Knowing Where to Look: Caption-Indexed Verification for Streaming Video Understanding

Mingu Kang Suahn Bae Junseok Kim Jiho Chang
Seoul National University
{gms5560, snubbb, kim.junseok, hoyaho03}@snu.ac.kr

Abstract

*Streaming video systems must build memory before knowing what users will later ask. This temporal separation makes caption-based compression risky: a caption can faithfully summarize an event while discarding the fine-grained visual fact required by a future query. We propose **Caption-Indexed Verification (CIV)**, a training-free proposal-verification framework for streaming video understanding. CIV treats online captions as semantic indices into the stream, not as final answer evidence. At query time, retrieved captions produce an answer proposal and select a small set of visual or cached verifier states. A frozen Video-LLM then verifies, corrects, or regenerates the proposal using the selected evidence, without relying on the retrieved captions as grounding context. By separating evidence localization from answer verification, CIV uses captions to decide where to look and visual evidence to decide what to believe. This provides an efficient alternative to full visual replay while avoiding the de-grounding failure mode of caption-only reasoning.*

1. Introduction

Streaming video understanding requires a system to maintain memory over an evolving visual stream before the downstream information need is known [2, 12, 18]. This setting introduces a fundamental asymmetry between *when memory is written* and *when evidence is needed*. At streaming time, the system must decide what to store, summarize, or index without access to future queries. At query time, however, the user may ask about fine-grained visual facts whose relevance was not apparent when the corresponding segment was first observed. Such facts include object color, small background entities, exact counts, visible text, spatial relations, or the temporal order of actions. Consequently, a memory representation that is sufficient for describing the stream may still be insufficient for answering or verifying a future question. The core challenge is therefore not only long visual context, but unknown future evidence demand: the system must preserve a way to recover evidence whose importance is revealed only after the stream has been processed [3, 10].

This challenge distinguishes streaming video understanding from standard offline video question answering [2, 9, 19]. In offline Video-QA, the model receives a fixed video together with a known question, allowing frame sampling, retrieval, or visual attention to be conditioned directly on the query [5, 16, 19]. In streaming scenarios, by contrast, the visual stream is observed before the question is available [2, 3, 12]. The system must therefore construct memory under query uncertainty. This changes the role of compression: compressing a video segment into a compact representation is not merely a matter of preserving what is salient at the moment, but of retaining access to details that may become salient only under a future query. A streaming memory can fail even when its summaries are locally accurate, because future questions may target information that was visually present but not retained in the memory representation.

Caption-based memory offers an appealing interface for long and streaming videos [1, 7, 15]. Captions are compact, temporally organized, searchable with language queries, and compatible with existing language-model reasoning pipelines. These properties make captions a natural mechanism for indexing long visual histories. However, caption memory also introduces a subtle failure mode when the retrieved captions are treated as the sole basis for answering. A caption is a selective textual abstraction of a visual segment; it preserves some semantic content while omitting many visual details. If downstream reasoning is performed only over retrieved caption text, the system cannot distinguish between a fact that was absent from the video and a fact that was present visually but omitted by the captioner. This limitation is especially problematic in streaming settings, where the omitted detail may become relevant only after the caption has already been written.

We argue that the limitation of caption memory arises less from the use of captions themselves than from assign-

ing captions the wrong functional role. Captions should not be expected to serve as lossless containers for all future-relevant visual facts. Instead, they are better understood as *semantic indices* into the video stream [7]. Under this view, a caption need not contain the final answer. It only needs to make the relevant temporal region searchable, so that the system can later recover visual or cached evidence associated with that region. Captions can therefore support two query-time operations: they can identify candidate evidence locations, and they can provide semantic context for forming a provisional answer. The final decision, however, should be grounded in evidence that remains connected to the original video rather than in caption text alone.

This paper develops this separation between *proposal* and *verification* for streaming video understanding. We propose that caption memory should be used before final answering, not as final answer evidence. Given a query, retrieved captions can guide the system toward relevant temporal segments and produce an answer-level hypothesis. The system should then open a small set of selected visual or cached evidence associated with those segments and use a Video-LLM to verify the hypothesis. This design preserves the advantages of caption-based memory as a lightweight and language-searchable index, while reducing reliance on captions as the final grounding substrate. It also avoids processing the entire accumulated stream at query time, since verification is performed only over selected evidence.

