GROUNDING ALL MOMENTS IN LONG VIDEOS VIA RETRIEVAL, RERANKING AND REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Video temporal grounding aims to localize video segments that match a language query. However, existing methods typically assume a single occurrence per query, which is inadequate for long videos where the same event may appear multiple times. Existing benchmarks are also limited in this setting, as repeated valid moments are often missing from the annotations. To address this problem, we introduce **AO-Bench**, a human-refined benchmark for arbitrary-occurrence long video temporal grounding, where the goal is to localize all temporal spans that satisfy a query. We further propose a three-stage framework that **retrieves** candidate clips, **reranks** them, and **refines** their temporal boundaries. Our analysis shows that reranking is the key bottleneck, as it requires distinguishing true query-matching events from visually similar distractors. Therefore, we train the reranking module on **AOVTG-66K** generated by a dynamic multi-agent consensus process, in which multiple MLLM agents independently assess candidate clips, exchange feedback, and iteratively revise their decisions, while weakly supported or inconsistent predictions are filtered out to produce robust consensus labels. Experiments on AO-Bench show that **ReGround** achieves leading multi-occurrence grounding performance while remaining competitive in the single-occurrence setting, with lower inference cost than the multi-agent pipeline.

1 INTRODUCTION

Video temporal grounding (VTG) aims to localize the video segments that correspond to a natural-language query. It is a core capability for long-video understanding and retrieval (Fu et al., 2025; Wu et al., 2024; Zhou et al., 2025). Recent progress in multimodal large language models (MLLMs) (Bai et al., 2025a;b; Wang et al., 2024a) and video representation learning (Bolya et al., 2025; Tschannen et al., 2025; Wang et al., 2022; 2025) has substantially improved VTG performance. However, current VTG benchmarks still largely evaluate a simplified setting in which each query is assumed to correspond to a single temporal instance in a video. In realistic long videos, this assumption often fails: the queried event may recur multiple times. Consequently, models that correctly retrieve multiple valid occurrences can be penalized by evaluation protocols recognizing only one instance.

One response to this mismatch is to rewrite queries with restrictive modifiers so that each query refers to exactly one occurrence (Zhang et al., 2025b). Although this reduces ambiguity, it changes the task definition and moves away from the generic queries that users naturally issue in practice. We instead study a more realistic setting, which we call **arbitrary-occurrence long video temporal grounding**: given a natural-language query, the goal is to localize *all* valid occurrences of the queried event in a long video. This setting requires both high recall over repeated events and high precision in rejecting temporally or semantically similar distractors.

To support this setting, we construct **AO-Bench**, a benchmark designed for arbitrary-occurrence grounding. Starting from existing VTG data, we perform a human-in-the-loop auditing and refinement process to identify and correct annotation issues that are particularly harmful in the multi-occurrence setting, including missing event instances, boundary misalignment, and ambiguous query-video pairs. The resulting benchmark contains two subsets with complementary characteristics: **Ego4D-AO**, which emphasizes highly repetitive egocentric activities, and **MomentSeeker-AO**, which covers more diverse domains, including egocentric and exocentric videos from movies, sports, and surveillance.

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.

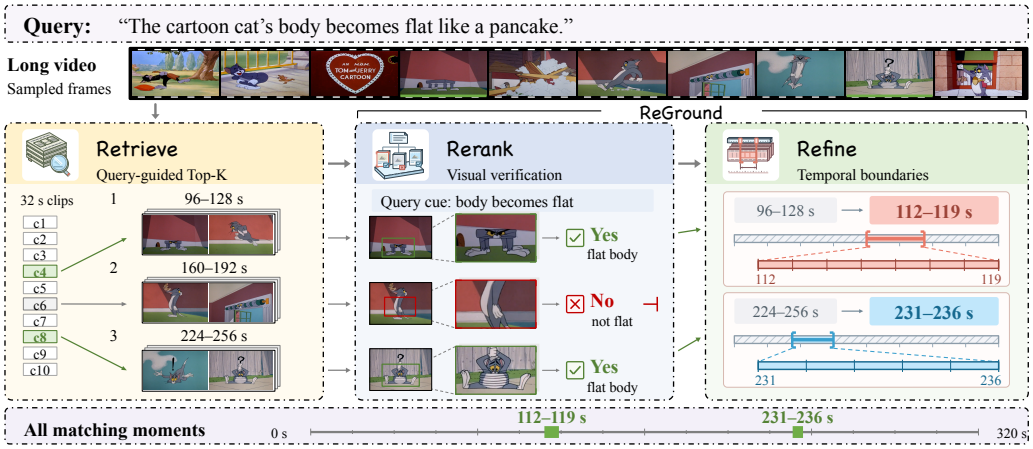


Figure 1: **Illustration of the Retrieve-Rerank-Refine paradigm.** An input long video is first sliced into consecutive clips. The retrieval stage selects candidate clips, and the reranking stage evaluates their relevance to the query through binary "Yes/No" decisions. The refinement stage then predicts temporal boundaries within the retained clips.

Arbitrary-occurrence grounding over hour-long videos is challenging because the system must identify a high-recall candidate set while avoiding a large number of hard negatives. We address this challenge with a **3R framework (Retrieve, Rerank, and Refine)**, as shown in Figure 1. In the retrieval stage, the model scans the video and proposes candidate clips with high recall. In the refinement stage, it adjusts the temporal boundaries of selected clips. Our analysis indicates that the main failure mode lies between these two steps: existing systems often retrieve plausible candidates but fail to reliably distinguish true matches from visually confusing negatives. We therefore focus on the intermediate reranking stage and propose **multi-agent consensus reranking (MACR)**, which uses multiple MLLM agents to evaluate candidate clips and aggregates their judgments through multi-round interaction, including peer reflection and an agent-expulsion mechanism, to suppress inconsistent or unsupported predictions. The output is a smaller set of candidates with substantially improved precision, which is then passed to the boundary refinement stage.

While adopting multi-agent consensus improves reranking quality, iteratively invoking multiple agents incurs prohibitive inference latency. To improve inference efficiency, we synthesize a large-scale instruction-tuning dataset (**AOVTG-66K**) and fine-tune an open-source MLLM to obtain **ReGround**, which jointly performs reranking and boundary refinement in a single forward pass. In the reranking comparison, the distilled model exceeds MACR on most reported metrics while avoiding iterative multi-agent inference. On AO-Bench, it achieves leading multi-occurrence grounding performance while remaining competitive in the single-occurrence setting.

2 Towards Reliable Arbitrary-Occurrence VTG Data

2.1 Problem Formulation

In arbitrary-occurrence video temporal grounding, a query may match multiple valid instances in one video. Given an untrimmed video $\mathcal{V}$ and a natural-language query $\mathcal{Q}$, the goal is to predict a set of temporal segments $\mathcal{M} = \{M_1, M_2, \ldots, M_S\}$, where each $M_i = (t_i^s, t_i^e)$ denotes one valid occurrence of the queried event, and $S \geq 1$ is not known in advance. Unlike standard single-moment grounding, this formulation requires recovering all valid occurrences while avoiding spurious matches.

2.2 Benchmark Construction

Existing long-video grounding datasets often contain annotation issues that worsen in the arbitrary-occurrence setting, including missing repeated instances, imprecise boundaries, and queries with

insufficient visual support. To address these issues, we refine the datasets through a human-in-the-loop auditing pipeline while preserving the original semantic intent whenever possible.

Query criteria. We retain only queries that support unambiguous visual grounding. Specifically, each query must satisfy three conditions. First, the query must be semantically aligned with the visual content, so that the described event can be matched to the target segment without avoidable ambiguity. Second, the event frequency must remain moderate: we exclude events that occur so frequently, or occupy so much of the video, that the task degenerates into coarse presence detection rather than temporal localization. Third, the event must be visually observable. Queries whose resolution depends primarily on audio, dialogue, or off-screen inference are removed.

Boundary annotation rules. Temporal boundaries are annotated according to query semantics. For static entities, the annotated segment spans the continuous interval during which the queried entity is visibly present. For dynamic actions, the segment spans the execution of the action itself, excluding irrelevant lead-in and follow-through frames. For compositional queries, the boundaries follow the logical structure of the description: conjunctions require the interval in which all specified conditions hold simultaneously, whereas disjunctions admit multiple valid segments, each annotated separately.

Auditing pipeline. Annotators first scan each video to understand its global event distribution and the context of the original annotation. They then review the video more carefully to identify and correct errors according to a fixed taxonomy (qualitative examples are in Appendix B.5):

- **Boundary error:** the query is valid, but the windows are loose or shifted; they are corrected.
- **Query ambiguity:** the query inaccurately describes the visible content; its wording is minimally revised, with boundaries adjusted if needed.
- **Incomplete occurrences:** some valid instances are unannotated; annotators search the full video and add all missing occurrences.
- **Unobservable event:** the query relies on non-visual or unreliable visual evidence and is removed.
- **Ill-posed event frequency:** the event is too frequent or brief for meaningful temporal grounding; the query is removed.

