ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230630

Turn on the Right Track: Weakly Supervised Video Moment Retrieval with Self-Improving Query Reconstruction

Yiming Zhong[a], Haifeng Sun[a], Jiachang Hao[a], Jing Wang[a], Cheng Zhou[b], Qi Qi[a;*], Jingyu Wang[a] and Jianxin Liao[a]

[a]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[b]China Mobile Research Institute

Abstract. Existing weakly-supervised temporal sentence grounding methods typically regard query reconstruction as the pretext task in place of the absent temporal supervision. However, their approaches suffer from two flaws, i.e. insignificant reconstruction and discrepancy in alignment. Insignificant reconstruction indicates the randomly masked words may not be discriminative enough to distinguish the target event from unrelated events in the video. Discrepancy in alignment indicates the incorrect partial alignment built by query reconstruction task. The flaws undermine the reliability of current reconstruction-based methods. To this end, we propose a novel Self-improving Query ReconstrucTion (SQRT) framework for weakly-supervised temporal sentence grounding. To deal with insignificant reconstruction, we devise a key words mining strategy to determine the important words for language grounding. To attain better moment-query alignment, we introduce inter-sample contrast to tackle the partial alignment built by query reconstruction. The self-improving framework utilizes query reconstruction for language grounding and alleviates the discrepancy in alignment, thus turning on the right track. Experiments on two popular datasets show that SQRT achieves state-of-the-art performance on Charades-STA and comparable performance to the state-of-the-art on ActivityNet Captions.

1 Introduction

Video moment retrieval, also known as temporal sentence grounding, aims to determine the moment that best corresponds to the given natural language query sentence in an untrimmed video. Many works [6, 8, 30, 14, 19, 23, 26] solve this task in a fully-supervised manner, where the temporal boundary annotation for every query sentence is required. However, the expensive annotation cost hinders their further application to real-world scenarios. Weakly supervised temporal sentence grounding (WTSG) alleviates this problem by requiring only video-level sentence annotations without temporal boundaries and has gained more attention.

In WTSG, previous reconstruction-based methods [11, 18, 33] regard the query reconstruction task as the pretext task in place of the absent ground truth supervision during training. Given a masked query sentence and candidate proposals, these methods mine the relation between video and text by reconstructing the original query sentence. During inference, they regard the reconstruction errors as the measurement of semantic similarity between the query and proposals and rank proposals according to reconstruction errors for final predictions.

(a) Insignificant reconstruction

(b) Discrepancy in alignment

Figure 1. a) Insignificant reconstruction indicates the masked words may not be discriminative enough to distinguish the target event from unrelated events in the video. Key words for language grounding are highlighted in red. b) Discrepancy in alignment. Three words are masked to perform query reconstruction. The reconstruction-based model focuses on mining the elements in visual modality that correspond to the masked words (i.e. the man, the frisbee, and the dog), leading to inaccurate grounding results.

However, this formula suffers from two flaws: *insignificant reconstruction* and *discrepancy in alignment*. Insignificant reconstruction indicates the randomly masked words may not be discriminative enough to distinguish the target event from unrelated events in the video. As illustrated in Figure 1(a), the masked words 'The' and 'cabinet' are insignificant to determine the target moment compared

* Corresponding Author. Email: qiqi8266@bupt.edu.cn

with the key words (highlighted in red). This may lead to misalignment between video and text during training and inference. Discrepancy in alignment indicates the incorrect partial alignment built by query reconstruction task. The query reconstruction task focuses on mining the elements in visual modality that correspond to the masked words, while the grounding task aims to determine the video segment that best corresponds to the overall semantics of the query sentence. As illustrated in Figure 1(b), the reconstruction errors respond on segments related to masked words 'man', 'frisbee', and 'dog', failing to ground from the global semantics of the query. Therefore, when performing the query reconstruction task, the moment that yields the least reconstruction error does not necessarily match the *entire query* best. It is inaccurate to perform language grounding purely depending on the query reconstruction error. Thus, it remains an open problem on how to utilize query reconstruction in temporal sentence grounding.

To deal with insignificant reconstruction, we evaluate the effectiveness of reconstruction by mining the words that correspond to the critical visual elements of the event. To this end, we devise a key words mining strategy to determine the important words to language grounding. To attain better moment-query alignment, we introduce inter-sample contrast to tackle the partial alignment built by query reconstruction. In WTSG, the reconstruction objective forces the model to focus on mining local visual elements corresponding to the masked words. The model is insensitive to sentence-level semantics due to the inability to utilize semantic cues between samples. Hence, we introduce inter-sample contrast and explicitly model sentence-level features. We mine similarity cues by pairing a video with a diversity of cross-sample unmatched queries in the corpus as negative pairs. Paired with the query reconstruction task, intra-sample and inter-sample contrast are fully utilized to infer the semantics of the entire query. It improves grounding performance without extra labels, turning the optimization on the right track.

Taking these into account, we propose a simple Self-improving Query ReconstrucTion (SQRT) framework to better utilize query reconstruction for WTSG. The framework consists of a query reconstruction branch and a grounding branch. The reconstruction branch performs query reconstruction and generates pseudo-labels for the grounding branch. The grounding branch localizes the target moment and self-improves. During query reconstruction, we choose the set of words whose reconstruction errors with larger variance across all moments as key words. With key words masked, the corresponding reconstruction errors are more reliable as semantic measurement. In the training phase, the key words reconstruction errors are converted to soft pseudo-labels for intra-sample contrast. The grounding branch improves itself through intra-sample contrast and inter-sample contrast. To model the overall semantics of query text explicitly, we devise a query-enhanced proposal interaction module. The self-improving framework utilizes query reconstruction for language grounding and alleviates the discrepancy in alignment.

