TOWARDS MULTIMODAL LIFELONG UNDERSTANDING: A MULTI-SCALE BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal systems can access increasingly long video histories, but their ability to connect observations of the same subject across days and months remains insufficiently evaluated. We formalize the **Lifelong Horizon**, distinguishing recorded duration from the physical temporal span of a persistent subject's history. To evaluate this setting, we introduce **MM-Lifelong**, a benchmark with 181.1 hours of video and 1289 questions across Day, Week, and Month scales. Its Month subset contains 105.6 hours of footage spanning 51 days. Every question is paired with manually annotated clue intervals. We evaluate answer quality and temporal grounding separately, using question-specific rubrics to accommodate valid alternative answers and partial correctness. Experiments with multimodal large language models (MLLMs) and agents reveal uneven gains from larger input budgets, with the best MLLM scoring 26.53 out of 100. Among agents, answer quality and grounding rankings diverge. Error analysis identifies lost answer-relevant details, confusion between event instances, and unresolved conflicting observations. These findings motivate memory systems that preserve event-specific evidence and verification that checks its support for the final answer.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have expanded the amount of video that models can process within a single context Yang et al. (2025a); Anil et al. (2023). In parallel, agentic systems use retrieval, external memory, and tools to access observations beyond that context Zhang et al. (2025); Zuo et al. (2025). These developments create an opportunity to study multimodal understanding over a subject's extended history. Across days and months, similar activities recur, attributes and locations change, and observations are interrupted by recording gaps. Answering questions about such histories requires identifying specific event instances and relating evidence across recording sessions. We study this capability as **multimodal lifelong understanding**.

Existing benchmarks provide only partial coverage of this setting. Long-video benchmarks such as Video-MME Fu et al. (2025) and LongVideoBench Wu et al. (2024) evaluate understanding within individual recordings. Multi-video benchmarks Peng et al. (2025); Zhu et al. (2025) broaden the input to multiple clips, which need not follow a persistent subject along a shared timeline. EgoLife Yang et al. (2025b) and TeleEgo Yan et al. (2025) establish longitudinal evaluation over several days, providing an important foundation for studying subject-specific histories. Month-scale histories extend this setting to recurring events and changing contexts distributed across substantially longer physical timelines. Evaluating them requires distinguishing *observational duration*, the amount of recorded footage, from *physical temporal span*, the elapsed time covered by those observations. We formalize this distinction through the **Lifelong Horizon**: sustained observation of a persistent subject across time.

To evaluate this setting, we introduce **MM-Lifelong**, a benchmark comprising 181.1 hours of video and 1289 questions across Day-scale gameplay, Week-scale egocentric life, and Month-scale livestreams. The Month subset contains 105.6 hours of footage distributed across 51 days, exposing a substantial separation between observed playback and physical temporal span (Figure 1(a)). Questions probe both the recovery of specific events from extensive histories and the integration of observations from disjoint intervals. Every question is paired with manually annotated clue intervals that identify its supporting evidence. We evaluate answer quality and temporal localization separately, using

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.

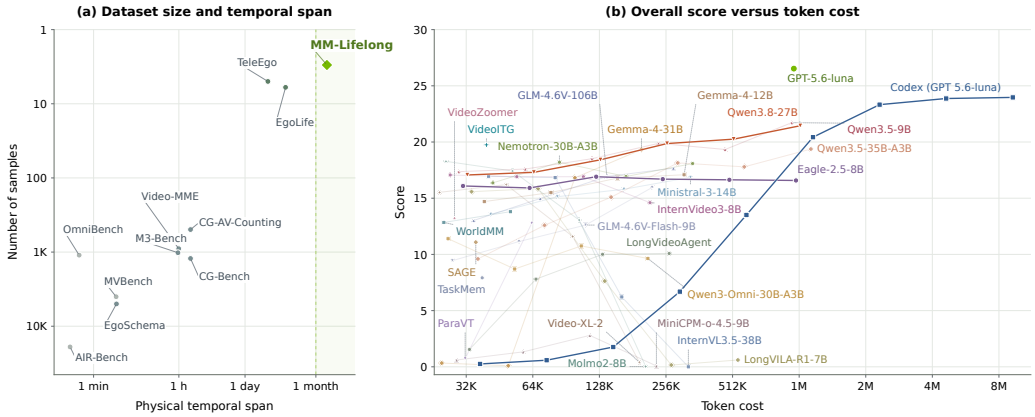


Figure 1: **Overview of MM-Lifelong. (a)** Dataset sample count and maximum physical temporal span; MM-Lifelong covers 51 days. **(b)** Overall score versus token cost for 18 MLLMs and 8 agents.

question-specific rubrics to accommodate valid alternative answers and partial correctness. This design supports examining both what a system answers and where it can locate the relevant evidence.

We evaluate end-to-end MLLMs and agentic systems without benchmark-specific training. The best evaluated MLLM scores 26.53 out of 100, and increasing input budgets produces uneven gains across models (Figure 1(b)). Among agents, stronger temporal grounding does not consistently accompany better answers. Analysis of representative failures identifies three recurring problems: loss of answer-relevant details, confusion between event instances, and failure to resolve conflicting observations. These findings identify concrete requirements for lifelong multimodal systems: preserving event-specific evidence, maintaining its temporal associations, and verifying the observations supporting an answer.

2 Related Work

Multimodal Understanding Benchmarks. Multimodal evaluation has progressed from static single-image tasks (e.g., MMMU Yue et al. (2024), MMBench Liu et al. (2024)) to dynamic video understanding. Early video benchmarks focused on short-term recognition Li et al. (2024); Ning et al. (2026), while more recent works such as VideoMME Fu et al. (2025) and LongVideoBench Wu et al. (2024) reach hour-long durations. Multi-video benchmarks Peng et al. (2025); Zhu et al. (2025) extend the horizon further but typically aggregate disjointed clips that share no continuous timeline. Building on pioneering egocentric datasets Grauman et al. (2022); Huang et al. (2024); He et al. (2025); Pei et al. (2025), EgoLife Yang et al. (2025b) pushes into longitudinal first-person recording but stays within a single shared apartment. MM-Lifelong complements these efforts with 105.6 hours of continuous live broadcasts spanning 51 days, in which physical timeline far exceeds observed playback.

Benchmarks for Long-Context Memory. In text, benchmarks like LongBench Bai et al. (2024) and BABILong Kuratov et al. (2024) use massive contexts for state tracking but lack visual input. Multimodal counterparts either rely on discrete images (e.g., Mem-Gallery Bei et al. (2026)) or focus on short-term streaming responsiveness Yang et al. (2025c), leaving the long-horizon, continuously evolving multimodal setting largely untested. MM-Lifelong fills this gap with "Needle-in-a-Lifestream" and multi-hop tasks that probe whether models can track a single subject's state across weeks of unobserved gaps.

Working and Persistent Memory. Two complementary lines of work address the burden of long-context video. *Within the model*, the field is moving from KV-cache compression and eviction Chen et al. (2025b); Xiao et al. (2024) to architectural changes that decouple memory footprint from sequence length, such as linear and hybrid attention Yang et al. (2025a); Blakeman et al. (2025) and latent or conditional memory DeepSeek-AI et al. (2024); Cheng et al. (2026). *Outside the model*, agentic systems transcend finite context via recursive tool use and external memory orchestration Anil et al. (2023); Google (2025); Chhikara et al. (2025), supported by retrieval mechanisms Faysse et al.

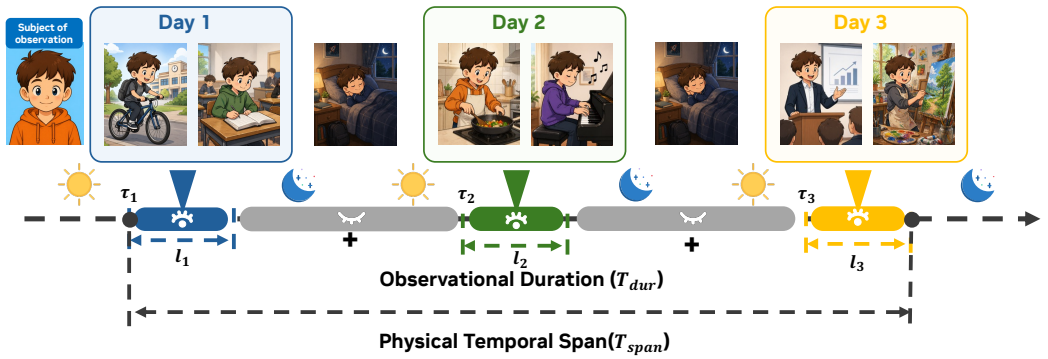


Figure 2: **Problem formalization of Multimodal Lifelong Understanding.** A persistent subject is observed across multiple recording sessions on different days, with sleep cycles and off-camera periods between sessions leaving unobserved gaps. Each recorded clip has a playback length l_i and a real-world start time τ_i. Two temporal quantities characterize the dataset: *Observational Duration* $T_{dur} = \sum_i l_i$ (total playback, summed across clips) and *Physical Temporal Span* $T_{span} = (\tau_N + l_N) - \tau_1$ (chronological horizon from first to last clip). On lifelong timelines spanning days to months, continuous capture is infeasible, so $T_{span} \gg T_{dur}$ arises naturally and forces models to reason across large unobserved gaps.

(2025); Wan & Yu (2025); Rege et al. (2026) and video-specific agents that perform frame selection and persistent visual memory Wang et al. (2026); Chen et al. (2025a); Jin et al. (2025); Chen et al. (2026a); Wang et al. (2025); Chen et al. (2026b); Yu et al. (2026); Huang et al. (2025). Existing benchmarks remain largely text-centric or limited to discrete visual tasks. MM-Lifelong evaluates both directions on the same continuous, 100+ hour stream, characterizing context use within MLLMs and evidence access within agentic systems.

3 Multimodal Lifelong Understanding

Multimodal lifelong understanding extends long-video understanding to streams whose physical timeline far exceeds what the model can actually observe, so that the model must reason across stretches of real time it never saw. This section makes this precise: we define the relevant temporal quantities and the conditions under which a task qualifies as lifelong rather than merely long.

A multimodal dataset is a discrete sampling of a continuous physical timeline, consisting of N video clips. Each clip has a playback length l_i and a real-world starting timestamp τ_i. The pair (l_i, τ_i) separates two temporal quantities that benchmarks typically conflate: how much footage the model sees, and what real-world span that footage covers.

We characterize the temporal scale of a dataset by two metrics (Figure 2):

- **Observational Duration (T_{dur}):** The sum of the playback lengths of all observed clips: $T_{dur} = \sum_{i=1}^{N} l_i$.
- **Physical Temporal Span (T_{span}):** The chronological horizon from the start of the first clip to the end of the last: $T_{span} = (\tau_N + l_N) - \tau_1$.

Long-video benchmarks typically span hours, allowing clips to be packed densely so that $T_{span} \approx T_{dur}$. Lifelong datasets, however, span days, weeks, or months: at this scale continuous capture is no longer feasible, since sleep cycles, off-camera periods, and the cost of round-the-clock recording inevitably create gaps. As a result, $T_{span} \gg T_{dur}$ arises naturally on lifelong streams, with the unobserved gaps representing real-world time that simply cannot be recorded. The model must therefore infer what happened during stretches it never saw, rather than locate adjacent frames it did see.

$T_{span} \gg T_{dur}$ alone is not sufficient: a five-minute clip from a year-long timeline already satisfies $T_{span}/T_{dur} = 10^5$, yet is just short-form recognition with arbitrary timestamps. Three additional constraints on the dataset rule out such constructions:

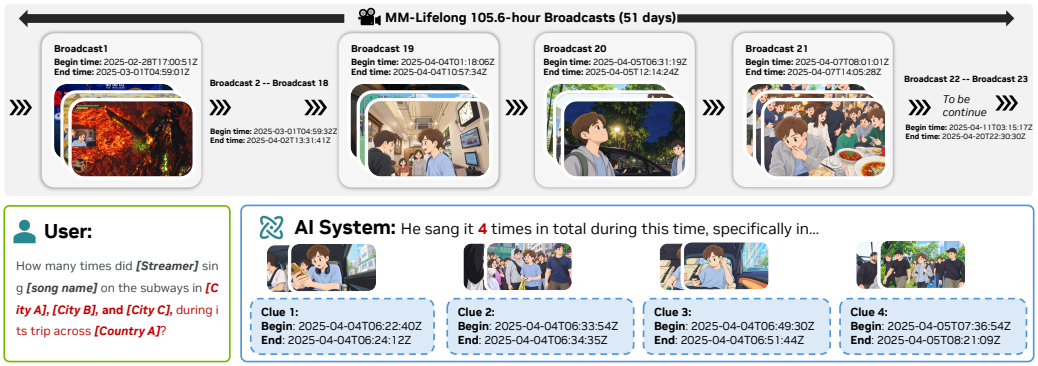


Figure 3: **(a)** The Live Stream subset of **MM-Lifelong** comprises 105.6 hours of broadcast footage spanning 51 days. **(b)** A multi-clue (multi-hop) reasoning question with a long temporal certificate: the task requires identifying all occurrences where the streamer sings a specific song on subways across multiple cities. Answering correctly requires persistent memory and multi-event inference across more than 10 hours of livestream data. People are depicted as fictional characters.

Table 1: **Multi-scale design of MM-Lifelong.** Each scale tracks one subject's evolving state: a gamer, wearer, or streamer. At the month scale, T_{span} exceeds that of released video corpora by 2–3 orders of magnitude.

Scale	Domain	Subject	T_{dur}	T_{span}	Data Source Description
Day	Gamer's Journey	Protagonist	23.6h	∼24h	Complete narrative walkthrough tracking the avatar's inventory and skill progression.
Week	Egocentric Life	Wearer	51.9h	∼7d	Continuous first-person recording of daily routines and household interactions from Ego-Life Yang et al. (2025b).
Month	Live Stream	Streamer	105.6h	∼51d	Unscripted IRL stream tracking the influencer's travel across cities and social events.

1. **Daily-scale observation** ($T_{dur} \geq 12h$)**:** each recording covers a full diurnal routine (roughly a person's active period), not isolated fragments.
2. **Cross-day temporal span** ($T_{span} \geq 24h$)**:** the span includes at least one unobserved interval such as a sleep cycle or off-camera period. A 23-hour recording does not qualify; a 26-hour span with a sleep cycle does.
3. **Persistent subject:** a single identity is tracked continuously across T_{span}; otherwise the task collapses to aggregating across unrelated subjects. Appearance, location, and activity may change, but the agent does not.

A task satisfying all three lies in the *Lifelong Horizon*.

Continuous multimodal footage at this scale is rare in practice. The most ambitious longitudinal efforts (EgoLife ∼7 days, TeleEgo ∼3 days) reach the day scale at most, leaving the month-scale Lifelong Horizon largely unpopulated. Section 4 introduces the recordings and annotation pipeline through which we populate this horizon at three scales (Day, Week, Month).

4 The MM-Lifelong Benchmark

We introduce **MM-Lifelong**, a *Multi-Scale Benchmark* that operationalizes the lifelong setting defined in Section 3. Since continuous lifetime recording is infeasible at scale, MM-Lifelong approximates this setting through three complementary scales (Day, Week, Month) that together cover the temporal properties of the Lifelong Horizon.

4.1 Dataset Construction

We describe the multi-scale design, video collection, and clue-grounded annotation of MM-Lifelong.

Multi-scale design. Lifelong complexity requires content variation as well as duration. Our three recording domains span distinct ratios of Observational Duration (T_{dur}) to Physical Span (T_{span}; Table 1). *Day* and *Week* provide dense observations with minimal interruption, while *Month* captures sparse long-term evolution with large unobserved gaps ($T_{span} \gg T_{dur}$). We exclude year-scale data because of contamination through frontier-model parametric knowledge; Appendix A.4 provides the analysis and examples.

Why depth over breadth. We prioritize temporal depth over multiple subjects per scale for three reasons. (i) Lifelong reasoning requires *intra-subject* state accumulation: tracking appearance, knowledge, and routines across unobserved gaps. Shallow clips from many subjects test a different capability, covered by multi-video benchmarks Peng et al. (2025); Zhu et al. (2025). (ii) Whereas EgoLife Yang et al. (2025b) and TeleEgo Yan et al. (2025) each cover one egocentric domain, our three scales follow different subjects in disjoint domains (gamer, wearer, streamer), preventing a single-persona shortcut across all three. (iii) At fixed annotation density, recruiting and clue-grounding multi-week streams from $N > 1$ subjects requires roughly N times the annotator-months. Existing cross-domain coverage and the benchmark's unsaturated performance favor temporal depth before this costly expansion.

