Temporal Sentence Grounding with Relevance Feedback in Videos

Jianfeng Dong[1]   **Xiaoman Peng**[1]   **Daizong Liu**[2]* **Xiaoye Qu**[3]

Xun Yang[4]   **Cuizhu Bao**[1]* **Meng Wang**[5]

[1]Zhejiang Gongshang University
[2]Peking University
[3]Huazhong University of Science and Technology
[4]University of Science and Technology of China
[5]Hefei University of Technology

Abstract

As a widely explored multi-modal task, Temporal Sentence Grounding in videos (TSG) endeavors to retrieve a specific video segment matched with a given query text from a video. The traditional paradigm for TSG generally assumes that relevant segments always exist within a given video. However, this assumption is restrictive and unrealistic in real-world applications where the existence of a query-related segment is uncertain, easily resulting in erroneous grounding. Motivated by the research gap and practical application, this paper introduces a new task, named Temporal Sentence Grounding with Relevance Feedback (TSG-RF) in videos, which accommodates the possibility that a video may or may not include a segment related to the query. This task entails localizing precise video segments that semantically align with the query text when such content is present, while delivering definitive feedback on the non-existence of related segments when absent. Moreover, we propose a novel Relation-aware Temporal Sentence Grounding (RaTSG) network for addressing this challenging task. This network first reformulates the TSG-RF task as a foreground-background detection problem by investigating whether the query-related semantics exist in both frame and video levels. Then, a multi-granularity relevance discriminator is exploited to produce precise video-query relevance feedback and a relation-aware segment grounding module is employed to selectively conduct the grounding process, dynamically adapting to the presence or absence of query-related segments in videos. To validate our RaTSG network, we reconstruct two popular TSG datasets, establishing a rigorous benchmark for TSG-RF. Experimental results demonstrate the effectiveness of our proposed RaTSG for the TSG-RF task. Our source code is available at https://github.com/HuiGuanLab/RaTSG.

1   Introduction

Grounding target content described by users in videos is a fundamental capability that facilitates various multimedia applications, such as intelligent robotic service [1], video on demand [2], and metaverse [3]. Following this demand, temporal sentence grounding in videos (TSG) [4] has recently become a research hotspot, attracting wide attention from researchers in various fields [5–9]. Despite the significant advancements made in this field, the existing TSG task is overly idealistic. It aims to identify segments that are semantically relevant to a given query from a given long video, assuming that relevant segments always exist in the given video [10–12]. However, this setting limits the applicability of TSG in real-world scenarios as the given video may not contain the query-related

---

*Corresponding authors.
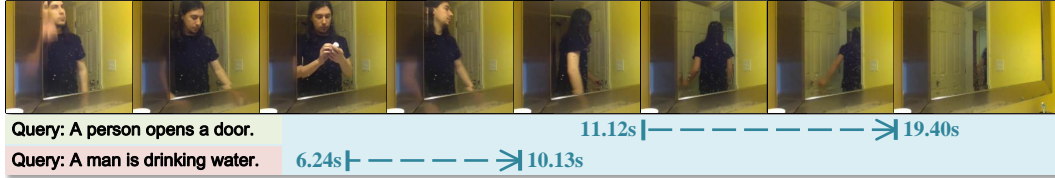
(a) Temporal Sentence Grounding (TSG) in videos

(b) Temporal Sentence Grounding with Relevance Feedback (TSG-RF) in videos

Figure 1: The difference between the TSG and TSG-RF tasks. TSG *always* predicts the start and end boundaries of the grounded segments, even in the absence of video content relevant to the given query text. By contrast, TSG-RF provides relevance feedback on whether exist query-related content in the given video, and *selectively* predicts the start and end boundaries of the grounded segments, according to the presence or absence of query-related segments in videos.

contents, resulting mismatched or wrong grounding results. As illustrated in Figure 1(a), given a video and a query, existing TSG methods always predict the start and end times of segments, even when the video lacks content relevant to the given query text. This suggests a significant gap between the literature and the real world.

To fill the gap, this paper makes the first attempt to introduce an expanded but challenging task for TSG, named Temporal Sentence Grounding with Relevance Feedback (TSG-RF) in videos. This new task aims to deliver more flexible grounding results for a given query and a video. Specifically, as illustrated in Figure 1(b), if relevant segments are present in the video, the task is to pinpoint the video segment that is semantically relevant to the given query description. Conversely, if no relevant segments exist, it should be explicitly conveyed to the user that "there are no target segments related to the query in the given video, and thus no corresponding segments can be localized."

To tackle this new task, it is necessary to first review the failure reason of traditional TSG methods [13–16]. We argue the essential issue of previous TSG methods is their inability to discern whether grounding results are warranted based on the presence or absence of relevant segments. To alleviate this issue, a straightforward approach to address the TSG-RF task involves a two-stage process: Firstly, a relevance discriminator is trained separately to assess the relevance between the query text and the video, providing relevant feedback. Then, query-video samples identified as relevant by the discriminator are processed using models designed for the TSG task to predict the target segment. Despite the structural simplicity of this integrated approach, it requires training two distinct models, leading to the redundant consumption of computational resources, increased storage demands, and suboptimal overall inference efficiency. Therefore, how to design an end-to-end relevance-aware TSG-RF method is an emerging problem.

Based on the above considerations, we propose to address the TSG-RF task by developing a multi-task learning model, *i.e*, the Relation-aware Temporal Sentence Grounding (RaTSG) network. This model promotes mutual enhancement between relevance discrimination and video segment grounding through shared knowledge, thus facilitating a unified and efficient approach to both tasks. As the essence of relevance discrimination lies in accurately measuring the semantic partial relevance between the query text and the video, we propose to first explore the measurement of partial semantic relevance between text and video at two different granularities: fine-grained and coarse-grained. Specifically, to capture fine-grained partial relevance, we adopt the concept of multiple instance learning [17–19], treating the entire video as a bag with each frame as an instance. Following [20], a video is considered partially related to the text if it contains at least one foreground frame; if all frames are background and irrelevant, the video is deemed unrelated to the text. Considering the semantic incompleteness of individual video frames, a coarse-grained perspective is also adopted to learn the relevance between the global video-level semantics and the full query text. Then, a multi-granularity relevance discriminator is designed to integrate the discriminative capabilities of both granularities to

predict the final relevance score. As for the selective grounding process based on the relevance score, existing segment grounding heads always make grounding for any query-video input, even when no relevant segments are present in the video, which is not suitable for the TSG-RF task. Therefore, we design a relation-aware segment grounding module that takes the previously predicted relevance information of each query-video pair as prior knowledge, and selectively conducts grounding.

We find a recent concurrent work [21] that shares a similar motivation with us. However, it simply treats the irrelevant video contents from an open-set perspective and implicitly utilizes cross-dataset evaluation. Instead, we implement original in-domain data to explicitly learn the potential irrelevant relationship. Further, existing TSG datasets can not be directly used to evaluate our proposed network for TSG-RF, since the datasets assume that a given query text can always localize a corresponding relevant segment from the given video. Hence, we further reconstruct two commonly used datasets, Charades-STA [22] and ActivityNet Captions [23], by including samples without grounding results where the video does not contain the content relevant to the query.

The main contributions of this paper are summarized as follows:

- We formalize a novel TSG task to explore more flexible segment grounding in videos, *i.e.*, Temporal Sentence Grounding with Relevance Feedback (TSG-RF) in videos, which advances user-specified TSG towards more practical applications.

- Targeting the TSG-RF task, we propose a novel Relation-aware Temporal Sentence Grounding (RaTSG) network, which mainly depends on a multi-granularity relevance discriminator and a relation-aware segment grounding module. The multi-granularity relevance discriminator is devised to predict the relevance feedback based on fine-grained and coarse-grained relevance between query text and videos, and the relation-aware segment grounding module selectively predicts the start and end boundaries of the grounded segment.

- To facilitate the evaluation of the TSG-RF task, we reconstruct two commonly used TSG datasets and establish appropriate performance evaluation metrics to meet the setting of TSG-RF. Extensive experiments conducted on these reconstructed datasets demonstrate the effectiveness of the proposed model.

