Supplement for Vid-Group: Temporal Video Grounding Pretraining from Unlabeled Videos in the Wild

Peijun Bao[1], Chenqi Kong[1], Siyuan Yang[1], Zihao Shao[2],
Xinghao Jiang[3], Boon Poh Ng[1], Meng Hwa Er[1], Alex Kot[1,4],
[1]Nanyang Technological University [2]Peking University
[3]Shanghai Jiaotong University [4]Shenzhen MSU-BIT University
peijun001@e.ntu.edu.sg chenqi.kong@ntu.edu.sg

1. Additional Ablation Studies

Impact of cleaning ratio. As the Vid-Group dataset is collected in a scalable way with minimal human intervention, it inevitably includes errors such as idle videos and mismatched video-query pairs. To address this, the ReCorrect algorithm incorporates a label-cleaning module in the semantics-guided refinement phase, where the cleaning ratio determines the percentage of data samples filtered out. Fig. 1 illustrates the impact of the cleaning ratio on zero-shot performance. Performance enhances as the cleaning ratio increases from 0% to 30%, as samples with pseudo-label errors are removed. However, performance declines when the cleaning ratio exceeds 50%. The curve also indicates that a cleaning ratio between 20% and 40% yields satisfactory results.

The influence of λ. The hyperparameter λ defines the weight of the loss term that balances the two components of the pretraining loss. Fig. 2 illustrates the effect of varying λ on performance metrics for temporal video grounding. As λ increases, the overall metrics improve because higher weights are assigned to pseudo temporal boundaries derived from the memory consensus correction. Satisfactory performance is achieved when λ is between 0.7 and 0.9, as these values balance the loss terms and encourage the exploration of diverse pseudo-label sources, yielding slightly better metrics than $\lambda = 1.0$.

The impact of α_1 and α_2. The hyperparameters α_1 and α_2 determine the magnitude of shrinking and expanding temporal boundaries in semantics-guided refinement. Fig. 3 summarizes the impact of α_1 and α_2 on zero-shot inference performance, represented by Avg R@m, the average value of R@m for $m = \{0.3, 0.5, 0.7\}$. The left plot in Fig. 3 illustrates that the performance peaks when α_1 around 0.24. The right plot reveals that as α_2 increases, the performance first improves gradually before declining. A range of 0.88 to 0.92 for α_2 achieves consistently strong results.

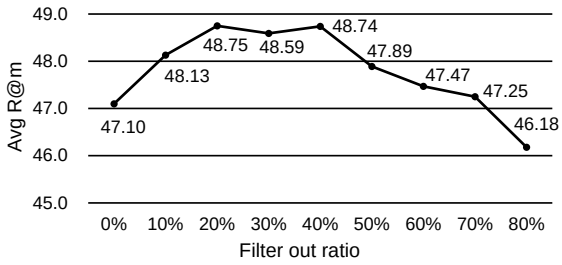


Figure 1. Ablation studies on the cleaning ratio.

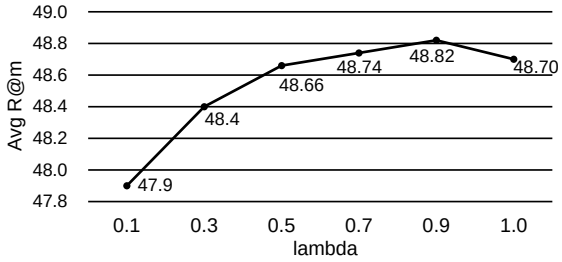


Figure 2. The influence of λ.

2. Vid-Group Dataset

Texual Prompts to Generate Pseudo Labels. When generating pseudo labels for Vid-Group dataset, each video is first uniformly sampled into n_{v2f} frames, which are then concatenated into a single image. The prompts illustrated in Fig. 4 are utilized for GPT-4o to generate pseudo labels for the video, with n_{v2f} set to 8. Regular expressions are then applied to extract the pseudo labels from GPT-4o's output, structuring them into sentence-boundary pairs. Finally, the extracted temporal boundaries, initially represented as frame indices, are converted into start and end timestamps based on each video's duration.

