International Journal of Computer Vision (2026) 134:53
https://doi.org/10.1007/s11263-025-02599-w

Large-Scale Pre-Trained Models Empowering Phrase Generalization in Temporal Sentence Localization

Yang Liu[1] · Minghang Zheng[1] · Qingchao Chen[2] · Shaogang Gong[3] · Yuxin Peng[1]

Received: 16 January 2025 / Accepted: 24 November 2025

Abstract

Video temporal sentence localization aims to localize a target moment in videos given language queries. We observe that existing models suffer from a sheer performance drop when dealing with phrases contained in the sentence. It reveals the limitation that existing models lack sufficient understanding of the semantic phrases in the query. To address this problem, we fully exploit the temporal constraints between phrases within the same sentence and attempt to transfer knowledge from externally pre-trained large models to help the model better accomplish phrase-level localization. Firstly, we propose a phrase-level Temporal Relationship Mining (TRM) framework that employs the temporal relationship between the phrase and the whole sentence to better understand each semantic entity (e.g. verb, subject) in the sentence. Specifically, we propose the consistency and exclusiveness constraints between phrase and sentence predictions to improve phrase-level prediction quality and use phrase-level predictions to refine sentence-level ones. Then, we extend the TRM framework with phrase-level training (TRM-PT) using the large-scale pre-trained models to generate fine-grained pseudo-labels for the phrase. To mitigate the negative impact of the label noise, we further propose to iteratively optimize the pseudo-labels. Finally, to enhance the understanding of verb phrases, we utilize a language model to infer changes in the scene's state before and after the occurrence of verb phrases and align them with the visual content. Experiments on the ActivityNet Captions and Charades-STA datasets show the effectiveness of our method on both phrase and sentence temporal localization and enable better model interpretability and generalization when dealing with unseen compositions of seen concepts. The code is available at https://github.com/minghangz/trm.

Keywords Video moment retrieval · Temporal video grounding · Cross-modal video retrieval · Large vision language model · Video understanding

MSC Classification: Primary: 68T45 · Secondary: 68T07

1 Introduction

Video temporal sentence localization has become an important research problem due to its potential for a wide range of practical applications, requiring intelligent systems to identify the start and end timestamps of segments (i.e., moments) with respect to any given language queries in an untrimmed video. Using free-form natural language as queries allows users to freely search for content without being limited

Communicated by Gunhee Kim.

✉ Yang Liu
yangliu@pku.edu.cn

Minghang Zheng
minghang@pku.edu.cn

Qingchao Chen
qingchao.chen@pku.edu.cn

Shaogang Gong
s.gong@qmul.ac.uk

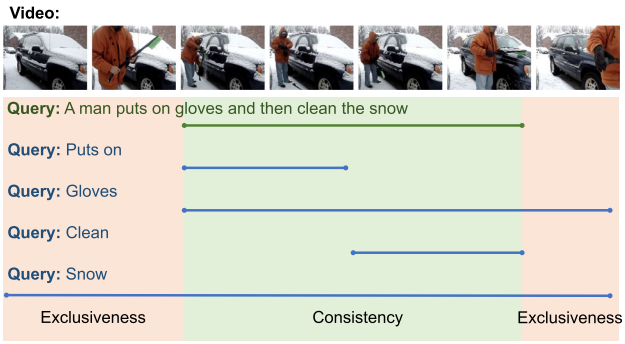
Yuxin Peng
pengyuxin@pku.edu.cn

[1] Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China

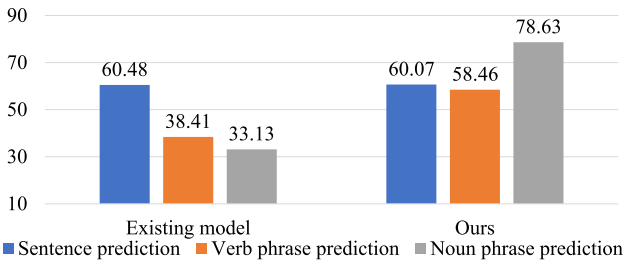
[2] National Institute of Health Data Science, Peking University, Beijing 100871, China

[3] School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, England, United Kingdom

Published online: 10 January 2026



(a) Sentence and phrase level prediction.



(b) Performance on Charades-STA (IoU@0.3).

Fig. 1 (a) The sentence-level (in green) and phrase-level (in blue) prediction. We make two assumptions about the relationship between phrases and sentences: 1) Consistency: for each phrase, the phrase-level prediction should overlap the sentence ground truth (in green); 2) Exclusivity: for each video clip that does not intersect with the sentence ground truth (in red), at least one phrase's prediction does not overlap it. (b) shows the evaluation results of the existing model (Wang et al., 2022) and our method on the Charades-STA (R@1, IoU=0.3) when using sentences or phrases as queries

to predefined classes, giving sentence localization greater application potential. The model is expected to understand the visual and language concepts and their compositions to achieve robust performance.

In recent decades, fully supervised approaches have demonstrated consistent advancement when handling complete sentences as queries. However, real-world human-generated queries vary a lot in terms of specificity. Consequently, it is important for the model to effectively handle various query types in order to be competent for real-world applications, including both complete sentences (highlighted in green in Fig.1(a)) and short phrases (indicated in blue in Fig.1(a)). However, our empirical observations reveal that even the most up-to-date open-source models, trained using sentence annotations, struggle to deal with phrase-level queries, as evident in Fig.1(b). To evaluate the performance of the existing method (Wang et al., 2022), we conduct experiments on the Charades-STA dataset and observe a significant decline in prediction accuracy. Specifically, when confronted with simpler verb and noun phrase queries, the IoU@0.3 metric is dropped by 22.07% and 27.35%, respectively.

Usually, a word or group of words forms a syntactic constituent with a single grammatical function (ie. verb, subject, or object), representing a more straightforward semantic meaning than sentences (no need to understand their compositions). The typical failure in much more straightforward scenarios reveals the following problems. *First, existing models tend to capture the annotation bias in the benchmark but lack sufficient understanding of the intrinsic relationship between simple visual and language concepts.* Consequently, existing models may easily fail when the unrealistic assumption of the in-distribution test setting does not hold, i.e., incapable of generalizing to novel combinations of visual entities and text, which is also revealed by Otani et al. (2020); Yuan et al. (2021); Li et al. (2022). *Second, the models' interpretability and robustness are questioned* since they fail to deal with simple (atomic) concepts, even though they achieve decent results in sentence-level prediction tasks. This may hinder the application of these methods in real scenarios.

Inspired by the insights mentioned earlier, we incorporate phrase-level prediction into the design of temporal localization models. To help the model better accomplish phrase-level localization and avoid the high annotation cost and subjective annotation bias of fine-grained phrases, we both fully exploit the temporal constraints between phrases within the same sentence and attempt to transfer knowledge from externally pre-trained large models. On the one hand, the sentence-level annotation imposes constraints on the internal phrase localization. Thus, we propose a phrase-level Temporal Relationship Mining (TRM) framework to improve the phrase temporal localization using sentence-level supervision only. The three key ideas underpinning this framework are as follows. *Firstly*, drawing inspiration from the effective utilization of Multiple Instance Learning (MIL) in weakly supervised temporal sentence localization, we train the model to distinguish between matched and unmatched video-phrase pairs, all without the need for phrase-level annotations. *Secondly*, in order to consider the constraints of sentence-level annotations on phrase-level predictions, we exploit the temporal localization relationship relevant to the phrase and the whole sentence and follow the two design principles -*consistency* and *exclusiveness*. Specifically, *consistency* requires every phrase-level prediction should share a period with the annotated sentence-level ground truth. As shown in Fig 1(a), all predictions of the phrases "puts on", "gloves", "clean" and "snow" should overlap with the sentence ground truth annotation (in green). *Exclusiveness* requires that every period not intersect the sentence ground truth (as shown in red boxes in Fig 1(a)) is at least excluded from one phrase-level prediction (not intersect at least one phrase prediction).

On the other hand, inspired by the robust generalization capabilities exhibited by recent large-scale pre-trained models, we further extend the TRM framework to TRM

with Phrase-level Training (TRM-PT) which employs a pre-trained model to generate fine-grained pseudo-labels for phrases within sentence queries. We require the model to learn both sentence-level and phrase-level localization and propose a method of pseudo-label refinement and sample re-weighting to mitigate the negative impact of the noise in the phrase pseudo-labels on the model. Meanwhile, we find that existing large-scale pre-trained vision-language models (VLMs) have a poorer understanding of actions in videos compared to their understanding of static states due to image-based pre-training. For instance, we separately evaluate the performance of pseudo labels for verb phrases and noun phrases on the Charades-STA dataset obtained from a pre-trained VLM (Rasheed et al., 2023). We observed a notably lower accuracy in pseudo labels for verb phrases (R@0.5, 32.47% vs. 56.84%). We further noticed that actions in videos often coincide with changes in scene states; for example, 'sitting down' leads to a transition of a person in the video from standing to sitting. Understanding and localizing these static scene states before and after such actions can help the model better understand the action. Therefore, we propose to leverage large-scale language models to predict the states of scenes before and after the occurrence of action phrases and align them with visual content before and after the action. This allows the model to leverage the VLM's strength in understanding static descriptions, thereby improving its ability to localize the action that connects them.

Our contributions are summarized as follows: **(1)** We highlight the importance of phrases in video temporal localization and exploit the temporal relationship relevant to phrases and the whole sentence. **(2)** We propose to generate pseudo-labels for phrases using a large-scale pre-trained visual-language model, and enhancing the model's phrase-level prediction performance by training on these automatically generated labels. **(3)** We propose phrase-level Temporal Relationship Mining (TRM) framework to investigate phrase-level prediction using sentence-level supervision only, which proposes the consistency and exclusiveness constraints to regularize the training process. **(4)** We propose to utilize a large-scale language model to infer changes in the scene's state before and after the occurrence of verb phrases and align them with the visual content to enhance the model's understanding of verb phrases. **(5)** Experiments on Charades-STA and ActivityNet Captions demonstrate our method's ability to improve phrase-level performance while performance in sentence-level settings remains stable, achieving better generalization performance.

Compared with our conference version (Zheng et al., 2023), which first highlighted the challenge of phrase-level generalization, this journal extension investigates how the powerful capabilities of modern large pre-trained models can be harnessed to further enhance phrase-level localization performance. We introduce a new framework, TRM-PT, with several key contributions not explored in the original TRM: In terms of methodology, 1) we are the first to leverage large-scale pre-trained vision-language models (VLMs) to generate fine-grained pseudo-labels for phrases and improve the performance of the model's phrase-level prediction by learning from these pseudo-labels. It is worth noting that even without using pre-trained models, our approach has achieved the best phrase-level predictive performance. Introducing large-scale pre-trained models can further enhance phrase-level predictive performance (details are discussed in Sec. 5.3). 2) We identified that VLMs often struggle with understanding dynamic actions compared to static states. To address this, we propose a novel approach using a large language model (LLM) to infer the scene's state before and after a verb phrase occurs. By aligning these inferred states with the corresponding visual content, we provide richer supervision to improve the model's understanding of verbs. 3) We propose a method of sample re-weighting and pseudo-label optimization to reduce the negative impact of phrase-level pseudo-labels on the model. Experimentally, 1) we verify that our method can be applied to different proposal generation strategies, making it possible to extend to long videos. 2) We conduct new experiments to demonstrate that utilizing pseudo-labels of phrases can not only improve the phrase localization performance but also improve the robustness of sentence-level prediction in cases of limited data and high-noise labels. 3) We conduct a controlled experiment to quantitatively analyze the impact of noisy pseudo-labels on our model's performance. 4) We carry out additional ablation studies on hyperparameters and model architecture to further demonstrate the effectiveness of our method. Additionally, we perform empirical analysis to investigate the impact of various visual features on the model's performance.

2 Related Work

2.1 Temporal Sentence Localization

Since its inception in the work of TALL (Gao et al., 2017), this task has garnered significant attention recently (Li et al., 2023; Lin et al., 2023; Jang et al., 2023; Fang et al., 2023). Previous approaches have generally fallen into two categories: one group generates candidate proposals and subsequently ranks them using multi-modal features (Zhang et al., 2020), while the other leverages multi-modal features directly to make timestamp predictions (Zhang et al., 2020). More recent research has begun to delve into the fine-grained aspects of vision and language. For instance, with regard to vision information, DORi (Rodriguez-Opazo et al., 2021) and MARN (Liu et al., 2022a) take into account object features within the video, leading to enhancements in model performance. Conversely, concerning language fea-

tures, LGI (Mun et al., 2020) generates sub-query features to implicitly consider fine-grained textual attributes, thus elevating sentence localization performance. MMN (Wang et al., 2022) trains models to discern between matched and unmatched video-sentence pairs sourced from both intra-video and inter-video contexts. Additionally, MGSL-Net (Liu et al., 2022b) employs memory to bolster uncommon samples during the training process. EMB (Huang et al., 2022) introduces elastic boundaries to address uncertainties in temporal boundaries. Meanwhile, VISA (Li et al., 2022) examines the distribution of various entities and assesses compositional generalization through the Charades-CG and ActivityNet-CG dataset splits, where novel compositions of seen phrases emerge in the test split. DeCo (Yang et al., 2023) learns a coarse-to-fine compositional representation for compositional temporal grounding. However, in Table 1 and Table 2, we evaluate the phrase-level performance of recent open-source methods and observe that existing methods exhibit subpar performance when dealing with simpler phrases as queries, indicating a lack of genuine comprehension regarding the inherent connection between vision and language. In this paper, we present a unified framework capable of handling both sentence and phrase queries concurrently, leading to performance improvements.