Video Diversity and Collection. The *Gaming* (synthetic) and *Egocentric* (first-person routine) subsets cover specialized scenarios. *Live Stream* adds unscripted open-world content, from *indoor chatting, gaming, and reaction videos* to *outdoor vlogs, sports, chaotic events, and singing performances*. This diversity tests robustness across visual contexts as well as temporal scales. The raw collection totals 211 GB of video.

4.2 Annotation Protocol

To support reproducible evaluation and future extensions, we adopt a *clue-grounded annotation strategy* inspired by CG-Bench Chen et al. (2025c). Unlike traditional QA pairs that provide only the final answer, we explicitly annotate the *clue intervals*, the specific video segments containing the visual evidence required for reasoning. This grounded approach both facilitates automated evaluation (Section 4.5) and leaves room for future interpretability studies that inspect which evidence a model attended to.

Task Definition. Building on the clue-grounded framework, we design two task families that probe complementary aspects of lifelong understanding. **Type I: Needle-in-a-Lifestream** targets single-occurrence details embedded in long streams; models must identify unique, low-frequency events within 100+ hour streams (e.g., *"exact moment the camera dropped"*), testing precise localization and robustness to surrounding noise. **Type II: Multi-Hop Reasoning** requires aggregating information across disjoint intervals separated by hours or days (e.g., *"outfit change between check-in and dinner"*), which requires maintaining a persistent state and performing logical inference rather than single-shot retrieval.

Quality Control. We apply a three-stage pipeline to ensure dataset integrity. **1) Distribution Enforcement:** Instead of random sampling, we curate data to ensure diverse clue durations and validate *temporal certificates* following EgoSchema Mangalam et al. (2023), reserving a subset for *ultra-long dependencies ($> 10h$)* to force cross-session inference. **2) Multi-Stage Verification:** All triplets undergo a two-stage validation: manual expert cross-checks eliminate ambiguity, and a GPT-5-based Singh et al. (2026) filter removes questions answerable by common sense, so that answers depend on the visual evidence rather than prior knowledge. **3) Human-Human Agreement:** Three independent evaluators (not the original annotators) re-answer each question with full video access. Their per-question average is reported as the "Human" row in Table 4 ($> 80\%$ accuracy across all splits), confirming that the annotated answers are reproducible by independent humans.

4.3 Dataset Statistics

As shown in Table 2, MM-Lifelong comprises 181.1 hours of footage across three domains, with 1289 questions grounded by 1810 *clue intervals*, the short evidence segments that directly verify each answer. Each question is also annotated with a *temporal certificate* Mangalam et al. (2023), the broader video region a human must inspect to reach the answer. Roughly $\mathbf{30}\%$ of questions have a temporal certificate exceeding one hour (267 in 1–10h, 127 beyond 10h), forcing multi-hour

Table 2: Statistics of the MM-Lifelong dataset.

Statistics	Number
Total Duration	181.1 hours
Total Questions	1289
* Avg. Question Length	26.79 words
* Avg. Answer Length	4.80 words
Total Clue Intervals	1810 (100%)
* Short (<90s)	1039 (57.40%)
* Medium (90-540s)	550 (30.39%)
* Long (>540s)	221 (12.21%)
* Avg. Clue Duration	362.26s

Statistics	Number
Total Temporal Certificate	1289 (100%)
* Short (<10m)	500 (38.79%)
* Medium (10m-1h)	395 (30.64%)
* Long (1h-10h)	267 (20.71%)
* **Ultra-long (>10h)**	**127 (9.85%)**
Questions by Split	dev/val/test
* Total	266/623/400
* Gamer's Journey (Day)	0/0/200
* Egocentric Life (Week)	0/0/200
* Live Stream (Month)	266/623/0

Table 3: **Multimodal datasets across short-context, long-context, and lifelong settings. Max. Dur (T_{dur})** is playback duration; **Max. Span (T_{span})** is the underlying physical timeline. MM-Lifelong combines the *Lifelong Horizon* ($T_{span} \gg T_{dur}$), with unobserved gaps over a nearly two-month span, and manual, clue-grounded annotation (**Clue**) of continuous audio-visual streams.

Dataset	Modalities	#Samples	Max. Dur	Max. Span	Anno.	QA	Clue
I. Short-Context Multimodal Dataset							
MMMU Yue et al. (2024)	Image	11.5k	0	0	M	11.5k	✗
AIR-Bench Yang et al. (2024)	Audio	19k	19.4s	19.4s	A&M	19k	✗
OmniBench Li et al. (2025)	Audio+Image	1.1k	30s	30s	A&M	1.1k	✗
MVBench Li et al. (2024)	Video	4.0k	2.95m	2.95m	A	4.0k	✗
II. Long-Context Multimodal Dataset							
EgoSchema Mangalam et al. (2023)	Video	5.0k	3.0m	3.0m	A&M	5.0k	✗
Video-MME Fu et al. (2025)	Video	900	59.6m	59.6m	M	2.7k	✗
M3-Bench Long et al. (2026)	Video	1020	57.5m	57.5m	M	4.9k	✗
CG-Bench Chen et al. (2025c)	Video	1219	1.75h	1.75h	M	12.2k	✓
CG-AV-Counting Lu et al. (2026)	Audio+Video	497	1.75h	1.75h	M	1.0k	✓
III. Lifelong Multimodal Dataset							
EgoLife Yang et al. (2025b)	Audio+Video	6	51.9h	∼7d	A&M	3.0k	✗
TeleEgo Yan et al. (2025)	Audio+Video	5	14.4h	∼3d	A&M	3.3k	✗
MM-Lifelong (Ours)	Audio+Video	**3**	**105.6h**	**∼51d**	**M**	**1.3k**	✓

cross-session reasoning even though individual clue intervals remain short. The dataset spans **11 question categories** (Counting 24.2%, Causal Reasoning 14.6%, and Entity Recognition 14.4% being the largest) and **8 video domains** (Chatting, Outdoor Sports, and Singing/Music each $\sim 25\%$); full distributions are in Appendix A.2 (Figure 5).

4.4 Benchmark Splits

We establish a split protocol that separates system development from evaluation under both temporal and domain shifts. Detailed statistics for each split are provided in Table 2.

Domain Generalization. We reserve the Day-scale and Week-scale subsets exclusively for testing. Their unseen subjects and environments serve as held-out tests of out-of-distribution generalization.

Temporal Partitioning. For the Month-scale subset, we avoid a naive random split to prevent temporal data leakage. In lifelong streams, random sampling often places development and validation examples in close temporal proximity, allowing systems to exploit local environmental context. We sort all question-answer-clue triplets chronologically by their clue positions, assigning the first 30% to the development split and the remaining 70% to validation. The development split exposes labels for prompt, retrieval, and system design, while validation measures generalization to later, unseen segments of the same stream. The reported benchmark evaluation remains training-free and does not use the development examples for benchmark-specific model training.

4.5 Evaluation Protocol

We benchmark models on answer quality and clue-level temporal localization.

Building the answer rubrics. A single reference answer cannot cover all valid phrasings or distinguish a partially correct response from a wrong one. We therefore build a separate rubric for every question. Human annotators first specify the facts that a correct answer must contain, valid alternatives, acceptable numerical or temporal ranges, and conditions that make an answer incorrect. GPT-5.6-sol then examines this specification, the annotated video evidence, and example responses from diverse MLLMs and agents. It adds only evidence-supported paraphrases and aliases and defines what scores from 0 to 5 mean for that question. The example responses help reveal common phrasings and errors, but do not determine the correct answer. We freeze all rubrics before scoring the benchmark.

Scoring model answers. For each model, we evaluate each prediction against the corresponding frozen rubric. GPT-5.6-luna serves as the rubric-based verifier, assessing the requested facts, omissions, and contradictions using the question, reference answer, and scoring criteria. Each prediction receives $s \in \{0, 1, 2, 3, 4, 5\}$, with missing predictions assigned zero. For a split with N questions, we report $100 \sum_i s_i/(5N)$. This preserves partial correctness while applying the same question-specific criteria to every model. Appendix B.2 provides the grading template and a worked example.

Reference Grounding (Ref@N). Exact interval overlap is overly sensitive when short clues must be localized within streams lasting many hours. We divide the video timeline into non-overlapping bins of N seconds and mark the bins touched by the predicted and reference intervals. Ref@N is the intersection-over-union of these two bin sets, $\text{Ref@N}(P, G) = |P \cap G|/|P \cup G| \times 100$. We use N=300s by default. Smaller bins require more precise localization, whereas larger bins allow greater temporal tolerance.

4.6 Comparison with Existing Multimodal Datasets

Table 3 positions MM-Lifelong against representative multimodal datasets across short-context, long-context, and lifelong settings. Within the lifelong category, MM-Lifelong differs from the closest prior efforts (EgoLife Yang et al. (2025b), TeleEgo Yan et al. (2025)) along four dimensions: *temporal scope* (the Month subset reaches month-scale T_{span}), *annotation* (manual clue-grounded throughout), *evaluation metric* (Ref@N for sparse-clue grounding), and *evaluation protocol* (a chronological dev/val split at the Month scale, with Day and Week as held-out cross-domain tests). Combined with multi-domain coverage across gamer, wearer, and streamer subjects, these properties make MM-Lifelong the first benchmark to operationalize $T_{span} \gg T_{dur}$ at the month scale with manual, clue-grounded annotations.

5 Evaluation Results

5.1 Evaluation Setup

We evaluate MM-Lifelong in a *training-free* setting, comparing existing systems within two technical routes without benchmark-specific training.

End-to-end MLLMs answer directly from sampled video frames. We include general-purpose models such as *GPT-5.6-luna* (OpenAI, 2026b) and *Qwen3.8-27B* (Qwen Team, 2026b), and video-specialized models such as *Eagle-2.5* (Chen et al., 2025b). Grounding intervals are predicted directly.

Agentic methods access the stream through selection, retrieval, memory, or iterative inspection. Examples include *Codex* (OpenAI, 2026a), *VideoITG* (Wang et al., 2026), and *WorldMM* (Yeo et al., 2026). Grounding uses reported intervals, selected frames, or tool traces.

Table 4 reports answer scores and Ref@300 for all methods, with protocols and compute details in Appendices B and F.

5.2 Results and Analysis

Table 4 summarizes answer scores and temporal grounding across the three evaluation splits.

Longer inputs bring uneven gains for end-to-end MLLMs. Increasing the token budget from 32K to 1M improves *Qwen3.5-9B* by 4.36 points, but *Eagle-2.5-8B* by 0.49 points in Figure 1(b). Eagle peaks at 128K rather than at the largest tested budget. Expanding input capacity therefore does not

Table 4: **Performance comparison on val@month, test@week, and test@day of MM-Lifelong.** Score evaluates answers; Ref@300 evaluates grounding in separate runs. Human reports accuracy in the Score columns. Best and second-best results within each method group are bold and underlined.

		Overall		Val@Month		Test@Week		Test@Day	
System	**Token cost**	**Score**	**Ref@300**	**Score**	**Ref@300**	**Score**	**Ref@300**	**Score**	**Ref@300**
Human (Acc.)	–	88.45	38.43	82.70	33.50	96.10	42.40	98.70	49.80
Multimodal Large Language Models (MLLMs)									
Eagle-2.5-8B (Chen et al., 2025b)	1M	16.58	0.55	16.53	0.03	19.50	1.69	13.80	1.01
Gemma-4-12B-IT (Gemma Team et al., 2026)	256K	17.09	0.43	18.56	0.19	20.70	1.21	8.90	0.39
Gemma-4-31B-IT (Gemma Team et al., 2026)	256K	19.34	2.74	20.35	1.99	22.00	7.03	13.50	0.81
GLM-4.6V-Flash-9B (V Team et al., 2025)	256K	15.95	2.55	14.25	2.24	22.00	5.21	15.20	0.88
GLM-4.6V-106B (V Team et al., 2025)	256K	17.67	3.72	17.66	3.68	22.90	6.51	12.50	1.03
GPT-5.6-luna (OpenAI, 2026b)	1M	**26.53**	**6.95**	**26.29**	4.11	**30.70**	**14.64**	**23.10**	**8.12**
InternVideo3-8B-Instruct (Yan et al., 2026)	128K	16.91	0.20	17.85	0.08	19.40	0.50	11.50	0.25
Ministral-3-14B (Liu et al., 2026a)	256K	16.85	1.80	17.18	0.82	18.20	5.67	14.50	0.97
Nemotron-3-Nano-Omni-30B-A3B (NVIDIA, 2026)	256K	18.08	0.16	18.17	0.00	23.60	0.52	12.30	0.25
Qwen3-Omni-30B-A3B (Xu et al., 2025)	32K	11.40	0.26	12.20	0.25	8.20	0.54	12.10	0.02
Qwen3.5-9B (Qwen Team, 2026a)	1M	21.72	2.58	24.08	2.02	23.70	5.00	12.40	1.90
Qwen3.5-35B-A3B (Qwen Team, 2026a)	1M	19.37	3.27	19.61	2.32	25.30	3.40	12.70	6.13
Qwen3.8-27B (Qwen Team, 2026b)	1M	21.45	5.75	21.73	**4.80**	27.80	9.45	14.20	5.01
VideoChat3-4B (Li et al., 2026)	104K	18.26	0.34	18.68	0.26	24.00	0.73	11.20	0.17
Video-XL-2 (Qin et al., 2025)	128K	11.59	0.70	13.00	0.75	12.20	0.51	6.60	0.72
VideoXL-Pro-3B (Liu et al., 2025)	32K	8.54	0.12	10.79	0.00	5.90	0.63	4.20	0.00
Agents									
Codex (GPT 5.6-luna) (OpenAI, 2026a)	8M	**23.97**	**13.84**	**22.38**	**12.65**	**25.10**	**18.18**	**27.80**	**13.17**
LongVideoAgent (Liu et al., 2026b)	8M	10.09	3.46	9.60	3.31	14.90	4.50	6.80	2.86
ParaVT (Yang et al., 2026)	8M	12.77	0.16	14.09	0.00	12.30	0.82	9.10	0.00
SAGE (Jain et al., 2026)	8M	11.09	4.03	11.20	3.80	9.30	3.22	12.50	5.53
TaskMem (Zou et al., 2026)	8M	7.92	6.51	7.93	5.39	6.10	5.68	9.70	10.85
VideoITG (Wang et al., 2026)	8M	19.75	1.95	20.22	1.17	21.50	4.05	16.50	2.28
VideoZoomer (Ding et al., 2026)	8M	13.27	0.37	13.26	0.04	16.00	1.16	10.60	0.63
WorldMM (Yeo et al., 2026)	8M	13.80	5.52	13.58	5.54	16.20	5.05	12.10	5.92

ensure effective evidence use. Most MLLMs also score lower on Day than Week despite shorter footage, suggesting that content and evidence demands matter alongside duration.

Strong grounding does not imply strong answering for agents. *TaskMem* ranks second on Ref@300 at 6.51 but last on answer score at 7.92. *VideoITG* ranks second in answering at 19.75 despite a Ref@300 of 1.95. This reversal separates temporal localization from recovering the facts needed for an answer. Locating relevant intervals is insufficient if their details are lost or misused, while answer correctness alone does not establish accurate localization. The results motivate agents that maintain explicit links between answer claims and supporting intervals.

5.3 Error Analysis

We analyze the errors of representative MLLMs and agents to identify what limits lifelong understanding. Figure 4 highlights recurring failures in evidence retention, event identification, and verification.

Loss of answer-relevant evidence. A scene-level summary can remain accurate while losing the details needed to reconstruct what happened. Missing evidence is the largest category, accounting for 46.8–50.5% of MLLM errors, 75.1% of VideoITG errors, and 75.0% of WorldMM errors. Inspected sampling records show brief state changes disappearing while surrounding scenes remain available. This challenges scene similarity as a compression criterion: visually redundant moments can contain distinct facts. Such gaps can be hard to detect from a summary that remains internally coherent. Future memory systems should retain state changes and source links for query-time recovery. They should also track observation gaps, so that a missing record is not mistaken for evidence that an event never occurred.