2   Related Work

2.1   Temporal Sentence Grounding and Highlight Detection in Videos

The Temporal Sentence Grounding in Videos (TSG) task aims to retrieve video segments that match user-input natural language descriptions. Recent efforts for TSG can be typically grouped into proposal-based and proposal-free methods. Proposal-based methods [22, 24–39] follow a two-stage paradigm: generating proposals by dividing the video into clips and then aligning textual semantics with visual features. Proposal-free methods [40–44, 10, 13, 45] offer an end-to-end paradigm, predicting target segments without proposals, enhancing computational efficiency. Although efficient, these methods struggled with segment-level feature capture. To combine strengths, Xiao et al. [45] used an anchor-free approach to generate candidates, then matched them with query statements using anchor-based methods. However, the above methods cannot be directly applied to the TSG-RF task proposed in this paper. They will still predict the start and end times of segments and output retrieved segments, even in the absence of video content relevant to the query text.

Highlight detection task is similar to TSG, which mainly aims to identify the most interesting or important segments within a video based on a given natural language query, focusing on segments that are salient or engaging [46–49]. Recent works typically tackle this task using multi-modal inputs, advanced transformers, or large-scale pretraining techniques. For instance, Moment-DETR [46] introduce a transformer-based approach that simplifies highlight detection by treating it as a set prediction task, eliminating traditional proposal steps. QD-DETR [48] enhances video-text understanding with cross-attention and negative pairs to improve relevance predictions. UniVTG [49] unifies diverse temporal annotations and enables large-scale pretraining to improve generalization across video grounding tasks. Notably, similar to TSG, highlight detection also assumes that every video contains highlights. In contrast, our proposed TSG-RF task not only grounds relevant segments but also accounts for cases where no relevant segments are present.

2.2 Cross-Modal Semantic Similarity Learning

This paper extends TSG by incorporating relevance feedback to discern the semantic connection between text and video, introducing techniques for learning cross-modal semantic similarity [50–53]. The prevailing approach aligns text and video semantics by learning a shared space, using a metric to assess similarity. For instance, Miech et al. [50] respectively represent text and video into a feature vector, and compute their similarity via cosine similarity metric. Croitoru et al. [52] use distillation techniques to combine knowledge from multiple pretrained models, enhancing cross-modal similarity learning. Current mainstream methods involve learning multiple shared spaces, with relevance computed as a weighted sum of similarities across these spaces. Li et al. [54] employ a multi-space, multi-loss learning framework, enhancing cross-modal similarity by leveraging complementary encoders. While these techniques assess the whole text-video similarity, TSG-RF requires learning partial relevance, where the text relates to only some frames within the video. For non-retrievable samples, the text is irrelevant to any frame. Therefore, the focus is on learning partial relevance.

3 The Proposed Method

3.1 Problem Definition and Overview

TSG-RF task. Given a video $V = [v_1, v_2, \ldots, v_n]$ and a query text $Q = [q_1, q_2, \ldots, q_m]$ as inputs, the TSG-RF task is asked to predict relevance feedback on whether exist query-related content in the given video. Besides, if the query has relevance with the video, it is asked to conduct grounding to identify the precise boundary indices of video segments that are semantically consistent with the query text in the video V. Otherwise, the grounding results are ignored. Compared to the traditional TSG task that always predicts grounding results for all videos, this TSG-RF task is more challenging as it requires not only distinguishing the videos containing and not containing query-relevant segments but also generating explicit non-existence signals for non-relevant videos besides the grounding results.

Overall framework. Figure 2 illustrates the framework of our proposed RaTSG network for the TSG-RF task. Given an input video V and a query text Q encoded by the attention-based mechanisms, the multi-granularity relevance discriminator first captures fine-grained and coarse-grained semantic-aware relevance between the video and query text at the frame and segment levels. Subsequently, it integrates the relevance determination capabilities of these two granularities to generate the relevance discrimination score P_v for distinguishing whether the video has a semantic-related content of the query. Then, a relation-aware segment grounding module takes the multimodal enhanced video features as the basic representation for generating the grounding probability distribution P_s of the start boundary and P_e of the end boundary, while exploiting relevance feedback signal as additional input to produce the separate final results of irrelevant and relevant segments. In the following sections, we will provide detailed illustration of each component.

3.2 Preprocessing of Text and Video Features

Given an input video and a query text, we first use a pre-trained video feature extractor, such as I3D [55], to extract the video feature sequence $V = [v_1, v_2, \ldots, v_n] \in \mathbb{R}^{n \times d_v}$, where d_v is the feature dimension. Subsequently, the query text features are extracted using a text feature extractor, like GloVe [56], producing word embeddings $Q = [q_1, q_2, \ldots, q_m] \in \mathbb{R}^{m \times d_q}$, where d_q represents the feature dimension. Both the video feature sequence V and the query text embeddings Q are first mapped into the same dimension d through a fully connected layer respectively, resulting in $V \in \mathbb{R}^{n \times d}$ and $Q \in \mathbb{R}^{m \times d}$. To enhance the video semantic with the query contexts, we utilize an attention mechanism to integrate word-wise query features Q with the frame-wise video feature V for generating text-guided enhanced video feature V_q [13]. To comprehend the whole semantic of the query for latter global-level reasoning, we use a self-attention layer to perform the intra-modal information interaction on the text feature sequences Q to obtain the sentence-level query representation h_Q. The final contextual multi-modal representation V_q' can be obtained by concatenating the frame-wise video feature V_q with sentence-level query feature h_Q. More details can be found in Appendix A.1.
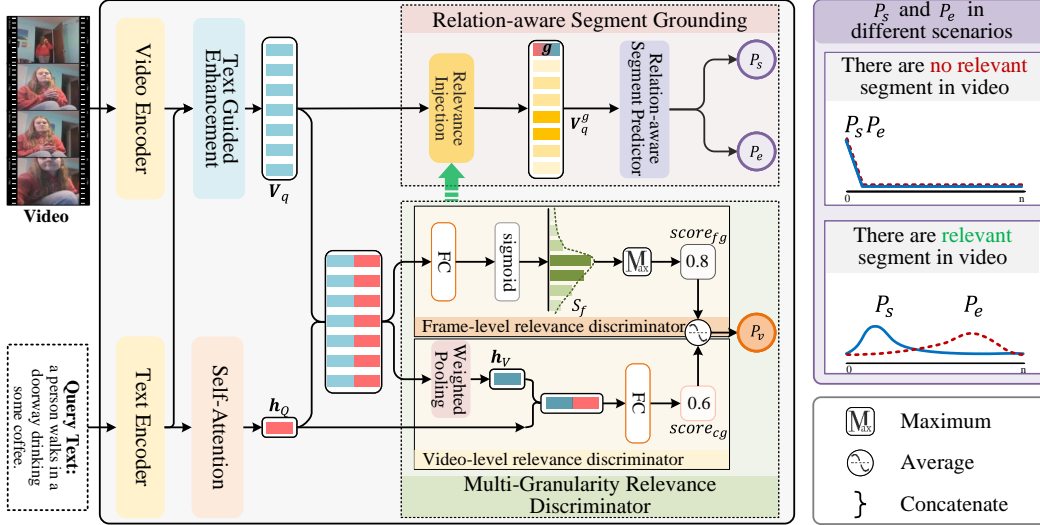
Figure 2: The overall framework of our proposed RaTSG for addressing TSG-RF task. It mainly depends on a multi-granularity relevance discriminator that is employed to learn query-video relevance, and a relation-aware segment grounding module that is used to selectively perform grounding.

3.3 Multi-Granularity Relevance Discriminator

Before generating the grounding results for each query-video pair, we need to first investigate the semantic relevance between them. To determine whether a given video contains segments relevant to a query text, the key lies in assessing the existence of relevance between the text semantic and the video content. Here, we delve into capturing the multi-grained query-video relevance by matching the global sentence-level textual contexts with frame-level and video-level video semantics, respectively.

