Supplementary Materials
SHINE: Saliency-aware HIerarchical NEgative Ranking for Compositional Temporal Grounding

Zixu Cheng*[1], Yujiang Pu*[2], Shaogang Gong[1], Parisa Kordjamshidi[2], and Yu Kong[2] (*Equal Contribution)

[1] Queen Mary University of London, London, UK
{zixu.cheng, s.gong}@qmul.ac.uk
[2] Michigan State University, East Lansing, US
{puyujian, kordjams, yukong}@msu.edu

1 Supplementary Experiments

In this section, we provide more supplementary experimental results to explore: (1) different masking ratios for hard negative construction; (2) different ranking margins in coarse- and fine-grained saliency ranking loss; (3) different variants of fine-grained saliency ranking loss; (4) more details of different $\mathcal{L}_{base}$ in Moment-DETR and QD-DETR; and (5) performance on a classic temporal grounding benchmark.

1.1 Different Masking Ratios for Hard Negative Construction

In this subsection, we explore the performance when applying different masking ratios to generate hard negative queries. We report the results of QD-DETR with our methods on ActivityNet-CG in Tab. 1. From the results we can observe that: (1) Larger masking ratios can yield better performance in both Test-Trivial and Novel-Word splits. (2) The model's performance in the Novel-Composition split is more sensitive to different masking ratios. (3) 'QD-DETR + Ours' achieves the best overall performance on the three test splits with the masking ratios of 10%, 30%, and 50%, which we keep as our default settings in all experiments.

Table 1: Effect of different masking ratios on the ActivityNet-CG dataset. r_1 to r_3 denote the progressive masking ratio for hard negative construction.

r_1	r_2	r_3	Test-Trivial			Novel-Composition			Novel-Word		
			R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
0.10	0.30	0.50	43.76	**25.98**	42.86	**29.56**	**14.37**	**32.44**	27.60	**13.11**	30.98
0.25	0.50	0.75	43.49	25.63	42.82	28.42	12.94	31.20	27.40	12.73	30.66
0.30	0.60	0.90	**44.04**	25.95	**43.13**	29.41	13.51	32.02	**28.11**	12.78	**31.30**

Table 2: Ablation of different margins in the Novel-Composition split of Charades-CG.

(a) Intra and inter margins in $\mathcal{L}_{cr}$

h_1	h_2	R1@0.5	R1@0.7	mIoU
0.2	1.0	47.91	24.35	42.76
0.5	1.0	47.04	26.03	42.29
1.0	1.0	48.63	24.35	42.98
1.0	2.0	**50.23**	**27.69**	**44.14**

(b) Relative ranking margins in $\mathcal{L}_{fr}$

m_0	m_1	m_2	m_3	R1@0.5	R1@0.7	mIoU
0.10	0.10	0.10	0.10	49.27	25.54	43.93
0.20	0.20	0.20	0.20	48.63	27.31	43.15
0.25	0.25	0.25	0.25	**50.23**	**27.69**	**44.14**
0.50	0.50	0.50	0.50	47.18	25.57	41.83
0.05	0.05	0.15	0.25	49.27	24.46	43.48
0.05	0.10	0.15	0.20	49.07	**25.86**	43.59
0.10	0.30	0.50	0.70	48.02	25.10	42.50
0.25	0.50	0.75	1.0	**50.00**	24.87	**43.78**

1.2 Different Ranking Margins in Coarse- and Fine-grained Saliency Ranking loss

We also explore the effectiveness of different margins in the coarse- and fine-grained saliency ranking loss. For the coarse-grained ranking margin, Tab. 2a shows that increasing intra-margin h_1 and inter-margin h_2 enhances the model performance, peaking when $h_1 = 1.0$ and $h_2 = 2.0$. Regarding the fine-grained ranking margins m_0 to m_3, Tab. 2b shows the highest results achieved by setting all margins in $\mathcal{L}_{fr}$ to 0.25. By choosing the appropriate margins, our method leads the model to learn nuances between different negative queries through $\mathcal{L}_{cr}$ and $\mathcal{L}_{fr}$, which enables the model to alleviate irrational saliency responses and improves the compositional generalizability. We then provide more comparative results under the setting of different fine-grained ranking margins. It can be seen that: (1) As the four margin values increase, the overall performance of the Novel-Composition split initially rises and then falls, with the best result achieved at 0.25. (2) The R1@0.7 metric is more sensitive to changes in margin values, indicating that appropriate margin values are beneficial for precise localization.