2.2 Multiple Instance Learning

Multiple Instance Learning (MIL) has been widely applied in computer vision, including tasks such as content-based image retrieval (Song et al., 2013), object localization and segmentation (Xu et al., 2015), computer-aided diagnosis and detection (Xu et al., 2014), among others. While (Huang et al., 2021; Yang et al., 2021; Huang et al., 2021; Zheng et al., 2022a, b; Huang et al., 2023) have employed MIL to tackle the challenge of weakly supervised temporal sentence localization, where training data comprises only videos and natural language queries, no prior research has explored its use in addressing the problem of phrase-level video temporal localization. It's worth noting that treating phrase-level prediction directly as a weakly supervised task and introducing MIL overlooks the constraints imposed by sentence-level annotations on phrase-level predictions. Hence, we delve into the relationship between phrase-level predictions and sentence-level annotations, introducing the concepts of consistency and exclusivity as key assumptions in our approach.

2.3 Large-Scale Pre-trained Visual-Language Models

Large-scale pre-trained visual-language models (VLMs) (Radford et al., 2021; Li et al., 2022; Zeng et al., 2022; Sun et al., 2019; Lei et al., 2021; Xu et al., 2021; Ma et al., 2022; Rasheed et al., 2023; Weng et al., 2023) have demonstrated strong generalization capability in various multi-modal tasks. For example, CLIP (Radford et al., 2021) maps the visual and text modalities to the same feature space by minimizing the cosine distance between matched image-text pairs. BLIP (Li et al., 2022) combines image-text contrastive learning with other pre-training tasks such as image caption generation and image-text matching, enabling the generative ability of the model. X-CLIP (Ma et al., 2022), Video Fine-tuned CLIP (Rasheed et al., 2023), and Open-VCLIP (Weng et al., 2023) fine-tune the CLIP model on video-text pairs, making the model more sensitive to video inputs.However, these visual-language pre-trained models are often trained on image-text data or trimmed video-text data, which makes them insensitive to the transitional parts between events in untrimmed videos.

Inspired by the powerful generalization ability of VLMs, some methods attempt to use them for other tasks. For instance, ReCLIP (Subramanian et al., 2022) and CPL (Liu et al., 2023) respectively utilize VLMs in zero-shot and weakly supervised visual grounding tasks to find the most relevant visual regions for a given query. VDI (Luo et al., 2023) propose visual-dynamic injection to empower the image-text pre-training models' ability to capture the video changes. SPL (Zheng et al., 2023), in zero-shot video temporal sentence localization tasks, uses VLMs to generate pseudo-queries along with corresponding pseudo-events. In contrast to them, we propose using pre-trained models to generate pseudo-labels for phrases to assist fully supervised video temporal sentence localization tasks.

Our phrase-level training is inspired by methods like SPL (Zheng et al., 2023) but differs fundamentally in its problem setting and methodology. While SPL targets zero-shot localization, we operate in a fully supervised setting to solve a different problem: learning fine-grained phrase semantics from coarse sentence-level annotations. Methodologically, we identify and solve key limitations of VLM-based pseudo-labeling. We find this approach is inherently more robust for simple phrases than for the complex sentences in SPL; our analysis on Charades-STA shows VLM-generated pseudo-label accuracy is significantly higher for phrases (e.g., 56.84% for nouns) than for sentences (27.14%). Crucially, this revealed that VLMs struggle with dynamic verb phrases (32.47% accuracy) compared to static noun phrases (56.84%). To address this critical weakness, we introduce a core innovation absent in SPL: we leverage an LLM to infer static scene states before and after an action, providing a targeted supervisory signal to improve verb understanding.

2.4 Phrase in Temporal Sentence Localization

Phrase-level features offer models a richer set of fine-grained textual representations and find extensive utility in vision-language tasks like video grounding (Rohrbach et al., 2016), and video captioning (Ryu et al., 2021; Zhang

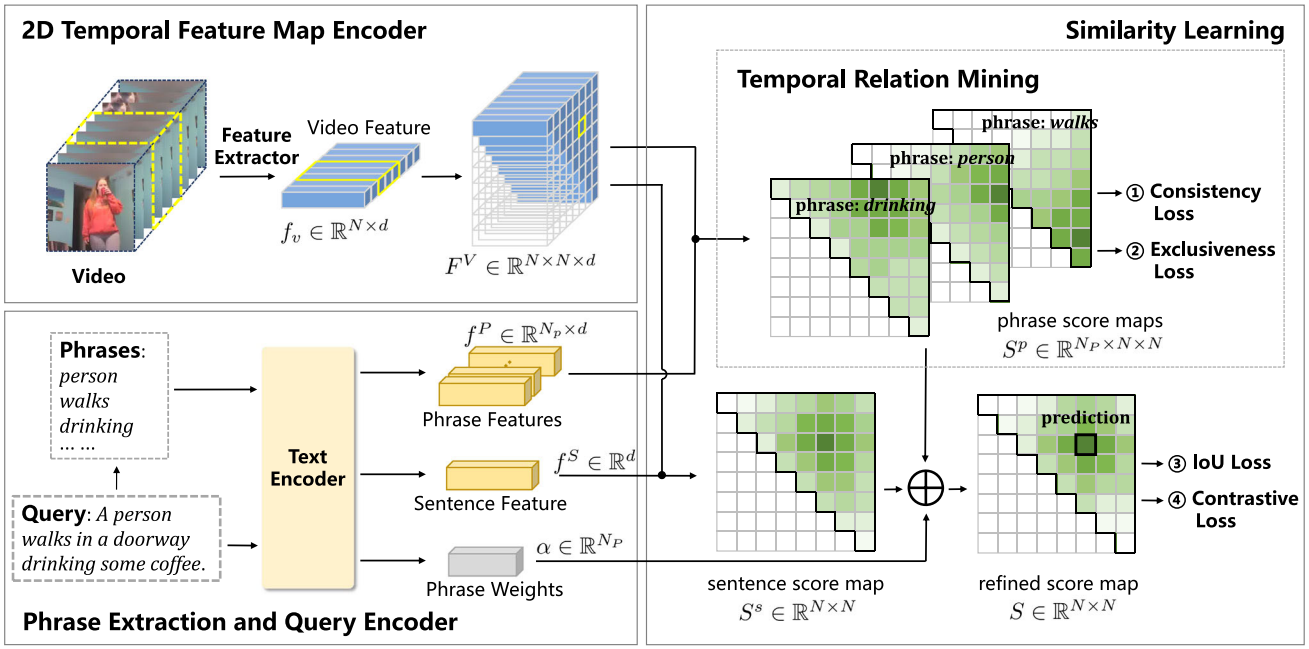


Fig. 2 Our proposed TRM model framework focuses on the temporal relationship between a sentence and its phrases. Our model consists of three modules: a video encoder extracts video features and generates a 2D temporal map; a query encoder extracts both sentence-level and phrase-level features and a similarity learning module to mine the temporal relationship of phrases and sentences based on our two constraints (consistency and exclusiveness) and leverage sentence-level contrastive learning. We apply the phrase-level constraint loss considering the intrinsic relationship between sentences and phrases.

et al., 2019). LGI (Mun et al., 2020) was the pioneering work that harnessed sub-query features, albeit it fused them early on to obtain fine-grained sentence features, without directly pinpointing phrases or considering the connection between phrase localization and sentence understanding. Consequently, LGI's results still exhibited a dramatic drop (mIoU drops by 16.34% on ActivityNet Captions) when handling phrase queries, as indicated in Table2. Subsequently, PLPNet(Li et al., 2022) made the notable stride of directly addressing the challenge of phrase localization, elevating phrase-level localization through the use of contrastive learning. However, it did not impose additional constraints on phrase-level and sentence-level predictions, neglecting the inherent relationship between the video segments corresponding to a sentence and its constituent phrases. In this paper, we present a unified framework capable of handling both sentence and phrase queries concurrently, enhancing the performance of both tasks. We introduce constraints that operate from the perspective of prediction results, allowing the TRM model to directly supervise predicted phrase-level timestamps without the need for extra phrase-level annotations. As far as our knowledge extends, we are the first to explicitly explore the temporal relationship between phrase-level and sentence-level predictions. This setting aligns more closely with real-world application scenarios and empowers the model to generalize to novel combinations of phrases.

3 TRM: Temporal Relationship Mining

3.1 Overview

Figure 2 illustrates the overall architecture of our proposed Temporal Relationship Mining (TRM) framework. We first extract video representation and generate a 2D Temporal Map (Zhang et al., 2020). Meanwhile, the query encoder generates phrases and extracts text features for both phrases and sentences. To represent the similarity between the text and each video proposal, we generate score maps using the 2D temporal map and the text feature for sentences and all phrases. Due to the lack of phrase-level annotation, we explored the *consistency* and *exclusiveness* relationship between phrases and sentences as the loss function to regularize the training process and improve the accuracy of phrase score maps. Since the phrase-level score maps can provide more fine-grained information for the sentence, we use them to refine the sentence score map with a weighted sum option as well, and the weight of each phrase represent its importance. Finally, we optimize the refined sentence score map with an IoU regression loss and a contrastive learning loss.

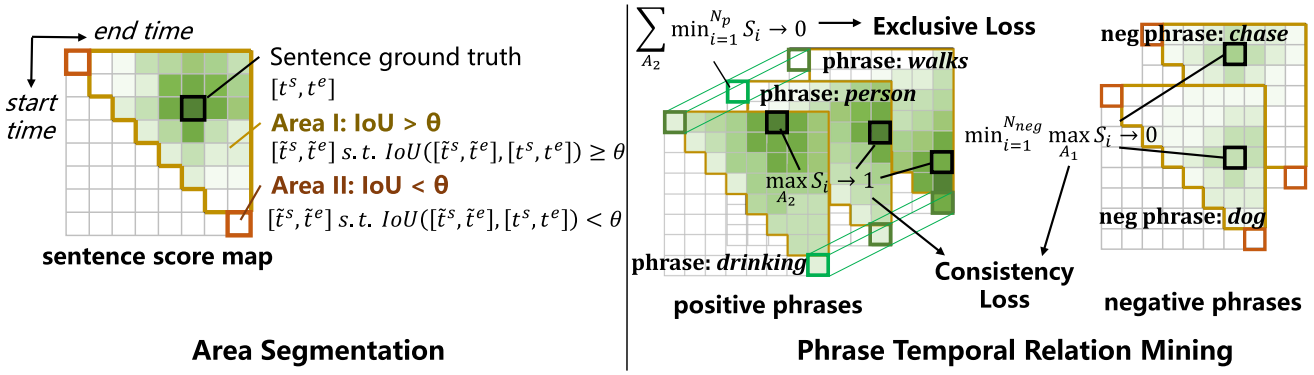


Fig. 3 The specific process of proposal segmentation and implementation of our consistency and exclusiveness principles

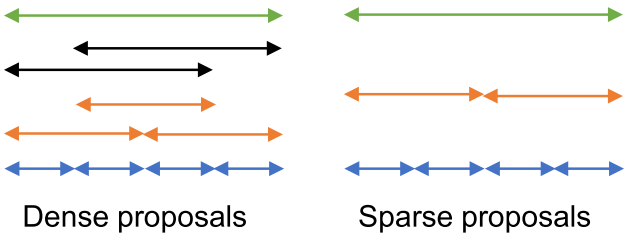


Fig. 4 Different strategies to generate proposals

3.2 Model Architecture

3.2.1 Video Encoder

The video encoder aims at extracting video features and generating a 2D temporal map for similarity learning. We extract features from the input video and encode them as a 2D temporal adjacent feature map following MMN (Wang et al., 2022). To process an input video, we initially divide it into smaller video segments, each consisting of an equal number of frames. Then we extract the clip-level visual feature with a pre-trained CNN model. We can obtain N clip-level features $\{f_i^V\}_{i=1}^N \in \mathbb{R}^{N \times d}$, where N is the number of clips and d is the feature dimension. Then, we build up the 2D proposal feature map $F^V \in \mathbb{R}^{N \times N \times d}$ following MMN (Wang et al., 2022), where proposal $F_{i,j}^V$ represents the video candidate starting from the i-th clip and ending with the j-th clip.

Note that we follow previous works (Wang et al., 2022; Zhang et al., 2020), constructing a dense 2D temporal map with $N \times N$ proposals, to facilitate direct comparison with prior works. This introduces a computational complexity of $O(N^2)$ with respect to the number of video clips N. However, our proposed TRM framework is agnostic to the proposal generation strategy. The dense map used for our main experiments can be replaced with more efficient sparse proposal generation methods, such as those based on hierarchical segment trees (Mu et al., 2024; Pan et al., 2023) as shown in Fig. 4, which can reduce the number of proposals and bring the complexity closer to $O(N)$. This adaptation does not alter the core principles of our methods. We provide a detailed analysis in our ablation studies in Section 6.1, Table 6, and Fig. 7.

3.2.2 Query Encoder

The query encoder aims to generate fine-grained phrases for a sentence and extract both sentence and phrase-level text features. More specifically, given a query sentence S, we first parse N_p phrases $[p_1, p_2, ..., p_{N_p}]$ using pre-trained SRL-BERT(Shi & Lin, 2019). SRLBERT assigns semantic role labels to each word in the sentence, while we only keep the semantic roles with more than 1000 occurrences in the training set as phrases. Then, we use a pre-trained DistilBERT (Sanh et al., 2019) model following MMN (Wang et al., 2022) to extract the features of sentences and phrases at the same time. Phrases provide fine-grained information to the sentence, and the sentence provides global information to phrases. Therefore, we further interact sentence and phrase features through a single-layer transformer encoder (Vaswani et al., 2017). The final sentence feature and phrase features are represented as $f^S \in \mathbb{R}^d$ and $f^P \in \mathbb{R}^{N_p \times d}$ respectively.

3.2.3 Similarity Learning Module

To learn the semantic relevance of each sentence and phrase with each temporal proposal, we generate score maps for both sentence and phrases according to the similarity of text and video features. In order to improve the quality of phrase score maps, we propose two assumptions of consistency and exclusivity to constrain the phrase score maps. Since phrases provide finer-grained semantic information for sentences, we use the phrase score maps to refine the sentence score map so that it can summarize the attentional information for each phrase. We use a weighted sum option over the phrase score maps and leverage phrase weights to describe the importance of different phrases. Finally, we optimize the refined sentence

score map with an IoU regression loss and a contrastive learning loss.