Additional Training Data Samples Visualization. Fig. 5

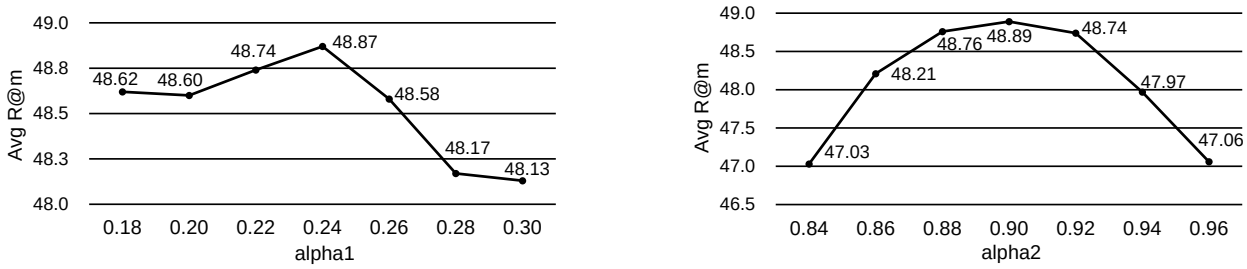


Figure 3. The impact of the hyperparameters α_1 and α_2.

Given a video evenly sampled into [n_v2f] image frames:
1. Analyze the events and human actions within these frames and partition them into segments based on the actions observed. Merge the frames with similar actions into the same segment. Provide a detailed description of the human actions and any related objects.
2. Discard any segments that do not contain explicit actions or where the action is uncertain.
Please output the results in json format. The json structure should be (start, end, description), where 'start' is the start frame, 'end' is the end frame of the segment.

Figure 4. The prompt that guides GPT-4o to generate pseudo labels. We then use regular expressions to extract the pseudo labels from GPT-4o's output, and finally convert the temporal boundaries from frame indices to timepoint format.

Table 1. Performance comparison of state-of-the-art methods in fully-supervised settings.

Method	Charades STA				ActivityNet Captions			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
UnLoc [4] ICCV23	–	60.80	38.40	–	–	48.00	30.20	–
MESM [3] AAAI24	–	61.24	38.04	–	–	–	–	–
BAM-DETR [2] ECCV24	72.93	59.95	39.38	52.33	–	–	–	–
SimBase [1] arxiv24	77.77	66.48	44.01	56.15	63.98	49.35	30.48	47.07
SimBase + GPT-4o Pretraining	**78.79** (+1.02)	**68.20** (+1.72)	**44.09** (+0.08)	**56.96** (+0.81)	**64.72** (+0.74)	**49.18** (-0.17)	**30.67** (+0.19)	**47.42** (+0.35)
SimBase + ReCorrect (Ours)	**78.55** (+0.78)	**68.39** (+1.91)	**45.78** (+1.77)	**57.42** (+1.27)	**65.12** (+1.14)	**49.45** (+0.10)	**30.73** (+0.25)	**47.59** (+0.52)

presents data samples from Vid-Group, consisting of untrimmed videos paired with pseudo annotations for sentence queries and temporal boundaries. Vid-Group contains a rich variety of visual and semantic content, which includes: 1) both individual and group activities, 2) a range of visual styles, such as thermal imaging and anatomical representations, 3) a variety of activity types, including underwater exploration and animal behaviors, and 4) a broad temporal distribution.

Dataset Availability and Copyright. The full Vid-Group dataset is released at https://github.com/baopj/Vid-Group, under the CC BY 4.0 license. In order to comply with legal requirements, we provide YouTube links to the videos instead of distributing the video files. Users are advised to download the videos independently and to strictly adhere to YouTube's Terms of Service and all applicable copyright policies when accessing or using the video content, as well as adhering to the license terms governing the annotations.

Dataset Statistics. Fig. 6a illustrates the duration distribution of the videos in Vid-Group. The average duration of the videos is 72.44 seconds. The duration slot of 55 to 60 seconds has the maximum videos, and the slot between the 65 to 240 seconds has balanced distribution of video numbers. Fig. 6b and 6c summarize the distribution of start and end timepoints for temporal boundaries in the Vid-Group Dataset, where the start and end times are normalized relative to the duration of their respective untrimmed videos. It exhibits a generally balanced distribution for both start and end timepoints, with slightly more annotations between 0 to 0.1 and fewer between 0.8 to 1.0 for the start time.

