TickTock: Teaching and Evaluating Video LLMs to Tell Time at Frame-Level Precision

Anonymous authors
Paper under double-blind review

Abstract

Recent advanced video large language models show temporal awareness: they can relate events to when they occur and answer multi-hop temporal questions. Despite this potential, they remain far from frame-level precision, especially for complex events whose boundaries unfold within a fraction of a second. A key obstacle is that their training labels and benchmark references are mostly recorded in whole seconds, so neither supervision nor evaluation can tell whether a prediction points to the right frame. In this work, we introduce **TickTock**, a frame-anchored annotation framework for teaching and evaluating frame-level temporal grounding. Instead of guessing a timestamp, the annotator, either a human or advanced MLLMs, selects the boundary frame from a numbered filmstrip, and returns its exact timestamp, so every label is traceable to a real frame. Building on this framework, we construct **TickTock-5K**, a training set of 5,000 videos spanning 11 domains from short videos and sports to lectures and screen tutorials, and release refined annotations for the *TimeLens-100K* and *TimeLens2-93K* training corpora. For evaluation, we build **TickTock-Refined**, a frame-level re-annotation from *TimeLens-Bench* and *VUE-TR*, complemented by **TickTock-Point500**, 500 human-annotated moments snapped to decoded frames. Across 28 open and closed video-LLMs, the strongest model places 79.2% of moments within five seconds but no model exceeds 21.6% within 0.1 s, a gap that IoU-based metrics largely conceal. Frame-anchored references reorder models and track human point accuracy more closely. Together, TickTock and its data and evaluation protocol lay a foundation for frame-level temporal grounding in Video-LLMs.

1 Introduction

Video LLM have advanced rapidly, from early instruction-tuned systems (Maaz et al., 2024; Zhang et al., 2023; Liu et al., 2023) to general models such as Qwen3-VL (Bai et al., 2025) and Gemini (Comanici et al., 2025; Google, 2025) that describe, answer questions about, and reason over long videos. A growing line of work equips them with temporal grounding, the ability to say *when* a described event happens. Specialized models such as TimeChat, VTimeLLM, LITA, VTG-LLM, and Grounded-VideoLLM learn temporal tokens or timestamp outputs (Ren et al., 2024; Huang et al., 2024a;b; Guo et al., 2025; Wang et al., 2025a); reinforcement-learning approaches further sharpen localization (Wang et al., 2025b; Chen et al., 2025); and benchmarks built on Charades-STA, ActivityNet Captions, and QVHighlights (Gao et al., 2017a; Krishna et al., 2017; Lei et al., 2021a), recently curated as TimeLens-Bench (Zhang et al., 2026) and extended to long videos by VUE-TR (Team et al., 2025a;b), measure progress by interval overlap. However, this progress is measured at roughly one-second resolution. Models typically sample one to a few frames per second, and the references they are trained and evaluated on are almost always stored in whole seconds (98–100% of endpoints in TimeLens-Bench and VUE-TR). The sub-second digits that models output are therefore largely estimates, and at times *hallucinations*, rather than times read off an observed frame. Yet a video is a sequence of frames, and nothing between two consecutive frames is ever recorded; the frame is therefore the **finest** temporal resolution a video can support, typically 1/30 to 1/60 of a second. In this work, we raise the ground truth of temporal grounding from whole seconds to this limit, *frame-level accuracy*, anchoring every label and reference to a decoded frame. This precision matters for real-world understanding: recovering how an object moves, when a hand first touches a tool, or when a risk begins requires knowing the exact frame, not the nearest second.

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.

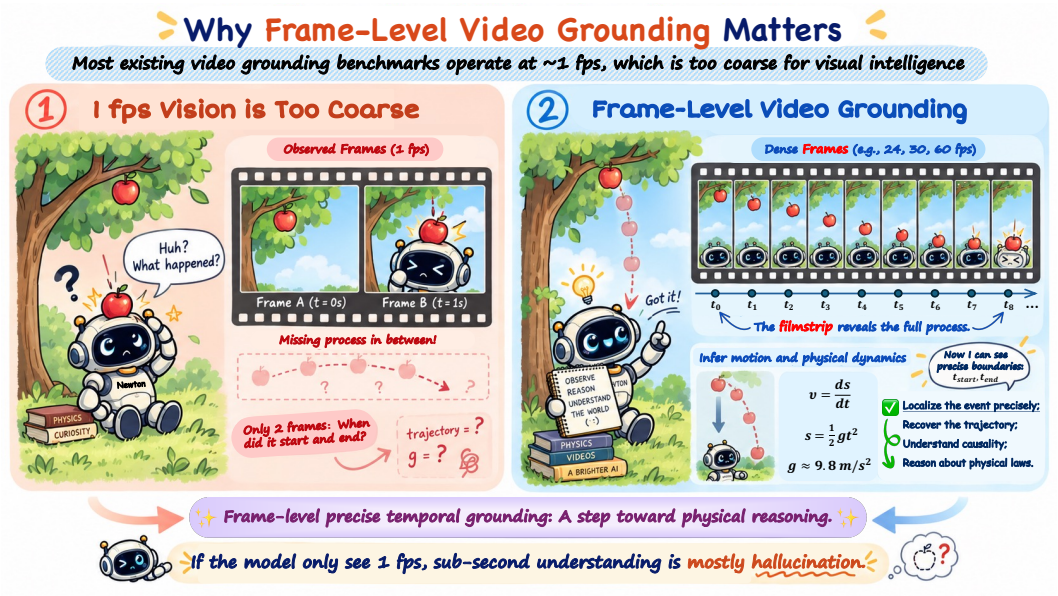

Figure 1: Seeing an event and locating its boundary require different temporal evidence. Sparse snapshots show the apple before and after it falls; denser frames reveal the motion and contact that determine the event boundary.

Figure 1 illustrates why this is hard. Imagine Newton watching an apple fall, but seeing the world at one frame per second. An apple dropping two meters reaches the ground in about 0.64 s, so he would see it on the branch and then on the ground, with at most one frame in between. He could say roughly when it fell, but could not recover its trajectory, let alone estimate $g \approx 9.8\,\mathrm{m/s^2}$ from $s = \frac{1}{2}gt^2$. Now hand him a filmstrip of the same fall at 30 fps: with consecutive frames laid side by side, each showing the apple a little lower, with the exact frame where it touches the ground. From these he can read off position against time and recover g. Three problems stand in the way. *(i) Sparse observation.* At the one to a few frames per second that models and benchmarks typically use, the motion and the moment of contact happen between observed frames, so any sub-second answer is a guess. *(ii) Ambiguous coordinates.* Even with dense frames, a timestamp such as 3.42 s identifies the moment of contact only if we know which decoded frame it names; frame rates, stream offsets, and clip origins can map the same number to different frames, and existing labels are mostly stored in whole seconds. *(iii) Insensitive evaluation.* Interval overlap (IoU), the standard metric, barely changes when a boundary moves by a few frames, so a benchmark cannot tell a frame-accurate answer from one that is only roughly right.

To bridge this gap, we introduce **TickTock**, a coarse-to-fine, frame-anchored annotation framework for teaching and evaluating frame-level temporal grounding, which targets each problem in turn. It first takes a second-level boundary, either from an existing label or from an MLLM, and then refines it on the frames themselves. Against sparse observation, the video around that coarse boundary is decoded at its native frame rate and every frame within ± 1 s is laid out, in order and with a visible ID, on a numbered filmstrip, exactly the view we would hand Newton. Against ambiguous coordinates, the annotator, an MLLM or a human, never writes a time: it selects the boundary frame's ID, and the video decoder returns that frame's exact presentation timestamp, which removes transcription errors and offset mismatches by construction (Figure 2). If the event runs past the window or the choice falls on its edge, the filmstrip is rebuilt over ± 3 s with more pages rather than smaller frames, so precision does not cost coverage. Against insensitive evaluation, references are anchored to frames and complemented by point-level targets scored by absolute time error. Because an MLLM can serve as the annotator, the framework scales from benchmark references to training corpora. With TickTock we build four resources: **TickTock-5K**, 5,000 videos from 11 domains for supervised fine-tuning of temporal grounding, including queries with repeated occurrences; refined annotations for the *TimeLens-100K* and *TimeLens2-93K* training corpora, whose original endpoints are whole seconds, covering 169,176 intervals; **TickTock-Refined**, frame-level references for 8,019 queries from TimeLens-Bench and the visual subsets of VUE-TR, a drop-in replacement for evaluating interval grounding at frame precision; and **TickTock-Point500**, 500 human-annotated moments

snapped to decoded frames, for evaluating when a specific moment occurs, such as the trigger for a proactive assistant.

Our comprehensive evaluation of 28 open and closed Video-LLMs shows that current models find the right second but seldom the right frame. The strongest model places 79.2% of human moments within five seconds, yet in our nine-model diagnostic cohort no model exceeds 21.6% within 0.1 s. Among predictions that IoU counts as correct (≥ 0.5), only 16–24% of boundaries fall within 0.1 s of the frame-anchored reference. Part of the reported performance does not come from the video at all: given only the filename and the query, ***without the video input***, all nine models we tested still obtain 7–17 mIoU on Charades. Rescored on TickTock-Refined, the same predictions change rank, and the resulting rankings agree more closely with human point accuracy.

Our contributions are as follows:

- We propose **TickTock**, an annotation framework that raises temporal ground truth from whole seconds to frame-level accuracy; because annotators select frames instead of writing times, every label is traceable to a decoded frame, free of transcription and coordinate errors, and the same protocol scales from human to MLLM annotators.
- We construct **TickTock-5K**, a source-balanced SFT dataset of 5,000 videos across 11 domains (39,530 queries, 50,295 intervals), and release refined annotations for the *TimeLens-100K* and *TimeLens2-93K* corpora (169,176 intervals), both for supervised fine-tuning of interval grounding, including events that occur more than once.
- We build two benchmarks: **TickTock-Refined**, a drop-in, frame-anchored replacement for TimeLens-Bench and VUE-TR to evaluate interval grounding at frame precision, and **TickTock-Point500**, 500 human-annotated moments to evaluate when a specific moment occurs, such as the trigger for a proactive assistant.
- Evaluating 28 open and closed Video-LLMs, we find that models localize seconds but not frames (at best 79.2% of moments within 5 s), that IoU conceals this gap and credits answers produced without the video, and that frame-anchored references yield rankings that better agree with human point accuracy.

2 Related Work

Temporal grounding benchmarks. Temporal grounding asks a model to localize the moment described by a query (Gao et al., 2017b; Zhang et al., 2020), with later formulations covering multiple moments (Lei et al., 2021b) and event boundaries (Wang et al., 2022). Standard benchmarks build on Charades-STA, ActivityNet Captions, and QVHighlights (Gao et al., 2017a; Krishna et al., 2017; Lei et al., 2021a); TimeLens-Bench re-curates these three (Zhang et al., 2026), and VUE-TR extends evaluation to long videos with repeated events and official modality tags (Team et al., 2025a;b). These benchmarks score predictions by interval overlap against references stored almost entirely in whole seconds. TickTock-Refined keeps their queries but re-annotates the references at frame level, and TickTock-Point500 adds point targets scored by absolute time error.

Temporal grounding in Video-LLMs. Video-LLMs acquire temporal grounding through timestamp-aware instruction tuning, as in TimeChat, VTimeLLM, LITA, VTG-LLM, and Grounded-VideoLLM (Ren et al., 2024; Huang et al., 2024a;b; Guo et al., 2025; Wang et al., 2025a), and increasingly through reinforcement learning with IoU-based rewards (Wang et al., 2025b; Li et al., 2025; Wu et al., 2026; Chen et al., 2025). Controlled studies compare how models should express time, as text numerals, special tokens, or continuous outputs (Jin et al., 2026), and TimeRefine learns to correct boundaries coarse-to-fine (Wang et al., 2026). All of these methods are trained and evaluated against second-level labels; TickTock provides frame-level ground truth for both.

Frame indexing and temporal supervision. NumPro overlays frame numbers on each frame so that a model can refer to moments by index, and uses them both for prompting and for fine-tuning (Wu et al., 2025). We tried the analogous approach, burning timestamps into the video and asking Gemini 3.1 Pro Preview to ground 523 visual VUE-TR queries. The overlay did not help: mIoU moved from 41.92 to 41.37, and the number of empty answers rose from 15 to 52. A plausible reason is that videos with printed timestamps are rare in MLLM training data, so models do not learn

to read the overlay, and video interfaces further resample and downscale frames until small labels are hard to see; the model therefore still estimates the time rather than reading it. TickTock takes a different route: instead of annotating the video, it lays out every native frame of a local window, each with an enlarged ID, on a static filmstrip passed to the MLLM as an image, lets the model select a frame, and resolves that frame's time through the decoder, so the answer is a decoded frame rather than a transcribed number. For training data, LLaVA-Video synthesizes video instructions at scale (Zhang et al., 2025), Seq2Time derives temporal labels from sequence positions (Deng et al., 2025), and VideoTG-R1 filters incomplete annotations (Dong et al., 2025). TickTock-5K instead keeps each label on its source video's own time axis, retains every occurrence of a query, and records the provenance of every interval.

3 TickTock: Frame-Anchored Data And Benchmarks

3.1 Coarse-To-Fine Frame Anchoring

TickTock represents every temporal coordinate by a decoded frame. A video stream provides, for each native frame j, an integer presentation timestamp p_j, a rational time base τ, and a stream start p_0, so a frame index determines its time exactly:

$$t_j = (p_j - p_0)\,\tau, \qquad t_j^{\text{local}} = t_j - t_0, \tag{1}$$

where t_0 is the origin of a clip or annotation window. Annotation proceeds from coarse to fine (Figure 2). A second-level boundary is first obtained, either from an existing label or from an MLLM. The video around it is then decoded at its native frame rate, and every frame within ± 1 s is laid out in order on a numbered filmstrip (contact sheet), 24 frames per sheet. The annotator, an MLLM or a human, selects the start and end frame IDs, and Eq. 1 returns their times. If the annotator reports that the event continues beyond the window, picks an edge frame, or is not confident, the window widens to ± 3 s at the same frame rate, using more sheets rather than smaller frames. The instruction states the boundary rule, in the example the first and last illuminated frames, so each decision can be checked against its unlit neighbors. We store the native index, PTS, time base, stream start, and clip origin with every label, and reject unknown IDs, reversed intervals, and out-of-video coordinates.

Two design choices follow from failures we observed. First, the time must come from the same decoded frame as the pixels: an earlier OpenCV-based pipeline paired some images with the preceding frame's timestamp, an error that a frame-rate grid check cannot detect. Second, the annotator reads a filmstrip image rather than a timestamped video, so that every candidate frame is shown once at full cell size (Appendix A.1).

3.2 Training Data

TickTock-5K is a supervised fine-tuning set for interval grounding. It contains 5,000 source videos, each with its full event list, drawn from 14 MLLM event-discovery collections that we group into 11 content categories spanning daily and creator videos, sports, film commentary, lectures, and screen tutorials (Figure 3). Together they hold 39,530 queries and 50,295 intervals, and each query keeps *all source-annotated* occurrences, so the set also supervises events that recur within a video. Records must have finite, ordered, nonnegative endpoints, one video reference, and a well-formed query. Videos are identified by URI or source ID across categories, and a video found in several is kept only in the scarcest one. Within each category, videos are ranked by a seeded hash and the selection takes a prefix of that order: the seven large categories contribute 589–590 videos each, action highlights 500, and the three small ones (GUI tutorials, software walkthroughs, and basketball games, 372 videos) are kept whole. Every example stores its source file, line, record hash, and video identity, so it can be traced to its raw annotation and extended without re-ranking. The inherited event timestamps retain their source precision; this selection does not establish frame-level boundary accuracy.

