VIDEOLIVE: BENCHMARKING STATEFUL REAL-TIME AUDIO-VISUAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have demonstrated strong video understanding capabilities, yet they are mostly evaluated on answers given at isolated moments. A correct answer at one moment does not show whether a model can keep taking correct and timely interactive actions as the scene evolves. Real-time interaction therefore requires a model to generate answers and to decide when to **ANSWER**, when to **WAIT**, and when to **REPAIR** an answer that has become stale. We introduce VIDEOLIVE, a *stateful streaming* benchmark that evaluates real-time audio-visual interaction through a unified causal protocol. VIDEOLIVE continuously requires an agentic MLLM to make decisions from the accumulated evidence available up to the current moment, autonomously choosing whether to answer, wait, or repair, and evaluates its capability across the full lifecycle of an answer, from forming an initial answer through retention to update. We evaluate 29 model configurations across four lifecycle types: Retention, Replacement, Withdrawal, and Formation. The results show that forming a correct initial answer does not guarantee timely subsequent updates: the best configuration completes **23.50%** of answer lifecycles, and across all 29 configurations only **21.11%** of valid initial answers were revised correctly and in time after the state changed. Even when an independent probe identifies the new activity, **44.26%** of the public answers from incremental interfaces remain incorrect two seconds later, revealing a substantial gap between content understanding and continuous interactive decision-making. We further propose **SPOT**, a post-training method that guides MLLMs to learn when to publish, retain, or update an answer by supervising their replies conditioned on the displayed answer; it improves MMDuet2 on every VIDEOLIVE metric, raising LCS from 3.25 to 8.50 and Overall from 15.44 to 25.72, and improves response timing on three held-out streaming benchmarks. We hope VIDEOLIVE helps move research toward reliable real-time interaction.

1 INTRODUCTION

Multimodal large language models (MLLMs) support increasingly rich questions about actions, events, and their relationships across video sequences (Mangalam et al., 2023; Fu et al., 2025). Recent streaming systems extend this interaction to continuously arriving video, allowing assistants to follow activities as they unfold (Chen et al., 2024; Qian et al., 2025), which opens opportunities for assistants that accompany users through everyday tasks, where each response is part of an ongoing interaction. In such settings an assistant must respond from the evidence that has arrived so far, and the user depends on the response while the task is still underway.

In real-time interaction, a response stays visible after it is generated while the scene keeps evolving, so a correct answer can become outdated. In Figure 1, "Pouring water" correctly describes someone preparing a drink, then becomes stale when the person starts stirring, yet remains visible until the assistant revises it. Existing streaming benchmarks evaluate changing answers, response timing, and multi-turn interaction (Lin et al., 2024; Huang et al., 2025; Niu et al., 2025; Wang et al., 2025c; Xun et al., 2025; Shi et al., 2026), but those in Table 1 score each response when it is produced and do not check whether a displayed answer remains correct afterward or is repaired in time. Each stage of that lifecycle is a decision: answering too early publishes a guess, revising without cause discards a correct answer, and repairing requires recognizing that the displayed answer no longer holds.

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.

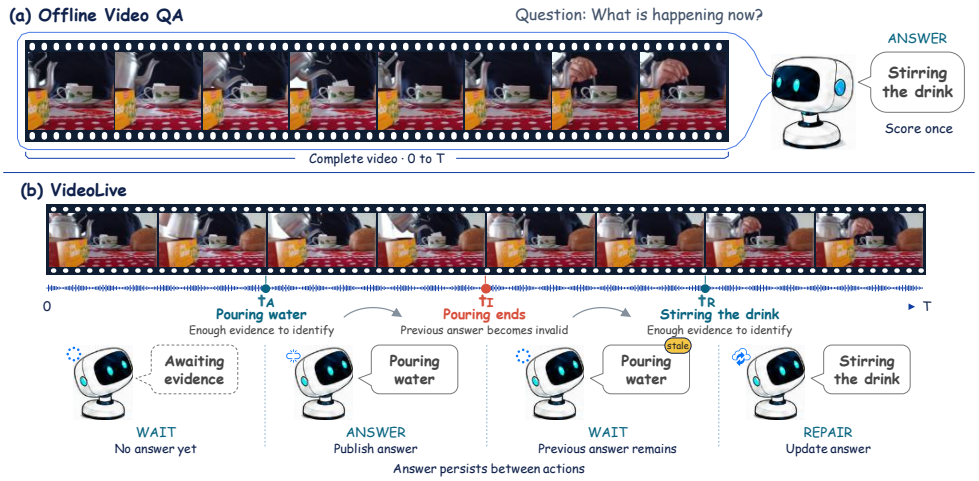


Figure 1: From complete-video answers to persistent answer states. Frames from one Perception Test video show a person pouring water and then stirring the drink; both panels ask "What is happening now?" **a**, Offline video question answering processes the complete clip and evaluates a single response. **b**, VIDEOLIVE evaluates when the agent answers and how its public answer evolves as evidence arrives: the agent publishes "Pouring water," leaves it visible during WAIT, and repairs it to "Stirring the drink." The answer becomes stale when pouring ends and remains exposed.

We introduce VIDEOLIVE, a streaming benchmark for evaluating the lifecycle of a *persistent public answer*, which remains visible until explicitly updated. As audio-visual media arrive, agents choose every 500 ms to WAIT, ANSWER, or REPAIR using the evidence available so far, under one causal protocol for incremental interfaces and causal-prefix adapters. Lifecycle Completion Score (LCS) measures a valid initial answer followed by retention or timely repair; Valid Answer Time (VAT) measures the proportion of supported time with a correct public answer. Phase-aligned probes separately score Understanding and Temporal decision.

We construct Core-400 from official temporal annotations in Perception Test and CaptainCook4D (Patraucean et al., 2023; Peddi et al., 2024). Each of its 400 distinct videos carries a current-activity question and a reference-answer timeline, with NONE when no activity in the question's scope occurs. The 400 trajectories divide equally among Retention, which requires keeping a correct answer, Replacement, which requires repairing it when a new activity begins, Withdrawal, which requires publishing NONE once the activity ends, and Formation, which requires answering NONE before updating to the activity once it starts.

Across 29 configurations, the highest LCS is 23.50, and only 21.11% of valid initial answers in Changed trajectories are repaired correctly within two seconds. Recognizing an activity does not imply maintaining an answer: pooled over the 16 incremental interfaces, 44.26% of public answers shown two seconds after a change remain incorrect even when an independent probe identifies the new activity. Updating away from an active answer is harder than reporting a newly started activity: pooled over all configurations, 17.40% of valid initial answers are repaired in time in Replacement and Withdrawal, against 26.63% in Formation.

We further propose SPOT (State-Paired Online Tuning), a post-training method that pairs the same evidence with different displayed answers and supervises the agent's native reply to each, so that the agent learns to read its own public answer before choosing silence, an answer, or a repair. Applied to MMDuet2, it raises LCS from 3.25 to 8.50 and Overall from 15.44 to 25.72, and improves response timing on the held-out StreamingBench, RIVER, and OVO-Bench.

Our contributions are as follows:

- We introduce **VIDEOLIVE and an evaluation framework.** Core-400 balances 400 videos across Retention, Replacement, Withdrawal, and Formation, over which MLLMs WAIT, ANSWER, or REPAIR under one causal streaming protocol.

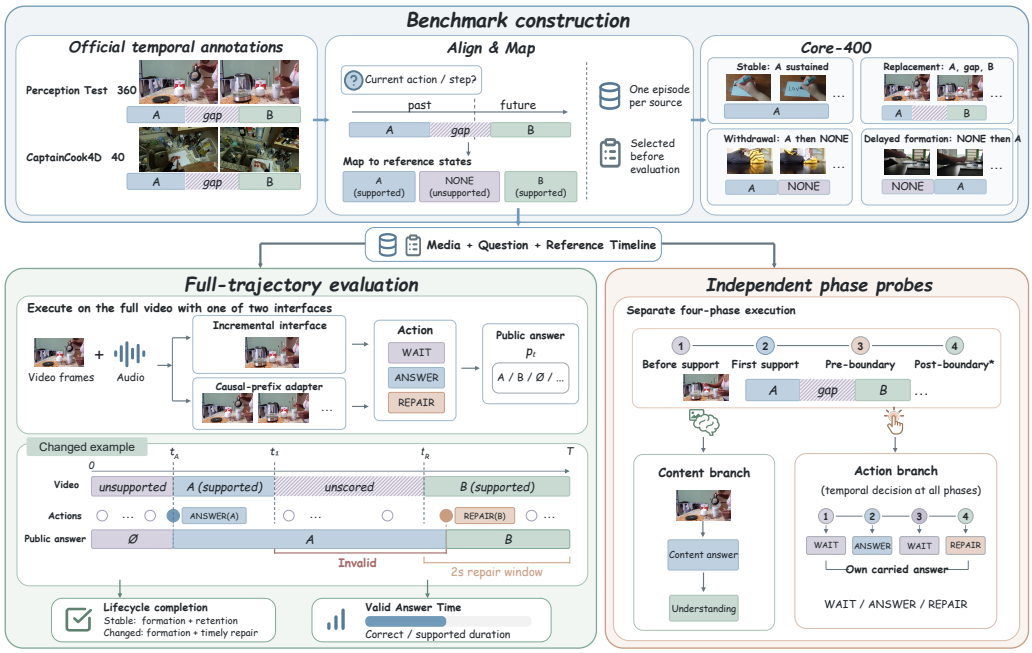


Figure 2: VIDEOLIVE overview. Causally delivered media elicit WAIT, ANSWER, or REPAIR, which determine the public answer visible over time. Source annotations define an invalidation boundary t_I and a post-change support boundary t_R; the repair window starts at t_R. Full trajectories yield lifecycle endpoints and VAT. The Changed example uses pouring and stirring as A and B.

- We find that **current MLLMs struggle both to form correct answers and to keep them valid over time.** Across 29 configurations, recognizing a new activity rarely leads to a timely repair, and the displayed answer itself shapes the next decision.
- We propose **SPOT,** a post-training method in which supervision is conditioned on the displayed answer. It improves every VIDEOLIVE metric on MMDuet2 and improves response timing on three held-out streaming benchmarks.

2 RELATED WORK

Video understanding over time. Video benchmarks assess understanding across durations and temporal contexts, from long-context questions in EgoSchema to videos of varied lengths and modalities in Video-MME (Mangalam et al., 2023; Fu et al., 2025). Online benchmarks tie questions to the evidence available at a particular moment: OVBench tests perception and memory from different temporal perspectives (Huang et al., 2025), and RTV-Bench revisits questions as their answers change with the scene (Xun et al., 2025). The same stream can thus support different answers over time, which determines when an assistant must update its response.

Response timing and interaction. Streaming interaction also requires choosing when to respond. OVO-Bench includes tasks that require waiting for sufficient evidence, alongside past-event retrieval and current perception (Niu et al., 2025). RIVER evaluates retrospective memory, live perception, and proactive responses with content and temporal alignment scores (Shi et al., 2026). StreamingBench includes timestamped questions, sequential dialogue, and proactive output (Lin et al., 2024). Omni-iMMI brings together multi-turn audio-visual reasoning, proactive alerts, and turn-taking (Wang et al., 2025c). ProactiveVideoQA scores multi-turn proactive replies with PAUC, which accumulates reply quality over each reference window (Wang et al., 2025b); its replies accumulate within a fixed-answer window, so an earlier reply is never scored as outdated. VIDEOLIVE instead follows each published answer until it is updated, evaluating its timing, retention, repair, and valid duration (Table 1).

Streaming systems and evaluation protocols. Streaming systems retain context as new media arrive and decide when to generate a response. VideoLLM-online learns frame-aligned dialogue (Chen

Table 1: Task coverage and evaluation of streaming video benchmarks. **Changing answers**: the same question can have different reference answers as the scene evolves. **Multi-response queries**: one user request supports successive responses as the video unfolds. Timing includes response-time and readiness scores. $\geq$ denotes query-count lower bounds with additional polling requests. For VIDEOLIVE, the count is the number of decision points in one full Core-400 run.

		Task coverage		Response-level		Answer validity over time		
Benchmark	**# Response requests**	**Changing answers**	**Multi-response queries**	**Content**	**Timing**	**Answer retention**	**Answer repair**	**Valid answer time**
StreamingBench	4,352–4,976	✓	×	✓	✓	×	×	×
OVBench	7,005	✓	×	✓	×	×	×	×
OVO-Bench	3,035	✓	✓	✓	✓	×	×	×
OmniMMI	≥2,290	✓	✓	✓	✓	×	×	×
RTV-Bench	4,608	✓	×	✓	×	×	×	×
RIVER	≥4,401	✓	✓	✓	✓	×	×	×
VIDEOLIVE (ours)	**10,343**	✓	✓	✓	✓	✓	✓	✓

et al., 2024), Dispider disentangles perception, decision, and reaction with asynchronous generation (Qian et al., 2025), and MMDuet2 learns response timing through multi-turn reinforcement learning (Wang et al., 2025a). StreamingEval standardizes how such systems are compared, specifying causal access, memory budgets, accuracy, throughput, first-token latency, and memory use (Tang et al., 2026). VIDEOLIVE places both interface types under shared clock and update rules.

3 VIDEOLIVE

VIDEOLIVE converts temporal annotations into reference-answer timelines and evaluates the answer visible to users as media arrive as shown in Figure 2. This formulation captures whether agents publish answers, maintain answers, or revise them correctly, when evidence becomes obvious.

3.1 MANAGING AN ANSWER AS EVIDENCE CHANGES

The *public answer state* p_t starts empty ($\varnothing$) and persists between decisions. ANSWER(y) publishes the first answer, REPAIR(y) replaces it, and WAIT preserves it. NONE is a published answer indicating that no activity in the question's declared scope is active.

Each trajectory follows one of four lifecycle types: Retention (A-to-A), Replacement (A-to-B), Withdrawal (A-to-NONE), and Formation (NONE-to-A); we refer to the last three as Changed trajectories. The initial reference is supported from t_A. In Retention trajectories, its support lasts until episode end T, and retention is evaluated from the anchor t_H, two seconds before that support ends. In Changed trajectories, the initial support ends at t_I, and the next reference becomes supported at $t_R \geq t_I$. Scoring covers $[t_A, t_I)$ and $[t_R, T]$, with the repair window starting at t_R (Appendix A).

3.2 CORE-400 CONSTRUCTION

Core-400 uses official action intervals from Perception Test and procedure-step intervals from CaptainCook4D (Patraucean et al., 2023; Peddi et al., 2024). Each trajectory pairs a current-activity question with a video segment and reference-answer timeline. The question specifies the activity scope: the 63 source action classes for Perception Test or the complete step vocabulary of the named recipe for CaptainCook4D. We determine reference support from the labels active in the complete source annotations: an active reference requires continuous, unique label coverage throughout its support interval, and NONE requires an empty active-label set. These intervals define the four lifecycle types. The question is posed at time zero, and a local timeline is set for media.

Deterministic selection yields 400 distinct source videos, with 100 trajectories per type. Each type contains 90 Perception Test and 10 CaptainCook4D videos. Together, these sources provide 84.59 minutes of video, with median episode lengths of 6.25 and 48.29 seconds, respectively (Figure 3). Appendix C details selection and coverage.

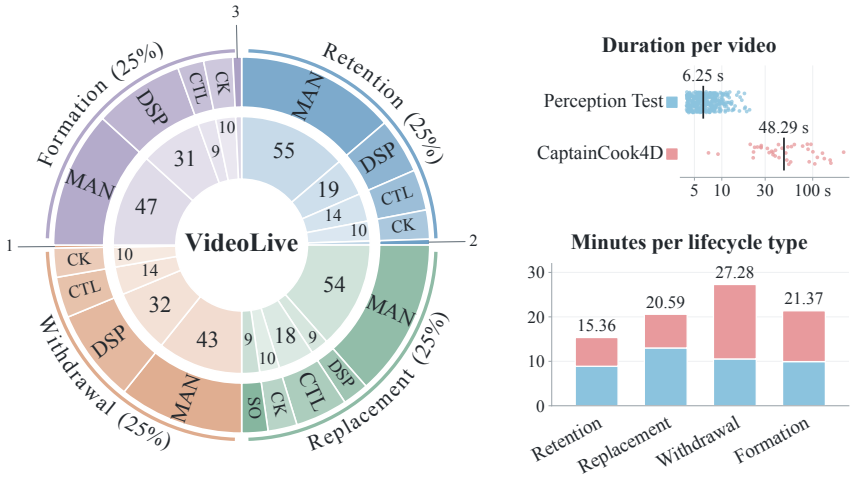


Figure 3: Core-400 composition. Core-400 contains 400 videos (84.59 minutes), 100 (25%) per lifecycle type. Left: each type is subdivided into five content families, with video counts inside and family codes outside; callouts mark small sectors. MAN, manipulation; DSP, display or gestures; CTL, access or controls; CK, cooking; SO, simulated or other. Right, top: duration of each video by source on a log scale, with ticks at the medians. Right, bottom: total minutes per lifecycle type, split by source. Perception Test supplies 90% of the videos and CaptainCook4D the longer 10%; each contributes about half of the total duration.

3.3 Unified Causal Streaming Protocol

Each trajectory is driven by a media clock: video arrives at 2 FPS, with aligned 500 ms audio chunks for audio-capable systems and decision opportunities every 500 ms, and at time t only media timestamped at or before t are available. Systems interact through one of two interface classes: incremental interfaces ingest new media and retain their native session, cache, or memory, while causal-prefix adapters receive the growing video prefix, question, timestamp, and current public answer at each decision (Appendix B). The evaluator maps responses to ANSWER, REPAIR, or WAIT, including silence as WAIT, and applies each action to p_t. Lifecycle scoring uses media time.

3.4 Evaluation Metrics and Setting

Initial answers and completion. Initial-answer success is the percentage of trajectories whose first ANSWER is semantically correct within $[t_A, T]$ for Retention or $[t_A, t_I)$ for Changed trajectories.

Retention completion requires a valid initial answer by the retention anchor t_H and a correct public answer over the following two seconds. Changed completion requires a valid initial answer followed by a correct REPAIR within $[t_R, t_R + 2\,\text{s}]$, so NONE-to-A requires ANSWER(NONE) followed by REPAIR(A). The Lifecycle Completion Score (LCS) is the percentage of trajectories meeting their completion criterion, and Adaptivity is this percentage over Changed trajectories. Stability is the percentage of Retention trajectories whose public answer is correct at t_H and stays correct for two seconds, whether or not the first answer was valid (Appendix D).

Valid Answer Time (VAT). VAT is the percentage of supported duration with a semantically correct public answer, averaged equally over trajectories. VAT complements LCS: it credits correct repairs that arrive after the two-second window, and it continues to score the public answer after a timely repair. Thus, VAT distinguishes reaching a correct answer from maintaining a valid public answer.

Understanding and temporal decision. Understanding measures semantic accuracy at three supported probe phases, querying the current activity with the previous public answer omitted. Temporal decision measures action accuracy at up to four phases, carrying the answer between decisions. The target is WAIT before reference support. During support, an empty state requires ANSWER, a semantically correct published answer requires WAIT, and an incorrect published answer requires REPAIR. An ANSWER or REPAIR is correct only if the resulting public answer is also semantically correct. Both scores average applicable phases within each trajectory (Appendices D and E).

Table 2: VideoLive performance (0–100; higher is better). LCS: Lifecycle Completion Score; VAT: Valid Answer Time (%); Decision: Temporal decision; Overall averages LCS, VAT, Understanding, and Decision. Always choosing WAIT scores 22.00 on Decision; always answering NONE scores 25.00 on Understanding. Bold and underlining mark the best and second-best scores within each interface group; $^{\text{close}}$ marks closed-source models, and the Qwen Omni-Flash models use their Realtime versions. + SPOT (ours) is MMDuet2 after post-training (Section 5).