In summary, our contributions are three-fold: (1) We reveal two major flaws in the query reconstruction task used in WTSG, which limits its further application in temporal sentence grounding. (2) We propose a self-improving query reconstruction framework for weakly supervised temporal sentence grounding as a simple approach to deal with the flaws. (3) Experiments on two popular datasets Charades-STA and ActivityNet Captions demonstrate the effectiveness of our approach. We achieve new state-of-the-art performance on Charades-STA and comparable performance to the state-of-the-art on ActivityNet Captions for weakly-supervised temporal sentence grounding.

2 Related Work

Fully-supervised temporal sentence grounding. In the fully-supervised setting, the accurate start and end timestamps paired with the sentence query are provided for model training. The task is first proposed in CTRL [6], where visual and textual features are fused to regress the temporal offset to a certain candidate clip. The 2D Temporal Adjacent Networks (2DTAN) [30] organize the candidate moments in a 2D temporal map manner and model the relations between temporally adjacent moments. LGI [14] proposes Sequential Query Attention Network(SQAN) to extract semantic phrase features and conduct local-global video-text interactions to directly regress the target timestamps.

However, annotating temporal boundaries for every query sentence is laborious and expensive. Also, the manual annotations may be subjective and inconsistency exists between different annotators, as reported in [15]. These issues hinder the application of fully-supervised methods to real-world scenarios.

Weakly-supervised temporal sentence grounding. In weakly-supervised temporal sentence grounding, only video-level query texts are required during training. Existing works generally tackle this task in an MIL-based or reconstruction-based paradigm. MIL-based methods [7, 32, 12] regard matched video-query pairs as positive bags and unmatched ones as negative bags to learn the latent alignment between video and text. Reconstruction-based methods [11, 18, 33] assume the proposal with the least reconstruct error matches the original query best. SCN [11] introduces semantic completion task which serves as guidance for the scoring process. CPL [33] develops proposal contrast between positives and intra-video negatives by reconstruction error.

Recent studies [10, 17, 27] in natural language understanding show that the random mask strategy is sub-optimal during the pre-training of language models. However, the effect of query reconstruction task has never been studied in WTSG. We argue that the reconstruction-based methods suffer from the aforementioned flaws, which undermine their reliability in temporal sentence grounding. To deal with the flaws, we propose a simple Self-improving Query ReconstrucTion (SQRT) framework to better utilize query reconstruction for WTSG.

3 Method

Given an untrimmed video V and a free-form query sentence Q, temporal sentence grounding aims to determine the video segment with temporal boundary $b = (t_s, t_e)$ that best matches the semantics of the query sentence Q.

3.1 Overall Framework

As shown in Figure 2, SQRT consists of a query reconstruction branch (Rec-branch) and a grounding branch.

The Rec-branch takes video features and masked query features as input to perform visual-based query reconstruction and generates pseudo-labels for intra-sample contrast learning of the grounding branch. It can be implemented as any visual-based query reconstruction model with reconstruction error as output. We choose CPL [33] as the implementation of Rec-branch.

The grounding branch aims to localize the target moment and predicts moment scores. It learns through intra-sample contrast and inter-sample contrast during training. The grounding branch consists of four key components: multimodal fusion transformer, query-enhanced proposal interaction, scoring network, and self-improving