We also release **refined annotations for the TimeLens training corpora**. The released endpoints of TimeLens-100K (Zhang et al., 2026) and TimeLens2-93K (Zhu et al., 2026) are all integer seconds. We provide MLLM-refined boundaries for 169,176 intervals (85,936 and 83,240), kept side



Figure 2: Coarse-to-fine frame anchoring, read from ① to ④. ① Sources: multi-domain videos with MLLM event lists are deduplicated and balanced into TickTock-5K; each label provides a second-level boundary. ② Decode: around that boundary, every native frame is decoded and laid out on a numbered filmstrip, shown here for a speaker whose light ring switches on and off. ③ Select: the annotator, an MLLM or a human, returns the start and end frame IDs (#7272, #7469), the first and last lit frames; if the event runs past the ± 1 s window, the filmstrip is rebuilt over ± 3 s. ④ Resolve: the decoder maps each ID to its presentation timestamp, giving $[145.440, 149.380]$ s, and invalid IDs or intervals are rejected. The frame-ID lookup is used for TickTock-Refined; the current TickTock-5K release retains the source event timestamps.

by side with the originals; after refinement, only 14.1% and 6.1% of endpoints remain whole seconds (Appendix Figure 18). These refinements come from an earlier timestamped-sheet version of the pipeline, so we mark them as model-refined rather than frame-anchored. Queries with partially retained occurrences are flagged rather than dropped. Appendix B gives per-category counts and checks.

3.3 Benchmarks And Metrics

TickTock-Refined evaluates interval grounding with frame-anchored references while preserving the query text and video for each retained example. It re-annotates TimeLens-Bench and the visual-query subsets of VUE-TR v1/v2 with the protocol above, retaining 8,019 of 10,693 queries after boundary, content, and modality screening: 2,498 on Charades, 3,384 on ActivityNet, 1,227 on QVHighlights, and 369 and 541 on the two VUE versions (Table 4). VUE provides official modality tags, and we keep only visual queries, since a filmstrip cannot verify an acoustic boundary; TimeLens has no such tags, so Gemini screening followed by GPT-6 Astra review removes 50 audio-dependent queries and 42 unresolved ones. The references and the screening are model-produced and PTS-validated rather than verified by humans query by query. Refined and original scores use the same predictions on the same queries; on VUE-TR-V2, the 425 retained queries with matching archived inputs are scored with the official interval-set scorer (Appendix C.2).

TickTock-Point500 evaluates when a specific moment occurs, such as the trigger for a proactive assistant. It contains 500 distinct videos (about 24 hours), each with one query and a human reference moment from the Omni Interaction collection. From 1,500 recorded interactions on 1,128 videos, 691 videos are visual-only candidates, and we fix one query on each of 500 videos across five task types (Figure 5). The target follows the request: an onset, a completion, a recurrence, or a trigger. Explicit answer times are removed from 34 queries, while on-screen clock readings that define the

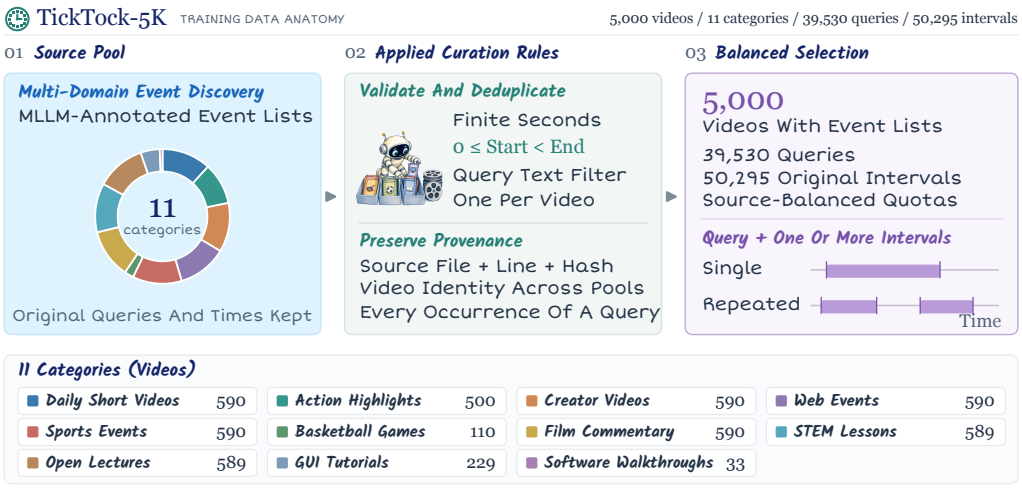

Figure 3: TickTock-5K, read left to right. **01** Source pool: MLLM-annotated event lists from 11 content categories, shown as a ring. **02** Curation: endpoints are validated, each video is kept once across collections, and every label keeps its source file, line, and hash. **03** Selection: source-balanced quotas yield 5,000 videos with 39,530 queries and 50,295 intervals, where a query may have one or several occurrences. The bottom row lists the number of videos in each category.

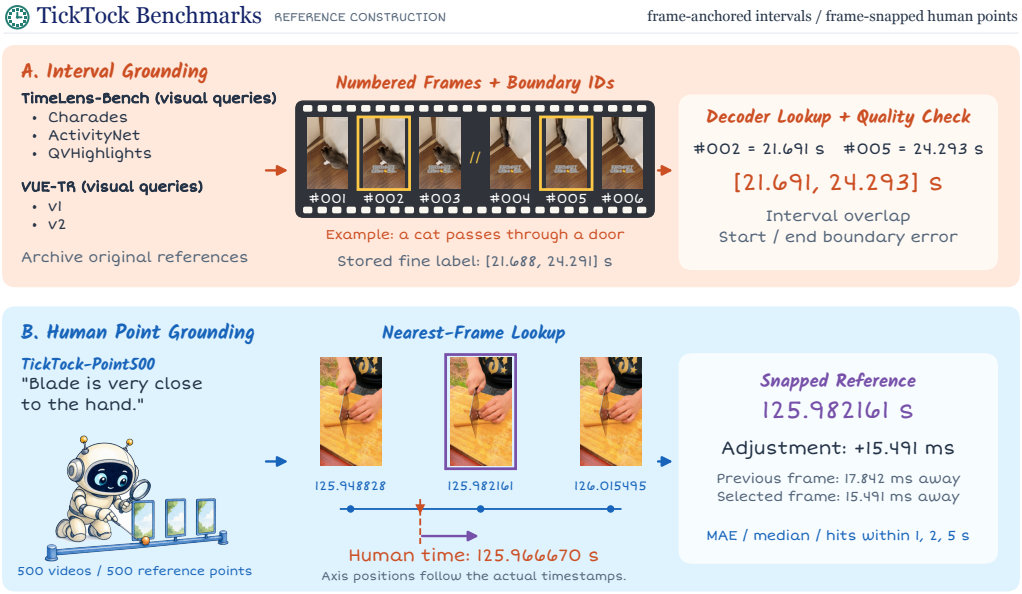

Figure 4: Frame-anchored benchmark references, each row read left to right. (A) TickTock-Refined: an interval query from the source benchmark is laid out on numbered frames, the annotator selects the boundary IDs (#002 and #005), and the decoder maps them to $[21.691, 24.293]$ s, replacing the stored label $[21.688, 24.291]$ s. (B) TickTock-Point500: a human-annotated moment at 125.966670 s is snapped to the nearest decoded frame at 125.982161 s, a shift of 15.491 ms, without changing what the annotator meant. Each record keeps its original annotation and decoder mapping.

target are kept. We keep the human interpretation and only anchor its coordinate: for human time t_i^{human} and decoded frame times $\{\tau_{ij}\}_j$, the reference is the nearest frame,

$$t_i = \tau_{ij_i^*}, \qquad j_i^* = \arg\min_j \left|\tau_{ij} - t_i^{\text{human}}\right|, \tag{2}$$

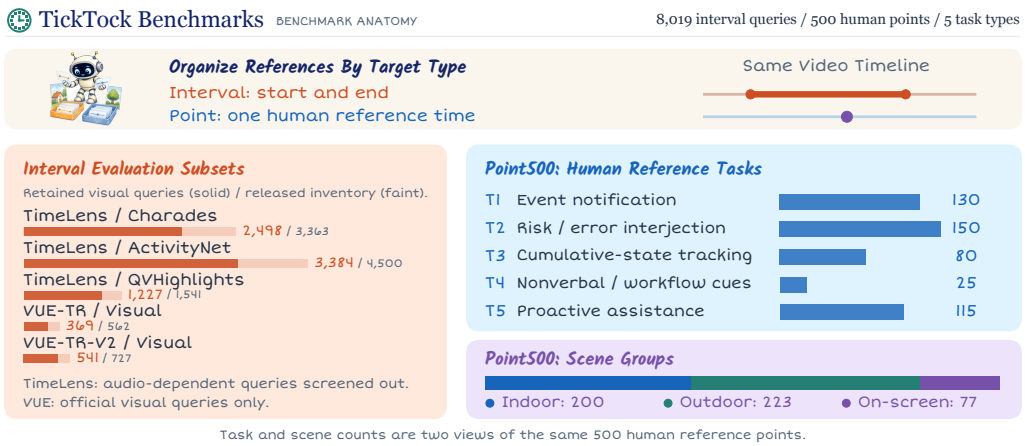

Figure 5: Benchmark composition. Top: intervals mark an event's start and end, whereas points mark a single human reference time on the same timeline. Left: TickTock-Refined subsets, with solid bars for the 8,019 queries retained after frame-anchored refinement and visual-only screening and faint bars for the released inventory (TimeLens: audio-dependent queries removed; VUE: official visual tags); Table 4 reports distinct videos. Right: TickTock-Point500 by task type (T1–T5) and by scene; both views cover the same 500 moments.

with ties resolved to the earlier frame (Figure 4B). The mean adjustment is 0.834 ms and the maximum 32.3 ms. Models see the full video, the query with its request context, and a search start a_i; a prediction is legal if $a_i \leq \hat{t}_i \leq D_i$. We report accuracy within 1, 2, and 5 s and the full tolerance curve.

All metrics keep the full query denominator. Missing, abstained, and illegal answers count as failures, and no prediction is clipped or rounded into range. Mean absolute error over legal answers is reported only as a diagnostic, because models differ in which queries they answer legally (Appendix C).

4 Seconds, Not Frames: How Current Models Localize

We evaluate 28 open and closed Video-LLMs on TickTock-Point500 and TickTock-Refined (Table 1): TimeLens and TimeLens2 (Zhang et al., 2026; Zhu et al., 2026), Qwen3-VL, Qwen3.5, Qwen3.6, Qwen3.8, and their Omni variants (Bai et al., 2025; Team, 2026; Qwen Team, 2026b;c;a;d), MiMo-V2.6 (Xiaomi MiMo Team, 2026), DeepSeek V4 Vision (DeepSeek-AI, 2026), GLM-5.3-Flash (Z.ai, 2026), Gemini 2.5 through 3.8 (Comanici et al., 2025; Google, 2025; Google DeepMind, 2026a;b;c;d), GPT-6 (OpenAI, 2026a;b), Muse Spark (Meta AI, 2026), and Doubao Seed (ByteDance Seed, 2026); exact API identifiers are in Appendix C.5. A nine-model cohort (seven Gemini routes, Qwen3-VL-8B, and TimeLens2-4B) has full tolerance curves and filename-only controls. Inputs are visual only, at 4 fps with at most 2,048 frames for Point500 and TimeLens and 1 fps for VUE, and no answer is chosen by its accuracy (Appendix C).

On the refined interval benchmarks, GPT-6 Astra leads Charades at 66.54 mIoU and Gemini 3.7 Flash leads ActivityNet and QVHighlights at 68.35 and 76.20; on VUE-v1, Doubao Seed 2.0 Lite leads hosted models at 61.60 and TimeLens2-8B leads open models at 59.89. Scores against the original references on the same queries are in Appendix Table 5.

Models reach the right second, not the right frame. Point accuracy climbs steadily with tolerance (Figure 6c). Across all 28 models, the best reaches 79.2% within five seconds and 54.6% within one second; in the nine-model cohort, accuracy within 0.1 s falls to 3.8–21.6%. The dominant error is not format: answers that are well formed but more than five seconds away account for 20.4–36.6% of queries, more than the illegal-answer rate of every model in the cohort (Figure 6a). All nine diagnostic models remain above the metadata priors; the strongest prior, returning the sup-

Table 1: Point and interval grounding (%). Point500: accuracy within $\pm 1/2/5$ s, MAE over legal answers, and illegal rate (count). TimeLens and VUE: refined-reference mIoU, with the invalid rate (count) for VUE; Char., ANet, and QVH have 2,498, 3,384, and 1,227 queries, VUE-v1 369, and VUE-v2 425. Invalid answers, abstentions, and execution failures count as misses. Shaded bold and underlined values rank first and second within each group. †RT: Realtime input at up to 1 fps with a 240 s context. Model versions and full metrics are in Appendix C.5.

	Point500 ($N = 500$)					TimeLens Refined (Vis.)			VUE-v1 (Vis.)		VUE-v2 (Vis.)	
Model	±1 s	±2 s	±5 s	MAE (s)	Illegal ↓	Char.	ANet	QVH	mIoU	Invalid ↓	mIoU	Invalid ↓
Open-Source Models												
TimeLens-8B	23.6	31.8	41.0	42.51	20.2 (101)	58.21	57.98	67.06	47.28	3.5 (13)	32.17	5.2 (22)
TimeLens2-2B	22.2	29.8	41.6	36.65	14.8 (74)	54.35	56.70	68.40	54.39	2.4 (9)	42.80	1.2 (5)
TimeLens2-4B	28.6	41.0	53.8	27.68	13.6 (68)	58.35	60.96	70.19	56.88	2.2 (8)	48.46	0.5 (2)
TimeLens2-8B	36.2	48.2	61.2	17.04	8.0 (40)	**58.83**	**61.19**	**72.81**	**59.89**	**0.3** (1)	**50.58**	**0.0** (0)
Qwen3-VL-8B	37.6	46.6	58.4	20.29	5.0 (25)	53.23	54.07	65.09	28.19	12.2 (45)	12.36	10.4 (44)
Qwen3.5 Flash	42.6	56.4	72.2	12.45	2.8 (14)	48.12	55.83	51.24	25.82	44.2 (163)	15.82	46.8 (199)
Qwen3.5 Plus	**46.4**	**60.2**	**76.6**	**9.11**	0.2 (1)	46.29	54.82	46.39	25.52	51.2 (189)	14.52	63.8 (271)
Qwen3.6 Flash	42.6	57.6	73.0	12.06	1.4 (7)	46.93	55.39	55.71	26.88	43.1 (159)	15.56	41.6 (177)
Qwen3.8-27B	40.2	57.8	68.8	10.83	5.6 (28)	48.98	51.54	59.75	34.55	3.5 (13)	21.54	2.4 (10)
MiMo V2.6 Pro	35.0	49.0	63.2	16.89	1.6 (8)	50.02	49.48	66.84	37.10	0.8 (3)	24.20	0.5 (2)
MiMo V2.6 Flash	32.4	49.0	61.6	18.62	2.8 (14)	49.84	51.60	64.78	37.21	**0.3** (1)	24.16	0.5 (2)
DeepSeek V4 Vision	17.2	27.8	46.6	20.13	**0.0** (0)	35.32	32.20	41.16	38.12	4.6 (17)	20.21	2.8 (12)
GLM-5.3 Flash	37.6	54.2	69.6	12.25	5.0 (25)	51.08	49.02	63.72	51.31	3.8 (14)	34.12	7.5 (32)
Closed-Source Models												
Gemini 2.5 Flash	36.6	46.6	58.2	13.70	12.2 (61)	51.95	59.16	65.73	25.56	15.4 (57)	13.16	20.2 (86)
Gemini 2.5 Pro	47.0	57.0	71.0	12.92	2.2 (11)	56.50	65.19	73.47	42.28	6.0 (22)	35.89	1.9 (8)
Gemini 3 Flash Preview	48.8	58.2	71.0	9.93	2.8 (14)	61.35	68.09	75.65	47.12	4.6 (17)	38.63	3.3 (14)
Gemini 3.1 Pro Preview	43.4	54.8	67.0	15.60	3.2 (16)	60.83	65.03	74.22	56.95	1.6 (6)	49.47	1.4 (6)
Gemini 3.5 Flash	50.6	60.0	72.0	8.92	3.2 (16)	62.57	66.31	74.24	50.30	3.0 (11)	42.01	2.6 (11)
Gemini 3.7 Flash	**54.6**	**63.4**	77.6	7.61	0.6 (3)	62.93	**68.35**	**76.20**	49.66	4.3 (16)	47.83	2.1 (9)
Gemini 3.8 Flash	54.2	62.6	**79.2**	8.73	0.4 (2)	62.34	66.83	72.94	53.46	3.8 (14)	47.86	2.8 (12)
Qwen3.8 Max	47.8	62.6	75.0	8.55	1.4 (7)	47.23	54.61	46.02	54.49	0.3 (1)	40.81	0.5 (2)
Qwen3.5 Omni Plus RT†	37.0	51.2	68.4	13.97	1.2 (6)	39.91	34.79	35.99	29.69	9.2 (34)	15.11	4.2 (18)
Qwen3.8 Omni Flash	47.8	59.2	73.8	10.75	0.8 (4)	47.07	57.36	69.59	52.67	0.8 (3)	40.79	0.5 (2)
GPT-6 Astra	46.6	56.0	66.4	**5.44**	**0.0** (0)	**66.54**	60.24	58.66	57.60	**0.0** (0)	50.50	**0.0** (0)
GPT-6 Sol	39.2	51.4	61.8	7.02	0.2 (1)	60.87	53.32	47.09	55.04	**0.0** (0)	44.14	**0.0** (0)
GPT-6 Luna	19.0	31.8	47.2	14.99	0.2 (1)	50.88	38.73	39.93	42.59	0.5 (2)	23.60	0.5 (2)
Muse Spark 1.1	48.2	61.6	76.0	7.90	0.2 (1)	56.04	58.14	68.64	61.32	1.1 (4)	46.80	1.9 (8)
Doubao Seed 2.0 Lite	46.8	59.2	72.2	12.08	1.4 (7)	55.96	64.87	75.41	**61.60**	**0.0** (0)	**52.39**	**0.0** (0)