We use a dual-annotator setup throughout the re-audit. Inter-annotator agreement for error classification and temporal boundary annotation is reported in Appendix B.3.

Source datasets. We apply this pipeline to two long-video grounding benchmarks: Ego4D-NLQ (Grauman et al., 2022) and MomentSeeker (Yuan et al., 2026). Ego4D-NLQ captures repetitive daily activities in unconstrained egocentric videos, making it well suited for studying recurring events. MomentSeeker covers broader domains, including egocentric and exocentric videos from movies, sports, and surveillance. Together, they support arbitrary-occurrence grounding evaluation across repetitive and heterogeneous long-video settings.

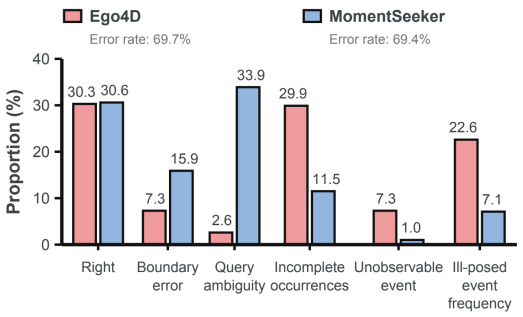


Figure 2: **Distribution of annotation error types.** Comparison of error proportions in Ego4D and MomentSeeker datasets.

2.3 Empirical Analysis

Although the source datasets are human-annotated and widely used, our audit reveals limitations in their single-occurrence annotations. By applying our auditing pipeline, we introduce **AO-Bench**, a unified arbitrary-occurrence evaluation suite comprising two high-fidelity subsets: **Ego4D-AO** and **MomentSeeker-AO**. Summary statistics are provided in Table 1, with further details in Appendix B.2.

Error statistics and analysis. Figure 2 shows high annotation error rates in both source datasets, with dominant error types reflecting their characteristics. Repetitive daily routines in Ego4D-NLQ

Table 1: Detailed statistics of the original datasets and our proposed arbitrary-occurrence benchmarks. "-SO" denotes the single-occurrence subsets, while "-MO" denotes the multi-occurrence subsets. All averages are query-weighted; overall video durations are computed from rounded SO/MO subset means. The last column reports total annotated duration per query.

Benchmark	# Videos	Avg. Video Duration (s)	# Queries	Avg. Windows per Query	Avg. Total Dur. / Query (s)
Ego4D-NLQ	415	499.75	4552	1.00	10.75
Ego4D-AO	134	504.47	1056	1.91	20.91
⌞ *Ego4D-AO-SO*	125	503.57	605	1.00	10.35
⌞ *Ego4D-AO-MO*	117	505.67	451	3.13	35.08
MomentSeeker(>= 128s)	187	664.88	837	1.00	14.32
MomentSeeker-AO	183	720.70	761	1.33	16.84
⌞ *MomentSeeker-AO-SO*	168	686.59	657	1.00	14.02
⌞ *MomentSeeker-AO-MO*	57	936.15	104	3.41	34.62
AO-Bench	**317**	**595.03**	**1817**	**1.67**	**19.21**
⌞ *AO-Bench-SO*	293	598.85	1262	1.00	12.26
⌞ *AO-Bench-MO*	174	586.34	555	3.18	34.99

often lead to *incomplete occurrences*, whereas diverse narratives in MomentSeeker contribute to *query ambiguity*. Further details are provided in Appendix B.

Crucially, if left unaddressed, these pervasive inaccuracies fundamentally distort evaluation metrics. Advanced models capable of retrieving exhaustive multi-occurrences are unfairly penalized, as their correct predictions are erroneously counted as false positives against incomplete ground truths. Our refined AO-Bench rectifies this "false penalty" phenomenon, providing a trustworthy metric to guide the future trajectory of long-video understanding research.

3 METHODOLOGY

We address arbitrary-occurrence video temporal grounding in long untrimmed videos with a **3R framework** (retrieve, rerank, refine). The pipeline retrieves a small candidate set, filters hard negatives, and predicts precise temporal boundaries for retained clips. Our analysis identifies reranking as the main bottleneck (Sec. 4.3). We therefore introduce multi-agent consensus reranking (MACR) and train a single efficient reranker on its relevance labels and temporal windows for scalable inference.

3.1 THE RETRIEVE-RERANK-REFINE PARADIGM

Directly grounding all occurrences of a query in a long untrimmed video is difficult because relevant events may be sparse, repeated, and visually confusable. To better characterize this setting and identify where existing systems fail, we decompose the task into three sequential stages.

Given an untrimmed video $\mathcal{V}$ and a query $\mathcal{Q}$, we extract N overlapping clips from $\mathcal{V}$,

$$\mathcal{C} = \{c_1, c_2, \ldots, c_N\}.$$

The pipeline sequentially applies a retriever $\mathcal{M}_{\text{ret}}$, a reranker $\mathcal{M}_{\text{rank}}$, and a boundary refiner $\mathcal{M}_{\text{ref}}$.

Retrieval. The retrieval stage selects a high-recall candidate set from the full clip collection $\mathcal{C}$. Given $\mathcal{Q}$, the retriever assigns each clip c_i a coarse relevance score, and the top-K clips are retained:

$$\mathcal{C}_{\text{ret}} = \text{top-k}\big(\mathcal{M}_{\text{ret}}(\mathcal{C}, \mathcal{Q})\big).$$

We consider three retrieval instantiations: (i) *embedding-based retrieval* ranks clips using cross-modal similarity between clip and query representations; (ii) *traditional VTG models* produce dense temporal proposals with confidence scores, from which each coarse clip inherits the highest proposal score contained within its span; (iii) *MLLM-based VTG models* are used with an iterative prediction-and-masking strategy: after each predicted segment is obtained, the corresponding region is masked and the model is prompted again to produce additional candidates.

Reranking. The retrieved set $\mathcal{C}_{\text{ret}}$ typically contains most true occurrences, but also includes visually similar hard negatives. The reranking stage aims to improve precision by rejecting hard negatives

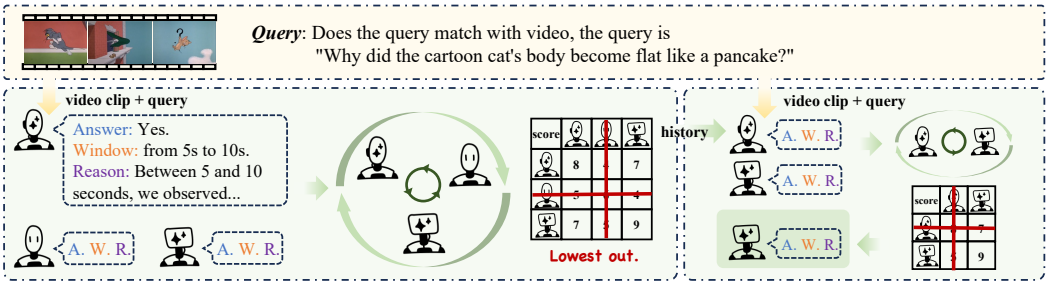


Figure 3: **Illustration of the Multi-Agent Consensus Reranking (MACR) process.** Given a candidate video clip and query, agents independently generate answers, temporal windows, and reasoning, then score each other's responses in a peer-review matrix. The lowest-scoring agent is removed, and the remaining agents iteratively refine their outputs to reach a final consensus.

while retaining valid candidates. For each $c_i \in \mathcal{C}_{\text{ret}}$, the reranker predicts whether the queried event is genuinely present:

$$\mathcal{C}_{\text{rank}} = \{\, c_i \in \mathcal{C}_{\text{ret}} \mid \mathcal{M}_{\text{rank}}(c_i, \mathcal{Q}) = 1 \,\}.$$

We instantiate the reranking model exclusively using *MLLM-based rerankers*, framing the decision as a binary video question-answering task. Specifically, the model is prompted to explicitly answer whether the queried event occurs in the candidate clip. We evaluate proprietary MLLMs (Gemini-3-Flash and GPT-4.1) and open-weight MLLMs (Qwen3-VL-8B and Qwen3-VL-32B) as individual rerankers, alongside MACR, two simpler aggregation strategies, and the distilled ReGround (Table 4).

Refinement. To recover the precise temporal boundaries, each surviving clip is then processed with a boundary refinement model. For every $c_i \in \mathcal{C}_{\text{rank}}$, the refiner predicts one or more grounded segments corresponding to the query within that clip. The final output is:

$$\mathcal{M} = \bigcup_{c_i \in \mathcal{C}_{\text{rank}}} \mathcal{M}_{\text{ref}}(c_i, \mathcal{Q}).$$

The refinement module can be instantiated with traditional temporal grounding models, open-weight multimodal language models, or proprietary MLLMs.

