VERIFYGROUND: LEARNING VERIFICATION-GUIDED EVIDENCE ACQUISITION FOR VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video temporal grounding requires visual evidence that identifies the queried event and resolves its temporal boundaries. Many existing multimodal large language model (MLLM)-based approaches predict temporal intervals from fixed global observations. Although these observations provide broad temporal coverage, they may be insufficient for accurate event retrieval and boundary refinement. We introduce VerifyGround, an interactive temporal grounding framework in which verification guides subsequent observation. The model first proposes a candidate interval, then inspects the corresponding video segment to verify its event match and boundary support. It acquires additional visual evidence from alternative temporal regions when the event hypothesis lacks support and gathers denser boundary evidence when the event is supported but its temporal extent remains uncertain. Verification thus guides prediction correction through targeted evidence acquisition, determining where and how to observe next. To learn this policy, we construct GroundStep-30K for cold-start supervised fine-tuning and introduce process-aware reinforcement learning to address the difficulty of assigning credit to intermediate corrections from final outcomes alone. Training combines the final localization objective with action-specific progress supervision, crediting each decision turn for the improvement produced by its corrective action. VerifyGround achieves state-of-the-art temporal grounding performance on Charades, ActivityNet, and QVHighlights, with a 3.4 points mIoU improvement over TimeLens-8B on QVHighlights. It also improves mIoU by 1.32 points over the same baseline on TACoS under cross-dataset zero-shot evaluation.

1 INTRODUCTION

Video Temporal Grounding (VTG) aims to determine the start and end timestamps of a target event within a video given a natural language query (Gao et al., 2017), serving as a crucial foundation for video retrieval, video-evidenced question answering, and video editing (Pramanick et al., 2025; Liu et al., 2026b). Although Multimodal Large Language Models (MLLMs) have demonstrated robust capabilities in video semantic understanding (OpenAI, 2025; Gemini Team, 2025), precise temporal grounding remains constrained by insufficient visual observations. Specifically, the target event may occupy only a small fraction of the video and be interleaved with other semantically similar activities. Furthermore, even when the correct event is identified, its onset and offset often manifest as gradual state transitions, necessitating finer-grained local observations for accurate delineation. Consequently, localization errors primarily stem from the inadequate observation of key events or boundary evidence. Therefore, a critical challenge in achieving accurate temporal grounding lies in how to dynamically supplement the requisite evidence based on the current grounding state.

Existing methods improve grounding through better temporal representations and evidence organization. TimeLens (Zhang et al., 2026) predicts temporal intervals from pre-sampled visual observations, while F2G (Zheng et al., 2026b) introduces an explicit evidence pool to support event selection and boundary estimation through an Identify-then-Measure formulation. However, both operate on visual evidence acquired before prediction, without obtaining additional observations in response to verification of the current hypothesis. Global sampling provides broad temporal coverage, but does not guarantee sufficient evidence for a specific localization decision: distinguishing similar 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.

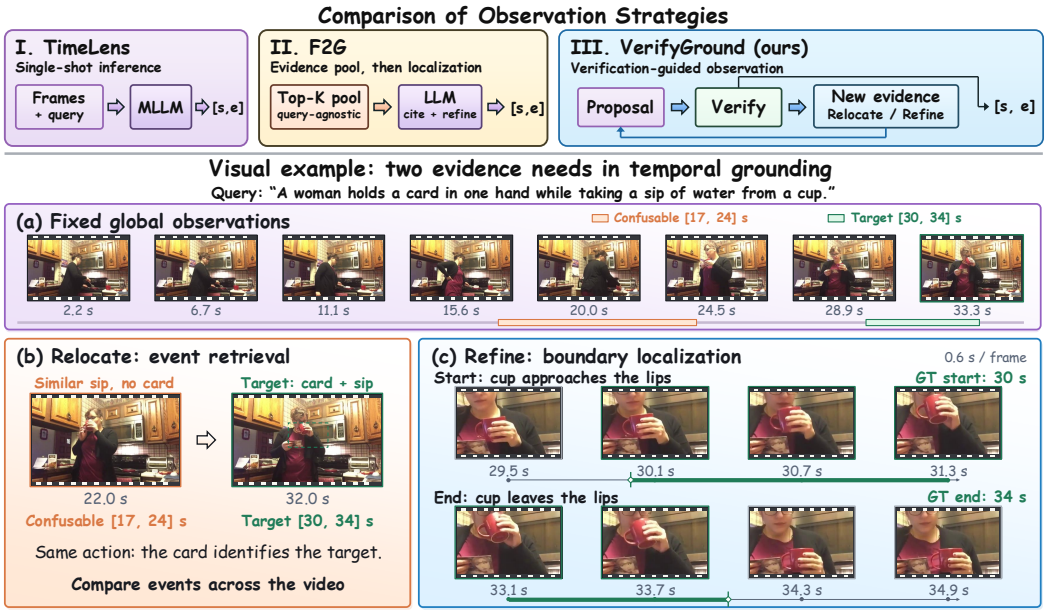

Figure 1: **Observation strategies and evidence needs.** Top: pre-sampled observations (TimeLens), a pre-built evidence pool (F2G), and verification-guided evidence acquisition (VerifyGround). Bottom: (a) global observations; (b) event disambiguation; (c) boundary refinement. Ground-truth intervals are shown for reference.

may require local action details, while precise boundaries may fall between sampled frames. When existing observations cannot resolve event ambiguity or boundary uncertainty, subsequent reasoning and boundary adjustment remain constrained by the same evidence. The challenge is therefore not only how to use available evidence, but also how to determine what and where to observe next based on the current localization hypothesis.

Interactive video models have begun to support localization refinement through candidate verification and local re-observation (Liu et al., 2026b; Zhang et al., 2025). Building on this direction, we introduce **VerifyGround**, a verification-guided framework for VTG. The model first proposes an initial interval from global observations, then extracts the corresponding segment to verify its event hypothesis. If the hypothesis is unsupported, `Relocate` uses the global context and interaction history to propose a new interval and acquires observations from that region for verification. If the event is supported but its boundaries remain uncertain, `Refine` acquires denser observations around the predicted start and end times to update the interval. The model terminates with `Answer` when both the event and its boundaries are sufficiently supported. Verification thus not only assesses the current prediction, but also determines the region and manner of subsequent observation, closing the loop between verification, targeted evidence acquisition, and localization updates.