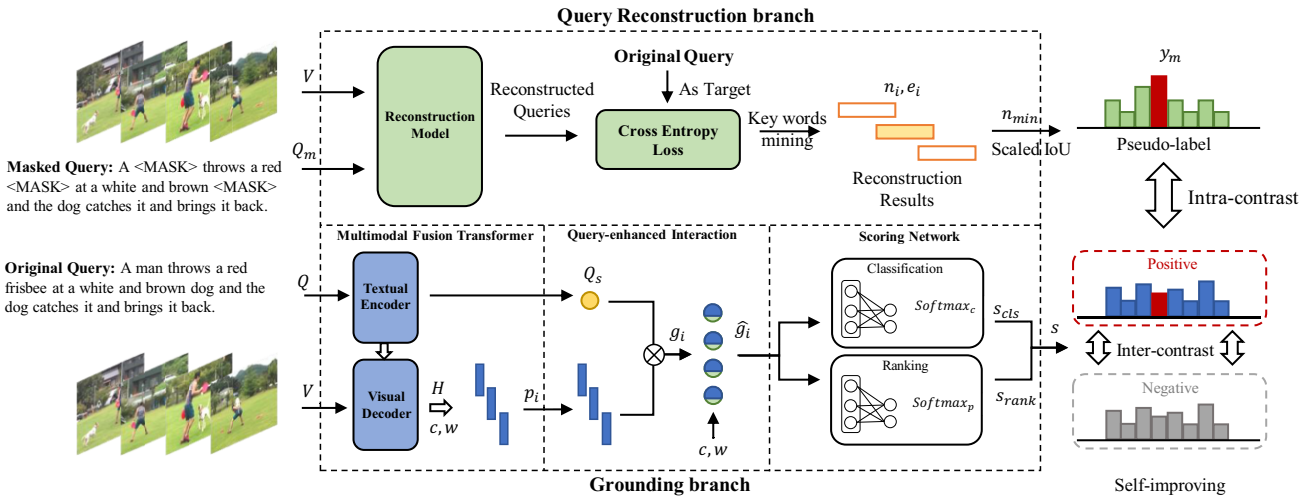

Figure 2. An overview of our Self-improving Query ReconstrucTion (SQRT) framework. SQRT consists of a query reconstruction branch (Rec-branch) and a grounding branch. The query reconstruction branch performs query reconstruction and outputs reconstruction results for the grounding branch. The grounding branch fuses video and query features through a query-enhanced proposal interaction module and conducts self-improving through intra-sample contrast from pseudo-labels and inter-sample contrast.

module. The multimodal fusion transformer performs an initial multi-modal fusion of video features and query features and predicts learnable Gaussian masks to generate proposals dynamically. The proposal features are passed to the query-enhanced proposal interaction module to conduct high-level feature interaction. The scoring network takes the enhanced proposal features as input and predicts proposal scores. The framework improves itself by performing intra-sample contrast and inter-sample contrast.

The two branches are parallel and are trained separately. We adopt a two-phase training strategy. The Rec-branch is trained in the first stage. In the second phase, the Rec-branch is frozen and the grounding branch is trained.

3.2 Reconstruction branch

The Rec-branch takes video features and masked query features as input to perform query reconstruction. It mines multimodal similarity cues by query reconstruction. Key words mining strategy is applied during reconstruction. The Rec-branch produces predicted masked words, their corresponding reconstruction errors, and temporal segments. The implementation of the Rec-branch is not the main concern of this paper, thus we leave it in our supplemental material.

Key Words Mining Strategy. The query reconstruction is a key part of mining the video-text correlation. However, the random mask approach in existing works may end up reconstructing the words that are not helpful in distinguishing the query-related moment from others. To this end, we propose a key words mining strategy to find out the essential words in language modality for temporal sentence grounding.

Intuitively, the words that are essential for distinguishing the target moment from others may contribute more to certain moments during reconstruction while less to others. Based on this, we aim to find out the group of words which exhibit greater variance on reconstruction errors along temporal dimension. To mine the words that are beneficial to language grounding, we generate L groups of candidate masked words in a query. In each group, we randomly mask one-third of the words in a query and ensure groups are different from each other. The nouns, verbs, and adjectives are masked with a higher probability. For each word group, we perform query reconstruction according to the visual features within K candidate proposals of different time spans. The reconstruction error is implemented as the cross entropy loss between the ground truth probability on vocabulary list p_{gt} and the predicted probability p_l^k. The reconstruction errors ϵ_l^k are normalized with a softmax function along temporal dimension. For a given group of masked words, we compute the variance over the candidate proposal dimension. The group of words $\{w_{l_{key}}\}$ with maximal variance are selected as the key words.

$$\epsilon_l^k = \text{CrossEntropy}(p_{gt}, p_l^k), \tag{1}$$

$$\hat{\epsilon}_l^k = \text{softmax}(\epsilon_l^k), \tag{2}$$

$$l_{key} = \arg\max_{l \in L}(\text{Variance}(\hat{\epsilon}_l^k)), \tag{3}$$

$$l = 1, 2, ..., L; k = 1, 2..., K \tag{4}$$

After the key words are selected, the candidate proposal with the least reconstruction error on the key words set is regarded as the high-confidence segment n_i. The pseudo-labels for self-improving are generated based on n_i.

3.3 Multimodal Fusion Transformer

We use transformer [22] to perform the initial multi-modal fusion of the video feature and query feature in the grounding branch.

Feature Extraction. For visual modality, we use pre-trained 3D CNN (eg. C3D [21]) to extract video features $V = \{v_1, v_2, ..., v_i\}_{i=1}^T \in \mathbb{R}^{T \times D_v}$, where T is the number of frames and D_v is the dimension of video features. For language modality, we use GloVe [16] to embed word sequence into query features $Q = \{w_1, w_2, ..., w_i\}_{i=1}^N \in \mathbb{R}^{N \times D_w}$ for fair comparisons, where N is the number of words and D_w is the dimension of query features. A [CLS] token is inserted at the beginning of the query features to aggregate sentence-level features. The video features and query features are projected to the same dimension D.

Multimodal Fusion. We fuse the extracted video features and query features using a transformer. Specifically, it consists of a textual encoder and a visual decoder. We append a [CLS] token v_{cls} to the

video features to obtain the fused global video feature. The textual encoder $Enc_t(\cdot)$ extracts contextual information in language modality. Then the visual decoder $Dec_v(\cdot)$ takes the encoded query features along with the video features V as input to get the fused feature $H = \{h_1, h_2, ..., h_{cls}\} \in \mathbb{R}^{(T+1)\times D}$:

$$H = Dec_v(V, Enc_t(Q)). \tag{5}$$

Proposal Generation. Following CPL [33], we use learnable Gaussian masks to generate proposals dynamically according to visual content from video and semantic information from text for both branches. As the [CLS] token h_{cls} gathers global context information from both modalities, we apply a fully-connected layer to predict the center and width of candidate proposals. Then, a Gaussian function is applied to generate masks for proposals:

$$c, w = \text{Sigmoid}(FC(h_{cls})), \tag{6}$$

$$m_k = \frac{1}{\sqrt{2\pi}(w_k/\sigma)} exp(-\frac{(i/T - c_k)^2}{2(w_k/\sigma)^2}), \tag{7}$$

$$k = 1, ..., K; i = 1, 2, ..., T \tag{8}$$

where σ is the hyperparameter of the Gaussian function and K is the number of candidate proposals.

