ZERO- AND MULTI-EVENT TEMPORAL GROUNDING WITH MULTIMODAL LARGE LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Temporal Grounding (VTG) aims to localize video segments matching a text query. Conventional VTG methods assume that each query corresponds to exactly one segment, an assumption violated in real-world scenarios where events may occur zero, once, or multiple times. We address this gap by formalizing Any-Moment Temporal Grounding (AMTG), which requires models to predict an arbitrary number of temporal segments, including none. To facilitate research on AMTG, we construct AnyMoment dataset, which includes queries with zero, single, and multiple ground-truth segments. Recently, reinforcement learning (RL) has emerged as an effective paradigm for optimizing multimodal large language models (MLLMs) for VTG. However, AMTG poses two additional challenges: recognizing distinct occurrences of the queried event from complex video content, and providing fine-grained learning signals for individual event predictions, as conventional sequence-level RL objectives assign a single advantage to the entire response and may therefore penalize correct predictions together with incorrect ones. To address these challenges, we propose Event-wise Grounding Optimization (EGO), an RL-based framework for AMTG. First, EGO employs a bivariate difficulty-aware sampling strategy that jointly characterizes the difficulty of recognizing event occurrences and locating their temporal extents, enabling RL to focus on samples that are challenging yet learnable. Second, we propose an occurrence-aware chain-of-thought reasoning strategy that identifies fine-grained descriptions of individual event occurrences and transforms them into discriminative subqueries, facilitating the localization of multiple similar or closely related events. Third, we introduce fine-grained event-level credit assignment, which computes event-specific advantages within each rollout group and assigns them to the corresponding output tokens, explicitly reinforcing correct event predictions while penalizing erroneous ones. Experiments show that EGO achieves state-of-the-art performance on AnyMoment and public multi-segment VTG benchmarks, demonstrating its effectiveness across diverse temporal grounding scenarios.

1 INTRODUCTION

Video Temporal Grounding (VTG) aims to localize video segments corresponding to a natural language query. Conventional VTG benchmarks and methods (Gao et al., 2017; Krishna et al., 2017; Li et al., 2025b; Wang et al., 2025) typically assume that each query corresponds to *exactly one* target event. In real-world applications, however, a queried event may be absent or recur at different points in a video. As illustrated in Fig. 1(a), "The man drinks water" matches multiple distinct moments, whereas "The woman drinks water" has no matching moment. A model operating under the single-event assumption may miss valid occurrences or predict a segment for an absent event. In this work, we generalize VTG to support an arbitrary number of output events, including zero. We term this task **Any-Moment Temporal Grounding (AMTG)**.

A few recent efforts have extended VTG beyond single-event retrieval. QV-M2 (Cao et al., 2025) adds multi-event queries to QVHighlights (Lei et al., 2021), while MUSEG (Luo et al., 2025), OMTG (Xu et al., 2026) and TimeLens2 (Zhu et al., 2026) develop datasets and reinforcement learning (RL) methods for grounding multiple occurrences. However, **these datasets do not jointly support queries with zero, single, or multiple matching events.** They lack queries with zero matching events, which leaves models without the ability to reject non-matching results. Beyond data coverage,

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.

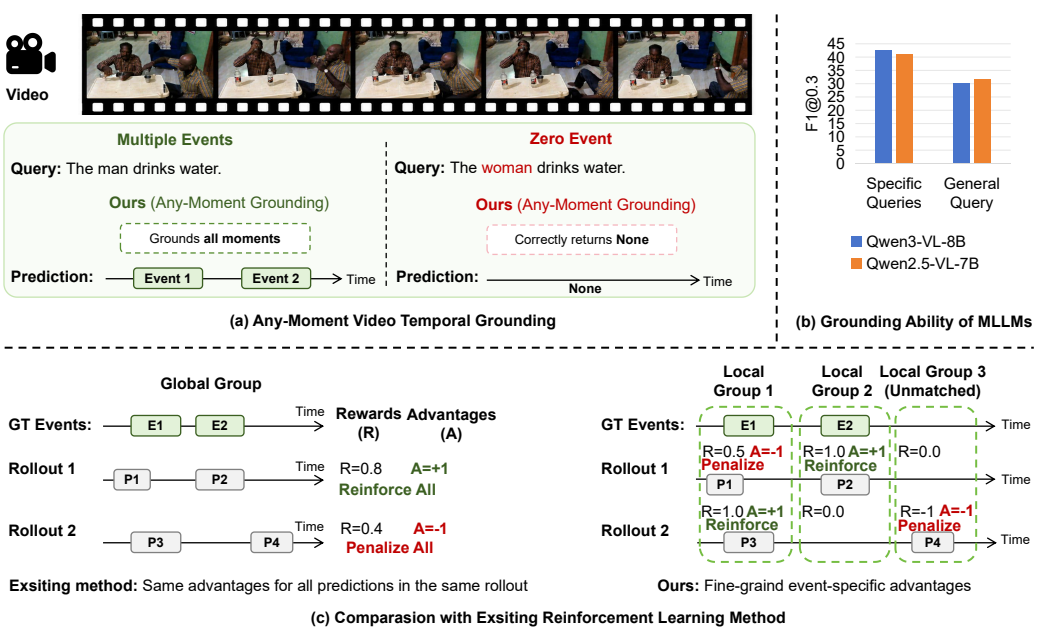

Figure 1: **(a)** Any-Moment Temporal Grounding supports an arbitrary number of output events, including zero. **(b)** MLLMs achieve higher grounding performance (F1@0.3) with occurrence-specific queries than with a general query covering multiple occurrences. **(c)** Existing reinforcement learning methods applies the same advantage to all predictions within a rollout, potentially reinforcing erroneous predictions (P1) or penalizing correct ones (P3). Our event-level credit assignment compares predictions within local groups associated with individual ground-truth events, with a separate group for unmatched predictions, providing event-specific learning signals.

existing RL objectives typically assign a shared advantage to all event predictions within a response, rather than providing fine-grained learning signals for individual event predictions. Specifically, MUSEG, OMTG, and TimeLens2 (Luo et al., 2025; Xu et al., 2026; Zhu et al., 2026) optimize multi-event grounding by aggregating prediction quality into a response-level reward, from which a shared advantage is assigned to the generated sequence. A single response, however, can contain both accurate and erroneous event predictions. This shared advantage cannot distinguish which individual predictions should be reinforced or penalized. The left side of Fig. 1(c) shows the consequence: P1 receives positive feedback despite inaccurate localization because its rollout scores well overall, while the accurate P3 is penalized because its rollout receives a lower reward.

To address zero-event grounding, **we construct the AnyMoment dataset, covering zero, single, and multiple event occurrences, and introduce a rejection reward that explicitly encourages the model to reject a query when its target event is absent.** AnyMoment is built with an automated caption-centric annotation pipeline for multi-event samples, while challenging zero-event queries are constructed by minimally modifying valid descriptions so that they remain semantically similar but no longer match the video. To further optimize multi-event grounding, we propose **Event-wise Grounding Optimization (EGO), an RL-based framework for AMTG that makes individual event occurrences explicit in both the reasoning process and the optimization signal.** First, AMTG requires the number of predicted events to be correct, and the time boundaries of each event to be precise. For some samples, it is difficult to identify all occurrences, while for others, it is difficult to determine their time boundaries accurately. Therefore, we propose *bivariate difficulty-aware sampling* to jointly characterize the difficulty of recognizing event occurrences and locating their temporal boundaries, selecting challenging yet learnable samples for RL. Secondly, we find that MLLMs achieve better localization with specific queries describing individual occurrences than with a general query covering them all, as shown in Fig. 1(b). Therefore, we propose *occurrence-aware chain-of-thought (CoT) reasoning* to identify the distinctive details of each matching occurrence and express them as a discriminative subquery before predicting its temporal boundaries. Details such as participant attributes, surrounding actions, and scene context help distinguish otherwise

similar occurrences. When no occurrence matches, the model explains the mismatch and returns an empty prediction. Finally, to distinguish the contributions of individual event predictions, we further introduce *event-level credit assignment* that compares predictions across rollouts within local groups associated with individual ground-truth events and treats unmatched predictions separately. It computes event-specific advantages and assigns them to the corresponding output tokens. As shown on the right of Fig. 1(c), this allows accurate and erroneous predictions within the same response to receive different learning signals, directly addressing the credit assignment problem of a shared sequence-level advantage. For a missed occurrence, its learning signal is instead assigned to the response-ending tokens, so stopping before all matching moments are reported can be penalized.

Our contributions are: (1) We formalize AMTG and construct the AnyMoment dataset encompassing queries with zero, single, and multiple ground-truth events, together with a rejection reward for learning to handle absent events. (2) We propose EGO, an RL-based framework combining bivariate difficulty-aware sampling, occurrence-aware CoT that generates discriminative subqueries for localization, and event-level credit assignment that provides fine-grained learning signals for individual predictions. (3) Extensive experiments demonstrate state-of-the-art performance on AnyMoment test set and public multi-event VTG benchmarks.

2 Related Work

Video Temporal Grounding Datasets Conventional benchmarks, including ActivityNet Captions (Krishna et al., 2017), Charades-STA (Gao et al., 2017), TACoS (Regneri et al., 2013), and DiDeMo (Anne Hendricks et al., 2017), primarily evaluate single-moment retrieval. QVHighlights (Lei et al., 2021) and QV-M2 (Cao et al., 2025) support multiple relevant moments but remain focused on Vlog and News videos. More recent efforts expand multi-segment supervision through action localization annotations in MUSEG (Luo et al., 2025) and automated video annotation in OMTG (Xu et al., 2026) and TimeLens2-93K (Zhu et al., 2026). Despite this progress, they still focus on queries with at least one matching segment. These datasets therefore provide limited support for learning when to reject a query. NExT-VMR (You et al., 2025) includes both multi-event and zero-event queries. However, its data is not publicly accessible, limiting reproducibility and community reuse; its reliance on manually annotated relation tuples also makes expansion to new videos labor-intensive. To bridge this gap, we introduce the AnyMoment dataset encompassing queries with zero, single, and multiple ground-truth events, together with a rejection reward for learning to handle absent events.

