TimeLens2: Generalist Video Temporal Grounding with Multimodal LLMs

Yuhan Zhu[*1,2] Changlian Ma[*1,2] Xiangyu Zeng[*1,2] Xinhao Li[*1] Zhiqiu Zhang[*3,2]
Songze Li[6,2] Jun Zhang[1] Tianxiang Jiang[5,2] Yuandong Yang[1] Ziang Yan[4,2]
Zikang Wang[3,2] Xinyu Chen[1,2] Haoran Chen[1,2] Shaowei Zhang[3,2] Limin Wang[1,2,✉]

[1]Nanjing University [2]Shanghai AI Laboratory [3]Shanghai Jiao Tong University
[4]Zhejiang University [5]University of Science and Technology of China [6]Fudan University

ABSTRACT

Video multimodal large language models (MLLMs) can describe what happens in a video, but rarely identify when the supporting evidence occurs. We study generalist video temporal grounding, in which one model predicts a variable-cardinality set of evidence intervals across video lengths, domains, query forms, and viewpoints. Existing training strategies are misaligned with this set-valued task: long-video labels often rely on brittle one-pass annotation, while reinforcement-learning rewards either fail to distinguish non-overlapping predictions or require fragile segment matching. TimeLens2 treats temporal evidence as an interval set throughout supervision and optimization. TimeLens2-93K constructs reliable multi-span supervision through caption-derived proposals, independent localization, cross-agent consensus, semantic verification, and boundary refinement. Our temporal Wasserstein reward computes exact one-dimensional W_1 between uniform distributions over merged interval supports, providing dense, matching-free feedback under unequal cardinalities and equivalent fragmentation; temporal IoU complements it with precise-overlap feedback. Across seven benchmarks, TimeLens2-2B outperforms all size-matched baselines on every benchmark, while the 4B and 8B variants achieve state-of-the-art performance, surpassing open-source models with up to 397B parameters. The 2B, 4B, and 8B variants improve over their Qwen3-VL backbones by 14.2, 13.0, and 18.1 mIoU points, respectively.

https://mcg-nju.github.io/TimeLens2
https://github.com/MCG-NJU/TimeLens2
https://huggingface.co/collections/MCG-NJU/timelens2

1 Introduction

Video multimodal large language models (MLLMs) promise to make growing video archives searchable through language [26, 52, 21, 2, 51]. Yet answering *what* happened is not enough: users must also know *when* the supporting evidence appears. Without temporal support, users must still search the full timeline, and even correct descriptions remain unverifiable. Temporal grounding is therefore the video analogue of citation: it makes outputs traceable by requiring precise, potentially disjoint evidence intervals rather than only fluent responses [8, 15, 31, 13, 41].

Temporal grounding is not a single-regime task: supporting evidence may occupy only seconds of an hour-long video, recur at disjoint moments, and appear in third- or first-person footage, while queries may be descriptions or questions. We therefore study *generalist temporal grounding*, where a single MLLM with a unified output interface localizes single- or multi-interval evidence across video lengths, domains, query forms, and viewpoints.

Two structural mismatches stand in the way. The first is in supervision. Temporal evidence is a set of intervals, yet long-video labels are often produced as a single global annotation decision. Because relevant moments are sparse and similar distractors dominate the timeline, a one-pass annotator may confuse similar occurrences, miss repeated evidence, or produce imprecise boundaries [29, 50, 53]. Meanwhile, many existing datasets emphasize short videos, declarative captions, and single-span answers [8, 15, 12]. Long-video annotation is therefore not merely a scaling problem; it is an evidence-verification problem.

[*] Equal contribution.
✉ Corresponding author: lmwang@nju.edu.cn.

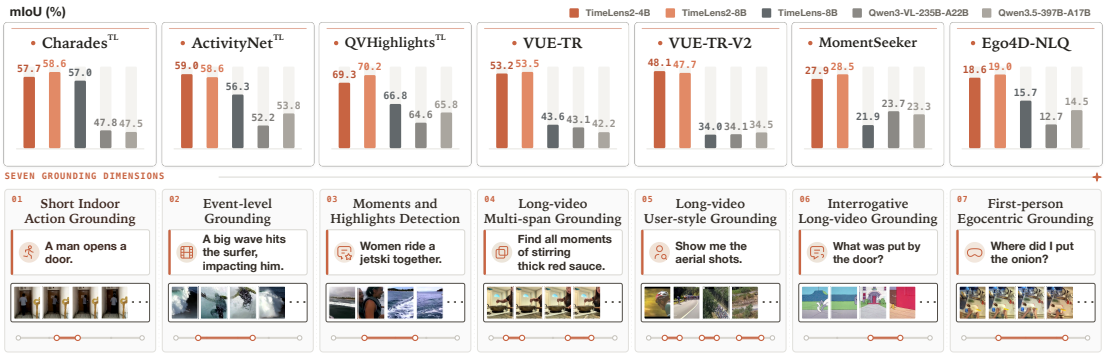

Figure 1. TimeLens2 as a generalist video temporal-grounding model. Top: mIoU across seven benchmarks covering diverse evaluation dimensions. **Bottom:** Representative queries from each benchmark.

Even with reliable supervision, optimization remains mismatched with temporal evidence: next-token prediction learns timestamp generation effectively, but lacks an explicit interval-level objective [31, 13, 50, 53]. Reinforcement learning with temporal IoU (tIoU) better aligns training with evaluation [44, 22, 18, 3], but assigns zero to all disjoint predictions, leaving near misses indistinguishable from distant errors. For multi-span targets, one-to-one matching [25] is also fragile under fragmentation or unequal cardinality.

We introduce TimeLens2, a generalist temporal-grounding MLLM that treats evidence as a first-class interval set throughout training. The central idea is to make evidence *verifiable* when constructing supervision and *geometry-aware* when optimizing predictions.

TimeLens2-93K turns one brittle global annotation decision into a sequence of increasingly focused ones. Hierarchical, time-stamped captions first provide long-range context for proposing declarative queries and coarse single- or multi-span evidence; complementary grounding agents then independently relocalize that evidence from the video. Temporal consensus and semantic verification remove unstable or mismatched instances, after which local refinement sharpens only the surviving boundaries. By progressively narrowing both the search space and the error type, the pipeline preserves repeated evidence while making long-video supervision more reliable.

Once reliable interval labels are available, training separates capability acquisition from geometric calibration. Long-context supervision first teaches the model to search full videos and generate variable-cardinality interval sets; GRPO then directly calibrates where those intervals lie [32]. tIoU measures how much predicted support is already correct, but not how a disjoint prediction should move. Our temporal Wasserstein reward supplies this missing geometry by computing the exact one-dimensional W_1 distance between uniform distributions over merged predicted and target supports, providing dense, matching-free guidance for near misses and unequal-cardinality outputs while remaining invariant to equivalent fragmentation.

Figure 1 summarizes performance across seven benchmarks: TimeLens2-2B, TimeLens2-4B, and TimeLens2-8B reach 44.5, 47.7, and 48.0 average mIoU, respectively. The 2B, 4B, and 8B variants improve over their Qwen3-VL backbones by 14.2, 13.0, and 18.1 points [2]; the 4B model also surpasses Qwen3.5-397B-A17B [37] on every benchmark by 7.5 points on average. In data ablations, declarative-only TimeLens2-93K raises question-form grounding on MomentSeeker from 15.3 to 25.8 mIoU [48]. Temporal Wasserstein restores informative preferences for 75.8% of all-zero-tIoU GRPO groups.

Our contributions are threefold:

- We introduce TimeLens2-93K, a staged evidence-verification pipeline that produces reliable single- and multi-span temporal grounding supervision for long videos.
- We propose a matching-free temporal Wasserstein reward based on exact one-dimensional W_1 over merged interval support, providing graded, fragmentation-invariant feedback for disjoint and unequal-cardinality predictions.
- We demonstrate compact 2B, 4B, and 8B generalist models across seven benchmarks covering long-video, multi-span, question-form, and egocentric grounding.

2 Related Work

Temporal Grounding Models. Video temporal grounding has evolved from dedicated localization architectures to generative MLLMs. Classical methods match moment proposals or regress boundaries [8, 55]; recent systems instead express temporal evidence through a language interface. TimeChat and VTimeLLM introduce temporal instruction tuning and boundary-aware training [31, 13], while LITA and Grounded-VideoLLM develop timestamp-aware representations [14, 41]. TimeSuite features grounding-centric instruction tuning to strengthen both long-video QA and temporal localization [50], whereas TimeLens systematically studies data quality and training recipes [53]. In parallel, general video representations and embeddings broaden transferable video evidence search [40, 24, 56]. This shift makes grounding a general MLLM capability. Yet supervision has not kept pace: transfer across video lengths, evidence cardinalities, viewpoints, and query forms remains unresolved. TimeLens2 is designed around this broader regime.

Temporal Grounding Data. The data landscape reflects the same progression. Early benchmarks primarily pair short or domain-specific videos with a single descriptive moment [30, 12, 8, 15]. Later datasets expand individual axes of difficulty, including subtitle-aware retrieval, repeated evidence, long-form video, and egocentric search [16, 17, 34, 11]. MLLM recipes then convert timestamped annotations into instruction-following conversations [31, 13, 50, 29], often scaling through dense captions, procedural videos, or weak narration alignment [54, 27, 35, 57]. Scale, however, does not by itself resolve supervision quality: in long videos, coarse alignment readily becomes a wrong occurrence, a missed repeat, or an imprecise boundary. TimeLens2-93K targets this bottleneck with consensus-based localization and semantic verification, yielding reliable single- and multi-span supervision across long, diverse videos.

Temporal Grounding Optimization. A parallel mismatch appears in optimization. Supervised instruction tuning teaches the syntax of timestamps through token likelihood, not the quality of the localized evidence. GRPO/RLVR methods make grounding verifiable through format and temporal-overlap rewards, and subsequent work extends this recipe with improved reward design, data selection, refusal, curricula, and multi-segment reasoning [32, 44, 22, 18, 3, 49, 6, 45, 25, 19]. Nevertheless, overlap supplies geometry only after prediction and target intersect. MUSEG introduces pairwise NGIoU for disjoint segments, but its one-to-one matching remains sensitive to fragmentation, merges, and unequal cardinality [25]. The issue is therefore not merely reward sparsity; it is also the choice of representation for set-valued evidence. Wasserstein geometry provides a useful precedent: NWD represents spatial boxes by Gaussian surrogates and applies pairwise W_2 in supervised tiny-object detection [42]. Our objective instead treats the merged support of a variable-cardinality interval set as a one-dimensional distribution and computes its exact W_1 distance.

3 TimeLens2

3.1 TimeLens2-93K: Scalable Temporal Grounding Data Construction

Long videos pose a fundamental supervision challenge: full-video context is needed to teach evidence search, yet sparse evidence and growing distractors make precise annotation unreliable. We construct TimeLens2-93K from long, diverse web videos and represent each example as $(v, q, \mathcal{Y})$, where $\mathcal{Y} = \{[s_k, e_k]\}_{k=1}^{K}$ is a variable-cardinality set of supporting intervals. This formulation unifies single- and multi-span grounding while preserving the context in which the evidence must be found.