For every video-query pair, we predict K candidate proposals. To reduce highly overlapped redundant proposals, a regularization term $\mathcal{L}_{div}$ is applied to encourage temporal diversity between candidate proposals thus increasing the recall rate:

$$\mathcal{L}_{div} = \|mm^\top - \lambda I\|_F^2, \tag{9}$$

where λ is the hyperparameter controlling the extent of overlap between proposals, I is the identity matrix, and $\|\cdot\|_F$ is the Frobenius norm.

The proposal feature p is obtained by mean-pooling on the fused video feature within the proposal.

3.4 Query-enhanced Proposal Interaction

For the grounding branch, we need to measure similarities based on the overall semantics of the query to achieve more accurate grounding. To this end, we propose a query-enhanced proposal interaction module.

To encourage interaction between high-level features from both modalities, we perform proposal-sentence fusion. We use the feature of [CLS] token after text encoding as the sentence-level query feature Q_s. Every proposal feature is fused with query feature Q_s using Hadamard product to get the enhanced proposal feature g_i. To make the proposal features more discriminative, we inject temporal information into proposal features by concatenating every proposal feature with its normalized center and width along the channel dimension. The query-enhanced proposal feature $\hat{g}_i$ is obtained by:

$$g_i = \boldsymbol{W_3}(\boldsymbol{W_1} p_i \odot \boldsymbol{W_2} Q_s), \tag{10}$$

$$\hat{g}_i = \text{concat}(g_i, c, w) \tag{11}$$

where $\boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{W_3}$ are learnable parameters.

Scoring Network. After sufficient multi-modal fusion, we aim to produce a reliable proposal score for every candidate proposal. We adopt a two-branch network consisting of a classification network and a ranking network to predict scores. The classification network evaluates the semantic similarity between the proposal and query text. It consists of an MLP followed by a softmax function applied along the channel dimension, resulting in the classification score $s_{cls} \in \mathbb{R}^{K\times 2}$. The ranking network ranks the candidate proposals and gives their relative matching scores. It has the same structure as the classification network except that the softmax function is applied along the proposal dimension to obtain the ranking score $s_{rank} \in \mathbb{R}^{K\times 1}$. The final score s for candidate proposals is computed by the Hadamard product of the two scores:

$$s_{cls} = Softmax_c(MLP(\hat{g}_i)). \tag{12}$$

$$s_{rank} = Softmax_p(MLP(\hat{g}_i)). \tag{13}$$

$$s = s_{cls} \odot s_{rank}. \tag{14}$$

3.5 Self-improving

To deal with the discrepancy in the alignment of reconstruction-based methods, we propose a self-improving scheme based on reconstruction results. The grounding branch conducts self-improving by learning from pseudo labels generated through reconstruction results and performs inter-sample contrast simultaneously.

Pseudo Label Generation. To utilize the correlation cues from reconstruction results, we propose to learn from soft labels generated from reconstruction results.

The reconstruction results contain several temporal segments n_i with corresponding reconstruction errors e_i. We select the segment with the least reconstruction error as the reconstruction highest response segment n_{min}. To generate soft labels, we compute the IoU score o_m between n_{min} and all candidate proposals, then scale the IoU scores to obtain the final pseudo labels y_m:

$$y_m = \begin{cases} 0, & o_m \le t_{min} \\ \frac{o_m - t_{min}}{t_{max} - t_{min}}, & t_{min} < o_m < t_{max} \\ 1, & o_m \ge t_{max} \end{cases}, \tag{15}$$

where t_{min} and t_{max} are IoU threshold.

Inter-sample Contrast. To enhance the ability to measure the similarities between video and overall semantics of query text, we introduce inter-sample contrast loss to contrast between matched video-query pairs and unmatched pairs. For a given video, we randomly select a query sentence within the batch to form the unmatched negative pair. The inter-sample contrast loss is defined as:

$$\mathcal{L}_{inter} = -\sum(log(S_p(V, Q_p) + log(S_n(V, Q_n))), \tag{16}$$

where the $S(V, Q) = \sum_{i=1}^{K} s_i(p_i, Q)$ is the score aggregation function, and S_p, S_n indicates the aggregated scores on the matched and unmatched channels, respectively.

3.6 Training and Inference

Training. The Rec-branch and grounding branch are optimized independently. There are two phases in training. The Rec-branch is trained in the first stage. In the second phase, the Rec-branch is frozen and the grounding branch is trained. The objective of the Rec-branch is the reconstruction error with respect to the original query. More details are shown in our supplemental material. For the grounding branch, the final loss consists of three parts: the inter-sample contrast loss $\mathcal{L}_{inter}$ for video-level contrast, intra-sample contrast loss $\mathcal{L}_{intra}$ for leveraging reconstruction results, and diversity loss $\mathcal{L}_{div}$ for diversifying learnable proposals. The intra-sample contrast loss is implemented as the cross entropy loss between the proposal score $s_i(p_i, Q)$ and the pseudo labels generated from the self-improving module:

$$\mathcal{L}_{intra} = \sum_{i=1}^{K} \text{CrossEntropy}(y_i, s_i). \quad (17)$$

The overall loss for the grounding branch is formulated as the weighted sum of the three loss term:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{inter} + \alpha_2 \mathcal{L}_{intra} + \alpha_3 \mathcal{L}_{div}. \quad (18)$$

Inference. When training is finished, the Rec-branch is discarded and the grounding branch is used for inference.