Single-Segment Video Temporal Grounding Traditional VTG methods localize a target moment through proposal-based matching (Gao et al., 2017; Xiao et al., 2021; Liu et al., 2021; Bhowmik et al., 2023) or direct boundary prediction (Yuan et al., 2019; Mun et al., 2020; Zhang et al., 2020; Pan et al., 2023; Mu et al., 2024). MLLM-based approaches such as VTimeLLM (Huang et al., 2024) and TimeChat (Ren et al., 2024) formulate grounding as timestamp generation, while Time-R1 (Wang et al., 2025), VideoTG-R1 (Dong et al., 2025), and VideoChat-R1.5 (Yan et al., 2025) improve localization through reinforcement learning with verifiable rewards. These approaches primarily operate under a single-event formulation, which does not account for rejecting absent events or retrieving all matching occurrences. We address these requirements through AMTG, allowing zero, single, or multiple output segments, and develop EGO to train MLLMs for this setting.

Multi-Segment Video Temporal Grounding Recent works have move to multi-segment video temporal grounding. FlashMMR (Cao et al., 2025) refines and verifies multiple candidate moments, while GVMR's BCANet (You et al., 2025) uses boundary-aware cross-attention and evidence scores to support multi-event retrieval. More recent MLLM-based methods optimize multi-segment grounding through RL. MUSEG (Luo et al., 2025) scores predictions through temporal segment matching; OMTG (Xu et al., 2026) combines temporal and cardinality rewards with caption-based reasoning rewards; and TimeLens2 (Zhu et al., 2026) complements set-level temporal IoU with a Wasserstein reward. Despite their different reward designs, these RL methods aggregate prediction quality at the response level and assign a shared advantage to the generated sequence. Such supervision cannot distinguish the contributions of accurate and erroneous event predictions within the same response. Our EGO addresses this limitation by assigning event-specific advantages to individual predictions, enabling fine-grained credit assignment within each response.

3 THE ANYMOMENT DATASET

We introduce **AnyMoment**, a benchmark for Any-Moment Temporal Grounding (AMTG), with explicit support for zero, single, and multiple matching moments.

Table 1: Target-count composition of temporal grounding datasets.

Dataset	Domain	Zero	Single	Multi
Anet-Captions (Krishna et al., 2017)	Activity	0	72.0K	0
Charades-STA (Gao et al., 2017)	Activity	0	16.1K	0
TACoS (Regneri et al., 2013)	Cooking	0	18.2K	0
DiDeMo (Anne Hendricks et al., 2017)	Open	0	41.2K	0
QVHighlights (Lei et al., 2021)	Vlog/News	0	5.8K	2.9K
QV-M2 (Cao et al., 2025)	Vlog/News	0	0.74K	1.47K
OMTG (Xu et al., 2026)	Open	0	1.37K	54.82K
AnyMoment-HACS (ours)	**Action**	**13.7K**	**0.3K**	**17.3K**
AnyMoment-InternVid (ours)	**Open**	**12.3K**	**7.3K**	**7.3K**
AnyMoment-Total (ours)	**Open**	**26.0K**	**7.5K**	**24.6K**

Data Annotation. The videos of AnyMoment are sourced from two sources: the human-verified action detection dataset HACS (Zhao et al., 2019), and the open-domain dataset InternVid (Wang et al., 2023). HACS (Zhao et al., 2019) provides precise temporal action annotations but only closed-set labels. **For multi-event data**, we prompt an MLLM (Comanici et al., 2025) to rewrite each action label into a free-form query that semantically summarizes the action and is required to cover all its occurrences in the video. We further sample videos from InternVid (Wang et al., 2023) to broaden visual coverage. We prompt the MLLM to produce a single query that summarizes the recurring events and returns all timestamps where the event occurs. **For zero-event data**, we apply a minimal-edit strategy: for each valid query, an MLLM makes the smallest semantic change needed to produce a query that no longer matches the video. This yields rejection samples that are semantically close to positives and therefore require fine-grained visual discrimination. **For quality control**, we crop each annotated segment and use an MLLM to verify its semantic consistency with the query, while also checking the unannotated portions for missed occurrences; samples that fail either check are removed. **For the test set**, we manually annotate a test set across zero-, single-, and multi-events categories, sampling from our data sources with full manual inspection.

Data Statistics and Analysis. As shown in Table 1, our AnyMoment contains 58.1K queries, including 26.0K zero-, 7.5K single-, and 24.6K multi-segment queries. AnyMoment therefore supports rejection, conventional single-event grounding, and multi-event grounding within one benchmark. The two subsets remain complementary: AnyMoment-HACS emphasizes repeated human actions, whereas AnyMoment-InternVid contributes broader open-domain coverage. More details and statistics are reported in Appendix B.

4 METHOD

4.1 OVERVIEW AND TASK FORMULATION

Task Definition. Given an untrimmed video V and a natural language query Q, Any-Moment Temporal Grounding (AMTG) requires the model to output a set of temporal segments $\mathcal{S} = \{S_i = (s_i, e_i)\}_{i=1}^{K}$, where s_i and e_i denote the start and end timestamps of a matching occurrence. Unlike conventional VTG, which assumes $K = 1$, AMTG allows zero, one, or multiple segments. We denote the ground-truth set by $\mathcal{S}^* = \{S_k^*\}_{k=1}^{K^*}$ and the predicted set by $\hat{\mathcal{S}}$.

Overview. Successfully solving AMTG requires both recognizing distinct occurrences of the queried event and accurately locating their temporal boundaries. We propose EGO to connect these capabilities through occurrence-aware reasoning and targeted optimization. As illustrated in Fig. 2, EGO comprises three components. First, *bivariate difficulty-aware sampling* (Sec. 4.2) estimates understanding difficulty from whether the grounding outputs contain the correct number of intervals, and localization difficulty from how well those intervals align with ground truth. It selects moderately challenging examples for RL and uses the remaining data for a supervised fine-tuning (SFT) cold start. Next, *occurrence-aware CoT reasoning* (Sec. 4.3) uses this SFT phase to teach the model to generate specificity-enriched subqueries and localize the described occurrences. Finally, *RL with event-level credit assignment* (Sec. 4.4) combines event-specific advantages with global response-level feedback, allowing individual predictions within one response to receive distinct learning signals.

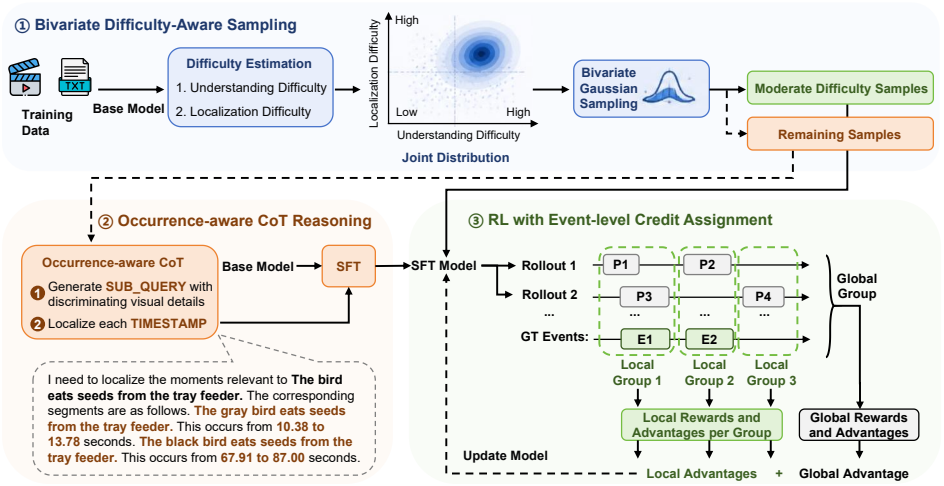

Figure 2: **Overview of Event-wise Grounding Optimization (EGO).** (1) Bivariate difficulty-aware sampling selects moderately challenging samples for RL and uses the remaining examples for SFT. (2) Occurrence-aware CoT reasoning generates discriminative subqueries and localizes each described occurrence. (3) RL with event-level credit assignment combines local event-specific advantages with a global response-level advantage to update the policy.

4.2 Bivariate Difficulty-Aware Data Sampling

Following recent RL-based VTG methods (Wang et al., 2025), our training pipeline consists of a CoT SFT cold start followed by RL post-training. AMTG requires the number of predicted events to be correct, and the time boundaries of each event to be precise. For some samples, it is difficult to identify all occurrences, while for others, it is difficult to determine their time boundaries accurately. To select examples for RL, existing approaches (Wang et al., 2025; Zhang et al., 2025; Dong et al., 2025) estimate difficulty using a single scalar, such as zero-shot temporal IoU. Such a measure does not distinguish between the two challenges. We therefore estimate them separately and jointly model them to select informative RL examples.

Difficulty Estimation. For each training pair (V, Q) with ground-truth set $\mathcal{S}^*$, we sample N responses from the zero-shot prompt asking the base model to return all matching intervals. Response n yields an interval set $\hat{\mathcal{S}}_n$, and its size $\hat{K}_n = |\hat{\mathcal{S}}_n|$ is the predicted number of occurrences. Understanding difficulty d_u is the fraction of responses for which $\hat{K}_n$ differs from K^*. Localization difficulty d_l measures how poorly those same interval sets align with $\mathcal{S}^*$. Formally,

$$d_u = 1 - \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[\hat{K}_n = K^*], \qquad d_l = 1 - \frac{1}{N}\sum_{n=1}^{N} \mathcal{I}(\hat{\mathcal{S}}_n, \mathcal{S}^*). \tag{1}$$

Here $\mathcal{I}(\hat{\mathcal{S}}_n, \mathcal{S}^*)$ is the mean temporal intersection over union (IoU) of one-to-one matched predicted and ground-truth intervals. To calculate $\mathcal{I}$ for a predicted set $\hat{\mathcal{S}} = \{\hat{S}_j\}$, we use the Hungarian algorithm to find the one-to-one assignment σ^* that maximizes total temporal IoU with $\mathcal{S}^* = \{S_k^*\}$. We then average the IoUs of the assigned interval pairs:

$$\sigma^* = \arg\max_{\sigma} \sum_{(j,k)\in\sigma} \mathrm{IoU}(\hat{S}_j, S_k^*), \qquad \mathcal{I}(\hat{\mathcal{S}}, \mathcal{S}^*) = \frac{1}{|\sigma^*|} \sum_{(j,k)\in\sigma^*} \mathrm{IoU}(\hat{S}_j, S_k^*). \tag{2}$$

We provide the data difficulty distributions in the Appendix C.2.

