SPARE: SHARED PERCEPTION AND ADAPTIVE RETRIEVAL OF EVIDENCE FOR STREAMING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video understanding requires answering questions from the observed video, with evidence drawn from recent observations or earlier events. Historical context can improve recall but may interfere with judgments based on recent frames. Its usefulness depends on both when it is introduced and which clips are selected, while relevance alone does not guarantee sufficient evidence for an answer. We propose SPARE, a framework that combines adaptive historical retrieval with visual encoder sharing. A lightweight gate predicts whether recent observations provide sufficient evidence to answer the question. When additional evidence is predicted to be necessary, a multimodal embedding model retrieves relevant historical clips to supplement the current context. LoRA instruction tuning trains the understanding model to use recent observations and retrieved history jointly and abstain when the evidence remains insufficient. To reduce redundant visual parameters, the understanding and retrieval branches share a visual encoder while retaining separate language backbones. Four linear adapters align the shared features with the retrieval model through distillation at the visual token and final retrieval embedding levels. With visual-only inputs, SPARE achieves 77.78% on the evaluated Real-time and Backward task groups of OVO-Bench and 73.31% on StreamingBench. Compared with Qwen3-VL-8B-Instruct evaluated at 2 fps, SPARE improves accuracy by **20.68** percentage points on OVO-Bench and **14.15** percentage points on StreamingBench.

1 INTRODUCTION

Recent multimodal large language models (MLLMs) have advanced visual perception and video event understanding (Bai et al., 2025; Qwen Team, 2026b; An et al., 2026). Streaming video understanding requires answering questions using only the video observed up to the current moment. Some questions can be answered from recent observations, whereas others require evidence from earlier events. Supporting both requires access to history while keeping the answer context and its processing cost bounded. Existing methods retain history through video KV caches (Di et al., 2025; Zhang et al., 2026), compressed visual memory (Zeng et al., 2025; Xie et al., 2026; Fu et al., 2026), or online-updated fast weights outside the attention context (Chen et al., 2026).

Historical context can supply evidence missing from the current view. A pretrained MLLM using only recent frames provides a competitive streaming baseline (Shen et al., 2026). Historical observations can improve recall of earlier events but may reduce the accuracy of current-scene perception (Shen et al., 2026; Chen et al., 2026). When recent observations contain sufficient evidence, additional history may distract the model from the information needed to answer the question. Video retrieval-augmented generation selects question-relevant historical content (Ataallah et al., 2024; Chen et al., 2025). Retrieved clips may depict the objects or scenes mentioned in a question without containing the specific event or detail required to answer it. The challenge is to obtain useful historical evidence while preserving accurate judgments grounded in current observations.

We investigate whether historical context can improve backward tracing while preserving real-time perception, using Qwen3-VL-8B-Instruct with different input frame selections (Figure 1). Adding uniformly sampled history to Recent-4 slightly improves backward tracing but reduces real-time

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.

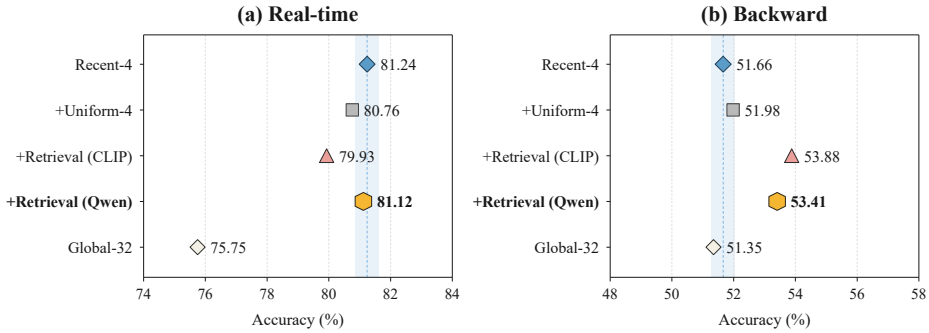


Figure 1: Effect of historical context on OVO-Bench (Niu et al., 2025). We report question-level accuracy within (a) Real-time (837 questions) and (b) Backward (631 questions). All configurations use Qwen3-VL-8B-Instruct with identical resolution and prompts, varying only input frame selection and count. Recent-4 uses up to four recent frames; +Uniform-4 and +Retrieval add up to four historical frames via uniform sampling or retrieval with CLIP ViT-L/14 or Qwen3-VL-Embedding-2B. Both retrieval variants retrieve history for every question without gating. Global-32 uniformly samples up to 32 frames from the observed video prefix. Blue dashed lines indicate Recent-4; bold entries highlight Qwen-based retrieval.

accuracy. CLIP retrieval raises backward accuracy from 51.66% to 53.88%, while lowering real-time accuracy from 81.24% to 79.93%. With the same eight-frame budget, Qwen3-VL-Embedding-2B retrieval achieves 53.41% backward accuracy and 81.12% real-time accuracy. It yields a smaller backward gain than CLIP but largely preserves real-time performance.