1.3 Different Variants of Fine-grained Saliency Ranking Loss

Moreover, we also explore an absolute ranking regime with incremental margins from m_0 to m_3, denoted as Eq. (1), where we fix the observations in the second through fourth constraints to be $d(S_p, S^1_{hn})$. From Tab. 3 we notice that this variant can also achieve considerable improvements, while our vanilla version still achieves optimal results, which we attribute to its finer constraints on adjacent hard negative queries to capture nuances. Therefore, we adopt the fine-grained ranking loss of relative distance as the default setting in all experiments.

$$\begin{aligned}\mathcal{L}'_{fr} = &\max(0, m_0 + d(Y, S_p) - d(S_p, S^1_{hn})) \\ &+ \max(0, m_1 + d(S_p, S^1_{hn}) - d(S_p, S^2_{hn})) \\ &+ \max(0, m_2 + d(S_p, S^1_{hn}) - d(S_p, S^3_{hn})) \\ &+ \max(0, m_3 + d(S_p, S^1_{hn}) - d(S_p, S_n)),\end{aligned} \tag{1}$$

Table 3: Absolute ranking margins in $\mathcal{L}'_{fr}$ in the Novel-Composition split of the Charades-CG Dataset.

m_0	m_1	m_2	m_3	R1@0.5	R1@0.7	mIoU
0.05	0.10	0.15	0.20	47.62	25.19	42.49
0.10	0.15	0.30	0.45	48.40	**26.67**	43.15
0.10	0.20	0.30	0.40	47.73	25.51	43.21
0.10	0.30	0.50	0.70	**49.45**	25.22	**44.02**
0.25	0.50	0.75	1.0	47.18	24.23	42.07

1.4 Details of Different $\mathcal{L}_{base}$ in Moment-DETR and QD-DETR

In our experiments, we retain all the original loss functions in our two baseline models, Moment-DETR [3] and QD-DETR [5], with the exception of the saliency loss. We replaced the original saliency loss with the proposed $\mathcal{L}_{intra}$ due to its better performance. Specifically, for moment retrieval, the $\mathcal{L}_{base}$ in Moment-DETR includes the span loss (L1 Loss + GIoU loss), as defined in Eq. (2) and combined with the classification loss (Negative Log-Likelihood loss). In addition, we borrow the negative pair loss (Eq. (3)) from QD-DETR to compute the negative saliency loss for an easy negative query.

$$\mathcal{L}_{span} = \lambda_{L1}||m - \hat{m}|| + \lambda_{GIoU}\mathcal{L}_{GIoU}(m - \hat{m}) \tag{2}$$

$$\mathcal{L}_{neg} = -\log(1 - S_{neg}) \tag{3}$$

Therefore, the $\mathcal{L}_{base}$ of moment-DETR can be formulated as follow:

$$\mathcal{L}_{base_{\text{moment-DETR}}} = \lambda_{neg}\mathcal{L}_{neg} + \sum_{i=1}^{N}[-\lambda_{\text{cls}}\log\hat{p}_{\hat{m}}(c_i) + L_{span}] \tag{4}$$

where m and $\hat{m}$ denotes the ground truth and predicted moments, respectively, N denotes the number of moment queries, and $\lambda_{neg}, \lambda_{cls}, \lambda_{L1}$ and $\lambda_{GIoU} \in \mathbb{R}$ are weights balancing the loss. In terms of QD-DETR [5], an extra contrastive ranking loss is included to learn the precisely segmented saliency levels, denoted as:

$$\mathcal{L}_{cont} = -\sum_{r=1}^{R}\log\left(\frac{\sum_{x\in X^{\text{r}}_{\text{pos}}}\exp\left(\frac{S(x)}{\tau}\right)}{\sum_{x\in(X^{r}_{\text{pos}}\cup X^{\text{r}}_{\text{neg}})}\exp\left(\frac{S(x)}{\tau}\right)}\right) \tag{5}$$

where R denotes the maximum rank value, X_r^{pos} and X_r^{neg} denotes the positive/negative set in the r^{th} iteration, and S denotes the saliency score. The overall $\mathcal{L}_{base}$ of moment-DETR can be formulated as:

$$\mathcal{L}_{base_{\text{QD-DETR}}} = \lambda_{neg}\mathcal{L}_{neg} + \lambda_{cont}\mathcal{L}_{cont} + \sum_{i=1}^{N}[-\lambda_{\text{cls}}\log\hat{p}_{\hat{m}}(c_i) + L_{span}] \tag{6}$$

where $\lambda_{cont} \in \mathbb{R}$ is the contrastive loss weights. For more details, please refer to their original papers [3, 5].