Frame-level relevance discriminator. Determining frame-level query-video relevance can be treated as a foreground-background frame discrimination problem. Theoretically, for a video containing query-related content, the foreground frames located within the start and end index boundaries should exhibit strong relevance to the query text. Conversely, for a video lacking relevant segments, no foreground frames in the video should relate to the text. To detect foreground frames from the query text, we design a foreground frame detector that finely learns the similarity between the text and each video frame. Specifically, we feed the fused multi-modal feature sequence V_q' into a fully connected layer to predict a sequence of query-to-frame similarity scores. These scores are then mapped to a range between 0 and 1 using the sigmoid function, resulting in the foreground frame prediction score sequence $S_f = \text{sigmoid}(\text{FC}(V_q')) \in \mathbb{R}^n$ where each score indicates the confidence of a frame belonging to the foreground. To train the frame-level relevance discriminator, a binary cross-entropy loss L_{frame} (Appendix A.2) is employed on each frame, which encourages the discriminator to output large scores for foreground frames while small scores for background frames.

To obtain the relevance between text and video at a fine-grained frame-level, we apply the concept of multiple instance learning [17–19]. To be specific, the video is treated as a bag with video frames as instances, categorized into foreground (positive) and background (negative) frames. Thus, a video containing at least one foreground frame is considered a positive bag, indicating partial relevance to the query text. In contrast, a video composed entirely of background frames is deemed a negative bag, showing no relevance. Following [20], we utilize a max-pooling operator to transfer the instance scores to bag scores, obtaining the fine-grained query-video discrimination score as $score_{fg} = \max(S_f)$. The low fine-grained discrimination score suggests that there are no foreground frames in the video, while a high score indicates the presence of at least one foreground frame.

Video-level relevance discriminator. In addition to capturing the frame-level relevance, we also attempt to perceive the semantic relevance from a more informationally complete video level. To generate the comprehensive video-level visual representation, we need to aggregate all query-related frame-wise features into a global one to determine its semantic similarity with the sentence contexts. Specifically, we apply a frame-wise weighted sum operation on the text-guided enhanced video feature V_q to obtain the query-related video representation $h_V \in \mathbb{R}^d$, where we assign weights to

5

each frame by normalizing the previously obtained frame-level similarity scores (reflect the relative relevance score between the text and each frame) using the softmax function. Then, we fuse the sentence-level textual feature h_Q with the global video feature h_v to obtain the video-level relation signal vector g, which is then input into a fully connected layer with a sigmoid function to obtain the final video-level similarity score $score_{cg} = \text{sigmoid}(\text{FC}(g))$, which globally describes the relevance between the text and the entire video. To train the video-level relevance discriminator, a binary cross-entropy loss L_{video} (Appendix A.3) is employed on each query-video sample, which encourages the discriminator output large scores for samples with grounding results while small scores for that without grounding results.

Multi-level relevance prediction. The previously discussed two relevance discriminators assess the relevance between query and video at the frame and video levels respectively, yielding fine-grained relevance score ($score_{fg}$) and coarse-grained relevance score ($score_{cg}$). Considering that both granularity levels are deemed equally important for determining the relevance between text and video, we calculate the average of both fine-grained and coarse-grained relevance scores as the final multi-granularity relevance score P_v as:

$$P_v = (score_{fg} + score_{cg})/2. \tag{1}$$

3.4 Relation-Aware Segment Grounding

After obtaining the query-video relevance score, we aim to design a flexible segment grounding module to selectively predict the start and end boundaries of the target segments, according to the presence or absence of query-related segments in videos. However, existing segment grounding heads designed for TSG tasks [13, 15, 57, 16] are not suitable for the TSG-RF task. They always generate high-confidence start/end boundaries of segments even when there are no relevant segments in the video. To this end, we introduce a relation-aware segment grounding module that incorporates previously obtained relevance feedback during the segment grounding process, which accommodates the possibility that there may be no relevant segments within the video.

Specifically, to address situations where the video contains no relevant segments, we prepend a special token to the general video feature sequence, representing index 0. This token serves as both the starting and ending boundary indices for samples that do not have grounding results, indicating that the query text does not align with any start and end boundaries in the given video. Therefore, it is essential to introduce a contextual relevance-aware representation as the special feature token to ensure that the segment predictor can effectively handle the absence of relevant segments in the video. Considering the video-level relation signal vector g obtained from the video-level relevance discriminator contains enough relevance-aware knowledge to determine whether the segment exists or not, we utilize the video-level relation signal vector g as the special token in this segment predictor. Consequently, the range of boundary indices for relevant segment grounding is $[1, n]$. Thus, as shown in Figure 2, if a sample is determined to lack relevant segments, the ground-truth boundary index is assigned as $A = [0, 0]$; if a sample is deemed to have grounding results, the ground-truth boundary index is defined as $A = [a^s, a^e]$, where $1 \leq a^s < a^e \leq n$.

To predict the probability distributions of start and end boundaries, following [13], we use two unidirectional LSTMs with two feed-forward layers. The boundary prediction loss function $L_{boundary}$ is represented as:

$$L_{boundary} = -\frac{1}{2}(\sum Y_s \log(P_s) + \sum Y_e \log(P_e)), \tag{2}$$

where Y_s and Y_e are the one-hot vector representations of the ground-truth start (a_s) and end (a_e) boundary indices, respectively. P_s and P_e indicate the predicted start and end boundary probability distribution, respectively.

3.5 Training and Inference

In summary, the total training loss can be defined as:

$$L_{total} = L_{boundary} + \beta L_{frame} + \gamma L_{video}, \tag{3}$$

where β and γ are hyperparameters used to balance the importance of the three losses.

6

During the inference phase, the multi-granularity relevance score P_v is first compared with a threshold $m = 0.5$ to determine whether query-related content exists or not in the given video:

$$\begin{cases} \text{Has grounding result,} & \text{if } P_v \geq m \\ \text{No grounding result,} & \text{if } P_v < m \end{cases}.$$ (4)

For samples judged to have grounding results, the model predicts the start ($\hat{a}_s$) and end ($\hat{a}_e$) boundary indices of the target segment. Specifically, we first compute the joint probability distribution according to the start and the end boundary probability distributions P_s, P_e generated by the relation-aware segment grounding module. The predicted start ($\hat{a}_s$) and end ($\hat{a}_e$) boundary indices are obtained by maximizing the joint probability:

$$(\hat{a}_s, \hat{a}_e) = \arg \max_{\hat{a}_s, \hat{a}_e} (P_s^T P_e), \quad 0 \leq \hat{a}_s \leq \hat{a}_e \leq n.$$ (5)

4 Evaluation

4.1 Experimental Setup

Datasets. To verify the viability of our proposed model for TSG-RF, samples without grounding results where the video does not contain the content relevant to the query are required. Therefore, we reconstruct two commonly used datasets, Charades-STA [22] and ActivityNet Captions [23], by adding samples without grounding results (Details of how we obtain them are illustrated in the Appendix A.5.). For each sample in both validation and test sets, we add a corresponding sample without grounding result, resulting in 1:1 ratio of samples with and without grounding results.

The Charades-STA [22] dataset comprises 6,672 videos, where the training set contains 12,408 video-text sample pairs, while the test set comprises 3,720 pairs. Since the original Charades-STA does not have a validation set, we randomly halve the original test samples to form a validation set and a test set. The ActivityNet Captions [23] dataset consists of approximately 20,000 videos featuring diverse and open-content videos. Consistent with previous work [43] for the TSG task, the training set includes 37,421 video-text pairs, while the validation and test sets contain 17,505 and 17,031 samples, respectively. After reconstruction, the validation and test sets were augmented with an equal number of samples without grounding results, doubling the total number of sample pairs to 35,010 and 34,062, respectively.

For ease of reference, we name the corresponding reconstructed datasets as *Charades-RF* and *ActivityNet-RF*, respectively.

Performance metric. As our proposed TSG-RF task requires models to provide relevance feedback indicating samples with or without grounding results, we use accuracy ("Acc") to measure the ability of relevance feedback. For measuring the grounding ability, referring to the previous TSG works [13, 43], we utilize the "R{n}@{m}" and "mIoU". "R{n}@{m}" indicates the percentage of query texts in the top-*n* segments where at least one instance has an Intersection over Union (IoU) greater than *m*. "mIoU" represents the mean IoU across all test samples. It is worth noting that due to the inclusion of samples with no grounding results, we redefine the calculation of Intersection over Union (IoU) in four specific scenarios: (1) When the sample is predicted to have no grounding results but actually has grounding results, the IoU is set to 0. (2) When both the prediction and the ground truth indicate that the sample has no grounding results, the IoU is set to 1. (3) When the sample is predicted to have grounding results but actually has no grounding results, the IoU is set to 0. (4) When both the prediction and the ground truth indicate that the sample has grounding results, the IoU is calculated as the Intersection over Union between the predicted segment and the ground truth.

