CHRONOSKV: DUAL-CLOCK TEMPORAL ENCODING FOR STREAMING VIDEO LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video LLMs process unbounded inputs by retaining a bounded visual cache and retrieving a sparse subset at question time. We demonstrate that this answer-time path suffers from a twofold information bottleneck: first, internal query–key attention locates the evidence little better than uniform sampling; second, standard position re-indexing preserves local token order while entirely collapsing absolute elapsed time. To resolve both limitations in a training-free framework, we introduce **ChronosKV**. ChronosKV pairs a lightweight contrastive vision-language retriever, delivering high-coverage evidence at constant query latency, with a **dual-clock temporal encoding**. By decoupling Rotary Position Embedding (RoPE) frequency bands band-by-band, ChronosKV routes high-frequency components to a compact ordinal track for local order, while steering low-frequency components to an absolute-time track for elapsed intervals. To evaluate what existing benchmarks miss, we introduce **Interval-QA**, a diagnostic suite isolating event order, recency, and relative metric distance. Across online streaming and long-video suites, ChronosKV attains the highest average accuracy of the compared systems, and restores the long-range elapsed time that previous position assignments discard.

1 INTRODUCTION

Video large language models (LLMs) answer questions about pre-recorded clips with striking competence (Li et al., 2025a; Bai et al., 2025; Zhang et al., 2025a), yet their most compelling applications, such as egocentric assistants, robotic perception, and live environmental monitoring, receive video as an unbounded stream. Because frames arrive continuously and questions can be posed at any moment, re-encoding the entire history becomes computationally prohibitive once it outgrows the context window. To circumvent this, a prominent line of work equips frozen video LLMs with a streaming KV cache (Song et al., 2024; Chen et al., 2024; Zhang et al., 2024a; Di et al., 2025; Chen et al., 2026b; Ning et al., 2025; Chen et al., 2026a). In this paradigm, video is ingested causally and exactly once into a bounded memory store. At question time, the system bypasses full re-reading by retrieving a sparse subset of cached entries, bounding per-question compute regardless of stream length. This cache-then-retrieve framework forms the basis of our study.

Our analysis reveals a fundamental bottleneck in streaming video architectures: video pipelines systematically lose temporal and contextual fidelity at answer time. During causal prefill at ingest, visual representations and their relative timing are preserved precisely, with each cache entry conditioned continuously on its preceding history. Information degradation occurs entirely during answer-time assembly, where a sparse subset of memory is extracted, scattered, and re-indexed into a prompt. This reconstruction process exposes two tightly coupled architectural dependencies: ***which*** historical entries are retrieved, and ***where*** they are mapped along the model position axis.

Consider first the retrieval step (Figure 1a,b). The prevailing strategy relies on **internal query–key attention**, scoring cached entries by how heavily the question attends to them within the LLM's native space (Di et al., 2025; Chen et al., 2026b;a). Because attending over every cached token at question time is too costly, systems pool each frame or block into a summary key. We discover that query attention over such pooled keys serves as a remarkably weak relevance judge: on our baseline pipeline (Chen et al., 2026a), it barely outperforms uniform random sampling in capturing annotated evidence windows. In contrast, an **external retriever** scoring frames against the question

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.

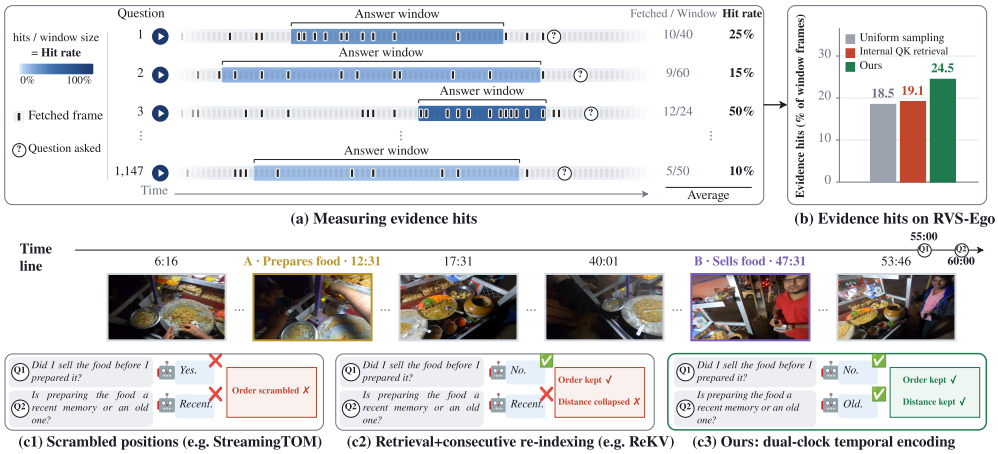


Figure 1: Where streaming video QA fails: retrieval precision and temporal positioning. **(a, b)** Evidence coverage on RVS-Ego (1,147 queries, 192 frames/query, the budget used throughout on RVS): internal QK scores 19.1%, barely outperforming uniform sampling (18.5%), whereas our external retriever achieves 24.5%. **(c)** Temporal assignment for events at 12:31 (A) and 47:31 (B), tested on order (Q1) and recency (Q2). **(c1)** Scrambled positions destroy order and distance. **(c2)** Consecutive re-indexing preserves order but collapses temporal scale. **(c3)** ChronosKV faithfully preserves both order and elapsed distance.

via a contrastive vision-language encoder captures substantially more evidence (Figure 1b) with negligible query-time overhead.

The second failure lies in **temporal positioning**, where elapsed time is entirely erased (Figure 1c). Suppose an agent prepares food at 12:31 (A) and sells it at 47:31 (B), and a user subsequently asks about their order (Q1) or recency (Q2). Existing paradigms offer broken compromises: **Scrambled positions** (used in our baseline pipeline (Chen et al., 2026a)) destroy even relative order (panel c1). Conversely, preserving true stream coordinates maintains order and distance, but can leave adjacent fetched entries far apart at gaps the backbone never encountered during training (Di et al., 2025), pushing spans past the trained context window on long streams. To avoid this, nearly all surveyed systems fall back to consecutive re-indexing (Di et al., 2025; Chen et al., 2026b; Pang et al., 2026; Zhang et al., 2026; Tiwari et al., 2026). While this preserves order, the 35-minute gap between events collapses into a single-step increment, indistinguishable from adjacent frames, leaving the model structurally blind to elapsed time (Q2, panel c2).

To resolve this dilemma, we introduce ChronosKV, anchored by **dual-clock temporal encoding** (panel c3). Because Rotary Position Embeddings (RoPE) map channel pairs across geometric frequencies, attention reads only relative phase differences. High-frequency bands resolve fine-grained local offsets, whereas low-frequency bands drift steadily over long horizons. ChronosKV assigns two position tracks to each retrieved entry: a compact index for fast bands that preserves token-level order, and a rescaled true arrival time for slow bands, so that position gaps follow true elapsed time. Because this modification alters only query-time position indices, backbone weights, model architecture, and stored KV caches remain untouched, making the framework training-free and plug-and-play for RoPE backbones.

Evaluating whether a repaired encoding truly restores temporal distance requires looking beyond existing streaming benchmarks. Current suites evaluate temporal competence almost exclusively in the ordering sense; even dedicated temporal-logic challenges confine themselves to relational operators such as before, after, until, and overlap (Xu et al., 2026), never addressing how far apart two events are. To bridge this evaluation gap, we introduce **Interval-QA**, a controlled diagnostic suite built on fifteen in-the-wild streams with densely annotated events. We hand every position assignment the same frames, those that show the events, so the layouts differ only in where these frames sit; what changes in the answers is then what the positions alone contribute to order, recency, and relative distance (§ 4).

Our contributions are fourfold.

- We demonstrate that internal QK retrieval performs comparably to uniform random sampling, and replace it with a lightweight external retriever that extracts significantly more evidence at negligible query latency.
- We propose **ChronosKV**, which couples this retriever with a dual-clock encoding that repairs the broken temporal coordinates of retrieved KV entries via frequency-decoupled position tracks, restoring order and elapsed time to the RoPE backbone without training.
- We release a rigorous diagnostic suite (**Interval-QA**) spanning fifteen in-the-wild streams and three novel question families (order, recency, and relative distance) to evaluate order and elapsed time on long streams.
- Evaluated against five prior streaming and offline systems under a unified protocol, ChronosKV achieves the highest average accuracy on both LLaVA-OneVision-7B and LongVA-7B; Interval-QA and component ablations separate what the retriever and the dual clock each contribute.