3. Additional Performance Comparisons

Fully-supervised learning. Table 1 shows that adding both GPT-4o Pretraining and ReCorrect to SimBase [1] in a fully-supervised setting enhance performance. Both methods show an R@0.5 improvement of over 1.5 points on Charades and around 1 point on ActivityNet for R@0.3 compared to SimBase. Compared to GPT-4o Pretraining, ReCorrect further boosts performance by approximately 2

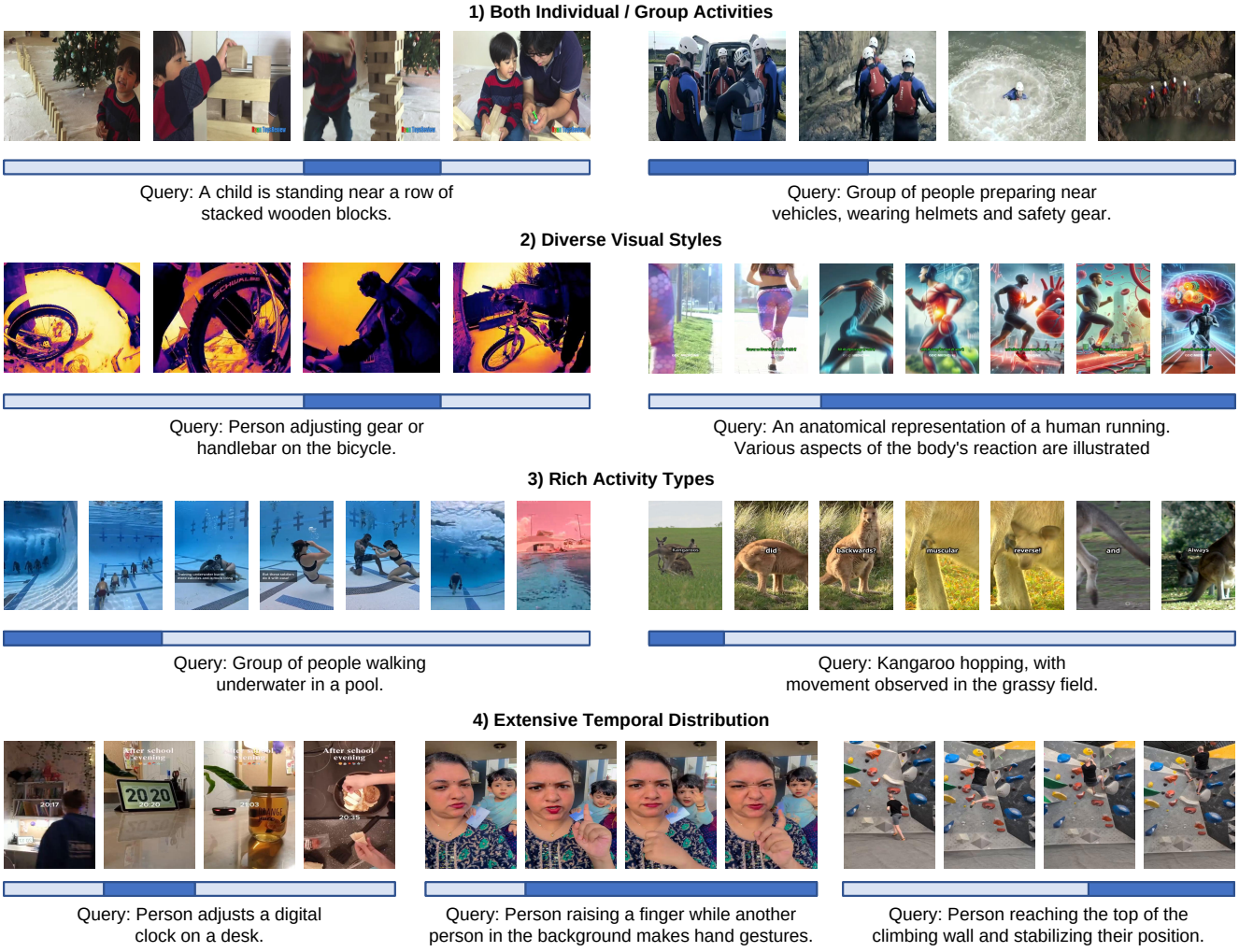
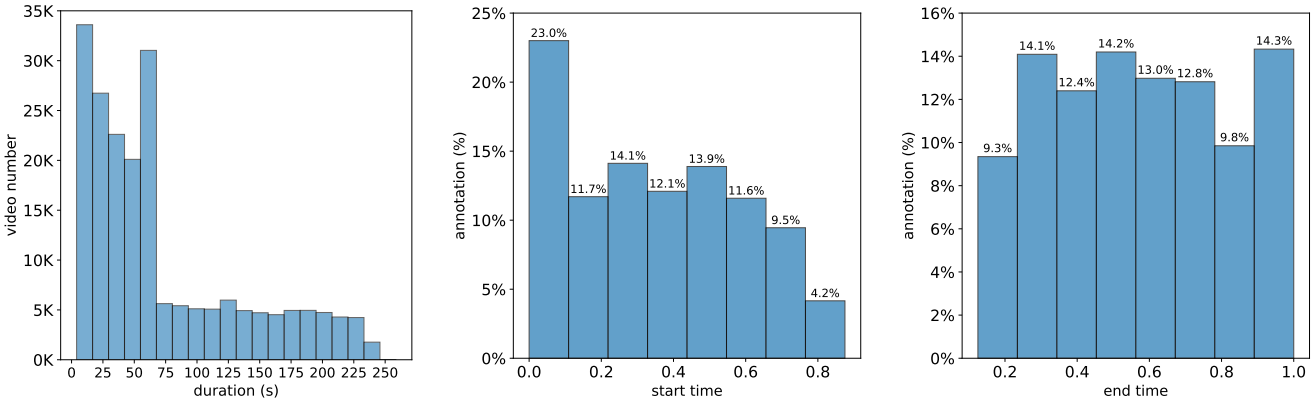