To learn this policy, we first construct a dataset GroundStep-30K for supervised fine-tuning. Using initial predictions from the base model and teacher verification, we construct trajectories covering direct answers, relocation, boundary refinement, and mixed corrections. We then introduce process-aware reinforcement learning to address the difficulty of assigning credit to individual corrective actions from final outcomes alone. Progress rewards measure the change in localization before and after each action: `Relocate` is rewarded for improved proximity between the proposal and ground-truth centers, whereas `Refine` is rewarded for reduced start- and end-boundary errors. Each progress reward is assigned only to the decision turn that executes the corresponding action and is combined with the final localization reward for policy optimization. Center proximity also provides a positional learning signal even when the predicted and ground-truth intervals do not overlap. Thus, verification guides subsequent observation, the action type determines how progress is evaluated, and the resulting progress supervises the corresponding decision turn.

Our contributions are threefold:

- We address the mismatch between fixed visual observations and the changing evidence needs of temporal grounding through verification-guided evidence acquisition. By translating hypothesis verification into relocation or boundary refinement, our framework acquires targeted observations of alternative candidate regions or event boundaries.
- To address the difficulty of credit assignment, we design an action-specific, progress learning mechanism. We initialize the policy with GroundStep-30K and attribute retrieval and boundary refinement progress to the corresponding decision turns, providing targeted supervision for intermediate corrections while retaining the final localization objective.
- We demonstrate improved grounding accuracy on Charades, ActivityNet, and QVHighlights. Zero-shot evaluation on TACoS supports cross-dataset generalization, while action ablations further support the effectiveness of targeted observation and process supervision.

2 Related Work

2.1 MLLM-based Video Temporal Grounding

MLLM-based video temporal grounding has advanced along two complementary directions: improving temporal representations and reorganizing the localization process. VTimeLLM develops boundary awareness through instruction tuning (Huang et al., 2023), while TimeChat and Grounded-VideoLLM strengthen visual–temporal alignment through timestamp-aware visual representations and discrete timestamp tokens, respectively (Ren et al., 2023; Wang et al., 2024a). TimePLE further models a joint distribution over valid temporal intervals rather than predicting endpoints independently (Zeng et al., 2026). As a strong baseline, TimeLens combines high-quality data, interleaved timestamp encoding, and thinking-free reinforcement learning to improve grounding performance (Zhang et al., 2026). However, its predictions still rely on pre-sampled visual inputs, which may omit evidence needed to disambiguate similar events or resolve precise boundaries.

Beyond representation learning, HawkEye recursively narrows the search range, whereas TimeRefine iteratively corrects timestamps within an autoregressive response (Wang et al., 2024c;b). VideoMind, VITAL, and DiaVTG introduce local verification, tool-mediated observation, or multi-turn correction (Liu et al., 2026b; Zhang et al., 2025; Huang et al., 2026), moving grounding toward interactive evidence acquisition. In particular, VideoTemp-o3 jointly models temporal grounding and video question answering, using on-demand cropping, dense sampling, and multi-turn reflection to acquire new evidence and correct localization (Liu et al., 2026a). F2G separates event identification from boundary measurement through a learned video-wide candidate evidence pool and explicit evidence citation (Zheng et al., 2026b); however, its candidate-conditioned decoding operates on pre-encoded visual features rather than verification-driven new observations. In contrast to F2G's organization of existing evidence and VideoTemp-o3's unified clip-cropping interaction, VerifyGround uses verification feedback to distinguish the evidence needs of event retrieval and boundary refinement, adapting both where and how to observe next.

2.2 Post-Training and Process Supervision for Video Grounding

Post-training methods improve temporal grounding through task-specific rewards and teacher supervision. Time-R1 applies reinforcement learning with verifiable rewards to VTG (Wang et al., 2025), while VideoChat-R1 studies rule-based rewards for spatio-temporal perception (Li et al., 2025a). TimeLens explores thinking-free RLVR recipes (Zhang et al., 2026). MUSEG combines multi-segment matching with timestamp-consistency supervision (Luo et al., 2025), and TimeLens2 introduces temporal Wasserstein rewards for interval-set prediction (Zhu et al., 2026). Beyond reward design, Video-OPD provides dense token-level teacher feedback on student-generated trajectories through on-policy distillation (Li et al., 2026). These approaches improve supervision for grounding, while a complementary question for interactive policies is how to evaluate localization progress and attribute it to individual corrective decisions.

Related methods supervise reasoning, evidence, or intermediate decisions. TaRO (Zheng et al., 2026a) rewards reasoning sensitivity to frame-order perturbations near event boundaries, while SER (Xia et al., 2026) evaluates the semantic relevance and localization quality of explicit spatio-temporal evidence. Video-MTR (Xie et al., 2025) rewards marginal improvements in turn-level

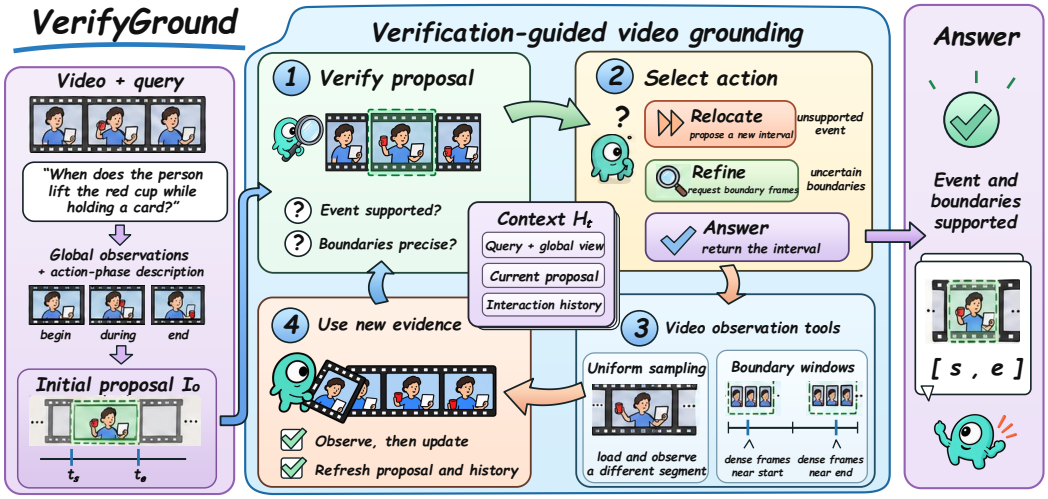

Figure 2: **VerifyGround overview.** Starting from an action-phase-guided proposal, the policy verifies visual support and selects `Relocate`, `Refine`, or `Answer`. New segment or boundary observations update the proposal and interaction history until the policy returns a final interval.

frame retrieval for video question answering, providing a related precedent for progress-based supervision. More generally, GiGPO (Feng et al., 2025) combines episode- and step-level relative advantages for multi-turn agent training. To optimize iterative correction, VerifyGround combines the final IoU reward with action-specific progress in proposal-center proximity and boundary accuracy, measured before and after each correction and assigned to the corresponding decision turn.

