Semantic Collaborative Re-Mining for Temporal Sentence Grounding with Relevance Feedback

Xianzheng Liu[1,2,3], Mingyao Zhou[4], Hao Sun[1,2,3,†], Wei Xie[1,2,3,†]
[1]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
Central China Normal University, Wuhan, China
[2]School of Computer Science, Central China Normal University, Wuhan, China
[3]Academy of Frontier Interdisciplinary Research, Central China Normal University, Wuhan 430079, China
[4]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan, 430072, China
xianzhengliu@mails.ccnu.edu.cn, zhoumingyao@whu.edu.cn, haosun@ccnu.edu.cn, XW@mail.ccnu.edu.cn

***Abstract*—Temporal sentence grounding with relevance feedback (TSG-RF) evaluates the relevance of a query to a video. If relevant, it predicts related segments, otherwise, it determines that the query is irrelevant. Existing methods usually simply use two separate modules for evaluation and localization after query and video feature extraction. However, these methods fail to suppress interference from irrelevant video content, which often causes erroneous relevance judgments for given queries due to neglected semantic synergy. To leverage the intrinsic connections, we propose a semantic collaborative re-mining (SCR) method. Firstly, a prediction module was built to generate an initial segment containing semantic information of the module. Secondly, an enhancement strategy is designed to reduce query-irrelevant information in video features through the initial segment, generating enhanced visual features for modality interaction. Finally, a multi-scale relevance evaluation module is built to facilitate module interaction through enhanced visual features, optimizing the final feedback results. Experimental results on Charades-RF and Activity-RF datasets demonstrate that SCR alleviates the limitations of existing methods on TSG-RF.**

***Index Terms*—Temporal Sentence Grounding, Relevance Feedback, Semantic Collaborative Re-mining**

I. Introduction

The objective of temporal sentence grounding (TSG) in videos is to retrieve the most pertinent video clips in response to users' natural language queries. This task requires the integration of computer vision and natural language understanding technologies. TSG is a crucial yet challenging task for its potential applications in video surveillance [1], [2] and robotic services [3], [4].

However, in real-world scenarios, the video may contain no content corresponding to user queries, leading to inaccurate or erroneous localization. Temporal Sentence Grounding with Relevance Feedback (TSG-RF) addresses this limitation by explicitly identifying and handling irrelevant queries, thus gaining increasing attention for its practical applicability.

Existing relevance feedback solutions [5] typically first use a discrimination module to predict the relevance of queries. If relevant, it generates a predicted segment through a prediction module. Otherwise, it provides irrelevant feedback. These methods typically focus on improving the discriminative ability of the evaluation module and the accuracy of the prediction module. The extracted text and video features are directly input into the evaluation and prediction module, outputting the final results. However, the mutual relationship between the two modules is ignored.

To exploit the mutual relationship, we propose a semantic collaborative re-mining (SCR) method. Similar to the previous method [5], we preprocess the extracted features to generate text-guided enhanced video features. First, we design a semantic association locator to predict segments. This module processes text-guided enhanced video features and outputs initial segments aligned with the query. Second, to achieve semantic interaction between prediction and evaluation modules, we introduce a semantic association enhancement strategy. Specifically, we apply weight decay to frames outside the initial segments. The obtained semantic association enhanced video features are passed to the evaluation module. This sharing mechanism utilizes information provided by the locator, guiding the evaluation module to focus on query-related content, thereby achieving implicit communication between the modules. Finally, we develop a multi-scale relevance evaluator to fully leverage the guidance information from the locator. This module processes the semantic association enhanced video features, further focuses on query-related information through a filter, and ultimately outputs feedback results at both frame level and video level.

In summary, the main contributions of this paper are as follows:

- We propose an SCR network, which employs the inherent relationship between prediction and evaluation modules to enhance the relevance feedback of the final output.
- Semantic association enhancement strategy guides the evaluator to focus on query-related content through a sharing mechanism, achieving semantic interaction between the two modules. This helps to optimize the evaluation of relevance.
- We introduce a multi-scale relevance evaluator with a filter to capture query-related content, and give relevance feed-

† Corresponding authors.

back by integrating the frame scale and video scale.
The source code and supporting data for this paper are available at https://github.com/SINLiu11/SCR.

