VICOS: VERIFY-THEN-CORRECT RETRIEVAL AND CONTEXT MOSAIC FOR AGGRESSIVE VIDEO TOKEN COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form video question answering requires vision-language models (VLMs) to process thousands of visual tokens, making inference expensive due to the quadratic cost of self-attention. Training-free visual token compression reduces this cost, but performance can degrade substantially under aggressive budgets. We identify two related problems: **mis-selection**, where the limited visual budget is spent on answer-irrelevant frames, and **discarded-context**, where unselected frames retain substantial query-aligned visual signal but are removed entirely. We propose **ViCoS** (*Verify-then-Correct with Context Mosaic*), a training-free framework that addresses both problems while keeping the final answer-stage visual-token budget fixed. **VerCoRe** self-verifies the current retrieved frames and, when the evidence is insufficient, generates corrective visual concepts to refine CLIP retrieval. **CoMo** preserves discarded temporal context by packing the remaining candidate frames into a single low-resolution mosaic. Under a budget of K images, ViCoS uses $K-1$ full-resolution focus frames and one mosaic; at $K=1$, it reduces to CoMo alone. Across two backbones and five video QA benchmarks, ViCoS achieves the highest compressed accuracy among the compared methods in all 30 evaluated backbone–benchmark–budget settings. On VideoVista with InternVL3-8B, ViCoS achieves 78.80 accuracy at a 12.5% visual-token budget, only 0.04 points below the full $K=8$ reference while using a single encoder image.

1 INTRODUCTION

Recently, long-video understanding has become increasingly important in a wide range of applications including surveillance, autonomous driving, embodied agents, sports and media analysis, and video question answering. In performing the task, video vision-language models (VVLMs) that jointly process visual observations and natural-language queries have been popularly used (Lin et al., 2024; Zhang et al., 2023a; Maaz et al., 2024; Li et al., 2024a; Zhang et al., 2024b; Bai et al., 2025).

One well-known drawback of VVLMs is the high computational and memory cost to process large amounts of visual information in a video. In fact, to preserve fine-grained visual details in an image, modern VVLMs should use thousands of tokens. In Qwen2.5-VL (Bai et al., 2025), for example, a single image can use up to 16,384 visual tokens. Therefore, in video understanding, the overall computational cost increases substantially as the number of processed frames grows. This issue becomes particularly important in practical long-form videos, such as tutorials, vlogs, sports broadcasts, and surveillance footage. In such videos, query-relevant events may appear only briefly within the entire video. To avoid missing important evidence, VVLMs must process a large number of frames spanning the video. In covering such a broad temporal range, temporally adjacent frames are often processed together, and similar scenes or background information may repeatedly appear across these frames. A substantial portion of the video may also consist of scenes that are not directly relevant to the query. Consequently, VVLMs spend considerable computational resources processing query-irrelevant scenes and redundant visual information. For the reasons mentioned, visual token compression techniques that do not include unnecessary or redundant visual tokens while preserving the visual evidence to answer the query are of great importance for long-video understanding.

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.