	Full trajectories					Phase probes		
System	LCS	Initial answer	Stability	Adaptivity	VAT	Understanding	Decision	**Overall**
Human	19.00	40.00	28.00	22.67	43.04	40.67	42.51	36.30
Incremental streaming interfaces								
MiniCPM-o	3.50	13.25	16.00	0.00	14.62	23.42	19.35	15.22
OmniLive	0.00	1.00	0.00	0.00	0.37	16.25	22.06	9.67
JoyAI-VL-Interaction	4.00	26.00	17.00	0.00	11.67	26.25	26.21	17.03
Aero Realtime	4.00	14.75	16.00	0.00	11.17	19.25	10.98	11.35
StreamOPD-4B	7.50	22.75	31.00	0.33	23.72	24.25	23.71	19.79
ThinkStream-3B	1.75	7.75	7.00	0.00	4.05	4.33	23.38	8.38
VST-7B	4.75	11.50	19.00	0.00	11.17	12.25	6.08	8.56
StreamingVLM	0.25	1.25	1.00	0.00	0.28	0.42	22.31	5.81
InfiniteVL	0.00	0.00	0.00	0.00	0.00	2.83	21.92	6.19
StreamForest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mage-VL (streaming)	0.00	0.00	0.00	0.00	0.00	27.58	22.00	12.40
LiveStarPro	5.50	24.50	21.00	3.00	18.90	26.83	19.88	17.78
MOSS-VL-Realtime	2.00	7.50	8.00	0.00	6.57	5.50	7.08	5.29
Qwen3.5-Omni-Flash$^{\text{close}}$	10.78	27.57	**47.00**	0.67	24.16	24.90	25.96	21.45
Qwen3.8-Omni-Flash$^{\text{close}}$	**14.79**	**38.10**	38.00	**12.71**	**36.82**	**37.26**	**42.34**	**32.80**
MMDuet2	3.25	12.75	13.00	0.00	8.95	21.33	28.23	15.44
+ SPOT (ours)	8.50	22.25	27.00	2.67	22.60	31.00	40.77	25.72
Causal-prefix streaming adapters								
Gemma 4 E2B-it	3.00	6.25	12.00	0.00	5.52	8.33	21.50	9.59
Gemma 4 E4B-it	2.00	7.75	9.00	0.00	7.59	9.33	24.54	10.87
Gemma 4 12B-it	3.25	15.00	17.00	1.00	15.80	21.92	31.77	18.18
Qwen3.6-27B	3.75	29.00	9.00	3.67	21.96	35.25	18.92	19.97
Qwen3.8-27B	17.00	35.00	**58.00**	6.00	29.76	31.17	37.62	28.89
Molmo2-8B	4.00	11.50	18.00	0.00	10.70	19.08	19.08	13.22
Mage-VL	4.00	26.00	16.00	0.00	16.86	22.17	17.21	15.06
MOSS-VL-Instruct	6.25	18.00	25.00	0.00	15.76	32.08	16.38	17.62
GPT-5.6 Luna$^{\text{close}}$	13.25	29.00	34.00	13.00	35.12	34.50	23.58	26.61
GPT-5.6 Terra$^{\text{close}}$	20.75	38.75	45.00	19.00	41.22	38.00	30.00	32.49
GPT-6 Luna$^{\text{close}}$	9.25	31.75	37.00	7.67	29.49	35.83	20.21	23.70
GPT-6 Sol$^{\text{close}}$	20.75	36.75	48.00	20.00	42.14	**40.67**	34.27	34.46
Gemini 3.8 Flash$^{\text{close}}$	**23.50**	**40.50**	24.00	**26.00**	**45.04**	39.92	**37.69**	**36.54**

Overall score. Overall is the arithmetic mean of LCS, VAT, Understanding, and Temporal decision.

We evaluate 16 incremental interfaces and 13 causal-prefix adapters, each with its supported modalities and state mechanism, on full trajectories and independent phase probes. Qwen Realtime models keep audio-video sessions; GPT models receive up to 50 frames per request, and Gemini 3.8 Flash receives prefixes processed at 1 FPS. Every configuration sees all 400 videos. Qwen3.8-27B judged open-text answers with deterministic decoding, and confidence intervals use 10,000 source-video bootstrap resamples stratified by lifecycle type. Fourteen participants deciding at 500 ms pauses reached 19.00% lifecycle completion and 43.04% VAT (Appendices E, L, E.6, G).

4 Experiments

4.1 Overall Lifecycle Performance

Completing an answer lifecycle remained difficult across both interfaces (Table 2). Among causal-prefix adapters, Gemini 3.8 Flash reached LCS 23.50, followed by GPT-6 Sol and GPT-5.6 Terra at 20.75. Qwen3.8-Omni-Flash-Realtime led the incremental interfaces, but its 38.10% initial-answer success fell to LCS 14.79. Retention and updating also showed distinct difficulty patterns: Qwen3.5-

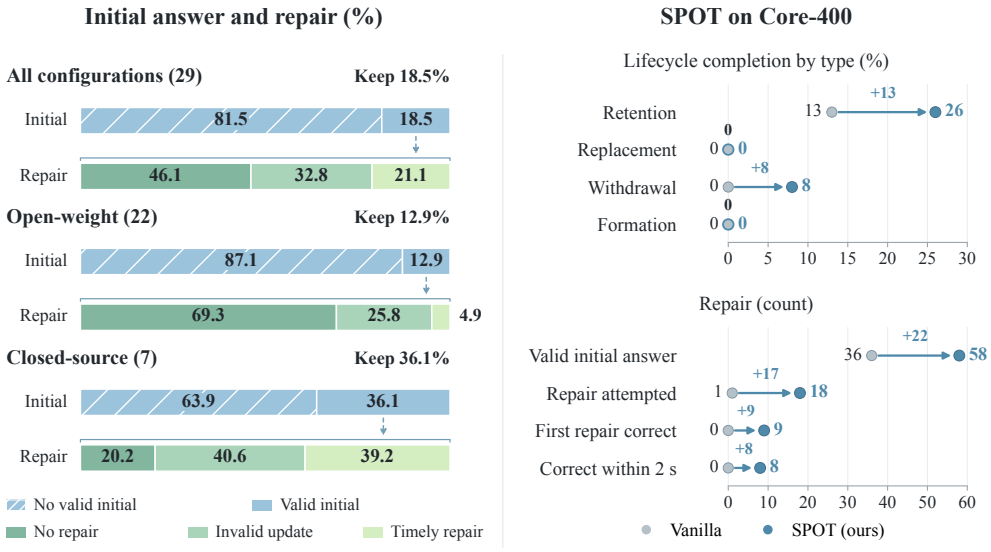


Figure 4: Left: initial-answer and repair outcomes on Changed trajectories, for all 29 configurations, the 22 open-weight, and the 7 closed-source systems; in each row the top bar gives initial-answer outcomes and the bottom bar the repair outcomes of those valid initial answers. Right: SPOT on Core-400, by lifecycle type and by repair stage. Appendix Figure 7 reports per-system breakdowns.

Omni-Flash-Realtime scored 47.00 in Stability but 0.67 in Adaptivity, StreamOPD-4B 31.00 and 0.33, and Qwen3.8-Omni-Flash-Realtime 38.00 and 12.71.

4.2 Initial Answers and Post-change Updating

Among Changed trajectories with a valid initial answer, Qwen3.8-Omni-Flash-Realtime repaired correctly within two seconds in 33.33%, LiveStarPro in 10.84%, Qwen3.5-Omni-Flash-Realtime in 3.17%, and StreamOPD-4B in 1.64%, while Gemini 3.8 Flash reached 58.65%.

Figure 4 extends this to all configurations. Pooling all 29 yields 1,606 valid initial answers among 8,698 Changed-trajectory evaluations: 741 issued no qualifying repair, 526 produced incorrect or premature updates, and 339 repaired correctly within two seconds on 21.11%. Updating away from an active answer was harder: 17.40% of valid initial answers were repaired in time in Replacement and Withdrawal, against 26.63% in Formation. Both failure modes appear on one clip in Figure 5.

Recognizing the new activity did not ensure that the public answer followed. We paired each Changed trajectory's post-boundary content probe with its public answer two seconds after t_R. On the 16 incremental interfaces the probe identified the updated reference in 389 trajectories, yet 104 of the 235 that published an answer (44.26%; 95% CI $[37.78, 51.43]$) showed an incorrect one (Appendix F.7). The carried answer also shapes the next decision: supplying the correct pre-transition answer raised Qwen3.8-27B's Retention accuracy from 63% to 93% but lowered that of Qwen3.6-27B on Withdrawal from 49% to 21% (Table 9).

These losses are specific to holding an answer over time rather than to understanding the clips. Querying the same 400 videos only at their endpoints gave Qwen3.6-27B 41% accuracy, for both the complete video and a terminal prefix, while its sequential interaction completed 3.75% of lifecycles and 21.96% VAT. What varies between the two protocols is not whether the answer is determined, but whether it is published when publishing is due.

5 SPOT: Learning to Act on the Public Answer

The correct action depends on both the evidence and the displayed answer: the same frame calls for WAIT when that answer is correct, ANSWER when nothing is displayed, and REPAIR when it is

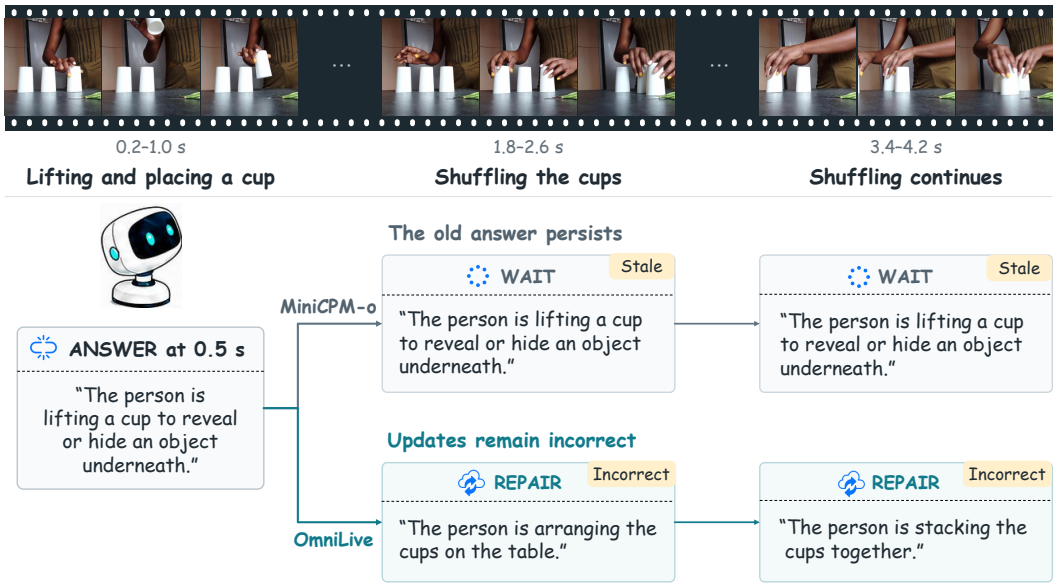


Figure 5: Public-answer maintenance on a cup-shuffling clip. MiniCPM-o retains a stale answer, while OmniLive repairs its answer twice but remains incorrect.

stale. Supervision along a reference-following history rarely shows the last two cases, because the displayed answer is usually correct: 86.0% of its targets are WAIT and 0.6% ANSWER (Appendix Table 29). SPOT (State-Paired Online Tuning) therefore keeps the standard reply cross-entropy but conditions each supervised reply on the displayed answer.

State-conditioned targets. Let r_t be the reference at decision time t, with $r_t = \perp$ outside reference support and distinct from a supported NONE reference, and let p be the displayed answer, with $p = \varnothing$ for an empty display. The target reply is

$$y^\star(r_t, p) = \begin{cases} \texttt{NO REPLY}, & r_t = \perp \text{ or } E(p, r_t) = 1, \\ r_t, & \text{otherwise}, \end{cases} \tag{1}$$

where $E(p, r_t) = 1$ when the labels are identical, with NONE labels equivalent and neither equated with an empty display. MMDuet2 receives p in its prompt as the current public answer, in training and at evaluation alike. Silence maps to WAIT, and a published reply to ANSWER or REPAIR depending on whether an answer is displayed, so $y^\star$ reproduces the Temporal decision targets (Section 3.4) without a separate action head.

State pairing. At each lifecycle anchor, the display a reference-following agent would show, $p_t^\star$, is paired with one alternative $\tilde{p} \neq p_t^\star$ labeled $y^\star(r_t, \tilde{p})$: an incorrect answer from the question's label vocabulary, an empty display, NONE, the preceding reference, or the correct answer. The two instances share the question, time, and history, so they differ only in the displayed answer. With the same 21,048 supervision instances per epoch as reference-following supervision, publication targets rise from 14.0% to 41.5%.

Native streaming replay. Training runs through MMDuet2's own incremental interface on the evaluation clock: 2 FPS, one decision per frame, and a persistent key–value cache. Paired instances are scored without committing their replies; the reference-following reply is then committed once, so the model conditions on the history a correct agent would have produced. We train once with LoRA for five epochs on 2,276 trajectories and evaluate the final checkpoint (Appendix N).

Results. SPOT improved MMDuet2 on every metric in Table 2, raising LCS from 3.25 to 8.50, above every open-weight incremental interface, VAT from 8.95% to 22.60%, Understanding from 21.33 to 31.00, and Initial-answer success from 12.75% to 22.25%; Retention completion rose from 13 to 26 and repair attempts from 1 to 18. On the held-out streaming benchmarks of Table 3 it

Table 3: SPOT and a plain SFT control on held-out streaming benchmarks; question or evaluation-point counts in parentheses. Early: queries answered more than 2 s early (% of queries), Cov means content coverage, CRR and SSR report No-F1 score, see Appendix N.

	StreamingBench Proactive (250)			RIVER (551)	OVO-Bench	
	Timing F1	Precision	Early↓	Cov	CRR (240)	SSR (629)
MMDuet2	35.97	38.71	50.80	37.57	48.28	47.50
+ Ordinary SFT	42.56	50.27	40.76	40.75	47.57	46.23
+ SPOT	**43.50**	**53.18**	**26.80**	**64.79**	**53.06**	**50.16**

improved response timing and reduced premature responses: timing F1 rose from 35.97 to 43.50, and premature responses fell from 50.80% to 26.80% of queries.

A plain SFT control trained on the same trajectories for the same five epochs, repeating every natural state twice to reach the same 21,048 supervision instances, also transfers, yet SPOT leads it on all six metrics of Table 3. With instance counts matched, the two differ in what is supervised: SPOT fits every natural state together with an alternative displayed answer, so the policy is trained on the answer states it meets in deployment rather than on the teacher's correct history alone, and that advantage holds on benchmarks whose clips never appear in training.

6 Discussion

Answer maintenance as a capability. Recognition of the new activity preceded a correct public answer in fewer than half of the cases, and missing repairs outnumbered incorrect ones (Section 4.2). Figure 6 illustrates this mismatch for MMDuet2: six clips switch to positive preference, yet all six choose WAIT and have incorrect public answers at 0, 0.5, and 1 s. Candidate preference therefore does not guarantee a correct public update. SPOT, which treats the displayed answer as explicit state, improves initial answering and retention.

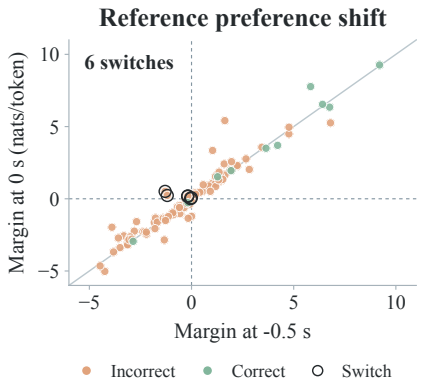


Figure 6: Result of MMDuet2 on 100 Replacement clips. Axes compare updated-minus-old mean token log-likelihood margins at -0.5 and 0 s. Colors indicate public-answer correctness at 0 s.

The displayed answer as an input. The intervention in Section 4.2 shows that this cuts both ways. A stale answer is both a symptom of a missed update and a state that enters the context determining the next decision, so a mistimed WAIT tends to persist instead of resolving on its own. Because the system sees the display it produced rather than the evidence it ignored, it has little reason to reopen a decision it has already made. A system that never inspects its own display therefore cannot distinguish a correct answer from one that has quietly expired, and treats both as equally good grounds.

What lifecycle scoring adds. Static video question answering reports whether a system can reach a correct answer; it does not report whether that answer survives contact with the next two seconds of media. The four lifecycle types separate failing to answer from failing to withdraw an expired answer, and VAT weights each correct answer by the time it was available.

7 Conclusion

We propose VideoLive, which evaluates how real-time audio-visual MLLMs form, retain, and repair a persistent public answer under causal input. Under our setting, the strongest configuration completes 23.50% of lifecycles, and only 21.11% of valid initial answers are repaired in time once the scene changes. SPOT, which conditions supervision on the displayed answer, raises MMDuet2's LCS from 3.25 to 8.50 and Overall from 15.44 to 25.72, and extends to other streaming benchmarks.

AI USE STATEMENT

AI tools assisted with literature discovery, framework and experiment design, implementation, data processing, analysis and interpretation, figure preparation, and manuscript drafting and editing. Semantic evaluation used Qwen3.8-27B, with GLM-5.3 comparisons on a stratified sample. The authors reviewed the AI-assisted work and take responsibility for the manuscript, methods, and results.

REPRODUCIBILITY STATEMENT

Appendix C describes Core-400 construction, temporal alignment, and causal media delivery. Appendices B and E specify model interfaces, inference settings, and semantic evaluation. Metric definitions and uncertainty estimation appear in Appendix D, and context interventions and decoder readouts are detailed in Appendix H. Appendix M describes the evaluation materials and their availability.

REFERENCES

Alibaba Cloud. Qwen-Omni-Realtime model access. Alibaba Cloud Model Studio documentation, 2026. URL https://help.aliyun.com/zh/model-studio/omni-realtime-model-access.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_VideoLLM-online_Online_Video_Large_Language_Model_for_Streaming_Video_CVPR_2024_paper.html.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026. URL https://arxiv.org/abs/2601.10611.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026. URL https://arxiv.org/abs/2604.27393.

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

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

Yiran Guan, Liang Yin, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Video streaming thinking: VideoLLMs can watch and think simultaneously. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2603.12262.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: OVBench and VideoChat-Online. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3328–3338, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Huang_Online_Video_Understanding_OVBench_and_VideoChat-Online_CVPR_2025_paper.html.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024. URL https://arxiv.org/abs/2411.03628.

Zikang Liu, Longteng Guo, Handong Li, Ru Zhen, Xingjian He, Ruyi Ji, Xiaoming Ren, Yanhao Zhang, Haonan Lu, and Jing Liu. Thinking in streaming video. *arXiv preprint arXiv:2603.12938*, 2026. URL https://arxiv.org/abs/2603.12938.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2308.09126.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How far is your video-LLMs from real-world online video understanding? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18902–18913, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Niu_OVO-Bench_How_Far_is_Your_Video-LLMs_from_Real-World_Online_Video_CVPR_2025_paper.html.

OpenAI. GPT-5.6 system card. OpenAI Deployment Safety Hub, 2026a. URL https://deploymentsafety.openai.com/gpt-5-6-preview.

OpenAI. GPT-6 system card. OpenAI Deployment Safety Hub, 2026b. URL https://deploymentsafety.openai.com/gpt-6-astra.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1852. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8540fba4abdc7f9f7a7b1cc6cd60e409-Abstract-Datasets_and_Benchmarks.html.

Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, Nicholas Ruozzi, Yu Xiang, and Vibhav Gogate. CaptainCook4D: A dataset for understanding errors in procedural activities. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4307. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f4a04396c2ed1342a5d8d05e94cb6101-Abstract-Datasets_and_Benchmarks_Track.html.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video LLMs with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24045–24055, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Qian_Dispider_Enabling_Video_LLMs_with_Active_Real-Time_Interaction_via_Disentangled_CVPR_2025_paper.html.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026a. URL https://arxiv.org/abs/2604.15804.

Qwen Team. Qwen3.6-27B. Hugging Face model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.6-27B.

Qwen Team. Qwen3.8-27B. Hugging Face model card, 2026c. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Yansong Shi, Qingsong Zhao, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. RIVER: A real-time interaction benchmark for video LLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xmtvHH62Ic.