2 PRELIMINARIES

Streaming Ingest and Bounded Memory Construction. A video stream delivers frames $f_1, f_2, \ldots$ causally at a fixed rate. Following standard streaming frameworks (Chen et al., 2026a), each frame f_i is encoded into tokens z_i, compressed to a budget of G tokens via causal temporal reduction (CTR) to form a group g_i, and prefilled against prior context:

$$f_i \xrightarrow[\text{tower+proj}]{\text{encode}} z_i \xrightarrow[\text{keep } G]{\text{CTR}} g_i \xrightarrow[\text{prior context}]{\text{prefill}} \mathcal{M}_i = \{(K_i^\ell, V_i^\ell)\}_{\ell=1}^{L} \rightarrow \mathcal{M}.$$

This produces key-value matrices $K_i^\ell, V_i^\ell \in \mathbb{R}^{G\times d}$ per layer ℓ, attending within a sliding window with a sink prefix (Xiao et al., 2024b). Frame entry $\mathcal{M}_i$ is stored in a quantized active memory store $\mathcal{M} = \{\mathcal{M}_i\}_{i\le t}$ (t is elapsed frames) and is never re-encoded.

Question-Time Retrieval and Decoding. Upon query q at time t, streaming pipelines perform sparse cache retrieval (Di et al., 2025; Chen et al., 2026b;a). Standard internal QK retrieval scores query states $\bar{q}^\ell$ against summary cache keys $\bar{k}_i^\ell$, selects a top-B subset $\mathcal{R}^\ell$, and dequantizes entries with assigned positions p for final generation:

$$q \xrightarrow[\text{no visual tokens}]{\text{probe}} \{\bar{q}^\ell\}_{\ell\le L} \xrightarrow[s_i^\ell=\langle \bar{q}^\ell, \bar{k}_i^\ell\rangle]{\text{score}} \mathcal{R}^\ell = \text{top}_B\{s_i^\ell\}_{i\le t} \xrightarrow[\text{assign } p]{\text{dequantize}} \text{prefill}\big(q \,\big|\, \{\mathcal{M}_i\}_{i\in\mathcal{R}^\ell}\big) \rightarrow \text{answer}$$

where $\bar{q}^\ell$ denotes layer-ℓ query states averaged over question tokens, B defines a layer-wise budget, and p denotes the position mapping defined in Eq. (2).

Positional Mechanics and Baseline Limitations. RoPE (Su et al., 2024) encodes positions by rotating d-dimensional head pairs in $d/2$ complex planes at geometric frequencies $\theta_j = \theta_{\text{base}}^{-2j/d}$. Attention depends strictly on relative phase differences over the positional gap $m - n$:

$$\left\langle e^{\mathrm{i}m\theta_j} a,\ e^{\mathrm{i}n\theta_j} b \right\rangle = \mathrm{Re}\left[e^{\mathrm{i}(m-n)\theta_j}\, a\bar{b} \right]. \tag{1}$$

Sparse retrieval requires assigning an answer-time position $p(x)$ to every fetched token (x). Let $\tau(x)$ be the arrival time, $m(x)$ a scrambled index, and $\rho(x)$ the retrieval rank. Existing systems rely on three canonical assignments:

$$p(x) = \begin{cases} m(x) & \textit{scrambled} \text{ (order and distance lost)}, \\ \tau(x) & \textit{true} \text{ (faithful distance, but OOD gaps) (Han et al., 2024)}, \\ \rho(x) & \textit{ordered} \text{ (order kept, elapsed time collapsed)}. \end{cases} \tag{2}$$

While scrambled indices disrupt sequence order and true coordinates induce catastrophic out-of-distribution gaps, re-indexing collapses the temporal scale entirely.

Problem Definition. Given a streaming video backbone with memory $\mathcal{M}$ and query q, we seek a training-free operator $(q, \mathcal{M}) \mapsto (\mathcal{R}, \{p_j(x)\})$ that: (i) returns a retrieved set $\mathcal{R}$ that covers as much of the evidence as possible (Eq. 7) under strict latency budgets, and (ii) reconstructs spectral positions $\{p_j(x)\}$ to jointly preserve local token order and global temporal distances, without altering backbone weights, modifying active caches, or inducing phase aliasing.

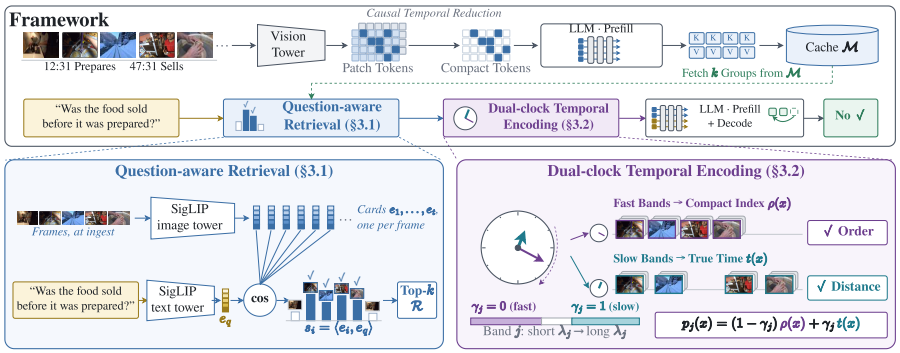


Figure 2: **Overview of ChronosKV. Top:** Ingest and quantized cache $\mathcal{M}$ remain untouched; retrieval and positional encoding intervene strictly at query time. **Bottom left:** A SigLIP vision tower encodes incoming frames into lightweight cards. At query time, text-to-card similarity retrieves the top-k frame groups from $\mathcal{M}$ in chronological order. **Bottom right:** The dual clock routes fast rotary bands to a compact ordinal track and slow bands to an absolute temporal track, preserving both local sequence order and global elapsed time.

3 Proposed Method

We introduce ChronosKV, a training-free framework that preserves standard streaming pipelines (Chen et al., 2026a) via two query-time modules (Figure 2): (i) Question-conditioned retrieval (§ 3.1), replacing internal QK scoring with a lightweight contrastive vision-language encoder; and (ii) Dual-clock temporal encoding (§ 3.2), superseding heuristic positions (Eq. 2) by splitting rotary embeddings into an ordinal and an absolute-time track. Operating exclusively during context assembly, ChronosKV leaves backbone weights and KV caches untouched.

3.1 Question-conditioned retrieval

Internal QK retrieval forces historical caches to evaluate their own relevance, introducing a fundamental tension between costly dense attention and severe information loss from pooled keys. We formulate relevance estimation as instance matching within a dedicated contrastive vision-language representation space.

A card per frame. Upon arrival and prior to temporal compaction, each raw frame f_i is mapped by the vision encoder of a dual-encoder model (SigLIP; Zhai et al., 2023) to a unit-norm embedding termed its *card*: $e_i = \phi_{\text{img}}(f_i)$. This lightweight card repository $\{e_1, \ldots, e_t\}$ constitutes the sole modification to the streaming ingest pipeline.

Matching at answer time. Given an incoming query q, we project it via the text encoder of the same SigLIP model: $e_q = \phi_{\text{txt}}(q)$. Relevance scoring is computed via direct inner product over the card repository:

$$s_i = \langle e_i, e_q \rangle, \quad \mathcal{R} = \text{top}_k\{s_i\}_{i \le t}, \tag{3}$$

where $\mathcal{R}$ denotes the indices of the k highest-scoring frames. The corresponding memory entries $\{\mathcal{M}_i : i \in \mathcal{R}\}$ are fetched, dequantized on demand, and reassembled chronologically for subsequent position assignment (§ 3.2).

Why an external judge. Offloading relevance estimation from internal attention yields two distinct structural advantages: (i) *One score per frame*: a unified visual card set serves all layers consistently, avoiding the multi-layer representation dispersion inherent to internal query probes; and (ii) *Superior localization*: external contrastive scoring captures a significantly larger proportion of ground-truth evidence windows (§ 5.3).

3.2 Dual-clock temporal encoding

The failure modes of standard position assignments (Eq. 2) stem from enforcing a uniform scalar coordinate $p(x)$ across heterogeneous rotary frequency bands. ChronosKV resolves this by estab-

lishing a frequency-domain mixture of position manifolds, enabling each rotary pair j to sample from the track dictated by its geometric wavelength $\lambda_j = 2\pi/\theta_j$.