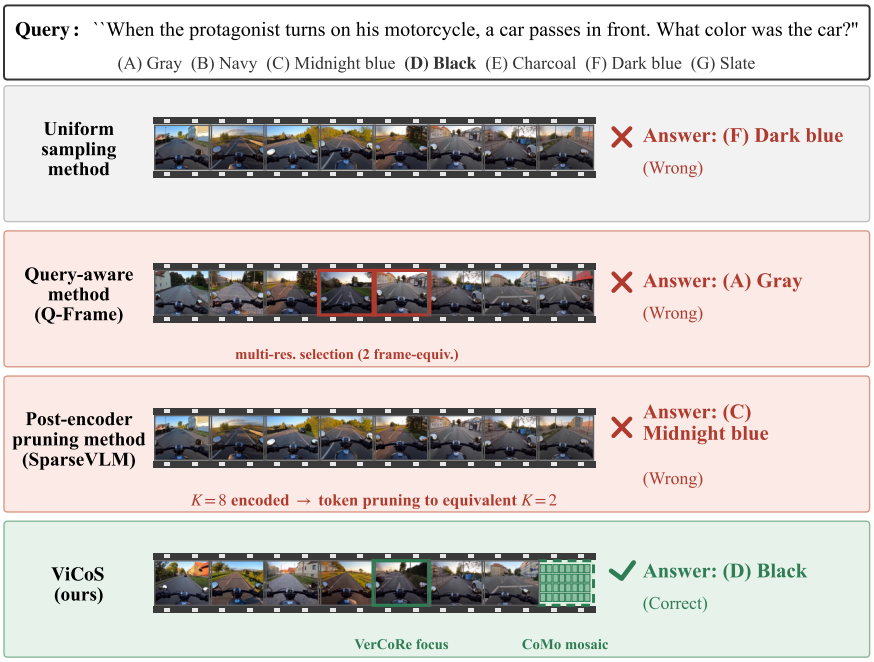


Figure 1: Representative comparison at a $K{=}2$ frame-equivalent budget on CG-Bench. Q-Frame performs query-aware selection with multi-resolution allocation whose frame-equivalent token budget equals 2 (see Appendix H); SparseVLM encodes $K{=}8$ uniformly sampled frames and prunes visual tokens down to the equivalent K-image budget; ViCoS uses VerCoRe to pick the relevant focus frame and CoMo to preserve the remaining temporal context under the same target frame-equivalent budget. Filmstrips show a common temporal reference sampled from the $F{=}32$ candidate pool; Q-Frame performs its multi-resolution selection over the full pool.

Over the years, various approaches to reduce the computational cost of VVLMs have been proposed. In (Li et al., 2024a; Bai et al., 2025), uniform sampling techniques have been proposed to limit the number of images used as inputs to the vision encoder. Another approach to reducing computational cost is to select only the frames that are relevant to the input query. For this purpose, query-aware frame-selection methods such as Q-Frame (Zhang et al., 2025a) have been proposed. In this scheme, candidate frames are ranked according to their relevance to the input query, and only a small number of informative frames are selected. Also, to reduce redundant visual tokens, post-encoder pruning techniques such as SparseVLM (Zhang et al., 2025b), FastV (Chen et al., 2024b), and VisionZip (Yang et al., 2024) have been proposed. It has been shown that these techniques perform well under moderate visual-token budgets, but their performance degrades considerably at aggressive compression scenarios due to the loss of valuable visual evidence.

The key observation we have made through extensive experiments is that important visual evidence may be lost either under overly aggressive compression or when the limited visual budget is allocated to frames containing unnecessary information. Essentially, the loss of answer-relevant visual evidence stems from two distinct but related problems: *mis-selection* and *discarded-context*. The *mis-selection* refers to selecting frames that contain unnecessary information. Most frame-selection methods compare each candidate frame with the input query and select frames with high visual–text similarity. However, some words in the query may describe surrounding context or actions rather than the answer itself. These words can give high similarity scores to irrelevant frames. As a result, frames without the evidence needed to answer the query may be selected instead of more useful frames. This is problematic under a limited visual budget because such incorrect selections waste the few available frame slots on irrelevant information. For example, when a query asks what object

a person picked up after opening a door, frames showing the door or the opening action may be ranked higher than the frame that actually shows the object.

The *discarded-context* refers to the problem of losing potentially useful information when all unselected frames are discarded. Even when highly relevant frames are selected, less relevant frames may still contain auxiliary information needed to determine the correct answer. Since conventional frame-selection methods discard these frames entirely, the information contained in them cannot be used by the model. This is problematic under aggressive compression because potentially useful temporal or contextual evidence is lost together with the discarded frames. For example, in questions about the order or duration of an action, not only the selected key frames but also the surrounding frames before and after them may provide important temporal context.

Motivated by these observations, we propose a training-free framework for addressing both problems, called Verify-then-Correct with Context Mosaic (**ViCoS**), which consists of two key evidence-preserving ingredients: Verify-then-Correct Retrieval (**VerCoRe**) and Context Mosaic (**CoMo**). VerCoRe tackles the mis-selection by enabling the target video-LLM to verify the current retrieval result. When the retrieved frames do not provide sufficient evidence, the model generates corrective visual concepts, which are used to update the CLIP retrieval scores and re-rank the candidate frames. In this way, retrieval is refined using both the query and the visual evidence observed from the video. This enables VerCoRe to correct misleading query-based retrieval and allocate the limited frame budget to more answer-relevant visual evidence. As shown by the green-highlighted frame in Figure 1, VerCoRe can redirect retrieval from an irrelevant candidate to the frame that actually contains the passing car queried by the user.