Table 4: Performance (%) of state-of-the-art methods on the Charades-STA dataset. Our results are shown in **bold**. 'RL': reinforcement learning methods. 'PB': proposal-based methods. 'PF': proposal-free methods. † indicates the results of our re-implementation using the officially released code. - indicates results are not available.

Setting	Method	Feature	R1@0.5	R1@0.7	mIoU
RL	TSP-PRL [8]	C3D	37.39	17.69	37.22
PB	2D-TAN [13]	VGG	39.70	23.31	-
	MMN [7]	VGG	47.31	27.28	-
	FVMR [2]	I3D	55.01	33.74	-
	MS-2D-TAN [12]	I3D	56.64	36.21	-
PF	VLSNet [11]	C3D	47.31	30.19	45.15
	LGI [6]	I3D	59.46	35.48	51.38
	HISA [9]	I3D	61.10	39.70	53.57
	UnLoc [10]	CLIP	60.80	38.40	-
	UniVTG [4]	SF+CLIP	60.19	38.55	52.17
	Moment-DETR [3]	SF+CLIP	53.63	31.37	-
	Moment-DETR† [3]	SF+CLIP	53.23	31.21	46.74
	Moment-DETR+Ours [3]	SF+CLIP	**56.85**	**32.96**	**48.90**
	QD-DETR [5]	I3D	50.67	31.02	-
	QD-DETR† [5]	I3D	59.22	36.72	50.50
	QD-DETR+Ours [5]	I3D	**61.72**	**40.48**	**52.55**

1.5 Performance on Charades-STA

To further validate the applicability and compatibility of our method, we conducted additional experiments on a widely used temporal grounding benchmark, *i.e.*, Charades-STA [1]. The training set contains 12,404 queries with 5,336 videos and the test set consists of 3,720 queries with 1,334 videos, which is used for evaluating IID (Independent and Identically Distributed) generalization capability.

From Tab. 4 we can observe that: (1) Our method can consistently improve the performance of two baselines, *i.e.*, Moment-DETR and QD-DETR, with 3.62% and 2.5% absolute gain in R1@0.5, respectively. (2) QD-DETR with our method achieves the new state-of-the-art results in both R1@0.5 and R1@0.7, while achieving comparable results to HISA [9] in mIoU. Notably, our reproduced results for QD-DETR† are significantly higher than those reported in their paper. Since they did not provide detailed training settings, we followed Moment-DETR's hyperparameters in addition to the learning rate, which is consistent with that in our submission.

2 More Qualitative Results

In this section, we provide more visual examples to demonstrate the effectiveness and superiority of our method.

2.1 Visualizations of Saliency Scores

First, we provide more visualization of the saliency scores in ActivityNet-CG. We observe that the existing work has difficulty recognizing hard negative queries, showing irrational saliency responses. For instance, in Fig. 1a, the hard negative query "*She jumps along the road and onto a grass pit*" is even more salient than the positive query "She runs down the track and into a sand pit". The irrational responses lead to unprecise moment localization since there is no corresponding moment of this negative query in the video. In contrast, our approach consistently improves the model's ability to distinguish between different words in positive and hard negative queries and yield hierarchical responses, thereby achieving better moment localization and compositional generalization.