Score Map Generation. For the sentence, we perform 1×1 convolution operation on visual feature map F and perform a linear projection on text features f^S respectively to project the features of two modalities into the same dimension d^H. The final representations of sentence features $f^S_{iou} \in \mathbb{R}^{d^H}$ and visual features $F^V_{iou} \in \mathbb{R}^{N \times N \times d^H}$ are:

$$f^S_{iou} = \mathrm{FC}_{iou}\left(f^S\right), F^V_{iou} = \mathrm{Conv}_{iou}\left(F^V\right) \tag{1}$$

where FC$(\cdot)$ is a fully connected network and Conv$(\cdot)$ is an 1×1 convolution. Then we regard the cosine similarity of f^S_{iou} and F^V_{iou} as sentence-level score map: $S^s = F^{V\mathsf{T}}_{iou} f^S_{iou} \in \mathbb{R}^{N \times N}$, in which $S^s_{i,j}$ represents the similarity score between the sentence and the proposal from the i-th video clip to the j-th video clip.

Temporal Relation Mining. In previous works (Wang et al., 2022; Zhang et al., 2020), the sentence score map is directly used to predict the timestamps. However, it dismisses the fine-grained phrases inside the query, and has poor performance when the query is a single phrase. To solve this problem, we build phrase score maps and mine the temporal relationship between the phrases and the sentence. Due to the lack of phrase-level annotation data, we impose constraints between the phrase score maps for training purposes. We have the following two hypotheses considering the relationship between phrases and sentences:

1. *consistency*: For paired sentences and videos, every phrase-level prediction should share a period with the annotated sentence-level ground truth. For unpaired sentences and videos, at least one phrase-level prediction does not share a period with the annotated ground truth.
2. *exclusiveness*: Each frame outside the ground truth is not contained in at least one phrase-level prediction result.

In detail, we first obtain the text feature $f^P_{i,iou} \in \mathbb{R}^{d^H}$ for the i-th phrase through Eq (1). Then we regard the cosine similarity as moments' estimation score map S^p of each phrase: $S^p_i = F^{V\mathsf{T}}_{iou} f^P_{i,iou} \in \mathbb{R}^{N \times N}$. Inspired by Multiple Instance Learning, we also randomly sample unmatched phrases in a batch and compute their score map $\hat{S}^p$. Based on the degree of intersection with the sentence ground truth, we divide all proposals into two subsets. As shown in the left half of Fig. 3, all the proposals in Area I have an IoU with the ground-truth moment large than a certain threshold θ, while the opposite is true for all proposals in Area II.

Our consistency loss ensures that each phrase-level prediction should be located in Aera I, which is illustrated in Fig.3. That is: for each phrase score map, the max score (marked by black) in Area I should be 1. Our consistency loss also requires that for a negative sentence, there should be at least one phrase that mismatches any proposal in Area I, which is represented in Fig.3 as $\min_{i=1}^{N_{neg}} \max_{A_1} S_i \rightarrow 0$. The consistency loss can be described as follows:

$$\begin{aligned}\mathcal{L}_{con} = &\max_{i=1}^{N_p}\left(L_f\left(\max_{(s,t)\in A_1} S^p_i[s,t], 1\right)\right) \\ &+ \min_{i=1}^{N_p}\left(L_f\left(\max_{(s,t)\in A_1} \hat{S}^p_i[s,t], 0\right)\right)\end{aligned} \tag{2}$$

where $\mathcal{L}_f$ is the focal loss (Lin et al., 2017) to balance the positive and negative samples, A_1 represents Area I, and A_2 represents Area II.

Our exclusiveness loss requires that each proposal in Area II should mismatch at least one phrase of the query sentence. That is: as shown in Fig. 3, at least one of the phrase's scores should be 0 (i.e. the minimum score marked by green should be 0) for all the proposals in Area II. The exclusiveness loss can be described as follows:

$$\mathcal{L}_{ex} = \frac{1}{|A_2|}\sum_{(s,t)\in A_2} L_f\left(\min_{i=1}^{N_p}(S^p_i[s,t]), 0\right) \tag{3}$$

Sentence Score Map Refinement.. Since the phrase-level score maps can provide more fine-grained information for the sentence, we use them to refine the original sentence score map $S^s \in R^{N \times N}$. We gain the final sentence score map $S \in R^{N \times N}$ by aggregating the score maps of the sentence and all of its phrases, which is shown as follows:

$$\alpha = \mathrm{softmax}(\mathrm{MLP}([p_1, p_2, ..., p_{N_p}])) \tag{4}$$

$$S = S^s + \sum \alpha_i S^p_i \in R^{N \times N} \tag{5}$$

where $\alpha \in \mathbb{R}^{N_p}$ is the phrase weights that describe the importance of different phrases $p_1, p_2, ..., p_{N_p}$, MLP denotes a multilayer perception with a output layer of 1-dimension.

To supervise the sentence score map, we apply the binary cross entropy loss to regress the IoU score of each proposal. Following (Zhang et al., 2020), we adopt a scaled IoU value y_i as the supervision scale, but not a hard binary score. Then the binary cross entropy loss can be expressed as

$$\mathcal{L}_{iou} = -\frac{1}{C}\sum_{i=1}^{C}(y_i \log S_i + (1-y_i)\log(1-S_i)), \tag{6}$$

where C is the number of proposals.

Sentence-level Contrastive Learning. Following MMN (Wang et al., 2022), we also use contrastive learning to

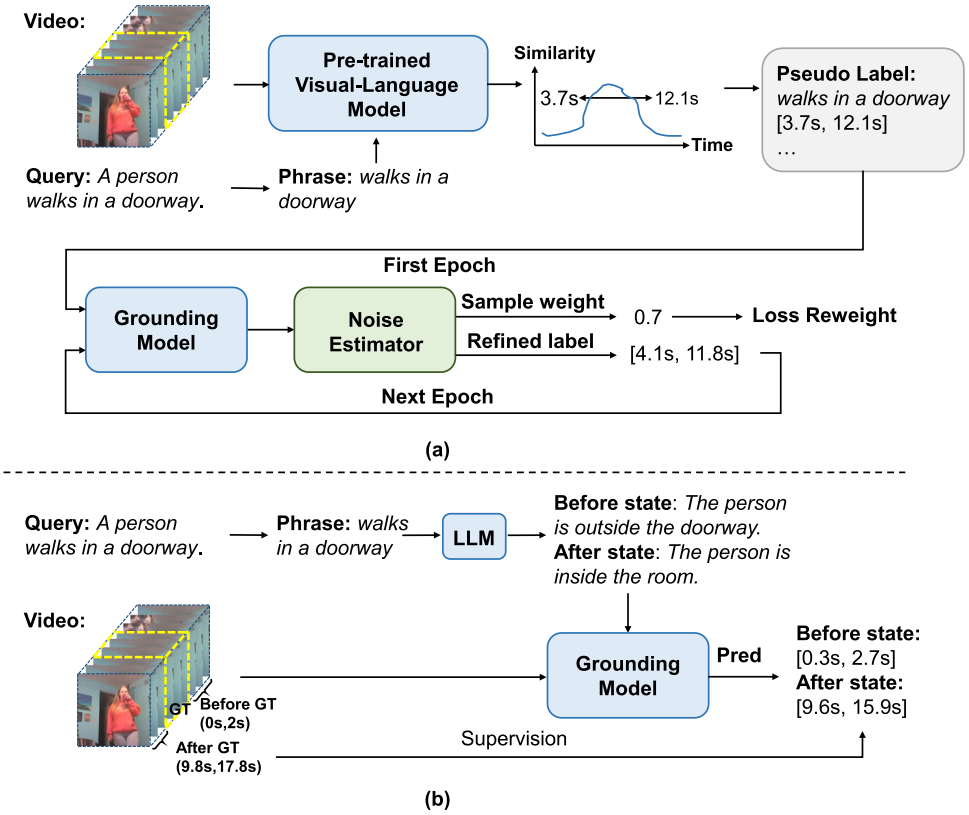


Fig. 5 The phrase-level training pipeline. (a) Due to the lack of phrase-level annotation, we extract phrases from sentences and then utilize a pre-trained visual-language model to calculate the similarity between phrases and video segments. Then, we generate phrase pseudo-labels based on visual-phrase similarity and use them to train the model. To mitigate the negative impact of noise in the pseudo-labels, we propose a noise estimator to reduce the weight of the loss function for high-noise samples and refine the pseudo-labels based on the model's predictions. (b) To improve the understanding of actions in videos, we propose to leverage large-scale language models to predict the states of scenes before and after the occurrence of action phrases. We use video clips before and after the occurrence of ground-truth as time labels for the state before and after the action phrase, and we use these state descriptions from the large language model to train the model

provide more supervised signals to the model. We collect positive and negative sentence-video pairs within and between videos, and use noise contrastive estimation (Oord et al., 2018) to estimate two conditional distributions $p(s|v)$ and $p(v|s)$. The former represents the probability that a sentence s matches the video v when giving v, and the latter represents the probability that a video v matches the sentence s when giving s. We adopt the contrastive loss to help capture better information between modalities as follows:

$$\mathcal{L}_{cont} = -\left(\sum_{s\in\mathbb{S}} \log p(v_s|s) + \sum_{v\in\mathbb{V}} \log p(s_v|v)\right) \tag{7}$$

where $\mathbb{S}$, $\mathbb{V}$ are the sets of training sentences and video in a batch, v_s is the video that matches the sentence s, and s_v is the sentence that matches the video v.

The total loss of our model is as follows.

$$\mathcal{L}_{sent} = \mathcal{L}_{iou} + \mathcal{L}_{cont} + \mathcal{L}_{con} + \mathcal{L}_{ex} \tag{8}$$

Given the lack of phrase-level annotations, we can still optimize the understanding of phrases during training with the constraints between the whole sentence and phrases.

4 TRM-PT: Temporal Relationship Mining with Phrase-Level Training

Using only the TRM in Sec. 3.2, our model has already achieved the best phrase-level localization performance. However, the consistency and exclusive constraints proposed in Sec. 3.2 rely on sentence-level annotations to implicitly supervise phrase-level predictions, leaving the phrase-level prediction results still lacking explicit supervision. Recently, the strong generalization abilities of large-scale pre-trained models have inspired their use in generating supervision for various tasks. Therefore, we introduce the Temporal Relationship Mining with Phrase-level Training (TRM-PT) framework, shown in Fig.5, which extends TRM in two sig-

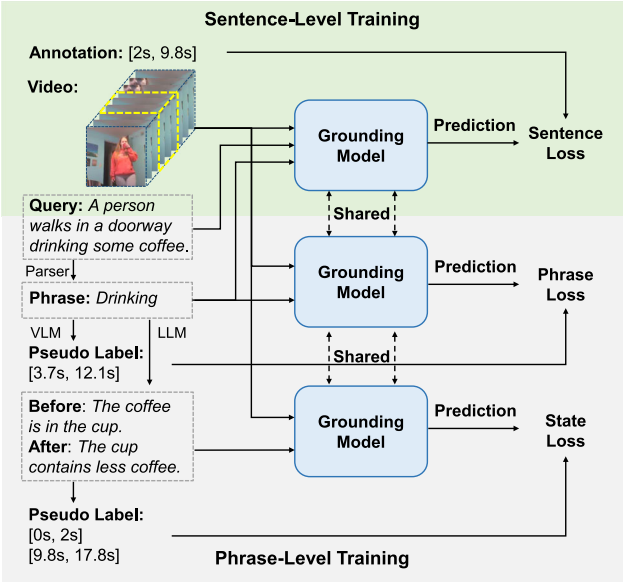


Fig. 6 Our proposed TRM-PT pipeline which utilizes sentence-level annotations and pre-trained vision-language models to assist in phrase-level predictions. We refer to the training process of the preliminary version of TRM described in Sec. 3.2 as sentence-level training, which use annotated videos and sentence queries to train the model. For phrase-level training, due to a lack of phrase-level annotations, we extract phrases from sentences and generate pseudo-labels from pre-trained vision-language models to train the model. We also use a large-scale language model to infer changes in the scene's state before and after the occurrence of verb phrases and align them with the visual content. This helps the model understand what the states typically look like before and after an action, thereby better assisting the model in locating verbs

nificant ways. First, we utilize large pre-trained models to generate phrase-level pseudo-labels, providing the explicit supervision that was previously missing. To address the noise inherent in these generated labels, we propose a Noise Estimator that re-weights training samples and iteratively refines the pseudo-labels. Second, we find that existing pre-trained vision-language models often have a poorer understanding of dynamic actions compared to static states. We observe that actions frequently coincide with changes in scene states (e.g., 'sitting down' involves a transition from standing to sitting). Understanding these states can help the model better localize the action itself. We, therefore, leverage a large-scale language model (LLM) to predict the states of scenes before and after an action phrase occurs, and we align these state descriptions with the corresponding visual content. This helps the model learn what the scene typically looks like before and after an action, providing a richer, more robust supervisory signal.

4.1 Overview

As we can see in Fig. 5 (a), due to the lack of phrase-level annotation, we first extract phrases from sentences and then utilize a pre-trained visual-language model to calculate the similarity between phrases and video segments. Then, we generate phrase pseudo-labels based on visual-phrase similarity and use them to train our TRM model. To mitigate the negative impact of noise in the pseudo-labels, we propose a noise estimator. On the one hand, it reduces the weight of the loss function for high-noise samples, and on the other hand, it refines the pseudo-labels based on the model's predictions. The refined pseudo-labels will be used for the training of the next epoch. Through iterative refinement, the quality of phrase-level pseudo-labels is improved and the model can learn from those pseudo-labels.