CoMo addresses the discarded-context by preserving information from unselected candidate frames. Rather than removing these frames entirely, CoMo arranges them in temporal order within a single low-resolution mosaic. Under a budget of K encoder images, ViCoS allocates $K-1$ slots to full-resolution focus frames and one slot to CoMo, preserving coarse temporal context without increasing the final image budget. This allows the model to retain useful peripheral evidence that would otherwise be discarded, complementing the selected focus frames under the same visual budget. For example, in Figure 1, CoMo preserves the temporal context surrounding the motorcycle start and the passing car, providing the broader event sequence needed to interpret the selected focus frame correctly.

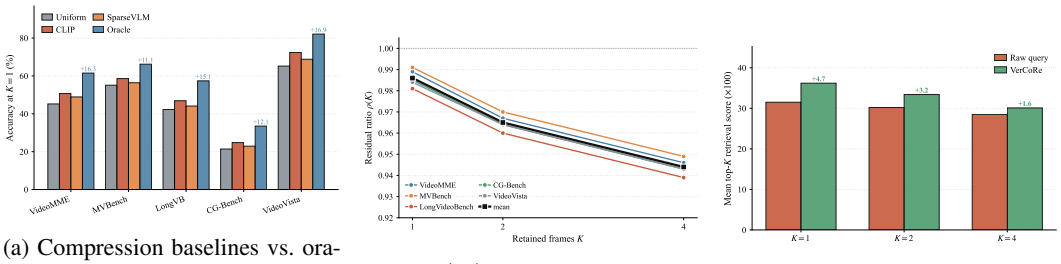
We evaluate ViCoS on Qwen2.5-VL-7B (Bai et al., 2025) and InternVL3-8B (Zhu et al., 2025) across five video QA benchmarks: VideoMME (Fu et al., 2024), MVBench (Li et al., 2024b), LongVideoBench (Wu et al., 2024), CG-Bench (Chen et al., 2024a), and VideoVista (Li et al., 2024d). Across 50%, 25%, and 12.5% visual-token budgets, ViCoS consistently improves over representative frame-selection and post-encoder pruning baselines. On VideoVista with InternVL3-8B, ViCoS achieves **78.80** accuracy at the 12.5% budget, only **0.04** points below the full $K=8$ reference of 78.84.

Our contributions are summarized as follows:

- We identify two key problems in aggressive training-free video compression: *mis-selection*, which causes the limited visual budget to be spent on frames lacking answer-relevant evidence, and *discarded-context*, which removes potentially useful information contained in unselected frames.

- To address these two problems, we propose **ViCoS**, a training-free framework consisting of *VerCoRe* and *CoMo*. *VerCoRe* refines retrieval through self-verification and corrective visual concepts, enabling more effective allocation of high-resolution frame slots. *CoMo*, in turn, preserves otherwise discarded temporal context within a single low-resolution mosaic. Together, they enable more effective use of a highly constrained visual budget without increasing the image budget at the final answer stage.

- We evaluate ViCoS on two video-LLM backbones across five video QA benchmarks and three aggressive visual-token budgets. ViCoS consistently outperforms representative frame-selection and post-encoder pruning baselines, including at the highly compressed 12.5% setting.



(a) Compression baselines vs. oracle at $K{=}1$.

(b) $\rho(K)$ decreases with K.

(c) Retrieval-score shift.

Figure 2: Motivation for ViCoS. **(a)** Practical compressors remain below an oracle at $K{=}1$, indicating substantial selection headroom. **(b)** The residual relevance ratio $\rho(K)$ decreases with K, showing that discarded frames retain more query-aligned signal at aggressive budgets. **(c)** VerCoRe increases the top-K retrieval score relative to raw-query scoring. The $K{=}1$ result is diagnostic only.

2 Analysis and Motivation

Before introducing ViCoS, we examine why existing training-free video compression methods become unreliable under aggressive visual-token budgets. Our analysis focuses on two questions: (i) whether the limited visual budget is often allocated to suboptimal evidence, and (ii) whether the frames excluded from that budget still contain query-aligned visual information. Together, these analyses motivate the two design principles behind ViCoS: refining the retrieval scoring before allocating high-resolution frame slots, and preserving coarse context from frames that cannot be retained at full resolution. Throughout this section, we use InternVL3-8B (Zhu et al., 2025) as the target video-LLM and report diagnostic results on the five main benchmarks (VideoMME, MVBench, LongVideoBench, CG-Bench, VideoVista), using VideoMME (Fu et al., 2024) as the primary probe. Notation used in this section (candidate pool $\{f_1, \ldots, f_F\}$, selected index set $\mathcal{S}_K$, similarity vector $\mathbf{s}(q)$, and the raw-query CLIP score $s_i(q)$) is introduced in Appendix C.

