

APPENDIX

A TRAINING DATA

Table 6 summarizes the statistics of the training datasets. For the QVHighlights (Lei et al., 2021) training split, which contains 9,996 examples, we only keep videos longer than 120 seconds. For the PLM-Video (Cho et al., 2025) multiple-choice split, we perform a quality check to remove examples that cannot be correctly answered using the full video, but can be correctly answered when restricted to the cropped segment defined by the clue duration. This ensures that the clue duration indeed provides sufficient information for identifying the correct video segment.

After Stage I training, we employ the model trained from Stage I for offline data filtering. Specifically, we generate $n = 8$ responses per example and discard those without meaningful reward signals. For question-answering ability, we exclude examples for which all generated responses answer the question correctly, as they lack discriminative signals. For temporal grounding ability, we filter out examples with low response variance. In particular, we retain only examples where the responses yield a sufficiently strong relative reward signal, quantified by the difference between the maximum IoU and the mean IoU across responses:

$$\delta = \max_{1 \leq i \leq n} \text{IoU}_i - \frac{1}{n} \sum_{i=1}^n \text{IoU}_i \quad (6)$$

We filter out examples with $\delta < 0.1$.

Table 6: **Statistics of training data.** NExT-GQA and ActivityNet in seconds stage are sampled from the first stage by filtering reward variation based on the first-stage model.

	Dataset	#Queries	Video Len.	Moment Len.
Stage I	NExT-GQA (Xiao et al., 2024)	3,358	43.9s	8.5s
	ActivityNet (Krishna et al., 2017)	4,727	177.3s	48.3s
	QVHighlights (Lei et al., 2021)	7,218	150.0s	34.1s
Stage II	ActivityNet (Krishna et al., 2017)	1,395	220.87	88.4s
	NExT-GQA (Xiao et al., 2024)	1,004	50.2s	7.1s
	PLM-Video (Cho et al., 2025)	5,333	808.6s	26.1s

B IMPLEMENTATION DETAILS

We adopt Qwen2.5-VL-7B (Bai et al., 2025) as the base model. The maximum number of video tokens is set to 8,192, with videos sampled at 1 fps during training. The minimum video frame resolution is $16 \times 28 \times 28$ pixels, and the maximum is $768 \times 28 \times 28$, allowing the number of tokens per frame to be adaptively adjusted under the video context budget. The maximum response length is capped at 512 tokens. Due to computational resource limitations, we conduct RL training in two stages. In the first stage, we train on 20K short-video GQA examples from NExT-GQA (Xiao et al., 2024), ActivityNet (Krishna et al., 2017), and QVHighlights (Lei et al., 2021). In the second stage, we train on the *yt1b_mcqa* split from PLM-Video (Cho et al., 2025), combined with the short-video data sampled from the first stage, for a total of 7K examples. The statistics of training data are shown in Appendix A and Table 6. All experiments are performed on NVIDIA A100 GPUs (80GB), with a global batch size of 64.

During inference, we evaluate all models at 1 FPS with a context size of 8,192 on the short-video benchmarks NExT-GQA (Xiao et al., 2024) and ReXTime (Chen et al., 2024a). For long-video benchmarks: CG-Bench (Chen et al., 2025a), VideoMME (Fu et al., 2025), MLVU (Zhou et al., 2025), and LVBench (Wang et al., 2024). We also uniformly sampled a maximum of 256 frames and set the context size to 16,384.

Table 7: A simple example to demonstrate the GRPO’s uniform credit assignment problem.

Response	R_{IoU}	R_{Acc}	A_{IoU}	A_{Acc}	R_{Sum}	A_{Sum}
1	0.0	1.0	-1.40	+0.82	1.0	+0.06
2	0.5	0.0	+0.44	-1.22	0.5	-1.54
3	0.4	1.0	+0.07	+0.82	1.4	+1.34
4	0.8	0.0	+1.55	-1.22	0.8	-0.58
5	0.2	1.0	-0.66	+0.82	1.2	+0.70

Table 8: Ablation on the number of generated responses G per prompt during GRPO training.