Confusion between event instances. Semantic relevance does not establish that a retrieved event is the one being queried. Among MLLM errors, instance confusion accounts for 15.2–17.7%, versus 6.4–6.9% for relational reasoning. Inspected cases show locally accurate observations supplying the wrong premise when bound to another occurrence or role. Describing that occurrence in greater detail can strengthen a misleading match without establishing its identity. Further reasoning may then elaborate the error rather than correct it. This motivates separating candidate retrieval from instance

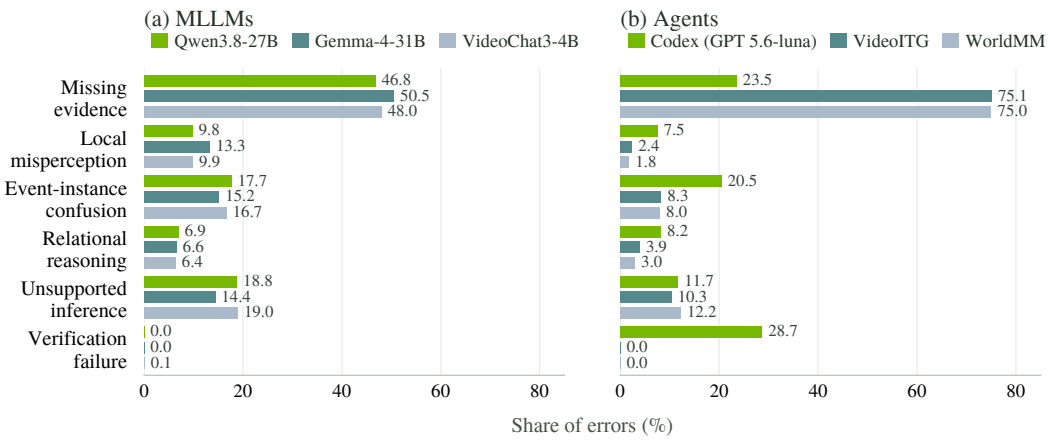


Figure 4: Error composition for three MLLMs on the left and three agents on the right. Percentages are computed within each method's predictions scoring below 4, with one primary label per prediction.

verification: preserve event identities, temporal anchors, and provenance, and verify these bindings before combining events. These bindings need to survive summarization as well as retrieval.

Failure to resolve conflicting evidence. Verification can fail even after uncertainty has been explicitly recognized. Verification failure accounts for 28.7% of Codex's errors, and unsupported inference for 11.7%. Inspected traces show conflicting reports from identical frames followed by a definite answer. The final choice leaves the contradiction unresolved. Because these reports share the same observations, their repetition cannot be counted as independent support. Re-querying the same frames may refine a reading, but cannot supply an observation missing from those frames. Controllers should retain competing explanations, inspect observations that distinguish them, and base stopping on evidential support rather than the completion of another tool call. Unresolved conflicts must remain attached to the claim through final answer generation.

5.4 Verifier–Human Alignment

We compare automatic and human grading on 400 responses, with 100 from each of two MLLMs and two agents. Following LongShOTBench (Kurpath et al., 2026), three volunteers grade each response using the frozen question-specific rubrics.

Table 5: Agreement with human grading.

Metric	Eval.–human	Human–human
Weighted $\kappa \uparrow$	0.63	0.67
MAE $\downarrow$	0.69	0.58
Pearson $r \uparrow$	0.75	0.81

Quadratic-weighted κ; MAE in 0–5 points.
Eval.–human: adjudicated consensus.
Human–human: mean over three rater pairs.

We find good agreement between the verifier and human consensus, with a weighted κ of 0.63 and Pearson r of 0.75 in Table 5. On the 0–5 scale, the mean absolute difference from human consensus is 0.69 points, while the corresponding difference between human annotators is 0.58 points. This agreement supports using the verifier for rubric-based evaluation of free-form answers in MM-Lifelong.

6 Conclusion

We introduce **MM-Lifelong**, a benchmark for multimodal understanding over subject-specific histories at Day, Week, and Month scales. It distinguishes observational duration from physical temporal span and connects questions to manually annotated evidence. Our evaluation reveals uneven gains from larger input budgets and differing system strengths in answering and temporal grounding. Error analysis highlights the need to retain event-specific details, maintain correct event identities, and resolve conflicting observations. These findings motivate memory systems that preserve explicit links between answer claims and supporting evidence. The current benchmark covers one primary subject per scale and a maximum span of 51 days. Future work should broaden subject coverage, extend temporal horizons, and explicitly evaluate reasoning about unobserved intervals. We release the benchmark, scoring rubrics, and baselines to support these efforts.

AI USE STATEMENT

Generative AI tools assisted the screening of MM-Lifelong questions for language-only answerability and automated answer grading. They also assisted manuscript drafting and revision. The authors reviewed the AI-assisted work and take full responsibility for the benchmark, evaluation procedures, analyses, manuscript claims, and released artifacts.

REFERENCES

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 40750–40784, 2026. doi: 10.18653/v1/2026.acl-long.1892. URL https://aclanthology.org/2026.acl-long.1892/.

Aaron Blakeman, Aarti Basant, Abhinav Khattar, Adithya Renduchintala, Akhiad Bercovich, Aleksander Ficek, Alexis Bjorlin, Ali Taghibakhshi, Amala Sanjay Deshmukh, Ameya Sunil Mahabaleshwarkar, et al. Nemotron-h: A family of accurate and efficient hybrid mamba-transformer models. *arXiv preprint arXiv:2504.03624*, 2025.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20237–20246, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Chen_LVAgent_Long_Video_Understanding_by_Multi-Round_Dynamical_Collaboration_of_MLLM_ICCV_2025_paper.html.

Chunliang Chen, Ming Guan, Xiao Lin, Jiaxu Li, Luxi Lin, Qiyi Wang, Xiangyu Chen, Jixiang Luo, Changzhi Sun, Dell Zhang, and Xuelong Li. Telemem: Building long-term and multimodal memory for agentic ai, 2026a. URL https://arxiv.org/abs/2601.06037.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Max Ehrlich, et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/833295cb9278a3ba973842a94ea68e3c-Abstract-Conference.html.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. CG-Bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/70fa5df8e3300dc30bf19bee44a56155-Abstract-Conference.html.

Guoxin Chen, Zile Qiao, Xuanzhong Chen, Donglei Yu, Haotian Xu, Wayne Xin Zhao, Ruihua Song, Wenbiao Yin, Huifeng Yin, Liwen Zhang, Kuan Li, Minpeng Liao, Yong Jiang, Pengjun Xie, Fei

Huang, and Jingren Zhou. Iterresearch: Rethinking long-horizon agents with interaction scaling, 2026b. URL https://arxiv.org/abs/2511.07327.

Xin Cheng, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Zhewen Hao, Han Zhang, Yu-Kun Li, Huishuai Zhang, Dongyan Zhao, and Wenfeng Liang. Conditional memory via scalable lookup: A new axis of sparsity for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 4968–4990, 2026. doi: 10.18653/v1/2026.acl-long.226. URL https://aclanthology.org/2026.acl-long.226/.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025*, pp. 2993–3000, 2025. doi: 10.3233/FAIA251160.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. VideoZoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/221b2665260a582f7eb821b9871e8001-Abstract-Conference.html.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/99e9e141aafc314f76b0ca3dd66898b3-Abstract-Conference.html.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Gemma Team et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Google. Gemini deep research demo | using ai to learn new topics in depth. YouTube video, 2025. URL https://www.youtube.com/watch?v=buwMJxvW7wI.

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

Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. EgoExoBench: A benchmark for first- and third-person view video understanding in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a974bceb4d1af7d46673d1876482bd4f-Abstract-Datasets_and_Benchmarks_Track.html.

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the ECCV (ECCV)*, pp. 754–769, 2018.

Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *CVPR*, pp. 14024–14034, 2020.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Dong Lu, Yali Wang, Limin Wang, and Yu Qiao. Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In *CVPR*, 2024.

Yifei Huang, Jilan Xu, Baoqi Pei, Lijin Yang, Mingfang Zhang, Yuping He, Guo Chen, Xinyuan Chen, Yaohui Wang, Zheng Nie, et al. Vinci: A real-time smart assistant based on egocentric vision-language model for portable devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–33, 2025. doi: 10.1145/3749513.

Jitesh Jain, Jialuo Li, Zixian Ma, Jieyu Zhang, Chris Dongjoo Kim, Sangho Lee, Rohun Tripathi, Tanmay Gupta, Christopher Clark, and Humphrey Shi. SAGE: Training smart any-horizon agents for long video reasoning with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41478–41488, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jain_SAGE_Training_Smart_Any-Horizon_Agents_for_Long_Video_Reasoning_with_CVPR_2026_paper.html.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. Videomem: Enhancing ultra-long video understanding via adaptive memory management. *arXiv preprint arXiv:2512.04540*, 2025.

Yury Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

Mohammed Irfan Kurpath, Jaseel Muhammad Kaithakkodan, Jinxing Zhou, Sahal Shaji Mullappilly, Mohammad Almansoori, Noor Ahsan, Beknur Kalmakhanbet, Sambal Shikhar, Rishabh Lalla, Jean Lahoud, Mariette Awad, Fahad Shahbaz Khan, Salman Khan, Rao Muhammad Anwer, and Hisham Cholakkal. A benchmark for omni-modal reasoning in long videos. *arXiv preprint arXiv:2512.16978v2*, 2026. URL https://arxiv.org/abs/2512.16978v2.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pp. 22195–22206, 2024.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, et al. VideoChat3: Fully open video MLLM for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026. URL https://arxiv.org/abs/2607.14935.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, et al. OmniBench: Towards the future of universal omni-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2c7c4a12a9dcf7dace9896d08154c705-Abstract-Datasets_and_Benchmarks_Track.html.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026a. URL https://arxiv.org/abs/2601.08584.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. LongVideoAgent: Multi-agent reasoning with long videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 40404–40416, 2026b. doi: 10.18653/v1/2026.acl-long.1876. URL https://aclanthology.org/2026.acl-long.1876/.

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-XL-Pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*, 2025. URL https://arxiv.org/abs/2503.18478.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PMz29A7Muq.

Lidong Lu, Guo Chen, Zhu Wei, Zhiqi Li, Yicheng Liu, and Tong Lu. AV-Reasoner: Improving and benchmarking clue-grounded audio-visual counting for MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33477–33487, 2026. URL https://openaccess.thecvf.com/CVPR2026?day=2026-06-07.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-Bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *Computational Visual Media*, 12(1):71–84, 2026. doi: 10.26599/CVM.2025.9450516. URL https://www.sciopen.com/article/10.26599/CVM.2025.9450516.

NVIDIA. Nemotron 3 Nano Omni: Efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026. URL https://arxiv.org/abs/2604.24954.

OpenAI. Codex: Open-source coding agent, 2026a. URL https://github.com/openai/codex. Accessed September 2026.

OpenAI. GPT-5.6 Luna: Model documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna.

Baoqi Pei, Yifei Huang, Jilan Xu, Yuping He, Guo Chen, Fei Wu, Jiangmiao Pang, and Yu Qiao. EgoThinker: Unveiling egocentric reasoning with spatio-temporal CoT. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3e9412a9c1d93810ef3ef7825115016b-Abstract-Conference.html.

Tianhao Peng, Haochen Wang, Yuanxing Zhang, Noah Wang, et al. Mvu-eval: Towards multi-video understanding evaluation for multimodal llms. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4dbb159f2c68b359b3b6ed7dd039481c-Abstract-Datasets_and_Benchmarks_Track.html.

Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Video-xl-2: Towards very long-video understanding through task-aware kv sparsification. *arXiv preprint arXiv:2506.19225*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026a. URL https://qwen.ai/blog?id=qwen3.5.

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

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 46575–46602, 2026. doi: 10.18653/v1/2026.acl-long.2161. URL https://aclanthology.org/2026.acl-long.2161/.

Aaditya Singh et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2026. URL https://arxiv.org/abs/2601.03267.

V Team, Wenyi Hong, Wenmeng Yu, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. URL https://arxiv.org/abs/2507.01006.

Xueyao Wan and Hang Yu. Mmgraphrag: Bridging vision and language with interpretable multimodal knowledge graphs. *arXiv preprint arXiv:2507.20804*, 2025.

Piaohong Wang, Motong Tian, Jiaxian Li, Yuan Liang, Yuqing Wang, Qianben Chen, Tiannan Wang, Zhicong Lu, Jiawei Ma, Yuchen Eleanor Jiang, and Wangchunshu Zhou. O-mem: Omni memory system for personalized, long horizon, self-evolving agents, 2025. URL https://arxiv.org/abs/2511.13593.

Shihao Wang, Guo Chen, De-An Huang, Zhiqi Li, Minghan Li, Guilin Liu, Jan Kautz, Jose M. Alvarez, Lei Zhang, and Zhiding Yu. VideoITG: Multimodal video understanding with instructed temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24640–24650, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_VideoITG_Multimodal_Video_Understanding_with_Instructed_Temporal_Grounding_CVPR_2026_paper.html.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5e5fd18f863cbe6d8ae392a93fd271c9-Abstract-Conference.html.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. URL https://arxiv.org/abs/2509.17765.

Jiaqi Yan, Ruilong Ren, Jingren Liu, Shuning Xu, Ling Wang, Yiheng Wang, Xinlin Zhong, Yun Wang, Long Zhang, Xiangyu Chen, et al. Teleego: Benchmarking egocentric ai assistants in the wild. *arXiv preprint arXiv:2510.23981*, 2025.

Ziang Yan, Sheng Xia, Jiashuo Yu, et al. InternVideo3: Agentify foundation models with multimodal contextual reasoning. *arXiv preprint arXiv:2606.12195*, 2026. URL https://arxiv.org/abs/2606.12195.

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

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. EgoLife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28885–28900, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_EgoLife_Towards_Egocentric_Life_Assistant_CVPR_2025_paper.html.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1979–1998, 2024. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. SVBench: A benchmark with temporal multi-turn dialogues for streaming video understanding. In *International Conference on Learning Representations*, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/693ce820ea7e0f0c70a2a833cbdf7ccc-Abstract-Conference.html.

Zuhao Yang, Kaichen Zhang, Sudong Wang, Keming Wu, Zhongyu Yang, Bo Li, Xiaojuan Qi, Shijian Lu, Xingxuan Li, and Lidong Bing. ParaVT: Taming the tool prior paradox for parallel tool use in agentic video reinforcement learning. *arXiv preprint arXiv:2605.20342*, 2026. URL https://arxiv.org/abs/2605.20342.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yeo_WorldMM_Dynamic_Multimodal_Memory_Agent_for_Long_Video_Reasoning_CVPR_2026_paper.html.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic memory: Learning unified long-term and short-term memory management for large language model agents, 2026. URL https://arxiv.org/abs/2601.01885.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8190b210e9808e54ee16263b673a847d-Abstract-Conference.html.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, et al. Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning. *arXiv preprint arXiv:2508.19542*, 2025.

Tao Zou, Yichen He, Tian Qiu, Yuan Lin, and Hang Li. Task-focused memorization for multimodal agents. *arXiv preprint arXiv:2605.31075*, 2026. URL https://arxiv.org/abs/2605.31075.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. Videolucy: Deep memory backtracking for long video understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Supplementary Material

Contents

Page

A Benchmark Construction . . . 17
A.1 Video . . . 17
A.2 Annotation . . . 19
A.3 Why Sub-sampling Cannot Substitute for Continuous Capture . . . 21
A.4 Data Contamination . . . 21
A.4.1 Web Search Impact . . . 21
A.4.2 Year-scale Data Exploration . . . 21

B Experiments . . . 22
B.1 Evaluation Prompts . . . 22
B.2 Detailed Evaluation Protocols . . . 22
B.3 Effect of Temporal Bin Width . . . 25

C Qualitative Examples . . . 26

D Limitations . . . 41

E Ethics Statement . . . 41

F Compute Resources . . . 42

A BENCHMARK CONSTRUCTION

A.1 VIDEO

Table 6 presents the metadata for each video clip, including the start time, end time, and clip duration. For the Live Stream Subset, we report absolute UTC timestamps, as the original livestreams provide access to their true broadcast times. For the Gamer's Journey Subset, clips correspond to consecutive gameplay segments spanning multiple in-game chapters and are concatenated end-to-end. We therefore treat the entire sequence as continuous gameplay within a single day and report relative timestamps accordingly. For the Egocentric Life Subset, where the exact dates are unavailable, we report timestamps using relative day index combined with absolute time-of-day.