As summarized in Figure 2, our pipeline separates candidate construction (Steps 1–3) from label determination (Steps 4–6). Hierarchical, time-stamped captions yield declarative queries and coarse single- or multi-span proposals; independent agents then relocalize the proposed evidence directly from short video clips. Temporal consensus and semantic verification reject unstable or mismatched instances, after which local refinement sharpens only the surviving boundaries. This staged factorization lets the model learn from full-video context while producing labels through controlled local decisions. The resulting corpus contains 23,793 videos and 93,232 grounding instances, including 12,091 instances with multiple supporting intervals.

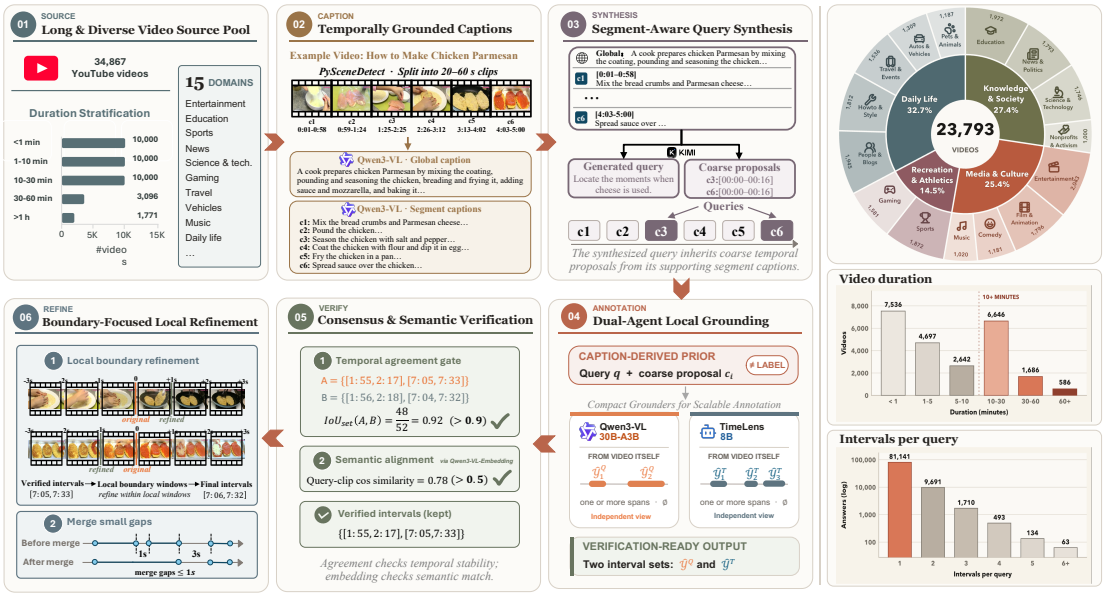

Figure 2. TimeLens2-93K construction pipeline and corpus profile. Steps 1–3 construct caption-derived queries and coarse proposals; Steps 4–6 relocalize proposals with independent agents, verify interval sets, and refine their boundaries. The right column summarizes the retained corpus.

Long and diverse video source pool. To make temporal search difficulty a deliberate design axis rather than an incidental consequence of web sampling, we stratify 34,867 YouTube videos across five duration ranges, from under one minute to over one hour, and diverse visual domains. Duration controls the search horizon and distractor load, while domain breadth diversifies the evidence to be grounded.

Hierarchical temporally grounded captions. Directly generating queries and localizing their evidence over a long video would conflate video understanding, query synthesis, and localization in one brittle step. We therefore convert each video into a hierarchical, time-stamped description that serves as a semantic index for subsequent query and proposal generation. A global caption captures subjects, setting, actions, and temporal progression to disambiguate recurring events; segment captions enumerate visually verifiable actions, object states, and interactions within bounded intervals. We construct this index by partitioning v into semantically coherent clips $\mathcal{C}(v) = \{c_i = [t_i, t_{i+1}]\}_{i=1}^{N}$ using content-based PySceneDetect boundaries constrained to 20–60 seconds. Compared with uniform slicing, these boundaries better preserve visual continuity while keeping clips short enough for reliable captioning. Qwen3-VL-235B-A22B [2] generates both caption levels, emphasizing observable events and temporal order.

Segment-aware query synthesis. Query synthesis and evidence proposal are naturally coupled: once a query is derived from time-stamped segment descriptions, its candidate support is already encoded in the same hierarchy. Rather than discard this alignment and recover it through a separate full-video retrieval stage, we condition Kimi-K2.5 [36] on the global and segment captions to jointly generate a non-redundant declarative query and select all segments whose captions support it. These segments form a coarse proposal $\mathcal{P}(q) \subseteq \mathcal{C}(v)$, which may include disjoint clips when evidence recurs. Caption-derived proposals provide temporal priors for downstream agent annotation.

Dual-agent local grounding. Caption-derived proposals narrow the search space but are not labels: captions may contain unsupported content, and segment boundaries rarely match true event boundaries. We therefore relocalize each proposal from the video itself. For each query-proposal pair (q, c_i), Qwen3-VL-30B-A3B [2] and TimeLens-8B [53] independently return one or more intervals at one-second resolution or an empty set. Both are strong yet efficient, while their distinct inductive biases provide two views for consensus. Empty outputs reject unsupported proposals; multi-interval

outputs preserve repeated evidence within a clip. We map clip-relative timestamps to the original timeline and aggregate them into $\hat{\mathcal{Y}}^{\mathrm{Q}}$ and $\hat{\mathcal{Y}}^{\mathrm{T}}$. This converts a coarse text match into a verifiable visual decision with explicit boundaries.

Cross-agent consensus and semantic verification. Independent localization makes temporal labels testable: reliable evidence should be recovered across annotators with different inductive biases. Yet agreement alone cannot rule out a shared semantic error. We therefore verify two distinct properties: temporal reproducibility and query–evidence alignment. For temporal consensus, let $\mathrm{merge}(\cdot)$ denote interval union and $|\cdot|$ total duration. Comparing merged supports for multi-span outputs:

$$\mathrm{IoU}_{\mathrm{set}}(\mathcal{A}, \mathcal{B}) = \frac{|\,\mathrm{merge}(\mathcal{A}) \cap \mathrm{merge}(\mathcal{B})|}{|\,\mathrm{merge}(\mathcal{A}) \cup \mathrm{merge}(\mathcal{B})|}.$$

We retain instances with $\mathrm{IoU}_{\mathrm{set}}(\hat{\mathcal{Y}}^{\mathrm{Q}}, \hat{\mathcal{Y}}^{\mathrm{T}}) > 0.9$ and keep the Qwen annotation as canonical; post-hoc fusion could create boundaries predicted by neither model. To verify semantic validity, we encode each retained target clip and its query with Qwen3-VL-Embedding [20] and require a normalized text–video cosine similarity of at least 0.5. Thus, consensus tests whether a localization is reproducible, while embedding verification tests whether the reproduced evidence is relevant.

Boundary-focused local refinement. Once consensus and semantic verification establish the event identity, the remaining uncertainty lies at its visual transitions. We therefore cast refinement as local change-point detection: for each retained boundary, Qwen3-VL-235B-A22B observes a ± 3-second neighborhood and predicts a refined transition point. For multi-interval labels, we merge adjacent spans separated by at most one second, treating such gaps as boundary jitter or brief occlusion rather than semantic breaks. Restricting refinement to verified local windows keeps this stronger model practical at scale.

The resulting cascade uses global context to propose evidence and concentrated local computation to verify and refine high-confidence interval-set labels.

3.2 TimeLens2 Model: Learning Generalist Video Temporal Grounding

Verified interval labels improve supervision, but a generative MLLM must still learn to search long contexts and express interval sets before being optimized by interval-level metrics. We therefore separate capability acquisition from geometric calibration: long-context supervised training learns evidence search and output conventions, while reinforcement learning directly optimizes decoded interval sets against non-differentiable temporal objectives.

Long-context supervised grounding. Temporal grounding is a search problem before it is a boundary-estimation problem: target-centered training removes the distractors that require evidence selection. We therefore fine-tune Qwen3-VL [2] on long-context examples drawn primarily from TimeLens2-93K and TimeLens-100K [53]. All samples use $(v, q, \mathcal{Y})$, where $\mathcal{Y}$ contains one or more supporting intervals; the official Ego4D-NLQ training split [11] is added to broaden supervision to first-person video. Retaining full-video context forces the model to isolate sparse evidence among competing events, jointly learning evidence search and interval-set generation.

Instruction and response-format diversity. Temporal grounding should be invariant to how evidence is requested and serialized. A single prompt and timestamp format can entangle localization with surface-form imitation, limiting cross-benchmark transfer. For each training example, we independently sample the grounding instruction, answer syntax, and timestamp encoding. The same $\mathcal{Y}$ is rendered in diverse single- or multi-span formats, including JSON, natural language, key–value or range styles, and alternative timestamp conventions. This encourages protocol-invariant interval semantics without synthetic paraphrases that might shift the evidence target.

Rollout-guided hard-sample mining. After SFT, GRPO should focus on what the current policy still fails to localize. Uniform prompt sampling wastes updates on solved examples, while static difficulty heuristics may not reflect model-specific failures. We therefore generate multiple off-policy completions from the SFT checkpoint for each example in TimeLens2-93K and TimeLens-100K, using their mean tIoU as an empirical estimate of model competence. Examples with lower scores receive

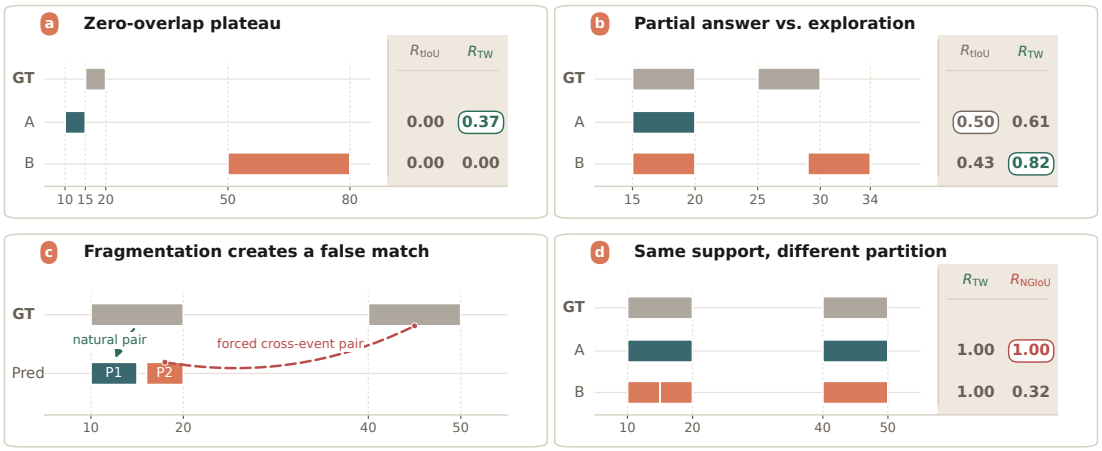

Figure 3. Overlap plateaus and matching artifacts in temporal interval-set rewards. **(a)** tIoU ties the near miss A and distant error B at zero, whereas R_{TW} ranks them by temporal proximity. **(b)** tIoU favors the single-moment prediction A, while R_{TW} favors B, which also recovers part of the second target moment. MUSEG addresses these failures with NGIoU under one-to-one matching [25], but introduces new artifacts: **(c)** fragmentation forces P2 into a false cross-event match; **(d)** identical merged support receives different R_{NGIoU} scores solely because the prediction is partitioned differently.

higher sampling weights during GRPO [32]. This converts the model's own failure distribution into an adaptive curriculum, concentrating optimization on unresolved localization errors rather than mirroring the dataset distribution.