G	NExT-GQA				RexTime			
	Acc	mIoU	R@0.3	R@0.5	Acc	mIoU	R@0.3	R@0.5
2	69.6	33.7	50.3	27.5	58.5	36.9	49.0	37.1
4	69.8	35.2	52.6	29.6	58.8	40.1	53.2	40.6
8	70.7	37.6	55.6	33.8	62.0	43.2	56.5	44.1

C LIMITATION OF GRPO IN UNIFORM CREDIT ASSIGNMENT

In Table 7, we present a simple example with two rewards, R_{IoU} and R_{Acc} , across five responses to illustrate the limitations of GRPO arising from naïve reward summation and uniform credit assignment. For example, the first response attains the lowest temporal grounding reward, $R_{IoU}^{(1)} = 0$, yet its overall advantage under standard GRPO is positive, $A_{Sum}^{(1)} = +0.06$. In contrast, response 4 achieves much better temporal grounding, $R_{IoU}^{(4)} = 0.8$, but receives a lower advantage, $A_{Sum}^{(4)} = -0.58$. Due to uniform credit assignment, all tokens in response 1 are reinforced by the positive advantage, while all tokens in response 4 are penalized. This hides the contribution of tokens that support more accurate temporal grounding.

In contrast, computing separate advantages for each reward, A_{IoU} and A_{Acc} , provides a clearer view of each task’s contribution. By selectively assigning these decoupled advantages to the corresponding tokens, our approach TokenAdv, updates the policy to increase the probability of tokens that positively impact their respective tasks.

D EXPERIMENTS

D.1 TEMPORAL GROUNDING COVERAGE

In addition to IoU, we evaluate how well the predicted segment covers the ground-truth clue span using Intersection-over-Ground truth (IoG), defined as $IoG = \frac{|\mathcal{I}_{pred} \cap \mathcal{I}_{gt}|}{|\mathcal{I}_{gt}|}$, where \mathcal{I}_{pred} is the predicted temporal span and \mathcal{I}_{gt} is the ground-truth clue span. We report mean IoG (mIoG) as the average IoG across instances. IoG directly measures coverage of the ground truth and thus verifies whether temporal grounding captures the key frames relevant to the query, particularly informative for the finer-grained zoom-in.

We present results in Table 3. For ReXTime (Chen et al., 2024a), only IoU and accuracy are available via server-side evaluation, and the ground-truth clue spans are not released; consequently, we compute and report mIoG on the validation set for comparison. Our model improves mIoG by 1.2% on NExT-GQA (Xiao et al., 2024) and by 3.3% on ReXTime (Chen et al., 2024a). For very long videos such as CG-Bench (Chen et al., 2025a), mIoU can be less informative because the larger denominator depresses scores. Considering both mIoU and mIoG shows that our model not only localizes the relevant moments but also achieves strong coverage of key frames.

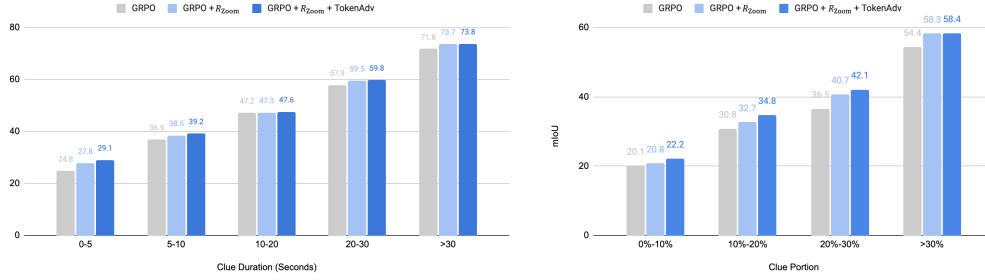


Figure 3: **Temporal grounding robustness analysis on NExT-GQA.** Left: mIoU results across different ground-truth clue durations. Right: mIoU results across different clue proportions (ground-truth clue duration relative to the total video duration).