2.1 Frame Selection Is a Major Bottleneck at Aggressive Budgets

We first measure the headroom available from improved frame selection. For each query, we compare four settings: (i) uniform sampling, (ii) raw-query CLIP top-K retrieval, (iii) SparseVLM attention-based token pruning over $K{=}8$ uniformly sampled frames, and (iv) an oracle selector that uses ground-truth answer information to choose the single most favorable frame among the $F{=}32$ candidates. For each candidate frame, we run the video-LLM with that frame as the sole visual input and select the frame that maximizes the log-probability of the ground-truth answer option. Figure 2a shows that, at aggressive budgets, all three practical baselines remain substantially below the oracle. The oracle is not intended as a realizable method; rather, it estimates the performance headroom available when more informative evidence is selected. These results indicate that frame selection remains a major bottleneck under aggressive compression: post-encoder pruning cannot compensate when the relevant visual content is absent from the encoded frame set.

2.2 Mis-Selection: Raw-Query Retrieval Can Misrank Relevant Frames

To analyze this problem, we consider a *raw-query CLIP retrieval baseline* that directly embeds the full question as the retrieval anchor, $a_{\text{raw}} = \phi_t(q)$. Natural-language questions often contain linguistic content that is only weakly aligned with the visual concepts required for retrieval, so the raw-query embedding may rank visually salient but answer-irrelevant frames above the evidence needed to answer the question. We compare raw-query scoring against VerCoRe's final running score using the same underlying patch-level MaxSim scorer. As summarized in Figure 2c, VerCoRe raises the mean top-K MaxSim score under the refined retrieval anchors by 4.7, 3.2, and 1.6 points at $K{=}1, 2, 4$, respectively (see Appendix D for the numeric table). This score is used as a diagnostic of visual-text alignment; downstream utility is evaluated separately through answer accuracy in Table 4. The largest diagnostic gain arises at $K{=}1$, where top-1 retrieval is maximally sensitive to ranking errors. We use this setting only to probe retrieval quality; in the final ViCoS pipeline, $K{=}1$ is handled by CoMo alone.

Observation F1. *Under aggressive budgets, the retrieval scoring directly determines which visual evidence receives the limited high-resolution budget. Replacing the raw query with visually focused content words improves the retrieval score, and verification-conditioned correction provides an additional gain, especially at small K.*

2.3 Discarded-Context: Unselected Frames Retain Query-Aligned Signal

Improving the selector alone does not address a second source of information loss: even when the selected frames are highly query-relevant, the remaining candidates can still retain substantial query-aligned visual signal. To quantify this, we rank the $F{=}32$ candidate frames according to the raw-query CLIP similarity $\mathbf{s}(q)$, letting $\mathcal{S}_K$ denote the selected top-K set and $\bar{\mathcal{S}}_K$ the discarded set, and define the residual signal ratio

$$\rho(K) = \frac{\text{mean}\big(\text{top-}K\big(\mathbf{s}(q)|_{\bar{\mathcal{S}}_K}\big)\big)}{\text{mean}\big(\mathbf{s}(q)|_{\mathcal{S}_K}\big)}. \tag{1}$$

A value close to 1 indicates that the highest-ranked frames outside the selected set remain nearly as query-aligned as the retained frames. Figure 2b shows that $\rho(K)$ rises from 94.4% at $K{=}4$ to 98.6% at $K{=}1$ (numeric table in Appendix E), so highly query-aligned frames increasingly remain outside the selected set as the budget tightens. $\rho(K)$ measures query alignment rather than complementary answer information, so its purpose is diagnostic rather than causal: it does not by itself prove that discarded frames answer the question, but establishes that meaningful query-aligned signal exists outside the selected set. Empirical support for preserving discarded context comes from the $K{=}1$ results in Section 4.1, where ViCoS reduces to CoMo alone and matches or outperforms Q-Frame on all five benchmarks on both backbones.