Dual-frequency regimes. For two tokens separated by a true temporal interval $\Delta = \tau(y) - \tau(x)$ (such as the boundary gap between adjacent retrieved entries), rotary pair j undergoes a phase rotation $\Delta\varphi_j = 2\pi\Delta/\lambda_j$, splitting behavior into two distinct regimes:

$$\Delta\varphi_j = \frac{2\pi\Delta}{\lambda_j} \;:\quad \begin{cases} \text{aliased}, & \lambda_j \ll \Delta, \\ \text{monotone}, & \lambda_j \gg \Delta. \end{cases} \tag{4}$$

When $\lambda_j \ll \Delta$, high-frequency bands undergo severe phase wrapping (aliasing) across retrieval gaps, destroying local order signals. Conversely, slow bands ($\lambda_j \gg \Delta$) act as linear time-to-phase transducers, operating within a fraction of a single rotation cycle to directly reflect metric distance Δ. Consequently, fast bands require a dense, contiguous layout-preserving token order, whereas slow bands require faithful absolute arrival timestamps within pre-trained capacity bounds.

Dual tracks, one clock. ChronosKV constructs a piecewise smooth position mapping. Track one provides a *compact* layout $\rho(x)$ via consecutive reindexing (Eq. 2), restoring dense local intervals for high-frequency bands. Track two provides a scaled absolute arrival time:

$$t(x) \;=\; \alpha\,\tau(x), \qquad \alpha = \min\!\big(1,\; T_\alpha/\mathrm{span}\big), \qquad \mathrm{span} = \tau_{\max} - \tau_{\min}, \tag{5}$$

where $\tau_{\min}$ and $\tau_{\max}$ bound the retrieved subset. True metric separation is preserved when the span fits within target cap T_α, scaling linearly otherwise. The query coordinate anchors to the trailing context boundary.

Each rotary dimension pair dynamically blends these tracks based on its geometric wavelength:

$$p_j(x) \;=\; (1-\gamma_j)\,\rho(x) + \gamma_j\,t(x), \qquad \gamma_j = \mathrm{clip}_{[0,1]}\frac{\ln\lambda_j - \ln\Lambda_{\mathrm{lo}}}{\ln\Lambda_{\mathrm{hi}} - \ln\Lambda_{\mathrm{lo}}}\,. \tag{6}$$

Bands with wavelengths below Λ_{lo} access exclusively the compact ordinal track, those above Λ_{hi} rely solely on the absolute time track, and intermediate bands transition smoothly via a log-wavelength ramp. Hyper-parameters Λ_{lo}, Λ_{hi}, and T_α govern this spectral allocation (Appendix D).

Cost Analysis. At query time, each fetched key vector undergoes a lightweight, elementwise rotation to $p_j(x)$ instead of $\rho(x)$. Consequently, inference latency remains invariant, requiring zero modifications to model weights or active cache states.

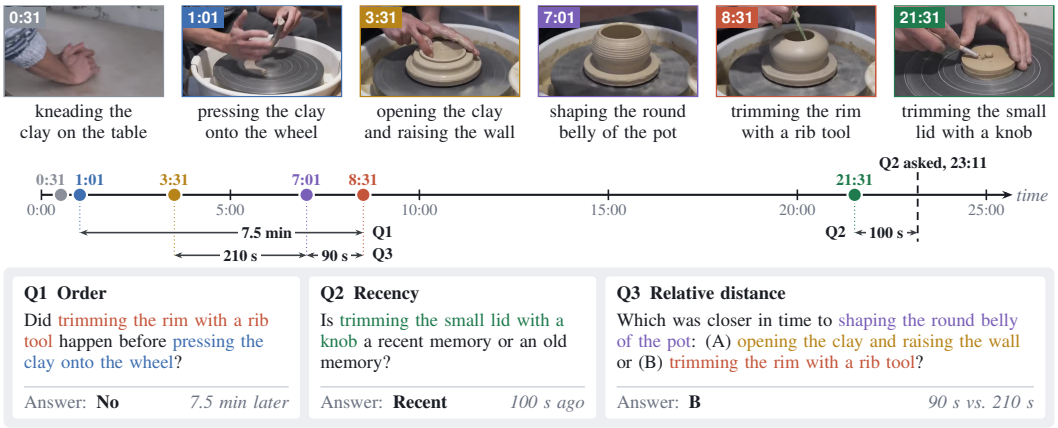


Figure 3: **Interval-QA evaluation on an in-the-wild video stream** (pottery, 25 min). Six annotated events are mapped to true timestamps alongside a multi-event evaluation query. *Order* queries evaluate relative precedence; *recency* queries measure proximity to the query horizon (dashed line); *relative distance* queries determine metric closeness relative to a reference anchor.

4 Interval-QA: A Challenging Benchmark for Elapsed Time

Existing streaming video benchmarks (RVS, StreamingBench, OVO-Bench) test content recognition and discrete event ordering. However, standard sequence-ordering tests remain invariant to temporal

scale: a model can perfectly order retrieved clips while having zero awareness of elapsed time. We introduce Interval-QA, a new benchmark that addresses this diagnostic gap.

Curation Principles. Interval-QA is formulated to evaluate whether a model can decode continuous temporal intervals from position geometry alone. Comprising 15 unscripted, long-form streams spanning 20 to 89 minutes, the benchmark features irregular event spacing and verified visual boundaries. To isolate positional encoding efficacy from retrieval variance, the comparison targets the positional layout alone under controlled inputs: an oracle provides identical frame snippets for each event window across all models, ensuring that retrieval cannot confound the evaluation.

Streams and Question Taxonomy. Our corpus comprises 15 unscripted video streams totaling 10.8 hours and 589 distinct events annotated by a frontier multimodal model(Claude), with a median event window of 12 seconds. Interval-QA constructs three distinct evaluation families:

(i) *Order:* Binary precedence queries (e.g., *"Did centering the pot occur before trimming the foot?"*), safeguarded by polarity-reversed controls.

(ii) *Recency:* Absolute temporal proximity tests differentiating immediate past events from deep history (≤ 2 minutes vs. ≥ 40 minutes separation).

(iii) *Relative Distance:* Triplets of a reference event and two candidates, one before and one after it, asking which candidate lies closer; order alone cannot answer, so the two elapsed distances must be compared.

Scope and utility. Interval-QA comprises 3,640 questions over 15 long videos (Table 1). Order queries probe sequence, whereas recency queries and relative-distance triplets probe elapsed time. Its dense event annotations also serve as ground-truth references for evaluating retriever coverage (§5.3).

Table 1: Composition of Interval-QA: video duration, event density, and query types.

Corpus		*Queries*	
Videos	15	Order	1,200
Length / video	20–89 min	Recency	1,648
Total length	10.8 h	Relative distance	792
Events	589	**Total**	**3,640**

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate across three video QA settings: (i) *Online streaming QA* uses RVS-Ego and RVS-Movie (Zhang et al., 2024a), where questions arrive at arbitrary timestamps and are answered from the cache built so far, following the ReKV protocol (Di et al., 2025) (judged by `gpt-3.5-turbo-0125`, reporting accuracy and the mean judge score). (ii) *Offline long-video QA* spans Video-MME (Fu et al., 2025) (w/o subtitles) and MLVU dev (Zhou et al., 2025), with videos fully ingested before querying. (iii) *Interval-QA* follows § 4, evaluated via exact-word matching against ground truth (Appendix A). Sections 5.2 and 5.3 report the public sets, except the coverage study of Table 4, which uses Interval-QA; § 5.4 is entirely on Interval-QA. Streams are sampled at 0.5 fps, dropping to 0.2 fps for Video-MME and MLVU videos exceeding 30 minutes; RVS and Interval-QA maintain 0.5 fps.

Baselines and Protocol. To ensure fair comparison, we independently rerun all baselines using official implementations under a unified setup[1]: ReKV (Di et al., 2025), LiveVLM (Ning et al., 2025), StreamingTOM (Chen et al., 2026a), HERMES-4K (Zhang et al., 2026), and DSCache (Pang et al., 2026). For offline benchmarks, plain backbones receive uniformly sampled frames via `lmms-eval` (32 frames for LLaVA-OneVision, 128 for LongVA). ChronosKV preserves baseline ingestion identically (G=50 tokens/frame post-compaction, a 14-token sink prefix, and 4-bit quantization), differing solely in query-time retrieval and dual-clock mechanisms (§ 3.1, 3.2), with the cache footprint matched to StreamingTOM and a retrieval budget never above its default. Reproducibility details are deferred to Appendix D.