D.2 THE NUMBER OF GENERATED RESPONSES

We investigate the impact of the number of generated responses G per prompt during GRPO training, as this hyperparameter directly influences the diversity and quality of the policy optimization process. As presented in Table 8, increasing G from 2 to 8 consistently improves performance across both datasets and all evaluation metrics. Based on these results, we adopt $G = 8$ for all main experiments, as it provides the optimal balance between computational efficiency and performance gains.

D.3 DIVIDE-AND-CONQUER

We study the impact of window size (Table 9) and the number of predicted temporal spans aggregated in the divide-and-conquer strategy. Because this approach requires scanning every sliding window during the coarse-grained pass, it introduces an average $\times 2.3$ increase in inference cost. Nevertheless, it improves performance across all long-video benchmarks by an average of +6.4%, demonstrating that our temporal zoom-in with higher spatial resolution provides substantial benefits for long video understanding. (Table 10) shows the impact of number of aggregated temporal spans with top answer confidence.

Table 9: Window size ablation.

Window Size	VideoMME (<i>long w sub.</i>)	MLVU	LVBench
128	67.4	72.1	47.6
256	68.7	73.4	48.1
384	68.4	72.4	48.5

Table 10: Number of aggregated temporal spans with top answer confidence.

# Aggregated Spans	VideoMME (<i>long w sub.</i>)	MLVU	LVBench
3	73.2	66.8	47.5
4	73.4	68.7	48.1
5	73.2	68.4	47.9

D.4 COARSE-TO-FINE VIDEO UNDERSTANDING ON GVQA

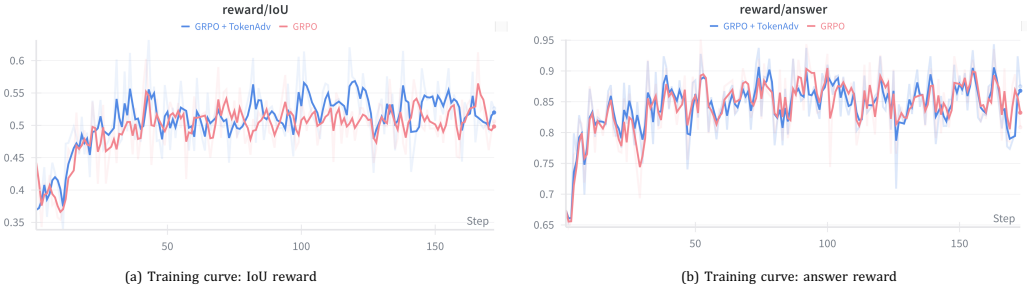
We further evaluate short-form GVQA answer accuracy on NExT-GQA and ReXTime through temporal zoom-in, as reported in Table 11. Both benchmarks consist of short videos, where the model

Table 11: Grounded question answering (GQA) results on NExT-GQA (Xiao et al., 2024) and ReXTime (Chen et al., 2024a) with temporal zoom-in.

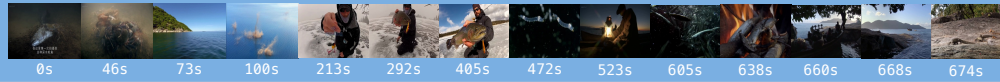
Models	NExT-GQA				ReXTime			
	Acc	mIoU	R@0.3	R@0.5	Acc	mIoU	R@0.3	R@0.5
Zoom-Zero	70.7	37.6	55.6	33.8	62.0	43.2	56.5	44.1
Zoom-Zero + Coarse-to-fine	71.4	N/A	N/A	N/A	62.8	N/A	N/A	N/A

Table 12: Spatial and temporal resolution in coarse-to-fine video understanding.