4 Experiment

4.1 Datasets

To evaluate the effectiveness of our proposed method, we conduct experiments on two popular benchmarks for temporal sentence grounding: Charades-STA [6] and ActivityNet Captions [9]. Both datasets are publicly available. For Charades-STA, the numbers of video-query pairs of training/validation/testing set are 12408/-/3720. For ActivityNet Captions, the numbers of video-query pairs of training/validation/testing set are 37417/17505/17031.

4.2 Evaluation Metrics

Following previous work [6, 13, 11], we adopt the 'R@n, IoU=m' metric to evaluate the performance of SQRT, where n means the top-n retrieval results and m is the pre-defined threshold. The metric represents the percentage of moments whose Intersection over Union (IoU) with the ground truth segment is greater than threshold m in top-n prediction.

4.3 Implementation Details

Data Preprocessing. We downsample the video at 8 frames per second and extract the video features using a pre-trained video backbone. For visual features, we use the pre-trained I3D [1] model to extract frame features for Charades-STA and C3D [21] model for ActivityNet Captions. For language features, we set the dimension of word features as 300 and initialize them with GloVe [16] word2vec, following previous works for fair comparisons. The maximum number of words in a query sentence is set to 20.

Model Settings. The dimension of hidden states in transformer D is set to 256 and there are 3 layers and 4 attention heads. The number of learnable candidate proposals K is set to 8 for both datasets. For hyperparameters, we set $\alpha_1 = \alpha_2 = \alpha_3 = 1$, $\sigma = 9$, $t_{max} = 0.7$, $t_{min} = 0.3$ for both datasets. The parameter λ in $\mathcal{L}_{div}$ is set to 0.146 for Charades-STA and 0.06 for ActivityNet Captions. The number of groups of candidate masked words in key words mining L is set to 3 for Charades-STA and 5 for ActivityNet Captions. The seed for generating random numbers is set to 8. During training, we use Adam optimizer and initial learning rate 1e-4. We train the reconstruction branch for 30 epochs and the grounding branch for 50 epochs on both datasets using NVIDIA RTX 3090 GPU.

4.4 Comparisons with State-of-the-Art Methods

We compare the performance of our SQRT with existing works in WTSG, as shown in Table 1 and Table 2. The best performance is highlighted in bold, and the second best is underlined. On Charades-STA, our model achieves superior performance on all metrics of R@1 and two of the three metrics of R@5. For the top-1 metric 'R@1, IoU={0.3, 0.5, 0.7}', our SQRT achieves performance gain by 4.91%, 3.1% and 2.53% compared with the state-of-the-art, respectively. It indicates that our model performs better in fine-grained grounding on the top-1 result. On ActivityNet Captions, our SQRT outperforms existing methods on the important metric of 'R@1, IoU={0.5, 0.7}' by 2.05% and 3.3%, respectively. This means our SQRT achieves more accurate grounding on the top-1 result on this challenging dataset. By setting λ=0.1 to encourage diversity, our performance on 'R@5, IoU={0.5, 0.7}' surpasses that of the state-of-the-art methods, which achieve similar results at the expense of many redundant proposals.

Table 1. Performance comparison on Charades-STA.

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
TGA [13]	32.14	19.94	8.84	86.58	65.52	33.51
SCN [11]	42.96	23.58	9.97	95.56	71.80	33.87
CTF [3]	39.80	27.30	12.90	-	-	-
WSRA [5]	50.13	31.20	11.01	86.75	70.50	39.02
CCL [32]	-	33.21	15.68	-	73.50	41.87
VLANet [12]	45.24	31.83	14.17	95.72	82.82	33.33
BAR [28]	44.97	27.04	12.23	-	-	-
MARN [18]	48.55	31.94	14.81	90.70	70.00	37.40
RTBPN [31]	60.04	32.26	13.24	97.48	71.85	41.18
WSTAN [24]	43.39	29.35	12.28	93.04	76.13	41.53
LoGAN [20]	51.67	34.68	14.54	92.74	74.30	39.11
VCA [25]	58.58	38.13	19.57	**98.08**	78.75	37.75
LCNet [29]	59.60	39.19	18.87	94.78	80.56	45.24
CPL [33]	66.4	49.24	22.39	96.99	84.71	52.37
SQRT	**71.31**	**52.34**	**25.14**	96.2	**85.56**	**53.93**

Table 2. Performance comparison on ActivityNet Captions.