Observation F2. *At aggressive budgets, the highest-ranked discarded candidates retain substantial query-aligned visual signal, and this residual signal is largest when only a small number of frames are preserved. This motivates retaining a low-cost representation of the discarded temporal context rather than eliminating it entirely.*

3 Method

The two problems affect different parts of the compression process: mis-selection determines which frames receive full-resolution capacity, whereas discarded-context determines what is lost from the remaining frames. ViCoS addresses them with two complementary mechanisms: **VerCoRe** refines the CLIP retrieval scoring to counter mis-selection (F1, Sec. 2.2), and **CoMo** packs unselected candidate frames into a single mosaic image to mitigate discarded-context (F2, Sec. 2.3). Crucially, the mosaic occupies the last of the K encoder slots, so the final answer stage receives exactly K encoder images, matched with Uniform-K; Q-Frame follows the same K frame-equivalent budget through its multi-resolution fixed-token allocation, and SparseVLM/FastV/VisionZip realize the budget through post-encoder token pruning. Figure 3 shows the full pipeline.

3.1 Overview

Given a video V, a query q, and a target encoder budget of K images, ViCoS uniformly samples $F{=}32$ candidate frames, and (for $K \geq 2$) applies VerCoRe (Sec. 3.2) to iteratively refine per-candidate retrieval scores using visual concept sets, from which it selects the top-$(K{-}1)$ focus frames $\mathcal{S} \subseteq \{1, \ldots, F\}$; the remaining $F{-}|\mathcal{S}|$ candidates are packed into a single grid image g by CoMo (Sec. 3.3). The final ordered image sequence $\mathcal{X}{=}(f_{i_1}, \ldots, f_{i_{K-1}}, g)$ places focus frames in temporal order followed by the mosaic in the last slot, so $|\mathcal{X}|{=}K$. At $K{=}1$, VerCoRe is skipped and the mosaic packs all $F{=}32$ candidates.

3.2 VerCoRe: Verify-then-Correct Retrieval

Motivation. Raw-query CLIP retrieval can be weakened by linguistic content that is only weakly aligned with the visual concepts needed for retrieval (F1, Sec. 2.2). VerCoRe initializes retrieval with a compact visual concept set $\mathcal{C}^{(0)}$ and iteratively refines the running retrieval scores using verification-conditioned corrective concept sets. Unlike one-shot query rewriting, each refinement is