Benchmark	VideoMME Long		MLVU		LVBench	
Duration	2386s		651s		4101s	
	Coarse	Fine	Coarse	Fine	Coarse	Fine
Avg frames	256	136	253	86	256	154
Avg FPS	0.1	1.0	0.2	1.0	0.06	1.0
Avg tokens/frame	54	76	64	190	64	62

Figure 4: Training curve: IoU reward R_{IoU} and answer reward R_{Acc} comparison with baseline GRPO and our improved GRPO with TokenAdv.

Question: Arrange the following events from the video in the correct chronological order: (1)The action of the lures is shown underwater as several different fish go after the lures; (2)Several men show off the different lures they are using for ice fishing; (3)The video ends with the closing credits and Graphics shown on the screen; (4)An introduction comes onto the screen for a video about fishing lures.
Options: (A) 3->2->1->4 (B) 4->2->1->3 (C) 1->2->3->4 (D) 2->1->3->4



Output: <think>First, the video introduces the topic of fishing lures with an opening screen <time> (20.3, 49.7) </time>. Next, several men showcase the different lures they are using for ice fishing <time> (283.5, 304.2) </time>. Then, the action of the lures is demonstrated underwater as various fish chase after them <time> (593.1, 611.4) </time>. Finally, the video concludes with closing credits and graphics <time> (673.8, 680.2) </time>.</think>\n<answer>B</answer><glue>[(20.3, 49.7), (283.5, 304.2), (593.1, 611.4), (673.8, 680.2)]</glue>

Figure 5: A qualitative example for long video understanding.

can preserve most temporal context and details within the context budget. In this setting, the zoom-in paradigm improves performance by 0.7% on NExT-GQA and 0.8% on ReXTime.

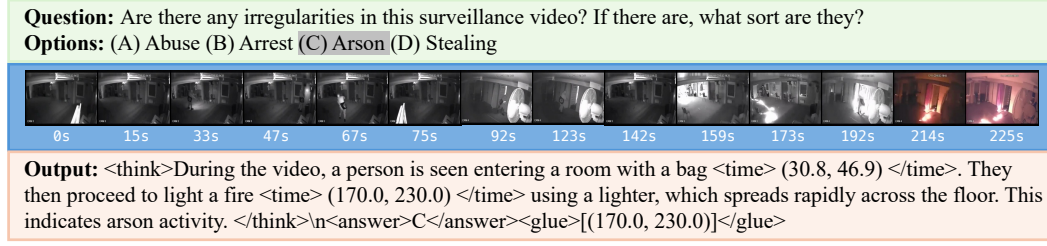


Figure 6: A qualitative example for long video understanding.

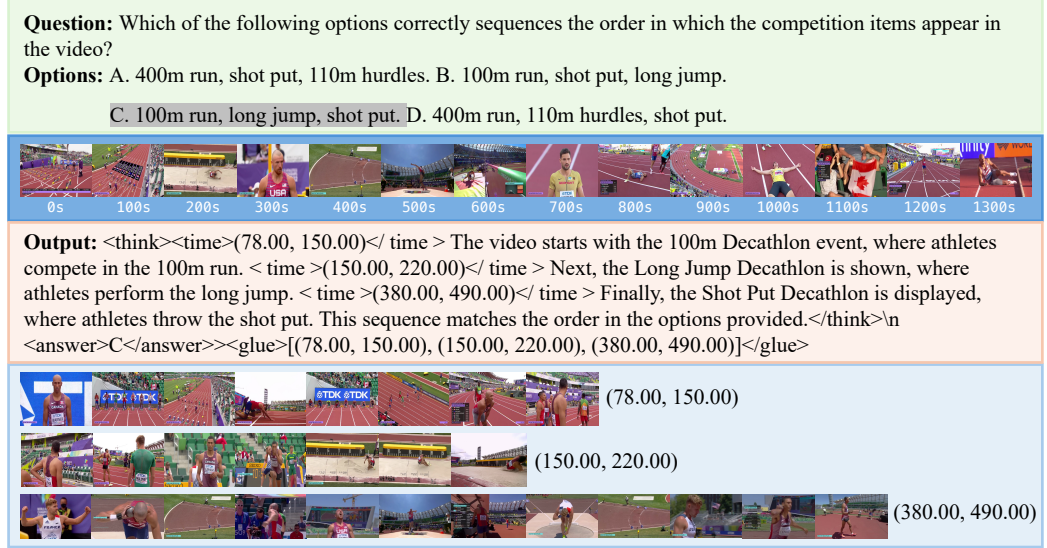


Figure 7: A qualitative example for long video understanding.

