

Think While Watching: Online Streaming Segment-Level Memory for Multi-Turn Video Reasoning in Multimodal Large Language Models

Lu Wang^{1,2}, Zhuoran Jin¹[†], Yupu Hao¹, Yubo Chen¹, Kang Liu¹,
Yulong Ao², and Jun Zhao¹[†]

¹ The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
wanglu2026@ia.ac.cn, zhuoran.jin@nlpr.ia.ac.cn, haoyupu2023@ia.ac.cn,
yubo.chen@nlpr.ia.ac.cn, kliu@nlpr.ia.ac.cn, jzhao@nlpr.ia.ac.cn

² Beijing Academy of Artificial Intelligence (BAAI), Beijing, China
aoyulong@outlook.com

Abstract. Multimodal large language models (MLLMs) have demonstrated strong performance in offline video understanding tasks, but most remain constrained to offline inference or exhibit weak online reasoning ability, rendering online multi-turn interaction over continuously arriving video streams challenging. Existing streaming approaches adopt an interleaved perception-generation paradigm, which precludes concurrent perception and generation and induces early memory decay with growing streams, degrading long-range dependency modeling. We propose **Think While Watching**, a memory-anchored streaming video reasoning framework that maintains continuous segment-level memory during multi-turn interaction. We construct a three-stage, multi-round, chain-of-thought (CoT) dataset with a stage-matched training strategy while enforcing strict causality in streaming reasoning via a segment-level streaming causal mask and streaming positional encoding. At inference, we design an efficient pipeline that overlaps watching and thinking and adaptively selects the optimal attention backend. We evaluate our method under single-round and multi-round streaming input protocols. Based on Qwen3-VL, we improve single-round accuracy by 2.6% on Streaming-Bench and 3.79% on OVO-Bench. In the multi-round protocol, we maintain performance while reducing output tokens by 56%. **Code is available at:** https://github.com/wanglu-cs/Think_While_Watching.

Keywords: Streaming Video Reasoning · Segment-level Memory · Input-Output Parallelism · Multi-turn Interaction

1 Introduction

Video understanding and reasoning are becoming central capabilities for multimodal assistants [15, 29]. While multimodal large language models (MLLMs)

[†] Corresponding authors.

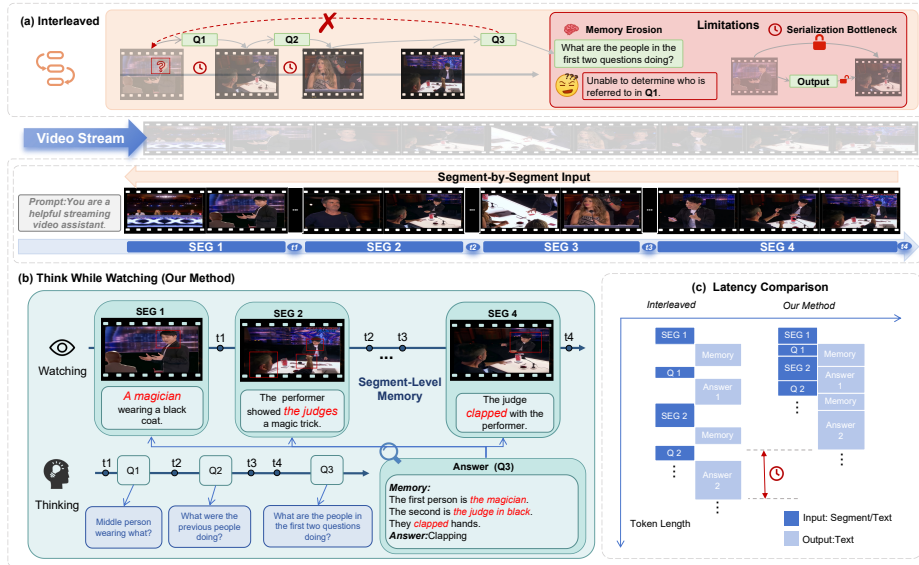


Fig. 1: Overview of Think While Watching. (a) **Interleaved baseline.** Video perception and answer generation are executed sequentially, which can cause **memory erosion** where early memory is forgotten and **serialization bottleneck** where generation stalls further input ingestion. (b) **Think While Watching (ours).** The video frames are processed in segments (SEG 1 to SEG 4) to build a continuous **segment-level memory**. During streaming, questions are answered online by retrieving implicitly relevant memories while continuing to watch. (c) **Latency comparison.** A schematic timeline showing that interleaved processing accumulates queueing delay, while our decoupled design parallelizes segment processing and answering to reduce latency.

have achieved strong performance in offline video benchmarks where the full video is available before inference in a single-turn setting [8, 9, 37, 54], many high impact scenarios are inherently **streaming**: live broadcasting [2, 51], monitoring, robotic assistants [23, 44], and other streaming scenarios. In these settings, users may ask questions at any time, and the assistant must answer in real-time while staying faithful to the visual evidence observed so far, especially under **multi-turn** interaction, where later questions often depend on earlier memories.

A widely adopted approach for streaming MLLMs is to interleave perception and generation [1, 2, 26]. Although this reduces delay compared to the offline approaches, it remains fundamentally serialized: text decoding blocks further video ingestion. This interleaved pattern leads to two phenomena. First, **Memory Erosion**: Multi-turn subsequent questions frequently refer back to earlier questions or earlier visual cues, but interleaving generation mode tends to erode long-range capability. The failure case in Fig. 1(a) makes this explicit: a later query about the first two questions becomes unanswerable because the model forgets who Q1 refers to. This issue is also validated by our experimental re-

sults: for the Qwen3-VL-4B Thinking model, accuracy in the online multi-round setting drops 40.39% compared to the offline setting, highlighting the severe challenge of maintaining long-term temporal consistency. Second, **Serialization Bottleneck**: As illustrated on the top right of Fig. 1(a), once the model starts generation, the decoder effectively locks the streaming sequence, directly harming responsiveness in dynamic streams. The root cause is that autoregressive models use unified positional encoding, so new inputs must align with generated outputs whose length is unknown, forcing ingestion to pause and causing a serialization bottleneck. The Fig. 1(c) further visualizes this effect: under interleaved processing, as the number of rounds accumulates, the input keeps piling up, leading to increasing end-to-end latency. **To mitigate Memory Erosion**, we make memory writing an explicit online behavior. For each observed segment, the model writes a memory note and appends it to a memory bank; when a question arrives, the model answers by implicitly integrating the relevant notes via the attention mechanism in Fig. 1(b). **To break the Serialization Bottleneck**, we assign independent positional encodings to decouple input and output streams at inference time, enabling input-output parallelism so the model can keep watching while thinking and thus reduce latency.

We propose **Think While Watching**, a memory-anchored streaming video reasoning framework for online multi-turn interaction. We represent a continuously arriving video as a sequence of segments and maintain a persistent **segment-level memory** throughout the dialogue. To make the framework practical, we design corresponding training and inference procedures. On the training side, we construct a three-stage, multi-round chain-of-thought (CoT) dataset with training matched to each stage, together with a **segment-level streaming causal mask** and **streaming positional encoding**, which jointly enforce strict causality throughout streaming reasoning. On the inference side, we design an efficient pipeline that overlaps watching with thinking. Our implementation is inspired by CPU process scheduling [35] in operating systems: we organize inference as a **multi-stage pipeline** on the Fig. 1(c) and decouple continuous visual ingestion from text decoding via a dual KV cache [25, 38, 39], enabling parallelism between perception and generation and mitigating serialization.

We evaluate Think While Watching under two streaming input protocols: **single-round**, where the input contains many arriving segments but the assistant answers one question, and **multi-round**, where the input contains many arriving segments and the assistant answers multiple questions over time. In the Qwen3-VL family, we improve single-round accuracy by 2.6% on StreamingBench and 3.79% on OVO-Bench. In the multi-round protocol, we preserve performance while reducing output tokens by 56%.

Our contributions are as follows:

- We propose **Think While Watching**, which maintains **segment-level memory** as a persistent state and answers each query by implicitly retrieving and integrating relevant memories, improving multi-turn consistency and enabling decoupled perception and generation; we further provide a practical **training and inference pipeline** with **three-stage training**, **stream-**

ing segment masking, and **streaming positional encoding** for causal segment-level modeling and a dual KV cache at inference time to support parallelism between perception and generation.

- We construct a three-stage, stage-aligned streaming CoT dataset with multi-round dialogues to support the proposed training strategy.
- On Qwen3-VL, we improve single-round accuracy by 2.6% on StreamingBench and 3.79% on OVO-Bench, while in multi-round streaming, we reduce output tokens by 56% without accuracy drop.

2 Related Work

2.1 Offline Video Understanding

Offline video MLLMs are improved by structured perception and cognition pipelines and temporal reasoning designs [8, 10, 14, 16, 37, 41, 54], and by reinforcement learning for complex temporal reasoning [5, 9]. Most of these methods assume the full video is available before answering, leaving causal online multi-turn interaction less explored.

2.2 Online Streaming Video Understanding

Benchmarks. StreamingBench evaluates the gap between offline models and streaming video understanding [24], while OVO-Bench emphasizes real-world online video understanding [31]. Recent work further studies streaming along with active perception and multi-turn interaction [12, 21, 22, 47, 50, 56, 59].

Interleaved perception and generation. Many streaming systems alternate visual ingestion and text decoding, as in VideoLLM-online [1] and StreamChat [26], and scale streaming supervision for online interaction [2, 46, 48]. This coupling limits input-output parallelism and does not fully address the need to model long-range dependencies across multiple turns.

Memory and efficient inference for long horizon streaming. For efficiency, one line reduces redundant visual tokens in streaming videos [3, 40, 45, 52, 57, 58]. Another line reuses historical context via KV cache retrieval and compression [4, 7, 20, 28, 30, 55]. Persistent memory and long-term multimodal agent memory further support evidence reuse across long streams [27, 49, 53]. Our work emphasizes stable segment-level memory for multi-turn streaming and an inference design that keeps perception and generation decoupled.

3 Preliminary

This section introduces the online multi-turn streaming video question answering setting studied in this work. A video is observed sequentially as a stream of segments, while a user may ask questions at arbitrary segment boundaries. The central requirement is strict streaming causality: at each turn, the system must produce its response using only the video content observed so far and the dialogue history, without accessing any future segments.