[1]Streaming systems ingest at 0.5 fps under the ReKV evaluation protocol; online answers are scored by the `gpt-3.5-turbo-0125` judge, offline sets by multiple-choice accuracy.

Table 2: **Online and offline video QA.** *Online*: RVS-Ego and RVS-Movie; accuracy (%) with GPT score subscripts. *Offline*: Video-MME (Short, Med., Long, w/o subtitles) and MLVU (dev); accuracy (%). Avg. denotes the macro-average per setting.

			Online (RVS)			Offline				
Method	Retr.	Positions	Ego	Movie	Avg.	Short	Med.	Long	MLVU	Avg.
LLaVA-OV-7B										
Uniform (32 fr.)	–	–	–	–	–	70.33	56.56	48.67	64.58	60.04
ReKV	internal	consec.	$61.50_{4.06}$	$53.54_{3.59}$	$57.52_{3.82}$	70.78	55.11	48.56	68.32	60.69
LiveVLM	internal	original	$62.05_{4.05}$	$54.80_{3.64}$	$58.42_{3.84}$	68.56	55.33	50.11	67.08	60.27
HERMES-4K	none	consec.	$62.12_{4.07}$	$58.53_{3.70}$	$60.33_{3.89}$	69.56	54.67	48.44	63.77	59.11
DSCache	none	consec.	$61.98_{4.04}$	$49.71_{3.52}$	$55.84_{3.78}$	69.22	51.44	46.67	63.12	57.61
StreamingTOM	internal	scrambled	$61.71_{4.04}$	$58.79_{3.69}$	$60.25_{3.87}$	**70.89**	56.11	50.33	67.20	61.13
Ours	external	dual clock	$\mathbf{64.98_{4.10}}$	$\mathbf{63.15_{3.82}}$	$\mathbf{64.06_{3.96}}$	**70.89**	**58.11**	**51.00**	**68.49**	**62.12**
LongVA-7B										
Uniform (128 fr.)	–	–	–	–	–	62.44	**54.33**	45.56	59.06	55.35
StreamingTOM	internal	scrambled	$65.32_{4.02}$	$54.86_{3.51}$	$60.09_{3.76}$	**63.44**	51.78	44.22	59.20	54.66
Ours	external	dual clock	$\mathbf{67.24_{4.05}}$	$\mathbf{58.69_{3.62}}$	$\mathbf{62.96_{3.83}}$	63.33	53.00	**47.00**	**59.48**	**55.70**

Table 3: **Ablation of question-conditioned retrieval.** Accuracy (%) comparing internal QK versus external SigLIP retrieval at matched token budgets under a fixed dual-clock configuration.

		RVS		VM			
Model	Retrieval	Ego	Movie	Short	Medium	Long	Avg.
LLaVA-OV-7B	Internal QK	64.51	60.21	70.89	56.22	50.33	60.43
	SigLIP (ours)	$\mathbf{64.98}_{+0.5}$	$\mathbf{63.15}_{+2.9}$	$70.89_{=}$	$\mathbf{58.11}_{+1.9}$	$\mathbf{51.00}_{+0.7}$	$\mathbf{61.63}_{+1.2}$
LongVA-7B	Internal QK	66.28	58.06	63.33	52.78	44.89	57.07
	SigLIP (ours)	$\mathbf{67.24}_{+1.0}$	$\mathbf{58.69}_{+0.6}$	$63.33_{=}$	$\mathbf{53.00}_{+0.2}$	$\mathbf{47.00}_{+2.1}$	$\mathbf{57.85}_{+0.8}$

5.2 Main results

Online Streaming QA. ChronosKV outperforms all baselines on online streaming benchmarks (Table 2), averaging 64.1 vs. 60.3 for the strongest competitor (HERMES-4K). While baselines cluster tightly on RVS-Ego (61.5–62.1), ChronosKV advances to 65.0; on RVS-Movie, where baselines degrade sharply (49.7–58.8), ChronosKV maintains superior robustness (63.2). These gains transfer consistently to LongVA-7B, lifting its average from 60.1 to 63.0.

Offline Long-Video QA. On offline benchmarks (Table 2), ChronosKV averages 62.1, outperforming StreamingTOM (61.1) and ReKV (60.7), with a $+1.0$ point gain on LongVA-7B. While uniform-frame baselines remain competitive on shorter videos, where uniform sampling incurs little loss, ChronosKV achieves a higher overall average across both backbones while operating in an online stream. On Video-MME Long, where fixed frame budgets sample sparsely, ChronosKV surpasses uniform sampling by $+2.3$ and $+1.4$ points. On Video-MME Short, all models converge at 69–71%, as 1–2 minute clips fit entirely within the retrieval budget, leaving no deficit for retrieval or positioning to resolve.

5.3 Ablation Study

Impact of question-conditioned retrieval. Table 3 isolates the retrieval mechanism by swapping the external retriever for internal QK scoring while holding all else fixed. The SigLIP retriever consistently improves average accuracy on both backbones (+1.2 for LLaVA-OneVision, +0.8 for LongVA) across benchmarks with video durations exceeding the retrieval budget; on Video-MME Short, clips fit entirely within the budget, causing both retrievers to select identical frames.

Effect of Retriever Evidence Coverage. Interval-QA annotates the visibility window of every event; thus, retrieval recall can be measured as *evidence coverage*:

$$\mathrm{cov}(K) = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{|E(q)|} \sum_{e \in E(q)} \frac{|\mathcal{R}_K(q) \cap W_e|}{|W_e|}, \tag{7}$$

where Q is the set of questions, $\mathcal{R}_K(q)$ the K frames fetched for question q, $E(q)$ the events it designates, and W_e the temporal window of event e. Coverage is evaluated per event, whereas Figure 1 uses the per-question answer windows annotated in RVS-Ego. Internal QK retrieval scores frames independently across layers; we report layer-averaged results.

As shown in Table 4, internal QK retrieval tracks uniform sampling across all budgets, whereas the external retriever scales ahead from the smallest budget upward.
At the operational budget (K=192), it captures 45.6% of evidence frames, i.e., roughly triple the coverage of alternative baselines.

Table 4: Evidence coverage (%) on Interval-QA versus retrieval budget K (Eq. 7).

	Retrieval budget K				
Retriever	8	16	32	64	192
Uniform sampling	0.7	1.2	2.7	6.0	16.2
Internal QK (LLaVA-OV)	1.4	2.0	2.8	5.2	15.1
Internal QK (LongVA)	1.5	2.0	2.9	4.8	16.2
External SigLIP (ours)	**10.0**	**14.7**	**20.5**	**27.9**	**45.6**
vs. baseline	*×6.7*	*×7.4*	*×7.1*	*×4.7*	*×2.8*

5.4 Interval-QA: Dissecting Positional Encoding

Table 5 isolates the dual clock on our Interval-QA benchmark under a fixed frame budget; results on public streaming and offline splits appear in Appendix B.4.
Ordered positions restore sequence only. Consecutive re-indexing boosts the sequence order family on LLaVA-OneVision ($54.8 \rightarrow 62.5$) but leaves elapsed-time dimensions unchanged (recency: $50.4 \rightarrow 49.6$, relative distance: $51.9 \rightarrow 52.0$), confirming that sequence compaction alone discards absolute temporal gaps.

Table 5: **Interval-QA** accuracy (%) of four position-assignment strategies on identical retrieved frames.

Positions	Order	Recency	Rel. dist.	Avg.
LLaVA-OV-7B				
Scrambled	54.8	50.4	51.9	52.4
Ordered	62.5	49.6	52.0	54.7
True positions	54.4	56.7	49.9	53.7
Dual clock (ours)	**65.5**	**61.5**	**54.2**	**60.4**
LongVA-7B				
Scrambled	54.3	50.2	50.4	51.6
Ordered	56.7	53.1	51.0	53.6
True positions	53.8	55.4	50.5	53.2
Dual clock (ours)	**59.2**	**62.4**	**52.7**	**58.1**

The dual clock recovers elapsed time. The dual clock raises recency accuracy to 61.5 (LLaVA-OneVision) and 62.4 (LongVA), gains of $+11.9$ and $+9.3$ points over ordered positions, while preserving sequence order (65.5 vs. 62.5). Relative distance improves less (+2.2 and +1.7 points): the model must read two intervals and compare them, a harder readout than the single coarse judgment that recency requires, which dilutes the gain the time signal brings.