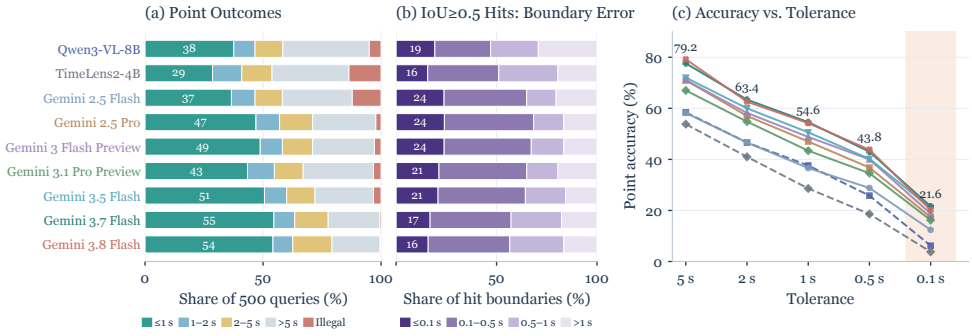

Figure 6: Seconds, not frames (nine-model cohort). (a) Point500 outcomes by error bin; labels give accuracy within 1 s. (b) Boundary errors of Charades predictions counted as hits (IoU ≥ 0.5) against frame-anchored references; labels give the share within 0.1 s. (c) Point500 accuracy at decreasing tolerances, labeled with the best model.

plied search start, reaches 17.2% within two seconds, against 41.0–63.4% for the models (Appendix Table 10).

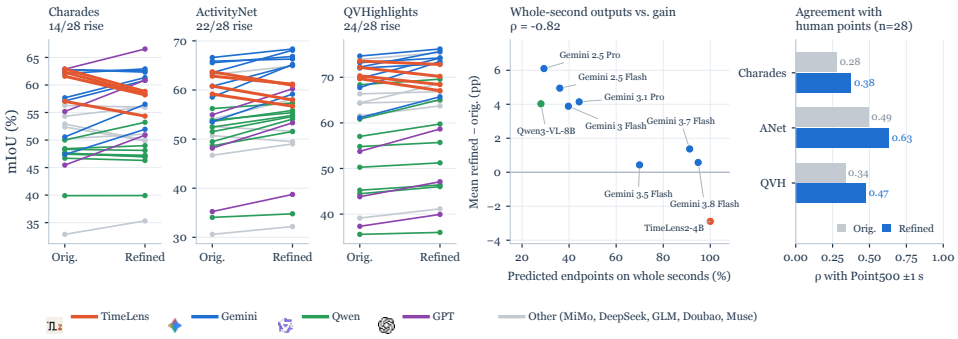

Figure 7: Reference sensitivity on the retained TimeLens queries. Left: mIoU of the same predictions under released and frame-anchored references for all 28 models. Fourth: mean change against the share of whole-second endpoints (nine-model cohort). Right: rank correlation with Point500 accuracy within 1 s under each reference.

IoU hides the gap and partly rewards priors. Moving only the start of a reference interval of length L inward by δ gives $\text{IoU} = 1 - \delta/L$, so a 0.1 s error on a five-second event still scores 0.98. The nine-model diagnostic cohort shows the same insensitivity: among Charades answers counted as correct at $\text{IoU} \geq 0.5$ against frame-anchored references, only 16–24% of boundaries lie within 0.1 s, and 16–29% are off by more than a second (Figure 6b). IoU also credits answers produced without the video: given only the filename and the query, all nine cohort models keep 7–17 mIoU on the original Charades references, and TimeLens2-4B keeps 14.19 of its 58.11 while answering $[1, 3]$ s for 35.2% of queries (Appendix Figure 15).

Reference changes alter the rankings. We compare released and frame-anchored references using the same accepted queries and predictions for all 28 models. Scores rise for 22 of 28 models on ActivityNet and 24 of 28 on QVHighlights. In the nine-model diagnostic cohort, score changes are negatively associated with the share of predicted endpoints on whole seconds ($\rho = -0.82$, $p = 0.007$). TimeLens2-4B loses 2.9 mIoU points on average across the three subsets, while Gemini 2.5 Pro and Qwen3-VL-8B gain 4–6. Across all 28 models, Spearman correlation with Point500 accuracy within 1 s rises from 0.28 to 0.38 on Charades, 0.49 to 0.63 on ActivityNet, and 0.34 to 0.47 on QVHighlights (Figure 7). This is agreement between benchmark rankings; it does not by itself establish that model-produced interval boundaries match independent human judgments. Point and interval targets also rank models differently, as do Point500 task types (Appendix D).

5 Conclusion

The frame is the finest temporal resolution a video supports. TickTock anchors annotation to it: annotators, human or MLLM, select frames on numbered filmstrips, and the decoder supplies each timestamp. We release TickTock-5K and model-refined TimeLens annotations for training, and TickTock-Refined and TickTock-Point500 for evaluation. Across 28 Video-LLMs, models find the right second far more often than the right frame, whole-second IoU conceals the difference, and frame-anchored references reorder models toward human point accuracy. We hope these resources move Video-LLMs toward frame-level grounding of complex events.

Limitations. TickTock-Refined references and the TimeLens audio screening are model-produced and PTS-validated, not verified by humans query by query; the refined TimeLens training annotations come from an earlier version of the pipeline; frame anchoring fixes coordinates, not the semantic ambiguity of when an event begins; and this version reports no model trained on TickTock-5K.

Reproducibility Statement

The supplementary archive holds raw responses, reference versions, decoder mappings, scorer hashes, and per-query metrics for every number; no answer is selected by its accuracy. Appendices C and D give prompts, parsers, and population definitions.

AI Use Statement

Generative AI plays two roles in this work. First, as part of the research protocol, multimodal LLMs serve as annotators and as evaluated models: an MLLM selects boundary frame IDs to produce the TickTock-Refined references and the refined TimeLens training annotations, the multi-domain source labels underlying TickTock-5K were produced by MLLMs, and MLLMs screen queries for audio dependence; these roles are described in Sections 3 and 4, and all model-produced labels are marked as such rather than as human verification. Second, as research assistance, we used generative AI tools to help implement analysis and data-processing code, and polish and translate the writing. The authors directed these tasks, reviewed all AI-assisted outputs, and made the final decisions on the research questions, methods, and interpretation of results. Formulating mathematical claims and writing proofs are not applicable to this work. We take full responsibility for the content of this paper, including all text, claims, data, and artifacts produced with the aid of generative AI.

References

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

ByteDance Seed. Seed2.0. https://seed.bytedance.com/en/seed2, 2026.

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.

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.

DeepSeek-AI. DeepSeek API documentation: Vision (deepseek-v4-flash-vision-exp). https://api-docs.deepseek.com/guides/vision/, 2026.

Andong Deng, Zhongpai Gao, Anwesa Choudhuri, Benjamin Planche, Meng Zheng, Bin Wang, Terrence Chen, Chen Chen, and Ziyan Wu. Seq2Time: Sequential knowledge transfer for video LLM temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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.

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, 2017a.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017b.

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

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

Google DeepMind. Gemini 3.5 Flash model card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026b.

Google DeepMind. Gemini 3.7 Flash model card. https://deepmind.google/models/model-cards/gemini-3-7-flash/, 2026c.

Google DeepMind. Gemini 3.8 Flash model card. https://deepmind.google/models/model-cards/gemini-3-8-flash/, 2026d.

Yongxin Guo, Jingyu Liu, Mingda Li, et al. VTG-LLM: Integrating timestamp knowledge into video LLMs for enhanced video temporal grounding. *arXiv preprint arXiv:2405.13382*, 2025.

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

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

Shengji Jin, Yuanhao Zou, Victor Zhu, Zhengping Ji, and Chen Chen. How should video LLMs output time? an analysis of efficient temporal grounding paradigms. In *CVPR 2026 Workshop*, 2026.

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

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

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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.

Meta AI. Introducing Muse Spark 1.1. https://ai.meta.com/blog/introducing-muse-spark-meta-model-api/, 2026.

OpenAI. GPT-6 Astra system card. https://deploymentsafety.openai.com/gpt-6-astra, 2026a.

OpenAI. Introducing GPT-6 Sol and Luna. https://openai.com/index/introducing-gpt-6-sol-and-luna/, 2026b.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026a.

Qwen Team. Qwen3.6-Plus: Towards real world agents. https://qwen.ai/blog?id=qwen3.6, 2026b.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8, 2026c.

Qwen Team. Qwen3.8-Omni-Flash: Omni senses. agentic delivery. https://qwen.ai/blog?id=qwen3.8-omni-flash, 2026d.

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

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

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*, 2025a.

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*, 2025b.

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, 2025a. URL https://arxiv.org/abs/2410.03290.

Xizi Wang, Feng Cheng, Ziyang Wang, Huiyu Wang, Md Mohaiminul Islam, Lorenzo Torresani, Mohit Bansal, Gedas Bertasius, and David Crandall. TimeRefine: Temporal grounding with time refining video LLM. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

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

Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. GEB+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference on Computer Vision*, pp. 709–725. Springer Nature Switzerland, 2022.

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.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13754–13765, 2025.

Xiaomi MiMo Team. Introducing MiMo-V2.6 series. https://mimo.xiaomi.com/mimo-v2-6, 2026.

Z.ai. GLM-5.3-Flash: Frontier intelligence, flash cost. https://z.ai/blog/glm-5.3-flash, 2026.

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.

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.

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

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. TimeLens2: Generalist video temporal grounding with multimodal LLMs. *arXiv preprint arXiv:2607.17423*, 2026.

APPENDIX CONTENTS

A Frame Anchoring: Interface And Resolution **16**

A.1 Why A Contact Sheet Instead Of A Timestamped Video 16

A.2 Overlap, Grids, And Frame Resolution. 16

A.3 Point500 Release Version. 16

B Data Provenance And Construction Details **16**

B.1 TickTock-5K Selection 16

B.2 Filtered TimeLens Records And Annotation Completeness 17

B.3 Multi-Domain Event-Discovery Export 18

B.4 Limits Of Model-Refined Training Annotations 18

B.5 Point500 Preservation Checks And Evaluation Separation 19

C Baseline Evaluation Details **19**

C.1 Response Snapshot And Model Identities 19

C.2 Completed TickTock-Refined References 20

C.3 Full Visual VUE-v1 Evaluation 21

C.4 Native-Frame Windows And Contact-Sheet Examples 22

C.5 Additional Model Protocols And Coverage 26

C.6 Per-Query IoU Distributions 29

C.7 Video Inputs And Generation Settings 30

C.8 Recorded Answer Retries. 31

C.9 Task Prompts And Parsing 31

C.10 Point Reference Versions And Category Results 31

C.11 Recovered VUE Scope And Media Exceptions. 32

C.12 Reference Eligibility And Modality Screening 32

D Benchmark Diagnostics And Robustness **32**

D.1 Temporal Tolerance And Error Structure. 32

D.2 Task Composition And Cross-Benchmark Rankings. 32

D.3 Temporal-Prior Controls 33

D.4 Sensitivity To The Recorded Retry Protocol. 34

D.5 Filename-And-Query Ablation 34

D.6 Frozen Predictions And Analysis Units 35

D.7 Tolerance And Denominator Sensitivity 35

D.8 First Responses And Completed Retries 35

E Training Annotation Diagnostics **36**

E.1 Timestamp Distributions Before And After Refinement. 36

Table 2: Filtering the model-refined TimeLens training annotations. Counts are intervals except the final query row; these corpora are separate from TickTock-5K.

Count	TimeLens-100K	TimeLens2-93K
Input intervals	96,544	108,491
Null	10,523	25,194
Zero length	84	57
Reversed	1	0
Retained	85,936	83,240
Queries	85,936	73,586

A Frame Anchoring: Interface And Resolution

A.1 Why A Contact Sheet Instead Of A Timestamped Video

Frame anchoring requires the annotator to identify a frame, not to read or write a time. Two interfaces could expose frame identities to an MLLM: numbers burned into the video, as in NumPro (Wu et al., 2025), whose released code passes the numbered frames through the video channel, or a static contact sheet passed as an image. We use the latter. Local frames are tiled chronologically, each with a legible ID in a separate label band, and the selected ID is resolved to its decoded PTS by Eq. 1.

The contact sheet makes every candidate frame available through the image interface and associates it with a visible ID. Window expansion adds pages at the same cell size rather than reducing spatial resolution. This specifies the annotation interface; it does not establish an accuracy advantage over a matched video-input baseline.

A.2 Overlap, Grids, And Frame Resolution

For a reference interval of length L, moving only its start inward by $\delta \in [0, L]$ gives $\mathrm{IoU} = 1 - \delta/L$. On a five-second event, shifts of 100 ms and $1/30$ s yield IoUs of 0.980 and 0.993, both indistinguishable from a correct answer under R1@0.7. An exact match to an integer-second reference likewise certifies agreement with a storage convention rather than with a native frame; the per-query IoU distributions in Appendix C.6 show both complete misses and exact coordinate matches.

If predictions are restricted to a regular sampled grid with spacing h, nearest-grid assignment has maximum interior quantization error $h/2$. This bounds that assignment rule, not every model: a model may interpolate between inputs or use other cues. Frame anchoring removes coordinate ambiguity while leaving the semantic choice of frame to be evaluated, which is why TickTock reports point error alongside overlap.

A.3 Point500 Release Version

TickTock-Point500 is released as version `FrameSnap-v1`, which stores each original human time, the selected frame, and the displacement. TickTock-Refined preserves the source queries and keeps only those whose intervals all pass the refinement checks, with VUE restricted to its official visual-query subsets. The scoped derivative preserves accepted records byte for byte and archives the IDs excluded by modality. The retained inventory and completed rescoring appear in Appendix C.2.