In Fig. 5 (b), to improve the model's understanding of actions in the video, we train the model to understand and localize the static scene states before and after the action in the query. Specifically, we prompt LLM to predict the states of scenes before and after the occurrence of action phrases and use the state descriptions as the queries to train the grounding model. We use the video segments with a fixed duration before and after the annotation to supervise the grounding results of the state descriptions before and after the action phrase, respectively. In our original TRM model, the IoU loss and contrastive loss only require the target segment to be semantically close to the sentence query, while it does not have explicit semantic constraints on the segments before and after the target. In contrast, our TRM-PT model explicitly requires the semantics of the segments before and after the target to correspond to the state descriptions before and after the action occurs, thereby providing a stronger supervisory signal. This helps the model understand what the states before and after the action typically look like, thereby better assisting the model in understanding of verb phrases.

We refer to the training process of the preliminary version of TRM described in Sec. 3.2 as sentence-level training. In Fig. 6, we show how phrase-level training collaborates with our sentence-level training. On one hand, we use the annotated videos and sentence queries to train the model as described in Sec. 3.2 (i.e. the sentence-level training in Fig. 6). On the other hand, for phrase-level training in Fig. 6), due to a lack of phrase-level annotations, we first extract phrases from sentences and generate pseudo-labels using a pre-trained visual-language model. We also use LLM to generate state descriptions before and after the action phrases as additional training data to improve the model's understanding of action phrases. To obtain a unified model capable of performing both phrase-level and sentence-level localization simultaneously, we share the model weights between the phrase-level and sentence-level training.

4.2 Pseudo Label Generation

In this step, we first generate phrase-level queries from sentence-level queries using pre-trained SRLBERT(Shi &

Lin, 2019) similar to the sentence-level training. Then, we generate pseudo-labels for each phrase-level query. Inspired by the powerful generalization ability of recent large-scale pre-trained vision-language models, we propose to use the pre-trained model to generate pseudo-labels for phrases. Specifically, for each phrase, we utilize a pre-trained VLM to extract text features $F^p \in \mathbb{R}^{N_p \times D}$, where D represents the feature dimension and N_p is the number of extracted phrases. Similarly, for the video, we the VLM to extract the video features $F^v \in \mathbb{R}^{N \times D}$, where N is the number of frames. Since the VLM text and visual feature spaces are well aligned, we can directly measure the relevance between the query and the video frame by calculating the cosine similarity of their respective features:

$$Sim = \frac{F^p F^{v\intercal}}{\|F^p\|\|F^v\|} \in \mathbb{R}^{N_p \times N} \tag{9}$$

Inspired by SPL (Zheng et al., 2023), we aim to generate high-quality pseudo-labels where the videos inside each label are highly relevant to the query while those outside are less relevant. To do this, we enumerate possible pseudo-labels $l_1, l_2, ..., l_{N_l}$ using the sliding window, where N_l is the total number of pseudo-labels. For each pseudo-label l_i, we compute the average similarity between the query and videos inside l_i, as well as the average similarity between the query and videos outside l_i. The difference between these two average similarities is then used as a quality score for the pseudo-label l_i:

$$Q_{ik} = \frac{1}{\|l_k\|} \sum_{j \in l_k} Sim_{ij} - \frac{1}{N - \|l_k\|} \sum_{j \notin l_k} Sim_{ij} \tag{10}$$

where Q_{ik} is the quality of the i-th phrase to the k-th pseudo-label proposal, S_{ij} is the relevance of the i-th phrase and the the j-th frame, and $\|l_k\|$ is the number of frames in the pseudo-label proposal l_k. Finally, the pseudo-label with the maximum quality score is selected to supervise the model training:

$$y_i = l_{\hat{k}}, \hat{k} = \arg\max_k Q_{ik} \tag{11}$$

This process allows us to generate pseudo-labels that maximize within-label similarity and minimize between-label similarity to the query.

4.3 Noise Estimator

After obtaining phrase-level pseudo-labels, we can directly use Eq.(8) to train the model. However, the pseudo-events may not be accurate enough, and the noise of the pseudo-labels may have a negative impact on the model. Inspired by SPL (Zheng et al., 2023), we use a noise estimator to estimate the pseudo-label noise and reduce the weight of the loss function for high-noise samples and refine the pseudo-labels based on the model's predictions. Specifically, if the model is confident in its own prediction and its prediction is close to the pseudo label, we consider the pseudo label to have low noise. Therefore, we define the cleanliness of a pseudo-label as $c = \alpha S_p + (1 - \alpha) IoU_p$, where p is the model's prediction, S_p is the sentence score of the prediction defined in Eq.(5), IoU_p is the intersection-over-union (IoU) between the prediction p and the pseudo-label, and α is a hyperparameter. We require that samples with a high level of cleanliness will have a higher training weight. Therefore, we use c to re-weight the loss of each sample. The final phrase-level training loss is:

$$\mathcal{L}_{phrase} = \sum_{i \in \mathbb{B}} c_i (\mathcal{L}_{iou} + \mathcal{L}_{cont} + \mathcal{L}_{con} + \mathcal{L}_{ex}) \tag{12}$$

where $\mathbb{B}$ is a set of training samples in a batch, $\mathcal{L}_{iou}$, $\mathcal{L}_{cont}$, $\mathcal{L}_{con}$, and $\mathcal{L}_{ex}$ are the same training loss as the sentence-level training in Eq.(8).

By the sample reweighting, we can avoid the negative impact of noisy samples on the model as much as possible. However, this alone is still insufficient because merely using sample reweighting, the model still cannot learn more correct samples. Thus, we further introduce the sample refinement to dynamically refine the pseudo-labels during the training process to improve their quality. Specifically, we can choose a new pseudo-label proposal with the highest cleanliness score c as the new pseudo-label for the next epoch training. We select the $\hat{k}$-th proposal as the refined pseudo-label, where $\hat{k} = \arg\max_k (\alpha S_k + (1 - \alpha) IoU_k)$. The model can refine the noisy label to the correct one if it has enough confidence in predicting the right label. The refined phrase-level pseudo-labels will be used in the training of the next epoch.

4.4 State Queries

As existing large-scale visual language models are trained with image-text pairs or trimmed video-text pairs, they have a poorer understanding of actions in an untrimmed video compared to their understanding of static states. Actions in videos often coincide with changes in scene states, and understanding these static scene states before and after such actions can help the model understand the action. Therefore, we propose to leverage large-scale language models to predict the states of scenes before and after the occurrence of action phrases and use these state descriptions as additional training data to improve the model's understanding of action phrases.

Specifically, we use the pre-trained SRLBERT(Shi & Lin, 2019) to extract phrases in the sentence as we described in

Sec. 3.2.2. As we focus on improving the model's understanding of actions, we only keep the verb phrases according to the semantic role label predicted by SRLBERT. Then, for each verb phrase, we require LLM to describe the status changes of objects before and after the verb. As a verb may correspond to multiple possible states before and after its occurrence, we require LLM to describe all possible states of the object as comprehensively as possible. The state description before the verb occurs should match the visual content before the target segment in the video, and the state description after the verb occurs should match the visual content after the target segment in the video. Therefore, for the target segment (st, en), we set the ground-truth for the state description before the verb occurs as $(st - \tau, st)$, and the ground-truth for the state description after the verb occurs as $(en + \tau, en)$, where τ is a hyperparameter. At this step, we do not use VLM to generate pseudo-labels for status descriptions because the ground truth of the sentence provides a more accurate prior. The verbs in the sentence occur within the ground truth video segment so that the segment before the ground truth should represent the state before the action, while the segment after the ground truth should represent the state after the action. Finally, we use the state descriptions and their corresponding ground-truth as additional samples to train the TRM model. In the TRM model, since the state before/after the verb occurs may correspond to multiple descriptions, we take the mean of the text features of all descriptions as the text feature of that state. The text features of sentences, phrases, and state descriptions will further interact through a layer of transformer encoder. Through this interaction, the model can obtain helpful information from the state descriptions to aid in the localization of sentences and phrases. For example, the localization results of sentences and phrases should be situated within the video segments corresponding to the state descriptions. The final state loss is:

$$\mathcal{L}_{state} = \frac{1}{2}(\mathcal{L}_{before} + \mathcal{L}_{after}) \tag{13}$$

where $\mathcal{L}_{before}$ and $\mathcal{L}_{after}$ are the loss for the state description before and after the verb occurs calculated by Eq.(8). Since the ground-truth for state descriptions is generated based on the ground truth of sentences, this helps the model more easily learn the relationship between sentence queries and state queries. Therefore, we did not introduce noise estimation in the pseudo labels for state descriptions to ensure that the segments corresponding to the sentences are situated within the video segments corresponding to the state descriptions.

4.5 Training and Inference

4.5.1 Training

The total loss of our model is as follows.

$$\mathcal{L} = \mathcal{L}_{sent} + \beta\mathcal{L}_{phrase} + \gamma\mathcal{L}_{state} \tag{14}$$

where β and γ are the hyper-parameters to balance three losses.

4.5.2 Inference

At the inference time, when given a sentence query, we can not only obtain the refined score maps S through Eq(5) to make sentence-level predictions but also use the phrase score map S^p to make phrase-level predictions, which is demonstrated in the qualitative results in Section 7. When given a single phrase query, we can treat it as a sentence (as the text encoders for phrase and sentence are shared). In this case, the score maps of the sentence and phrase are the same and both can be used to output phrase predictions.

5 Experiments

5.1 Dataset

Charades-STA Charades-STA (Gao et al., 2017) originates from Charades (Sigurdsson et al., 2016) dataset, containing indoor videos with sentence queries and corresponding annotations. There are 12,408 and 3,720 video-query pairs for training and testing respectively. Our sentence-level results are reported on the test split.

ActivityNet Captions ActivityNet Captions (Krishna et al., 2017) contains 20K videos, with 37,417/17,505/17,031 video-query pairs in the train /val_1/val_2 split. We adopt standard splits and report the sentence-level results on the val_2 split.

5.2 Experiment Settings

Evaluation Metric. Following (Gao et al., 2017), we adopt the "R@1, IoU $= m$" and mIoU (the mean average IoU) metrics to evaluate the model's performance. Specifically, this metric evaluates the percentage of predicted moments that have the temporal Intersection over Union (IoU) larger than the threshold m, and m is set to $\{0.3, 0.5, 0.7\}$.

Evaluation for phrase. When evaluating the performance of phrases, we use a single phrase rather than a complete sentence as the query, in which case the score map of the sentence

Table 1 Sentence-level and Phrase-level prediction accuracy on Charades-STA

Method		feature	Sentence prediction				Phrase prediction			
			IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
SAP Chen and Jiang (2019)	AAAI 19	VGG	—	27.42	13.36	—				
MAN Zhang et al. (2019)	CVPR 19		—	41.24	20.54	—				
LGI Mun et al. (2020)	CVPR 20		57.20	40.70	20.13	38.75				
FVMR Gao and Xu (2021)	ICCV 21		—	42.36	24.14	—				
DRN Zeng et al. (2020)	CVPR 20		—	42.90	23.68	—				
SSCS Ding et al. (2021)	ICCV 21		—	43.15	25.54	—				
CBLN Liu et al. (2021)	CVPR 21		—	43.67	24.44	—				
CPN Zhao et al. (2021)	CVPR 21		**64.41**	46.08	25.06	**43.90**				
G2L Li et al. (2023)	ICCV 23		—	47.91	28.42	—				
2D-TAN Zhang et al. (2020)	AAAI 20		57.31	42.8	23.25	—	45.15	23.22	10.14	—
MMN Wang et al. (2022)	AAAI 22		60.48	47.45	27.15	—	38.41	22.19	10.1	—
SPL Zheng et al. (2023)	ACL 23		60.73	40.70	19.62	40.47	39.14	21.46	8.17	27.69
PLPNet Li et al. (2022)	ICMR 23		57.82	41.88	20.56	39.12	46.24	22.94	7.69	28.46
PTAN Wei et al. (2024)	ICMR 24		61.16	45.13	24.68	41.69	47.29	26.62	12.10	30.59
TRM (ours)	AAAI 23	VGG	60.67	47.77	28.01	42.77	57.03	33.69	11.86	35.82
TRM-PT (ours)			61.57	**48.13**	**28.97**	42.81	**58.21**	**34.65**	**12.85**	**36.75**

Table 2 Sentence-level and phrase-level prediction accuracy on ActivityNet Captions

Method		Feature	Sentence prediction				Phrase prediction			
			IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
DORi Rodriguez-Opazo et al. (2021)	WACV 21	C3D	57.89	41.49	26.41	42.78				
BPNet Xiao et al. (2021)	AAAI 21		58.98	42.07	24.69	42.11				
VSLNet Zhang et al. (2020)	ACL 20		63.16	43.22	26.16	43.19				
DeNet Zhou et al. (2021)	CVPR 21		61.93	43.79	—	—				
CPN Zhao et al. (2021)	CVPR 21		62.81	45.10	28.10	45.70				
DRN Zeng et al. (2020)	CVPR 20		—	45.45	24.36	—				
SeqPAN Zhang et al. (2021)	ACL 21		61.65	45.50	28.37	45.11				
FIAN Qu et al. (2020)	MM 20		64.10	47.90	29.81	—				
CBLN Liu et al. (2021)	CVPR 21		66.34	48.12	27.60	—				
SMIN Wang et al. (2021)	CVPR 21		—	48.46	30.34	—				
MGSL-Net Liu et al. (2022b)	AAAI 22		—	**51.87**	31.42	—				
BMRN Seol et al. (2023)	CVPR 23		—	48.47	31.15	—				
G2L Li et al. (2023)	ICCV 23		—	51.68	**33.35**	—				
MS-DETR Jing et al. (2023)	ACL 23		62.12	48.69	31.15	46.82				
SnAG Mu et al. (2024)	CVPR 24		-	48.55	30.56	-				
LGIMun et al. (2020)	CVPR 20		58.48	41.65	24.1	41.48	35.39	21.07	9.76	25.14
2D-TANZhang et al. (2020)	AAAI 20		59.45	44.51	27.38	—	51.71	42.19	32.22	—
MIGCNZhang et al. (2021)	TIP 21		60.03	44.94	27.85	43.59	42.25	33.75	16.37	30.9
RaNetGao et al. (2021)	EMNLP 21		60.96	45.59	28.67	44.82	47.44	37.51	27.58	38.45
MMNWang et al. (2022)	AAAI 22		65.05	48.59	29.26	—	51.91	42.27	32.88	—
SPL Zheng et al. (2023)	ACL 23		50.24	27.24	15.03	35.44	34.13	18.69	9.46	21.57
PLPNet Li et al. (2022)	ICMR 23		56.92	39.20	20.91	39.53	50.10	38.12	25.24	37.96
PTAN Wei et al. (2024)	ICMR 24		61.11	47.58	31.30	45.41	50.43	41.86	33.74	42.47
TRM (ours)	AAAI 23	C3D	66.41	50.44	31.18	47.68	52.46	42.84	33.68	43.29
TRM-PT (ours)			**66.92**	51.54	31.85	**47.76**	**53.79**	**44.01**	**34.21**	**44.23**