Guowei Tang, Tianwen Qian, Huanran Zheng, Yifei Wang, and Xiaoling Wang. StreamingEval: A unified evaluation framework towards realistic streaming video understanding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5945–5960. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.295. URL https://aclanthology.org/2026.findings-acl.295/.

Hongyuan Tao, Bencheng Liao, Shaoyu Chen, Haoran Yin, Qian Zhang, Wenyu Liu, and Xinggang Wang. InfiniteVL: Synergizing linear and sparse attention for highly-efficient, unlimited-input vision-language models. *arXiv preprint arXiv:2512.08829*, 2025. URL https://arxiv.org/abs/2512.08829.

Pengyu Wang, Chenkun Tan, Shaojun Zhou, Qirui Zhou, Yanxin Chen, Xingyang He, Huazheng Zeng, Jijun Cheng, Chenghao Wang, Xiaomeng Qian, Pengfei Wang, Zhan Huang, Shanqing Gao, Wei Huang, Longjun Cao, Wu Ran, Jie Liu, Changtai Zhu, Hongkai Wang, Yixian Tian, Chenghao Liu, Zhen Ye, Xinghao Wang, Botian Jiang, Guoguo Feng, Zhaoye Fei, Ruixiao Li, Mingshu Chen, Yang Gao, Qinyuan Cheng, Shimin Li, and Xipeng Qiu. MOSS-VL technical report. *arXiv preprint arXiv:2608.15045*, 2026. URL https://arxiv.org/abs/2608.15045.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. MMDuet2: Enhancing proactive interaction of video MLLMs with multi-turn reinforcement learning. *arXiv preprint arXiv:2512.06810*, 2025a. URL https://arxiv.org/abs/2512.06810.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. ProactiveVideoQA: A comprehensive benchmark evaluating proactive interactions in video large language models. *arXiv preprint arXiv:2507.09313*, 2025b. URL https://arxiv.org/abs/2507.09313.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. OmniMMI: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025c. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_OmniMMI_A_Comprehensive_Multi-modal_Interaction_Benchmark_in_Streaming_Video_Contexts_CVPR_2025_paper.html.

Keming Wu, Baoyi Wang, Kaichen Zhang, Xiang An, Zuhao Yang, Sudong Wang, Haowei Zhu, Tingxuan Huang, Hongcheng Gao, and Bin Wang. StreamOPD: A post-training recipe with spatio-temporal cue gating for streaming video understanding. *arXiv preprint arXiv:2608.16320*, 2026. URL https://arxiv.org/abs/2608.16320.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. StreamingVLM: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*, 2025. URL https://arxiv.org/abs/2510.09608.

Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, Yixin Liu, Hanbo Zhang, Ying Ma, and Xuming Hu. RTV-Bench: Benchmarking MLLM continuous perception, understanding and reasoning through real-time video. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0600. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19e4ea30dded58259665db375885e412-Abstract-Datasets_and_Benchmarks_Track.html.

Senqiao Yang, Kaichen Zhang, Zhaoyang Jia, Jinghao Guo, Yifei Shen, Xinjie Zhang, Xiaoyi Zhang, Haoqing Wang, Xiao Li, Peng Zhang, Xiang An, Yin Xie, Zhening Liu, Xun Guo, Jiahao Li, Shicheng Zheng, Jinglu Wang, Zongyu Guo, Wenxuan Xie, Zihan Zheng, Yuxuan Luo, Bin Li, and Yan Lu. Mage-VL: An efficient codec-native streaming multimodal foundation model. *arXiv preprint arXiv:2607.24904*, 2026a. URL https://arxiv.org/abs/2607.24904.

Zhenyu Yang, Kairui Zhang, Bing Wang, Shengsheng Qian, and Changsheng Xu. LiveStarPro: Proactive streaming video understanding with hierarchical memory for long-horizon streams. *arXiv preprint arXiv:2606.17798*, 2026b. URL https://arxiv.org/abs/2606.17798.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Xiangyu Zeng, Yifei Li, Haowen Hou, Zheming Liang, Congcong Wang, Kaiwen Tuo, Jun Zhang, Yuhan Zhu, Yuhang Cao, Shenglong Ye, Shuai Xie, Shuhuan Gu, Haoyang Huang, Qingyi Si, Nan Duan, and Jiaqi Wang. JoyAI-VL-Interaction: Real-time vision-language interaction intelligence. *arXiv preprint arXiv:2606.14777*, 2026. URL https://arxiv.org/abs/2606.14777.

Z.ai. GLM-5.3. Hugging Face model card, 2026. URL https://huggingface.co/zai-org/GLM-5.3.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. StreamForest: Efficient online video understanding with persistent event memory. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2509.24871.

Kaichen Zhang, Wei Huang, Keming Wu, Bo Li, and Xiaojuan Qi. Aero realtime: Fully aligned input-output streams for low-latency streaming multimodal generation. *arXiv preprint arXiv:2608.08469*, 2026. URL https://arxiv.org/abs/2608.08469.

Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, Qipeng Guo, Haodong Duan, Xin Chen, Han Lv, Zheng Nie, Min Zhang, Bin Wang, Wenwei Zhang, Xinyue Zhang, Jiaye Ge, Wei Li, Jingwen Li, Zhongying Tu, Conghui He, Xingcheng Zhang, Kai Chen, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2.5-OmniLive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*, 2024. URL https://arxiv.org/abs/2412.09596.

A Task Formulation and Trace Semantics

A.1 Causal media arrivals and public actions

An episode consists of timestamped media arrivals

$$X = \{x_i = (m_i, \tau_i)\}_{i=1}^{N}, \tag{2}$$

where m_i is a video frame, audio chunk, or interface-specific media event and τ_i is its availability time in the episode timebase. At decision time t, the harness may deliver only $X_{\leq t} = \{x_i : \tau_i \leq t\}$. The official system maintains an internal state s_t according to its released inference interface; any internal representation is admissible.

At each decision opportunity, the system produces one of three protocol actions:

$$a_t \in \{\text{WAIT},\ \text{ANSWER}(y),\ \text{REPAIR}(y)\}. \tag{3}$$

WAIT leaves the public state unchanged. ANSWER creates the first public answer. REPAIR replaces an existing public answer. A REPAIR emitted before the first public ANSWER is recorded as an illegal transition, while the public state remains empty. The empty public state is distinct from a published NONE answer.

Let p_t denote the public answer state after processing actions through time t. This state is empty until an ANSWER, equals the text of the most recent ANSWER or REPAIR thereafter, and persists across WAIT decisions. Evaluation follows the trajectory of p_t across the episode.

A.2 LIFECYCLE ANNOTATIONS

Each record defines a reference answer y_A, the start t_A of its annotated support, and an episode end T. A Retention record additionally defines an evaluation boundary t_H and horizon H while y_A remains valid. A Changed record defines the invalidation time t_I, the onset $t_R \geq t_I$ of annotated support for the updated reference, and that reference y_B. The evaluable answer intervals are

$$\mathcal{V}_{\text{retention}} = [t_A, T], \tag{4}$$

and

$$\mathcal{V}_{\text{changed}} = [t_A, t_I) \cup [t_R, T], \qquad r(t) = \begin{cases} y_A, & t \in [t_A, t_I), \\ y_B, & t \in [t_R, T]. \end{cases} \tag{5}$$

For Retention trajectories, $r(t) = y_A$ throughout $\mathcal{V}_{\text{retention}}$. The Changed gap $[t_I, t_R)$ is outside the scoring domain, and repair eligibility begins at t_R. Figure 2 illustrates the timing structure.

The four core transitions are Retention (A-to-A), Replacement (A-to-B), Withdrawal (A-to-NONE), and Formation (NONE-to-A). NONE denotes an empty active-label set within the activity scope specified by the question. It is a published reference answer, with the same initial-answer and update requirements as an active label.

Core-400 derives its support boundaries from complete official action or procedure-step intervals, as detailed in Appendix C. The initial support starts at t_A. A Retention trajectory places t_H two seconds before that support ends and fixes $H = 2000$ ms. For A-to-B, the initial support ends at t_I, and the next support starts at t_R. For Withdrawal and Formation, the boundary between an active support and a NONE support defines $t_I = t_R$.

A.3 CAUSAL AND OBSERVATIONAL SCOPE

The trace records delivery times, decision times, public actions, model decisions, and runtime metadata. For every decision at t, the recorded delivery sequence contains no media item with $\tau_i > t$. The released source revision, delivery order, and action events define the evaluated interface at the harness boundary.

B MODEL INTERFACES AND MEDIA ACCESS

B.1 PRIMARY INTERFACE CLASSES

VIDEOLIVE uses two primary interface classes. A *Released Incremental Interface* accepts video frames, audio chunks, or audio-visual events as media arrive. The interface may be a local session, a WebSocket service, or a server API. A *Causal-Prefix Streaming Adapter* supplies the media available through the current decision time and carries the preceding public answer into the next call. Diagnostic variants change the evidence window, cadence, modality, or query schedule for sensitivity analysis.

The classes describe media access and state continuity independently of model architecture. Table 4 lists input modalities, temporal state mechanisms, and public-action mappings for sixteen incremental configurations.

B.2 MEDIA DELIVERY AND ACTION MAPPING

Each evaluated configuration uses its released checkpoint through the corresponding incremental interface or causal-prefix adapter. The evaluation delivers timestamped media and records model outputs and public actions.

The harness may schedule already-arrived media, map model output into the public WAIT/ANSWER/REPAIR protocol, persist traces, and record resources. An unmodified configuration preserves the model forward pass, KV or cache policy, attention mask, silence threshold, response policy, and official query, rollback, and commit lifecycle. Each intervention is evaluated as a distinct configuration.

Table 4: Input modalities and temporal state mechanisms of the sixteen incremental interfaces.

System	Official entry point	Source modalities	Mechanism annotation	Harness boundary
MiniCPM-o	official streaming prefill and generation entry points	audio, video, text	persistent official streaming state with speculative rollback on silence	timestamped media delivery and public-action recording
MMDuet2	official `ProactiveInference Client.inference`	video, text	official client owns conversation history, KV state, masks, and generation	adapter delegates inference; history and KV remain client-owned
OmniLive	official `Session` and `Model` process/query path	video, text	committed segments plus replaceable active segment; reactive query at tick	one-frame delivery with memory acknowledgement through the official query path
JoyAI-VL-Interaction	released OpenAI-compatible server session	video, text	server-side session state with one causal frame per turn	public silence and response tokens mapped to WAIT and state updates
Aero Realtime	released `run_aero.py` real-time path	audio, video, text	aligned persistent audio-video session with lexical-burst emission	completed lexical bursts sampled on the common decision clock
StreamOPD-4B	released recent-window streaming path	video, text	four-frame causal window with spatio-temporal cue gating	one response per native update; intervening ticks carry the state
ThinkStream-3B	released watch–think–speak inference path	video, text	persistent dynamic KV state with a private thinking branch	content after the response marker becomes the public update
VST-7B	released streaming worker	video, text	streaming thought memory updated with arriving frames	current-memory query exposed on each native update
Streaming VLM	released persistent-KV inference path	video, text	incremental visual KV with cached prefix state	one streaming response exposed per native update
InfiniteVL	released streaming-inference path	video, text	persistent recurrent state with a cloned query cache	branch query reads a clone of the current state; the main cache persists
StreamForest	released online-video evaluation path	video, text	causal-prefix event memory with local-window tokens	query-conditioned memory response exposed as the public update
Mage-VL	released `inference_streaming.py` path	video, text	chronological segment gate with generation on selected segments	StreamMind emissions aligned to the common clock and mapped to public updates
LiveStarPro	released Pro streaming-inference path	video, text	ShortTermMemory with recursive event-tree retrieval	causal frame delivery with proactive captions and answers mapped to public updates
MOSS-VL-Realtime	released online-inference path	video, text	persistent text KV and cumulative visual cross-attention KV	timestamped causal frames with silence and lexical responses mapped to public actions
Qwen3.5-Omni-Flash^{close}	Realtime WebSocket session	audio, video, text	persistent audio-video session; manually triggered text responses	causal media chunks advance between responses; probe queries replay media in a fresh session
Qwen3.8-Omni-Flash^{close}	Realtime WebSocket session	audio, video, text	persistent audio-video session; manually triggered text responses	causal media chunks advance between responses; probe queries replay media in a fresh session

B.3 Shared inputs and supported modalities

Configurations share the questions and reference timelines on overlapping videos. Media sampling follows the interface settings, with audio supplied to interfaces that support it. Public actions use the common content and timing metrics. Appendix J measures audio contribution for MiniCPM-o and Aero Realtime through paired audio-video and video-only runs.

Both interface classes follow the same causal media clock and public-state transition rules. The sensitivity experiments vary one delivery or query condition at a time.

C Data Collection and Evaluation Protocol

C.1 Held-out Core-400 cohort

Core-400 contains one episode from each of 400 distinct source videos. Its four lifecycle types each contain 90 Perception Test and 10 CaptainCook4D videos. Perception Test contributes 42.35 minutes and CaptainCook4D contributes 42.24 minutes, accounting for 50.1% and 49.9% of the total observation time. Scored reference supports occupy 41.37 and 40.64 minutes, respectively, including intervals with NONE as the reference answer. Episode durations range from 4.10 to 218.93 seconds. The selected supports cover 49 action labels in Perception Test and 41 procedure-step labels in CaptainCook4D, with NONE counted separately. Each label count uses its source's annotation vocabulary (Patraucean et al., 2023; Peddi et al., 2024).

All 100 Retention records use a 2000 ms retention horizon, with reference states taken from their source annotations.

Content families in Figure 3. The figure groups existing labels for descriptive visualization, without changing trajectory selection, reference answers, or evaluation tasks. Each trajectory receives one content anchor: the pre-change active reference for Retention and Withdrawal, and the post-change active reference for Replacement and Formation. Perception Test labels are grouped into object manipulation (including motion, material changes, and marking), display and gestures (including showing and browsing), access and controls (including opening, closing, and switches), and simulated/other actions. The last group contains the explicit pretending-action labels and the single source label "Other." CaptainCook4D anchors form the cooking-procedure group. The four lifecycle rows have counts (MAN, DSP, CTL, CK, SO) of $(55, 19, 14, 10, 2)$, $(54, 9, 18, 10, 9)$, $(43, 32, 14, 10, 1)$, and $(47, 31, 9, 10, 3)$, respectively. The released figure data include the label-to-family mapping and one row per trajectory.

Selection precedes model execution and media inspection. Source-video grouping contributes at most one primary record per source. The 100-per-transition allocation exceeds the prespecified Wilson planning minimum of 93 independent sources for a two-sided 95% interval with approximately ten percentage points of half-width under the worst-case proportion.

Each record is aligned to its question-end timestamp, making the first subsequent media point time zero. CaptainCook4D trajectories measure maintenance or transition of a local procedure step.

C.2 Interval-to-trajectory mapping

For Perception Test, we partition the full source timeline at every annotated action boundary and merge adjacent segments with identical active-label sets. A single-label segment supports that action; an empty segment supports NONE, including leading and trailing segments. Every positive-length annotation contributes to the active set. Support intervals require one active label or an empty set throughout. CaptainCook4D candidates use the official procedure-step mapping and satisfy the same coverage and active-label conditions against all annotated steps.

Each activity or NONE support lasts at least two seconds. Retention candidates contain at least four seconds of continuous support, with t_H two seconds before its end. The initial-answer window before the evaluation boundary and the subsequent retention or repair window each contain at least four observations on the 500 ms grid. Perception Test A-to-B candidates join different single-label segments, either adjacent or separated by an empty interval of at most 750 ms. The interval $[t_I, t_R)$ lies outside their scored answer supports. A-to-NONE and NONE-to-A use an empty interval as an explicit reference support. Perception Test Withdrawal clips this NONE support to its first two seconds; Formation uses its final two seconds.

The question lists the activity scope in alphabetical order: all 63 source action classes for Perception Test, or the complete official step vocabulary for the named CaptainCook4D recipe. It asks which activity in that scope is happening now, accepts semantically equivalent descriptions, and states the publication rules in Section 3.1.

Selection assigns source videos to transition types by deterministic maximum flow, with one trajectory per video and source-specific quotas of 90 and 10 for each type. Within an assigned video–type pair, a fixed SHA-256 ordering selects the candidate. Each selected episode is clipped and aligned to question end, with all reference boundaries expressed in the same local timebase. The selected media retain their original frames and aligned audio.

C.3 Causal delivery and decisions

For the 22 Core-400 configurations, video is delivered at 2 frames per second. Where a system accepts source audio, audio is chunked at 500 ms. The decision grid is 500 ms, and the common harness exposes a 128-token generation budget. Released clients use their documented generation settings, which are recorded in the model specifications. API sampling and generation settings appear in Appendix L.

At each tick, the harness first delivers every newly available media unit and waits for the system-specific acknowledgement defined by the adapter. It then requests or observes a decision through the official interface. The event log records media time, wall-clock time, raw model output, parsed decision, public action, latency, and system metadata. Malformed or illegal model decisions remain in the trace.

C.4 Cohort coverage

The 29 configurations contribute 11,598 complete lifecycle trajectories and 46,392 phase records, including 34,794 applicable content queries and 45,002 applicable action queries. All configurations cover the complete 400 videos, apart from one video that both Qwen Realtime configurations reject at the provider. Each full trajectory contributes to initial-answer success and VAT; Retention and Changed trajectories contribute to Stability and Adaptivity, respectively.

D Formal Metric Definitions

Let $C(y, r) \in \{0, 1\}$ indicate whether public text y is semantically correct for reference r, with $C(\varnothing, r) = 0$. For open text, Qwen3.8-27B supplies the binary semantic evaluation.

D.1 Time-valid initial answer

For an initial ANSWER at time t_0 with text y_0, semantic correctness alone is insufficient. In a Retention episode, the answer is time-valid when $t_0 \geq t_A$. In a Changed episode, it is time-valid only in $[t_A, t_I)$. The initial endpoint is

$$I_{\text{init}} = C(y_0, y_A)\, \mathbf{1}[t_0 \in \mathcal{T}_A], \tag{6}$$

where $\mathcal{T}_A = [t_A, T]$ for Retention and $\mathcal{T}_A = [t_A, t_I)$ for Changed. No ANSWER, a premature answer, or a semantically incorrect answer receives zero.

D.2 Retention completion and Stability

Retention completion, which enters LCS, requires a valid initial answer by t_H and a correct public answer throughout the following observation horizon:

$$I_{\text{retention@}H} = I_{\text{init}}\, \mathbf{1}[t_0 \leq t_H]\, \mathbf{1}[C(p_t, y_A) = 1 \text{ for all } t \in [t_H, \min(T, t_H + H)]], \tag{7}$$

with the indicator set to zero when no time-valid initial answer exists by t_H. Stability scores retention separately from the first answer, over the half-open window $\mathcal{W} = [t_H, t_H + H)$:

$$I_{\text{stab}} = \mathbf{1}[C(p_t, y_A) = 1 \text{ for all } t \in \mathcal{W}]. \tag{8}$$

An empty or incorrect answer at t_H fails, as does any incorrect state inside $\mathcal{W}$; an incorrect first answer corrected before t_H can succeed. The window excludes its endpoint because a publication at $t_H + H$ is never displayed inside it. Stability averages I_{stab} over all assigned Retention trajectories.

D.3 Correct repair within two seconds

For Changed episodes, a successful primary repair requires a time-valid, semantically correct initial answer and at least one REPAIR whose time lies in $[t_R, t_R + 2000\,\text{ms}]$ and whose text is correct for y_B:

$$I_{\text{repair@}2s} = I_{\text{init}}\, \mathbf{1}\left[\exists\, \text{REPAIR}(y)_t :\ 0 \leq t - t_R \leq 2000,\ C(y, y_B) = 1\right]. \tag{9}$$

A repair emitted before annotated post-change support begins at t_R is premature. The primary Changed endpoint requires a valid initial answer followed by a timely correct repair.

D.4 Valid Answer Time and diagnostics

Let $\mathcal{V} = [t_A, T]$ for Retention trajectories and $\mathcal{V} = [t_A, t_I) \cup [t_R, T]$ for Changed trajectories. Valid Answer Time measures the fraction of this supported duration with a semantically correct public answer:

$$\text{VAT} = 100\, \frac{\int_{t \in \mathcal{V}} C(p_t, r(t))\, dt}{\int_{t \in \mathcal{V}} dt}. \tag{10}$$

Table 5: VideoLive scores and their evaluation populations. All scores use a 0–100 scale. Phase scores average applicable observations within each trajectory before equal-weight aggregation over trajectories.