Temporal Wasserstein reward. An effective RL reward must rank imperfect localizations by their progress toward the target, not merely recognize existing overlap. Yet tIoU, the dominant reward in prior video temporal grounding RL [44, 53, 22], is better suited to evaluation than reward shaping. In Figure 3(a), it assigns zero to both the near miss A and the distant error B, discarding their temporal geometry. In the multi-moment example of Figure 3(b), tIoU favors the partial answer A over B (0.50 vs. 0.43), even though B covers both target moments. When a GRPO group falls on such a plateau, tIoU provides no relative learning signal after mean centering. We therefore complement it with a temporal Wasserstein reward, R_{TW}, which ranks A above B in (a) and B above A in (b) by measuring the transport needed to align predicted temporal mass with the target support.

Accordingly, each sampled completion is parsed into a predicted interval set $\hat{\mathcal{Y}}$ and scored by three complementary terms:

$$R(\hat{\mathcal{Y}}, \mathcal{Y}) = R_{\mathrm{tIoU}}(\hat{\mathcal{Y}}, \mathcal{Y}) + R_{\mathrm{TW}}(\hat{\mathcal{Y}}, \mathcal{Y}) - \mathbf{1}_{\mathrm{invalid}}.$$

The invalid-output indicator penalizes responses that cannot be parsed into valid intervals. For valid predictions, set-level tIoU rewards the target support already recovered:

$$R_{\mathrm{tIoU}}(\hat{\mathcal{Y}}, \mathcal{Y}) = \frac{|\operatorname{merge}(\hat{\mathcal{Y}}) \cap \operatorname{merge}(\mathcal{Y})|}{|\operatorname{merge}(\hat{\mathcal{Y}}) \cup \operatorname{merge}(\mathcal{Y})|}.$$

tIoU measures how much support is already correct, but not how an incorrect prediction should move. To expose this missing geometry, we lift each non-empty interval set $\mathcal{A}$ to a uniform temporal distribution over its merged support:

$$\mu_{\mathcal{A}}(t) = \frac{\mathbf{1}\{t \in \operatorname{merge}(\mathcal{A})\}}{\ell(\mathcal{A})}.$$

Here $\ell(\mathcal{A}) = |\operatorname{merge}(\mathcal{A})|$. This representation makes the reward depend on where evidence lies, rather than how the same support is partitioned into spans. The 1-Wasserstein distance measures the transport needed to align the predicted and target mass; in one-dimensional time, it has the exact CDF form

$$W(\hat{\mathcal{Y}}, \mathcal{Y}) = W_1(\mu_{\hat{\mathcal{Y}}}, \mu_{\mathcal{Y}}) = \int_{\mathbb{R}} \left| F_{\hat{\mathcal{Y}}}(t) - F_{\mathcal{Y}}(t) \right| \, dt,$$

where $F_{\mathcal{A}}(t) = \int_{-\infty}^{t} \mu_{\mathcal{A}}(\tau)\, d\tau$. We convert the distance into a scale-normalized similarity:

$$R_{\mathrm{TW}}(\hat{\mathcal{Y}}, \mathcal{Y}) = \exp\left(-\frac{W(\hat{\mathcal{Y}}, \mathcal{Y})}{|\operatorname{merge}(\mathcal{Y})| + \epsilon}\right),$$

where the target-duration denominator makes the score comparable across target event scales without allowing overly broad predictions to enlarge their own normalization factor.

Discussion. Boundary- or center-based losses can supply the geometry missing from tIoU for a single interval, but multi-span grounding additionally requires correspondences between predicted and target intervals. MUSEG handles this with one-to-one matching followed by pairwise NGIoU [25]; the matching itself, however, is fragile. Fragmentation sends P2 to the wrong target in Figure 3(c), while predictions with identical merged support receive NGIoU scores of 1.0 and 0.32 solely because they are partitioned differently in (d). R_{TW} instead compares merged temporal mass, avoiding correspondence and providing dense, partition-invariant credit. Thus, tIoU anchors support overlap, R_{TW} supplies geometry, and the parse penalty enforces valid outputs.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate TimeLens2 on the seven temporal grounding benchmarks: the TimeLens-Bench [53] re-annotations of Charades-STA [8], ActivityNet Captions [15], and QVHighlights [17]; the vision subsets of the long-video VUE-TR [38] and VUE-TR-V2 [39]; the text-only query subset of MomentSeeker [48] for question-form localization; and the official validation split of the egocentric Ego4D-NLQ [11].

Evaluation metrics. We report mean temporal Intersection-over-Union (mIoU) and Recall@1 at tIoU thresholds of 0.3, 0.5, and 0.7. mIoU measures average boundary overlap, whereas R1@threshold is the fraction of top predictions meeting the specified overlap. The main paper reports mIoU and R1@0.5; R1@0.3 and R1@0.7 are deferred to the supplementary material (Table 11).

Implementation details. Unless otherwise specified, TimeLens2 uses Qwen3-VL-2B-Instruct, Qwen3-VL-4B-Instruct, or Qwen3-VL-8B-Instruct [2]. Following the two-stage recipe in Section 3.2, SFT uses TimeLens2-93K, TimeLens-100K [53], and the official Ego4D-NLQ-v2 training split [11], whose answer windows are converted to the shared $(v, q, \mathcal{Y})$ format. For the 4B and 8B variants, we train for one epoch on packed sequences of up to 100K tokens, with a global batch size of 256 and a learning rate of 5×10^{-6} under cosine decay. For the 2B variant, we instead use a 160K-token packed length, a global batch size of 128, and a learning rate of 10^{-5}.

For RL, we draw eight off-policy rollouts per example from the SFT checkpoint on TimeLens2-93K and TimeLens-100K, up-weighting examples with low mean tIoU. GRPO then uses eight generations per prompt, a global batch size of 64, a learning rate of 10^{-6}, and a KL coefficient of 0.04; group rewards are mean-centered without variance scaling. Videos are sampled at 2 fps, capped at 512 frames, and processed with a 16K-token budget. The default reward is $R_{\mathrm{tIoU}} + R_{\mathrm{TW}} - \mathbf{1}_{\mathrm{invalid}}$; all RL experiments and ablations retain the fixed $-\mathbf{1}_{\mathrm{invalid}}$ penalty for unparsable outputs. The visual encoder remains frozen in both stages, while the language model is updated.

4.2 Comparison with State-of-the-Art

Temporal grounding. Table 1 shows that TimeLens2 performs consistently across model scales and benchmark regimes. At 2B, TimeLens2-2B outperforms all three size-matched open-source baselines on all seven benchmarks in both metrics, reaching 44.5 average mIoU—14.2 points above Qwen3-VL-2B [2] and 2.4 points above TimeLens-8B [53]. TimeLens2-8B surpasses all prior methods in R1@0.5 on all seven benchmarks and in mIoU on six, trailing Gemini 2.5 Pro on QVHighlights by only 0.2 points. Notably, TimeLens2-4B already outperforms every prior model, including substantially larger and proprietary systems, on six of seven benchmarks in both metrics. Its advantage is largest where temporal search is hardest: over the Qwen3-VL-4B backbone, it improves mIoU by 19.6 points on VUE-TR and 27.8 on VUE-TR-V2, while gaining 12.6 on question-form MomentSeeker and 7.2 on

Table 1. Comparison with proprietary and open-source models on temporal video grounding benchmarks. Metrics are R1@0.5 and mIoU (%). [TL] denotes evaluation on TimeLens-Bench [53]; † marks the VUE-TR vision subsets, and ‡ the MomentSeeker text-only query subset.

Size	Model	Charades[TL]		ActivityNet[TL]		QVHighlights[TL]		VUE-TR†		VUE-TR-V2†		MomentSeeker‡		Ego4D-NLQ	
		R1@0.5	mIoU	R1@0.5	mIoU	R1@0.5	mIoU	R1@0.5	mIoU	R1@0.5	mIoU	R1@0.5	mIoU	R1@0.5	mIoU
Proprietary models															
	GPT-5 [33]	42.0	40.5	44.9	42.9	60.4	56.8	–	–	19.5	20.0	–	–	–	–
	Gemini 3 Pro [9]	–	–	–	–	–	–	–	–	41.3	39.7	–	–	–	–
	Gemini 2.5 Pro [5]	61.1	52.8	64.2	58.1	75.9	70.4	20.5	21.9	–	–	–	–	–	–
Open-source models															
2B	Marlin-2B [28]	51.7	46.5	40.5	37.9	48.5	46.8	23.6	24.8	19.9	19.2	15.2	16.3	7.8	8.9
2B	Qwen3.5-2B [37]	46.8	46.0	52.8	48.9	65.7	60.6	27.2	27.8	21.5	21.6	12.8	15.6	10.1	11.2
2B	Qwen3-VL-2B [2]	44.4	43.4	41.0	39.9	53.4	52.2	29.9	31.4	22.5	24.1	9.8	12.4	8.1	9.3
2B	**TimeLens2-2B**	60.6	53.4	62.0	55.4	72.8	67.0	50.7	51.1	44.5	43.8	21.9	24.3	16.5	16.8
4B	InternVL3.5-4B [43]	14.1	16.0	12.7	14.9	15.8	17.7	14.1	16.5	6.0	8.5	2.1	3.2	0.7	2.0
4B	Qwen3-VL-4B [2]	53.8	49.4	54.8	50.7	66.4	62.3	33.0	33.6	19.9	20.3	13.5	15.3	10.6	11.4
4B	Molmo2-4B [4]	31.1	34.7	40.3	40.9	62.6	60.8	42.5	43.1	30.9	32.1	14.7	19.5	7.7	10.0
4B	VideoChat3-4B [23]	64.9	56.1	60.5	54.6	72.5	67.0	47.8	47.9	40.4	40.2	24.4	25.9	13.9	14.6
4B	**TimeLens2-4B**	66.6	57.7	65.5	59.0	75.2	69.3	52.4	53.2	49.5	48.1	27.0	27.9	18.1	18.6
7B+	Qwen3.5-397B-A17B [37]	47.8	47.5	59.1	53.8	71.6	65.8	42.7	42.2	35.5	34.5	22.3	23.3	13.3	14.5
7B+	Qwen3-VL-235B-A22B [2]	50.8	47.8	57.5	52.2	70.2	64.6	42.5	43.1	34.5	34.1	21.2	23.7	11.5	12.7
7B+	Qwen3.5-35B-A3B [37]	50.4	48.2	58.6	52.6	72.6	66.0	46.3	44.6	37.1	35.3	20.2	22.6	12.0	13.0
7B+	Qwen3-VL-30B-A3B [2]	46.5	48.1	51.1	49.4	67.6	63.2	40.4	42.5	32.3	32.6	19.6	22.6	10.4	11.5
7B+	Vidi-1.5-9B [38]	18.3	22.0	40.2	39.3	58.4	55.9	45.0	47.0	31.9	34.3	14.8	18.6	6.5	8.4
7B+	LLaVA-OneVision-2-8B [1]	59.6	52.6	57.9	52.4	70.0	65.7	40.8	41.3	35.0	34.5	18.0	19.6	12.0	12.8
7B+	Qwen3-VL-8B [2]	49.3	47.2	51.1	48.0	62.9	59.4	23.8	25.5	13.1	14.0	9.4	9.8	4.8	5.3
7B+	InternVideo3-8B [47]	61.7	53.2	51.4	46.3	62.6	59.2	39.4	41.1	32.4	32.6	17.9	19.3	7.3	8.6
7B+	TimeLens-8B [53]	65.8	57.0	62.0	56.3	71.7	66.8	42.9	43.6	34.4	34.0	20.4	21.9	14.6	15.7
7B+	Video-o3-7B [51]	34.2	38.6	41.0	38.9	49.5	47.4	22.1	25.1	12.9	15.3	9.8	13.0	1.9	3.1
7B+	VideoChat-Flash-7B [21]	37.9	39.7	21.8	24.8	30.6	32.7	15.4	17.3	9.5	10.1	5.9	7.2	1.5	2.5
7B+	MiMo-VL-7B [46]	42.6	39.6	38.7	35.5	42.6	41.5	17.5	19.1	10.1	10.4	3.8	6.0	0.4	1.0
7B+	TimeSuite-7B [50]	35.5	38.1	17.5	19.8	16.9	21.7	10.3	13.2	5.3	7.3	3.3	5.9	0.6	1.4
7B+	**TimeLens2-8B**	68.0	58.6	65.6	58.6	76.1	70.2	51.6	53.5	49.1	47.7	27.3	28.5	18.4	19.0