We instantiate this principle in **Caption-Indexed Verification (CIV)**, a training-free proposal-verification framework for streaming video understanding. During streaming, CIV stores each temporal segment as a caption-memory item consisting of a caption, lightweight metadata, and a recoverable evidence handle. At query time, CIV retrieves relevant caption-memory items and uses their textual content to generate a provisional answer. The same retrieved items provide candidate handles to visual or cached evidence, from which CIV selects a budgeted subset for verification. A frozen Video-LLM then receives the query, the provisional answer, and the resolved evidence, and produces the final answer by accepting, revising, or regenerating the proposal. Crucially, the retrieved captions are not included in the final verification prompt. Thus, CIV uses captions to determine where evidence is likely to reside, while final answer generation is grounded in selected visual or cached evidence.

Our contributions are threefold:

1. We identify **unknown future evidence demand** as a central challenge in streaming video understanding: memory must be constructed before the system knows which visual details future queries will require.
2. We formulate the **caption-as-index** principle, in which captions serve as semantic access points for evidence localization and proposal generation rather than as final answer evidence.
3. We introduce **Caption-Indexed Verification (CIV)**, a training-free framework that retrieves caption-indexed memory, generates provisional answers, and verifies them using a frozen Video-LLM over selected visual or cached evidence.

2. Related Work

Streaming video understanding has recently emerged as a distinct paradigm that challenges the conventional offline formulation of Video-LLMs. In standard video-language benchmarks, models are typically evaluated on fixed videos or pre-selected keyframes with the user query provided upfront. Streaming systems, by contrast, must continuously ingest an ongoing visual stream and support asynchronous interactions regarding evidence that may have occurred far back in the past. This shift has catalyzed the development of frameworks that maintain evolving memories of past observations, schedule compute budgets online, or adapt offline Video-LLMs into online assistants for long-horizon interactions [2, 6, 11, 13, 14, 17]. These pioneering works establish that streaming video understanding is not merely long-video understanding scaled up; rather, it demands specialized memory mechanisms capable of preserving access to historical visual evidence under incremental observation and delayed querying.

To avoid the prohibitive computational cost of repeatedly processing the entire accumulated stream, a parallel line of research focuses on improving inference-time efficiency without modifying or retraining the underlying core Video-LLM. Rather than recalculating full visual tokens at every step, these approaches optimize the inference interface around a frozen foundation model via memory compression, recurrent state maintenance, or query-aware retrieval. For instance, key-value (KV) cache retrieval methods extract query-relevant past visual states from historical context, while recurrent streaming methods propagate compact visual descriptors across successive clips to eliminate redundant processing [3, 4, 13]. The foundational premise shared by these efficiency-driven methods is that only a small, query-specific subset of the past visual stream is necessary to address any given user interaction.

Language-based memory provides a highly intuitive and structured interface for executing this selective retrieval. Clip-level captions and textual event summaries are inherently compact, chronologically organized, and seamlessly searchable via text-based queries, making them highly attractive for long-horizon streaming architectures. Consequently, several recent frameworks adopt captions or text summaries as a lightweight memory footprint to localize past events and facilitate downstream question answering [4, 17]. Despite their efficiency, however, caption-based

memories suffer from a fundamental bottleneck: text descriptions are inherently lossy abstractions of complex visual evidence. They routinely discard fine-grained visual detail, such as precise object attributes, counts, localized text, or transient spatial-temporal relations, that may appear trivial during captioning but become critical for future verification. When retrieved captions are treated as the definitive context for final answering, the system is forced to reason over the artifacts of text generation rather than the raw visual reality.

3. Method

We introduce **Caption-Indexed Verification (CIV)**, a training-free framework for streaming video understanding. CIV represents an incoming video stream with a caption-indexed memory and answers user queries through a propose–select–verify procedure. During streaming, each temporal segment is converted into a compact memory item consisting of a natural-language caption, structured metadata, and a recoverable evidence handle. At query time, captions and metadata are used to retrieve relevant memory items and generate a provisional answer. The retrieved items also define candidate evidence locations, from which CIV selects a budgeted subset and resolves the corresponding handles into model-compatible visual or cached evidence. A frozen Video-LLM then produces the final answer by verifying the provisional answer against the resolved evidence.

The central separation in CIV is between *semantic indexing* and *final grounding*. Captions provide a language-searchable interface to the stream and support proposal generation, but they are not used as grounding context in the final verification prompt. Instead, final answer generation is conditioned on the query, the caption-derived proposal, and the selected visual or cached evidence.

Figure 1 summarizes the full CIV pipeline. We next formalize the streaming caption memory, caption-indexed retrieval, proposal generation, evidence selection, and final visual verification stages.

