LEARNING TO GROUND WITHOUT FORGETTING: REINFORCEMENT LEARNING FOR LONG-VIDEO TEMPORAL UNDERSTANDING IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video temporal grounding introduces complexities not found in standard short-video settings, as the extended temporal horizon amplifies localization error and increases event sparsity. A central challenge in adopting reinforcement learning (RL) for this task is preserving pre-trained knowledge while actively exploring novel strategies for generating the precise, structured timestamps needed to resolve event sparsity and ambiguity. To address this, we propose *Token-aware KL Regularization*, which selectively relaxes the KL-divergence on numeric and structural tokens to encourage exploration during early training. To further mitigate the *sparsity* of key events, we introduce a dense reward called *Center Distance Reward* (CenDist). Combined with our automatically constructed *SceneTG* dataset that facilitates effective RL training, our model, LONGVTG-R1, achieves the highest mIoU among the compared methods at similar training scales across three long-video temporal grounding benchmarks. Beyond grounding, our model generalizes to long-video question answering (QA), serving either as an explicit grounding module or, more importantly, as a unified system that jointly performs grounding and QA. Finally, LONGVTG-R1 transfers to multi-segment grounding through recursive single-segment inference, without multi-segment fine-tuning.

1 INTRODUCTION

Understanding long videos, where models must reason over hours of untrimmed content, has become an essential goal for multimodal large language models (MLLMs). A central step is temporal grounding (Gao et al., 2017; Zhang et al., 2023; Liu et al., 2025), which localizes fine-grained video segments corresponding to a natural language query (*e.g.*, "Find the moment when the driver releases the steering wheel"). Despite its importance, temporal grounding in long videos remains a significant bottleneck due to two factors: (1) the **sparsity** of query-relevant events in long videos indicates that even a 1% shift in the total duration can cause a drastic miss of the targets; and (2) the **ambiguity** of language queries, which requires distinguishing visually similar and recurring events by leveraging long video context. As temporal grounding demands highly specific timestamps, such **sparsity** and **ambiguity** challenge MLLMs to *precisely* comprehend long videos.

Recent advances in reinforcement learning (RL)-based grounding, such as TimeZero (Wang et al., 2025b), TVG-R1 (Chen et al., 2025), VideoChat-R1 (Li et al., 2025a), and Time-R1 (Wang et al., 2025a), demonstrate that reinforcement post-training can be a competitive and data-efficient alternative to traditional large-scale supervised fine-tuning (SFT) (Ma et al., 2023; Meinardus et al., 2024; Lu et al., 2024) for temporal grounding. However, precise grounding in long videos remains challenging. When extended to the long-video regime, existing RL strategies can still struggle with localization, as the increased search space and sparse rewards make it difficult for models to converge on precise timestamps.

This gap suggests that simply scaling short-video RL frameworks is insufficient for the complexities of long horizons. Adopting an RL approach for long-video temporal grounding requires a fundamental shift: the model must explore novel strategies for generating precise, structured timestamps to address event sparsity and ambiguity, while preserving its vast pre-trained knowledge for semantic perception and language understanding, which forms the foundation of its reasoning. Conventional

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

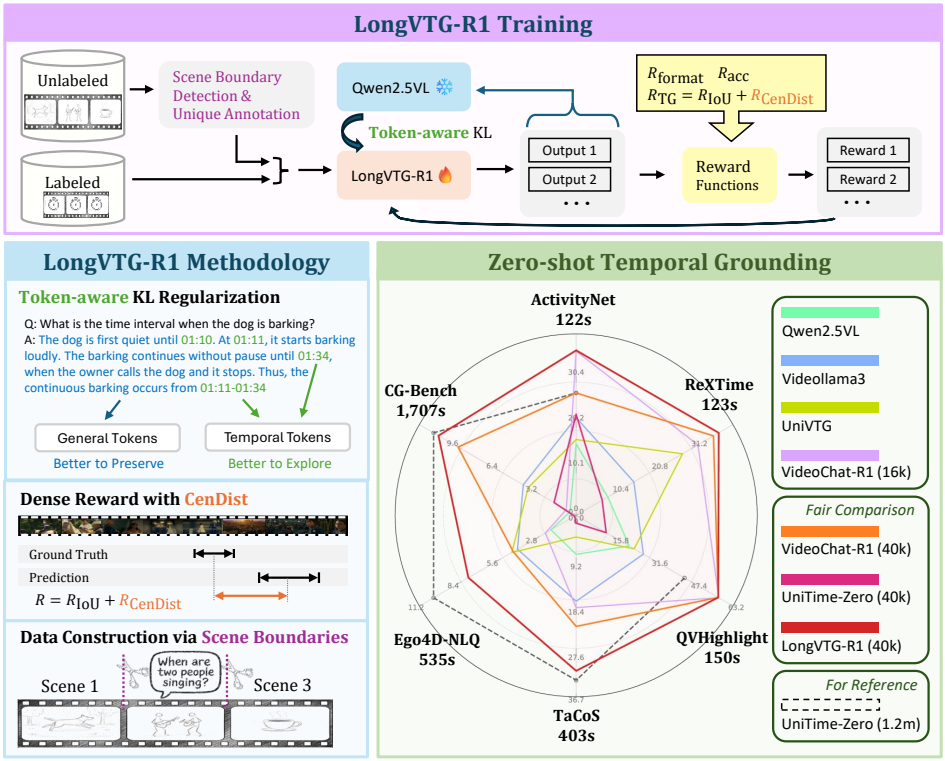

Figure 1: Overview of LongVTG-R1. **(Top)** To enable effective RL fine-tuning for long-video temporal grounding, we apply GRPO-based RL fine-tuning with our proposed token-aware KL regularization, and introduce CenDist reward to provide denser learning signals. **(Bottom-Left)** Our main technical contributions are: (i) token-aware KL regularization to encourage exploration of temporal tokens, (ii) dense reward design with CenDist to reduce reward sparsity, and (iii) the SceneTG dataset constructed via scene boundaries for scene-aligned grounding supervision. **(Bottom-Right)** Our LongVTG-R1 achieves performance comparable to the SFT-based UniTime-Zero while requiring 31x less training data, and improves long-video grounding over the matched-data VideoChat-R1 baseline.

KL-divergence regularization, as used in PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024), applies a uniform penalty coefficient across token types and can therefore constrain timestamp exploration. In temporal grounding as illustrated in Figure 1, *the exploration-critical tokens are clearly identifiable: those representing timestamps*. We therefore propose *Token-aware KL Regularization*, which selectively relaxes the penalty on numeric and structural tokens (*e.g.*, digits, separators, and tags) to encourage exploration, while maintaining stronger penalties on general language tokens to preserve core perceptual capabilities.

Besides the requirement of exploration, the inherent *sparsity* of the reward in temporal grounding further hinders MLLMs from receiving meaningful feedback during exploration toward precise grounding. This sparsity arises because the standard Intersection-over-Union (IoU) reward assigns zero reward for predictions that are temporally close but non-overlapping (Rezatofighi et al., 2019), which is a common scenario in long videos. To provide a denser learning signal, we introduce the *Center Distance Reward* (CenDist), an auxiliary reward that encourages predictions to approach ground truth segment centers. Beyond alleviating sparsity, we also enhance the quality of reward signals by addressing the *ambiguity* problem, particularly targeting the noisy patterns in existing datasets (Xia et al., 2023; Dong et al., 2025a; Zhang et al., 2025b) that often contain imprecise boundaries and queries matching multiple segments. To provide high-quality training data with minimized grounding ambiguity, we automatically construct the *SceneTG* dataset through a "propose-then-annotate" paradigm, which leverages scene boundary detection and a uniqueness filter to produce temporally sharp and less ambiguous data for RL. Although SceneTG only focuses on scene boundaries, it makes RL more effective and leads to improvements in real-world benchmarks.

Drawing insights from above, we present a model strong at long-video temporal grounding, named LongVTG-R1 (Figure 1), which also demonstrates generalization to long-video question-answering (QA) and multi-segment temporal grounding. In summary, our contributions are as follows:

- We propose token-aware KL regularization that improves exploration on timestamp digits while retaining regularization on language tokens to support general video QA ability. Although introduced in the context of long-video temporal grounding, our analysis suggests that it could potentially be applicable to other RL fine-tuning methods, and we hope it can benefit other tasks in future work.
- We introduce an auxiliary CenDist reward, which provides denser learning signals and alleviates the reward sparsity challenge in long-video temporal grounding.
- We design an automatic "propose-then-annotate" paradigm for constructing temporal grounding datasets, leading to the SceneTG dataset with sharp boundaries and less ambiguous queries that facilitate effective RL training.
- We demonstrate successful RL fine-tuning for long-video temporal grounding, with the resulting model, LongVTG-R1, achieving notable gains on long-video benchmarks, improving long-video QA in both ground-then-answer and unified settings, and transferring to multi-segment grounding under a simple recursive protocol without multi-segment fine-tuning.

2 Related Work

Video Temporal Grounding. Video Temporal Grounding (VTG) localizes temporal segments in untrimmed videos given natural-language queries. Early approaches largely rely on frozen visual-language features followed by regression or matching modules (Gao et al., 2017; Carreira & Zisserman, 2017; Devlin et al., 2019). More recent methods move toward end-to-end or generative formulations, including detection-style grounding with transformers, query-enrichment pipelines in video LLMs, and autoregressive MLLM-based timestamp generation (Lei et al., 2021; Wang et al., 2024a; Pramanick et al., 2025; Ren et al., 2024; Chen et al., 2024b; Guo et al., 2025; Zhang et al., 2025a). Existing MLLM VTG systems primarily target a single interval (Wang et al., 2024a; Zeng et al., 2024; Li et al., 2025a;b), while explicit timestamp-aware multi-segment grounding has only recently begun to receive dedicated treatment (Luo et al., 2025). Some recent tasks also couple answer prediction with temporal evidence rather than pure localization, such as visually grounded video QA and clue-grounded QA (Xiao et al., 2024; Chen et al., 2024a). Complementary long-video VTG methods further explore training-free hierarchical alignment or linear-time temporal compression outside the MLLM post-training setting (Xu et al., 2026; An & Grauman, 2025). Distance-aware objectives such as GIoU (Rezatofighi et al., 2019) and DIoU (Zheng et al., 2020) augment IoU with normalized distance terms, but in long videos, their enclosing-interval normalization can become weak.

Recent work also revisits supervision and RL design. TimeLens (Zhang et al., 2025b) studies annotation quality and thinking-free RL, OmniVTG (Zheng et al., 2026a) combines broader semantic coverage with self-correction training, and TaRO (Zheng et al., 2026b) uses caption-guided exploration and a temporal-sensitivity reward. We instead study selective KL regularization and dense localization feedback with scene-based supervision. Appendix Table 8 records their published results and evaluation differences.