This comparison highlights the importance of retrieved clip quality: the value of historical evidence should be assessed by both its contribution to backward tracing and its effect on current-scene judgments. Global-Uniform-32 performs worse than Recent-4 on both task groups, showing that broader coverage alone does not ensure better answers. The experimental findings motivate considering both whether additional history is needed and which clips can usefully supplement recent observations.

To address these two considerations, we propose SPARE, a streaming video understanding framework that combines evidence sufficiency prediction with question-conditioned historical retrieval. A lightweight head assesses whether recent observations suffice to answer: if so, the model answers directly; otherwise, a multimodal embedding model retrieves historical clips by their embedding similarity to the question to augment the current context. LoRA fine-tuning trains the understanding model to use recent and retrieved evidence jointly and abstain when it remains insufficient.

To reduce redundant visual parameters and encoding computation, the two branches share the understanding model's visual encoder while retaining separate language backbones. Four linear adapters map shared visual features into the retrieval model's input space, removing its standalone visual tower. The adapters are distilled from a pretrained multimodal embedding model with the same backbone architecture as the understanding model, using supervision on both visual-token features and final retrieval embeddings.

We evaluate SPARE on OVO-Bench (Niu et al., 2025) and StreamingBench (Lin et al., 2026) with visual-only inputs. SPARE achieves 77.78% on the evaluated Real-time and Backward tasks of OVO-Bench and 73.31% on StreamingBench. As shown in Table 1, SPARE outperforms Qwen3-VL-8B-Instruct evaluated at 2 fps by 20.68 percentage points on OVO-Bench and 14.15 on StreamingBench.

Ablations show that on-demand retrieval substantially reduces history use while retaining nearly the same accuracy as always retrieving. With the same maximum number of input frames, combining recent frames with retrieved history outperforms using recent frames alone. Efficiency analyses show a mean per-question peak answering context of 2.87K tokens and stable response latency across the tested observed frame counts.

Our main contributions are summarized as follows:

- We propose SPARE, a streaming video understanding framework that predicts whether recent observations provide sufficient evidence and retrieves history based on the question when needed. LoRA fine-tuning trains the understanding model to use available evidence and abstain when it remains insufficient.
- We share the visual encoder between understanding and retrieval through four linear adapters trained with distillation on visual token features and final retrieval embeddings. This removes the retrieval branch's standalone visual encoder while retaining both language backbones.
- We demonstrate improvements over native Instruct models on OVO-Bench and Streaming-Bench using only visual inputs. Ablations examine the effects of encoder sharing, LoRA tuning, and input frame allocation.

2 Related Works

2.1 Streaming Video Understanding

Offline video understanding assumes a fully observed sequence, whereas streaming methods maintain access to past observations as new frames arrive. They use visual-token pruning, KV-state selection, recurrent encoding, or hierarchical memory (Yao et al., 2025; Di et al., 2025; Qian et al., 2024; Zeng et al., 2025; Xie et al., 2026); other systems decouple continuous perception from response generation (Qian et al., 2025). Recent-window baselines show that history can improve recall while impairing current-scene perception (Shen et al., 2026), and post-training can improve performance even under a fixed recent-window protocol (Wu et al., 2026).

2.2 Image–Text Matching and Multimodal Embeddings

CLIP learns comparable image–text embeddings through dual encoders and symmetric InfoNCE, using paired examples as positives and other batch elements as negatives (Radford et al., 2021). SigLIP instead applies independent sigmoid losses, avoiding global batch normalization (Zhai et al., 2023). Multimodal embedding models extend this approach to instruction-conditioned retrieval over text, natural and document images, and video (Jiang et al., 2025; Li et al., 2026). Multi-task contrastive learning, hard negatives, and reranker distillation adapt similarity to different relevance criteria; related work also studies video-conditioned answerability (Yoon et al., 2025). We use pretrained multimodal embeddings to index historical clips once and rank them for later queries, with visual alignment adapting the shared encoder to the retrieval space.

3 Method

3.1 Overview

SPARE combines video understanding with historical retrieval to answer questions about an observed video stream. Figure 2 summarizes the memory construction and adaptive answering pipeline. At query time t, SPARE answers q_t from the observed video prefix $\mathcal{V}_{\leq t}$. The current context $\mathcal{C}_t$ contains the most recent n frames, while earlier clips populate an incrementally updated visual memory $\mathcal{M}_t$. The MLP retrieval gate (labeled Retrieval Head in Figure 2) assesses whether the current context provides sufficient evidence to answer the question. If the evidence is insufficient and history is available, we construct a fresh input with the retrieved historical clips preceding the current frames, preserving their temporal order for a second answering pass. The same LoRA adapter in the understanding model is enabled in both passes, allowing the model to answer or abstain based on the supplied evidence.

The understanding and retrieval components can operate independently with their native visual encoders and language backbones. SPARE connects them through the clips selected from memory and optionally shares their visual encoder to remove the parameter overhead of a second tower. Figure 3 contrasts the native and shared configurations. In the shared configuration, each model retains its own language backbone and runs the shared encoder on its own inputs. Lightweight adapters map the shared visual features into the retrieval model's input space. We describe the visual memory, retrieval decision, and answer supervision below.