3.1. Streaming Caption Memory

Let a streaming video be observed as an incremental sequence of temporal segments:

$$V_{1:T} = \{v_1, v_2, \ldots, v_T\}, \tag{1}$$

where each segment v_i corresponds to a short temporal interval of frames. As each segment arrives, CIV constructs a memory item

$$m_i = (c_i, z_i, h_i). \tag{2}$$

Here, c_i is a natural-language caption describing the salient semantic content of segment v_i. The caption can be generated from the segment together with localized temporal context,

$$c_i = C(v_i; \rho_i), \tag{3}$$

where C denotes an off-the-shelf captioning module and ρ_i denotes optional local context, such as a short running summary or neighboring segment information. This context is used only to produce a coherent segment-level description and is not treated as final evidence for downstream question answering.

The term z_i denotes lightweight structured metadata associated with the segment. Examples include timestamps, detected entities, human actions, OCR cues, scene attributes, camera/source identifiers, or other backend-specific annotations. We write the textual representation of a memory item as

$$x_i = \mathrm{Serialize}(c_i, z_i), \tag{4}$$

where $\mathrm{Serialize}(\cdot)$ converts captions and metadata into a retrieval-compatible textual form.

The handle h_i specifies how evidence associated with segment v_i can later be recovered. The form of h_i is backend-dependent. It may correspond to a timestamp window, frame indices, source-video metadata, a pointer to cached visual features, a pointer to visual tokens, or a mapping to recoverable KV states. Thus, the caption memory separates the searchable representation of a segment from the evidence representation consumed by the verifier. After observing T segments, the accumulated memory is

$$\mathcal{M}_T = \{m_i\}_{i=1}^{T}. \tag{5}$$

In $\mathcal{M}_T$, captions serve as semantic indices for retrieval, while handles preserve access to the underlying visual or cached evidence.

3.2. Caption-Indexed Retrieval

Given a user query q, CIV retrieves relevant memory items from $\mathcal{M}_T$ using the textual representations $\{x_i\}_{i=1}^{T}$. Let q_r denote the retrieval query constructed from q. For open-ended dialogue, q_r may include the current query and relevant dialogue context. For multiple-choice tasks with candidate answers $\mathcal{A} = \{a_1, \ldots, a_K\}$, CIV can augment the retrieval query with the serialized answer set:

$$q_r = \mathrm{Serialize}(q, \mathcal{A}). \tag{6}$$

This encourages retrieval of segments that are discriminative with respect to the candidate answers.

Let $S_r(q_r, m_i)$ be the retrieval score between the retrieval query and memory item m_i:

$$S_r(q_r, m_i) = S_r(q_r, x_i). \tag{7}$$

The scoring function can be instantiated with dense retrieval, lexical matching, reranking, or a hybrid retrieval

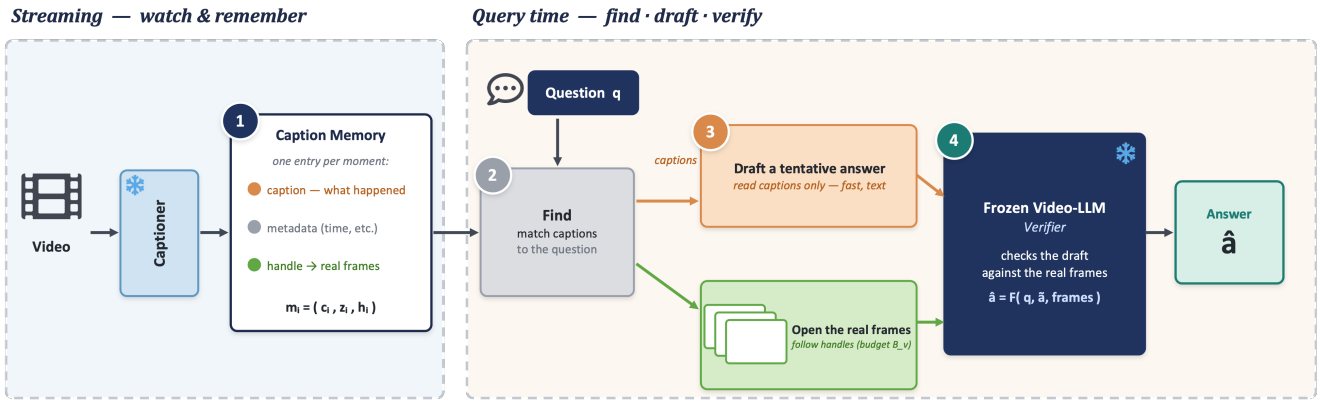


Figure 1. **Overview of Caption-Indexed Verification (CIV).** CIV stores each streaming segment as a memory item $m_i = (c_i, z_i, h_i)$, where c_i is a searchable caption, z_i contains lightweight metadata, and h_i points to recoverable visual or cached evidence. Given a query q, CIV retrieves relevant caption-memory items, generates a provisional answer $\tilde{a}$, selects a budgeted set of evidence handles, and resolves them into verifier-compatible evidence $E_{B_v}(q)$. The frozen Video-LLM produces the final answer from q, $\tilde{a}$, and $E_{B_v}(q)$. Retrieved captions guide proposal generation and evidence selection, but are excluded from the final verification prompt.