Long-Video MLLMs. Long-video understanding has become a distinct subfield of video-language modeling. Recent systems extend MLLMs to long contexts through temporal compression, grounded tuning, hierarchical reasoning, or search-based mechanisms (Ren et al., 2024; Li et al., 2024; Zeng et al., 2024; Liu et al., 2025; Pan et al., 2025). Benchmarks such as LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025), LVBench (Wang et al., 2024b), and CG-Bench (Chen et al., 2024a) evaluate this regime explicitly. We draw on this perspective and treat long-video temporal grounding as more than a simple length extrapolation of standard VTG.

Reinforcement Fine-tuning for Video Understanding. Reinforcement learning with verifiable or rule-based rewards has recently improved reasoning in both LLMs and MLLMs (Christiano et al., 2017; Ouyang et al., 2022; Shao et al., 2024; Wen et al., 2025; Meng et al., 2025; Yan et al., 2025). In video, TimeZero (Wang et al., 2025b), Video-R1 (Feng et al., 2025), TVG-R1 (Chen et al.,

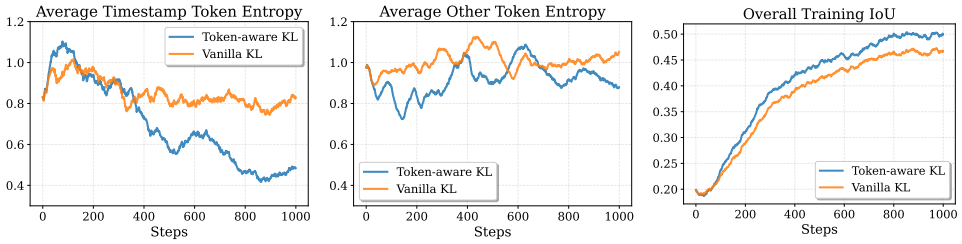

Figure 2: **Training dynamics under token-aware KL and vanilla KL.** Token-aware KL maintains higher entropy on timestamp-number tokens early in training while the entropy of other tokens remains comparatively stable. We report policy entropy as an empirical proxy for exploration, following DAPO (Yu et al., 2025).

2025), VideoChat-R1 (Li et al., 2025a), and Time-R1 (Wang et al., 2025a) show that GRPO-style post-training can improve temporal reasoning or grounding. However, as we later show in Tables 3, simply extending the strong VideoChat-R1 baseline with curated long-video training data still leaves substantial room. Recent VTG post-training recipes further emphasize the importance of high-quality data for stable optimization (Chen et al., 2025). Our work targets this gap directly by improving RL for long-video temporal grounding from both the algorithmic side and the data side.

3 Methodology

3.1 Preliminary

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is a reinforcement post-training method adapted to video grounding by VideoChat-R1 (Li et al., 2025a) and Time-R1 (Wang et al., 2025a). Given a video and query q, the rollout policy π_{old} samples G completions $o_i = (o_{i,1}, \ldots, o_{i,|o_i|})$. We update π_θ using these completions and a separate frozen reference π_{ref} for regularization. A rule-based reward has three components,

$$R = R_{\text{format}} + R_{\text{acc}} + R_{\text{tg}}, \tag{1}$$

where R_{format} rewards valid formatting, R_{acc} measures answer correctness for grounded-QA training samples such as NExT-GQA (Xiao et al., 2024), and R_{tg} measures localization quality. For pure grounding without answer labels, $R_{\text{acc}} = 0$. Group-normalizing the completion rewards gives one advantage $\hat{A}_i$ per completion, shared by all its tokens.

With the normalized advantages, GRPO optimizes the policy π_θ by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\{o_i\}_{i=1}^{G} \sim \pi_{\text{old}}(\cdot|q)} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \Big(\min\big(r_{i,t}(\theta)\hat{A}_i,\ \text{clip}(r_{i,t}(\theta),\, 1-\varepsilon,\, 1+\varepsilon)\hat{A}_i\big) - \beta\, D_{\text{KL}}\big(\pi_\theta(\cdot \mid q, o_{i,<t}) \,\|\, \pi_{\text{ref}}(\cdot \mid q, o_{i,<t})\big)\Big) \Bigg]. \tag{2}$$

where $r_{i,t}(\theta) = \pi_\theta(o_{i,t} \mid h_{i,t}) / \pi_{\text{old}}(o_{i,t} \mid h_{i,t})$ and $h_{i,t} = (q, o_{i,<t})$. The old policy is held fixed during the update; it is distinct from the KL reference. Suppressing the common prefix, the regularizer is

$$D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big) = \sum_{v \in \mathcal{V}} \pi_\theta(v) \log \frac{\pi_\theta(v)}{\pi_{\text{ref}}(v)}. \tag{3}$$

3.2 Token-aware KL Regularization

Longer video grounding increases the search space of timestamps. Early in training, most sampled timestamps in long videos miss the target interval entirely, so rewards are often zero or nearly uninformative. In this regime, exploration over timestamp-number tokens becomes especially important: if KL regularization keeps numeric outputs too tightly anchored to the reference model,

the policy rarely discovers better temporal hypotheses. The standard KL term in Equation 3 is effective for preventing abrupt policy drift, but its uniform penalty treats semantically different token subsets in the same way. We call a token subset *exploration-critical* if the pretrained policy has high uncertainty or low task accuracy on that subset, and if errors on that subset dominate task reward. In temporal grounding, timestamp-number tokens satisfy this definition: precise numeric outputs are not reliably produced by the pretrained model and directly determine localization reward.

This motivates relaxing KL on numeric outputs while retaining it on ordinary language. We first describe the fixed-subset regularizer that motivates this choice, then specify the completion-position mask used in training.

Fixed-subset definition. For a non-empty subset S of the token vocabulary and distributions p, q, define

$$D_{\mathrm{KL}}^{S}(p \,\|\, q) \;\triangleq\; q(S) - p(S) + \sum_{o \in S} \Big(p(o) \log \tfrac{p(o)}{q(o)} \Big). \tag{4}$$

This is the generalized KL divergence between the restrictions of p and q to S, including the mass correction $q(S) - p(S)$. Appendix F proves non-negativity and equality to zero exactly when the distributions match on S. These properties concern the regularizer; retention of downstream QA ability is assessed empirically.

Token-aware KL for Temporal Grounding. In implementation, a binary mask $m_{i,t}$ identifies numeric tokens throughout the completion and tokens forming opening/closing `think`, `answer`, and `timestamp` tags. It does not exempt all text enclosed by these tags. Let $\beta_{i,t} = (1 - m_{i,t})\beta_{\text{text}} + m_{i,t}\beta_{\text{time}}$, where β_{time} names the relaxed coefficient, including numeric tokens outside timestamps. At each sampled token, we replace the uniform KL penalty with

$$\mathcal{K}_{i,t} = \beta_{i,t}\big(e^{\Delta_{i,t}} - \Delta_{i,t} - 1\big), \qquad \Delta_{i,t} = \log \frac{\pi_{\text{ref}}(o_{i,t} \mid h_{i,t})}{\pi_{\theta}(o_{i,t} \mid h_{i,t})}. \tag{5}$$

The final setting uses $\beta_{\text{text}} = 0.04$ and $\beta_{\text{time}} = 0$. For a fixed subset and on-policy samples, this estimator includes the mass correction in Equation 4; Appendix D gives the identity and contextual masking rules. The actual completion-dependent mask is not a single fixed vocabulary partition. Figure 2 illustrates exploration dynamics, while Appendix Table 11 reports sensitivity to the relaxed coefficient.

3.3 Dense Reward Design with CenDist

Token-aware KL addresses reward sparsity from the exploration side by making timestamp search less conservative early in RL. We complement it with a denser temporal reward. IoU is the dominant evaluation metric in temporal grounding (Lei et al., 2021; Regneri et al., 2013), but it provides zero signal when the predicted segment is close to the target yet does not overlap with it. This near-miss case is common in long videos, especially early in training when relevant events occupy only a small portion of the full timeline.

Previous work such as GIoU (Rezatofighi et al., 2019) and DIoU (Zheng et al., 2020) augments overlap-based objectives with a distance term. In long videos, however, the relevant failure mode is specific: when both the predicted and ground-truth intervals are short and far apart, the enclosing interval is approximately determined by their separation, so the normalized distance term becomes close to a constant over many incorrect predictions. This weakens the reward signal precisely in the long-range regime where dense feedback is most needed. We instead use a duration-normalized center-distance reward, R_{CenDist}, which measures the distance between the centers of the predicted and ground-truth intervals and normalizes it by the full video duration. Let $[l_{\text{gt}}, r_{\text{gt}}]$ and $[l, r]$ be the ground-truth and predicted intervals, respectively:

$$R_{\text{CenDist}} = 1 - \frac{\left| \frac{l_{\text{gt}} + r_{\text{gt}}}{2} - \frac{l + r}{2} \right|}{\text{Duration}}. \tag{6}$$

The final reward is $R_{\text{tg}} = \alpha \cdot R_{\text{IoU}} + (1 - \alpha) \cdot R_{\text{CenDist}}$. The center term supplies feedback even without overlap; its empirical benefit need not increase monotonically with video duration.

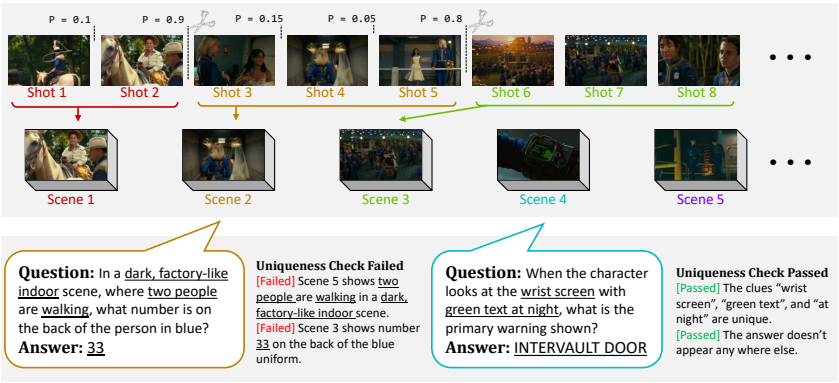

Figure 3: **Overview of SceneTG construction.** We begin with unlabeled videos and first apply ShotCoL to detect scene boundaries. Next, we generate question-answer pairs for each identified scene. Finally, we apply a uniqueness check to filter out queries that remain valid for multiple scenes.

3.4 SceneTG: Reducing Ambiguity in Long-Video Grounding Supervision

Existing grounding datasets often contain imprecise temporal boundaries, unclear queries, and repeated events that admit multiple plausible matches (Dong et al., 2025a; Gao et al., 2017; Dong et al., 2025b; Zhang et al., 2025b). These issues are especially problematic for RL, where ambiguous supervision can translate directly into noisy rewards.