Table 3 Compositional generalization results on ActivityNet-CG dataset. † denotes the results relying on external detector knowledge

	Method		Test-Trivial			Novel-Composition			Novel-Word		
			IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU
Weakly-supervised	WSLL Duan et al. (2018)	NeurIPS 18	11.03	4.14	15.07	2.89	0.76	7.65	3.09	1.13	7.10
RL-based	TSP-PRL Wu et al. (2020)	AAAI 20	34.27	18.80	37.05	14.74	1.43	12.61	18.05	3.15	14.34
Proposal-free	LGI Mun et al. (2020)	CVPR 20	43.56	23.29	41.37	23.21	9.02	27.86	23.10	9.03	26.95
	VLSNet Zhang et al. (2020)	ACL 20	39.27	23.12	42.51	20.21	9.18	29.07	21.68	9.94	29.58
	VISA† Li et al. (2022)	CVPR 22	47.13	29.64	44.02	31.51	16.73	35.85	30.14	15.90	35.13
	DeCo Yang et al. (2023)	CVPR 23	47.38	28.43	46.03	28.69	12.98	32.67	-	-	-
Proposal-based	TMN Liu et al. (2018)	ECCV 18	16.82	7.01	17.13	8.74	4.39	10.08	9.93	5.12	11.38
	2D-TAN Zhang et al. (2020)	AAAI 20	44.50	26.03	42.12	22.80	9.95	28.49	23.86	10.37	28.88
	SPL Zheng et al. (2023)	ACL 23	28.41	17.43	34.87	19.45	7.63	21.59	22.58	10.49	27.31
	PTAN Wei et al. (2024)	ICMR 24	50.66	34.45	48.75	31.77	16.41	34.07	31.22	15.99	34.32
	TRM (Ours)	AAAI 23	**55.22**	35.06	**51.85**	33.80	16.86	35.80	**35.49**	17.68	**37.50**
	TRM-PT (Ours)		55.04	**35.21**	51.37	**35.01**	**18.01**	**36.75**	35.21	**17.84**	37.35

Table 4 Compositional generalization results on Charades-CG dataset. † denotes the results relying on external detector knowledge. The dark row indicates the results using I3D features fine-tuned on the Charades dataset. ∗ indicates the results reproduced using the same features as ours with officially released code

	Method		Test-Trivial			Novel-Composition			Novel-Word		
			IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU
Weakly-supervised	WSLL Duan et al. (2018)	NeurIPS 18	15.33	5.46	18.31	3.61	1.21	8.26	2.79	0.73	7.92
RL-based	TSP-PRL Wu et al. (2020)	AAAI 20	39.86	21.07	38.41	16.30	2.04	13.52	14.83	2.61	14.03
Proposal-free	LGI Mun et al. (2020)	CVPR 20	49.45	23.80	45.01	29.42	12.73	30.09	26.48	12.47	27.62
	VLSNet Zhang et al. (2020)	ACL 20	45.91	19.80	41.63	24.25	11.54	31.43	25.60	10.07	30.21
	VISA† Li et al. (2022)	CVPR 22	53.20	26.52	47.11	**45.41**	**22.71**	**42.03**	42.35	20.88	40.18
	DeCo Yang et al. (2023)	CVPR 23	58.75	28.71	49.06	47.39	21.06	40.70	-	-	-
Proposal-based	TMN Liu et al. (2018)	ECCV 18	18.75	8.16	19.82	8.68	4.07	10.14	9.43	4.96	11.23
	2D-TAN Zhang et al. (2020)	AAAI 20	48.58	26.49	44.27	30.91	12.23	29.75	29.36	13.21	28.47
	SPL Zheng et al. (2023)	ACL 23	44.21	22.78	39.63	23.41	8.54	18.63	24.14	11.63	25.98
	PTAN Wei et al. (2024)	ICMR 24	62.73	39.24	53.48	47.53	26.09	42.09	53.81	34.68	47.71
	PTAN* Wei et al. (2024)	ICMR 24	47.19	26.87	43.08	33.41	16.12	32.82	40.29	22.45	37.91
	TRM (Ours)	AAAI 23	55.38	34.08	48.48	40.98	20.81	37.13	**44.60**	26.33	41.04
	TRM-PT (Ours)		**56.32**	**34.84**	**49.24**	41.14	22.43	39.02	44.43	**26.47**	**41.14**

and phrase is the same and both can be used to output predictions. Due to the lack of phrase-level annotations, we adopt the action annotation used for the Temporal Action Localization task and use the action names as the query phrases. Although we only tested with verbs, our model can handle arbitrary phrases. To prove this, we also use the object annotations on the Charades-STA dataset provided by Yuan et al. (2017). We collect the common noun phrases in the sentences, and get the time of the first appearance and the last disappearance of the object in the object annotation as the noised noun phrase ground truth timestamps. We report the evaluation results of our model when using noun phrases as queries in the ablation section. It is worth noting that we only use the phrase-level annotations for evaluating the model's performance on phrases, and avoid using them in the training process. So our experiment setting is fair compared with others.

Implementation Details. For the 2D temporal feature map encoder, we use exactly the same settings with 2D-TAN (Zhang et al., 2020) and MMN (Wang et al., 2022) for fair comparisons. We use the VGG (Simonyan & Zisserman, 2015) features for the Charades-STA dataset and C3D features (Tran et al., 2015) for the ActivityNet Captions dataset, and the number of sampled clips N is 16 for Charades-STA and 64 for ActivityNet Captions. For the text encoder, we use the HuggingFace (Wolf et al., 2019) implementation of

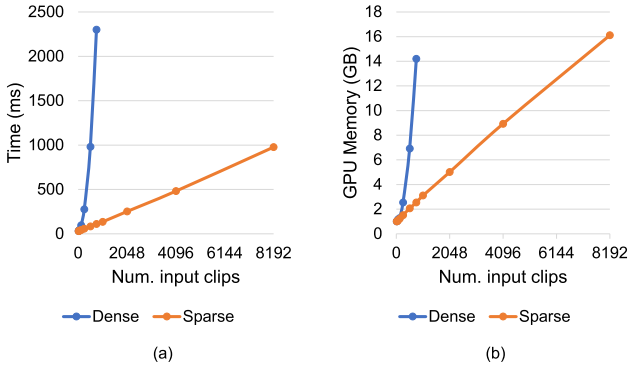


Fig. 7 Scalability analysis on videos of increasing length. (a) Inference time per video. (b) Peak GPU memory usage. The dense proposal method (blue) shows quadratic scaling, while the sparse proposal method (orange) scales near-linearly

DistilBERT (Sanh et al., 2019) with pre-trained model following MMN (Wang et al., 2022). The hyper-parameter τ is set to 4 for both datasets. The threshold θ for dividing proposals Area 1 and Area 2 in Fig. 3 is set to 0.1. We use BLIP-2 (Li et al., 2022) to evaluate the similarity between video frames and phrases to generate phrase pseudo labels. We use AdamW (Loshchilov & Hutter, 2019) optimizer with learning rate 1×10^{-4} and batch size 12 for Charades, learning rate 1×10^{-4} and batch size 20 for ActivityNet Captions. The learning rate of DistilBERT is 1/10 of our main model.

5.3 Comparison with Other Methods

This part compares state-of-the-art models and TRM's ability to deal with sentence-level and phrase-level prediction. On both Charades-STA and ActivityNet Captions datasets, we use sentences and verb phrases (obtained from action labels used for the temporal action localization task) as queries respectively. We assess the phrase-level localization performance of various recent approaches, provided their code is publicly accessible. For a fair comparison, all methods use C3D (Tran et al., 2015) features on ActivityNet Captions and VGG (Simonyan & Zisserman, 2015) features on Charades-STA [1]. 'TRM' represents our TRM model with only sentence-level training and 'TRM-PT' represents our TRM model with both sentence-level and phrase-level training.

As shown in Table 1 and Table 2, TRM achieves comparable results when using completed sentences as queries and achieves an absolute advantage when using verb phrases as queries. All the existing methods we reproduced have a sheer drop when using phrases as queries. For example, on the Charades-STA dataset, the phrase-level prediction performance of MMN on the metric 'IoU=0.3' is dropped by 22.07% compared with the sentence-level performance. This reveals that existing models lack sufficient understanding of the intrinsic relationship between simple visual and language concepts. When introducing phrase information, the gap is narrowed to 3.64%, which demonstrates the effectiveness of our method. We also compare our method with SPL (Zheng et al., 2023), which generates pseudo-labels for zero-shot localization. As we can see, our TRM-PT framework significantly outperforms SPL across both datasets. This underscores the effectiveness of our approach, which is specifically designed to handle phrase-level learning and address the weaknesses of VLMs in understanding verbs by inferring the state changes before and after the verb.

We can also find that when using phrase-level training to provide more supervision signals, the phrase-level performance can be further improved. This demonstrates the effectiveness of our phrase-level training, which generates phrase-level pseudo-labels and estimates the noise in the pseudo-labels to refine them. Although our method shows less significant improvement in sentence-level prediction performance, the experiments in Fig 8(a) demonstrate that our method performs better when the amount of training data is limited.

5.4 Compositional Generalization

VISA (Li et al., 2022) constructed the ActivityNet-CG and Charades-CG datasets by resplitting the ActivityNet and Charades-STA datasets to validate the model's compositional generalization. Both of them consist of the training split, the test-trivial split, the novel-composition split, and the novel-word split, where the test-trivial split has the same distribution as the training split, the novel-composition split includes unseen compositions of seen phrases, and the novel-word split includes unseen words. We use ActivityNet-CG and Charades-CG datasets to evaluate the generalization performance of our model in Table 3 and Table 4.

The Novel-Composition split includes novel compositions of seen phrases, which evaluate the model's ability to understand phrases and to generalize to novel compositions of phrases. This is most relevant to our research question, and our TRM-PT model shows a clear and significant improvement over the original TRM. 1) As shown in Table 3, on the ActivityNet-CG dataset, our method achieves the best performance across all splits. Notably, on the Novel-Composition split, our TRM-PT model outperforms the next best method, VISA (Li et al., 2022), by 3.5% on the metric '$IoU = 0.5$'. This is significant because our model was not specifically designed for compositional generalization, yet by learning fine-grained phrase-level concepts and their relationship to the sentence, it exhibits superior generaliza-

[1] For a fair comparison, some methods utilizing object-level features (Rodriguez-Opazo et al., 2021; Liu et al., 2022a) or large-scale pre-trained models (Luo et al., 2023; Wang et al., 2023) have not been included in the Table 2 and Table 1.

Table 5 Ablation studies on the influence of phrase and score map and the implementation of our hypotheses in our TRM model

Method			Sentence prediction			Verb phrase prediction			Noun phrase prediction		
Phrase	Consistency	Exclusiveness	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
✗	✗	✗	60.48	47.45	27.15	38.41	22.19	10.01	33.13	8.17	3.15
✔	✗	✗	59.84	46.65	26.99	41.13	22.63	10.60	35.41	7.36	2.68
✔	✔	✗	60.22	46.56	27.31	56.69	30.85	10.85	71.12	51.67	8.57
✔	✗	✔	60.13	45.89	27.80	38.90	22.11	10.46	36.88	8.63	3.01
✔	✔	✔	**60.67**	**47.77**	**28.01**	**57.03**	**33.69**	**11.86**	**78.25**	**57.10**	**10.17**

Table 6 Comparison of our TRM model and baseline with dense vs. sparse proposals on the Charades-STA dataset. We test the inference time and GPU memory for each video

Proposals	Method	Sentence Prediction			Phrase Prediction			Time (ms)	GPU Memory (GB)	
		R@0.3	R@0.5	R@0.7	R@0.3	R@0.5	R@0.7		Proposals	Others
Dense	MMN Wang et al. (2022)	60.48	47.45	27.15	38.41	22.19	10.1	31.17	0.0005	1.04
	TRM (Ours)	**60.67**	**47.77**	**28.01**	**57.03**	**33.69**	**11.86**	31.25	0.0005	1.04
Sparse	MMN Wang et al. (2022)	57.14	45.21	24.13	36.14	18.93	7.23	30.41	0.0001	1.04
	TRM (Ours)	**58.03**	**45.96**	**25.87**	**56.14**	**31.17**	**10.21**	30.75	0.0001	1.04

Table 7 Ablation studies on the effectiveness of focal loss in our TRM model

Method	sentence prediction			verb phrase prediction			noun phrase prediction		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
w/o focal loss	**60.81**	45.54	25.54	56.28	29.79	11.13	71.59	52.88	**11.67**
w/ focal loss	60.67	**47.77**	**28.01**	**57.03**	**33.69**	**11.86**	**78.25**	**57.10**	10.17

tion when encountering new combinations of old phrases. 2) Moreover, the performance gain of TRM-PT over TRM on the Novel-Composition split is particularly noteworthy. This demonstrates that improving the model's foundational understanding of individual phrases, a direct result of our phrase-level pseudo-labeling and state-change modeling, is crucial for enhancing its ability to generalize to new combinations of those phrases. 3) On the Charades-CG dataset (Table 4), DeCo and PTAN report results using I3D features fine-tuned on the Charades dataset using verb annotations, which provides an unfair comparison with our methods and those of others in the table, as other methods can not access these verb annotations. For a fair and direct comparison, we reproduce the results of PTAN using the official code with the same visual features (methods marked by *). As indicated in the table, our TRM and TRM-PT models achieve better performance than DeCo and PTAN in the novel-composition split. These empirical results prove that learning phrase-level predictions and the temporal relationships between phrases and sentences help the model generalize to both novel words and novel compositions of seen concepts.

