TROVE: LEARNING TO RE-OBSERVE VIDEOS FOR TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video temporal grounding requires global context to identify a queried occurrence and fine-grained local evidence to resolve its boundaries. Uniform frame sampling provides temporal coverage, but distributes additional observations across both relevant and irrelevant regions. We propose TROVE, a framework that learns temporal grounding through targeted re-observation. A shared policy first predicts a coarse interval from sparse global frames, then refines it using dense local evidence selected by that interval together with sparse global support. We train this trajectory through Grounding Initialization followed by Joint Re-Observation Learning. A Geometry-to-Metric Reward Curriculum supplies graded feedback for disjoint predictions before emphasizing overlap quality and evaluation thresholds. Contrastive Re-Observation Calibration (CRC) trains the first action using both its localization reward and the difference between targeted and uniform second-observation rewards. The uniform comparator is used only during training. On TimeLens-Bench, TROVE with a 4B backbone improves mIoU over the native model by 11.7, 8.5, and 4.5 points on the Charades, ActivityNet, and QVHighlights subsets, respectively. Experiments across three model scales, training and observation ablations, and boundary analyses demonstrate the value of learning temporal localization together with targeted evidence acquisition.

1 INTRODUCTION

Video temporal grounding (VTG) connects a natural-language description to the precise moment it describes, providing the temporal coordinates needed to retrieve and reason about video content (Gao et al., 2017; Hendricks et al., 2017). Multimodal large language models (MLLMs) approach this task by generating start and end timestamps from sampled video frames. Accurate grounding requires two complementary forms of evidence: full-video context distinguishes the queried occurrence from other activities, while fine-grained local observations reveal when that occurrence begins and ends. Integrating these observations within a finite frame budget makes evidence allocation part of the grounding problem.

Uniformly sampled frames provide a fixed visual context for a single grounding prediction (Figure 1, left). This sampling preserves temporal coverage, but increasing its density distributes additional frames across both relevant and irrelevant regions. For native Qwen3.5-4B, increasing inference from 2 to 4 fps raises the mean input from 202.5 to 302.6 frames, while mIoU changes from 50.6 to 50.3 (Appendix A, Figure 6, left). Denser uniform observation thus does not necessarily improve localization. The challenge is to allocate a limited frame budget to obtain the temporal evidence most useful for the query while retaining full-video context.

Adaptive and sequential video observation makes evidence allocation depend on the query or on intermediate predictions (Wang et al., 2025d; Zou et al., 2026; Zhang et al., 2026a; Liu et al., 2026). Many such methods target video question answering, with representative workflows incorporating selected clips into the reasoning context (Figure 1, middle). Although this interaction can also support grounding, evidence sufficient to identify an event may leave its start and end unresolved, and optional clip requests need not revisit these boundaries. This motivates an explicit refinement step that uses the coarse interval to acquire finer boundary evidence.

We introduce TROVE (**T**argeted **R**e-**O**bservation for **V**ideo t**E**mporal grounding), which couples evidence acquisition with temporal localization through two calls to a shared policy (Figure 1, right).

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.

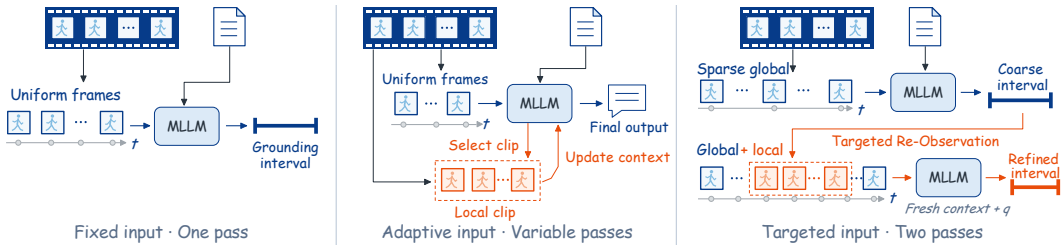


Figure 1: **Visual evidence acquisition.** *Left:* a standard video MLLM predicts from a fixed, uniformly sampled input. *Middle:* adaptive and sequential video observation, illustrated by a representative workflow that selects temporal clips and updates the model's context with their visual evidence (Zhang et al., 2026a; Liu et al., 2026). *Right:* TROVE uses a coarse interval to construct a second observation with sparse global support and dense local detail, then refines the interval in a fresh context with the same policy.

Global Grounding first predicts a coarse interval from sparse full-video samples. Targeted Re-Observation expands that interval into a contextual window and acquires dense frames from the source video. The second call combines these local frames with sparse global support under a per-call frame budget and predicts the refined interval. It uses a fresh context: the coarse prediction directs refinement through the visual evidence it selects, while global support retains the surrounding activity context. The first prediction thus serves both as a localization estimate and as an observation decision whose value depends on the resulting refinement.

Training this trajectory requires feedback for both interval accuracy and observation utility. Temporal intersection over union (tIoU), when used as a reward, assigns zero to every prediction disjoint from the target, regardless of how close it is to the correct interval. Figure 6 (right) in Appendix A illustrates this ambiguity by translating an equal-length interval: temporal generalized IoU (tGIoU) distinguishes these predictions through their separation from the target. The localization signal must remain informative before overlap is established and support precise boundaries as predictions improve. The first action also needs feedback on the localization achieved after re-observation, beyond the overlap of its initial interval.

We train the policy through Grounding Initialization followed by Joint Re-Observation Learning. Two learning mechanisms support this progression. The *Geometry-to-Metric Reward Curriculum* first uses tGIoU to distinguish disjoint predictions, then shifts toward a smooth objective emphasizing overlap quality and evaluation thresholds. *Contrastive Re-Observation Calibration* (CRC) evaluates the observation decision through a training-only uniform comparator. The targeted–uniform reward difference calibrates the first action's downstream return alongside its own localization reward, while the refined action learns directly from its grounding outcome. This couples the ability to propose a useful interval with the ability to exploit the evidence it reveals.

Our contributions are:

- We introduce TROVE, a two-step grounding framework in which a shared policy uses coarse localization to select dense local evidence and refines temporal boundaries while retaining global context.
- We develop a Geometry-to-Metric Reward Curriculum that provides graded feedback for disjoint intervals before shifting toward overlap quality and evaluation thresholds.
- We propose CRC, which trains the initial prediction using both its localization reward and the targeted–uniform difference in downstream grounding rewards, explicitly supervising its observation choice.

2 RELATED WORK

Video MLLMs for VTG. Video MLLMs improve timestamp generation through temporal representations and grounding supervision. Timestamp-aware frames, temporal tokens, explicit time information, and frame identifiers expose temporal position to the language model (Ren et al., 2024;

Wang et al., 2025a; Guo et al., 2025a; Wu et al., 2025). Boundary-aware and grounded tuning provide localization supervision (Huang et al., 2024a; Zeng et al., 2025), while TRACE models causal dependencies between events (Guo et al., 2025b). Reinforcement-learning approaches further optimize interval predictions with verifiable rewards (Wang et al., 2025c; Li et al., 2025; Chen et al., 2025). A common setting is prediction from a fixed, uniformly sampled visual input. TROVE couples localization with further evidence acquisition: the predicted interval selects a second observation, and training optimizes the resulting grounding trajectory.

Adaptive and sequential video observation. Observation methods differ in whether they select frames or tokens, or request additional video segments during inference. AdaFrame selects frames using observation history and global memory (Wu et al., 2019), while query-adaptive approaches refine relevant frame clusters or validate candidate frames before sampling their neighborhoods (Wang et al., 2025d; Zou et al., 2026). Text-guided token pruning and sampling reduce the visual representation at the token level (Huang et al., 2024b; Zhang et al., 2025; Fan et al., 2026). For clip-level interaction, VITAL acquires temporal segments through timestamp-based video tools (Zhang et al., 2026a). VideoTemp-o3 (Liu et al., 2026) couples temporal grounding with video question answering through on-demand clipping: the model can answer directly or request segments and append their visual content to the dialogue for further reasoning. REVISOR (Li et al., 2026a) proposes a review interval and resamples it more densely while retaining its initial reasoning trace. TROVE organizes re-observation as an explicit grounding step: the initial interval constructs a fresh context combining dense local frames with sparse global support, and action-wise returns connect evidence selection to the refined grounding outcome.

3 METHOD

TROVE couples temporal localization with evidence acquisition through a two-step trajectory (Figure 2). We first describe how the initial prediction constructs the observation for refinement, then develop the rewards used to learn interval quality and observation utility.

3.1 GROUNDING AS SEQUENTIAL OBSERVATION

Given a raw video V of duration D and a natural-language query q, VTG predicts a temporal interval aligned with the target $y^* = (s^*, e^*)$. The model observes a source-time-ordered sequence $\mathcal{O} = \{(x_i, t_i)\}_{i=1}^{N}$, where x_i is a frame acquired at time t_i. The grounding operation is formulated as

$$\hat{y} = (\hat{s}, \hat{e}) = f_\theta(q, D, \mathcal{O}). \tag{1}$$

Here f_θ denotes the grounding policy with parameters θ, and $\hat{s}$ and $\hat{e}$ are the predicted start and end times. Each invocation is constrained by a frame budget, and $B_\Sigma = \sum_j |\mathcal{O}^{(j)}|$ measures cumulative model-input frames, counting a frame each time it is supplied to the model.