Score	Population	Definition
Overall	Four summary scores	Arithmetic mean of Lifecycle Completion Score, Valid Answer Time, Understanding, and Temporal decision, each weighted at 25%.
Lifecycle Completion Score	Evaluated trajectories	Percentage completing valid formation and the applicable Retention or Changed repair requirement.
Initial-answer success	Evaluated trajectories	Percentage whose first ANSWER is semantically correct within the initial support interval.
Stability	Retention trajectories	Percentage with a correct public answer at t_H that remains correct throughout $[t_H, t_H + 2\,\mathrm{s})$.
Adaptivity	Changed trajectories	Percentage with a valid initial answer and at least one semantically correct REPAIR within two seconds of t_R.
Valid Answer Time	Evaluated trajectories	Mean percentage of supported duration occupied by a correct public answer, over $[t_A, T]$ for Retention and $[t_A, t_I) \cup [t_R, T]$ for Changed.
Understanding	Three supported phases per trajectory	Mean semantic content accuracy across first-sufficient, pre-boundary, and post-boundary probes.
Temporal decision	Up to four applicable phases per trajectory	Mean per-trajectory decision accuracy over applicable phases. The target uses reference support, publication status, and the preceding answer's semantic correctness; an ANSWER or REPAIR also requires a semantically correct resulting answer.

The public answer p_t retains the latest ANSWER or REPAIR text until the next publication or episode end. At each supported interval, it is scored against the reference $r(t)$ for that interval. The gap $[t_I, t_R)$ contributes to neither integral. Empty or incorrect public answers contribute zero to the numerator. The reported VAT is the equal-weight mean of the evaluated trajectory-level percentages.

VAT accumulates correct duration throughout the supported intervals, including correct states reached after an erroneous first answer or a late repair. Changed completion is determined by a valid initial answer and the occurrence of a correct repair within two seconds; subsequent public answers contribute to VAT according to their duration and correctness. Appendix F reports a closed-label exact-match variant for matcher sensitivity, together with outcome partitions and repair-deadline sensitivity.

D.5 Lifecycle summary and capability dimensions

The VideoLive Lifecycle Completion Score averages the applicable lifecycle indicator across the assigned cohort. Retention trajectories contribute $I_{\mathrm{retention@}H}$ and Changed trajectories contribute $I_{\mathrm{repair@}2s}$. With 100 Retention and 300 Changed trajectories, $\mathrm{LCS} = 0.25R + 0.75A$, where R is the Retention completion rate and A is Adaptivity on the 0–100 scale. Overall is the equal mean of LCS, VAT, Understanding, and Temporal decision. Table 5 summarizes every score and its population.

D.6 Aggregation and uncertainty

Initial-answer success and lifecycle completion average the evaluated trajectories. Stability uses the Retention subset; Adaptivity uses the Changed subset. Understanding averages the three supported content probes within each trajectory. Temporal decision averages the applicable action probes within each trajectory. Both then average trajectories equally. The complete 400-video cohort contributes 1,552 applicable action observations: 352 before support and 400 at each supported phase; the 399-video Qwen Realtime cohort contributes 1,549. Valid Answer Time averages each trajectory's correct-duration fraction. Overall averages LCS, VAT, Understanding, and Temporal decision and is recomputed within each bootstrap draw.

Intervals use 10,000 source-video bootstrap draws, stratified by the four lifecycle types (NumPy generator seed 20260922). Each trajectory has a distinct source video. All observations from a sampled video are drawn together, and paired conditions share the same draws. Pooled rates divide

summed success counts by the corresponding total number of evaluated pairs. For pooled results across systems, each draw retains all systems evaluated on that video. The 8,698 system–Changed-trajectory pairs correspond to 300 source-video clusters, with 100 per transition type. Reported intervals are percentile intervals.

D.7 Score uncertainty intervals

With these procedures, Tables 6 and 7 report point estimates and 95% intervals for all 29 model configurations and all eight main-table scores, using the cohorts specified in the evaluation setting.

E Model and Evaluation Details

The incremental interfaces are MiniCPM-o (Cui et al., 2026), MMDuet2 (Wang et al., 2025a), OmniLive (Zhang et al., 2024), JoyAI-VL-Interaction (Yao et al., 2026), Aero Realtime (Zhang et al., 2026), StreamOPD-4B (Wu et al., 2026), ThinkStream-3B (Liu et al., 2026), VST-7B (Guan et al., 2026), StreamingVLM (Xu et al., 2025), InfiniteVL (Tao et al., 2025), StreamForest (Zeng et al., 2025), Mage-VL (Yang et al., 2026a), LiveStarPro (Yang et al., 2026b), MOSS-VL-Realtime (Wang et al., 2026), Qwen3.5-Omni-Flash-Realtime (Qwen Team, 2026a), and Qwen3.8-Omni-Flash-Realtime (Alibaba Cloud, 2026). The local causal-prefix adapters use Gemma 4 (Gemma Team, 2026), Qwen3.6-27B (Qwen Team, 2026b), Qwen3.8-27B (Qwen Team, 2026c), Molmo2-8B (Clark et al., 2026), Mage-VL, and MOSS-VL-Instruct; the hosted adapters are GPT-5.6 (OpenAI, 2026a), GPT-6 (OpenAI, 2026b), and Gemini 3.8 Flash (Google DeepMind, 2026) (Appendix L).

E.1 Released incremental interfaces

MiniCPM-o is evaluated through its released streaming prefill and generation entry points with audio, video, and text inputs, including the official speculative-generation rollback behavior. MMDuet2 delegates each inference step to the released `ProactiveInferenceClient`, which maintains history, KV state, keep masks, and generation. OmniLive uses the released `Session` and `Model` classes, including their frame queue, memory revision, query backup, and memory selection. These interfaces consume causal media and expose the state transitions required by the common trace contract.

Each of the fourteen released incremental interfaces processes 400 complete lifecycle trajectories and 1,600 phase-aligned observations. Each system retains the session, KV cache, recurrent state, segment gate, recent window, or event-memory mechanism listed in Table 4. Model responses remain in the trace when empty, malformed, or incorrect. Aero Realtime additionally completes a paired 400-trajectory video-only condition.

E.2 Causal-prefix streaming adapters

Full-trajectory evaluation advances over the same 500 ms decision grid used by the released incremental interfaces. At each decision point, the adapter retrieves every 2 FPS frame whose media timestamp has arrived and submits that growing prefix through the model's video inference endpoint. The prompt contains the question, decision timestamp, and current public answer. The model returns one WAIT, ANSWER, or REPAIR action, after which the benchmark updates the public answer and advances the causal clock. Each decision uses newly constructed visual inputs, and the public answer persists across the complete trajectory.

The eight local causal-prefix configurations also complete the matched-prefix analysis on all 400 trajectories. A content-readout call receives one selected causal prefix with no previous public answer. The action calls visit the four selected phases chronologically and receive the benchmark's current public answer in the prompt. The public state follows the same WAIT, ANSWER, and REPAIR transition rule as the full-trajectory evaluation. Each local configuration contributes 400 lifecycle trajectories and 1,600 phase-aligned observations. The hosted configurations follow the cohorts and media settings in Appendix L.

Table 6: Lifecycle scores and 95% confidence intervals.

System	LCS	Initial-answer success	Stability	Adaptivity
Incremental streaming interfaces				
MiniCPM-o	3.50 [2.00, 5.25]	13.25 [10.00, 16.50]	16.00 [9.00, 24.00]	0.00 [0.00, 0.00]
MMDuet2	3.25 [1.75, 5.00]	12.75 [9.75, 16.00]	13.00 [7.00, 20.00]	0.00 [0.00, 0.00]
OmniLive	0.00 [0.00, 0.00]	1.00 [0.25, 2.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
JoyAI-VL-Interaction	4.00 [2.25, 6.00]	26.00 [22.49, 29.75]	17.00 [10.00, 25.00]	0.00 [0.00, 0.00]
Aero Realtime	4.00 [2.25, 6.00]	14.75 [11.50, 18.25]	16.00 [9.00, 24.00]	0.00 [0.00, 0.00]
StreamOPD-4B	7.50 [5.25, 9.75]	22.75 [18.75, 26.75]	31.00 [22.00, 40.00]	0.33 [0.00, 1.00]
ThinkStream-3B	1.75 [0.50, 3.00]	7.75 [5.25, 10.50]	7.00 [2.00, 12.00]	0.00 [0.00, 0.00]
VST-7B	4.75 [3.00, 6.75]	11.50 [8.50, 14.50]	19.00 [12.00, 27.00]	0.00 [0.00, 0.00]
StreamingVLM	0.25 [0.00, 0.75]	1.25 [0.25, 2.50]	1.00 [0.00, 3.00]	0.00 [0.00, 0.00]
InfiniteVL	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
StreamForest	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Mage-VL (streaming)	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
LiveStarPro	5.50 [3.50, 7.75]	24.50 [20.75, 28.25]	21.00 [13.00, 29.00]	3.00 [1.33, 5.00]
MOSS-VL-Realtime	2.00 [0.75, 3.50]	7.50 [5.00, 10.00]	8.00 [3.00, 14.00]	0.00 [0.00, 0.00]
Qwen3.5-Omni-Flash^{close}	10.78 [8.27, 13.28]	27.57 [23.56, 31.58]	47.00 [37.00, 57.00]	0.67 [0.00, 1.67]
Qwen3.8-Omni-Flash^{close}	14.79 [11.53, 18.30]	38.10 [33.33, 42.86]	38.00 [29.00, 47.00]	12.71 [9.03, 16.39]
Causal-prefix streaming adapters				
Gemma 4 E2B-it	3.00 [1.50, 4.75]	6.25 [4.00, 8.50]	12.00 [6.00, 19.00]	0.00 [0.00, 0.00]
Gemma 4 E4B-it	2.00 [0.75, 3.50]	7.75 [5.25, 10.50]	9.00 [4.00, 15.00]	0.00 [0.00, 0.00]
Gemma 4 12B-it	3.25 [1.75, 5.00]	15.00 [11.75, 18.50]	17.00 [10.00, 25.00]	1.00 [0.00, 2.33]
Qwen3.6-27B	3.75 [2.00, 5.75]	29.00 [24.50, 33.25]	9.00 [4.00, 15.00]	3.67 [1.67, 6.00]
Qwen3.8-27B	17.00 [14.00, 20.25]	35.00 [30.75, 39.25]	58.00 [48.00, 68.00]	6.00 [3.67, 8.67]
Molmo2-8B	4.00 [2.25, 6.00]	11.50 [8.50, 14.50]	18.00 [11.00, 26.00]	0.00 [0.00, 0.00]
Mage-VL	4.00 [2.25, 6.00]	26.00 [22.00, 30.00]	16.00 [9.00, 24.00]	0.00 [0.00, 0.00]
MOSS-VL-Instruct	6.25 [4.25, 8.50]	18.00 [14.50, 21.75]	25.00 [17.00, 34.00]	0.00 [0.00, 0.00]
GPT-5.6 Luna^{close}	13.25 [10.25, 16.50]	29.00 [25.00, 33.25]	34.00 [25.00, 43.00]	13.00 [9.33, 16.67]
GPT-5.6 Terra^{close}	20.75 [17.00, 24.50]	38.75 [34.25, 43.25]	45.00 [35.00, 55.00]	19.00 [14.67, 23.33]
GPT-6 Luna^{close}	9.25 [6.50, 12.00]	31.75 [27.75, 35.75]	37.00 [28.00, 47.00]	7.67 [4.67, 10.67]
GPT-6 Sol^{close}	20.75 [17.00, 24.75]	36.75 [32.25, 41.25]	48.00 [38.00, 58.00]	20.00 [15.67, 24.33]
Gemini 3.8 Flash^{close}	23.50 [19.50, 27.50]	40.50 [36.25, 44.75]	24.00 [16.00, 33.00]	26.00 [21.33, 30.67]

LCS: Lifecycle Completion Score. Estimates and 95% intervals use each model's evaluated cohort. Intervals use 10,000 source-video resamples stratified by lifecycle type. ^{close}Closed-source models; Qwen-Omni-Flash uses Realtime.

Table 7: Capability scores and 95% confidence intervals.

System	Valid Answer Time (%)	Understanding	Temporal decision	Overall
Incremental streaming interfaces				
MiniCPM-o	14.62 [11.91, 17.39]	23.42 [20.58, 26.33]	19.35 [17.35, 21.40]	15.22 [13.45, 17.10]
MMDuet2	8.95 [6.93, 11.12]	21.33 [18.83, 24.00]	28.23 [26.44, 30.00]	15.44 [14.06, 16.95]
OmniLive	0.37 [0.08, 0.77]	16.25 [13.92, 18.58]	22.06 [21.25, 22.81]	9.67 [9.00, 10.33]
JoyAI-VL-Interaction	11.67 [9.48, 14.01]	26.25 [23.50, 29.08]	26.21 [23.75, 28.77]	17.03 [15.28, 18.92]
Aero Realtime	11.17 [8.84, 13.71]	19.25 [16.58, 22.00]	10.98 [9.08, 13.00]	11.35 [9.51, 13.35]
StreamOPD-4B	23.72 [20.71, 26.73]	24.25 [21.67, 26.83]	23.71 [21.46, 26.02]	19.79 [17.77, 21.83]
ThinkStream-3B	4.05 [2.67, 5.57]	4.33 [3.25, 5.58]	23.38 [22.25, 24.50]	8.38 [7.54, 9.29]
VST-7B	11.17 [8.79, 13.78]	12.25 [10.00, 14.67]	6.08 [4.94, 7.25]	8.56 [7.07, 10.21]
StreamingVLM	0.28 [0.01, 0.80]	0.42 [0.08, 0.83]	22.31 [21.31, 23.31]	5.81 [5.40, 6.35]
InfiniteVL	0.00 [0.00, 0.00]	2.83 [1.83, 3.92]	21.92 [21.06, 22.73]	6.19 [5.84, 6.54]
StreamForest	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Mage-VL (streaming)	0.00 [0.00, 0.00]	27.58 [24.75, 30.50]	22.00 [21.19, 22.75]	12.40 [11.63, 13.17]
LiveStarPro	18.90 [16.21, 21.61]	26.83 [24.08, 29.58]	19.88 [17.79, 21.96]	17.78 [15.75, 19.87]
MOSS-VL-Realtime	6.57 [4.72, 8.56]	5.50 [4.33, 6.75]	7.08 [5.77, 8.50]	5.29 [4.29, 6.37]
Qwen3.5-Omni-Flash^{close}	24.16 [21.15, 27.24]	24.90 [22.14, 27.74]	25.96 [24.46, 27.44]	21.45 [19.59, 23.30]
Qwen3.8-Omni-Flash^{close}	36.82 [33.70, 40.03]	37.26 [34.34, 40.27]	42.34 [39.95, 44.74]	32.80 [30.56, 35.05]
Causal-prefix streaming adapters				
Gemma 4 E2B-it	5.52 [3.84, 7.37]	8.33 [6.42, 10.42]	21.50 [20.27, 22.75]	9.59 [8.32, 10.94]
Gemma 4 E4B-it	7.59 [5.76, 9.62]	9.33 [7.42, 11.42]	24.54 [23.33, 25.77]	10.87 [9.73, 12.15]
Gemma 4 12B-it	15.80 [13.37, 18.38]	21.92 [19.42, 24.50]	31.77 [29.96, 33.62]	18.18 [16.58, 19.91]
Qwen3.6-27B	21.96 [20.10, 23.90]	35.25 [32.33, 38.17]	18.92 [16.85, 21.04]	19.97 [18.43, 21.55]
Qwen3.8-27B	29.76 [26.88, 32.68]	31.17 [28.25, 34.08]	37.62 [35.85, 39.38]	28.89 [26.77, 31.02]
Molmo2-8B	10.70 [8.30, 13.28]	19.08 [16.67, 21.50]	19.08 [17.33, 20.92]	13.22 [11.66, 14.89]
Mage-VL	16.86 [14.48, 19.35]	22.17 [19.67, 24.75]	17.21 [15.67, 18.77]	15.06 [13.60, 16.57]
MOSS-VL-Instruct	15.76 [13.09, 18.57]	32.08 [29.50, 34.75]	16.38 [14.56, 18.33]	17.62 [15.90, 19.44]
GPT-5.6 Luna^{close}	35.12 [32.45, 37.89]	34.50 [31.67, 37.42]	23.58 [21.33, 25.85]	26.61 [24.44, 28.87]
GPT-5.6 Terra^{close}	41.22 [38.21, 44.23]	38.00 [34.92, 41.00]	30.00 [27.60, 32.42]	32.49 [30.03, 34.97]
GPT-6 Luna^{close}	29.49 [26.62, 32.44]	35.83 [32.92, 38.67]	20.21 [18.10, 22.35]	23.70 [21.67, 25.78]
GPT-6 Sol^{close}	42.14 [39.01, 45.25]	40.67 [37.67, 43.50]	34.27 [31.77, 36.75]	34.46 [31.92, 37.01]
Gemini 3.8 Flash^{close}	45.04 [42.40, 47.59]	39.92 [37.08, 42.67]	37.69 [35.50, 39.83]	36.54 [34.23, 38.84]

LCS: Lifecycle Completion Score. Estimates and 95% intervals use each model's evaluated cohort. Intervals use 10,000 source-video resamples stratified by lifecycle type. ^{close}Closed-source models; Qwen-Omni-Flash uses Realtime.

E.3 SEMANTIC EVALUATION

Each scored ANSWER or REPAIR is matched to its reference text and decision type through the recorded action event. Qwen3.8-27B supplies binary semantic judgments with deterministic decoding. Procedure-step pronouns are resolved using the preceding annotated step when assessing the referenced object. Appendix I reports cross-model and human agreement on open-text judgments.

E.4 MATCHED-PREFIX CONTENT–ACTION ANALYSIS

The matched-prefix analysis samples four phases on the 500 ms decision grid. Evidence-before is the grid point at or immediately before $\max(0, t_A - 500\,\mathrm{ms})$ and is applicable when $t_A > 0$. First-sufficient is the first grid point at or after t_A. For Changed trajectories, pre-boundary is the grid point at or immediately before $\max(t_{\mathrm{first}}, t_I - 500\,\mathrm{ms})$, where t_{first} is the first-sufficient time, and post-boundary is the first grid point at or after t_R. Retention trajectories use the same pre-boundary rule with t_H in place of t_I. Their post-boundary observation is the first grid point at or after $t_H + 500\,\mathrm{ms}$, capped at the retention-horizon end.

At each supported phase, the content query asks for the current action or step with the previous public answer omitted. Qwen3.8-27B judges its semantic correctness. Understanding is the mean of the three supported probe outcomes within each trajectory, averaged equally over the evaluated cohort.

The action sequence visits all four phases and carries its own public answer between probe decisions. Incremental interfaces use the session and probe settings in Appendices B and L; causal-prefix adapters include the carried answer in the next prompt. The target is WAIT before initial support. During support, an empty public state requires ANSWER; an existing answer requires WAIT when semantically correct and REPAIR when incorrect. Qwen3.8-27B judges the preceding and resulting public answers separately against the current reference and question scope, accepting equivalent wording. Published NONE is a nonempty state. Action accuracy measures whether the parsed action matches its target; public-answer accuracy measures the semantic correctness of the answer exposed after the decision. A decision is correct when its action matches the target and, for ANSWER or REPAIR, the resulting public answer is correct. Temporal decision averages applicable observations within each trajectory, then averages the trajectory scores. Table 8 reports counts at each phase for every configuration.

The supplied-state intervention evaluates Qwen3.6-27B and Qwen3.8-27B on all 400 Core-400 trajectories, at the same final-phase query times as their independent action probes. Carried uses the public answer produced by preceding probe decisions. Supplied initializes the answer to the correct pre-transition reference, including published NONE for Formation; Retention receives its supported reference answer. Both conditions use the same causal video prefix, lifecycle prompt, and deterministic decoding.