egocentric Ego4D-NLQ. These gains across long videos, query forms, and viewpoints indicate that TimeLens2 improves temporal evidence search rather than merely fitting dataset-specific boundary patterns.

Qualitative analysis. The central challenge in temporal grounding is not event recognition alone, but selecting query-relevant evidence from long, repetitive, or fragmented timelines. Figure 4 makes this distinction visible. In a 93.7-minute video, TimeLens2 retrieves a brief event near 4,750 seconds, while the baselines drift, miss the target, or produce an invalid timestamp. When the evidence recurs, it recovers all five sparse handcuff spans without adding plausible but false windows. The question-form example further separates semantic relevance from answer-bearing evidence: TimeLens2 isolates the brief OCR start list needed to answer the query rather than grounding the broader competition scene. In the egocentric case, it identifies the actual object removal while every baseline selects a later, visually similar interaction. Together, these cases show that TimeLens2 learns evidence-directed temporal search: it remains complete across disjoint spans while resisting distractors, even when the evidence is late, sparse, or defined by the query's intent.

4.3 Ablation Study

Effect of TimeLens2-93K. We ask whether scaling TimeLens2-93K teaches temporal search itself or primarily broadens its coverage. To isolate supervision from RL, we fine-tune Qwen3-VL-4B on different data mixtures. Table 2 reveals a two-phase scaling curve. The first 5% of TimeLens2-93K delivers the largest gain, raising average mIoU from 34.7 to 42.8; 20% reaches 45.3. The full corpus improves the average more modestly to 45.8, but its gains concentrate on harder settings such as VUE-TR-V2 and Ego4D-NLQ. Thus, a small high-confidence subset establishes the basic evidence-search behavior, while broader coverage improves robustness to long contexts and ambiguous boundaries. This behavior also transfers across query forms: although TimeLens2-93K contains

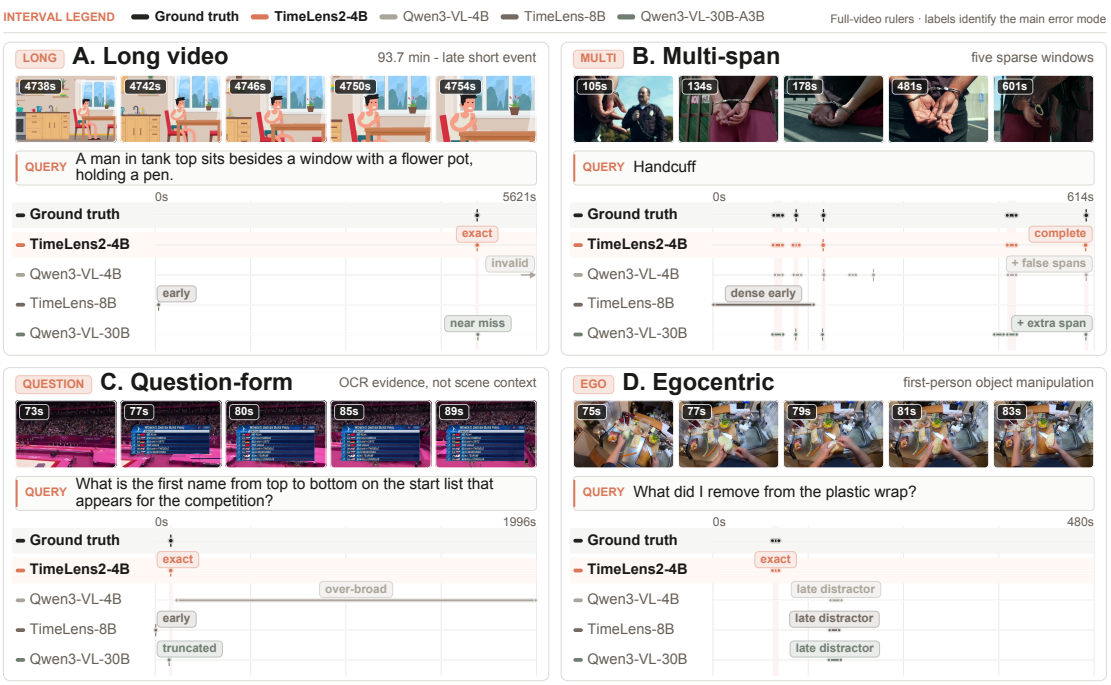

Figure 4. Qualitative comparison on temporal grounding. Each panel shows five target frames, the query, and a full-video ruler from 0 to the video duration; colored spans follow the shared legend. Dense labels summarize many short windows after merging small display gaps.

only declarative queries, it raises question-form MomentSeeker from 15.3 to 25.8 mIoU, indicating that the model learns to search for evidence rather than imitate a query template.

Table 2. Effect of TimeLens2-93K in supervised training. Qwen3-VL-4B mIoU (%) under different data recipes; Avg. is unweighted. Sizes are videos / QA turns in thousands, with QA turns counted from multi-turn annotations. Shading marks the core corpus and final mixture.

Supervised setting	Videos / QA (K)	Evaluation mIoU (%)							
		Charades^{TL}	ActivityNet^{TL}	QVHighlights^{TL}	VUE-TR†	VUE-TR-V2†	MomentSeeker‡	Ego4D-NLQ	Avg.
Backbone									
Qwen3-VL-4B	–	49.4	50.7	62.3	33.6	20.3	15.3	11.4	34.7
Scaling TimeLens2-93K									
5% TimeLens2-93K	1.2 / 4.7	53.0	55.7	67.5	48.3	38.0	22.8	14.2	42.8
20% TimeLens2-93K	4.8 / 18.6	53.4	57.2	68.8	**51.1**	44.4	**26.2**	15.8	45.3
50% TimeLens2-93K	11.9 / 46.6	53.4	57.4	68.6	50.5	45.2	26.1	15.9	45.3
100% TimeLens2-93K	23.8 / 93.2	54.7	57.7	68.8	50.3	**46.0**	25.8	17.2	45.8
Dataset comparison and mixture									
TimeLens-100K	19.4 / 96.6	53.0	51.6	61.3	39.0	33.8	22.9	14.5	39.4
TimeLens2-93K + TimeLens-100K	43.2 / 189.8	55.9	58.4	69.0	50.3	45.3	25.8	17.3	46.0
TimeLens2-93K + TimeLens-100K + Ego4D-NLQ	**44.5 / 203.7**	**56.3**	**58.9**	**69.2**	50.9	45.6	25.8	**17.8**	**46.4**

Scale alone does not explain the gain: TimeLens-100K has a comparable number of QA turns but reaches only 39.4 average mIoU, with substantially weaker long-video and QVHighlights results. Adding it to TimeLens2-93K yields a smaller improvement from 45.8 to 46.0, mainly on the TimeLens-Bench-style datasets; Ego4D-NLQ then raises the average to 46.4 and strengthens egocentric grounding. TimeLens2-93K therefore provides the transferable search capability, while TimeLens-100K and Ego4D-NLQ add complementary benchmark and viewpoint coverage.

Impact of progressive label curation. Effective curation should remove distinct supervision errors, not merely shrink the dataset. Under a fixed Qwen3-VL-4B SFT recipe, Table 3 reveals a quality-over-quantity pattern. The raw annotators tie at 42.0 average mIoU yet differ by dataset, making agreement informative. Temporal consensus cuts the Qwen-anchored pool from 735.4K to 174.2K QA turns while raising mIoU to 43.4; semantic verification leaves 93.2K and reaches 44.1. Removing temporal instability and semantic mismatch therefore improves performance with only one-eighth as many labels.

With evidence identity established, boundary noise becomes the dominant bottleneck. Refining the same 93.2K labels yields the largest stage gain, adding 1.7 points to reach 45.8 mIoU without additional data. The full cascade improves over the raw Qwen labels by 3.8 points overall, including 6.8 on ActivityNet and 8.7 on QVHighlights. Although refinement loses 0.5 points on VUE-TR and MomentSeeker, the progression exposes a clear factorization of annotation error: consensus tests temporal reproducibility, semantic verification tests relevance, and local refinement resolves boundary uncertainty.

Table 3. **Impact of progressive label curation.** Qwen3-VL-4B mIoU (%) under a fixed SFT recipe. Parentheses report stage-wise changes in the Qwen-anchored cascade. The cascade applies temporal consensus, semantic verification, and boundary refinement. Shading marks the Qwen anchor and final labels.