Table 6: Temporal metadata of video clips across subsets. Each clip is annotated with its begin and end time, using relative timestamps or absolute UTC timestamps.

Table 7: Video Timeline Accumulation of Gamer's Journey Subset.

#	Begin	End	Dur.(s)
1	Day 1 00:00:00	Day 1 00:13:41	821
2	Day 1 00:13:41	Day 1 00:54:14	2,433
3	Day 1 00:54:14	Day 1 01:46:40	3,146
4	Day 1 01:46:40	Day 1 02:16:39	1,799
5	Day 1 02:16:39	Day 1 03:15:07	3,508
6	Day 1 03:15:07	Day 1 04:10:32	3,325
7	Day 1 04:10:32	Day 1 04:54:35	2,643
8	Day 1 04:54:35	Day 1 06:02:21	4,066
9	Day 1 06:02:21	Day 1 06:41:32	2,351
10	Day 1 06:41:32	Day 1 07:02:28	1,256
11	Day 1 07:02:28	Day 1 07:46:41	2,653
12	Day 1 07:46:41	Day 1 08:37:39	3,058
13	Day 1 08:37:39	Day 1 09:29:30	3,111
14	Day 1 09:29:30	Day 1 10:20:29	3,059
15	Day 1 10:20:29	Day 1 11:14:59	3,270
16	Day 1 11:14:59	Day 1 11:51:49	2,210
17	Day 1 11:51:49	Day 1 12:22:52	1,863
18	Day 1 12:22:52	Day 1 13:08:51	2,759
19	Day 1 13:08:51	Day 1 13:58:56	3,005
20	Day 1 13:58:56	Day 1 14:33:17	2,061
21	Day 1 14:33:17	Day 1 15:33:47	3,630
22	Day 1 15:33:47	Day 1 16:20:36	2,809
23	Day 1 16:20:36	Day 1 17:32:13	4,297
24	Day 1 17:32:13	Day 1 18:25:02	3,169
25	Day 1 18:25:02	Day 1 19:13:35	2,913
26	Day 1 19:13:35	Day 1 20:08:42	3,307
27	Day 1 20:08:42	Day 1 20:42:44	2,042
28	Day 1 20:42:44	Day 1 21:21:19	2,315
29	Day 1 21:21:19	Day 1 22:08:13	2,814
30	Day 1 22:08:13	Day 1 22:47:38	2,365
31	Day 1 22:47:38	Day 1 23:35:14	2,856

Table 8: Video Timeline Accumulation of Live Stream Subset.

#	Begin time	End time	Dur.(s)
1	02-28T17:00:51Z	03-01T04:59:01Z	42,900
2	03-01T04:59:32Z	03-01T06:52:59Z	6,813
3	03-01T17:00:16Z	03-01T21:10:17Z	14,740
4	03-02T18:00:12Z	03-02T22:00:22Z	14,421
5	03-05T22:16:34Z	03-05T23:13:35Z	3,430
6	03-06T17:31:51Z	03-06T21:00:34Z	12,535
7	03-07T14:19:28Z	03-07T16:11:38Z	6,740
8	03-08T13:29:35Z	03-08T15:02:49Z	5,604
9	03-08T17:32:44Z	03-08T18:12:22Z	2,390
10	03-12T17:15:50Z	03-12T20:06:57Z	10,259
11	03-14T17:30:22Z	03-14T20:46:29Z	11,777
12	03-16T16:00:20Z	03-16T20:14:46Z	15,203
13	03-19T16:00:45Z	03-19T18:55:28Z	10,466
14	03-24T06:00:05Z	03-24T12:19:50Z	22,799
15	03-26T01:29:56Z	03-26T07:39:35Z	22,191
16	03-28T05:45:53Z	03-28T10:35:38Z	17,396
17	03-31T06:31:41Z	03-31T11:54:51Z	19,402
18	04-02T07:00:25Z	04-02T13:31:41Z	23,485
19	04-04T01:18:06Z	04-04T10:57:34Z	34,777
20	04-05T06:31:19Z	04-05T12:14:24Z	20,594
21	04-07T08:01:01Z	04-07T14:05:28Z	21,702
22	04-11T03:15:17Z	04-11T08:25:16Z	18,173
23	04-20T16:00:13Z	04-20T22:30:30Z	22,504

Table 9: Video Timeline Accumulation of Egocentric Life Subset.

#	Begin datetime	End datetime	Dur.(s)
1	Day 1 11:09:42.08	Day 1 22:05:49.11	39,367
2	Day 2 10:44:25.06	Day 2 22:58:25.00	44,040
3	Day 3 11:17:27.02	Day 3 22:51:33.07	41,646
4	Day 4 10:48:20.00	Day 4 22:24:34.18	41,774
5	Day 5 11:00:31.00	Day 5 23:29:46.08	44,955
6	Day 6 09:49:33.00	Day 6 22:16:59.01	44,846
7	Day 7 11:56:08.17	Day 7 20:16:15.08	30,007

A.2 ANNOTATION

Definition and Examples. The definitions and number of QA types are listed in Table 10, and examples are given in Table 11.

Table 10: Statistics and definitions of question categories. The categories are sorted by the total number of samples.

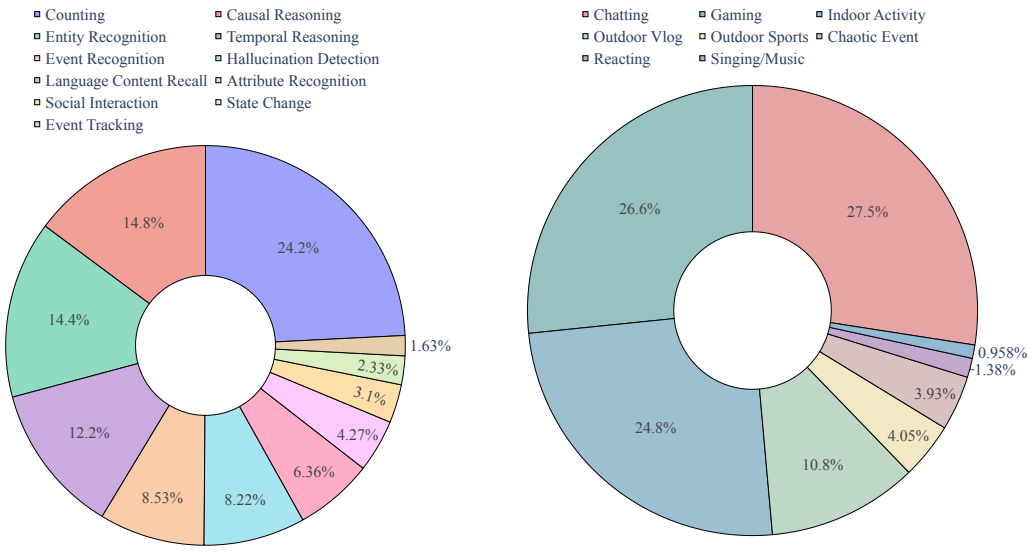
Category	Definition	Day	Week	Month	Total
Counting	For the object(s) or event(s) mentioned, ask how many times they appear or repeat.	40 (20.00%)	59 (29.50%)	213 (23.96%)	312 (24.20%)
Causal Reasoning	For an event mentioned, ask about its cause or the result it leads to.	20 (10.00%)	17 (8.50%)	151 (16.99%)	188 (14.58%)
Entity Recognition	Identify a specific entity (object, person, or place) referenced in the question.	31 (15.50%)	23 (11.50%)	132 (14.85%)	186 (14.43%)
Temporal Reasoning	Ask about temporal order, chronological sequencing, or duration of events.	31 (15.50%)	36 (18.00%)	85 (9.56%)	152 (11.79%)
Hallucination Det.	Given a set of statements, ask which statements are correct or which are wrong.	31 (15.50%)	10 (5.00%)	71 (7.99%)	112 (8.69%)
Event Recognition	Perform a recognition or identification of an event mentioned in the question.	11 (5.50%)	17 (8.50%)	80 (9.00%)	108 (8.38%)
Lang. Content Recall	Ask about specific linguistic content (speech/ASR or on-screen text/OCR).	6 (3.00%)	15 (7.50%)	63 (7.09%)	84 (6.52%)
Attribute Recognition	Ask about the attribute (e.g., color, appearance, size) of an object or person.	6 (3.00%)	8 (4.00%)	40 (4.50%)	54 (4.19%)
Social Interaction	Ask about relationships, social roles, or the nature of interactions between people.	0 (0.00%)	9 (4.50%)	31 (3.49%)	40 (3.10%)
State Change	For an object or scene mentioned, ask about how its state changes.	12 (6.00%)	0 (0.00%)	18 (2.02%)	30 (2.33%)
Event Tracking	Ask when an event happened or when an object was acquired (retrospective tracing).	12 (6.00%)	6 (3.00%)	5 (0.56%)	23 (1.78%)

Annotation Cost. Eight annotators were recruited to label the three subsets. Due to variations in temporal length, annotation costs differed across the data: the average time per sample was 19 minutes for the day-scale (Gamer's Journey), 33 minutes for the week-scale (Egocentric Life), and 51 minutes for the month-scale (Live Stream) subsets.

Distribution Visualizations. Figure 5 visualizes the distribution of question categories and video clip domains, complementing the textual summary in Section 4.3.

Table 11: QA example of each category.

Question sub-category	QA pair example
Counting	Q: In the first live stream after returning from both the China trip and the Mongolia trip, how many times did [The Streamer] hear the song 'Sunshine Rainbow Little White Horse' while browsing videos on Discord? A: 8 times
Event Recognition	Q: In the game FRAGPUNK played by [The Streamer], in the sixth round of the new match after the final score was 4:2, what event occurred that shocked [The Streamer]? A: [The Streamer] and [The Streamer]'s teammates were all killed by an invisible enemy.
Language Content Recall	In Day 5, when the first-person protagonist arrives at the supermarket, what is written on the front of their tablet's case? A: Scholar.
State Change	Q: What change occurred in the clothing of the character [The Streamer] was watching in the second video before hosting the talent show by [The Streamer]? A: Red-Black
Causal Reasoning	Q: During [The Streamer]'s visit to Hong Kong, why did he still feel shocked after talking to a little boy upon leaving the gaming area and arriving at the parking lot? A: The 12-year-old boy had an exceptionally mature voice.
Event Tracking	Q: During [The Streamer]'s trip to Mongolia, [The Streamer] was kissed on the cheek by a male fan in the car leaving the museum. When was the last time he was kissed by a male fan? A: While walking on the street in Chengdu.
Temporal Reasoning	Q: In the event involving [The Streamer] and others linking microphones, what is the correct sequence? 1. A lady who changed many pairs of high heels and played football with them; 2. Three children, one of whom looks especially like Messi; 3. A chubby guy who challenged himself to eat a pizza within one minute. A: 3, 2, 1
Social Interaction	Q: In [The Streamer]'s first live stream after returning from his trips to China and Mongolia, who appeared most frequently, even throughout the entire video, when he watched Coco's video about his China trip on Discord? A: The man in the red floral shirt.
Hallucination Detection	Q: During [The Streamer]'s visit to China, which of the following statements are correct? 1. In a park in Chongqing, he played tennis for a while. 2. In a basketball court in Shanghai, he played basketball with Jackson Wang for a while. 3. In Yu Garden, Shanghai, he played a translation game with a Chinese guy, translating from English to Chinese. 4. At the end of his Shanghai visit, he greeted a guy with an injured ankle. A: 1, 4
Attribute Recognition	Q: After the player enters the Flaming Mountains chapter, what are the player's health and mana values, respectively, before the first challenge against Yinhu? A: 660, 360
Entity Recognition	Q: On Day 1, what was the dessert made after dinner? A: Strawberry Cream Cupcake



(a) Distribution of question categories.

(b) Distribution of video clip domains.

Figure 5: **Distribution of question categories (left) and video clip domains (right) in MM-Lifelong.**

A.3 Why Sub-sampling Cannot Substitute for Continuous Capture

A natural question is whether the regime $T_{span} \gg T_{dur}$ at the day, week, or month scale can be obtained simply by sub-sampling existing long videos. It cannot. Sub-sampling can only enlarge the span up to the source's own physical duration. The longest commonly-released videos (movies, broadcasts, lectures) span only a few hours of physical time, so any sub-sampling protocol over them yields T_{span} at most a few hours. Operationalizing $T_{span} \gg T_{dur}$ at the day, week, or month scale therefore requires *continuous* physical-day-, week-, or month-scale capture of a single subject. Such data is not produced by post-hoc sub-sampling; it must be collected longitudinally.

A.4 Data Contamination

We investigated the risks of data contamination arising from the integration of search engine tools. Specifically, we analyzed year-scale data to assess the current state of contamination in large-scale evaluation.

A.4.1 Web Search Impact

We evaluated the impact of web search capabilities on a livestream subset by comparing the latest model from the Gemini family of multimodal models Anil et al. (2023) in both offline and web-enabled configurations.

Prompting Strategy To facilitate efficient testing, we batched multiple queries into a single request, instructing the system to process them via a deep research workflow and return responses in a structured JSON format. Below is an example prompt containing 100 questions:

```
[
   { "index": 1, "question": "{question 1}" },
   ...
   { "index": 100, "question": "{question 100}" }
]

Task: Thoroughly research and analyze these questions. Provide
    ↪ individual answers in the following JSON format:
[
   { "index": 1, "answer": "{answer 1}" },
   ...
   { "index": 100, "answer": "{answer 100}" }
]
```

Results and Analysis Without web access, this Gemini model achieved an accuracy score of 5.54. Enabling web search raised the score to 11.79, more than doubling accuracy through external information retrieval.

This gain also reflects the risk of dataset contamination. Since livestream content is often documented or discussed online, it becomes difficult to distinguish long-context reasoning from indirect exposure to ground-truth data. Evaluation of long-video understanding should therefore control external knowledge access and clearly separate *closed-book* from *open-book* settings.

A.4.2 Year-scale Data Exploration

To explore year-scale data curation, we curated a dataset spanning the 19-year career of a professional athlete. This dataset comprises 140 match videos from the same competition, totaling approximately 582 GB and 236.42 hours of footage.

We attempted to annotate 10 QA pairs, each requiring specific temporal verification. However, the annotation cost was nearly prohibitive: verifying a single QA pair took substantially longer than at the month scale. Even with semi-automatic annotation tools, human annotators were still required to perform exhaustive manual verification of clue intervals, yielding minimal efficiency gains.

External metadata (news reports, match statistics) sped up annotation by letting annotators focus on specific segments, but it introduced contamination: frontier models such as GPT-5 and Gemini had already internalized this well-known information during pre-training and could answer a subset of questions from text-only knowledge. When equipped with web search, these models answered nearly all year-scale questions correctly, leaving little room to evaluate pure video-based reasoning.

As a result, we excluded the year-scale from the benchmark. These findings suggest that as external information becomes increasingly accessible, the evaluation of long-video understanding risks collapsing into a test of textual retrieval rather than visual reasoning. Future benchmarks should prioritize "non-Googleable" visual tasks to keep evaluation grounded in temporal-visual reasoning rather than parametric recall.

B Experiments

B.1 Evaluation Prompts

End-to-End MLLMs.

```
# QA Prompts for End-to-End MLLMs.
Answer the following question based on the video with a concise answer.\
    ↪ nQuestion: '{HERE IS THE QUESTION}'

# Grounding Prompts for Video Agents and Video-LLMs.
Find time intervals in the video when the query occurs. Query: '{HERE
    ↪ IS THE QUESTION}' Provide all possible intervals in seconds.
    ↪ Format for each interval: 'xx.xx seconds - xx.xx seconds'.
    ↪ Multiple intervals are linked by' and '. Output the intervals
    ↪ only, do not output anything else.
```

B.2 Detailed Evaluation Protocols

Ref@N evaluation code. We provide Python code to provide a better understanding of the evaluation of reference grounding.

```
from typing import List, Tuple, Set

Interval = Tuple[float, float]

def Ref_N(
   intervals_a: List[Interval],
   intervals_b: List[Interval],
   total_seconds: float,
   bucket_size: float = 300.0,
) -> float:
   def intervals_to_buckets(intervals: List[Interval]) -> Set[int]:
      buckets: Set[int] = set()
      for s, e in intervals:
         # clamp
         s = max(0.0, s)
         e = min(total_seconds, e)
         if s >= e:
            continue

         start = int(s // bucket_size)
         end = int((e - 1e-9) // bucket_size)
         buckets.update(range(start, end + 1))
      return buckets

   buckets_a = intervals_to_buckets(intervals_a)
   buckets_b = intervals_to_buckets(intervals_b)

   if not buckets_a and not buckets_b:
```

```
        return 0.0

    return len(buckets_a & buckets_b) / len(buckets_a | buckets_b)
```

Rubric-based grading. We evaluate each model prediction against the frozen rubric for its question, using GPT-5.6-luna as the rubric-based verifier. The grading inputs are the question, reference answer, rubric, and prediction. The rubric specifies the required facts, accepted expressions and aliases, incorrect claims, and conditions for assigning each score. The verifier assesses the prediction against these criteria and provides an integer score from 0 to 5 with a brief justification. Video evidence is used to establish the rubric during construction; grading uses the resulting textual criteria.