pipeline over captions and metadata. CIV retrieves a subset of memory items under a caption-context budget B_c:

$$\mathcal{R}(q) \in \arg\max_{\mathcal{S}\subseteq\mathcal{M}_T} \sum_{m_i\in\mathcal{S}} S_r(q_r, m_i) \quad \text{s.t.} \quad \sum_{m_i\in\mathcal{S}} \ell(x_i) \le B_c, \tag{8}$$

where $\ell(x_i)$ denotes the text length of the serialized caption-metadata representation. In practice, this budgeted retrieval objective can be approximated by ranking memory items according to S_r and adding items until the context budget is reached.

The retrieved set $\mathcal{R}(q)$ provides two intermediate objects. First, it defines the caption context used for proposal generation:

$$\mathcal{C}_R(q) = \mathrm{Order}\big(\{(i, c_i, z_i) \mid m_i \in \mathcal{R}(q)\}\big), \tag{9}$$

where $\mathrm{Order}(\cdot)$ denotes either temporal ordering or relevance ordering. Second, it defines the candidate evidence handles:

$$\mathcal{H}_R(q) = \{h_i \mid m_i \in \mathcal{R}(q)\}. \tag{10}$$

Thus, the same caption-indexed retrieval step supplies both the textual context for proposing an answer and the candidate visual locations for subsequent verification.

3.3. Proposal Generation

CIV generates a provisional answer from the query and the retrieved caption context:

$$\tilde{a} = G\big(q, \mathcal{C}_R(q)\big), \tag{11}$$

where G denotes a lightweight proposal module. The module G can be instantiated as an off-the-shelf language model, a prompting procedure, or the language component of the target model operating only on text. No additional visual evidence is opened in this stage.

The proposal input is constructed by serializing the retrieved caption context together with the query:

$$P_G(q) = \mathrm{Prompt}_G\big(q, \mathcal{C}_R(q)\big), \tag{12}$$

and the provisional answer is generated as

$$\tilde{a} = G(P_G(q)). \tag{13}$$

For multiple-choice tasks, G may score each candidate answer and produce

$$\tilde{a} = \arg\max_{a_k\in\mathcal{A}} s_G(a_k \mid q, \mathcal{C}_R(q)), \tag{14}$$

where s_G denotes the proposal score assigned to candidate a_k. For open-ended tasks, G produces a free-form provisional response.

The proposal $\tilde{a}$ is an answer-level hypothesis derived from caption-indexed context. It summarizes the retrieved semantic information into a compact candidate response, but it is not treated as final output. Final grounding is deferred to the verifier, which evaluates the proposal using resolved visual or cached evidence.

3.4. Evidence Selection and Resolution

The retrieved memory items may correspond to more evidence than the Video-LLM verifier can process. CIV therefore selects a budgeted subset of candidate evidence handles from $\mathcal{H}_R(q)$. Let $\kappa(h_i)$ denote the verifier-side cost of resolving handle h_i, measured in backend-specific units such as temporal windows, sampled frames, visual tokens, feature blocks, or KV segments. CIV selects a subset of

retrieved memory items by optimizing an evidence utility under budget B_v:

$$\mathcal{S}_{B_v}(q) \in \arg\max_{\mathcal{S}\subseteq\mathcal{R}(q)} U(q,\tilde{a},\mathcal{S}) \quad \text{s.t.} \quad \sum_{m_i\in\mathcal{S}} \kappa(h_i) \le B_v. \tag{15}$$

The corresponding selected evidence handles are

$$\mathcal{H}_{B_v}(q) = \{h_i \mid m_i \in \mathcal{S}_{B_v}(q)\}. \tag{16}$$

The utility function $U(q,\tilde{a},\mathcal{S})$ specifies which retrieved segments should be opened for verification. It can incorporate retrieval relevance, answer-candidate relevance, metadata overlap, temporal coverage, and diversity among selected segments. A generic form is

$$U(q,\tilde{a},\mathcal{S}) = \sum_{m_i\in\mathcal{S}} u(q,\tilde{a},m_i) + \lambda\Delta(\mathcal{S}), \tag{17}$$

where $u(q,\tilde{a},m_i)$ is an item-level relevance score and $\Delta(\mathcal{S})$ encourages coverage or diversity over time. This formulation is intentionally backend-agnostic: implementations may approximate the selection objective with top-ranked retrieval results, heuristic reranking, temporal non-maximum suppression, prompt-based scoring, or learned rerankers.