True global positions fall short. Raw global positions yield marginal recency gains at the cost of sequence order, trailing the dual-clock design by 5–7 points overall. Since LongVA operates comfortably within its 224k context window, this deficit is not a range limitation. Rather, it stems from high-frequency band aliasing: while true global distances preserve slow temporal bands (explaining the mild recency boost), retrieval gaps distort high frequencies. This removes the order signal and disturbs how the model reads the fetched entries, so every family suffers. The dual-clock formulation avoids this by pairing a dense local layout for fast bands with absolute temporal coordinates for slow ones.

5.5 Efficiency Study

Table 6 profiles RVS-Ego. The external retriever adds negligible overhead: its SigLIP pass ingests 33.8 fps (29.6 ms per frame), far above the 0.5 fps stream rate, so frames never queue. At query time, it adds only 10 ms to the first token and 30 ms to the full answer. Furthermore, the encoder consumes a static 1.7 GB per GPU, while each 2.3 KB frame card accounts for under 1% of the corresponding frame's key–value cache footprint.

Table 6: **Computational efficiency** on RVS-Ego (10 streams, 17,181 frames, 1,465 queries; LLaVA-OV-7B, one 80 GB GPU). Δ is ours minus internal QK.

Metric	Internal QK	**Ours**	Δ
Ingest (ms/frame)	23.4	29.6	+6.2
First token (s)	0.245	0.255	+0.010
Full answer (s)	0.83	0.86	+0.03
Memory (GB)	20.7	22.4	+1.7

6 Related Work

Streaming Video LLMs. While offline video models leverage ample compute for multi-pass keyframe selection or agentic re-encoding (Tang et al., 2025; Ataallah et al., 2024; Wang et al., 2024; 2025), streaming assistants operate under strict real-time constraints: continuous frame arrival, asynchronous queries, zero latency tolerance, and irreversible history pruning. Early streaming paradigms rely on fixed-size memory buffers (Song et al., 2024; Zhang et al., 2024a; Qian et al., 2024b) or heuristic triggers (Chen et al., 2024; Qian et al., 2025). We build upon visual KV-cache retention paired with question-time retrieval (Xiao et al., 2024a; Di et al., 2025; Chen et al., 2026b; Ning et al., 2025; Chen et al., 2026a; Zhang et al., 2026; Pang et al., 2026), decoupling inference complexity from stream length while preserving ingest dynamics.

Retrieval Signals. Sub-second relevance scoring demands robust token matching. Standard pipelines rely on internal query–key attention over pooled cache keys (Di et al., 2025; Chen et al., 2026b;a), a zero-overhead signal. Although ReKV notes parity between internal attention and external CLIP scoring, and offline pipelines, which face no latency budget, default to external vision-language encoders (Ataallah et al., 2024; Wang et al., 2024), prior work evaluates end accuracy without diagnosing the underlying retrieval mechanism. We bridge this gap by measuring whether internal QK scores localize annotated evidence windows, and find that they do so little better than uniform sampling. By offloading relevance to a contrastive vision-language space (SigLIP), we decouple perceptual matching from causal generation within strict streaming budgets.

Rotary Positional Mechanics. RoPE (Su et al., 2024) maps positions via channel phase rotations $\Delta\phi_j = 2\pi\Delta/\lambda_j$: high-frequency bands ($\lambda_j \ll \Delta$) track local order, while low-frequency bands ($\lambda_j \gg \Delta$) capture long-range span. Because frozen backbones degrade outside training bounds (Han et al., 2024), context extension typically employs global interpolation (Chen et al., 2023), wavelength ramps (Peng et al., 2024), dimension scaling (Ding et al., 2024), or neighborhood remaps (Jin et al., 2024; An et al., 2024). Retrieval introduces a conflicting constraint as retrieved frames span temporal gaps far exceeding training horizons. While LiveVLM preserves raw global coordinates (Ning et al., 2025), most frameworks re-index entries into a compact, order-preserving range (Di et al., 2025; Chen et al., 2026b; Pang et al., 2026; Zhang et al., 2026; Tiwari et al., 2026), sacrificing absolute time for local order. Rather than introducing a novel encoding, ChronosKV exploits spectral decoupling: fast bands map to a compact index for local order, while slow bands ingest absolute arrival time, aligning with video-specific rotary designs that assign temporal axes to low-frequency pairs to prevent long-horizon aliasing (Wei et al., 2025; Li et al., 2025b).

Temporal Benchmarks. Existing benchmarks measure broad comprehension (Zhang et al., 2024a; Lin et al., 2026; Niu et al., 2025; Fu et al., 2025; Zhou et al., 2025; Mangalam et al., 2023; Wu et al., 2024), event sequencing (Liu et al., 2024; Zhang et al., 2024b; Shangguan et al., 2025; Swetha et al., 2025; Xu et al., 2026), or needle-in-a-haystack depth (Zhang et al., 2025a). Suites targeting chronological reasoning, such as temporal grounding (Gao et al., 2017; Krishna et al., 2017) or unconstrained timestamp queries (Chandrasegaran et al., 2024; Grauman et al., 2022; Yang et al., 2025), rely on specialized timestamp tokens or grounding heads (Ren et al., 2024; Huang et al., 2024a; Qian et al., 2024a; Huang et al., 2024b). To our knowledge, no benchmark isolates whether a general streaming backbone's intrinsic rotary positions encode temporal distance. Interval-QA (§ 4) fills this void by evaluating precision-interval decoding from position geometry alone.

7 Conclusion

We present ChronosKV, a training-free framework for streaming video LLMs that pairs an external contrastive retriever, which finds far more of the evidence than internal QK scoring, with a dual-clock temporal encoding that restores elapsed time to the retrieved entries. By decoupling relevance offloading from positional assembly, ChronosKV maximizes evidence coverage under latency constraints while reconciling local sequence order with global elapsed time. Extensive evaluations show that ChronosKV establishes state-of-the-art results across benchmarks, proving that frozen backbones harbor untapped temporal awareness for scalable, long-context video understanding.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing the method and evaluation code, formatting the dataset under our guidance (Section 4), refining research hypotheses, interpreting results, editing text for readability, and generating figure scripts. We did not use generative AI tools to conceive the core research idea, design experiments, or formulate the theoretical framework. The authors independently verified and tested all AI-assisted code and thoroughly reviewed and revised it, taking full responsibility for the final text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The code and the Interval-QA benchmark (video links, event annotations, and questions) will be released upon publication. The evaluation protocol and the hyperparameters of every reported number are given in Appendices E and D; the construction and scoring of Interval-QA are described in § 4 and Appendix A.

ETHICS STATEMENT

This work uses public video benchmarks and, for Interval-QA, publicly available YouTube videos that we annotated ourselves; no new data were collected from people, and the benchmark is released as video links and annotations, leaving the videos with their uploaders. The method is training-free and adds no new data to the backbones, so it inherits their known limitations and biases without introducing new ones. We see no direct negative societal application beyond those of streaming video understanding in general.

REFERENCES

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. Training-free long-context scaling of large language models. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.17463.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos. In *ECCV*, 2024. arXiv:2407.12679.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-hour video-language understanding. In *NeurIPS Datasets and Benchmarks*, 2024. arXiv:2411.04998.

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 *CVPR*, 2024.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. StreamingTOM: Streaming token compression for efficient video understanding. In *CVPR*, 2026a. arXiv:2510.18269.

Yilong Chen, Xiang Bai, Zhibin Wang, Chengyu Bai, Yuhan Dai, and Ming Lu. StreamKV: Streaming video question-answering with segment-based KV cache retrieval and compression. In *AAAI*, 2026b.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video KV-cache retrieval. In *ICLR*, 2025.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.13753.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *CVPR*, 2025. arXiv:2405.21075.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *ICCV*, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. LM-Infinite: Zero-shot extreme length generalization for large language models. In *NAACL*, 2024.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *CVPR*, 2024a. arXiv:2311.18445.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. LITA: Language instructed temporal-localization assistant. In *ECCV*, 2024b. arXiv:2403.19046.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. LLM maybe LongLM: Self-extend LLM context window without tuning. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2401.01325.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. arXiv:2408.03326.

Haoran Li, Yingjie Qin, Baoyuan Ou, Lai Xu, and Ruiwen Xu. HoPE: Hybrid of position embedding for long context vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b. arXiv:2505.20444.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. In *ICASSP*, pp. 12147–12151, 2026. arXiv:2411.03628.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of ACL*, 2024.

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

Zhenyu Ning, Guangda Liu, Qihao Jin, Chengwei Li, Wenchao Ding, Minyi Guo, and Jieru Zhao. LiveVLM: Efficient online video understanding via streaming-oriented KV cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025.