Label curation stage	Videos / QA (K)	Evaluation mIoU (%)							
		Charades^{TL}	ActivityNet^{TL}	QVHighlights^{TL}	VUE-TR†	VUE-TR-V2†	MomentSeeker‡	Ego4D-NLQ	Avg.
Raw Qwen3-VL-30B-A3B labels	28.9 / 735.4	51.7	50.9	60.1	47.4	43.2	24.9	15.7	42.0
Raw TimeLens-8B labels	33.6 / 999.2	54.4	51.5	59.2	48.5	40.3	24.7	15.7	42.0
Qwen-anchored curation cascade									
+ Temporal consensus	26.4 / 174.2	53.1 (+1.4)	52.9 (+2.0)	62.0 (+1.9)	49.5 (+2.1)	44.1 (+0.9)	25.8 (+0.9)	16.6 (+0.9)	43.4 (+1.4)
+ Semantic verification	23.8 / 93.2	53.3 (+0.2)	53.4 (+0.5)	63.0 (+1.0)	50.8 (+1.3)	45.1 (+1.0)	26.3 (+0.5)	16.6 (+0.0)	44.1 (+0.6)
+ Boundary refinement	**23.8 / 93.2**	**54.7** (+1.4)	**57.7** (+4.3)	**68.8** (+5.8)	50.3 (-0.5)	**46.0** (+0.9)	25.8 (-0.5)	**17.2** (+0.6)	**45.8** (+1.7)

Effect of long-context training. Longer contexts help only if the model can search them without absorbing more distractors. Holding the data, backbone, and optimization fixed, we vary the maximum packed length. Table 4 shows average mIoU rising from 44.4 at 16K to 46.4 at 100K, but unevenly: Charades and QVHighlights peak by 32K, whereas VUE-TR-V2, MomentSeeker, and Ego4D-NLQ gain 4.0, 1.5, and 3.3 points by 100K. VUE-TR's late 1.8-point jump from 64K to 100K further suggests a context-threshold effect for long-range examples. These asymmetries indicate that long-context training expands the effective temporal search horizon rather than uniformly benefiting every task. We therefore use 100K for final SFT.

Table 4. **Effect of long-context training.** Qwen3-VL-4B mIoU (%) with fixed supervised data and varying maximum packed length. Shading marks the final recipe.

Max. packed length	Charades^{TL}	ActivityNet^{TL}	QVHighlights^{TL}	VUE-TR†	VUE-TR-V2†	MomentSeeker‡	Ego4D-NLQ	Avg.
16K	56.1	57.0	68.2	49.1	41.6	24.3	14.5	44.4
32K	**56.3** (+0.2)	57.5 (+0.5)	**69.2** (+1.0)	48.8 (-0.3)	44.2 (+2.6)	25.1 (+0.8)	16.6 (+2.1)	45.4 (+1.0)
64K	56.1 (-0.2)	58.5 (+1.0)	**69.2** (+0.0)	49.1 (+0.3)	45.5 (+1.3)	25.7 (+0.6)	17.4 (+0.8)	45.9 (+0.5)
100K	**56.3** (+0.2)	**58.9** (+0.4)	**69.2** (+0.0)	**50.9** (+1.8)	**45.6** (+0.1)	**25.8** (+0.1)	**17.8** (+0.4)	**46.4** (+0.5)

Effect of instruction and response-format diversity. A generalist grounder should recover the same evidence regardless of request phrasing or timestamp serialization; otherwise, it may learn the interface rather than the task. Holding the TimeLens2-93K examples, Qwen3-VL-4B backbone, and SFT recipe fixed, we independently vary requests and renderings using the 27 instructions and 28 response specifications in Table 10.

Table 5 shows that rendering diversity alone raises average mIoU from 44.9 to 45.4, while request diversity reaches 45.0. Together they reach 45.8 (+0.9), perform best on six of seven benchmarks, and yield the largest gains on VUE-TR (+3.0) and QVHighlights (+1.3). The joint gain exceeds the sum of the isolated gains by 0.3 points. Thus, response-format diversity is the stronger standalone lever, but their positive interaction suggests that varying both sides of the interface helps disentangle evidence localization from surface realization and improves transfer across prompting protocols.

Table 5. **Effect of instruction and response-format diversity.** Qwen3-VL-4B mIoU (%) in a 2×2 ablation with fixed examples and SFT recipe. "Single" fixes the request or numeric interval-array rendering; "Diverse" samples from the full pools. Parentheses in Avg. show gains over the fully fixed baseline; shading marks the baseline and fully diverse setting.

Grounding request	Target rendering	Charades^{TL}	ActivityNet^{TL}	QVHighlights^{TL}	VUE-TR†	VUE-TR-V2†	MomentSeeker‡	Ego4D-NLQ	Avg.
Single	**Single**	54.3	57.0	67.5	47.3	45.8	25.7	16.6	44.9
Single	Diverse	**54.8**	57.4	68.5	50.0	45.4	25.3	16.3	45.4 (+0.5)
Diverse	Single	54.7	57.2	67.3	48.2	45.7	25.3	16.5	45.0 (+0.1)
Diverse	**Diverse**	54.7	**57.7**	**68.8**	**50.3**	**46.0**	**25.8**	**17.2**	**45.8** (+0.9)

Ablation on temporal reward design. R_{TW} is only as informative as the temporal distribution used to represent an interval set. We therefore ask whether the reward should preserve full interval occupancy or compress it into endpoints, centers, or boundaries. Holding all other RL settings fixed, including the parse penalty, we compare the four mass models in Figure 5. As shown in Table 6, every variant improves average mIoU over R_{tIoU}, from 47.0 to 47.4–47.7, confirming that distance-aware credit robustly complements overlap. Uniform support is the only variant that improves every benchmark and achieves the best average, 47.7.

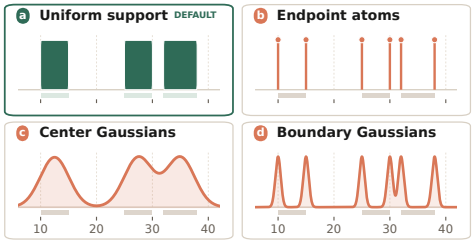

Figure 5. **Temporal mass models.** The same intervals $[[10, 15], [25, 30], [32, 38]]$ are converted into four probability distributions for the Wasserstein reward.

Table 6. **Ablation on temporal reward design.** mIoU (%) when varying only R_{temporal} in $R = R_{\text{temporal}} - \mathbf{1}_{\text{invalid}}$; all other RL settings are fixed. Parentheses in Avg. show gains over the tIoU baseline; shading marks the baseline and default uniform-support design.

Temporal reward	Temporal distribution	Charades^{TL}	ActivityNet^{TL}	QVHighlights^{TL}	VUE-TR†	VUE-TR-V2†	MomentSeeker‡	Ego4D-NLQ	Avg.
R_{tIoU}	–	57.4	58.1	67.2	52.0	47.8	27.7	18.5	47.0
$R_{\text{tIoU}} + R_{\text{TW}}$	**Uniform support**	**57.7**	**59.0**	69.3	53.2	**48.1**	27.9	18.6	**47.7** (+0.7)
$R_{\text{tIoU}} + R_{\text{TW}}$	Endpoint atoms	57.2	58.9	68.4	**53.9**	47.6	27.6	18.5	47.4 (+0.4)
$R_{\text{tIoU}} + R_{\text{TW}}$	Center Gaussians	57.4	**59.0**	**69.4**	53.2	47.4	**28.3**	18.8	47.6 (+0.6)
$R_{\text{tIoU}} + R_{\text{TW}}$	Boundary Gaussians	57.1	**59.0**	69.2	52.9	46.9	28.0	**18.9**	47.4 (+0.4)

Uniform support is more consistent because it preserves both where the evidence lies and how long it lasts, matching the span occupancy measured by mIoU. Endpoint atoms and boundary Gaussians emphasize transitions, while center Gaussians mainly encode location; each discards part of the interval interior or duration. Such summaries may favor individual datasets but are less stable across tasks. The broader lesson is that dense credit alone is insufficient: the reward representation should preserve the geometry of the evaluation object. We therefore use uniform support by default.

Comparison with matched NGIoU. NGIoU also provides graded feedback for non-overlapping intervals, making it a natural alternative to R_{TW}. We replace only R_{TW} with R_{NGIoU}, using MUSEG's one-to-one matching [25]. As shown in Table 7, NGIoU reaches 47.1 average mIoU, compared with 47.7 for temporal Wasserstein. The gap is larger on the multi-interval VUE-TR benchmarks: 1.4 points on VUE-TR and 1.0 on VUE-TR-V2. NGIoU depends on pairwise assignments, which can change when predictions split, merge, or differ in number. In contrast, R_{TW} compares merged temporal support directly and is invariant to equivalent fragmentations. Thus, dense feedback alone is insufficient; robust multi-interval rewards should avoid fragile correspondences.

Table 7. **Temporal Wasserstein vs. matched NGIoU.** We vary only the auxiliary dense reward; all other RL settings are fixed. R_{NGIoU} uses MUSEG's start-time-sorted one-to-one matching [25]. The Avg. gain is relative to matched NGIoU; shading distinguishes the comparator and our default reward.

Temporal reward	Charades^{TL}	ActivityNet^{TL}	QVHighlights^{TL}	VUE-TR†	VUE-TR-V2†	MomentSeeker‡	Ego4D-NLQ	Avg.
$R_{\text{tIoU}} + R_{\text{TW}}$	**57.7**	**59.0**	**69.3**	**53.2**	**48.1**	27.9	**18.6**	**47.7** (+0.6)
$R_{\text{tIoU}} + R_{\text{NGIoU}}$	**57.7**	58.8	67.7	51.8	47.1	**28.2**	18.5	47.1

Diagnostic ablation on zero-overlap misses. Average gains establish efficacy, but not mechanism: R_{TW} may help for reasons unrelated to its distance-aware signal. We therefore examine the 4,332 benchmark predictions from RL-tIoU that have no overlap with the target and group them by their

minimum temporal distance to it. If R_{TW} supplies the intended geometry, recovery should decay with distance.

Table 8. **Where does temporal Wasserstein help?** We stratify the 4,332 valid zero-tIoU predictions from RL-tIoU by distance to the target. Gap is normalized by total ground-truth duration; recovery is the fraction of corresponding examples on which adding R_{TW} yields positive tIoU. Shading follows the distance gradient.

Zero-overlap misses under R_{tIoU}			After adding R_{TW}	
Distance	Normalized gap	Cases	Recovered (%) ↑	mIoU (%) ↑
Near	$\leq 1\times$	644	**21.9**	**7.5**
Mid	1–$5\times$	808	15.8	4.0
Far	$> 5\times$	2,880	5.7	1.4
All zero-overlap misses		**4,332**	**10.0**	**2.8**

Table 8 shows a clear distance effect. Adding R_{TW} recovers positive overlap for 21.9% of near misses, 15.8% of mid-distance misses, and only 5.7% of far misses; mIoU similarly falls from 7.5 to 4.0 and 1.4. These zero-overlap cases also account for most of the overall improvement from 47.0 to 47.7 in Table 6. Thus, R_{TW} helps mainly where tIoU provides no signal, giving more credit to near misses than distant errors.