The selected handles are then resolved into verifier-compatible evidence:

$$E_{B_v}(q) = \mathrm{Resolve}\big(\mathcal{H}_{B_v}(q)\big). \tag{18}$$

The resolution function depends on the target Video-LLM backend. For frame-based models, it may decode the selected temporal windows and sample frames. For feature-based models, it may load cached visual features or visual tokens. For cache-based models, it may recover the corresponding KV states. The output $E_{B_v}(q)$ is the evidence representation that will be supplied to the verifier.

3.5. Visual Verification

The final answer is produced by a frozen Video-LLM verifier F. The verifier receives the query, the provisional answer, and the resolved evidence:

$$\hat{a} = F\big(q,\tilde{a},E_{B_v}(q)\big). \tag{19}$$

Equivalently, the verifier prompt is constructed as

$$P_F(q) = \mathrm{Prompt}_F\big(q,\tilde{a},E_{B_v}(q)\big). \tag{20}$$

CIV imposes the following verifier-input constraint:

$$\mathcal{C}_R(q) \notin P_F(q). \tag{21}$$

That is, the retrieved caption context is not provided to the verifier as textual grounding evidence. Captions influence the final stage only indirectly, through the proposal $\tilde{a}$ and through the evidence handles selected for resolution.

The verifier evaluates whether the proposal is supported by the selected evidence. If the evidence supports the proposal, the verifier may preserve it. If the evidence contradicts the proposal or supports only part of it, the verifier may revise the answer. If the proposal is unsupported, the verifier may generate a new answer from the resolved evidence. For multiple-choice tasks, the verifier outputs a final candidate answer:

$$\hat{a} \in \mathcal{A}. \tag{22}$$

For open-ended tasks, the verifier generates a natural-language response.

Through this verification step, CIV changes the role of caption memory from final answer storage to evidence localization. Caption memory identifies where relevant evidence is likely to reside, while the frozen Video-LLM determines the final answer from the selected visual or cached evidence.

4. Experiments

4.1. Experimental Setup

Benchmarks and evaluation protocol. We evaluate Caption-Indexed Verification (CIV) under both offline and streaming video understanding protocols, so that the method is tested both as a long-video reasoning system and as an online memory mechanism. For the offline setting we adopt two complementary benchmarks. MLVU [19] targets long-horizon understanding over extended visual contexts and therefore directly probes whether caption-indexed memory can surface the task-relevant evidence that is scattered across a long video. VideoMME [5] contributes a broader evaluation that spans a wide range of video domains, durations, and question types, and thus measures whether the same mechanism generalizes beyond a single long-video regime rather than overfitting to one benchmark. For the streaming setting we use the Backward Tracing subset of OVO-Bench [9], which evaluates online understanding under a strict temporal constraint: a query issued at a given playback time may draw only on information observed up to that moment, and no future frames are accessible. Because CIV is fundamentally a memory-based method whose central aim is to recover evidence from previously observed segments, Backward Tracing is the most faithful probe of our design, since it isolates retrospective evidence retrieval from the already-observed history. We therefore report Backward average accuracy and omit the overall and real-time/forward scores, as those regimes emphasize forward prediction and instantaneous perception, which lie outside the scope of memory-based retrospective verification. On OVO-Bench, CIV operates on a 1 fps stream and incrementally constructs its caption-indexed memory online from the observed frames. At query time it retrieves the relevant caption-memory items, selects

candidate evidence handles, and resolves the corresponding stream segments for verification, all while respecting the online constraint that only pre-query observations are available.

Backbones and inference configurations. All experiments use LLaVA-OneVision [8] as a *frozen* Video-LLM backbone, and we apply no task-specific fine-tuning at any stage; CIV is realized entirely through its memory, retrieval, and verification procedure on top of this fixed model. Unless otherwise noted, we evaluate LLaVA-OneVision-7B across MLVU, VideoMME, and OVO-Bench Backward Tracing. Our main results report the full CIV pipeline, which we denote **Dual**. Dual first retrieves caption-memory items to form an initial answer proposal and to select a small set of candidate verifier states; the selected handles are then resolved into visual or cached evidence, and the frozen Video-LLM verifies, corrects, or regenerates the proposal against this resolved evidence. In this design the retrieved captions act purely as an index that decides where to look and what to propose, and they are never passed to the verifier as final evidence. For the ablation study we additionally consider a **Caption** configuration that shares exactly the same caption-indexed memory and retrieval pipeline but answers directly from the retrieved caption context, without resolving any visual or cached evidence. Comparing Dual against Caption thus isolates the contribution of the visual verification stage from that of semantic retrieval.

