A BIMODAL PERSPECTIVE ON VIDEO GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video temporal grounding requires localizing events within videos based on natural language queries. Although temporal Intersection over Union (tIoU) is the de facto metric for both evaluation and Reinforcement Learning from Verifiable Rewards (RLVR), we show that applying it uniformly across event durations creates two previously underexplored failure modes. First, for short events, small absolute boundary shifts cause disproportionately large changes in tIoU, so temporally close predictions can receive misleadingly low scores. Second, when tIoU is directly optimized as an RLVR reward, a model can increase its expected reward by inflating predicted intervals; consequently, tIoU-based RLVR improves long-event localization while degrading short-event performance. These observations motivate a duration-aware framework that treats short and long events as distinct temporal regimes. We evaluate them separately with distance-based and overlap-based criteria, respectively. For training, we combine distance-IoU with boundary accuracy to provide smoother supervision for short events while preserving effective overlap supervision for long events and discouraging interval inflation. We further introduce a multi-rollout query sampler that uses duration-aware success criteria to retain hard-yet-solvable examples, together with BiVTG-100K, a refined dataset with improved coverage and annotation quality for short events. Experiments on the manually re-annotated Charades, ActivityNet, and QVHighlights benchmarks show that the proposed framework consistently improves duration-aware grounding and mitigates the short-event failures of tIoU-based RLVR.

1 INTRODUCTION

Video temporal grounding (Gao et al., 2017; Grauman et al., 2022; Krishna et al., 2017; Lei et al., 2021; Oncescu et al., 2021; Regneri et al., 2013) is a fundamental problem in video understanding: given a natural-language query, the goal is to localize the corresponding temporal segment in a video. It underpins a wide range of downstream applications, including agentic video question answering (Zeng et al., 2026b; Yang et al., 2025; Shi et al., 2026; Zhang et al., 2025b; Lin et al., 2026; Ye et al., 2025; Meng et al., 2025b), procedural planning (Song et al., 2023; Zhou et al., 2018; Chang et al., 2020; Li et al., 2024b), and egocentric assistance (Tian et al., 2025; Grauman et al., 2022).

Recent Multimodal Large Language Models (MLLMs) are reshaping temporal grounding from specialized architectures (Lei et al., 2021; Lin et al., 2023; Mu et al., 2024; Gordeev et al., 2026; Moon et al., 2023) toward unified video-language formulations (Meinardus et al., 2024; Guo et al., 2024; Ren et al., 2024; Huang et al., 2024; Zeng et al., 2024; Wu et al., 2025; Chen et al., 2024; Li et al., 2025d; Ge et al., 2025; Li et al., 2025a). Within this trend, reinforcement learning, particularly Reinforcement Learning from Verifiable Rewards (RLVR) (Zhang et al., 2026; Wang et al., 2025; Chen et al., 2025; Yue et al., 2025; Li et al., 2025c), has emerged as a way to improve temporal reasoning by directly optimizing task-level outcomes. Its effectiveness, however, depends critically on whether the verifiable reward faithfully measures localization quality.

Despite recent progress, existing approaches use temporal Intersection over Union (**tIoU**), or minor variants, for both training reward signal and evaluation metric across widely varying event durations. However, as illustrated in Figure 1, a small temporal shift can eliminate overlap, assigning near-correct predictions for short events. Motivated by this limitation, we propose to evaluate short- and long-event localization performance separately, using criteria suited to their respective temporal regimes, *i.e.*, absolute boundary errors for short events and interval overlap for long events.

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.

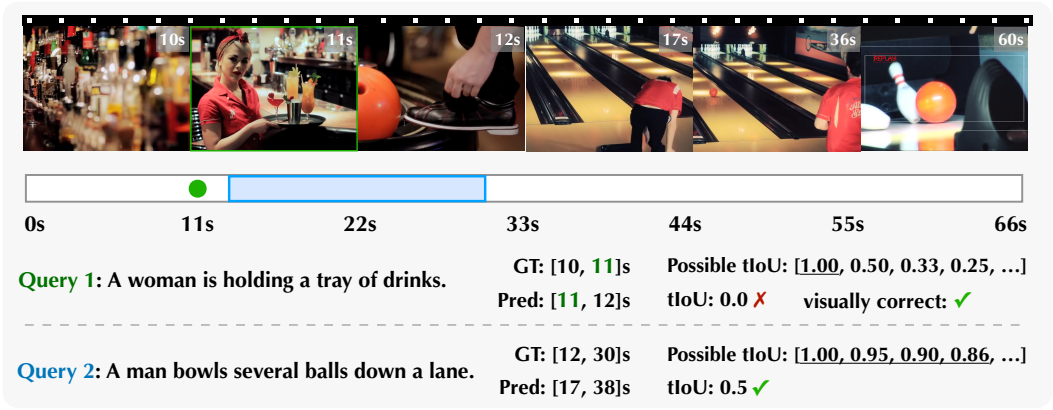


Figure 1: **Duration-dependent limitations of tIoU.** Unlike long events spanning extended **temporal intervals**, short events may occur within only a few frames or near a specific **timestamp**. For such events, small temporal shifts can sharply reduce tIoU, leading to misleading evaluation and biased hard sample selection. When used as an RLVR reward, tIoU can further encourage interval inflation, a form of reward hacking, which increases overlap possibility without improving localization precision.

Our second observation concerns RLVR training: when tIoU is used directly as a reward, models can exploit it by predicting excessively long intervals. Such interval inflation increases the likelihood of non-zero overlap and can therefore improve the expected tIoU reward without enhancing grounding precision, as shown in Figure 2. We therefore conduct experiments to find an optimized reward shaping that combines distance-IoU (dIoU) with a boundary-error term, jointly accounting for interval overlap, center alignment, and boundary precision while discouraging interval inflation.

Training sample selection presents a related challenge. Prior methods (Zhang et al., 2026; Wang et al., 2025) retain low-tIoU queries from a single greedy decoding pass, often selecting unsolvable cases with uniformly poor rollouts and little reward variance for Group Relative Policy Optimization (GRPO). Our multi-rollout sampler instead selects *hard-yet-solvable* queries with both successful and failed predictions, using distance-based success criteria for short events and tIoU for long events.

Finally, we find that these algorithmic issues are compounded by data quality. To support duration-aware RLVR, we introduce **BiVTG-100K**, a refined dataset that removes audio-dependent shortcuts, resolves ambiguous queries, and improves coverage of underrepresented short-event instances. Built on this rectified pipeline, our model **BiVTG-8B** achieves state-of-the-art performance on the manually re-annotated TimeLens-Bench across Charades, ActivityNet, and QVHighlights.

Our contributions are threefold: **(i) Duration-dependent diagnosis.** We are the first to show that tIoU can mischaracterize near-correct short-event predictions and induce interval-inflation reward hacking under RLVR. **(ii) Duration-aware reward shaping and sampling.** Motivated by the diagnosis, we introduce stratified evaluation, a hybrid dIoU and boundary error reward shaping, and a multi-rollout sampler that selects hard-yet-solvable queries with regime-appropriate success criteria. **(iii) Refined data and strong grounding performance.** We propose the **BiVTG-100K** and show that **BiVTG-8B** consistently improves duration-aware localization on three manually re-annotated benchmarks.

2 METHOD

We first review RL-based temporal grounding, then analyze its duration-dependent limitations and introduce bimodal reward shaping, sample selection, and data refinement.