B Data Provenance And Construction Details

B.1 TickTock-5K Selection

TickTock-5K is a per-pool prefix of a seeded hash ranking over deduplicated source videos, and can be extended without re-ranking. Records and times are copied verbatim from the source collections; no query or timestamp is modified.

TickTock-5K holds 50,295 intervals and 39,530 queries, about ten intervals per example, but density varies widely across categories: from 1.6 intervals per daily short video to 128 per basketball game (Table 3).

Table 3: TickTock-5K by content category. Each example is one source video; queries group identical event descriptions within an example, and each query keeps all of its source intervals. The two tutorial categories are kept whole.

Category	Videos	Queries	Intervals
Daily Short Videos	590	933	954
Sports Events	590	1,528	2,185
Web Events	590	3,357	3,783
Film Commentary	590	4,317	4,759
Creator Videos	590	2,128	2,489
STEM Lessons	589	6,786	7,622
Action Highlights	500	1,065	3,462
Open Lectures	589	5,613	6,195
GUI Tutorials	229	2,708	2,954
Software Walkthroughs	33	1,622	1,844
Basketball Games	110	9,473	14,048
Total	5,000	39,530	50,295

Figure 8 shows how differently the categories are structured. Short-video categories mostly contain one or a few events per video, lectures and tutorials a handful of steps, and basketball games and software walkthroughs dense event lists. Interval durations span more than three orders of magnitude.

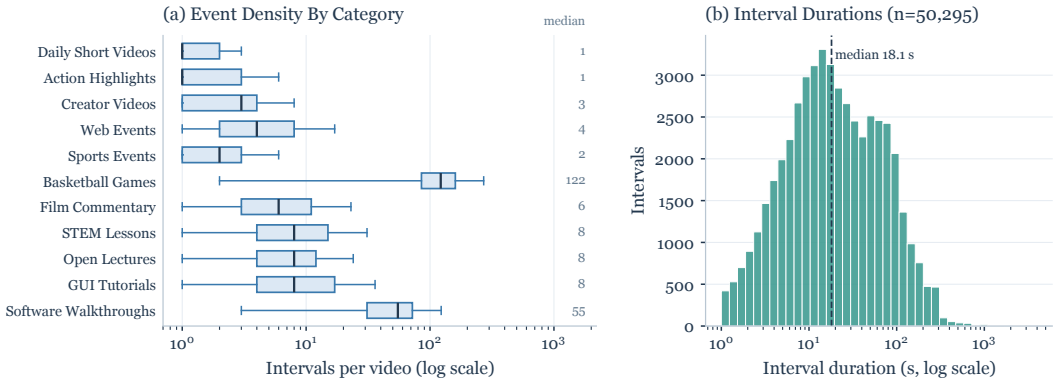

Figure 8: TickTock-5K anatomy. (a) Intervals per video in each content category (boxes span the interquartile range; medians at right). (b) Distribution of all 50,295 interval durations on a logarithmic axis.

B.2 Filtered TimeLens Records And Annotation Completeness

The current training-pool version is `TimeLens-Refined-ValidSpans-v2`. Its input snapshot contains 96,544 TimeLens-100K and 108,491 TimeLens2-93K refinement records. These count available refinement outputs, not the original released corpus size. The TimeLens2 source release has 93,232 queries and 108,718 intervals; a refinement record is identified by video, query, and interval index. Filtering removes 35,717 records with a null endpoint, 141 zero-length intervals, and one reversed interval. No negative or unparseable non-null endpoint remains in the inspected source snapshot. The 169,176 retained intervals form 85,936 TimeLens-100K queries and 73,586 TimeLens2 queries. Video identifiers are unique only within each source inventory; these counts do not establish cross-source video disjointness.

Of the retained TimeLens2 records, 3,518 contain 6,571 string endpoints. There are 4,911 decimal strings with an `s` suffix and 1,660 plain decimal strings. Conversion checks that the numeric value has the same decimal value as the source string. Original strings are stored in `refined_start_raw` and `refined_end_raw`; a separate archive also preserves the retained source JSONL lines byte for byte. No extra decimal places are introduced. The earlier timestamp-sheet pipeline labels displayed frames to three decimals, so preserving the returned value does not establish six-decimal annotation accuracy.

Content judgments do not filter this version: 18,081 retained records have `content_correct` false or missing. We retain that metadata for later analyses. Source-confidence fields are also preserved without a threshold. Unlike a whole-query exclusion rule, interval filtering keeps the valid occurrences of a query even when another occurrence is removed. The query export stores its interval list, retained source-record IDs, the expected source record count, and `all_source_intervals_retained`. For TimeLens2, the expected list comes from the original refinement-input inventory. Missing records are distinguished from records rejected by this filter in the audit. There are 4,065 partial query groups containing 5,166 retained intervals. The completeness flag concerns the source inventory, not independently verified semantic completeness.

The export preserves source order and provides separate record-level and query-level files. Verification checks all 205,035 input records, each rejection reason, decimal conversion, query membership, original fields, raw retained lines, and file hashes. It does not inspect video files or establish new train/test splits. These TimeLens records are not part of TickTock-5K, and TimeLens-Bench test references are kept separate from them.

B.3 Multi-Domain Event-Discovery Export

The multi-domain archive uses version 1 of the valid-span export. All 14 source files match the downloaded snapshot by SHA256; their raw copies are unchanged. They are grouped into the 11 content categories of Table 3 while retaining per-file provenance. Source records contain event lists, so record, query, and interval counts differ.

All endpoints are seconds strings in the source event lists. Each parses to a finite nonnegative value without changing its decimal value, and every interval has positive duration. The export stores numeric endpoints and their original strings, along with the source file, row, event index, task, video URI, and unmodified event description. Grouped query records retain every source interval. There are no rejected intervals or partial query groups in this snapshot; the 25 empty records comprise eight film-commentary and 17 action-highlight records. No event or point target is inferred from an empty record.

These are inherited event-discovery labels, distinct from the newly proposed TimeLens refinement times. We do not apply content-confidence filters or relabel events. Structural verification checks every exported event against the source text and every query's interval membership. Media accessibility, video-duration bounds, frame alignment, semantic accuracy, and cross-pool overlap require separate checks before these labels are used for training.

B.4 Limits Of Model-Refined Training Annotations

The earlier TimeLens training-refinement pipeline extracts timestamped frames around a supplied boundary and then subsamples the contact sheet to at most 25 cells. A nominal three-second window at 30 fps can therefore display frames roughly a tenth of a second apart. Requesting three decimal places does not recover the omitted frames. The pipeline also stores a separate content decision; a successful API request or the presence of two numbers does not imply that the interval is ordered, in bounds, or semantically correct.

Generated proposals retain their original intervals, proposed intervals, content decisions, and provenance. The current cleaned TimeLens pool enforces finite, nonnegative, ordered endpoints and preserves the content decisions without using them for selection. Its structural validity does not establish correct media alignment or semantic accuracy. Independent boundary validation is needed for claims of frame-level annotation accuracy. The original test annotations are archived separately from these training refinements and from the new TickTock-Refined benchmark release.

B.5 Point500 Preservation Checks And Evaluation Separation

Figure 9 summarizes the release. Most videos last between one and ten minutes. Reference moments cluster near the start and the end of a video but cover the whole timeline. Snapping leaves 155 references unchanged and moves the rest by a few milliseconds.

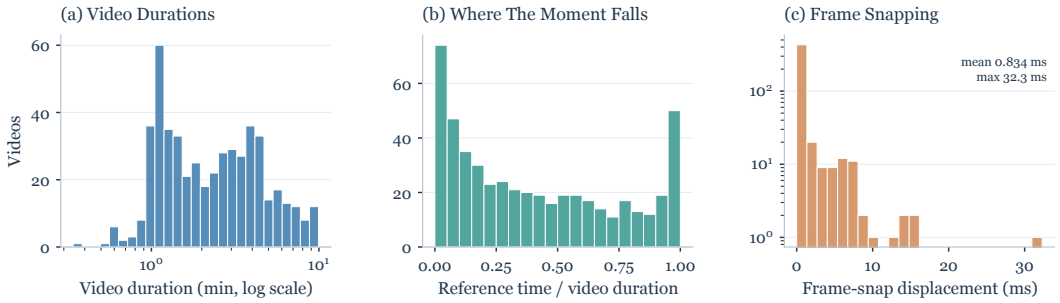

Figure 9: TickTock-Point500 anatomy. (a) Video durations. (b) Position of each reference moment relative to its video's duration. (c) Displacement introduced by snapping each human time to its nearest decoded frame (logarithmic count axis).

Point500 retains the previously frozen selection of 500 distinct videos, with 130, 150, 80, 25, and 115 queries in the five task categories T1–T5. The construction removes explicit video-answer times from 34 query descriptions. An on-screen clock reading is preserved as a visual target, and two references to images are replaced with objects already specified in the same source annotation. The query-edit audit records these changes. The archived human timestamp is the original trigger time divided by 1,000 to express seconds. The derived scoring reference follows Equation 2.

Validation checks the record count, video links, input/reference separation, legal reference times, equality of archived human times with the source, and nearest-frame selection for all 500 samples. We additionally decoded 13 difficult videos from the beginning through a separate PyAV interface, including all six videos with nonzero stream start timestamps; the selected frames agreed. One source reference equals a rounded search-start time. Its closest frame is 3.333 microseconds earlier, so that record's search lower bound is adjusted by the same amount to keep its reference legal. The original bound remains in the archive. All other inputs, and all query and context text, are unchanged. These checks validate frame alignment and preservation; semantic timing accuracy still depends on the inherited annotations. Deduplication within Point500 does not establish disjointness from training corpora. Claims about newly trained models require a separate cross-corpus overlap audit. No point labels are constructed from interval starts, ends, or midpoints in this benchmark.

C Baseline Evaluation Details

C.1 Response Snapshot And Model Identities

The original nine-model diagnostic analyses use the response snapshot recorded on September 24, 2026 at 12:12 UTC. The refined tables use the reference snapshot described in Appendix C.2; completed results from later runs are added as described in Appendix C.5. Already completed cells of the original cohort keep their frozen predictions. We retain only newline-committed records from each append-only response ledger. Successful returned answers are preserved regardless of their correctness. Identical duplicate records are collapsed by query ID; conflicting answers for one ID trigger an error rather than a reference-based selection. Table 9 counts returned answers, including outputs that subsequently fail parsing or temporal-range checks. Returned-answer counts therefore differ from the legal-output counts in Table 1.

The seven Gemini API route IDs are `gemini-2.5-flash`, `gemini-2.5-pro`, `gemini-3-flash-preview`, `gemini-3.1-pro-preview`, `gemini-3.5-flash`, `gemini-3.7-flash`, and `gemini-3.8-flash`. Every retained response has a returned model-version field equal to its requested route and has cross-model fallback disabled. These identifiers describe the evaluated API routes. The local models are Qwen3-VL-8B-Instruct and

TimeLens2-4B; the manifests retain checkpoint configuration and tokenizer hashes, weight-shard sizes, and inference-code hashes.

C.2 Completed TickTock-Refined References

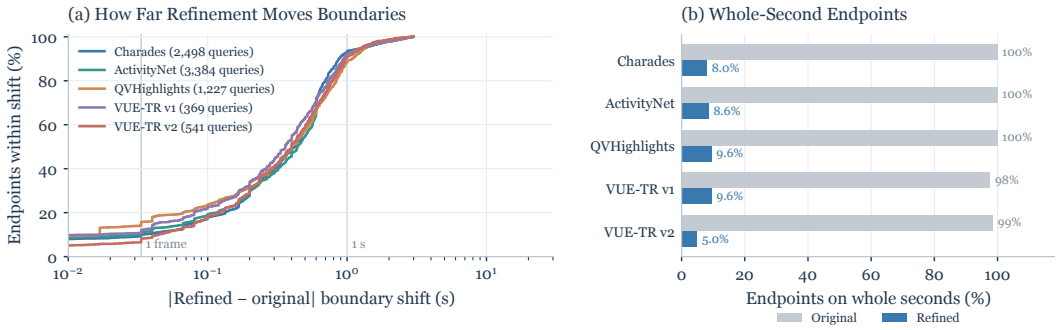

Figure 10: What refinement changes. (a) Cumulative distribution of the shift between each original endpoint and its frame-anchored replacement, per subset. (b) Share of endpoints on whole seconds before and after refinement.

Figure 10 shows the effect on the references themselves. The median endpoint moves by about 0.4 s in every subset, and almost every original endpoint lies on a whole second, against 5–10% after refinement.

The active release restricts the refinement snapshot frozen on September 24, 2026 at 21:45:43 UTC to officially tagged visual VUE queries and TimeLens queries retained after Gemini screening and GPT-6 Astra review. It retains 8,019 complete queries. An MLLM selects boundary frame IDs, the decoder maps them to presentation timestamps, and a separate content-verification request checks the event. These references have not received independent, query-by-query human verification. A query with several occurrences is retained only if every interval passes the checks; rejected or unresolved intervals are not replaced by their original annotations. The export contains 8,457 intervals and passes 16,914 endpoint PTS checks. The source snapshot is preserved; scope filtering does not change any accepted frame ID or timestamp.

For each released start and end time, the annotator first receives all native frames within ± 1 s, arranged in numbered contact sheets. It selects the first event frame for the start and the last event frame for the end. An unresolved boundary, a selection at an interior window edge, or confidence below 0.7 triggers a second search within ± 3 s of the original endpoint. Windows are clipped to the video bounds. A missing event or a boundary that remains unresolved is excluded. Both endpoints must pass, and a separate content-verification request must support the event with confidence at least 0.75. Thus interval refinement can change the semantic boundary by seconds; Point500 instead preserves the human reference moment and only snaps its coordinate to the nearest decoded frame. Appendix C.4 shows frame excerpts from the uploaded contact sheets, window expansion, and frame-to-time lookup for three archived examples.

Video counts are deduplicated within each subset, not summed across overlapping benchmarks. Losing a query does not remove its video if another query on that video survives. For example, Charades loses 865 queries but only 143 video IDs. Point500 keeps all 500 videos and queries after timestamp snapping. Before quality filtering, the official VUE modality tags exclude 1,036 v1 and 873 v2 audio or audiovisual queries. These scope exclusions are separate from the 193 and 186 visual queries removed by refinement checks. The archived VUE-v2 score cohort already consists of 425 visual queries, so modality filtering does not change its scores. After quality filtering, TimeLens screening removes another 50 audio-dependent and 42 unresolved queries, leaving 7,109 across its three subsets. Appendix C records these exclusions.

We rescore the archived parsed predictions without changing the answers, prompts, media, or the audited parser decisions. The inherited snapshot includes the Point500/VUE parser repairs in Appendix C.5. Original and refined scores in Table 5 use identical retained query IDs. The original IoUs reproduce the archived scores with zero per-query mismatches. A numerical entry requires a

Table 4: Completed refinement inventory and overlap with archived evaluation inputs. Videos are distinct released video IDs within each benchmark; a video remains if at least one complete query passes. Source counts include missing media and restrict VUE to official visual queries. Removed queries fail quality checks or are excluded by the subsequent TimeLens modality screen, including unresolved cases. Intervals belong to retained queries, and Scored counts evaluation queries. VUE-v1 now covers all 369 retained visual queries; VUE-v2 scores cover 425 of its 541 retained queries. The remaining VUE-v2 queries are outside the archived evaluation inputs and are not counted as model failures.

	Videos		Queries				
Benchmark	Source	Retained	Source	Retained	Removed	Intervals	Scored
Charades	1,313	1,170	3,363	2,498	865	2,498	2,498
ActivityNet	1,455	1,387	4,500	3,384	1,116	3,384	3,384
QVHighlights	1,511	1,206	1,541	1,227	314	1,227	1,227
VUE-TR-v1	268	209	562	369	193	586	369
VUE-TR-v2	510	420	727	541	186	762	425

finalized outcome for every matched query; unfinished inference remains a dash. Verified exhausted execution retries follow Appendix C.5 and contribute zero overlap. Abstentions and invalid answers also contribute zero overlap, with abstentions excluded from invalid-answer counts. A score change here measures the effect of the reference change on fixed predictions, not an improvement in the model.