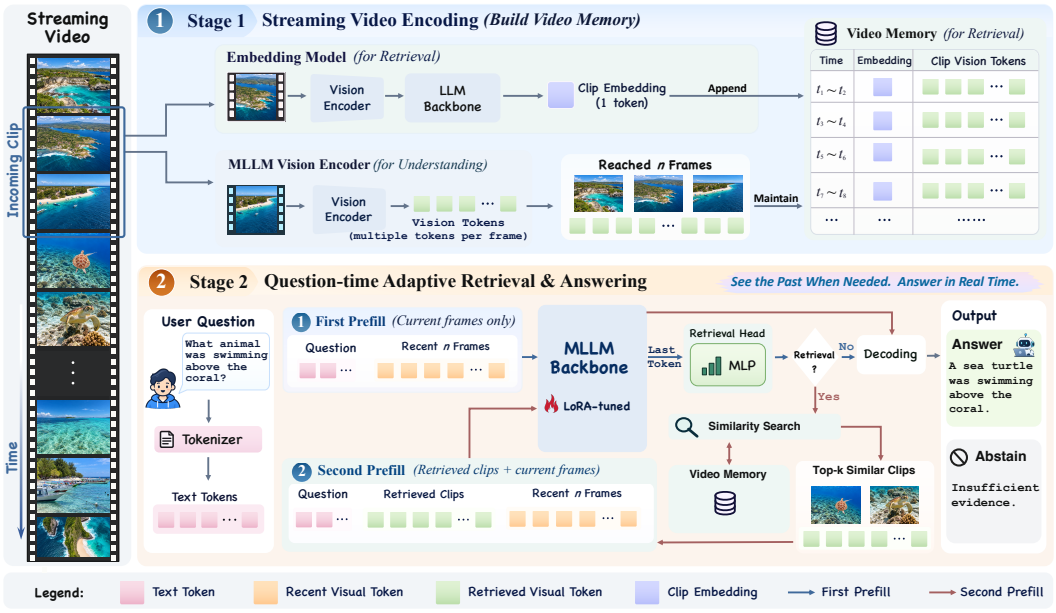


Figure 2: Overview of SPARE. **Memory construction:** the two branches share visual-encoder weights and index historical clips by embeddings and timestamps. **Adaptive answering:** a gate reads the current-input prefill state and triggers historical retrieval when evidence is insufficient. The retrieved clips and current window form a fresh second-pass input; otherwise, the first-pass answer is retained. LoRA tuning supports answering or abstention from the supplied evidence.

3.2 Shared-Encoder Visual Memory Retrieval

To select historical clips more relevant to the question, we adopt Qwen3-VL-Embedding-2B (Li et al., 2026) as the retrieval model in our system, based on its stronger video retrieval performance (Appendix B.1).

Clip-level visual memory. Qwen3-VL-Embedding maps historical video clips and textual queries into a common retrieval space. We divide the stream into non-overlapping Δ-second clips $X_i = \{x_{i,1}, \ldots, x_{i,m}\}$ with intervals $\tau_i = [b_i, e_i)$, sampled at $f_s = 1$ FPS with $m = \Delta f_s$ frames. Each completed clip is encoded as

$$\mathbf{z}_i = \mathcal{N}\big(E_{\text{vis}}(X_i)\big), \qquad \mathcal{N}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}. \tag{1}$$

where E_{vis} is the complete clip-to-embedding path and $\mathbf{z}_i \in \mathbb{R}^d$. For current-context start time s_t, the eligible memory is

$$\mathcal{M}_t = \{(\mathbf{z}_i, \tau_i, \pi_i) \mid e_i \le s_t\}, \tag{2}$$

where π_i indexes the original visual clip, separating search vectors from answering inputs and excluding overlap with the current window. The text path E_{text} ranks eligible clips by cosine similarity:

$$\begin{aligned} \mathbf{z}_{q_t} &= \mathcal{N}\big(E_{\text{text}}(q_t)\big), \\ s_{t,i} &= \mathbf{z}_{q_t}^{\top} \mathbf{z}_i, \\ \mathcal{I}_t &= \text{TopK}_{i:e_i \le s_t}(s_{t,i}). \end{aligned} \tag{3}$$

TopK selects at most K clips, returning all if fewer are available. Their visual content and timestamps form $\widehat{\mathcal{R}}_t = \{(X_i, \tau_i) \mid i \in \mathcal{I}_t\}$, ordered chronologically for answer generation.

Shared visual encoder. To remove the retrieval branch's standalone visual tower, we reuse the visual encoder of the Instruct model, as described in Figure 3. Its features must be mapped into the representation space expected by the retrieval language model. We learn this mapping by distilling the original Embedding-2B model into a student with the shared encoder. Four independent linear