II. Related Work

A. Temporal Sentence Grounding in Videos

In the traditional TSG setting, it is always assumed that each given query content has a matching segment in the video. TCSF [6] explores the relationship between adjacent frames by aggregating three low-level visual features and residuals. LCNnet [7] explores fine-grained relationships in query-video pairs through self-supervised learning, ultimately improving parsing capabilities for complex temporal scenarios. SCAnet [8] matches text features and video features at multi-granularity, achieving cross-modal alignment by integrating local word-phrase semantics with global sentence-clip contexts. CPL [9] and CNM [10] generate possible proposals through a learnable Gaussian function and use irrelevant clips in the same video as negative samples for comparative learning. However, the assumption that the query content exists in the video limits their applicability in real-world scenarios.

III. Method

A. Overall Framework

Figure 1 shows the SCR network framework we designed for TSG-RF. First, for a given uncut video and query text, after extracting their features, the semantic association locator uses text-guided enhanced video features to generate the starting boundary position probability distribution P_s and the ending one P_e. Then, using the semantic association enhancement strategy, the frames outside the predicted segment are down-weighted to obtain the semantic association enhanced video features. Finally, the multi-scale relevance evaluator uses the video features to give a final relevance feedback at frame scale and video scale. The following sections explain each module in detail.

B. Feature Extraction of Videos and Queries

For a given text query, we extract word features and embed them using the pre-trained GloVe [11] for each word in it. For a given uncut video, we extract the video feature sequence using a pre-trained feature extraction network (such as I3D [12]). We align video and query features into a shared space of dimension d through separate fully connected layers, obtaining video features $\mathbf{V_p} \in \mathbb{R}^{N\times d}$ and text features $\mathbf{W} \in \mathbb{R}^{M\times d}$. Then, we inject word-level text clues into the video features through a cross-attention module to generate text-guided enhanced video representations $\mathbf{V_e} \in \mathbb{R}^{N\times d}$.

C. Semantic Association Locator

After obtaining the text-guided enhanced video features, we designed a semantic association locator to predict the probability distribution of the start and end boundary position indexes. Specifically, we follow [13] and use two cascaded LSTM networks to model $\mathbf{V_e}$ to predict the probability distribution of the start and end positions, respectively. That is, the start predictor first receives the input sequence while the end predictor is then modeled on the start predictor's output. Each predictor consists of an LSTM layer and two fully-connected layers, followed by softmax function to yield probability distributions over start and end index. The loss function is defined as L_{index}.

$$L_{\text{index}} = -\frac{1}{2}\left(\sum Y_s \log\left(P_s\right) + \sum Y_e \log\left(P_e\right)\right), \quad (1)$$

Where Y_s and Y_e are one-hot vector representations of the actual start (a_s) and end (a_e) boundary indices, respectively. P_s and P_e represent the predicted start and end boundary probability distributions, respectively.

D. Semantic Association Enhancement Strategy.

To exploit the semantic information provided by the locator, we adopt semantic association enhancement strategy. This strategy uses the aligned video features $\mathbf{V_p}$ to guide the scoring module to focus on relevant areas by suppressing the influence of irrelevant frames. Specifically, as shown in Figure 1, P_s and P_e are used to take the start index $\hat{a_s} = \arg\max P_s(t)$ and the end index $\hat{a_e} = \arg\max P_e(t)$ of the segment. Subsequently, the features of the video frame in $[\hat{a_s}, \hat{a_e}]$ remain unchanged, and others are multiplied by the scale coefficient μ ($\mu < 1$) for weight attenuation. Then we obtain the semantically associated enhanced video feature $\mathbf{V_s} \in \mathbb{R}^{N\times d}$. We use a self-attention network to extract a sentence-level semantic feature $\mathbf{h_w} \in \mathbb{R}^{1\times d}$. By connecting $\mathbf{V_s}$ with $\mathbf{h_w}$, we construct a joint representation $\mathbf{V_u} \in \mathbb{R}^{(N+1)\times d}$.

