Hierarchical Event Memory for Accurate and Low-latency Online Video Temporal Grounding

Minghang Zheng[1] Yuxin Peng[1] Benyuan Sun[3] Yi Yang[3] Yang Liu[1,2*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]State Key Laboratory of General Artificial Intelligence, Peking University
[3]Central Media Technology Institute, Huawei
{minghang,pengyuxin,yangliu}@pku.edu.cn
{sunbenyuan,yangyi16}@huawei.com

In this supplementary material, we provide additional implementation details. In Section 1, we provide more implementation details of our method and evaluation metrics. In Section 2, we provide more visualization results.

1. Experiment Details

1.1. Feature Extraction.

We follow Gan et al. [1] and adopt the C3D features [6] on the TACoS [4] and ActivityNet Captions [2] datasets and CLIP [3] features on the MAD [5] datasets for a fair comparison. Specifically, on the TACoS and ActivityNet Captions datasets, we use the C3D model to extract the RGB features of 16 consecutive frames with a stride of 16, resulting in a final input frame rate of 1.875 FPS. We use the GloVe 6B word embeddings and a 5-layer transformer encoder to extract the sentence features. On the MAD dataset, we use CLIP-B-32 to extract both the visual and text features. The input frame rate is 5 FPS on the MAD dataset. The pre-trained C3D and CLIP models are frozen during training.

1.2. Training Details.

Due to the long duration of the complete videos, we segment the video sequences into segments with a maximum length of L for efficient training. For ActivityNet and MAD datasets, $L = 1024$, while for TACoS dataset, $L = 2304$. For each video segment, we sequentially enumerate short-term windows, construct event proposals within the current window, and update the memory. When we obtain all event proposals from the video, we perform predictions and gradient updates.

*Corresponding author

1.3. Evaluation Details.

We follow Gan et al. [1] to use an offline evaluation (i.e. evaluate after all videos have been input). *For baseline models* that predict the probability s_i, e_i of an event starting or ending at time i, we follow the baseline model, enumerate candidates (i, j) from clip $i-1$ to j, calculate the probability: $c_{ij} = s_i * e_j$. The final top-n predictions are the top-n candidates with the highest scores c_{ij} after Non-maximum suppression (NMS). *For our model without future prediction*, we obtain the score c_i and the regressed time boundaries (p_i^s, p_i^e) for the i-th proposal through the classifier and regressor. We directly output the top-n predictions with the highest scores c_i after NMS. Through future predictions, we can obtain the probability c_t^f that the target event will occur in the near future at time t, as well as the predicted start time of the event $t + o_t^f$. *For our model with future prediction*, we modify the start time predictions from event proposals to the future prediction. Specifically, for the i-th proposal and time stamp t, the modified prediction is $(t + o_t^f, p_i^e)$, and its score is $c_t^f * c_i$. We enumerate i and t, and output the top-n predictions with the highest scores $c_t^f * c_i$ after NMS.

For the start and end times evaluation, assuming we have the true positive samples $\{(p_i^s, p_i^e, t_i^s, t_i^e)\}_{i=1}^{N_p}$, where p_i^s and p_i^e are the predicted start and end times, t_i^s, t_i^e are the times at which the model outputs the start/end predictions and N_p is the number of true positive predictions. The corresponding ground truth is $\{g_i^s, g_i^e\}_{i=1}^{N_p}$, where g^s and g^e are the start and end time. A sample is considered a true positive if and only if the model's top-1 prediction has an IoU with the ground truth greater than the threshold m. We define the start time delay (SD) and end time delay (ED) as:

$$\text{SD} = \frac{1}{N_p}\sum_{i=1}^{N_p}(t_i^s - g_i^s), \text{ED} = \frac{1}{N_p}\sum_{i=1}^{N_p}(t_i^e - g_i^e) \quad (1)$$

We evaluate delays at different values of $m(m =$

$0.3, 0.5, 0.7$) and report the average delay.

2. More Qualitative Results

In Figure 1 and 2, we provide additional visualization results. It can be observed that the baseline model successfully locates the shorter target event in Figure 1a. However, it incorrectly predicts the longer target time in Figure 1b. In contrast, our method performs better in both cases. Specifically, in Figure 1b, our prediction based on event proposals provides a more accurate estimate of the event's end time. However, the prediction of the event's start time based on future predictions remains insufficiently accurate. This indicates the difficulty of predicting the start time of a long event that has not yet begun. Figure 2 shows a failure case. The query text is complex and includes multiple events, such as marathons, live music, and live group yoga, resulting in the target event having a longer duration. Both our method and the baseline only identified the first half of the target event, leading to incomplete localization results. This indicates that our method still has shortcomings in locating complex events.