Our analysis of this diagnostic pipeline (detailed in Sec. 4.3) identifies reranking as the main bottleneck. Reranking errors include both retaining hard negatives and rejecting valid candidates: the former introduces spurious matches, while the latter removes true occurrences that subsequent refinement cannot recover. These complementary failure modes motivate the reranking framework introduced next.

3.2 Multi-Agent Consensus Reranking

To improve the reranking accuracy, we propose **Multi-Agent Consensus Reranking (MACR)**, a committee-based framework in which multiple multimodal large language model (MLLM) agents jointly assess whether a retrieved clip contains the queried event. MACR is motivated by the observation that single-agent reranking is often brittle in the presence of visually similar hard negatives. Instead of relying on one model response, MACR aggregates structured judgments from multiple agents through iterative verification, peer scoring, and committee pruning.

Given a candidate clip $c \in \mathcal{C}_{\text{ret}}$ and a query $\mathcal{Q}$, MACR initializes a committee of L MLLM agents, $\mathcal{A}^{(0)} = \{A_1, A_2, \ldots, A_L\}$. At each round t, every active agent produces a structured assessment of the clip. The agents then score one another's responses, and the lowest-scoring agent is removed. This process repeats until a final consensus decision is obtained.

Phase 1: structured action and attribute verification. Each agent first evaluates the candidate clip independently under a grounding-oriented prompt that instructs the agent to decompose the query into its essential semantic attributes, such as actions, objects, participants, or scene conditions, and to verify these attributes only from visible evidence in the clip.

For agent $A_l \in \mathcal{A}^{(t)}$, the round-$t$ output is:

$$O_l^{(t)} = \big(r_l^{(t)},\, y_l^{(t)},\, w_l^{(t)}\big),$$

where $r_l^{(t)}$ is a textual rationale describing attribute-level evidence, $y_l^{(t)} \in \{0, 1\}$ is a binary decision indicating whether the query is satisfied, and $w_l^{(t)}$ is a set of temporal evidence windows cited by the agent. The full response set at round t is denoted by $\mathcal{O}^{(t)} = \{O_l^{(t)} \mid A_l \in \mathcal{A}^{(t)}\}$.

Phase 2: peer scoring and committee pruning. After all agents produce their responses, MACR performs a peer-scoring step to estimate the reliability of each agent. Conditioned on the clip c, the query $\mathcal{Q}$, and the response set $\mathcal{O}^{(t)}$, every agent $A_i \in \mathcal{A}^{(t)}$ evaluates each response $O_j^{(t)}$ using a scoring prompt, yielding a scalar score $s_{i\to j}^{(t)}$. The aggregated reliability score of agent A_j is then computed by summing up all the scalar scores it receives from the committee members, and the agent with the lowest aggregated score is subsequently removed from the committee:

$$S^{(t)}(A_j) = \sum_{A_i \in \mathcal{A}^{(t)}} s_{i\to j}^{(t)}, \qquad \mathcal{A}^{(t+1)} = \mathcal{A}^{(t)} \setminus \left\{ \arg\min_{A_j \in \mathcal{A}^{(t)}} S^{(t)}(A_j) \right\}$$

Phase 3: iterative refinement. The remaining agents are then given the previous round's reasoning context, including peer rationales and scores, and are prompted to revise their judgments. Formally, for each surviving agent $A_l \in \mathcal{A}^{(t+1)}$, MACR generates an updated response $O_l^{(t+1)}$ conditioned on the clip, query, and historical context from round t. In this way, later rounds incorporate both independent visual verification and feedback from the committee.

The verification–scoring–pruning cycle repeats until only one agent remains. The final surviving agent's decision is used as the reranking output. Compared with single-agent reranking, MACR reduces sensitivity to individual hallucinations and encourages decisions that are better grounded in shared visual evidence. However, this iterative process introduces substantial inference cost, which motivates the distillation strategy described next.

3.3 Distilling Consensus into a Unified Reranker

MACR's iterative multi-agent inference introduces substantial latency. We fine-tune the student on the committee-generated relevance labels and temporal windows. This transfers the committee's outputs to a single open-source MLLM for efficient inference.

Data synthesis. For a given text query $\mathcal{Q}$, we first employ our retrieval model to extract the top-K candidate clips $\{c_k\}_{k=1}^{K}$. We then process each $(c_k, \mathcal{Q})$ pair through the MACR pipeline. The final consensus reached by the surviving agents provides a binary relevance label $y^* \in \{0, 1\}$ and a set of refined temporal windows w^*. By applying this process across a large set of queries, we construct a high-quality instruction-tuning dataset $\mathcal{D}_{\text{distill}} = \{(c_i, \mathcal{Q}, y_i^*, w_i^*)\}_{i=1}^{N}$. Strict video ID checks ensure that AOVTG-66K and AO-Bench share no source videos.

Distillation process. We frame the distillation as supervised fine-tuning (SFT) on a base MLLM Θ_{student}. Given a clip c_i and a query $\mathcal{Q}$, Θ_{student} is trained to predict the MACR-derived binary decision y_i^* and the set of temporal windows w_i^*:

$$(\hat{y}_i, \hat{w}_i) = \Theta_{\text{student}}(c_i, \mathcal{Q}), \qquad \hat{y}_i \in \{0, 1\}, \qquad \hat{w}_i \;\; = \big\{(\hat{t}_{i,k}^{s}, \hat{t}_{i,k}^{e})\big\}_{k=1}^{K_i}.$$

Here, K_i is the number of predicted windows in clip c_i. For $\hat{y}_i = 0$, we set $K_i = 0$ and $\hat{w}_i = \emptyset$; for $\hat{y}_i = 1$, $K_i \geq 1$. This training enables the student to perform binary relevance classification and multi-window temporal localization in a single forward pass. Distilling the committee's decisions into one model reduces inference time (see Table 9 in the Appendix).

4 Experiments

In this section, we first compare the overall grounding performance of our framework with existing temporal grounding models and proprietary MLLMs. Subsequently, we conduct a decoupled evaluation of the three individual stages to identify the main bottlenecks. Finally, we evaluate existing

Table 2: **Overall grounding performance.** SO and MO denote the single- and multi-occurrence evaluation settings, respectively. SO is evaluated using R1@0.3, R1@0.5, R1@0.7, and mIoU. MO is evaluated using set-level mIoU, mIoP, mIoG, and R1@0.3, together with instance-level matching metrics. All metrics are reported in %. Best results are in **bold**, and second-best results are underlined.

Method	AO-Bench-SO				AO-Bench-MO						
					Set-level				*Instance-level* (IoU=0.3)		
	R1@0.3	R1@0.5	R1@0.7	mIoU	mIoU	mIoP	mIoG	R1@0.3	Prec.	Rec.	F1
Video Temporal Grounding Models											
UniversalVTG (An et al., 2026)	29.20	20.29	12.41	21.43	15.72	39.21	21.53	19.64	42.00	15.18	21.74
UniTime (Li et al., 2026b)	36.53	22.94	11.67	24.43	11.91	42.21	14.19	12.36	38.55	14.12	20.15
TimeLens (Zhang et al., 2025b)	36.90	25.73	13.46	26.11	14.79	33.95	20.11	19.13	34.84	12.88	18.38
TimeLens2 (Zhu et al., 2026)	**43.86**	**30.09**	**16.07**	**30.67**	19.59	42.51	27.60	26.71	**43.48**	17.05	23.51
Proprietary Models											
Doubao-Seed-1.6-vision-0815 (ByteDance, 2025)	35.00	22.50	12.50	25.92	17.47	32.48	28.65	20.63	39.55	21.15	25.77
Gemini-3-Flash (Google, 2025)	38.33	25.83	10.00	26.30	13.72	23.15	25.80	17.46	15.71	18.26	15.80
GPT-4.1 (OpenAI, 2025)	17.50	7.50	3.33	11.88	9.34	14.36	21.15	6.35	11.51	5.69	7.14
ReGround	41.73	25.02	12.31	27.76	**21.51**	**45.81**	**30.83**	**30.36**	40.44	**27.06**	**29.89**

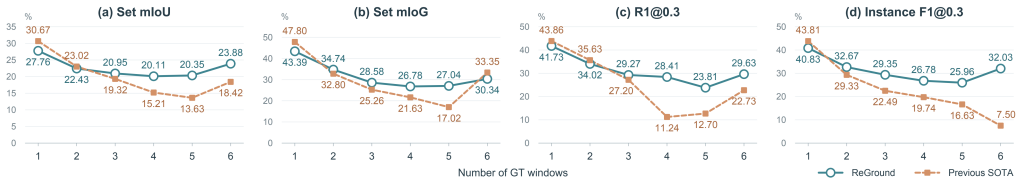


Figure 4: **Grounding performance by occurrence count.** Comparison of ReGround and TimeLens2 ("Previous SOTA" in the legend). The horizontal axis denotes the number of ground-truth windows per sample, with one window corresponding to SO and multiple windows to MO. From left to right, the panels report set-level mIoU, set-level mIoG, R1@0.3, and instance-level F1@0.3, all in %.

reranking models and conduct ablation studies of our MACR framework. Ablation studies on the choice of retrieval and grounding models are provided in Appendices C.2 and C.3, respectively. A full-data ablation of teacher supervision is provided in Appendix C.1.