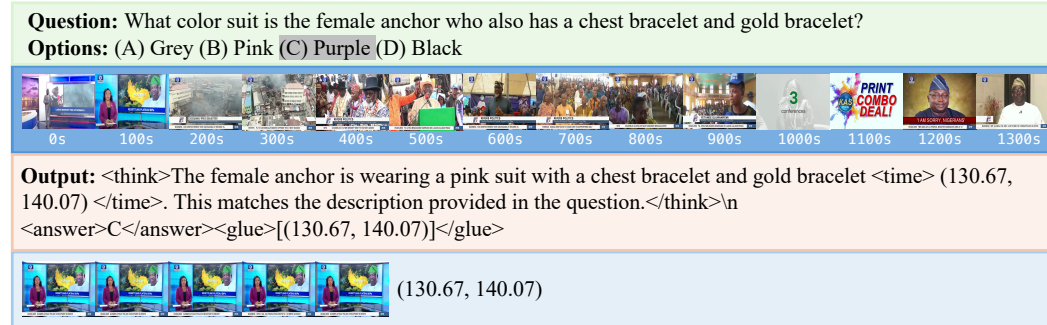


Figure 8: A qualitative example for long video understanding.

E LIMITATION AND FUTURE DIRECTION

Our current approach performs only a single round of zoom-in during both training and inference. We did not explore iterative or recursive zooming due to computational constraints. However, multi-stage zooming could further refine temporal grounding by progressively narrowing the search space and focusing on increasingly fine-grained visual cues.

Another limitation is that the zoom-in process is enforced rather than adaptive. Ideally, the model itself should decide whether, when, and how many times to zoom in, guided by the task objective. A goal-oriented, multi-step zooming policy could potentially yield more efficient and faithful grounding.

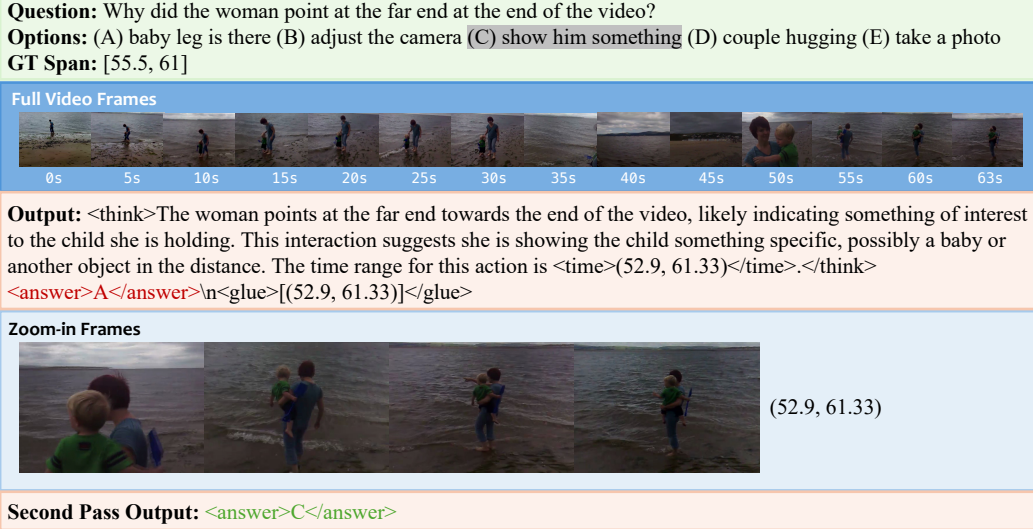


Figure 9: A qualitative example for grounded videoQA with temporal zoom-in.