E. Multi-scale Relevance Evaluator

Frame-scale Relevance Evaluator. To evaluate frame-scale relevancy, we use V_u to predict a sequence of frame-scale relevance scores $S_r = sigmoid\left(FC\left(V_u\right)\right) \in \mathbb{R}^N$. To guide the evaluator to focus on query-related content, we choose to retain the top k highest-scoring frames and zero the rest. Subsequently, following [14]–[17], we use max pooling to compute the query-frame correlation score $score_f = max\left(S_r\right)$, where a higher score indicates at least one relevant frame, and a lower score suggests none.

Video-scale Relevance Evaluator. We perform a frame-level weighted sum operation on V_s to obtain a video-scale representation $h_v \in \mathbb{R}^d$. Subsequently, we fuse the sentence-level text feature h_w with the global video feature h_v to generate a video-level relation signal vector $g = FC\left(\left[h_w, h_v\right]\right) \in \mathbb{R}^d$. This vector is used to compute the video-level correlation score $score_v = sigmoid\left(FC\left(g\right)\right)$.

Multi-scale relevance prediction. Following [5], we calculate the average of the frame-scale and video-scale relevance scores as the final multi-scale relevance score P_v.

F. Model Training and Inference

We apply binary cross-entropy loss L_{frame} for each frame and L_{video} for each query-video pair. In summary, the total training loss can be defined as:

$$L_{\text{total}} = L_{\text{index}} + \beta L_{\text{frame}} + \gamma L_{\text{video}}. \quad (2)$$