2.1 PRELIMINARIES

Task Formulation. Given a video $\mathcal{V}$ and a language query q describing an event occurring in the video, the goal of video temporal grounding is to localize the corresponding temporal interval $\hat{\mathcal{T}} = [\hat{t}_{\text{start}}, \hat{t}_{\text{end}}]$ that best matches the queried event.

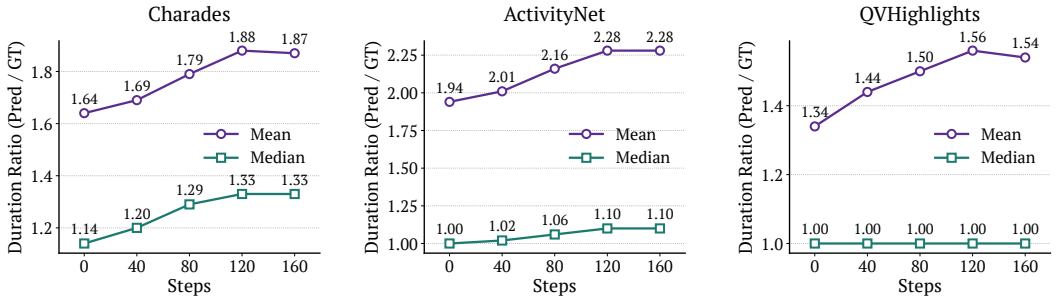


Figure 2: Median and mean ratios of predicted-to-ground-truth span duration across three benchmarks during RLVR training with tIoU reward. The model consistently predicts longer intervals, especially for short events and short videos of Charades, revealing a reward hacking behavior induced by tIoU.

Training Pipeline. Recent approaches (Zhang et al., 2026; Wang et al., 2025) tackle this task by adapting Multimodal Large Language Models (MLLMs) through a two-stage training pipeline. In the first stage, the model is fine-tuned on an instruction-tuning dataset $\mathcal{D}_{\text{SFT}}$ to acquire basic temporal grounding capability. In the second stage, the model is further optimized via Reinforcement Learning with Verifiable Rewards (RLVR), typically implemented using Group Relative Policy Optimization (GRPO) (Shao et al., 2024).

Concretely, at each RL step, a batch of queries is sampled from a curated dataset $\mathcal{D}_{\text{RL}} \subset \mathcal{D}_{\text{SFT}}$, and G candidate predictions $\{o_i\}_{i=1}^{G}$ are generated per query. The advantage $\hat{A}_i$ for each prediction is computed by normalizing its reward within the group, and the training objective maximizes the normalized advantage while regularizing deviation from the reference:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}\hat{A}_i - \beta\,\mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]\right].$$

Reward and Evaluation Protocol. A key foundation in existing methods is to adopt temporal Intersection over Union (tIoU) as a unified measure for both training signal and evaluation protocol. Given a predicted interval $\hat{\mathcal{T}}$ and ground-truth interval $\mathcal{T} = [t_{\text{start}}, t_{\text{end}}]$, tIoU is defined as:

$$\text{tIoU}(\mathcal{T}, \hat{\mathcal{T}}) = \frac{\hat{\mathcal{T}} \cap \mathcal{T}}{\hat{\mathcal{T}} \cup \mathcal{T}} = \frac{\max\left(0,\ \min(\hat{t}_{\text{end}}, t_{\text{end}}) - \max(\hat{t}_{\text{start}}, t_{\text{start}})\right)}{\max(\hat{t}_{\text{end}}, t_{\text{end}}) - \min(\hat{t}_{\text{start}}, t_{\text{start}})}.$$

During the RLVR stage, tIoU directly serves as the reward signal, providing supervision for temporal alignment beyond token-level correctness. For evaluation, a prediction is typically considered acceptable if tIoU $\geq \lambda$ (*e.g.*, $\lambda = 0.7$), and mean tIoU (mIoU) is computed by directly averaging tIoU scores over all samples, forming the basis of standard benchmarking protocols.

2.2 Rethinking Temporal Grounding from a Bimodal Perspective

Temporal grounding instances span a broad range of durations, from transient moments lasting only a few frames to activities that extend over tens of seconds. As illustrated in Fig. 1, the query *"A woman is holding a tray of drinks"* corresponds to a highly localized event around the 11-second mark with a duration of roughly one second, whereas *"A man bowls several balls down a lane"* occupies a much longer interval from 12 to 30 seconds. This variation in temporal scale is not incidental; it fundamentally affects how localization quality should be measured and optimized.

A central consequence is that tIoU is poorly behaved for short events. When the ground-truth interval is brief, even a small temporal offset can eliminate overlap entirely. For example, predicting $[11, 12]$ for a ground truth $[10, 11]$ yields tIoU $= 0$, despite being temporally close and semantically plausible. Under tIoU, such near misses are treated identically to completely incorrect predictions. By contrast, for long events, the same absolute boundary deviation typically induces only a modest change in overlap, so tIoU remains comparatively smooth and informative.

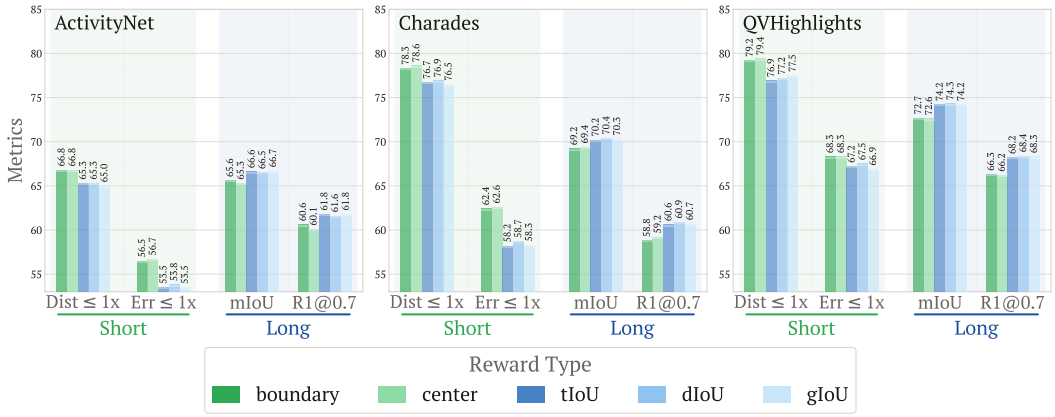


Figure 3: Comparison of the effect of **point-based** and **interval-based** rewards on short and long events. Point-based rewards (boundary, center) improve short event pinpointing, whereas interval-based rewards (tIoU, dIoU, gIoU) enhance long event localization.

This scale-dependent behavior has implications beyond evaluation. In the standard SFT-to-RLVR pipeline, it induces two coupled failure modes, which we further address in Sec. 2.3 and Sec. 2.4.

(i) Reward collapse and hacking. For short events, tIoU sparsity collapses reward variance and weakens GRPO optimization, while simultaneously encouraging over-extended predictions that increase the chance of non-zero overlap. **(ii) Unreliable hard sample estimation.** Using low-tIoU samples as "hard" examples becomes noisy under this regime, as semantically reasonable near misses are often assigned near-zero scores, thereby biasing the estimation of difficulty and model capability.