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 *CVPR*, 2025.

Zhanzhong Pang, Dibyadip Chatterjee, Fadime Sener, and Angela Yao. Decouple and cache: KV cache construction for streaming video understanding. *arXiv preprint arXiv:2605.01858*, 2026.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.00071.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *International Conference on Machine Learning (ICML)*, 2024a. arXiv:2402.11435.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. arXiv:2405.16009.

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 *CVPR*, 2025. arXiv:2501.03218.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, 2024. arXiv:2312.02051.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *ICLR*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, et al. MovieChat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Sirnam Swetha, Hilde Kuehne, and Mubarak Shah. TimeLogic: A temporal logic benchmark for video QA. *arXiv preprint arXiv:2501.07214*, 2025.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2502.21271.

Santiram Tiwari, Nishant Sinha, and Kunal Kislay. Chronostitch: Training-free composition of visual KV memories for long-horizon temporal reasoning. *arXiv preprint arXiv:2607.19547*, 2026.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *ECCV*, 2024. arXiv:2403.10517.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *CVPR*, 2025. arXiv:2405.19209.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. VideoRoPE: What makes for good video rotary position embedding? In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.05173.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS Datasets and Benchmarks*, 2024.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. InfLLM: Training-free long-context extrapolation for LLMs with an efficient context memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. arXiv:2402.04617.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, 2024b.

Zhaoyang Xu, Xusheng He, Wei Liu, Zhenyang Li, and Jianlong Wu. Timelogic challenge @ CVPR 2026: Strong MLLMs meet evidence-seeking agents for temporal-logic video question answering. *arXiv preprint arXiv:2606.01631*, 2026.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. EgoLife: Towards egocentric life assistant. In *CVPR*, 2025. arXiv:2503.03803.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-VStream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024a.

Haowei Zhang, Shudong Yang, Jinlan Fu, See-Kiong Ng, and Xipeng Qiu. HERMES: KV cache as hierarchical memory for efficient streaming video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2601.14724.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing LMMs over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024b.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2025a. arXiv:2406.16852.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b. arXiv:2410.02713.

Junjie Zhou, Yan Shu, Bo Zhao, et al. MLVU: Benchmarking multi-task long video understanding. In *CVPR*, 2025. arXiv:2406.04264.

APPENDIX

A INTERVAL-QA PROTOCOL DETAILS

Evaluation protocol. Interval-QA follows the online evaluation pipeline under the ReKV evaluation framework: streams are ingested at 0.5 fps, queries are posed at their corresponding temporal timestamps as bare prompts enforcing a strict answer format, and decoding proceeds greedily. Because all positional configurations share an identical set of retrieved frames, performance variations isolate positional encoding effects exclusively. Questions span three core diagnostic axes, order (*yes/no*), recency (*recent/old*), and relative distance (*A/B*), and are evaluated via exact-word string matching against ground-truth labels without relying on an external LLM judge.

Temporal distribution across question families. Figure 4 illustrates the temporal intervals characterizing each family. The order benchmark spans event pairs separated by 2 to 84 minutes (median 12.5 minutes), probing both proximal and distal interactions. The recency benchmark deliberately separates event pairs by an order of magnitude (2 minutes versus $40+$ minutes). Because the mapping from descriptive terms (*recent*, *old*) to absolute elapsed time is intrinsic to the backbone, precise duration decoding is inherently bounded by the frame rate and the pre-training position range. Consequently, a coarse temporal gap provides the most direct probe of whether models utilize global clock signals; this structural separation is clearly reflected in model performance (Table 5). Conversely, the relative-distance benchmark evaluates proximity relative to the past rather than absolute time, positioning candidate events within tighter windows. While conceptually straightforward, this family presents two inherent challenges: pairwise distance comparison is inherently more difficult than binary classification, and slow rotary bands, optimized for extensive spans, exhibit minimal phase rotation over shorter intervals, leading to lower absolute accuracies across all variants.

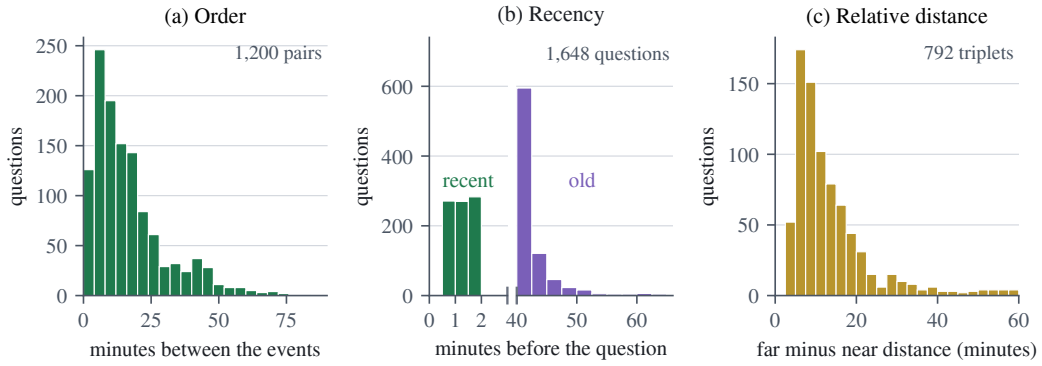


Figure 4: **Temporal distribution across Interval-QA evaluation families.** **(a)** Inter-event duration (minutes) for event-order queries. **(b)** Elapsed time (minutes) separating target events from the query timestamp for recent versus old instances in recency queries. **(c)** Absolute difference in temporal separation (minutes) between distant and proximal candidates relative to a reference event in relative-distance queries.

B Extended Experiments

This section provides supplemental evaluations, including phrasing robustness tests, qualitative visual comparisons, analyses of common failure modes, and broad positional ablations across streaming and offline benchmarks.

B.1 Reverse polarity on the order family

To verify robustness against phrasing bias, we evaluate an inverted-polarity formulation for every order query (asking *after* rather than *before*, with labels flipped) across the same 1,200 pairs (Table 7). While baseline ordered positions degrade by 5.7 points under polarity reversal, the dual-clock formulation exhibits superior stability, dropping by only 2.5 points. This demonstrates that anchoring slow temporal frequencies substantially reinforces structural sequence signals against linguistic variations.

Table 7: **Polarity robustness on the order family** (LLaVA-OneVision, 1,200 questions per split). Accuracy (%) when the same temporal relation is phrased forward (*before*) or inverted (*after*); Δ is the change from *before* to *after*.

	Query polarity		
Positions	*before*	*after*	Δ
Scrambled	54.8	51.8	−3.0
Ordered	62.5	56.8	−5.7
True positions	54.4	55.3	+0.9
Dual clock (ours)	**65.5**	**63.0**	−2.5

B.2 Qualitative Study

Figure 5 presents qualitative samples across three video streams, one question from each family, with the answers of the four positional layouts on LLaVA-OneVision. In these examples the dual clock answers all three families correctly. Consecutive re-indexing preserves order but not elapsed time: on recency and relative distance it gives the same answers as the scrambled layout. True global positions retain some of the distance information but lose order and are less stable.

B.3 Failure Modes Analysis

Figure 6 highlights failure cases common to all layouts—typically arising when target events share visually indistinct semantics (e.g., nearly identical cabin perspectives or subtle artistic shifts). Because these errors occur uniformly across all position layouts, failures stem from perceptual ambiguity rather than positional degradation, confirming that visual evidence grounding remains a primary frontier for future improvement.

B.4 Position ablation on streaming and offline benchmarks

Table 8 isolates the impact of positional configurations under a fixed external retriever. Across both backbones, performance gains over scrambled baselines derive primarily from restoring sequence order. Incorporating the slow temporal clock yields minimal variation on standard benchmarks, as these suites do not explicitly test absolute elapsed time; consequently, the dual-clock and ordered-position configurations perform comparably. The distinct advantages of the dual-clock architecture manifest specifically in long-range temporal reasoning tasks (Table 5).

Table 8: **Positional ablation** across streaming and offline benchmarks, reporting accuracy (%). All variants employ an identical external retriever, differing solely in the position assignment strategy; Avg. denotes macro-averaging across all five evaluation suites.