The target under Supplied is WAIT for Retention and REPAIR for all Changed types. Public-answer correctness assesses the answer exposed after the decision using Qwen3.8-27B semantic judgments. Supplied action-target accuracy is 200/400 (50%) for Qwen3.6 and 148/400 (37%) for Qwen3.8. Table 9 reports answer correctness, paired gains and losses, and 95% intervals from 10,000 source-video bootstrap resamples stratified by lifecycle type. A carried answer may already match the post-transition reference; injecting its predecessor can introduce an additional updating requirement.

E.5 AUDIO ABLATION PROTOCOL

Aero Realtime was evaluated with audio and with equal-duration silence supplied to its native audio input; MiniCPM-o was evaluated with audio-video and with its audio input removed. Each comparison uses the same 400 Core-400 trajectories, with matched video, questions, media timestamps, and decision schedules. Initial-answer success, Stability, and Adaptivity use 400, 100 Retention, and 300 Changed trajectories, respectively. Table 26 reports the paired results.

E.6 HUMAN PERFORMANCE

Fourteen participants divided the 400 Core-400 videos among them, so that every video in the benchmark contributes one human trajectory: 360 from Perception Test and 40 from CaptainCook4D, with

Table 8: Independent phase probes: correct / applicable queries.

	Content			Action			
System	First	Pre	Final	Before	First	Pre	Final
Incremental streaming interfaces							
MiniCPM-o	84/400	131/400	66/400	155/352	37/400	71/400	37/400
MMDuet2	78/400	110/400	68/400	344/352	13/400	51/400	42/400
OmniLive	67/400	82/400	46/400	351/352	0/400	1/400	1/400
JoyAI-VL-Interaction	95/400	160/400	60/400	210/352	61/400	87/400	50/400
Aero Realtime	83/400	89/400	59/400	37/352	57/400	51/400	28/400
StreamOPD-4B	118/400	122/400	51/400	144/352	59/400	110/400	56/400
ThinkStream-3B	0/400	39/400	13/400	345/352	20/400	9/400	0/400
VST-7B	41/400	64/400	42/400	68/352	14/400	10/400	3/400
StreamingVLM	5/400	0/400	0/400	341/352	6/400	5/400	1/400
InfiniteVL	20/400	7/400	7/400	348/352	1/400	1/400	0/400
StreamForest	0/400	0/400	0/400	0/352	0/400	0/400	0/400
Mage-VL (streaming)	113/400	115/400	103/400	352/352	0/400	0/400	0/400
LiveStarPro	118/400	125/400	79/400	0/352	121/400	115/400	72/400
MOSS-VL-Realtime	40/400	12/400	14/400	60/352	9/400	24/400	15/400
Qwen3.5-Omni-Flash^{close}	84/399	140/399	74/399	305/352	17/399	62/399	23/399
Qwen3.8-Omni-Flash^{close}	148/399	205/399	93/399	342/352	91/399	151/399	78/399
Causal-prefix streaming adapters							
Gemma 4 E2B-it	22/400	36/400	42/400	315/352	4/400	13/400	11/400
Gemma 4 E4B-it	28/400	47/400	37/400	352/352	1/400	19/400	20/400
Gemma 4 12B-it	94/400	108/400	61/400	352/352	18/400	74/400	59/400
Qwen3.6-27B	127/400	205/400	91/400	34/352	97/400	115/400	44/400
Qwen3.8-27B	100/400	181/400	93/400	339/352	24/400	144/400	81/400
Molmo2-8B	46/400	116/400	67/400	200/352	28/400	43/400	28/400
Mage-VL	123/400	106/400	37/400	212/352	27/400	22/400	9/400
MOSS-VL-Instruct	127/400	182/400	76/400	165/352	22/400	50/400	18/400
GPT-5.6 Luna^{close}	115/400	197/400	102/400	0/352	107/400	167/400	90/400
GPT-5.6 Terra^{close}	135/400	219/400	102/400	43/352	124/400	203/400	93/400
GPT-6 Luna^{close}	126/400	200/400	104/400	23/352	109/400	132/400	53/400
GPT-6 Sol^{close}	142/400	230/400	116/400	90/352	116/400	218/400	104/400
Gemini 3.8 Flash^{close}	132/400	235/400	112/400	133/352	111/400	225/400	118/400

Content queries assess understanding; action queries assess temporal decision. An action query counting ANSWER or REPAIR as correct also requires a correct resulting public answer. First, pre, and final denote first support, pre-boundary, and the final probe. The Retention final probe evaluates retention. Before denotes the pre-support action probe. Main-table scores average applicable probes within each trajectory before averaging trajectories.

Table 9: Public-answer correctness after a supplied-state query on Core-400. Carried uses the preceding phase-probe answer; Supplied uses the correct predecessor, including published NONE for Formation. Differences are Supplied minus Carried in percentage points, with 95% source-video bootstrap intervals.

Model	Type	Carried	Supplied	Wrong→correct	Correct→wrong	Difference [95% CI]
Qwen3.6-27B	Retention	23/100	37/100	21	7	+14.00 [+4.00, +24.00]
Qwen3.6-27B	Replacement	9/100	22/100	14	1	+13.00 [+6.00, +20.00]
Qwen3.6-27B	Withdrawal	49/100	21/100	5	33	-28.00 [-39.00, -17.00]
Qwen3.6-27B	Formation	4/100	0/100	0	4	-4.00 [-8.00, -1.00]
Qwen3.6-27B	All Changed	62/300	43/300	19	38	-6.33 [-10.67, -2.00]
Qwen3.6-27B	All trajectories	85/400	80/400	40	45	-1.25 [-5.25, +2.75]
Qwen3.8-27B	Retention	63/100	93/100	30	0	+30.00 [+21.00, +39.00]
Qwen3.8-27B	Replacement	11/100	8/100	2	5	-3.00 [-8.00, +2.00]
Qwen3.8-27B	Withdrawal	2/100	2/100	1	1	+0.00 [-3.00, +3.00]
Qwen3.8-27B	Formation	5/100	4/100	1	2	-1.00 [-5.00, +2.00]
Qwen3.8-27B	All Changed	18/300	14/300	4	8	-1.33 [-3.67, +1.00]
Qwen3.8-27B	All trajectories	81/400	107/400	34	8	+6.50 [+3.75, +9.25]

100 trajectories for each of Retention, Replacement, Withdrawal, and Formation. Each participant saw a disjoint set of videos and worked independently. The interface displayed a short current-action or current-step question and played each clip sequentially, pausing at 500 ms intervals and at the final

frame for a decision. The participant could retain the current answer, publish an answer, or update it, including publishing NONE. Answers persisted until updated. Replay was restricted to the observed prefix, and response time during each pause was unrestricted. Across all participants the submissions contained 11,824 decisions: 400 initial publications, 1,760 updates, and 9,664 decisions to retain the existing state.

Qwen3.8-27B judged the original Chinese responses against the question-scoped reference states. The participant-facing questions used abbreviated wording without the activity-scope lists supplied to the models. Initial-answer success, retention, repair, and valid-answer duration were evaluated on the media timeline. Retention completion required a valid initial answer and retention; Changed completion required a valid initial answer and timely repair. Table 10 reports the pooled results. Initial-answer success was 40.00%; Stability and Adaptivity were 28.00% and 22.67%, respectively. Retrospective sampling of the submitted answer traces at the three supported phases gave Understanding 40.67 and Temporal decision 42.51. These are equal means of the three phase accuracies: content uses 400 observations per phase, and actions use 340, 400, and 400 recorded decisions. Together with LCS and VAT, their equal-weight mean is 36.30.

Table 10: Human performance on all 400 videos, pooled over 14 participants. Initial-answer success and completion columns give successful trajectories; VAT is the mean per-trajectory Valid Answer Time. Completion requires a valid initial answer and the applicable retention or repair criterion.

Transition	Videos	Initial-answer success	Completion	VAT (%)
Retention	100	24	8	44.00
Replacement	100	16	12	33.82
Withdrawal	100	24	16	41.90
Formation	100	96	40	52.42
Overall	400	160	76	43.04

F Lifecycle Diagnostics

F.1 Initial-answer success and conditional repair

Across 300 Changed trajectories per system, MiniCPM-o formed a time-valid correct answer on 39 cases, MMDuet2 on 36, and OmniLive on three; none completed a correct repair within two seconds. Qwen3.8-27B formed correctly on 88 Changed trajectories and completed 18. Conditional repair therefore assesses a selected subset whose size varies across systems (Table 11).

F.2 Lifecycle outcomes across systems

The taxonomy assigns each of 8,698 system–Changed-trajectory pairs across 29 configurations to its earliest unresolved stage. Initial answering fails in 7,092 pairs: 4,816 semantic mismatches, 2,233 missing answers, and 43 semantically correct but mistimed answers. The remaining 1,606 pairs have a valid initial answer (18.46%). Among them, 741 have no qualifying repair, 526 have incorrect or premature updates, and 339 complete a correct repair within two seconds. The conditional repair rate is 339/1,606 (21.11%); the completion rate over all evaluated Changed pairs is 339/8,698 (3.90%).

By transition, conditional repair is 63/432 (14.58%) for Replacement, 104/528 (19.70%) for Withdrawal, and 172/646 (26.63%) for Formation. Replacement and Withdrawal together give 167/960 (17.40%), 9.23 points below Formation (95% CI $[3.31, 15.05]$; 10,000 source-video draws stratified by transition, seed 2026092502). Formation minus Replacement alone is 12.04 points ($[4.25, 19.06]$). Configurations differ in where their valid initial answers occur: Qwen3.8-27B formed a valid initial NONE answer on 1 of 100 Formation trajectories, so its Formation completion is zero while its Replacement completion is 14%.

F.3 Diagnostic cohort definitions

The paired MiniCPM-o/MMDuet2 cache analysis, boundary shifts, representative cases, and the sequential-terminal comparison in Table 17 use a complementary event-centered cohort: 400 distinct

Table 11: Initial answers and timely repair on Changed trajectories.

System	Valid initial answers	Timely repairs	Conditional repair (%)	All Changed (%)
Incremental streaming interfaces				
MiniCPM-o	39	0	0.00	0.00
MMDuet2	36	0	0.00	0.00
OmniLive	3	0	0.00	0.00
JoyAI-VL-Interaction	88	0	0.00	0.00
Aero Realtime	43	0	0.00	0.00
StreamOPD-4B	61	1	1.64	0.33
ThinkStream-3B	23	0	0.00	0.00
VST-7B	27	0	0.00	0.00
StreamingVLM	4	0	0.00	0.00
InfiniteVL	0	0	—	0.00
StreamForest	0	0	—	0.00
Mage-VL (streaming)	0	0	—	0.00
LiveStarPro	83	9	10.84	3.00
MOSS-VL-Realtime	22	0	0.00	0.00
Qwen3.5-Omni-Flash^{close}	63	2	3.17	0.67
Qwen3.8-Omni-Flash^{close}	114	38	33.33	12.71
Causal-prefix streaming adapters				
Gemma 4 E2B-it	13	0	0.00	0.00
Gemma 4 E4B-it	22	0	0.00	0.00
Gemma 4 12B-it	43	3	6.98	1.00
Qwen3.6-27B	92	11	11.96	3.67
Qwen3.8-27B	88	18	20.45	6.00
Molmo2-8B	29	0	0.00	0.00
Mage-VL	88	0	0.00	0.00
MOSS-VL-Instruct	45	0	0.00	0.00
GPT-5.6 Luna^{close}	94	39	41.49	13.00
GPT-5.6 Terra^{close}	125	57	45.60	19.00
GPT-6 Luna^{close}	108	23	21.30	7.67
GPT-6 Sol^{close}	120	60	50.00	20.00
Gemini 3.8 Flash^{close}	133	78	58.65	26.00

Conditional repair divides timely correct repairs by valid initial answers; a dash indicates a zero denominator. All Changed uses every evaluated Changed trajectory. Cohorts follow the evaluation setting.

source videos, with 100 trajectories per lifecycle type. These clips total 75.92 minutes and pair selected temporal action or procedure-step intervals with short current-activity questions. Paired conditions share clips, reference timelines, and evaluation ticks; the cache comparison uses its 300 Changed videos. The Core-400 terminal queries, supplied-state interventions, and MiniCPM-o context probes use the complete-interval, activity-scoped construction in Section 3.

F.4 Representative trajectories from the diagnostic cohort

The four cases in Figure 8 illustrate retention, stale persistence, incorrect updating, and timely repair, from top to bottom. The retention case is the lexicographically first MMDuet2 Retention trajectory with one public action and successful retention. The stale-answer case is the shortest MiniCPM-o Changed trajectory with a fully classified public-action trace, a correct initial answer, at least five seconds of observation after new evidence, and no post-evidence REPAIR. The OmniLive case shows repeated object and action errors during a tomato-drying step. The final case is the earliest MMDuet2 trajectory meeting the two-second repair criterion. Table 13 reports the resulting timing anchors.

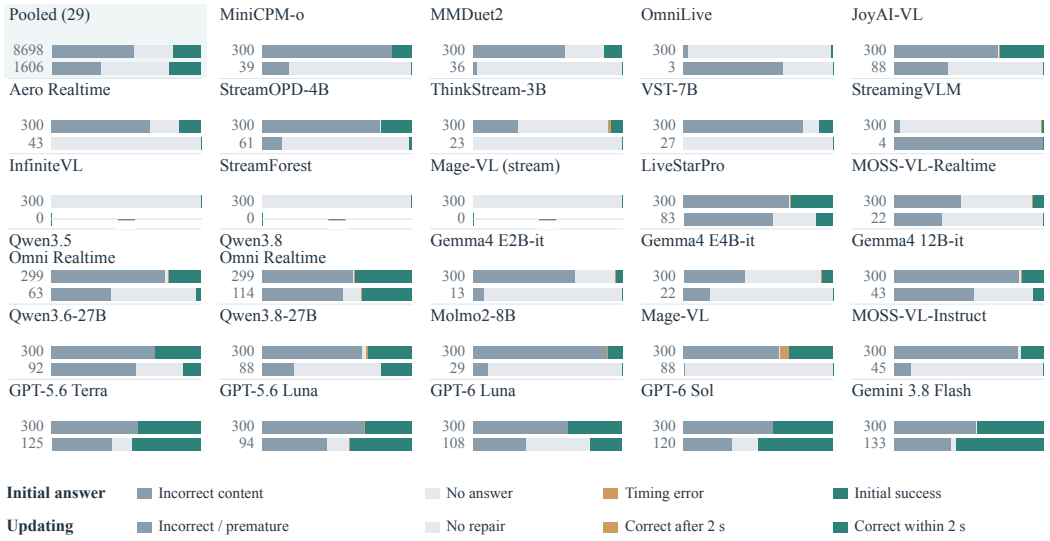


Figure 7: Initial-answer success and updating outcomes for each system. Upper bars partition initial-answer outcomes; lower bars condition on a valid initial answer. Counts give the denominators; each bar spans 0–100%. The pooled panel aggregates 8,698 Changed-trajectory evaluations across all 29 configurations. Dashes indicate undefined updating proportions.

F.5 Reporting-window sensitivity on diagnostic trajectories

Repair remains rare across reporting deadlines from one to five seconds. Counting every assigned trajectory, correct-repair counts at one, two, and five seconds are respectively 0, 0, and 1 for MiniCPM-o; 0, 1, and 1 for MMDuet2; and 0, 1, and 1 for OmniLive. These counts cover 300 Changed trajectories with at least one post-evidence second, 293 with at least two seconds, and 91 with at least five seconds. Trajectories that end before a deadline without a correct repair count as unsuccessful.

F.6 Valid Answer Time matcher sensitivity

Table 14 compares the semantic event-state matcher used in the main profile with closed-label exact matching. The semantic score assigns each public action its binary correctness decision and integrates that decision until the next public action or supported-state boundary. It yields positive duration on 26 of the 29 configurations and spans 0.00–45.04%. Exact matching is sensitive to surface-form differences between free-form model responses and dataset labels.

F.7 Content recognition and public answers on incremental interfaces

For every Changed trajectory, the independent post-boundary content probe queries the current activity at the first decision at or after t_R, with the previous public answer omitted. When this probe is semantically correct for the updated reference, we read the full-trajectory public answer at $t_R + 2\,\mathrm{s}$ and classify it as correct, incorrect, or not yet published, using the same Qwen3.8-27B judgments as the lifecycle scores. The probe and the full trajectory are separate executions on the same video and media clock. Table 15 reports the 16 incremental interfaces. Of 389 recognized trajectories, 131 show a correct public answer, 104 an incorrect one, and 154 no answer. Mage-VL (streaming), OmniLive, and MMDuet2 account for 129 of the unpublished cases. Among published answers, the incorrect share is 44.26% (95% CI $[37.78, 51.43]$; 10,000 draws over 170 source videos, seed 2026092503).

F.8 Terminal-answer queries on Core-400

Both the complete-video query and the fresh terminal-prefix query achieved 41% accuracy. They differ in prompts and input organization and share the complete observed video at the endpoint. Sequential interaction on the same videos yielded 3.75% lifecycle completion and 21.96% Valid Answer Time.

Table 12: Lifecycle completion by transition type (%).

System	Retention	Replacement	Withdrawal	Formation
Incremental streaming interfaces				
MiniCPM-o	14.00	0.00	0.00	0.00
MMDuet2	13.00	0.00	0.00	0.00
OmniLive	0.00	0.00	0.00	0.00
JoyAI-VL-Interaction	16.00	0.00	0.00	0.00
Aero Realtime	16.00	0.00	0.00	0.00
StreamOPD-4B	29.00	1.00	0.00	0.00
ThinkStream-3B	7.00	0.00	0.00	0.00
VST-7B	19.00	0.00	0.00	0.00
StreamingVLM	1.00	0.00	0.00	0.00
InfiniteVL	0.00	0.00	0.00	0.00
StreamForest	0.00	0.00	0.00	0.00
Mage-VL (streaming)	0.00	0.00	0.00	0.00
LiveStarPro	13.00	2.00	1.00	6.00
MOSS-VL-Realtime	8.00	0.00	0.00	0.00
Qwen3.5-Omni-Flash^{close}	41.00	1.00	0.00	1.00
Qwen3.8-Omni-Flash^{close}	21.00	5.00	21.21	12.00
Causal-prefix streaming adapters				
Gemma 4 E2B-it	12.00	0.00	0.00	0.00
Gemma 4 E4B-it	8.00	0.00	0.00	0.00
Gemma 4 12B-it	10.00	1.00	2.00	0.00
Qwen3.6-27B	4.00	2.00	3.00	6.00
Qwen3.8-27B	50.00	14.00	4.00	0.00
Molmo2-8B	16.00	0.00	0.00	0.00
Mage-VL	16.00	0.00	0.00	0.00
MOSS-VL-Instruct	25.00	0.00	0.00	0.00
GPT-5.6 Luna^{close}	14.00	5.00	8.00	26.00
GPT-5.6 Terra^{close}	26.00	5.00	20.00	32.00
GPT-6 Luna^{close}	14.00	5.00	4.00	14.00
GPT-6 Sol^{close}	23.00	12.00	13.00	35.00
Gemini 3.8 Flash^{close}	16.00	10.00	28.00	40.00

Each column reports the completion percentage within its lifecycle type and evaluated cohort. ^{close}Closed-source models; Qwen-Omni-Flash uses Realtime.

Table 13: Timing anchors for the representative event-centered diagnostic trajectories. Times use the common media clock. A dash indicates an inapplicable anchor or absence of a correct repair.

Panel	System	Transition	Initial (s)	New evidence (s)	Correct repair (s)	Repair latency (s)	Trace outcome
a	MMDuet2	Retention	7.000	–	–	–	Retained through the boundary and 2 s horizon.
b	MiniCPM-o	A-to-B	0.500	2.061	–	–	No update through episode end.
c	OmniLive	A-to-B	3.000	9.098	–	–	Repeated object and action errors.
d	MMDuet2	NONE-to-A	1.500	1.964	3.000	1.036	Complete Changed lifecycle.

F.9 Terminal-answer access on event-centered trajectories

The paired difference between sequential maintenance and fresh-prefix readout is -4.75 percentage points (95% CI, -9.00 to -0.75). Formation accuracy falls from 53% to 42%; Replacement falls from 48% to 45% and Retention from 44% to 39%. Withdrawal remains at 1% in both conditions and at 0% with the complete video.