2.3 Bimodal Reward Shaping for RLVR Training

Distance-Based Metrics for Short Events. To better characterize short-event localization, we introduce distance-based metrics that remain informative even when interval overlap is unstable. Let

$$c = (t_{\text{start}} + t_{\text{end}})/2, \qquad \hat{c} = (\hat{t}_{\text{start}} + \hat{t}_{\text{end}})/2$$

denote the centers of the ground-truth and predicted intervals, respectively. We define the center distance and boundary error as

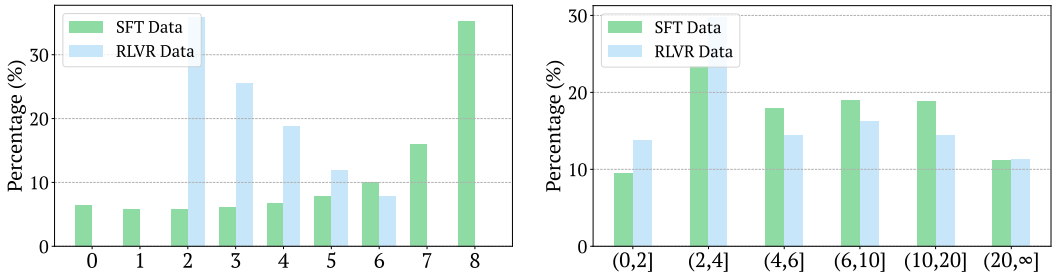
$$\text{Dist} = |\hat{c} - c|, \qquad \text{Err} = |\hat{t}_{\text{start}} - t_{\text{start}}| + |\hat{t}_{\text{end}} - t_{\text{end}}|. \tag{1}$$

A short-event prediction is considered acceptable when both Dist and Err fall below predefined thresholds, analogous to thresholding tIoU at $\lambda = 0.7$ for long events.

Reward Shaping Investigation. Most prior RLVR methods (Zhang et al., 2026; Wang et al., 2025; Li et al., 2025c; Chen et al., 2025) use tIoU or closely related variants as the training reward. Beyond its sparsity on short events, we find that directly optimizing such overlap-based rewards systematically encourages temporal over-expansion. As shown in Fig. 2, the mean and median ratios between predicted and ground-truth durations on the test split, *i.e.*, $|\hat{\mathcal{T}}|/|\mathcal{T}|$, increase consistently throughout RLVR training. This reflects a form of reward hacking: by producing longer intervals, the model raises the chance of overlapping with the ground-truth span and thus receiving non-zero reward. While beneficial under the training metric, this strategy substantially harms short-event localization. Prior work largely overlooks this failure mode because evaluation is not stratified by temporal scale.

These observations motivate reward functions that provide smoother and more scale-sensitive supervision. We consider two classes of rewards: *interval-based* rewards, which extend overlap-based objectives with additional geometric structure, and *point-based* rewards, which directly supervise boundaries and centers without requiring overlap.

Interval-based Rewards. To preserve the reward signal under non-overlap and penalize center misalignment, we extend Generalized IoU (Rezatofighi et al., 2019) and Distance IoU (Zheng et al.,



(a) Accepted rollout counts of 8 total rollouts.

(b) Event duration distribution (seconds).

Figure 4: The solvability and event duration distribution of queries in the SFT and RLVR datasets.

2020) to the temporal domain:

$$\text{gIoU} = \text{tIoU} - \frac{|\mathcal{C}| - |\mathcal{U}|}{|\mathcal{C}|}, \qquad \text{dIoU} = \text{tIoU} - \frac{(\hat{c} - c)^2}{|\mathcal{C}|^2},$$

where $|\mathcal{C}| = \max(\hat{t}_{\text{end}}, t_{\text{end}}) - \min(\hat{t}_{\text{start}}, t_{\text{start}})$ is the length of the smallest enclosing interval and $|\mathcal{U}|$ is the temporal union. Both rewards are normalized to $[0, 1]$.

Point-based Rewards. To supervise boundary precision and center alignment independently of overlap, we define

$$r_{\text{boundary}} = \frac{1}{2}\exp\left(-\alpha\varepsilon_{\text{start}}^2\right) + \frac{1}{2}\exp\left(-\alpha\varepsilon_{\text{end}}^2\right), \qquad r_{\text{center}} = \frac{1}{1 + \left(\varepsilon_{\text{center}}/\gamma\right)^2},$$

where

$$\varepsilon_{\text{start}} = \frac{|\hat{t}_{\text{start}} - t_{\text{start}}|}{|\mathcal{T}|}, \quad \varepsilon_{\text{end}} = \frac{|\hat{t}_{\text{end}} - t_{\text{end}}|}{|\mathcal{T}|}, \quad \varepsilon_{\text{center}} = \frac{|\hat{c} - c|}{|\mathcal{T}|}$$

are errors normalized by the ground-truth duration, where α and γ are decay parameters.

Key Findings. Fig. 3 reveals a clear regime-dependent pattern once short and long events are evaluated separately: point-based rewards are more effective for short-event localization, whereas interval-based rewards are better suited for long events. Notably, training with interval-based rewards alone degrades short-event performance relative to the SFT baseline, consistent with the reward-hacking behavior discussed above and analyzed further in Sec. 3.3. Based on this observation, our final model uses a weighted sum of dIoU and boundary rewards, jointly constraining interval overlap, boundary precision, and center alignment, while reducing the reward hacking and reward collapse.

2.4 Bimodal-Aware Sample Selection for RLVR Training

Prior methods adopt a **one-pass**, **tIoU-based hard filtering** strategy (Zhang et al., 2026; Wang et al., 2025), where samples are selected based on a single greedy prediction and the tIoU metric that is unreliable for short events. This practice is sub-optimal, as it fails to capture the capability boundary of MLLMs as probabilistic models. Consequently, many low-tIoU queries identified as "hard" may fail to produce any acceptable rollouts even under repeated sampling, rendering them uninformative for GRPO training.

In contrast, we propose a novel **multi-rollout, partial-accept sampling** strategy under a **hybrid criterion** to construct the RLVR training subset $\mathcal{D}_{\text{RL}}$. Our design is guided by two principles: **1) metric consistency**, where acceptability is defined using duration-aware criteria that remain informative for both short and long events; and **2) algorithm alignment**, where sample selection explicitly preserves queries that induce meaningful reward variance across rollouts by retaining only *occasionally solvable* queries with both accepted and rejected rollouts.

Multi-Rollout Hybrid Criterion. To correctly identify acceptable rollouts across events of varying duration, we sample G rollouts per query q over the entire SFT dataset $\mathcal{D}_{\text{SFT}}$, and apply duration-

Table 1: **Statistics of training and evaluation datasets.** Vid. Avg.: video duration (s); Evt.: event duration (s); Std.: standard deviation; P50/P95: 50th/95th percentile of ascending event duration. The large P50–P95 gap highlights that there are both short and long events in video temporal grounding.