Backbone	Positions	RVS-Ego	RVS-Movie	VM-Short	VM-Medium	VM-Long	Avg.
LLaVA-OV-7B	Scrambled	61.50	61.47	70.89	56.33	50.22	60.08
LLaVA-OV-7B	Ordered	64.64	62.57	70.89	58.11	50.78	61.40
LLaVA-OV-7B	Dual clock (ours)	**64.98**	**63.15**	70.89	**58.11**	**51.00**	**61.63**
LongVA-7B	Scrambled	65.05	54.33	63.33	51.00	43.33	55.41
LongVA-7B	Ordered	66.21	57.38	63.33	**53.11**	**47.11**	57.43
LongVA-7B	Dual clock (ours)	**67.24**	**58.69**	63.33	53.00	47.00	**57.85**

C Textual timestamp injection baseline

As a training-free alternative to structural positional modification, we explicitly evaluate prepending textual timestamps to prompts, using the format introduced by LLaVA-Video (Zhang et al., 2025b). Under this protocol, questions are preceded by explicit string encodings denoting total stream duration, frame count, and absolute timestamps for each retrieved frame:

Textual time instruction prompt template

The video lasts for 2460.00 seconds, and 10 frames are sampled from it. These frames are located at 28.00s, 30.00s, …, 2402.00s, 2404.00s. Please answer the following questions related to this video.

As shown in Table 9, textual timestamp injection yields no consistent advantage over standard ordered positions (averaging 54.6 versus 54.7), exhibiting minor gains in recency alongside regressions in relative distance. Stating the time in words thus gives a general-purpose backbone no consistent gain, whereas the dual clock does, staying ahead on all three families (60.4 on average).

Table 9: **Performance comparison of textual timestamp injection** on Interval-QA (LLaVA-OV-7B), evaluating exact match accuracy (%) over identical retrieved frame subsets.

Positions	Order	Recency	Rel. distance	Avg.
Ordered	62.5	49.6	52.0	54.7
Ordered + text timestamps	62.1	51.9	49.9	54.6
Dual clock (ours)	**65.5**	**61.5**	**54.2**	**60.4**

Painting stream

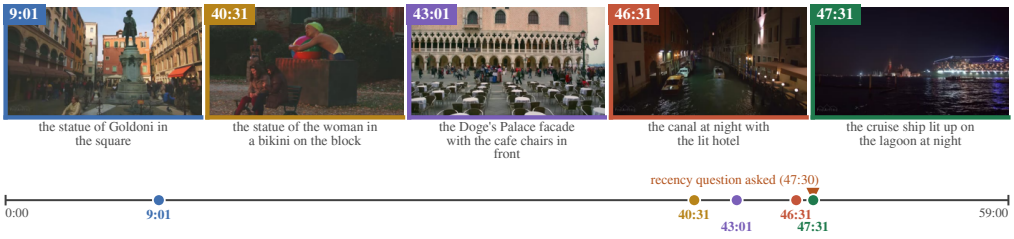


Order (asked at 66:06). Did holding up the purple space book happen before the host raising both hands?
Answer: Yes **ChronosKV: Yes** ✓ scrambled: No ✗ ordered: Yes ✓ true positions: No ✗

Recency (asked at 48:00). Is holding up the purple space book a recent memory or an old memory?
Answer: Old **ChronosKV: Old** ✓ scrambled: Recent ✗ ordered: Recent ✗ true positions: Old ✓

Relative distance (asked at 66:06). Which of these two happened closer in time to the full-canvas close-up with white unpainted eyes: (A) holding up the purple space book or (B) blending thick strokes across the forehead close-up?
Answer: B **ChronosKV: B** ✓ scrambled: A ✗ ordered: A ✗ true positions: B ✓

City-walk stream

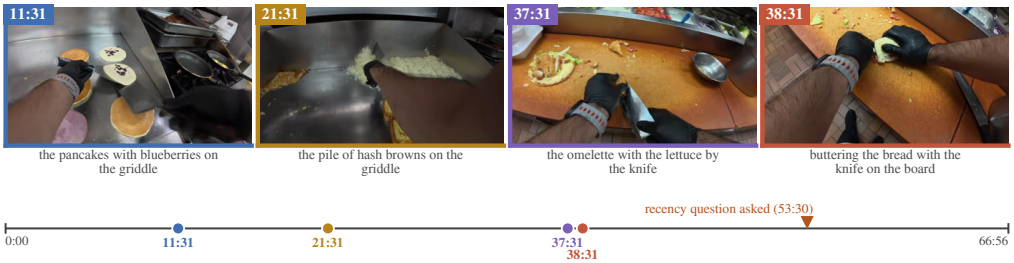


Order (asked at 59:00). Did the statue of Goldoni in the square happen before the cruise ship lit up on the lagoon at night?
Answer: Yes **ChronosKV: Yes** ✓ scrambled: No ✗ ordered: Yes ✓ true positions: Yes ✓

Recency (asked at 47:30). Is the canal at night with the lit hotel a recent memory or an old memory?
Answer: Recent **ChronosKV: Recent** ✓ scrambled: Old ✗ ordered: Old ✗ true positions: Old ✗

Relative distance (asked at 59:00). Which of these two happened closer in time to the statue of the woman in a bikini on the block: (A) the Doge's Palace facade with the cafe chairs in front or (B) the statue of Goldoni in the square?
Answer: A **ChronosKV: A** ✓ scrambled: B ✗ ordered: B ✗ true positions: A ✓

Kitchen stream



Order (asked at 66:56). Did the omelette with the lettuce by the knife happen before the pile of hash browns on the griddle?
Answer: No **ChronosKV: No** ✓ scrambled: Yes ✗ ordered: No ✓ true positions: Yes ✗

Recency (asked at 53:30). Is the pancakes with blueberries on the griddle a recent memory or an old memory?
Answer: Old **ChronosKV: Old** ✓ scrambled: Recent ✗ ordered: Recent ✗ true positions: Old ✓

Relative distance (asked at 66:56). Which of these two happened closer in time to the omelette with the lettuce by the knife: (A) the pile of hash browns on the griddle or (B) buttering the bread with the knife on the board?
Answer: B **ChronosKV: B** ✓ scrambled: A ✗ ordered: A ✗ true positions: B ✓

Figure 5: **Qualitative comparison of positional layouts on Interval-QA** (LLaVA-OneVision). Each block illustrates a video stream evaluated across all three diagnostic families, with color-coded markers linking individual frames to their respective temporal coordinates on the stream.