Figure 8: Representative event-centered diagnostic trajectories. Each row aligns video frames with the reference, emitted actions, and the carried public answer. A and B denote the initial and updated references described below the frames; gray gaps lack reference support. Circles denote ANSWER and diamonds denote REPAIR. Green and orange indicate correct and incorrect content, respectively. The public-answer strip is white before publication and hatched during unscored intervals. Dashed lines mark updated reference support; brackets span the two-second repair window.

G Boundary and Parser Sensitivity

G.1 Lifecycle-boundary sensitivity on diagnostic trajectories

The common-boundary perturbation in Table 18 shifts all lifecycle boundaries by the same signed offset while leaving model actions and semantic judgments unchanged. At -500, 0, and $+500$ ms, MiniCPM-o's time-valid initial count is 62, 63, and 20; MMDuet2's is 67, 75, and 82; and OmniLive's is 10, 10, and 8. Retention counts over the same offsets are 12, 12, and 5; 21, 22, and 22; and 1, 1, and 0. Across the full ± 1000 ms grid, the largest observed initial, Retention, and Changed-repair counts remain 82/400, 22/100, and 1/300.

G.2 Lifecycle-score sensitivity

Across 29 configurations (Table 19), Retention weights of 0.50 and 0.75 give rank correlations of 0.977 and 0.940 with the reported LCS ranking. The retention-only score gives 0.871. The Changed-only score gives 0.785, with seventeen configurations tied at zero.

Table 14: Valid Answer Time under semantic and closed-label exact matching.

System	Semantic (%)	Exact (%)
Incremental streaming interfaces		
MiniCPM-o	14.62	11.74
MMDuet2	8.95	0.23
OmniLive	0.37	0.00
JoyAI-VL-Interaction	11.67	7.05
Aero Realtime	11.17	9.22
StreamOPD-4B	23.72	22.85
ThinkStream-3B	4.05	2.18
VST-7B	11.17	11.17
StreamingVLM	0.28	0.02
InfiniteVL	0.00	0.00
StreamForest	0.00	0.00
Mage-VL (streaming)	0.00	0.00
LiveStarPro	18.90	0.00
MOSS-VL-Realtime	6.57	5.95
Qwen3.5-Omni-Flash^{close}	24.16	23.31
Qwen3.8-Omni-Flash^{close}	36.82	26.42
Causal-prefix streaming adapters		
Gemma 4 E2B-it	5.52	5.05
Gemma 4 E4B-it	7.59	6.73
Gemma 4 12B-it	15.80	14.46
Qwen3.6-27B	21.96	12.91
Qwen3.8-27B	29.76	27.74
Molmo2-8B	10.70	9.51
Mage-VL	16.86	10.29
MOSS-VL-Instruct	15.76	14.81
GPT-5.6 Luna^{close}	35.12	22.21
GPT-5.6 Terra^{close}	41.22	19.32
GPT-6 Luna^{close}	29.49	8.40
GPT-6 Sol^{close}	42.14	10.29
Gemini 3.8 Flash^{close}	45.04	32.20

Both matchers use the same public-answer trace and supported intervals for each trajectory.

G.3 Boundary representation diagnostic

The diagnostic in Table 20 uses every MiniCPM-o trajectory that succeeds on Retention. The $t_A - 500$ ms and t_A prefixes contain the same sampled frames for all twelve trajectories because their initial support begins at 0–100 ms on the 2 FPS grid. Their pooled visual representations consequently coincide. The $t_A + 500$ ms prefix adds one frame in every case, producing a mean cosine distance of 0.0059 and changing the diagnostic ANSWER/WAIT action on three trajectories. Last-layer vision-transformer attention is recorded within each frame; its spatial concentration changes by -0.0002 on average from t_A to $t_A + 500$ ms, with a 95% interval of -0.0041 to 0.0030. The result separates changes in the scoring clock from changes in the model's sampled visual evidence.

G.4 Parser sensitivity

Analysis of the recorded raw outputs finds 1,439 outputs left unclassified by the action parser, concentrated in 21 MiniCPM-o trajectories, with no such output for MMDuet2 and eight across three OmniLive trajectories. A conservative embedded-object extractor identifies 13 MiniCPM-o WAIT objects, four MiniCPM-o REPAIR-like objects from one trajectory, and one OmniLive WAIT object. Normalizing the WAIT objects leaves the public-answer endpoints unchanged. The REPAIR-like

Table 15: Public answers two seconds after t_R on Changed trajectories where the independent content probe identifies the updated reference.

System	Probe identifies new reference	Public answer correct	Public answer incorrect	No public answer
MiniCPM-o	21	11	10	0
MMDuet2	43	4	18	21
OmniLive	32	0	1	31
JoyAI-VL-Interaction	12	1	11	0
Aero Realtime	40	20	7	13
StreamOPD-4B	26	18	8	0
ThinkStream-3B	7	4	0	3
VST-7B	19	9	9	1
StreamingVLM	0	0	0	0
InfiniteVL	6	0	0	6
StreamForest	0	0	0	0
Mage-VL (streaming)	77	0	0	77
LiveStarPro	56	36	20	0
MOSS-VL-Realtime	6	2	3	1
Qwen3.5-Omni-Flash^{close}	18	11	6	1
Qwen3.8-Omni-Flash^{close}	26	15	11	0
Total	389	131	104	154

Table 16: Qwen3.6-27B terminal-answer accuracy on 400 Core-400 videos, with 100 per lifecycle type. Both queries access the full video at its endpoint. Values are percentages.

Query	All	Retention	Replacement	Withdrawal	Formation
Complete video	41	67	39	7	51
Fresh terminal prefix	41	67	47	0	50

fragments fall outside the evaluated action grammar. The remaining outputs are empty, free text, or incomplete structured fragments under this extractor.

H Evidence Delivery and Context Use

H.1 Visual delivery and public-answer validity

Each condition covers the same 400 Core-400 trajectories. MiniCPM-o, MMDuet2, and OmniLive receive sustained visual delays of 500 or 1000 ms. The frame at time t is replaced by the latest sampled frame at or before $\max(0, t - d)$. Each model also receives a one-second local hold around t_R on Changed trajectories and the retention anchor on Retention trajectories. Audio, questions, decision times, and reference boundaries retain their baseline schedules. Figure 9 summarizes the paired VAT changes.

MiniCPM-o's VAT decreases from 14.62% at baseline to 13.28% under 500 ms lag and 12.16% under 1000 ms lag. Its one-second lag reduces VAT by 2.46 points on all 400 trajectories and by 2.24 points on the Changed subset; local-hold differences are -0.57 and -0.87 points, respectively. MMDuet2's one-second-lag differences are -1.55 and -2.22 points, and its all-trajectory local-hold difference is -0.12 points. OmniLive's baseline VAT is 0.37%, placing its small absolute changes near the metric's floor. MMDuet2 supplementary conditions use accelerated attention and a 512-token generation limit; the baseline uses the released inference implementation and a 128-token limit. Delay conditions use bounded SDPA decoder prefill; local hold and low-frequency sampling use native-aligned masked-tail decoding. These contrasts include implementation differences.

Intervals use 10,000 source-video bootstrap draws, stratified by lifecycle type. For the 400-trajectory VAT endpoint, two-sided sign-flip tests use 10,000 draws with add-one correction. Holm adjustment covers the fifteen comparisons in Table 22: nine visual-delivery conditions, four cadence conditions, and two audio contrasts. Changed-subset intervals describe paired effects over 300 source videos. MiniCPM-o's one-second-lag adjusted p is 0.258; all fifteen adjusted values exceed 0.05.

Table 17: Qwen3.6-27B terminal-answer accuracy under three access protocols on 400 paired event-centered diagnostic trajectories. Values are percentages.

Protocol	**Overall**	Retention	Replacement	Withdrawal	Formation
Complete video, one-shot	35.75	45.00	47.00	0.00	51.00
Terminal prefix, fresh state readout	36.50	44.00	48.00	1.00	53.00
Sequential prefixes, terminal public state	31.75	39.00	45.00	1.00	42.00

Table 18: Temporal-boundary sensitivity under a common signed offset to the prespecified lifecycle boundaries. Offset 0 is the primary result; all nonzero rows reuse identical model actions and semantic judgments and are descriptive only. Cells report success/assigned (rate), with source-video-cluster bootstrap 95% intervals below.

System	Offset (ms)	Initial valid	Retention@H	Repair@2s
MiniCPM-o	-1000	54/400 (13.50%) [10.25, 17.00]	10/100 (10.00%) [5.00, 16.00]	1/300 (0.33%) [0.00, 1.00]
MiniCPM-o	-500	62/400 (15.50%) [12.25, 19.00]	12/100 (12.00%) [6.00, 18.00]	1/300 (0.33%) [0.00, 1.00]
MiniCPM-o	+0	63/400 (15.75%) [12.25, 19.25]	12/100 (12.00%) [6.00, 18.00]	0/300 (0.00%) [0.00, 0.00]
MiniCPM-o	+500	20/400 (5.00%) [2.75, 7.25]	5/100 (5.00%) [1.00, 10.00]	1/300 (0.33%) [0.00, 1.00]
MiniCPM-o	+1000	5/400 (1.25%) [0.25, 2.50]	2/100 (2.00%) [0.00, 5.00]	0/300 (0.00%) [0.00, 0.00]
MMDuet2	-1000	63/400 (15.75%) [12.50, 19.75]	17/100 (17.00%) [10.00, 24.00]	0/300 (0.00%) [0.00, 0.00]
MMDuet2	-500	67/400 (16.75%) [13.25, 20.50]	21/100 (21.00%) [13.00, 30.00]	0/300 (0.00%) [0.00, 0.00]
MMDuet2	+0	75/400 (18.75%) [15.25, 22.75]	22/100 (22.00%) [14.00, 31.00]	1/300 (0.33%) [0.00, 1.00]
MMDuet2	+500	82/400 (20.50%) [16.75, 24.50]	22/100 (22.00%) [14.00, 31.00]	1/300 (0.33%) [0.00, 1.00]
MMDuet2	+1000	48/400 (12.00%) [8.75, 15.50]	11/100 (11.00%) [5.00, 18.00]	1/300 (0.33%) [0.00, 1.00]
OmniLive	-1000	10/400 (2.50%) [1.00, 4.25]	1/100 (1.00%) [0.00, 3.00]	0/300 (0.00%) [0.00, 0.00]
OmniLive	-500	10/400 (2.50%) [1.00, 4.25]	1/100 (1.00%) [0.00, 3.00]	0/300 (0.00%) [0.00, 0.00]
OmniLive	+0	10/400 (2.50%) [1.00, 4.25]	1/100 (1.00%) [0.00, 3.00]	1/300 (0.33%) [0.00, 1.00]
OmniLive	+500	8/400 (2.00%) [0.75, 3.50]	0/100 (0.00%) [0.00, 0.00]	0/300 (0.00%) [0.00, 0.00]
OmniLive	+1000	8/400 (2.00%) [0.75, 3.50]	0/100 (0.00%) [0.00, 0.00]	0/300 (0.00%) [0.00, 0.00]

H.2 MiniCPM-o context probes on Core-400

The analysis covers 400 distinct source videos and 1,784 decision times. Probes are taken after current-input ingestion and before native reply generation, at the observed ticks corresponding to offsets of -500, 0, 500, 1000, and 2000 ms from the first decision at or after the reference anchor. Changed trajectories use post-change support; Retention trajectories use the retention anchor. Layers 12, 24, and 36 are recorded under intact context, history masking, new-input masking, and random-context masking, giving 21,408 layer–time–condition records.

Table 19: LCS rank sensitivity to the Retention weight across 29 configurations.

Retention weight	0.00	0.25	0.50	0.75	1.00
Spearman ρ	0.785	1.000	0.977	0.940	0.871

Changed weight is one minus Retention weight. Correlations compare each weighted score with the reported LCS ranking; ties receive average ranks.

Table 20: MiniCPM-o diagnostic on all 12 retention successes. Representation distance is cosine distance between pooled visual prefixes; intervals use 1,000 source-trajectory bootstrap resamples.

Prefix comparison	Same frames	Mean distance	95% CI	ANSWER / WAIT
$t_A - 500$ ms to t_A	12/12	0.0000	[0.0000, 0.0000]	4 / 8
t_A to $t_A + 500$ ms	0/12	0.0059	[0.0040, 0.0078]	7 / 5

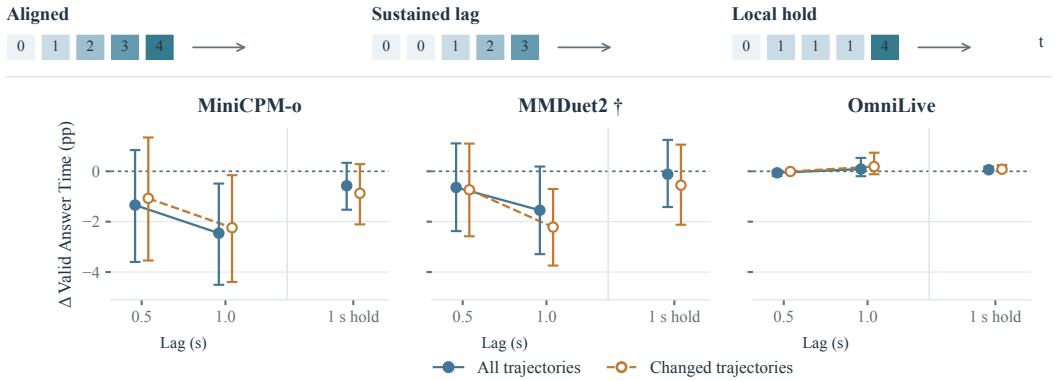


Figure 9: Visual delivery and Valid Answer Time. Frame identities illustrate sustained lag and local hold. Points show paired changes from baseline, with 95% source-video bootstrap intervals; filled and open markers denote all 400 and 300 Changed trajectories. Hold is a separate intervention. †MMDuet2 comparisons also change the inference implementation (Appendix H.1).

Candidate preference is the post-transition reference's mean token log likelihood minus the pre-transition reference's mean token log likelihood. Retention compares the correct supported answer with NONE. The three layer records share one output-level preference score, which is counted once per video, time, and condition. Each video contributes its mean effect over observed times; each lifecycle type then averages its 100 videos equally. Intervals use 10,000 paired source-video bootstrap draws. All intervals in this subsection are descriptive marginal 95% intervals.

Masks block access to equal numbers of historical, new-input, or randomly selected contextual positions during candidate scoring. The cached representations and native reply path remain intact. New input comprises the current prefill, including its media and associated text. Random positions are sampled from the combined history and new-input context. Figure 10 compares the two targeted

Table 21: Visual-delivery and cadence conditions on 400 trajectories per configuration (0–100).

Condition	LCS	Initial-answer success	Stability	Adaptivity	VAT
MiniCPM-o baseline	3.50	13.25	16.00	0.00	14.62
1 FPS / 1000 ms	4.50	15.25	17.00	1.33	13.87
500 ms lag	2.25	10.50	15.00	0.00	13.28
1000 ms lag	2.50	10.75	14.00	0.33	12.16
1000 ms hold	3.50	13.25	16.00	0.00	14.05
4 FPS / 250 ms	2.75	11.00	20.00	0.00	16.11
MMDuet2 baseline	3.25	12.75	13.00	0.00	8.95
500 ms lag	2.75	12.25	11.00	0.00	8.31
1000 ms lag	3.50	12.50	15.00	0.00	7.40
1 FPS / 1000 ms	3.00	13.25	12.00	0.00	8.14
1000 ms hold	3.00	13.00	13.00	0.00	8.83
OmniLive baseline	0.00	1.00	0.00	0.00	0.37
1000 ms lag	0.00	0.75	0.00	0.00	0.46
1000 ms hold	0.00	1.00	0.00	0.00	0.43
500 ms lag	0.00	0.75	0.00	0.00	0.31
1 FPS / 1000 ms	0.25	0.50	1.00	0.00	0.51

Table 22: Paired Valid Answer Time differences from the corresponding baseline (percentage points, 95% intervals).

Condition	All 400	Changed 300	p_{Holm}
MiniCPM-o: 1 FPS / 1000 ms	-0.75 [-2.77, +1.28]	-1.19 [-3.44, +1.06]	1.0000
MiniCPM-o: 500 ms lag	-1.34 [-3.60, +0.84]	-1.07 [-3.54, +1.34]	1.0000
MiniCPM-o: 1000 ms lag	-2.46 [-4.51, -0.49]	-2.24 [-4.39, -0.15]	0.2576
OmniLive: 1000 ms lag	+0.09 [-0.20, +0.53]	+0.19 [-0.12, +0.74]	1.0000
OmniLive: 1000 ms hold	+0.06 [+0.00, +0.19]	+0.08 [+0.00, +0.25]	1.0000
OmniLive: 500 ms lag	-0.06 [-0.19, +0.03]	-0.01 [-0.07, +0.05]	1.0000
MiniCPM-o: 1000 ms hold	-0.57 [-1.53, +0.34]	-0.87 [-2.11, +0.28]	1.0000
MMDuet2: 500 ms lag†	-0.64 [-2.37, +1.11]	-0.73 [-2.58, +1.10]	1.0000
OmniLive: 1 FPS / 1000 ms	+0.14 [-0.15, +0.48]	+0.10 [-0.26, +0.51]	1.0000
MMDuet2: 1000 ms lag†	-1.55 [-3.29, +0.19]	-2.22 [-3.74, -0.70]	1.0000
MiniCPM-o: 4 FPS / 250 ms	+1.49 [-0.88, +3.78]	+0.45 [-1.97, +2.78]	1.0000
Aero Realtime: audio-video − video-only	-0.06 [-0.44, +0.38]	-0.08 [-0.59, +0.49]	1.0000
MiniCPM-o: video-only − audio-video	+1.80 [+0.49, +3.18]	+2.16 [+0.49, +3.95]	0.1305
MMDuet2: 1 FPS / 1000 ms†	-0.80 [-2.61, +1.02]	-1.04 [-3.04, +0.92]	1.0000
MMDuet2: 1000 ms hold†	-0.12 [-1.42, +1.25]	-0.55 [-2.12, +1.06]	1.0000

Two-sided source-video sign flips (10,000 draws, add-one correction); Holm adjustment covers all fifteen listed contrasts on the 400-trajectory endpoint. Changed-subset intervals are descriptive. † MMDuet2 conditions use accelerated attention and a 512-token generation limit; these contrasts include implementation differences from baseline.

masks with this random control. System tokens and the appended diagnostic query remain accessible under every mask. Figure 11 reports each mask's effect relative to intact context.

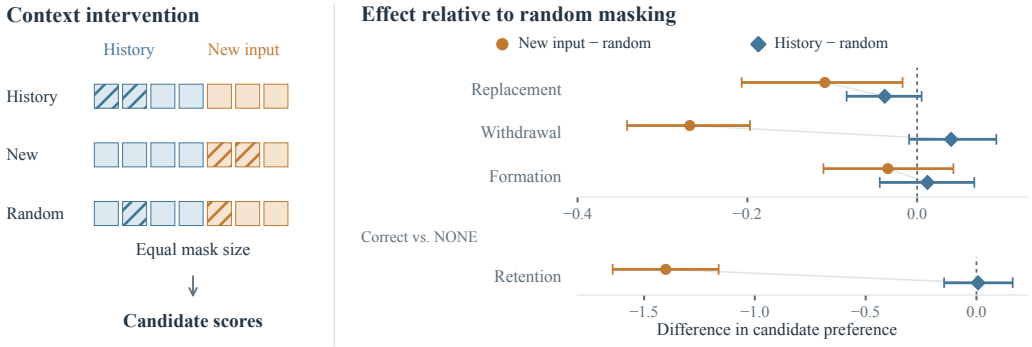


Figure 10: MiniCPM-o context sensitivity. Equal-sized masks act on a cache copy during candidate scoring. Points compare targeted and random masks, averaging probe times within each of 100 videos per type; bars show paired 95% source-video bootstrap intervals. Retention compares the supported answer with NONE on a separate scale; Changed types compare post- and pre-transition answers.