The Test-Trivial and Novel-Word splits evaluate different capabilities. The Test-Trivial split shares the same distribution as the training set, while the Novel-Word split contains words not seen during training, thus testing open-vocabulary understanding. 1) As we can see, our TRM and TRM-PT achieve the best performance on most of the metrics in the test-trivial and novel-word splits, demonstrating the effectiveness of our methods. 2) When comparing our TRM-PT with our TRM, the TRM-PT demonstrates less improvement on the two splits. This is because the Test-Trivial and Novel-Word splits evaluate different capabilities, which are not the primary targets of our TRM-PT extension. We test the variance of the TRM-PT performance, where the variances for IoU=0.5, IoU=0.7, and mIoU are 0.37, 0.41, and 0.52, respectively on the Test-Trivial splits. As shown in Table 3, the performance fluctuations of TRM-PT compared to TRM are within the range of variance. Therefore, we consider this reasonable.

6 Ablation Studies

We conduct ablative experiments on the Charades-STA dataset to analyze the effectiveness of our model design.

Table 8 Ablation studies on the phrase extraction of our TRM model on ActivityNet Captions dataset

Method	Sentence prediction				Phrase prediction			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
SRL	**66.59**	50.03	**31.52**	47.99	**52.33**	**44.91**	**33.12**	**43.29**
Sub-sentence	66.37	**50.57**	31.02	**48.01**	52.26	42.96	32.87	42.58

Table 9 Ablation studies on different methods to aggregate phrase score map on our TRM model

Method	sentence prediction			verb phrase prediction			noun phrase prediction		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
Average	60.27	47.19	26.95	56.27	**33.84**	**12.06**	77.23	**57.48**	9.71
Weighted sum	**60.67**	**47.77**	**28.01**	**57.03**	33.69	11.86	**78.25**	57.10	**10.17**

Table 10 Ablation studies on different features of our TRM-PT model on Charades-STA dataset

Method	Sentence prediction				Phrase prediction			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
VGG	60.67	47.77	28.01	42.77	57.03	33.69	11.86	35.82
CLIP	62.39	48.49	28.66	43.67	59.01	34.89	12.75	37.84
I3D	**67.31**	**55.73**	**33.33**	**46.42**	**60.97**	**36.72**	**16.67**	**39.57**

Table 11 Ablation studies on different features of our TRM-PT model on ActivityNet Captions dataset

Method	Sentence prediction				Phrase prediction			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
CLIP	61.63	45.18	26.53	44.24	50.87	41.97	33.29	42.27
C3D	**66.41**	**50.44**	**31.18**	**47.68**	**52.46**	**42.84**	**33.68**	**43.29**

Table 12 Ablation study on the threshold θ for Area Segmentation on the Charades-STA dataset

Threshold θ	Sentence prediction				Phrase prediction			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
0.0	60.54	47.61	27.86	42.65	57.01	33.14	11.64	35.46
0.1	**60.67**	**47.77**	**28.01**	**42.77**	**57.03**	**33.69**	**11.86**	**35.82**
0.3	60.11	47.54	27.98	42.47	56.87	32.87	11.54	35.12
0.5	60.17	47.13	27.51	42.09	55.74	31.47	10.89	34.73

Table 13 Ablation studies on the phrase-level training and noise estimator of our TRM-PT model on the Charades-STA dataset

Phrase Loss	Noise Estimator	State Loss	Sentence prediction				Phrase prediction			
			IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
✗	✗	✗	60.67	47.77	28.01	42.77	57.03	33.69	11.86	35.82
✔	✗	✗	60.28	46.93	27.72	42.58	57.38	33.24	11.38	35.57
✔	✔	✗	60.07	47.43	27.84	42.69	57.93	34.12	12.69	36.21
✔	✔	✔	**61.57**	**48.13**	**28.97**	**42.81**	**58.21**	**34.65**	**12.85**	**36.75**

Table 14 Design choices of state queries of our TRM-PT model on the Charades-STA dataset

Method		sentence prediction				phrase prediction			
		IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
State description from	sentence	59.76	47.61	28.23	41.89	57.13	33.24	12.01	35.17
	verb	**61.57**	**48.13**	**28.97**	**42.81**	**58.21**	**34.65**	**12.85**	**36.75**
State ground-truth from	phrase pseudo-label	59.74	46.98	27.43	41.13	56.37	32.12	11.76	35.09
	VLM	60.83	47.11	27.94	41.74	57.71	33.11	12.01	35.14
	sentence annotation	**61.57**	**48.13**	**28.97**	**42.81**	**58.21**	**34.65**	**12.85**	**36.75**
Inference	w/ state description	61.37	**48.24**	28.69	42.73	**58.33**	34.41	**12.87**	36.58
	w/o state description	**61.57**	48.13	**28.97**	**42.81**	58.21	**34.65**	12.85	**36.75**

Table 15 Performance comparison on rare and common phrases on the Charades-STA dataset. "Rare" refers to phrases with low frequency in the training set

Method		Rare phrase prediction			Common phrase prediction		
		IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
2D-TAN Zhang et al. (2020)	AAAI 20	41.53	20.84	8.64	45.39	23.87	10.71
MMN Wang et al. (2022)	AAAI 22	35.89	20.14	8.87	39.01	22.54	10.31
PTAN Wei et al. (2024)	ICMR 24	45.11	24.67	10.47	47.51	26.89	12.53
TRM (ours)	AAAI 23	55.56	31.37	9.46	57.98	33.90	12.07
TRM-PT (ours)		**57.21**	**33.87**	**11.59**	**58.34**	**34.73**	**12.96**

Table 16 Performance comparison on rare and common phrases on the ActivityNet Captions dataset. "Rare" refers to phrases with low frequency in the training set

Method		Rare phrase prediction			Common phrase prediction		
		IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
2D-TAN Zhang et al. (2020)	AAAI 20	49.21	39.87	30.54	51.98	42.67	32.78
MMN Wang et al. (2022)	AAAI 22	49.47	40.87	31.04	52.11	42.39	33.12
PTAN Wei et al. (2024)	ICMR 24	47.83	38.96	31.14	50.79	42.07	34.14
TRM (ours)	AAAI 23	49.97	40.89	31.04	52.83	42.91	33.74
TRM-PT (ours)		**52.41**	**42.84**	**33.09**	**53.84**	**44.43**	**34.37**

6.1 Ablations on Sentence-Level Training

In this section, we evaluate the effectiveness of our design in the sentence-level training. All these experiments do not involve the phrase-level training (TRM-PT), but are designed to validate the necessity of our proposed modules within the sentence-level training scheme.

Effectiveness of temporal relationship mining. As shown in Table 5, comparing the first and second rows, we find that simply introducing fine-grained phrase features without considering the relationship between phrase and sentence-level predictions has limited performance improvement for phrase prediction. From the third row, we see that consistency loss can greatly improve the performance of phrase prediction. From the fourth row, it can be seen that training with only exclusiveness loss has a negative impact on the model. This is because only the exclusivity loss is incomplete because the all-zero scores map of phrases is a set of trivial solutions. From the fifth row, we can see that the consistency loss and exclusiveness loss together can further improve the performance of both sentences and phrases. The results show that exploiting the consistency and exclusiveness constraints of phrase-level predictions and sentence-level predictions can regularize the training process, thus alleviating the ambiguity of each phrase localization.

Ablation on Proposal Density and Scalability. Our framework is agnostic to the specific method for generating proposals. To investigate the trade-offs, we evaluate two strategies: a dense approach using sliding windows and a sparse approach employing a hierarchical segment tree. We assess their impact on model performance, inference speed, and GPU memory usage. (1) As shown in Table 6, on the

Table 17 Performance on the Charades-STA dataset with label noise

Method	sentence prediction				phrase prediction			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
MMN Wang et al. (2022)	54.23	43.01	23.10	37.43	36.15	19.12	9.12	28.96
TRM (ours)	54.69	43.93	23.93	38.02	44.57	22.11	10.71	31.24
TRM-PT (ours)	**55.39**	**44.70**	**24.70**	**38.74**	**45.01**	**22.47**	**10.91**	**31.54**

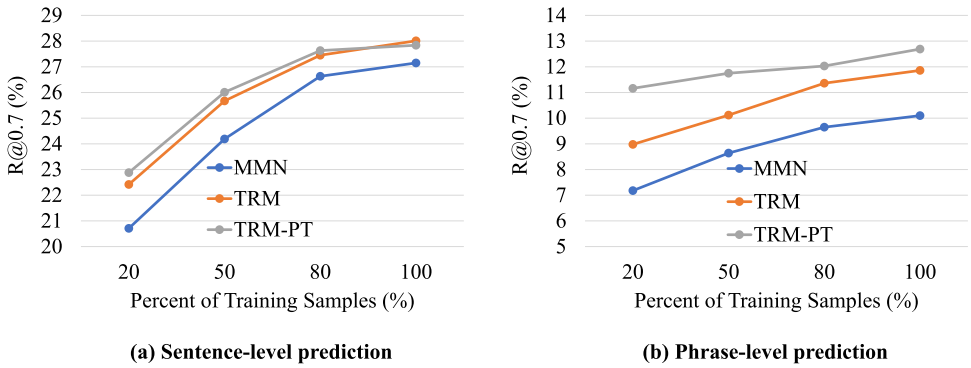


Fig. 8 Performance on the Charades-STA dataset with limited training data

Charades-STA dataset, the dense proposal method significantly outperforms the sparse one. Given the small number of video clips in this dataset, the computational overhead of the dense method is manageable, making its superior performance the deciding factor. Consequently, we adopt dense proposals for our main experiments. (2) Our experiments in Fig. 7 show that as the number of input clips increases, the resource consumption of the sparse proposal method scales near-linearly, while the dense method scales quadratically. To illustrate the real-world implications for long videos, consider a benchmark like the MAD (Soldan et al., 2022) dataset, where videos average 110 minutes. Sampling one clip per second would yield $N = 6600$. Projecting from our analysis in Fig. 7, the GPU memory usage with a sparse approach would be less than 12GB, which is feasible on modern hardware. This confirms our method's applicability to hour-long videos. (3) The number of phrases (N_p) introduces a negligible computational overhead. In practice, the number of phrases extracted per sentence is very small. Specifically, we split phrases based on subject-verb-object structure. Charades mostly consist of simple sentences, resulting in an average of three phrases (subject, verb and object). ActivityNet may contain sentences with multiple verbs and objects, leading to an average of 5 phrases. As shown in Table 6, when comparing our full TRM model against the MMN baseline (which does not use phrases), the increase in overhead is negligible on the Charades-STA dataset. This demonstrates that the processing of phrases has a very small impact on the overall resource requirements.

Effectiveness of focal loss. We have used the focal loss (Lin et al., 2017) in our consistency loss $\mathcal{L}_{con}$ and exclusiveness loss $\mathcal{L}_{ex}$ to balance the positive and negative samples during training. Table 7 shows the effectiveness of the focal loss. In the first row, we use the BCE loss in the consistency and exclusiveness loss instead of focal loss. As we can see, focal loss improves the performance of both sentence prediction and phrase prediction.

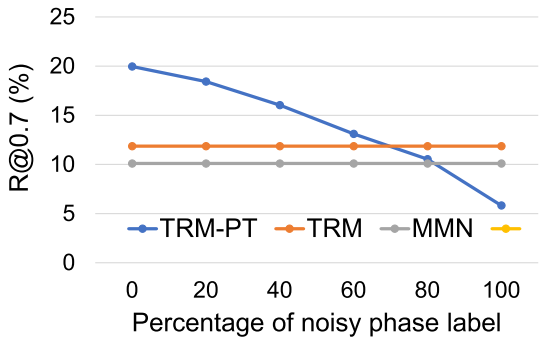


Fig. 9 Ablation study on the robustness of our TRM-PT model to the phrase-level label noise on the Charades-STA dataset. We replace pseudo-labels with ground-truth phrase annotations and introduce controlled noise

We also find that focal loss can significantly improve the performance of noun phrases. This may be because the distribution of nouns in the training set is more imbalanced compared to the distribution of verbs, and in such cases, focal loss can yield greater benefits.