The first invocation predicts a coarse interval $\tilde{y}$ from sparse global frames $\mathcal{O}_g$. That interval determines a targeted observation $\mathcal{O}_{\mathrm{tgt}}$ for the second invocation, which produces the final prediction. Both calls share f_θ, allowing the same grounding capability to guide observation and interpret the newly acquired evidence.

3.2 SPARSE-TO-DENSE RE-OBSERVATION

Global Grounding. To identify the queried occurrence before selecting a local region, Global Grounding uniformly samples the complete video:

$$\mathcal{O}_g = \mathrm{Lin}(V, [0, D], B_g), \qquad \tilde{y} = (\tilde{s}, \tilde{e}) = f_\theta(q, D, \mathcal{O}_g). \tag{2}$$

Here Lin selects source-ordered samples uniformly over the specified interval, and B_g is the global sample count induced by nominal rate r_g, subject to the frame cap. The resulting interval identifies the region to revisit at higher temporal resolution.

Re-Observation Window Construction. Boundary refinement requires observing the transitions around the initial prediction. We therefore expand a valid coarse interval to include neighboring

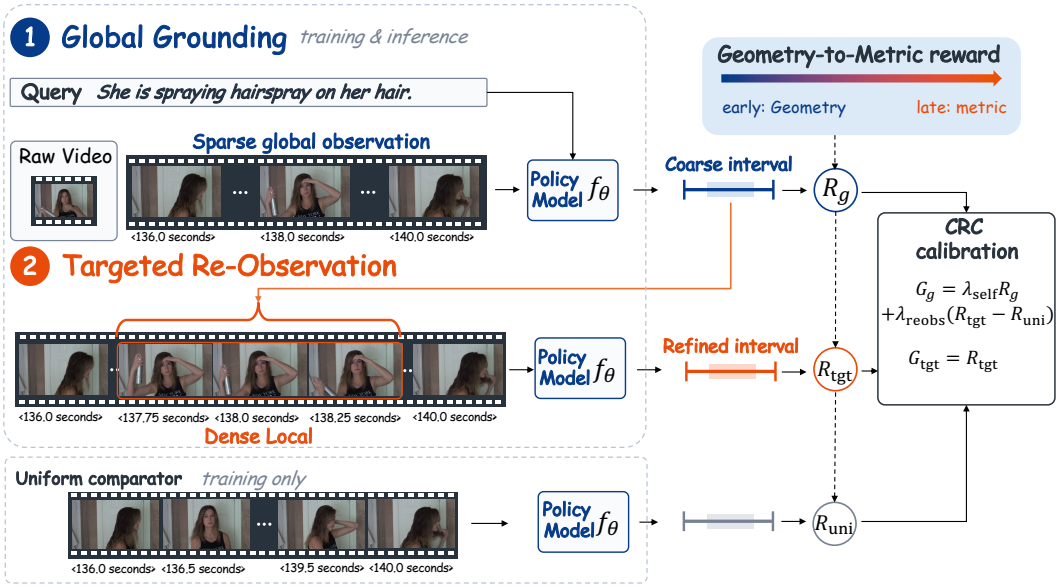


Figure 2: **Overall framework of TROVE.** A shared policy first grounds the query from sparse global frames, then refines its prediction using dense local evidence and sparse global support. Geometry-to-Metric rewards score the outputs; CRC calibrates the global return using the targeted–uniform reward difference. Real-video thumbnails and source-time labels illustrate sampling, with omitted frames indicated by ellipses.

context. For a span $[\bar{s}, \bar{e}]$ clipped to the source boundaries, let $\ell = \bar{e} - \bar{s}$ and $[z]_+ = \max(z, 0)$. The padding and window are

$$p = \min\left\{\max\left(\frac{D}{K}, \frac{[L_{\min} - \ell]_+}{2}\right), \frac{[\rho D - \ell]_+}{2}\right\},$$
$$W(\tilde{y}, D) = [\max(0, \bar{s} - p),\ \min(D, \bar{e} + p)]. \tag{3}$$

Here D/K provides duration-relative context, while $L_{\min}$ encourages a sufficient span for short predictions. The ratio ρ limits window expansion to control the number of additional frames sampled at the fixed local rate. The window is constructed directly from the predicted timestamps and source duration.

Targeted Re-Observation. To refine the interval without discarding full-video context, we combine sparse global samples with dense candidates acquired inside the window:

$$\mathcal{G} = \text{Global}(V; B_g), \qquad \mathcal{L} = \text{Local}(V, W; B_\ell(W)),$$
$$\mathcal{O}_{\text{tgt}} = \text{BudgetMerge}(\mathcal{G}, \mathcal{L}; B_t), \quad \hat{y}_{\text{tgt}} = f_\theta(q, D, \mathcal{O}_{\text{tgt}}). \tag{4}$$

Here $B_g \le B_t$ counts global samples, $B_\ell(W)$ counts local candidates according to window length and nominal rate r_ℓ, and B_t is the uniform-equivalent frame budget induced by nominal rate r_b. BudgetMerge retains the global samples, removes duplicate source indices, and evenly selects local candidates within the remaining budget. The merged frames are ordered by source time, forming one sequence with increased temporal density around the predicted event. Original frame indices and source fps are retained for the native video processor's timestamp encoding, so both observations use full-video time coordinates.

Sparse global support preserves surrounding activities and competing occurrences, while local frames acquired from the raw video expose finer boundary evidence. The second call receives a fresh context containing the query, duration, and merged observation; the first response and its visual tokens are not appended. The initial prediction therefore guides refinement through the evidence it selects, while the final interval is predicted in full-video coordinates.

3.3 LEARNING TEMPORAL GROUNDING

Training follows the same progression from localization to observation-guided refinement. *Grounding Initialization* establishes interval prediction from frames uniformly sampled at nominal rate r_b. *Joint Re-Observation Learning* then trains both actions along the prediction-driven trajectory. Both stages use the Geometry-to-Metric reward to score interval quality; joint learning additionally uses CRC to connect the initial prediction to its downstream observation outcome.

From geometric recovery to metric alignment. An overlap-only reward cannot distinguish disjoint predictions even when their distances from the target differ substantially. To provide geometric feedback in this region, we use temporal generalized IoU, adapting the correction of Rezatofighi et al. (2019) to intervals. For $u = \mathrm{tIoU}(\hat{y}, y^*)$, it is defined as $g = u - (|C| - |\hat{y} \cup y^*|)/|C|$, where C is the smallest enclosing interval. The enclosing-gap term orders disjoint spans by their temporal separation, extending the learning signal beyond the region of positive overlap.

For precise localization, we combine continuous overlap quality with the thresholds used in evaluation:

$$R_{\mathrm{metric}}(u) = \eta u + (1-\eta) \sum_{j=1}^{m} w_j \sigma\left(\frac{u - \tau_j}{T}\right), \qquad \sum_j w_j = 1, \tag{5}$$

Here σ is the sigmoid function, η balances overlap and threshold terms, w_j weights threshold τ_j, and T controls transition sharpness. The overlap term distinguishes predictions within a recall band, while the sigmoid terms smoothly emphasize transitions across the evaluation thresholds.

Geometry-to-Metric Reward Curriculum. To connect geometric recovery with metric alignment, we linearly transfer reward weight between the two objectives over update boundaries $k_0 < k_1$:

$$\beta_k = \mathrm{clip}\left(\frac{k_1 - k}{k_1 - k_0}, 0, 1\right), \qquad R_k(\hat{y}, y^*) = \beta_k g + (1 - \beta_k) R_{\mathrm{metric}}(u). \tag{6}$$

As β_k decreases, supervision moves from graded geometric feedback to overlap quality and high-threshold localization. Each training stage starts its own update index.

3.4 LEARNING THE OBSERVATION DECISION

The first action determines both an interval and the evidence available to the second action. We supervise its observation role using the downstream grounding outcome, with uniform observation providing an online reference for the same query and policy:

$$\mathcal{O}_{\mathrm{uni}} = \mathrm{Uniform}(V; B_t), \qquad \hat{y}_{\mathrm{uni}} = f_\theta(q, D, \mathcal{O}_{\mathrm{uni}}). \tag{7}$$

For each valid global prediction, we sample a separate comparator response from a deterministic uniform frame set at nominal rate r_b. The targeted and uniform branches share the policy, query, fresh-context protocol, and frame ceiling B_t. The targeted branch allocates frames according to $\tilde{y}$ within this ceiling.