History masking differs little from random masking in all four types. New-input masking has a more negative effect on Replacement and Withdrawal, with differences of -0.109 and -0.268 and descriptive 95% intervals $[-0.207, -0.017]$ and $[-0.342, -0.197]$; on Formation, the difference is -0.034 ($[-0.110, 0.043]$). Retention shows a larger contrast, -1.402: history and random masks each raise preference by about 1.21, while the new-input mask lowers it by 0.197. On Formation, all three conditions lower preference by 0.236–0.282.

Figure 12 shows how attention is distributed across system tokens, history, new input, and the appended diagnostic query. Figure 13 measures the hidden representation's cosine distance from its intact-context counterpart at the same decision time. Distances are averaged over times within each video, then over videos. The curves separate layer-dependent representation changes from output-level candidate preference. Attention mass describes allocation during the diagnostic query; the masking contrasts measure its response to changed context access.

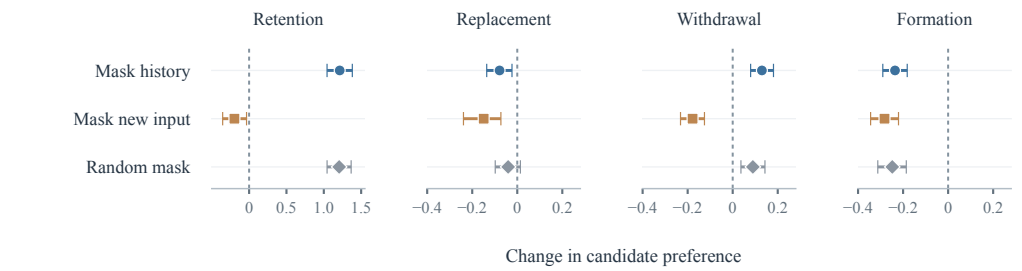


Figure 11: Candidate-preference effects relative to intact context. Each point averages probe times within source video and then the 100 videos of one lifecycle type; bars give 95% source-video bootstrap intervals. Retention compares the supported answer with NONE; the other types compare post- and pre-transition answers. Native public actions use the intact cache.

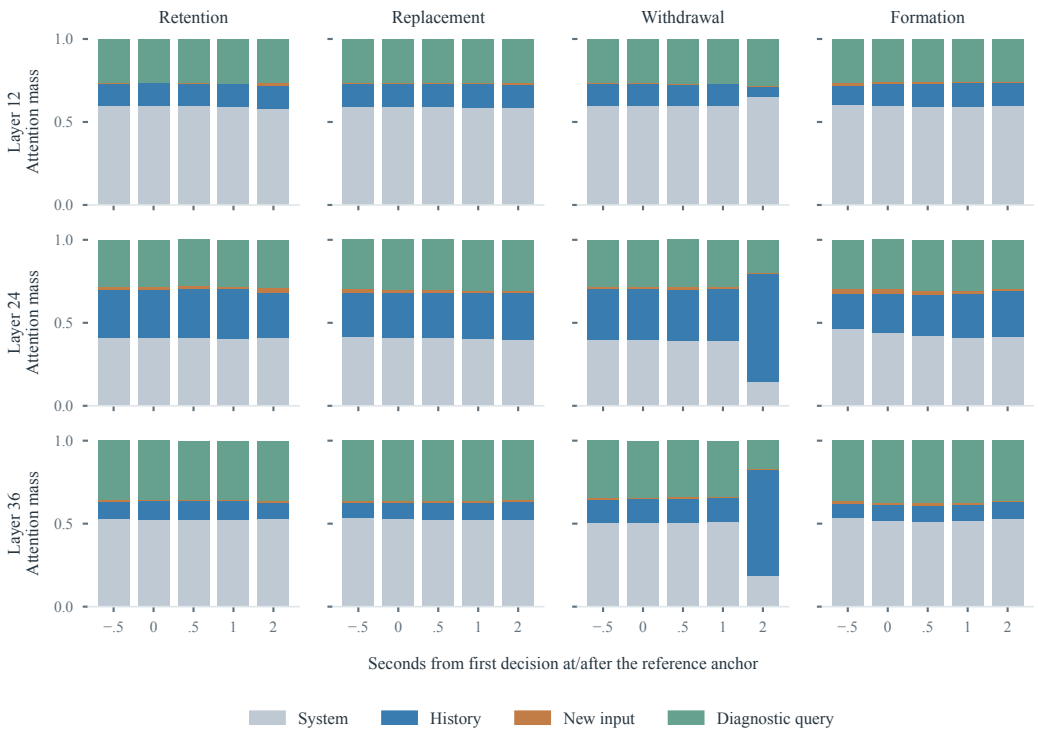


Figure 12: MiniCPM-o diagnostic-query attention on Core-400 under intact context. Each bar averages the mass assigned to four context groups at one layer and anchor-relative decision time. Layer and time remain separate; every contributing video has equal weight within a bar. The available videos vary by offset because probes remain within each episode.

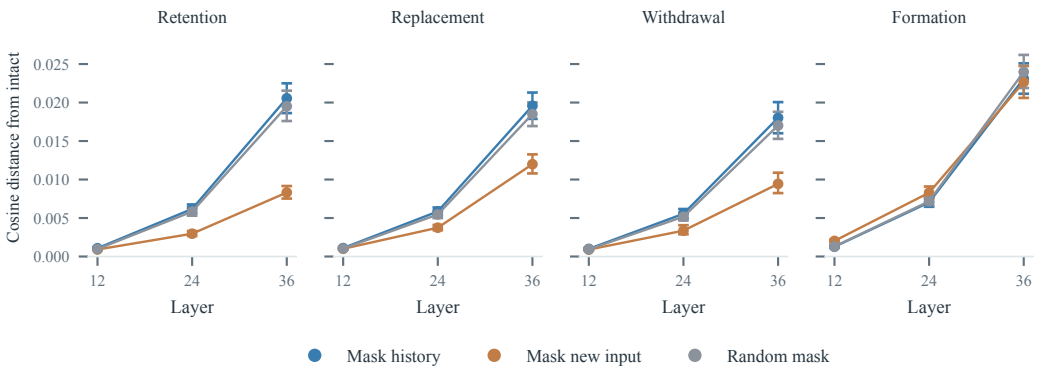


Figure 13: MiniCPM-o representation changes under context masking. Points show mean cosine distance from the same-time intact representation at each recorded layer; bars are 95% source-video bootstrap intervals. Each type contains 100 videos, with repeated times averaged within video.

H.3 Paired native readouts on event-centered clips

Paired cache probes. MiniCPM-o and MMDuet2 processed the same 300 Changed event-centered diagnostic trajectories through their native incremental interfaces, using audio-video and video input, respectively. Let $t_\star = 500\lceil t_R/500 \rceil$ ms. At $t_\star + \{-500, 0, 500, 1000, 2000\}$ ms within the observed decision range, a copy of the current key–value cache received the question and a short-answer instruction. Each model contributed 1,389 readouts. The complete pre-change and post-change reference answers were scored as candidate continuations. Their post-change-minus-pre-change mean token log likelihood defined the preference margin; the sequence-sum difference provided a length-sensitive comparison.

The intact cache was compared with three masks sampling equal numbers of historical, latest-input, or random contextual positions. The sample size was the minimum of 128 and the two temporal groups' token counts. Historical context comprised preceding media and text. Latest input comprised the current prefill's media, timing text, and question; the random control sampled their combined contextual positions. System and appended diagnostic-query positions remained accessible. Masks changed attention access during candidate scoring over fixed cached representations. All conditions used the same diagnostic attention kernel and physical token positions within each model; MMDuet2 also retained its native three-axis position offsets. Native reply generation continued from the intact session. Public-answer correctness referred to the answer exposed after that tick's action, including text carried through WAIT.

The primary contrast was history-mask minus latest-input-mask margin at $t_\star$, separately for each model and transition. It covered 100 Replacement, 100 Withdrawal, and 98 Formation trajectories per model; two Formation trajectories had an empty historical group. All 300 contributed intact readouts. Means and 95% percentile intervals used 10,000 source-video bootstrap samples. Two-sided paired source-video sign-flip tests used 10,000 draws with add-one correction and Holm adjustment across the six contrasts. Secondary controls and time comparisons used descriptive bootstrap intervals. Repeated ticks and heads were grouped within their source video. Likelihood units followed each model's tokenizer, and effects were evaluated within model.

Context sensitivity across transitions. MiniCPM-o's history-mask minus latest-input-mask margin was positive on Replacement and Withdrawal and negative on Formation. Withdrawal remained significant after six-comparison Holm adjustment (Table 23). MMDuet2's corresponding contrasts were $+0.003$, $+0.019$, and $+0.019$, with all three intervals spanning zero. On Withdrawal, its history, latest-input, and random-context masks changed the margin relative to intact context by $+0.0045$, -0.0141, and $+0.0007$, respectively.

Preference and public updating. At $t_\star$, MiniCPM-o preferred the updated reference on 47 Replacement, 11 Withdrawal, and 89 Formation trajectories; six, zero, and 24 of these also exposed a

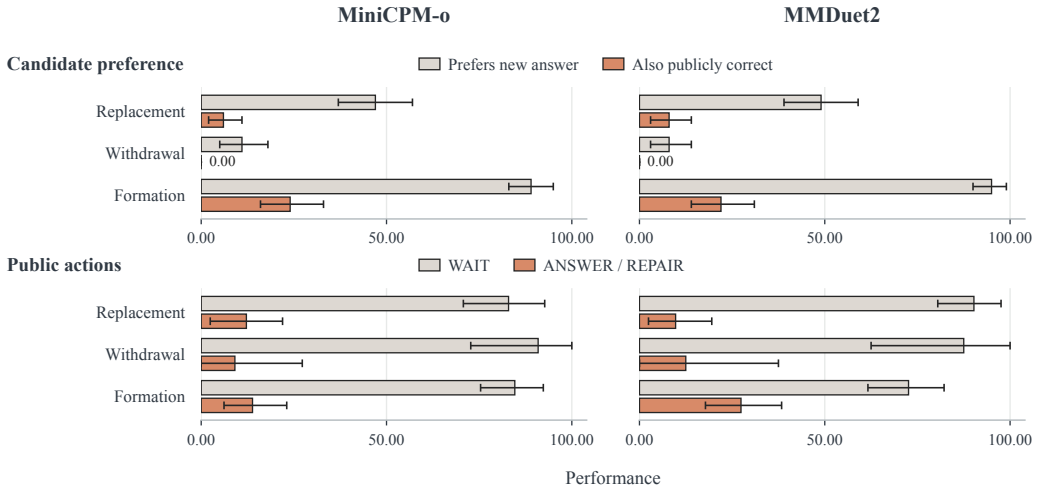


Figure 14: Candidate preference and public actions at $t_\star$. Top: updated-reference preference and its intersection with a correct public answer, each divided by 100 trajectories per transition. Bottom: action frequencies conditional on positive preference and an incorrect public answer. Conditional sample sizes are 41/11/65 for MiniCPM-o and 41/8/73 for MMDuet2, in displayed transition order. Parse errors remain in these denominators. All axes show percentages with 95% source-video bootstrap intervals within the respective cohorts.

Table 23: Context-mask effects at the first eligible post-change decision. The contrast is history-mask minus new-input-mask preference for the new reference, in mean token log-likelihood units. Intervals use 10,000 source-video bootstrap samples; adjusted p values use paired sign flips and Holm correction across all six comparisons.

Transition	n	Mean difference	95% CI	Adjusted p
MiniCPM-o				
Replacement	100	$+0.34818$	$[+0.07539, +0.65433]$	0.0900
Withdrawal	100	$+0.77614$	$[+0.57906, +0.97784]$	0.0006
Formation	98	-0.43856	$[-0.77258, -0.11392]$	0.0510
MMDuet2				
Replacement	100	$+0.00288$	$[-0.01070, +0.01717]$	0.6875
Withdrawal	100	$+0.01860$	$[-0.00003, +0.03741]$	0.1539
Formation	98	$+0.01899$	$[-0.00390, +0.04231]$	0.2294

correct public answer. MMDuet2's preference counts were 49, eight, and 95, with corresponding intersections of eight, zero, and 22. Each transition contained 100 trajectories.

Figure 14, bottom, conditions on positive preference and an incorrect public answer in the intact readout. The resulting cohort sizes are 41/11/65 for MiniCPM-o and 41/8/73 for MMDuet2. WAIT counts are 34/10/55 and 37/7/53; ANSWER or REPAIR counts are 5/1/9 and 4/1/20. The remaining two MiniCPM-o Replacement cases and one Formation case are parse errors, retained in the denominator. Descriptive 95% percentile intervals resample the distinct source videos within each fixed conditional cohort 10,000 times. Both action rates share the same resamples, with seeds 20260915–20260920 in model–transition order.

The top-row intersections differ from total public-answer accuracy: MMDuet2 exposes three semantically correct Withdrawal answers, none paired with positive preference. Public correctness also includes text carried through WAIT and does not require a valid initial answer or a qualifying REPAIR. Among Replacement trajectories with positive preference and an incorrect public answer at $t_\star$, 40/41 MiniCPM-o and 41/41 MMDuet2 trajectories remain incorrect at all three decisions through $t_\star + 1000$ ms. These observations use the same initial cohorts. All groups have 100 observations at

offsets 0, 500, and 1000 ms; at 2000 ms, only 83 Replacement, 13 Withdrawal, and 95 Formation trajectories remain observable per model.

At $t_\star - 500$ ms, before t_R, positive-margin counts were 46/100, 10/100, and 88/98 for MiniCPM-o, and 44/100, 9/100, and 94/98 for MMDuet2. Across the sampled grid, a nonpositive pre-boundary margin became positive at a post-boundary tick on 10/4/4 MiniCPM-o trajectories and 14/3/0 MMDuet2 trajectories. Much of the Formation preference was therefore present before annotated post-change support began.

Restricting the paired comparison to $t_\star - 500$ ms and $t_\star$ identifies two MiniCPM-o and six MMDuet2 Replacement trajectories with a nonpositive-to-positive margin change. Every one selected WAIT at $t_\star$ and exposed an incorrect answer at $t_\star$, $t_\star + 500$ ms, and $t_\star + 1000$ ms. The all-trajectory WAIT counts at $t_\star$ were 90/100 and 92/100. The corresponding counts among positive-preference, incorrect-answer trajectories were 34/41 and 37/41. Candidate-score changes and native actions were paired within source video; all Replacement trajectories meeting this two-tick sign-change definition were included.

Figure 15 shows these Replacement readouts around the transition.

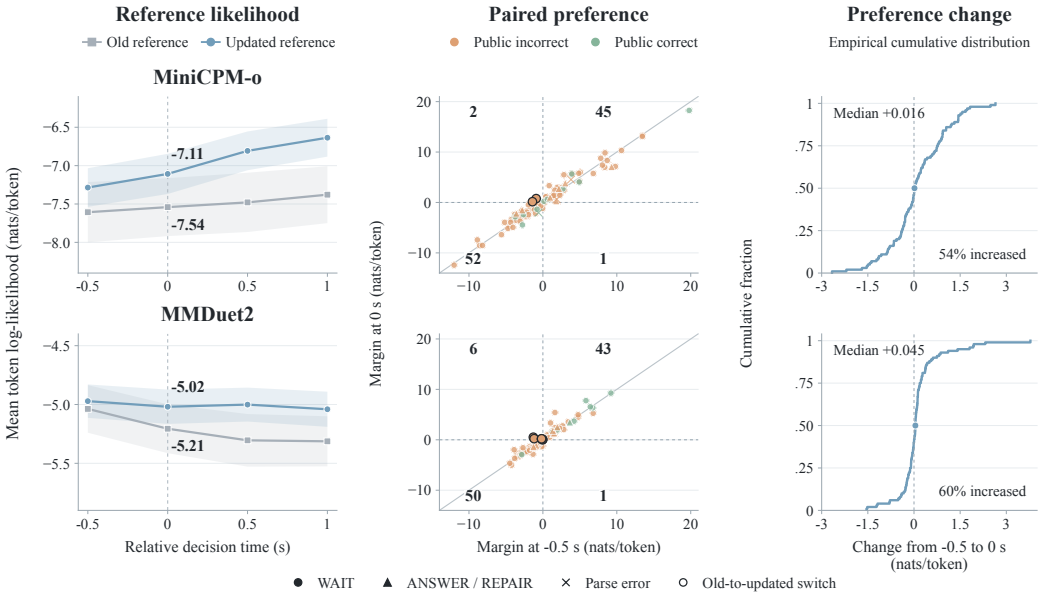


Figure 15: Updated-reference preference and public answers on the 100 Replacement clips per model. Left, mean token log-likelihood of the preceding and updated references around the transition. Middle, margin at the preceding decision against the margin at the first post-support decision; color gives public-answer correctness, the marker gives the native action, and an open marker is a video whose preference switched from the preceding to the updated reference. Right, empirical cumulative distribution of the margin change over the same interval.

Reference ranking also depends on length normalization. On Replacement, the signs of the mean-token and sequence-sum margins agree on 49/100 MiniCPM-o and 42/100 MMDuet2 trajectories. Sequence sums are length-sensitive; the primary readout retains the mean-token definition throughout.

Layer and time structure. Figure 16 summarizes the final diagnostic-query token's attention and hidden representations. For each trajectory and layer, attention density was the head-averaged mass divided by the corresponding context group's token count. The plotted quantity was the mean $\log_2$ ratio of latest-input to historical density. MiniCPM-o assigned greater density to recent positions in early and final layers and to history in intermediate layers; MMDuet2 assigned greater density to recent positions through most layers. At $t_\star$, the layer-18 mean log-ratios ranged from $+0.48$ to $+0.67$ across transitions for MMDuet2 and from -1.56 to -0.92 for MiniCPM-o.

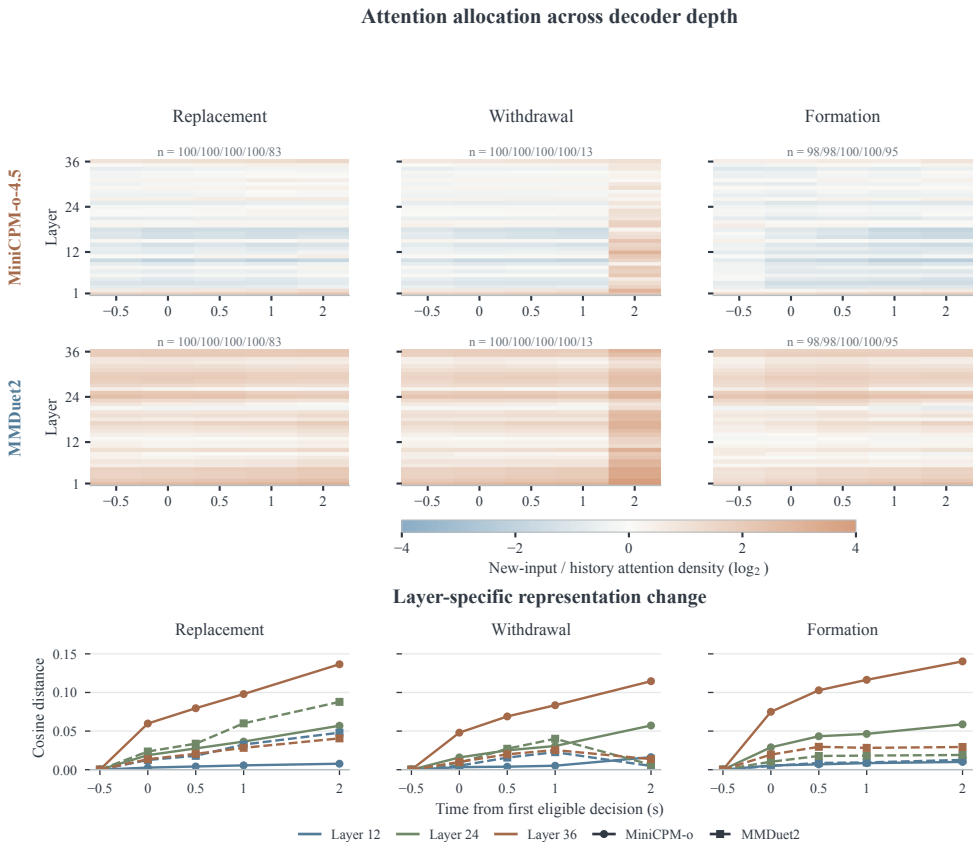


Figure 16: Decoder depth and temporal context in MiniCPM-o and MMDuet2. Top: mean log-ratio of new-input to historical attention density at the final diagnostic-query token, for all 36 decoder layers. Density is head-averaged attention mass divided by the group's available token count. Positive values indicate denser attention to recent input; heatmaps share a color scale. Bottom: mean cosine distance from the same trajectory's pre-boundary query representation at layers 12, 24, and 36; layer 36 includes final normalization. Colors identify layers, and line styles and markers identify models. Curves share a vertical scale. Both panels show descriptive source-video means; counts follow the five sampled times, with paired pre-boundary representations used for cosine distances.