Split	Source	#Videos	#Events	Vid. Avg.	Vid. Std.	Evt. Avg.	Evt. P50	Evt. P95
Train	`cosmo_cap` (Wang et al., 2024a)	9,517	42,344	122.9	42.0	12.2	7.0	39.0
	`didemo` (Anne Hendricks et al., 2017)	2,751	11,542	57.9	27.4	12.1	6.0	43.0
	`hirest` (Zala et al., 2023)	126	565	200.5	101.6	13.9	8.0	47.4
	`hirest_grounding` (Zala et al., 2023)	201	619	198.6	88.6	14.9	9.0	45.2
	`hirest_step` (Zala et al., 2023)	326	947	193.7	92.3	15.1	9.0	51.4
	`internvid_vtime` (Huang et al., 2024)	4,642	20,882	72.2	49.2	9.5	6.0	29.0
	`queryd` (Oncescu et al., 2021)	1,436	4,651	138.5	87.1	10.0	6.0	30.0
	`plm_rtloc` (Cho et al., 2025)	9,299	17,444	17.2	3.7	7.6	6.0	18.0
	Total	28,298	98,994	76.0	64.0	10.8	6.0	33.0
Test	`activitynet` (Krishna et al., 2017; Zhang et al., 2026)	1,455	4,500	134.9	60.6	22.5	6.0	118.1
	`charades` (Gao et al., 2017; Zhang et al., 2026)	1,313	3,363	29.6	7.1	5.5	3.0	18.0
	`qvhighlights` (Lei et al., 2021; Zhang et al., 2026)	1,511	1,541	149.6	2.6	31.3	15.0	150.0

conditioned acceptability criteria as defined in Eq. 1:

$$\text{accept}(o_i) = \begin{cases} \mathbf{1}[\text{Err}(o_i) < \delta], & \text{if } |\mathcal{T}_q| < T_{\text{short}}, \\ \mathbf{1}[\text{tIoU}(o_i) > \lambda], & \text{otherwise.} \end{cases} \tag{2}$$

For long events, a rollout o_i is accepted if its tIoU with the ground truth exceeds a threshold λ, as the continuous nature of tIoU makes it a reliable quality indicator in this regime. For short events, however, tIoU is too sparse to distinguish near-correct predictions from incorrect ones, so we instead accept a prediction if its boundary error (*i.e.*, Err) falls below a threshold δ. This hybrid criterion ensures that short-event predictions are evaluated on a meaningful scale, correcting the false-negative misclassifications that tIoU-only selection introduces.

Sampling Occasionally Solvable Queries. The GRPO algorithm derives its learning signal from the contrast between accepted and rejected rollouts within each group. Consequently, queries for which all G rollouts are either universally accepted (trivially easy) or universally rejected (prohibitively difficult) provide no informative training signal and are therefore excluded from $\mathcal{D}_{\text{RL}}$.

Concretely, we sample queries from $\mathcal{D}_{\text{SFT}}$ to construct $\mathcal{D}_{\text{RL}}$ with probability proportional to the number of accepted rollouts, $|o_{\text{accepted}}| = \sum_{i=1}^{G} \text{accept}(o_i)$, modulated by a temperature parameter τ:

$$\Pr(q \in \mathcal{D}_{\text{RL}}) \propto \begin{cases} \exp\left(\dfrac{\tau}{|o_{\text{accepted}}|}\right), & \text{if } N_{\min} \le |o_{\text{accepted}}| \le N_{\max}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This formulation preferentially selects queries that remain challenging yet solvable for the SFT model. Fig. 4a illustrates the resulting distributions of accepted rollout counts of queries in both the original SFT dataset and the sampled RLVR subset.

2.5 Denoising and Re-balancing Training Data

We build our training data pipeline upon the prior state-of-the-art, TimeLens (Zhang et al., 2026), which has refined samples from publicly available datasets (Wang et al., 2024a; Anne Hendricks et al., 2017; Zala et al., 2023; Huang et al., 2024; Oncescu et al., 2021) to produce TimeLens-100K (BSD 3-Clause License). These datasets are uniformly sampled based on video duration to ensure sampled videos are approximately distributed within 0 to 240 seconds, with a small portion of longer videos included. Upon closer examination, however, we identify several issues.

Audio-dependent queries persist in TimeLens-100K, which are inherently unanswerable from visual information alone and require audio cues for correct resolution. Training vision-language models on such samples, as in previous work, not only forces the model to resort to rote memorization during the SFT stage, but also introduces spurious difficulty signals during hard sample mining in the RLVR stage, with only vision and language modalities.

Ambiguous multi-occurrence queries describe events that occur multiple times throughout the video, yet the provided ground-truth annotation specifies only a single temporal interval. These

Table 2: **Performance Comparison.** Our model is built upon Qwen3-VL-8B (Bai et al., 2025), consistent with the baseline method (TimeLens-8B) under the same context budget.

Model	ActivityNet-TimeLens				Charades-TimeLens				QVHighlights-TimeLens			
	Short		Long		Short		Long		Short		Long	
	Dist≤1x	Err≤1x	mIoU	R1@.7	Dist≤1x	Err≤1x	mIoU	R1@.7	Dist≤1x	Err≤1x	mIoU	R1@.7
Proprietary Models												
Gemini-3.1-Pro	68.5	45.9	66.7	59.4	79.9	51.6	71.3	63.2	76.5	55.6	73.3	68.0
Seed-2.0-Pro	73.6	58.3	71.3	65.7	85.2	56.3	74.4	69.3	79.6	65.0	77.5	73.5
Open-Source Models												
Qwen3.5-9B (Qwen Team, 2026)	65.2	44.0	60.5	51.8	81.8	32.8	64.6	50.7	74.2	55.3	67.0	60.9
Qwen3-VL-8B (Bai et al., 2025)	59.0	42.1	56.5	48.9	73.2	42.0	66.1	55.4	68.3	52.8	65.5	59.9
TimeLens-8B (Zhang et al., 2026)	63.4	49.1	62.0	54.2	77.0	58.3	69.3	58.3	75.3	62.2	69.9	63.3
BiVTG-8B	**66.6**	**55.9**	**66.0**	**60.9**	**78.4**	**61.3**	**69.5**	**59.6**	**78.9**	**68.1**	**73.1**	**67.0**
	+3.2	+6.8	+4.0	+6.7	+1.4	+3.0	+0.2	+1.3	+3.6	+5.9	+3.2	+3.7

cases require either reformulating the query to specify the target occurrence (*e.g.*, "the first time") or revising the query itself. Additionally, some single-occurrence annotations are imprecise, with ground-truth intervals that do not correspond to the actual event boundaries in the video.

Under-representation of short events and videos. The proportion of samples involving short-duration events or short videos remains insufficient, introducing a distributional imbalance that may hinder the model's ability to generalize to bimodal scenarios.

To address these issues, we conduct a re-examination of the dataset and supplement it with a subset of the samples drawn from PLM-Video-Human (Cho et al., 2025), which originate from the SA-V dataset (Ravi et al., 2025). The resulting curated dataset, as shown in Tab. 1, maintains a total number of event queries comparable to that of TimeLens-100K, while improving data quality and balance.

3 Experiments

3.1 Experimental Setup

Benchmarks and Evaluation Metrics. For evaluation, we adopt the manually re-annotated Charades (Gao et al., 2017), QVHighlights (Lei et al., 2021), and ActivityNet (Krishna et al., 2017) introduced by TimeLens (Zhang et al., 2026) to exclude errors such as no event occurrence, duplicate queries, and inaccurate annotations. To address the evaluation deficiencies identified in Sec. 2.2, we explicitly separate short and long events during evaluation using a threshold of $T_{\text{short}} = 5$ seconds, and the threshold analysis is further discussed in the Appendix. For short events, we report the proportion of predictions whose boundary error (Err) and center distance (Dist) are within one ground-truth duration, namely $\leq 1 \cdot |\mathcal{T}|$. For long events, we report Recall@1 with tIoU $= 0.7$ and mean tIoU. In addition, to facilitate comparison with prior work, we also report conventional evaluation metrics computed over the entire dataset without distinguishing event duration, including Recall@1 with tIoU $= 0.3, 0.5, 0.7$ and mean tIoU.