4.1 Experimental Setups

Datasets. To address arbitrary-occurrence video temporal grounding, we curate a diverse subset from existing corpora (Grauman et al., 2022; Regneri et al., 2013; Tang et al., 2019; Zala et al., 2023; Zhou et al., 2018) spanning varied scenes, durations, and query semantics. We process this subset with MACR to construct **AOVTG-66K**, an instruction-tuning dataset for reranking and temporal grounding. Detailed curation procedures and dataset statistics are provided in Appendix D. We evaluate on AO-Bench, comprising Ego4D-AO and MomentSeeker-AO, and report results separately under the single-occurrence (SO) and multi-occurrence (MO) settings.

Implementation details. Our PyTorch implementation uses `Qwen3-VL-8B` (Bai et al., 2025a) as the base model. We fine-tune ReGround for one epoch using AdamW (Loshchilov & Hutter, 2017) with batch size 8, learning rate 10^{-5}, and linear warmup over the first 3% of steps. The vision encoder is frozen, while the LLM is adapted with LoRA (Hu et al., 2022) (rank = 64, alpha = 16). By default, UniTime retrieves the top-5 32-second clips for reranking by MACR or the distilled ReGround. MACR uses Gemini-3-Flash, GPT-4.1, and Qwen3-VL-32B. See Appendix A.1 for inference details.

Evaluation metrics. For *set-level evaluation*, we represent the unions of predicted and ground-truth windows as two binary temporal masks. We compute IoU as the ratio of their intersection duration to their union duration and report its mean across samples (mIoU). For *instance-level evaluation*, we apply strict one-to-one window matching at a temporal IoU threshold of 0.3. Full metric definitions are provided in Appendix A.2.

4.2 Main Results on Arbitrary-Occurrence Video Temporal Grounding

Evaluation results. Table 2 shows that **ReGround achieves stronger MO grounding while maintaining competitive SO performance.** In the SO setting, it ranks second in both R1@0.3 and

Table 3: **Oracle component analysis on Ego4D-AO.** Retrieval, reranking, and refinement are replaced with oracle modules individually and in combination. The four row groups correspond to the baseline, one oracle module, two oracle modules, and the all-oracle configuration. Named models indicate the non-oracle modules retained in each configuration. All scores are reported in %.

Retrieve	Rerank	Refine	R-Acc	R-Rec	C-Prec	GT-Rec	SO mIoU	MO mIoU	O-mIoU
UniTime	Qwen	UniTime	73.67	42.22	50.78	31.30	23.96	17.20	21.08
Oracle	Qwen	UniTime	36.38	36.30	100.00	36.32	34.32	18.16	27.41
UniTime	*Oracle*	UniTime	100.00	100.00	100.00	72.30	35.72	29.40	33.03
UniTime	Qwen	*Oracle*	73.67	42.22	50.78	31.30	32.09	32.61	32.31
UniTime	*Oracle*	*Oracle*	100.00	100.00	100.00	72.30	83.89	71.03	78.42
Oracle	Qwen	*Oracle*	36.38	36.30	100.00	36.32	37.34	34.09	35.95
Oracle	*Oracle*	UniTime	100.0	100.0	100.0	100.0	58.42	55.03	56.96
Oracle	*Oracle*	*Oracle*	100.0	100.0	100.0	100.0	100.0	100.0	100.0

mIoU, reaching 41.73% and 27.76%, respectively. In the MO setting, it achieves the best results in six of the seven reported metrics, including a set-level mIoU of 21.51% and R1@0.3 of 30.36%. The leading mIoP (45.81%) and mIoG (30.83%) indicate that ReGround better excludes irrelevant temporal content while covering more of the target events. It also recovers more distinct occurrences: instance-level recall reaches 27.06%, compared with 17.05% for TimeLens2 (Zhu et al., 2026), while F1 improves from 23.51% to 29.89%. These gains demonstrate more complete occurrence recovery with a better precision–recall balance.

Effect of occurrence count. Figure 4 compares the grounding performance of ReGround and TimeLens2 across different numbers of ground-truth windows. Both methods obtain lower scores in the MO setting relative to their respective SO results; however, ReGround exhibits smaller reductions and less variation across the evaluated MO subsets, particularly in R1@0.3 and instance-level F1. These results demonstrate that **ReGround achieves more balanced grounding performance, combining competitive SO results with superior MO localization and reduced sensitivity to variations in the number of ground-truth windows.**

4.3 Ablation study with oracle components

Experimental setup & metrics. We evaluate oracle substitutions on Ego4D-AO using UniTime for retrieval and refinement and Qwen for reranking (Table 3). *Oracle retrieval* supplies clips covering all annotated occurrences, *oracle reranking* uses ground-truth relevance labels, and *oracle refinement* provides ground-truth boundaries for retained positive clips. Retrieval and reranking use the same matching criterion: a candidate clip is positive if its IoG with at least one ground-truth window exceeds 0.3. Detailed experimental settings and metric definitions are provided in Appendix A.3.

Evaluation results. Single-module substitutions yield relatively close O-mIoU scores (27.41–33.03%), as errors in the remaining stages continue to affect the final predictions. With two oracle modules, retaining only the original reranker yields 35.95% O-mIoU, compared with 78.42% and 56.96% when retaining only the original retriever or refiner, respectively. These results identify **reranking as the main bottleneck**: valid candidates rejected at this stage cannot be recovered by subsequent refinement.

4.4 Reranking Performance and MACR Ablation

Evaluation metrics. Reranking uses the same IoG-based ground-truth labels as retrieval (Section 4.3). At the *clip level*, we report classification accuracy, precision, recall, and F1 to assess candidate selection. At the *system level*, we fix UniTime as the retriever and refiner and compare the resulting SO mIoU, MO mIoU, and MO instance-level F1. Aggregation settings are detailed in Appendix A.4.

Evaluation results. Table 4 shows that ReGround improves the precision–recall balance of existing rerankers, exceeding the best baseline by 6.34 percentage points in R-F1. MACR also outperforms both simpler aggregation strategies, improving SO and MO mIoU over Self-Consistency by 4.88 and 3.20 percentage points, respectively. **These results support MACR's advantage over simpler aggregation and demonstrate the distilled ReGround's effectiveness in candidate selection and downstream grounding.**

Table 4: **Reranking performance and MACR ablation on Ego4D-AO.** The three blocks report existing MLLM rerankers, MACR versus simpler aggregation, and the distilled ReGround, respectively. UniTime is fixed for retrieval and refinement. All scores are reported in %. Best and second-best results across all rows are shown in **bold** and underlined, respectively.

Reranking Strategy	R-Acc	R-Precision	R-Recall	R-F1	SO mIoU	MO mIoU	MO I-F1
Existing MLLM rerankers							
Gemini-3-Flash	70.42	45.70	60.28	51.99	22.70	18.09	27.16
GPT-4.1	69.74	45.42	**65.53**	52.97	22.56	19.19	28.28
Qwen3-VL-8B	73.76	50.87	42.39	46.24	24.00	17.29	25.25
Qwen3-VL-32B	73.21	49.81	55.02	52.28	23.47	19.41	28.04
MACR Ablation							
Self-Consistency	75.02	52.99	55.03	53.99	20.34	17.51	28.55
Plain Majority Voting	73.92	51.00	53.03	51.99	19.79	16.67	27.66
MACR (Ours)	76.56	55.71	58.46	57.05	25.22	20.71	**30.95**
ReGround (Ours)	**78.84**	**60.79**	57.89	**59.31**	**26.82**	**21.05**	30.43

5 Related Work

Long video temporal grounding benchmarks. Extensive efforts have been dedicated to VTG benchmarks across diverse domains, including cooking (TACoS; Regneri et al., 2013), movies (MAD; Soldan et al., 2022), egocentric activities (Ego4D-NLQ; Grauman et al., 2022), and multi-domain scenarios (MomentSeeker; Yuan et al., 2026). Despite their diversity, these datasets typically operate under a restrictive **single-occurrence assumption**. Recent studies by Zhang et al. (2025b) and Li et al. (2026b) highlight that omitting recurring events severely distorts evaluation. To mitigate this, TimeLens (Zhang et al., 2025b) employs query-rewriting with strict modifiers to enforce single-match constraints. However, such text manipulation deviates from generic real-world user behavior. Meanwhile, MultiHop-EgoQA (Chen et al., 2025) and QVHighlights (Lei et al., 2021) address grounded multi-hop video QA and query-based moment retrieval with highlight detection, respectively, but both operate on relatively short video clips. We focus on **arbitrary-occurrence grounding**, annotating all valid occurrences in long, multi-domain videos to evaluate both unique and recurring events.