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
WS-DEC [4]	41.98	23.34	-	-	-	-
SCN [11]	47.23	29.22	14.88	71.45	55.69	31.81
EC-SL [2]	44.29	24.16	-	-	-	-
CTF [3]	44.30	23.60	-	-	-	-
CCL [32]	50.12	31.07	-	77.36	61.29	-
BAR [28]	49.03	30.73	-	-	-	-
MARN [18]	47.01	29.95	-	72.02	57.49	-
RTBPN [31]	49.77	29.63	-	79.89	60.56	-
WSTAN [24]	52.45	30.01	11.77	79.38	63.42	39.29
VCA [25]	50.45	31.00	-	71.79	53.83	-
LCNet [29]	48.49	26.33	-	**82.51**	62.66	-
CPL [33]	**55.73**	31.37	13.57	63.05	41.13	22.13
SQRT(λ=0.06)	53.15	**33.42**	**18.18**	70.68	54.53	33.95
SQRT(λ=0.1)	50.43	27.55	15.90	78.42	**63.45**	**43.05**

Table 3. Performance on the new testing split.

Method	R1@			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
Rec-only	57.80	28.88	11.03	36.45
Rec-grounding	53.91	33.98	17.49	36.22
SQRT	54.29	35.79	22.12	38.90

4.5 Ablation Study

We conduct ablation study on the key components of SQRT.

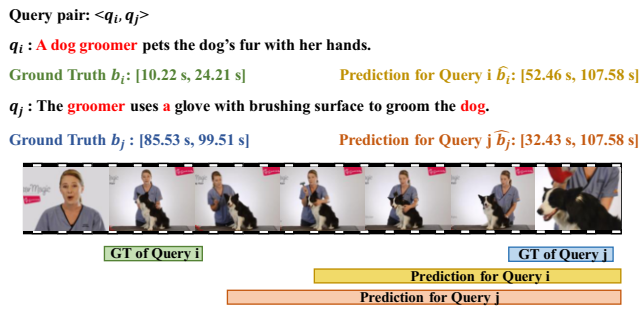

Figure 3. Illustration of validation of discrepancy in alignment. Given a video, a query pair consists of two queries with different semantics and share some common words. Models are validated on all such query pairs. For the reconstruction-based model, the common words are masked during reconstruction. The masked words are in red.

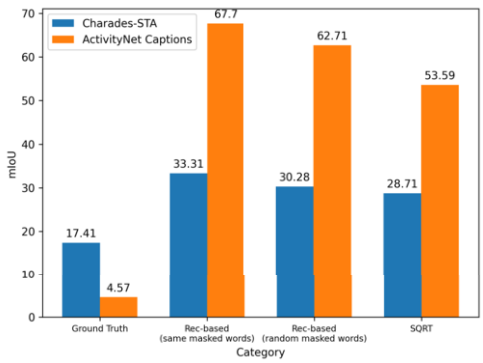

Figure 4. Validation of discrepancy in alignment on two datasets.

Table 4. Ablation study of key words mining strategy.

Method	R1@			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
Random mask	69.98	51.23	24.48	45.63
Mask nouns	68.43	49.81	24.57	44.84
Key words mining	71.31	52.34	25.14	46.39

Table 5. Ablation study of self-improving strategy.

Method	R1@			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
w/o pseudo labels	37.43	26.60	11.43	25.62
w/o inter-sample contrast	66.78	48.01	23.53	43.73
Full	71.31	52.34	25.14	46.39

Table 6. Ablation study of query-enhanced proposal interaction.

Method	R1@			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
w/o enhanced	69.54	49.78	22.42	45.28
Hadamard product	69.60	50.41	22.55	45.04
NLB	68.97	49.49	23.72	45.24
product+concat	71.31	52.34	25.14	46.39

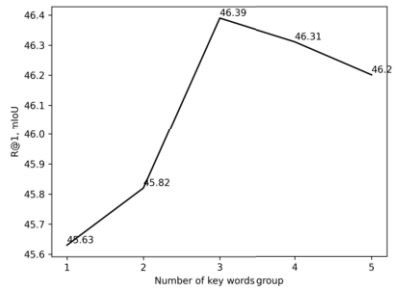

Figure 5. Ablation study of number of groups of candidate masked words.

4.5.1 *Validation of discrepancy in alignment.*

We experimentally show how the masked words affect temporal sentence grounding in the reconstruction-based model. We choose the pre-trained CPL [33] as the reconstruction-based model used in this experiment and validate on two popular datasets Charades-STA and ActivityNet Captions.

As illustrated in Figure 3, for every video V in the test set, we select query pairs $< q_i, q_j >$ from its query set Q satisfying: (1) the semantics of q_i and q_j are different, that is $b_i \neq b_j$, where b_i represents the ground truth temporal boundary of q_i. (2) q_i and q_j share some common words, i.e. $C_{ij} = \{W_i\} \cap \{W_j\} \neq \emptyset$, where $\{W_i\}$ represents the word set of query q_i. The reconstruction-based model predicts moments and ranks them according to the reconstruction errors. During query reconstruction, *same words* C_{ij} are masked for every query pair $< q_i, q_j >$ to observe the effect of masked words. For every query pairs $< q_i, q_j >$ and corresponding predicted moments $< \hat{b}_i, \hat{b}_j >$, the IoU of ground truths $IoU(b_i, b_j)$ and predicted moments $IoU(\hat{b}_i, \hat{b}_j)$ are collected and the mean IoU are calculated, as shown in Figure 4.