Contrastive Re-Observation Calibration. Let $R_g = R_k(\tilde{y}, y^*)$, $R_{\mathrm{tgt}} = R_k(\hat{y}_{\mathrm{tgt}}, y^*)$, and $R_{\mathrm{uni}} = R_k(\hat{y}_{\mathrm{uni}}, y^*)$. CRC combines direct localization feedback with the targeted–uniform difference in downstream grounding rewards:

$$\Delta_{\mathrm{reobs}} = \mathbb{I}[\mathcal{C}](R_{\mathrm{tgt}} - R_{\mathrm{uni}}), \quad G_g = \lambda_{\mathrm{self}} R_g + \lambda_{\mathrm{reobs}} \Delta_{\mathrm{reobs}}, \quad G_{\mathrm{tgt}} = R_{\mathrm{tgt}}. \tag{8}$$

The indicator $\mathbb{I}[\mathcal{C}]$ activates calibration when all three responses are valid and the targeted branch is executed rather than uniform fallback. The self term rewards an accurate initial interval; the contrast term increases its return when targeted observation outperforms the uniform reference and decreases it in the reverse case. The second action receives its own grounding reward. The comparator supplies a scalar training reward and contributes no policy-loss tokens.

Action-wise optimization. For each query, we sample M global responses and their induced second observations. We compute advantages separately for the global and refined actions within this query group:

$$A_{h,i} = \frac{G_{h,i} - \mu_h}{\sigma_h + \epsilon}, \qquad h \in \{g, \mathrm{tgt}\}. \tag{9}$$

Here μ_h and σ_h are the return mean and standard deviation over available responses of action type h. Each refined response is evaluated in the context induced by its corresponding global prediction. The resulting advantages supervise the two sets of response tokens through clipped GRPO (Shao et al., 2024), updating one shared policy. Appendix B gives the token-level objective and complete algorithms.

4 EXPERIMENTS

We evaluate TROVE on TimeLens-Bench across model scales, with ablations of observation strategies, joint training, the Geometry-to-Metric reward curriculum, and CRC. We further analyze the evidence selected by initial predictions and how refinement changes boundary accuracy and success at different overlap thresholds.

4.1 EXPERIMENTAL SETUP

Data and evaluation. We evaluate on TimeLens-Bench (Zhang et al., 2026b), which refines annotations from Charades-STA, ActivityNet Captions, and QVHighlights (Gao et al., 2017; Krishna et al., 2017; Lei et al., 2021). The three subsets contain 3,363, 4,500, and 1,541 queries, respectively. We follow the TimeLens-Bench top-1 evaluation protocol and report R@0.3, R@0.5, R@0.7, and mIoU for each subset.

Model and training. We use Qwen3.5-2B, Qwen3.5-4B, and Qwen3.5-9B (Qwen Team, 2026) and train on the same 2,500-example TimeLens-100K subset with a frozen visual encoder and the ROLL framework (Wang et al., 2025b). Ablations and trajectory analyses use the 4B model. Grounding Initialization uses GRPO with eight responses per prompt. Joint Re-Observation Learning starts from the initialized policy and updates both grounding actions, with $\lambda_{\mathrm{self}} = \lambda_{\mathrm{reobs}} = 0.5$. Both stages use the geometry-to-metric transition from updates $k_0 = 10$ to $k_1 = 20$. The metric target uses $\eta = 0.9$, equally weighted thresholds $\{0.3, 0.5, 0.7\}$, and $T = 0.05$.

Observation protocol. We set the nominal rates to $r_g = 1$, $r_\ell = 4$, and $r_b = 2$ fps, with a 448-frame cap per call. Window construction uses $K = 32$, $L_{\min} = 2$ s, and $\rho = 0.34$. Appendices C–E provide the training-data, implementation, and evaluation details.

4.2 BENCHMARK PERFORMANCE AND MODEL SCALE

With the 4B backbone, TROVE achieves 57.7, 57.6, and 69.2 mIoU on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens, respectively (Table 1). Relative to native Qwen3.5-4B, it improves mIoU by 11.7, 8.5, and 4.5 points and R@0.7 by 15.3, 11.1, and 5.8 points. The gains in both metrics show that the improvement extends to precise temporal alignment, where predicted intervals must closely match the target. It also exceeds the reported TimeLens-8B results by 2.5, 4.4, and 3.7 mIoU points and by 3.2, 6.3, and 4.9 R@0.7 points across the three subsets. This performance is attained at a smaller model scale, using 2,500 training examples and a frozen visual encoder. The complete learned re-observation trajectory adds 1.4, 2.4, and 2.6 mIoU points over Grounding Initialization with uniform inference. It also improves R@0.7 by 1.6, 3.4, and 3.1 points, showing that the learned trajectory increases the fraction of predictions meeting a strict overlap criterion even after initialization has strengthened interval prediction.

REVISOR (Li et al., 2026a) does not use re-observation for VTG: its official VTG code performs fixed-input, single-pass inference. VideoTemp-o3 (Liu et al., 2026) permits further clip requests, but 91.4% of queries terminate after one call in our evaluation, leaving re-observation unused for most queries. Most additional rounds occur on QVHighlights; across the full benchmark, VideoTemp-o3 processes an average of 231.8 cumulative input frames per query. Thus, support for clip requests alone does not ensure that additional evidence is used for VTG. TROVE explicitly learns and executes

a second grounding step using evidence selected by the initial interval. The fixed two-call trajectory allocates denser observations to the predicted region while retaining sparse global context. Joint training rewards the initial prediction for the grounding utility of its selected observation and trains the second prediction to refine the interval. This makes evidence selection and boundary refinement explicit learning targets for the shared policy. Both calls reuse the same parameters, so refinement requires no separate model. The gains over initialization support this shared-policy design within the same backbone.

Model scale. Table 1 shows that Grounding Initialization improves mIoU and R@0.7 for every backbone on all three subsets. Re-observation then adds 3.3 mIoU points on both Charades and ActivityNet at 2B, and 2.2 and 2.7 points at 9B. These gains extend to R@0.7, with improvements of 2.3 and 3.8 points at 2B, and 0.5 and 1.9 points at 9B, supporting more precise interval localization. On QVHighlights, the 4B and 9B variants both reach 69.2 mIoU, exceeding their initialized policies by 2.6 and 2.3 points, while the 2B variant retains the initialization gains with a further 0.1-point increase. The consistent gains over initialization in both metrics support learning to re-observe across model capacities: the learned trajectory builds on the initial grounding capability to improve both overall overlap and precise localization.

Table 1: **TimeLens-Bench performance (%, higher is better).** † denotes our reproduction; other external results follow the protocols reported by TimeLens (Zhang et al., 2026b). Bold denotes the best listed value. At each model scale, Native and Grounding Initialization use uniform 2-fps inference; TROVE uses both training stages and targeted inference. Appendix Table 3 includes additional external methods.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
Model	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	mIoU
GPT-5	59.3	42.0	22.0	40.5	57.4	44.9	30.4	42.9	72.4	60.4	46.4	56.8
Gemini-2.5-Pro	74.1	61.1	34.0	52.8	72.3	64.2	47.1	58.1	**84.1**	**75.9**	**61.1**	**70.4**
Qwen2.5-VL-7B	59.7	37.8	16.6	39.3	44.1	31.0	16.1	31.4	41.5	27.8	15.2	31.6
TimeLens-7B	70.5	55.6	28.4	48.8	62.8	51.0	32.6	46.2	74.1	62.7	43.1	56.0
Qwen3-VL-8B	69.2	53.4	27.5	48.3	62.1	51.2	34.4	46.8	74.2	64.6	49.3	59.4
TimeLens-8B	76.6	63.0	35.2	55.2	68.9	58.4	40.6	53.2	80.2	71.6	55.5	65.5
VideoTemp-o3†	58.9	37.7	17.3	39.8	55.1	41.4	25.0	40.0	70.2	57.2	40.4	53.1
REVISOR†	52.2	30.6	14.0	35.1	46.0	29.1	14.6	32.0	63.3	46.9	30.0	46.0
Qwen3.5-2B	63.9	41.3	21.5	42.9	57.6	45.1	29.9	43.0	71.2	60.6	47.0	57.0
+ Initialization	67.8	51.9	30.9	48.7	62.7	52.8	37.7	49.2	80.4	71.5	56.5	66.2
+ TROVE	71.8	57.8	33.2	52.0	66.6	56.6	41.5	52.5	80.4	71.8	56.6	66.3
Qwen3.5-4B	69.6	45.1	23.1	46.0	65.3	52.0	35.8	49.1	80.5	69.3	54.6	64.7
+ Initialization	79.1	64.9	36.8	56.3	70.2	60.9	43.5	55.2	80.7	71.8	57.3	66.6
+ TROVE	81.0	66.7	**38.4**	57.7	72.3	63.2	46.9	57.6	82.9	74.2	60.4	69.2
Qwen3.5-9B	70.7	45.6	24.0	47.2	65.4	51.6	35.7	49.4	77.9	67.4	53.9	63.8
+ Initialization	80.6	64.1	37.0	56.5	71.8	61.9	45.6	56.5	81.6	72.5	57.1	66.9
+ TROVE	**82.6**	**69.4**	37.5	**58.7**	**74.4**	**66.1**	**47.5**	**59.2**	83.3	75.3	58.9	69.2

4.3 TRAINING AND OBSERVATION

Table 2 compares uniform, similarity-guided, and targeted observation for the initialized and jointly learned 4B policies. Targeted improves mIoU over both uniform rates on every subset with either policy. Its mean cumulative input B_Σ is 271.4 frames with Initialization and 264.1 with Joint Learning, compared with 274.3 for uniform 3 fps. Targeted re-observation thus improves grounding with fewer input frames on average than the denser uniform protocol.