To address this, we design SceneTG around two principles: **(i) boundary clarity** and **(ii) content uniqueness**. Guided by these principles, we introduce SceneTG, a novel "propose-then-annotate" paradigm for constructing less ambiguous temporal-grounding supervision from unlabeled long-form videos, as illustrated in Figure 3. The paradigm is fully automatic, and dataset-scale statistics are deferred to Appendix J.

Scene Boundaries as Candidates. To improve boundary clarity, we start from natural scene breaks rather than from free-form captions. This reverses the usual "caption-then-localize" workflow: we first identify candidate segments based on scene cuts and only then generate queries for them. Scenes are natural temporal units of long-form video, grouping multiple shots into a coherent block by location, characters, or intent. In practice, we reproduced ShotCol (Chen et al., 2021) and score shot transitions with a scene-boundary detector and merge shots between consecutive scene cuts to form candidate segments.

Uniqueness Filtering. For each proposed scene, we automatically generate question-answer pairs that highlight discriminative details such as unique objects, fine-grained actions, or context-specific interactions. We then apply a local uniqueness check against nearby scenes: retained pairs should neither have all question clues satisfied nor contain the answer in neighboring segments. Pairs that fail this test are discarded, so the retained queries are less likely to admit multiple valid temporal matches.

These steps produce less ambiguous data that is more constrained than open-world benchmarks and therefore yields cleaner rewards for RL optimization. To verify that this benefit comes from the annotation paradigm rather than from a particular video source, Appendix Table 13 and Figure 4 also evaluate *ActivityNet + OurAnno*, obtained by taking ActivityNet videos and re-annotating them with the same paradigm. Re-annotating ActivityNet already improves agreement over the original ActivityNet annotations. This supports our claim that the method is broadly applicable and reduces annotation ambiguity for training.

4 Experiments

Implementation Details. We build on Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and fine-tune it with GRPO on a mixture of Charades-STA (Gao et al., 2017), NExT-GQA (Xiao et al., 2024), and our SceneTG, following the mixed pure-grounding/grounded-QA recipe of VideoChat-R1 (Li et al.,

Table 1: Zero-shot temporal grounding performance (mIoU) on out-of-domain benchmarks. UniTime-Zero (1,260k samples) is shown only as a large-scale reference, and we additionally reproduce VideoChat-R1 and UniTime-Zero with the same 40k training samples when possible. Best results among the listed methods trained with approximately 40k samples or fewer are highlighted in bold; additional comparisons appear in Appendix Table 7.

Model	Training Data	Long Videos			Short Videos		
		TaCoS 403s	Ego4D 535s	CG-Bench 1707s	ActivityNet 122s	ReXTime 123s	QVHighlight 150s
Qwen2.5-VL (Bai et al., 2025)	–	7.9	1.8	0.5	15.9	8.4	21.4
VideoLLaMA 3 (Zhang et al., 2025a)	210k	17.4	4.2	4.3	21.8	15.3	27.1
UniVTG (Lin et al., 2023)	–	4.4	4.6	3.9	16.9	28.2	23.4
VTG-LLM (Guo et al., 2025)	–	5.3	1.4		17.9		8.6
VTimeLLM (Huang et al., 2024)	–	8.1	2.5	1.6	29.3	20.1	
TimeChat (Ren et al., 2024)	–	3.0	1.3		21.5	11.7	14.3
TimeSuite (Zeng et al., 2024)	349k	5.7	0.9	2.1			
VideoChat-R1 (Li et al., 2025a)	16k	18.7	2.2	0.8	36.6	32.3	**57.5**
LongVTG-R1	40k	**31.8**	**7.7**	**11.2**	**36.8**	**37.8**	57.3
VideoChat-R1 (fair comparison)	40k	22.5	4.5	9.6	27.3	36.4	57.3
UniTime-Zero (fair comparison)	40k	1.6	0.11	1.8	22.5	6.9	12.1
UniTime-Zero (Li et al., 2025b)	1,260k	33.4	10.2	11.6	27.3		43.7

2025a). Unless otherwise specified, the main model uses the approximately 40k-example setting and is trained for about 90 hours on 8 NVIDIA B200 GPUs. In the final model, we use the mixed IoU+CenDist reward with $\alpha = 0.8$ together with token-aware KL. Appendix D provides the full training mixture, prompts, reward parsing rules, and optimization settings.

4.1 Temporal Grounding Evaluation

We report mIoU on both in-domain and zero-shot out-of-domain grounding benchmarks. Table 2 shows that LongVTG-R1 outperforms all reported baselines on the in-domain datasets and improves over the matched-data VideoChat-R1 reproduction by 1.7 mIoU on Charades and 2.6 on NExT-GQA, using VideoChat-R1 as a strong RL grounding baseline.

Table 2: In-domain temporal grounding results (mIoU). The fair-comparison row reproduces VideoChat-R1 with the same 40k training mixture used by LongVTG-R1.

Model	In-domain Evaluation	
	Charades	NExT-GQA
Video-XL2-8B (Qin et al., 2025)	54.2	–
VideoLLaMA 3 (Zhang et al., 2025a)	59.9	–
Time-R1 (Wang et al., 2025a)	58.1	–
VideoChat-R1 (Li et al., 2025a)	60.8	32.4
VideoChat-R1 (fair comparison)	59.8	35.6
LongVTG-R1	**61.5**	**38.2**

On the out-of-domain benchmarks in Table 1, LongVTG-R1 achieves the best mIoU among the listed methods trained with approximately 40k samples or fewer on all three long-video benchmarks: TaCoS (31.8), Ego4D-NLQ (7.7), and CG-Bench (11.2). Relative to the matched-data VideoChat-R1 reproduction, the gains are +9.3, +3.2, and +1.6 mIoU, respectively. On shorter benchmarks, performance remains competitive while still improving over the matched-data baseline on ActivityNet and ReXTime. Compared with the 1.26M-sample UniTime-Zero reference, our model narrows the gap substantially on TaCoS and CG-Bench while using far less data. Appendix Table 7 additionally lists partial comparisons to TVG-R1 (Chen et al., 2025), ED-VTG (Pramanick et al., 2025), and HiTeA (Xu et al., 2026), whose papers report only subsets of benchmarks or thresholds. Recent source-reported results are summarized separately in Appendix Table 8.

Table 3 explicitly separates data curation from algorithm design. Adding SceneTG to VideoChat-R1 already improves transfer, especially on CG-Bench (0.8 $\rightarrow$ 9.2 at 16k). Increasing the curated data from 16k to 40k yields a smaller additional gain for the baseline. However, under the same 16k or 40k SceneTG mixtures, LongVTG-R1 still outperforms VideoChat-R1 by large margins. This indicates that SceneTG improves the training distribution, while

Table 3: Disentangling SceneTG data curation from RL algorithm design on zero-shot long-video grounding (mIoU). SceneTG improves baseline transfer, while LongVTG-R1 yields additional gains under the same SceneTG training mixtures.

Model	Training Setting	TaCoS	Ego4D	CG-Bench
VideoChat-R1 (Li et al., 2025a)	16k, w/o SceneTG	18.7	2.2	0.8
VideoChat-R1	16k, w/ SceneTG	21.2	4.4	9.2
VideoChat-R1	40k, w/ SceneTG	22.5	4.5	9.6
LongVTG-R1	16k, w/ SceneTG	28.9	7.1	10.3
LongVTG-R1	40k, w/ SceneTG	**31.8**	**7.7**	**11.2**

Table 4: Long-video QA accuracy (%) in direct, ground-then-answer, and unified settings. Non-reference scores show percentage-point deltas: blue/up for gains and red/down for declines; reference rows omit deltas. All deltas are relative to direct Qwen2.5-VL except direct UniTime-Zero (restricted), paired separately with Qwen2-VL as an older-backbone reference. Matched-data SFT uses the same Qwen2.5-VL backbone and grounding-data mixture as LongVTG-R1, with restricted decoding to separate forgetting from output-format collapse.

Setting	Ground Model	QA Model	LongVideoBench 473s	MLVU 651s	VideoMME (Long) 2,386s	LVBench 4,101s
Direct	–	Qwen2-VL	55.6	54.9	51.1	37.8
	–	UniTime-Zero (restricted)	26.0 ↓29.6	36.8 ↓18.1	30.1 ↓21.0	26.6 ↓11.2
	–	Qwen2.5-VL (Bai et al., 2025)	60.5	68.6	53.4	41.3
	–	Matched-data SFT (restricted)	48.5 ↓12.0	53.5 ↓15.1	44.2 ↓ 9.2	34.6 ↓ 6.7
	–	VideoChat-R1 (Li et al., 2025a)	59.2 ↓ 1.3	69.6 ↑ 1.0	54.3 ↑ 0.9	40.4 ↓ 0.9
	–	Time-R1 (Wang et al., 2025a)	58.0 ↓ 2.5	66.0 ↓ 2.6	54.4 ↑ 1.0	36.5 ↓ 4.8
	–	LongVTG-R1	58.7 ↓ 1.8	69.1 ↑ 0.5	55.0 ↑ 1.6	42.2 ↑ 0.9
Ground-then-answer	Qwen2.5-VL	Qwen2.5-VL	55.5 ↓ 5.0	55.8 ↓12.8	48.4 ↓ 5.0	33.3 ↓ 8.0
	UniTime-Zero		61.2 ↑ 0.7	70.7 ↑ 2.1	50.4 ↓ 3.0	42.9 ↑ 1.6
	VideoChat-R1		57.8 ↓ 2.7	65.6 ↓ 3.0	52.3 ↓ 1.1	42.6 ↑ 1.3
	Time-R1		53.4 ↓ 7.1	53.4 ↓15.2	44.1 ↓ 9.3	33.3 ↓ 8.0
	LongVTG-R1		63.1 ↑ 2.6	**70.8** ↑ 2.2	54.2 ↑ 0.8	47.1 ↑ 5.8
Unified	Qwen2.5-VL	Qwen2.5-VL	55.5 ↓ 5.0	55.8 ↓12.8	48.4 ↓ 5.0	33.3 ↓ 8.0
	VideoChat-R1	VideoChat-R1	58.6 ↓ 1.9	66.3 ↓ 2.3	52.9 ↓ 0.5	41.4 ↑ 0.1
	Time-R1	Time-R1	54.9 ↓ 5.6	56.7 ↓11.9	46.0 ↓ 7.4	33.9 ↓ 7.4
	LongVTG-R1	LongVTG-R1	**63.3** ↑ 2.8	70.6 ↑ 2.0	**54.9** ↑ 1.5	**47.5** ↑ 6.2

token-aware KL and CenDist improve how the model
learns from that distribution rather than only benefiting from the data alone.