Implementation details. The implementation details are presented in Appendix A.6.

4.2 Comparison with Baseline Methods

As models specifically designed for TSG-RF are non-existing, we compare our proposed model with models targeted at conventional TSG, and adapt them to TSG-RF by adding an extra relevance discriminator for relevance feedback. Specifically, we select four traditional TSG models, including VSLNet [13], SeqPAN [14], EAMAT [15], ADPN [16], UniVTG [49], QD-DETR [48] considering their source code are released thus ensuring fair and replicable comparisons. For the relevance

Table 1: Performance comparison on Charades-RF and ActivityNet-RF dataset. Model^{++} denotes that the baseline model adapted to TRF-RF by utilizing an additional trained relevance discriminator, using two cascaded models: a relevance discrimination model and a segment grounding model. Our proposed RaTSG, a unified model for both relevance discrimination and segment grounding, achieves the best performance with a very lightweight model for TRF-RF.

Method	Charades-RF					ActivityNet-RF					Params (M)
	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU	
VSLNet	50.00	33.74	27.31	17.72	24.69	50.00	31.06	21.88	12.82	22.27	**1.16**
UniVTG	50.00	35.81	30.03	16.67	24.96	50.00	30.89	21.67	11.29	21.35	41.35
QD-DETR	50.00	35.16	29.46	19.27	25.31	50.00	26.50	19.15	11.07	18.99	7.07
ADPN	50.00	35.62	28.44	19.87	25.98	50.00	30.72	20.74	12.38	22.05	2.27
SeqPAN	50.00	35.35	29.57	20.51	26.14	50.00	31.85	22.65	13.34	22.86	1.19
EAMAT	50.00	37.12	30.59	20.86	27.27	50.00	31.10	20.80	12.07	22.07	94.12
VSLNet^{++}	71.94	61.40	56.77	49.65	54.67	81.60	66.15	58.37	50.64	58.65	5.34
UniVTG^{++}	71.94	62.58	58.55	48.79	54.65	81.60	66.15	58.36	49.46	58.00	45.53
QD-DETR^{++}	71.94	62.18	58.20	50.96	55.13	81.60	62.43	56.13	49.27	55.97	11.25
ADPN^{++}	71.94	62.26	57.23	51.16	55.41	81.60	65.85	57.41	50.28	58.47	6.45
SeqPAN^{++}	71.94	62.12	58.01	51.61	55.49	81.60	66.77	58.98	51.11	59.11	5.37
EAMAT^{++}	71.94	63.55	59.17	51.96	56.23	81.60	66.13	57.36	49.93	58.45	98.30
RaTSG (ours)	**76.85**	**68.17**	**61.91**	**54.19**	**59.93**	**84.27**	**69.02**	**60.68**	**52.88**	**61.15**	1.27

discriminator, we utilize a binary classification model which is trained separately to assess the relevance between the query text and the video (For the details of the relevance discriminator, please refer to Appendix B.1). Then, query-video samples identified as relevant by the discriminator are processed using models designed for the TSG task to predict the target segment. For ease of reference, we denote the corresponding models associated with an extra relevance discriminator as VSLNet^{++}, SeqPAN^{++}, EAMAT^{++}, ADPN^{++}, UniVTG^{++}, and QD-DETR^{++} respectively.

Table 1 summarizes the performance comparison and model parameters on the Charades-RF and ActivityNet-RF dataset. It is worth noting that traditional TSG models lack the ability to discriminate relevance. These models assume all samples have grounding results, making them incapable of correctly handling samples without grounding results. Hence, this often leads to mismatched grounding prediction, resulting in low recall and mIoU performance for TSG-RF. Additionally, since the test set has an equal ratio (1:1) of samples with and without grounding results, the relevance prediction accuracy of these models is only 50%. Additionally, the enhanced versions of these baseline models, namely VSLNet^{++}, SeqPAN^{++}, EAMAT^{++}, ADPN^{++}, UniVTG^{++}, and QD-DETR^{++}, include a relevance discriminator, which results in consistent performance gains compared to their counterparts without the discriminator. However, these enhanced models require separate and independent training of the relevance discriminator and video grounding components, leading to redundant use of computational resources and increased model size. In contrast, our proposed RaTSG model provides a more lightweight and comprehensive solution by integrating the discrimination and grounding modules, achieving the best performance.

4.3 Ablation Studies

4.3.1 Effectiveness of the Multi-Granularity Relevance Discriminator

To validate the effectiveness of the multi-granularity relevance discriminator, we compare it with degraded models that use only coarse-grained or fine-grained discrimination scores. As shown in Table 2, single-granularity models fail to perceive the various degrees of partial relevance between text and video, therefore performing worse than the multi-granularity one.

4.3.2 Effectiveness of the Relation-aware Segment Grounding

To assess the viability of the relation-aware segment grounding module, we conduct ablations on it by replacing the relation signal vector g with a randomly initialized one. As shown in Table 3, the relation-aware segment grounding consistently outperforms the random counterpart. The result demonstrates the benefit of using previously predicted relevance information of each query-video pair as prior knowledge for TSG-RF. Additionally, in order to further explore how relation features are

Table 2: The effectiveness of the multi-granularity discriminator on Charades-RF.

Granularity		Acc	R1@0.3	R1@0.5	R1@0.7	mIoU
coarse	fine					
✗	✓	75.35	67.34	60.91	53.84	59.27
✓	✗	75.73	67.63	60.48	53.25	59.18
✓	✓	**76.85**	**68.17**	**61.91**	**54.19**	**59.93**

Table 3: The effectiveness of the relation-aware segment grounding on Charades-RF.

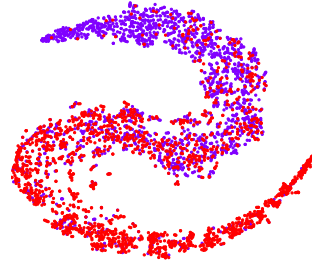| Relation-aware	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU
✗	76.40	66.18	59.62	51.96	57.82
✓	**76.85**	**68.17**	**61.91**	**54.19**	**59.93**

Figure 3: t-SNE visualization of relation signal vectors of all test samples on Charades-RF, where red dots represent relevant samples and blue points represent irrelevant samples of queries and videos.

Table 4: The performance comparison of relevance discrimination with and without the segment grounding module on Charades-RF.

Segment Grounding	Acc
✗	75.59
✓	**76.85**

Table 5: The performance comparison of segment grounding with and without the relevance discriminator on Charades-STA.

Discriminator	R1@0.3	R1@0.5	R1@0.7	mIoU
✗	67.47	54.62	35.43	49.37
✓	**74.19**	**56.61**	**37.47**	**53.02**

Table 6: Our proposed multi-granularity relevance discriminator and relation-aware segment grounding enhance can be jointly used to enhance traditional TSG methods for TSG-RF tasks.

Method	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU
EAMAT	50.00	37.12	30.59	20.86	27.27
EAMAT[++]	71.94	63.55	59.17	51.96	56.23
EAMAT+Ours	**76.37**	**67.47**	**62.02**	**54.83**	**59.58**

learned, we visualize the relation signal feature g by t-SNE [58] in Figure 3. It is demonstrated that dots with the same color are relatively more clustered than those with different colors. Overall, the result shows the good discrimination ability of the learned relation feature in our model for relevant and irrelevant samples.

4.3.3 Mutual Enhancement between Relevance Discrimination and Segment Grounding

Our proposed RaTSG is a unified model for both relevance discrimination and segment grounding tasks, trained in a multi-task learning manner. To explore whether the two tasks mutually enhance each other, we evaluate the counterparts with the corresponding segment grounding module or relevance discriminator removed. As shown in Table 4, removing the segment grounding module influences the performance of relevance discrimination. Similarly, as shown in Table 5, removing the relevance discriminator degrades the performance of segment grounding. The results allow us to conclude that the relevance discriminator and the segment grounding in our model mutually enhance each other, which also demonstrates the effectiveness of our unified dual-branch framework design.