3.1 Streams and Turns

Segmented stream. We represent a video stream as an ordered sequence of segments

$$\mathbf{S}_{1:T} \triangleq \langle \mathbf{S}_1, \dots, \mathbf{S}_T \rangle, \quad (1)$$

where each $\mathbf{S}_t$ denotes a contiguous chunk of frames. Segments arrive in temporal order, and the system processes them online.

Multi-turn questioning. We consider an interaction with R question and answer turns. At turn $r \in \{1, \dots, R\}$, the user asks a question q_r after the system has observed a prefix of the stream. Let $\tau_r \in \{1, \dots, T\}$ denote the index of the latest observed segment when q_r is issued. Equivalently, q_r is asked after ingesting the segment prefix.

$$\mathbf{S}_{1:\tau_r} \triangleq \langle \mathbf{S}_1, \dots, \mathbf{S}_{\tau_r} \rangle. \quad (2)$$

Since questions arrive over time, the indices are nondecreasing:

$$1 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_R \leq T. \quad (3)$$

The dialogue history before turn r is

$$\mathcal{H}_{r-1} \triangleq \langle \langle q_1, a_1 \rangle, \dots, \langle q_{r-1}, a_{r-1} \rangle \rangle. \quad (4)$$

Under strict causality, the answer a_r at turn r is conditioned only on the observed video prefix $\mathbf{S}_{1:\tau_r}$, the question q_r , and the dialogue history $\mathcal{H}_{r-1}$.

3.2 Streaming Protocols

We consider two online evaluation protocols that share the same segmented stream $\mathbf{S}_{1:T}$ but differ in the number of question turns.

Single-round streaming. Only one question is asked, so $R = 1$. The system processes segments online and produces a single output for the question asked at τ_1 . We denote the model output as $\langle \pi_1, a_1 \rangle$, where π_1 is an optional intermediate rationale such as chain of thought and a_1 is the final answer.

Multi-round streaming. Multiple questions are asked, so $R > 1$, at different times with nondecreasing τ_r . At each turn r , the system must answer online using only the stream prefix $\mathbf{S}_{1:\tau_r}$ and the dialogue history $\mathcal{H}_{r-1}$, producing an output pair $\langle \pi_r, a_r \rangle$.

3.3 Streaming Unit Notation

To describe training and inference in a single causal formulation, we serialize a streaming interaction as an interleaved sequence of received units and a one to one aligned sequence of generated units.

Received units. Let the received unit sequence be

$$\mathbf{R}_{1:U} \triangleq \langle R_1, \dots, R_U \rangle, \quad (5)$$

where each R_u is either a visual segment unit S_t that contains the content of $\mathbf{S}_t$, or a question unit Q_r that contains the text q_r . We write $R_u \in \{S, Q\}$ to indicate the unit type. Let $\text{idx}[\cdot]$ return the arrival index in $\mathbf{R}_{1:U}$, so $\text{idx}[S_t]$ is the index u where segment t appears, and $\text{idx}[Q_r]$ is the index u where question r appears.

Generated units. For each received unit R_u , the model generates exactly one output unit C_u in the same order, forming

$$\mathbf{C}_{1:U} \triangleq \langle C_1, \dots, C_U \rangle. \quad (6)$$

If $R_u = S_t$, then C_u is a memory note denoted m_t . If $R_u = Q_r$, then C_u is the question answering output that contains the rationale π_r and answer a_r .

Token lengths and visual grids. For any text unit Y in $\{Q_1, \dots, Q_R, C_1, \dots, C_U\}$, let $L[Y]$ denote its text token length. For any segment unit S_t , let its visual token grid sizes be $\langle T_t, H_t, W_t \rangle$. Here T_t is the number of visual tokens along the temporal axis, H_t is the height axis, and W_t is the width axis, defined by the vision encoder token grid for this segment. We will also use a unit span function $\Delta[R_u]$ that assigns each received unit a nonoverlapping input position span:

$$\Delta[R_u] = \begin{cases} \max\{T_u, H_u, W_u\}, & R_u \in \{S\}, \\ L[R_u], & R_u \in \{Q\}. \end{cases} \quad (7)$$

4 Method

A simple streaming design interleaves perception and generation but is inherently serial: autoregressive decoding halts further input ingestion, and the interleaving pattern mismatches LLM pretraining format. Think While Watching generates segment-level memory notes online and decouples perception and generation, enabling interaction in real time across multiple turns. Details are shown in Fig. 2.

4.1 Segment-level Memory Notes

To mitigate Memory Erosion, our method maintains a persistent segment-level memory bank as the online state for multi-turn streaming. For each arriving segment $\mathbf{S}_t$, the model writes exactly one memory note and appends it to the bank. The memory bank after observing the segment prefix $\mathbf{S}_{1:t}$ is defined as

$$\mathcal{M}_t \triangleq \{\langle i, m_i \rangle\}_{i=1}^t. \quad (8)$$

Each note m_t is a compact text unit grounded in $\mathbf{S}_t$ that records reusable evidence, including key entities and attributes, salient actions and interactions, scene changes, and short-range temporal relations. We denote the memory-writing function implemented by the MLLM backbone with parameters θ as $\text{Mem}_\theta[\cdot]$, and write

$$m_t = \text{Mem}_\theta[\mathbf{S}_t], \quad C_{\text{idx}[S_t]} = m_t. \quad (9)$$

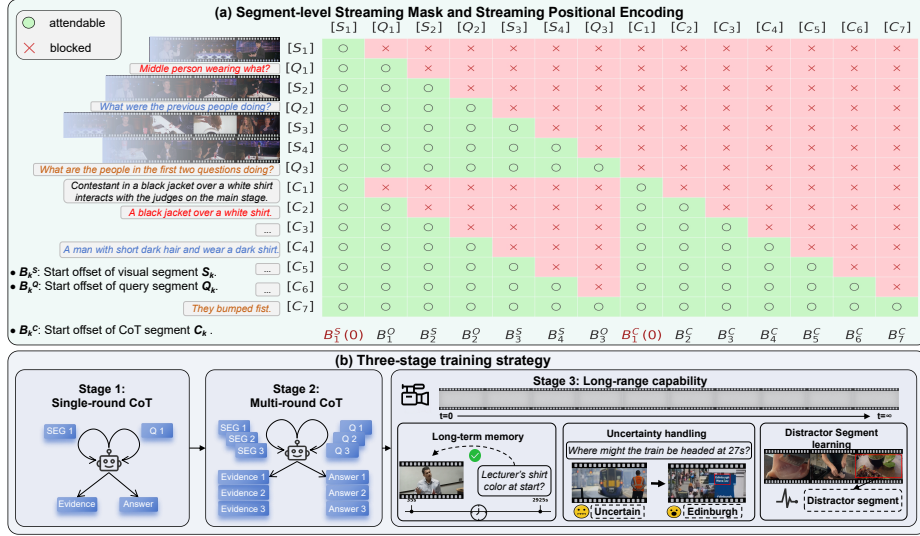


Fig. 2: Training components of Think While Watching. (a) segment-level streaming attention mask and streaming positional encoding: example input stream $\mathbf{R} = \langle S_1, Q_1, S_2, Q_2, S_3, S_4, Q_3 \rangle$ with generated outputs $\mathbf{C} = \langle C_1, \dots, C_7 \rangle$. Green indicates the source prefix available at time step u , which C_u is allowed to attend to. Red masks all future segments to prevent information leakage. For positional encoding, we use separate position indices for the input and output streams. (b) Three-stage training strategy: single-round CoT for streaming input adaptation, multi-round CoT for multi-turn interaction, and long-range capability training for long-term memory, uncertainty handling, and distractor learning.

When a question q_r is issued after observing segment τ_r , the model answers by conditioning on the current question, the dialogue history, and the available memory prefix, while letting attention implicitly select relevant notes:

$$\langle \pi_r, a_r \rangle \sim p_\theta[\pi_r, a_r \mid q_r, \mathcal{H}_{r-1}, \mathcal{M}_{\tau_r}]. \quad (10)$$

4.2 Streaming Architecture

Prefix and suffix formatting. To match the standard SFT format of LLMs, we serialize each training example as a source prefix followed by a target suffix. The source prefix contains the entire received unit stream $\mathbf{R}_{1:U}$, while the target suffix contains the aligned generated stream $\mathbf{C}_{1:U}$. Without additional constraints, this serialization would leak future received units to earlier generated units.

Segment-level streaming attention mask. We feed the concatenated sequence $\langle R_1, \dots, R_U, C_1, \dots, C_U \rangle$ with a segment-level mask M^{seg} to enforce streaming causality. Let A denote the segment that contributes query tokens and B denote the segment that contributes key and value tokens, with $A, B \in$

$\{R_1, \dots, R_U, C_1, \dots, C_U\}$. The mask is defined as:

$$M^{\text{seg}}[A, B] = \begin{cases} \mathbb{I}[v \leq u], & A = R_u, B = R_v, \\ \mathbb{I}[v \leq u], & A = C_u, B = R_v, \\ \mathbb{I}[k \leq u], & A = C_u, B = C_k, \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Here u is the arrival index of the querying unit A . For the attended unit B , we use v if B is a received unit R_v and k if B is a generated unit C_k . The first three cases in Eq. (11) enforce streaming causality: the received stream is causal in arrival order, each generated unit C_u can attend to the received prefix up to step u , and generated units are causal with access only to $C_{1:u}$. All remaining connections are masked, including $R_u \rightarrow C_k$ and $C_u \rightarrow R_v$ for $v > u$. We obtain token-level masks by expanding M^{seg} and applying standard causal masking within each C_u . As shown in Fig. 2(a), C_1 attends only to S_1 , C_2 attends to $\langle S_1, Q_1 \rangle$, and C_3 attends to $\langle S_1, Q_1, S_2 \rangle$.