Ablation of phrase extraction. The query sentence on ActivityNet Captions is more complex and usually contains multiple verbs. So we also tried to divide a long sentence into multiple sub-sentences based on verbs as phrases for our training (Table 18 provides some examples). As shown

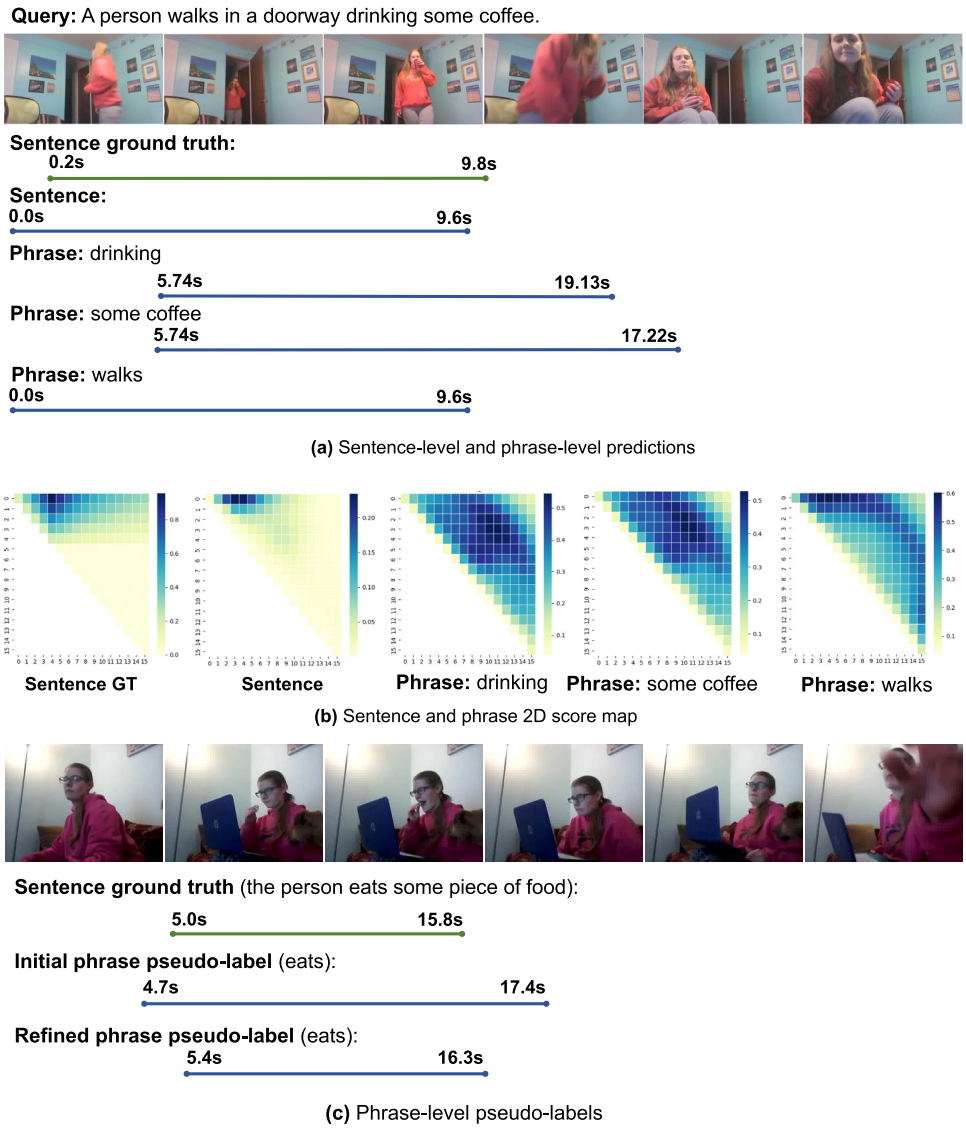


Fig. 10 Qualitative results of our TRM-PT model on Charades-STA

in Table 8, 'SRL' means we extract semantic role labels of verbs and proto-patients as our phrases, which is the same as what we do on the Charades-STA dataset. 'Sub-sentence' means we extract sub-sentence as phrases. The two phrase extraction methods have similar performance on sentence-level prediction. Still, the more fine-grained extraction of phrases according to the semantic rule label performs better on phrase-level prediction. Thus, we use 'SRL' in all experiments.

Ablation of phrase weight. To verify the effectiveness of phrase weights, we use average pooling to aggregate the phrase score map and the results are shown in Table 9. As we can see, predicting the weight of each phrase as its importance and aggregating the phrase score map achieves better performance. This demonstrates assigning different weights to different phrases can improve the performance of sentence-level prediction slightly. This is because different phrases have varying levels of importance within a sentence, so the score maps of different phrases should also influence the sentence-level score map to different extents.

Ablation of different features. As shown in Table 10 and Table 11, we use the VGG (Simonyan & Zisserman, 2015), C3D (Tran et al., 2015), I3D (Carreira & Zisserman, 2017), CLIP (Radford et al., 2021) to extract the visual features. As we can see, for both datasets, we find that visual features have a greater impact on performance. We find that the I3D and C3D feature, which is more sensitive to action, achieves superior performance, which indicates that the action information is important. On the Charades-STA dataset, the CLIP features perform better than VGG features, but not as well as I3D features. This suggests that the CLIP model, which is trained with large-scale image-text pairs, has better generalization performance than VGG. But in video tasks, it is not

Table 18 Some examples of our extracted phrases during training

Dataset	Sentence	Phrases
Charades-STA	Person drinking a glass of water.	"drinking", "a glass of water"
	A person is putting a book on a shelf.	"putting", "a book"
	Person picks up a plate holding a sandwich.	"pick", "a sandwich", "holding", "a plate"
ActivityNet	She pours various liquids into a mixer and shakes the mixture all together.	"pours", "various liquids", "shakes", "the mixture all"
	We see the holographic man on the wall with the paper.	"see", "the footprints on the platform"
	Both women stand talking to the camera while presenting the braid just made.	"stand", "talking", "presenting", "the braid just made", "made", "the braid"
ActivityNet (sub-sentence)	She pours various liquids into a mixer and shakes the mixture all together.	"She pours various liquids into a mixer", "She shakes the mixture all together"
	We see the holographic man on the wall with the paper.	"We see the holographic man on the wall with the paper"
	Both women stand talking to the camera while presenting the braid just made.	"Both women stand", "Both women talking to the camera", "Both women presenting the braid just made", "the braid just made"

as good as the I3D features that have been pre-trained with video action detection tasks.

Ablations of hyper-parameters. We conduct an ablation study to analyze the impact of the threshold θ, which is crucial for dividing proposals into Area I and Area II for our consistency and exclusiveness losses. As shown in Table 12, setting θ=0.1 achieves the best overall performance across both sentence and phrase-level predictions. A smaller threshold (e.g., 0.0) is too lenient and may incorrectly classify negative proposals as positive, while a larger threshold (e.g., 0.5) is too strict and may discard proposals that have a meaningful overlap with the ground truth. Therefore, we set θ=0.1 for all experiments.

6.2 Ablations on Phrase-Level Training

In this section, we evaluate the effectiveness of our design in the phrase-level training. All the experiments are conducted on our TRM model with both sentence-level and phrase-level training.

Effectiveness of phrase-level training. Table 13 shows the effectiveness of the three designs in our phrase-level training: the phrase loss, the noise estimator, and the state loss. As we can see, when only introducing the phrase-level training while not considering the noise in the pseudo-labels, both the sentence-level and phrase-level performance are dropped. This is because the label noise in the pseudo-labels negatively impacts the model performance. When the noise estimator is introduced, the performance of both sentence prediction and phrase prediction is improved. When further introducing the state loss, both the performance of sentence-level prediction and phrase-level prediction are improved, demonstrating the effectiveness of our proposed phrase-level training.

Design choices of state queries. We conduct ablation experiments on the design of state queries. (1) As shown in Table 14, we first use the LLM to infer the state descriptions before and after the entire query sentence, instead of individual verbs. We find that the model's performance declines, which may be due to the complexity of the full sentences, leading to decreased reliability in the LLM's predictions of the state descriptions before and after the sentence. (2) We also use pseudo-labels corresponding to verb phrases to determine the ground truth for the state query. We find that the model's performance declined compared to using sentence annotations to determine the state query's ground truth. This is because errors may still exist in the pseudo-labels corresponding to verb phrases, even with our noise estimator. Since verb phrases are contained within sentence queries, determining the state query's ground truth based on sentence annotations is a better choice. (3) We also use VLM to infer pseudo-labels for state descriptions, rather than relying on sentence annotations. We find that the performance decreased, which is likely due to the potential errors in the pseudo-labels inferred by VLMs. Since actions must occur within the video segments corresponding to the sentence, the video segment before the sentence ground-truth must correspond to the state descriptions before the actions, and vice versa. Therefore, the sentence ground truth can provide more accurate state query labels. (4) During inference, we experiment with whether to use the state description as an additional input, provided to the model along with the query sentence. The experimental results show that the performance of both approaches was similar. This may be because using the state

description as training data provided a stronger supervisory signal, constraining the semantics of the segments before and after the target segment, which made the visual features of the target segment more discriminative. Therefore, even without providing the state description during inference, the model's performance still improved. However, using the state description incurs additional costs for LLM API calls, so we prefer not to use the state description during inference.

s**Comparison on rare phrases.** To provide a deeper analysis of the performance gain from our TRM-PT, we conducted a new, more fine-grained analysis by splitting the phrases in the test sets of Charades-STA and ActivityNet Captions into 'rare' and 'common' categories based on their frequency of appearance in the training set sentence queries. As shown in Table 15 and 16, on the rare phrases, TRM-PT provides a substantial performance boost over the TRM while on common phrases, the performance improvement is less pronounced. This is because the TRM primarily relies on sentence-level annotations to provide implicit supervision for the constituent phrases. Consequently, its ability to learn representations for a given phrase is heavily influenced by the number of sentence annotations containing that phrase. For rare phrases, this supervision is sparse. Our proposed phrase-level training (TRM-PT) directly addresses this limitation by using external VLMs to generate phrase-level pseudo-labels, effectively supplementing the training data and compensating for the shortcomings of TRM on these rare concepts. Since common phrases constitute the majority of the test data, the significant gains on rare phrases are diluted in the overall average, leading to the modest improvements observed in the main tables (Table 1) and Table 2.

Comparison when using limited training data. By introducing additional phrase-level pseudo-labels through phrase-level training, we find that the model performance with limited training data is improved. As shown in Fig. 8, we train the model using a subset of training data from the Charades-STA dataset and measure the model's sentence-level and phrase-level performance under different training data quantities in Fig. 8(a) and Fig. 8(b) respectively. It can be observed that our method outperforms the baseline across various data quantities. On the other hand, as the data quantity gradually decreases, the advantages of TRM-PT become increasingly evident. This suggests that the introduction of phrase-level pseudo-labels to some extent augments the training data, enabling better adaptation to scenarios with limited data.

Comparison when using noised annotation. We also find that introducing phrase-level training to some extent helps improve the model's robustness under noisy sentence annotation conditions. As shown in Table 17, we artificially introduced random offsets to the start and end times of annotated video segments in the Charades-STA dataset to create scenarios with noisy labels and evaluate the model's performance. It can be observed that under this setup, both our sentence-level and phrase-level localization outperform the baseline. We also observed that the performance of the phrase-level training method with pseudo-labels is better. This is because the generation of phrase-level pseudo-labels is less affected by noise in sentence-level annotations and can influence sentence-level localization through consistency and exclusive constraints, making it more robust.

Analysis on Pseudo-Labels Noise. (1) To analyze the noise on the pseudo-labels, we ask annotators to check the quality of pseudo-labels. We randomly selected 50 videos with 120 queries from the Charades-STA dataset and found that 51.6% of the pseudo-labels were accurate. As shown in Table 1, even with the presence of noise in these pseudo-labels, the TRM-PT model trained with them surpasses previous methods and achieves state-of-the-art performance on phrase-level prediction. This demonstrates the effectiveness of our TRM-PT framework in training models using phrase pseudo-labels. (2) To systematically evaluate the impact of pseudo-label noise on model performance, we have conducted a new controlled experiment on the Charades-STA dataset. Instead of using pseudo-labels generated by a VLM, we utilized the ground-truth temporal annotations for verb and noun phrases (from action localization and object detection tasks) as an "oracle". We then synthetically introduced noise into these ground-truth phrase labels to simulate scenarios with varying levels of pseudo-label corruption. Specifically, we randomly selected a certain percentage of the phrase annotations and perturbed their start and end times. The offset for each boundary was randomly sampled from a uniform distribution $U(-0.1 \times D, 0.1 \times D)$, where D is the duration of the video. We trained our TRM-PT model on these noisy labels and evaluated its performance on both sentence and phrase localization. The results are shown in Fig. 9. As illustrated, the performance of our model gracefully degrades as the noise ratio in the phrase labels increases. Notably, even with 60% of the phrase labels being noisy, our TRM-PT maintains better performance than TRM and MMN, showcasing its resilience. This experiment validates the effectiveness of our proposed Noise Estimator, which successfully mitigates the negative impact of noisy samples during training. The results confirm that our framework is robust to the inherent noise present in automatically generated pseudo-labels, which is a critical aspect for real-world applications.

7 Qualitative Results

In Fig. 10, we provide a visualization of the predictions and score maps of our model and the baseline (without phrase) on Charades-STA Dataset. In Table 18, we provide some

examples of our extracted phrases during training. As we can see, our prediction for the sentence matches the ground truth (in green) well. Also, our TRM understands that the entire sentence consists of three phrases: the phrase 'drinking', 'some coffee', and 'walks'. We also shows the 2D score map of all the sentence and phrases. All the predictions and scoremaps satisfy our constraints of consistency and exclusiveness. In contrast, while the baseline model also performs well in the sentence-level prediction, the predictions for the phrases 'drinking' and 'walks' are less accurate and violate the constraint of consistency, i.e., there is no overlap with the ground truth of the sentence. We also note that the score map predicted by the baseline model has multiple peaks on phrases, which indicates that the baseline model is not confident in the prediction of phrase. Fig. 10(c) also provides visualizations of some phrase-level pseudo-labels. It can be observed that the initial pseudo-labels generated using pre-trained large-scale models offer approximate locations of the video segments corresponding to phrase queries, but they are still not accurate enough. In contrast, the pseudo-labels obtained after noise estimation and refinement exhibit higher accuracy.