4.2 General Long-Video QA Evaluation

Beyond temporal grounding, we examine whether reinforcement fine-tuning on grounding affects long-video question answering (QA). We evaluate on LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025), VideoMME (Long) (Fu et al., 2025), and LVBench (Wang et al., 2024b) in three settings: direct answering, ground-then-answer, and a unified single-model setup.

Direct QA and forgetting. We first ask whether a grounding-trained model can still answer directly from the full video. For clarity, Table 4 shows only the matched-data SFT result under restricted decoding; Appendix Table 10 reports the vanilla/restricted SFT trajectory on LongVideoBench. The colored deltas in Table 4 make the matched-data comparison explicit: relative to direct Qwen2.5-VL, restricted SFT drops by 12.0, 15.1, 9.2, and 6.7 points on LongVideoBench, MLVU, VideoMME, and LVBench, respectively. The older Qwen2-VL/UniTime-Zero pair shows a qualitatively similar forgetting pattern, again with uniformly negative deltas. All three RL methods improve on VideoMME and show smaller direct-QA changes than matched-data SFT in this evaluation. The difference among the three RL methods is relatively minor, while LongVTG-R1 demonstrates a slight advantage on the two longest benchmarks, VideoMME and LVBench.

Grounding as an intermediate step. We next test grounding as an explicit intermediate step: a grounding model localizes a relevant segment, and Qwen2.5-VL answers from that clip. LongVTG-R1 improves over direct Qwen2.5-VL on all four benchmarks, reaching 63.1 (+2.6) on LongVideoBench, 70.8 (+2.2) on MLVU, 54.2 (+0.8) on VideoMME, and 47.1 (+5.8) on LVBench, while outperforming the reproduced RL baselines.

Unified grounding and QA. Finally, we evaluate whether one model can localize and answer jointly. The unified LongVTG-R1 setting also yields positive deltas on all four benchmarks relative to direct Qwen2.5-VL (+2.8, +2.0, +1.5, and +6.2), slightly outperforms the two-model strategy on LongVideoBench, VideoMME, and LVBench, and remains comparable on MLVU. This indicates that the same policy can support both localization and answer generation while improving QA accuracy over direct answering in the evaluated settings.

Table 5: Zero-shot multi-segment grounding under the MUSEG (Luo et al., 2025) protocol on THUMOS (Idrees et al., 2017). We report average F1 (%) over IoU thresholds 0.1, 0.3, 0.5, and 0.7. All results are from our evaluations. Rows marked "+ recursive" use the recursive wrapper; unmarked rows do not. The matched-40k baseline and our policy share the training mixture and no-answer supervision; neither uses multi-segment fine-tuning.

Training Type	Model	THUMOS14	THUMOS15
Pretrained	Qwen2.5-VL-7B (Bai et al., 2025)	24.9	23.4
Single-segment Training	Video-R1 (Feng et al., 2025)	3.5	3.4
Single-segment Training	VideoChat-R1 (Li et al., 2025a)	14.3	13.4
Single-segment Training	VideoChat-R1 (40k) + recursive	28.62	26.85
Single-segment Training	LongVTG-R1 + recursive	**33.45**	**31.52**
Multi-segment Training	MUSEG-7B (Luo et al., 2025)	29.7	29.3

Table 6: Reduced-scale component ablations on four benchmarks. The final row uses $R_{\text{tg}} = 0.8R_{\text{IoU}} + 0.2R_{\text{CenDist}}$. Bold indicates the best result in each column.

Method	TaCoS		Ego4D-NLQ		NExT-GQA		CG-Bench	
	mIoU	R@0.3	mIoU	R@0.3	mIoU	R@0.3	mIoU	R@0.3
GRPO (VideoChat-R1)	18.0	27.3	3.7	4.7	34.6	52.9	7.77	9.27
GRPO w/o KL	17.7	26.8	3.6	4.6	34.7	54.0	7.55	9.50
+ Token-aware	18.4	27.8	4.9	6.1	34.8	54.4	7.82	9.84
+ Token-aware + GIoU	20.7	30.8	5.5	6.8	34.9	53.5	8.16	9.97
+ Token-aware + CenDist	**21.8**	**31.6**	**5.7**	**7.1**	**35.5**	**55.1**	**8.17**	**10.00**

4.3 Recursive Multi-Segment Grounding

We evaluate transfer from single- to multi-segment grounding without multi-segment fine-tuning. Our policy learns from single-segment examples and negative examples teaching a no-answer sentinel (Appendix D). At inference, it repeatedly predicts and removes one segment until no relevant evidence remains (Appendix I).

With the same 40k training mixture, no-answer supervision, and recursive protocol, LongVTG-R1 scores 33.45/31.52 on THUMOS14/15 versus VideoChat-R1's 28.62/26.85, gains of 4.83/4.67 F1 points (Table 5). Appendix Table 15 compares prior checkpoints and stopping judges. These results support the complete training recipe without isolating QA retention as the cause.

4.4 Ablation Study

To save compute, we ablate token-aware KL and CenDist using 30% of the training data and one-third of the original resolution. The full-scale long-video gap is larger: with the same 40k SceneTG mixture, LongVTG-R1 raises CG-Bench mIoU from VideoChat-R1's 9.6 to 11.2 (Table 3), with the gap continuing to widen during full-scale training (Appendix Figure 5).

Across four reduced-scale benchmarks (Table 6), token-aware KL improves TaCoS and Ego4D-NLQ mIoU over GRPO by 0.4 and 1.2 points. With token-aware KL fixed, the default IoU+CenDist mixture exceeds GIoU by 1.1 points on TaCoS, 0.2 on Ego4D-NLQ, and 0.6 on NExT-GQA, while remaining close on CG-Bench (8.17 versus 8.16). Appendix Tables 11 and 12 report parameter sensitivity.

5 Conclusion

We combine token-aware KL regularization, the dense CenDist reward, and SceneTG data construction to improve long-video temporal grounding while supporting direct and grounded QA. Without multi-segment fine-tuning, the resulting LongVTG-R1 also transfers to multi-segment grounding through recursive single-segment inference.

AI USE STATEMENT

We used AI tools to assist with writing and polishing the manuscript and with retrieval and discovery of related literature. An AI assistant also helped prepare the manuscript in the ICLR 2027 format. During revision, AI assistance was used to interpret review feedback, suggest experimental controls, and clarify mathematical notation and proof boundary cases. The authors supplied the additional experimental results. Separately, AI models are components of the research methodology, including the automated SceneTG annotation pipeline described in the method and Appendix J. The authors are responsible for the final content of this work.

ETHICS STATEMENT

Appendix B discusses potential benefits, surveillance and privacy risks, errors in temporal localization, and copyright and terms-of-use considerations for the video assets used in SceneTG. Appendix A describes limitations of the proposed approach.

REPRODUCIBILITY STATEMENT

Appendix D documents the training mixture, optimization settings, and reward implementation. Appendix F provides the mathematical proof, Appendix G lists evaluation prompts, and Appendix J describes the constructed data. Additional experimental results and controlled analyses appear in Appendices C and E. The underlying SceneTG videos are non-public; Appendix J documents this access limitation, and Table 13 evaluates the annotation procedure on public ActivityNet videos.

REFERENCES

Joungbin An and Kristen Grauman. Hieramamba: Video temporal grounding via hierarchical anchor-mamba pooling. *arXiv preprint arXiv:2510.23043*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. CG-Bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024a.

Ruizhe Chen, Tianze Luo, Zhiting Fan, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. Datasets and recipes for video temporal grounding via reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 983–992, 2025.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024b.

Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9796–9805, 2021.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Lu Dong, Haiyu Zhang, Han Lin, Ziang Yan, Xiangyu Zeng, Hongjie Zhang, Yifei Huang, Yi Wang, Zhen-Hua Ling, Limin Wang, et al. Videotg-r1: Boosting video temporal grounding via curriculum reinforcement learning on reflected boundary annotations. *arXiv preprint arXiv:2510.23397*, 2025a.

Lu Dong, Haiyu Zhang, Hongjie Zhang, Yifei Huang, Zhen-Hua Ling, Yu Qiao, Limin Wang, and Yali Wang. Weakly supervised temporal sentence grounding via positive sample mining. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025b.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

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*, pp. 5267–5275, 2017.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3302–3310, 2025.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

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.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025a.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. UniTime-Zero: Universal video temporal grounding with generative multi-modal large language models. *arXiv preprint arXiv:2506.18883*, 2025b.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*, 2025.

Weiheng Lu, Jian Li, An Yu, Ming-Ching Chang, Shengpeng Ji, and Min Xia. Llava-mr: Large language-and-vision assistant for video moment retrieval. *arXiv preprint arXiv:2411.14505*, 2024.

Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. Museg: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding. *arXiv preprint arXiv:2505.20715*, 2025.

Kaijing Ma, Xianghao Zang, Zerun Feng, Han Fang, Chao Ban, Yuhan Wei, Zhongjiang He, Yongxiang Li, and Hao Sun. Llavilo: Boosting video moment retrieval via adapter-based multimodal modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2798–2803, 2023.

Boris Meinardus, Anil Batra, Anna Rohrbach, and Marcus Rohrbach. The surprising effectiveness of multimodal large language models for video moment retrieval. *arXiv preprint arXiv:2406.18113*, 2024.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Junwen Pan, Rui Zhang, Xin Wan, Yuan Zhang, Ming Lu, and Qi She. Timesearch: Hierarchical video search with spotlight and reflection for human-like long video understanding. *arXiv preprint arXiv:2504.01407*, 2025.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24297–24308, 2025.

Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Video-xl-2: Towards very long-video understanding through task-aware kv sparsification. *arXiv preprint arXiv:2506.19225*, 2025.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. LVBench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024b.

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*, 2025a.

Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. TimeZero: Temporal video grounding with reasoning-guided lvlm. *arXiv preprint arXiv:2503.13377v1*, 2025b.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Kun Xia, Le Wang, Sanping Zhou, Gang Hua, and Wei Tang. Learning from noisy pseudo labels for semi-supervised temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10160–10169, 2023.

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*, pp. 13204–13214, 2024.

Xinyi Xu, Hongsong Wang, Guo-Sen Xie, Caifeng Shan, and Fang Zhao. Hitea: Hierarchical temporal alignment for training-free long-video temporal grounding. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, et al. Task preference optimization: Improving multimodal large language models with vision task alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29880–29892, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Temporal sentence grounding in videos: A survey and future directions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (8):10443–10465, 2023.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal LLMs. *arXiv preprint arXiv:2512.14698*, 2025b.

Minghang Zheng, Zihao Yin, Yi Yang, Yuxin Peng, and Yang Liu. OmniVTG: A large-scale dataset and training paradigm for open-world video temporal grounding. *arXiv preprint arXiv:2604.25276*, 2026a.

Minghang Zheng, Zihao Yin, Yi Yang, Yuxin Peng, and Yang Liu. Temporal-aware reasoning optimization for video temporal grounding. *arXiv preprint arXiv:2606.09248*, 2026b.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12993–13000, 2020.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13691–13701, 2025.

