liao, Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding, in: European Conference on Computer Vision. Cham: Springer Nature Switzerland, Springer, 2022, pp. 130–147.

M. Zhai, C. Li, C. Jing, Y. Wu, Synthesizing counterfactual samples for overcoming moment biases in temporal video grounding, in: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer International Publishing, 2022, pp. 436–448.

D. Liu, P. Zhou, Z. Xu, H. Wang, R. Li, Few-shot temporal sentence grounding via memory-guided semantic learning, IEEE Transactions on Circuits and Systems for Video Technology (2022).

M. Zheng, Y. Huang, Q. Chen, Y. Peng, Y. Liu, Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15555–15564.

J. Chen, W. Luo, W. Zhang, L. Ma, Explore inter-contrast between videos via composition for weakly supervised temporal sentence grounding, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 267–275.

J. Gao, C. Xu, Learning video moment retrieval without a single annotated video, IEEE Transactions on Circuits and Systems for Video Technology 32 (2021) 1646–1657.

M. Zheng, Y. Huang, Q. Chen, Y. Liu, Weakly supervised video moment localization with contrastive negative sample mining, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 3517–3525.

Y. Zhao, Z. Zhao, Z. Zhang, Z. Lin, Cascaded prediction network via segment tree for

temporal video grounding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4197–4206.

Y. Wang, M. Liu, Y. Wei, Z. Cheng, Y. Wang, L. Nie, Siamese alignment network for weakly supervised video moment retrieval, IEEE Transactions on Multimedia (2022) 1–13.

J. Huang, Y. Liu, S. Gong, H. Jin, Cross-sentence temporal and semantic relations in video activity localisation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7199–7208.

H. S. Nawaz, Z. Shi, Y. Gan, A. Hirpa, J. Dong, H. Zheng, Temporal moment localization via natural language by utilizing video question answers as a special variant and bypassing nlp for corpora, IEEE Transactions on Circuits and Systems for Video Technology 32 (2022) 6174–6185.

J. Gao, C. Sun, Z. Yang, R. Nevatia, Tall: Temporal activity localization via language query, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5267–5275.

T. Chen, W. Wang, K. Han, H. Xu, Sagcn: Semantic-aware graph calibration network for temporal sentence grounding, IEEE Transactions on Circuits and Systems for Video Technology (2022).

X. Sun, J. Gao, Y. Zhu, X. Wang, X. Zhou, Video moment retrieval via comprehensive relation-aware network, IEEE Transactions on Circuits and Systems for Video Technology (2023).

X. Sun, X. Long, D. He, S. Wen, Z. Lian, Vsrnet: End-to-end video segment retrieval with text query, Pattern Recognition 119 (2021) 108027.

J. Wu, G. Li, S. Liu, L. Lin, Tree-structured policy based progressive reinforcement learning for temporally language grounding in video, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12386–12393.

J. Wang, L. Ma, W. Jiang, Temporally grounding language queries in videos by contextual boundary-aware prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12168–12175.

Z. Lin, Z. Zhao, Z. Zhang, Q. Wang, H. Liu, Weakly-supervised video moment retrieval via semantic completion network, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 11539–11546.

F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.

J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing, 2014, pp. 1532–1543.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems (2017) 1–11.

C. Ju, H. Wang, J. Liu, C. Ma, Y. Zhang, P. Zhao, J. Chang, Q. Tian, Constraint and union for partially-supervised temporal sentence grounding, arXiv:2302.09850 (2023).

M. Gao, R. Socher, C. Xiong, Weakly supervised natural language localization networks, 2019.

S. Chen, Y. Zhao, Q. Jin, Q. Wu, Fine-grained video-text retrieval with hierarchical graph reasoning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10638–10647.

J. Xiao, A. Yao, Z. Liu, Y. Li, W. Ji, T.-S. Chua, Video as conditional graph hierarchy for multi-granular question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 2804–2812.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

Y. Yuan, L. Ma, J. Wang, W. Liu, W. Zhu, Semantic conditioned dynamic modulation for temporal sentence grounding in videos, Advances in Neural Information Processing Systems 32 (2019).

X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, J. Huang, Weakly supervised dense event captioning in videos, Advances in Neural Information Processing Systems 31 (2018).