4.2. Quantitative Results

Streaming video understanding. Table 1 reports Backward Tracing results on OVO-Bench alongside previously published systems. With a frozen LLaVA-OneVision-7B backbone and a 1 fps streaming input, CIV reaches 57.53% Backward average accuracy, corresponding to a $+9.35$-point improvement over the published LLaVA-OneVision-7B reference of 48.18%. This is a substantial gain given that CIV introduces no additional training and changes only how evidence is stored, retrieved, and verified at inference time. We emphasize that the comparison is not a controlled same-backbone evaluation, since the listed systems differ in both their backbones and their frame policies; the table is therefore intended to situate CIV relative to existing methods rather than to assert a strictly matched comparison. Within this context, CIV remains competitive with strong proprietary and open systems and, in particular, exceeds the memory-based OASIS baseline (57.21%) while relying only on a frozen backbone. Combined with the offline results, this demonstrates that the same caption-indexed verification mechanism transfers from offline long-video reasoning to the online streaming regime, where evidence must be recovered from a causally restricted history.

Table 1. Comparison on OVO-Bench Backward Tracing. We report Backward average accuracy. Published systems are reported under their original settings. Our CIV result uses a frozen LLaVA-OneVision-7B backbone with a 1 fps streaming input. The value in parentheses denotes the improvement over the published LLaVA-OneVision-7B reference row.

System	Frames / Policy	Backward Avg.
Gemini 1.5 Pro	1 fps	62.54
GPT-4o	64	60.75
Qwen2-VL-72B	64	56.95
LLaVA-Video-7B	64	53.74
Qwen2-VL-7B	64	43.73
Flash-VStream-7B	1 fps	27.38
VideoLLM-online-8B	2 fps	16.24
LLaVA-OneVision-7B	64	48.18
OASIS	0.5 fps	57.21
CIV	1 fps	**57.53 (+9.35)**

Offline video understanding. Table 2 compares CIV with a set of recent offline-to-online memory systems on MLVU and VideoMME, all built on the LLaVA-OneVision-7B backbone and evaluated with the full Dual pipeline. Despite operating on a frozen backbone without any task-specific fine-tuning, CIV attains 68.0% on MLVU and 61.5% on VideoMME. On MLVU this places CIV essentially on par with the strongest reported result (68.5%) while clearly surpassing the remaining memory systems, and on VideoMME CIV achieves the best accuracy among all compared methods, improving over the next best system by a clear margin. We note that the competing systems differ substantially in their frame sampling policies and memory budgets, so the comparison is not perfectly controlled; nonetheless, CIV remains competitive or better on both benchmarks while keeping its backbone entirely fixed. Taken together, these results indicate that caption-indexed verification is effective not only for long-horizon reasoning, as measured by MLVU, but also across the more heterogeneous videos and question types in VideoMME, which supports the generality of the approach rather than a benchmark specific effect.

4.3. Qualitative Results

Figures 2a and 2b present two representative streaming cases that illustrate why retrieval alone is insufficient, and show that the same failure mode recurs across different kinds of fine-grained facts.

In the first case (Figure 2a), the online caption correctly localizes the relevant action but describes it only with the generic phrase "filling it." Asked what was placed in the pan, the caption-only baseline cannot recover the missing detail from the textual memory and selects a plausible but incorrect option (oil). CIV instead treats the caption as an

Table 2. Comparison on offline video understanding (MLVU and VideoMME) in the offline-to-online setting with a LLaVA-OneVision-7B backbone. Competing systems are listed under their originally reported settings; "–" denotes a result not reported. Best result per column is in **bold**, second best is underlined.