3 Method

We introduce VerifyGround, which uses proposal verification to guide adaptive video observation for event retrieval and boundary localization. We first describe the interactive grounding framework (Section 3.1), then present the construction of supervised trajectories (Section 3.2), and finally introduce process rewards for retrieval correction and boundary refinement (Section 3.3).

3.1 Verification-Guided Interactive Grounding

Given a video V and a natural-language query q, VTG aims to predict the temporal interval corresponding to the query. Conditioned on global video observations O^g, the model first decomposes the query into three action phases, `begin`, `during`, and `end`, and generates their descriptions D_0. It then predicts an initial proposal I_0, formulated as $D_0 \sim \pi_\theta(\cdot \mid O^g, q), \quad I_0 \sim \pi_\theta(\cdot \mid O^g, q, D_0)$.

This structured representation encourages event-centric reasoning grounded in the video, providing fine-grained action cues for initial localization rather than directly mapping overall semantics to timestamps. The system then returns the proposed segment for verification. At turn t, let H_t denote the context comprising the query, global observations, current proposal, and interaction history. The model selects an action

$$a_t \sim \pi_\theta(\cdot \mid H_t), \quad a_t \in \{\texttt{Relocate}, \texttt{Refine}, \texttt{Answer}\}. \tag{1}$$

For a non-terminal action, the proposal and context are updated as $(I_{t+1}, H_{t+1}) = \mathcal{T}_\theta(H_t, V, a_t)$, where $\mathcal{T}_\theta$ denotes the action-dependent proposal update and visual observation process. If the current segment does not support the query, the model executes `Relocate`: it proposes a new interval grounded in the global context and validates it by supplementing with new evidence. If the target event is supported but its boundaries remain imprecise, the model executes `Refine`, acquiring temporally dense observations near the predicted boundaries to adjust the interval. Once both the event and its boundaries are sufficiently supported, the model executes `Answer` and returns the current interval. Interaction continues until the model answers or reaches the turn limit. By using proposal

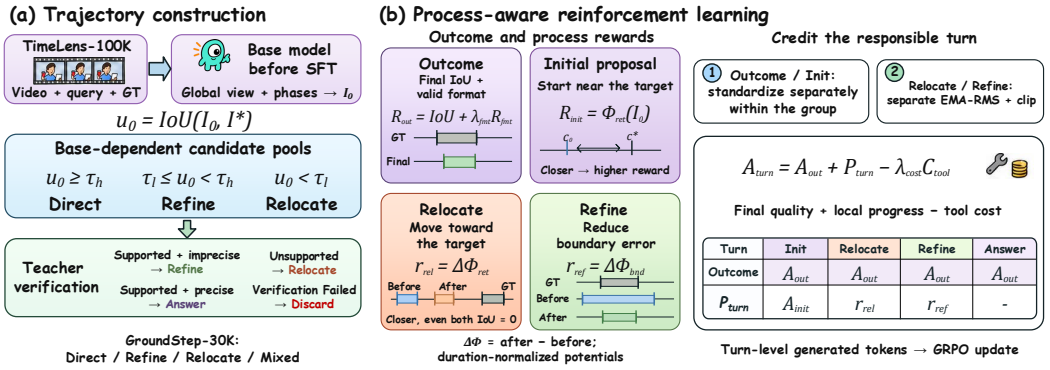

Figure 3: **Training pipeline.** (a) Base-model predictions and teacher verification produce GroundStep-30K trajectories. (b) GRPO combines outcome supervision with turn-local initial-proposal and corrective signals, minus tool costs. `Relocate` rewards gains in center proximity; `Refine` rewards gains in boundary accuracy.

verification to distinguish retrieval errors from boundary errors, VerifyGround allocates different visual evidence to the two types of correction and supports targeted supervision and optimization.

3.2 Supervised Trajectory Construction

As shown in Fig. 3(a), we combine IoU-based stratification with teacher verification to construct GroundStep-30K, a cold-start dataset of 30K trajectories from TimeLens-100K, mitigating potential annotation noise from its automated pipeline. For each example (V, q, I^*), the base model before fine-tuning generates action-phase descriptions and an initial proposal I_0 from the query and global observations. We compute $u_0 = \mathrm{IoU}(I_0, I^*)$, where I^* is the annotated interval. Given $\tau_l < \tau_h$, examples satisfying $u_0 \geq \tau_h$, $\tau_l \leq u_0 < \tau_h$, and $u_0 < \tau_l$ enter the Direct, Refine, and Relocate candidate pools, respectively.

We use a teacher model to filter candidates according to the correctness of the proposed event and the accuracy of its boundaries. `Direct` trajectories are retained only when both the event and its boundaries are correct. `Refine` trajectories are retained only when the target event is correctly identified, with the teacher then guided to construct boundary-refinement steps. Candidates failing their respective checks are discarded. For `Relocate` candidates, the teacher is directly guided to perform relocation. Subsequent verification and correction produce trajectories containing a single corrective action type or a mixture of both; persistent relocation failures trigger annotation-assisted reflection and recovery. GroundStep-30K thus covers correct termination, boundary refinement, and retrieval correction. Detailed filtering and recovery procedures are provided in Appendix A.

3.3 Process-Aware Reinforcement Learning

After supervised fine-tuning, we further optimize the grounding policy through reinforcement learning. For each input, we sample G multi-turn trajectories from the current policy, recording the initial proposal, the intervals before and after each action, and the corresponding decision turns. Let I_i^F denote the final prediction of trajectory i. Its outcome reward is

$$R_i^{\mathrm{out}} = \mathrm{IoU}(I_i^F, I^*) + \lambda_{\mathrm{fmt}} R_i^{\mathrm{fmt}}, \tag{2}$$

where R_i^{fmt} measures adherence to the required formats for tool calls and final answers. Following group relative policy optimization (GRPO), we normalize rewards within the trajectory group for the same input to obtain the outcome advantage A_i^{out}. Relying solely on this trajectory-level signal has two limitations. First, broadcasting the outcome reward across the trajectory does not distinguish whether an individual turn improves localization, limiting credit assignment to intermediate actions. Second, for difficult examples, all final predictions in a group may fail to overlap the ground truth and receive zero IoU, leaving no relative localization signal from the IoU component. We therefore introduce process rewards based on retrieval and boundary progress, measuring whether predictions move toward the target interval and providing continuous feedback for effective intermediate steps.