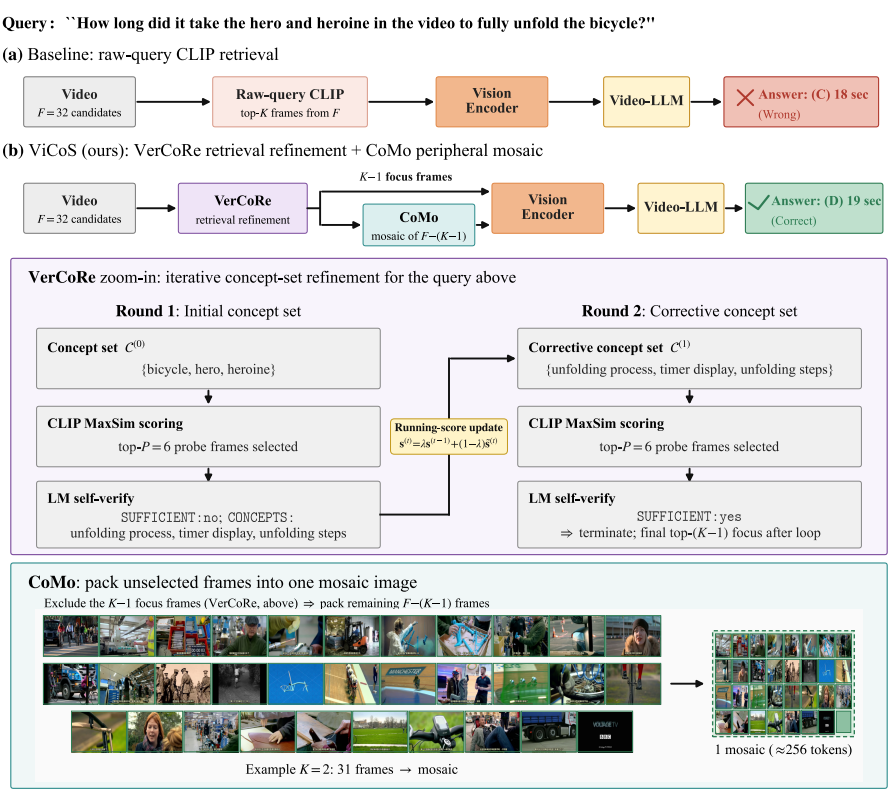


Figure 3: Overview of ViCoS. For $K \geq 2$, VerCoRe iteratively refines the retrieval scores to select $K-1$ full-resolution focus frames, while CoMo packs the remaining candidates into one temporally ordered mosaic occupying the final encoder slot. At $K=1$, VerCoRe is skipped and CoMo alone forms the single input image. For final answer generation, the vision encoder receives exactly K images. Concept sets are abbreviated for visualization.

conditioned on the current retrieval result, so the generated concepts are tied to an observed retrieval failure.

Base representation. We initialize $\mathcal{C}^{(0)}$ by parsing q with a lightweight rule that keeps content words (common nouns, proper nouns, verbs, and adjectives) and drops stopwords and function words. This yields a small set of textual concepts ($|\mathcal{C}^{(0)}| \leq 8$); if no content word remains, we fall back to the raw query.

Iterative multimodal verification. We perform at most T_A verification-and-correction rounds, with $T_A=2$ by default. The initial frame scores are $\mathbf{s}_i^{(0)} = \frac{1}{|\mathcal{C}^{(0)}|} \sum_{c \in \mathcal{C}^{(0)}} \max_p \cos(\phi_v^{\text{patch}}(f_i)_p, \phi_t(c))$ using CLIP ViT-L/14@336 patches. At each round t, the video-LLM verifies the current top-P retrieved frames ($P=6$). If the evidence is judged sufficient, refinement terminates. Otherwise, the model emits a corrective concept set $\mathcal{C}^{(t)}$, from which new MaxSim scores $\tilde{\mathbf{s}}^{(t)}$ are computed and blended with the running score:

$$\mathbf{s}^{(t)} = \lambda\, \mathbf{s}^{(t-1)} + (1-\lambda)\, \tilde{\mathbf{s}}^{(t)}, \tag{2}$$

with $\lambda=0.5$ in all experiments. The verification prompt template is provided in Appendix L, and the end-to-end pseudocode is provided in Appendix M.

Applicability. VerCoRe contributes when there is at least one focus slot to allocate, i.e. at $K \geq 2$; at $K=1$ the entire budget is the CoMo mosaic and VerCoRe is not invoked. The video-LLM acts as its own retrieval verifier, so no external multimodal refiner is required.

Table 1: Main accuracy comparison across two backbones, five benchmarks, and three visual-token budgets. Bold denotes the best compressed result for each benchmark and budget.