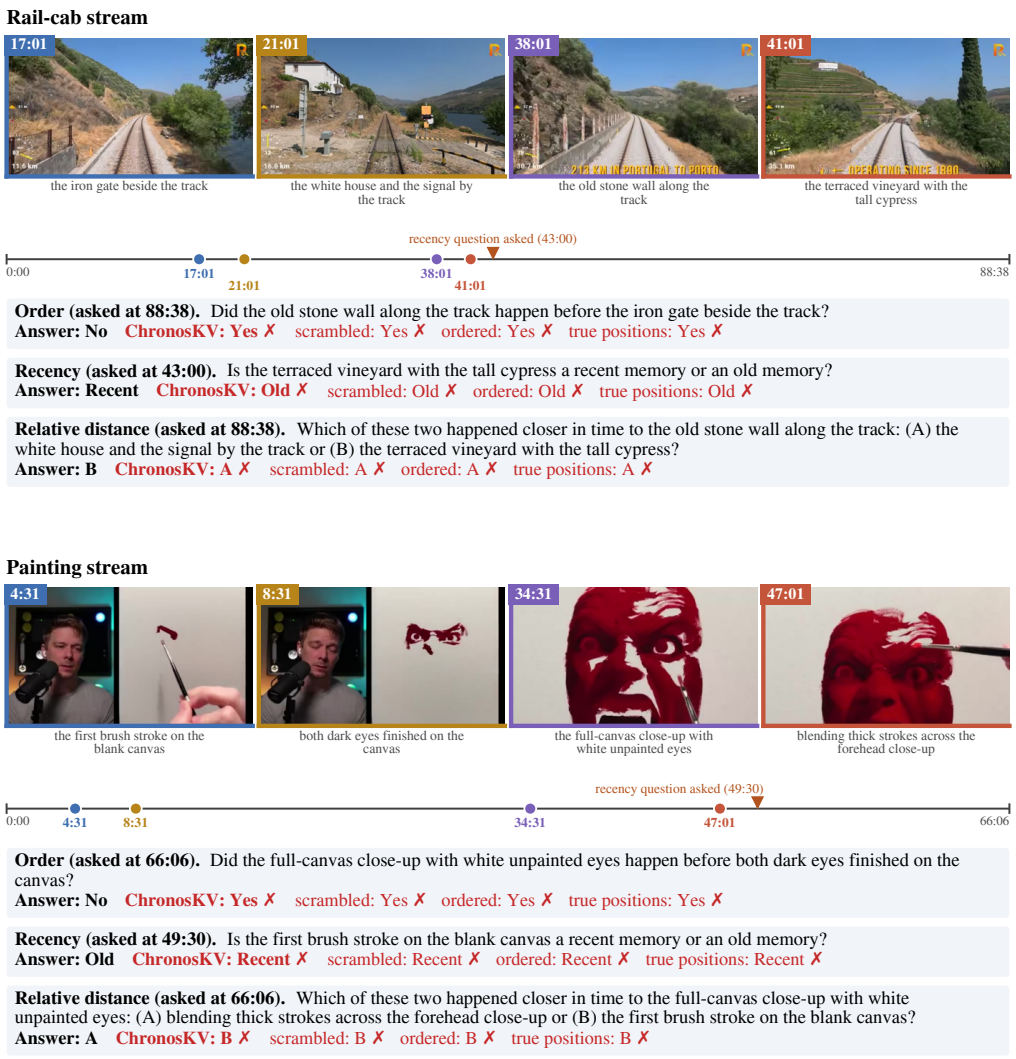


Figure 6: **Representative Interval-QA failure cases shared across all positional layouts** (LLaVA-OneVision), illustrating instances where visually indistinct target events prevent successful grounding. Formatted identically to Figure 5.

D Reproducibility Details

D.1 Implementation Details

We employ two zero-shot 7B vision-language backbones: LLaVA-OneVision-7B (Li et al., 2025a) and LongVA-7B (Zhang et al., 2025a). Frame and query feature cards are extracted via SigLIP-SO400M (Zhai et al., 2023) at a resolution of 384 pixels. During inference, the top-k scoring frames (k=192 for online streaming benchmarks; k=251 for offline long-video suites) are retrieved and projected into the dual-clock layout according to Eq. 6. Experiments are conducted on NVIDIA A100 80 GB GPUs; latency and memory profiling are executed sequentially on a single GPU.

D.2 HYPER-PARAMETER CONFIGURATION

Table 10 specifies the band separation thresholds ($\Lambda_{\text{lo}}/\Lambda_{\text{hi}}$) and frame budgets ($k$) utilized across all benchmarks. The temporal track threshold T_α is set to 28k tokens for LLaVA-OneVision (aligning with its 32k trained Qwen2 context limit) and 224k tokens for LongVA. All ingestion parameters mirror baseline configurations ($G{=}50$ tokens per frame, a 14-token sink prefix, a $4{,}800$-token sliding window, and 4-bit KV cache quantization).

Table 10: Band threshold configurations and retrieved frame budgets per dataset and backbone architecture.

Backbone	Dataset	$\Lambda_{\text{lo}}/\Lambda_{\text{hi}}$ (tokens)	k
LLaVA-OV-7B	RVS-Ego	1.8M / 2.3M	192
LLaVA-OV-7B	RVS-Movie	8k / 32k	192
LLaVA-OV-7B	Video-MME (all splits)	1.8M / 2.3M	251
LLaVA-OV-7B	MLVU	1.8M / 2.3M	251
LLaVA-OV-7B	Interval-QA	8k / 32k	oracle
LongVA-7B	RVS-Ego	950M / 1.5B	192
LongVA-7B	RVS-Movie	6k / 24k	192
LongVA-7B	Video-MME, MLVU	56k / 224k	251
LongVA-7B	Interval-QA	100k / 200k	oracle

E STREAMING EVALUATION SCORING PROTOCOLS

Open-ended streaming benchmarks (RVS-Ego, RVS-Movie) are evaluated by comparing free-form generations against ground-truth references using an automated LLM judge. To ensure strict methodological parity, we adopt the official ReKV evaluation protocol: `gpt-3.5-turbo-0125` at temperature 0 guided by the verbatim Video-ChatGPT prompt. This prompt mandates a binary correctness verdict alongside a 5-point integer score. Exact accuracy reflects the proportion of affirmative verdicts. All reported runs use uniform stream ingestion (0.5 fps), bare question phrasing without formatting hints, a maximum generation length of 256 tokens, and identical evaluation inputs ($1{,}465$ and $1{,}905$ queries for RVS-Ego and RVS-Movie, respectively). Because published literature often employs disparate judge versions, we re-evaluate all competing baselines under this unified protocol. The system prompt is detailed below:

Judge prompt for online streaming benchmarks

System prompt:
You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:
—
##INSTRUCTIONS:
- Focus on the meaningful match between the predicted answer and the correct answer.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.

User prompt:
Please evaluate the following video-based question-answer pair:
Question: {question}
Correct Answer: {reference}
Predicted Answer: {prediction}
Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {`'pred': 'yes', 'score': 4.8`}.

E.1 ANALYSIS OF JUDGE SCORING DYNAMICS

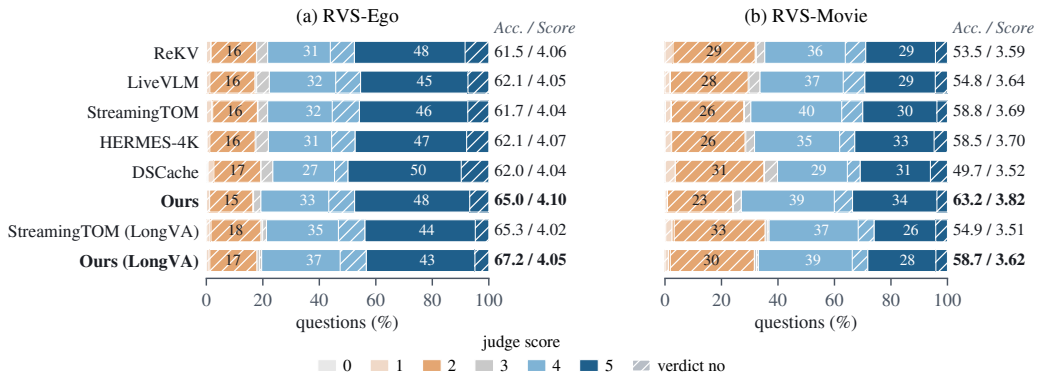


Figure 7: **Distribution of LLM judge scores across online streaming benchmarks**, paired alongside aggregate accuracy and mean score metrics for each evaluated configuration. Hatched segments are the answers whose verdict is *no*.

Figure 7 shows the judge's scores for every streaming system of Table 2 on the two RVS sets, with hatching on the answers it judged wrong. The score follows the verdict: an answer judged correct almost always receives a 4 or a 5, and one judged wrong receives a 3 or below more often than not. The hatched ends of the 4 and 5 bars are wrong answers that still received a high score; they take a similar share for every system, ours included. Mean score and accuracy therefore move together across systems.

E.2 Evaluation Repeatability

To quantify evaluation variance, LLaVA-OneVision baselines and ChronosKV configurations were independently re-judged under identical settings (Table 11). Across evaluation passes, verdict consistency reaches 97–98%, with absolute accuracy fluctuations remaining under 0.6 points, verifying that performance disparities significantly exceed evaluation noise.

Table 11: Re-judging consistency across two independent execution passes on LLaVA-OneVision streaming benchmarks, reporting accuracy (%) and verdict agreement share (%).

	RVS-Ego			RVS-Movie		
	Pass 1	Pass 2	Agreement	Pass 1	Pass 2	Agreement
StreamingTOM	61.7	61.9	98.4	58.8	59.4	96.9
ChronosKV	65.0	65.1	98.4	63.2	62.9	98.1
Performance Margin	+3.3	+3.1	–	+4.4	+3.5	–

F Limitations

ChronosKV gives a frozen backbone back its sense of elapsed time, but only at the resolution its rotary positions can carry, and that resolution is coarse. What Interval-QA verifies is order, recent against old, and nearer against farther; it never asks the model how many minutes passed, and we make no claim that it could say. The slow rotary bands measure time in tokens, so what a phase means in seconds depends on the ingest frame rate and on the tokens kept per frame, and a backbone that was never trained to name a duration has had no reason to calibrate them. We also tried stating the elapsed time in the prompt as text instead: on Interval-QA it brought no consistent gain over ordered positions and fell well short of the dual clock. We have also tested only backbones with one-dimensional RoPE; those with multi-axis rotary positions, such as Qwen2.5-VL, place the temporal axis on a different set of pairs; the dual clock can be carried over, though not without effort. Multi-axis backbones and finer temporal scales are the two directions we leave for future work.