Streaming positional encoding with MRoPE. We build on Multimodal Rotary Positional Embeddings MRoPE [42], but decouple the input and output to support parallel reasoning. Specifically, the input stream follows the standard cumulative offset scheme, while the output stream independently starts positional encoding from 0. We use B to represent the base offset and compute the start offsets of the k -th visual segment S_k input, the k th question Q_k input, and the k th generated unit C_k output :

$$B_k = \begin{cases} B_k^S = \sum_{u < \text{id}\mathbb{X}[S_k]} \Delta[R_u], \\ B_k^Q = \sum_{u < \text{id}\mathbb{X}[Q_k]} \Delta[R_u], \\ B_k^C = \begin{cases} 0, & k = 1, \\ \sum_{i=1}^{k-1} L[C_i], & k \geq 2. \end{cases} \end{cases} \quad (12)$$

In this design, B_k^S and B_k^Q are computed only from the received input prefix, while B_k^C is computed only from previously generated tokens. Therefore, even when the output length is still unknown during decoding, newly arriving input segments can always be assigned correct input positions.³

³ MRoPE extends RoPE to multimodal tokens by applying rotary positional encoding along modality-specific axes. For a visual segment unit S_k with token grid size $\langle T_k, H_k, W_k \rangle$, a token at local grid coordinate $\langle t, h, w \rangle$ is assigned global coordinates $\langle t + B_k^S, h + B_k^S, w + B_k^S \rangle$, where the start offset B_k^S is given in Eq. (12). For a text unit $Y \in \{Q_k, C_k\}$ with length $L[Y]$, a token at local position n uses $n + B_k^Q$ for question inputs and $n + B_k^C$ for generated outputs. The contribution of each received unit to the input position budget is determined by the unit span $\Delta[R_u]$ in Eq. (7).

Table 1: Dataset statistics across three training stages. Stages 1&2 are built from VideoChatOnline-IT short videos, while Stage 3 contains long-range multi-round dialogues from YouTube. Video duration is reported in seconds (min/avg/max).

Stage	Source	Instances	Rounds	Avg. rounds	Frames	Min (s)	Avg (s)	Max (s)
1	VideoChatOnline-IT	5,160	5,160	1.00	≤ 64	8.18	79.40	3550.10
2	VideoChatOnline-IT	2,752	8,513	3.09	≤ 64	2.00	400.92	3443.97
3	YouTube	1,500	6,000	4.00	100-300+	600.12	1697.30	3595.03

4.3 Streaming Training

Three-stage training. We fine-tune the MLLM in three stages: Stage 1 learns to write segment-level memory notes and answer single-round queries. Stage 2 scales to multi-round dialogues. Stage 3 targets long-range behaviors on long videos, including long-term evidence recall, uncertainty handling, and distractor segment learning where we insert irrelevant frames as distractors. In Fig. 2(b), Stage 3 covers three long-horizon behaviors: **long-term memory** for recalling early evidence in late queries, **uncertainty handling** for deferring commitment when evidence is not yet observable, and **distractor robustness** for ignoring irrelevant segments during streaming.

Three-stage Streaming CoT Dataset Generation. Streaming CoT [17, 19] datasets for MLLMs are extremely scarce. Multi-round streaming CoT with memory notes is largely absent. We therefore synthesize a three-stage dataset that matches our three-stage training. Table 1 summarizes the dataset statistics across the three training stages.

Stage 1 and Stage 2 short video streaming CoT. We use VideoChatOnline-IT [18] as the source pool and sample up to 64 frames per instance. Stage 1 constructs 5,160 single-round instances from temporal perception subsets. Stage 2 converts 8,513 short video QA instances into 2,752 multi-round dialogues by grouping questions over the same video prefix. For both stages, we use GPT-5.2 to generate memory-anchored CoT based on the original dataset QAs.

Stage 3 long-range streaming CoT. We collect long videos from YouTube using 500 keywords spanning three categories: tutorial for procedural content, lecture for explanatory content, and longform for continuous recordings. We then generate 1,500 long video instances with balanced input lengths of 100 to 200 frames, 200 to 300 frames, and 300 or more frames, and each instance contains 3 to 5 rounds. QA and CoT generation follow the same procedure as in Stage 1 and Stage 2. Details of the dataset and the prompt can be found in Appendix E.

Quality inspection. We enforce the core constraints in Table A5 during synthesis, and additionally verify that each example contains exactly $S + Q$ output items.

4.4 Streaming Inference

Parallel reasoning. For real-time deployment, we adopt a dual KV cache implementation that decouples continuous source ingestion from autoregressive decoding. This engineering pattern is common in recent streaming systems [13, 38]. Meanwhile, we keep the same segment-level streaming mask and streaming MRoPE at inference time to ensure consistency with training.

Adaptive attention backend. During decoding, our streaming mask is not always a standard causal mask: queries must attend to an allowed source prefix while remaining causal over the generated suffix, so the query and key lengths can differ ($q_{\text{len}} \neq k_{\text{len}}$). We therefore choose the attention backend adaptively—using Flash Attention [6] when the pattern is standard causal, and otherwise applying an explicit streaming mask with memory-efficient attention [34]. Specifically, we use Flash Attention for source prefilling ($q_{\text{len}} = k_{\text{len}}$) and for autoregressive steps ($q_{\text{len}} = 1$), and switch to memory-efficient attention when $1 < q_{\text{len}} < k_{\text{len}}$ to enforce the custom streaming mask. This preserves segment-level causality while keeping inference fast.

Memory-bounded deployment. To verify the efficiency of our method on long videos under limited memory, we further evaluate a training-free visual-token eviction variant in Appendix H, where old visual tokens are evicted while instruction, question, and memory-note tokens are retained.

5 Experiments

5.1 Datasets and Setup.

We evaluate online streaming performance on StreamingBench [24] and OVO-Bench [31]. StreamingBench focuses on streaming video understanding and includes four subsets. OVO-Bench emphasizes real-world video understanding under three subsets. More details of datasets are in Appendix E.

Evaluation protocols. We evaluate models under both offline and online protocols. In the offline Batch setting, all sampled frames from the full video are given as a single input, and the model answers after observing the complete video. In the online protocol, we consider single-turn and multi-turn interaction. For single-turn evaluation, each video is segmented by the benchmark-provided question timestamps into consecutive intervals $[0, t_1]$, $[t_1, t_2]$, $\dots$; any segment longer than 60s is further split into 30s chunks. For multi-turn evaluation, we use the same segmentation strategy, while requiring the model to answer multiple questions online as segments arrive continuously.

Backbones and checkpoints. We evaluate our method with Qwen3-VL backbones at 2B, 4B, and 8B scales. We use the Instruct model for training and compare its performance with the Thinking model. TWW is used to denote our method in the following. Stage 2 and Stage 3 refer to the checkpoints obtained after training up to the second and third stages.

5.2 Baselines.

Offline evaluation. We evaluate Gemini 1.5 Pro [36] and GPT-4o [32] as representative closed-source MLLMs. Qwen3-VL-Instruct and Qwen3-VL-Thinking are used as open-source baselines. We also report results for our $TWW_{\text{Batch},S2}$ and $TWW_{\text{Batch},S3}$ checkpoints, corresponding to Stage 2 and Stage 3, evaluated under the offline batch protocol.

Online evaluation. For online evaluation, $\text{Instruct}_{\text{online}}$ and $\text{Thinking}_{\text{online}}$ run Qwen3-VL-Instruct and Qwen3-VL-Thinking, respectively, under the multi-turn protocol in Sec. 5.1. We further evaluate our checkpoints under streaming settings: $TWW_{\text{single-turn},S2}$ and $TWW_{\text{single-turn},S3}$ follow the single-round streaming protocol, while $TWW_{\text{multi-turn},S2}$ and $TWW_{\text{multi-turn},S3}$ follow the multi-round streaming protocol. Finally, Interleaved alternates between ingesting one segment and decoding text, coupling perception and generation as a naive streaming baseline. More online baselines and details are in Appendix D.

5.3 Metrics.

We use accuracy to evaluate performance on each benchmark and each evaluation regime. We also report Avg Tokens, the average number of generated output tokens per query. Token Reduce, denoted by $\Delta\%$, the percentage reduction of Avg Tokens compared with the Thinking baseline of the same backbone size and Avg Frames, the average number of processed frames per query. For latency, we report TTFT, time to first token, measured as the number of tokens processed before the first answer token is produced.

5.4 Main Results.

Tables 2 and 3 report results on two streaming benchmarks, StreamingBench and OVO-Bench. We summarize the key findings below. We further analyzed the types of errors in Appendix G.

- **Naive streaming inference collapses without streaming aligned training, highlighting the difficulty of multi-turn streaming.** On StreamingBench, directly running $\text{Instruct}_{\text{online}}$ and $\text{Thinking}_{\text{online}}$ only 21.47% and 18.13% overall compared with 56.67% and 58.52% with the Qwen3-VL-4B in the offline batch setting. A similar drop is observed on OVO-Bench: 21.45% and 16.21% versus 50.32% and 50.70%, showing that multi-turn streaming is nontrivial and requires streaming aligned supervision.
- **Streaming aligned supervision improves accuracy.** With the Qwen3-VL-4B, $TWW_{\text{single-turn},S3}$ improves overall accuracy from 58.52% to 60.04% on StreamingBench and from 50.70% to 55.02% on OVO-Bench compared with the Thinking baseline.
- **Long video training strengthens streaming behavior.** Stage 3 generally improves upon Stage 2. For example, on OVO-Bench with the 4B backbone, $TWW_{\text{single-turn},S3}$ improves from 54.51% to 55.02%.

Table 2: StreamingBench results with accuracy. Columns left of the double bar report **performance**, higher is better, while columns on the right report **efficiency** for open-source models only. Δ is computed against the same backbone Thinking baseline. Avg Frames is 148.35 for the single-turn protocol and 62.58 for the multi-turn protocol.