A Limitations and Future Work

A first limitation concerns scale. Due to computational constraints, our model is trained on a much smaller data regime than UniTime-Zero, which uses approximately 1.2 million samples. We therefore do not claim to match the highest-data setting, and future work should study how our method behaves under substantially larger-scale training.

A second limitation concerns generality. Although token-aware KL is applicable to discrete-action RL in principle, our empirical study is focused on temporal grounding. We do not claim that the same token partition or reward design will transfer unchanged to other domains, and broader validation is left to future work.

A third limitation concerns supervision scope. Although Section 4.3 shows strong zero-shot transfer to multi-segment grounding under recursive inference, our RL training still relies primarily on single-segment annotations and a simple sentinel-based test-time stopping protocol. Explicit multi-segment supervision could further improve recall, stopping behavior, and robustness on multi-event videos.

B Broader Impact

This work studies long-video temporal grounding and grounded question answering for multimodal language models. A potential positive impact is improved access to long-form video content: better temporal localization can help users search educational videos, audit long recordings more efficiently, retrieve relevant evidence for question answering, and reduce the amount of irrelevant video that must be manually reviewed. The main technical focus of the paper is also on data efficiency and reduced forgetting, which may lower the cost of adapting video-language systems for research use.

At the same time, stronger long-video grounding can be misused in sensitive settings. For example, a system that localizes queried moments in long surveillance footage, body-camera videos, or private archives could make targeted monitoring and large-scale indexing easier. Performance errors can also matter: overconfident but incorrect localization may cause users to overlook relevant evidence or trust the wrong video segment. In addition, our SceneTG pipeline uses long-form video assets and automated filtering, so any future release should consider copyright, terms of use, and whether portions of the source material require additional restrictions or redaction.

We therefore view this work as best suited for research on long-video understanding rather than for unconstrained high-stakes deployment. If code, models, or derived annotations are released in the future, a responsible release should document intended use, limitations, and known failure modes, and should consider access controls or other restrictions if the released assets materially increase misuse risk.

C Full Experimental Results

Table 7 reports the complete temporal grounding results across all benchmarks discussed in the paper. The overall pattern is consistent with the main-text summary: generic video-language models remain competitive on several shorter or easier settings, but their grounding quality drops substantially on long-video benchmarks such as Ego4D-NLQ and CG-Bench. By contrast, the RL-based grounding models improve markedly on these harder regimes, and LongVTG-R1 improves on VideoChat-R1 across TaCoS, Ego4D-NLQ, and CG-Bench, while the much larger-data UniTime-Zero remains a strong reference. We also include recent partial baselines such as TVG-R1, ED-VTG, and HiTeA when their papers report only a subset of datasets or thresholds (Chen et al., 2025; Pramanick et al., 2025; Xu et al., 2026).

Recent published baselines. Table 8 records results reported by TimeLens, OmniVTG, and TaRO, separately from our evaluations. TimeLens entries use original benchmark annotations, not its refined TimeLens-Bench. Its ActivityNet split is described as test in the source, whereas Table 7 uses val; we therefore keep these source-reported results separate. Training data, backbone versions, and evaluation implementations also differ. No mIoU is inferred from recall scores, and these results are not matched-data reproductions.

Table 7: Full temporal grounding results. Panel (a) reports ActivityNet, ReXTime, and QVHighlight; panel (b) reports TaCoS, Ego4D-NLQ, and CG-Bench. Each method occupies one row per panel, with the metric specified in each column. Dashes denote unavailable comparable results. In particular, Time-R1's ActivityNet results are R@0.3/R@0.5/R@0.7, with no mIoU reported here. HiTeA reports R@0.1/R@0.3/R@0.5, so its R@0.7 cells are unavailable.

(a) Short-video benchmarks

Method	ActivityNet (val)				ReXTime				QVHighlight (val)			
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
Qwen2.5-VL (Bai et al., 2025)	15.9	20.9	12.1	6.5	8.36	10.8	6.84	4.67	21.4	27.6	17.0	9.23
VideoLLaMA 3 (Zhang et al., 2025a)	21.8	28.2	15.6	7.5	15.3	19.3	9.23	4.8	27.1	29.7	17.9	11.6
UniVTG (Lin et al., 2023)	16.86	11.1	4.06	–	28.17	41.34	26.88	–	23.38	14.13	4.42	–
Mr. BLIP (Meinardus et al., 2024)	–	–	–	–	–	–	–	–	–	–	–	–
VTG-LLM (Guo et al., 2025)	17.86	12.32	6.74	–	–	–	–	–	8.6	4.13	1.55	–
VTimeLLM (Huang et al., 2024)	29.3	–	–	–	20.14	28.84	17.41	–	–	–	–	–
TimeChat (Ren et al., 2024)	21.53	16.38	8.36	–	11.65	14.42	7.61	–	14.25	8.32	4.26	–
TimeMaker (Chen et al., 2024b)	22.03	16.56	9.28	–	–	–	–	–	21.3	12.32	9.16	–
TimeSuite (Zeng et al., 2024)	–	–	–	–	–	–	–	–	–	–	–	–
ED-VTG (Pramanick et al., 2025)	35.2	52.1	33.1	16.0	–	–	–	–	–	–	–	–
HiTeA (Xu et al., 2026)	–	–	–	–	–	–	–	–	–	–	–	–
UniTime-Zero (Li et al., 2025b)	27.31	22.77	14.14	–	–	–	–	–	43.71	41.03	31.48	–
TVG-R1 (Chen et al., 2025)	–	–	–	–	28.20	41.04	24.54	11.07	–	–	–	–
Time-R1 (Wang et al., 2025a)	–	58.6	39.0	21.4	–	–	–	–	–	–	–	–
VideoChat-R1 (Li et al., 2025a)	36.6	33.4	–	17.7	32.3	42.2	32.7	23.3	57.5	78.0	62.3	44.3
LongVTG-R1	36.8	53.2	34.0	18.0	37.8	50.0	36.1	25.8	57.3	75.7	59.3	44.5

(b) Long-video benchmarks

Method	TaCoS				Ego4D-NLQ (val)				CG-Bench			
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
Qwen2.5-VL (Bai et al., 2025)	7.92	11.8	5.97	1.67	1.83	1.83	1.01	0.59	0.52	0.57	0.33	0.23
VideoLLaMA 3 (Zhang et al., 2025a)	17.4	24.8	14.2	5.6	4.24	6.27	2.84	1.26	4.25	5.43	2.8	1.03
UniVTG (Lin et al., 2023)	4.4	5.17	1.27	–	4.63	6.48	3.48	–	3.86	–	–	–
Mr. BLIP (Meinardus et al., 2024)	17.94	24.59	14.32	–	5.37	6.49	3.2	–	–	–	–	–
VTG-LLM (Guo et al., 2025)	5.27	6.87	2.92	–	1.36	1.71	0.46	–	–	–	–	–
VTimeLLM (Huang et al., 2024)	8.08	9.3	3.9	–	2.52	1.67	0.77	–	1.58	–	–	–
TimeChat (Ren et al., 2024)	2.95	3.77	1.6	–	1.3	1.67	0.79	–	–	–	–	–
TimeMaker (Chen et al., 2024b)	–	–	–	–	–	–	–	–	–	–	–	–
TimeSuite (Zeng et al., 2024)	5.71	6.75	2.5	–	0.94	0.88	0.43	–	2.09	–	–	–
ED-VTG (Pramanick et al., 2025)	12.7	14.5	6.0	2.3	–	–	–	–	–	–	–	–
HiTeA (Xu et al., 2026)	19.79	29.08	16.10	–	8.12	10.39	6.04	–	–	–	–	–
UniTime-Zero (Li et al., 2025b)	33.38	50.06	31.54	–	10.18	14.67	7.38	–	11.63	–	–	–
TVG-R1 (Chen et al., 2025)	–	–	–	–	–	–	–	–	–	–	–	–
Time-R1 (Wang et al., 2025a)	26.33	39.56	24.71	12.00	2.06	2.68	1.30	0.63	3.69	4.61	2.86	1.63
VideoChat-R1 (Li et al., 2025a)	18.7	28.6	16.2	6.17	2.18	2.68	1.29	0.57	0.82	1.03	0.47	0.2
LongVTG-R1	31.8	46.4	31.6	16.1	7.71	10.0	6.22	2.99	11.15	13.97	7.23	3.03

Table 8: Recent source-reported VTG results (%). Panel (a): original ActivityNet and QVHighlights annotations; panel (b): long-video results from TaRO. TimeLens numbers come from its Tables 8–9, OmniVTG from Table 2, and TaRO from Tables 1 and 3. Model size/version is explicit; unavailable metrics are dashes. These are literature references under the source papers' protocols, not new evaluations by us.

(a) ActivityNet and QVHighlights

Method	ActivityNet				QVHighlights			
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
TimeLens-7B (Zhang et al., 2025b)	37.7	53.5	35.2	19.7	57.4	78.4	63.6	43.7
OmniVTG-7B (Zheng et al., 2026a)	–	60.3	39.8	21.4	–	82.8	67.0	47.3
TaRO (Qwen2.5-VL-7B) (Zheng et al., 2026b)	–	60.6	39.8	21.4	–	82.6	69.4	48.8
TaRO (Qwen3-VL-8B) (Zheng et al., 2026b)	–	59.7	40.6	23.9	–	82.7	68.5	51.7

(b) TaCoS and Ego4D-NLQ

Method	TaCoS				Ego4D-NLQ			
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
TaRO (Qwen2.5-VL-7B) (Zheng et al., 2026b)	–	37.9	22.5	8.60	–	8.19	4.55	1.93
TaRO (Qwen3-VL-8B) (Zheng et al., 2026b)	–	52.3	36.5	22.5	–	13.80	7.86	4.17

D Training Details and Reproducibility Notes

This section summarizes the concrete training configuration used for LongVTG-R1 and the reduced-scale ablations. We keep the details that materially affect reproduction and omit machine-specific filesystem paths from our internal training environment.

Training mixtures. The final full-scale LongVTG-R1 run corresponds to the "40k" setting in the main text and uses 41,713 traces after dataset-level caps and shuffling. Table 9 aggregates this mixture

by source family. Positive and negative traces are both included; negative traces use the no-answer sentinel described below. The reduced-scale ablations use a 12,000-trace mixture with the same overall source pattern but smaller caps (3,780 pure-grounding SceneTG traces, 1,020 grounded-QA SceneTG traces, 4,800 Charades-STA traces, and 2,400 NExT-GQA traces).

Table 9: Training mixture used by the final full-scale LONGVTG-R1 run. We refer to this 41,713-trace mixture as the "40k" setting in the main text for brevity.