Retrieval progress. Standard IoU assigns zero to all non-overlapping proposals and therefore cannot distinguish their proximity to the target. We use proposal-center proximity as a proxy for coarse-grained retrieval progress and define the retrieval potential as

$$\Phi_{\mathrm{ret}}(I) = \exp\left(-\frac{|c - c^*|}{\max(d^*, d_{\min})}\right), \tag{3}$$

where c and c^* are the centers of the predicted and ground-truth intervals, respectively, $d^* = e^* - s^*$, and $d_{\min} > 0$ is a lower bound on the normalization duration. This function provides a continuous proximity signal even for non-overlapping proposals. The initial-proposal reward is

$$R_i^{\mathrm{init}} = \Phi_{\mathrm{ret}}(I_{i,0}). \tag{4}$$

It is standardized independently within the trajectory group to obtain A_i^{init}, which is assigned only to the turn that generates the initial proposal. Let $I_{i,k}^-$ and $I_{i,k}^+$ denote the intervals before and after the k-th action of the corresponding type. The `Relocate` reward is

$$r_{i,k}^{\mathrm{rel}} = \Phi_{\mathrm{ret}}(I_{i,k}^+) - \Phi_{\mathrm{ret}}(I_{i,k}^-). \tag{5}$$

Boundary progress. For `Refine`, we normalize the start- and end-time errors by the target duration and define the boundary potential as

$$E_{\mathrm{bnd}}(I, I^*) = \frac{|s - s^*| + |e - e^*|}{2\max(d^*, d_{\min})}, \quad \Phi_{\mathrm{bnd}}(I) = \exp\left(-\frac{E_{\mathrm{bnd}}(I, I^*)}{\tau_b}\right), \tag{6}$$

where $I = [s, e]$ and τ_b controls the scale of the boundary potential. The corresponding reward is

$$r_{i,k}^{\mathrm{ref}} = \Phi_{\mathrm{bnd}}(I_{i,k}^+) - \Phi_{\mathrm{bnd}}(I_{i,k}^-). \tag{7}$$

Duration normalization measures boundary errors relative to event length, rather than evaluating short events solely through absolute errors. Both process rewards retain their signs.

Turn-level credit assignment. To accommodate scale differences between the two process rewards, we combine the group-relative outcome advantage with non-centered process rewards. For each $a \in \{\mathrm{rel}, \mathrm{ref}\}$, we maintain an exponential moving average m_a of squared rewards and compute

$$\widetilde{r}_{i,k}^{a} = \mathrm{clip}\left(\frac{r_{i,k}^{a}}{\sqrt{m_a + \epsilon}}, -\kappa, \kappa\right), \tag{8}$$

where $\epsilon > 0$ ensures numerical stability and $\kappa > 0$ is the clipping threshold. This normalization adjusts reward magnitude without changing its sign. We denote the local process advantage assigned to turn t by $P_{i,t}$, incorporating the initial-proposal signal or the corresponding `Relocate` or `Refine` reward. The total advantage for that turn is

$$A_{i,t} = A_i^{\mathrm{out}} + P_{i,t} - \lambda_{\mathrm{cost}} C_{i,t}^{\mathrm{tool}}, \tag{9}$$

where $C_{i,t}^{\mathrm{tool}}$ counts tool calls at turn t and λ_{cost} controls their cost. We assign $A_{i,t}$ to the model-generated tokens in the corresponding turn and use it for GRPO policy optimization. This provides local feedback for retrieval correction and boundary refinement while retaining the final localization objective, alleviating the coarse credit assignment of trajectory-level rewards.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate VerifyGround on Charades-TimeLens (Gao et al., 2017), ActivityNet-TimeLens (Krishna et al., 2017), QVHighlights-TimeLens (Lei et al., 2021), and TACoS (Regneri et al., 2013). The first three constitute TimeLens-Bench (Zhang et al., 2026) and contain 3,363, 4,500, and 1,541 queries, respectively. We use the revised TimeLens annotations rather than the original benchmark annotations, following a single-span grounding protocol. For TACoS, we use the test annotations distributed with VideoMind (Liu et al., 2026b), comprising 4,001 queries over 25 videos. TACoS is evaluated in a zero-shot setting, without using its videos for post-training.

Table 1: Results on the three TimeLens-Bench subsets. R_τ denotes R1@τ. All scores are percentages; higher is better. Best and second-best results are highlighted in **bold** and underline.

Model	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
	$R_{0.3}$	$R_{0.5}$	$R_{0.7}$	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-2.5-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 General-purpose VLMs												
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
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
Qwen3.5-9B	68.8	52.1	26.3	47.0	63.6	52.5	35.7	48.3	75.7	65.8	50.7	60.1
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
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
Open-source VTG-specific methods												
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
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
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	70.4	57.2	28.7	49.9	64.7	54.8	35.8	50.1	75.9	66.9	51.4	60.6
F2G	74.2	63.1	36.3	54.9	69.8	59.4	42.3	54.1	79.6	70.4	54.5	64.1
VerifyGround (Ours)	**77.5**	**64.8**	**37.1**	**56.4**	**70.5**	**62.1**	**45.3**	**56.3**	**83.2**	**74.9**	**59.8**	**68.9**

Baselines and metrics. We compare VerifyGround with three categories of methods: (i) proprietary models, including GPT-4o (OpenAI, 2024), GPT-5 (OpenAI, 2025), Gemini-2.5-Flash, and Gemini-2.5-Pro (Gemini Team, 2025); (ii) open-source general-purpose VLMs, such as Qwen3.5-9B (Qwen Team, 2026), MiMo-VL-7B (Xiaomi LLM-Core Team, 2025), VideoChat-Flash-7B (Li et al., 2024), and VideoChat-R1-7B (Li et al., 2025a); (iii) open-source VTG-specific methods, encompassing TRACE-uni (Guo et al., 2025), TimeSuite (Zeng et al., 2025), Grounded-VideoLLM (Wang et al., 2024a), Time-R1-7B (Wang et al., 2025), TimeLens-8B (Zhang et al., 2026), VideoTemp-o3 (Liu et al., 2026a), and F2G (Zheng et al., 2026b). We report Recall@1 at temporal intersection-over-union thresholds of 0.3, 0.5, and 0.7 (R1@0.3/0.5/0.7), together with mean IoU (mIoU). For the implementation details regarding training and hyperparameter settings, please refer to Appendix B.