		Performance						Efficiency		
Regime	Method	SQA↑	OmniSource↑	Realtime↑	Proactive↑	Overall↑	Δ↑	Avg Tokens↓	Token Reduce↑	Δ%
Closed source models										
Offline	Gemini 1.5 Pro [36]	54.80	67.80	77.39	45.10	70.26	-	-	-	-
	GPT-4o [32]	32.80	50.95	74.54	56.86	64.10	-	-	-	-
Open source models										
Online	Flash-VStream-7B [55]	26.80	26.00	23.23	1.96	24.04	-	-	-	-
	VideoLLM-online-8B [1]	30.80	28.45	35.99	3.92	32.48	-	-	-	-
	Dispider-7B [33]	34.80	35.66	67.63	25.34	53.12	-	-	-	-
	StreamAgent-7B [49]	39.60	36.26	74.28	28.90	57.02	-	-	-	-
Qwen3-VL-2B										
Offline	Instruct	37.60	31.60	68.36	29.60	52.24	-1.17	-	-	-
	Thinking	34.00	31.73	70.02	36.80	53.41	+0.00	1232.91	0.00	-
	TWW _{Batch,S2}	44.00	33.13	69.16	33.20	53.75	+0.34	1012.35	17.89	-
	TWW _{Batch,S3}	44.80	33.20	69.20	36.00	54.00	+0.59	1102.34	10.59	-
Online	Instruct _{online}	9.20	28.80	21.28	13.20	22.67	-30.74	-	-	-
	Thinking _{online}	8.40	11.60	12.92	19.20	12.58	-40.83	832.23	32.50	-
	TWW _{single-turn,S2}	47.20	34.20	71.84	32.80	55.76	+2.35	923.58	25.09	-
	TWW _{single-turn,S3}	48.00	34.27	72.00	34.40	56.00	+2.59	930.23	24.55	-
	TWW _{multi-turn,S2}	42.40	31.33	69.20	33.60	53.11	-0.30	285.42	76.85	-
	TWW _{multi-turn,S3}	45.20	31.47	69.24	34.80	53.40	-0.01	300.20	75.65	-
Qwen3-VL-4B										
Offline	Instruct	37.20	38.47	71.36	38.40	56.67	-1.85	-	-	-
	Thinking	46.40	36.53	74.50	42.80	58.52	+0.00	689.22	0.00	-
	TWW _{Batch,S2}	42.40	39.60	71.84	39.60	57.67	-0.85	594.28	13.77	-
	TWW _{Batch,S3}	44.00	39.67	71.88	40.80	57.87	-0.65	620.35	9.99	-
Online	Instruct _{online}	12.80	21.53	22.88	15.60	21.47	-37.05	-	-	-
	Thinking _{online}	21.20	21.47	16.92	7.20	18.13	-40.39	482.24	30.03	-
	TWW _{single-turn,S2}	46.00	40.67	74.36	41.20	59.71	+1.19	558.12	19.02	-
	TWW _{single-turn,S3}	46.80	41.00	74.48	43.20	60.04	+1.52	570.68	17.20	-
Online	TWW _{multi-turn,S2}	40.80	39.20	71.20	40.00	57.11	-1.41	291.86	57.65	-
	TWW _{multi-turn,S3}	43.60	39.33	71.28	40.80	57.40	-1.12	302.56	56.10	-
Qwen3-VL-8B										
Offline	Instruct	44.40	37.53	73.60	36.00	57.87	-0.35	-	-	-
	Thinking	45.60	35.47	74.46	44.80	58.21	+0.00	759.30	0.00	-
	TWW _{Batch,S2}	52.80	38.93	74.56	38.00	59.44	+1.23	708.16	6.70	-
	TWW _{Batch,S3}	53.20	39.07	74.64	40.00	59.67	+1.45	720.30	5.10	-
Online	Instruct _{online}	17.20	21.47	25.12	17.60	23.04	-35.17	-	-	-
	Thinking _{online}	14.80	13.00	17.56	16.40	15.82	-42.39	573.75	24.41	-
	TWW _{single-turn,S2}	54.00	40.07	77.60	39.60	61.67	+3.46	651.74	14.13	-
	TWW _{single-turn,S3}	54.40	40.67	77.68	41.60	62.04	+3.83	660.92	12.92	-
	TWW _{multi-turn,S2}	48.80	37.40	74.32	38.40	58.60	+0.39	288.64	61.97	-
	TWW _{multi-turn,S3}	50.00	37.47	74.40	40.00	58.82	+0.61	290.82	61.68	-

- **Multi-turn segment-level memory yields a strong accuracy&efficiency tradeoff.** Under the multi-turn protocol, TWW_{multi-turn,S3} maintains competitive accuracy while substantially reducing decoding tokens. With the 4B backbone, it achieves 57.40% on StreamingBench with an average of 302.56 tokens, reducing token usage by 56.10%. On OVO-Bench, it obtains 51.80% with 255.91 tokens on average, reducing token usage by 45.80%.

5.5 Analysis.

Generalization to Offline Video Understanding. Although our training is designed for streaming scenarios, we also evaluate whether the learned be-

Table 3: OVO-Bench results. Avg Frames is 63.23 for the single-turn protocol and 25.47 for the multi-turn protocol.

		Performance					Efficiency			
Regime	Method	Backward↑	Realtime↑	Forward↑	Overall↑	Δ↑	Avg Tokens↓	Token Reduce↑	Δ%	
Closed source models										
Offline	Gemini 1.5 Pro [36]	69.32	62.54	57.15	63.00	-	-	-		
	GPT-4o [32]	64.46	60.75	53.40	59.54	-	-	-		
Open source models										
Online	Flash-VStream-7B [55]	27.38	28.37	45.09	33.61	-	-	-		
	Dispider-7B [33]	36.06	54.55	34.72	41.78	-	-	-		
	StreamForest-7B [53]	52.02	61.20	53.49	55.57	-	-	-		
	StreamAgent-7B [49]	41.70	61.30	45.40	49.40	-	-	-		
Qwen3-VL-2B										
Offline	Instruct	37.29	56.65	51.51	49.97	+2.23	-	-		
	Thinking	41.78	57.27	45.04	47.74	+0.00	590.25	0.00		
	TWW _{Batch,S2}	43.25	56.52	44.72	47.67	-0.07	518.64	12.13		
	TWW _{Batch,S3}	44.37	56.87	45.31	48.30	+0.56	530.32	10.15		
Online	Instruct _{online}	11.89	16.01	26.87	20.76	-26.98	-	-		
	Thinking _{online}	8.56	19.47	20.29	17.63	-30.11	478.74	18.89		
	TWW _{single-turn,S2}	44.58	57.85	47.35	49.67	+1.93	456.28	22.70		
	TWW _{single-turn,S3}	45.96	58.18	47.86	50.31	+2.57	470.20	20.34		
	TWW _{multi-turn,S2}	40.85	55.42	44.34	46.67	-1.07	278.52	52.81		
	TWW _{multi-turn,S3}	42.16	55.79	44.54	47.15	-0.59	280.32	52.51		
Qwen3-VL-4B										
Offline	Instruct	44.32	62.39	46.29	50.32	-0.38	-	-		
	Thinking	50.78	61.17	45.08	50.70	+0.00	472.18	0.00		
	TWW _{Batch,S2}	52.45	62.18	47.37	52.51	+1.81	412.37	12.67		
	TWW _{Batch,S3}	53.88	62.37	47.86	53.11	+2.41	430.51	8.83		
Online	Instruct _{online}	13.63	20.79	24.95	21.45	-29.25	-	-		
	Thinking _{online}	14.42	14.10	18.06	16.21	-34.49	360.63	23.62		
	TWW _{single-turn,S2}	53.92	64.25	49.55	54.51	+3.81	358.45	24.09		
	TWW _{single-turn,S3}	55.47	64.52	49.78	55.02	+4.32	378.64	19.81		
	TWW _{multi-turn,S2}	48.65	60.85	47.67	51.51	+0.81	251.36	46.77		
	TWW _{multi-turn,S3}	49.13	61.17	47.86	51.80	+1.10	255.91	45.80		
Qwen3-VL-8B										
Offline	Instruct	42.63	62.86	51.02	52.54	-1.28	-	-		
	Thinking	54.24	61.34	49.63	53.82	+0.00	390.42	0.00		
	TWW _{Batch,S2}	55.82	63.15	51.83	55.78	+1.96	325.82	16.55		
	TWW _{Batch,S3}	57.37	63.68	52.01	56.34	+2.52	340.22	12.86		
Online	Instruct _{online}	15.69	25.57	21.12	21.22	-32.60	-	-		
	Thinking _{online}	11.25	19.59	18.51	17.30	-36.52	330.67	15.30		
	TWW _{single-turn,S2}	56.45	64.52	52.78	56.78	+2.96	276.19	29.26		
	TWW _{single-turn,S3}	57.05	64.76	52.97	57.07	+3.25	290.07	25.70		
	TWW _{multi-turn,S2}	52.05	63.25	52.85	55.55	+1.73	224.68	42.45		
	TWW _{multi-turn,S3}	52.77	63.68	53.29	56.05	+2.23	227.33	41.77		

haviors can transfer to offline video understanding tasks. We evaluate on two offline benchmarks: **Video-MME** [11] and **LV-Bench** [43], following their official evaluation settings. Table 4 shows that streaming training also benefits offline evaluation. In particular, TWW_{single-turn,S3} improves Video-MME from 68.89% to **73.41%** and LV-Bench from 53.47% to **57.68%**, showing that long-range streaming supervision transfers effectively to offline long video reasoning. Unless otherwise stated, the following analyses use Qwen3-VL-4B on StreamingBench.

Long Range Attention Analysis. We analyze how far back the model consults history when generating answers by aggregating the last layer attention from answer tokens to historical tokens and grouping the attended history by

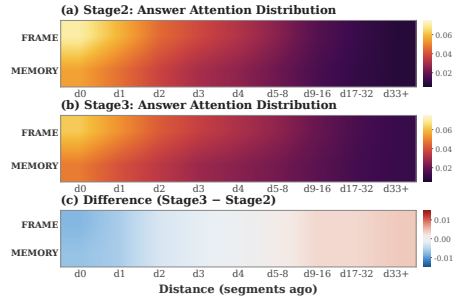


Fig. 3: Answer attention vs segment distance on $TWW_{\text{multi-turn}}$. After Stage 3, attention mass shifts from near-history to more distant segments.

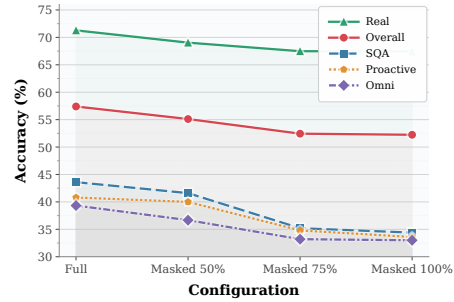


Fig. 4: Ablation under frame masking on $TWW_{\text{multi-turn},S3}$. Overall represents the accuracy rate. The remaining curves represent the results of the subsets.

Table 4: Offline video understanding results on Video-MME and LV-Bench.