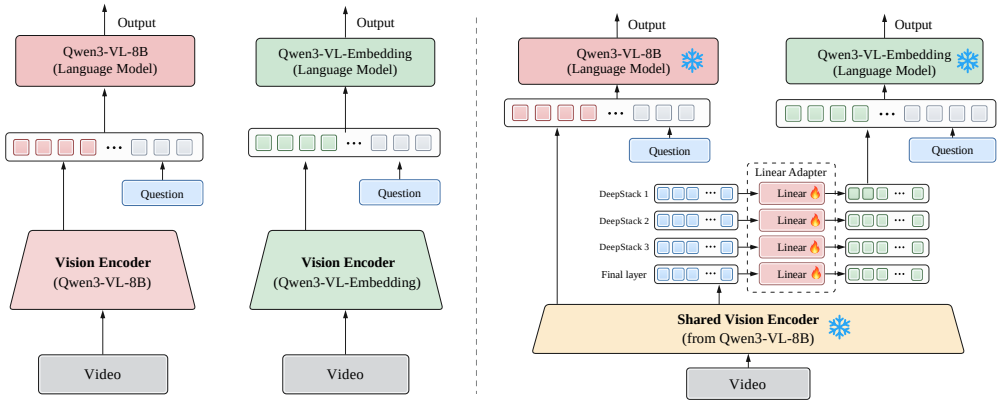


Figure 3: Shared visual encoding for understanding and retrieval. Left: separate visual encoders serve the two language models. Right: SPARE reuses the understanding model's visual encoder and adapts its three DeepStack merger features and final merger features through four linear adapters for the retrieval branch. Each branch retains its own language backbone.

adapters read the features immediately before the shared mergers' output projections: three DeepStack branches and the final branch. They map these features into the retrieval model's input space while preserving the understanding branch's original projections. Let $\mathbf{H}_j$ be the shared merger features, $\mathbf{T}_j$ the corresponding teacher outputs, and $A_{\eta,j}$ the adapters, with $j = 0, 1, 2$ denoting DeepStack and $j = 3$ the final branch. The student features and semantic embeddings are

$$\begin{aligned} \mathbf{S}_j &= A_{\eta,j}(\mathbf{H}_j), \qquad j \in \{0,1,2,3\}, \\ \mathbf{e}_s &= G_\psi(\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3), \\ \mathbf{e}_t &= G_\psi(\mathbf{T}_0, \mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3). \end{aligned} \tag{4}$$

Here, G_ψ is the frozen retrieval language model with its original representation extraction procedure; both paths use identical text inputs and visual injection schemes. To preserve the pretrained retrieval representation, we align both the visual features supplied to G_ψ and the final clip embeddings:

$$\mathcal{L}_{\text{distill}} = \sum_{j=0}^{3} w_j \Big[d_{\cos}\big(\mathbf{S}_j, \text{sg}(\mathbf{T}_j)\big) + 0.1\ \text{SmoothL1}\ \big(\mathbf{S}_j, \text{sg}(\mathbf{T}_j)\big) \Big] + d_{\cos}\big(\mathbf{e}_s, \text{sg}(\mathbf{e}_t)\big). \tag{5}$$

Here, w_j weights the visual alignment loss for branch j; we set $(w_0, w_1, w_2, w_3) = (0.1, 0.2, 0.3, 0.4)$ in our experiments. The distance $d_{\cos}$ averages $1 - \cos(\cdot, \cdot)$ between corresponding vectors over tokens for visual features and over samples for final embeddings. SmoothL1 averages over all elements, and sg stops gradients. Only the four adapters are updated. At inference, the student implements E_{vis} in Eq. equation 1: the shared encoder, four adapters, and G_ψ produce $\mathbf{e}_s(X_i)$, and $\mathbf{z}_i = \mathcal{N}(\mathbf{e}_s(X_i))$ is stored in memory. The teacher's visual tower is removed; the text path and both language backbones remain unchanged.

3.3 On-Demand Retrieval for Streaming Video Understanding

At query time, the model first processes the question and current frames. The retrieval gate uses the prefill state to estimate whether this current evidence is insufficient, before any answer tokens are generated:

$$\mathbf{h}_t = \big[F^{(L)}_{\theta,\omega}\big(\mathcal{P}(q_t, \mathcal{C}_t, \varnothing)\big)\big]_{\text{last}}, \qquad p_t = \sigma\big(g_\gamma(\mathbf{h}_t)\big). \tag{6}$$

Here, $\mathcal{P}$ constructs the prompt, and $F_{\theta,\omega}$ is the L-layer understanding model with frozen base parameters θ and LoRA parameters ω. The state $\mathbf{h}_t$ is taken from the last valid input token after the final normalization. The retrieval gate g_γ is a two-layer MLP with layer normalization, GELU, and training-time dropout; σ maps its logit to a probability. Its input reflects the question and current frames, without retrieved content or generated answers. With a default threshold of 0.5, retrieval is triggered when $p_t \geq 0.5$ and historical memory is nonempty:

$$r_t = \mathbb{I}[p_t \geq 0.5]\, \mathbb{I}[|\mathcal{M}_t| > 0]. \tag{7}$$

$\mathbb{I}[\cdot]$ is the indicator function. This decision selects the context for the final response, as shown in Figure 2. If $r_t = 0$, the model returns its response from the current window, with $\mathcal{R}_t = \varnothing$. If $r_t = 1$, retrieval supplies $\mathcal{R}_t = \widehat{\mathcal{R}}_t$, and a second prefill processes the retrieved history followed by current frames, retaining timestamps. The response from this context becomes the final output.