8 Conclusion

In this work, we propose the phrase-level Temporal Relationship Mining (TRM) framework considering both phrase and sentence queries, making the first attempt to mine the phrase-proposal relation in the temporal localization task. We develop a method to constrain phrase-level prediction in training, tackling the lack of phrase-level annotation. We propose the consistency and exclusiveness constraints of phrase-level and sentence-level predictions to regularize the training process, thus alleviating the ambiguity of each phrase prediction. We also propose to use the pre-trained model to generate fine-grained pseudo-labels for phrases and use the noise estimator to mitigate the negative impact of the label noise. Finally, to enhance the understanding of verb phrases in the model, we utilize a large-scale language model to infer changes in the scene's state before and after the occurrence of verb phrases and align them with the visual content. Experimental results on Charades-STA and ActivityNet Captions indicate that our model surpasses other models in phrase-level prediction while sentence-level results remain stable, demonstrating our model's competence, interpretability, and generalization performance.

Funding This work was supported by the grants from the National Natural Science Foundation of China (62372014, 61925201, 62132001, 62432001) and Beijing Natural Science Foundation (4252040, L247006).

Data Availability All datasets used in this work are publicly available. ActivityNet Captions (Krishna et al., 2017) dataset is available at https://cs.stanford.edu/people/ranjaykrishna/densevid. Charades-STA (Gao et al., 2017) dataset is available at https://prior.allenai.org/projects/charades.

Code Availability The code is available at https://github.com/minghangz/trm.

References

Wang, Z., Wang, L., Wu, T., Li, T., & Wu, G. (2022). Negative sample matters: A renaissance of metric learning for temporal grounding. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pp. 2613–2623.

Otani, M., Nakashima, Y., Rahtu, E., & Heikkilä, J. (2020). Uncovering hidden challenges in query-based video moment retrieval. In: 31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020.

Yuan, Y., Lan, X., Wang, X., Chen, L., Wang, Z., & Zhu, W. (2021). A closer look at temporal sentence grounding in videos: Dataset and metric. In: Proceedings of the 2nd International Workshop on Human-centric Multimedia Analysis, pp. 13–21.

Li, J., Xie, J., Qian, L., Zhu, L., Tang, S., Wu, F., Yang, Y., Zhuang, Y., & Wang, X.E. (2022). Compositional temporal grounding with structured variational cross-graph correspondence learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 3022–3031.

Rasheed, H.A., Khattak, M.U., Maaz, M., Khan, S.H., & Khan, F.S. (2023). Fine-tuned CLIP models are efficient video learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 6545–6554.

Zheng, M., Li, S., Chen, Q., Peng, Y., & Liu, Y. (2023). Phrase-level temporal relationship mining for temporal sentence localization. In: Williams, B., Chen, Y., Neville, J. (eds.) Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pp. 3669–3677.

Gao, J., Sun, C., Yang, Z., & Nevatia, R. (2017). TALL: temporal activity localization via language query. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 5277–5285.

Li, H., Cao, M., Cheng, X., Li, Y., Zhu, Z., & Zou, Y. (2023). G2L: semantically aligned and uniform video grounding via geodesic and game theory. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 11998–12008.

Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., & Shou, M.Z. (2023). Univtg: Towards unified video-language temporal grounding. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 2782–2792.

Jang, J., Park, J., Kim, J., Kwon, H., & Sohn, K. (2023). Knowing where to focus: Event-aware transformer for video grounding. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 13800–13810.

Fang, X., Liu, D., Zhou, P., & Nan, G. (2023). You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In: IEEE/CVF Conference on

Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 2448–2460.

Zhang, S., Peng, H., Fu, J., & Luo, J. (2020). Learning 2d temporal adjacent networks for moment localization with natural language. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pp. 12870–12877.

Zhang, H., Sun, A., Jing, W., & Zhou, J.T. (2020). Span-based localizing network for natural language video localization. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proc. of ACL, Online, pp. 6543–6554.

Rodriguez-Opazo, C., Marrese-Taylor, E., Fernando, B., Li, H., & Gould, S. (2021). Dori: Discovering object relationships for moment localization of a natural language query in a video. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1079–1088.

Liu, D., Qu, X., Zhou, P., & Liu, Y. (2022). Exploring motion and appearance information for temporal sentence grounding. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pp. 1674–1682.

Mun, J., Cho, M., & Han, B. (2020). Local-global video-text interactions for temporal grounding. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 10807–10816.

Liu, D., Qu, X., Di, X., Cheng, Y., Xu, Z., & Zhou, P. (2022). Memory-guided semantic learning network for temporal sentence grounding. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pp. 1665–1673.

Huang, J., Jin, H., Gong, S., & Liu, Y. (2022). Video activity localisation with uncertainties in temporal boundary. In: European Conference on Computer Vision, pp. 724–740. Springer.

Yang, L., Kong, Q., Yang, H., Kehl, W., Sato, Y., & Kobori, N. (2023). Deco: Decomposition and reconstruction for compositional temporal grounding via coarse-to-fine contrastive ranking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 23130–23140.

Song, X., Jiao, L. C., Yang, S., Zhang, X., & Shang, F. (2013). Sparse coding and classifier ensemble based multi-instance learning for image categorization. *Signal Processing,93*(1), 1–11.

Xu, H., Venugopalan, S., Ramanishka, V., Rohrbach, M., & Saenko, K. (2015). A multi-scale multiple instance video description network. ArXiv preprint arXiv:abs/1505.05914

Xu, Y., Zhu, J.-Y., Chang, E.I.-C., Lai, M., & Tu, Z. (2014). Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis,18*(3), 591–604.

Huang, J., Liu, Y., Gong, S., & Jin, H. (2021). Cross-sentence temporal and semantic relations in video activity localisation. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pp. 7179–7188.

Yang, W., Zhang, T., Zhang, Y., & Wu, F. (2021). Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing,30*, 3252–3262.

Zheng, M., Huang, Y., Chen, Q., & Liu, Y. (2022). Weakly supervised video moment localization with contrastive negative sample mining. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pp. 3517–3525.

Zheng, M., Huang, Y., Chen, Q., Peng, Y., & Liu, Y. (2022). Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 15534–15543.

Huang, Y., Yang, L., & Sato, Y. (2023). Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 18908–18918.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proc. of ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763.

Li, J., Li, D., Xiong, C., & Hoi, S.C.H. (2022). BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) Proc. of ICML. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900.

Zeng, Y., Zhang, X., & Li, H. (2022). Multi-grained vision language pre-training: Aligning texts with visual concepts. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) Proc. of ICML. Proceedings of Machine Learning Research, vol. 162, pp. 25994–26009.

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). Videobert: A joint model for video and language representation learning. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 7463–7472.

Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., & Liu, J. (2021). Less is more: Clipbert for video-and-language learning via sparse sampling. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, pp. 7331–7341.

Xu, H., Ghosh, G., Huang, P.-Y., Arora, P., Aminzadeh, M., Feichtenhofer, C., Metze, F., & Zettlemoyer, L. (2021). VLM: Task-agnostic video-language model pre-training for video understanding. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, pp. 4227–4239.

Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., & Ji, R. (2022). X-CLIP:: End-to-end multi-grained contrastive learning for video-text retrieval. ArXiv preprint arXiv:abs/2207.07285

Weng, Z., Yang, X., Li, A., Wu, Z., & Jiang, Y. (2023). Open-vclip: Transforming CLIP to an open-vocabulary video model via interpolated weight optimization. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proc. of ICML. Proceedings of Machine Learning Research, vol. 202, pp. 36978–36989.

Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., & Rohrbach, A. (2022). ReCLIP: A strong zero-shot baseline for referring expression comprehension. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proc. of ACL, Dublin, Ireland, pp. 5198–5215.

Liu, Y., Zhang, J., Chen, Q., & Peng, Y. (2023). Confidence-aware pseudo-label learning for weakly supervised visual grounding. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 2816–2826.

Luo, D., Huang, J., Gong, S., Jin, H., & Liu, Y. (2023). Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 23045–23055.

Zheng, M., Gong, S., Jin, H., Peng, Y., & Liu, Y. (2023). Generating structured pseudo labels for noise-resistant zero-shot video sentence localization. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proc. of ACL, Toronto, Canada, pp. 14197–14209.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2016). Grounding of textual phrases in images by reconstruction. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 817–834. Springer.

Ryu, H., Kang, S., Kang, H., & Yoo, C.D. (2021). Semantic grouping network for video captioning. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 2514–2522.

Zhang, J., & Peng, Y. (2019). Hierarchical vision-language alignment for video captioning. In: MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25, pp. 42–54. Springer.

Li, S., Li, C., Zheng, M., & Liu, Y. (2022). Phrase-level prediction for video temporal localization. In: International Conference on Multimedia Retrieval (ICMR), pp. 360–368.

Mu, F., Mo, S., & Li, Y. (2024). Snag: Scalable and accurate video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18930–18940.

Pan, Y., He, X., Gong, B., Lv, Y., Shen, Y., Peng, Y., & Zhao, D. (2023). Scanning Only Once: An End-to-end Framework for Fast Temporal Grounding in Long Videos . In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13721–13731. IEEE Computer Society, Los Alamitos, CA, USA.

Shi, P., & Lin, J.J. (2019). Simple bert models for relation extraction and semantic role labeling. ArXiv preprint arXiv:abs/1904.05255

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv preprint arXiv:abs/1910.01108

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In: Guyon, I., Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Proc. of NeurIPS, pp. 5998–6008.

Lin, T., Goyal, P., Girshick, R.B., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 2999–3007.

Oord, A.v.d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. ArXiv preprint arXiv:abs/1807.03748

Chen, S., & Jiang, Y.-G. (2019). Semantic proposal for activity localization in videos via sentence query. *Proceedings of the AAAI Conference on Artificial Intelligence,33*(01), 8199–8206.

Zhang, D., Dai, X., Wang, X., Wang, Y., & Davis, L.S. (2019). MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 1247–1257.

Gao, J., & Xu, C. (2021). Fast video moment retrieval. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pp. 1503–1512.

Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., & Gan, C. (2020). Dense regression network for video grounding. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 10284–10293.

Ding, X., Wang, N., Zhang, S., Cheng, D., Li, X., Huang, Z., Tang, M., & Gao, X. (2021). Support-set based cross-supervision for video grounding. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pp. 11553–11562.

Liu, D., Qu, X., Dong, J., Zhou, P., Cheng, Y., Wei, W., Xu, Z., & Xie, Y. (2021). Context-aware biaffine localizing network for temporal sentence grounding. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, pp. 11235–11244.

Zhao, Y., Zhao, Z., Zhang, Z., & Lin, Z. (2021). Cascaded prediction network via segment tree for temporal video grounding. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, pp. 4197–4206.

Wei, Z., Jiang, X., Wang, Z., Shen, F., & Xu, X. (2024). Ptan: Principal token-aware adjacent network for compositional temporal grounding. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. ICMR '24, pp. 618–627. Association for Computing Machinery, New York, NY, USA.

Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., & Xiao, J. (2021). Boundary proposal network for two-stage natural language video localization. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 2986–2994.

Zhou, H., Zhang, C., Luo, Y., Chen, Y., & Hu, C. (2021). Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, pp. 8445–8454.

Zhang, H., Sun, A., Jing, W., Zhen, L., Zhou, J.T., & Goh, S.M.R. (2021). Parallel attention network with sequence matching for video grounding. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, pp. 776–790.

Qu, X., Tang, P., Zou, Z., Cheng, Y., Dong, J., Zhou, P., & Xu, Z. (2020). Fine-grained iterative attention network for temporal language localization in videos. In: MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020, pp. 4280–4288.

Wang, H., Zha, Z., Li, L., Liu, D., & Luo, J. (2021). Structured multi-level interaction network for video moment localization via language query. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, pp. 7026–7035.

Seol, M., Kim, J., & Moon, J. (2023). Bmrn: Boundary matching and refinement network for temporal moment localization with natural language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5570–5578.

Jing, W., Sun, A., Zhang, H., & Li, X. (2023). MS-DETR: Natural language video localization with sampling moment-moment interaction. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proc. of ACL, Toronto, Canada, pp. 1387–1400.

Zhang, Z., Han, X., Song, X., Yan, Y., & Nie, L. (2021). Multi-modal interaction graph convolutional network for temporal language localization in videos. *IEEE Transactions on Image Processing,30*, 8265–8277.

Gao, J., Sun, X., Xu, M., Zhou, X., & Ghanem, B. (2021). Relation-aware video reading comprehension for temporal language grounding. In: Moens, M.-F., Huang, X., Specia, L., Yih, S.W.-

t. (eds.) Proc. of EMNLP, Online and Punta Cana, Dominican Republic, pp. 3978–3988.
Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., & Huang, J. (2018). Weakly supervised dense event captioning in videos. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Proc. of NeurIPS, pp. 3063–3073.
Wu, J., Li, G., Liu, S., & Lin, L. (2020). Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pp. 12386–12393.
Yang, L., Kong, Q., Yang, H.-K., Kehl, W., Sato, Y., & Kobori, N. (2023). Deco: Decomposition and reconstruction for compositional temporal grounding via coarse-to-fine contrastive ranking. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 23130–23140. https://doi.org/10.1109/CVPR52729.2023.02215
Liu, B., Yeung, S., Chou, E., Huang, D.-A., Fei-Fei, L., & Niebles, J.C. (2018). Temporal modular networks for retrieving complex compositional activities in videos. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 552–568.
Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A.K. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. ArXiv preprint arXiv:abs/1604.01753
Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J.C. (2017). Dense-captioning events in videos. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 706–715.
Yuan, Y., Liang, X., Wang, X., Yeung, D., & Gupta, A. (2017). Temporal dynamic graph LSTM for action-driven video object detection. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 1819–1828.
Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) Proc. of ICLR.
Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 4489–4497.
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., & Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. ArXiv preprint arXiv:abs/1910.03771
Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In: Proc. of ICLR.
Wang, L., Mittal, G., Sajeev, S., Yu, Y., Hall, M., Boddeti, V.N., & Chen, M. (2023). Protégé: Untrimmed pretraining for video temporal grounding by video temporal grounding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 6575–6585.
Soldan, M., Pardo, A., Alcázar, J.L., Caba, F., Zhao, C., Giancola, S., & Ghanem, B. (2022). Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5026–5035.
Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 4489–4497.
Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 4724–4733.