Method	Video-MME				LV-Bench						
	Short	Medium	Long	Overall	ER	EU	KIR	TG	Rea	Sum	Overall
Thinking	77.78	66.67	62.22	68.89	54.22	51.62	62.07	46.82	55.50	40.35	53.47
Instruct	78.22	66.78	62.89	69.30	56.15	53.78	65.86	50.45	60.00	43.86	56.19
TWW _{Batch,S3}	78.89	67.11	64.00	70.00	58.07	54.24	66.21	51.36	62.00	47.37	57.39
TWW _{multi-turn,S3}	79.00	67.11	62.33	69.48	56.44	54.39	66.55	50.91	61.50	45.61	56.81
TWW _{single-turn,S3}	83.22	70.11	66.89	73.41	57.33	54.70	67.59	51.36	62.50	52.63	57.68

segment distance $d = \tau_r - i$, where τ_r denotes the index of the latest observed segment when answering the r th question, and $i \leq \tau_r$ denotes the index of a historical segment being attended to; $d = 0$ corresponds to the most recent segment, and larger d indicates segments further in the past. We separately measure attention to **FRAME** tokens, which are visual tokens from prior segments, and **MEMORY** tokens, which are accumulated memory note tokens written for those segments. Figure 3 shows that the Stage 2 checkpoint exhibits a strong recency bias, whereas Stage 3 reallocates attention mass from near history buckets to more distant buckets. The shift is more pronounced on **MEMORY** tokens than on visual tokens, consistent with the intended role of memory notes as a compact long-range state for multi-turn interaction.

Ablation Studies. We conduct ablations to isolate the roles of the memory bank, visual inputs, and segmentation granularity. Removing the memory bank causes a clear accuracy drop from 57.40% to 52.35% in Table 5, confirming that memory notes serve as an effective persistent state in multi-round streams. For visual ablations, Fig. 4 shows monotonic degradation as more frames are masked. Performance remains stable under moderate corruption, indicating that segment-level memory notes provide a stabilizing signal once written; under severe corruption, accuracy approaches the no-memory regime, as unreliable visual evidence impairs memory writing. For segmentation granularity, Table 5 shows a clear accuracy&efficiency tradeoff: longer segments (120s/60s) reduce

Table 5: Ablations on StreamingBench. a/b represents the maximum segment duration and the chunk duration.

Category	Setting	Acc↑	Avg Tok↓
Memory	with notes	57.40	302.56
Memory	without notes	52.35	330.73
Segment	60s/30s	57.40	302.56
Segment	120s/60s	55.33	230.46
Segment	30s/15s	57.20	380.50

Table 6: TTFT on StreamingBench. Overall accuracy and time-to-first-token of Qwen3-VL-4B for batch, interleaved streaming, and our TWW_{multi-turn,S3} pipeline.

Method	Overall Acc↑	TTFT↓
Thinking	58.52	31203.69
Interleaved	55.35	2304.28
TWW _{multi-turn,S3}	57.40	2304.28

the average decoding length from 302.56 to 230.46 tokens but lower accuracy by 2.07%, whereas shorter segments (30s/15s) preserve accuracy but increase decoding length to 380.50 tokens (+25.8%) due to more frequent memory updates. **TTFT Analysis.** Table 6 reports overall accuracy and TTFT on StreamingBench with the 4B backbone. Compared with batch Thinking, our streaming pipeline reduces TTFT by **92.6%**, from 31203.69 to 2304.28 tokens, while maintaining comparable accuracy. The interleaved baseline achieves a similar TTFT for a multi-turn but is consistently less accurate.

Theoretical Latency Analysis. Our method decouples ingestion from decoding, largely eliminating decoder-induced ingestion backlog, thereby avoiding backlog explosion as ρ (arrival rate over processing rate) approaches 1 and significantly reducing latency. The complete derivation is in Appendix C.

Supplementary Experiments. To further demonstrate the effectiveness of our method, Appendix H further examines TWW under practical and diagnostic settings, including memory-bounded inference with visual-token eviction, stage-matched versus mixed training, cross-backbone transfer, adaptive segment boundaries, and recovery from noisy early evidence.

6 Conclusion

We presented **Think While Watching**, a memory-anchored streaming video reasoning framework for multi-turn interaction over continuously arriving streams. Our approach maintains **segment-level memory notes** as a persistent state, enforces strict causality through a **segment-level streaming causal mask** and **streaming positional encoding**, and enables practical real-time deployment via a dual KV cache pipeline with adaptive attention backends. Experiments on StreamingBench and OVO-Bench validate the proposed method’s effectiveness, consistently improving online accuracy while maintaining strong efficiency.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.U24A20335), and the independent research project of the Key Laboratory of Cognition and Decision Intelligence for Complex Systems.

References

1. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J., Gao, Z., Mao, D., Shou, M.Z.: Videollm-online: Online video large language model for streaming video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 18407–18418. IEEE (2024). <https://doi.org/10.1109/CVPR52733.2024.01742>, <https://doi.org/10.1109/CVPR52733.2024.01742>
2. Chen, J., Zeng, Z., Lin, Y., Li, W., Ma, Z., Shou, M.Z.: Livecc: Learning video LLM with streaming speech transcription at scale. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 29083–29095. Computer Vision Foundation / IEEE (2025). <https://doi.org/10.1109/CVPR52734.2025.02708>, https://openaccess.thecvf.com/content/CVPR2025/html/Chen_LiveCC_Learning_Video_LLM_with_Streaming_Speech_Transcription_at_Scale_CVPR_2025_paper.html
3. Chen, X., Tao, K., Shao, K., Wang, H.: Streamingtom: Streaming token compression for efficient video understanding. CoRR **abs/2510.18269** (2025). <https://doi.org/10.48550/ARXIV.2510.18269>, <https://doi.org/10.48550/arXiv.2510.18269>
4. Chen, Y., Bai, X., Wang, Z., Bai, C., Dai, Y., Lu, M., Zhang, S.: Streamkv: Streaming video question-answering with segment-based KV cache retrieval and compression. CoRR **abs/2511.07278** (2025). <https://doi.org/10.48550/ARXIV.2511.07278>, <https://doi.org/10.48550/arXiv.2511.07278>
5. Chen, Y., Huang, W., Shi, B., Hu, Q., Ye, H., Zhu, L., Liu, Z., Molchanov, P., Kautz, J., Qi, X., Liu, S., Yin, H., Lu, Y., Han, S.: Scaling RL to long videos. CoRR **abs/2507.07966** (2025). <https://doi.org/10.48550/ARXIV.2507.07966>, <https://doi.org/10.48550/arXiv.2507.07966>
6. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness (2022), <https://arxiv.org/abs/2205.14135>
7. Di, S., Yu, Z., Zhang, G., Li, H., Zhong, T., Cheng, H., Li, B., He, W., Shu, F., Jiang, H.: Streaming video question-answering with in-context video kv-cache retrieval. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), <https://openreview.net/forum?id=8g9fs6mdEG>
8. Fei, H., Wu, S., Ji, W., Zhang, H., Zhang, M., Lee, M., Hsu, W.: Video-of-thought: Step-by-step video reasoning from perception to cognition. CoRR **abs/2501.03230** (2025). <https://doi.org/10.48550/ARXIV.2501.03230>, <https://doi.org/10.48550/arXiv.2501.03230>
9. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. CoRR **abs/2503.21776** (2025). <https://doi.org/10.48550/ARXIV.2503.21776>, <https://doi.org/10.48550/arXiv.2503.21776>
10. Feng, Y., Yan, Z., Jia, Y., Chen, E.Q., Qin, J.: Training-free dense video captioning with large-scale pretrained models. Machine Intelligence Research pp. 1–16 (2026)
11. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., Sun, X.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville,

- TN, USA, June 11-15, 2025. pp. 24108–24118. Computer Vision Foundation / IEEE (2025). <https://doi.org/10.1109/CVPR52734.2025.02245>, 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
12. Fu, S., Yang, Q., Li, Y., Peng, Y., Lin, K., Wei, X., Hu, J., Xie, X., Zheng, W.: Vispeak: Visual instruction feedback in streaming videos. CoRR **abs/2503.12769** (2025). <https://doi.org/10.48550/ARXIV.2503.12769>, <https://doi.org/10.48550/arXiv.2503.12769>
 13. Guan, Y., Yin, L., Liang, D., Ju, J., Luo, Z., Luan, J., Liu, Y., Bai, X.: Video streaming thinking: Videollms can watch and think simultaneously. CoRR **abs/2603.12262** (2026). <https://doi.org/10.48550/ARXIV.2603.12262>, <https://doi.org/10.48550/arXiv.2603.12262>
 14. Han, C., Fan, J., Wu, N., Dai, J., Bao, H., Lu, X.: Object-centric video prediction with mask-guided spatiotemporal diffusion. Machine Intelligence Research pp. 1–11 (2026)
 15. Hao, Y., Cao, P., Jin, Z., Liao, H., Chen, Y., Liu, K., Zhao, J.: Evaluating personalized tool-augmented llms from the perspectives of personalization and proactivity. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 21897–21935 (2025)
 16. Hu, G., He, B., Zhang, H.: Compositional prompting video-language models to understand procedure in instructional videos. Machine Intelligence Research **20**(2), 249–262 (2023)
 17. Hu, Y., Yang, Z., Wang, S., Qian, S., Wen, B., Yang, F., Gao, T., Xu, C.: Streamingcot: A dataset for temporal dynamics and multimodal chain-of-thought reasoning in streaming videoqa. CoRR **abs/2510.25332** (2025). <https://doi.org/10.48550/ARXIV.2510.25332>, <https://doi.org/10.48550/arXiv.2510.25332>
 18. Huang, Z., Li, X., Li, J., Wang, J., Zeng, X., Liang, C., Wu, T., Chen, X., Li, L., Wang, L.: Online video understanding: A comprehensive benchmark and memory-augmented method. CoRR **abs/2501.00584** (2025). <https://doi.org/10.48550/ARXIV.2501.00584>, <https://doi.org/10.48550/arXiv.2501.00584>
 19. Jin, Z., Zhu, K., Yuan, H., Hao, Y., Cao, P., Chen, Y., Liu, K., Zhao, J.: Look light, think heavy: What multimodal chain-of-thought reasoning can and cannot do. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8567–8608 (2026)
 20. Kim, D., Yang, S., Shin, W., Kim, J.Y.: V-rex: Real-time streaming video llm acceleration via dynamic kv cache retrieval (2025), <https://arxiv.org/abs/2512.12284>
 21. Lee, D., Mukherjee, S., Kveton, B., Rossi, R.A., Lai, V.D., Yoon, S., Bui, T., Dernoncourt, F., Bansal, M.: Streamgaze: Gaze-guided temporal reasoning and proactive understanding in streaming videos (2025), <https://arxiv.org/abs/2512.01707>
 22. Li, J., Huang, S., Jin, Z., Zhang, C., Cao, P., Chen, Y., Liu, K., Zhao, J.: Mmr-life: Piecing together real-life scenes for multimodal multi-image reasoning. arXiv preprint arXiv:2603.02024 (2026)
 23. Li, Y., Li, J., Yu, K., Xiao, X., Liu, D., Wang, T., Tang, R.: Personalize your large vision-language models with in-context prompt tuning. arXiv preprint arXiv:2605.31513 (2026)
 24. Lin, J., Fang, Z., Chen, C., Wan, Z., Luo, F., Li, P., Liu, Y., Sun, M.: Streamingbench: Assessing the gap for mllms to achieve streaming video understanding (2024), <https://arxiv.org/abs/2411.03628>