With the jointly learned policy, Targeted improves mIoU over CLIP-guided observation by 1.7, 1.4, and 0.4 points on Charades, ActivityNet, and QVHighlights, respectively. These comparisons support using the grounding policy's own interval prediction to select useful evidence for refinement.

Joint learning further improves targeted mIoU by 1.6 and 2.4 points on ActivityNet and QVHighlights, while maintaining similar performance on Charades. The corresponding gains are 0.6 and 1.5

points under uniform 3-fps inference. The larger gains with targeted inference support training the policy along the observation trajectory it follows at test time.

Table 2: **Training and observation strategies (4B, %).** CLIP-guided uses similarity-based selection; Targeted uses two-call re-observation. Mean B_Σ over 9,404 queries is shown as selection + final grounding, or grounding alone for Uniform. Stage means and cumulative totals are rounded independently. Implementation details are given in Appendix C.1.

Policy	Inference	Charades		ActivityNet		QVHighlights		Mean B_Σ (frames)
		mIoU	R@0.5	mIoU	R@0.5	mIoU	R@0.5	
Initialization	Uniform 2 fps	56.3	64.9	55.2	60.9	66.6	71.8	202.5
	Uniform 3 fps	57.6	65.9	55.2	61.0	66.1	71.9	274.3
	CLIP-guided	56.1	64.4	55.0	60.6	67.0	71.8	211.6 + 199.6
	Targeted	57.9	66.4	56.0	61.6	66.8	72.2	101.2 + 170.2
Joint Learning	Uniform 2 fps	56.2	65.1	55.7	60.9	68.1	73.1	202.5
	Uniform 3 fps	57.5	66.2	55.8	61.0	67.6	73.4	274.3
	CLIP-guided	56.0	63.7	56.2	61.4	68.8	74.0	211.6 + 199.6
	Targeted	57.7	66.7	57.6	63.2	69.2	74.2	101.2 + 163.0

4.4 Learning Objectives

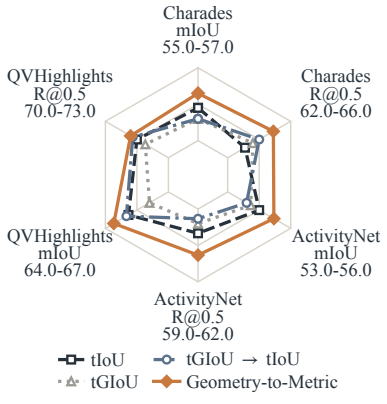


Figure 3: **Grounding Initialization rewards** (4B). Uniform 2-fps evaluation; spoke labels show axis-specific percentage ranges.

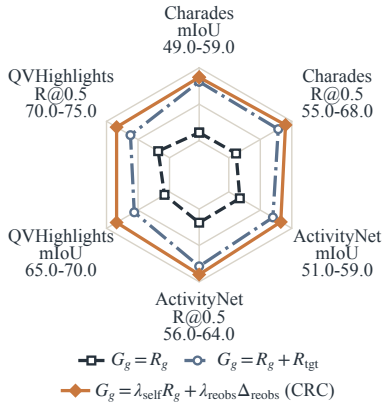


Figure 4: **Global-action calibration** (4B). Refined grounding under three definitions of G_g; CRC uses a training-only uniform comparator.

Effect of the reward curriculum. The Geometry-to-Metric curriculum achieves the highest mIoU among the four Grounding Initialization reward configurations on all three subsets (Figure 3). It improves over tIoU-only training by 0.4, 0.7, and 0.7 mIoU points. The metric endpoint is particularly effective: replacing the terminal tIoU objective in the tGIoU-to-tIoU schedule adds 0.7, 1.3, and 0.6 mIoU points and 0.9, 1.5, and 0.1 R@0.5 points. The complete reward therefore combines the useful geometric ordering of early predictions with an endpoint that rewards overlap quality and evaluation-threshold transitions.

The training rollouts illustrate these complementary roles. During geometry-only updates 0–9, tGIoU provides graded feedback to 11.6% of valid rollouts that tIoU maps to zero. After the transition, at updates 21–99, 91.2% of rollouts overlap the target, with 63.7% and 45.8% reaching tIoU thresholds 0.5 and 0.7. The curriculum follows this progression from locating the target region to improving boundary precision.

Effect of re-observation calibration. Figure 4 compares three returns for Global Grounding. Moving from the self-only return R_g to $R_g + R_{\text{tgt}}$ adds supervision from the grounding outcome of the selected observation, improving final mIoU by 7.0, 4.2, and 2.4 points. CRC then calibrates this downstream outcome using the targeted–uniform reward difference, yielding a further 0.6, 1.0, and

1.4 mIoU points and 1.5, 0.9, and 1.1 R@0.5 points. The two successive gains show the value of supervising the first prediction through the grounding it enables and calibrating that outcome against uniform observation.

4.5 Observation and Refinement Analysis

Observation utility and boundary localization. A coarse interval can select useful evidence before its boundaries are accurate enough as a final answer. Among 3,692 Global Grounding predictions with tIoU below 0.5, 45.5% already induce windows covering the full target. Refinement brings 39.0% of these 1,680 cases to tIoU at least 0.5, illustrating how the initial prediction supports grounding through its observation decision. The resulting corrections also improve endpoint accuracy (Figure 5a). Refinement reduces mean boundary error relative to both Initialization and Global Grounding on every subset, improving over Initialization by 0.1, 0.6, and 0.6 seconds on Charades, ActivityNet, and QVHighlights.

Net refinement gains. Figure 5b tracks predictions crossing each evaluation threshold after refinement across all 9,404 queries. At R@0.7, 1,059 predictions become successful and 543 lose success, producing a net gain of 516; net gains are positive at every threshold on every subset. Of the new R@0.7 successes, 696 begin in $[0.5, 0.7)$, showing how re-observation converts approximate localization into more precise grounding. The improvements also include substantial changes: 732 predictions move from tIoU below 0.65 to at least 0.75, versus 340 in the reverse direction. Together, these results connect targeted evidence acquisition to lower boundary errors and net improvements in localization quality. Appendix G provides real-video examples of successful refinement and the behavior of compound and short events.

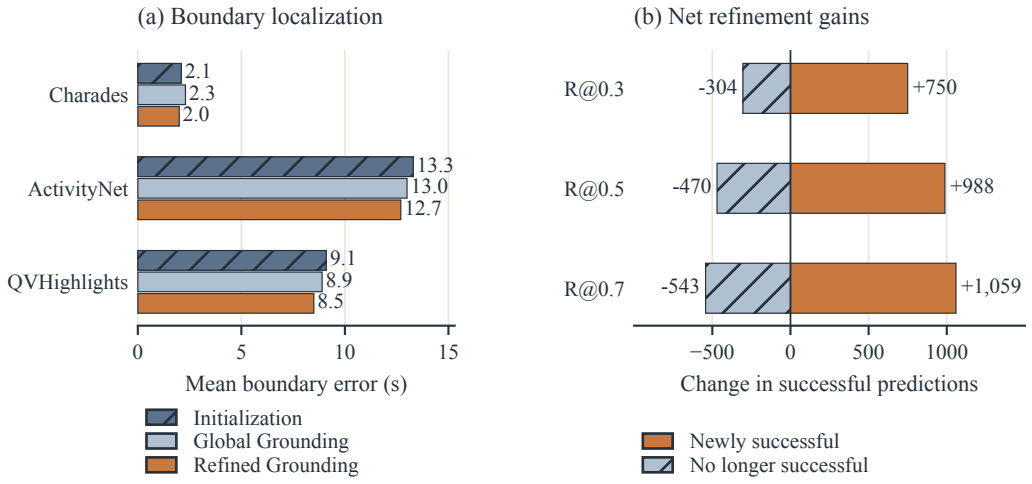


Figure 5: **Observation utility and refinement gains** (4B; 9,404 queries). (a) Mean absolute endpoint error in seconds. (b) Predictions newly meeting each tIoU threshold after refinement (right) and no longer meeting it (left), relative to Global Grounding. Labels give query counts; unchanged success status is omitted.