Separating filtering from reference changes. For Qwen3-VL-8B, original-reference mIoU on the full Charades inventory is 46.63. Retaining the 2,498 accepted queries changes it to 50.00, a filtering effect of $+3.37$ points. Replacing only their reference intervals changes it to 53.23, a further $+3.22$ points. The corresponding filtering and reference effects are $+3.15$ and $+4.60$ on ActivityNet, and $+2.54$ and $+4.27$ on QVHighlights. Every prediction and parser outcome is unchanged, and both sets of IoUs reproduce from the archived intervals.

There is no universal two-point bound on this reference effect: IoU depends on displacement relative to event duration. On retained Charades queries, the median original interval is 4 s and the median absolute endpoint change is 0.425 s. Of 2,498 queries, 309 move at least one endpoint by more than 1 s; none exceeds 3 s. Even among the other 2,189 queries, the mean change is $+2.78$ mIoU points. Overall, 1,495 query scores rise, 867 fall, and 136 stay unchanged. These changes quantify sensitivity to model-annotated references; they do not establish improved localization against independent human boundaries.

C.3 Full Visual VUE-v1 Evaluation

The evaluation covers all 369 accepted visual queries on 209 videos, with 586 refined reference intervals. Inputs follow the recorded VUE protocol: silent video at a nominal 1 fps, capped at 2,048 frames, and a query asking for all relevant intervals. Omni Realtime uses the streaming configuration in Appendix C.5. Reference timestamps are excluded from the model inputs. Both score columns in Table 8 use the same predictions and query IDs, with the released or refined references. The repaired parser and abstention policy match the current VUE-v2 tables.

Before scoring, raw-frame synchronization errors were repaired for requests that failed locally before reaching the API. The repair preserves the sampled video, frame layout, prompt, and all previously returned answers. Initial API executions allow five attempts; the separate Omni Realtime repair is documented in Appendix C.5. Finalized execution failures, invalid answers, and abstentions receive zero overlap under the full denominator. All 28 model configurations in Table 8 have a finalized outcome for every query. The main paper reports 28 model rows; this table also retains the 02-23 Qwen3.5-35B-A3B route and the undated Qwen3.5-397B-A17B route. Qwen3-VL-8B has 369 returned answers and 28.19 refined mIoU, with 45 invalid answers and one abstention. Omni Realtime has 241 returned answers and 128 execution failures after repair; its refined mIoU is 29.69. Its 34 invalid answers and seven abstentions are separate from those execution failures.

Table 5: Interval grounding on the retained TimeLens-Bench queries (mIoU, %). *Orig.* and *Refined* score the same predictions on the same query IDs using released and frame-anchored references, respectively. Retained counts are given in Section 3.3. Bold marks the best score per column. Hosted models use the input protocols in Appendix C.5.

	Charades 3,363 → 2,498 queries		ActivityNet 4,500 → 3,384 queries		QVHighlights 1,541 → 1,227 queries	
Model	Orig. GT	Refined GT	Orig. GT	Refined GT	Orig. GT	Refined GT
Open-Source Models						
TimeLens-8B	61.60	58.21	60.79	57.98	69.69	67.06
TimeLens2-2B	57.03	54.35	59.17	56.70	70.32	68.40
TimeLens2-4B	62.32	58.35	63.70	60.96	72.15	70.19
TimeLens2-8B	**62.90**	58.83	62.82	61.19	73.56	72.81
Qwen3-VL-8B	50.00	53.23	49.47	54.07	60.82	65.09
Qwen3.5 Flash	48.31	48.12	53.59	55.83	50.27	51.24
Qwen3.5 Plus	46.70	46.29	52.45	54.82	45.24	46.39
Qwen3.6 Flash	47.66	46.93	53.84	55.39	54.81	55.71
Qwen3.8-27B	48.43	48.98	48.63	51.54	57.03	59.75
MiMo V2.6 Pro	52.90	50.02	50.70	49.48	66.36	66.84
MiMo V2.6 Flash	52.38	49.84	52.55	51.60	64.35	64.78
DeepSeek V4 Vision	32.87	35.32	30.60	32.20	39.13	41.16
GLM-5.3 Flash	50.22	51.08	46.70	49.02	61.56	63.72
Closed-Source Models						
Gemini 2.5 Flash	47.36	51.95	53.50	59.16	61.13	65.73
Gemini 2.5 Pro	50.56	56.50	58.55	65.19	67.74	73.47
Gemini 3 Flash Preview	57.67	61.35	63.53	68.09	72.23	75.65
Gemini 3.1 Pro Preview	57.10	60.83	60.76	65.03	69.78	74.22
Gemini 3.5 Flash	62.75	62.57	65.87	66.31	73.21	74.24
Gemini 3.7 Flash	62.09	62.93	**66.61**	**68.35**	**74.66**	**76.20**
Gemini 3.8 Flash	62.81	62.34	65.55	66.83	72.02	72.94
Qwen3.8 Max	47.54	47.23	51.52	54.61	44.46	46.02
Qwen3.5 Omni Plus RT	39.90	39.91	34.06	34.79	35.56	35.99
Qwen3.8 Omni Flash	47.43	47.07	56.22	57.36	68.36	69.59
GPT-6 Astra	62.87	**66.54**	54.95	60.24	53.75	58.66
GPT-6 Sol	55.18	60.87	48.16	53.32	43.81	47.09
GPT-6 Luna	45.45	50.88	35.25	38.73	37.32	39.93
Muse Spark 1.1	54.32	56.04	54.02	58.14	64.36	68.64
Doubao Seed 2.0 Lite	56.28	55.96	63.12	64.87	73.93	75.41

The original-reference diagnostics use the full released TimeLens inventories or the recovered 538-query VUE-v2 subset. They cannot be subtracted from refined scores computed on different query sets.

C.4 Native-Frame Windows And Contact-Sheet Examples

What the annotator receives. Each boundary request contains all decoded frames in its local window, at the source video's native rate, as separate JPEG contact-sheet pages. A page holds up to 24 frames in four columns and six rows. Each 320×240-pixel cell contains an aspect-preserving image within 320×212 pixels and a 28-pixel ID band. A full page is 1280×1440 pixels, encoded at JPEG quality 88. IDs run consecutively across pages within one request and restart for a new window. All pages for that window are sent together; start and end boundaries use separate requests.

Expanding from ± 1 to ± 3 s adds pages at the same cell size and native rate. A six-second window at 30 fps has approximately 180 frames and needs eight pages per endpoint. Exact counts follow decoded PTS and include any frames on the window bounds. Windows exceeding 1,000 native frames fail the media check instead of being subsampled. The annotator is Gemini 3.1 Pro Preview, with temperature 0, a 4,096-token output limit, and low thinking. These image requests do not

Table 6: TickTock-Refined TimeLens results, in percent. Header counts show released → retained queries (Q) and videos (V); each metric includes every retained query. Bold marks the maximum in each column.

	Charades Q: 3,363 → 2,498 V: 1,313 → 1,170			**ActivityNet** Q: 4,500 → 3,384 V: 1,455 → 1,387			**QVHighlights** Q: 1,541 → 1,227 V: 1,511 → 1,206		
Model	mIoU	R1@.5	R1@.7	mIoU	R1@.5	R1@.7	mIoU	R1@.5	R1@.7
Open-Source Models									
TimeLens-8B	58.21	64.17	40.19	57.98	62.85	46.42	67.06	72.86	57.21
TimeLens2-2B	54.35	57.37	36.71	56.70	61.47	44.21	68.40	74.08	59.41
TimeLens2-4B	58.35	63.77	41.07	60.96	66.58	49.85	70.19	75.96	61.53
TimeLens2-8B	58.83	65.45	41.15	61.19	66.70	50.86	72.81	78.32	65.69
Qwen3-VL-8B	53.23	54.92	33.79	54.07	56.53	43.00	65.09	68.70	56.32
Qwen3.5 Flash	48.12	46.20	22.94	55.83	58.01	40.99	51.24	49.06	35.45
Qwen3.5 Plus	46.29	44.36	19.86	54.82	56.44	39.10	46.39	42.95	29.58
Qwen3.6 Flash	46.93	43.96	22.06	55.39	57.57	41.78	55.71	54.77	41.24
Qwen3.8-27B	48.98	46.40	26.90	51.54	51.89	38.27	59.75	60.31	48.49
MiMo V2.6 Pro	50.02	52.76	31.47	49.48	52.87	39.95	66.84	72.62	59.90
MiMo V2.6 Flash	49.84	52.72	31.51	51.60	56.00	40.75	64.78	69.68	57.29
DeepSeek V4 Vision	35.32	32.27	21.10	32.20	31.62	21.51	41.16	42.79	27.06
GLM-5.3 Flash	51.08	52.96	34.79	49.02	51.36	37.38	63.72	67.56	54.69
Closed-Source Models									
Gemini 2.5 Flash	51.95	53.76	35.35	59.16	62.32	51.00	65.73	67.24	57.46
Gemini 2.5 Pro	56.50	59.73	39.23	65.19	68.59	58.69	73.47	75.88	66.91
Gemini 3 Flash Preview	61.35	65.81	45.44	68.09	73.49	**61.41**	75.65	80.11	69.11
Gemini 3.1 Pro Preview	60.83	66.81	46.12	65.03	69.77	56.94	74.22	78.16	66.91
Gemini 3.5 Flash	62.57	68.94	47.28	66.31	72.28	57.39	74.24	78.81	67.40
Gemini 3.7 Flash	62.93	70.10	46.44	**68.35**	**75.06**	60.05	**76.20**	81.01	**70.74**
Gemini 3.8 Flash	62.34	69.10	44.68	66.83	73.37	58.10	72.94	77.10	66.18
Qwen3.8 Max	47.23	44.76	19.86	54.61	56.38	39.39	46.02	42.38	30.40
Qwen3.5 Omni Plus RT	39.91	36.95	17.05	34.79	32.00	18.76	35.99	33.74	18.99
Qwen3.8 Omni Flash	47.07	44.84	21.38	57.36	60.08	44.24	69.59	73.92	60.64
GPT-6 Astra	**66.54**	**74.98**	**55.40**	60.24	64.89	56.83	58.66	61.21	54.93
GPT-6 Sol	60.87	66.37	46.96	53.32	57.65	47.90	47.09	49.63	42.54
GPT-6 Luna	50.88	52.40	34.39	38.73	40.81	29.64	39.93	42.79	33.33
Muse Spark 1.1	56.04	60.69	41.91	58.14	61.82	51.33	68.64	71.23	63.00
Doubao Seed 2.0 Lite	55.96	58.37	34.43	64.87	70.30	54.55	75.41	**81.42**	68.22

explicitly set `mediaResolution`; the low-resolution video setting used by baseline inference does not describe these annotation requests.

Expansion after an outside-window decision. ActivityNet query `v_kgvbU_3jEy8/1` asks for "The screen pauses, and the player in white clothes is circled." Its released interval is $[65, 69]$ s. At 30 fps, the initial end window $[68, 70]$ s contains 61 frames on three pages. The white circles remain visible throughout, so the model returns `OUTSIDE`. Expanding to $[66, 72]$ s supplies 181 frames on eight pages. The model selects ID 137 at 70.533333 s, the last frame with the circles; ID 138 at 70.566667 s has no circles and the footage resumes. The retained interval is $[65.533333, 70.533333]$ s. Figure 11 shows the complete sixth page of the expanded request, containing IDs 121–144 in their original order. For display, we remove sheet padding and retypeset the ID/time labels.

Starts resolved within the initial window. Two further cases expose a visible transition between consecutive frames (Figure 12). ActivityNet query `v_W30cufYc_ZI/0` targets the title "Beginning Of the Dive Nassau, Bahamas 12/4/08." Its initial start window $[1, 3]$ s contains 51 frames on three pages at 25 fps. The model returns ID 36 at 2.400000 s, where the title appears; ID 35 at 2.360000 s

Expanded End Window: Complete Contact-Sheet Page

[66, 72] s | 30 fps | Page 6 / 8 | IDs 0121–0144 Gemini: #0137



Figure 11: Complete contact-sheet page containing Gemini's returned end. After the initial $[68, 70]$ s window returns `OUTSIDE`, the expanded $[66, 72]$ s request supplies 181 native frames on eight pages. This is page six, with all 24 frames, IDs 121–144, in the original four-column order. Gemini selects ID 137 at 70.533333 s, marked in teal; the circles disappear in ID 138, one frame later. All eight pages were sent together.

Table 7: VUE-TR-v2 visual-query reference comparison. Refinement retains 541 of 727 eligible queries on 420 of 510 videos; both score columns use the same 425 matched queries. Original and refined mIoU use the same predictions. All R1 metrics use refined references. Invalid and Abstained are counts; both receive zero overlap.

Model	Orig. mIoU	Ref. mIoU	R1@.3	R1@.5	R1@.7	Invalid	Abstained
Open-Source Models							
TimeLens-8B	31.79	32.17	44.24	32.24	22.35	22	0
TimeLens2-2B	42.73	42.80	59.29	43.76	29.41	5	0
TimeLens2-4B	49.20	48.46	63.29	48.94	36.47	2	0
TimeLens2-8B	50.55	50.58	67.53	52.24	38.59	0	0
Qwen3-VL-8B	11.45	12.36	15.76	12.47	6.82	44	7
Qwen3.5 Flash	15.51	15.82	21.18	16.71	10.35	199	0
Qwen3.5 Plus	13.69	14.52	19.76	15.53	9.65	271	0
Qwen3.6 Flash	15.18	15.56	21.41	15.29	11.06	177	2
Qwen3.8-27B	20.68	21.54	28.24	22.35	16.24	10	1
MiMo V2.6 Pro	23.92	24.20	33.65	24.24	12.94	2	8
MiMo V2.6 Flash	24.40	24.16	33.88	22.35	13.65	2	7
DeepSeek V4 Vision	19.99	20.21	27.06	19.06	12.00	12	8
GLM-5.3 Flash	33.63	34.12	46.12	34.59	23.29	32	3
Closed-Source Models							
Gemini 2.5 Flash	12.49	13.16	18.35	13.41	6.59	86	8
Gemini 2.5 Pro	34.92	35.89	47.29	36.47	26.35	8	3
Gemini 3 Flash Preview	38.52	38.63	51.06	38.12	26.35	14	0
Gemini 3.1 Pro Preview	48.77	49.47	64.24	51.06	38.82	6	3
Gemini 3.5 Flash	42.16	42.01	57.18	43.76	30.59	11	1
Gemini 3.7 Flash	45.67	47.83	61.18	50.12	41.18	9	0
Gemini 3.8 Flash	46.76	47.86	61.88	49.18	39.53	12	0
Qwen3.8 Max	40.41	40.81	53.18	42.12	29.18	2	0
Qwen3.5 Omni Plus RT	14.02	15.11	20.24	15.29	10.82	18	9
Qwen3.8 Omni Flash	39.58	40.79	54.35	42.35	27.76	2	1
GPT-6 Astra	48.44	50.50	61.41	52.00	**43.29**	0	100
GPT-6 Sol	42.86	44.14	54.82	45.41	37.41	0	104
GPT-6 Luna	23.18	23.60	31.29	23.06	16.71	2	123
Muse Spark 1.1	45.68	46.80	62.59	49.65	35.53	8	10
Doubao Seed 2.0 Lite	**52.54**	**52.39**	**72.00**	**56.00**	37.88	0	3

has no title. The returned PTS is 216,000 with time base $1/90{,}000$ s and zero stream start. For the first occurrence of VUE-TR-v2 query 1045, "Aerial landscape views," the initial start window $[80, 82]$ s contains 48 frames on two pages at $24{,}000/1{,}001$ fps. ID 25 at 81.039292 s shows an indoor interview; returned ID 26 at 81.081000 s is the first aerial frame. Its PTS is 1,945,944 with time base $1/24{,}000$ s. Both starts are resolved without expansion; end boundaries are processed separately.