Long video temporal grounding methods. Existing methods generally follow two paradigms. *Traditional architectures* have evolved from simple pooling and sliding windows (Lei et al., 2021; Ramakrishnan et al., 2023; Zhang et al., 2020) to context-aware coarse-to-fine pipelines (Hannan et al., 2024; Hou et al., 2023; Pan et al., 2023) and efficient linear-time state-space models (An & Grauman, 2025; An et al., 2026). In parallel, *MLLM-based methods* use cross-modal reasoning to predict boundaries via temporal tokens (Huang et al., 2024), timestamp tags (Chen et al., 2024; Zhang et al., 2025a), or regression heads (Pramanick et al., 2025). To address hour-long contexts, recent MLLMs adopt hierarchical retrieval (Li et al., 2026b) or recursive zooming mechanisms (Hannan et al., 2025). Despite progress, both paradigms remain biased toward single-occurrence localization, leaving a gap in multi-occurrence scenarios. We bridge this gap with a diagnostic **Retrieve-Rerank-Refine** paradigm to identify stage-wise bottlenecks and advance arbitrary-occurrence grounding.

6 Conclusion

In this paper, we study arbitrary-occurrence video temporal grounding in long videos. To address incomplete annotations in existing datasets, we introduce AO-Bench, refined through human auditing for exhaustive long-video evaluation. Our diagnostic Retrieve-Rerank-Refine paradigm identifies reranking as the critical performance bottleneck. We address it with Multi-Agent Consensus Reranking, a dynamic collaboration framework that effectively filters hallucinations. We distill MACR's decisions into ReGround, a single efficient MLLM that achieves leading multi-occurrence grounding performance while remaining competitive in the single-occurrence setting. Our benchmark and efficient distillation paradigm provide a rigorous foundation for advancing comprehensive and scalable long-video understanding.

AI USE STATEMENT

As part of the research methodology, we used AI to check whether source queries matched their annotated video segments when constructing the instruction-tuning dataset AOVTG-66K. We used AI in MACR for clip reranking and for generating the relevance labels and temporal windows used to train ReGround (Section 3 and Appendix D). We also used generative AI to improve the clarity and readability of the manuscript. The authors reviewed the AI-assisted manuscript edits for accuracy and take full responsibility for the final content of this work, including the datasets, manuscript, and reported results.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix B describe the query-selection criteria, temporal annotation rules, dual-annotator auditing procedure, and inter-annotator agreement for AO-Bench. Appendix D details the data sources, filtering, separation of training and evaluation videos, and supervision generation for AOVTG-66K. Section 3 describes the Retrieve-Rerank-Refine framework, MACR, and distillation into ReGround, with MACR prompts provided in Appendix E. Section 4.1 and Appendix A provide the training settings, inference procedures, and evaluation metrics. We plan to publicly release AOVTG-66K, AO-Bench and code upon acceptance.

REFERENCES

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

Joungbin An, Agrim Jain, and Kristen Grauman. Universalvtg: A universal and lightweight foundation model for video temporal grounding. *arXiv preprint arXiv:2604.08522*, 2026.

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

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

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

ByteDance. Seed1.6 tech introduction. https://seed.bytedance.com/en/seed1_6, 2025. Accessed: September 2025.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *AAAI*, volume 39, pp. 2159–2167, 2025.

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

Siran Chen, Qinglin Xu, Yue Ma, Yu Qiao, and Yali Wang. Attentive snippet prompting for video retrieval. *TMM*, 26:4348–4359, 2023.

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

Google. A new era of intelligence with gemini 3, 2025. URL https://blog.google/products/gemini/gemini-3.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pp. 18995–19012, 2022.

Tanveer Hannan, Md Mohaiminul Islam, Thomas Seidl, and Gedas Bertasius. Rgnet: A unified clip retrieval and grounding network for long videos. In *ECCV*, pp. 352–369. Springer, 2024.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. Revisionllm: Recursive vision-language model for temporal grounding in hour-long videos. In *CVPR*, pp. 19012–19022, 2025.

Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wk Chan, Chong-Wah Ngo, Mike Zheng Shou, and Nan Duan. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. In *ACL*, pp. 8013–8028, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *ECCV*, pp. 202–218. Springer, 2024.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *NeurIPS*, 34:11846–11858, 2021.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026a.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. *Advances in Neural Information Processing Systems*, 38:64426–64455, 2026b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

OpenAI. Gpt-4.1 system card. https://openai.com/index/gpt-4-1/, 2025.

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. In *ICCV*, pp. 13767–13777, 2023.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal llms. In *ICCV*, pp. 24297–24308, 2025.

Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Spotem: Efficient video search for episodic memory. In *ICML*, pp. 28618–28636. PMLR, 2023.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *TACL*, 2013.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *CVPR*, pp. 5026–5035, 2022.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, 2019.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

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

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

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 37:28828–28857, 2024.

Huaying Yuan, Jian Ni, Zheng Liu, Yueze Wang, Junjie Zhou, Zhengyang Liang, Bo Zhao, Zhao Cao, Ji-Rong Wen, and Zhicheng Dou. Momentseeker: A task-oriented benchmark for long-video moment retrieval. *Advances in Neural Information Processing Systems*, 38, 2026.

Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *CVPR*, 2023.

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

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, pp. 6543–6554, 2020.

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

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *CVPR*, pp. 13691–13701, 2025.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.

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.

A Experimental Details

A.1 Implementation Details for Main Results

Inference strategies. To fairly evaluate arbitrary-occurrence grounding, we formalize the inference mechanisms as follows: (1) For traditional *temporal grounding models*, we adapt them to output multiple windows by employing Top-K proposal selection or utilizing confidence scores as thresholding filters (τ). (2) For *proprietary MLLMs*, we prompt them to directly generate multiple temporal windows in an end-to-end manner. (3) In contrast, *our approach* adopts a dedicated three-stage **Retrieve-Rerank-Refine** pipeline. Specifically, we first segment videos into 32-second clips and retrieve the top-5 candidates via UniTime. For the training-free variant, we leverage MACR (powered by advanced agents including Gemini-3-Flash, GPT-4.1, and Qwen3-VL-32B) for reranking, and reuse UniTime to refine the final temporal boundaries. For the distilled variant, the retrieved candidates are fed into our proposed ReGround, which unifies the reranking and fine-grained grounding processes into a single efficient step.

A.2 Evaluation Metrics for Main Results

Single-occurrence evaluation. In the SO setting, we follow standard temporal grounding evaluation and report R1@τ for $\tau \in \{0.3, 0.5, 0.7\}$ and mIoU. Temporal IoU is the intersection duration divided by the union duration; R1@τ measures the fraction of samples localized at IoU $\geq \tau$, while mIoU averages IoU across samples. These metrics assess localization success at different overlap thresholds and overall temporal alignment.

Multi-occurrence set-level evaluation. We represent the unions of predicted and ground-truth windows as two binary temporal masks. For each sample, we normalize the intersection duration by the union, predicted, and ground-truth durations, then average these ratios across samples to obtain mIoU, mIoP, and mIoG, respectively. These metrics assess overall temporal alignment, temporal precision, and coverage. R1@0.3 measures the fraction of samples with set-level IoU ≥ 0.3.

Multi-occurrence instance-level evaluation. We match predicted and ground-truth windows at a temporal IoU threshold of 0.3 under a strict one-to-one constraint, excluding both windows in each accepted pair from further matching. Precision and recall quantify the proportions of matched predictions and ground-truth instances, respectively, while F1 is their harmonic mean. These metrics assess the ability to recover individual occurrences while limiting false positives and missed events, complementing duration-based set-level evaluation.

A.3 Experimental Details for Oracle Analysis

Experimental setup. We evaluate oracle substitutions on Ego4D-AO using UniTime for retrieval and refinement and Qwen for reranking as the baseline (Table 3). *Oracle retrieval* supplies candidate clips selected using ground-truth annotations to cover all target occurrences. *Oracle reranking* uses ground-truth relevance labels to retain positive candidates and reject negatives. *Oracle refinement* replaces temporal predictions for retained positive candidates with the corresponding ground-truth boundaries. The latter two stages operate on the candidates provided by preceding stages and cannot recover omitted occurrences. We replace one, two, or all three modules while retaining the baseline models for the remaining stages.

Evaluation metrics. Retrieval and reranking use the same matching criterion: a candidate clip is positive if its IoG with at least one ground-truth window exceeds 0.3, where IoG is the intersection duration divided by the ground-truth window duration.

Using these labels, R-Acc measures classification accuracy over input clips, and R-Rec measures the fraction of positive input clips retained by the reranker. C-Prec is the fraction of selected clips satisfying this criterion, while GT-Rec is the fraction of ground-truth windows matched by at least one selected clip. In Table 3, both are measured after reranking to assess the precision and occurrence coverage of the retained candidates. SO mIoU, MO mIoU, and O-mIoU average temporal IoU over the SO subset, the MO subset, and all evaluation samples, respectively, measuring final grounding quality.