Source family	Output type	Traces
SceneTG (pure grounding)	timestamp only	18,002
SceneTG (grounded QA)	answer + timestamp	6,745
Charades-STA	timestamp only	13,208
NExT-GQA	answer + timestamp	3,758
Total	mixed	41,713

Optimization and video preprocessing. We initialize from Qwen2.5-VL-7B-Instruct and train with GRPO for one epoch using DeepSpeed ZeRO-3 with CPU offloading, bf16, gradient checkpointing, FlashAttention 2, AdamW, per-device batch size 1, and gradient accumulation 1 on 8 GPUs. Each prompt samples 8 completions at temperature 1.0, so the GRPO group size is 8. The main run uses the standard GRPO learning rate of 1×10^{-6}, maximum prompt/completion lengths of 8192/1024 tokens, and data seed 42. We sample up to 768 frames across the full input video (512 in reduced-scale ablations). For inputs longer than 384 seconds (256 in ablations), we linearly rescale the model-facing time axis to that duration, with training annotations transformed accordingly and predicted timestamps mapped back for evaluation. This coordinate rescaling does not truncate the video coverage. The final run also uses a three-times larger pixel budget than the reduced-scale ablations, matching the "one-third resolution" description in Section 4.4.

Training prompts and no-answer handling. Each example first inserts the query into a short dataset-specific stem: raw event text for pure-grounding SceneTG traces, a Charades localization stem for Charades-STA, or a grounded-QA stem of the form "Locate the video clip that is relevant to the question: ..." for NExT-GQA and grounded-QA SceneTG traces. The trainer then appends one of the following shared output protocols.

Training prompt for pure temporal grounding.

```
[EVENT]

Output your thought process within the <think> </think> tags.

Then, provide the start and end times (in seconds, precise to one decimal
↪  places) in the format "start time to end time" within the <timestamp>
↪  </timestamp> tags.
```

Training prompt for grounded QA with temporal evidence.

```
[EVENT]

Output your thought process within the <think> </think> tags and a
↪  capital letter indicating the best option within the <answer>
↪  </answer> tags.

Then, provide the timestamp of video clip that support your answer in the
↪  format "start time to end time" (in seconds, precise to one decimal
↪  places) within the <timestamp> </timestamp> tags.
```

To teach the no-answer sentinel, all negative examples and a random 50% of the remaining training prompts append the suffix

```
If no relevant video clip exists, indicate this by returning "<timestamp>
↪  -1 to -2 </timestamp>".
```

This sentinel is only a training-time and recursive-inference option; the standard single-segment evaluation setting in the paper assumes that at least one relevant clip exists.

Reward parsing and aggregation. We parse the last `<timestamp>` block in each sampled completion and extract the last numeric pair matching "x to y" (or "x and y"). For positive examples,

R_{IoU} is the interval IoU and $R_{\text{CenDist}} = \max(0, 1 - |c_{\text{pred}} - c_{\text{gt}}|/\text{Duration})$, where c_{pred} and c_{gt} are the predicted and ground-truth segment centers. Reversed intervals receive a small penalty of -0.1 under the temporal reward. For negative examples, reversed intervals are treated as valid no-answer outputs and receive 0.5 under both temporal rewards, while ordinary positive-length intervals receive 0. The format reward checks for the required `<think>` and `<timestamp>` structure, and grounded-QA examples additionally require an `<answer>` block. QA accuracy is computed from the first alphabetic character inside `<answer>`, so "A" and "(A)" are treated equivalently.

In the final full-scale LongVTG-R1 configuration, the scalar reward is

$$R = 1.0\, R_{\text{format}} + 0.8\, R_{\text{IoU}} + 0.2\, R_{\text{CenDist}} + 1.0\, R_{\text{acc}},$$

where R_{acc} is used only for grounded-QA examples and is zero otherwise. Rewards are normalized within each group of 8 sampled completions before the GRPO update. Auxiliary indicators such as whether the ground truth is positive/negative and whether the model predicted a positive/negative interval are logged for diagnostics but are not weighted in the final scalar reward.

Token-aware KL in code. The final implementation applies a KL coefficient of 0.04 to unmasked completion positions and zero to exempt positions. The exemption covers numeric tokens throughout the completion, including numbers in reasoning text, rather than only numbers inside timestamp tags. Decimal points and minus signs are included when identified as part of a number. Tokens forming the opening and closing `<think>`, `<answer>`, and `<timestamp>` tags are also exempt; the tag exemption does not extend to their enclosed text.

For a sampled token v at prefix h, let $\Delta = \log \pi_{\text{ref}}(v \mid h) - \log \pi_\theta(v \mid h)$. The implemented KL penalty is $\exp(\Delta) - \Delta - 1$, multiplied by the position's KL coefficient. Thus the estimator includes the likelihood-ratio correction, rather than only the log-ratio term. For positive distributions p, q and a fixed retained subset S, its expectation satisfies

$$\mathbb{E}_{v\sim p}\left[\mathbf{1}_{\{v\in S\}}\left(\frac{q(v)}{p(v)} - \log\frac{q(v)}{p(v)} - 1\right)\right] = q(S) - p(S) + \sum_{v\in S} p(v)\log\frac{p(v)}{q(v)}.$$

This identity explains the linear correction in Equation 4 for a fixed subset. The actual implementation identifies numeric and structural-tag positions from the generated completion; this operational mask should be distinguished from a single fixed vocabulary partition.

E Additional Controlled Analyses

This section collects controlled analyses of training, regularization, reward design, and annotation quality. Table 10 isolates the effect of direct SFT on the same backbone and data mixture. Tables 11 and 12 provide sensitivity analyses for token-aware KL and the mixed temporal reward, respectively. Table 13 and Figure 5 further clarify the roles of SceneTG quality and training scale.

Controlled SFT degrades the pretrained model in two distinct ways. First, vanilla decoding can collapse in format after enough SFT updates, which explains the very low vanilla scores at 3.2k and 11.2k. Second, even after removing this formatting confound with restricted decoding, Table 10 still shows a monotonic drop from 60.5 to 48.5 on LongVideoBench, indicating genuine loss of long-video QA ability rather than only output-format failure. The main QA table reports the final restricted-decoding comparison across all four direct-QA benchmarks, where the same checkpoint also reaches 53.5 on MLVU, 44.2 on VideoMME, and 34.6 on LVBench. Together these are the controlled comparisons referred to in the main text when we state that RL better preserves QA ability in our setting.

We next examine whether the benefit of token-aware KL is tied to a single hyperparameter choice. Table 11 shows a consistent trend as β_{time} is reduced while β_{text} is fixed: relaxing KL on timestamp-number tokens leaves NExT-GQA roughly unchanged but steadily improves CG-Bench, where timestamps account for a much smaller fraction of the full video duration. This is consistent with the intended role of timestamp-number tokens as the subset that most needs additional freedom during RL optimization.

Table 10: Controlled SFT on LongVideoBench using the same Qwen2.5-VL backbone and training mixture. "Restricted" decoding forces the model to output one uppercase answer letter, which separates format collapse from actual QA degradation.

Model	Vanilla	Restricted
Qwen2.5-VL	60.5	60.5
SFT 1.6k	58.2	58.2
SFT 3.2k	0.15	54.5
SFT 11.2k	4.7	48.5

Table 11: β_{time} sensitivity for token-aware KL. Reducing KL on timestamp-number tokens consistently improves grounding, with the clearest gains on the longer CG-Bench benchmark.

β_{text}	β_{time}	CG-Bench		NExT-GQA	
		mIoU	R@0.3	mIoU	R@0.3
0.04	0.04	7.77	9.27	34.6	52.9
0.04	0.02	7.78	9.43	34.8	54.1
0.04	0.00	7.82	9.84	34.8	54.4

Table 12: α sensitivity for the mixed temporal grounding reward. Lowering α increases the weight on CenDist, which helps the longer CG-Bench benchmark but slightly hurts the shorter NExT-GQA benchmark.

Reward	CG-Bench		NExT-GQA	
	mIoU	R@0.3	mIoU	R@0.3
+ Token-aware + GIoU	8.16	9.97	34.9	53.5
+ Token-aware + CenDist ($\alpha = 0.8$)	8.17	10.00	35.5	55.1
+ Token-aware + CenDist ($\alpha = 0.5$)	8.42	11.33	34.6	53.1

Table 12 complements the reduced-scale ablation in the main paper by making the reward interpolation explicit. The default setting uses the mixed reward R_{tg} with $\alpha = 0.8$, not CenDist alone. Lowering α increases the contribution of CenDist and yields a clearer benefit on the longer CG-Bench benchmark, especially in R@0.3, while slightly hurting the shorter NExT-GQA benchmark. These results illustrate a benchmark-dependent trade-off in the reward mixture.

Finally, we provide two additional analyses for SceneTG. Table 13 reports the human verification study used to motivate our ambiguity claim. Two annotators localized the relevant temporal interval from each video–query pair without access to the dataset ground truth. Each annotator watched each video at least twice. To test whether the annotation method is broadly applicable rather than tied to our movie/TV source videos, we also apply the same propose-then-annotate pipeline to ActivityNet videos and denote the result as *ActivityNet + OurAnno*. Relative to the original ActivityNet annotations on the same video pool, this re-annotation substantially improves both inter-annotator agreement and alignment between human judgments and dataset ground truth. SceneTG further achieves the highest human–GT IoU in the long-video regime it targets. Figure 4 visualizes the human–GT IoU comparison and highlights the same trend. Together they support reduced annotation ambiguity under the evaluated human-verification protocol.

Figure 5 addresses a different question: whether the modest reduced-scale ablation gaps understate the final algorithmic difference. The answer is yes. As more curated SceneTG data are used, both VideoChat-R1 and LongVTG-R1 improve on CG-Bench, but the margin between them also widens. This indicates that the benefit of our algorithmic changes is not a small-data artifact; instead, the advantage becomes more visible as the long-video RL setting is scaled up.

F Proof of the Properties of Token-aware KL

Definition. For any non-empty subset S of the token space and discrete distributions p, q on the same domain, define the *Token-aware KL*

$$D_{\text{KL}}^{S}(p \,\|\, q) \;\triangleq\; q(S) - p(S) + \sum_{o \in S} \Big(p(o) \log \tfrac{p(o)}{q(o)} \Big). \tag{7}$$

Basic properties.

Proposition 1. *For any non-empty S and any discrete p, q:*

1. *(*Non-negativity*)* $D_{\text{KL}}^{S}(p \,\|\, q) \geq 0$.

Table 13: Human verification of annotation ambiguity and pipeline generality. *ActivityNet + OurAnno* is obtained by taking ActivityNet videos and re-annotating them with the same propose-then-annotate pipeline used for SceneTG. SceneTG is designed to reduce ambiguity for RL training rather than to claim perfect event boundary detection. We report inter-annotator agreement and alignment between human annotations and dataset ground truth.