System	FPS	Frames (budget)	Memory	MLVU	VideoMME
ReKV	0.5	64 (12k)	Local 15,000	**68.5**	–
Infinipot-V	768	KV (6k)	6,000	–	–
StreamKV	0.5/0.2	KV (6k)	6,000	66.9	59.4
StreamingTOM	0.5/0.2	240 (12k)	Local 4,800 (exc. 19,200)	67.9	59.9
MuKV	0.5	64 (6k)	–	67.8	61.2
HERMES	1	KV (6k)	6,000	–	58.4
CIV (Ours)	1	KV (6k)	6,000	68.0	**61.5**

index: it opens the corresponding frame at 4:35 and reads directly from the resolved visual evidence that the pan is being filled with water. The second case (Figure 2b) exhibits the same pattern for object identity. The caption refers to the dropped item only as "sliced vegetables," so the caption-only baseline guesses an incorrect vegetable (onion), whereas CIV opens the frame at 3:55 and grounds its answer in the visual evidence, which reveals that the object is in fact eggplant.

Both examples crystallize the central motivation behind CIV. Captions reliably indicate *where* the relevant evidence resides, whereas pinning down *what* the answer actually is typically demands direct visual grounding. The two cases further show that the lost detail is not tied to a single attribute type: it may be the substance being poured or the identity of a manipulated object. By decoupling semantic retrieval from evidence verification, CIV recovers precisely the details that caption memory omits or underspecifies, turning otherwise unanswerable captions into correct, visually grounded predictions.

4.4. Ablation Study

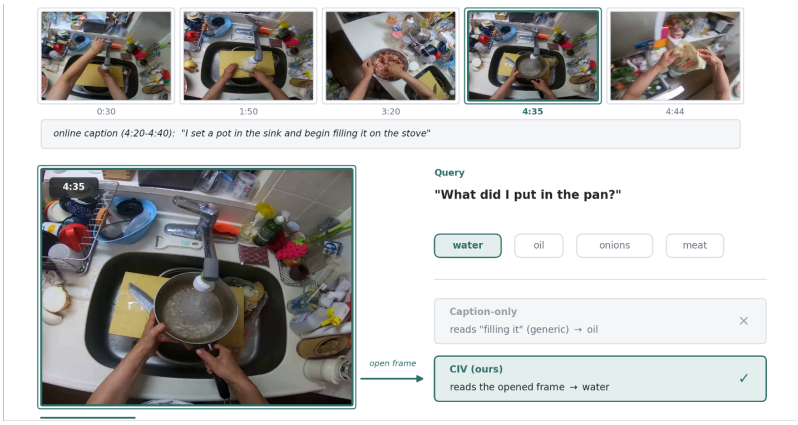
We isolate the contribution of the visual verification stage by comparing the Caption and Dual configurations under an identical caption-indexed memory and retrieval pipeline; the two differ only in whether the retrieved captions are resolved into visual or cached evidence before the final answer is produced. As reported in Table 3, on OVO-Bench Backward Tracing the Dual pipeline improves Backward average accuracy from 50.48% to 57.53% with the frozen LLaVA-OneVision-7B backbone, an absolute gain of $+7.05$ points. This gap is informative. The Caption configuration already benefits from the same semantic localization and proposal information, yet answering directly from the retrieved caption text is fundamentally limited by the lossy nature of textual summaries, which tend to preserve coarse scene semantics while discarding the fine-grained visual attributes that many questions ultimately depend on. By resolving the selected evidence handles and re-examining them with the frozen Video-LLM, Dual recovers exactly these omitted details and corrects proposals that the caption text alone cannot support. The consistent improvement therefore corroborates the central design principle of CIV, namely the separation between semantic retrieval, which decides where to look, and visual verification, which decides what the final answer is. This behavior is also consistent with the qualitative analysis in Figure **??**, where caption-only answering fails on a fine-grained object identity that is resolved only once the indexed frame is opened and verified.

Table 3. Verification ablation on OVO-Bench Backward Tracing. **Caption** answers only from retrieved caption memory, whereas **Dual** verifies caption-derived proposals with selected visual or cached evidence. We report Backward average accuracy, and Δ denotes the absolute improvement of Dual over Caption in percentage points.

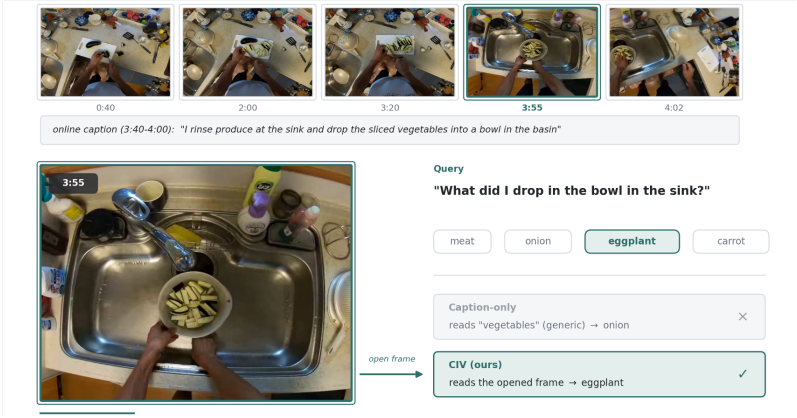
Backbone	Caption	Dual	Δ
OV-7B	50.48	57.53	+7.05