Difficulty-Based Sampling. Examples that are already easy for the base model offer limited scope for improvement, whereas consistently unsuccessful examples may provide little reward variation for relative policy optimization. We exclude examples with $d_u = d_l = 1$ from the RL candidate pool and sample the remaining examples in proportion to a bivariate Gaussian density $\mathcal{N}(\mu, \Sigma)$ over (d_u, d_l). Its center lies in the intermediate-to-high difficulty region, prioritizing examples that are challenging

yet learnable across both dimensions. Selected examples form the RL subset, while the remaining training examples provide the CoT SFT cold start shown in Fig. 2.

4.3 Occurrence-Aware Chain-of-Thought Reasoning

Our diagnostic study in Fig. 1(b) shows that decomposing a general multi-event query into specificity-enriched single-event queries improves localization. This motivates a CoT schema that uses the model's semantic understanding to formulate more precise localization targets. Rather than grounding all similar occurrences directly from one general query, the model describes the distinctive visual context of each occurrence and then localizes that description.

CoT Format. For positive queries, the reasoning consists of successive subquery–timestamp pairs, followed by a final answer collecting the predicted intervals. Each subquery preserves the original event semantics while adding discriminative details. The response follows the structure

```
<think> I need to localize the moments relevant to {QUERY}.
The corresponding segments are as follows.  {SUB_QUERY_1}
This occurs from {START_1} to {END_1} seconds.  ... </think>
<answer>[[{START_1}s, {END_1}s], ...]</answer>
```

An example is shown in Fig. 2. For zero-segment queries, the model explains the visual–textual mismatch and returns `<answer>[]</answer>`, grounding rejection in the observed content.

CoT Data and Fine-Tuning. We construct reasoning annotations using an MLLM (Comanici et al., 2025). For positive samples, it receives the video, original query, and ground-truth intervals, and generates a discriminative description for each interval using details such as participant attributes, background context, and surrounding actions. These descriptions are paired with the annotated timestamps. For negative samples, the annotator explains why the minimally edited query does not match the video. We fine-tune the base model on these responses using the training examples not selected for RL. This establishes a prior for occurrence-specific description, temporal localization, and evidence-based rejection.

4.4 RL with Event-Level Credit Assignment

After SFT, we optimize the policy using Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Conventional GRPO assigns a shared advantage to an entire response. As illustrated in Fig. 1(c), this can penalize accurate predictions along with errors or reinforce erroneous predictions in an otherwise successful response. We address this mismatch by forming local comparison groups for individual events across rollouts and combining their advantages with global response-level feedback, as shown in Fig. 2.

Local Group Construction. For each video–query pair, we sample M rollouts. We use $n \in \{1, \ldots, M\}$ to index a rollout and $g \in \mathcal{G} = \{1, \ldots, K^*, \emptyset\}$ to index a local group. Each ground-truth event $g \leq K^*$ defines one group; $g = \emptyset$ is a separate group for predictions that match no ground-truth event. Within each rollout, we assign predicted intervals to ground-truth intervals one-to-one using the Hungarian algorithm with temporal IoU as the matching similarity. Let $\hat{S}_{n,j}$ be prediction j in rollout n, and let U_n be its unmatched predictions. Each group receives one score from each rollout:

$$r_{g,n} = \begin{cases} \mathrm{IoU}(S_g^*, \hat{S}_{n,j}), & \text{if prediction } j \text{ matches event } g, \\ 0, & \text{if event } g \text{ is missed,} \end{cases} \quad g \in \{1, \ldots, K^*\},$$
$$r_{\emptyset,n} = \begin{cases} -1, & \text{if } U_n \neq \emptyset, \\ 0, & \text{if } U_n = \emptyset. \end{cases} \tag{3}$$

Thus, a matched event is scored by its temporal IoU, a missed event contributes zero, and the unmatched group penalizes a rollout with any extra predictions.

Local Advantages and Token Assignment. For each group g, we compare its M scores across rollouts and compute a group-relative advantage:

$$a_{g,n} = \frac{r_{g,n} - \bar{r}_g}{s_g + \epsilon}, \qquad \bar{r}_g = \frac{1}{M} \sum_{n=1}^{M} r_{g,n}, \qquad s_g^2 = \frac{1}{M-1} \sum_{n=1}^{M} (r_{g,n} - \bar{r}_g)^2. \tag{4}$$

This comparison gives each event its own learning signal instead of letting the quality of other predictions determine its advantage. For a matched prediction, we assign $a_{g,n}$ to its subquery and timestamp in <think> and the corresponding interval in <answer>; an unmatched prediction receives the advantage of the $\emptyset$ group. If event g is missed, its advantage is assigned to the termination tokens </think> and </answer>: the model ended its response while that event was still unpredicted.

Global Rewards and Advantage Fusion. In addition to local event scores, we evaluate each rollout with a response-level global reward:

$$\mathcal{R}^{\text{global}} = \mathcal{R}_{\text{format}} + \mathcal{R}_{\text{match}} + \mathcal{R}_{\text{count}} + \mathcal{R}_{\text{cons}} + \mathcal{R}_{\text{rej}}. \tag{5}$$

The format term $\mathcal{R}_{\text{format}}$ enforces the required <think> and <answer> structure. Following MUSEG (Luo et al., 2025), $\mathcal{R}_{match}$ provide temporal-alignment feedback using one-to-one segment matching. The count term $\mathcal{R}_{\text{count}}$ checks the number of intervals in the final answer. $\mathcal{R}_{\text{cons}}$ checks whether reasoning-stage timestamps agree with that answer. To explicitly guide the model to learn the zero-event decision in the global reward, we also propose a rejection reward $\mathcal{R}_{\text{rej}}$ that encourages the model to accurately reflect the emptiness status of the ground truth, rejecting when it should, and predicting segments when it should not. Let K^* and $\hat{K}_n$ denote the numbers of ground-truth and predicted segments for rollout n. The rejection reward is

$$\mathcal{R}_{rej} = \mathbb{I}(K^* = 0 \text{ and } \hat{K} = 0) \text{ or } (K^* > 0 \text{ and } \hat{K} > 0)$$

The details of the global rewards are provided in the Appendix C.6. Finally, we normalize $\mathcal{R}^{\text{global}}$ across the M rollouts to obtain the advantages A_n^{global}, which we average with the local advantage for group g:

$$A_{g,n} = \frac{A_n^{\text{global}} + a_{g,n}}{2}. \tag{6}$$

We apply $A_{g,n}$ to the tokens associated with group g in the policy update. Event-level feedback thus distinguishes predictions within one response, while global feedback preserves overall response quality.

5 Experiments

5.1 Datasets and Evaluation

Evaluation Benchmarks. We evaluate EGO on the *AnyMoment-HACS* and *AnyMoment-InternVid* test subsets, which assess grounding when the queried event may have zero, one, or multiple occurrences. To assess transfer beyond our benchmark, we also test zero-shot performance on the multi-event *QV-M2* test set (Cao et al., 2025). To separately verify the effectiveness of our method, we further train our model on the OMTG (Xu et al., 2026) dataset and evaluate on the multi-event *OMTG Bench*. We also evaluate the zero-shot performance on the single-event *Charades-TimeLens* (Zhang et al., 2025) benchmark.

Evaluation Metrics. For the zero- and multi-event grounding, we report mean temporal IoU (mIoU), temporal F1 (F1@0.5/0.7), and rejection F1 (RejF1). On OMTG Bench, we follow its official metrics (Xu et al., 2026): count accuracy (C-Acc), F1@0.3/0.5/0.7, and mIoU. On Charades-TimeLens, we report tIoU and R1@0.5/0.7. More details are in the Appendix D.

Implementation Details. We train EGO on AnyMoment using Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and Qwen3-VL-8B-Instruct (Yang et al., 2025) as base models. For the OMTG experiment, we follow OMTG (Xu et al., 2026) in using Qwen3-VL-4B, but train EGO on OMTG-46K with our event-level local advantages. Exact hyperparameters and prompts are deferred to the Appendix C.

5.2 Comparison with State-of-the-Art Methods

Comparison on zero-event and multi-event benchmarks. (1) Table 2 shows that EGO improves localization and rejection on both AnyMoment subsets across model backbones. EGO-7B's substantially higher RejF1 than MUSEG and Time-R1 demonstrates the ability to recognize absent events.

[1]OMTG-4B was trained on both OMTG-46K and TimeLens-100K, whereas our EGO-4B model was trained only on OMTG-46K.