Dataset	Avg Duration	Avg GT/Duration	IoU (HumanA, HumanB)	IoU (Human, GT)
SceneTG	319.7s	7.51%	71.2	79.0
ActivityNet + OurAnno	153.2s	7.13%	76.0	76.5
ActivityNet	153.2s	7.46%	60.9	58.8

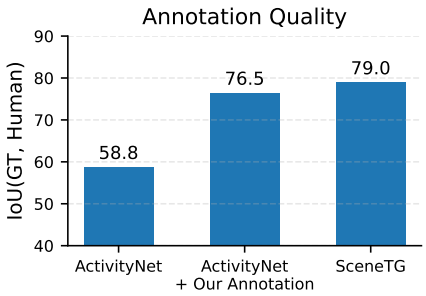

Figure 4: **Human–GT annotation quality.** Applying our annotation pipeline to ActivityNet videos already improves substantially over the original ActivityNet annotations without changing the underlying video pool, and SceneTG further achieves the highest human–GT IoU in our long-video setting.

2. *(*Identity on S*)* $D_{\mathrm{KL}}^{S}(p \,\|\, q) = 0$ *if and only if* $p(o) = q(o)$ *for all* $o \in S$.

Proof. Fix q and write $m \triangleq p(S) \in [0,1]$ and $q_S \triangleq q(S) \in [0,1]$. We use the conventions $0 \log(0/q) = 0$ (including $q = 0$) and $p \log(p/0) = +\infty$ for $p > 0$. If $m = 0$, the divergence equals $q_S \geq 0$, with equality exactly when $q_S = 0$; this is precisely agreement on S. It remains to consider $m > 0$. On S, factor the *shape* of p and q from their total mass:

$$r(o) \triangleq \frac{p(o)}{m}, \quad o \in S, \qquad s(o) \triangleq \frac{q(o)}{q_S}, \quad o \in S,$$

where r and (when $q_S > 0$) s are distributions on S. We first consider the case $q_S > 0$ and then treat the boundary case $q_S = 0$.

Case 1: $q_S > 0$**.** Using $p(o) = m\, r(o)$ and $q(o) = q_S\, s(o)$ for $o \in S$,

$$\begin{aligned}
\sum_{o \in S} p(o) \log \frac{p(o)}{q(o)} &= \sum_{o \in S} m\, r(o) \Big(\log m + \log r(o) - \log q_S - \log s(o) \Big) \\
&= m \log \frac{m}{q_S} \;+\; m \sum_{o \in S} r(o) \log \frac{r(o)}{s(o)} \\
&= m \log \frac{m}{q_S} \;+\; m\, \mathrm{KL}(r \| s).
\end{aligned}$$

Therefore

$$D_{\mathrm{KL}}^{S}(p \| q) = \underbrace{\left(q_S - m + m \log \tfrac{m}{q_S} \right)}_{\phi_{q_S}(m)} + m\, \mathrm{KL}(r \| s). \tag{8}$$

For fixed m, the second term is minimized uniquely at $r = s$ with value 0. Hence

$$\inf_{p|_S :\, p(S) = m} D_{\mathrm{KL}}^{S}(p \| q) \;=\; \phi_{q_S}(m) = q_S - m + m \log \tfrac{m}{q_S}.$$

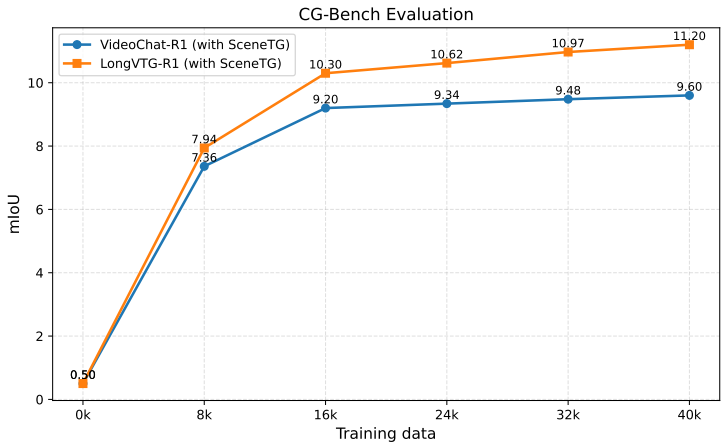

Figure 5: **CG-Bench scaling with SceneTG.** Both VideoChat-R1 and LongVTG-R1 improve as more SceneTG data is used, but the gap between them also widens. This indicates that the algorithmic advantage of LongVTG-R1 becomes more pronounced as scale increases.

Now minimize $\phi_{q_S}(m)$ over $m \in [0, 1]$. A direct calculation shows

$$\phi'_{q_S}(m) = \log \frac{m}{q_S}, \qquad \phi''_{q_S}(m) = \frac{1}{m} > 0 \ \ (m > 0),$$

so ϕ_{q_S} is strictly convex on $(0, 1]$ with the unique minimizer at $m = q_S$, where $\phi_{q_S}(q_S) = 0$. Combining the two steps, the minimizing restriction of p to S is uniquely determined by

$$m^\star = q_S \quad \text{and} \quad r^\star = s \iff p(o) = q(o) \ \ \forall\, o \in S,$$

and the minimum value is 0. Since (8) is a sum of a nonnegative divergence and a strictly convex nonnegative function of m, it follows that $D^S_{\mathrm{KL}}(p\|q) \geq 0$, with equality if and only if $p|_S = q|_S$.

Case 2: $q_S = 0$**.** Then $q(o) = 0$ for all $o \in S$. If $p(o) > 0$ for some $o \in S$, the logarithmic term is $+\infty$ and hence $D^S_{\mathrm{KL}}(p\|q) = +\infty > 0$. If instead $p(o) = 0$ for all $o \in S$ (equivalently $m = 0$), every term in (4) is 0, so $D^S_{\mathrm{KL}}(p\|q) = 0$. Thus non-negativity holds, and equality occurs iff $p|_S = q|_S$ (both identically zero on S).

The two cases together prove both claims. □

G Evaluation Prompts

The evaluation tasks in this paper follow two output protocols: pure temporal grounding and grounded QA with temporal evidence. We therefore organize the prompt templates by output format rather than only by dataset name.

Evaluation Prompt for Pure Temporal Grounding. Used for Charades-STA, ActivityNet, QVHighlight, TaCoS, and Ego4D-NLQ.

```
To accurately pinpoint the event "[EVENT]" in the video, determine the
↪  precise time period of the event.

Output your thought process within the <think> </think> tags.

Then, provide the start and end times (in seconds, precise to one decimal
↪  place) in the format "start time to end time" within the <timestamp>
↪  </timestamp> tags.
```

For Ego4D-NLQ, [EVENT] is the natural-language query.

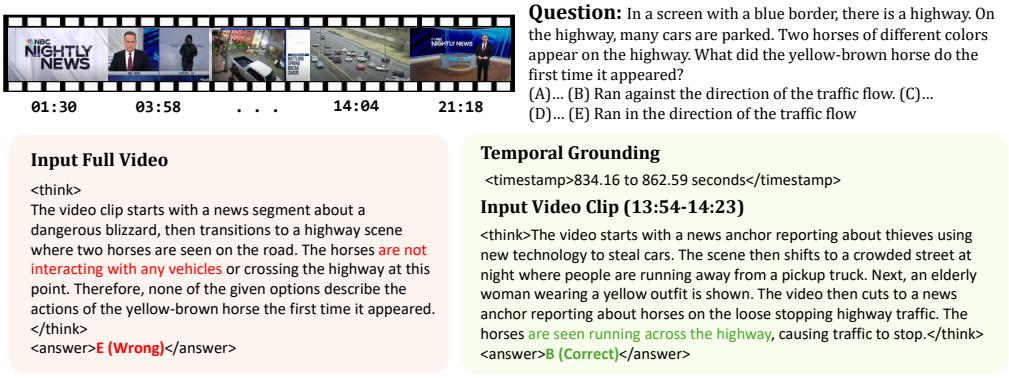

Figure 6: **Qualitative analysis for unified grounding and QA.** The input video lasts over 20 minutes and includes various news segments. Directly answering from the full video (left) yields an incorrect answer. Grounding first (right) localizes the relevant clip (13:54–14:23), where the horses are running across the highway, and enables the correct answer.

Evaluation Prompt for Grounded QA with Temporal Evidence. Used for NExT-GQA, ReXTime, and CG-Bench; the same dual-output structure is also used for the grounded-QA portion of our training mixture.

```
Locate the video clip that is relevant to the question:
↪  "[QUESTION][OPTIONS (if applicable)]"

Output your thought process within the <think> </think> tags and a
↪  capital letter indicating the best option within the <answer>
↪  </answer> tags.

Then, provide the timestamp of the video clip that supports your answer
↪  in the format "start time to end time" (in seconds, precise to one
↪  decimal place) within the <timestamp> </timestamp> tags.
```

The recursive multi-segment prompt differs from the standard single-segment setting because it introduces an explicit stop sentinel. We therefore defer it to Appendix Section I, which also summarizes the rollout protocol. With matched training data, no-answer supervision, and recursive inference, LongVTG-R1 outperforms VideoChat-R1 by 4.83 and 4.67 F1 points on THUMOS14 and THUMOS15, respectively (Table 15).

H Qualitative Analysis

We include two qualitative case studies to complement the quantitative results. Figure 6 shows how grounding improves long-video QA by localizing the relevant clip before answering, while Figure 7 highlights cleaner temporal boundary alignment for single-segment grounding. Displayed timestamps refer to the original video timeline.

Figure 6 presents a representative long-video QA example in which direct answering from the full video is distracted by irrelevant context, whereas grounding first isolates the key evidence and enables the correct answer. The video lasts more than 20 minutes and contains multiple news segments; after localizing the relevant interval (13:54–14:23), the model focuses on the clip where the horses are running across the highway and answers correctly. This qualitative behavior matches the quantitative improvements of LongVTG-R1 in the ground-then-answer and unified settings in Table 4.