4.2 Main results

Comparison with State-of-the-Art Methods. As shown in Table 1, VerifyGround achieves the best performance across all three TimeLens-Bench subsets, consistently outperforming concurrent VTG-specific methods. On ActivityNet, our method attains 56.3 mIoU, surpassing F2G and TimeLens-8B by up to 2.2 points and 3.1 points, respectively. On QVHighlights, VerifyGround reaches 68.9 mIoU, exceeding TimeLens-8B and F2G by 3.4 points and 4.8 points, and even approaching the proprietary model Gemini-2.5-Pro (mIoU: 68.9 vs. 70.4; R1@0.7: 59.8 vs. 61.1). These consistent gains are attributed to the verification-guided decoupling of event retrieval and boundary localization. This design enables VerifyGround to acquire targeted visual evidence for each error type, thereby satisfying their distinct evidence requirements within a unified framework.

As shown in Table 2, VerifyGround demonstrates strong out-of-distribution generalization on TACoS. Our method achieves 42.74 mIoU and outperforms previous SOTA methods. VerifyGround consistently improves across all metrics, achieving its largest gain of 3.34 points at R1@0.5, while also exceeding TimeLens-8B and TaRO by 1.25 and 7.44 points at the stricter R1@0.7 threshold.

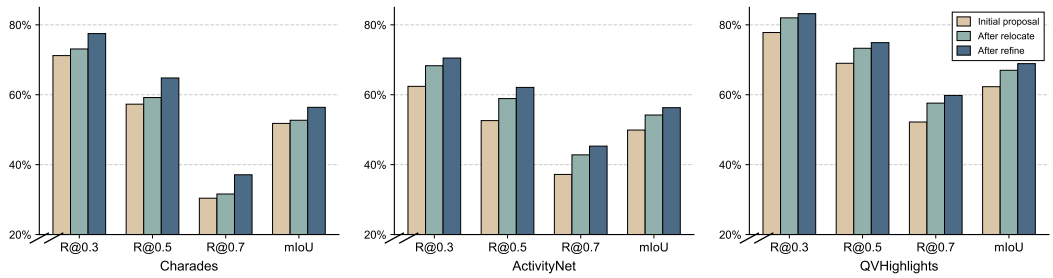

Figure 4: Stage-wise gains over the initial proposal. This demonstrates the complementary nature of the two operations, where Relocate primarily aids event retrieval in ambiguous contexts, and Refine significantly improves boundary precision for short-duration events.

This robust transferability is supported by the process-aware reinforcement learning paradigm: by supplementing the final IoU objective with action-specific progress signals, the learned correction policy generalizes effectively to videos with markedly different temporal scales and annotation distributions without task-specific re-tuning.

Stage-wise correction analysis. Figure 4 illustrates the stage-wise improvements from Relocate and Refine. On Charades, Refine contributes a larger mIoU gain than Relocate (+2.7 vs. +0.9). This pattern is consistent with the sensitivity of short action segments to localization errors, where even small deviations can substantially reduce IoU. On QVHighlights, Relocate yields larger gains across all four metrics (+3.7 mIoU compared with +2.9 from Refine). This is consistent with the benefit of explicitly comparing candidate events to distinguish semantically similar occurrences. Averaged equally across the three datasets, Relocate improves R@0.3 by 4.0 points, compared with 2.6 points from Refine, while Refine achieves its largest average recall gains at the stricter thresholds of R@0.5 and R@0.7 (+3.5 and +3.2). Together, these results support the complementary roles of the two operations and the effectiveness of providing targeted observations for event retrieval and boundary refinement. Additional qualitative case studies are provided in the appendix C.

Table 2: Zero-shot results on TACoS.

Model	$R_{0.3}$	$R_{0.5}$	$R_{0.7}$	mIoU
VideoChat-R1.5-7B (2025)	28.60	16.20	6.17	-
Time-R1-7B (2025)	36.80	21.50	7.90	-
ED-VTG (2025)	14.50	6.00	2.30	12.70
UniTime-Zero (2025b)	50.06	31.54	-	33.38
HiTeA (2026)	29.08	16.10	-	19.79
VideoMind (2026b)	49.50	36.20	21.40	34.40
TaRO (2026a)	52.30	36.50	22.50	-
TimeLens-8B (2026)	56.80	43.52	28.69	41.42
Qwen3-VL-8B (2025)	38.60	30.50	20.40	-
VerifyGround (Ours)	**58.59**	**46.86**	**29.94**	**42.74**

4.3 Ablation Studies

We organize the ablations around three questions: which training signals improve grounding, which corrective actions are necessary, and how additional interaction affects accuracy and cost. The metrics reported in the experiments are all averages across three datasets.

Post-training and action-specific credit assignment. Table 3 separates the effects of outcome optimization and intermediate supervision. Starting from the same SFT checkpoint, we compare outcome-only GRPO using the final IoU and format reward in Eq. 2 with the full objective in Eq. 9, and individually remove the initial-proposal, Relocate, and Refine rewards. Outcome-only GRPO improves mIoU from 52.7 to 57.8 and R1@0.5 from 57.3 to 65.1, demonstrating the benefit of optimizing final localization quality beyond SFT. Adding action-specific supervision further raises these scores to 60.6 and 67.2. Removing the initial-proposal, Relocate, or Refine reward reduces mIoU by 1.6, 1.9, and 1.1 points, respectively, with corresponding declines in R1@0.5. Thus, each

signal contributes to the full model's performance: the initial reward encourages proposals near the target, while the corrective rewards guide retrieval progress and boundary improvement.

Table 3: Post-training and reward ablations.

Training configuration	mIoU ↑	$R_{0.5}$ ↑
SFT only	52.7	57.3
Outcome-only GRPO	57.8	65.1
Full	**60.6**	**67.2**
Full w/o initial reward	59.0	66.3
Full w/o Relocate reward	58.7	64.9
Full w/o Refine reward	59.5	66.6

Table 4: Action-space ablations. One-pass retains event decomposition but performs no tool interaction.