4.3.4 Enhancing traditional TSG methods to work for TSG-RF tasks.

To further investigate the scalability of our proposed multi-granularity relevance discriminator and relation-aware segment grounding modules, we also conduct experiments to explore whether they can be adapted for traditional TSG methods to enable them to work for the TSG-RF task. Specifically, we integrate these components into an existing traditional TSG model, EAMAT, for comparison. As shown in Table 6, EAMAT with our devised multi-granularity relevance discriminator and relation-aware segment grounding outperforms the original EAMAT and the enhanced EAMAT[++] with an extra relevance discriminator. These results demonstrate the adaptability and effectiveness of our proposed modules for enhancing traditional TSG methods to work for the TSG-RF task.

4.4 Analysis of Grounding Examples

Figure 4 visualizes grounding examples obtained by our proposed RaTSG and VSLNet^{++}. The results show that our RaTSG effectively handles samples with and without grounding results. For the first sample with grounding results, RaTSG localizes video segments more accurately than VSLNet^{++}. Furthermore, foreground frame prediction scores obtained by RaTSG are more reasonable. We attribute it to the fact that RaTSG is trained using samples without grounding results, allowing the model to learn the similarity characteristics between background frames and the text, thus enhancing its ability to distinguish foreground and background frames in videos.

For the second sample without grounding results in Figure 4, RaTSG consistently predicts low foreground frame scores, providing accurate relevance feedback. Besides, the relation-aware segment grounding module assigns a high probability of the special index of 0, indicating no grounding result. In contrast, VSLNet^{++} provides incorrect relevance feedback for this sample, showing high foreground frame prediction scores and resulting in incorrect grounding segments.

In Figure 5, we illustrate two bad examples to discuss the limitations of our model. In the first example, our model incorrectly judges the relevance feedback due to the lack of audio cues which are crucial for identifying the action of sneezing. In the second example, our model misinterprets the temporal sequence of actions, mistaking the closing action for the opening action. These examples demonstrate that our proposed model struggles with handling audio-related actions and temporal-sensitive content. However, such limitation can be alleviated by integrating audio features and temporal modeling in the video representation moudel.

Figure 4: Visualization of grounding examples obtained by our proposed RaTSG and the VSLNet^{++} baseline.
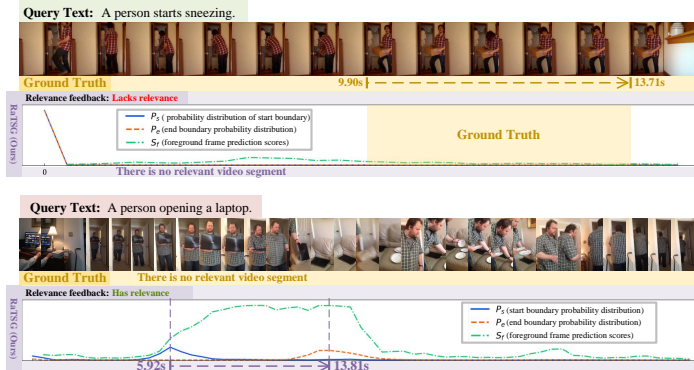

Figure 5: Visualization of two bad examples obtained by our proposed RaTSG.

5 Conclusions

This paper breaks through the bottleneck of existing TSG works that can not handle the cases where query-related segments do not exist, by introducing a more realistic and natural task, *i.e.*, TSG-RF. By incorporating the multi-granularity relevance discriminator into the segment grounding model with further mutual enhancement designs, our proposed method can effectively and efficiently localize precise video segments that closely match the query text for videos containing relevant content, while providing clear feedback indicating the absence of related segments for videos not containing relevant content. Besides, two constructed datasets for TSG-RF are contributed.

Acknowledgements

References

[1] Dionisis Andronas, George Apostolopoulos, Nikos Fourtakas, and Sotiris Makris. Multi-modal interfaces for natural human-robot interaction. *Procedia Manufacturing*, 54:197–202, 2021.

[2] Xue Song, Baohan Xu, and Yu-Gang Jiang. Predicting content similarity via multimodal modeling for video-in-video advertising. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2): 569–581, 2020.

[3] Gareth W Young and Oliver Dawkins. Multimodality: Exploring sensibility and sense-making beyond the metaverse. In *International Conference on Human-Computer Interaction*, pages 307–322. Springer, 2023.

[4] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19 (2):1–33, 2023.

[5] Zijian Zhang, Zhou Zhao, Zhu Zhang, Zhijie Lin, Qi Wang, and Richang Hong. Temporal textual localization in video via adversarial bi-directional interaction networks. *IEEE Transactions on Multimedia*, 23:3306–3317, 2020.

[6] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4280–4288, 2020.

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

[8] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. Benchmarking micro-action recognition: Dataset, method, and application. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 6238–6252, 2024.

[9] Qi Zheng, Jianfeng Dong, Xiaoye Qu, Xun Yang, Yabing Wang, Pan Zhou, Baolong Liu, and Xun Wang. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–21, 2023.

[10] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*, 2019.

[11] Xin Sun, Xuan Wang, Jialin Gao, Qiong Liu, and Xi Zhou. You need to read again: Multi-granularity perception network for moment retrieval in videos. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1022–1032, 2022.

[12] Zhe Xu, Da Chen, Kun Wei, Cheng Deng, and Hui Xue. Hisa: Hierarchically semantic associating for video temporal grounding. *IEEE Transactions on Image Processing*, 31:5178–5188, 2022.

[13] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020.

[14] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Parallel attention network with sequence matching for video grounding. *arXiv preprint arXiv:2105.08481*, 2021.

[15] Shuo Yang and Xinxiao Wu. Entity-aware and motion-aware transformers for language-driven action localization in videos. *arXiv preprint arXiv:2205.05854*, 2022.

[16] Houlun Chen, Xin Wang, Xiaohan Lan, Hong Chen, Xuguang Duan, Jia Jia, and Wenwu Zhu. Curriculum-listener: Consistency-and complementarity-aware audio-enhanced temporal sentence grounding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3117–3128, 2023.

[17] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[18] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

[19] Zhi-Hua Zhou and Min-Ling Zhang. Neural networks for multi-instance learning. In *Proceedings of the International Conference on Intelligent Information Technology, Beijing, China*, pages 455–459. Citeseer, 2002.

[20] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 246–257, 2022.

[21] Xiang Fang, Wanlong Fang, Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Renfu Li, Zichuan Xu, Lixing Chen, Panpan Zheng, et al. Not all inputs are valid: Towards open-set video moment retrieval using language. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

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

[23] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[24] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[25] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 15–24, 2018.

[26] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019.

[27] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8199–8206, 2019.

[28] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021.

[29] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020.

[30] Daizong Liu, Xi Ouyang, Shuangjie Xu, Pan Zhou, Kun He, and Shiping Wen. Saanet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing*, 413:145–157, 2020.

[31] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. *arXiv preprint arXiv:2109.06398*, 2021.

[32] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1841–1851, 2020.

[33] Daizong Liu, Shuangjie Xu, Xiao-Yang Liu, Zichuan Xu, Wei Wei, and Pan Zhou. Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2100–2108, 2021.

[34] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Zichuan Xu, Haozhao Wang, Xing Di, Weining Lu, and Yu Cheng. Transform-equivariant consistency learning for temporal sentence grounding. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(4):1–19, 2024.

[35] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Guoshun Nan, Pan Zhou, Zichuan Xu, Lixing Chen, He Yan, and Yu Cheng. Filling the information gap between video and query for language-driven moment retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4190–4199, 2023.

[36] Xiang Fang, Zeyu Xiong, Wanlong Fang, Xiaoye Qu, Chen Chen, Jianfeng Dong, Keke Tang, Pan Zhou, Yu Cheng, and Daizong Liu. Rethinking weakly-supervised video temporal grounding from a game perspective. In *European Conference on Computer Vision. Springer*, 2024.

[37] Daizong Liu, Xiaoye Qu, Xiang Fang, Jianfeng Dong, Pan Zhou, Guoshun Nan, Keke Tang, Wanlong Fang, and Yu Cheng. Towards robust temporal activity localization learning with noisy labels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16630–16642, 2024.