Implementation Details. We build our model on Qwen3-VL-8B-Instruct (Bai et al., 2025), and sample video frames at 2 FPS. We set the minimum number of tokens per frame `min_tokens` to 64, and the maximum number of tokens for the entire video `total_tokens` to 14336; all the data preprocessing and configurations are the same as TimeLens-8B (Zhang et al., 2026). In training, we freeze the vision encoder while updating the projection layer and LLM. For supervised fine-tuning (SFT), we use a batch size of 128 and a learning rate of 0.00001. For reinforcement learning (RL), the number of unique prompts per batch is 8, each prompt samples $G = 8$ rollouts, and the KL coefficient β is set to 0. The experiments are conducted on NVIDIA H100 (80GB) GPUs.

3.2 Comparison with State-of-the-arts

Table 2 reports performance under our duration-aware metrics across three benchmarks in TimeLens-Bench. Our model consistently outperforms the previous state-of-the-art open-source model,

Table 3: **Main Ablation Studies.** We analyze the impact of modifications at each stage, separately for short and long events. Within each group, **bold** and underline denote the best and second-best values per column, respectively; red highlights notable performance degradation.

SFT	Sampling Strategy	Reward Shaping	ActivityNet		Charades		QVHighlights	
			Short	Long	Short	Long	Short	Long
			Dist≤1x↑	R1@.7↑	Dist≤1x↑	R1@.7↑	Dist≤1x↑	R1@.7↑
TimeLens-100K	-	-	**67.8**	42.5	**80.5**	41.5	78.6	46.5
+ Denoising	-	-	66.4	60.0	78.2	56.2	78.3	**67.1**
+ PLM-RTLoc-17K	-	-	67.0	**60.5**	78.2	**58.7**	**79.4**	66.7
TimeLens-100K	One-Pass Hard (tIoU)	tIoU	63.4	54.2	**77.0**	58.3	75.3	63.3
Ours-100K	One-Pass Hard (tIoU)	tIoU	62.3	60.3	74.1	**61.9**	73.6	**70.2**
Ours-100K	Some-Accept (tIoU)	tIoU	64.9	60.9	76.6	60.0	76.4	69.3
Ours-100K	Some-Accept (Hybrid)	tIoU	**65.3**	**61.8**	76.7	60.6	**76.9**	68.2
Ours-100K	Ours	tIoU	65.3	**61.8**	76.7	60.6	76.9	68.2
Ours-100K	Ours	dIoU	65.3	61.6	76.9	**60.9**	77.2	**68.4**
Ours-100K	Ours	gIoU	65.0	**61.8**	76.5	60.7	77.5	68.3
Ours-100K	Ours	boundary	**66.8**	60.6	78.3	58.8	79.2	66.3
Ours-100K	Ours	center	**66.8**	60.1	**78.6**	59.2	**79.4**	66.2
Ours-100K	Ours	dIoU, boundary	66.6	60.9	78.4	59.6	78.9	67.0

TimeLens-8B, across all three benchmarks and under both short-event and long-event metrics. The gains are particularly pronounced on short-event metrics, where our model improves the proportion of predictions satisfying Err $\leq 1 \times |\mathcal{T}|$ by +6.8, +3.0, and +5.9 points on ActivityNet, Charades, and QVHighlights, respectively, validating the effectiveness of our model in localizing brief, temporally precise events. For long events, consistent improvements in both mIoU and R1@0.7 further confirm that our approach does not sacrifice long-event performance at the expense of short-event gains.

3.3 Ablation Studies

Effect of Refining SFT Dataset. As shown in Table 3, filtering and rewriting TimeLens-100K primarily improve model performance on long events, while the introduction of PLM-RTLoc (Cho et al., 2025) compensates for the short-video benchmark, Charades, as well as short-event localization on the other two long-video benchmarks, and enriches diversity for the subsequent RLVR stage.

Effect of Sampling Strategy in RLVR. As shown in the first two rows of the sampling strategy part, we observe that **applying RLVR with only the tIoU reward on the SFT model inherently degrades short-event localization while boosting long-event grounding**, as the model tends to predict longer intervals to hack the tIoU reward on short-event queries. Building on this, we first introduce our proposed sampling strategy, which selects samples via multiple rollouts under hybrid criteria for short/long events, retaining only queries for which the model produces occasionally acceptable predictions. Our sampling strategy achieves consistently superior performance.

Effect of Reward Shaping in RLVR. Regarding reward design, we ablate several reward formulations and find that point-based rewards (boundary error and center distance) yield better timestamp regression for short events, while interval-based rewards (tIoU, gIoU, dIoU) provide stronger generalization for long-event localization. We therefore adopt a combination of dIoU and boundary error as our final reward, which jointly captures holistic interval quality and enforces precise boundary and center regression constraints, achieving a favorable balance across both event types.

4 Related Work

4.1 Video Temporal Grounding Methods

Video temporal grounding is a fundamental task in video understanding (Gao et al., 2017; Grauman et al., 2022; Krishna et al., 2017; Lei et al., 2021; Oncescu et al., 2021; Regneri et al., 2013). Previous approaches (Lei et al., 2021; Lin et al., 2023; Mu et al., 2024; Gordeev et al., 2026; Moon et al., 2023) build task-specific architectures on pre-trained vision-language encoders (Carreira & Zisserman,

Table 4: Comparison with previous works under conventional metrics. The results are only for reference because of the aforementioned deficiencies of tIoU on short events.

Model	ActivityNet-TimeLens				Charades-TimeLens				QVHighlights-TimeLens			
	R1@.3	R1@.5	R1@.7	mIoU	R1@.3	R1@.5	R1@.7	mIoU	R1@.3	R1@.5	R1@.7	mIoU
Proprietary Models												
Gemini-2.5-Pro (Comanici et al., 2025)	72.3	64.2	47.1	58.1	74.1	61.1	34.0	52.8	84.1	75.9	61.1	70.4
Gemini-3.1-Pro (Google DeepMind, 2026)	72.8	61.0	42.3	55.4	77.1	61.8	34.1	54.1	81.8	72.0	57.3	67.1
Seed-2.0-Pro (Seed, 2025)	77.2	67.4	47.6	60.0	81.1	64.4	36.5	56.7	85.8	77.5	61.3	70.3
Open-Source Models												
Time-R1-7B (Wang et al., 2025)	44.8	31.0	19.0	33.1	57.9	32.0	16.9	36.6	65.8	51.5	36.1	49.2
TRACE (Guo et al., 2024)	43.4	33.9	22.0	32.7	37.2	21.8	9.6	27.1	49.7	39.1	28.1	39.0
TimeSuite (Zeng et al., 2024)	27.1	17.5	8.6	19.8	56.3	35.5	18.0	38.1	27.1	16.9	9.9	21.7
Qwen3.5-9B (Qwen Team, 2026)	67.3	55.0	36.6	50.8	71.3	48.6	24.3	47.6	75.3	65.4	50.8	61.1
Qwen3-VL-8B (Bai et al., 2025)	62.1	51.2	34.4	46.8	69.2	53.4	27.5	48.3	74.2	64.6	49.3	59.4
Video-OPD (Li et al., 2026)	60.5	45.6	35.8	47.3	73.1	45.8	32.4	52.0	73.8	60.3	50.4	61.0
TimeLens-8B (Zhang et al., 2026)	68.9	58.4	40.6	53.2	76.6	63.0	35.2	**55.2**	80.2	71.6	55.5	65.5
BiVTG-8B	**71.7**	**62.4**	**44.6**	**56.0**	**77.1**	**63.7**	**35.7**	55.0	**82.6**	**73.7**	**57.8**	**68.1**
	+2.8	+4.0	+4.0	+2.8	+0.5	+0.7	+0.5	-0.2	+2.4	+2.1	+2.3	+2.6