Figure 5. Data samples from Vid-Group, including 1) both individual and group activities, 2) diverse visual styles including thermal imaging and anatomical representations, 3) a wide range of activity types such as underwater and animal behaviors, and 4) an extensive temporal distribution.



(a) (b) (c)

Figure 6. a) The distribution of video durations in the Vid-Group dataset. b) Temporal distribution of start timepoint for temporal boundaries, normalized by the duration of the corresponding untrimmed videos. c) The distribution of end timepoint.

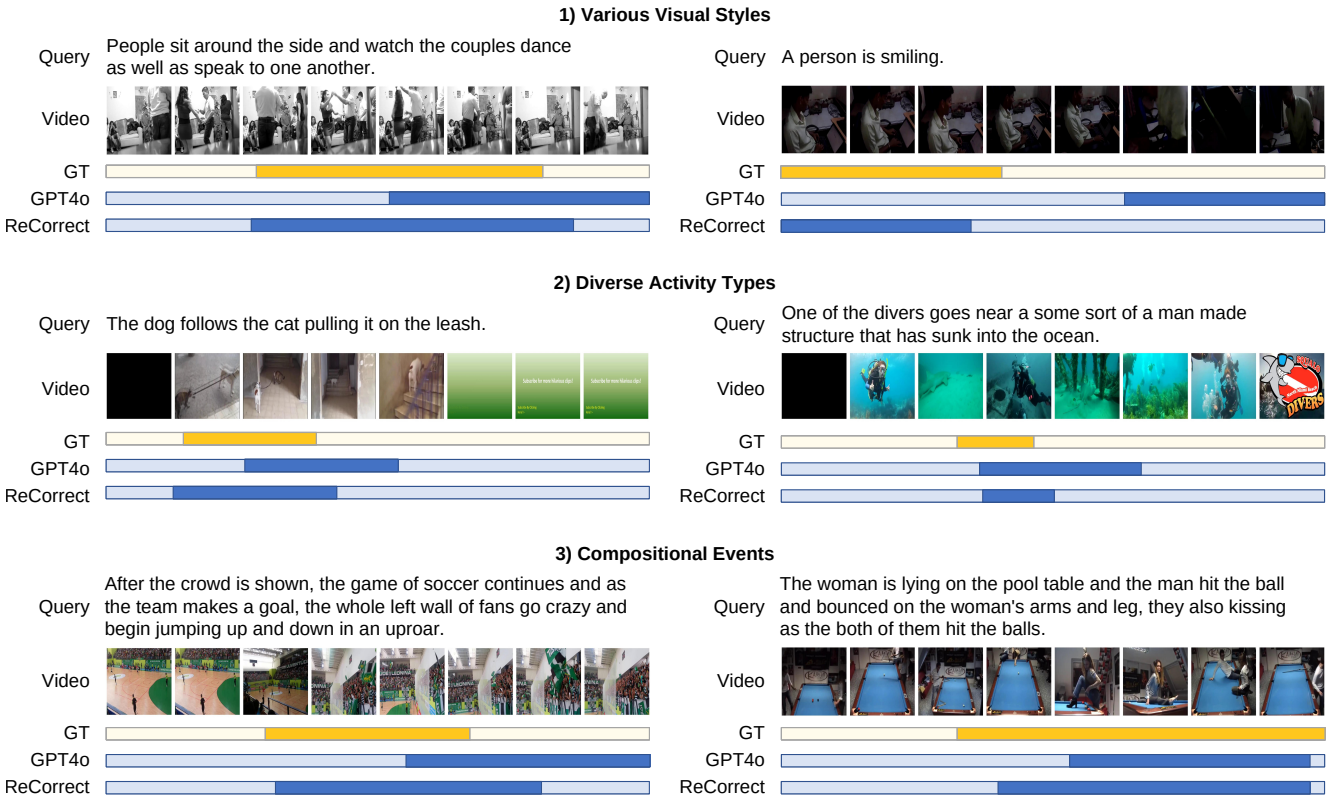


Figure 7. Qualitative comparison of zero-shot inference between GPT-4o pretraining and our ReCorrect algorithm. Our zero-shot ReCorrect demonstrates its powerful capability in temporal video grounding across: 1) various visual conditions, such as black-and-white movie segments and low-light scenarios; 2) diverse activity types, including animal behavior and underwater scenes; and 3) compositional events comprising multiple sub-events and necessitating temporal reasoning. Here "GT" indicates ground truth. The darker yellow rectangle represents the ground-truth temporal boundaries, while the darker blue rectangle denotes the model's prediction.

points at R@0.7 on Charades and 0.5 points at R@0.3 on ActivityNet. The margin of improvement between ReCorrect and GPT-4o Pretraining is narrower here than in zero-shot settings, as the fine-tuning datasets provide 12.8K and 37.4K manual labels, respectively.