A.4 Details of Reranking and MACR Ablation

We compare existing MLLM rerankers, MACR, simpler aggregation strategies, and the distilled ReGround in Table 4. UniTime is used for retrieval and refinement, and candidate relevance labels follow the IoG-based criterion in Section 4.3. MACR uses Gemini-3-Flash, Qwen3-VL-32B, and GPT-4.1 with the iterative consensus procedure described in Section 3.2. The simpler aggregation baselines are configured as follows.

Plain Majority Voting. We take the majority of the relevance decisions from Gemini-3-Flash, Qwen3-VL-32B, and GPT-4.1 for each candidate clip and query.

Self-Consistency. We run Gemini-3-Flash three times for each candidate clip and query and take the majority of its three relevance decisions.

B Details of AO-Bench

In this section, we provide a comprehensive overview of our proposed arbitrary-occurrence long video temporal grounding benchmark, **AO-Bench**. We detail the source datasets chosen for refinement, the final statistical distributions post-refinement (Section B.2), and the rigorous annotation interface and manual used to guarantee data quality.

B.1 Source Datasets

We construct AO-Bench by systematically auditing two of the most representative existing long video temporal grounding benchmarks, forming two distinct subsets:

- **Ego4D-AO:** Derived from the widely used Ego4D-NLQ (Grauman et al., 2022) dataset. Egocentric videos naturally capture unscripted, long-form daily human activities (e.g., cooking, cleaning, working). These routines are inherently repetitive, making this subset a perfect testbed for evaluating a model's ability to retrieve highly frequent, visually similar recurring events. Specifically, we build upon the validation set of Ego4D-NLQ v2. Because the complete validation split is excessively large and would result in prohibitively high evaluation costs, we uniformly sample 139 videos. All queries associated with these sampled videos are retained, yielding a total of 1,552 initial queries for our auditing pipeline.
- **MomentSeeker-AO:** Derived from the highly diverse MomentSeeker (Yuan et al., 2026) dataset. This subset encompasses a broad spectrum of domains, including exocentric movies, sports broadcasts, surveillance footage, and varied egocentric views. The diverse narrative structures and camera motions challenge models to perform arbitrary-occurrence grounding across significant domain shifts. While the original MomentSeeker dataset contains both short and long videos, our work focuses exclusively on long-video understanding. Therefore, we establish a strict duration lower bound of 128 seconds, collecting and annotating only the data from videos that exceed this threshold.

B.2 Detailed Statistics

Following our rigorous refinement pipeline, AO-Bench presents a significantly different distribution compared to its source datasets, accurately reflecting the arbitrary-occurrence nature of real-world videos. Detailed dataset statistics are summarized in Table 1. To better analyze model performance on events with different frequency properties, we further divide our refined benchmarks (Ego4D-AO and MomentSeeker-AO) into two sub-categories: the **-SO** (Single Occurrence) subset where the queried event happens exactly once, and the **-MO** (Multi Occurrence) subset where the event recurs multiple times within the same video.

Overall scale. Our refined AO-Bench consists of 317 untrimmed long videos (134 from Ego4D-AO and 183 from MomentSeeker-AO), with a query-weighted average duration of approximately 595.03s. Across these videos, the benchmark comprises 1,817 highly verified natural language queries (1,056 in Ego4D-AO and 761 in MomentSeeker-AO), including 1,262 in the SO subset (69.5%) and 555 in the MO subset (30.5%). The benchmark contains over 3,000 annotated temporal windows, with approximately 1.67 windows and 19.21s of total annotated duration per query.

The combined SO and MO subsets cover 293 and 174 videos, respectively. These video sets overlap because the same video can contribute queries to both subsets. Combined statistics are reported in Table 1; aggregate averages are query-weighted and computed from the rounded source statistics. Overall video durations for each source benchmark and AO-Bench are computed from the corresponding SO/MO subset means and query counts.

Arbitrary-Occurrence distribution. The statistics in Table 1 clearly highlight the severe single-occurrence bias in the original datasets, where the average number of windows per query was artificially constrained to exactly 1.00. In stark contrast, our refined benchmark reveals the true recurring nature of events in long videos:

- **Egocentric repetitiveness (Ego4D-AO):** Daily unscripted activities are highly repetitive. After our exhaustive auditing, the average windows per query nearly doubled to 1.91. Strikingly, 42.7% of the queries (451 out of 1056) belong to the multi-occurrence (*-MO*) subset, where a single queried event occurs an average of 3.13 times per video.
- **Exocentric & diverse domains (MomentSeeker-AO):** Even in more narrative or edited videos, recurring events are common. The average occurrences per query increased to 1.33. Within the *-MO* subset, events recur an average of 3.41 times. Interestingly, these multi-occurrence events tend to appear in significantly longer videos (averaging 936.15s), heavily challenging models to maintain stable retrieval capabilities over extended temporal contexts.

Furthermore, the average total annotated duration per query naturally increases in the multi-occurrence setting (e.g., jumping from 10.35s in *Ego4D-AO-SO* to 35.08s in *Ego4D-AO-MO*). This indicates that accurately grounding arbitrary occurrences requires models to comprehensively capture a much larger span of scattered temporal evidence, significantly elevating the task difficulty.

B.3 Inter-Annotator Agreement (IAA) on the Re-audit

To quantitatively validate our human-in-the-loop refinement, we employed a **dual-annotator setup** throughout the process. The Inter-Annotator Agreement (IAA) was directly computed between the two independent annotators, covering both the categorical error classification and the continuous temporal boundary annotation tasks. The results are summarized below:

Table 5: **Inter-annotator agreement on the re-audit.** Agreement is computed between the two independent annotators.

Annotation Task & Metric	**Ego4D-AO**	**MomentSeeker-AO**
Error Type Classification (Overall Agreement) (%)	82.5%	84.2%
Temporal Boundary Annotation (Human-Human tIoU) (%)	66.8%	69.4%

- **Error Type Classification:** Given the complexity of our fine-grained error taxonomy, achieving an absolute overall agreement exceeding 82% represents a strong and reliable consensus among annotators. This confirms that our annotation guidelines are clear and the identified errors are highly objective rather than arbitrary guesses.
- **Temporal Boundary Annotation:** Defining the exact start and end timestamps of a queried event in unconstrained long videos is inherently subjective and essentially a **needle-in-a-haystack** localization challenge. In this demanding context, achieving a human-human tIoU baseline approaching 70% is exceptionally high, thoroughly validating the robustness of our boundary annotations.

B.4 Annotation Interface

We present our annotation interface in Fig. 5.

B.5 Qualitative Examples

We present the qualitative examples of error types in existing benchmarks in Fig. 6.

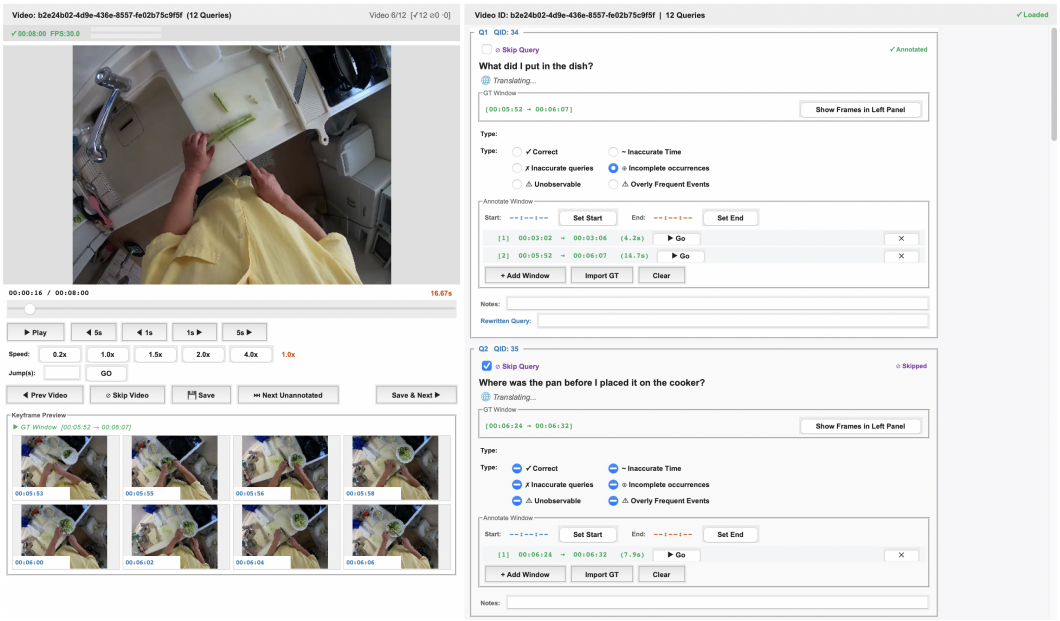


Figure 5: The customized human-in-the-loop annotation interface used for refining AO-Bench.