The same LoRA adapter supports answering or abstention in both passes (Section 3.4). The gate decides whether to seek history; the understanding model decides whether the supplied evidence supports an answer, since retrieval may still leave the question unanswerable.

To train the gate, we label whether the current frames alone support an answer. For this current-context answerability $a_t^{\text{cur}} \in \{0, 1\}$, the target is $r_t^* = 1 - a_t^{\text{cur}}$. We freeze the understanding model and optimize the gate on cached states with unweighted binary cross-entropy:

$$\mathcal{L}_{\text{gate}} = -\mathbb{E}_{(\mathbf{h}_t, r_t^*) \sim \mathcal{D}_{\text{gate}}} \left[r_t^* \log p_t + (1 - r_t^*) \log(1 - p_t) \right]. \tag{8}$$

Here, $\mathcal{D}_{\text{gate}}$ contains cached states and current-evidence labels; the loss is implemented with logits. We train the gate on original-model states, then adapt it to LoRA-enabled states using the same labels and objective, so its inputs match the answering model at deployment. Appendices A.1.2 and A.2 provide label construction and training details.

3.4 LoRA Instruction Tuning

LoRA instruction tuning (Hu et al., 2022) teaches the model to answer or abstain based on combined current and retrieved evidence, whereas the gate assesses current frames alone. Sources cover *current-answerable* questions, *history-assisted* questions requiring earlier evidence, and *insufficient-evidence* questions unsupported by the supplied current and historical observations. We pair each question with one clip retrieved offline and the current window, in chronological order with timestamps, sampling each at 1 FPS with up to four frames. We reassess answerability on this combined input because retrieval can change the source label. Sampling and consistency filtering retain 173 examples (Appendix A.2).

For example i, $x_i = \mathcal{P}(q_i, \mathcal{C}_i, \mathcal{R}_i, \mathcal{O}_i)$ contains the question, timestamped frames, instructions, and options $\mathcal{O}_i$ when present. The combined-evidence label a_i is 1 for answerable inputs and 0 otherwise, selecting the target:

$$y_i^* = \begin{cases} b_i^{\text{ans}} \oplus e, & a_i = 1, \\ b_i^{\text{abs}} \oplus e, & a_i = 0, \end{cases} \tag{9}$$

where $e =$ `<|im_end|>` and $\oplus$ denotes concatenation. For multiple-choice questions, b_i^{ans} is the supported option letter and b_i^{abs} is the existing abstention-option letter. For open-ended questions, they are the reviewed short answer and `Unable to answer.`, respectively. The answerability label determines the target but is absent from the input.

These input–target pairs form $\mathcal{D}_{\text{LoRA}} = \{(x_i, y_i^*)\}_{i=1}^N$, on which we minimize the mean answer-only loss:

$$\begin{aligned} \mathcal{L}_{\text{LoRA}}(\omega) &= \frac{1}{|\mathcal{D}_{\text{LoRA}}|} \sum_{(x_i, y_i^*) \in \mathcal{D}_{\text{LoRA}}} \ell_i(\omega), \\ \ell_i(\omega) &= -\frac{1}{T_i} \sum_{j=1}^{T_i} \log p_{\theta,\omega}\big(y_{i,j}^* \mid x_i, y_{i,<j}^*\big). \end{aligned} \tag{10}$$

Here, T_i includes the end-of-turn token. Teacher forcing uses full-vocabulary probabilities, equal weight per example, and no loss on prompt, visual, or padding positions. We update only language-model LoRA parameters ω for one epoch, freezing the base parameters θ, visual encoder, retriever, and gate. We then freeze LoRA and adapt the gate to LoRA-enabled first-pass states; both answering passes use the same adapter. Appendix A.2 provides prompts, masking, optimization, and gate-adaptation details.

Table 1: Measured results on OVO-Bench and StreamingBench (%). For SPARE, $a + b$ denotes the frame counts in the first and retrieval-triggered second answering passes. Native and shared ViT refer to the retrieval branch's visual encoder. Bold denotes the best result and underlining the second-best result in each column, including ties.