5 Conclusion

TROVE learns temporal grounding as a trajectory in which an initial interval guides evidence acquisition for refinement. Its shared policy combines sparse global context with dense local observations, while the Geometry-to-Metric curriculum and CRC train interval prediction and observation utility together. Experiments across TimeLens-Bench subsets and model scales demonstrate improved grounding, supported by training–inference ablations, reduced boundary errors, and net gains at strict overlap thresholds. These results highlight the central benefit of re-observation: an approximate first prediction can select the evidence needed for a more precise final answer.

AI USE STATEMENT

During the preparation of this manuscript, large language models (LLMs) were utilized in a strictly limited and supportive capacity. Their application was confined to refining grammar, enhancing stylistic flow, and improving overall readability, without any modification to the underlying technical content. Crucially, LLMs played no role in conceptualizing the research, designing the methodology, executing experiments, interpreting data, or formulating conclusions. The entirety of the technical contributions presented herein remains the exclusive work of the authors.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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*, 2025b.

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. doi: 10.18653/v1/2025.emnlp-industry.66. URL https://aclanthology.org/2025.emnlp-industry.66/.

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.

Rong Fan, Kaiyan Xiao, Minghao Zhu, Liuyi Wang, Kai Dai, and Zhao Yang. GroundVTS: Visual token sampling in multimodal large language models for video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10408–10418, 2026.

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. 5277–5285, 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, 2025a.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. TRACE: Temporal grounding video LLM via causal event modeling. In *International Conference on Learning Representations*, pp. 89594–89618, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/df027cf11469e746ef94d583f9f5537f-Paper-Conference.pdf.

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.

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, 2024a.

Kai Huang, Hao Zou, Ye Xi, Bochen Wang, Zhen Xie, and Liang Yu. IVTP: Instruction-guided visual token pruning for large vision-language models. In *European Conference on Computer Vision*, pp. 214–230, 2024b.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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. QVHighlights: Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021.

Jiaze Li, Hao Yin, Wenhui Tan, Jingyang Chen, Boshen Xu, Yuxun Qu, Yijing Chen, Jianzhong Ju, Zhenbo Luo, and Jian Luan. REVISOR: Beyond textual reflection, towards multimodal introspective reasoning in long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5059–5069, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_REVISOR_Beyond_Textual_Reflection_Towards_Multimodal_Introspective_Reasoning_in_Long-Form_CVPR_2026_paper.html.

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*, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. VideoChat-Flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, pp. 109089–109117, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b14d7175755b180dc2163e15e3110cb6-Paper-Conference.pdf.

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, Kaiyu Tang, Bin Wen, Fan Yang, Tingting Gao, Han Li, Yinwei Wei, and Xuemeng Song. VideoTemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. *arXiv preprint arXiv:2602.07801*, 2026. URL https://arxiv.org/abs/2602.07801.

OpenAI. Introducing GPT-5. OpenAI announcement, 2025. August 7, 2025.

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, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 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. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 959–975, 2025a. doi: 10.18653/v1/2025.findings-emnlp.50. URL https://aclanthology.org/2025.findings-emnlp.50/.

Weixun Wang, Shaopan Xiong, Gengru Chen, Wei Gao, Sheng Guo, Yancheng He, Ju Huang, Jiaheng Liu, Zhendong Li, Xiaoyang Li, et al. Reinforcement learning optimization for large-scale learning: An efficient and user-friendly scaling library. *arXiv preprint arXiv:2506.06122*, 2025b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 83330–83364, 2025c. doi: 10.52202/085713-2793. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7801b29c93b599b8d0c44138596bdeed-Paper-Conference.pdf.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3272–3283, 2025d.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13754–13765, 2025.

Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S. Davis. AdaFrame: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1278–1287, 2019.

Xiaomi LLM-Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, et al. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *International Conference on Learning Representations*, pp. 38057–38081, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5e8309c9ca683e11672e3dbcd4b87776-Paper-Conference.pdf.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_Thinking_With_Videos_Multimodal_Tool-Augmented_Reinforcement_Learning_for_Long_Video_CVPR_2026_paper.html.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_TimeLens_Rethinking_Video_Temporal_Grounding_with_Multimodal_LLMs_CVPR_2026_paper.html.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 74840–74857, 2025.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *International Conference on Learning Representations*, pp. 11302–11329, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1332893b662f655660c9abdf793230cf-Paper-Conference.pdf.

A Observation Density and Interval Geometry

Figure 6 supplements the motivation in Section 1: the left panel examines uniform sampling density, and the right panel illustrates the interval geometry used by the grounding reward.

Aggregate metrics for the FPS sweep. The native-model FPS sweep in Figure 6 (left) reports mIoU over all $N = 9{,}404$ TimeLens-Bench queries:

$$\mathrm{mIoU}_{\mathrm{all}} = \frac{100}{N} \sum_{i=1}^{N} \mathrm{tIoU}(\hat{y}_i, y_i^*), \tag{10}$$

where $\hat{y}_i$ and y_i^* are the predicted and ground-truth intervals for query i. Each query has equal weight; equivalently, the subset mIoUs are weighted by their query counts (4,500 for ActivityNet, 3,363 for Charades, and 1,541 for QVHighlights), rather than averaged equally across the three subsets. The horizontal axis likewise averages model-input frame counts over these 9,404 queries.

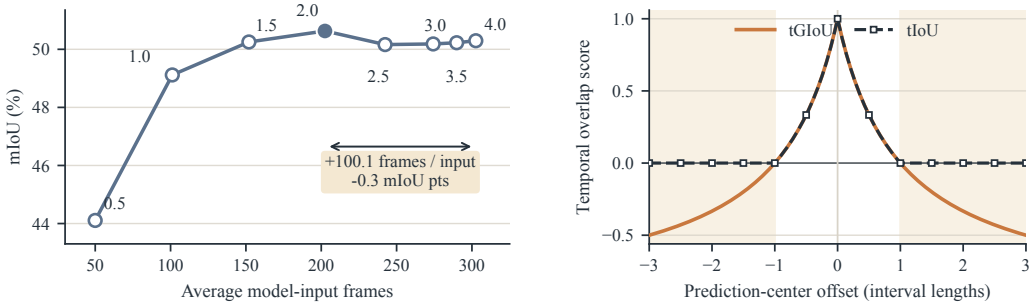


Figure 6: **Motivating targeted observation and geometry-aware learning.** *Left:* uniform densification of native Qwen3.5-4B inputs on 9,404 TimeLens-Bench queries yields diminishing localization returns; capped inputs retain full-video coverage. *Right:* translating an equal-length prediction relative to its target illustrates how tGIoU distinguishes disjoint intervals that all receive zero tIoU.

B Full Method Derivation and Algorithms

Observation budget. For a trajectory with J calls, let $n_j = |\mathcal{O}^{(j)}|$ denote the input frame count of call j. The cumulative count is $B_\Sigma = \sum_{j=1}^{J} n_j$, as defined in Section 3.1. Let N be the number of source frames, F the processor frame factor, and

$$B(D; r) = \lfloor \min\{N, B_{\max}, \max(rD, B_{\min})\} \rfloor_F, \qquad \lfloor z \rfloor_F = F \lfloor z/F \rfloor, \tag{11}$$

be the count induced by nominal rate r. Using the notation of Section 3.2, set $B_t = B(D; r_b)$, $B_g = \min\{B_t, B(D; r_g)\}$, and $B_\ell(W) = \max\{F, \lfloor |W| r_\ell \rfloor_F\}$. Then $\mathcal{G} = \mathrm{Lin}(V, [0, D], B_g)$ and $\mathcal{L} = \mathrm{Lin}(V, W, B_\ell(W))$, and

$$\mathrm{BudgetMerge}(\mathcal{G}, \mathcal{L}; B_t) = \mathrm{SortUnique}(\mathcal{G} \cup \mathrm{Even}(\mathcal{L} \setminus \mathcal{G}, B_t - B_g)). \tag{12}$$

$\mathrm{Even}(A, n)$ keeps all elements when $|A| \le n$ and otherwise selects n source-ordered, evenly spaced elements. Equation 11 is instantiated after rounding the processor minimum and maximum to the required frame factor. The targeted sampler first builds full-video and window-local linspace candidates, removes duplicates, and evenly downsamples only the local candidates to the residual budget $B_t - B_g$, ensuring that its frame count does not exceed the uniform budget $B(D; r_b)$.

Boundary handling. Equation 3 expands short valid proposals with local context, retains proposals already longer than ρD, and intersects every resulting window with the video range $[0, D]$. If Global Grounding does not produce a usable interval, the second call uses the uniform observation, preserving full-video coverage.

Action eligibility and advantages. CRC is active when Global Grounding produces a usable interval, targeted re-observation is executed, and both second-observation responses are valid; otherwise $\Delta_{\text{reobs}} = 0$. Global and targeted responses retain their own grounding rewards, while the uniform comparator supplies only its scalar reward. Returns are normalized separately for the two policy actions as in Equation 9.