Method	VideoMME				MVBench				LongVB				CG-Bench				VideoVista			
	100%	50%	25%	12.5%	100%	50%	25%	12.5%	100%	50%	25%	12.5%	100%	50%	25%	12.5%	100%	50%	25%	12.5%
InternVL3-8B-Instruct																				
Uniform	58.96	55.22	50.00	46.93	65.20	63.08	58.20	50.35	56.10	51.31	49.96	46.45	38.00	30.93	30.97	31.33	78.84	78.72	77.66	75.69
Raw-query CLIP	–	57.00	54.26	50.74	–	59.78	56.23	53.08	–	55.87	54.45	52.13	–	33.23	35.17	33.60	–	77.82	76.39	75.49
Q-Frame	–	55.22	53.52	45.44	–	64.60	61.68	52.05	–	51.16	49.74	46.00	–	32.10	29.53	30.73	–	77.33	77.49	74.75
SparseVLM	–	58.26	56.70	53.33	–	63.15	58.50	53.90	–	53.85	50.41	48.24	–	32.47	31.67	30.67	–	78.96	78.43	77.74
FastV	–	59.22	58.22	56.19	–	72.50	70.30	62.98	–	54.45	52.88	50.41	–	32.67	31.93	31.30	–	79.17	78.47	78.76
VisionZip	–	48.81	45.19	42.81	–	47.12	43.58	42.88	–	45.62	44.35	43.46	–	29.50	28.33	27.53	–	76.23	75.20	74.22
ViCoS (ours)	–	**59.26**	**58.26**	**56.22**	–	**73.15**	**71.45**	**66.25**	–	**56.17**	**54.90**	**53.18**	–	**34.10**	**35.60**	**34.20**	–	**79.21**	**78.55**	**78.80**
Qwen2.5-VL-7B-Instruct																				
Uniform	52.70	49.56	43.78	41.89	60.30	58.85	53.45	47.70	55.10	50.64	44.05	41.36	34.00	30.33	29.17	27.30	76.00	75.20	75.08	70.22
Raw-query CLIP	–	52.19	49.11	46.37	–	52.68	49.83	48.35	–	53.70	50.94	50.04	–	31.60	29.87	29.77	–	75.37	74.84	70.55
Q-Frame	–	48.85	47.11	41.63	–	55.53	52.50	50.20	–	49.74	48.39	41.59	–	30.73	29.60	26.60	–	75.49	74.67	69.93
SparseVLM	–	51.74	50.52	48.07	–	58.30	55.10	49.10	–	50.79	48.39	45.25	–	31.60	30.63	30.23	–	76.02	76.55	76.31
FastV	–	51.96	51.19	50.11	–	63.05	61.55	57.38	–	51.61	50.64	46.90	–	31.70	31.37	30.53	–	75.53	75.94	76.51
VisionZip	–	47.52	45.48	44.78	–	56.73	52.83	49.42	–	45.10	44.95	44.58	–	30.80	30.17	29.67	–	75.53	75.78	75.82
ViCoS (ours)	–	**56.11**	**54.19**	**51.30**	–	**64.43**	**62.25**	**58.13**	–	**54.67**	**54.30**	**50.94**	–	**32.33**	**32.80**	**30.87**	–	**76.39**	**77.12**	**77.70**

3.3 CoMo: Context Mosaic

Under aggressive budgets, only a small number of candidates can be retained as full-resolution focus frames, while Section 2.3 shows that frames outside the retained set can still carry substantial query-aligned visual signal. CoMo preserves a coarse representation of this peripheral temporal context by arranging the unselected frames $\mathcal{U} = \{1, \ldots, F\} \setminus \mathcal{S}$ on an $R \times G$ grid (R=8, G=4 by default) in row-major temporal order and rendering the result as a single RGB image g at 448×448. The mosaic g occupies one of the K encoder slots, so the final answer stage receives exactly K encoder images. Grid geometry, cell resolution, and the prompt tag used to distinguish the mosaic from full-resolution focus frames are detailed in Appendix K.

Pseudocode for the end-to-end pipeline is provided as Algorithm 1 in Appendix M.

4 Experiments

We evaluate ViCoS on two video-LLM backbones (Qwen2.5-VL-7B (Bai et al., 2025) and InternVL3-8B (Zhu et al., 2025)) across five diverse video QA benchmarks (VideoMME (Fu et al., 2024), MVBench (Li et al., 2024b), LongVideoBench (Wu et al., 2024), CG-Bench (Chen et al., 2024a), VideoVista (Li et al., 2024d)) spanning short- to long-form settings, at three visual-token budgets (50%, 25%, 12.5%). We use K=8 encoder images as the 100% reference budget, so $K \in \{4, 2, 1\}$ correspond to $\{50\%, 25\%, 12.5\%\}$ of the reference visual-token budget respectively. Throughout the paper, we use K to denote the target frame-equivalent visual-token budget. For Uniform and ViCoS, K is also the number of encoder images; Q-Frame realizes the same K frame-equivalent budget through multi-resolution allocation, while SparseVLM/FastV/VisionZip realize it through post-encoder token pruning. The candidate pool from which all pre-encoder selection methods draw is fixed at F=32 uniformly sampled frames per video. We compare against Q-Frame (Zhang et al., 2025a) for pre-encoder frame selection and SparseVLM (Zhang et al., 2025b), FastV (Chen et al., 2024b), and VisionZip (Yang et al., 2024) for token-level compression, together with Uniform and raw-query CLIP references. All methods are evaluated under matched target frame-equivalent budgets. Full setup details, including hyperparameters, baseline implementations, and evaluation protocol, are provided in the appendix.