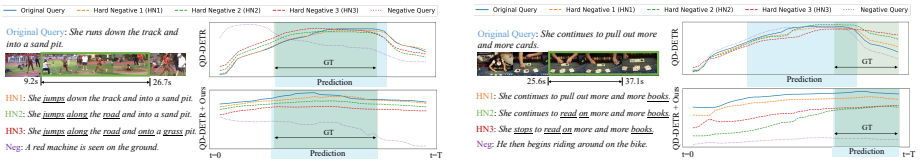


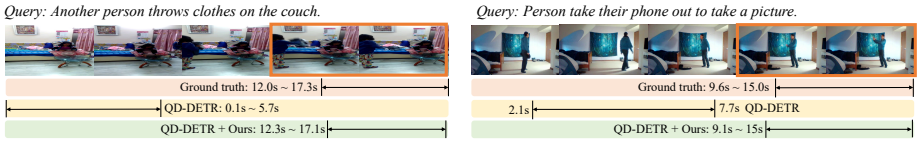
Fig. 1: Visualization of saliency scores given different query sentences in ActivityNet-CG. Existing work has difficulty in recognizing hard negative queries, showing irrational saliency responses. Our approach consistently improves the model's ability to distinguish between different primitives in a query sentence and achieves better compositional generalization.

However, in the failure case Fig. 1b, we find that our method sometimes fails to distinguish the subtle differences in the Hard Negative 2 and 3 but still responds rationally to the Positive, Hard Negative 1, and the Negative queries. We assume that our method struggles with longer videos and sentences in ActivitiNet-CG.

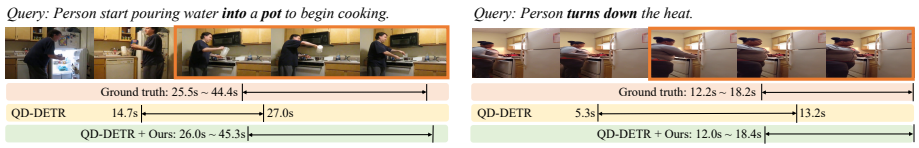
In these saliency score visualizations, we find that the gap of saliency scores derived from positive and negative queries of our method is more discriminative than that of the baseline model. We suggest that it is the inter loss $\mathcal{L}_{inter}$ in the coarse-grained ranking loss $\mathcal{L}_{cr}$ that plays a role in improving the model's ability to discriminate between positive and negative samples. Moreover, the magnitude of the saliency score within the GT is less significant than that of the baseline. We assume this is due to the larger gap between the positive and negative samples leading to less pronounced saliency score variation in the positive sample. However, this does not impair the model's ability to localize moments of positive samples accurately.

2.2 Visualizations of Temporal Grounding

We offer more visualization results of the moment predictions in Charades-CG and ActivityNet-CG in Fig. 2 and Fig. 3. Our method can enhance the exist-



(a) Two Samples from the Test-Trivial split



(b) Two Samples from the Novel-Composition split

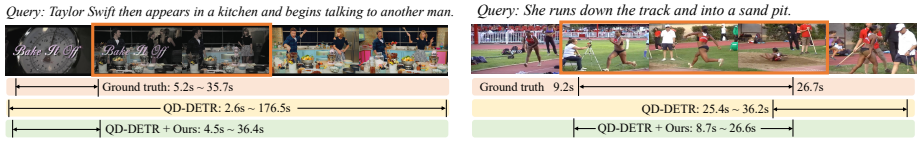
(c) Two Samples from the Novel-Word split

Fig. 2: Qualitative comparisons between QD-DETR and QD-DETR+Ours on samples from different test splits of Charades-CG.

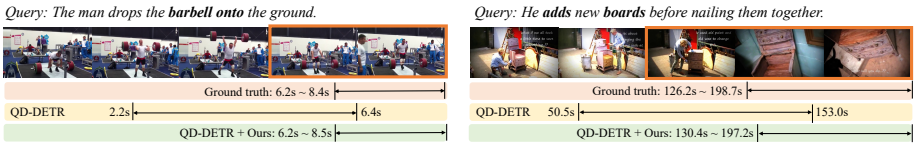
ing work to generalize to Novel-Composition and Novel-Word testing as well as predict the moment more precisely in the IID Test-Trivial split.