Available actions	mIoU ↑	$R_{0.5}$ ↑
One-pass	57.2	63.8
Without Relocate	58.5	64.7
Without Refine	58.9	65.2
Full	**60.6**	**67.2**

Complementary roles of Relocate and Refine. Table 4 evaluates the corrective actions under matched total visual-token budgets, with each variant retrained through SFT and RL from the same pretrained backbone. The one-pass baseline retains event decomposition but uses its entire visual budget for a single global observation. The full policy outperforms this baseline by 3.4 points in both mIoU and R1@0.5, supporting the value of adaptive observation rather than simply more visual input. Removing Relocate reduces mIoU and R1@0.5 by 2.1 and 2.5 points, respectively, while removing Refine causes drops of 1.7 and 2.0 points. These degradations are consistent with their complementary roles: Relocate revises unsupported event hypotheses by inspecting alternative segments, whereas Refine acquires denser boundary observations for an already identified event. Neither action alone recovers the full performance, supporting the explicit separation of retrieval correction and boundary refinement according to their distinct evidence requirements.

Visual budget allocation and interaction efficiency. We match TimeLens's 14k visual-token budget (14,336 tokens), distributing it between the initial global observation and subsequent tool observations rather than adding visual input. Each tool observation has a 1,568-token cap, using 1 fps (up to 8 frames) for verification or 2 fps (up to 16 frames) for refinement. We vary the inference-time turn limit (Table 5), reallocating budgets from removed tool observations to the initial input. Despite reducing the initial allocation, moving from three to four turns improves mIoU/$R_{0.5}$ by 2.3/1.3 points. The four-turn policy also exceeds the separately trained, budget-matched one-pass baseline (57.2/63.8) by 3.4 points on both metrics. Together, these results support allocating a limited visual budget across multi-turn, on-demand observations rather than spending it entirely upfront. A fifth turn yields only 0.3/0.1 additional points while increasing mean interaction length from 3.32 to 3.48 turns, favoring four turns as the default accuracy–efficiency trade-off.

Table 5: Turn budgets at 14,336 tokens.

Max. turns	3	**4**	5
Initial tokens	11,200	9,632	8,064
Tool tokens	$1{,}568 \times 2$	$1{,}568 \times 3$	$1{,}568 \times 4$
mIoU ↑	58.3	60.6	60.9
$R_{0.5}$ ↑	65.9	67.2	67.3
Mean turns	2.91	3.32	3.48

5 Conclusion

We presented VerifyGround, which couples verification-guided evidence acquisition with action-specific, turn-level supervision for video temporal grounding. Rather than using verification solely to assess an existing prediction, VerifyGround uses it to guide subsequent observation: inspecting alternative regions when an event hypothesis is unsupported and acquiring denser boundary evidence when a supported event remains imprecisely localized. GroundStep-30K and process-aware reinforcement learning support learning this policy through correction demonstrations and progress rewards assigned to the corresponding decision turns. Results on TimeLens-Bench, zero-shot evaluation on TACoS, and controlled ablations support the effectiveness of this approach. Our findings highlight the value of learning not only to predict temporal intervals, but also to acquire evidence for unresolved localization decisions and attribute the resulting progress to individual corrective actions.

AI USE STATEMENT

Generative AI tools assisted with literature search and synthesis, refinement of methodological ideas and experimental designs, implementation and debugging, interpretation of experimental results, and manuscript drafting, translation, editing, and LaTeX formatting. We also used a vision-language teacher to generate and verify synthetic training trajectories for GroundStep-30K, as described in our trajectory-construction procedure. The authors are responsible for checking AI-assisted text, code, analyses, and citations and take full responsibility for the final content, claims, and artifacts of this work.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide comprehensive implementation details in Appendix B, including training configurations, hyperparameter settings, and process-reward implementation specifics for both supervised fine-tuning and reinforcement learning stages. Furthermore, we will publicly release the GroundStep-30K dataset used for cold-start SFT, as well as the curated data splits and sampling protocols used for RL training, upon acceptance. These resources, together with the methodological descriptions in the main paper and appendix, are intended to enable full replication of our results and facilitate future research on verification-guided temporal grounding.

REFERENCES

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-Group Policy Optimization for LLM Agent Training, 2025. URL https://arxiv.org/abs/2505.10978.

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 (ICCV)*, pp. 5267–5275, 2017. URL https://openaccess.thecvf.com/content_ICCV_2017/papers/Gao_TALL_Temporal_Activity_ICCV_2017_paper.pdf.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261v1.

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 (ICLR)*, 2025. URL https://openreview.net/forum?id=14fFV0chUS.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to Grasp Video Moments, 2023. URL https://arxiv.org/abs/2311.18445.

Hongyu Huang, Junyi Yang, Sipeng Yang, and Xiaogang Jin. DiaVTG: multi-turn reasoning framework for video temporal grounding. *The Visual Computer*, 42:474, 2026. doi: 10.1007/s00371-026-04645-1. URL https://link.springer.com/article/10.1007/s00371-026-04645-1.

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 (ICCV)*, pp. 706–715, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Krishna_Dense-Captioning_Events_in_ICCV_2017_paper.html.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11846–11858. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/62e0973455fd26eb03e91d5741a4a3bb-Paper.pdf.

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, 2026. URL https://arxiv.org/abs/2602.02994.

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, 2024. URL https://arxiv.org/abs/2501.00574v1.

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, 2025a. URL https://arxiv.org/abs/2504.06958.

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 *Advances in Neural Information Processing Systems*, volume 38, pp. 64426–64455. Curran Associates, Inc., 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5d2e24df9cfaad3189833b819c40b392-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, 2026a. URL https://arxiv.org/abs/2602.07801.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. VideoMind: A Chain-of-LoRA agent for temporal-grounded video reasoning. In *International Conference on Learning Representations (ICLR)*, 2026b. URL https://arxiv.org/abs/2503.13444.

Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. MUSEG: Reinforcing Video Temporal Understanding via Timestamp-Aware Multi-Segment Grounding, 2025. URL https://arxiv.org/abs/2505.20715.

OpenAI. GPT-4o system card. System card, OpenAI, August 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-5 system card. System card, OpenAI, August 2025. URL https://openai.com/index/gpt-5-system-card/.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24297–24308, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Pramanick_Enrich_and_Detect_Video_Temporal_Grounding_with_Multimodal_LLMs_ICCV_2025_paper.html.

Qwen Team. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Qwen Team. Qwen3.5-9B. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B.

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. doi: 10.1162/tacl_a_00207. URL https://aclanthology.org/Q13-1003/.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding, 2023. URL https://arxiv.org/abs/2312.02051.

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, 2024a. URL https://arxiv.org/abs/2410.03290.