Limitations. As demonstrated in Table 6 of the main text, our method is slower than the baseline model. However, our method can still meet the requirements for real-time prediction (with a speed greater than the frame rate of typical videos). Additionally, although our hierarchical event memory can retain long-term historical information, the length of this historical information is still limited by the memory size and the number of event scales L (with the longest proposal length being $2^{(L-1)}$).

References

[1] Tian Gan, Xiao Wang, Yan Sun, Jianlong Wu, Qingpei Guo, and Liqiang Nie. Temporal sentence grounding in streaming videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4637–4646, 2023. 1

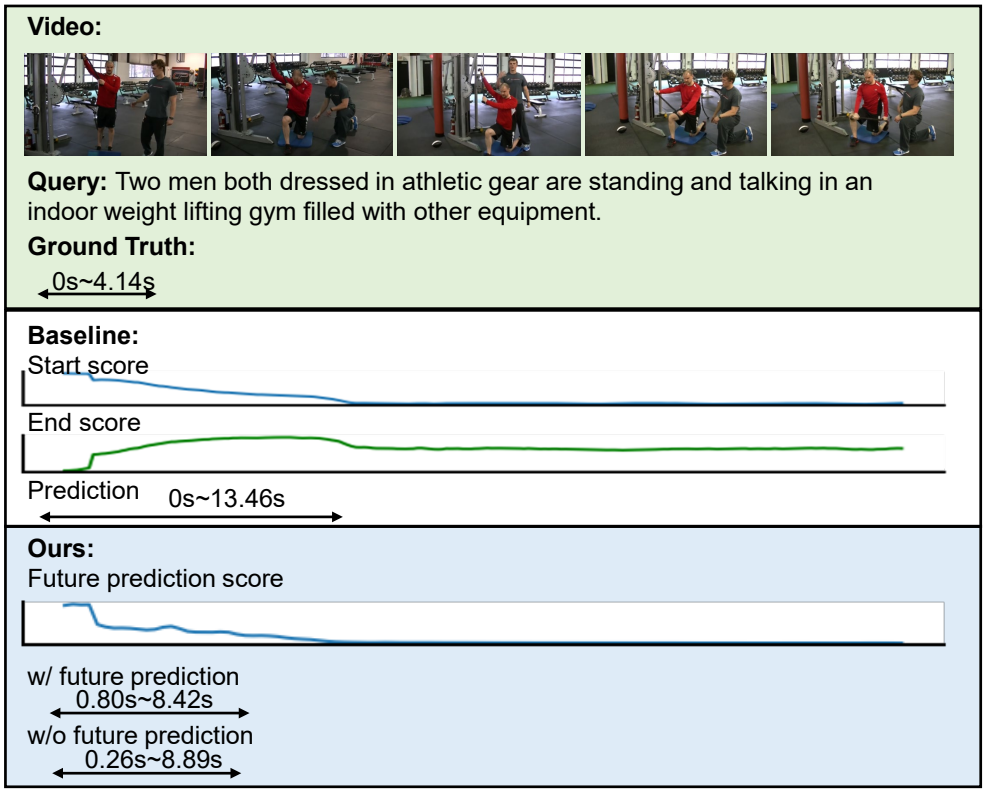
[2] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