5. Conclusion

We presented Caption-Indexed Verification (CIV), a training-free proposal-verification framework for streaming video understanding that reassigns the functional role of caption memory. Instead of treating retrieved captions as the grounding context for final answering, CIV uses them as semantic indices that localize candidate evidence within the stream and seed an initial answer proposal, deferring the final decision to a frozen Video-LLM that resolves a small, budgeted set of visual or cached evidence and verifies, revises, or regenerates the proposal against it. By excluding the retrieved captions from the verification prompt, CIV retains the efficiency of a lightweight, language-searchable memory while avoiding the de-grounding failure mode that arises when lossy textual summaries are used as the sole basis for answering. This separation directly addresses the



(a) Substance identity (water vs. oil).



(b) Object identity (eggplant).

Figure 2. Qualitative examples of Caption-Indexed Verification. In both cases the online caption localizes the relevant segment but is too generic to answer the query, so caption-only answering fails, whereas CIV opens the indexed frame and grounds the final answer in the resolved visual evidence. (a) The caption describes the action only as "filling it," so caption-only answering selects an incorrect option (oil), while CIV opens the frame at 4:35 and reads the answer as water. (b) The caption describes the dropped item only as generic "vegetables," so caption-only answering predicts an incorrect object (onion), while CIV opens the frame at 3:55 and identifies it as eggplant. Together they illustrate the guiding principle that captions say *where* to look, while the resolved frame says *what* the answer is.

defining difficulty of streaming memory, namely that a system must commit to what it stores before it knows which visual facts a future query will require.

Our experiments support this design across both offline and streaming protocols. With a frozen LLaVA-OneVision-7B backbone, CIV attains 68.0% on MLVU and 61.5% on VideoMME, and reaches 57.53% Backward average accuracy on OVO-Bench Backward Tracing, a $+9.35$ point improvement over the corresponding reference. A verification ablation isolates the role of visual grounding, showing that resolving and re-examining selected evidence raises Backward average accuracy from 50.48% to 57.53% relative to answering directly from caption text. These findings are consistent with the central premise of CIV, that captions are well suited to deciding where to look, whereas visual evidence remains necessary for deciding what to believe.

We regard the present study as an initial validation of the caption-as-index principle rather than a complete characterization of the framework. The evaluation relies on a single backbone family and single-run estimates, and it compares CIV against a caption-only baseline and against published systems under their original settings rather than against re-implemented baselines under a strictly matched protocol; moreover, although CIV is motivated as an efficient alternative to full visual replay, its query-time cost has not yet been quantified. Future work will therefore evaluate CIV against streaming Video-LLMs, caption-memory methods, visual-token selection, KV-cache retrieval, and full visual replay on a common frozen backbone, extend the analysis to additional backbones, benchmarks, and repeated runs, report the evidence budget together with the resulting accuracy-versus-cost trade-off, and refine the retrieval, proposal, selection, and verification components while attributing residual errors to caption coverage, retrieval, evidence selection, or the frozen verifier.

References

[1] 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. *arXiv preprint arXiv:2407.12679*, 2024. 1

[2] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[3] 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 (ICLR)*, 2025. 1, 2

[4] Vaggelis Dorovatas, Soroush Seifi, Gunshi Gupta, and Rahaf Aljundi. Recurrent attention-based token selection for efficient streaming video-llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2

[5] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. 1, 5

[6] Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[7] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. In *ACL Findings*, 2025. 1, 2

[8] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[9] Yifei Li, Junbo Niu, 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 (CVPR)*, 2025. 1, 5

[10] Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. Oasis: On-demand hierarchical event memory for streaming video reasoning. *arXiv preprint arXiv:2604.17052*, 2026. 1

[11] Jihao Liu, Zhiding Yu, Shiyi Lan, Shihao Wang, Rongyao Fang, Jan Kautz, Hongsheng Li, and Jose M. Alvarez. Streamchat: Chatting with streaming video. *arXiv preprint arXiv:2412.08646*, 2024. 2

[12] Zhenyu Ning, Guangda Liu, Qihao Jin, Chengwei Li, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025. 1

[13] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[14] Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2

[15] Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*, 2024. 1

[16] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. 1

[17] Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[18] Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding, 2025. 1

[19] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 5