[38] Daizong Liu, Xiang Fang, Xiaoye Qu, Jianfeng Dong, He Yan, Yang Yang, Pan Zhou, and Yu Cheng. Unsupervised domain adaptative temporal sentence localization with mutual information maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3567–3575, 2024.

[39] Zeyu Xiong, Daizong Liu, Xiang Fang, Xiaoye Qu, Jianfeng Dong, Jiahao Zhu, Keke Tang, and Pan Zhou. Rethinking video sentence grounding from a tracking perspective with memory network and masked attention. *IEEE Transactions on Multimedia*, 2024.

[40] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171, 2018.

[41] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12168–12175, 2020.

[42] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019.

[43] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020.

[44] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166, 2019.

[45] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2986–2994, 2021.

[46] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

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

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

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

[50] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[51] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4065–4080, 2021.

[52] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021.

[53] Jianfeng Dong, Minsong Zhang, Zheng Zhang, Xianke Chen, Daizong Liu, Xiaoye Qu, Xun Wang, and Baolong Liu. Dual learning with dynamic knowledge distillation for partially relevant video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11302–11312, 2023.

[54] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. Sea: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 23:4351–4362, 2020.

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

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

[57] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182, 2019.

[58] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[59] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

[60] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*, 2016.

[61] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

[62] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, page 2, 2019.

[63] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[64] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

A More Technical Details

A.1 Video and Text Encoders

Considering the need for intra-modal interaction among features, this paper implements a single layer of the simplest Transformer encoder to achieve self-attention. Since the Transformer primarily models global information interactions and does not capture local region information within the feature sequence, four convolutional layers are added prior to the multi-head attention mechanism in the Transformer to enhance the capability for local area interaction.

A.1.1 Self-Attention Mechanism

To capture the complete semantic information and obtain sentence-level features of the query text, a self-attention mechanism is employed to aggregate the word embedding sequence. Specifically, the sentence-level feature h_Q of the query text can be calculated with the word-wise query feature Q as follows:

$$h_Q = Q^T \cdot \text{softmax}(QW), \tag{6}$$

where $h_Q \in \mathbb{R}^d$, and $W \in \mathbb{R}^{d \times 1}$ is a learnable weight matrix that represents the importance of each word in the sentence.

A.1.2 Text-Guided Video Feature Enhancement

To enhance the video semantic with the query contexts, we employ a query-video bidirectional attention mechanism [59, 60], which allows for simultaneous focus on the interaction between text words and video frames. The query-video bidirectional attention mechanism calculates attention scores from two directions: from video to text and from text to video. Both directions of attention matrices are derived from a common similarity matrix $S \in \mathbb{R}^{n \times m}$, which can be expressed as:

$$s_{ij} = \alpha(V_i, Q_j), \tag{7}$$

where $s_{ij} \in \mathbb{R}$ represents the similarity score between the i-th video feature vector V_i and the j-th word vector Q_j. The function $\alpha(\cdot)$ is a trilinear function [61] used to measure the similarity between the two input vectors V_i and Q_j:

$$\alpha(v, q) = W_0[v; q; v \odot q], \tag{8}$$

where W_0 represents a trainable weight matrix, and $\odot$ denotes element-wise multiplication. Normalizing the rows of the common similarity matrix S yields the attention weights from video to text $A^{v2q} = [a_1^{v2q}, a_2^{v2q}, \ldots, a_n^{v2q}] \in \mathbb{R}^{n \times m}$. Similarly, normalizing the columns of S gives the attention weights from text to video $A^{q2v} = [a_1^{q2v}, a_2^{q2v}, \ldots, a_m^{q2v}] \in \mathbb{R}^{m \times n}$. Subsequently, using the attention weights A^{v2q}, the weighted text feature sequence Q is used to construct the video feature sequence $C^v = A^{v2q} \cdot Q \in \mathbb{R}^{n \times d}$. Similarly, utilizing the attention weights A^{q2v}, the text feature sequence reconstructed from video features $C^{q2v} = A^{q2v} \cdot V' \in \mathbb{R}^{m \times d}$ is obtained. The reconstructed text feature sequence C^{q2v} is then mapped back into the video encoding space using the attention weights A^{v2q}, producing a text-aware video feature sequence $C^q = A^{v2q} \cdot C^{q2v} \in \mathbb{R}^{n \times d}$. Finally, through a simple feedforward network, the original video feature sequence V, the video feature sequence reconstructed from text features C^v, and the text-aware video feature sequence C^q are fused to obtain the text-guided enhanced video feature V_q:

$$V_q = \text{FFN}([V; C^v; V \odot C^v; V \odot C^q]), \tag{9}$$

where $V_q \in \mathbb{R}^{n \times d}$; FFN$(\cdot)$ denotes the feedforward network; $\odot$ indicates element-wise multiplication.

A.1.3 Multi-Modal Feature Fusion

Although the bidirectional attention mechanism facilitates cross-modal interaction and information fusion between query text and video, this process primarily enables local interaction between individual words and video frames. Therefore, we further integrate sentence-level features h_Q, which contain complete semantic information, with the text-guided enhanced video feature V_q to generate the final contextual multi-modal representation V_q'. This is achieved through the feature concatenation operation:

$$V_q' = \text{FC}([h_Q, v_{q1}, h_Q; v_{q2}, \ldots, h_Q; v_{qn}]), \tag{10}$$

where $[;]$ denotes concatenation along the feature dimension, and FC$(\cdot)$ represents a fully connected layer that maps the features to a d-dimensional space.

A.2 Supervision of the Frame-level Relevance Discriminator

Considering that the similarity score between text and each video frame depends on the fully connected layer's understanding of foreground and background frames, this paper introduces a binary cross-entropy loss L_{frame} to

supervise the learning of the rrame-level relevance discriminator. This supervision ensures that the discriminator correctly learns the differences between foreground frames and background frames. The loss function is defined as follows:

$$L_{frame} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_{f_i}\log(s_{f_i}) + (1 - y_{f_i})\log(1 - s_{f_i})\right), \tag{11}$$

where y_{f_i} is the i-th element of the binary sequence Y_f, and s_{f_i} is the i-th element of the foreground frame prediction score sequence S_f, with n being the length of sequence. The binary sequence Y_f is constructed as follows: for a video feature sequence of length n, a_s and a_e denote the start and end boundary indices of the target segment, respectively. Frames within these indices are labeled as 1 (indicating foreground frames), while those outside are labeled as 0 (indicating background frames).

A.3 Supervision of the Video-level Relevance Discriminator

The sentence-level query feature h_Q and the query-related video representation h_V are concatenated to produce a video-level relation signal vector $g \in \mathbb{R}^d$, represented as $g = \text{FC}([h_Q; h_V])$. To train the video-level relevance discriminator and ensure that the video-level relation signal vector accurately learns the semantic relationship between the query text and video segments, this paper introduces a video-level relation constraint L_{video}. Specifically, the video-level relation signal vector g is input into a fully connected layer with a sigmoid function to obtain the video-level similarity score $score_{cg} = \text{sigmoid}(\text{FC}(g))$, which globally describes the relevance between the text and the video. The video-level relation constraint is calculated as:

$$L_{video} = -(y_h\log(score_{cg}) + (1 - y_h)\log(1 - score_{cg})), \tag{12}$$

where y_h is the ground truth label indicating whether a sample has retrievable results or not.