Model	Frames / sampling	*StreamingBench*				*OVO-Bench*										
						Real-Time Visual Perception							*Backward Tracing*			
		RTVU	OSU	CU	All	OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	All
Offline Video LLMs																	
LLaVA-OneVision2-8B	16	74.20	37.90	44.80	59.60	72.48	61.47	71.55	63.48	65.35	60.87	65.87	56.57	64.86	11.29	44.24	55.05
InternVL2-8B	32 / ≤64	67.60	35.20	36.90	53.58	69.80	59.63	67.24	50.56	71.29	57.61	62.69	52.53	58.11	25.81	45.48	54.08
Dispider-7B (offline)	16/clip	68.72	42.20	29.90	54.20	57.05	49.54	62.07	44.94	60.40	51.63	54.27	48.82	55.41	4.30	36.18	45.22
Online / Streaming Video LLMs																	
Flash-VStream-7B	1 fps	53.16	29.40	23.90	41.38	41.61	53.21	48.28	29.21	60.40	42.93	45.94	41.75	41.22	1.08	28.01	36.98
TimeChat-Online-7B	1/0.5/0.2 fps	73.92	39.70	32.40	57.09	73.83	47.71	69.83	46.63	71.29	60.33	61.60	56.90	64.19	6.99	42.69	52.15
StreamForest-7B	1 fps	77.52	38.20	37.30	59.84	73.15	53.21	72.41	52.25	63.37	61.96	62.72	58.59	66.22	29.03	51.28	57.00
HERMES-7B	1 fps	78.88	49.50	44.20	64.64	88.59	66.97	80.17	54.49	72.28	71.74	72.37	47.14	63.51	39.25	49.97	61.17
StreamTTT-4B	2 fps	81.28	55.70	46.70	67.91	91.28	80.73	80.17	56.18	**82.18**	77.72	78.04	**60.94**	58.11	60.75	59.93	68.99
Open-Source Video LLMs																	
Qwen3-VL-8B-Instruct	2 fps	74.84	41.00	38.10	59.16	81.21	62.39	75.00	58.99	71.29	63.04	68.65	56.57	68.24	11.83	45.55	57.10
+ SIMPLESTREAM	4	80.68	62.80	58.50	71.78	93.29	83.49	81.90	66.29	78.22	**84.24**	81.24	52.53	56.76	52.69	53.99	67.61
+ SPARE w/ native ViT	4+8	83.08	**64.50**	**62.70**	**74.42**	95.97	84.40	**86.21**	70.22	74.26	83.15	**82.37**	58.92	66.89	43.01	56.28	69.32
+ SPARE w/ shared ViT	4+8	**83.20**	63.80	62.20	74.22	**96.64**	**85.32**	85.34	70.22	71.29	82.61	81.91	59.93	68.24	42.47	56.88	69.39
+ SPARE w/ shared ViT & LoRA	4+8	82.60	61.10	62.30	73.31	95.97	**85.32**	83.62	**70.79**	73.27	82.61	81.93	60.27	**70.27**	**90.32**	**73.62**	**77.78**

4 EXPERIMENTS

4.1 TRAINING SETTINGS

We use Qwen3-VL-8B-Instruct for understanding and Embedding-2B for retrieval. Four visual adapters are aligned on 64K image and 64K video entries with both backbones frozen. The initial retrieval gate is trained on cached states with independent current-evidence assessments and mixed question formats. LoRA instruction tuning then updates the understanding model for one epoch on 173 audited examples (78 multiple-choice and 95 open-ended), paired with retrieved history. Finally, we re-extract LoRA-enabled states and adapt the gate for one epoch on 15k audited questions using unweighted binary cross-entropy. The source videos used for gate training and adaptation and LoRA instruction tuning do not overlap with those used for evaluation in OVO-Bench or StreamingBench. Appendices A.2 and A.1 provide data construction, prompts, and optimization settings.

4.2 EVALUATION SETTINGS

Benchmarks. We evaluate streaming video understanding on OVO-Bench (Niu et al., 2025) and StreamingBench (Lin et al., 2026) with visual-only inputs. For OVO-Bench, we evaluate 1,468 questions covering Real-time Visual Perception and Backward Tracing, excluding Forward Responding. These two groups assess understanding of current observations and recall of earlier events. We compute the macro-average accuracy across tasks within each group and average the two group scores for the overall result. For StreamingBench, we evaluate all 4,500 questions from 900 videos, covering real-time visual, omni-source, and contextual understanding. We report accuracy for each category and compute the overall score by weighting the three category accuracies by question count. Appendix A.3 provides further dataset details.

Compared models. Table 1 includes offline models (Bai et al., 2025; An et al., 2026; Chen et al., 2024), Dispider in offline mode (Qian et al., 2025), streaming models (Yao et al., 2025; Zeng et al., 2025; Zhang et al., 2026; Chen et al., 2026), and the same-backbone, four-frame recent-window SimpleStream baseline (Shen et al., 2026). All are re-evaluated on the same question sets with visual-only inputs. Frame budgets and sampling strategies remain model-specific, and computational budgets are not matched.

Inference settings. SPARE samples at 1 FPS, anchored at the end of each interval. The first pass reads up to four current frames. In the shared-ViT LoRA configuration, the same LoRA-enabled model supplies the MLP gate's prefill state and generates the answer. Our evaluation implementation completes a current-only candidate answer. When the gate triggers at threshold 0.5 and history is

Table 2: Retrieval policies with the same shared ViT, LoRA adapter, and retrieval gate on OVO-Bench (%). All is the mean of the Real-time and Backward task-macro accuracies. History use is the fraction of questions whose final responses use historical frames.

Policy	Real-time	Backward	All	History use
Recent-only	81.06	68.00	74.53	0.00
Uniform history	81.05	70.70	75.87	99.46
Always retrieve	81.16	73.22	77.19	99.46
Gated retrieve	**81.93**	**73.62**	**77.78**	40.74

available, Top-1 retrieval selects an earlier non-overlapping four-second clip; a fresh second pass reads up to four historical and four current frames chronologically and replaces the candidate with its response. Otherwise, the first-pass response is returned.