Table 2: **Comparison with multimodal large language models on zero-event and multi-event video temporal grounding benchmarks.**

Method	Size	AnyMoment-HACS				AnyMoment-InternVid				QV-M2 (Cao et al., 2025)		
		mIoU	F1@0.5	F1@0.7	RejF1	mIoU	F1@0.5	F1@0.7	RejF1	mIoU	F1@0.5	F1@0.7
Qwen2.5-VL-7B-Instruct as base model												
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	7B	23.0	24.9	14.3	46.1	33.0	36.4	20.4	74.4	21.5	26.6	16.3
VideoChat-R1.5 (Yan et al., 2025)	7B	20.2	23.4	11.6	0.0	30.1	31.3	14.7	0.0	9.5	11.4	5.2
Time-R1 (Wang et al., 2025)	7B	21.7	26.7	15.3	18.4	34.1	37.7	21.3	32.1	25.5	32.2	22.1
MUSEG (Luo et al., 2025)	7B	33.6	39.5	23.2	0.0	40.2	44.3	24.9	0.3	31.0	36.2	26.6
MUSEG (Luo et al., 2025) w/ Ours Data	7B	35.3	41.9	25.7	24.8	43.5	47.5	30.7	55.2	33.4	38.6	29.3
EGO-7B (Ours)	7B	**41.3**	**47.9**	**31.5**	**75.4**	**50.4**	**57.6**	**42.4**	**87.1**	**38.0**	**46.0**	**32.9**
Qwen3-VL-8B-Instruct as base model												
Qwen3-VL-8B-Instruct (Yang et al., 2025)	8B	21.4	24.0	14.0	48.3	40.5	47.7	32.7	80.6	21.9	28.1	19.0
TimeLens-8B (Zhang et al., 2025)	8B	20.8	26.6	16.1	0.0	33.3	38.3	21.1	0.0	27.3	35.1	25.7
TimeLens2-8B (Zhu et al., 2026)	8B	24.0	29.5	18.1	0.0	36.7	39.4	23.0	0.0	30.6	37.5	28.6
EGO-8B (Ours)	8B	**43.3**	**50.0**	**35.0**	**75.0**	**56.4**	**64.2**	**52.7**	**93.5**	**42.8**	**51.9**	**36.3**

Table 3: **Comparison on the multi-event OMTG Bench.** C-Acc is count accuracy, F1 is temporal F1, and mIoU is average temporal IoU.

Method	C-Acc	F1@0.3	F1@0.5	F1@0.7	mIoU
Seed-1.8	38.12	67.13	54.67	38.79	56.81
Gemini-2.5-Pro	50.94	55.72	43.57	27.97	43.24
Gemini-3-Pro	30.63	58.30	47.75	29.89	47.63
VideoChat-R1-7B (Li et al., 2025a)	0.00	32.07	19.70	10.42	24.93
VideoChat-R1.5-7B (Yan et al., 2025)	0.31	28.41	15.53	9.85	27.96
Time-R1-7B (Wang et al., 2025)	0.00	28.94	18.73	10.00	24.11
UniTime (Li et al., 2025b)	0.00	35.27	30.15	23.58	37.12
TimeLens-8B (Zhang et al., 2025)	0.00	39.14	32.76	22.58	32.38
OMTG-4B (Xu et al., 2026)	55.63	73.46	65.40	48.96	61.24
EGO-4B (Ours)	**59.06**	**75.72**	**67.03**	**50.86**	**62.43**

Table 4: **Zero-shot evaluation on single-event Charades-TimeLens.** Gray results indicate methods trained on TimeLens data.

Method	tIoU	R1@0.5	R1@0.7
OMTG-4B (Xu et al., 2026) [1]	50.5	58.3	32.0
TimeLens-7B (Zhang et al., 2025)	48.8	55.6	28.4
TimeLens-8B (Zhang et al., 2025)	55.2	63.0	35.2
VideoChat-Flash-7B (Li et al., 2024)	39.7	37.9	17.8
VideoChat-R1-7B (Li et al., 2025a)	33.7	30.8	11.7
Time-R1-7B (Wang et al., 2025)	36.6	32.0	16.9
Qwen2.5-VL-7B (Bai et al., 2025)	39.3	37.8	16.6
Qwen3-VL-4B (Yang et al., 2025)	45.6	47.5	24.1
Qwen3-VL-8B (Yang et al., 2025)	48.3	53.4	27.5
EGO-4B (Ours)	**49.9**	**55.8**	**31.5**

(2) EGO also transfers zero-shot to QV-M2, an independent multi-event benchmark, and achives the best performance. (3) The comparison among MUSEG, MUSEG w/ OursData, and EGO separates the effect of our data from the effect of our method. Adding AnyMoment training data to MUSEG consistently improves its grounding and rejection performance, while EGO achieves further gains on AnyMoment, supporting contributions from both the dataset and the method.

Comparison on OMTG Bench. To separately verify EGO's effectiveness beyond AnyMoment, Table 3 compares models on multi-event OMTG Bench using the same Qwen3-VL-4B backbone. EGO outperforms OMTG-4B across all reported metrics. This is because our local advantages assign separate learning signals to individual event predictions, avoiding shared credit for correct and erroneous predictions in the same response. The ablation below tests this mechanism more directly.

Comparison on single-event benchmarks. Table 4 shows that EGO-4B improves its Qwen3-VL-4B backbone on every Charades-TimeLens metric and also outperforms Time-R1, without training on TimeLens data. The gray baselines use TimeLens training data, so their results are not directly comparable as zero-shot evaluations. Although EGO is designed for AMTG, these gains show that event-wise training also benefits single-event localization.

5.3 Ablation Studies

We conduct ablations on AnyMoment-InternVid using Qwen2.5-VL-7B-Instruct as the base model.

Effect of bivariate data sampling. The first block of Table 5 shows that bivariate sampling outperforms univariate sampling on every metric. Its larger gains in F1@0.7 and RejF1 suggest that jointly considering understanding and localization difficulty primarily helps recover and localize present events more accurately, while further improving rejection.

Effect of CoT and RL training stages. The second block of Table 5 shows that CoT SFT and RL each improve the base model across all metrics, with RL alone giving the larger gains. Their combination performs best, improving localization further over RL alone. This demonstrates the effectiveness of both our occurrence-aware CoT reasoning and the RL with event-level credit assignment.

Table 5: Ablation studies on the AnyMoment-InternVid dataset.

Model	mIoU	F1@0.5	F1@0.7	RejF1
Univariate Sample	49.2	56.3	40.3	81.9
Bivariate Sample	**50.4**	**57.6**	**42.4**	**87.1**
Qwen2.5-VL-7B-Instruct	33.0	36.4	20.4	74.4
+ CoT	42.0	49.1	33.8	81.4
+ RL	48.9	55.4	40.5	86.5
+ CoT + RL	**50.4**	**57.6**	**42.4**	**87.1**
Local advantages only	47.3	55.2	39.8	83.6
Global advantage only	50.1	57.0	41.0	84.3
Global + local advantages	**50.4**	**57.6**	**42.4**	**87.1**

Table 6: Detailed analysis of our local advantages on the AnyMoment-InternVid dataset.

# GT Events	Model	mIoU	F1@0.5	F1@0.7
1	Global advantage only	59.7	**67.5**	49.5
	Global + local advantages	**59.7**	66.5	**50.2**
2	Global advantage only	**43.9**	50.0	35.7
	Global + local advantages	43.5	**50.7**	**37.1**
3	Global advantage only	42.9	49.1	34.5
	Global + local advantages	**45.5**	**53.2**	**38.1**
≥ 4	Global advantage only	38.9	46.6	29.8
	Global + local advantages	**41.6**	**52.1**	**32.1**

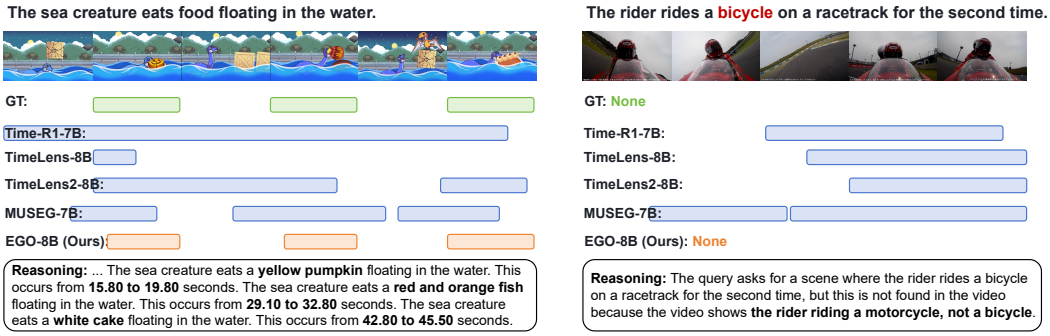

Figure 3: Qualitative visualizations on the AnyMoment-InternVid test set.

Effect of local and global advantages. (1) The final block of Table 5 shows that local advantages alone do not replace the response-level supervision provided by global rewards. Local scores primarily provide event-wise temporal feedback and penalize unmatched predictions, but they do not directly enforce the required output format, the predicted event count, consistency between reasoning and the final answer, or the zero-event rejection decision. (2) Combining global and local advantages performs best overall. The breakdown by event count in Table 6 shows where local advantages matter most: performance is comparable for single-event queries, while multi-event queries benefit increasingly as the number of events grows. For a single event, a response-level advantage largely reflects that event's quality. When several events occur, correct and erroneous predictions can coexist in one response, and event-level advantages assign them distinct learning signals.

Qualitative analysis. Figure 3 provides two representative cases from AnyMoment-InternVid. For the query with three occurrences, EGO-8B recovers all three event windows and aligns them closely with the ground truth, whereas the baselines miss occurrences or merge them into overly broad intervals. For the absent-event query, EGO correctly outputs "None" and identifies the bicycle–motorcycle mismatch, while the other methods hallucinate temporal segments. These examples illustrate event-level training improves both multi-event grounding and zero-event rejection.

6 Conclusion

We presented Any-Moment Temporal Grounding (AMTG), a formulation of video temporal grounding in which a query may correspond to zero, one, or multiple temporal events. To support this setting, we introduced AnyMoment, a dataset that combines recurring-event localization with absent-event rejection, and Event-wise Grounding Optimization (EGO), which integrates bivariate difficulty-aware sampling, occurrence-aware reasoning, and local event-level credit assignment with a global objective. Across AnyMoment, QV-M2, OMTG Bench, and Charades-TimeLens, EGO consistently improves localization and rejection over baselines and transfers to unseen benchmarks. Stratified ablations further show that local advantages become more valuable as the number of target events increases, demonstrating the importance of assigning fine-grained event-level credit for reliable multi-event video temporal grounding.