In Figure 4, we can observe that the mIoU of ground truth moments corresponding to the query pairs are relatively low on the two datasets, which means the semantics of the two queries are different. However, when the masked words are the same between two queries during query reconstruction, the reconstruction-based model tends to give highly overlapped predictions (higher mIoU), compared with the random mask. This phenomenon shows that the query reconstruction task focuses on mining the elements in visual modality that correspond to the masked words, failing to perform language grounding from the global semantics of the query. There is a discrepancy in alignment built by query reconstruction errors and the alignment expected by temporal sentence grounding. The mIoU of SQRT's predictions is closer to the mIoU of ground truths. This indicates that SQRT is better at grounding according to the overall semantics of the query. But the mIoU is still larger than the ground truth's, especially on ActivityNet Captions. We infer that this may be due to the tendency of Gaussian masks to predict long moments on ActivityNet Captions and the greater complexity of queries in ActivityNet Captions.

4.5.2 *Effectiveness of key words mining strategy.*

We evaluate the effectiveness of our proposed key words mining strategy by adopting a different masking approach in the reconstruction model of SQRT on the Charades-STA dataset.

As shown in Table 4, we adopt a different masking approach in reconstruction model in SQRT. *Random mask* means randomly masked

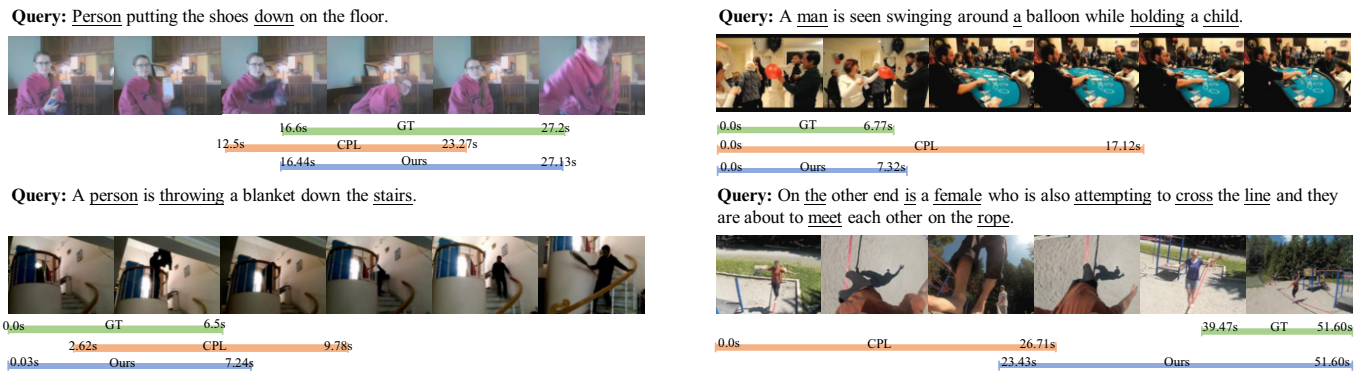

Figure 6. Qualitative examples of our method. Examples on the left are from Charades-STA. Examples on the right are from the ActivityNet Captions dataset. The masked words in CPL are underlined.

one-third of the words in query sentences, nouns, and verbs are masked with a higher probability. *Mask nouns* means randomly mask nouns of the query sentence. We observe that *Random mask* achieves higher mIoU than *Mask nouns*, which indicates that not all nouns are helpful to grounding. Our key words mining strategy achieves the best among the three. This indicates some key words contribute more to distinguish the target moment from others. It is beneficial to build correct video-language correlation during model training.

We also evaluate the effect of the number of groups of candidate masked words in key words mining on Charades-STA. As shown in Figure 5, there is performance gain when the number of groups of key words increases from 1 to 3. The computation cost grows higher as the number rises to 4 and the gain starts to shrink. Note that when the number equals 1 it equals the random mask approach.

4.5.3 *Effectiveness of self-improving.*

We evaluate the effect of our proposed self-improving scheme by ablating the grounding branch and grounding-oriented self-improving objective.

We build a new testing split based on the testing set of ActivityNet Captions, which requires a good comprehension of the overall semantics of the query for better grounding. Specifically, we select queries with complex semantics according to pre-defined rules to form the new split. More details about the split are shown in our supplemental material. We evaluate the performance of reconstruction-based models and our SQRT on this split, as seen in Table 3. *Rec-only* is the baseline of our Rec-branch. *Rec-grounding* is the grounding model trained with the video features produced by Rec-branch. We can see that SQRT performs worst in terms of mIoU while our model performs best on the challenging split. This demonstrates that the video features captured by the reconstruction-based model are more suitable for text generation rather than accurate grounding, and the pure reconstruction-based model performs worse when grounding from the overall semantics of the complex query.

We ablate the grounding-oriented self-improving objective on Charades-STA in several settings, as Table 5 shows. We can see that the performance drops significantly without learning from reconstruction results. *w/o inter* means training the grounding branch without inter-sample contrast loss. This setting means the grounding branch purely learns from the pseudo labels generated from reconstruction results, leading to inferior performance. The full model performs best compared with these settings, which indicates the importance of self-improving on self-generated pseudo labels and leveraging video-level labels from inter-sample contrast.

4.5.4 *Effectiveness of query-enhanced proposal interaction.*

We evaluate the effectiveness of the proposed query-enhanced proposal interaction module on Charades-STA by replacing it with several proposal-query fusion modules, as shown in Table 6. *w/o enhanced* means predicting proposal scores without proposal-query interaction. We offer several proposal-query interaction variations including Hadamard product, Non-Local-Block, and the proposed interaction in our SQRT. We can see that performing high-level multimodal interaction and injecting temporal information are beneficial for acquiring query semantics and context modeling thus achieving better grounding, and the *product+concatenation* in our SQRT performs best among several interaction variations.