We aggregate the per-question scores as

$$100 \times \frac{\sum_i s_i}{5N},$$

where N is the full split size. Missing or unusable answers receive zero. For auxiliary strict-accuracy analyses, only scores of 4 or 5 count as correct.

Verifier prompt. For readability, we present the grading instructions in single-prediction form below. The reference answer states the target response, while the rubric defines acceptable alternatives and the conditions for full or partial credit. The prediction is the model response being assessed. The output contains its score and a short explanation tied to the rubric. The worked example on the next page illustrates the distinction between an omitted detail and an incorrect claim.

System prompt

```
Evaluate one model prediction using the supplied question, reference answer,
and frozen question–specific rubric.

Check which required facts the prediction provides, which it omits, and which
it contradicts. Interpret equivalent expressions using the accepted phrasings
and aliases in the rubric.

Apply the rubric's scoring conditions, including any requirements for exact
values, temporal tolerances, set membership, or item order. Distinguish omissions
from incorrect claims as specified in the rubric. Do not revise the rubric,
introduce additional requirements, or use other models' answers as evidence.

Assign one integer score from 0 to 5 according to the question–specific rules.
Briefly explain which rubric conditions support that score.

Return exactly one JSON object with two fields:
"score": an integer from 0 through 5;
"reason": a brief rubric–based justification.
Do not include markdown or any text outside the JSON object.
```

User-message template

```
QUESTION:
{question}
REFERENCE ANSWER:
{reference_answer}
FROZEN RUBRIC:
{frozen_rubric_json}
MODEL PREDICTION:
{model_prediction}

Evaluate this prediction according to the supplied rubric.
```

The four placeholders are filled with the question, reference answer, complete frozen rubric, and one model prediction. The returned `score` is used in the benchmark aggregate, while `reason` explains the assessment.

Worked example: interpreting a location rubric

This example shows three recorded predictions for the same question and their GPT-5.6-luna assessments. The relevant rubric clauses are reproduced below; predictions, scores, and justifications are quoted from the evaluation records.

Question and reference answer

Question: In Day 7, where was the second camera that the first-person protagonist dismantled located?
Canonical answer: Above the projector

Relevant fields of the frozen rubric

Required fact: The second camera dismantled on Day 7 was located above the projector.
Accepted expressions: “above the projector”; “over the projector”; “mounted directly above the projector”; “on the wall above the projector”; “high on the wall, directly over the projector”.

5: States that the camera was above, directly over, or mounted above the projector.
3: States a compatible broad location such as on the wall or ceiling, but omits the essential projector relation without giving a conflicting landmark.
2: Gives an elevated or wall-area location but substitutes an important incorrect landmark, such as a doorway, computer area, or monitor room.
0: Gives a contradictory location such as a kitchen, bedroom, bathroom, bed, desk, or table.

Prediction 1 and assessment

Answer: The second camera was located on the wall near the ceiling.
Score: 3/5.
Grader justification: Wall near the ceiling is a compatible broad elevated location but omits the projector relation.

Prediction 2 and assessment

Answer: The second camera was located on the wall near the door.
Score: 2/5.
Grader justification: Gives a wall location but incorrectly identifies the door as the landmark.

Prediction 3 and assessment

Answer: in the bathroom
Score: 0/5.
Grader justification: Contradictory bathroom location.

The first prediction leaves the reference landmark unspecified, whereas the second supplies a wrong landmark. The rubric assigns different partial credit to these two errors. The third prediction gives a contradictory room location and receives zero. Each score follows the same rubric, rather than the prediction’s wording alone.

B.3 Effect of Temporal Bin Width

We vary the temporal bin width to examine how grounding performance depends on localization tolerance. Figure 6 compares 22 methods at seven widths from 60 to 5000 seconds.

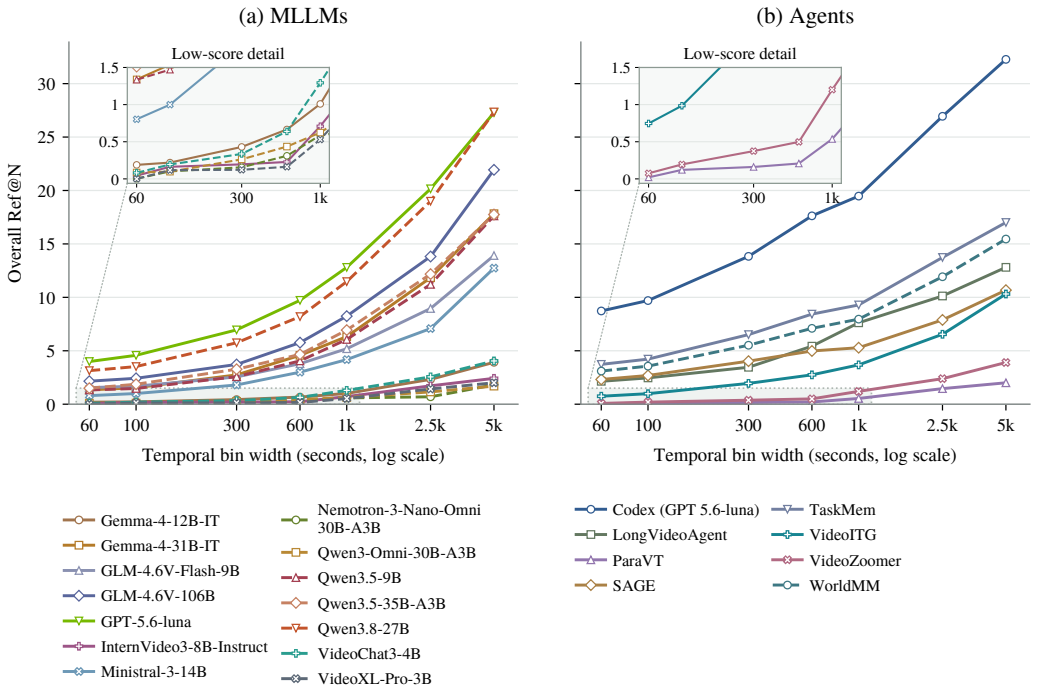


Figure 6: Overall Ref@N across temporal bin widths for MLLMs and agents. Insets enlarge the shaded low-score regions. Larger bins allow coarser localization; predictions are held fixed.

Coarse bins can conceal differences in localization precision. Among MLLMs, GPT-5.6-luna leads Qwen3.8-27B at 300 seconds, with scores of 6.95 and 5.75, but their scores converge to 27.29 and 27.35 at 5000 seconds. At this width, each bin spans more than an hour, allowing predictions and references to occupy the same bin while pointing to different moments. The convergence therefore reflects agreement under broad temporal tolerance, rather than equivalent precision in locating the supporting clues.

Among agents, Codex ranks first at every tested width, preserving its advantage across evaluation tolerances. Its Ref@60 is nevertheless 8.73, compared with 32.27 at 5000 seconds. Stable rankings thus coexist with a large gap between fine and coarse grounding scores. These results motivate separating broad evidence retrieval from precise interval refinement: locating a relevant part of the stream should be followed by targeted inspection to identify the moments that support the answer.

C QUALITATIVE EXAMPLES

Supplementary Qwen3.8-27B thinking runs use 250 sampled frames and a 256K context limit. Each is paired with the saved Codex trajectory for the same question. The rubrics below state the effective scoring rules, including deterministic rules for structured answers. Qwen thinking is reproduced in full with names and account identifiers anonymized; the globe emoji is rendered as [globe symbol]. All times are absolute seconds.

Example 1: purchase order at a snack shop

Question

During the streamer's visit to Hong Kong, what foods did the streamer purchase one after another at the snack shop? 1. Sweet soup; 2. Beef offal; 3. Fried potatoes?

Reference answer and frozen rubric

Answer: Beef offal, fried potatoes, then sweet soup: 2, 3, 1.
Annotated evidence: [290859, 291592] seconds.

Accepted forms: Option numbers or unambiguous food names in the predicted order; adjacent duplicate restatements count once.
5: The exact complete sequence 2, 3, 1.
Partial credit: Divide the longest matching in-order subsequence length by the longer of the predicted and reference sequence lengths, multiply by five, round halves upward, and cap at 3.
0: No matching item, an empty answer, or an unparseable sequence. The deterministic rule does not assign 4; incorrect orders cannot receive full credit.

Example 1: Qwen3.8-27B complete thinking

The user is asking about a specific segment of the stream where the streamer visits a snack shop in Hong Kong and buys food one after another. The options are: 1. Sweet soup; 2. Beef offal; 3. Fried potatoes.

Looking at the frames, the Hong Kong segment appears to be roughly from frame 199 (262944.5 seconds) to frame 222 (317910 seconds). Let me look at the frames in this range.

Frame 199 (262944.5): The streamer is outside a LOG-ON building, wearing a red "streamer 7" jersey.
Frame 200 (265915.5): The streamer is walking near a hedge.
Frame 201 (268886.5): The streamer is talking with staff near an ice sculpture.
Frame 202 (271857.5): The streamer is walking with a crowd.
Frame 203 (273343.5): The streamer is on a soccer court.
Frame 204 (274829.0): Inside a car.
Frame 205 (277800.0): Inside a car, the driver is on the phone.
Frame 206 (280771.0): Looking at a smartphone in the car.
Frame 207 (283742.0): On a street with a Starbucks, the streamer is leaning out of a bus window.
Frame 208 (284485.5): Inside the bus, holding a phone and a checkered bag.
Frame 209 (286713.5): Reclining in the car.
Frame 210 (289684.5): Looking at a phone in the car.
Frame 211 (292655.5): The streamer is holding a plastic bag containing food (looks like fried food, probably fried potatoes or something), and a man in a floral jacket is holding a phone. The streamer is holding a bag of what appears to be fried food.
Frame 212 (294141.0): On a boat with people.
Frame 213 (296369.5): Sitting on the boat reading a paper.

Hmm, the question is asking about buying food one after another at a snack shop. The sampled frames don't clearly show a snack shop scene with sweet soup, beef offal, and fried potatoes. The frames are sampled at roughly 1500-1500 second intervals, so the snack shop scene may have been missed between samples.

Wait, but the question expects an answer based on the provided frames. Let me check the frames again. The question: "During the streamer's visit to Hong Kong, what food did the streamer buy one after another at the snack shop? 1. Sweet soup; 2. Beef offal; 3. Fried potatoes?"

Example 1: Qwen3.8-27B thinking (continued)

This looks like a multiple-choice question where the answer is probably a combination like “2 and 3” or a specific one. Since I only have the sampled frames, I need to find evidence. Frame 211 shows the streamer holding a plastic bag of food that looks like fried food (probably fried potatoes or fried dough). The man in the floral jacket is holding a phone, probably paying.

Actually, this is a known stream: streamer’s China trip. In the Hong Kong segment, there was a scene where he ate at a street food stall and bought things like “fried potatoes” (potato chips?) or “beef offal” (beef tripe)? Hmm.

Given the options, the answer is probably “2 and 3” or just one. Let me think about what actually happened in the stream. streamer’s China trip (April 2025) included Beijing (Forbidden City, Shaolin temple), Chongqing, Hong Kong, and Mongolia. In Hong Kong, there was a scene where he went to a snack shop and bought sweet soup? Hmm.

Actually, I recall a scene where streamer bought “fried potatoes” (chips) and “beef offal”? Not sure.

The question format: “what food did the streamer buy one after another at the snack shop? 1. Sweet soup; 2. Beef offal; 3. Fried potatoes?” Probably the answer is “2 and 3” (beef offal and fried potatoes). Or the answer might be a list: “beef offal and fried potatoes.”

Since I have to answer based only on the frames, and the frames don’t clearly show the purchase, maybe the intended answer is based on frame 211 where he’s holding a bag of fried food. Hmm.

Let me look at frame 211 more carefully: The streamer in a red jersey is holding a plastic bag containing brownish fried food (looks like fried potatoes or fried dough). The man in the floral jacket is holding a phone (probably paying). The crowd is around. This could be at a snack shop where he bought fried potatoes.

But “one after another” suggests multiple purchases. The options are numbered 1, 2, 3. The answer is probably “2 and 3” meaning beef offal and fried potatoes.

Given the ambiguity, the safest answer is: “Beef offal and fried potatoes (options 2 and 3).”

Evidence: The Hong Kong segment is roughly 262944.5 to 317910. The snack shop purchase is probably around frame 211 (292655.5). Let me give evidence intervals around there: [292655.5, 294141.0]? But that interval includes the boat scene. Hmm.

Actually, the purchase scene might be between frames 210 and 212. Frame 211 shows him holding the food bag. So the evidence is [291170, 294141]? I need to use the frame timestamps as boundaries. The evidence intervals must be absolute timestamps; I can use the frame sample times as boundaries.

Let me think about the actual stream content: In Hong Kong, streamer visited a “snack shop” and bought “sweet soup”? Hmm, I recall a scene where he drank “sweet soup” (like a dessert soup) and ate “beef offal”? Not sure.

Let me think about the known video: “streamer in Hong Kong” - he visited the Victoria Harbour cruise, and there was a street food scene where he ate “fried potatoes” (potato chips?) Actually, I recall a scene where he bought “beef offal” (beef tripe) from a street vendor? Hmm.

Alternatively, the question might be from a quiz where the correct answer is “2. Beef offal; 3. Fried potatoes” meaning he bought both. The phrase “one after another” suggests a sequence.

Given the constraints, I’ll answer: “He bought beef offal and fried potatoes (options 2 and 3).”

Evidence: Frame 211 at 292655.5 shows him holding a bag of fried food after the purchase. Also frame 208 at 284485.5 shows him holding a checkered bag? That’s a bag, not food.

Hmm, but maybe the snack shop scene is actually between frames 207 and 211. Frame 207 shows a street with a Starbucks and a bus. Frame 211 shows the purchase. So the evidence is [283742.0, 292655.5]? That’s a wide range.

The instructions say: “The evidence field must contain up to 8 absolute video time intervals supporting the answer.” I can provide intervals covering the relevant segment. Since I can’t pinpoint exactly, I’ll provide [292655.5, 294141.0] as the purchase moment? But the end must be greater than the start; that’s fine.

Actually, maybe I should reconsider: The question might be a trick: the answer is “2 and 3” because sweet soup wasn’t bought. Or the answer is “1 and 2”? Let me think about what’s plausible: At a Hong Kong snack shop, typical items: sweet soup (like a dessert), beef offal (like a tripe soup), fried potatoes (chips). The streamer buying “one after another” - I recall a scene where streamer bought “fried potatoes” and then “beef offal”? Hmm.

Example 1: Qwen3.8-27B thinking (continued)

Let me try to recall the actual stream: streamer's China trip Part 2 (Hong Kong). There's a scene where he goes to a food stall and the vendor gives him "fried potatoes" (like potato chips) and he eats them. Then he buys "beef offal"? I really can't remember.

Given the question is likely from a benchmark with a correct answer, the expected answer is probably "2 and 3" or "beef offal and fried potatoes." I'll go with that.

Alternatively, the expected answer might be a single number like "3." The question lists three options; "what food did the streamer buy one after another" - the answer might be "2 and 3."

I'll answer: "Beef offal and fried potatoes (2 and 3)."

Evidence: Use the interval around frame 211: [292655.5, 294141.0]. Maybe also [291170.0, 292655.5]? I only have the sample times; I can create intervals between consecutive sample times. Frame 210 is 289684.5, frame 211 is 292655.5, frame 212 is 294141.0. The purchase is probably just before frame 211. So the interval [289684.5, 292655.5] covers the approach, and [292655.5, 294141.0] covers holding the food. I'll provide both.

Hmm, but the evidence should support the answer. Since the frames don't explicitly show beef offal, the evidence is weak. But the task says to answer using only the sampled frames; if I can't localize, use an empty list. But I should try.