AI USE STATEMENT

In this work, we used generative AI tools for [assist with translation, generate synthetic data sets, clean and reformat dataset, and implement methods]. We have not used generative AI tools for [propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, and interpret results], and [help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, and assist in the writing of proofs] are not applicable to this work. Additionally, we used generative AI tools for [summarize or analyse existing literature]. We have reviewed all AI-assisted work. [LLM-helped translations have been checked by authors]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work uses publicly available video datasets and does not involve direct interaction with human subjects or the collection of new personally identifiable information. As with large-scale web video datasets and MLLM-assisted annotation, the resulting data and models may inherit biases from their source data or annotation models. We therefore encourage responsible use of the dataset and models for research purposes and caution against deployment in high-stakes applications without additional validation for privacy, fairness, and reliability.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproducibility of our work. Section 3 describes the construction of the AnyMoment dataset, and Section 4 presents the formulation and implementation of EGO. Further details are provided in the appendix: Appendix B reports dataset and annotation statistics and quality analysis; Appendix C provides implementation details for data partitioning, video input, SFT, RL training, difficulty modeling, and reward design; Appendix D specifies the evaluation metrics and diagnostic experiments; and Appendix E lists the prompts used for data construction, training, evaluation, and baseline reproduction. Our AnyMoment dataset, EGO model, and code will be open-sourced after the paper is accepted.

REFERENCES

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Aritra Bhowmik, Martin R Oswald, Pascal Mettes, and Cees GM Snoek. Revisiting proposal-based object detection. *arXiv preprint arXiv:2311.18512*, 2023.

Zhuo Cao, Heming Du, Bingqing Zhang, Xin Yu, Xue Li, and Sen Wang. When one moment isn't enough: Multi-moment retrieval with cross-moment interactions. *arXiv preprint arXiv:2510.17218*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, and et.al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

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

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

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

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: detecting moments and highlights in videos via natural language queries. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

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. *arXiv preprint arXiv:2501.00574*, 2024.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025a.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. In *NeurIPS*, 2025b.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. *arXiv preprint arXiv:2109.06398*, 2021.

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

Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18930–18940, 2024.

Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10810–10819, 2020.

Yulin Pan, Xiangteng He, Biao Gong, Yiliang Lv, Yujun Shen, Yuxin Peng, and Deli Zhao. Scanning only once: An end-to-end framework for fast temporal grounding in long videos. *arXiv preprint arXiv:2303.08345*, 2023.

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

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

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

Gunnar A Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2156–2165. IEEE, 2017.

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. *arXiv preprint arXiv:2503.13377*, 2025.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2986–2994, 2021.

Qi Xu, Yue Tan, Shihao Chen, Jiahao Meng, Anna Wang, Shunping Ji, Hao Fei, and Jason Li. Towards one-to-many temporal grounding. *arXiv preprint arXiv:2606.06294*, 2026.

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. *arXiv preprint arXiv:2509.21100*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qin You, Qilong Wu, Yicong Li, Wei Ji, Li Li, Pengcheng Cai, Lina Wei, and Roger Zimmermann. Generalized video moment retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 49207–49223, 2025.

Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9159–9166, 2019.

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

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12870–12877, 2020.

Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8668–8678, 2019.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, et al. Timelens2: Generalist video temporal grounding with multimodal llms. *arXiv preprint arXiv:2607.17423*, 2026.

APPENDIX CONTENTS

1. Additional results and qualitative visualizations
   (a) Performance by target count
   (b) Limitations
   (c) Visualization of Event-Level Credit Assignment
   (d) Dataset visualizations
   (e) Qualitative output comparisons
2. Dataset and annotation statistics
   (a) Train/test statistics
   (b) Target-count and duration analysis
   (c) Dataset quality
3. Method details
   (a) Data partition and subset composition
   (b) Data difficulty distribution
   (c) Video input
   (d) Supervised fine-tuning
   (e) Reinforcement learning
   (f) Global reward definitions
4. Evaluation metrics
5. Prompts
   (a) Prompts for data collection
   (b) Prompts for CoT data construction
   (c) Training and test prompt for EGO
   (d) Test prompt for other baselines

A ADDITIONAL RESULTS AND QUALITATIVE VISUALIZATIONS

A.1 PERFORMANCE BY TARGET COUNT

We further analyze model performance with respect to the number of ground-truth target segments. We report *exact count accuracy* to evaluate whether a model predicts the correct number of event occurrences, and F1@0.5 to jointly assess the completeness and temporal accuracy of the predicted segments. For zero-target queries, exact count accuracy corresponds to the correct-rejection rate, while F1@0.5 is reported only for queries containing at least one target event.

Table 7: Exact count accuracy (%) grouped by the number of ground-truth (GT) segments. "All" denotes the overall accuracy across all target counts. Bold denotes the best result within each base-model family.

Method	AnyMoment-HACS						AnyMoment-InternVid					
	GT=0	GT=1	GT=2	GT=3	GT≥4	All	GT=0	GT=1	GT=2	GT=3	GT≥4	All
# Samples	405	265	869	512	1,149	3,200	1,112	788	800	200	100	3,000
Qwen2.5-VL-7B-Instruct as base model												
Qwen2.5-VL-7B	30.1	26.8	50.7	11.3	6.8	24.1	59.9	55.7	55.4	19.5	15.0	53.4
VideoChat-R1.5-7B	0.0	58.5	33.9	11.3	3.0	17.0	0.0	63.8	42.8	14.0	8.0	29.4
Time-R1-7B	10.1	**100.0**	2.4	2.0	3.0	11.6	19.2	**100.0**	2.1	5.0	10.0	34.6
MUSEG-7B	0.0	71.3	22.8	**43.0**	11.0	22.9	0.4	77.2	38.8	**46.5**	24.0	34.6
EGO-7B	**63.0**	84.9	**59.1**	34.4	**13.7**	**41.5**	**78.7**	84.0	**55.5**	38.0	**29.0**	**69.5**
Qwen3-VL-8B-Instruct as base model												
Qwen3-VL-8B	32.1	86.4	10.4	11.3	8.0	18.7	83.1	96.6	17.5	22.5	23.0	63.1
TimeLens-8B	0.0	95.5	1.0	1.4	1.7	9.0	0.0	95.2	1.4	5.0	6.0	25.9
TimeLens2-8B	0.0	**98.5**	6.4	3.5	2.2	11.4	0.0	**99.7**	9.4	8.5	4.0	29.4
EGO-8B	**63.0**	76.2	**43.6**	**36.9**	**18.5**	**38.7**	**89.1**	83.1	**50.2**	**47.5**	**33.0**	**72.5**

Table 8: F1@0.5 (%) grouped by the number of ground-truth (GT) segments. "All" denotes the overall F1@0.5 across all non-zero-target queries. Zero-target queries are omitted from the stratified results because temporal localization is undefined for them. Bold denotes the best result within each base-model family.

Method	AnyMoment-HACS					AnyMoment-InternVid				
	GT=1	GT=2	GT=3	GT≥4	All	GT=1	GT=2	GT=3	GT≥4	All
Qwen2.5-VL-7B-Instruct as base model										
Qwen2.5-VL-7B	41.8	33.7	22.6	15.4	24.9	45.8	30.0	30.9	25.3	36.4
VideoChat-R1.5-7B	50.9	31.7	21.7	11.6	23.4	42.5	25.0	20.4	14.4	31.3
Time-R1-7B	64.5	39.2	21.2	10.9	26.7	55.6	27.3	20.8	14.4	37.7
MUSEG-7B	56.9	47.5	39.9	29.2	39.5	52.7	39.3	37.9	31.3	44.3
EGO-7B	**66.7**	**55.7**	**49.9**	**36.8**	**47.9**	**66.5**	**50.7**	**53.2**	**52.1**	**57.6**
Qwen3-VL-8B-Instruct as base model										
Qwen3-VL-8B	53.5	26.6	22.2	20.6	24.0	64.0	36.0	38.9	30.3	47.7
TimeLens-8B	62.5	36.6	22.0	12.7	26.6	55.0	29.6	18.9	14.9	38.3
TimeLens2-8B	**75.3**	44.5	21.8	10.0	29.5	64.4	24.2	16.2	9.6	39.4
EGO-8B	60.1	**51.1**	**51.9**	**46.0**	**50.0**	**72.0**	**57.3**	**64.0**	**59.6**	**64.2**

As shown in Table 7, both EGO variants achieve the highest overall exact count accuracy within their respective base-model families on both datasets. Several existing methods perform particularly well when GT=1 but degrade substantially for zero-target and multi-target queries. For example, Time-R1 and TimeLens2 achieve nearly perfect count accuracy for single-target queries, yet their performance drops sharply as the number of target events increases. This suggests that these methods retain a strong bias toward predicting a single segment. In contrast, EGO maintains substantially more balanced performance across different target counts, including zero-target queries.

The advantage of EGO becomes more pronounced when considering temporal localization quality. As shown in Table 8, EGO consistently improves F1@0.5 as the number of target events increases. For GT$\geq$ 4, EGO-7B surpasses the strongest competing method by 7.6 points on AnyMoment-HACS and 20.8 points on AnyMoment-InternVid, while EGO-8B improves by 25.4 and 29.3 points, respectively. Although exact counting remains challenging for densely repeated events, particularly on AnyMoment-HACS, these results demonstrate that EGO is substantially more effective at recovering multiple event occurrences while maintaining accurate temporal localization.