2017; Tran et al., 2015; Lin et al., 2022; Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025; Wang et al., 2022; 2024b) and optimize them using regression losses. With the advent of Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Qwen Team, 2026; Li et al., 2024a; Zhang et al., 2024; 2025a; Zhu et al., 2025), recent works address this task via unified instruction tuning using cross-entropy loss. (Meinardus et al., 2024; Guo et al., 2024; Ren et al., 2024; Huang et al., 2024; Zeng et al., 2024; Wu et al., 2025; Chen et al., 2024; Li et al., 2025d; Ge et al., 2025; Li et al., 2025a). Despite their differences, virtually all of these methods are optimized for and evaluated using temporal intersection-over-union (tIoU). In this paper, we expose the limitations of this protocol and propose improved training and evaluation practices.

4.2 Reinforcement Learning for Video Temporal Grounding

The Group Relative Policy Optimization (GRPO) algorithm and its variants have demonstrated strong potential for training large language models on specialized tasks in coding and mathematics (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025b; Zeng et al., 2026a). The Reinforcement Learning with Verifiable Rewards (RLVR) paradigm has since been extended to perception tasks that have deterministic outcomes, including OCR, counting, object detection, and segmentation (Jiang et al., 2025; Yu et al., 2025a; Liu et al., 2025b;c; Meng et al., 2025a). Accordingly, GRPO has also been adopted in video understanding tasks, including temporal grounding (Li et al., 2025b;c; Yue et al., 2025; Chen et al., 2025; Feng et al., 2025; Wang et al., 2025; Liu et al., 2025a; Zhang et al., 2025c; Liao et al., 2025). However, existing works select training samples for RLVR via single-pass inference, typically retaining those with low tIoU scores, which is suboptimal due to the inherent limitations of the tIoU metric. In this paper, we establish RLVR based on a hybrid evaluation protocol to reliably identify queries that are partially but not consistently solved within multiple rollouts.

5 Conclusion

In this work, we revisit video temporal grounding from a duration-aware perspective and identify two limitations of using tIoU uniformly across event durations: it can mischaracterize near-correct short-event predictions and induce interval-inflation reward hacking under RLVR. We further show that distance- and overlap-based rewards favor short and long events, respectively, revealing a trade-off between the two temporal regimes. Motivated by these findings, we introduce a duration-aware framework comprising stratified evaluation, a hybrid dIoU and boundary reward shaping, and a multi-rollout sampler that selects hard-yet-solvable queries using regime-appropriate criteria. We also construct BiVTG-100K with improved short-event coverage and annotation quality. Extensive experiments show that our framework mitigates the aforementioned limitations and improves grounding precision across different temporal scales. The consistent improvements across Charades, ActivityNet, and QVHighlights further demonstrate the effectiveness of the proposed bimodal framework.

AI USE STATEMENT

In this work, we used generative AI tools to assist in formulating mathematical claims, refining the research methodology and experimental design, implementing part of the methods, and cleaning and reformatting the dataset (TimeLens-100K). We also used these tools to polish portions of the manuscript and check citations. We did not use generative AI to generate synthetic datasets, develop theoretical models or conceptual frameworks, provide essential elements of mathematical proofs, write proofs, propose or refine hypotheses, perform translations, conduct qualitative or thematic analyses, or interpret the results.

All AI-assisted outputs were critically reviewed and verified by the authors. In particular, AI-polished text was carefully checked, citations were verified against the original sources, and any AI-assisted code or equations were independently validated. The authors take full responsibility for the final content of this work, including all text, claims, code, equations, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Because this work was conducted in collaboration with an industry partner, we cannot release the complete source code for SFT and RLVR training, model weights (BiVTG-8B), and the proposed dataset (BiVTG-100K) during the review process. Upon acceptance, we will publicly release these materials to support the reproducibility of our results.

REFERENCES

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 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.

Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pp. 334–350. Springer, 2020.

Delong Chen, Tejaswi Kasarla, Yejin Bang, Mustafa Shukor, Willy Chung, Jade Yu, Allen Bolourchi, Theo Moutakanni, and Pascale Fung. Action100m: A large-scale video action dataset. *arXiv preprint arXiv:2601.10592*, 2026.

Ruizhe Chen, Zhiting Fan, Tianze Luo, 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. *arXiv preprint arXiv:2507.18100*, 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*, 2024.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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.

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.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*, 2025.

Google DeepMind. Gemini 3.1 pro model card. Technical report, Google DeepMind, February 2026. URL `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf`.

Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 907–916, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, et al. Step3-vl-10b technical report. *arXiv preprint arXiv:2601.09668*, 2026.

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.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. *arXiv preprint arXiv:2510.12798*, 2025.

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

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8592–8603, 2025a.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-opd: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. *arXiv preprint arXiv:2602.02994*, 2026.

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*, 2024a.

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*, 2025b.

Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms. *arXiv preprint arXiv:2509.18056*, 2025c.

Zeqian Li, Qirui Chen, Tengda Han, Ya Zhang, Yanfeng Wang, and Weidi Xie. Multi-sentence grounding for long-term instructional video. In *European Conference on Computer Vision*, pp. 200–216. Springer, 2024b.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025d.

Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. Improved visual-spatial reasoning via r1-zero-like training. *arXiv preprint arXiv:2504.00883*, 2025.

Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Videoseek: Long-horizon video agent with tool-guided seeking. *arXiv preprint arXiv:2603.20185*, 2026.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

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*, 2025a.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025b.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025c.

Boris Meinardus, Hector Rodriguez, Anil Batra, Anna Rohrbach, and Marcus Rohrbach. Chrono: A simple blueprint for representing time in mllms. *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*, 2025a.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025b.

WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration for video temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023.

Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18930–18940, 2024.

Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2265–2269. IEEE, 2021.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 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.

Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. Technical report, Bytedance, 2025.

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.

Yudi Shi, Shangzhe Di, Qirui Chen, Qinian Wang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Weaver: End-to-end agentic system training for video interleaved reasoning. *arXiv preprint arXiv:2602.05829*, 2026.

Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Advances in Neural Information Processing Systems*, 36:38863–38886, 2023.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, December 2015.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. Cosmo: Contrastive streamlined multimodal model with interleaved pre-training. *arXiv preprint arXiv:2401.00849*, 2024a.

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

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024b.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13754–13765, 2025.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*, 2025.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8579–8591, 2025.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025a.

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*, 2025b.

