TempR1: Improving Temporal Understanding of MLLMs via Temporal-Aware Multi-Task Reinforcement Learning

Supplementary Material

A. More Detailed Formulations of each Task

Temporal Grounding (TG): Given a video V and a text query that describes an event in the video, the model is required to predict a temporal interval where the event occurs. Formally, the output O should follow the following format: $O =$ <answer>t_s to t_e</answer>, where t_s and t_e denote the start and end timestamps, respectively.

Dense Temporal Grounding (DTG): In DTG, the model must localize multiple temporal intervals corresponding to a series of event descriptions within a complex query. The predicted temporal intervals and the event descriptions are in one-to-one correspondence. The output format is: $O =$ <answer>t_{s1} to t_{e1}, t_{s2} to t_{e2}, ...</answer>.

Video Highlight Detection (VHD): This task requires locating multiple highlight segments from the video based on the description of the highlight event. The number of highlight segments is uncertain, but they all belong to the same highlight event. The output format is: $O =$ <answer>t_{s1} to t_{e1}, t_{s2} to t_{e2}, ...</answer>.

Grounded Video Question Answering (GVQA): Given a video and a question about the video content, the model must provide both (1) the correct answer option and (2) the supporting temporal evidence. This task requires reasoning over multiple time segments that justify the answer. The output format is: $O =$ <answer>A</answer><glue>t_s^1 to t_e^1, t_s^2 to t_e^2, ... </glue>, where A denotes the chosen answer option.

Temporal Action Localization (TAL): Given a predefined action label (e.g., "long jump" or "sailing"), the model needs to localize *all* temporal intervals in the video that contain that action. Unlike TG/DTG, the number of instances is not given a priori, so the model must infer both the instance count and corresponding temporal boundaries. The output format is identical to DTG:$O =$ <answer>t_s^1 to t_e^1, t_s^2 to t_e^2, ...</answer>.

B. Detailed Matching Algorithm Description

We provide an expanded description of the matching algorithm used to align predicted temporal intervals with ground-truth instances for computing the Type 3 temporal localization reward (e.g., TAL), as shown in Algorithm 1.

Algorithm 1 Matching between Predicted Intervals and Groundtruth Instances

Require: Predicted intervals $\{p_1, p_2, \ldots, p_m\}$ and ground-truth instances $\{g_1, g_2, \ldots, g_n\}$

Step 1: Compute IoU matrix $\mathbf{I} \in \mathbb{R}^{m \times n}$.
Step 2: Matching by Dynamic Programming.
Initialize DP table $\mathbf{D} \in \mathbb{R}^{(m+1)\times(n+1)} \leftarrow 0$
Initialize path record table $\mathbf{P}$ of the same size
for $i = 1$ to m **do**
  for $j = 1$ to n **do**
    $match \leftarrow \mathbf{I}[i, j]$ ▷ IoU between p_i and g_j
    $option_1 \leftarrow \mathbf{D}[i-1, j]$ ▷ Skip current prediction
    $option_2 \leftarrow \mathbf{D}[i, j-1]$ ▷ Skip current ground truth
    $option_3 \leftarrow \mathbf{D}[i-1, j-1] + match$ ▷ Match p_i and g_j
    $\mathbf{D}[i, j] \leftarrow \max(option_1, option_2, option_3)$
    if $\mathbf{D}[i, j] == option_3$ **then**
      Append (i, j) to $\mathbf{P}[i, j]$
    else if $\mathbf{D}[i, j] == option_1$ **then**
      $\mathbf{P}[i, j] \leftarrow \mathbf{P}[i-1, j]$
    else
      $\mathbf{P}[i, j] \leftarrow \mathbf{P}[i, j-1]$
    end if
  end for
end for
$\mathcal{M} \leftarrow \mathbf{P}[m, n]$ ▷ Retrieve final matching pairs
return $\mathcal{M}$

C. Training Corpus Details

We curate a large-scale multi-task training corpus for temporal video understanding by integrating data from multiple established datasets and benchmarks. The final corpus comprises approximately **60K** video samples spanning five representative temporal understanding tasks, offering diverse temporal structures and linguistic expressions for reinforcement learning. Table 1 summarizes the dataset composition.

Task	Source Datasets	# Samples
Temporal Grounding (TG)	Charades-STA [3], DiDeMo [1], TimeRFT [6]	25,742
Dense Temporal Grounding (DTG)	ActivityNet-Caption [4]	9,244
Video Highlight Detection (VHD)	QVHighlights [5]	7,038
Grounded Video QA (GVQA)	NExTGQA [7]	3,353
Temporal Action Localization (TAL)	ActivityNet-v1.3 [2], HACS [8]	14,990
Total	—	**60,367**

Table 1. Composition of the multi-task temporal understanding training corpus.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[3] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1

[4] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1

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

[6] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025. 1

[7] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 1

[8] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019. 1