A.2 Limitations

Despite the improvements brought by EGO, multi-event grounding remains challenging, particularly when the queried event occurs many times. As shown in Table 7, exact count accuracy decreases substantially as the number of target events increases. A similar trend can be observed in temporal localization (Table 8), indicating that completely recovering many repeated occurrences remains difficult even with event-level optimization. Moreover, supporting zero-event rejection introduces a trade-off for single-event queries. EGO may occasionally reject a query whose target is actually present, resulting in slightly lower counting or localization performance on some single-event subsets than methods specifically optimized under the assumption that a target segment always exists. Developing more reliable occurrence counting and better calibrated rejection is therefore an important direction for future work.

A.3 Visualization of Event-Level Credit Assignment

Figure 4 further illustrates how our event-level credit assignment provides fine-grained supervision within individual rollouts. The two examples are different rollouts sampled from the same training instance. As shown in Figure 4, the local advantages can simultaneously reinforce a well-localized event prediction while penalizing inaccurate or unmatched predictions in the same response. In contrast, the global advantage assigns a single learning signal to the whole rollout and therefore cannot distinguish good and bad event predictions within that response. The unmatched extra prediction receives a negative local advantage, while the correctly localized prediction is positively reinforced. Meanwhile, when a ground-truth occurrence is missed, there is no corresponding prediction token



(a) Rollout 1.



(b) Rollout 2.

Figure 4: Visualization of event-level credit assignment for two different rollouts of the same training sample. Green denotes positive advantages, whose corresponding tokens are reinforced, while red denotes negative advantages, whose tokens are penalized.

to which its learning signal can be assigned. We therefore assign the negative advantage of the missed occurrence to the response-termination tokens, i.e., `</think>` and `</answer>`, as shown in Figure 4 (b). This explicitly penalizes premature termination when valid occurrences remain unpredicted, encouraging the model to continue generating until all target events are covered.

A.4 Dataset Visualizations



Query: A man uses an *electric toothbrush* to brush a girl's teeth in a bathroom.

GT: 56.03s – 72.00s; 122.48s – 126.28s; 162.68s – 176.49s

CoT Annotation: I need to localize the moments relevant to A man uses an electric toothbrush to brush a girl's teeth in a bathroom. The corresponding segments are as follows. A man uses an electric toothbrush to brush a girl's teeth for the first time while she stands upright in a bathroom. This occurs from **56.03 to 72.00 seconds**. A man uses an electric toothbrush to brush a girl's teeth for the second time while she lies on his lap in a bathroom. This occurs from **122.48 to 126.28 seconds**. A man uses an electric toothbrush to brush a girl's teeth for the third time while she lies on his lap in a bathroom. This occurs from **162.68 to 176.49 seconds**.

Query: A man uses a *manual toothbrush* to brush a girl's teeth in a bathroom.

GT: None

CoT Annotation: The query seeks a scene where a man uses a manual toothbrush to brush a girl's teeth in a bathroom, but this is not found in the video because in the target segments, the man consistently uses an electric toothbrush to brush the girl's teeth, rather than a manual one.

(a) Visualization of our AnyMoment-HACS data.



Query: The woman is playing a fingerstyle melody on the *classical guitar*.

GT: 00.00s – 05.25s; 12.00s – 17.75s; 21.00s – 35.50s

CoT Annotation: I need to localize the moments relevant to The woman is playing a fingerstyle melody on the classical guitar. The corresponding segments are as follows. The woman is playing a fingerstyle melody on the classical guitar for the first time. This occurs from **0.00 to 5.25 seconds**. The woman is playing a fingerstyle melody on the classical guitar for the second time. This occurs from **12.00 to 17.75 seconds**. The woman is playing a fingerstyle melody on the classical guitar for the third time. This occurs from **21.00 to 33.50 seconds**.

Query: The woman is playing a fingerstyle melody on the *electric guitar*.

GT: None

CoT Annotation: The video does not show the woman playing a fingerstyle melody on the electric guitar, as she is instead playing a classical acoustic guitar, not an electric guitar.

(b) Visualization of our AnyMoment-InternVid data.

Figure 5: Dataset visualizations.

As illustrated in Figure 5, the AnyMoment dataset demonstrates two core characteristics: grounding any number of matching events and fine-grained visual disambiguation (rejection) capabilities. Taking the AnyMoment-HACS example, for a positive query ("electric toothbrush"), the dataset annotates all three non-contiguous occurrence intervals and provides high-quality structured Chain-of-Thought (CoT) annotations. The CoT uses specific visual context (e.g., the girl's posture changing from "stands upright" to "lies on his lap") to generate discriminative subqueries for individual occurrences before localizing them. This turns multi-event search into occurrence-specific localization steps.

More importantly, for negative queries containing subtle semantic substitutions (e.g., replacing "electric toothbrush" with "manual toothbrush", or "classical guitar" with "electric guitar"), the ground truth (GT) is strictly set to None. In these cases, the CoT annotation explicitly articulates the specific reasons why the video content mismatches the text query. This design forces the model to base its rejection decisions on genuine visual evidence rather than relying on linguistic priors, thereby effectively mitigating grounding hallucinations in fine-grained scenarios.



(a) AnyMoment-HACS.

(b) AnyMoment-InternVid.

Figure 6: Qualitative Comparisons on multi-segment data.



Figure 7: Qualitative Comparisons on zero-segment data.

A.5 Qualitative Output Comparisons

Figures 6 and 7 visualize model predictions on multi-segment and zero-segment queries, respectively. As we can see, for multi-segment grounding, existing methods often miss valid occurrences, merge multiple events into overly long intervals, or produce extra predictions. In contrast, EGO more consistently recovers all relevant moments with tighter temporal boundaries. For zero-segment queries, existing methods frequently hallucinate temporal segments even when the target event is absent. EGO correctly returns `None` in both examples, demonstrating the effectiveness of our zero-event data and rejection reward. Together, these visualizations show that EGO improves both complete multi-event localization and reliable rejection of absent events.

B Dataset and Annotation Statistics

B.1 Train/test statistics

Table 9: AnyMoment annotation statistics by source and split. A multi-target query has at least two ground-truth segments. "Moments" counts all positive temporal segments.

Source	Split	Videos	Hours	Queries	Zero	Single	Multi	Moments	Words / Query	Seconds / Moment
HACS	Train	13,798	610.4	28,060	13,280	0	14,780	57,626	15.1	22.1
HACS	Test	3,198	133.1	3,200	405	265	2,530	11,101	17.3	20.1
InternVid	Train	9,881	165.7	23,880	11,190	6,477	6,213	21,053	11.3	8.7
InternVid	Test	2,373	35.2	3,000	1,112	788	1,100	3,448	10.3	9.0
Total	**Train**	**23,679**	**776.2**	**51,940**	**24,470**	**6,477**	**20,993**	**78,679**	**13.3**	**18.5**
Total	**Test**	**5,571**	**168.2**	**6,200**	**1,517**	**1,053**	**3,630**	**14,549**	**13.9**	**17.5**

As shown in Table 9, the training set contains 51,940 queries, while the manually inspected test set contains 6,200 queries. HACS training positives intentionally focus on repeated actions, whereas its test split adds 265 single-target cases. InternVid supplies 6,477 single-target training queries and thereby broadens both the visual domain and the event-count distribution. Train and test video identifiers are disjoint for both sources.

B.2 Target-count and duration analysis

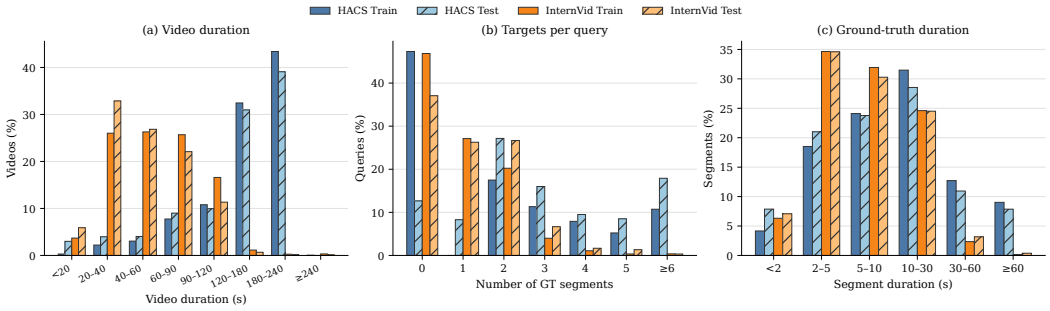

Figure 8: Distributions recomputed from the current annotations. Each series is normalized within its source and split. (a) Video duration uses unique video identifiers. (b) Query-level target counts group six or more segments in the final bin. (c) Ground-truth segment durations group segments of at least 60 seconds in the final bin.

Figure 8 summarizes the temporal statistics of AnyMoment. As shown in Fig. 8(a), HACS and InternVid cover complementary video durations: HACS mainly contains longer videos, while InternVid is dominated by shorter clips, with similar train/test distributions within each source. Figure 8(b) shows that HACS places greater emphasis on repeated events and contains more queries with many target segments, whereas InternVid is more concentrated on queries with zero to two targets. This complementary target-count distribution enables AnyMoment to cover both sparse and densely repeated events. As shown in Fig. 8(c), HACS also contains generally longer ground-truth segments, while InternVid is dominated by shorter events. The train/test distributions remain consistent across both sources. Overall, combining the two sources provides broad coverage in video length, target count, and event duration.

B.3 Dataset quality

Human agreement. Two annotators independently label the same samples. For zero-target samples, agreement is the percentage for which both annotators judge the event absent. For positive samples, we report the mean temporal IoU between their temporal annotations, following (Sigurdsson et al., 2017).

Table 10: Inter-annotator agreement by target type.

Target type	Agreement (%)
Zero	95.0
Single	80.1
Multi	77.4

As shown in Table 10, zero-target decisions are highly consistent. Temporal boundaries are inherently more ambiguous, especially when an event repeats, yet agreement remains high: 80.1% for single-target and 77.4% for multi-target samples. For context, (Sigurdsson et al., 2017) report 72.5% temporal-IoU agreement on Charades, so the positive-sample agreement in AnyMoment is higher under the same type of overlap measure.