4.3 Benchmark Performance

SPARE with shared ViT and LoRA achieves 77.78% on OVO-Bench (Table 1), outperforming SimpleStream by 10.17 percentage points and StreamTTT by 8.79. Compared with SimpleStream, backward tracing improves from 53.99% to 73.62%, while real-time perception rises from 81.24% to 81.93%. Hallucination detection shows the largest gain, from 52.69% to 90.32%. These results show improved backward tracing while preserving real-time perception relative to SimpleStream.

All three SPARE variants outperform SimpleStream on StreamingBench with visual-only inputs. Shared ViT with LoRA achieves 73.31%, exceeding SimpleStream by 1.53 percentage points, with higher real-time visual and contextual understanding scores but lower omni-source accuracy. Without LoRA, native and shared ViT achieve 74.42% and 74.22%. For the shared-ViT configuration, LoRA tuning and gate adaptation increase the OVO-Bench score from 69.39% to 77.78%, while the StreamingBench score decreases slightly from 74.22% to 73.31%.

Without LoRA, replacing the native retrieval encoder with shared ViT changes OVO-Bench accuracy from 69.32% to 69.39% and StreamingBench accuracy from 74.42% to 74.22%. Encoder sharing thus maintains comparable overall performance while removing the retrieval branch's standalone visual encoder.

5 Ablation Studies and Efficiency Analysis

5.1 Ablation Studies

Retrieval policy. Table 2 fixes the trained components and inference settings across four policies. Recent-only uses up to four current frames. Uniform history and Always retrieve add uniformly sampled history and the Top-1 clip, respectively, with historical frame counts matched per question at up to four. Gated retrieve adds the Top-1 clip only when the gate triggers.

Table 3: The impact of frames on OVO-Bench.

Current frames	Frames/ clip	Retrieved clips	Score (%)
2	–	–	72.47
4	–	–	74.53
8	–	–	74.87
4	2	1	76.70
4	4	1	**77.78**
4	8	1	77.02
4	16	1	77.44
4	4	2, consecutive	76.59
4	4	4, consecutive	76.64
4	4	2, Top-2	77.28
4	4	4, Top-4	76.93

Uniform history improves Backward accuracy from 68.00% to 70.70%, while Real-time accuracy remains nearly unchanged (81.06% versus 81.05%). With historical frame counts matched, Always retrieve further improves Backward accuracy to 73.22%, with 81.16% on Real-time, showing that selecting relevant clips primarily benefits backward tracing. Gated retrieve raises Real-time and Backward accuracy to 81.93% and 73.62%, improving the overall score from 77.19% to 77.78% while reducing history use from 99.46% to 40.74%. Thus, selective retrieval improves both task groups while using historical context less of-

ten. Appendix B.3 provides task scores, the comparison protocol, and the native-encoder control without LoRA.

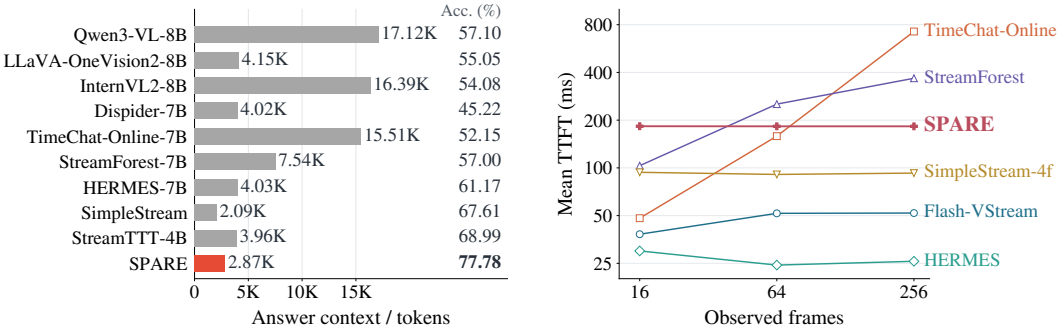
Frame allocation and clip selection. Table 3 fixes the shared ViT, LoRA adapter, and gate. Increasing current frames from four to eight yields only a small gain (74.53% to 74.87%). With the same maximum of eight input frames, four current frames plus one retrieved four-frame clip achieve 77.78%, a gain of 2.91 percentage points over eight current frames. Using eight or sixteen historical frames scores lower (77.02% and 77.44%), favoring four recent and four historical frames among the tested settings.

With four current frames and four frames per historical clip, Top-2 outperforms two consecutive clips ending with the highest-ranked clip (77.28% versus 76.59%); Top-4 likewise outperforms four consecutive clips (76.93% versus 76.64%). Both remain below single-clip retrieval (77.78%). Thus, relevance-based selection outperforms consecutive expansion, but additional clips provide no accuracy gain in these comparisons.

5.2 Efficiency Analysis

Answering-context length. SPARE uses a mean per-question peak answering context of 2.87K tokens, compared with 17.12K for Qwen3-VL-8B-Instruct (Figure 4(a)). For each question, we take the maximum context length across passes, then average over questions. Retrieved answering frames are included; retrieval memory is excluded. Relative to SimpleStream, SPARE uses an additional 0.78K tokens (2.09K to 2.87K) while improving accuracy from 67.61% to 77.78%.