Diagnostic on group-relative reward structure. GRPO learns from within-group preferences, so a dense reward helps only if it breaks ties among rollouts. With tIoU, a group of non-overlapping predictions can receive uniformly zero reward and, after mean-centering, no temporal learning signal. To isolate the effect of R_{TW}, we score the same valid rollout groups collected over 300 training steps with both R_{tIoU} and $R_{\text{tIoU}} + R_{\text{TW}}$.

Table 9. **Does temporal Wasserstein create group-relative signal?** We score the same valid rollout groups collected over 300 training steps with both rewards. Constant groups contain only one reward value; zero-tIoU rescue is the fraction of all-zero-tIoU groups that become non-constant after adding R_{TW}. Advantages are mean-centered without variance scaling.

Scoring reward	Constant groups ↓	Distinct/group ↑	$\text{Var}_g(R)$ ↑	$\mathbb{E}_g\lvert A\rvert$ ↑	Zero-tIoU rescue ↑
R_{tIoU}	13.8%	4.63	0.023	0.098	–
$R_{\text{tIoU}} + R_{\text{TW}}$	**3.6%**	**6.04**	**0.091**	**0.190**	**75.8%**

As shown in Table 9, adding R_{TW} reduces constant-reward groups from 13.8% to 3.6% and rescues 75.8% of all-zero-tIoU groups. Beyond tie-breaking, distinct rewards per group rise from 4.63 to 6.04, within-group variance increases by $4.0\times$, and mean absolute advantage nearly doubles. Because rewards are mean-centered without variance scaling, this extra spread passes directly into the policy update. Temporal Wasserstein thus turns otherwise silent groups into ranked training signals, explaining its strong fit with group-relative optimization.

5 Conclusion

Video MLLMs become reliable interfaces to video archives only when their answers are traceable to supporting moments. This requires trustworthy evidence labels and an objective that guides imperfect interval predictions. TimeLens2 addresses both. TimeLens2-93K replaces brittle global annotation with evidence proposal, independent localization, consensus and semantic verification, and boundary refinement; the temporal Wasserstein reward supplies dense, matching-free geometry when overlap is uninformative. Across seven benchmarks, this design enables compact 2B, 4B, and 8B models to outperform much larger open-source models in long-video, multi-span, question-form, and egocentric settings. More broadly, treating evidence consistently as an interval set from supervision through optimization turns temporal grounding from a specialized retrieval task into a native, auditable capability of generalist video MLLMs.

References

[1] Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

[2] 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*, 2025.

[3] Ruizhe Chen, Zhiting Fan, Tianze Luo, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. Datasets and recipes for video temporal grounding via reinforcement learning, 2025. URL https://arxiv.org/abs/2507.18100.

[4] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

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

[6] 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, 2025. URL https://arxiv.org/abs/2510.23397.

[7] Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Zerun Ma, Yuxuan Qiao, Mo Li, Tianhao Liang, Lin Zhu, Amit Agarwal, Xiaozhe Li, Shengyuan Ding, Jiazi Bu, Ziyu Liu, Zhangyang Qi, Yifei Li, Yuhang Zang, Zhe Chen, Lin Chen, Yuan Liu, Yubo Ma, Hailong Sun, Yifan Zhang, Shiyin Lu, Tack Hwa Wong, Weiyun Wang, Peiheng Zhou, Chaoyou Fu, Junbo Cui, Jixuan Chen, Enxin Song, Song Mao, Junming Lin, Xilin Wei, Jinsong Li, Zeyi Sun, Zhaowei Wang, Zicheng Zhang, Xiaoyi Dong, Junjun He, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2026. URL https://arxiv.org/abs/2407.11691.

[8] 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.

[9] Google. Gemini 3: News and announcements. https://blog.google/products-and-platforms/products/gemini/gemini-3-collection/, 2025.

[10] Google DeepMind. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro, 2026.

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

[12] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language, 2017. URL https://arxiv.org/abs/1708.01641.

[13] 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.

[14] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant, 2024. URL https://arxiv.org/abs/2403.19046.

[15] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

[16] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval, 2020. URL https://arxiv.org/abs/2001.09099.

[17] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

[18] Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*, 2025.

[19] Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-opd: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation, 2026. URL https://arxiv.org/abs/2602.02994.

[20] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. URL https://arxiv.org/abs/2601.04720.

[21] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024.

[22] 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, 2025. URL https://arxiv.org/abs/2504.06958.

[23] Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, Xinyu Chen, Haoran Chen, Zizheng Huang, Jun Zhang, Kun Ouyang, Lin Sui, Ziang Yan, Yicheng Xu, Chenting Wang, Yinan He, Hongjie Zhang, Yi Wang, Yu Qiao, Yali Wang, Ziwei Liu, Kai Chen, and Limin Wang. Videochat3: Fully open video mllm for efficient and generalist video understanding, 2026. URL https://arxiv.org/abs/2607.14935.

[24] Chunxu Liu, Jiyuan Yang, Ruopeng Gao, Yuhan Zhu, Feng Zhu, Rui Zhao, and Limin Wang. Reasoning guided embeddings: Leveraging mllm reasoning for improved multimodal retrieval, 2025. URL https://arxiv.org/abs/2511.16150.

[25] Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. Museg: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35549–35561. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1644. URL https://aclanthology.org/2026.acl-long.1644/.

[26] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

[27] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, 2019. URL https://arxiv.org/abs/1906.03327.

[28] NemoStation. Marlin-2B: A tiny vlm to extract structured information from videos. Hugging Face model repository, 2026. URL https://huggingface.co/NemoStation/Marlin-2B. Accessed July 18, 2026.

[29] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning, 2024. URL https://arxiv.org/abs/2402.11435.

[30] 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.

[31] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *ArXiv*, abs/2312.02051, 2023.

[32] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

[33] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[34] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5026–5035, 2022.

[35] 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, 2019. URL https://arxiv.org/abs/1903.02874.

[36] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[37] Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

[38] Vidi Team, Celong Liu, Chia-Wen Kuo, Dawei Du, Fan Chen, Guang Chen, Jiamin Yuan, Lingxi Zhang, Lu Guo, Lusha Li, et al. Vidi: Large multimodal models for video understanding and editing. *arXiv preprint arXiv:2504.15681*, 2025.

[39] Vidi Team, Celong Liu, Chia-Wen Kuo, Chuang Huang, Dawei Du, Fan Chen, Guang Chen, Haoji Zhang, Haojun Zhao, Lingxi Zhang, et al. Vidi2: Large multimodal models for video understanding and creation. *arXiv preprint arXiv:2511.19529*, 2025.

[40] Chenting Wang, Yuhan Zhu, Yicheng Xu, Jiange Yang, Lang Lin, Ziang Yan, Yali Wang, Yi Wang, and Limin Wang. Internvideo-next: Towards general video foundation models without video-text supervision, 2026. URL https://arxiv.org/abs/2512.01342.

[41] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models, 2025. URL https://arxiv.org/abs/2410.03290.

[42] Jinwang Wang, Chang Xu, Wen Yang, and Lei Yu. A normalized gaussian wasserstein distance for tiny object detection. *arXiv preprint arXiv:2110.13389*, 2021.

[43] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[44] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding, 2025. URL https://arxiv.org/abs/2503.13377.

[45] Tao Wu, Li Yang, Gen Zhan, Yabin Zhang, Yiting Liao, Junlin Li, Deliang Fu, Li Zhang, and Limin Wang. Tempr1: Improving temporal understanding of mllms via temporal-aware multi-task reinforcement learning, 2026. URL https://arxiv.org/abs/2512.03963.

[46] LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

[47] Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, Limin Wang, Yu Qiao, and Yi Wang. Internvideo3: Agentify foundation models with multimodal contextual reasoning, 2026. URL https://arxiv.org/abs/2606.12195.

[48] 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.

[49] Feng Yue, Zhaoxing Zhang, Junming Jiao, Zhengyu Liang, Shiwen Cao, Feifei Zhang, and Rong Shen. Tempo-r0: A video-mllm for temporal video grounding through efficient temporal sensing reinforcement learning, 2025. URL https://arxiv.org/abs/2507.04702.

[50] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. In *International Conference on Learning Representations*, volume 2025, pp. 38057–38081, 2025.

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

[52] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

[53] Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026.

[54] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos, 2017. URL https://arxiv.org/abs/1703.09788.

[55] Yuhan Zhu, Guozhen Zhang, Jing Tan, Gangshan Wu, and Limin Wang. Dual detrs for multi-label temporal action detection, 2024. URL https://arxiv.org/abs/2404.00653.

[56] Yuhan Zhu, Xiangyu Zeng, Chenting Wang, Xinhao Li, Chunxu Liu, Yicheng Xu, Ziang Yan, Yi Wang, and Limin Wang. Freeret: Mllms as training-free retrievers, 2026. URL https://arxiv.org/abs/2509.24621.

[57] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos, 2019. URL https://arxiv.org/abs/1903.08225.

A Supplementary Material

Contents

A.1 Instruction and Response-Format Diversity

Temporal grounding requires both locating relevant evidence and expressing intervals through a specified interface. Training with a single prompt–response convention can entangle localization with surface form, reducing robustness when the request phrasing or output format changes.

For each QA turn, we keep the video, synthesized query q, and target interval set $\mathcal{Y}$ fixed, then independently and uniformly sample one of 27 grounding requests and one of 28 response specifications. The request pool varies how localization is posed, while the response pool controls answer syntax, timestamp encoding, and multi-interval composition. A deterministic renderer expresses the same $\mathcal{Y}$ under the sampled specification. As summarized in Table 10, this factorization defines 756 possible interface pairings without query paraphrases that could shift the intended evidence. It therefore encourages interface-invariant temporal semantics and format-conditioned serialization, while enabling the controlled 2×2 analysis in Table 5.

Table 10. Factorized interface diversity used during SFT. Independent sampling from 27 grounding requests and 28 response specifications defines 756 possible pairings while the query and target intervals remain fixed. Counts in each response axis partition the same 28 specifications; examples render two intervals.

Component	Variation	Count	Representative realization
Grounding request	Question, command, or explicit task schema	27	"When does q happen?"; "Locate the segments where q occurs."; "Query: q. Return matching intervals."
Response specification	Format-conditioned interval rendering	28	Specifies syntax, timestamp encoding, and rules for one or multiple intervals
Composition of the 28 response specifications			
Surface syntax	Natural language / labeled fields / bracketed arrays / bare intervals	12 / 5 / 6 / 5	`It happens from 12.0s to 18.5s, and 31.0s to 36.0s.`; `[[12.0, 18.5], [31.0, 36.0]]`
Timestamp encoding	Decimal seconds / minute–second / hour–minute–second	26 / 1 / 1	`12.0s`; `00:12`; `00:00:12`
Multi-interval composition	Format-consistent joining	28	Repeated clauses or interval tokens joined by conjunctions, punctuation, or an outer array

A.2 Full Temporal Grounding Results

The main paper reports R1@0.5 and mIoU; Table 11 adds R1@0.3 and R1@0.7 wherever available. The progression from loose to strict overlap separates coarse evidence retrieval from boundary precision: a model may find the correct event at R1@0.3 yet fail to delimit it at R1@0.7. Through its 4B and 8B variants, TimeLens2 achieves the highest reported R1@0.3 and R1@0.7 on all seven benchmarks. The 2B variant extends this evaluation to a smaller scale, reaching average R1@0.3, R1@0.5, and R1@0.7 scores of 57.9, 47.0, and 32.0, respectively.