Feng Yue, Zhaoxing Zhang, Junming Jiao, Zhengyu Liang, Shiwen Cao, Feifei Zhang, and Rong Shen. Tempo-r0: A video-mllm for temporal video grounding through efficient temporal sensing reinforcement learning. *arXiv preprint arXiv:2507.04702*, 2025.

Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23056–23065, 2023.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026a.

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. In *ICLR*, 2024.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

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.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025b.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2026.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025c.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

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*, pp. 12993–13000, 2020.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

A DURATION THRESHOLD ANALYSIS

In this subsection, we want to establish the theoretical relationship between Err and tIoU. Our goal is to characterize how boundary errors translate into temporal overlap-based evaluation, and to identify under what temporal scale (e.g., a threshold T_{short}) the two metrics become comparable. Let the ground-truth interval be $\mathcal{T} = [t_s, t_e]$ with duration $|\mathcal{T}| = t_e - t_s$, and the predicted interval be $\hat{\mathcal{T}} = [\hat{t}_s, \hat{t}_e]$. We define the signed boundary errors as

$$\varepsilon_s = \hat{t}_s - t_s, \quad \varepsilon_e = \hat{t}_e - t_e, \quad \text{Err} = |\varepsilon_s| + |\varepsilon_e|.$$

The intersection and union between $\mathcal{T}$ and $\hat{\mathcal{T}}$ can be written as

$$|\mathcal{I}| = \max(0,\ |\mathcal{T}| - \max(0, \varepsilon_s) - \max(0, -\varepsilon_e)),$$

$$|\mathcal{U}| = |\mathcal{T}| + \max(0, -\varepsilon_s) + \max(0, \varepsilon_e).$$

Bounds on tIoU. For $\text{Err} = |\varepsilon_s| + |\varepsilon_e|$, we consider two extreme configurations.

Lower bound (inward contraction, $\varepsilon_s > 0, \varepsilon_e < 0$):

$$|\mathcal{I}| = \max(0,\ |\mathcal{T}| - \text{Err}), \quad |\mathcal{U}| = |\mathcal{T}|,$$

$$\text{tIoU}_{\min} = \max\left(0,\ 1 - \frac{\text{Err}}{|\mathcal{T}|}\right).$$

Upper bound (outward expansion, $\varepsilon_s < 0, \varepsilon_e > 0$):

$$|\mathcal{I}| = |\mathcal{T}|, \quad |\mathcal{U}| = |\mathcal{T}| + \text{Err},$$

$$\text{tIoU}_{\max} = \frac{|\mathcal{T}|}{|\mathcal{T}| + \text{Err}}.$$

Combining the two cases yields the following tight bound:

$$\boxed{\max\left(0,\ 1 - \frac{\text{Err}}{|\mathcal{T}|}\right) \ \le\ \text{tIoU}\ \le\ \frac{|\mathcal{T}|}{|\mathcal{T}| + \text{Err}}}.$$

Implications for Threshold Selection. Let $r = \text{Err}/|\mathcal{T}|$ denote the normalized error ratio. Then:

- If $r \le 1 - \lambda$, then $\text{tIoU} \ge \lambda$.
- If $1 - \lambda < r \le \frac{1-\lambda}{\lambda}$, the result becomes ambiguous, i.e., tIoU may be either above or below λ.
- If $r > \frac{1-\lambda}{\lambda}$, then $\text{tIoU} = 0$, and the prediction is guaranteed to fail.

For $\lambda = 0.7$, this implies:

- $r \le 0.3$: $\text{tIoU} \ge 0.7$ is guaranteed;
- $0.3 < r \le 0.43$: ambiguous regime;
- $r > 0.43$: $\text{tIoU} < 0.7$ is guaranteed.

Failure on short events. Assume a typical absolute boundary error δ (e.g., 1 second), then $r = \delta/|\mathcal{T}|$. As $|\mathcal{T}|$ decreases, r increases rapidly, leading to severe degradation in tIoU, which means that for short events, even small absolute errors can push the prediction into the failure regime.

B Additional Implementation Details

B.1 RLVR Sampling Pipeline

Hybrid acceptance criterion. As defined in Equation 2, the threshold for short events, T_{short}, is set to 5 seconds. For events with short durations ($|\mathcal{T}| < T_{\text{short}}$), we employ the error-sum criterion (Err), as tIoU becomes overly sensitive when $|\mathcal{T}|$ is small. In this regime, the acceptable error threshold δ is set to be less than 2 seconds. For long-duration events ($|\mathcal{T}| \geq T_{\text{short}}$), we instead adopt the standard tIoU criterion, with an acceptance threshold λ greater than 0.7.

Partial-acceptance hard mining. As noted in Equation 3, we retain only samples satisfying $|o_{\text{accepted}}| \in [2, 6]$. Samples that are considered excessively difficult or too easy are excluded. The retained samples contain both accepted and rejected rollouts, yielding a maximal training signal under the GRPO objective.

Duration-aware down-sampling. We further down-sample the dataset to target size $|\mathcal{D}_{\text{RL}}| = 1000$ using balanced duration buckets: $[0, 30), [30, 60), \ldots, [240, \infty)$. As in Equation 3, samples with fewer accepted rollouts (*i.e.*, harder instances) are assigned a higher probability. A small temperature τ sharpens this preference.

B.2 Prompt Templates

Prompts for TimeLens and Qwen-VL series

```
Please find the visual event described by the sentence "{query}",
determining its starting and ending times.
The format should be:

'The event happens in <start time> - <end time> seconds'.
```

Prompts for evaluating Gemini models

```
Please find the visual event described by the sentence "{query}",
determining its starting and ending times.
Answer in the format 'The event happens in MM:SS - MM:SS'.
```

Prompts for evaluating Doubao-2.0-Pro

```
Please find the visual event described by the sentence "{query}",
determining its starting and ending times.
Answer strictly in JSON format:

{"span": "[<start time in second>, <end time in second>]"}.
```

C Additional Exploration

Beyond the methods presented in the main paper, we conducted several additional investigations that did not yield consistent positive gains. First, we experimented with various chain-of-thought (CoT) responses for RLVR training, including CoT based on temporal segment descriptions and CoT

that reflects over multiple candidate prediction proposals. Neither yielded positive improvements, which is to some extent consistent with observations in prior work (Zhang et al., 2026; Huang et al., 2026), suggesting that video temporal grounding appears to be a predominantly perceptual task that benefits little from explicit multi-step reasoning. Second, building on the above, we further conducted rejection sampling for reinforcement fine-tuning (RFT) on top of the RLVR-trained model, which produced marginal but inconsistent gains over the RLVR checkpoint. Third, we explored data scaling using Action100M (Chen et al., 2026); however, due to distributional shift and data quality issues, this led to performance degradation on the target benchmarks. We report these negative results in the hope that they provide useful guidance for future research in this direction.

D Additional Experimental Results

D.1 Main Results under Shorter Threshold

In Tab. 5, we report the metrics under short/long event threshold T_{short} at 3 seconds, meaning that only events with durations of 1–2 seconds are treated as short events. Compared with TimeLens-8B, BiVTG-8B achieves consistent gains on nearly all short-event metrics across the three benchmarks, with especially notable improvements on distance- and error-based localization accuracy. The improvements on long events are generally smaller but remain competitive, suggesting that the proposed bimodal design primarily benefits fine-grained short-event grounding without sacrificing long-duration performance.