Spatial detail and verification limits. These examples were selected for abrupt visibility transitions: annotation overlays disappear, a title appears, or the requested view begins. They are qualitative illustrations, not a random sample for estimating boundary accuracy. We re-decoded the neighboring PTS positions to check the archived images against source frames. The football images occupy 283×212 pixels inside each uploaded cell; the other two occupy 320×180. Small movements and gradual action boundaries can still be difficult to resolve at these sizes, and provider-side image processing is unobserved. The separate content verifier sees 12 context frames and checks the event, not each boundary transition. Model confidence and content checks therefore do not certify one-frame accuracy across the benchmark. A comparison with larger cells and independent human boundary annotations remains to be performed.

Table 8: VUE-v1 on all 369 retained visual queries from 209 videos. Original and refined mIoU use identical predictions; recall uses refined references. Overlap and recall are percentages. Returned, Invalid, Abstained, and Failed are counts. Failed counts executions with no returned answer after their recorded stopping rule. Returned includes invalid answers and abstentions; Returned plus Failed is 369 for every row.

Model	Returned	Orig. mIoU	Ref. mIoU	R1@.5	R1@.7	Invalid	Abstained	Failed
Open-Source Models								
TimeLens-8B	369	46.88	47.28	46.88	39.30	13	0	0
TimeLens2-2B	369	53.43	54.39	54.20	44.17	9	0	0
TimeLens2-4B	369	56.28	56.88	57.18	47.43	8	0	0
TimeLens2-8B	369	59.45	59.89	59.89	51.22	1	0	0
Qwen3-VL-8B	369	27.83	28.19	29.27	21.95	45	1	0
Qwen3.5 Flash	369	25.06	25.82	24.66	22.22	163	0	0
Qwen3.5 Plus	369	25.19	25.52	27.37	21.68	189	0	0
Qwen3.6 Flash	357	26.59	26.88	27.10	21.14	159	0	12
Qwen3.8-27B	369	33.53	34.55	33.06	26.02	13	3	0
MiMo V2.6 Pro	365	36.86	37.10	36.04	27.91	3	6	4
MiMo V2.6 Flash	365	37.32	37.21	36.59	27.91	1	9	4
DeepSeek V4 Vision	369	37.48	38.12	37.13	29.54	17	5	0
GLM-5.3 Flash	357	50.66	51.31	51.49	41.46	14	0	12
Closed-Source Models								
Gemini 2.5 Flash	369	25.36	25.56	25.75	13.82	57	5	0
Gemini 2.5 Pro	369	42.05	42.28	42.55	30.89	22	2	0
Gemini 3 Flash Preview	369	46.20	47.12	46.88	35.50	17	0	0
Gemini 3.1 Pro Preview	369	55.82	56.95	56.64	48.51	6	1	0
Gemini 3.5 Flash	369	49.73	50.30	50.41	40.11	11	0	0
Gemini 3.7 Flash	369	48.63	49.66	51.76	41.46	16	0	0
Gemini 3.8 Flash	369	51.81	53.46	54.47	46.07	14	0	0
Qwen3.8 Max	369	53.45	54.49	54.74	43.63	1	1	0
Qwen3.5 Omni Plus RT	241	28.63	29.69	29.54	21.41	34	7	128
Qwen3.8 Omni Flash	369	50.89	52.67	52.30	40.65	3	0	0
GPT-6 Astra	369	56.16	57.60	56.91	50.41	0	57	0
GPT-6 Sol	369	54.25	55.04	54.74	47.97	0	47	0
GPT-6 Luna	369	41.74	42.59	41.73	33.60	2	47	0
Muse Spark 1.1	366	60.34	61.32	62.87	51.76	4	0	3
Doubao Seed 2.0 Lite	366	60.90	61.60	63.96	52.03	0	0	3

C.5 Additional Model Protocols And Coverage

The base response snapshot was started on September 25, 2026 at 00:25:30 UTC, with later execution retry and TimeLens2 continuation snapshots added afterward. A cell is finalized when every query in its fixed inventory has a terminal response or, for MiMo, Muse, Doubao, and Qwen Omni, a verified exhausted retry budget. GLM uses the recorded execution-failure stopping rule described below. Added TimeLens scores use the same 2,498/3,384/1,227 query sets as the initial models. All point rows use the same 500 FrameSnap-v1 references, and VUE-v2 rows use the same 425 matched visual queries out of 541 retained. Qwen3-VL-8B, TimeLens-8B, and the final released TimeLens2-2B, TimeLens2-4B, and TimeLens2-8B checkpoints have results on all six benchmarks. TimeLens-8B has returned answers for all 425 VUE-v2 queries. Table 9 lists returned-answer coverage for the added models and TimeLens continuations; Table 8 covers the full VUE-v1 evaluation.

The September 26 GLM-5.3 Flash snapshot finalizes all recorded no-answer executions as misses: 3/500 on Point500, 84/3,384 on ActivityNet, 7/1,227 on QVHighlights, and 34/425 on VUE-v2. Every query was attempted. The provider returned gateway timeouts, bad requests, image-parsing errors, and input URL or size errors. Some errors terminated requests before five attempts; the logs retain their actual attempt counts. These failures contribute zero accuracy and IoU, remain



Figure 12: Initial-window start selections with visible frame transitions. Each pair shows the preceding frame and the returned frame in teal. (a) A diving title first appears at 2.400000 s in ActivityNet. (b) An aerial view replaces an indoor interview at 81.081000 s in VUE-TR-v2. Images come from the archived requests; only sheet padding and display labels are changed.

in the full denominators, and are separate from illegal answers and abstentions. Point MAE uses legal returned answers. The first returned answer is retained regardless of validity or score. The previously completed Charades and VUE-v1 scores are unchanged.

The reported Omni models are Qwen3.8 Omni Flash and Qwen3.5 Omni Plus Realtime. Both have finalized results on Point500, the three TimeLens subsets, and both VUE benchmarks, for twelve model–benchmark results. Their VUE-v1 continuations preserve the two earlier diagnostic answers and evaluate the remaining 367 queries. Flash returns all 369 answers. After execution repair, Realtime has 241 returned answers and 128 finalized execution failures, with all 369 queries retained in the scoring denominator.

Tables group local checkpoints and hosted models by the availability of associated public visual weights. Hosted evaluations use API names. Qwen3.5 Flash, Qwen3.5 Plus, and Qwen3.6 Flash are associated with the public Qwen3.5-35B-A3B, Qwen3.5-397B-A17B, and Qwen3.6-35B-A3B models, respectively, and remain in the open-source group. This association does not establish that a provider serves identical weights or processing code. Qwen3.8 Max remains in the closed-source group for vision: its public Qwen3.8-2.4T-A95B counterpart accepts text only. The evaluated Omni routes have no identified matching public visual checkpoint. Models from the same family appear together.

The benchmark uses one configuration per reported Qwen variant: Qwen3.5 Flash, Qwen3.5 Plus, and Qwen3.6 Flash. Exact route versions are retained in the protocol below.

All hosted routes use ModelRouter. MiMo, Muse, Doubao, and the native-video Qwen routes receive the frozen silent video assets used by the original API evaluation. The Qwen Omni Realtime input is described below. GPT and GLM receive all sampled frames as timestamped image grids, with at most 50 JPEG sheets and image detail set to high. DeepSeek receives all sampled frames in at most 16 JPEG sheets, each capped at a 2,048-pixel long side, with high image detail. These input transports differ: provider-side video sampling and image resizing are not observable. Their scores describe the recorded routes and input protocols, rather than an isolated comparison of checkpoint weights.

MiMo, Muse, and Doubao use temperature 0 and a 32,768-token completion budget. DeepSeek and GLM use the same temperature and completion budget; GPT uses that completion budget with medium reasoning effort. GLM retains provider-default reasoning settings. The native-video Qwen routes use temperature 0, a 4,096-token output budget, and thinking disabled. These settings are retained from the recorded runs.

Qwen3.5 Omni Plus Realtime, abbreviated RT in the tables, uses a separate streaming configuration. We select at most one frame per second from the same frozen MP4, retaining source presentation timestamps, and send JPEG frames over WebSocket. Each Base64 JPEG is below 60 KiB to fit the gateway's 64 KiB message limit. A 16 kHz PCM silence stream carries the source timeline; no source audio is supplied. The first frame follows the required 100 ms of initial silence, while later timestamps and the total duration are preserved. Transport is accelerated to at most 20 frames per

Table 9: Prediction coverage for added models and the TimeLens continuations. The interval denominators are the retained and matched inventories in Table 4, after filtering. Counts include terminal abstentions, invalid answers, and provider refusals. Each denominator is fixed before scoring. Counts show actual returned answers; finalized execution failures receive zero score. Full VUE-v1 coverage is in Table 8.

Model	Point500	Charades	ActivityNet	QVHighlights	VUE-v2
Open-Source Models					
TimeLens-8B	500/500	2498/2498	3384/3384	1227/1227	425/425
TimeLens2-2B	500/500	2498/2498	3384/3384	1227/1227	425/425
TimeLens2-4B	500/500	2498/2498	3384/3384	1227/1227	425/425
TimeLens2-8B	500/500	2498/2498	3384/3384	1227/1227	425/425
Qwen3.5 Flash	500/500	2498/2498	3384/3384	1227/1227	425/425
Qwen3.5 Plus	500/500	2498/2498	3384/3384	1227/1227	425/425
Qwen3.6 Flash	500/500	2498/2498	3384/3384	1227/1227	425/425
Qwen3.8-27B	500/500	2498/2498	3384/3384	1227/1227	425/425
MiMo V2.6 Pro	499/500	2498/2498	3384/3384	1227/1227	420/425
MiMo V2.6 Flash	499/500	2498/2498	3384/3384	1227/1227	420/425
DeepSeek V4 Vision	500/500	2498/2498	3384/3384	1227/1227	425/425
GLM-5.3 Flash	497/500	2498/2498	3300/3384	1220/1227	391/425
Closed-Source Models					
Qwen3.8 Max	500/500	2498/2498	3384/3384	1227/1227	425/425
Qwen3.5 Omni Plus RT	497/500	2243/2498	3201/3384	1150/1227	149/425
Qwen3.8 Omni Flash	500/500	2498/2498	3384/3384	1227/1227	425/425
GPT-6 Astra	500/500	2498/2498	3384/3384	1227/1227	425/425
GPT-6 Sol	500/500	2498/2498	3384/3384	1227/1227	425/425
GPT-6 Luna	500/500	2498/2498	3384/3384	1227/1227	425/425
Muse Spark 1.1	495/500	2498/2498	3380/3384	1227/1227	390/425
Doubao Seed 2.0 Lite	495/500	2498/2498	3380/3384	1227/1227	390/425

wall-clock second. The query is supplied in the session instructions; generation uses temperature 0 and a 4,096-token budget, without an explicit thinking option. The provider's documented context limits are the most recent 240 s of video and 600 s of audio. Although the client sends the complete sampled input, earlier video context may be discarded by the server. RT therefore measures this streaming configuration, including its context limit.

The TimeLens2 continuations preserve existing Transformers answers and use SGLang only for missing executions. They keep the frozen prompts, video tensors, decoded timestamps, per-query seeds, and 4,096-token output budget, with temperature 0.01, top-p 0.001, and top-k 1. SGLang uses BF16 and continuous batching; its kernels can differ numerically from Transformers. Run contracts record the checkpoint, backend, preprocessing hashes, and generation settings. Already complete TimeLens2-4B cells retain their original prediction snapshot.

The Qwen3-VL-8B VUE-v1 continuation uses SGLang with tensor parallelism of two across four independent replicas. It preserves the two earlier answers and adds 367 responses with the frozen inputs, source timestamps, per-query seeds, temperature 0.7, top-p 0.8, top-k 20, and 4,096-token budget. Preflight checks verify tokenizer and video-tensor hashes against the original local configuration. All 369 queries have returned answers; there are no execution failures. Of these answers, 36 have truncated structures and nine have zero-length intervals, giving 45 invalid answers. One explicit empty prediction is an abstention. These outcomes all receive zero IoU; only the 45 invalid answers enter the invalid count.

We retain the first terminal answer in archived ledger order. Explicit no-match answers and provider refusals receive zero score and are excluded from illegal-answer counts. The added routes have no answer-format retry; the initial nine models retain the recorded retry protocol described below. No answer is selected or rewritten based on its overlap with a reference. The same repaired point/span validity checks apply, and TimeLens retains its first-interval extraction rule. Numeric strings are accepted when unambiguous; null endpoints, zero-length spans, and reversed spans are invalid.

The Point500 and VUE parser audit applies the same repairs to all models. It accepts singleton arrays around explicit point objects, including `event_timestamp`, redundant nesting around temporal pairs, and unquoted clock values in named timestamp fields. When explanatory text quotes format examples, a complete final answer block can supply the prediction. The audit preserves all previously valid predictions and does not change the original nine-model diagnostic results. Spatial `bbox_2d` and `point_2d` coordinates remain invalid; unpaired numeric sequences are not split into guessed intervals. Interval-only answers remain invalid for Point500, and no invalid interval is removed to salvage the rest of an answer.

For MiMo V2.6 Pro and Flash, Muse Spark 1.1, and Doubao Seed 2.0 Lite, each remaining execution failure receives five additional attempts with the same input, prompt, model route, and generation settings. A model answer ends retries even if it is illegal; illegal answers remain in the reported counts and receive zero score. Requests still without an answer after the fifth attempt also receive zero accuracy and IoU, while remaining separate from illegal answers and abstentions. Their point error is undefined and does not enter valid-answer MAE. Attempt logs retain query IDs, timestamps, processing stage, HTTP status, and error messages. Coverage counts report actual responses rather than relabeling exhausted failures as answers. For Muse and Doubao, each scored inventory contains 5/500 finalized execution failures on Point500, 4/3,384 on ActivityNet, and 35/425 on VUE-v2; Charades and QVHighlights have none. Each MiMo model has 1/500 finalized execution failures on Point500 and 5/425 on VUE-v2, with none on the three TimeLens subsets. These counts are separate from the illegal-answer counts in the tables.

The initial Omni runs allow at most five execution attempts per query, with the same first-answer and zero-score rules. Realtime exhausts this budget on 3/500 Point500 queries, 255/2,498 Charades queries, 183/3,384 ActivityNet queries, 77/1,227 QVHighlights queries, and 276/425 VUE-v2 queries. These failures remain in the accuracy and IoU denominators and are separate from illegal answers. Qwen3.8 Omni Flash has no exhausted execution failures across its six reported benchmarks.

Realtime VUE-v1 initially has 144 execution failures after five attempts per query. A separate repair retains the model, prompts, sampled frames, and source clock, and lowers concurrency to two sessions. It allows at most five additional attempts for 16 short-video connection failures, all of which return an answer on the first new attempt, and tests three representative long-video failures once each. The latter still fail. Of the 128 remaining failed queries, 118 have an audio/video count-mismatch error in their attempt history; the server's reported video count stops at 500 or 501. The other ten have startup, timeout, or connection errors without that diagnostic. These are execution failures, not model answers. They receive zero IoU under the full 369-query denominator. The final 241 returned answers contain 34 invalid answers and seven abstentions; refined mIoU is 29.69. Original attempts and previously returned answers are preserved in the repair archive.

The route IDs are `MiMo-V2.6-Pro`, `mimo-V2.6-flash`, `muse-spark-1.1`, `doubao-seed-2-0-lite-260428`, `gpt-6-astra`, `gpt-6-sol`, `gpt-6-luna`, `glm-5.3-flash`, and `deepseek-v4-flash-vision-exp`. The Qwen routes share the prefix `dashscope.`: `qwen3.8-27b`, `qwen3.5-plus-2026-02-15` (displayed as Qwen3.5 Plus), `qwen3.5-flash`, `qwen3.6-flash`, and `qwen3.8-max`. Qwen3.8 Max here denotes the hosted visual API, not a visual evaluation of its public text-only checkpoint. The Omni routes use the same `dashscope.` prefix: `qwen3.5-omni-plus-realtime` and `qwen3.8-omni-flash`. Requested route IDs, returned model IDs when available, generation settings, media hashes, raw answers, and scoring-code hashes are retained in the snapshot.