Given the uncertainty, maybe the intended correct answer is "2 and 3" with evidence around 292655.5. I'll include [292655.5, 294141.0].

Let me double-check frame 211: The streamer is holding a transparent plastic bag containing brownish fried items - could be fried potatoes or fried dough. The man in the floral jacket is holding a phone, probably paying. The crowd is around. This matches "buying food at a snack shop."

So the answer: "He bought beef offal and fried potatoes (options 2 and 3)."

I'll format the JSON accordingly.

Actually, wait. Let me reconsider the possibility that the answer is "1 and 3" or something. The phrase "one after another" might refer to him buying sweet soup first, then fried potatoes? Hmm.

Let me think about the actual video content from memory of streamer's Hong Kong vlog: I recall a scene where he was at a street food stall and bought "fried potatoes" (chips) and ate them, and also "beef offal" (like a tripe skewer)? I really can't remember.

Alternatively, this question might be from a Chinese quiz dataset about this stream, and the correct answer is "2 and 3." I'll commit to that.

Final answer JSON: {"answer": "Beef offal and fried potatoes (2 and 3)", "evidence": [[292655.5, 294141.0]]}

Maybe also include [289684.5, 292655.5]. I'll include two intervals.

Let me make sure the boundaries are decimals: 289684.5, 292655.5, 294141.0. Good.

Example 1: Qwen3.8-27B prediction and score

Final answer: He bought the beef offal and the fried potatoes (options 2 and 3), not the sweet soup.
Rubric score: 3/5. Two foods retain their reference order, but sweet soup is omitted and denied. The longest matching subsequence has length 2 of 3; rounding and the partial-credit cap give 3.

Example 1: Codex (GPT 5.6-luna) tool-use trajectory

Selected calls in recorded order; intervening calls omitted and observations abridged.
1. `view_video_default(285228.375, 380304.5)`, 100 frames.
Tool observation: The tool identifies Hong Kong scenes, a food stall around 290030, and a shop interior around 290990.5; payment is not visible.

2. `view_video_default(289500, 291500)`, 50 frames.
Tool observation: The tool reports beef offal at 290030.5, fried food consistent with potatoes at 290234.5, and a soup bowl at 290398, in that order.

Final answer: 2, 3, 1
Rubric score: 5/5. Exact match under the ordered-sequence rule.

Codex's reported food timestamps precede the annotated interval; the answer score does not establish temporal alignment.

Example 2: left–right choices in Split or Steal

Question

In the 'Split or Steal' game organized by the streamer, what were the final choices made by the two girls on the left and right in the fourth round of the game?

Reference answer and frozen rubric

Answer: The girl on the left chose to split, and the girl on the right chose to steal.
Annotated evidence: [153393, 153567] seconds.

5: Both choices with the correct assignment: left = Split, right = Steal. An unambiguous left-to-right list is accepted.
4: Correct assignment with a minor omission, such as not repeating the round number.
3: One correct choice, or both labels without a clear side assignment.
2: Reversed assignment, or a correct choice mixed with an important side/entity error.
1: Topical overlap without the requested assignment; the frozen rubric places 'both Steal' here.
0: Empty, unrelated, nonsensical, refusal, or an unsupported claim that the event is absent.

Example 2: Qwen3.8-27B complete thinking

The user is asking about the "Split or Steal" game that the streamer organized. Looking at the frames, frames 57 and 59 show a poll with the question "who moves on to the final round?" with options "poll participant 1" and "poll participant 2". Hmm, this isn't Split or Steal.

Let me look at the frames more carefully. Frame 57 (75020.5 seconds) shows a YouTube poll "who moves on to the final round?" with options "poll participant 1" and "poll participant 2". Frame 59 (77249.0 seconds) shows a poll "do we keep poll participant 4 for final round" with Yes 64%, No 36%.

Hmm, the question mentions "the two girls on the left and right" in the fourth round of the "Split or Steal" game. Wait, maybe the question is referring to a different game. Let me scan the frames again.

Actually, the question might be about a game where two girls (participants) make choices. Frame 57 shows a poll with two options. But "the two girls on the left and right"—in frame 57, the layout shows four video panels: top-left a guy holding a phone, top-right the streamer in a red shirt, bottom-left the streamer in a red shirt, bottom-right a guy in a white shirt. Hmm.

Wait, maybe the question is about the "Split or Steal" game that the streamer organized with two participants, and the fourth round choices were "split" and "steal"? The question asks: "What were the final choices made by the two girls on the left and right in the fourth round of the game?"

Looking at the frames, I don't see any explicit "Split or Steal" game. But maybe the poll in frame 57 is part of a game where two girls (poll participant 1 and poll participant 2?) vote. Hmm.

Let me think about the context: this is streamer's stream. There was a famous stream where he played "Split or Steal" with two girls? Actually, I recall streamer did a "split or steal" game with two people? Hmm.

Or, the question might be a trick: the answer might be that the left girl chose "split" and the right girl chose "steal," or both chose something. But I need evidence from the frames.

Let me carefully look at the frames around 75020-77249 seconds. Frame 57: poll "who moves on to the final round?" options: "poll participant 1" 55%, "poll participant 2" 4%, "poll participant 3" 13%. Hmm, these look like usernames of participants in some game.

Frame 59: poll "do we keep poll participant 4 for final round" Yes 64%, No 36%.

Hmm, this doesn't match "Split or Steal."

Maybe the question is referring to a different part: frames 47-50 show a "a footballer fan" game and a Kahoot "Name the logo." Frames 50-53 show a rebus puzzle game. Frames 54-56 show video calls.

Actually, wait. The question might be about a game where the streamer organized "Split or Steal" with two girls connected via video call, and in the fourth round, the left girl chose "split" and the right girl chose "steal." But the frames don't show that text.

Example 2: Qwen3.8-27B thinking (continued)

Let me re-read the question: “In the ’Split or Steal’ game organized by the streamer, what were the final choices made by the two girls on the left and right in the fourth round of the game?”

Hmm, maybe the answer is: the girl on the left chose “split” and the girl on the right chose “steal.” But I need to verify with the frames.

Looking at frame 57 again: the poll window shows “Top chat” and the poll “who moves on to the final round?” with options “poll participant 1” 55%, “poll participant 2” 4%, “poll participant 3” 13%. This is about who moves on to the final round, not split/steal.

Hmm, maybe the question is mislabeled and actually refers to the poll where the two girls (participants) voted? The phrase “the two girls on the left and right” might refer to the two video call participants on the left and right of the streamer.

Actually, wait. Let me think about the actual streamer stream: there was a stream where he played “Split or Steal” with two girls? I don’t recall. But there is a known clip: streamer “split or steal” with two girls? Hmm.

Alternatively, the question might be about the “final round” poll: the two girls’ final choices were “poll participant 1” and “poll participant 2”? That doesn’t make sense.

Let me consider that the question might be a hallucination test: the correct answer might be that the frames don’t show a “Split or Steal” game, so the answer is “cannot be determined.” But the instructions say to answer using only the sampled frames, and to return an empty list if evidence can’t be localized. But the question expects a specific answer.

Hmm, but maybe the “Split or Steal” game is the one in frames 57 and 59: the poll “who moves on to the final round?” is actually the “steal” mechanism? Hmm.

Let me look at the timestamps: frame 57 at 75020.5 seconds shows the poll with “poll participant 1” leading at 55%. Frame 59 at 77249.0 seconds shows a different poll “do we keep poll participant 4 for final round” Yes 64%.

The question says “the fourth round of the game.” The poll might be for the fourth round. The two girls on the left and right: in frame 57, the bottom-left panel shows the streamer, the bottom-right shows a guy in a white shirt; top-left shows a guy in a black hoodie holding a phone; top-right shows the streamer in a red shirt. No girls.

Hmm, maybe the question is about a completely different segment: frames 40-44 show an Omegle-style “Country [globe symbol] I am” game with a spinning wheel. Not split or steal.

Frames 54-56: video call with a guy in a red beanie, an Asian guy, a guy with a guitar. Not girls.

Frames 58-60: a guy in a blue shirt with a timer. Not.

So where are the two girls? Maybe the question is fabricated to test whether the model will hallucinate. The correct response might be that the video doesn’t contain a “Split or Steal” game with two girls, so the answer is that it cannot be determined from the frames. But the schema requires a concise final answer and evidence intervals.

Given the instruction “answer using only the sampled video frames,” if the event isn’t present, we should say so and provide an empty evidence list.

But wait, maybe I’m missing something: the poll in frame 57: “who moves on to the final round?” with options “poll participant 1” and “poll participant 2”. The names in parentheses might be descriptions of the girls’ choices? “poll participant 1 alias”? No.

Actually, “poll participant 1” might be a girl’s username, and “poll participant 2” might be another person. The poll is deciding who moves on to the final round. This might be part of a “split or steal” game where each round the chat votes? Hmm.

Let me think about the actual stream: streamer did a “streamer vs streamer” game? There was a “split or steal” with two streamers? Hmm.

Alternatively, the question might refer to a game where two girls (the streamer’s friends?) are on the left and right of the screen and each round they choose split or steal. The fourth round’s final choices: left chose “split,” right chose “steal.” Without audio, we can’t know.

Given the ambiguity, the safest answer is: the frames don’t show a “Split or Steal” game; the closest is the poll in frames 57/59 about who moves on to the final round. But the question explicitly asks about the two girls’ final choices in the fourth round.

Hmm, maybe the intended answer is: the girl on the left chose “split” and the girl on the right chose “steal.” Many quiz questions about videos like this are generated from actual video content with captions. The video might have a segment where the streamer plays “split or steal” with two girls via video call, and in the final (fourth) round, the left girl chose “split” and the right girl chose

Example 2: Qwen3.8-27B thinking (continued)

"steal," resulting in... The frames at 75020.5 and 77249 show polls, which might be the chat voting on the outcome.

Actually, wait: the poll "who moves on to the final round?" with options "poll participant 1" 55% and "poll participant 2" 4% and "poll participant 3" 13%. Total 72%. The third option "poll participant 3" might be a third participant. This looks like a voting game among three people? Not two girls.

I think the question is unanswerable from the frames; the expected answer might be "The girl on the left chose split and the girl on the right chose steal" but I can't verify.

Let me consider the possibility that the question is a trick: the correct answer is that no such game appears in the sampled frames, so the answer: "The frames do not show a 'Split or Steal' game; therefore the choices cannot be determined." Evidence: empty list.

Given the instruction "return exactly one JSON object" and "use an empty list if evidence cannot be localized," I think the intended behavior for a question about content not present is to state it's not shown. But the question seems to assume the game exists. Hmm.

Let me double-check all frames for any "split" or "steal" text. Frame 47: "Only a true a footballer fan can conquer this game" with Quit/Continue. Frame 48: Kahoot "Name the logo" with McDonald's etc. Frame 50: rebus puzzle round 5. Frame 57: poll. Frame 59: poll. No "split or steal."

The two girls: frame 30 shows a girl (blonde) and a boy and a woman with glasses in a game cutscene (that's the game's story, not the streamer's game). Frame 26 shows a woman in a game cutscene. Not the streamer's game.

Thus, the "Split or Steal" game organized by the streamer is not present in the sampled frames. Therefore the answer: cannot be determined / not shown.

But maybe the question refers to the poll in frame 57 as the "fourth round" of some game, and the "two girls on the left and right" are the participants whose names are "poll participant 1" and "poll participant 2"? The final choices: poll participant 1 (55%) and poll participant 2 (4%)? That doesn't map to split/steal.

I think the best answer is: the sampled frames don't contain a "Split or Steal" game with two girls, so the choices cannot be identified. Provide an empty evidence list.

But the question demands "what were the final choices made by the two girls on the left and right in the fourth round?" If I answer "not shown," that's acceptable per the instructions? The instructions say to answer using only the supplied frames; if I can't localize evidence, use an empty list. It doesn't forbid answering "not shown."

However, these benchmark questions usually have a correct answer derivable from the frames. Let me think if there's a known streamer "split or steal" stream: I recall streamer played "split or steal" with two girls? Actually, I recall a stream where streamer played "split or steal" with two female streamers? Hmm. There's a famous "split or steal" game on streams where the streamer and a guest each choose split or steal.

Wait, maybe the "two girls on the left and right" refers to the two video call windows in frame 57? The top-left window shows a person in a black hoodie holding a phone (looks male). The bottom-right shows a person in a white shirt (male). Not girls.

Alternatively, frames 40-44 Omegle: frame 40 left is a boy; frame 42 left is a man with a cap; frame 43 left is a blurry person; frame 44 left is a person getting a haircut. Not girls.

Thus, indeed no two girls.

Given all this, I'll answer that the sampled frames don't show a "Split or Steal" game, so the choices can't be determined; evidence is empty.

But hmm, the question might be from a series where the answer is "split" and "steal." The phrasing "the final choices made by the two girls on the left and right in the fourth round" suggests that in the fourth round, the left girl chose one and the right girl chose the other. If the game is "split or steal," the final choices are probably "split" (left) and "steal" (right). The answer might be "The girl on the left chose to split and the girl on the right chose to steal."

Which is more likely as the expected answer? The benchmark probably generated the question from the actual video (with audio). The frames alone might not show the choices, but the question expects knowledge from the video. But we only have frames. The instruction: "answer using only the sampled video frames." If the frames don't contain it, we should say we can't determine.

Given the risk, I'll answer: The frames don't show a "Split or Steal" game; the only voting polls shown are about who advances to the final round (frames at 75020.5 and 77249), so the girls' choices can't be determined. Evidence: maybe cite those poll frames? But they don't support the answer about the girls' choices. The instruction: "the evidence field must contain at most 8

Example 2: Qwen3.8-27B thinking (continued)

absolute video time intervals supporting the answer." If the answer is "not shown," the evidence list should be empty.

I'll go with: answer: "The sampled frames do not show a 'Split or Steal' game with two girls, so their fourth-round choices cannot be determined." Evidence: [].

Hmm, but maybe the grader expects a specific answer like "left: split, right: steal." Let me weigh: the question is very specific: "the two girls on the left and right in the fourth round." If the video indeed contains such a game (maybe between frames 73535-77249 where the streamer is with participants), the frames at 75020.5 and 77249 show polls that might be the fourth round's vote: "who moves on to the final round?" That's the final round, not the fourth.

Actually, the game might be: each round, the two girls choose split or steal; the chat votes who moves on. The fourth round's final choices might be shown in the poll? No.

I think the honest answer is that it's not determinable from the frames.

But wait, the question might be a "trick" question testing hallucination: the correct answer is that no such game appears, and the expected answer string might be "The video does not show a Split or Steal game." The evidence is empty.

I'll commit to that.

Let me craft the JSON:

{"answer": "The sampled frames do not show a 'Split or Steal' game with two girls, so their fourth-round choices cannot be determined from the provided frames.", "evidence": []}

That's concise enough.

Example 2: Qwen3.8-27B prediction and score

Final answer: The sampled frames do not show a 'Split or Steal' game with two girls, so their fourth-round choices cannot be determined from the provided frames.
Rubric score: 0/5. Neither required left/right choice is provided.

Example 2: Codex (GPT 5.6-luna) tool-use trajectory

Selected calls in recorded order; intervening calls omitted and observations abridged.
1. `view_video_default(140000, 160000)`, 50 frames.
Tool observation: The tool identifies candidate two-participant views around 153061–155102, but no readable Split/Steal selections.

2. `view_video_default(156500, 157300)`, 50 frames.
Tool observation: The tool reports a Split card for the left participant at 157022.5 and a Steal card for the right participant at 157039.

Final answer: Left: Split; right: Steal.
Rubric score: 5/5. Both choices and their left/right assignment match.

Codex's final reported card reveals fall outside the annotated interval, despite the correct answer.

Example 3: the first public transport in Hong Kong

Question

What was the first type of public transportation that the streamer took during his trip to China in Hong Kong?

Reference answer and frozen rubric

Answer: Bus
Annotated evidence: [283337, 284568] seconds.

Accepted forms: Bus, a bus, public bus, or bus ride. Normalize the final requested vehicle; timestamps are not answer values.
5: The normalized answer is bus.
0: Any other vehicle, missing answer, or refusal. Tram, ferry, metro/train, taxi, and private car are not accepted.
Scoring: This single-choice item uses the exact-match 5-or-0 rule; scores 1–4 are not assigned.