Table 11. Full temporal grounding results across IoU thresholds.

Model	Metric	Charades^{TL}	ActivityNet^{TL}	QVHighlights^{TL}	VUE-TR†	VUE-TR-V2†	MomentSeeker‡	Ego4D-NLQ
GPT-5 [33]	R1@0.3	59.3	57.4	72.4	–	28.0	–	–
	R1@0.5	42.0	44.9	60.4	–	19.5	–	–
	R1@0.7	22.0	30.4	46.4	–	11.6	–	–
	mIoU	40.5	42.9	56.8	–	20.0	–	–
Gemini 3 Pro [9]	R1@0.3	–	–	–	–	51.5	–	–
	R1@0.5	–	–	–	–	41.3	–	–
	R1@0.7	–	–	–	–	26.0	–	–
	mIoU	–	–	–	–	39.7	–	–
Gemini 2.5 Pro [5]	R1@0.3	74.1	72.3	84.1	29.1	–	–	–
	R1@0.5	61.1	64.2	75.9	20.5	–	–	–
	R1@0.7	34.0	47.1	61.1	10.3	–	–	–
	mIoU	52.8	58.1	70.4	21.9	–	–	–
Qwen3.5-397B-A17B [37]	R1@0.3	69.3	70.6	82.1	52.4	45.6	34.3	20.5
	R1@0.5	47.8	59.1	71.6	42.7	35.5	22.3	13.3
	R1@0.7	24.9	41.7	56.1	33.1	25.5	9.9	7.1
	mIoU	47.5	53.8	65.8	42.2	34.5	23.3	14.5
Qwen3-VL-235B-A22B [2]	R1@0.3	71.7	69.0	79.6	55.8	44.1	34.4	17.8
	R1@0.5	50.8	57.5	70.2	42.5	34.5	21.2	11.5
	R1@0.7	24.5	39.3	54.5	31.6	24.1	11.3	6.3
	mIoU	47.8	52.2	64.6	43.1	34.1	23.7	12.7
Qwen3.5-35B-A3B [37]	R1@0.3	69.5	69.4	81.6	56.2	47.0	33.2	18.2
	R1@0.5	50.4	58.6	72.6	46.3	37.1	20.2	12.0
	R1@0.7	27.3	40.1	56.5	32.4	24.9	9.7	6.9
	mIoU	48.2	52.6	66.0	44.6	35.3	22.6	13.0
Qwen3-VL-30B-A3B [2]	R1@0.3	70.3	65.7	79.3	53.0	42.7	33.6	16.1
	R1@0.5	46.5	51.1	67.6	40.4	32.3	19.6	10.4
	R1@0.7	25.1	36.5	52.6	32.2	21.8	9.2	5.5
	mIoU	48.1	49.4	63.2	42.5	32.6	22.6	11.5
Vidi-1.5-9B [38]	R1@0.3	28.7	51.8	68.7	57.3	43.9	27.0	11.4
	R1@0.5	18.3	40.2	58.4	45.0	31.9	14.8	6.5
	R1@0.7	10.1	28.2	46.2	36.0	22.2	6.9	3.4
	mIoU	22.0	39.3	55.9	47.0	34.3	18.6	8.4
LLaVA-OneVision-2-8B [1]	R1@0.3	73.1	65.6	78.2	51.2	46.3	28.8	18.2
	R1@0.5	59.6	57.9	70.0	40.8	35.0	18.0	12.0
	R1@0.7	34.0	40.9	57.0	30.1	22.8	8.2	6.4
	mIoU	52.6	52.4	65.7	41.3	34.5	19.6	12.8
Qwen3-VL-8B [2]	R1@0.3	68.5	63.8	74.4	33.0	19.6	13.6	7.4
	R1@0.5	49.3	51.1	62.9	23.8	13.1	9.4	4.8
	R1@0.7	26.2	35.0	49.3	17.5	8.3	4.5	2.5
	mIoU	47.2	48.0	59.4	25.5	14.0	9.8	5.3
InternVideo3-8B [47]	R1@0.3	75.8	60.0	72.2	51.0	40.8	28.8	12.3
	R1@0.5	61.7	51.4	62.6	39.4	32.4	17.9	7.3
	R1@0.7	32.8	33.4	48.9	32.0	21.7	7.3	3.2
	mIoU	53.2	46.3	59.2	41.1	32.6	19.3	8.6
TimeLens-8B [53]	R1@0.3	78.8	71.8	81.2	56.4	46.2	32.0	21.9
	R1@0.5	65.8	62.0	71.7	42.9	34.4	20.4	14.6
	R1@0.7	37.6	43.8	57.1	31.2	23.5	9.9	7.7
	mIoU	57.0	56.3	66.8	43.6	34.0	21.9	15.7
Video-o3-7B [51]	R1@0.3	58.7	53.2	62.7	32.8	21.4	18.2	4.2
	R1@0.5	34.2	41.0	49.5	22.1	12.9	9.8	1.9
	R1@0.7	16.2	24.4	34.4	14.5	7.0	3.9	0.7
	mIoU	38.6	38.9	47.4	25.1	15.3	13.0	3.1
VideoChat-Flash-7B [21]	R1@0.3	60.2	35.5	45.2	20.2	13.7	10.9	3.1
	R1@0.5	37.9	21.8	30.6	15.4	9.5	5.9	1.5
	R1@0.7	17.8	10.5	16.7	10.7	5.6	2.2	0.6
	mIoU	39.7	24.8	32.7	17.3	10.1	7.2	2.5
MiMo-VL-7B [46]	R1@0.3	57.9	49.3	57.1	25.1	14.7	6.5	0.9
	R1@0.5	42.6	38.7	42.6	17.5	10.1	3.8	0.4
	R1@0.7	20.5	22.4	28.4	11.2	4.8	1.6	0.1
	mIoU	39.6	35.5	41.5	19.1	10.4	6.0	1.0
TimeSuite-7B [50]	R1@0.3	56.3	27.1	27.1	15.8	9.4	6.9	1.6
	R1@0.5	35.5	17.5	16.9	10.3	5.3	3.3	0.6
	R1@0.7	18.0	8.6	9.9	7.0	3.2	1.2	0.3
	mIoU	38.1	19.8	21.7	13.2	7.3	5.9	1.4
InternVL3.5-4B [43]	R1@0.3	24.6	20.0	25.5	20.4	11.0	3.7	2.0
	R1@0.5	14.1	12.7	15.8	14.1	6.0	2.1	0.7
	R1@0.7	6.2	8.2	6.2	9.5	3.5	0.6	0.1
	mIoU	16.0	14.9	17.7	16.5	8.5	3.2	2.0
Qwen3-VL-4B [2]	R1@0.3	71.8	67.4	78.4	41.0	25.6	21.5	15.5
	R1@0.5	53.8	54.8	66.4	33.0	19.9	13.5	10.6
	R1@0.7	28.7	37.6	51.1	25.0	13.3	5.4	6.0
	mIoU	49.4	50.7	62.3	33.6	20.3	15.3	11.4
Molmo2-4B [4]	R1@0.3	44.4	50.8	73.7	52.0	41.7	25.2	11.8
	R1@0.5	31.1	40.3	62.6	42.5	30.9	14.7	7.7
	R1@0.7	18.5	29.9	51.8	33.7	20.8	8.0	4.5
	mIoU	34.7	40.9	60.8	43.1	32.1	19.5	10.0
VideoChat3-4B	R1@0.3	78.4	70.1	81.0	61.0	53.8	37.2	20.9
	R1@0.5	64.9	60.5	72.5	47.8	40.4	24.4	13.9
	R1@0.7	35.9	42.2	56.8	34.7	27.3	12.0	7.2
	mIoU	56.1	54.6	67.0	47.9	40.2	25.9	14.6
Marlin-2B [28]	R1@0.3	66.3	50.5	60.7	31.6	24.5	23.9	12.7
	R1@0.5	51.7	40.5	48.5	23.6	19.9	15.2	7.8
	R1@0.7	27.1	25.3	34.9	16.4	12.0	5.8	3.8
	mIoU	46.5	37.9	46.8	24.8	19.2	16.3	8.9
Qwen3.5-2B [37]	R1@0.3	67.3	64.5	74.8	34.7	28.7	22.2	15.9
	R1@0.5	46.8	52.8	65.7	27.2	21.5	12.8	10.1
	R1@0.7	24.5	35.9	50.5	17.0	13.7	6.1	5.1
	mIoU	46.0	48.9	60.6	27.8	21.6	15.6	11.2
Qwen3-VL-2B [2]	R1@0.3	62.7	53.7	67.8	41.9	33.7	17.3	13.3
	R1@0.5	44.4	41.0	53.4	29.9	22.5	9.8	8.1
	R1@0.7	23.4	26.5	39.3	18.5	13.8	3.2	3.8
	mIoU	43.4	39.9	52.2	31.4	24.1	12.4	9.3
TimeLens2-2B	R1@0.3	74.2	71.1	81.1	64.2	57.5	34.1	23.3
	R1@0.5	60.6	62.0	72.8	50.7	44.5	21.9	16.5
	R1@0.7	33.6	43.2	57.8	38.9	31.5	9.8	9.1
	mIoU	53.4	55.4	67.0	51.1	43.8	24.3	16.8
TimeLens2-4B	R1@0.3	80.3	74.6	83.3	66.5	62.9	38.8	26.0
	R1@0.5	66.6	65.5	75.2	52.4	49.5	27.0	18.1
	R1@0.7	37.5	47.2	60.0	41.1	35.5	12.5	10.0
	mIoU	57.7	59.0	69.3	53.2	48.1	27.9	18.6
TimeLens2-8B	R1@0.3	80.9	74.0	84.4	65.3	61.8	40.8	27.0
	R1@0.5	68.0	65.6	76.1	51.6	49.1	27.3	18.4
	R1@0.7	38.9	46.6	61.2	41.9	34.8	13.7	9.6
	mIoU	58.6	58.6	70.2	53.5	47.7	28.5	19.0