C.6 Per-Query IoU Distributions

We compare all nine models on the same 3,363 Charades queries, the completed TimeLens subset shared by the three model families. The analysis uses the original released references and the first parsed interval, matching the with-video baseline in Table 11. Each observation is a query's IoU; mIoU is the mean of this distribution.

Qwen3-VL-8B has 308 zero-overlap predictions and 22 exact interval matches, compared with 276 and 579 for TimeLens2-4B. Gemini 3.8 Flash has 216 and 486. Thus the strongest endpoint concentration is model-dependent: TimeLens2-4B places 17.2% of queries at IoU 1, while Qwen3-VL-8B



Figure 13: Per-query IoU on the same 3,363 Charades queries with released references. Orange and green bars count exact zeros and exact ones; blue bins hold $0 < \text{IoU} < 1$; dashed lines mark mIoU. The share of exact matches ranges from 0.7% to 17.2% across models.

places only 0.7% there. These are exact matches to the released coordinates; they do not measure agreement with the TickTock-Refined references scored separately in Table 6.

C.7 Video Inputs And Generation Settings

Point500 and TimeLens use a nominal rate of 4 fps; VUE uses 1 fps. The Omni Realtime configuration is an exception, with its sampling and context limits described in Appendix C.5. The base inputs cover the complete video with a cap of 2,048 frames. For local models, the sampler uses the source frame count and reported rate to choose indices. When the cap is exceeded, it chooses uniformly spaced interior indices across the video. It reads decoded presentation timestamps for the selected frames and assigns each two-frame temporal patch the mean of its endpoint timestamps, repeating the last frame when padding is needed. The spatial preprocessing uses a minimum area of 32^2 pixels and maximum areas of 640^2 for Point500/TimeLens and 480^2 for VUE. The total pixel parameter is $128{,}000 \times 32^2$. With n frames after temporal padding, the per-frame area cap is the smaller of the task maximum and $2 \times 128{,}000 \times 32^2/n$, before the processor's spatial rounding.

Gemini inputs are silent videos re-encoded with a uniform FFmpeg frame rate, a rational rate specification, a 90,000-unit video time scale, and CRF 18. The frame and spatial limits follow the same task budgets. Provider-internal sampling remains unobserved, and re-encoding does not guarantee that Gemini and the local models see identical source frames. Audio is excluded and no separate subtitle channel is supplied. Text already visible in video pixels is part of the visual input.

All Gemini routes use temperature 0, a 32,768-token maximum output budget, `MEDIA_RESOLUTION_LOW`, and the provider's default thinking setting. Qwen3-VL-8B uses temperature 0.7, top-p 0.8, and top-k 20. TimeLens2-4B uses temperature 0.01, top-p 0.001, and top-k 1. Both local configurations enable sampling and caching, set the repetition penalty to 1.0, and allow 4,096 new tokens. Their per-query seed is the SHA-256 hash of the query ID modulo 2^{31}. The original local runs use Transformers 4.57.3, `qwen-vl-utils` 0.0.14, and SDPA attention; later SGLang continuations are described in Appendix C.5. The comparison uses the recorded model settings; it does not isolate the effects of decoding parameters. The API runner retries network failures and HTTP 408, 429, and 5xx errors, with at most five attempts. It does not retry a returned answer because it is invalid or inaccurate.

C.8 Recorded Answer Retries

Each first-pass answer rejected by the original parser or temporal-range checks receives exactly one additional attempt. The second answer is used regardless of its reference error; answers that fail again count as incorrect. The selected retry IDs are frozen before the parser repair; all final answers use the repaired parser. Unselected responses, prompts, videos, and generation settings are unchanged. No retry is selected using the reference timestamp.

For VUE, Each answer rejected by the original first-pass parser receives one additional attempt with an explicit numeric-seconds instruction and a JSON response schema for arrays of two-number intervals. The selected retry IDs are frozen before the parser repair. The second response is scored regardless of reference overlap. Unselected responses are unchanged. All final answers use the repaired parser; video assets, frame budget, and interval-set IoU computation are retained. The first-pass response archive remains available.

C.9 Task Prompts And Parsing

The Point500 prompt supplies the query, available request context, video duration, and search-start time. It asks for "the single key or trigger moment that matches the query in its request context" and instructs the model to follow onset, completion, or recurrence only when specified. Earlier video frames remain available as context. The output must be one JSON object, {`"timestamp":` `<seconds>`}, on the original video's time axis. The repaired parser accepts a timestamp object, a scalar time, or one uniquely specified time in prose. Numeric strings, explicit time units, and clock notation are converted to seconds without rounding. Optional code fences and surrounding explanation are allowed; neither JSON syntax nor a fixed number of decimal places is required. Conflicting point candidates and interval answers are rejected. Range checks apply after parsing, including the query-specific search lower bound. Comparisons at the search-start and video-end boundaries use an absolute tolerance of 10^{-9} seconds, with zero relative tolerance. This prevents decimal values such as 3.06667 from being rejected against the stored bound 3.0666700000000002. Predictions and references retain their original values when computing time errors.

The TimeLens prompt is: "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'." We use the archived TimeLens parser and score the first parsed interval. The VUE prompt asks: "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]]." The repaired VUE parser accepts interval arrays, start/end objects, and explicit textual ranges, including unambiguous strings such as `551.651s` and `01:30.500`. The latter denotes 90.5 seconds. All intervals must have finite, nonnegative endpoints with start strictly before end. Any null, zero-length, reversed, or incomplete interval invalidates the whole answer. A missing outer list delimiter may be restored only when every inner interval and endpoint is already complete. Unlabelled numeric tuples with more than two entries are rejected rather than assigned an inferred minute/second layout. All intervals in valid answers are scored together. Explicit empty arrays, the sentinel -1, direct no-match declarations, and provider refusals are abstentions. They receive zero overlap under the full denominator and are excluded from invalid-answer counts. Other unparsable prose remains invalid. Point abstentions fail every tolerance; their time error is undefined and is excluded from valid-only MAE. The full 369-query VUE-v1 evaluation uses the same audited parser as the current VUE-v2 tables. The full prompt templates and parser source are included in the score archive; no language-model answer rewriting is used during scoring.

C.10 Point Reference Versions And Category Results

The original runs used the inherited human times. To match the benchmark definition in Section 3.3, we rescore the archived first-pass predictions against FrameSnap-v1. Query text, request context, and video duration agree between the two input releases. For one query, the snapped reference is 3.333 microseconds before the rounded original search-start bound; the derived release moves that bound earlier by the same amount. Its original prompt remains in the inference archive. This change alters no legal-output count for any of the nine evaluated models. This reference-only comparison used the same original parser on both releases. The subsequent parser repair is audited separately and retains the Point500 temporal-range checks.

The audit retains hashes of both reference releases, the saved predictions, the frozen scorer, and both score versions. All paper Point500 results use FrameSnap-v1. Table 12 reports category accuracies alongside additional temporal tolerances, using the same final predictions and answer-retry protocol as the main point table; the original responses and reference-comparison scores remain in the audit.

Category sizes range from 25 to 150 queries. Estimates describe this fixed selection, with precision that can differ across category sizes. Valid-only error diagnostics use each model's legal-output subset. Their rows therefore differ in which queries contribute to the error estimate.

C.11 Recovered VUE Scope And Media Exceptions

The original-reference VUE-v2 diagnostics use 538 visual queries from 406 recovered videos. Query `vue_tr_v2_1353` has damaged media and contributes zero overlap in that diagnostic cohort. The matched 425-query TickTock-Refined comparison excludes this query. The full refined VUE-v1 evaluation covers 369 queries on 209 videos; the two shared answers from the earlier media-availability check are preserved. These populations are reported separately.

C.12 Reference Eligibility And Modality Screening

The retained inventory is in Table 4. Approximately 98.2% and 98.9% of eligible VUE-TR and VUE-TR-V2 visual-query endpoints are integer seconds. This describes the stored coordinates, not their semantic accuracy. Only queries whose complete interval sets pass refinement enter the benchmark export.

The VUE scope follows its released `query_modality=vision` labels. TimeLens provides no comparable per-query labels. Gemini screens query text and selected video evidence; GPT-6 Astra reviews flagged cases using visual evidence. The final mask removes 50 audio-dependent queries and, separately, 42 unresolved cases: 24 below the review confidence threshold, seven explicitly uncertain, ten provider refusals, and one execution failure. Their IDs and exclusion reasons are logged. Charades removes 1 audio-dependent and 19 unresolved queries, ActivityNet 22 and 14, and QVHighlights 27 and 9. The 42 unresolved cases are excluded rather than relabeled as audio-dependent. Retained queries include candidates cleared by text screening; this is not an official modality annotation or exhaustive human verification of visual sufficiency.

D Benchmark Diagnostics And Robustness

We examine how tolerance, task composition, and answer validity affect the reported scores. All analyses reuse the saved responses and the reference versions from Section 4.

D.1 Temporal Tolerance And Error Structure

Figure 6(a) separates illegal answers from valid predictions at different distances from the human reference. Valid predictions more than five seconds away account for 20.4%–36.6% of all queries across the nine models, exceeding each model's illegal-answer rate. Fixing output validity alone would therefore leave many localization errors unresolved. At a 100-ms tolerance, accuracy ranges from 3.8% to 21.6%. Frame-aligned reference coordinates permit this measurement, but do not imply that the models observed every intervening frame or that the human event target is semantically unambiguous at this scale.

The highest observed accuracy depends on the tolerance. Gemini 3.7 Flash reaches 63.4% within two seconds, compared with 62.6% for Gemini 3.8 Flash. At five seconds, their order reverses: 77.6% and 79.2%, respectively. The complete curves and the fixed-tolerance scores in Appendix D make these differences visible without selecting a tolerance after inspecting the ranking.

D.2 Task Composition And Cross-Benchmark Rankings

Figure 14(a) subtracts each model's overall two-second accuracy from its accuracy on each Point500 task. T4 is below the overall score for all nine models, by 5.0–21.0 percentage points, while T5 is above it by 1.7–18.2 points. T4 contains only 25 videos, so its estimates are more sensitive to

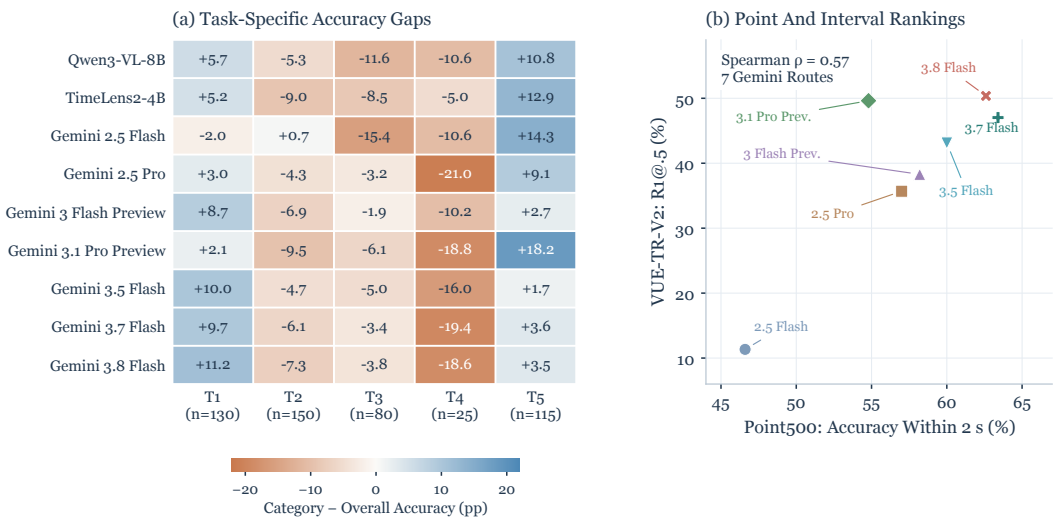

Figure 14: Rankings depend on the target. (a) Point500 accuracy within 2 s on each task minus the model's overall accuracy, in percentage points. (b) The seven Gemini routes on Point500 (accuracy within 2 s) and VUE-TR-V2 (R1@.5).

Table 10: Metadata-only priors on Point500. Each control uses the supplied search start and video duration, with no video pixels or query content. Accuracy is in percent over all 500 queries; MAE is in seconds. The uniform row gives analytic expectations. All controls stay within the allowed window.

Control	±1 s	±2 s	±5 s	MAE (s)	Illegal (%)
Search start	8.2	17.2	31.8	34.55	0.0
Window midpoint	1.8	4.4	10.0	50.64	0.0
Uniform time (expected)	3.0	5.8	12.7	55.35	0.0

individual examples than those of the larger categories. These are descriptive differences across the inherited task groups.

The seven common Gemini routes have a Spearman rank correlation of 0.57 between Point500 two-second accuracy and VUE R1@.5 (Figure 14(b)). The ranking differs between the two targets. The datasets differ in queries, target definition, and input budget, so this comparison does not isolate a single cause of the ranking change. It supports reporting point and interval localization separately.

D.3 Temporal-Prior Controls

The Point500 prompt includes a search-start time and video duration. We test three controls that use only these two values, without the video or query content: the search start, the midpoint of the allowed window, and a uniform time within that window. For the uniform control, we compute expected accuracy and MAE analytically for each query and then average over all 500 queries. No random model response is sampled.

Table 10 reports the three controls.

The search-start control reaches 17.2% within two seconds, versus 4.4% for the midpoint and 5.8% for uniform time. All nine models in the original diagnostic cohort exceed these controls, with two-second accuracy of 41.0%–63.4%. The relatively strong search-start result also shows that this supplied metadata carries a temporal prior. These controls measure its direct predictive value; a matched inference ablation would be needed to establish how much each model uses visual evidence.

Table 11: Filename-and-query control on all 3,363 original-reference Charades queries per model. mIoU is in percent; Drop is video minus filename-and-query mIoU in percentage points. No Interval counts responses without a parsed interval, including requests for the missing video. These are misses and remain in the denominator.

Model	Video	Filename + Query	Drop (pp)	No Interval
Open-Source Models				
TimeLens2-4B	58.11	14.19	43.92	0
Qwen3-VL-8B	46.63	9.77	36.86	0
Closed-Source Models				
Gemini 2.5 Flash	42.91	7.31	35.60	105
Gemini 2.5 Pro	46.49	11.42	35.06	4
Gemini 3 Flash Preview	53.71	12.37	41.34	2
Gemini 3.1 Pro Preview	53.00	13.14	39.86	2
Gemini 3.5 Flash	59.26	16.58	42.68	10
Gemini 3.7 Flash	58.89	15.09	43.80	1
Gemini 3.8 Flash	59.54	13.86	45.68	1

D.4 Sensitivity To The Recorded Retry Protocol

We rescore both the first response and the completed retry protocol with the same parser and references. On Point500, the change in two-second accuracy ranges from 0.0 to +2.2 percentage points. On VUE, the change in mean IoU ranges from -2.29 to $+2.51$ points. Retrying an answer rejected by the original parser can therefore improve or reduce the final score. The VUE retry also adds numeric-seconds instructions and a JSON response schema, so its effect includes that prompt change. Appendix D reports every model's first and final score, paired confidence intervals, and the effect of excluding illegal answers from the denominator.