A.4 Relation-Aware Segment Predictor

Considering the video-level relation signal vector g obtained from the video-level relevance discriminator contains enough relevance-aware knowledge to determine whether the segment exists or not, we utilize the video-level relation signal vector g as the special token in this segment predictor. Specifically, using the foreground frame prediction score sequence S_f produced by the Frame-level relevance discriminator, the foreground frames in the text-guided enhanced video feature V_q are enhanced, while the background features unrelated to the query text are diminished. The foreground-enhanced video feature sequence $\tilde{V}_q$ is calculated as:

$$\tilde{V}_q = V_q \odot S_f, \tag{13}$$

where $\odot$ denotes element-wise multiplication. Subsequently, the video-level relation signal vector g is concatenated to the front of the foreground-enhanced video feature sequence $\tilde{V}_q$, forming the input feature sequence $V_q^g = [g; \tilde{V}_q]$ for the relation-aware segment predictor, in dimension $\mathbb{R}^{(n+1)\times d}$. This sequence is then fed into a stacked two-layer LSTM network. The first layer of the LSTM generates the time-sequential feature sequence for predicting the start boundary $H^s = [h_0^s, h_1^s, \ldots, h_n^s]$ in $\mathbb{R}^{(n+1)\times d}$, and the second layer processes this output to generate the end boundary feature sequence $H^e = [h_0^e, h_1^e, \ldots, h_n^e] \in \mathbb{R}^{(n+1)\times d}$, as follows:

$$\begin{aligned} h_t^s &= \text{LSTM}_1(v_{qt}^g, h_{t-1}^s), \\ h_t^e &= \text{LSTM}_2(h_t^s, h_{t-1}^e), \end{aligned} \tag{14}$$

where $h_t^{s/e}$ denotes the t-th time-sequential feature vector in $H^{s/e}$, and v_{qt}^g represents the t-th video feature vector in V_q^g. LSTM_* indicates the corresponding layer of the LSTM network. The start and end time-sequential feature sequences H^s and H^e are each fed into their respective feedforward networks to obtain the boundary probability distributions $P_{s/e} \in \mathbb{R}^{(n+1)}$:

$$\begin{aligned} P_s &= \text{softmax}(FFN_{\text{start}}([H^s; V_q^g])), \\ P_e &= \text{softmax}(FFN_{\text{end}}([H^e; V_q^g])), \end{aligned} \tag{15}$$

where P_s represents the start boundary probability distribution, and P_e the end boundary probability distribution. The boundary prediction loss function $L_{boundary}$ is represented as:

$$L_{boundary} = -\frac{1}{2}\left(\sum Y_s\log(P_s) + \sum Y_e\log(P_e)\right), \tag{16}$$

where Y_s and Y_e are the one-hot vector representations of the ground-truth start (a_s) and end (a_e) boundary indices, respectively. P_s and P_e indicate the predicted start and end boundary probability distribution, respectively.
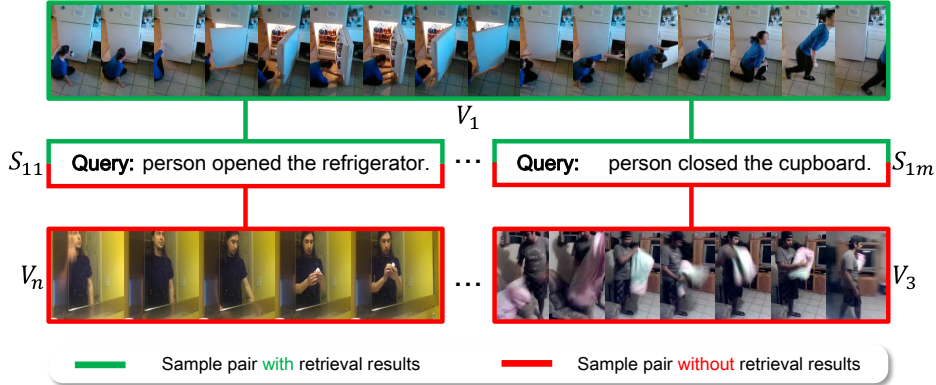
Figure 6: The organizational structure of reconstructed dataset.

Table 7: The detailed statistics of the different datasets.

Dataset	Domain	#Videos	#Anno.	#Anno-RF.
Charades-STA [22]	Indoor	6,672	12,408/1,860/1,860	-
ActivityNet Captions [23]	Open	19,207	37,421/17,505/17,031	-
Charades-RF	Indoor	6,672	12,408/3,720/3,720	-/1,860/1,860
ActivityNet-RF	Open	19,207	37,421/35,010/34,062	-/17,505/17,031

[1] #Anno. denotes the number of video-text annotation pairs in different sets (train/val/test).
[2] #Anno-RF. denotes the number of video-text annotation pairs without retrieval results in different sets (train/val/test).
[3] The original Charades-STA dataset lacked a validation set, so we randomly split the test set into two equal parts to create a validation set and a new test set

A.5 Dataset Reconstruction

Due to the absence of datasets specifically created for the TSG-RF task, this paper has reconstructed the validation and test sets of two widely used datasets in the TSG domain: Charades-STA [22] and ActivityNet Captions [23], to construct a testing environment for this task. In the original datasets, a single video corresponds to multiple query texts, as illustrated in Figure 6, where video V_1 corresponds to m query texts with grounding results $[S_{11}, S_{12}, \ldots, S_{1m}]$. For each text with grounding results, a corresponding sample without grounding results (i.e., the video does not contain segments relevant to the query text) is constructed.

Specifically, the reconstruction process is exemplified by the sample pair (S_{11}, V_1). By randomly selecting another video V_n from the video library to pair with the query text S_{11}, a sample pair without grounding results (S_{11}, V_n) is formed. To ensure the quality of reconstructed sample pairs and remove low-quality ones, this paper utilizes Large Language Models [62–64] for their deep and precise text modeling capabilities to achieve sample selection. Specifically, the BERT model [62] is used to extract features from the query text. Subsequently, the cosine similarity between the query text S_{11} and the query texts with grounding results from the randomly selected video $[S_{n1}, \ldots, S_{nm}]$ is calculated. If the maximum similarity is less than a threshold value (empirically set to 0.2), the reconstructed sample pair without grounding results (S_{11}, V_n) is added to the dataset. Otherwise, a new video is randomly selected until the condition is met. This methodology ensures the relevance and integrity of the dataset for accurately simulating the TSG-RF task.

As shown in Table 7, we name the corresponding reconstructed datasets as Charades-RF and ActivityNet-RF for the TSG-RF task. In these reconstructed datasets, we added an equal number of samples without grounding results to the validation and test sets. This augmentation ensures that the datasets more accurately reflect real-world scenarios where not all query texts correspond to relevant video segments.

A.6 Implementation Details

Each word in the query text is initialized using GloVe300d, which remains frozen during training. Visual features of videos are extracted using a pre-trained I3D network. The maximum video feature sequence length is set to 128. Sequences longer than this are uniformly downsampled to 128, while shorter sequences are zero-padded to the same length. During training, since the training set was not reconstructed and only contains samples with

grounding results, each batch includes randomly selected videos paired with the original query text to create samples without grounding results. In Equation 3, we empirically set $\beta = 6$ and $\gamma = 6$ to balance all loss functions at the start of training. For the threshold m in Equation 4, the value providing the highest accuracy on the validation set is chosen: 0.5 for Charades-RF and 0.3 for ActivityNet-RF. All experiments are conducted on a workstation with an NVIDIA GeForce RTX 3090Ti GPU and 256G RAM. Training our proposed model on the Charades-RF dataset takes approximately 2 hours, while training it on the ActivityNet-RF dataset takes approximately 5 hours.