Figure 7 presents a representative example from ActivityNet that illustrates the comparative behavior of our LongVTG-R1 and the baseline VideoChat-R1 on the temporal grounding task. This case highlights the intrinsic difficulty of grounding textual descriptions in videos containing repeated or visually similar events, which can easily mislead models toward incorrect temporal spans. While both methods must reason over duplicate actions, the baseline exhibits temporal drift across scene

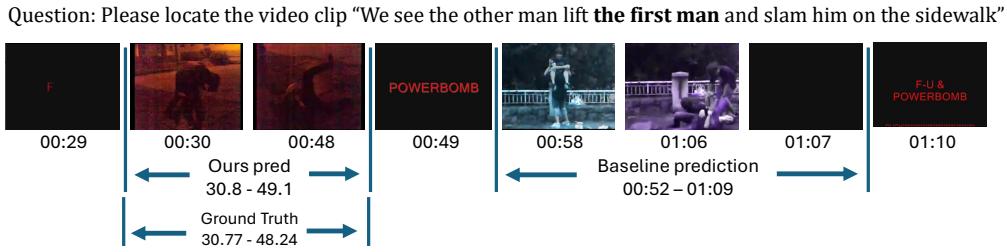

Figure 7: **Temporal Grounding Qualitative Analysis.** Comparison between LONGVTG-R1 and VideoChat-R1 on an ActivityNet example, highlighting (1) the difficulty posed by duplicate events, (2) the baseline's drift beyond scene boundaries versus our precise alignment, and (3) our model's preference for the first event when resolving "first vs. last" ambiguity.

boundaries, extending its predicted segment into regions where the screen has already turned black and end-screen text appears—evidence of misalignment with visual structure. In contrast, our model terminates before the cut and shows stronger sensitivity to scene transitions. When the query distinguishes between the "first" and "last" occurrence of an event, LONGVTG-R1 also selects the human-annotated first instance, whereas the baseline drifts to a later repetition. These examples illustrate stronger robustness to repeated events, scene cuts, and fine-grained temporal references.

I MULTI-SEGMENT DETAILS

This section details the recursive protocol in Section 4.3. Table 15 compares prior checkpoints, alternative stopping judges, and a matched-40k VideoChat-R1 trained with the same data and no-answer supervision as LONGVTG-R1. These comparisons evaluate transfer without multi-segment fine-tuning; they do not isolate QA retention as its cause. In the recursive example, timestamps within each round refer to the remaining video, while the final summary reports intervals on the original video timeline.

Evaluation Prompt for Recursive Multi-Segment Grounding (THUMOS14/15).

For recursive multi-segment evaluation, we reuse the single-segment grounding setting round by round, but explicitly allow the model to stop when no relevant moment remains in the current video. This stop sentinel is used only in the recursive multi-segment setting; the standard grounding evaluations in the paper always assume that at least one relevant clip exists.

```
Locate one remaining video clip that matches the event "[EVENT]" in the
↪  current video.

Output your thought process within the <think> </think> tags.

If a relevant clip remains, provide the start and end times (in seconds,
↪  precise to one decimal place) in the format "start time to end time"
↪  within the <timestamp> </timestamp> tags.

If no relevant clip remains, instead return "<timestamp> -1 to -2
↪  </timestamp>".
```

Figure 8 illustrates this interpretation qualitatively. LONGVTG-R1 identifies both relevant segments and stops once the remaining video no longer contains matching evidence, whereas VideoChat-R1 makes an incorrect first prediction and then keeps drifting into later rounds. This example illustrates localization and stopping behavior within the recursive inference protocol.

Under the matched training and recursive inference conditions in Table 15, VideoChat-R1 reaches 28.62/26.85 on THUMOS14/15, compared with 33.45/31.52 for LONGVTG-R1, a gain of 4.83/4.67 F1 points. Thus an advantage remains after matching training data and no-answer supervision. The hybrid rows separately characterize the effect of replacing the stopping judge. These results support

Table 14: Pseudo-code for the simple recursive multi-segment protocol used in Section 4.3. The same rollout is applied to all methods in Table 15.

Input: video V, query q, grounding model f, maximum rounds T
Initialize: remaining video $V^{(1)} \leftarrow V$, predicted segments $\mathcal{P} \leftarrow \emptyset$
  maintain a map M_t from the current video's timeline to the original timeline
for $t = 1, \ldots, T$ **do**
  query the model on the current video: $y_t \leftarrow f(V^{(t)}, q)$
  parse one predicted segment $\hat{s}_t$ from y_t
  if $\hat{s}_t = [-1, -2]$, **break**
  express $\hat{s}_t$ in seconds of the current video
  map both endpoints to the original timeline: $s_t \leftarrow M_t(\hat{s}_t)$
  append s_t to $\mathcal{P}$
  remove the predicted interval, obtaining $V^{(t+1)}$; update M_{t+1}
return $\mathcal{P}$

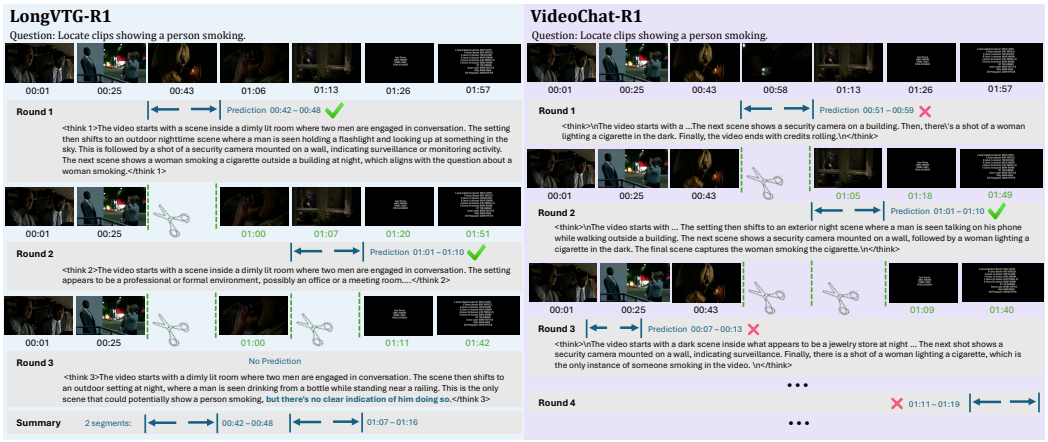

Figure 8: Qualitative comparison for multi-segment grounding via recursive prediction. LONGVTG-R1 identifies two relevant smoking segments and then correctly stops when no further evidence remains, while VideoChat-R1 makes an incorrect first prediction and continues drifting in later rounds instead of terminating.

the complete training recipe, without attributing the gains to QA retention or any one component alone.

J SCENETG STATISTICS

Figure 9 summarizes the key statistics of the SceneTG dataset. Starting from 20k unlabeled long-form videos, our propose-then-annotate pipeline generates roughly five candidate scenes per video (about 100k candidates in total) and retains 24k final annotations after uniqueness filtering. Most videos range from 100 to 1,000 seconds in length, and the annotated ground-truth segments constitute 3%–10% of the total video length. These statistics complement the main-text methodology discussion by quantifying the construction yield and the long-video regime targeted by SceneTG.

Source and availability. The source videos come from a non-public internal video collection. Source-level identifiers cannot be disclosed, and the underlying videos are not publicly released. We report aggregate statistics and the annotation procedure to document this setting. The same annotation model is used for question–answer generation and neighboring-scene uniqueness filtering. To assess the procedure on a public video source, Table 13 also compares original ActivityNet annotations with annotations produced by the same pipeline on the same video pool. This comparison measures annotation agreement, rather than downstream training gains.

Table 15: Recursive multi-segment grounding: average F1 (%) over IoU thresholds 0.1, 0.3, 0.5, and 0.7. All rows use the same recursive protocol and are evaluated by us. The matched-40k VideoChat-R1 uses the same training mixture, including no-answer examples and supervision, as LONGVTG-R1; other VideoChat-R1 rows use the prior checkpoint. "Same model" uses the proposer for stopping as well.

Segment proposer	**Continue/stop judge**	**THUMOS14**	**THUMOS15**
Prior single-model recursive baselines			
Video-R1 (Feng et al., 2025) + recursive	same model	11.9	14.4
VideoChat-R1 (Li et al., 2025a) + recursive	same model	5.26	4.96
Hybrid rollouts with pretrained continue/stop judge			
VideoChat-R1 (Li et al., 2025a) + recursive	Qwen2.5-VL-7B (Bai et al., 2025)	23.82	24.64
Time-R1 (Wang et al., 2025a) + recursive	Qwen2.5-VL-7B (Bai et al., 2025)	22.33	24.22
LONGVTG-R1 + recursive	Qwen2.5-VL-7B (Bai et al., 2025)	29.76	29.75
Matched training data and no-answer supervision			
VideoChat-R1 (matched 40k) + recursive	same model	28.62	26.85
LONGVTG-R1 + recursive	same model	**33.45**	**31.52**

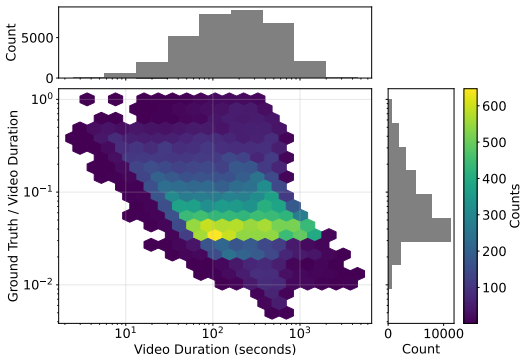

Figure 9: **SceneTG statistics** of video durations with ground-truth density.

Table 16: Existing assets used in this work. We report the license or terms of use according to the official project pages or dataset/model cards when available.

Asset	Use	License / terms	Notes
Qwen2.5-VL-7B-Instruct	Backbone	Apache-2.0	Used as the base MLLM; model card lists Apache-2.0.
Charades-STA / Charades	Train / eval	Non-Commercial License (Allen Institute for AI)	Charades license restricts redistribution and requires respectful treatment of subjects; Charades-STA is derived from Charades.
NExT-GQA	Train / eval	MIT License	Used for grounded QA; we cite the original paper and repository and follow the released data terms.
SceneTG	New data	Non-public internal source	Source videos and identifiers are not publicly released; aggregate statistics and the annotation procedure are described above.
ActivityNet / ActivityNet-Captions	Eval / re-annotation	CC BY 4.0 (Metadata) / Terms of Use (Videos)	Used for VTG evaluation and ActivityNet+OurAnno.
TaCoS	Eval	Non-Commercial / MPII TaCoS License	Used for temporal grounding evaluation.
Ego4D-NLQ / Ego4D	Eval	Ego4D License Agreement	Used for temporal grounding evaluation.
CG-Bench	Eval	CC BY-NC-SA 4.0	Used for temporal grounding evaluation.
QVHighlights	Eval	MIT License	Used for temporal grounding evaluation.
ReXTime	Eval	CC BY-NC-SA 4.0	Used for temporal grounding evaluation.
LongVideoBench	Eval	CC BY-NC-SA 4.0	Used for long-video QA evaluation.
MLVU	Eval	CC BY-NC-SA 4.0	Used for long-video QA evaluation.
Video-MME	Eval	CC BY-NC-SA 4.0	Used for long-video QA evaluation.
LVBench	Eval	CC BY-NC-SA 4.0	Used for long-video QA evaluation.
THUMOS14 / THUMOS15 protocol	Eval	Academic / Non-Commercial Use	Used for recursive multi-segment grounding evaluation.