4.6 *Qualitative Results*

We present some qualitative examples in Figure 6. The masked words in the query reconstruction-based model (CPL) are underlined. As shown in Figure 6, our method achieves more accurate grounding results, whereas CPL focuses on retrieving the segment containing missing elements corresponding to the masked words. This demonstrates the effectiveness of our self-improving scheme. In the second example of Figure 6, our method performs better in the face of a complex query, which means our method performs grounding from the overall semantics of the query. It indicates the effectiveness of the self-improving scheme and query-enhanced proposal interaction.

5 Conclusion

We reveal the flaws that undermine the reliability of reconstruction-based methods in WTSG and propose a novel Self-improving Query ReconstrucTion (SQRT) framework for WTSG. To deal with insignificant reconstruction, we propose a key words mining strategy. To attain better moment-query alignment, SQRT conducts self-improving by intra-sample contrast from pseudo labels generated from reconstruction results and inter-sample contrast among different video-text pairs. In the future, we will seek to extend the idea to other challenging weakly supervised tasks such as spatial-temporal grounding.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants (62171057, 62201072, 62071067, 62001054), in part by the Ministry of Education and China Mobile Joint Fund (MCM20200202), Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

References

[1] 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*, pp. 6299–6308, (2017).

[2] Shaoxiang Chen and Yu-Gang Jiang, 'Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8425–8435, (2021).

[3] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong, 'Look closer to ground better: Weakly-supervised temporal grounding of sentence in video', *arXiv preprint arXiv:2001.09308*, (2020).

[4] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang, 'Weakly supervised dense event captioning in videos', *Advances in Neural Information Processing Systems*, **31**, (2018).

[5] Zhiyuan Fang, Shu Kong, Zhe Wang, Charless Fowlkes, and Yezhou Yang, 'Weak supervision and referring attention for temporal-textual association learning', *arXiv preprint arXiv:2006.11747*, (2020).

[6] 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*, pp. 5267–5275, (2017).

[7] Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong, 'Wslln: Weakly supervised natural language localization networks', *arXiv preprint arXiv:1909.00239*, (2019).

[8] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao, 'Query-aware video encoder for video moment retrieval', *Neurocomputing*, **483**, 72–86, (2022).

[9] 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*, pp. 706–715, (2017).

[10] Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham, 'Pmi-masking: Principled masking of correlated spans', *arXiv preprint arXiv:2010.01825*, (2020).

[11] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu, 'Weakly-supervised video moment retrieval via semantic completion network', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11539–11546, (2020).

[12] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo, 'Vlanet: Video-language alignment network for weakly-supervised video moment retrieval', in *European conference on computer vision*, pp. 156–171. Springer, (2020).

[13] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury, 'Weakly supervised video moment retrieval from text queries', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11592–11601, (2019).

[14] Jonghwan Mun, Minsu Cho, and Bohyung Han, 'Local-global video-text interactions for temporal grounding', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10810–10819, (2020).

[15] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä, 'Uncovering hidden challenges in query-based video moment retrieval', *arXiv preprint arXiv:2009.00325*, (2020).

[16] 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)*, pp. 1532–1543, (2014).

[17] Nafis Sadeq, Canwen Xu, and Julian McAuley, 'Informask: Unsupervised informative masking for language model pretraining', *arXiv preprint arXiv:2210.11771*, (2022).

[18] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu, 'Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos', *arXiv preprint arXiv:2003.07048*, (2020).

[19] 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*, pp. 1022–1032, (2022).

[20] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer, 'Logan: Latent graph co-attention network for weakly-supervised video moment retrieval', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2083–2092, (2021).

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, 'Learning spatiotemporal features with 3d convolutional networks', in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, (2015).

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', *Advances in neural information processing systems*, **30**, (2017).

[23] Gongmian Wang, Xing Xu, Fumin Shen, Huimin Lu, Yanli Ji, and Heng Tao Shen, 'Cross-modal dynamic networks for video moment retrieval with text query', *IEEE Transactions on Multimedia*, **24**, 1221–1232, (2022).

[24] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li, 'Weakly supervised temporal adjacent network for language grounding', *IEEE Transactions on Multimedia*, (2021).

[25] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang, 'Visual co-occurrence alignment learning for weakly-supervised video moment retrieval', in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1459–1468, (2021).

[26] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu, 'Negative sample matters: A renaissance of metric learning for temporal grounding', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2613–2623, (2022).

[27] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen, 'Should you mask 15% in masked language modeling?', *arXiv preprint arXiv:2202.08005*, (2022).

[28] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin, 'Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos', in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1283–1291, (2020).

[29] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu, 'Local correspondence network for weakly supervised temporal sentence grounding', *IEEE Transactions on Image Processing*, **30**, 3252–3262, (2021).

[30] 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, pp. 12870–12877, (2020).

[31] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He, 'Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos', in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4098–4106, (2020).

[32] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al., 'Counterfactual contrastive learning for weakly-supervised vision-language grounding', *Advances in Neural Information Processing Systems*, **33**, 18123–18134, (2020).

[33] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu, 'Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15555–15564, (2022).