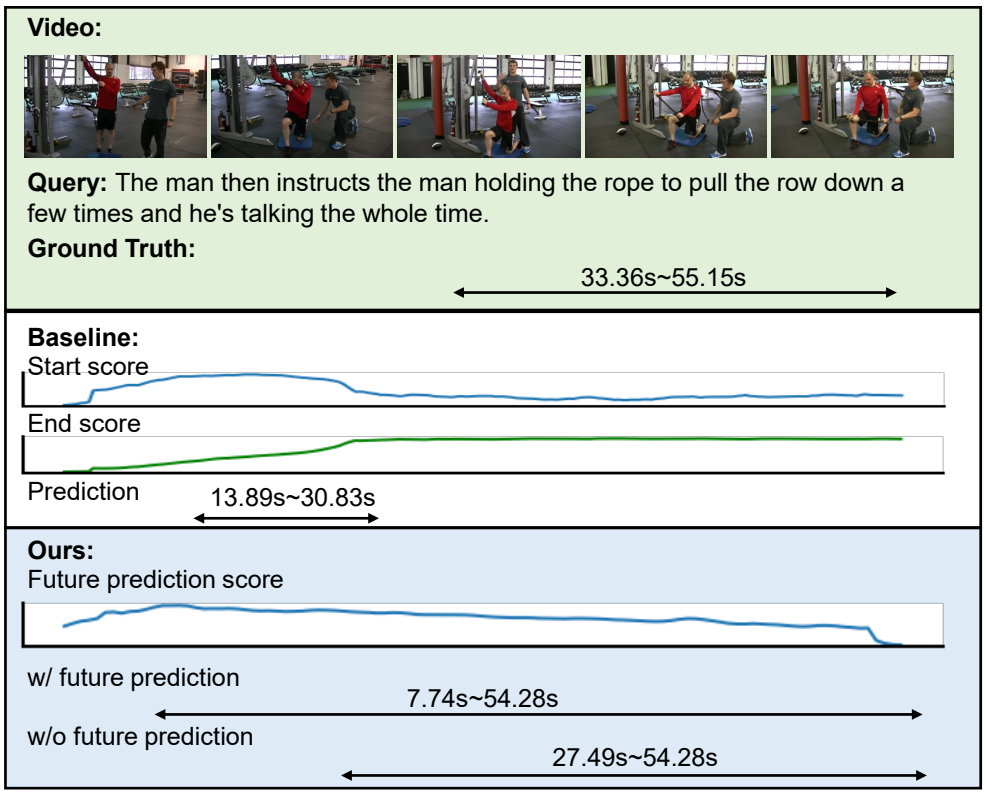
[4] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 184–195. Springer, 2014. 1

[5] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022. 1

[6] 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*, pages 4489–4497, 2015. 1



(a)



(b)

Figure 1. Qualitative results on ActivityNet Captions dataset.

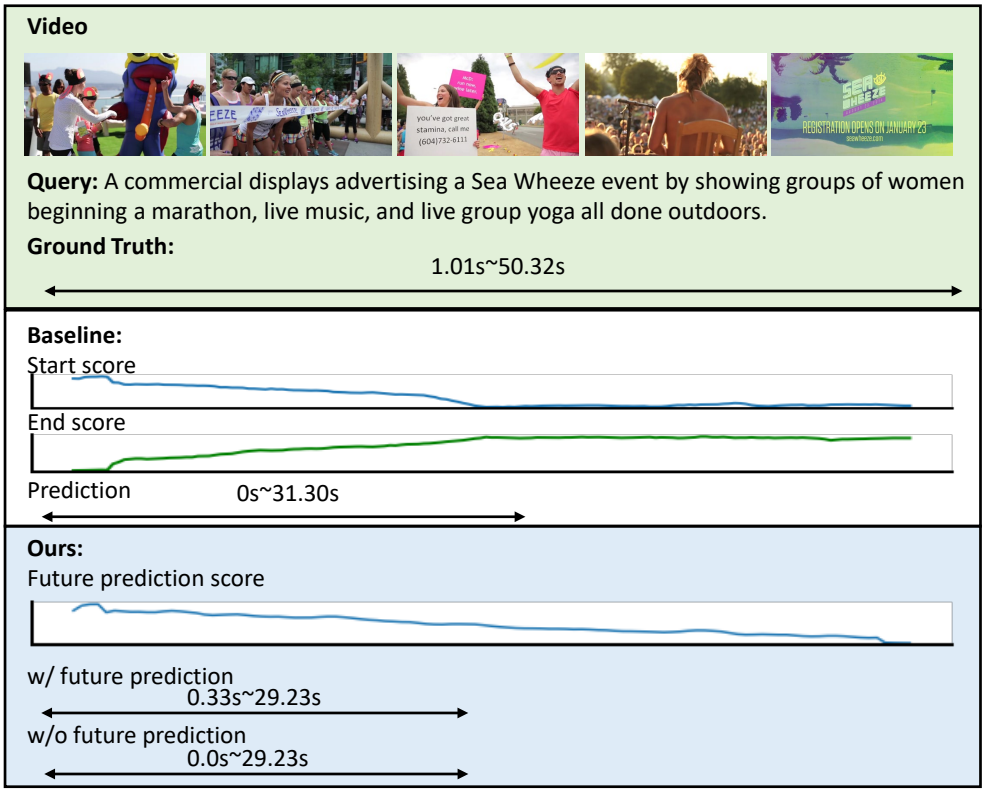


Figure 2. Qualitative results of failure cases on ActivityNet Captions dataset.