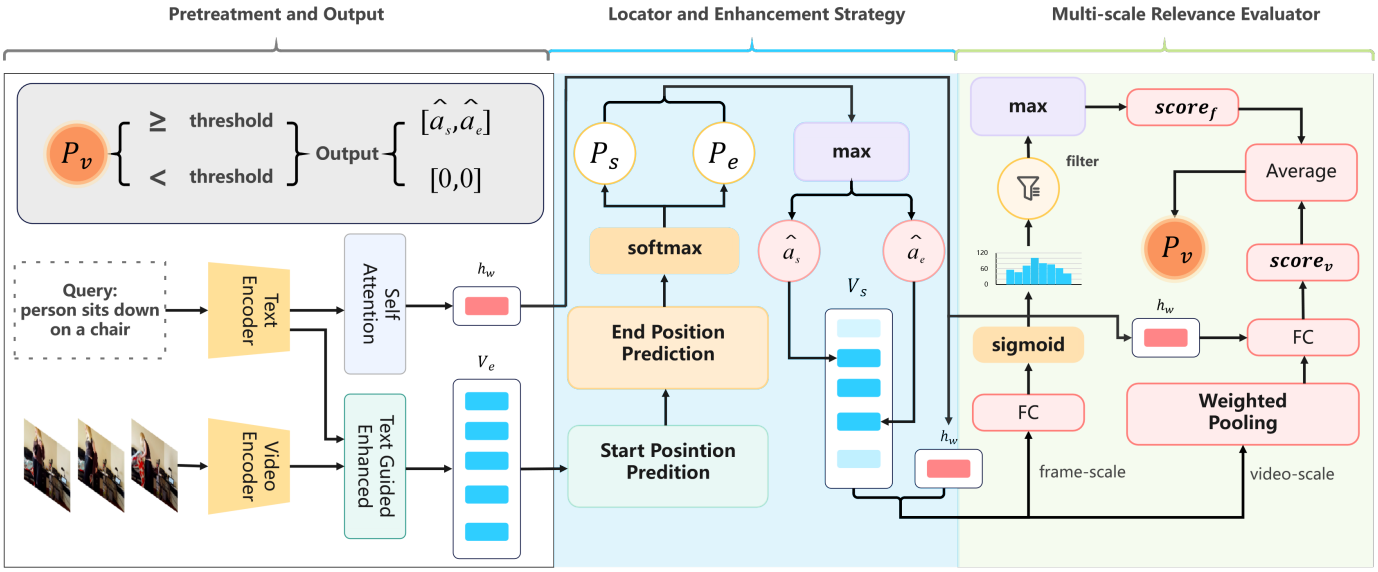


Fig. 1: SCR consists of several key steps. First, a frozen feature extractor is used to extract query and video features. Subsequently, semantic association locator processes text-enhanced video features. Finally, semantic association enhancement strategy passes this semantic information to the multi-scale relevance evaluator, and the evaluator gives scores. In addition, the faded blue squares below V_s in the figure represent the frame features after multiplying by the attenuation coefficient μ.

In the inference phase, we first calculate the joint probability distribution based on the start and end boundary probability distributions P_s, P_e. Second, the model predicts starting $\hat{a_s}$ and ending $\hat{a_e}$ boundary indexes of the target segment and adopts a semantic association enhancement strategy to output a multi-scale relevance score P_v. P_v is compared with the threshold n to determine whether there is query-related content in a given video:

$$\begin{cases} \text{Relevance query}, & \text{if } P_v \geq n \\ \text{Irrelevant query}, & \text{if } P_v < n \end{cases}. \tag{3}$$

Finally, the model outputs the target segment boundaries $[\hat{a_s}, \hat{a_e}]$ for relevant queries, and [0, 0] for irrelevant queries.

IV. Evaluation

A. Datasets and Metric

1) Datasets Set: To test the effectiveness of SCR, we conducted experiments on two publicly available datasets, ActivityNet-RF [5] and Charades-RF [5]. Charades-RF contains 12408 query-video pairs, and the test set and validation set contain 3720 query-video pairs, respectively. ActivityNet-RF contains 19290 videos, which are open and diverse in content. The training set contains 37421 query-video pairs, and the validation set and the test set contain 17505 and 17031 query-video pairs, respectively. The ratio of sample pairs with and without grounding results in the test and validation sets of these two datasets is 1:1.

2) Evaluation Metric: Following [13], [17]–[21], We use the Rn@m and mIoU metrics commonly used in TSG research to measure alignment ability, where Rn@m represents the proportion of queries in which at least one of the first n predicted segments has an IoU greater than m with the actual segment, and mIoU is the average IoU value of all test samples. The accuracy (Acc) is additionally used to evaluate the relevance judgment ability. It is worth noting that since samples without alignment results need to be considered, following [5], we redefine the IoU calculation. A time metric was introduced, defined as the approximate time required to train for 100 epochs.

B. Implementation Details

Each word in the query text is embedded with GloVe300d. The visual features of the video are extracted by a pre-trained I3D network, and the maximum length of the video feature sequence is set to 128. Sequences longer than 128 are uniformly downsampled to 128, otherwise, they are padded with zeros to the same length. During training, randomly selected videos are paired with the original query text to create samples without real results. In Equation 2, we set β to 6 and γ to 6. To achieve good performance on the validation set, the threshold n in Equation 3 is set to 0.5 on Charades-RF and 0.3 on ActivityNet-RF. For k in the filter and for μ in the semantic association enhancement strategy, we set them to 0.9.

C. Comparison with Existing Methods

To our knowledge, at the time of submission, RaTSG [5] remains the only publicly available dedicated dataset and method for TSG-RF. We compare our proposed model with recent models for TSG-RF and traditional models for TSG. Specifically, we selected six models with released source code, among which RaTSG [5] is the model for TSG-RF, VSLNet [13], SeqPAN [22], ADPN [23], UniVTG [24], and QD-DETR [25] are the models for traditional TSG. Following [5], we employ a separately trained binary classifier to evaluate