Xizi Wang, Feng Cheng, Ziyang Wang, Huiyu Wang, Md Mohaiminul Islam, Lorenzo Torresani, Mohit Bansal, Gedas Bertasius, and David Crandall. TimeRefine: Temporal Grounding with Time Refining Video LLM, 2024b. URL https://arxiv.org/abs/2412.09601.

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, 2025. URL https://arxiv.org/abs/2503.13377.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. HawkEye: Training Video-Text LLMs for Grounding Text in Videos, 2024c. URL https://arxiv.org/abs/2403.10228.

Sheng Xia, Zhengqin Lai, Tianxiang Jiang, Kanghui Tian, Shoujun Zhou, Bin Li, and Yi Wang. SER: Learning to Ground Video Reasoning with Semantic Evidence Rewards, 2026. URL https://arxiv.org/abs/2606.24726.

Xiaomi LLM-Core Team. MiMo-VL technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-MTR: Reinforced Multi-Turn Reasoning for Long Video Understanding, 2025. URL https://arxiv.org/abs/2508.20478.

Xinyi Xu, Hongsong Wang, Guo-Sen Xie, Caifeng Shan, and Fang Zhao. HiTeA: Hierarchical temporal alignment for training-free long-video temporal grounding. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=vIecIscDJf.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. VideoChat-R1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception, 2025. URL https://arxiv.org/abs/2509.21100.

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 (ICLR)*, 2025. URL https://openreview.net/forum?id=nAVejJURqZ.

Yuhui Zeng, Xinyu Mao, Xiaokun Liu, Xin Tao, Jinfa Huang, Jiayi Ji, and Xiawu Zheng. TimePLE: Rethinking Temporal Representation for Video Temporal Grounding, 2026. URL https://arxiv.org/abs/2607.23951.

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*, 2025. doi: 10.48550/arXiv.2508.04416. URL https://arxiv.org/abs/2508.04416.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2512.14698.

Minghang Zheng, Zihao Yin, Yi Yang, Yuxin Peng, and Yang Liu. Temporal-Aware Reasoning Optimization for Video Temporal Grounding, 2026a. URL https://arxiv.org/abs/2606.09248.

Zelin Zheng, Xinyan Liu, Ruixin Li, Antoni B. Chan, Guorong Li, Qingming Huang, and Laiyun Qing. Foresee-to-Ground: From predictive temporal perception to evidence-driven reasoning for video temporal grounding. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026b. URL https://arxiv.org/abs/2605.21973.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. TimeLens2: Generalist Video Temporal Grounding with Multimodal LLMs, 2026. URL https://arxiv.org/abs/2607.17423.

A Supervised Trajectory Construction Details

Teacher inputs and sample filtering. The teacher receives the query, timestamped global observations, action-phase descriptions, the current proposal, and the interaction history, with the same tool observations available to the student at each step. Ground-truth intervals are withheld except during the recovery procedure below. After IoU-based stratification, Direct candidates must pass verification of both target-event presence and start/end boundary accuracy before yielding direct answers. Refine candidates require only target-event verification, followed by refinement using dense boundary observations. Candidates from either pool that fail their respective checks are discarded rather than reassigned. Relocate candidates do not require the current interval to contain the target event; the teacher is directly guided to relocate.

Failure recovery and privileged-hint removal. If two consecutive `Relocate` actions both yield proposals with IoU below τ_l, we provide the annotated interval to the teacher on the third relocation attempt, prompting it to reflect on previous errors and localize the target interval. This converts persistent failures into recovery demonstrations through offline annotation-assisted supervision. When constructing student SFT examples, we remove the ground-truth hint from the input and ensure that the reflection retains no direct references to it, preventing disclosure of the privileged hint in the student-visible context.

B Implementation Details

Training configuration. VerifyGround is initialized from Qwen3.5-9B, with Qwen3.6-27B serving as the teacher for constructing GroundStep-30K. We perform supervised fine-tuning on GroundStep-30K for one epoch with a learning rate of 1×10^{-5}. Following the difficulty-aware selection principle of Time-R1, we select 3,000 moderately difficult video–query pairs for subsequent reinforcement learning. We optimize the policy using GRPO with a learning rate of 1×10^{-6} and employ early stopping during reinforcement learning. For each input, we sample $G = 8$ multi-turn trajectories from the current policy and set the format-reward coefficient to $\lambda_{\text{fmt}} = 0.2$. The maximum interaction length is four turns during training and by default at inference; the inference-time turn-budget study varies this limit while keeping the trained checkpoint fixed.

Process-reward implementation details. All temporal quantities used in the reward functions are measured in seconds. We use a shared duration floor of $d_{\min} = 1.0\,\text{s}$ for the retrieval and boundary potentials and set the boundary temperature to $\tau_b = 0.5$. The initial-proposal rewards are standardized within each input's trajectory group using their own mean and population standard deviation, independently of the outcome rewards; we add 10^{-6} to the standard deviation for numerical stability, yielding zero initial-proposal advantages when all rewards in the group are identical. For each corrective action type $a \in \{\text{rel}, \text{ref}\}$, we maintain a separate exponential moving average m_a of squared progress rewards, initialized to 1 and updated with decay $\beta_{\text{EMA}} = 0.99$ without bias correction. Each statistic is updated once per rollout batch using the mean squared reward over all completed actions of the corresponding type across training workers; zero-progress actions are included, and the statistic remains unchanged if no such action occurs. We use the updated statistic to rescale the current batch and hold it fixed during subsequent optimizer passes over that batch. Specifically, each signed progress reward is divided by $\sqrt{m_a} + \epsilon$, with $\epsilon = 10^{-6}$, and clipped to $[-\kappa, \kappa]$, with $\kappa = 3$. No mean subtraction is applied to the corrective rewards, so their signs are preserved. The initial-proposal, Relocate, and Refine signals enter the local advantage $P_{i,t}$ with unit weights after their respective normalization procedures. In the combined turn-level advantage, we set the tool-cost coefficient to $\lambda_{\text{cost}} = 0.1$ per counted tool call.