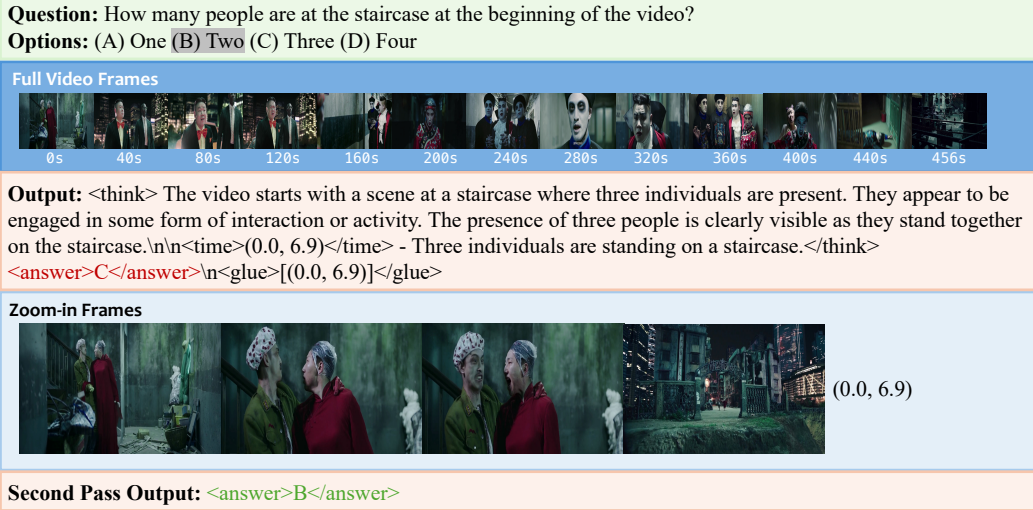


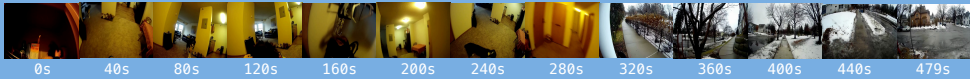
Figure 10: A qualitative example for long video understanding with coarse-to-fine zoom-in.

If a strong pretrained model with reasonable temporal grounding ability and exploration samples is sufficient, our framework could also be trained without explicit temporal interval annotations. Instead of relying on rule-based reward, i.e., R_{IoU} , the model could learn to verify whether key visual clues exist within its predicted temporal segments. This self-verification mechanism has the potential to mutually enhance answer accuracy and temporal grounding, especially in long-video scenarios where temporal annotations are often scarce to obtain.

Question: What did I take out of the cupboard?

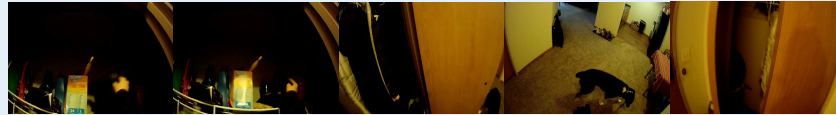
Options: (A) cotton hand gloves I took out of the cupboard (B) wool sweater (C) plastic hangers (D) metal spoon

Full Video Frames



Output: <think> The video shows a person opening a cupboard and taking out items. The specific item taken out is not clearly visible, but it appears to be a piece of clothing based on the context of the action. The options provided suggest that the item could be either a wool sweater or cotton hand gloves. However, without a clear view of the item, it's difficult to determine which one it is</think>\n<answer>B</answer>\n<glue>[(224.0, 251.0)]</glue>

Zoom-in Frames



Second Pass Output: <answer>A</answer>

Figure 11: A qualitative example for long video understanding with coarse-to-fine zoom-in.