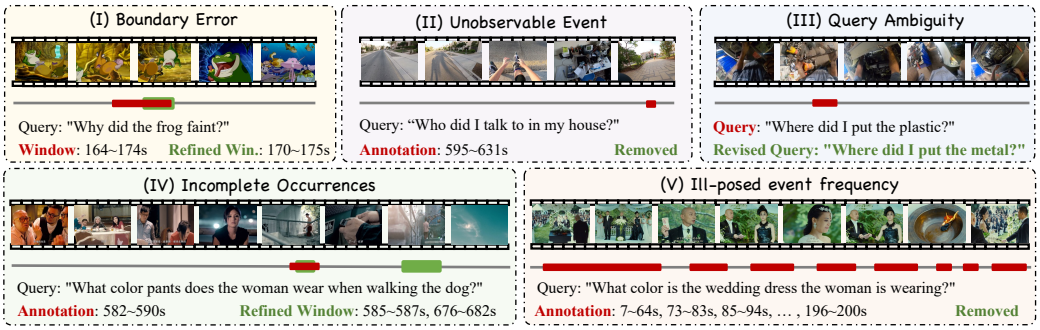


Figure 6: **Qualitative examples of annotation error types in existing datasets.** These include boundary error, query ambiguity, incomplete occurrences, unobservable event and ill-posed event frequency. Red and green bars denote the original annotations and the corrected ground truth, respectively. Through our rigorous manual refinement, these errors have been properly corrected.

C More Experimental Results

C.1 Full-Data Ablation of Teacher Supervision

We conduct a full-data ablation using all 65,657 clip–query pairs in AOVTG-66K for each of three supervision methods: (i) a single Qwen3-VL-32B teacher, (ii) majority voting over the independent predictions of Gemini-3-Flash, GPT-4.1, and Qwen3-VL-32B, and (iii) MACR using the same three agents. Each method labels the exact same training pairs, including negatives. We fine-tune a Qwen3-VL-8B student on each full training set using the same training budget and settings (Section 4.1). The student trained with MACR supervision is ReGround.

We evaluate the teachers and students on Ego4D-AO, using UniTime for retrieval and refinement with the same top-5 32-second candidate clips. Table 6 reports the results, with teacher and ReGround scores reproduced from Table 4.

Evaluation results. ReGround achieves the best student performance across all four metrics. Compared with single-teacher and majority-vote supervision, it improves R-F1 by 3.01 and 3.57 percentage

Table 6: **Full-data teacher-source ablation on Ego4D-AO.** All students use all 65,657 clip–query pairs in AOVTG-66K with the same training budget. The MACR-trained student is ReGround. Teacher results and ReGround results are reproduced from Table 4, with UniTime fixed for retrieval and refinement. All scores are in %.

Teacher method	Teacher performance				Student performance			
	R-F1	SO mIoU	MO mIoU	MO I-F1	R-F1	SO mIoU	MO mIoU	MO I-F1
Qwen3-VL-32B	52.28	23.47	19.41	28.04	56.30	24.58	19.24	28.77
Plain Majority Voting	51.99	19.79	16.67	27.66	55.74	22.22	18.42	28.99
MACR	57.05	25.22	20.71	30.95	59.31	26.82	21.05	30.43

points and MO mIoU by 1.81 and 2.63 points, respectively. With training samples, student backbones, and training budgets held fixed, these results support the benefit of MACR-generated supervision for both candidate selection and downstream grounding.

All three students improve R-F1 over their respective teachers. In particular, ReGround exceeds its MACR teacher by 2.26 percentage points in R-F1, 1.60 points in SO mIoU, and 0.34 points in MO mIoU, with a 0.52-point decrease in MO instance-level F1. These results show that the distilled student improves candidate selection while largely preserving the teacher's downstream grounding performance under the same retrieval and refinement settings.

C.2 Evaluation of the Retrieval Model

Table 7: **Clip retrieval performance.** *Prec.* and *Rec.* denote clip precision and GT recall (%), respectively, measuring the correctness of retrieved clips and the coverage of ground-truth moments. These metrics are introduced in Appendix C.2.

Retrieval Model	MomentSeeker-AO						Ego4D-AO					
	SO		MO		Overall		SO		MO		Overall	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Embedding Model												
ASP-CLIP (Chen et al., 2023)	8.10	36.99	15.77	30.99	9.15	34.88	7.11	34.05	19.51	35.39	12.41	34.99
InternVideo2-Stage2-6B (Wang et al., 2024b)	8.55	38.81	17.12	32.39	9.72	36.56	7.50	36.03	18.36	33.83	12.14	34.49
Qwen3-VL-Embedding-8B (Li et al., 2026a)	9.59	45.36	13.46	27.89	10.12	39.23	9.29	44.96	20.71	38.22	14.17	40.24
SigLip2-so400m-patch14-384 (Tschannen et al., 2025)	18.90	**83.26**	**30.58**	**60.56**	**20.50**	**75.30**	17.09	79.67	36.23	64.76	25.27	69.23
Temporal Grounding Model												
UniversalVTG (An et al., 2026)	10.14	43.23	14.42	26.76	10.72	37.45	9.79	44.30	21.20	36.80	14.66	39.05
UniTime (Li et al., 2026b)	**18.93**	80.49	27.94	51.83	20.18	70.38	**18.05**	**82.31**	**38.27**	**68.01**	**26.69**	**72.30**

Evaluation metrics. Retrieval uses the same IoG-based matching criterion as reranking (Section 4.3). A candidate clip matches a ground-truth window if their intersection duration divided by the ground-truth window duration exceeds 0.3. **Clip Precision** is the fraction of retrieved clips matching at least one ground-truth window, while **GT Recall** is the fraction of ground-truth windows matched by at least one retrieved clip. These metrics assess the precision of retrieved candidates and the coverage of target occurrences, respectively.

Evaluation results. As shown in Table 7, we find that: (i) most embedding models struggle to distinguish clips within the same video due to high visual redundancy, typically yielding sub-optimal performance (e.g., InternVideo2 achieves merely 12.1% overall precision and 34.5% recall on Ego4D-AO). (ii) in contrast, SigLip2 shows much stronger fine-grained vision-language alignment, achieving high overall recall on both MomentSeeker-AO and Ego4D-AO (75.3% and 69.2%, respectively). (iii) for UniversalVTG, although frame-level confidence scores are effective for temporal grounding, they are less suitable for clip-level retrieval, resulting in limited overall recall, e.g., 37.5% on MomentSeeker-AO and 39.1% on Ego4D-AO. (iv) ultimately, the MLLM-based VTG model UniTime achieves strong overall performance (top-tier recall, 72.3%, on Ego4D-AO) by leveraging a coarse-to-fine architecture and an iterative prediction-and-masking strategy. Since retrieval is not the main focus of this work and our framework can adopt any strong retriever, we use UniTime in the following experiments considering its overall performance across both benchmarks.

C.3 Evaluation of the Refinement Model

Table 8: **Boundary refinement performance.** Performance is measured by R1@.3 and mIoU (%) across SO, MO, and Overall scenarios. Best results are in **bold**, and second best are underlined.

Refinement Model	MomentSeeker-AO						Ego4D-AO					
	SO		MO		Overall		SO		MO		Overall	
	R1@.3	mIoU	R1@.3	mIoU	R1@.3	mIoU	R1@.3	mIoU	R1@.3	mIoU	R1@.3	mIoU
Temporal Grounding Specific Model												
UniTime	40.94	28.04	25.00	14.32	38.76	26.15	**40.00**	26.81	27.94	21.06	34.85	24.35
UniversalVTG	39.42	26.71	21.15	15.56	36.93	25.19	38.35	25.93	27.49	20.65	33.71	23.68
TimeLens	42.77	29.34	**26.92**	**20.32**	40.60	28.11	35.04	24.22	30.82	21.67	33.24	23.13
Vision Foundation Models / MLLMs												
Gemini-3-Flash	36.07	24.02	7.69	7.22	32.19	21.72	36.69	24.63	13.30	12.89	26.70	19.62
GPT-4.1	41.10	27.85	25.00	18.47	38.90	26.57	35.54	24.00	27.05	20.31	31.91	22.42
Doubao-Seed-1.6-vision-0815	43.84	29.74	25.96	18.51	41.39	28.20	39.50	**27.36**	30.82	22.70	**35.80**	**25.37**
Qwen3-VL-8B	34.70	24.83	23.08	17.56	33.11	23.84	28.76	20.26	20.40	16.47	25.19	18.64
Ours												
ReGround	**45.81**	**29.89**	**26.92**	18.11	**43.23**	**28.28**	37.69	25.92	**32.59**	**23.13**	35.51	24.73

Evaluation setup & metrics. To decouple fine-grained grounding from upstream stages, we introduce a controlled evaluation setup. Specifically, we use UniTime for retrieval and ReGround for reranking. We then feed only the correctly matched clips (true positives) into various refinement models to isolate and assess their pure boundary prediction capabilities. Performance is measured by Recall@1 at IoU=0.3 (R1@.3) and multi-window mIoU (%).