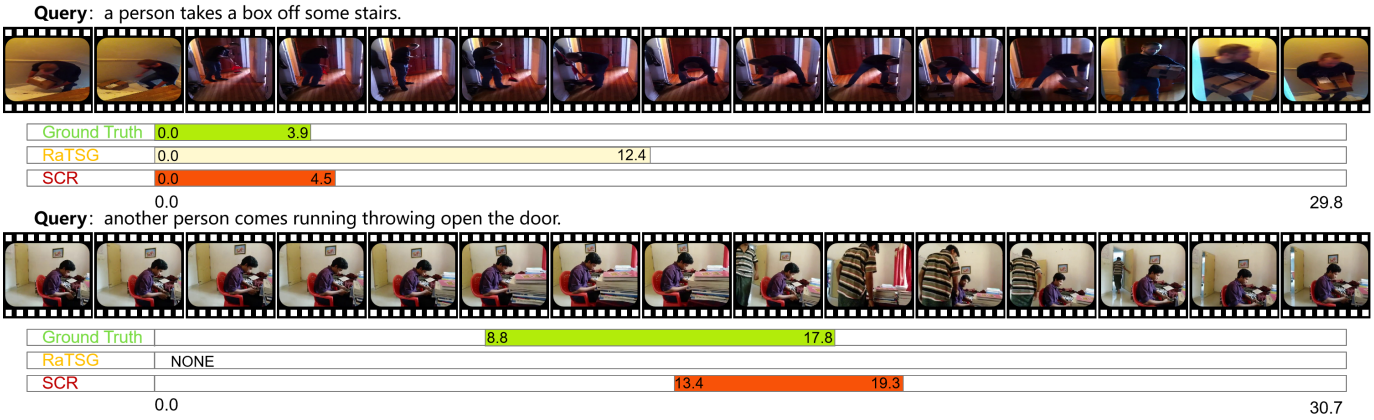


Fig. 2: Qualitative results of SCR on Charades-RF dataset.

TABLE I: Performance comparison on Charades-RF dataset.

Method	Charades-RF					
	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU	time(h)
VSLNet+	71.94	61.40	56.77	49.65	54.67	-
UniVTG+	71.94	62.58	58.55	48.79	54.65	-
QD-DETR+	71.94	62.18	58.20	50.96	55.13	-
ADPN+	71.94	62.26	57.23	51.16	55.41	-
SeqPAN+	71.94	62.12	58.01	51.61	55.49	-
RaTSG	75.30	67.37	61.67	54.62	60.34	**0.36**
SCR (ours)	**76.02**	**68.92**	**63.36**	**56.51**	**61.52**	0.42

TABLE II: Performance comparison on ActivityNet-RF dataset.

Method	ActivityNet-RF					
	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU	time(h)
VSLNet+	81.60	66.15	58.37	50.64	58.65	-
UniVTG+	81.60	66.15	58.36	49.46	58.00	-
QD-DETR+	81.60	62.43	56.13	49.27	55.97	-
ADPN+	81.60	65.85	57.41	50.28	58.47	-
SeqPAN+	81.60	66.77	58.98	51.11	59.11	-
RaTSG	**83.96**	68.26	59.74	52.10	60.44	**2.13**
SCR (ours)	83.21	**68.57**	**60.75**	**53.45**	**61.14**	2.29

query-video relevance, denoting the augmented versions of traditional models as VSLNet(+), SeqPAN(+), ADPN(+), UniVTG(+), and QD-DETR(+). To guarantee reliability, we compared our method against RaTSG under uniform configurations. Secondary comparisons utilized established results from existing literature. Table I and Table II show the performance and model parameter comparison of our model and previous models on Charades-RF and ActivityNet-RF datasets. After introducing semantic association enhancement operations, the SCR inference speed is slightly slower than the baseline method, but the difference is within an acceptable range. Besides, We can observe that: (1) Our method achieves the best results in all indicators on Charades-RF, which demonstrates the effectiveness of our model in dealing with TSG-RF. (2) Our method outperforms the existing state-of-the-art methods in recall and mIoU on ActivityNet-RF. However, there is still a certain gap between SCR and the state-of-the-art methods in terms of relevance prediction accuracy. One possible reason is that our method pays more attention to relevant frames, resulting in irrelevant queries being mistakenly regarded as relevant.