Stratified automatic-annotation accuracy. We compare the original automatic labels against the final human-adjusted labels on the test set. A sample is counted as correct when tIoU≥ 0.7, yielding an estimated 87% accuracy for the automatic labels used in training. Table 11 separates this estimate by target type, data source, and mean target duration.

Table 11: Automatic-annotation accuracy (%) under tIoU≥ 0.7, stratified by target type, source, and mean segment duration.

By target type

Type	Accuracy
Zero	96.5
Single	82.6
Multi	80.6

By source

Source	Accuracy
HACS	94.0
InternVid	86.9

By mean segment duration

Duration	Accuracy
$<$ 2s	79.0
2–$<$ 5s	83.5
5–$<$ 10s	89.6
$\geq$ 10s	94.9

The 96.5% zero-target accuracy supports the minimal-edit rejection pipeline. Multi-target accuracy is only 2.0 points below single-target accuracy, indicating that the automatic pipeline usually recovers repeated moments rather than only one salient occurrence. HACS is more accurate because it begins with expert temporal annotations, whereas InternVid trades some precision for open-domain coverage and scalability. Accuracy increases with duration because an identical boundary displacement causes a larger relative tIoU decrease for a short segment.

Boundary errors. We additionally measure absolute boundary differences between the original automatic annotations and the final adjusted labels in Table 12.

Table 12: Distribution (%) of absolute start- and end-boundary errors.

Absolute error	Start	End
$\leq$ 0.5s	64.8	64.8
$>$ 0.5–1s	19.3	19.3
$>$ 1–2s	9.6	9.6
$>$ 2–5s	5.0	5.0
$>$ 5s	1.3	1.3

Overall, 84.1% of both start and end boundaries are within one second of the final annotation, while only 1.3% differ by more than five seconds. Together with the human-agreement results, this suggests that most residual disagreement is localized near inherently ambiguous event boundaries rather than caused by missing or spurious events.

C METHOD DETAILS

C.1 DATA PARTITION AND SUBSET COMPOSITION

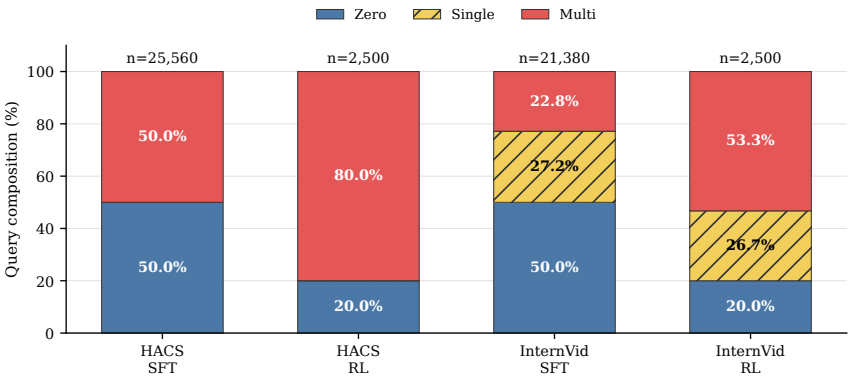

Figure 9: Query composition of the CoT SFT and RL data. Positive queries are summarized as single-target or multi-target. Totals above the bars give the number of queries.

Figure 9 summarizes the composition of the SFT and RL training data. For RL, we fix the proportion of zero-target samples at 20%. The remaining samples are drawn from non-zero-target queries, including both single- and multi-target cases, using our bivariate difficulty-aware sampler. The SFT data consist of the remaining training samples and provide complementary supervision before RL post-training.

C.2 DATA DIFFICULTY DISTRIBUTION

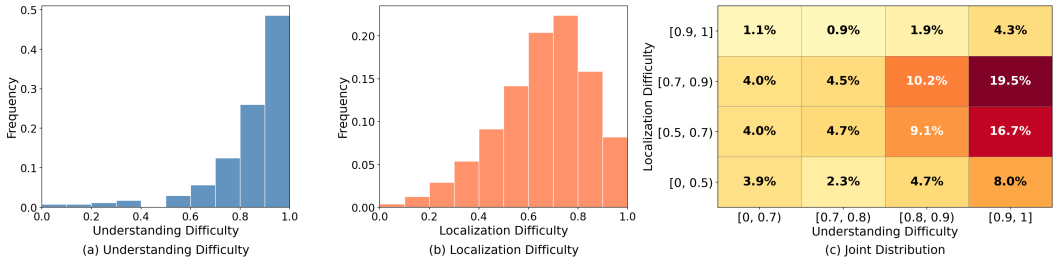

Figure 10: Distribution of understanding difficulty d_u, localization difficulty d_l, and their joint distribution for the RL candidate data. Larger values indicate harder examples.

Figure 10 empirically supports treating the two difficulty dimensions separately. Understanding difficulty is strongly skewed toward hard examples: 74.4% have $d_u \geq 0.8$, so the base model frequently predicts an incorrect occurrence count. Localization difficulty is more dispersed, with 72.7% of examples in the intermediate range $0.5 \leq d_l < 0.9$ and only 8.2% in the highest bin. The joint heatmap further shows that high count difficulty does not imply uniformly maximal localization difficulty: 36.2% of all examples have $d_u \geq 0.9$ but $0.5 \leq d_l < 0.9$, whereas only 4.3% fall in the joint top-right bin. A single scalar score would obscure this structure. The bivariate sampler can instead emphasize the populous challenging-but-learnable regions while excluding examples for which both dimensions are exactly one.

C.3 Video Input

For EGO, Time-R1, and MUSEG evaluation, videos are sampled at 2 FPS. We dynamically adjust the video resolution so that each video contributes approximately 3,584 visual tokens, allowing at most 768 frames per video. This design keeps the video-token budget comparable across methods.

C.4 Supervised Fine-Tuning

For the AnyMoment experiments, we initialize EGO-7B from Qwen2.5-VL-7B-Instruct and EGO-8B from Qwen3-VL-8B-Instruct, matching the two backbone groups in Table 2. SFT is performed with LoRA. We use LoRA rank 8, alpha 8, dropout 0.05, one training epoch, global batch size 32, learning rate 2×10^{-4}, warmup ratio 0.03, a cosine learning-rate schedule, bf16 training, gradient checkpointing, and DeepSpeed ZeRO-2. The vision tower, LLM base weights, and merger are frozen during SFT. LoRA is applied only to the language model modules.

C.5 Reinforcement Learning

RL starts from the SFT checkpoint and uses GRPO in VERL. We train for 2 epochs on 8 NPUs with train batch size 16, rollout number $N = 8$, actor learning rate 1×10^{-6}, and KL loss coefficient $\beta = 0.01$. The vision tower remains frozen during RL. RL training mixes HACS, InternVid, and Time-R1 (Wang et al., 2025) training samples.

C.6 Global Reward Definitions

For a rollout with ground-truth set $\mathcal{G} = \{g_i\}_{i=1}^{K}$ and parsed prediction set $\mathcal{P} = \{p_j\}_{j=1}^{M}$, the response-level reward in Eq. 5 is the unweighted sum

$$\mathcal{R}^{\text{global}} = \mathcal{R}_{\text{format}} + \mathcal{R}_{\text{match}} + \mathcal{R}_{\text{count}} + \mathcal{R}_{\text{cons}} + \mathcal{R}_{\text{rej}}. \tag{7}$$

The format reward is one only when the response follows the required `<think>...</think><answer>...</answer>` structure. The count and rejection terms are

$$\mathcal{R}_{\text{count}} = \mathbb{I}[K = M], \qquad \mathcal{R}_{\text{rej}} = \begin{cases} 1, & (K = 0 \wedge M = 0) \vee (K > 0 \wedge M > 0), \\ -1, & \text{otherwise.} \end{cases} \tag{8}$$

The signed rejection term explicitly penalizes a nonempty prediction for a zero-target query and an empty prediction for a positive query; the matching and count terms then refine the quality of positive predictions. For temporal matching, we solve maximum-weight one-to-one Hungarian matching σ^* and normalize by the larger set size:

$$\mathcal{R}_{\text{match}} = \begin{cases} \dfrac{\sum_{(i,j)\in\sigma^*} \text{IoU}(g_i, p_j)}{\max(K, M)}, & K > 0,\ M > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Finally, $\mathcal{R}_{\text{cons}} = 1$ only when the number of reasoning-stage timestamp pairs, and the number of final-answer intervals agree, and corresponding reasoning/final timestamps match within 10^{-3} seconds.

D Evaluation Metrics

For AnyMoment, rejection treats an empty ground-truth set as the positive class. A true positive is a zero-target query predicted as empty, a false positive is a positive-target query predicted as empty, and a false negative is a zero-target query with a nonempty prediction. We report

$$\text{RejP} = \frac{\text{TP}_{\text{rej}}}{\text{TP}_{\text{rej}} + \text{FP}_{\text{rej}}}, \quad \text{RejR} = \frac{\text{TP}_{\text{rej}}}{\text{TP}_{\text{rej}} + \text{FN}_{\text{rej}}}, \quad \text{RejF1} = \frac{2\,\text{RejP}\,\text{RejR}}{\text{RejP} + \text{RejR}}. \tag{10}$$

If $K > 0$, the sample contributes to localization metrics. F1@τ uses one-to-one greedy matching over all prediction-GT pairs whose IoU is at least threshold τ. Precision, recall, and F1 are

$$\text{Precision}_\tau = \frac{\text{TP}_\tau}{M}, \quad \text{Recall}_\tau = \frac{\text{TP}_\tau}{K}, \quad \text{F1}_\tau = \frac{2\text{Precision}_\tau\text{Recall}_\tau}{\text{Precision}_\tau + \text{Recall}_\tau}. \tag{11}$$

We report F1@0.5 and F1@0.7 in the main paper.

mIoU is computed by Hungarian bipartite matching and normalized by $\max(K, M)$:

$$\mathrm{mIoU} = \frac{\sum_{(i,j)\in\sigma^*} \mathrm{IoU}_{ij}}{\max(K, M)}. \tag{12}$$

All non-zero-sample metrics are averaged over non-zero samples only.