Token-level objective. Let π_θ denote the autoregressive policy underlying f_θ. For response token $a_{h,i,j}$ under action context $c_{h,i}$, let $r_{h,i,j}(\theta) = \pi_\theta(a_{h,i,j} \mid c_{h,i}, a_{h,i,<j}) / \pi_{\text{old}}(a_{h,i,j} \mid c_{h,i}, a_{h,i,<j})$. π_{old} is the behavior policy that generated the rollout, whereas π_{ref} is a frozen copy of the policy that initializes the corresponding training schedule and supplies KL regularization. The clipped GRPO objective is

$$\mathcal{L}(\theta) = -\mathbb{E}_{h,i,j}[\min(r_{h,i,j}A_{h,i}, \text{clip}(r_{h,i,j}, 1-\varepsilon, 1+\varepsilon)A_{h,i})] + \lambda_{\text{KL}}\mathbb{E}[D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})]\,. \quad (13)$$

CRC scope. CRC uses uniform observation as an online comparator: both second observations use the same query, policy, fresh-context protocol, and frame ceiling B_t, while the targeted construction never exceeds this budget. Their reward difference calibrates the downstream return of Global Grounding against the outcome obtainable from generic uniform observation. For each valid global prediction, a fresh comparator response is sampled from the deterministic uniform frame set before action-wise normalization.

Algorithm 1 Joint Re-Observation Learning: one policy update

Require: Minibatch $\mathcal{B} = \{(V, q, D, y^*)\}$; policy f_θ; update k;
rollout count M; rates r_g, r_ℓ, r_b; frame factor F
Ensure: Updated policy parameters θ

for all $(V, q, D, y^*) \in \mathcal{B}$ **do**
  $B_t \leftarrow B(D; r_b)$; $B_g \leftarrow \min\{B_t, B(D; r_g)\}$
  $\mathcal{O}_g \leftarrow \text{Lin}(V, [0, D], B_g)$; $\mathcal{G} \leftarrow \text{Global}(V; B_g)$
  $\mathcal{O}_{\text{uni}} \leftarrow \text{Uniform}(V; B_t)$
  for $i = 1, \ldots, M$ **do**
    $\tilde{y}_i \leftarrow f_\theta(q, D, \mathcal{O}_g)$; $\Delta_{\text{reobs},i} \leftarrow 0$
    if $\tilde{y}_i$ is valid **then**
      $W_i \leftarrow W(\tilde{y}_i, D)$ ▷ Equation 3
      $B_{\ell,i} \leftarrow \max\{F, \lfloor |W_i| r_\ell \rfloor_F\}$
      $\mathcal{L}_i \leftarrow \text{Local}(V, W_i; B_{\ell,i})$
      $\mathcal{O}_{\text{tgt},i} \leftarrow \text{BudgetMerge}(\mathcal{G}, \mathcal{L}_i; B_t)$
      $\hat{y}_{\text{uni},i} \leftarrow f_\theta(q, D, \mathcal{O}_{\text{uni}})$ ▷ Fresh comparator rollout
    else
      $\mathcal{O}_{\text{tgt},i} \leftarrow \mathcal{O}_{\text{uni}}$ ▷ Fallback; no comparator rollout
    end if
    $\hat{y}_{\text{tgt},i} \leftarrow f_\theta(q, D, \mathcal{O}_{\text{tgt},i})$ ▷ Fresh context
    $R_{g,i} \leftarrow R_k(\tilde{y}_i, y^*)$; $R_{\text{tgt},i} \leftarrow R_k(\hat{y}_{\text{tgt},i}, y^*)$
    if $\mathbb{I}[\mathcal{C}_i] = 1$ **then** ▷ Eligibility in Equation 8
      $R_{\text{uni},i} \leftarrow R_k(\hat{y}_{\text{uni},i}, y^*)$
      $\Delta_{\text{reobs},i} \leftarrow R_{\text{tgt},i} - R_{\text{uni},i}$
    end if
    $G_{g,i} \leftarrow \lambda_{\text{self}} R_{g,i} + \lambda_{\text{reobs}} \Delta_{\text{reobs},i}$; $G_{\text{tgt},i} \leftarrow R_{\text{tgt},i}$
  end for
  Compute $A_{h,i}$ over available rows of each $h \in \{g, \text{tgt}\}$ via Equation 9
end for
Update θ using global and targeted response tokens via Equation 13
return θ

All calls share the rollout policy; its parameters are updated only after collecting the minibatch. $\mathcal{C}_i$ is false for fallback or invalid responses. Comparator tokens never enter the policy loss.

Algorithm 2 TROVE Inference

Require: Video V, query q, duration D; policy f_θ;
rates r_g, r_ℓ, r_b; frame factor F
Ensure: Refined interval $\hat{y}_{\text{tgt}}$

$B_t \leftarrow B(D; r_b)$; $B_g \leftarrow \min\{B_t, B(D; r_g)\}$
$\mathcal{O}_g \leftarrow \text{Lin}(V, [0, D], B_g)$
$\tilde{y} \leftarrow f_\theta(q, D, \mathcal{O}_g)$
if $\tilde{y}$ yields a usable interval within $[0, D]$ after clipping **then**
  $W \leftarrow W(\tilde{y}, D)$ ▷ Equation 3
  $B_\ell(W) \leftarrow \max\{F, \lfloor |W| r_\ell \rfloor_F\}$
  $\mathcal{G} \leftarrow \text{Global}(V; B_g)$; $\mathcal{L} \leftarrow \text{Local}(V, W; B_\ell(W))$
  $\mathcal{O}_{\text{tgt}} \leftarrow \text{BudgetMerge}(\mathcal{G}, \mathcal{L}; B_t)$
else
  $\mathcal{O}_{\text{tgt}} \leftarrow \text{Uniform}(V; B_t)$ ▷ Fallback
end if
$\hat{y}_{\text{tgt}} \leftarrow f_\theta(q, D, \mathcal{O}_{\text{tgt}})$ ▷ Fresh context; no first answer
return $\hat{y}_{\text{tgt}}$ ▷ No output clamping before evaluation

The uniform comparator used by CRC and all training rewards are absent at inference.

C Implementation Details

Grounding Initialization uses Qwen3.5-4B with the observation protocol in Section 4.1, a frozen visual encoder, and GRPO with eight responses per prompt. For the 4B experiments, both training schedules are configured for 100 updates. Initialization uses a batch of 32 queries, yielding 256 response rollouts per update. The optimizer uses learning rate 10^{-6}, weight decay 10^{-2}, one PPO epoch, KL coefficient 10^{-2}, and bfloat16 arithmetic. Each training schedule starts its reward update index at zero and uses the geometry-to-metric transition from updates 10 to 20.

Joint Re-Observation Learning starts from the Grounding Initialization policy and uses return weights 0.5/0.5 with two policy-loss actions and one training-only comparator generation per valid Global Grounding prediction. Each joint update collects a batch of 32 trajectories in four query groups of eight rollouts each; every trajectory contains one global and one second grounding action. The sampler uses frame factor $F = 2$, minimum $B_{\min} = 4$, and maximum $B_{\max} = 448$ in Equation 11. Window construction uses 32 bins, a 2-second minimum span, and padding ratio 0.34. The strict 9,404-record evaluation observed no fallback, although inference retains the deterministic uniform fallback.

Window ratio and sampling budget. The value $\rho = 0.34$ can be understood through a nominal density allocation. For a window occupying a fraction $\alpha = |W|/D$, allocating r_g fps outside the window and r_ℓ fps inside it gives an average rate $r_g(1-\alpha) + r_\ell\alpha$. Matching this rate to the nominal budget r_b yields

$$\alpha_{\text{nom}} = \frac{r_b - r_g}{r_\ell - r_g} = \frac{1}{3}. \tag{14}$$

The configured $\rho = 0.34$ lies close to this one-third reference and controls window expansion in Equation 3. The discrete sampler retains global support and merges local candidates within the exact second-call frame budget through Equations 11 and 12.

C.1 CLIP-Guided Observation

This baseline replaces only inference-time observation selection, retaining the grounding checkpoint specified in Table 2 without additional training. The selector uses CLIP ViT-L/14 (Radford et al., 2021). Each video is divided into 32 equal-duration bins, sampled at nominal 2 fps with a 1-second minimum decoding span. CLIP scores each bin by its maximum query–frame cosine similarity. The selector ranks bins by score with adjacent-bin suppression, retains the first two candidates, and admits up to two more whose standardized scores are at least 0.5 and whose addition keeps the expanded-window union within 34% of the video duration. Selected bins receive one-bin context on each side; adjacent regions are merged and short windows expanded to 2 seconds. The final

grounding call uses the global–local sampling rates and frame budget of Targeted Re-Observation (Section 4.1). The sampler removes duplicates, preserves global support, and uniformly subsamples excess local candidates.