D. Ablation Study

Effectiveness of Multi-scale Correlation Evaluator. To verify the effectiveness of the multi-scale correlation evaluator, we compare it with downgraded models that only use frame-scale or video-scale evaluation scores. As shown in Table III, the single-scale model cannot perceive the different degrees of partial relevance between text and video, and thus performs worse than the multi-scale one.

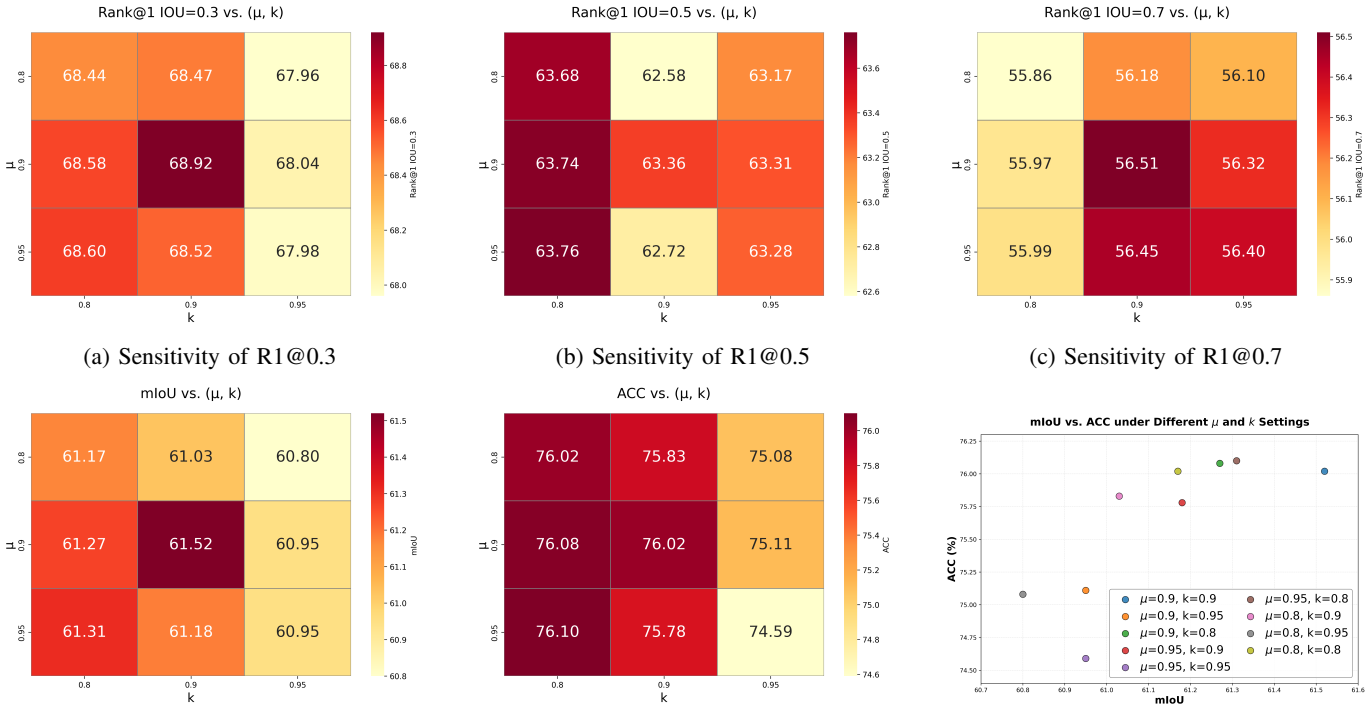
TABLE III: Effectiveness of multi-scale correlation evaluator on Charades-RF

Multi-scale		**Acc**	**R1@0.3**	**R1@0.5**	**R1@0.7**	**mIoU**
Frame	**Video**					
×	✓	73.12	67.12	62.55	**57.12**	61.04
✓	×	74.78	66.91	60.67	53.41	59.07
✓	✓	**76.02**	**68.91**	**63.36**	56.51	**61.52**

Effectiveness of Semantic Association Enhancement Strategy and the filter. As shown in Table IV, the model using the semantic association enhancement strategy and the filter outperforms the control on many metrics. The results show that using this strategy is beneficial because it forces the model to pay more attention to related frames, thereby giving higher confidence relevance scores. The control group over-focuses on irrelevant frames, resulting in a low confidence relevance score in the final output. The experimental results show that the filter we set is simple yet effective.