In the examples of the Charades-CG test-trivial split (referring to Fig. 2a), though the queries "Another person throws clothes on the couch" and "Person take their phone out to take a picture"contain no unseen compositions and words, the baseline method localizes a wrong moment while our method demonstrates more precise alignment with the ground truth. In the Novel-Composition query (Fig. 2b), "*Person start pouring water* ***into*** *a* ***pot*** *to begin cooking*" and "*Person* ***turns down*** *the heat*", our method achieves a more precise temporal localization aligned with the ground truth, indicating enhanced capability in generalizing to novel preposition-noun and verb-adverb combinations within the queries. It also demonstrates that our method improves the discrimination of the semantic meaning of prepositions and adverbs. When encountering the Novel-Word query (Fig. 2c) "*A person fixes their* ***hair*** *in a mirror*" and "*Person takes a towel from the* ***dryer***", our method can still generalize well to the novel word "hair" and "dryer".

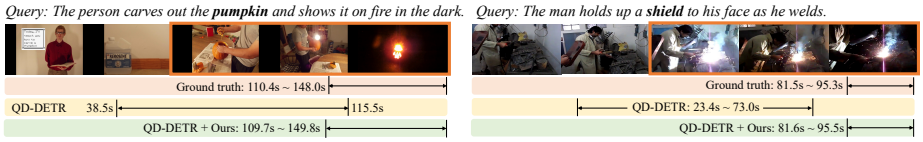
When dealing with the samples in ActivityNet-CG, our approach also shows its adaptability to longer videos and sentences. Despite the queries in Fig. 3a “Taylor Swift then appears in a kitchen and begins talking to another man” and "She runs down the track and into a sand pit" only containing known words and compositions, the baseline inaccurately identifies the relevant moment, whereas our method exhibits a more precise alignment with the ground truth. For the Novel-Composition queries (Fig. 3b) "*The man drops the* ***barbell onto*** *the ground*" and "*He* ***adds*** *new* ***boards*** *before nailing them together*",



(a) Two Samples from the Test-Trivial split



(b) Two Samples from the Novel-Composition split



(c) Two Samples from the Novel-Word split

Fig. 3: Qualitative comparisons between QD-DETR and QD-DETR+Ours on samples from different test splits of ActivityNet-CG.

our approach delivers better temporal accuracy, closely reflecting the ground truth. Upon encountering Novel-Word queries such as "*The person carves out the* ***pumpkin*** *and shows it on the fire in the dark*" and "*The man holds up a* ***shield*** *to his face as he welds*" in Fig. 3c, our method proficiently generalizes to the unseen primitives "***pumpkin***" and "***shield***".

The visualization results demonstrate that our approach successfully directs DETR-based models to leverage hierarchical negative samples, thereby improving their ability to generalize to unseen compositions and words.

3 Limitations and Future Work

Existing large language models are highly sensitive to prompt templates, and the negative queries generated by different templates may vary in effectiveness, *e.g.*, we noticed that some negative queries generated by the LLM still lack semantic feasibility and may include words that do not exist in the dictionary. It is worth exploring how to design effective prompt templates as well as data curation strategies to produce better negative queries. In addition, we only consider novel compositions in the query sentence without taking visual-level compositions into account, which might facilitate the creation of better compositional vision-language representations.

References

1. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
2. Gao, J., Xu, C.: Fast video moment retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1523–1532 (2021)
3. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems **34**, 11846–11858 (2021)
4. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2794–2804 (2023)
5. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23023–23033 (2023)
6. Mun, J., Cho, M., Han, B.: 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)
7. Wang, Z., Wang, L., Wu, T., Li, T., Wu, G.: Negative sample matters: A renaissance of metric learning for temporal grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2613–2623 (2022)
8. Wu, J., Li, G., Liu, S., Lin, L.: Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12386–12393 (2020)
9. Xu, Z., Chen, D., Wei, K., Deng, C., Xue, H.: Hisa: Hierarchically semantic associating for video temporal grounding. IEEE Transactions on Image Processing **31**, 5178–5188 (2022)
10. Yan, S., Xiong, X., Nagrani, A., Arnab, A., Wang, Z., Ge, W., Ross, D., Schmid, C.: Unloc: A unified framework for video localization tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13623–13633 (2023)
11. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Proceedings of the Annual Meeting of the Associationfor Computational Linguistics (ACL). pp. 6543–6554 (2020)
12. Zhang, S., Peng, H., Fu, J., Lu, Y., Luo, J.: Multi-scale 2d temporal adjacency networks for moment localization with natural language. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9073–9087 (2021)
13. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12870–12877 (2020)