B   More Experiments

B.1   Implement of the Enhanced Baseline with Relevance Feedback

Since there are currently no models specifically designed for the TSG-RF task, we select representative open-source models developed for the TSG task in recent years as baseline models to ensure fair and replicable comparisons, including VSLNet [13], SeqPAN [14], EAMAT [15], ADPN [16], UniVTG [49], QD-DETR [48]. While these baseline models demonstrate strong video segment grounding capabilities, they lack the ability to discern the relevance between the video and text in input samples, and thus assume that all samples contain retrievable results. Considering that directly training an additional relevance determination model is the most straightforward and efficient strategy to enable the above baseline models to discern relevance, this paper constructs a simple relevance determination model. By integrating this independently trained relevance determination model with the aforementioned baseline models, we compensate for their lack of relevance discernment capability, forming enhanced baseline models: VSLNet^{++}, SeqPAN^{++}, EAMAT^{++}, ADPN^{++}, UniVTG^{++}, and QD-DETR^{++}. Figure 7 illustrates the architecture of the simple relevance discrimination model. Specifically, the query text and video are processed through separate feature encoders. The encoded features of text and video are then concatenated along the sequence length dimension, with a CLS token added at the beginning to aid in learning the relevance between text and video. This concatenated sequence is subsequently input into a Transformer encoder to facilitate the interaction of information between feature vectors. Finally, the CLS token from the input sequence is processed through a fully connected layer to obtain the case discrimination score P_v.
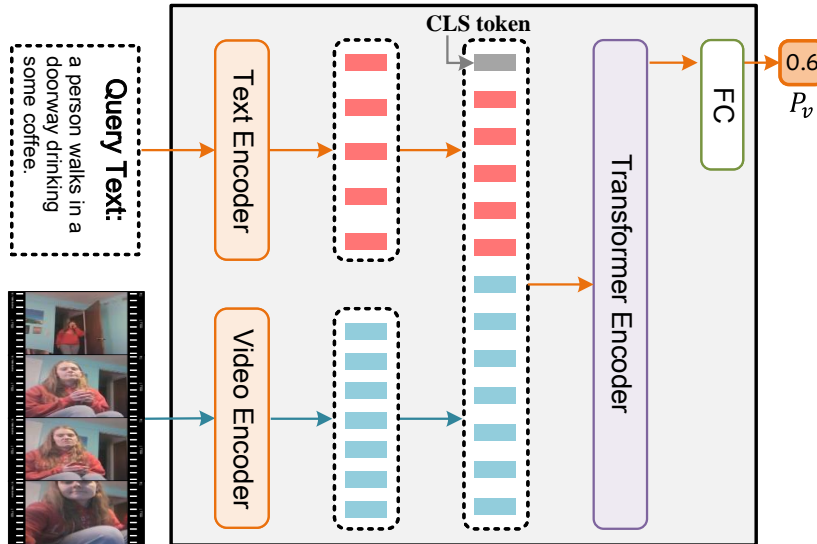

Figure 7: The architecture of the simple relevance discrimination Model.

B.2   Grouped Performance Comparison with Baseline Methods

To further explore model performance under unbalanced data conditions, the number of samples without grounding results in the test set was randomly decreased in proportion, yielding six groups of test sets with ratios of non-grounding to retrieval samples at 0, 0.2, 0.4, 0.6, 0.8, and 1.0, respectively. Observing Figure 8 from left to right, when the proportion of non-grounding samples is 0, the test set contains only grounding-result samples, testing all models on their TSG task performance. Starting from a 0.2 ratio, all models are assessed on their TSG-RF task performance. As the proportion of non-grounding samples in the test set increases, although there is a general improvement in the mIoU of all models, the increase exhibited by RaTSG is the most significant. This is attributed to RaTSG's multi-scale learning of the semantic relationships between text and video, effectively integrating relevance feedback during the segment grounding process. Particularly in test sets

with a high proportion of non-grounding samples, RaTSG demonstrates superior performance compared to other models, showcasing its exceptional ability in case discrimination and fine-grained grounding.
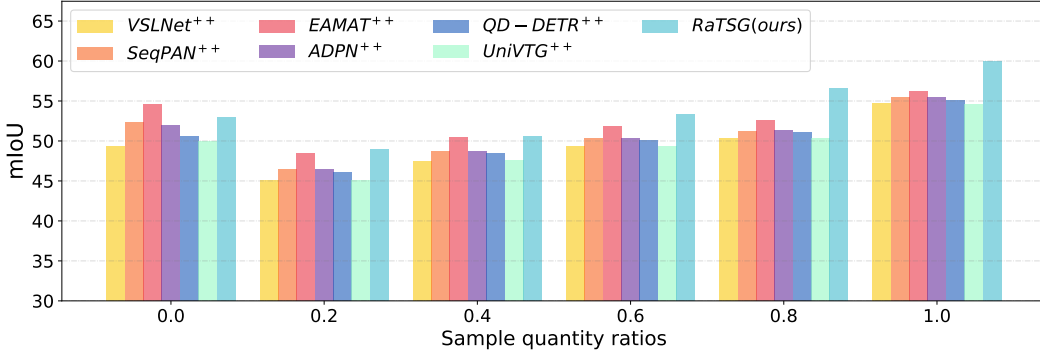

Figure 8: Performance Comparison across different models at varying sample quantity ratios.

B.3   Performance Comparison in the Context of TSG Task

As illustrated in Table 8, we present a performance comparison of various models on the Charades-STA and ActivityNet Captions datasets for the TSG task. While our model, RaTSG, does not outperform all other models in every metric, it remains highly competitive. On the Charades-STA dataset, RaTSG achieves an R1@0.3 of 74.19%, which is very close to the top-performing EAMAT at 74.25%. For R1@0.5, RaTSG scores 56.61%, just slightly behind SeqPAN's 59.14%. In terms of mIoU, RaTSG records 53.02%, comparable to ADPN's 51.96%. Similarly, on the ActivityNet Captions dataset, RaTSG attains an R1@0.3 of 61.46%, nearly matching ADPN's 61.46% and EAMAT's 63.20%. For R1@0.5, RaTSG scores 42.36%, close to SeqPAN's 45.31%. In terms of mIoU, RaTSG achieves 43.72%, demonstrating its competitiveness alongside other models. Although RaTSG is not consistently the highest performer, it demonstrates competitive results across various metrics. It is important to note that our model is specifically designed for the TSG-RF task, which includes the challenge of handling samples without grounding results. This specialized focus may slightly affect its performance on the standard TSG task, yet RaTSG remains highly competitive. This demonstrates the robustness and versatility of our approach, indicating its strong potential in more complex real-world scenarios.

Table 8: Comparison on Charades-STA and ActivityNet Captions in the context of TSG task.

Method	Charades-STA				ActivityNet Captions			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
VSLNet	67.47	54.62	35.43	49.37	62.12	43.76	25.64	44.54
SeqPAN	70.70	59.14	41.02	52.32	63.71	45.31	26.69	45.73
EAMAT	74.25	61.18	41.72	54.53	62.20	41.60	24.14	44.15
ADPN	71.24	56.88	39.73	51.96	61.46	41.49	24.78	44.12
QD-DETR	70.32	58.92	38.54	50.62	62.20	41.60	24.14	44.15
UniVTG	71.62	60.06	33.34	49.92	61.78	43.34	22.59	42.71
RaTSG(ours)	74.19	56.61	37.47	53.02	61.46	42.36	24.74	43.72

B.4   Comparison to CLIP as relevance discriminator

As shown in Table 9, we also try to compare our designed relevance discriminator with large-scale pre-trained vision-language models due to their strong cross-modal representation capabilities. To implement the comparison, we choose the CLIP model, a large-scale pre-trained vision-language model, to assess the relevance between the query text and video. Specifically, we first utilize the CLIP to measure the cosine similarity score between the query text and each frame of a video, and then aggregate the similarity scores over all frames to obtain the final relevance. In particular, we use two aggregate methods implemented by the CLIP model: (1) averaging all scores (CLIP-Avg) and (2) averaging on top-5 scores (CLIP-Top5). We incorporate CLIP with the SOTA method EAMAT to achieve the joint relevance feedback and grounding framework. The experimental results on the Charades-RF dataset are summarized in Table 9. We can find that CLIP-based EAMAT models achieve relatively lower performance, demonstrating that the CLIP model has a poor ability to predict relevance. We attribute it

to the fact that the CLIP model is simply pre-trained on common scenarios without additional fine-tuning on specific downstream task, thus severely suffering from domain shift issues on the target dataset.

Table 9: Comparison of relevance discrimination ability with CLIP on Charades-RF dataset.

Method	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU
EAMAT^{++}	71.94	63.55	59.17	51.96	56.23
EAMAT$^{CLIP-Avg}$	60.70	54.92	51.96	47.10	50.21
EAMAT$^{CLIP-Top5}$	62.52	55.75	52.47	46.94	50.25
RaTSG(Ours)	**76.85**	**68.17**	**61.91**	**54.19**	**59.93**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction of this paper accurately reflect its contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the paper discusses the limitations of the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all the necessary information to reproduce the main experimental results, ensuring that the main claims and conclusions can be independently verified.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data have been submitted as supplementary material in ZIP format.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: As mentioned in Section 4.1, this paper provides comprehensive details on the training and test settings, including data splits, hyperparameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not provide error bars, confidence intervals, or statistical significance tests for the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix A.6, we provide the type of compute workers (GPUs) and the size of memory. Additionally, we provide the training time of our proposed model on each dataset.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research fully adheres to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: As mentioned in Section 1, the TSG-RF task proposed in this paper improves upon the limitations of existing TSG tasks, making it more suitable for real-world applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.