E. Hyperparameter Analysis on Charades-RF

As shown in Figure 3(a-d), the performance of IoU and mIoU varies under different μ and k parameter settings. A



(a) Sensitivity of R1@0.3 (b) Sensitivity of R1@0.5 (c) Sensitivity of R1@0.7

(d) Sensitivity of mIoU (e) Sensitivity of ACC (f) Sensitivity of mIoU and ACC

Fig. 3: Hyperparameter Sensitivity Analysis on Charades-RF

TABLE IV: Effectiveness of Filter and Semantic Association Enhancement Strategy on Charades-RF

Component	Status	Metrics				
		ACC	R1@0.3	R1@0.5	R1@0.7	mIoU
Filter	×	74.46	68.01	62.77	**56.64**	61.12
	✓	**76.02**	**68.91**	**63.36**	56.51	**61.52**
Enhancement	×	75.86	68.63	62.82	**56.53**	61.27
	✓	**76.02**	**68.91**	**63.36**	56.51	**61.52**

comprehensive comparison of all indicators shows that the parameter combination $\mu = 0.9$ and $k = 0.9$ exhibits the best balance across all evaluations, with mIoU reaching its highest value (61.52). As shown in Figure 3(e), the model achieves the highest classification accuracy (76.10) when parameters $\mu = 0.95$ and $k = 0.8$. Within a certain parameter range, ACC shows a significant upward trend as k decreases and μ increases. We hypothesize that retaining frames with higher relevance to the video content can effectively enhance the model's discriminative ability. Assigning greater weight to the retained frames can further reduce the interference of low-relevance frames on the model's judgment, thereby improving overall performance. Figure 3(f) shows the distribution relationship of mIoU and ACC under different combinations of μ and k parameters through a scatter plot. Most experimental points are concentrated in the range of mIoU between 61.0 and 61.2 and ACC between 75.5 and 76.0. Considering the overall performance of mIoU and ACC, we finally selected $\mu = 0.9$ and $k = 0.9$ as the hyperparameters of the model.

F. Qualitative Results

Figure 2 shows two qualitative examples on Charades-RF dataset. Each example includes different queries and provides their respective temporal boundaries for ground truth. None indicates the query is irrelevant. In Figure 2, compared to RaTSG, SCR demonstrates more accurate and reasonable results in terms of correlation judgment and positioning accuracy. We consider that these performance improvements are attributed to the combination of our proposed modules.

V. Conclusions

In this work, we propose a novel method for TSG-RF, termed semantic collaborative re-mining (SCR). The proposed framework introduces a multi-scale correlation evaluator with a filter and employs a semantic association enhancement strategy to guide the model's attention to relevant frames. This design enables the model to leverage the intrinsic connections and deliver reliable feedback. Extensive experiments and ablation studies conducted on the Charades-RF and ActivityNet-RF datasets validate the effectiveness of SCR. In the future, we will explore how to more effectively balance the relationship between relevance judgment and positioning accuracy, and extend the semantic collaboration framework to complex multimodal video understanding scenarios to solve real-world application problems more efficiently.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under Grant 62377026, in part by the Fundamental Research Funds for the Central Universities under Grant JC2026PT-003, and in part by Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning under Grant 2025AISL011.