D.5 Filename-And-Query Ablation

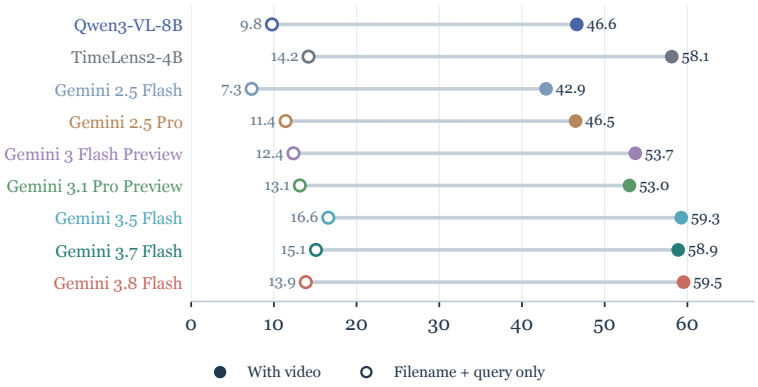

Figure 15: Charades mIoU on the 3,363 original-reference queries with the video (filled) and with only the video filename and the query (open).

We run each of the nine models on all 3,363 original-reference Charades queries from Figure 13. Each prompt contains only the video basename, such as `3MSZA.mp4`, the query, and the original interval-format instruction. It contains no media, image placeholders, full path, duration, FPS, source labels, or tool access. Runtime checks verify zero media parts for the APIs and no vision tokens or vision tensors for the local models. The first parsed interval is scored with the same frozen parser and IoU computation as the video baseline, without duration clipping. Explicit refusals and responses without a parsed interval receive zero; all runs have 3,363 returned responses.

Generation parameters match each archived video run. Gemini uses temperature 0 and a 32,768-token output limit, with provider-default thinking settings. Returned model versions are checked

against the requested route and fallback is disabled. Transport failures are retried, while completed answers are retained regardless of parsing or correctness. Rate-limit and concurrency changes are recorded in the execution history.

The local models use 4,096 output tokens. TimeLens2 uses temperature 0.01, top-p 0.001, and top-k 1; Qwen3-VL uses 0.7, 0.8, and 20. Text inputs are left-padded in batches of 16, with an independent CUDA random stream per query initialized from its original sample-ID seed. BF16 batching can change logits and individual responses relative to sequential inference: a three-query preflight matched all three TimeLens2 answers and two of three Qwen answers. The comparison preserves sampling settings without claiming bitwise equivalence between execution modes.

TimeLens2 returns 18 distinct intervals without video and achieves 48 exact matches to released coordinates, compared with 579 with video. Qwen3-VL returns 486 distinct intervals and achieves seven exact matches, compared with 22 with video. Nonzero overlap and occasional exact matches therefore occur without visual evidence. The experiment does not separate filename memorization from query or interval priors; doing so would require additional filename-masking or query-shuffling controls. All 30,267 paired predictions, raw responses, input hashes, and per-query scores are archived.

D.6 Frozen Predictions And Analysis Units

The following point and VUE diagnostics use Point500 FrameSnap-v1 and the recovered 538-query VUE-TR-V2 vision subset with original references. They retain that cohort separately from the 425-query refined main comparison. We freeze the final per-query scores, first-pass responses, retry IDs, reference files, and parser hashes before computing the diagnostics. The same repaired parser scores both first and final answers. No new inference calls or reference-based choices between responses enter these comparisons.

Point500 has one query per source video. For paired retry differences we resample its 500 videos with replacement and recompute the mean difference. For VUE, we resample the 406 source videos with replacement, retain all queries within every sampled video, and recompute the query-weighted difference over the resulting sample. We use 10,000 paired bootstrap replicates with seed 20260924 and report the 2.5th and 97.5th percentiles. These intervals describe variation across the fixed benchmark population, not variation across repeated model runs.

D.7 Tolerance And Denominator Sensitivity

Table 12 combines stricter and looser Point500 tolerances with task-specific results. The denominator remains 500 at every threshold. Figure 16(b) shows how restricting the denominator to valid predictions increases apparent accuracy. For example, Gemini 2.5 Flash rises from 46.6% to 53.1% within two seconds, and TimeLens2-4B rises from 41.0% to 47.5%. This is a denominator change with identical predictions. The primary accuracy metric counts illegal answers as incorrect; valid-only errors remain diagnostics.

The duration breakdown in Figure 16(a) contains 23, 288, 121, and 68 videos in the four bins. Performance is not uniformly monotone with duration across models. Video length, task composition, and effective sampling rate vary together in these runs; the breakdown does not estimate the causal effect of length or frame rate.

D.8 First Responses And Completed Retries

Table 13 retains the original retry selection. Each selected ID receives one additional answer, and the second response is used regardless of its accuracy. IDs were selected by the original parser before its repair; some selected first responses are valid under the repaired parser. We do not redefine the retry population after reparsing. Thus a replacement can either improve or harm the score, as shown in Figure 17.

For VUE, Gemini 2.5 Pro gains 2.51 points with a paired 95% interval of $[1.50, 3.59]$, while Gemini 3 Flash Preview loses 2.29 points with an interval of $[-3.80, -0.84]$. Both outcomes are retained. The comparison evaluates the recorded format-repair protocol with fixed retry IDs and does not select the better of two responses using the benchmark reference.

Table 12: Point500 diagnostics. Additional tolerance accuracies use all 500 queries. Accuracies within 1, 2, and 5 seconds appear in Table 1. Median absolute error uses valid predictions only; illegal counts are out of 500. Task columns report two-second accuracy over their full category denominators. T1: event notification; T2: risk/error interjection; T3: cumulative-state tracking; T4: nonverbal/workflow cues; T5: proactive assistance. Illegal answers count as incorrect. Bold marks category maxima.

	Extra Tolerances (s)					Diagnostics		Tasks (±2 s)				
Model	0.1	0.2	0.5	10	30	Median (s)	Illegal Count	T1 130	T2 150	T3 80	T4 25	T5 115
Open-Source Models												
Qwen3-VL-8B	6.2	11.4	25.8	67.4	79.8	2.33	25/500	52.3	41.3	35.0	36.0	57.4
TimeLens2-4B	3.8	8.4	18.6	62.4	72.4	2.35	68/500	46.2	32.0	32.5	36.0	53.9
Closed-Source Models												
Gemini 2.5 Flash	12.4	16.0	28.8	66.4	74.4	1.76	61/500	44.6	47.3	31.2	36.0	60.9
Gemini 2.5 Pro	17.4	25.6	36.8	78.4	85.8	1.24	11/500	60.0	52.7	53.8	36.0	66.1
Gemini 3 Flash Preview	18.8	26.2	40.0	79.4	87.6	1.00	14/500	66.9	51.3	56.2	**48.0**	60.9
Gemini 3.1 Pro Preview	16.2	22.8	34.6	73.8	81.8	1.33	16/500	56.9	45.3	48.8	36.0	**73.0**
Gemini 3.5 Flash	21.0	28.4	40.2	80.0	88.2	0.98	16/500	70.0	55.3	55.0	44.0	61.7
Gemini 3.7 Flash	21.6	30.0	43.0	84.8	92.6	0.77	3/500	73.1	**57.3**	**60.0**	44.0	67.0
Gemini 3.8 Flash	20.2	29.4	43.8	84.8	91.4	0.77	2/500	**73.8**	55.3	58.8	44.0	66.1

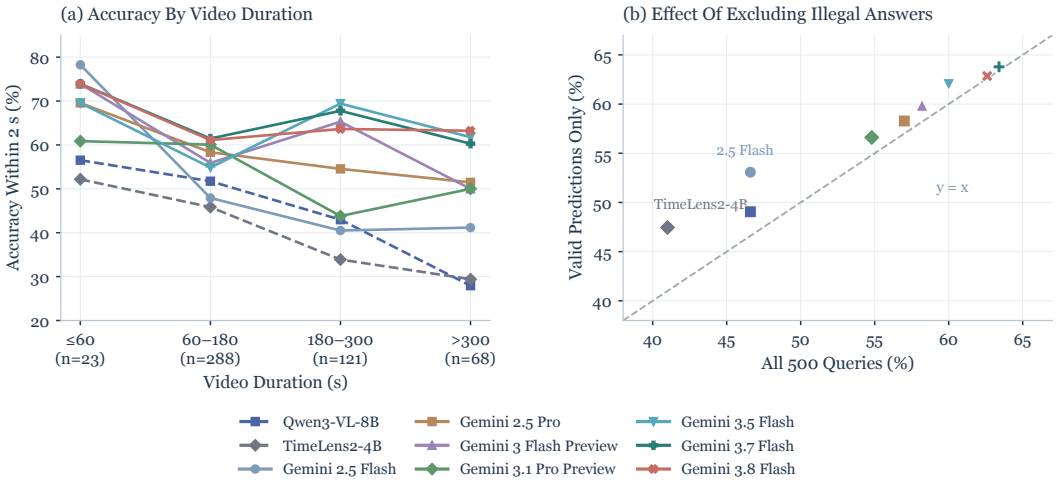

Figure 16: Point500 duration and denominator diagnostics. (a) Accuracy within two seconds in four video-duration bins, with each bin's full denominator. (b) Accuracy over all 500 queries versus accuracy over valid predictions only. The diagonal marks equality; removing illegal answers increases the score without changing any prediction.

E Training Annotation Diagnostics

E.1 Timestamp Distributions Before And After Refinement

We compare the released TimeLens training boundaries with their paired Gemini refinements in the valid-span pool. The comparison contains 85,936 TimeLens-100K intervals and 83,240 TimeLens2-93K intervals, with two boundaries per interval. Each before/after pair uses the same retained record.

All original endpoints in these paired records are integer seconds. After refinement, 14.1% of TimeLens-100K endpoints and 6.1% of TimeLens2-93K endpoints remain integers. Median absolute shifts are 0.333 s and 0.300 s; the maxima are 37.000 s and 7.155 s. Figure 18 measures how the annotation coordinates change. These archived proposals precede the correction of an OpenCV

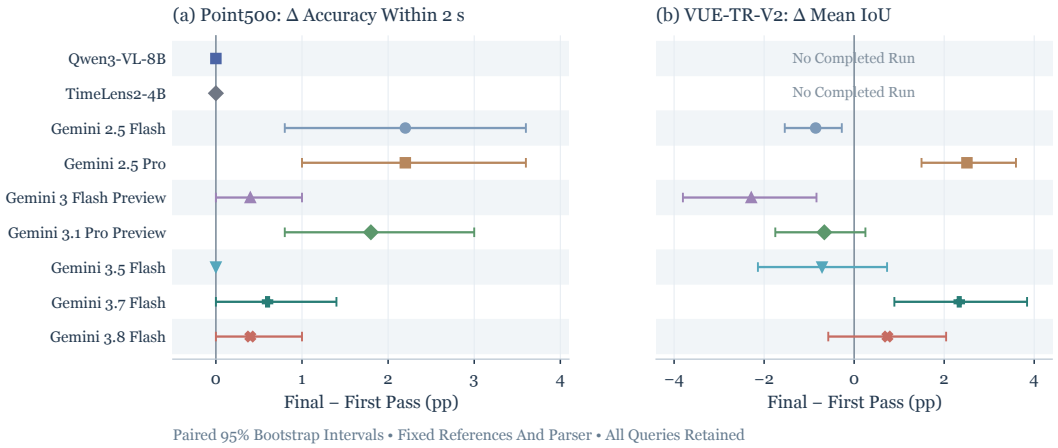

Figure 17: Effect of the completed answer-retry protocol. Points show the final minus first-pass score, in percentage points; bars show paired 95% bootstrap intervals over source videos. Point500 uses two-second accuracy and VUE uses mean IoU. The VUE retry changes output instructions and the JSON schema; these are protocol comparisons rather than an isolated sampling ablation.

Table 13: First-pass and final scores under the recorded retry protocol. Retries counts selected query IDs. Point500 uses accuracy within two seconds and all 500 queries; VUE-TR-V2 uses mean IoU and all 538 queries, including the damaged-media query. Scores are percentages, and differences with paired 95% bootstrap intervals are percentage points. Bootstrap units are source videos, retaining all VUE queries within each sampled video.

Model	Retries	First	Final	Δ [95% CI]
(A) Point500: Accuracy Within 2 s, $N = 500$				
Open-Source Models				
Qwen3-VL-8B	29	46.60	46.60	+0.00 [+0.00, +0.00]
TimeLens2-4B	73	41.00	41.00	+0.00 [+0.00, +0.00]
Closed-Source Models				
Gemini 2.5 Flash	90	44.40	46.60	+2.20 [+0.80, +3.60]
Gemini 2.5 Pro	34	54.80	57.00	+2.20 [+1.00, +3.60]
Gemini 3 Flash Preview	23	57.80	58.20	+0.40 [+0.00, +1.00]
Gemini 3.1 Pro Preview	31	53.00	54.80	+1.80 [+0.80, +3.00]
Gemini 3.5 Flash	25	60.00	60.00	+0.00 [+0.00, +0.00]
Gemini 3.7 Flash	9	62.80	63.40	+0.60 [+0.00, +1.40]
Gemini 3.8 Flash	8	62.20	62.60	+0.40 [+0.00, +1.00]
(B) VUE-TR-V2: Mean IoU, $N = 538$				
Closed-Source Models				
Gemini 2.5 Flash	44	12.32	11.47	−0.86 [−1.54, −0.27]
Gemini 2.5 Pro	41	31.21	33.72	+2.51 [+1.50, +3.59]
Gemini 3 Flash Preview	210	39.62	37.34	−2.29 [−3.80, −0.84]
Gemini 3.1 Pro Preview	46	47.08	46.42	−0.66 [−1.75, +0.25]
Gemini 3.5 Flash	158	40.74	40.02	−0.71 [−2.14, +0.73]
Gemini 3.7 Flash	131	41.47	43.80	+2.33 [+0.89, +3.84]
Gemini 3.8 Flash	145	44.38	45.11	+0.74 [−0.57, +2.04]

read-order error that could associate a frame with the preceding timestamp. They are not treated as verified frame matches or as outputs of the native-frame ID-selection protocol. Semantic boundary accuracy requires independent event references and is not established by the spread of fractional values.

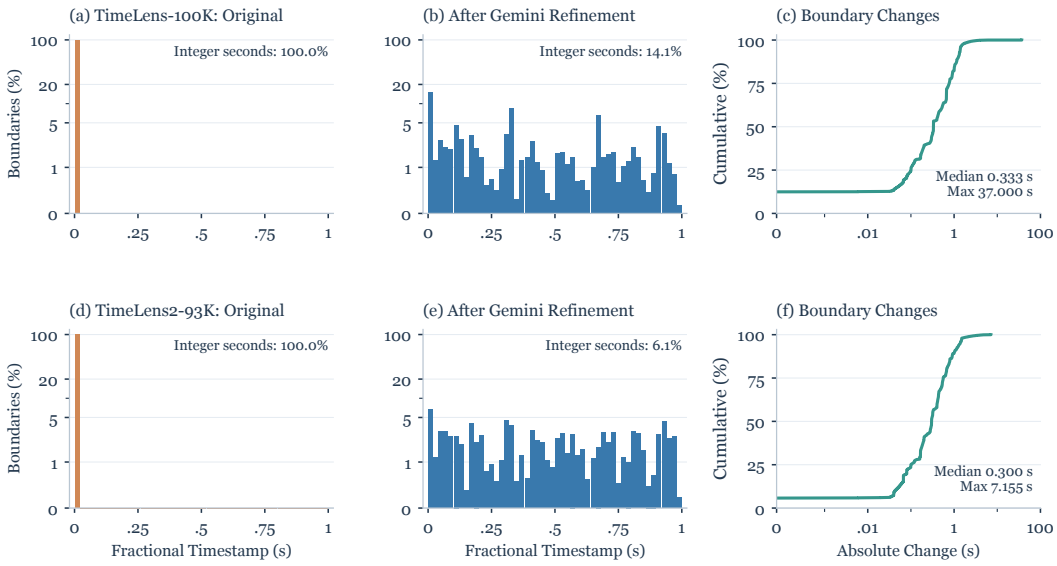

Figure 18: TimeLens training timestamps before and after Gemini refinement. Rows show TimeLens-100K and TimeLens2-93K, using 171,872 and 166,480 paired boundaries. Fractional parts use 20-ms bins; the vertical scale is logarithmic above 1%. Right panels show absolute boundary changes, including the full tail.