E PROMPTS

E.1 PROMPTS FOR DATA COLLECTION

We use Gemini-3-Pro (Comanici et al., 2025) for the data annotation and collection. The detailed prompts are as follows.

E.1.1 HACS-BASED ANNOTATION

The HACS branch starts from human-verified temporal action annotations. Since HACS labels are closed-set action categories, we convert each action label and its repeated temporal instances into a free-form natural-language query. The annotation prompt asks the MLLM to summarize only the behavior shared by all provided target segments.

```
You are a video content understanding and fine-grained description expert who
strictly follows instructions. Your task is to generate a set of high-precision
text descriptions for several target segments of <action_type> based on the
given video, for use in video segment retrieval and discrimination.

Generate a global description `global_query` that meets the following
requirements:
1. It must match all given segments and summarize only the shared observable
   behavior pattern across these segments.
2. It should be only 1 sentence.
3. Prefer the structure: subject(like person/people/animal/object) + action +
   key visible details.
```

E.1.2 INTERNVID-BASED ANNOTATION

The InternVid branch is designed to scale beyond human-action categories. We first identify subjects or objects that perform repeated activities, and then ask the MLLM to produce dense timestamped captions for each repeated event.

```
You are a senior data annotation expert proficient in generating high-quality
annotation data for the Video Temporal Grounding task.
Given a video {video_name}.mp4, you must output {video_name}.json. You need to
identify potential repeated events in the video. The final JSON file should
contain several `masklet_id` and their corresponding `dense_captions`:
1. `masklet_id`: Subjects appearing in the video, such as people, animals, or
   objects, which must perform repeated actions in the video.
2. `dense_captions`: All segments where this `masklet_id` performs an activity
   in the video. For each segment, provide all the start time (`start_time`),
   end time (`end_time`), a description of the event/activity (`caption`).
   - Caption format: "The <subject_category_name> <verb phrase>
     <goal/context>."
```

E.1.3 REJECTION QUERY CONSTRUCTION

For every positive query, we create semantically close hard negatives by minimally editing the query so that it no longer matches any target segment. The prompt also asks for an observable reason, which is later used as supervision for zero-segment CoT examples.

```
You are an expert in video content understanding and fine-grained video
annotation. Your task is to generate precise text descriptions for multiple
target video segments based on the input video.
You will be given the following inputs:
- 'video': the full original video
- 'segments': a list of target segment time intervals, <target_gt>; the 'i'-th
  segment corresponds to '[si, ei]'
- 'global_query': the shared description, <shared_global_query>, which applies
  to all target segments
Generate one 'global_rejection_query' and one corresponding
'global_rejection_reason'.

Objective: Create a description that is semantically close to 'global_query' but
does not match any of the target segments.

Annotation Guidelines:
- The 'global_rejection_query' must be derived from 'global_query'.
- Introduce one or a small number of key deviations from 'global_query'.
- The deviation must be natural, plausible, and visually meaningful.
- Do not use absurd, contradictory, or intentionally unreasonable changes.
- The resulting description must not match any of the given target segments.

Acceptable Deviation Types include changing the action direction or action
result, changing the key object, changing the scene condition, or changing the
interaction relationship.

Reason Requirement: Generate 'global_rejection_reason' to explain why
'global_rejection_query' does not match any segment in the video. The reason
must be concise, specific, and based on observable mismatch.
```

E.1.4 QUALITY CONTROL

We apply two automatic quality-control checks before training. First, each annotated segment is cropped and sent to an MLLM to verify whether the clip matches the query. Second, the complement of the annotated segment is checked for leakage, i.e., whether the same event appears outside the labeled windows.

Segment-query consistency prompt.

```
Given the video clip extracted from a long video and a query, determine how well
the video clip matches the query with a score.

Focusing solely on whether the described objects, motions, and relative order of
sub-events (if any) appear in this video clip. Do NOT consider where or how many
times these query sub-events occur in the original long video.

Scoring Criteria:
1. Exact Match (score = 1.0): The video clip fully depicts all the content
   described in the query.
2. Partial Match (0 < score < 1.0): The clip contains only some of the
   sub-events or details mentioned in the query.
3. No Match or Inconsistency (score = 0.0): The clip does not contain any of the
   described content, or the visual content contradicts the query.

Query: "{sentence_query}"

Provide a single score between 0 and 1 representing the degree of alignment.
And output a sentence describing the clip and the length of clip. The format is:
Score: <a number between 0 and 1>
Description: <a sentence>
```

Leakage-detection prompt.

```
Given a video clip and a query, compute the total amount of time (in seconds)
```

```
during which the described event occurs in this video clip.

The video clip is constructed by removing an interval from the original video:
- Original video duration: {duration} seconds
- Removed segment: [{start}, {end}] seconds
- Evaluated clip: segments [0, {start}) and ({end}, {duration}] concatenated
  together
- Current video duration: {new_duration} seconds

Task:
- Examine the concatenated video clip and identify all temporal intervals where
  the query event occurs.
- Sum up the durations (in seconds) of these detected intervals to get a total
  occurrence time.

Formatting requirements:
- If the query event does not occur at all in the video clip, return `0.0`.
- If the event occurs multiple times, sum up each occurrence's duration and
  return the total as a floating-point number.

Query: "{sentence_query}"

Provide the result as a single number indicating the total seconds of
query-event occurrence.
```

E.2 Prompts for CoT Data Construction

CoT sub-query construction. For positive examples, we generate a segment-specific query for each ground-truth segment. The generated sub-query preserves the global query as much as possible and adds only directly observable details needed to distinguish one occurrence from the others. If visual differences are insufficient, ordinal expressions such as "for the first time" are used.

```
For each target segment, inspect the visual content and rewrite `global_query`
into a segment-specific description, `segment_queries[i]`.

Objective: Generate one description for each segment that uniquely identifies
that segment among the target segments.

Annotation Guidelines:
- Each `segment_queries[i]` must be derived from `global_query`.
- Preserve the original wording and sentence structure of `global_query` as much
  as possible.
- Add only the minimum directly observable visual information needed to
  distinguish the segment from the other target segments.
- Include only information that is directly visible in the segment.
- Do not add inferred, subjective, causal, or non-visible information.
- Do not rewrite more than necessary.

Fallback Rule:
If some target segments cannot be reliably distinguished by direct observable
visual differences alone, use ordinal expressions such as "for the first time"
or "for the second time" according to chronological order.
```

The final supervised response uses the structured schema:

> <think> I need to localize the moments relevant to {QUERY}. The corresponding segments are as follows. {SUB_QUERY_1} This occurs from {START_1} to {END_1} seconds. ... </think><answer>[[START_1s, END_1s], ...]</answer>

For zero-segment examples, the rejection reason generated by the annotation pipeline is used as the reasoning content, and the final answer is `[]`.

E.3 Training and Test Prompt for EGO

During both training and evaluation of EGO, we use the following prompt:

```
To accurately pinpoint the event '{query}' in the video, determine the precise
time periods of the event.
The event may appear multiple times or zero times in the video.

Output your thought process within the <think> </think> tags, including analysis
of the video timeline with specific time ranges.

Then, provide the list of start and end times within the <answer> </answer> tags.
The time format MUST be precise to two decimal places and end with 's'
(e.g., xx.xxs). The output format should be '[[start1, end1], [start2, end2]]'.
For example: [[12.54s, 17.83s], [25.10s, 30.50s]] or [] if the event does not
appear.
```

E.4 Test Prompts for Other Baselines

For each baseline, we keep its original output style as much as possible and minimally modify the prompt to allow zero, single, or multiple segments.

MUSEG prompt.

```
You are given a video containing a series of actions. Watch the video carefully
and find all the visual events belonging to the action category: {query} The
event might appear multiple times or 0 times. First, output reasoning process in
<think> </think> tags. The reasoning process must REFER TO SPECIFIC TIMESTAMPS
TO TELL WHERE YOU GET THE INFORMATION FROM THE VIDEO. Then summarize your
reasoning process above and output selected segments like
"<answer>X.XX-X.XX</answer>". If there are multiple segments, separate them with
spaces like "<answer>X.XX-X.XX X.XX-X.XX</answer>". If the event does not
appear, output "<answer> </answer>". Your output format should be like
"<think>...</think><answer>...</answer>".
```

Time-R1 prompt.

```
To accurately pinpoint the event '{query}' in the video, determine the precise
time period of the event. The event might appear multiple times or 0 times.

Output your thought process within the <think> </think> tags, including analysis
with specific time ranges (xx.xx to xx.xx) in <timestep> </timestep> tags.

Then, provide the start and end times (in seconds, precise to two decimal
places) in the format 'start time to end time' within the <answer> </answer>
tags. If there are multiple segments, separate them with comma like
"<answer>X.XX to X.XX, X.XX to X.XX</answer>". If the event does not appear,
output "<answer> </answer>".
```

VideoChat-R1.5 prompts.

```
To accurately pinpoint the event "{query}" in the video, determine the precise time
period of the event. The event might appear multiple times or 0 times.

Provide the start and end times (in seconds, precise to two decimal places) in the format
of [(s1, e1), (s2, e2), ...] within the <answer> </answer> tags. For example: <answer>
[(12.54, 17.83)]</answer> or <answer>[]</answer> if the event does not appear.
```

TimeLens prompts.

```
Please find all visual events described by the sentence '{query}', determining their
starting and ending times. The format should be: 'The events happen in
[[<start time>, <end time>], ...] seconds'.
If the event does not occur, return '[]'.",
```

TimeLens2 prompts.

```
You are given a video.
Task: temporal retrieval.
Given the query: "{query}", return ALL time spans (in seconds) where the query is relevant.
Output format MUST be a JSON array of [start, end] pairs, e.g. [[0, 3.5], [10, 12]].
```