C Case Study

Relocate: Cross-Segment Event Retrieval and Relocation. As illustrated in Figure 5 (a) and 5(b), videos frequently contain multiple segments that are semantically similar to the query. For instance, in (a), the video encompasses both "unpacking and trying on clothes" and "looking through the container." The initial proposal [37, 78]s incorrectly anchors on the trying-on phase. By identifying the semantic inconsistency between the observed actions and the query "looking through the container," the model actively triggers Relocate, successfully localizing the target event in an earlier interval

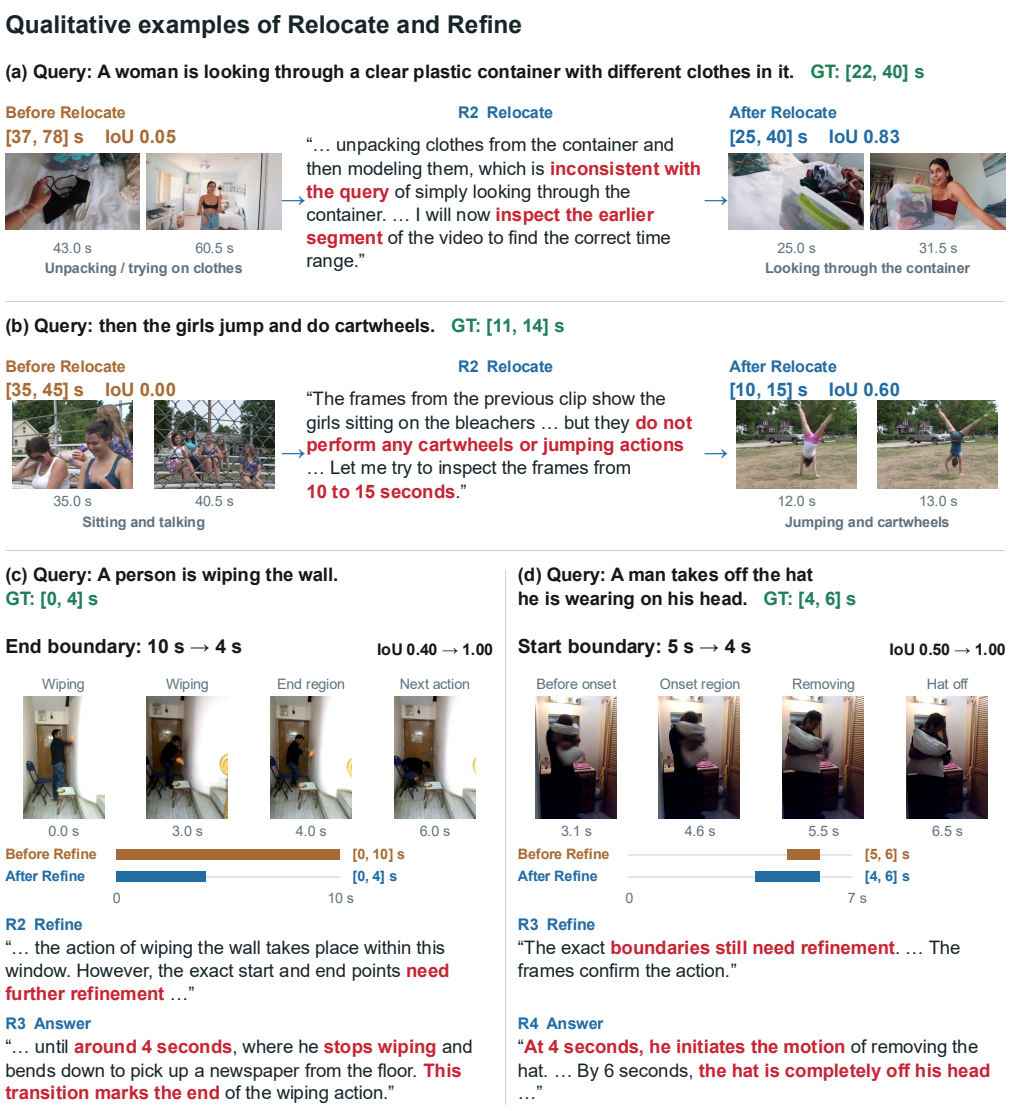

Figure 5: Qualitative examples of Relocate and Refine. (a)–(b) Relocate cases where the initial proposal anchors on a semantically similar but incorrect segment; the model identifies the event mismatch through verification and retrieves the correct temporal region, substantially improving IoU. (c)–(d) Refine cases where the target event is correctly identified but temporal boundaries remain imprecise; dense local observations near the predicted boundaries enable precise boundary correction to the ground truth. Ground-truth intervals are shown for reference.

[25, 40]s (improving IoU from 0.05 to 0.83). Similarly, in (b), the initial proposal falls on a segment where girls are sitting and talking. By verifying the absence of jumping or cartwheeling actions, the model rejects this hypothesis and retrieves the correct jumping segment at [10, 15]s (IoU from 0.00 to 0.60). These cases demonstrate that Relocate effectively leverages global context and interaction history to compare and disambiguate multiple candidate segments. However, we also observe two failure modes that Relocate struggles to address. First, when the target action is extremely brief (e.g., lasting less than 1–2 seconds), the uniform frame sampling during the initial global observation may completely miss the event, depriving the model of the visual cues necessary to form a valid relocation hypothesis. Second, when the query requires fine-grained action distinction (e.g., left-hand vs. right-hand operations) or multi-step causal reasoning, the inherent action understanding capabilities of current VLMs are insufficient to support reliable verification, highlighting the need for further advancements in foundational vision-language models.

Refine: Boundary Refinement and Short-Action Localization. As shown in Figure 5 (c) and (d), Refine excels in scenarios where the target event is correctly identified but its temporal boundaries remain ambiguous, or when the action duration is notably short. In (c), the model correctly identifies the "wiping the wall" action but initially predicts an overly delayed end time (10s vs. GT 4s). By acquiring dense frame observations near the predicted end boundary, the model captures the transition point where the person stops wiping and bends down to pick up a newspaper at around 4 seconds, correcting the IoU from 0.40 to 1.00. In (d), a similar start-boundary deviation (5s vs. GT 4s) is rectified by observing the precise onset frames of the hat removal motion. These cases validate the effectiveness of dense local observations in precisely delineating short actions or gradual boundaries. Nevertheless, a primary limitation of the current Refine mechanism is that the temporal window for dense sampling (i.e., the expansion range around the predicted boundaries) relies on manually preset, fixed hyperparameters (e.g., 2 fps, up to 16 frames). It lacks an adaptive adjustment mechanism conditioned on action type, video pacing, or initial prediction confidence. For extremely short actions, a fixed window might be excessively broad, introducing irrelevant frames; for slowly evolving actions, it might be insufficient to cover the complete transition interval. Developing context-aware mechanisms to dynamically determine sampling density and window size during inference remains a promising direction for future work.