HERO: Hierarchical Embedding-Refinement for Open-Vocabulary Temporal Sentence Grounding in Videos

Supplementary Material

Figure 1. Per-sample performance comparison between our method and the EMB baseline on Charades-OV dataset.

1. Additional Experimental Results

1.1. Result on Charades-CD

Table 1 summarizes the performance of different methods on the Charades-CD dataset. As shown, our proposed HERO achieves the best results across both evaluation metrics, reaching 51.59% at R1@0.5 and 30.84% at R1@0.7. These results outperform the second-best results, which are 49.36% and 27.76%, by absolute margins of 2.23% and 3.08%, respectively. Notably, HERO surpasses strong DETR-style models such as TR-DETR and QD-DETR, as well as prior alignment-based methods like VSLNet and LG, demonstrating its superior capability in handling distribution shifts inherent in the Charades-CD benchmark.

1.2. Result on ActivityNet Captions

Table 2 reports the performance of various methods on the ActivityNet Captions dataset. HERO achieves the best results, attaining 46.14% at R1@0.5 and 27.89% at R1@0.7, outperforming all prior approaches across both metrics. These results further validate that HERO not only excels in open-vocabulary scenarios but also delivers strong performance under standard grounding settings.

Table 1. Experimental results on the Charades-CD dataset. **Bold**/underlined fonts indicate the best/second-best results.

Method	R1@0.5	R1@0.7
2D-TAN [12]	35.88	13.91
LG [6]	42.90	19.29
DRN [10]	31.11	15.17
DCM [9]	45.57	22.70
Moment-DETR [3]	42.31	18.31
VSLNet [11]	44.41	24.83
EMB [2]	48.00	27.76
HLTI [1]	46.67	27.08
QD-DETR [5]	49.19	22.04
TR-DETR [7]	49.36	24.36
HERO (Ours)	**51.59**	**30.84**

Table 2. Experimental results on the ActivityNet Captions dataset. **Bold**/underlined fonts indicate the best/second-best results.

Method	R1@0.5	R1@0.7
QSPN [8]	27.70	13.60
DEBUG [4]	39.72	-
Moment-DETR [3]	36.22	21.15
VSLNet [11]	43.22	26.16
EMB [2]	44.81	26.07
QD-DETR [5]	38.11	21.47
TR-DETR [7]	33.14	17.88
HERO (Ours)	**46.14**	**27.89**

1.3. Ablation Study on λ_{RS} and λ_{CL}

Table 3 examines the effect of the relevance score loss (L_{RS}) and the contrastive learning loss (L_{CL}), with the base grounding loss L_{TSGV} fixed at a coefficient of 1. We first introduce L_{RS} while keeping other components unchanged. Results show that $\lambda_{RS} = 0.1$ achieves the best performance overall. Building upon this, we additionally introduce L_{CL} and observe that $\lambda_{CL} = 0.01$ yields the best R1@0.5, while $\lambda_{CL} = 0.1$ attains the best R1@0.7 with only a minor 0.27% decrease at R1@0.5. Therefore, we adopt $\lambda_{RS} = 0.1$ and $\lambda_{CL} = 0.1$ as our final setting, which provides a good trade-off across both evaluation thresholds.

2. Additional Visualization Results

Figure 1 presents per-sample qualitative comparisons between our HERO and the EMB [2] baseline on the

Table 3. Performance comparison across different values of λ_{RS} and λ_{CL}. **Bold** and underlined numbers indicate the best and second-best results, respectively.

Value	λ_{RS}		λ_{CL}	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7
0.01	42.98	25.07	**45.78**	23.73
0.1	**45.21**	**26.83**	45.51	**27.2**
1	44.23	25.78	44.65	26.59
10	44.95	25.63	45.6	26.34

Charades-OV dataset. The results illustrate the performance degradation of EMB under open-vocabulary conditions, often leading to incorrect grounding. In contrast, HERO consistently provides accurate localization, demonstrating enhanced robustness and generalization to novel linguistic expressions.

References

[1] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: a novel training framework for temporal grounding. In *Proceedings of European Conference on Computer Vision*, pages 130–147, 2022. 1

[2] Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. Video activity localisation with uncertainties in temporal boundary. In *Proceedings of European Conference on Computer Vision*, pages 724–740, 2022. 1

[3] Mohit Bansal Jie Lei, Tamara L Berg. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 1

[4] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5144–5153, 2019. 1

[5] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023. 1

[6] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 1

[7] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4998–5007, 2024. 1

[8] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9062–9069, 2019. 1

[9] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information retrieval*, pages 1–10, 2021. 1

[10] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 1

[11] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, and R. S. M. Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[12] 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*, pages 12870–12877, 2020. 1