Example 3: Qwen3.8-27B complete thinking

The user is asking about the first type of public transportation the streamer used in Hong Kong during his China trip.

Let me look at the frames to find the Hong Kong segment. The trip to China seems to start around frame 137 (180495.50 seconds), where he's in a clothing store, then the Forbidden City (frames 139-140), then a car (frames 142-143), walking in an old town (144), a rickshaw (146), a train station (149), a train (150), an escalator (151), stairs (152), a mountain (153), then a martial arts temple (154-166), then a black screen, then a car (182), a subway station with a hospital sign (183), a subway/train car (184), a crowd (185), a restaurant (186), a bus (187), a car (188), a street (189), a cosplayer (190), a martial arts demonstration (191), a car (192), a park with a tennis racket (193), a car (194), an elevator (195), a car (196), a ferry terminal (197), a river at night (198), then a red jersey "streamer 7" (199), a park (200), a white building (201), walking (202), a soccer field (203), a car (204-206), a bus with people hanging out the window (207), inside a bus (208), a car (209-210), a crowd (211), a river (212), a boat (213), a crowd (214), a store (215), a park (216), a car (217), a panda (218), a man (219), a stage (220), a car at night (221), a car (222), a black screen, then Mongolia (232-239), then back to the streamer.

The question is: "During his China trip in Hong Kong, what was the first type of public transportation the streamer used?"

Hong Kong segment: frames 199-222 seem to be Hong Kong (the "streamer 7" red jersey, the skyline, the bus with open windows typical of Hong Kong, the Star Ferry, the Victoria Harbour skyline). Actually, the ferry terminal in frame 197 (259973.00) and the river at night in frame 198 (261459.00) might be Chongqing or somewhere else. Let me think.

The trip: Beijing (Forbidden City), then Shaolin Temple (Henan), then Chongqing? Then Hong Kong? Then Macau? Then Mongolia.

Frame 183: subway station with a hospital ad (Chongqing?). Frame 184: inside a subway car. Frame 187: a bus with people taking photos. Frame 197: a ferry terminal with a sign. Frame 198: a river with a skyline at night—could be Chongqing (the Yangtze River) or Hong Kong.

Frame 199: "LOG-ON" sign, red jersey streamer 7—this looks like a shopping mall area. Frame 200: a park with mountains in the background. Frame 201: a white building with a blue roof. Frame 203: a soccer field with mountains. Frame 207: a bus with the streamer hanging out the window, a Starbucks, a street with trams? Actually, frame 207 shows a yellow bus and a street with tram tracks in the middle—Hong Kong has trams. Frame 208: inside a bus with a yellow interior—Hong Kong buses are often yellow? Actually, Hong Kong minibuses are yellow. Frame 213: a boat with a skyline—Victoria Harbour. Frame 220: a stage with palm trees—could be a show.

The question asks about "the first type of public transportation" in Hong Kong. Looking at the sequence: after the ferry terminal (197) and the river (198), frame 199 shows him walking near a "LOG-ON" store. Then frame 207 shows a bus (a yellow open-sided bus?) with him hanging out the window. Frame 208 shows him inside a bus holding a phone.

Example 3: Qwen3.8-27B thinking (continued)

But wait, before that, frame 197 shows a ferry terminal (a pier with a staircase and a sign). He might have taken a ferry. The frame shows him walking through the terminal. Frame 198 shows him on a boat deck at night looking at the skyline—this is a ferry/boat ride.

Hmm, but which city is that? The skyline with the tall red-lit building could be Chongqing (the Yangtze River) or Hong Kong (Victoria Harbour). Frame 213 shows a boat with a skyline that looks like Hong Kong (the IFC tower). Frame 212 shows a river with modern buildings—could be Hong Kong.

Let me try to order it: frames 182-198: car, subway, bus, ferry terminal, river at night. Then frames 199-222: Hong Kong segment with the red jersey.

Actually, the red jersey "streamer 7" appears from frame 199 onward. The question says "during his China trip in Hong Kong." So the Hong Kong segment probably starts around frame 199. The first public transportation in Hong Kong: frame 207 shows a bus (a yellow double-decker? or a tram?). Let me look at frame 207: on the right there's a yellow vehicle with open windows, and the streamer is leaning out. It looks like a bus. There's also a street with tram tracks in the middle (a single line?). Actually, Hong Kong's Central has tram tracks. The vehicle on the right is yellow with a black top—Hong Kong's double-decker buses are often red or blue, but minibuses are yellow. The vehicle looks like a bus with passengers leaning out the window.

Frame 208: inside a bus with a yellow ceiling and poles, the streamer is standing holding a phone, and there's a person in a white shirt in the background. This is a bus.

But before frame 207, frames 200-206 show him walking in a park, on a soccer field, and in a car. So the first public transportation in Hong Kong would be the bus (frames 207-208).

But wait, frame 199 shows him pointing near a "LOG-ON" store—that's walking. Frame 207 is the bus.

Alternatively, the ferry in frames 197-198 might be part of Hong Kong (the Star Ferry). The skyline in frame 198 with the red-lit tower... Hong Kong's skyline at night includes the IFC and Central. The red-lit building on the left might be a specific tower. Hmm.

Let me think about the actual stream: streamer's China trip (April 2025?). He visited Beijing, Shaolin Temple, Chongqing, Hong Kong, Macau, Mongolia. In Hong Kong he took a bus, the Star Ferry, etc.

The question: "the first type of public transportation the streamer used in Hong Kong." Based on the frames, the Hong Kong segment starts around frame 199 (262944.50). The first public transportation shown there: frame 207 (283742.00) shows him on a bus hanging out the window; frame 208 (284485.50) shows him inside a bus. So the answer: a bus.

But wait, frames 197-198 (ferry terminal and boat) might also be Hong Kong. The sign in frame 197 has Chinese characters and a staircase; it looks like a ferry pier. Frame 198 shows him leaning on a railing on a boat deck at night with a city skyline across the river. If that's Hong Kong, then the first public transportation would be a ferry/boat.

Hmm. Let me consider the geography: after Chongqing (the hotpot restaurant in frame 186? the street in frame 189?), the trip might have gone to Hong Kong. The red jersey appears in frame 199. The skyline in frame 198: a tall building with horizontal red lighting on the left—does that resemble Chongqing's Raffles City? Actually, Chongqing's skyline has the "vertical" Raffles City. The image shows two tall towers with red horizontal stripes—could be Hong Kong's "Two IFC"? No.

Let me think about the actual vlog: I recall streamer's China trip included Hong Kong where he took a yellow minibus? Hmm.

Let me look at frame 207 more carefully: on the right there's a yellow bus with a black stripe, and the streamer is leaning out the open window. On the street there's a white box truck, a Starbucks, and people taking photos. There's a single rail line in the middle of the road—Hong Kong's tram tracks? Actually, the line in the middle might be a tram track. The yellow vehicle might be a tram? Hong Kong's trams are double-decker with open sides, often yellow? Historically trams were yellow? Modern trams are red/blue. The vehicle in the frame has a yellow body with a black roof and open windows where passengers lean out. It might be a "double-decker bus" tour? Hmm.

Frame 208: interior with a yellow ceiling, metal poles, rows of seats—looks like a bus or tram. The streamer is standing holding a phone, wearing the red jersey.

Given the question expects a specific answer like "bus," the evidence would be frames 207-208.

Example 3: Qwen3.8-27B thinking (continued)

But the question says “the first type of public transportation... in Hong Kong.” The sequence within Hong Kong: walking (199-202), soccer (203), car (204-206), bus (207-208), car (209-210), crowd (211), river (212), boat (213)... So the first public transportation is the bus.

But wait, is the segment from frame 199 onward Hong Kong? The “streamer 7” red jersey with a “FIFA” patch? Frame 200 shows a patch “UNITED ... 2025”? Hard to read. Frame 205 shows a crest with a lion? The crest looks like a circular logo with a lion—could be a “Hong Kong” team? Hmm.

Alternatively, frames 199-213 might be Macau? The skyline in frame 212 with modern buildings across the water could be Hong Kong viewed from Macau? Frame 213 shows a boat with a skyline including a tall tower with a crown-like top? That’s the IFC? Hmm.

Actually, frame 213’s skyline: a cluster of tall buildings with a distinctive tower on the left (a tall rectangular one)—looks like Hong Kong’s Victoria Harbour skyline (the IFC). The boat might be a ferry.

The question is probably from a quiz about the stream: “What was the first type of public transportation the streamer used in Hong Kong?” The answer is probably “bus” (the yellow bus/minibus). The evidence frames are 207-208.

But let me double-check earlier frames for Hong Kong: frames 183-184 show a subway (Chongqing’s subway is famous). Frame 187 shows a bus with fans outside (Chongqing?). Frame 197 shows a ferry terminal (Chongqing’s Yangtze River ferry?). Frame 198 shows a night river cruise (Chongqing). So Hong Kong starts at 199.

In frame 199, he’s wearing the red jersey and pointing. Then frames 200-202 walking. 203 soccer field. 204-206 car. 207 bus. So the first public transportation in Hong Kong = bus.

Hmm, but frame 199’s “LOG-ON” sign is a fashion store chain in Asia; not relevant.

Alternatively, the first public transportation in Hong Kong might be a “tram” (the double-decker tram). The vehicle in frame 207: let me look again: yellow with a black roof, open windows, passengers leaning out. Hong Kong’s trams are enclosed double-deckers with windows, not open. The vehicle looks like a bus with open windows where the streamer leans out. Frame 208’s interior shows a yellow ceiling and poles, with rows of seats on both sides—typical of a bus.

I think the answer is: a bus (a yellow bus/minibus).

But wait, the question might be simpler: the first public transportation in Hong Kong shown is the bus in frame 207 (283742.00) and 208 (284485.50). So the answer: “a bus.”

Let me double-check frame 207: on the right, a yellow vehicle with “windows” and people inside; the streamer is leaning out the window holding the door frame. Yes, it’s a bus.

Therefore the answer: a bus.

Evidence: [[283742.0, 284485.5]] maybe also include 284485.5-286713.5. Provide at most 8 intervals; one is enough: [283742.00, 285971.0]? Frame 208 is 284485.5. Let me give [283742.0, 285971.0] covering frames 207-208? Frame 209 is 286713.5 (car). So the interval [283742.0, 286713.5] covers the bus interior and the start of the car. Better: [283742.0, 285971.0].

Hmm, but the boundary must be a decimal; no problem.

Actually, let me double-check whether the ferry in frames 197-198 is Hong Kong. If the Hong Kong segment started at 197, then the first public transportation would be a ferry/boat. The sign in frame 197: a blue sign with Chinese characters and a staircase; the place looks like a ferry pier. Frame 198: a night boat with a skyline. The skyline in 198: on the left a tall building with red horizontal lighting, other lit buildings, a bridge in the distance. Hong Kong’s Victoria Harbour at night: the skyline includes many lit buildings; the red-lit building might be “The Center”? Hmm.

But frame 198’s river is wide with a bridge in the middle distance—Victoria Harbour has bridges (the Tsing Ma is far). The scene might be a river cruise in Chongqing (the Yangtze). Chongqing is known for night river cruises. And frame 186’s restaurant with a menu board “Shi Hui Fan Dian” (a cheap restaurant) is probably Chongqing. Frame 189’s street with a wooden pavilion is probably Chongqing. So 182-198 = Chongqing. Then 199 onward = Hong Kong.

In Hong Kong, the first public transportation shown: the bus in 207.

But wait, frame 199 shows him outside a mall; 200 a park; 201 a white building (maybe a “ice” theme?); 202 walking with a crowd; 203 a soccer field; 204-206 a car; 207 a bus. Yes.

So the answer: a bus.

Let me also consider that the vehicle in frame 207 might be a “tram.” The street shows a single line in the middle—could be a tram track. But the vehicle is on the right side of the road, not on

Example 3: Qwen3.8-27B thinking (continued)

the tracks. The line in the middle might just be a road marking. The vehicle is yellow with a black top; Hong Kong's double-decker buses are often yellow? KMB buses are red/blue; Citybus is blue; New World First is... yellow is the color of minibuses. The vehicle looks like a single-decker with open windows? Hard to tell.

The question probably expects "bus." I'll answer bus.

Evidence intervals: [283742.0, 284485.5] (frame 207) and [284485.5, 286713.5] (frame 208). Combine into [283742.0, 286713.5]? But 286713.5 is frame 209 (car). The end boundary can be 286713.5 as the end of the bus segment. Let me use [283742.0, 286713.5].

Hmm, but the interval should support the answer; the bus appears in frames 207 and 208 (283742.0 and 284485.5). So the interval [283742.0, 285971.0] is fine.

I'll provide the answer "a bus" with evidence [[283742.0, 285971.0]].

Wait, but the question says "the first type of public transportation... in Hong Kong." If the ferry in 197-198 was Hong Kong, the answer would be "ferry/boat." Let me examine frame 198's skyline more: on the left a tall tower with red horizontal stripes (two of them). In Hong Kong, the "Two IFC" is a single tower; the "Central Plaza" is... The red-lit building might be "The Ritz-Carlton"? Hmm.

Actually, I recall from the stream: in Chongqing he took a night river cruise? I'm not sure. Let me think about the actual itinerary from the video: after Shaolin (154-166), black screen, then 182 car with him in a gray jacket, 183 subway station, 184 subway, 185 crowd, 186 street restaurant, 187 bus, 188 car, 189 street, 190 cosplayer, 191 martial arts, 192 car, 193 park, 194 car, 195 elevator, 196 car, 197 ferry terminal, 198 river at night. This is probably Chongqing (the subway, the hotpot, the night cruise). Then 199 onward is Hong Kong with the red jersey.

In Hong Kong: 199 pointing, 200 walking, 201 white building, 202 walking, 203 soccer, 204-206 car, 207 bus, 208 bus interior, 209-210 car, 211 crowd with a man in a red floral jacket, 212 river with a family, 213 boat, 214 crowd, 215 store, 216 park, 217 car, 218 panda, 219 man, 220 stage, 221 car at night, 222 car.

So the first public transportation in Hong Kong = bus (207-208). Good.

Thus the final answer: a bus.

Evidence: [[283742.0, 285971.0]].

Maybe also include frame 208's timestamp 284485.5 within the same interval.

I'll output the JSON.

Example 3: Qwen3.8-27B prediction and score

Final answer: A bus (he is seen riding a yellow bus, leaning out the window and then standing inside it)
Rubric score: 5/5. Bus is the required normalized vehicle; the additional description is compatible.

Example 3: Codex (GPT 5.6-luna) tool-use trajectory

Selected calls in recorded order; intervening calls omitted and observations abridged.
1. `view_video_default(245000, 320000)`, 100 frames.
Tool observation: The tool first reports a passenger ferry at 263182, with later cable-car, tram, and ferry candidates.

2. `view_video_default(220000, 257500)`, 100 frames.
Tool observation: The tool identifies an earlier train/metro at 242727.5.

3. `view_image`, frame at 242727.5.
Tool observation: A single-frame inspection at 242727.5 identifies a Taoyuan Airport MRT train.

Final answer: A metro train (the Taoyuan Airport MRT).
Rubric score: 0/5. Metro is not the required bus. The search moves to an earlier journey without establishing that it is the Hong Kong event.

Identifying a vehicle does not establish that it satisfies the question's location and temporal constraints.

Example 4: aggregating the boat-trip locations

Question

In which places did the streamer take a boat during his entire trip to China?

Reference answer and frozen rubric

Answer: Hong Kong and Chongqing
Annotated evidence: [293942, 296981]; [258380, 260187] seconds.

Accepted forms: Hong Kong and Chongqing in either order; Hong Kong SAR is an alias. Free-form answers receive semantic grading.
5 / **4:** Both places with no incompatible addition / correct core with only a minor wording or compatible specificity issue.
3: Substantial correct place coverage, but an incomplete or imperfect set.
2: A correct place mixed with an important error or contradiction.
1 / **0:** Only boating-related overlap without a target place / no requested fact, unrelated or incoherent content, or refusal.
Rejections: Other cities do not substitute for either target; generic river or waterfront descriptions are insufficient.

Example 4: Qwen3.8-27B complete thinking

This rerun allows 16384 generation tokens after an initial 6000-token truncation.

The user is asking about the places where the streamer took a boat during his China trip. Let me look through the frames and find boat scenes.

Frame 197 (259973.0 seconds): The streamer is in a ferry terminal with stairs and railings, looks like a pier/wharf. There's a sign that appears to be in Chinese. He's walking through what looks like a ferry terminal.