REFERENCES

[1] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. J. Burt, and L. E. Wixson, "A system for video surveillance and monitoring," 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:9132354

[2] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[3] D. Andronas, G. Apostolopoulos, N. Fourtakas, and S. Makris, "Multi-modal interfaces for natural human-robot interaction," *Procedia Manufacturing*, vol. 54, pp. 197–202, 2021, 10th CIRP Sponsored Conference on Digital Enterprise Technologies (DET 2020) – Digital Technologies as Enablers of Industrial Competitiveness and Sustainability. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2351978921001669

[4] A. C. Simões, A. Pinto, J. Santos, S. Pinheiro, and D. Romero, "Designing human-robot collaboration (hrc) workspaces in industrial settings: A systematic literature review," *Journal of Manufacturing Systems*, vol. 62, pp. 28–43, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278612521002296

[5] J. Dong, X. Peng, D. Liu, X. Qu, X. Yang, C. Bao, and M. Wang, "Temporal sentence grounding with relevance feedback in videos," *Advances in Neural Information Processing Systems*, vol. 37, pp. 43 107–43 132, 2024.

[6] X. Fang, D. Liu, P. Zhou, and G. Nan, " You Can Ground Earlier than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos ," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2023, pp. 2448–2460. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00242

[7] W. Yang, T. Zhang, Y. Zhang, and F. Wu, "Local correspondence network for weakly supervised temporal sentence grounding," *IEEE Transactions on Image Processing*, vol. 30, pp. 3252–3262, 2021.

[8] S. Yoon, G. Koo, D. Kim, and C. Yoo, "Scanet: Scene complexity aware network for weakly-supervised video moment retrieval," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13 530–13 540, 2023.

[9] M. Zheng, Y. Huang, Q. Chen, Y. Peng, and Y. Liu, " Weakly Supervised Temporal Sentence Grounding with Gaussian-based Contrastive Proposal Learning ," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2022, pp. 15 534–15 543. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01511

[10] M. Zheng, Y. Huang, Q. Chen, and Y. Liu, "Weakly supervised video moment localization with contrastive negative sample mining," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3517–3525.

[11] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162/

[12] J. Carreira and A. Zisserman, " Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset ," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2017, pp. 4724–4733. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.502

[13] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 6543–6554. [Online]. Available: https://aclanthology.org/2020.acl-main.585/

[14] J. Dong, X. Chen, M. Zhang, X. Yang, S. Chen, X. Li, and X. Wang, "Partially relevant video retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. ACM, Oct. 2022, p. 246–257. [Online]. Available: http://dx.doi.org/10.1145/3503161.3547976

[15] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370296000343

[16] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems*, vol. 10. MIT Press, 1997. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1997/file/82965d4ed8150294d4330ace00821d77-Paper.pdf

[17] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 870–12 877.

[18] Z. Lin, Z. Zhao, Z. Zhang, Q. Wang, and H. Liu, "Weakly-supervised video moment retrieval via semantic completion network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 11 539–11 546.

[19] H. Sun, M. Zhou, W. Chen, and W. Xie, "Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4998–5007.

[20] M. Zhou, W. Chen, H. Sun, W. Xie, M. Dong, and X. Lu, "Query-aware multi-scale proposal network for weakly supervised temporal sentence grounding in videos," *Knowledge-Based Systems*, vol. 304, p. 112592, 2024.

[21] M. Zhou, H. Sun, W. Xie, M. Dong, C. Wang, and M. Ye, "PC-net: Weakly supervised compositional moment retrieval via proposal-centric network," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=kQAnOaayIo

[22] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, and S. M. R. Goh, "Parallel attention network with sequence matching for video grounding," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 776–790. [Online]. Available: https://aclanthology.org/2021.findings-acl.69/

[23] H. Chen, X. Wang, X. Lan, H. Chen, X. Duan, J. Jia, and W. Zhu, "Curriculum-listener: Consistency- and complementarity-aware audio-enhanced temporal sentence grounding," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 3117–3128. [Online]. Available: https://doi.org/10.1145/3581783.3612504

[24] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou, " UniVTG: Towards Unified Video-Language Temporal Grounding ," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2023, pp. 2782–2792. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00262

[25] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, " Query -Dependent Video Representation for Moment Retrieval and Highlight Detection ," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2023, pp. 23 023–23 033. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.02205