Response latency. Figure 4(b) reports mean time to first token (TTFT) on OVO-Bench using one H100 80GB GPU and batch size one. Timing starts at question arrival and ends when the first token of the final response is available, including retrieval and processing needed to produce that token but excluding video preparation and state restoration. For latency measurements, SPARE reuses cached visual features and routes immediately after the gate's prefill decision, skipping first-pass answer generation when retrieval is triggered. Its mean TTFT remains near 183 ms at 16, 64, and 256 observed frames. Appendix A.3 provides the timing protocol.



(a) Answering-context length

(b) Response latency

Figure 4: Efficiency on OVO-Bench. **(a)** Mean per-question peak answering-context length and accuracy on all Real-time and Backward questions. **(b)** Mean time to first token on OVO-Bench's Real-time and Backward tasks at 16, 64, and 256 observed frames.

6 Conclusion

We introduced SPARE, a streaming video understanding framework combining evidence sufficiency prediction, historical retrieval, and LoRA tuning for answering or abstention. The understanding and retrieval components can operate independently or share a visual encoder through four distilled linear adapters, removing the retrieval branch's separate visual encoder. Experiments on OVO-Bench and StreamingBench show improved accuracy over the base model and its recent-window baseline with visual-only inputs.

AI USE STATEMENT

We used generative AI tools to assist with translation, language editing, literature search and reference formatting. We also used Qwen3.8-27B (Qwen Team, 2026a) to assess whether video evidence was sufficient to answer questions and to filter training examples for LoRA instruction tuning and retrieval-gate training. The data selection procedure retained examples only when two model assessments agreed, with additional checks on answer targets and multiple-choice options for LoRA instruction tuning, as detailed in Appendices A.2 and A.1. The authors take responsibility for the final manuscript, its claims, and the associated data and code, including content produced with AI assistance.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026. URL https://arxiv.org/abs/2605.25979.

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 *Computer Vision – ECCV 2024*, pp. 251–267. Springer, 2024. doi: 10.1007/978-3-031-73397-0_15. URL https://link.springer.com/chapter/10.1007/978-3-031-73397-0_15.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Heilbron_ActivityNet_A_Large-Scale_2015_CVPR_paper.html.

Joya Chen, Zeyun Zhong, and Mike Zheng Shou. StreamTTT: Reconciling real-time perception and long-term memory in streaming VLMs. *arXiv preprint arXiv:2608.13416*, 2026.

Tao Chen, Shaobo Ju, Qiong Wu, Chenxin Fang, Kun Zhang, Jun Peng, Hui Li, Yiyi Zhou, and Rongrong Ji. Towards effective long video understanding of multimodal large language models via one-shot clip retrieval. *arXiv preprint arXiv:2512.08410*, 2025. URL https://arxiv.org/abs/2512.08410.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. URL https://arxiv.org/abs/2412.05271.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.08691.

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 *International Conference on Learning Representations*, 2025.

Muxin Fu, Yifan Zhang, Wentao Zhang, Fangming Guo, Qian Chen, Guibin Zhang, Shuicheng Yan, and Bo An. StreamFlow: Dynamic memory flows for streaming video understanding. *arXiv preprint arXiv:2608.10949*, 2026.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.09344.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5803–5812, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Hendricks_Localizing_Moments_in_ICCV_2017_paper.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TE0KOzWYAF.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026. URL https://arxiv.org/abs/2601.04720.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 12147–12151. IEEE, 2026. URL https://ieeexplore.ieee.org/document/11463959.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Image_Retrieval_on_Real-Life_Images_With_Pre-Trained_Vision-and-Language_Models_ICCV_2021_paper.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. URL https://arxiv.org/abs/2507.04590.

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

Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. QuerYD: A video dataset with high-quality text and audio narrations. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021. URL https://www.robots.ox.ac.uk/~vgg/publications/2021/Oncescu21/.

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*, volume 37, 2024. doi: 10.52202/079017-3792.

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

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision – ECCV 2016*, pp. 510–526. Springer, 2016. doi: 10.1007/978-3-319-46448-0_31. URL https://link.springer.com/chapter/10.1007/978-3-319-46448-0_31.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Tang_COIN_A_Large-Scale_Dataset_for_Comprehensive_Instructional_Video_Analysis_CVPR_2019_paper.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

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

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. FluxMem: Adaptive hierarchical memory for streaming video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.02096.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. TimeChat-Online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025. URL https://arxiv.org/abs/2504.17343.

Eunseop Yoon, Hee Suk Yoon, Mark A. Hasegawa-Johnson, and Chang D. Yoo. Can video LLMs refuse to answer? alignment for answerability in video large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P9VdRQOyqu.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. StreamForest: Efficient online video understanding with persistent event memory. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2547.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. doi: 10.1109/ICCV51070.2023.01100.

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 (Volume 1: Long Papers)*, pp. 8411–8430. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.381. URL https://aclanthology.org/2026.acl-long.381/.

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

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8475–8489, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhao_MMVU_Measuring_Expert-Level_Multi-Discipline_Video_Understanding_CVPR_2025_paper.html.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12342. URL https://ojs.aaai.org/index.php/AAAI/article/view/12342.