Table 5: Performance Comparison (Duration threshold $T_{\text{short}} = 3$ seconds).

Model	ActivityNet-TimeLens				Charades-TimeLens				QVHighlights-TimeLens			
	Short		Long		Short		Long		Short		Long	
	Dist≤1x	Err≤1x	mIoU	R1@.7	Dist≤1x	Err≤1x	mIoU	R1@.7	Dist≤1x	Err≤1x	mIoU	R1@.7
Proprietary Models												
Gemini-3.1-Pro (Google DeepMind, 2026)	61.8	28.6	63.5	54.2	73.4	37.6	65.8	52.4	67.2	43.1	71.4	64.6
Seed-2.0-Pro (Seed, 2025)	68.5	48.1	67.7	59.9	80.5	42.0	68.7	56.6	74.0	54.2	75.1	69.3
Open-Source Models												
Qwen3.5-9B (Qwen Team, 2026)	58.2	30.1	57.9	47.2	75.5	10.2	60.2	39.3	68.6	42.3	65.2	57.7
Qwen3-VL-8B (Bai et al., 2025)	50.6	28.7	53.8	44.4	64.1	25.3	61.0	45.0	61.3	41.8	63.6	56.1
TimeLens-8B (Zhang et al., 2026)	57.3	39.3	59.6	51.2	70.7	48.5	**65.4**	**51.7**	69.1	52.1	69.0	61.8
BiVTG-8B	**61.8**	**48.1**	**62.8**	**55.7**	**72.8**	**51.9**	64.8	51.3	**72.7**	**59.3**	**71.9**	**64.4**
	+4.5	+8.8	+3.2	+4.5	+2.1	+3.4	-0.6	-0.4	+3.6	+7.2	+2.9	+2.6

D.2 More Ablation Studies

As shown in Table 6, we also investigate the effects of the mixing ratio between the boundary error reward and the IoU-based reward, the magnitude of the KL coefficient, and the inclusion of entropy regularization in the loss to prevent entropy collapse based on an immediate training configuration.

Table 6: **Ablation studies** of other hyperparameters on ActivityNet-TimeLens. We report Err≤1x (↑) and R1@0.7 (↑) for short and long events. Note that this only serves as an initial sensitivity analysis built on TimeLens-8B instead of our final configuration.

(a) **Reward composition.**

bound : tIoU	Err≤1x	R1@.7
0.5 : 0.4	55.6	60.9
0.7 : 0.2	55.9	60.9

(b) **KL coefficient β.**

β	Err≤1x	R1@.7
0.0	55.9	60.9
0.001	56.0	61.1

(c) **Entropy loss coefficient λ_{entropy}.**

λ_{entropy}	Err≤1x	R1@.7
0	55.9	60.9
0.01	55.8	60.9

D.3 Reward Values and Sampling Sensitivity

Table 7 illustrates boundary reward value smoothness of one example; Tables 8a and 8b report sampling sensitivity on ActivityNet-TimeLens. The first row of each sensitivity table is the default.

Table 7: Boundary reward values at different temporal deviations.

Scenario	GT duration (s)	Boundary error (s)	r_{boundary}
Minor deviation	2.0	1.0	0.8825
Boundary touching	2.0	2.0	0.6065
Significant deviation	2.0	4.0	0.1353
Extreme deviation	2.0	8.0	0.0003
Minor deviation	10.0	1.0	0.9900
Boundary touching	10.0	10.0	0.3679

Table 8: Sampling hyper-parameter sensitivity on ActivityNet-TimeLens.

(a) Sensitivity to prediction acceptance thresholds in Equation 2.

δ (s)	λ	Dist≤1x ↑	R1@.7 ↑
2	0.7	66.6	60.9
3	0.7	65.7	61.2
2	0.8	66.4	61.1

(b) Sensitivity to the acceptable rollout count range among eight rollouts per query.

$N_{\min}$	$N_{\max}$	Dist≤1x ↑	R1@.7 ↑
2	6	66.6	60.9
3	5	66.8	60.2
1	7	65.9	59.3

D.4 Performance Breakdown

Table 9 reports results by tIoU on ActivityNet-TimeLens. Δ denotes the change in percentage points from TimeLens-8B to BiVTG-8B.

Table 9: Distribution of prediction tIoU over all 4,500 queries. (Mean tIoU is 0.532 / 0.560 and median tIoU is 0.571 / 0.667 for TimeLens-8B / BiVTG-8B.)

	TimeLens-8B		BiVTG-8B		
tIoU bin	Count	%	Count	%	Δ
$[0.0, 0.1)$	972	21.6	936	20.8	-0.8
$[0.1, 0.2)$	194	4.3	136	3.0	-1.3
$[0.2, 0.3)$	243	5.4	203	4.5	-0.9
$[0.3, 0.4)$	308	6.8	281	6.2	-0.6
$[0.4, 0.5)$	163	3.6	138	3.1	-0.5
$[0.5, 0.6)$	403	9.0	408	9.1	$+0.1$
$[0.6, 0.7)$	386	8.6	393	8.7	$+0.1$
$[0.7, 0.8)$	315	7.0	295	6.6	-0.4
$[0.8, 0.9)$	424	9.4	467	10.4	$+1.0$
$[0.9, 1.0]$	1,092	24.3	1,243	27.6	$+3.3$

The two distributions exhibit similar overall shapes: both are bimodal, with substantial probability mass near complete misses ($\text{tIoU} < 0.1$) and at the high-quality end ($\text{tIoU} \geq 0.9$). However, **BiVTG consistently shifts probability mass toward the high-quality tail**. Specifically, it produces more predictions in $[0.9, 1.0]$ (27.6% vs. 24.3%) and fewer predictions in the low- to medium-quality range $[0.1, 0.5)$ (16.8% vs. 20.1% in total). These results suggest that BiVTG's improvement stems from converting more of TimeLens's low- and medium-quality predictions into high-quality ones.

E Limitations

This work has several limitations. First, to ensure fair comparison with baselines, we do not investigate the effect of input data preprocessing hyperparameters such as context length, input resolution, and frame sampling rate. Second, our bimodal framework does not account for complex events that manifest simultaneously as both a point-in-time occurrence and a temporal interval. Third, despite human re-annotation efforts, TimeLens-Bench (Zhang et al., 2026) still contains a non-trivial number of annotation errors. Finally, the temporal extent of certain events is inherently ambiguous; for

example, “person sits down on sofa” could refer to the instantaneous moment of sitting or to the entire duration of the person remaining seated on the sofa. Addressing these limitations constitutes promising directions for future work.

F BROADER IMPACTS

This work advances video temporal grounding, a fundamental capability underlying a wide range of real-world applications. On the positive side, improved temporal localization can enable more accurate retrieval of relevant moments from personal video recordings. It can also facilitate procedural understanding in educational and industrial settings, support more effective video-based information retrieval at scale, and contribute to the development of more capable agentic video question answering systems that assist users in navigating long-form video content efficiently. On the negative side, enhanced temporal grounding capabilities could be misused for large-scale surveillance, enabling automated and precise localization of individuals’ activities across extended video footage without their consent. Similarly, such technology could lower the barrier to extracting sensitive or private moments from video data, raising concerns around unauthorized monitoring and privacy violations. The improved ability to precisely locate events in video could also be exploited in the context of disinformation, for instance, by facilitating the targeted clipping and decontextualization of footage to manipulate public perception.