D Training Data

Both training schedules use the same 2,500-example subset of TimeLens-100K. Each example pairs a source video and a natural-language query with the video duration and one annotated temporal interval $[s^*, e^*]$ in seconds. The subset contains 2,500 unique query identifiers over 2,336 videos, and these fields provide the video, query, duration, and grounding target used by the training objective. TimeLens-100K is constructed by re-annotating videos sampled from CosMo-Cap, InternVid-VTime, DiDeMo, QuerYD, and HiREST (Zhang et al., 2026b).

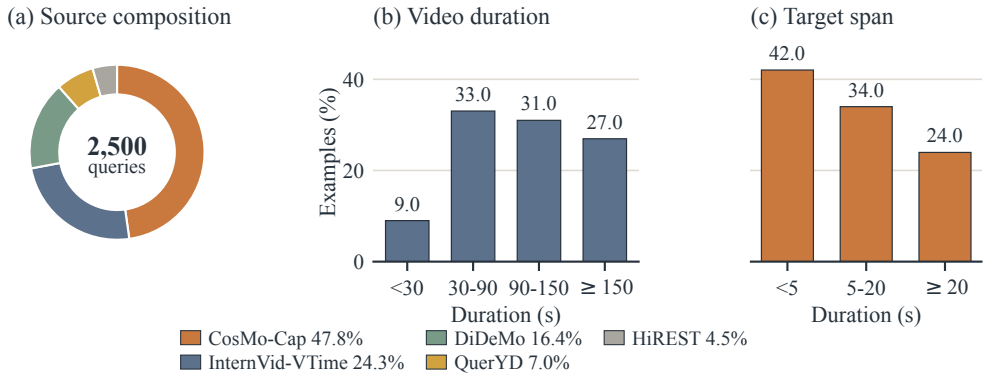


Figure 7: **Training-subset composition** ($n = 2{,}500$). (a) Source distribution; source labels are provided by TimeLens-100K. (b,c) Distributions of full-video duration and annotated target-span duration. Percentages are computed from the selected records.

The subset is selected from a 10,000-example training-only candidate pool using difficulty-stratified native Qwen3.5-4B rollouts. Four bins of mean tIoU over eight rollouts contribute 1,000, 750, 500, and 250 examples from $[0, 0.3)$, $[0.3, 0.5)$, $[0.5, 0.7)$, and $[0.7, 1.0]$, respectively, while preserving coverage over video duration, target length, and target-to-video duration ratio. No TimeLens-Bench examples or predictions enter the selection procedure, and the resulting sample identifiers are fixed across both training schedules and all evaluated model scales.

E Dataset, Prompts, and Evaluation

Charades-TimeLens contributes 3,363 queries, ActivityNet-TimeLens 4,500, and QVHighlights-TimeLens 1,541. All aggregates are recomputed from the same 9,404 official keys. Joint Re-Observation Learning additionally requires exactly one explicit finite interval satisfying $0 \le s < e \le D$, without repairing invalid responses, for the rewards and eligibility used in Equation 8. Official evaluation normalizes endpoint order but does not clamp predictions to video duration before scoring; degenerate intervals follow the official scorer, and all queries remain in the evaluation set.

Evaluation metrics. The TimeLens-Bench top-1 protocol scores one predicted interval $\hat{y}_i$ per query against its ground-truth interval y_i^*. Temporal IoU is $u_i = |\hat{y}_i \cap y_i^*| / |\hat{y}_i \cup y_i^*|$, where $|\cdot|$ denotes temporal duration. For a subset of N_s queries, we report

$$\mathrm{R@}\tau = \frac{100}{N_s} \sum_{i=1}^{N_s} \mathbb{I}[u_i \ge \tau], \qquad \mathrm{mIoU} = \frac{100}{N_s} \sum_{i=1}^{N_s} u_i, \tag{15}$$

with $\tau \in \{0.3, 0.5, 0.7\}$. Both metrics are percentages with equal weight per query within each subset; Appendix A defines the pooled mIoU used for the FPS sweep.

Prompt template. We follow the TimeLens input template (Zhang et al., 2026b), with the query substituted at runtime:

```
<video> 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'.
```

F FULL 2B/4B/9B RESULTS

Table 3: **Complete TimeLens-Bench comparison (%, higher is better).** † denotes our reproduction; results for the other 18 external methods are transcribed from TimeLens Table 1 (Zhang et al., 2026b), retaining their reported protocols. Qwen3.5 rows include Native, Grounding Initialization, and TROVE at 2B, 4B, and 9B. Native and Initialization use uniform 2-fps inference. All values use one-decimal precision; bold denotes the best listed value.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
Model	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	mIoU
Proprietary models												
GPT-4o	60.6	44.5	23.5	41.8	55.2	41.4	25.8	40.4	69.0	54.8	38.5	52.1
GPT-5	59.3	42.0	22.0	40.5	57.4	44.9	30.4	42.9	72.4	60.4	46.4	56.8
Gemini-2.0-Flash	66.4	53.5	27.1	46.7	62.9	54.0	37.7	49.3	76.2	66.4	48.3	60.8
Gemini-2.5-Flash	68.7	56.1	30.6	48.6	66.8	57.5	41.3	52.5	78.2	69.4	55.0	64.3
Gemini-2.5-Pro	74.1	61.1	34.0	52.8	72.3	64.2	47.1	58.1	**84.1**	**75.9**	**61.1**	**70.4**
Open-source models												
VideoChat-Flash-7B	60.2	37.9	17.8	39.7	35.5	21.8	10.5	24.8	45.2	30.6	16.7	32.7
VideoChat-R1-7B	51.9	30.8	11.7	33.7	35.0	23.9	11.3	25.0	29.3	19.1	9.4	21.5
Time-R1-7B	57.9	32.0	16.9	36.6	44.8	31.0	19.0	33.1	65.8	51.5	36.1	49.2
TRACE	37.2	21.8	9.6	27.1	43.4	33.9	22.0	32.7	49.7	39.1	28.1	39.0
TRACE-uni	38.2	22.9	10.4	28.1	44.3	35.1	22.6	33.6	49.9	40.0	29.2	39.8
TimeSuite	56.3	35.5	18.0	38.1	27.1	17.5	8.6	19.8	27.1	16.9	9.9	21.7
Grounded-VideoLLM	43.3	28.7	13.5	30.0	39.2	29.6	19.5	30.0	43.7	33.8	22.5	33.4
MiMo-VL-7B	57.9	42.6	20.5	39.6	49.3	38.7	22.4	35.5	57.1	42.6	28.4	41.5
Qwen2.5-VL-7B	59.7	37.8	16.6	39.3	44.1	31.0	16.1	31.4	41.5	27.8	15.2	31.6
TimeLens-7B	70.5	55.6	28.4	48.8	62.8	51.0	32.6	46.2	74.1	62.7	43.1	56.0
Qwen3-VL-235B-A22B	71.7	50.8	24.5	47.8	69.0	57.5	39.3	52.2	79.6	70.2	54.5	64.6
Qwen3-VL-8B	69.2	53.4	27.5	48.3	62.1	51.2	34.4	46.8	74.2	64.6	49.3	59.4
TimeLens-8B	76.6	63.0	35.2	55.2	68.9	58.4	40.6	53.2	80.2	71.6	55.5	65.5
VideoTemp-o3†	58.9	37.7	17.3	39.8	55.1	41.4	25.0	40.0	70.2	57.2	40.4	53.1
REVISOR†	52.2	30.6	14.0	35.1	46.0	29.1	14.6	32.0	63.3	46.9	30.0	46.0
Qwen3.5 backbones and our configurations												
Qwen3.5-2B	63.9	41.3	21.5	42.9	57.6	45.1	29.9	43.0	71.2	60.6	47.0	57.0
+ Initialization	67.8	51.9	30.9	48.7	62.7	52.8	37.7	49.2	80.4	71.5	56.5	66.2
+ TROVE	71.8	57.8	33.2	52.0	66.6	56.6	41.5	52.5	80.4	71.8	56.6	66.3
Qwen3.5-4B	69.6	45.1	23.1	46.0	65.3	52.0	35.8	49.1	80.5	69.3	54.6	64.7
+ Initialization	79.1	64.9	36.8	56.3	70.2	60.9	43.5	55.2	80.7	71.8	57.3	66.6
+ TROVE	81.0	66.7	**38.4**	57.7	72.3	63.2	46.9	57.6	82.9	74.2	60.4	69.2
Qwen3.5-9B	70.7	45.6	24.0	47.2	65.4	51.6	35.7	49.4	77.9	67.4	53.9	63.8
+ Initialization	80.6	64.1	37.0	56.5	71.8	61.9	45.6	56.5	81.6	72.5	57.1	66.9
+ TROVE	**82.6**	**69.4**	37.5	**58.7**	**74.4**	**66.1**	**47.5**	**59.2**	83.3	75.3	58.9	69.2