25. Lin, J., Tong, J., Wu, H., Zhang, J., Liu, J., Jin, X., Shen, X.: Speak while watching: Unleashing true real-time video understanding capability of multimodal large language models (2026), <https://arxiv.org/abs/2601.06843>
26. Liu, J., Yu, Z., Lan, S., Wang, S., Fang, R., Kautz, J., Li, H., Álvarez, J.M.: Streamchat: Chatting with streaming video. CoRR **abs/2412.08646** (2024). <https://doi.org/10.48550/ARXIV.2412.08646>, <https://doi.org/10.48550/arXiv.2412.08646>
27. Long, L., He, Y., Ye, W., Pan, Y., Lin, Y., Li, H., Zhao, J., Li, W.: Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. CoRR **abs/2508.09736** (2025). <https://doi.org/10.48550/ARXIV.2508.09736>, <https://doi.org/10.48550/arXiv.2508.09736>
28. Maeda, T., Cao, J., Ukita, N., Kitani, K.: Cacheflow: Fast human motion prediction by cached normalizing flow. CoRR **abs/2505.13140** (2025). <https://doi.org/10.48550/ARXIV.2505.13140>, <https://doi.org/10.48550/arXiv.2505.13140>
29. Men, T., Jin, Z., Cao, P., Chen, Y., Liu, K., Zhao, J.: Empowering gui agents via autonomous experience exploration and hindsight experience utilization for task planning. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 36090–36108 (2026)
30. Ning, Z., Liu, G., Jin, Q., Ding, W., Guo, M., Zhao, J.: Livevlm: Efficient online video understanding via streaming-oriented KV cache and retrieval. CoRR **abs/2505.15269** (2025). <https://doi.org/10.48550/ARXIV.2505.15269>, <https://doi.org/10.48550/arXiv.2505.15269>
31. Niu, J., Li, Y., Miao, Z., Ge, C., Zhou, Y., He, Q., Dong, X., Duan, H., Ding, S., Qian, R., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J.: Ovo-bench: How far is your video-llms from real-world online video understanding? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11–15, 2025. pp. 18902–18913. Computer Vision Foundation / IEEE (2025). <https://doi.org/10.1109/CVPR52734.2025.01761>, 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
32. OpenAI, :, Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A.T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C.J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D.P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F.P., Raso, F., Zhang, F., von Lohmann, F.,

- Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H.P., Ren, H., Chang, H., Chung, H.W., Kivlichan, I., O’Connell, I., O’Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J.G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J.Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J.W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Afak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljube, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M.J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M.O.T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Gode-ment, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R.G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunningham, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., Malkov, Y.: Gpt-4o system card (2024), <https://arxiv.org/abs/2410.21276>
33. Qian, R., Ding, S., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., Wang, J.: Dispidder: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 24045–24055. Computer Vision Foundation / IEEE (2025). <https://doi.org/10.1109/CVPR52734.2025.02239>, https://openaccess.thecvf.com/content/CVPR2025/html/Qian_Dispidder_Enabling_Video_LLMs_with_Active_

Real-Time Interaction via Disentangled CVPR_2025_paper.html

34. Rabe, M.N., Staats, C.: Self-attention does not need $o(n^2)$ memory (2022), <https://arxiv.org/abs/2112.05682>
35. Silberschatz, A., Galvin, P.B., Gagne, G.: Operating System Concepts, 10th Edition. Wiley (2018), <http://os-book.com/OS10/index.html>
36. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., Tacchetti, A., Gaffney, C., Daruki, S., Sercinoglu, O., Gleicher, Z., Love, J., Voigtlaender, P., Jain, R., Surita, G., Mohamed, K., Blevins, R., Ahn, J., Zhu, T., Kawintiranon, K., Firat, O., Gu, Y., Zhang, Y., Rahtz, M., Faruqui, M., Clay, N., Gilmer, J., Co-Reyes, J., Penchev, I., Zhu, R., Morioka, N., Hui, K., Haridasan, K., Campos, V., Mahdiah, M., Guo, M., Hassan, S., Kilgour, K., Vezzer, A., Cheng, H.T., de Liedekerke, R., Goyal, S., Barham, P., Strouse, D., Noury, S., Adler, J., Sundararajan, M., Vikram, S., Lepikhin, D., Paganini, M., Garcia, X., Yang, F., Valter, D., Trebacz, M., Vodrahalli, K., Asawaroengchai, C., Ring, R., Kalb, N., Soares, L.B., Brahma, S., Steiner, D., Yu, T., Mentzer, F., He, A., Gonzalez, L., Xu, B., Kaufman, R.L., Shafey, L.E., Oh, J., Hennigan, T., van den Driessche, G., Odooom, S., Lucic, M., Roelofs, B., Lall, S., Marathe, A., Chan, B., Ontanon, S., He, L., Teplyashin, D., Lai, J., Crone, P., Damoc, B., Ho, L., Riedel, S., Lenc, K., Yeh, C.K., Chowdhery, A., Xu, Y., Kazemi, M., Amid, E., Petrushkina, A., Swersky, K., Khodaei, A., Chen, G., Larkin, C., Pinto, M., Yan, G., Badia, A.P., Patil, P., Hansen, S., Orr, D., Arnold, S.M.R., Grimstad, J., Dai, A., Douglas, S., Sinha, R., Yadav, V., Chen, X., Gribovskaya, E., Austin, J., Zhao, J., Patel, K., Komarek, P., Austin, S., Borgeaud, S., Friso, L., Goyal, A., Caine, B., Cao, K., Chung, D.W., Lamm, M., Barth-Maron, G., Kagohara, T., Olszewska, K., Chen, M., Shivakumar, K., Agarwal, R., Godhia, H., Rajwar, R., Snider, J., Dotiwalla, X., Liu, Y., Barua, A., Ungureanu, V., Zhang, Y., Batsaikhan, B.O., Wirth, M., Qin, J., Danihelka, I., Doshi, T., Chadwick, M., Chen, J., Jain, S., Le, Q., Kar, A., Gurumurthy, M., Li, C., Sang, R., Liu, F., Lamprou, L., Munoz, R., Lintz, N., Mehta, H., Howard, H., Reynolds, M., Aroyo, L., Wang, Q., Blanco, L., Cassirer, A., Griffith, J., Das, D., Lee, S., Sygnowski, J., Fisher, Z., Besley, J., Powell, R., Ahmed, Z., Paulus, D., Reitter, D., Borsos, Z., Joshi, R., Pope, A., Hand, S., Selo, V., Jain, V., Sethi, N., Goel, M., Makino, T., May, R., Yang, Z., Schalkwyk, J., Butterfield, C., Hauth, A., Goldin, A., Hawkins, W., Senter, E., Brin, S., Woodman, O., Ritter, M., Noland, E., Giang, M., Bolina, V., Lee, L., Blyth, T., Mackinnon, I., Reid, M., Sarvana, O., Silver, D., Chen, A., Wang, L., Maggiore, L., Chang, O., Attaluri, N., Thornton, G., Chiu, C.C., Bunyan, O., Levine, N., Chung, T., Eltyshev, E., Si, X., Lillicrap, T., Brady, D., Aggarwal, V., Wu, B., Xu, Y., McIlroy, R., Badola, K., Sandhu, P., Moreira, E., Stokowiec, W., Hemsley, R., Li, D., Tudor, A., Shyam, P., Rahimtoroghi, E., Haykal, S., Sprechmann, P., Zhou, X., Mincu, D., Li, Y., Addanki, R., Krishna, K., Wu, X., Frechette, A., Eyal, M., Dafoe, A., Lacey, D., Whang, J., Avrahami, T., Zhang, Y., Taropa, E., Lin, H., Toyama, D., Rutherford, E., Sano, M., Choe, H., Tomala, A., Safranek-Shrader, C., Kassner, N., Pajarskas, M., Harvey, M., Sechrist, S., Fortunato, M., Lyu, C., Elsayed, G., Kuang, C., Lottes, J., Chu, E., Jia, C., Chen, C.W., Humphreys, P., Baumli, K., Tao, C., Samuel, R., dos Santos, C.N., Andreassen, A., Rakićević, N., Grewe, D., Kumar, A., Winkler, S., Caton, J., Brock, A., Dalmia, S., Sheahan, H., Barr, I., Miao, Y., Natsev, P., Devlin, J., Behbahani, F., Prost, F., Sun, Y., Myaskovsky, A., Pillai, T.S., Hurt, D., Lazaridou, A., Xiong, X., Zheng, C., Pardo, F., Li, X., Horgan, D., Stanton,