Qualitative Comparisons Fig. 7 illustrates a qualitative performance comparison between GPT-4o pretraining and our ReCorrect algorithm in zero-shot temporal video grounding. The results highlight the strengths of our zero-shot ReCorrect approach across three challenging scenarios: (1) handling diverse visual conditions, such as black-and-white movie segments and low-light scenarios; (2) effectively retrieving moments from diverse activity types, including animal behavior and underwater scenes; and (3) accurately reasoning about compositional events that involve multiple sub-events and require temporal understanding.

4. Additional Implementation Details

The frame number T and step size δ for the semantics-guided refinement are set to 256 and 5, respectively. The hyperparameters α_1 and α_2 are configured as 0.22 and 0.92, respectively. The loss weight λ is set to 0.7. For the temporal video grounding model, we adopt the same network architecture as the state-of-the-art fully-supervised model SimBase [1]. Details of the network architecture can be referred to [1] and can also be accessed through our implementation at https://github.com/baopj/Vid-Group.

References

[1] Peijun Bao and Alex Kot. Simbase: A simple baseline for temporal video grounding. *arXiv*, 2024. 2, 4

[2] Pilhyeon Lee and Hyeran Byun. Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In *ECCV*, 2024. 2

[3] Zhihang Liu, Jun Li, Hongtao Xie, Pandeng Li, Jiannan Ge, Sun-Ao Liu, and Guoqing Jin. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *AAAI*, 2024. 2

[4] Shengjia Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David A. Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *ICCV*, 2023. 2