Table 3 compares TROVE with the proprietary GPT-4o (Hurst et al., 2024), GPT-5 (OpenAI, 2025), Gemini-2.0-Flash (Zhang et al., 2026b), Gemini-2.5-Flash (Comanici et al., 2025), and Gemini-2.5-Pro (Comanici et al., 2025). The open-source comparison includes VideoChat-Flash-7B (Li et al., 2026b), VideoChat-R1-7B (Li et al., 2025), Time-R1-7B (Wang et al., 2025c), TRACE (Guo et al., 2025b), TRACE-uni (Guo et al., 2025b), TimeSuite (Zeng et al., 2025), Grounded-VideoLLM (Wang et al., 2025a), MiMo-VL-7B (Xiaomi LLM-Core Team et al., 2025),

Qwen2.5-VL-7B (Bai et al., 2025b), TimeLens-7B (Zhang et al., 2026b), TimeLens-8B (Zhang et al., 2026b), Qwen3-VL-235B-A22B (Bai et al., 2025a), Qwen3-VL-8B (Bai et al., 2025a), and our reproductions of VideoTemp-o3 (Liu et al., 2026) and REVISOR (Li et al., 2026a). The high-threshold results complement the mIoU gains across model scales in Table 1. Relative to initialization, 2B TROVE improves R@0.7 by 2.3 points on Charades and 3.8 on ActivityNet; the corresponding 9B improvements are 0.5 and 1.9 points (Table 3). On QVHighlights, both 4B and 9B reach 69.2 mIoU at the reported precision, with R@0.7 of 60.4 and 58.9, respectively.

G Real-Video Qualitative Examples

Figures 8–11 present two successful refinements and two diagnostic failures from aligned 4B evaluation records. Each case compares the uniform 2-fps Qwen3.5-4B prediction, Global Grounding, and TROVE against the ground-truth interval on a shared source-time axis. The successful cases were selected to show a progression from native inference to coarse localization and then refined grounding. The failures have interpretable visual causes: a query that joins separated sub-actions can be contracted to one constituent, while the gradual onset of a two-second event makes a one-second boundary shift visually plausible but costly under tIoU. These examples illustrate distinct behaviors rather than estimate their prevalence. All displayed frames are decoded from the corresponding source videos at the labeled times to visualize their content; they are not claimed to be the exact model-input frames.

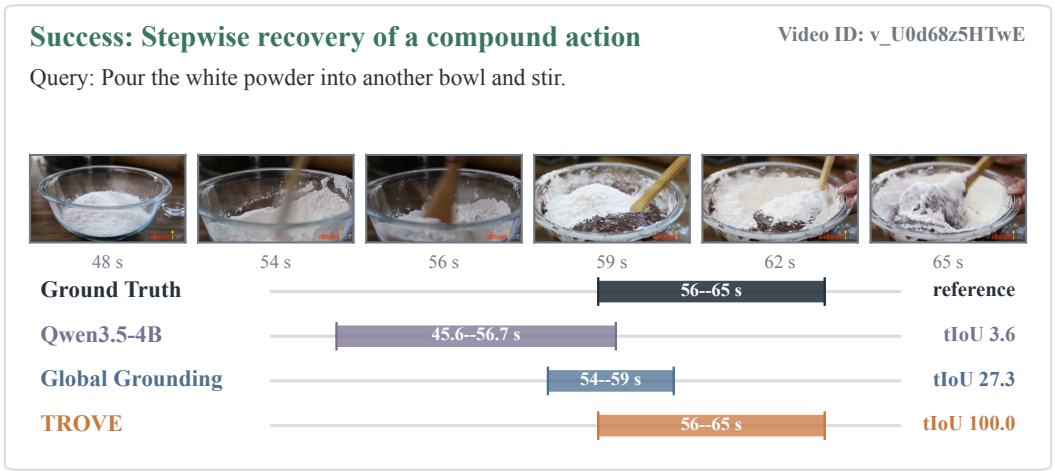


Figure 8: **Success: stepwise recovery of a compound action** (4B). Qwen3.5-4B ends near the start of the annotated pouring-and-stirring event. Global Grounding moves to the correct episode but covers only its early portion, whereas TROVE recovers the complete interval.

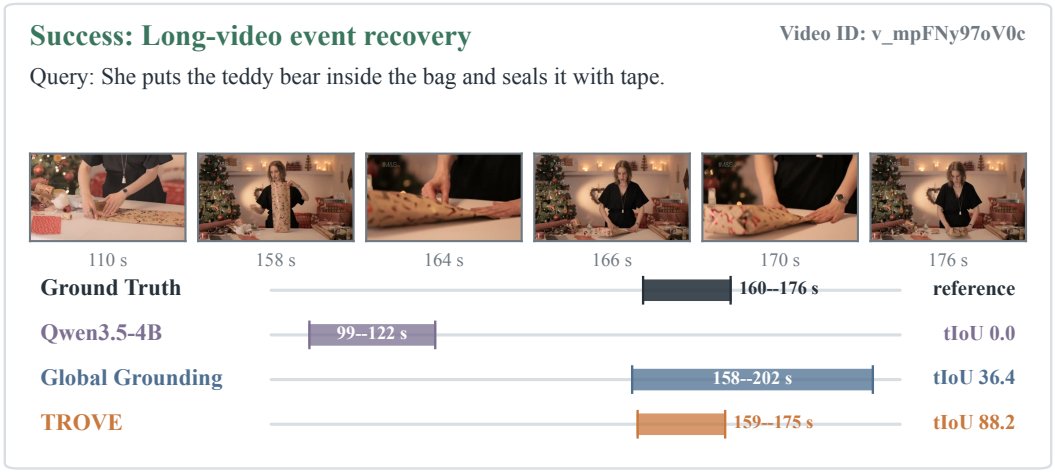


Figure 9: **Success: long-video event recovery** (4B). Qwen3.5-4B selects an earlier wrapping stage. Global Grounding identifies the correct episode but remains over-wide, while TROVE tightens the prediction around placing the teddy bear in the bag and sealing it.

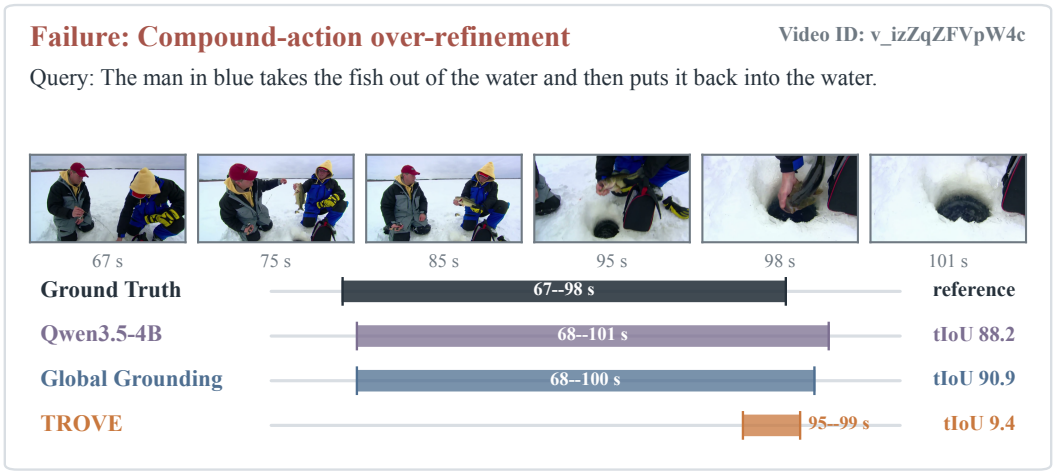


Figure 10: **Failure: compound-action over-refinement** (4B). Qwen3.5-4B and Global Grounding cover both taking the fish out and putting it back. The two constituents are separated in time, and the visually salient return to the water occurs near the end; TROVE retains only this later constituent rather than the full compound event.



Figure 11: **Failure: short-event boundary sensitivity** (4B). Global Grounding matches the two-second belly-pat annotation. The displayed half-second frames show a gradual transition into the motion before the annotated onset, making TROVE's one-second-earlier boundary visually plausible; because the reference lasts only two seconds, the small shift produces a pronounced tIoU decrease.

H Limitations and Future Directions

The current formulation uses a fixed refinement depth and nominal sampling rates. Figures 10 and 11 illustrate two challenges: refinement can retain only one constituent of a compound event, while a small boundary shift can substantially reduce overlap for a short event. Future work could check coverage of queried sub-events during refinement and use hierarchical re-observation to resolve boundaries at different temporal scales. Learning when to continue or stop based on expected localization improvement could allocate additional observations within a cumulative frame budget.

I Reproducibility Statement

The base models, training data, prompts, observation settings, optimization hyperparameters, and evaluation protocols are documented in Section 4.1 and Appendices C–E. Appendix B specifies the complete training and inference algorithms, including observation construction, action-wise optimization, and validity handling.

J Ethics Statement

The study uses established research benchmarks and introduces no new human-subject data. Temporal localization can support retrieval and accessibility but can also enable surveillance or sensitive-event search. Use of the datasets and source videos remains subject to their licenses and applicable privacy requirements.