J., Ambar, M., Xia, F., Lince, A., Wang, M., Mustafa, B., Webson, A., Lee, H., Anil, R., Wicke, M., Dozat, T., Sinha, A., Piqueras, E., Dabir, E., Upadhyay, S., Boral, A., Hendricks, L.A., Fry, C., Djolonga, J., Su, Y., Walker, J., Labanowski, J., Huang, R., Misra, V., Chen, J., Skerry-Ryan, R., Singh, A., Rijhwani, S., Yu, D., Castro-Ros, A., Changpinyo, B., Datta, R., Bagri, S., Hrafnkelsson, A.M., Maggioni, M., Zheng, D., Sulsky, Y., Hou, S., Paine, T.L., Yang, A., Riesa, J., Rogozinska, D., Marcus, D., Badawy, D.E., Zhang, Q., Wang, L., Miller, H., Greer, J., Sjos, L.L., Nova, A., Zen, H., Chaabouni, R., Rosca, M., Jiang, J., Chen, C., Liu, R., Sainath, T., Krikun, M., Polozov, A., Lepiau, J.B., Newlan, J., Cankara, Z., Kwak, S., Xu, Y., Chen, P., Coenen, A., Meyer, C., Tsihla, K., Ma, A., Gottweis, J., Xing, J., Gu, C., Miao, J., Frank, C., Cankara, Z., Ganapathy, S., Dasgupta, I., Hughes-Fitt, S., Chen, H., Reid, D., Rong, K., Fan, H., van Amersfoort, J., Zhuang, V., Cohen, A., Gu, S.S., Mohanane, A., Ilic, A., Tobin, T., Wieting, J., Bortsova, A., Thacker, P., Wang, E., Caveness, E., Chiu, J., Sezener, E., Kaskasoli, A., Baker, S., Millican, K., Elhawaty, M., Aisopos, K., Lebsack, C., Byrd, N., Dai, H., Jia, W., Wiethoff, M., Davoodi, E., Weston, A., Yagati, L., Ahuja, A., Gao, I., Pundak, G., Zhang, S., Azzam, M., Sim, K.C., Caelles, S., Keeling, J., Sharma, A., Swing, A., Li, Y., Liu, C., Bostock, C.G., Bansal, Y., Nado, Z., Anand, A., Lipschultz, J., Karmarkar, A., Proleev, L., Ittycheriah, A., Yeganeh, S.H., Polovets, G., Faust, A., Sun, J., Rustemi, A., Li, P., Shivanna, R., Liu, J., Welty, C., Lebron, F., Baddepudi, A., Krause, S., Parisotto, E., Soricut, R., Xu, Z., Bloxwich, D., Johnson, M., Neyshabur, B., Mao-Jones, J., Wang, R., Ramasesh, V., Abbas, Z., Guez, A., Segal, C., Nguyen, D.D., Svensson, J., Hou, L., York, S., Milan, K., Bridgers, S., Gworek, W., Tagliasacchi, M., Lee-Thorp, J., Chang, M., Guseynov, A., Hartman, A.J., Kwong, M., Zhao, R., Kashem, S., Cole, E., Miech, A., Tanburn, R., Phuong, M., Pavetic, F., Cevey, S., Comanescu, R., Ives, R., Yang, S., Du, C., Li, B., Zhang, Z., Iinuma, M., Hu, C.H., Roy, A., Bijwadia, S., Zhu, Z., Martins, D., Saputro, R., Gergely, A., Zheng, S., Jia, D., Antonoglou, I., Sadosky, A., Gu, S., Bi, Y., Andreev, A., Samangooei, S., Khan, M., Kocisky, T., Filos, A., Kumar, C., Bishop, C., Yu, A., Hodkinson, S., Mittal, S., Shah, P., Moufarek, A., Cheng, Y., Bloniarz, A., Lee, J., Pejman, P., Michel, P., Spencer, S., Feinberg, V., Xiong, X., Savinov, N., Smith, C., Shakeri, S., Tran, D., Chesus, M., Bohnet, B., Tucker, G., von Glehn, T., Muir, C., Mao, Y., Kazawa, H., Slone, A., Soparkar, K., Shrivastava, D., Cobon-Kerr, J., Sharman, M., Pavagadhi, J., Araya, C., Misiunas, K., Ghelani, N., Laskin, M., Barker, D., Li, Q., Briukhov, A., Houlsby, N., Glaese, M., Lakshminarayanan, B., Schucher, N., Tang, Y., Collins, E., Lim, H., Feng, F., Recasens, A., Lai, G., Magni, A., Cao, N.D., Siddhant, A., Ashwood, Z., Orbay, J., Dehghani, M., Brennan, J., He, Y., Xu, K., Gao, Y., Saroufim, C., Molloy, J., Wu, X., Arnold, S., Chang, S., Schrittwieser, J., Buchatskaya, E., Radpour, S., Polacek, M., Giordano, S., Bapna, A., Tokumine, S., Hellendoorn, V., Sottiaux, T., Cogan, S., Severyn, A., Saleh, M., Thakoor, S., Shefey, L., Qiao, S., Gaba, M., Yin Chang, S., Swanson, C., Zhang, B., Lee, B., Rubenstein, P.K., Song, G., Kwiatkowski, T., Koop, A., Kannan, A., Kao, D., Schuh, P., Stjerngren, A., Ghiasi, G., Gibson, G., Vilnis, L., Yuan, Y., Ferreira, F.T., Kamath, A., Klimenko, T., Franko, K., Xiao, K., Bhattacharya, I., Patel, M., Wang, R., Morris, A., Strudel, R., Sharma, V., Choy, P., Hashemi, S.H., Landon, J., Finkelstein, M., Jhakra, P., Frye, J., Barnes, M., Mauger, M., Daun, D., Baatarsukh, K., Tung, M., Farhan, W., Michalewski, H., Viola, F., de Chaumont Quitry, F., Lan, C.L., Hudson, T., Wang, Q., Fischer, F., Zheng, I., White, E., Dragan, A., baptiste Alayrac, J., Ni, E., Pritzel, A., Iwanicki, A., Isard, M., Bulanova, A., Zilka, L., Dyer, E., Sachan, D., Srinivasan, S.,

Muckenhirn, H., Cai, H., Mandhane, A., Tariq, M., Rae, J.W., Wang, G., Ayoub, K., FitzGerald, N., Zhao, Y., Han, W., Alberti, C., Garrette, D., Krishnakumar, K., Gimenez, M., Levskaya, A., Sohn, D., Matak, J., Iturrate, I., Chang, M.B., Xiang, J., Cao, Y., Ranka, N., Brown, G., Hutter, A., Mirrokni, V., Chen, N., Yao, K., Egyed, Z., Galilee, F., Liechty, T., Kallakuri, P., Palmer, E., Ghemawat, S., Liu, J., Tao, D., Thornton, C., Green, T., Jasarevic, M., Lin, S., Cotruta, V., Tan, Y.X., Fiedel, N., Yu, H., Chi, E., Neitz, A., Heitkaemper, J., Sinha, A., Zhou, D., Sun, Y., Kaed, C., Hulse, B., Mishra, S., Georgaki, M., Kudugunta, S., Farabet, C., Shafran, I., Vlasic, D., Tsitsulin, A., Ananthanarayanan, R., Carin, A., Su, G., Sun, P., V, S., Carvajal, G., Broder, J., Comsa, I., Repina, A., Wong, W., Chen, W.W., Hawkins, P., Filonov, E., Loher, L., Hirschschall, C., Wang, W., Ye, J., Burns, A., Cate, H., Wright, D.G., Piccinini, F., Zhang, L., Lin, C.C., Gog, I., Kulzhinskaya, Y., Sreevatsa, A., Song, S., Cobo, L.C., Iyer, A., Tekur, C., Garrido, G., Xiao, Z., Kemp, R., Zheng, H.S., Li, H., Agarwal, A., Ngani, C., Goshvadi, K., Santamaria-Fernandez, R., Fica, W., Chen, X., Gorgolewski, C., Sun, S., Garg, R., Ye, X., Eslami, S.M.A., Hua, N., Simon, J., Joshi, P., Kim, Y., Tenney, I., Potluri, S., Thiet, L.N., Yuan, Q., Luisier, F., Chronopoulou, A., Scellato, S., Srinivasan, P., Chen, M., Koverkathu, V., Dalibard, V., Xu, Y., Saeta, B., Anderson, K., Sellam, T., Fernando, N., Huot, F., Jung, J., Varadarajan, M., Quinn, M., Raul, A., Le, M., Habalov, R., Clark, J., Jalan, K., Bullard, K., Singhal, A., Luong, T., Wang, B., Rajayogam, S., Eisenschlos, J., Jia, J., Finchelstein, D., Yakubovich, A., Balle, D., Fink, M., Agarwal, S., Li, J., Dvijotham, D., Pal, S., Kang, K., Konzelmann, J., Beattie, J., Dousse, O., Wu, D., Crocker, R., Elkind, C., Jonnalagadda, S.R., Lee, J., Holtmann-Rice, D., Kallarackal, K., Liu, R., Vnukov, D., Vats, N., Invernizzi, L., Jafari, M., Zhou, H., Taylor, L., Prendki, J., Wu, M., Eccles, T., Liu, T., Kopparapu, K., Beaufays, F., Angermueller, C., Marzoca, A., Sarcar, S., Dib, H., Stanway, J., Perbet, F., Trdin, N., Sterneck, R., Khorlin, A., Li, D., Wu, X., Goenka, S., Madras, D., Goldshtein, S., Gierke, W., Zhou, T., Liu, Y., Liang, Y., White, A., Li, Y., Singh, S., Bahargam, S., Epstein, M., Basu, S., Lao, L., Ozturel, A., Crous, C., Zhai, A., Lu, H., Tung, Z., Gaur, N., Walton, A., Dixon, L., Zhang, M., Globerson, A., Uy, G., Bolt, A., Wiles, O., Nasr, M., Shumailov, I., Selvi, M., Piccinno, F., Aguilar, R., McCarthy, S., Khalman, M., Shukla, M., Galic, V., Carpenter, J., Villela, K., Zhang, H., Richardson, H., Martens, J., Bosnjak, M., Belle, S.R., Seibert, J., Alnahlawi, M., McWilliams, B., Singh, S., Louis, A., Ding, W., Popovici, D., Simicich, L., Knight, L., Mehta, P., Gupta, N., Shi, C., Fatehi, S., Mitrovic, J., Grills, A., Pagadora, J., Munkhdalai, T., Petrova, D., Eisenbud, D., Zhang, Z., Yates, D., Mittal, B., Tripuraneni, N., Assael, Y., Brovelli, T., Jain, P., Velimirovic, M., Akbulut, C., Mu, J., Macherey, W., Kumar, R., Xu, J., Qureshi, H., Comanici, G., Wiesner, J., Gong, Z., Ruddock, A., Bauer, M., Felt, N., GP, A., Arnab, A., Zelle, D., Rothfuss, J., Rosgen, B., Shenoy, A., Seybold, B., Li, X., Mudigonda, J., Erdogan, G., Xia, J., Simsa, J., Michi, A., Yao, Y., Yew, C., Kan, S., Caswell, I., Radebaugh, C., Elisseeff, A., Valenzuela, P., McKinney, K., Paterson, K., Cui, A., Latorre-Chimoto, E., Kim, S., Zeng, W., Durden, K., Ponnapalli, P., Sosea, T., Choquette-Choo, C.A., Manyika, J., Robenek, B., Vashisht, H., Pereira, S., Lam, H., Velic, M., Owusu-Afriyie, D., Lee, K., Bolukbasi, T., Parrish, A., Lu, S., Park, J., Venkatraman, B., Talbert, A., Rosique, L., Cheng, Y., Sozanschi, A., Paszke, A., Kumar, P., Austin, J., Li, L., Salama, K., Perz, B., Kim, W., Dukkipati, N., Baryshnikov, A., Kaplanis, C., Sheng, X., Chervonyi, Y., Unlu, C., de Las Casas, D., Askham, H., Tunyasuvunakool, K., Gimeno, F., Poder, S., Kwak, C., Miecznikowski, M., Mirrokni, V., Dimitriev, A., Parisi, A., Liu, D., Tsai,