Representation change was measured by cosine distance from the same trajectory's query-token state at $t_\star - 500$ ms. Among layers 12, 24, and 36, MiniCPM-o's mean distances were largest at layer 36 across all three transitions; MMDuet2's Replacement means were largest at layer 24. Layer 36 included final normalization.

I Cross-Model Semantic Agreement

The cross-model comparison samples 240 action-level items from the event-centered diagnostic cohort, 20 from each system-by-transition stratum of 1,284 MiniCPM-o, 416 MMDuet2, and 1,031 OmniLive public actions. Sampling targets ten Qwen-positive and ten Qwen-negative items per stratum; strata with fewer positives contribute every positive item and fill the remaining positions with negatives. GLM-5.3 (Z.ai, 2026) judges the blinded reference–prediction pairs under the same prompt and rubric. The analysis reports sample agreement, cross-model Cohen's κ_{cross}, inverse-sampling-weighted estimates for the 2,731-item source population, and endpoint sensitivity after substituting the sampled GLM-5.3 decisions. The unweighted sample coefficient is 0.861; applying the sampling weights gives 0.866. The weights correct unequal sampling probabilities across label strata. Human–Qwen agreement uses the separate notation κ_{human}. Table 24 reports agreement, and Table 25 reports the endpoint changes.

For either comparison, Cohen's coefficient is $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is observed agreement and p_e is agreement expected from the two raters' marginal label frequencies. For κ_{cross}, these frequencies use the stated sample or inverse-sampling weights. For κ_{human}, they use the 36 equally weighted human–Qwen pairs.

Table 24: Agreement between Qwen3.8-27B and GLM-5.3 on the stratified 240-item sample. The source-population and system rows apply the prespecified inverse sampling weights. The source-population interval is a 10,000-replicate trajectory-cluster bootstrap interval.

Scope	Items	Agreement	κ_{cross}	Label difference
Sample (unweighted)	240	93.33%	0.861	6.67%
Source population (weighted)	2,731	97.88% [96.24, 98.99]	0.866	2.12%
MiniCPM-o (weighted)	1,284	99.42%	0.959	0.58%
MMDuet2 (weighted)	416	88.57%	0.703	11.43%
OmniLive (weighted)	1,031	99.73%	0.951	0.27%

Table 25: Endpoint sensitivity after substituting the 240 sampled GLM-5.3 labels for the corresponding Qwen3.8-27B labels. Entries are percentage-point changes with 95% source-video-cluster bootstrap intervals. This panel covers formation, retention, and repair.

System	Initial	Retention@H	Repair@2s
MiniCPM-o	-0.75 [-1.75, 0.00]	-2.00 [-5.00, 0.00]	0.00 [0.00, 0.00]
MMDuet2	+0.25 [-0.75, 1.25]	-1.00 [-3.00, 0.00]	-0.33 [-1.00, 0.00]
OmniLive	-0.25 [-0.75, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]

I.1 Independent human semantic spot check

A human reviewer independently evaluated 36 action-level reference–prediction pairs, with three items drawn from each system-by-transition stratum. Each adjudication packet contained the question, reference answer, model answer, decision type, and timestamp. The Qwen3.8 label and rationale were hidden until the human decision was recorded. Human judgments agree with Qwen3.8 on 34/36 items (94.44%; Cohen's $\kappa_{\text{human}} = 0.875$). The two differences concern whether a static scene description satisfies a NONE reference and whether hand motion near an object entails movement of that object.

J Audio Contribution

Aero Realtime's audio-video and video-only conditions both complete 16/400 lifecycles, with 16/100 Retention successes and no Changed successes. Initial-answer success is 59/400 with audio and 60/400 with silence. The paired initial-answer success difference is -0.25 points (95% CI $[-1.25, 0.75]$), and the VAT difference is -0.06 points ($[-0.44, 0.38]$).

MiniCPM-o's audio-video condition scores 23.42 in Understanding, compared with 19.75 after removing audio. Its VAT is 14.62% with audio and 16.42% without it; lifecycle completion is 14/400 and 15/400, respectively. Audio thus raises Understanding without improving VAT or lifecycle completion.

K Input and Decision Frequency

MiniCPM-o is evaluated at 1 FPS with 1000 ms decisions, 2 FPS with 500 ms decisions, and 4 FPS with 250 ms decisions. Each condition uses 400 Core-400 trajectories; content and action probes retain the common four reference phases. The intervention changes input sampling and decision frequency together. VAT is 13.87%, 14.62%, and 16.11%, while LCS is 4.50, 3.50, and 2.75. Relative to 2 FPS/500 ms, the VAT differences are -0.75 points ($[-2.77, 1.28]$) and $+1.49$ points ($[-0.88, 3.78]$).

Table 26: Audio contribution on 400 matched trajectories per condition. Differences are audio-video minus video-only.

Metric	Audio-video	Video-only	Difference [95% CI]
Aero Realtime			
LCS	4.00	4.00	+0.00 [+0.00, +0.00]
Initial-answer success	14.75	15.00	-0.25 [-1.25, +0.75]
Stability	16.00	16.00	+0.00 [+0.00, +0.00]
Adaptivity	0.00	0.00	+0.00 [+0.00, +0.00]
VAT	11.17	11.23	-0.06 [-0.44, +0.38]
Understanding	19.25	19.08	+0.17 [-0.25, +0.58]
Temporal decision	10.98	10.98	+0.00 [+0.00, +0.00]
Overall	11.35	11.32	+0.03 [-0.12, +0.20]
MiniCPM-o			
Lifecycle Completion Score	3.50	3.75	-0.25 [-1.00, +0.50]
Initial-answer success	13.25	13.00	+0.25 [-0.75, +1.25]
Stability	16.00	16.00	+0.00 [-3.00, +3.00]
Adaptivity	0.00	0.67	-0.67 [-1.67, 0.00]
Valid Answer Time	14.62	16.42	-1.80 [-3.18, -0.49]
Understanding	23.42	19.75	+3.67 [+2.25, +5.17]
Temporal decision	19.35	20.21	-0.85 [-1.94, +0.19]
Overall	15.22	15.03	+0.19 [-0.44, +0.82]

Aero Realtime receives equal-duration silence; MiniCPM-o receives video with its audio input removed. Video and decision schedules are matched.

OmniLive's 1 FPS/1000 ms condition has LCS 0.25 and VAT 0.51%, compared with 0.00 and 0.37% at 2 FPS/500 ms. Tables 21 and 22 give the scores and paired differences.

MMDuet2's 1 FPS/1000 ms condition has LCS 3.00 and VAT 8.14%, compared with 3.25 and 8.95% at baseline. Understanding remains 21.33, and Temporal decision changes from 28.23 to 27.92. Its local-hold condition scores 20.83 and 27.60 on these two probe dimensions. The implementation settings are specified in Appendix H.1.

L Hosted Model Evaluation

L.1 Cohort and scoring

The hosted-model cohort is the complete Core-400 benchmark: 400 videos, 100 trajectories per lifecycle type, and 360 Perception Test / 40 CaptainCook4D videos, using the complete Core-400 questions and reference timelines. GPT-5.6 Terra, GPT-5.6 Luna, GPT-6 Luna, GPT-6 Sol, and Gemini 3.8 Flash each complete all 400 trajectories. Qwen3.5- and Qwen3.8-Omni-Flash-Realtime complete the same 399 videos after a provider content-filter rejection: 100 Retention, 100 Replacement, 99 Withdrawal, and 100 Formation trajectories.

Full trajectories and independent phase probes follow Appendix D. Understanding and Temporal decision average applicable probes within each trajectory, then average trajectories equally. Each GPT configuration and Gemini contribute 1,200 content and 1,552 applicable action probes, for 13,095 queries including 10,343 trajectory decisions; each Qwen Realtime configuration contributes 1,197 and 1,549, for 12,651 queries including 9,905 trajectory decisions. Open-text answers use deterministic Qwen3.8-27B semantic judgments, with thinking disabled and an 8192-token limit.

L.2 Input and generation settings

GPT image prefixes. All four GPT configurations receive a causal image prefix and text. Frames come from a 2 FPS cache, with up to 50 images per request: the eight most recent frames and 42 uniformly sampled from earlier history; shorter prefixes use all frames. Images have a 768-pixel long edge and JPEG quality 85. Each decision starts a fresh query whose prompt carries the preceding public answer.

GPT-5.6 Luna, GPT-5.6 Terra, GPT-6 Luna, and GPT-6 Sol use the corresponding model identifiers through Codex CLI (versions 0.155.1–0.157.0) with low reasoning effort; image detail and output allowance follow the client settings. Tools are disabled, as checked in the returned event records.

Gemini audio-video prefixes. The request identifier is `gemini-3.8-flash`. Each call receives the audio-video prefix ending at the query time. Video is assembled from a 2 FPS source cache, and the request specifies 1 FPS video processing with `MEDIA_RESOLUTION_LOW`. Audio is truncated at the same causal cutoff. Generation uses temperature 0, thinking level LOW, a 4096-token output limit, and JSON responses.

Qwen Realtime sessions. The request identifiers are `qwen3.5-omni-flash-realtime` and `qwen3.8-omni-flash-realtime`. Both use the same WebSocket protocol. A full trajectory maintains one incremental audio-video session, receiving a frame every 500 ms. Images use a 768-pixel long edge and initial JPEG quality 85, reduced when needed to meet a 250,000-byte Base64 image limit and a 300,000-byte event limit. Audio uses 16 kHz mono PCM, with each 500 ms interval sent in 100 ms chunks. Responses are manually triggered, text-only, and generated at temperature 0 with a 4096-token limit.

All configurations make decisions every 500 ms of media time and at the clip endpoint. Qwen advances the media clock after the preceding response completes; natural-trajectory frame delivery begins at 500 ms. An applicable zero-time probe uses the first frame. These scores measure behavior against media time. GPT and Gemini use a fresh prefix query for each probe; Qwen opens a fresh session and replays media to the probe time. Action probes carry their branch's own public answer between phases, while content probes query the current activity independently.

L.3 Paired comparisons

Tables 27 and 28 use 10,000 paired source-video bootstrap resamples, stratified by lifecycle type. The same sampled videos are used for both models and all metrics; Overall is recomputed from its four components in each resample. Intervals are 95% percentile intervals. At equal LCS, GPT-6 Sol's Overall exceeds Terra's by 1.97 points; it trails Gemini's by 2.08, with an interval spanning zero. Its differences from the two Luna configurations are larger.

Table 27: Paired score differences on shared videos. Values and 95% intervals are in points on the 0–100 scale; positive values favor the first named configuration.

Paired contrast	LCS difference [95% CI]	Overall difference [95% CI]
GPT-6 Sol − GPT-5.6 Luna	+7.50 [+3.50, +11.50]	+7.84 [+6.00, +9.70]
GPT-6 Sol − GPT-6 Luna	+11.50 [+7.50, +15.50]	+10.76 [+8.88, +12.81]
GPT-5.6 Luna − GPT-6 Luna	+4.00 [+0.50, +7.75]	+2.92 [+1.26, +4.56]
GPT-6 Sol − GPT-5.6 Terra	+0.00 [-3.75, +3.75]	+1.97 [+0.32, +3.64]
GPT-6 Sol − Gemini 3.8 Flash	-2.75 [-7.25, +1.75]	-2.08 [-4.30, +0.09]
Qwen3.8 Realtime − Qwen3.5 Realtime	+4.01 [+0.25, +7.77]	+11.35 [+9.43, +13.32]

Table 28: Qwen Realtime profile differences on the common 399 videos. Positive values favor Qwen3.8.

Metric	Difference [95% CI]	Metric	Difference [95% CI]
LCS	+4.01 [+0.25, +7.77]	Valid Answer Time	+12.65 [+9.42, +15.94]
Initial-answer success	+10.53 [+5.51, +15.54]	Understanding	+12.36 [+9.69, +15.04]
Stability	-9.00 [-19.00, +1.00]	Temporal decision	+16.37 [+14.06, +18.69]
Adaptivity	+12.04 [+8.36, +16.05]	Overall	+11.35 [+9.43, +13.32]

M Released Materials

The release records cohort definitions, model settings, trajectories, semantic judgments, and scoring code. Redistributable materials will be deposited in a persistent archive; source media remain under their original dataset terms.

N SPOT DETAILS

Objective and state pairing. For a trajectory with supervised set $\mathcal{S}$ of time–display instances, the loss averages the token-level cross-entropy of each target reply,

$$\mathcal{L} = \frac{1}{|\mathcal{S}|} \sum_{(t,p)\in\mathcal{S}} \frac{1}{|y^\star|} \sum_{j=1}^{|y^\star|} -\log P_\theta\big(y_j^\star \,\big|\, h_{<t}, x_t, q, p, y_{<j}^\star\big)\,, \tag{11}$$

where $h_{<t}$ is the native streaming history, x_t the media chunk at t, and q the question. Supervised times are lifecycle anchors on the 500 ms grid: the start, one second, the last decision before the evaluation boundary, the first decision after post-change support begins, and the last decision of the episode. Each epoch supervises 10,524 such anchors. A reference-following schedule supervises the teacher-consistent display $p_t^\star$, the answer a reference-following agent would show at t, twice per anchor. SPOT keeps this instance and pairs it with one alternative display $\tilde{p} \neq p_t^\star$, selected by a fixed hash and labeled $y^\star(r_t, \tilde{p})$. The paired displays comprise 3,256 incorrect answers drawn from the question's label vocabulary, 2,369 empty displays, 3,405 NONE displays, 1,111 preceding references, and 383 correct answers. Table 29 compares the resulting targets, which share the same number of supervision instances. Target replies are either `NO REPLY` or a reference label; the equivalence E in Section 5 uses exact matching on these controlled training labels, whereas evaluation judges free-text answers semantically.

Table 29: Targets per epoch under a reference-following schedule and SPOT, over the same 2,276 trajectories and 10,524 anchors. ANSWER and REPAIR are the protocol actions implied by each target reply and display.

	Reference-following	SPOT
Supervision instances	21,048	21,048
WAIT targets	18,092 (86.0%)	12,319 (58.5%)
ANSWER targets	116 (0.6%)	2,427 (11.5%)
REPAIR targets	2,840 (13.5%)	6,302 (29.9%)
Target tokens	111,120	126,000

Training data and replay. The training pool contains 2,276 trajectories from 1,000 Perception Test and 100 CaptainCook4D source videos, with 856 Retention, 334 Replacement, 628 Withdrawal, and 458 Formation trajectories; no source-video identifier overlaps with Core-400. The trajectories total 9.99 hours (median 6.79 s, maximum 430.60 s). Training replays each trajectory through MMDuet2's incremental interface at 2 FPS with one decision per frame and a persistent key–value cache. Paired instances share the teacher history and are scored without committing their replies; the teacher reply is then committed once. Each epoch replays 73,058 streaming turns; turns without supervision only extend the cached history.

Optimization. MMDuet2 is adapted with LoRA on the decoder self-attention query, key, value, and output projections (rank 16, $\alpha = 32$, no dropout; 7,372,800 trainable parameters). We use AdamW with a constant learning rate of 2×10^{-5} and gradient clipping at 1.0. The effective batch size is one trajectory for the first 1,792 visits and four thereafter, apart from two partial accumulations, giving 4,190 optimizer updates over five epochs (11,380 trajectory visits). The historical key–value cache is detached, so each supervised instance back-propagates through its current turn only; the cache itself still grows with the stream.

Training loss. Figure 17 shows the training loss. Over epochs 2–5, which share the same batch size, mean per-trajectory loss falls by 32.4%, from 0.0623 to 0.0421.

Core-400 evaluation. The trained model runs all 400 Core-400 trajectories once through MMDuet2's native interface with the settings of Section 3.4; semantic judging, lifecycle scoring, and phase probes are unchanged, and no checkpoint is selected on the test set. Original-model values are the MMDuet2 row of Table 2, and the right panel of Figure 4 breaks the change down by lifecycle type and repair stage. Stability is 27/100, whereas Retention completion, which enters LCS, is 26/100:

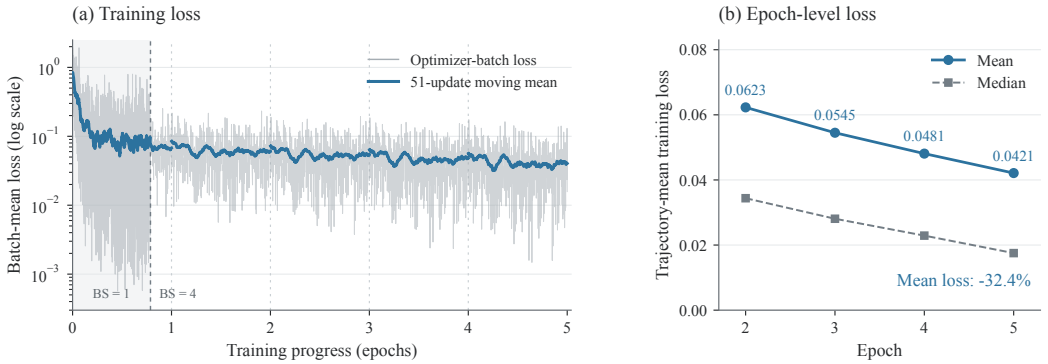


Figure 17: SPOT training loss. (a) Optimizer-batch means of per-trajectory losses (gray, log scale) and a 51-update moving mean computed within each epoch (blue); the effective batch size changes from 1 to 4 after 1,792 trajectory visits. (b) Mean and median per-trajectory losses for epochs 2–5, each with 2,276 visits at batch size 4.

in one trajectory an incorrect REPAIR is published exactly at $t_H + 2$ s, the endpoint excluded from the Stability window but included in the completion window. Timely correct repairs accumulate to 4, 6, 7, and 8 of the 300 Changed trajectories at deadlines of 0.5, 1, 1.5, and 2 s after post-change support begins.

StreamingBench Proactive Output. We evaluate the complete 250-question Proactive Output subtask through a causal-prefix polling interface shared by both models. Video is sampled at 2 FPS. From the query time, the model is polled once per second with at most 32 uniformly sampled frames available up to that time, starting each poll with an empty cache. The trigger prompt asks whether it is time to give the specified answer; the first reply containing "yes" triggers a content response at the same media time. Decoding is greedy with a 64-token limit. For $N = 250$ queries, let P count emitted first responses and TP count responses within ± 2 s of the annotated time. Timing precision is TP/P, timing accuracy is TP/N, and event-level timing F1 is $2TP/(N + P)$; the early-response rate uses all 250 queries as its denominator.

RIVER. We use the 551 RIVER Instant questions sourced from QVHighlights. Each model runs through its native incremental interface with a committed history, deciding once per second from the query time until the clip ends; the first non-WAIT reply is the notification. Content coverage is the percentage of all 551 questions whose notification a reference-text judge matches to the reference event, regardless of response time. Unanswered questions count as misses.

OVO-Bench CRR and SSR. We use the complete released Clues Reveal Responding (CRR, 240 evaluation points) and Sequential Steps Recognition (SSR, 629 points) subtasks. At each point, the model answers whether the clue has appeared or the step has been completed. No-F1 is the F1 score of "No" judgments against the annotated negatives, and the false-positive rate is the fraction of negative points answered "Yes" (96 in CRR and 206 in SSR). False positives fall from 41.67% to 18.75% in CRR and from 44.66% to 21.84% in SSR.