A.3 Additional Qualitative Examples

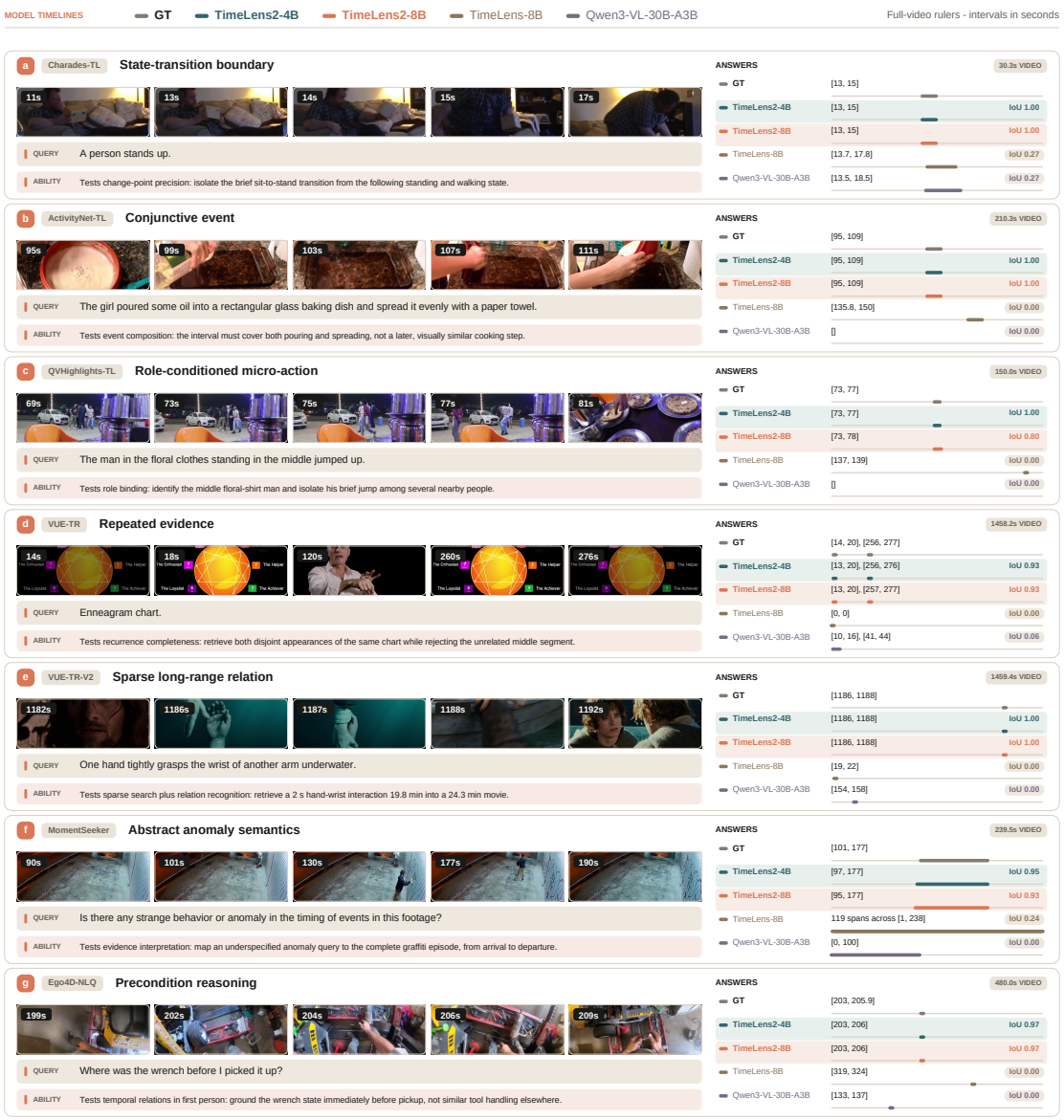

Figure 6. Additional qualitative results on all seven benchmarks. Each row pairs five frames near the target evidence with the query, the temporal reasoning challenge, and full-video timelines showing the ground truth and predictions from four models. The TimeLens2 variants consistently recover the intended evidence, whereas the baselines select distractors, miss events, overextend boundaries, or fragment their predictions.

These examples probe complementary temporal-reasoning capabilities. Charades-TL tests precise state-transition boundaries; ActivityNet-TL requires one interval covering both actions in a conjunction; QVHighlights-TL tests role-conditioned entity binding; and VUE-TR requires complete retrieval of recurring evidence. VUE-TR-V2 adds sparse search for a two-second relation nearly 20 minutes into a film. MomentSeeker requires mapping an underspecified anomaly query to the complete graffiti episode, rather than broad or fragmented coverage. Ego4D-NLQ tests precondition reasoning by locating the wrench immediately before pickup. Across all settings, TimeLens2 remains precise, complete, and resistant to visually similar distractors.

A.4 Temporal Grounding Evaluation Protocol

Benchmarks and evaluation subsets. We conduct all in-house evaluations with VLMEvalKit [7]. Within this framework, we implement evaluation adapters for the TimeLens-Bench re-annotations of Charades-STA, ActivityNet Captions, and QVHighlights; the vision-query subsets of VUE-TR and VUE-TR-V2; the text-only query subset of MomentSeeker; and the official Ego4D-NLQ-v2 validation split. We use visual frames only: audio and subtitles are not supplied to the model.

Standard video input configuration. For models evaluated, the input configuration is fixed per benchmark rather than tuned per model; the exact constructor arguments are given in Table 12. The video processor receives `min_pixels` and `max_pixels` as per-frame pixel-area bounds and `total_pixels` as the aggregate visual pixel budget. We first sample at the specified frame rate. If $\lfloor T \times \mathrm{fps} \rfloor$ exceeds `frames_limit`, we uniformly retain the capped number of frames over the entire duration T, and pass the resulting effective sampling rate to the model.

Model-specific input settings. All evaluated Qwen3-VL and Qwen3.5 variants, LLaVA-OneVision-2-8B, TimeLens-8B, VideoChat3-4B, InternVideo3-8B, Molmo2-4B, and TimeLens2-2B/4B/8B use the standard per-benchmark settings in Table 12. Every deviation is summarized in Table 13 by model and dataset group. These deviations are used when a model cannot accommodate the standard context or frame budget. All Qwen3.5 models except Qwen3.5-2B are evaluated with thinking enabled, which substantially improves temporal-grounding performance. For Qwen3.5-2B, we disable thinking because it frequently enters non-terminating reasoning loops, especially on long-video benchmarks. Thinking is disabled for every other evaluated model. Because Molmo2 follows generic instructions and prescribed output formats unreliably, we adopt the native spatiotemporal-grounding question template from its original work. Specifically, we use the prompt "Point the start and end of the event described by the sentence *query*." and parse the first two temporal coordinates from its native `<points coords="...">` or `<tracks coords="...">` response.

Prompt templates. The evaluation entry point uses the dataset-side prompts in Table 14.

A.5 Train–Test Overlap Audit

We canonicalized the complete list of training-source YouTube IDs and compared it with the video identifiers used by all seven evaluation subsets. Specifically, we removed the `v_` prefix from ActivityNet IDs, reduced QVHighlights clip names to their 11-character source IDs, and used the native YouTube IDs provided by VUE-TR, VUE-TR-V2, and the explicitly identified YouTube videos in MomentSeeker. We additionally searched each benchmark's `video`, `video_id`, and `src_video_path` fields for embedded training IDs. The intersection was *empty* for every benchmark. Charades and Ego4D use non-YouTube identifiers, while most MomentSeeker videos are distributed under renamed internal identifiers without a source-ID mapping; thus, this audit establishes *zero* overlap among recoverable source identifiers but cannot exclude content duplicates concealed by renaming.

Table 12. Standard input configuration. Exact arguments for the seven temporal grounding benchmarks.

Benchmark	Queries N	Temporal sampling		Per-frame area		Total area	Output
		FPS	`frames_limit`	`min_pixels`	`max_pixels`	`total_pixels`	
Charades-TimeLens	3,363	4	2,048	32^2	640^2	128000×32^2	one interval
ActivityNet-TimeLens	4,500	4	2,048	32^2	640^2	128000×32^2	one interval
QVHighlights-TimeLens	1,541	4	2,048	32^2	640^2	128000×32^2	one interval
VUE-TR (vision)	525	1	2,048	32^2	480^2	128000×32^2	all intervals
VUE-TR-V2 (vision)	715	1	2,048	32^2	480^2	128000×32^2	all intervals
MomentSeeker (text-only)	1,000	2	2,048	32^2	480^2	128000×32^2	all intervals
Ego4D-NLQ-v2 (val)	4,552	2	2,048	32^2	480^2	128000×32^2	one interval

Table 13. Model-specific overrides. Input configurations that deviate from Table 12. TL denotes the three TimeLens-Bench datasets.

Model	Dataset group	FPS	Frame cap	Min. `min_pixels`	Max. `max_pixels`	Total area `total_pixels`
Global overrides						
Video-o3-7B / MiMo-VL-7B	All seven	2	768	10×28^2	768×28^2	24576×28^2
VideoChat-Flash-7B	All seven	~1	512	448^2	448^2	–
TimeSuite-7B	All seven	–	128	224^2	224^2	–
Dataset-specific override						
InternVL3.5-4B	TL	1	32	28^2	448^2	12800×32^2
	VUE-TR	1	128	28^2	448^2	12800×32^2
	VUE-TR-V2 / MomentSeeker	1	128	32^2	480^2	12800×32^2
	Ego4D-NLQ	1	64	32^2	480^2	12800×32^2

Table 14. Dataset-side prompt templates. Exact templates used in the standard temporal grounding evaluation; `{query}` is replaced by the benchmark query.

Benchmark group	Prompt template
Charades-/ActivityNet-/ QVHighlights-TimeLens	`Please find the visual event described by the sentence '{query}', determining its starting and ending times. The format should be: 'The event happens in <start time> - <end time> seconds'.`

VUE-TR / VUE-TR-V2	`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]].`

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

Ego4D-NLQ-v2 (val)	`You are given a video.`
	`Task: temporal grounding.`
	`Given the query: "{query}", return the time span (in seconds) where the query is grounded in the video.`
	`Output format MUST be [[start, end]].`

A.6 Limitations and Future Directions

Limitations. TimeLens2-93K was constructed under finite financial and computational budgets. We therefore relied on cost-effective, predominantly open-weight models throughout the pipeline: Qwen3-VL-235B-A22B-Instruct for captioning, Kimi-K2.5 for query and proposal generation, Qwen3-VL-30B-A3B and TimeLens-8B for temporal annotation, and Qwen3-VL-235B-A22B-Instruct for final refinement. Although this pipeline already produces effective training data, its quality is inevitably bounded by the capabilities of these models. Replacing them with stronger proprietary multimodal models, such as Gemini 3.1 Pro [10], could substantially improve the richness, diversity, and temporal precision of the resulting annotations. We therefore believe that the current dataset is not the endpoint of this recipe, but a first demonstration with considerable room to scale.

Future directions. Beyond improving temporal grounding data with stronger models, our construction framework may naturally extend to long-context spatiotemporal grounding, where a model must determine not only *when* an event occurs, but also *where* relevant entities appear and how they evolve over time. This capability is particularly important for embodied intelligence: an agent operating over long horizons must connect past observations with the current scene, track changes in objects and environments, and ground its decisions in concrete visual evidence before completing downstream tasks. We hope our work can inspire larger and richer spatiotemporally grounded datasets, as well as models that turn long-video perception into persistent memory, grounded reasoning, and ultimately reliable action.