Evaluation results. As shown in Table 8, (i) ReGround achieves precise temporal grounding while maintaining accurate reranking. It consistently outperforms specialized video temporal grounding models such as UniTime, especially on the out-of-domain MomentSeeker-AO, achieving 28.3% overall mIoU compared with UniTime's 26.2%. (ii) ReGround also surpasses general proprietary MLLMs and remains highly competitive with the leading `Doubao-Seed-1.6-vision` model. Notably, ReGround achieves the best overall mIoU on MomentSeeker-AO with 28.3%, exceeding Doubao's 28.2%. These results confirm that jointly training reranking and grounding preserves fine-grained grounding accuracy while improving the efficiency of the overall pipeline.

C.4 Inference efficiency analysis

We compare the inference efficiency of **MACR + Refinement** and the distilled ReGround. MACR + Refinement first performs multi-agent consensus reranking and then applies a refinement model, while ReGround jointly predicts relevance and temporal boundaries in a single forward pass. For evaluation, we randomly sample 10 videos from each benchmark and report the average wall-clock inference time over 20 videos. Both methods use the same UniTime-retrieved candidates, and retrieval and video decoding time are excluded. All experiments are conducted under the same hardware and inference settings.

Table 9: **Inference time breakdown.** We report the average inference time for reranking and refinement. Retrieval and video decoding time are excluded. ReGround jointly performs reranking and refinement in one forward pass.

Method	Rerank Time	Refine Time	Total Time
MACR + Refinement	101.1s	1.5s	102.6s
ReGround	1.2s		1.2s

As shown in Table 9, MACR + Refinement incurs substantial inference latency, taking 102.6s per clip on average. The cost is dominated by iterative multi-agent reranking, which alone requires 101.1s, while the refinement stage only takes 1.5s. In contrast, ReGround removes online multi-agent deliberation and jointly performs reranking and refinement in a single forward pass, reducing the average inference time to 1.2s. This yields an approximately 85.5× speedup over MACR +

Refinement, demonstrating that distillation preserves the benefits of MACR while making arbitrary-occurrence grounding substantially more efficient.

D Details of Curating AOVTG-66K

To construct AOVTG-66K, we developed a rigorous pipeline that filters raw video data and generates high-fidelity temporal grounding annotations. The process is divided into three key phases:

D.1 Data Source and Filtering

We aggregate raw videos and their corresponding annotations from five large-scale datasets: **Ego4D-NLQ** (Grauman et al., 2022), **TACoS** (Regneri et al., 2013), **YouCook2** (Zhou et al., 2018), **COIN** (Tang et al., 2019), and **HiREST** (Zala et al., 2023). To meet the requirements of long-form video understanding, we filter the collection to include only videos with a total duration exceeding **128 seconds**.

Training–evaluation separation. During data construction, we strictly cross-check the source video IDs in AOVTG-66K against those in Ego4D-AO and MomentSeeker-AO. These ID sets are disjoint, ensuring that no evaluation video or derived clip appears in the training data.

D.2 The MACR Data Generation Pipeline

The curation follows a verification-then-generation strategy:

- **GT verification:** The original ground-truth (GT) windows and queries are processed by `gemini-3-flash`. The model determines if the query is logically and visually grounded in the segment; if they do not match, the pair is discarded.
- **Clip retrieval:** For verified queries, we utilize **UniTime** to retrieve the **top 10** most relevant clips across the entire video.
- **Consensus refinement:** Each segment-query pair is then processed through the **MACR pipeline** (Gemini-3-flash, GPT-4.1 and Qwen3-VL-32B). This multi-agent setup generates structured attributes, identifies visual evidence in specific frames, and performs peer review to ensure the final annotation is accurate.

D.3 Dataset Statistics

The final statistics of AOVTG-66K are summarized in Table 10. The dataset emphasizes diversity in video content and complexity in temporal reasoning.

Overall scale. AOVTG-66K contains 2,644 videos, 9,962 queries, and 65,657 segment–query pairs from five source datasets. The query-weighted mean video duration is approximately 461.96s, with approximately 2.77 windows per query. The dataset name reflects its approximately 66K segment–query pairs.

Table 10: Detailed statistics of the AOVTG-66K dataset. The dataset focuses on videos longer than 128s across diverse action and egocentric domains. Aggregate averages are computed from rounded source statistics, weighted by query counts.

Source Dataset	# Videos	Avg. Dur. (s)	# Queries	Avg. # Win.	# Seg-Query Pairs
Source Data Distribution					
Ego4D-NLQ	1212	496.35	6962	3.01	45158
COIN	652	220.88	750	2.83	5037
YouCook2	527	404.9	750	1.39	5148
TACoS	18	504.82	750	1.97	5162
HiREST	235	397.95	750	2.71	5152
Total / Avg.	**2644**	**461.96**	**9962**	**2.77**	**65657**

E PROMPT DESIGNS FOR MACR FRAMEWORK

Prompt for Stage 1: Structured Attribute Verification

System: You are a strict video relevance evaluator. Determine whether the video segment contains sufficient visible evidence to match the query or answer the question.

User: A user asked: "{query}"
=== CANDIDATE SEGMENT ({seg_block}) ===
I sampled {num_frames} frames at ∼1 fps (≤ {max_frames} total) from this segment. Each frame is labeled with its timestamp.

Rules:

1. Use only visible evidence in the sampled frames. Do not guess, infer, or rely on off-screen or missing moments.
2. First decompose the query into the required visual attributes that must match.
3. Check each required attribute separately. If any is missing, unclear, or uncertain, then `match=false`.
4. For actions, you must explicitly see the progression across multiple frames.
5. For questions, `match=true` only if visible evidence is sufficient; else `answer=null`.
6. `match_windows` must list all non-overlapping windows that tightly cover the evidence.

Format (JSON):

```
{
    "required_attributes": ["...", "..."],
    "attribute_assessment": [
      {
        "attribute": "...",
        "status": "matched | not_matched",
        "evidence": "frame/timestamp evidence or null"
      }
    ],
    "reason": "Concise comparison.",
    "match": true,
    "match_windows": [{"start": "...", "end": "..."}],
    "answer": null
}
```

Prompt for Stage 2: Peer Review and Scoring

System: You are a critical review expert in a multi-agent committee. Your goal is to identify hallucinations and logical inconsistencies in other agents' assessments.

User: Query: {query}
=== CANDIDATE SEGMENT ({seg_block}) ===
I sampled {num_frames} frames at ∼1 fps (≤ {max_frames} total) from this segment. Each frame is labeled with its timestamp.
Current Committee Responses: {collective_responses}
Task: 1. Review the `attribute_assessment` and `match_windows` of each agent. 2. Cross-reference their claims with the visual evidence. 3. Assign a reliability score (0.0 to 10.0) to each agent based on their logical consistency and alignment with video content.
Return Format (JSON): "peer_review": ["agent_id": "...", "critique": "...", "score": 8.5], "consensus_suggestion": "..."

Prompt for Stage 3: Iterative Consensus Refinement

System: You are a senior reasoning agent in a multi-agent committee. Based on the previous debate, peer critiques, and credibility scores, you must refine your final verdict.

User: A user asked: "{query}"
=== CONVERSATION HISTORY ===
The following is the structured debate from the previous round:

- **Your Previous Response:** {last_output}
- **Peer Critiques & Scores:** {review_feedback}
- **Expelled Agent:** {expelled_agent_id} (Their reasoning was deemed unreliable due to low peer scores).

=== CANDIDATE SEGMENT ({seg_block}) ===
I sampled {num_frames} frames at $\sim$1 fps ($\leq$ {max_frames} total) from this segment. Each frame is labeled with its timestamp.

Task:

1. Re-evaluate the video segment by critically considering the critiques you received.
2. Correct any hallucinations in your `attribute_assessment` or `match_windows` if peers pointed out valid discrepancies.
3. If you maintain your original stance despite peer disagreement, you must provide stronger, frame-specific evidence to justify your reasoning.
4. Reach a final consensus on whether the event occurs.

Format (JSON): [Same as the JSON structure defined in Stage 1]

AI USE STATEMENT

We used generative AI to rephrase selected questions in the dataset and benchmark; these questions are identified in the paper. We also used generative AI to improve the clarity and readability of the manuscript. The authors reviewed the AI-assisted content and take full responsibility for the final questions and manuscript.

REPRODUCIBILITY STATEMENT

Appendix X details the construction rules for the dataset and benchmark, and Appendix Y provides the experimental settings and implementation details. We plan to publicly release the dataset and benchmark upon acceptance.