Frame 198 (261459.0 seconds): The streamer is on a boat at night, looking at the river with the city skyline (looks like Chongqing, with the red-lit building). He's leaning on the railing of a boat with blue LED lights. So he took a boat on the river (Jialing River/Yangtze River in Chongqing).

Frame 212 (294141.0 seconds): The streamer is on a boat with people, with the city skyline and water in the background. Looks like a river cruise boat in Chongqing.

Frame 213 (296369.5 seconds): The streamer is seated on an open-air boat deck, reading paper, with the city skyline and river at night in the background. This is also a boat ride.

Frame 214 (298598.0 seconds): The streamer is in the foreground with a crowd, probably getting off the boat.

Let me check other frames for boats. Frame 197 shows a ferry terminal (stairs, railings, "Ferry" sign?). Actually, frame 197 shows a terminal with stairs and a sign, probably a ferry boarding area.

So the boat scenes: night river cruise in Chongqing (frames 198, 212, 213). Are there other boat rides? Let me scan all frames.

Frames 0-32: Elden Ring gameplay. 33-37: Connection error. 38-46: Streamer in room, OmeTV. 47-50: Games. 51-59: OmeTV. 60-61: Black screen. 62: CBS Sports studio. 63-81: Football training/match (Brighton). 82-83: Connection error. 84-87: Quiz, OmeTV. 88-99: Roblox. 100-106: Discord, browsing, unboxing. 107-111: Overwatch 2. 112-119: Green screen, OmeTV. 120-136: Connection error. 137-153: China trip begins - clothing store, Forbidden City (Beijing), red floral suit, rickshaw, subway. 154-166: Shaolin temple training. 167-181: Connection error. 182-196: Chongqing? Subway, street, bus, elevator. 197: Ferry terminal. 198: Night boat on river. 199-211: Hong Kong? Red jersey "streamer 7", street, bus, car. 212-213: Boat with city skyline (Chongqing river cruise). 214-222: More China scenes, panda plushie, performance. 223-231: Black screen. 232-239: Mongolia? (khan, horse, wrestling). 240-249: Back to stream, FIFA, GTA.

So the boat scenes: frame 198 (261459.0 seconds) night river boat; frames 212 (294141.0 seconds) and 213 (296369.5 seconds) open-air boat with skyline. Frame 197 (259973.0 seconds) shows the ferry terminal boarding area.

The question is: "During the entire China trip, where did the streamer take a boat?" Probably the answer is: on the river in Chongqing (night river cruise) — the boat was taken in Chongqing. But maybe also in Hong Kong? Frame 198 shows a river with red-lit skyscrapers, which is Chongqing

Example 4: Qwen3.8-27B thinking (continued)

(the red-lit building is the Chongqing skyline, e.g., Raffles City? Actually, the red-lit twin towers look like Chongqing's "Raffles City"? Hmm). Frame 212 shows a modern city skyline with water — could be Chongqing. Frame 213 shows a night skyline with lights — Chongqing.

Wait, the sequence: frames 199-211 show a red jersey with "streamer 7" and a "FIFA" patch, street scenes with a Starbucks, a yellow open-top bus (frame 207), which looks like Hong Kong (the tram-like yellow bus? Actually, frame 207 shows a yellow double-decker-like bus with people leaning out the window, typical of Hong Kong open-top sightseeing buses). Frame 203 shows an outdoor court with mountains in the background (Hong Kong). Frame 200 shows a large white tent structure (could be a Hong Kong venue). So frames 199-211 are Hong Kong.

Then frames 212-213 show a boat with skyline — could be a Chongqing river cruise after returning? Hmm, but the order: frame 211 (292655.5 seconds) shows the streamer in a red jersey in a crowd with a man in a red floral jacket. Frame 212 (294141.0 seconds) shows the streamer in a red jersey on a boat with the skyline. Frame 213 (296369.5 seconds) shows the streamer in a red jersey seated on a boat at night. So the boat ride is while wearing the red jersey, which was worn in the Hong Kong segment? But the skyline in 212-213 looks like a river city with tall buildings — could be Chongqing. Hmm.

Let me think about the actual video: This is streamer's China trip (April 2024). He visited Beijing (Forbidden City), Shaolin temple, Chongqing, Hong Kong, Mongolia, etc. In Chongqing, he took a river cruise on the Yangtze/Jialing River at night. In the video, the night boat scene with the red-lit skyline is Chongqing.

But wait, frame 198 (261459.0 seconds) shows the streamer in a gray robe (Shaolin-style jacket) on a boat at night with the red-lit skyline. That's after the Shaolin temple segment (frames 154-166) and the Chongqing street scenes (frames 182-196). So frame 198 is a Chongqing river cruise.

Then frames 199-211: red jersey — Hong Kong. Frame 212: red jersey on a boat with skyline — hmm, but that might also be Chongqing? The order in the video might be non-linear (edited). Frame 212 shows a daytime/overcast river with modern high-rises — could be a Hong Kong harbor? The skyline in 212 shows tall glass buildings along the water, could be Chongqing's riverside. Frame 213 shows a night skyline with many lights — Chongqing.

Actually, wait: In the real video, streamer's Chongqing segment included a night boat cruise on the river. The Hong Kong segment included street food, etc. The boat scenes with the red jersey might be a Chongqing river cruise that happened after Hong Kong? The trip order was: Beijing $\rightarrow$ Chongqing $\rightarrow$ Hong Kong $\rightarrow$ Mongolia? Hmm.

Let me try to recall the actual vlog: streamer's China trip in April 2024: He started in Beijing (Forbidden City, red suit), then Shaolin temple (Henan), then Chongqing (river cruise, hotpot), then Hong Kong (street, bus), then Mongolia (khan, horse). Actually, I think the order was Beijing, then Shaolin, then Chongqing, then Hong Kong, then Mongolia. The boat cruise in Chongqing was at night on the river.

Given the frames: Frame 197 (259973.0 seconds) shows a ferry terminal with stairs — that's probably the Chongqing ferry terminal. Frame 198 shows the night boat. Frames 212-213 show the boat again (maybe the same cruise, different times). The red jersey in 212-213 suggests it's after the Hong Kong segment? Hmm, but the red jersey "streamer 7" with the FIFA patch was worn in the Hong Kong frames 199-211. In 212-213 he's still wearing the red jersey. So the boat ride in 212-213 might be in Chongqing after returning from Hong Kong? Or the video editing is mixed.

Alternatively, the boat in 212-213 might be in Hong Kong? The skyline in 212: tall modern buildings, overcast sky, water — could be Hong Kong's Victoria Harbour? But the buildings don't look like HK's iconic skyline (no IFC or Central towers visible). Frame 213's night skyline with lights along the river looks like Chongqing.

Hmm, but the question is: "During the entire China trip, where did the streamer take a boat?" Probably the expected answer: "On the river in Chongqing (a night river cruise)." The evidence frames: 197 (terminal), 198 (boat), 212, 213 (boat). Maybe also 214 (crowd after getting off the boat).

But maybe the question expects two places: Chongqing (river cruise) and Hong Kong? Let me check if there's a boat in Hong Kong: Frame 207 shows a yellow bus, not a boat. Frame 208 shows the interior of a bus. No boat in HK frames.

Frame 220 (313453.5 seconds) shows a performance stage with palm tree background — that's a concert, not a boat.

Example 4: Qwen3.8-27B thinking (continued)

So the boat is only in Chongqing. But wait, frame 198's skyline: two red-lit towers — that's Chongqing's "Raffles City"? Actually, Chongqing's red-lit building might be the "Chongqing Raffles City" complex? Hmm, the red-lit twin towers in frame 198 look like the "Chongqing" skyline at night. Also frame 240 (359506.0 seconds) shows the same red-lit twin towers at night with the streamer posing with a woman in red — that's Chongqing.

So the answer: He took a boat on the river in Chongqing (a night cruise on the Yangtze/Jialing River). The question says "where did the streamer take a boat during the entire China trip?" Probably the answer: "In Chongqing, on the river (a night boat cruise)."

But maybe they want both: the ferry terminal and the river cruise are the same place. So the answer: Chongqing (river cruise).

Let me double-check frame 197: It shows a terminal with white beams, stairs, railings, and a sign with Chinese characters "Ferry..."? The sign reads "Ferry..."? Hard to read. It looks like a ferry boarding area. So that's the Chongqing ferry terminal.

Therefore evidence intervals: [259973.0, 261459.0]? Actually, the intervals should be ranges covering the boat scenes: 197: 259973.0; 198: 261459.0; 212: 294141.0; 213: 296369.5. I can give intervals like [259973.0, 261459.0], [294141.0, 296369.5]. But each interval must have end > start. I can use [259973.0, 262944.5]? Hmm, better to give small ranges around the frames: e.g., [259973.0, 261459.0] covers the terminal and boat. But the boat scene is at 261459.0; the next frame 199 at 262944.5 is a street. So the boat scene is roughly between 261459 and 262944.5. Similarly, 212-213 are between 294141 and 296369.5, and 214 at 298598 is a crowd.

Let me construct the evidence: [[259973.0, 262944.5], [294141.0, 298598.0]]. This covers the terminal + boat and the boat scenes.

But the question might expect the answer "Chongqing" as the place. Let me phrase it: "On the river in Chongqing (a night river cruise from the ferry terminal)."

Alternatively, the question might be interpreted as: "Where did he take a boat?" Answer: "In Chongqing, on the river (Jialing/Yangtze River cruise)."

I'll answer: He took a boat only in Chongqing, on a night river cruise (boarding at the ferry terminal).

Evidence: [[259973.0, 262944.5], [294141.0, 298598.0]].

Wait, but frame 212 shows an overcast daytime scene; frame 213 shows night. Both are on a boat. No problem.

Also, is there a boat in frame 220? No, that's a stage.

So the final answer: Chongqing (river cruise).

Let me double-check the red-lit twin towers in frame 198: That's probably the "Chongqing" skyline near the Raffles City complex? Actually, the red-lit towers look like the "Chongqing" skyline. Also frame 240 shows the same towers with the streamer and a woman — caption "a viewer has gifted 10 Tier 1 memberships" — that's a stream overlay, and the background is a still image of Chongqing.

OK.

So the JSON: {"answer": "Only in Chongqing – he boarded a ferry/river cruise boat on the city's river (night cruise).", "evidence": [[259973.0, 262944.5], [294141.0, 298598.0]]}

But the question says "where did the streamer take a boat during the entire China trip?" Maybe they expect the answer "Chongqing" as the single place. I'll include the detail.

Make sure the boundaries are decimals: 259973.0 etc. are fine.

Let me provide at most 8 intervals; I'll provide 2.

Final answer.

Example 4: Qwen3.8-27B prediction and score

Final answer: The streamer only took a boat in Chongqing, where he boarded a ferry/river cruise boat on the city's river (seen at the ferry terminal and on the open-deck night cruise with the lit skyline).
Rubric score: 2/5. Chongqing is correct, but 'only' explicitly excludes the required Hong Kong visit: a correct place mixed with an important contradiction.

Example 4: Codex (GPT 5.6-luna) tool-use trajectory

Selected calls in recorded order; intervening calls omitted and observations abridged.
1. `view_video_high_resolution(294000, 302000)`, 25 frames.
Tool observation: The high-resolution inspection identifies Hong Kong's Victoria Harbour using the skyline and ferry-pier views.

2. `view_video_high_resolution(258000, 267000)`, 20 frames.
Tool observation: The high-resolution inspection attributes the illuminated river-cruise setting to Chongqing.

3. `view_video_high_resolution(314000, 319000)`, 20 frames.
Tool observation: The tool interprets railings and waterfront views as a boat ride and identifies Shenzhen, which is then added to the answer.

Final answer: Chongqing, Hong Kong, and Shenzhen.
Rubric score: 3/5. Both required places are recovered, but Shenzhen is an erroneous addition.

Qwen attributes both boat segments to one city; Codex expands the answer set with an additional waterfront candidate.

D LIMITATIONS

(i) Single subject per scale. Each scale in our benchmark tracks one primary subject: one game character on the Day scale, one camera wearer on the Week scale, and one live streamer on the Month scale. This depth-over-breadth choice provides substantial temporal depth and ensures that solving the benchmark requires reasoning about a single agent's evolving state, rather than aggregating shallow clips across many unrelated subjects (which is already covered by existing multi-video benchmarks Peng et al. (2025); Zhu et al. (2025)). The three scales also draw from disjoint subjects in disjoint domains (gamer, wearer, streamer), so a model cannot solve all three by overfitting to one persona. We acknowledge that any per-scale claim still rests on a single subject's stream, and whether the failure modes we observe generalize to additional subjects at the same scale is an empirical question that requires further longitudinal recordings. Recruiting and clue-grounding multi-week streams from $N > 1$ subjects scales annotator-months roughly linearly in N, which is difficult to justify before the community has begun saturating the benchmark in its current form. Extending the per-scale subject pool is therefore left to future versions of the benchmark.

(ii) Month-scale upper bound on T_{span}. We cap the physical temporal span at the month scale. While this is sufficient to expose the central failure modes documented in this paper, genuinely lifelong understanding will ultimately require maintaining coherent state across much longer horizons: identity drift, recurring routines, and major life events that play out over years rather than weeks. Extending MM-Lifelong toward such horizons is an open direction left to future versions of the benchmark.

(iii) The benchmark does not explicitly probe unobserved-interval reasoning. A core property of the lifelong setting is that events occurring during unobserved gaps (sleep cycles, off-camera periods) still influence the observed timeline. Our current QA pairs test whether a model can connect observed events across such gaps via multi-hop reasoning, but the benchmark does not contain questions or annotations that directly probe the content of an unobserved interval itself. As a result, a model's ability to infer what happened during a gap is only indirectly assessed through whether its answers are consistent with the observed evidence on either side. Building probes that directly test gap-content inference requires both new task formulations (e.g., counterfactual queries about what happened during a gap) and ways of validating ground truth without access to the unrecorded footage. We see this as the most natural direction for follow-up work building on the released benchmark; finer-grained temporal analysis such as action segmentation Huang et al. (2020) and egocentric attention modeling Huang et al. (2018) may also be useful for capturing within-segment dynamics that complement inter-gap inference.

E ETHICS STATEMENT

Data subjects and consent. MM-Lifelong contains three data subjects: a synthetic game protagonist (no privacy concern), a camera wearer whose footage is sourced from EgoLife Yang et al. (2025b) under their original consent protocol, and a public live-streamer whose broadcasts are publicly accessible on the originating platform. As detailed in Section 4.6, we do *not* redistribute any raw video file; users obtain footage from the original platform under that platform's terms of use. This design keeps PII-bearing video content under the control of its origin platform, eliminating the need for additional face or voice processing on our side.

Third-party individuals and anonymization. Individuals other than the primary subject may incidentally appear in the videos. Because we do not redistribute visual content, we instead anonymize them at the *annotation* layer: all named individuals in our QA pairs and metadata are replaced by role tags (e.g., "[The Streamer]", "[Friend A]", "[Passerby]") before release. No real names, addresses, phone numbers, or identification documents appear in any released annotation.

Annotator compensation and IRB. Eight annotators, recruited as part-time research assistants from the authors' institution, were compensated at approximately 50/hour, well above the local minimum wage. Per institutional policy, this work does not require IRB review because annotators interact only with publicly available video content and do not involve any human-subject experimentation.

Potential misuse and safeguards. Long-horizon multimodal memory datasets carry inherent risks of enabling surveillance, stalking, or cross-session re-identification of named individuals. We adopt the following safeguards to limit such misuse: (i) we do *not* redistribute raw video files, eliminating direct access to visual biometric signals through our release; (ii) all named individuals in our annotations are anonymized via role tags, as described above; (iii) our Acceptable Use Policy, posted alongside the benchmark, explicitly prohibits downstream use for surveillance, stalking, or commercial re-identification of any individual; (iv) we release only text-level annotations and timestamps, with no face, voice, or other biometric embeddings; (v) we accept takedown and removal requests through HuggingFace Discussions and commit to responding within a reasonable timeframe.

F Compute Resources

Locally hosted MLLMs and agent backbones were evaluated on shared NVIDIA A100 80GB and H100 80GB GPU clusters. Closed-source systems were accessed through their official APIs. End-to-end cost varies substantially with context length, model size, serving framework, and the number of retrieval or tool calls made by an agent. We therefore report the hardware pool and evaluation settings, while avoiding a single runtime estimate that would not be comparable across system families.