- T., Shevlane, T., Kouridi, C., Garmon, D., Goedeckemeyer, A., Brown, A.R., Vijayakumar, A., Elqursh, A., Jazayeri, S., Huang, J., Carthy, S.M., Hoover, J., Kim, L., Kumar, S., Chen, W., Biles, C., Bingham, G., Rosen, E., Wang, L., Tan, Q., Engel, D., Pongetti, F., de Cesare, D., Hwang, D., Yu, L., Pullman, J., Narayanan, S., Levin, K., Gopal, S., Li, M., Aharoni, A., Trinh, T., Lo, J., Casagrande, N., Vij, R., Matthey, L., Ramadhana, B., Matthews, A., Carey, C., Johnson, M., Goranova, K., Shah, R., Ashraf, S., Dasgupta, K., Larsen, R., Wang, Y., Vuyyuru, M.R., Jiang, C., Ijazi, J., Osawa, K., Smith, C., Boppana, R.S., Bilal, T., Koizumi, Y., Xu, Y., Altun, Y., Shabat, N., Bariach, B., Korchemniy, A., Choo, K., Ronneberger, O., Iwuanyanwu, C., Zhao, S., Soergel, D., Hsieh, C.J., Cai, I., Iqbal, S., Sundermeyer, M., Chen, Z., Bursztein, E., Malaviya, C., Biadsy, F., Shroff, P., Dhillon, I., Latkar, T., Dyer, C., Forbes, H., Nicosia, M., Nikolaev, V., Greene, S., Georgiev, M., Wang, P., Martin, N., Sedghi, H., Zhang, J., Banzal, P., Fritz, D., Rao, V., Wang, X., Zhang, J., Patraucean, V., Du, D., Mordatch, I., Jurin, I., Liu, L., Dubey, A., Mohan, A., Nowakowski, J., Ion, V.D., Wei, N., Tojo, R., Raad, M.A., Hudson, D.A., Keshava, V., Agrawal, S., Ramirez, K., Wu, Z., Nguyen, H., Liu, J., Sewak, M., Petrini, B., Choi, D., Philips, I., Wang, Z., Bica, I., Garg, A., Wilkiewicz, J., Agrawal, P., Li, X., Guo, D., Xue, E., Shaik, N., Leach, A., Khan, S.M., Wiesinger, J., Jerome, S., Chakladar, A., Wang, A.W., Ornduff, T., Abu, F., Ghaffarkhah, A., Wainwright, M., Cortes, M., Liu, F., Maynez, J., Terzis, A., Samangouei, P., Mansour, R., Kepa, T., Aubet, F.X., Algymr, A., Banica, D., Weisz, A., Orban, A., Senges, A., Andrejczuk, E., Geller, M., Santo, N.D., Anklin, V., Merey, M.A., Baeuml, M., Strohman, T., Bai, J., Petrov, S., Wu, Y., Hassabis, D., Kavukcuoglu, K., Dean, J., Vinyals, O.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024), <https://arxiv.org/abs/2403.05530>
37. Tong, J., Mou, Y., Li, H., Li, M., Yang, Y., Zhang, M., Chen, Q., Liang, T., Hu, X., Zheng, Y., Chen, X., Zhao, J., Huang, X., Qiu, X.: Thinking with video: Video generation as a promising multimodal reasoning paradigm. CoRR **abs/2511.04570** (2025). <https://doi.org/10.48550/ARXIV.2511.04570>, <https://doi.org/10.48550/arXiv.2511.04570>
 38. Tong, J., Fan, Y., Zhao, A., Ma, Y., Shen, X.: Streamingthinker: Large language models can think while reading. CoRR **abs/2510.17238** (2025). <https://doi.org/10.48550/ARXIV.2510.17238>, <https://doi.org/10.48550/arXiv.2510.17238>
 39. Tong, J., Fu, J., Lin, Z., Fan, Y., Zhao, A., Su, H., Shen, X.: LLM as effective streaming processor: Bridging streaming-batch mismatches with group position encoding. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025. pp. 23497–23517. Association for Computational Linguistics (2025), <https://aclanthology.org/2025.findings-acl.1207/>
 40. Wang, C., Tan, X., Cao, J., Li, K., Chen, T.: Curvestream: Boosting streaming video understanding in mlms via curvature-aware hierarchical visual memory management. CoRR **abs/2603.19571** (2026). <https://doi.org/10.48550/ARXIV.2603.19571>, <https://doi.org/10.48550/arXiv.2603.19571>
 41. Wang, H., Fei, H., Dai, B., Gui, J.: Multimodal skeleton-based action representation learning via decomposition and composition. Machine Intelligence Research pp. 1–13 (2026)
 42. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model’s perception of the world

- at any resolution. CoRR **abs/2409.12191** (2024). <https://doi.org/10.48550/ARXIV.2409.12191>, <https://doi.org/10.48550/arXiv.2409.12191>
43. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., Ding, M., Tang, J.: Lvbench: An extreme long video understanding benchmark (2025), <https://arxiv.org/abs/2406.08035>
 44. Wang, Y., Li, Z., Qian, T., Zheng, H., Wang, Z., Fu, Y., Wang, X.: Streameqa: Towards streaming video understanding for embodied scenarios (2025), <https://arxiv.org/abs/2512.04451>
 45. Xiong, H., Yang, Z., Yu, J., Zhuge, Y., Zhang, L., Zhu, J., Lu, H.: Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), <https://openreview.net/forum?id=JbPb6RieNC>
 46. Xu, R., Xiao, G., Chen, Y., He, L., Peng, K., Lu, Y., Han, S.: Streamingvlm: Real-time understanding for infinite video streams. CoRR **abs/2510.09608** (2025). <https://doi.org/10.48550/ARXIV.2510.09608>, <https://doi.org/10.48550/arXiv.2510.09608>
 47. Xun, S., Tao, S., Li, J., Shi, Y., Lin, Z., Zhu, Z., Yan, Y., Li, H., Zhang, L., Wang, S., Liu, Y., Zhang, H., Ma, Y., Hu, X.: Rtv-bench: Benchmarking MLLM continuous perception, understanding and reasoning through real-time video. CoRR **abs/2505.02064** (2025). <https://doi.org/10.48550/ARXIV.2505.02064>, <https://doi.org/10.48550/arXiv.2505.02064>
 48. Yan, W., Dai, Y., Ran, Q., Li, H., Lin, W., Liao, H., Xie, X., Jin, T., Lian, J.: Proact-vl: A proactive videollm for real-time AI companions. CoRR **abs/2603.03447** (2026). <https://doi.org/10.48550/ARXIV.2603.03447>, <https://doi.org/10.48550/arXiv.2603.03447>
 49. Yang, H., Tang, F., Zhao, L., An, X., Hu, M., Li, H., Zhuang, X., Wang, B., Lu, Y., Zhang, X., Swikir, A., He, J., Ge, Z., Razzak, I.: Streamagent: Towards anticipatory agents for streaming video understanding. CoRR **abs/2508.01875** (2025). <https://doi.org/10.48550/ARXIV.2508.01875>, <https://doi.org/10.48550/arXiv.2508.01875>
 50. Yang, Z., Hu, Y., Du, Z., Xue, D., Qian, S., Wu, J., Yang, F., Dong, W., Xu, C.: Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), <https://openreview.net/forum?id=Hz4BYVY8YM>
 51. Yang, Z., Zhang, K., Hu, Y., Wang, B., Qian, S., Wen, B., Yang, F., Gao, T., Dong, W., Xu, C.: Livestar: Live streaming assistant for real-world online video understanding. CoRR **abs/2511.05299** (2025). <https://doi.org/10.48550/ARXIV.2511.05299>, <https://doi.org/10.48550/arXiv.2511.05299>
 52. Yao, L., Li, Y., Wei, Y., Li, L., Ren, S., Liu, Y., Ouyang, K., Wang, L., Li, S., Li, S., Kong, L., Liu, Q., Zhang, Y., Sun, X.: Timechat-online: 80% visual tokens are naturally redundant in streaming videos. CoRR **abs/2504.17343** (2025). <https://doi.org/10.48550/ARXIV.2504.17343>, <https://doi.org/10.48550/arXiv.2504.17343>
 53. Zeng, X., Qiu, K., Zhang, Q., Li, X., Wang, J., Li, J., Yan, Z., Tian, K., Tian, M., Zhao, X., Wang, Y., Wang, L.: Streamforest: Efficient online video understanding with persistent event memory. CoRR **abs/2509.24871** (2025). <https://doi.org/10.48550/ARXIV.2509.24871>, <https://doi.org/10.48550/arXiv.2509.24871>

54. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. CoRR **abs/2508.04416** (2025). <https://doi.org/10.48550/ARXIV.2508.04416>, <https://doi.org/10.48550/arXiv.2508.04416>
55. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Jin, X.: Flash-vstream: Efficient real-time understanding for long video streams. CoRR **abs/2506.23825** (2025). <https://doi.org/10.48550/ARXIV.2506.23825>, <https://doi.org/10.48550/arXiv.2506.23825>
56. Zhang, Y., Shi, C., Wang, Y., Yang, S.: Eyes wide open: Ego proactive video-llm for streaming video. CoRR **abs/2510.14560** (2025). <https://doi.org/10.48550/ARXIV.2510.14560>, <https://doi.org/10.48550/arXiv.2510.14560>
57. Zhao, Q., Li, Y., Sun, Q., Yan, Z.: Resilphase: Plug-and-play phase mapping and noise-resilient macro-trajectory extrapolation for diffusion acceleration (2026), <https://arxiv.org/abs/2606.26769>
58. Zhou, X., Arnab, A., Buch, S., Yan, S., Myers, A., Xiong, X., Nagrani, A., Schmid, C.: Streaming dense video captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 18243–18252. IEEE (2024). <https://doi.org/10.1109/CVPR52733.2024.01727>, <https://doi.org/10.1109/CVPR52733.2024.01727>
59. Zhu, K., Jin, Z., Yuan, H., Li, J., Tu, S., Cao, P., Chen, Y., Liu, K., Zhao, J.: Mmr-v: What’s left unsaid? a benchmark for multimodal deep reasoning in videos. arXiv preprint arXiv:2506.04141 (2025)