4.1 Main Results

ViCoS achieves the highest compressed accuracy among the compared methods in all 30 backbone–benchmark–budget settings (Table 1). Averaged across the 15 settings per backbone, ViCoS obtains **59.29** on InternVL3-8B and **55.57** on Qwen2.5-VL-7B, the highest average performance among the compared methods on both backbones. At the most aggressive 12.5% budget, ViCoS on InternVL3-8B reaches **78.80** on VideoVista, only 0.04 points below the full-budget K=8 reference (78.84). At K=1, ViCoS reduces to CoMo alone (Sec. 3.2–3.3), so improvements in this regime highlight the value of preserving peripheral temporal coverage rather than VerCoRe refinement.

Table 2: Budget-wise component ablation on InternVL3-8B, averaged over VideoMME, MVBench, and LongVideoBench at each budget. At $K=1$, VerCoRe is disabled by the applicability rule (Section 3.2): the $\times\times$ and $\checkmark\times$ variants both reduce to raw-query top-1 selection, and the $\times\checkmark$ and $\checkmark\checkmark$ variants both reduce to CoMo alone. The overall averages match Table 3.

VerCoRe	CoMo	50% ($K=4$)	25% ($K=2$)	12.5% ($K=1$)	Avg.
×	×	57.55	54.98	51.98	54.84
✓	×	57.56	54.81	51.98	54.78
×	✓	58.04	56.11	58.55	57.57
✓	✓	**62.86**	**61.54**	**58.55**	**60.98**

4.2 Ablation

We ablate each ViCoS component on InternVL3-8B, holding the CLIP backbone (ViT-L/14@336) and candidate pool ($F=32$) fixed.

Removing VerCoRe or CoMo (Tables 3 and 2). We compare four settings: (i) full ViCoS, (ii) ViCoS without VerCoRe (raw-query CLIP scoring used to pick the $K-1$ focus frames; CoMo unchanged), (iii) ViCoS without CoMo (VerCoRe used to pick top-K focus frames; the K-th slot is filled by an additional focus frame rather than the mosaic), (iv) neither, which reduces to raw-query CLIP top-K selection. At $K=1$, we disable VerCoRe for all ablation variants to preserve the same applicability rule as the full ViCoS pipeline (Sec. 3.2); hence the no-CoMo and neither variants both reduce to raw-query top-1 selection, and settings (iii) and (iv) coincide at the 12.5% budget. Full ViCoS achieves the strongest aggregate performance on all three benchmarks (Table 3). Table 2 further breaks the same ablation down by budget: CoMo provides the dominant contribution at $K=1$, where the entire budget is assigned to the mosaic and VerCoRe is not applicable, while VerCoRe further improves retrieval when focus slots are available at $K \geq 2$.

Retrieval strategy (Table 4). We compare four retrieval strategies on Qwen2.5-VL-7B/VideoMME while keeping the frame budget and answer-stage VLM fixed: (i) the raw query as anchor, (ii) content words extracted by our POS filter, (iii) concepts generated once by a text-only LLM from the query without video access or verification, and (iv) the running score produced by VerCoRe (content-word init + iterative video-conditioned score fusion). At the 50% budget, performance increases from 55.60 with the raw query to 55.80 with content-word filtering, 55.95 with one-shot LLM concepts, and 56.11 with VerCoRe. At the 25% budget, one-shot rewriting does not consistently improve over content-word filtering (53.81 vs. 54.10), whereas VerCoRe achieves the best result at 54.19. These results show that VerCoRe achieves the highest accuracy at both budgets, while one-shot query rewriting does not consistently improve over content-word filtering in this controlled setting.

Selection-stage efficiency is summarized in Table 5, with additional measurement details provided in Appendix J: at $K=2$ on VideoMME with $P=6$ probe frames, ViCoS uses on average 1.40 verifier calls due to early termination, adding a 5.8% latency overhead over Q-Frame while retaining the same nominal frame-equivalent answer-stage budget.

4.3 Additional Analysis

Budget matching across compressor families (Table 6). The methods in Table 1 realize the same target frame-equivalent budget at different compression stages. Uniform and ViCoS reduce encoder inputs directly to K images; Q-Frame uses its multi-resolution fixed-token allocation, so its raw encoder-image count is 7, 5, and 1 at the three budgets, corresponding to a frame-equivalent budget of 4, 2, and 1; and post-encoder baselines (SparseVLM, FastV, VisionZip) encode $K=8$ frames before reducing the retained visual-token budget. ViCoS therefore uses no more encoder images at the final answer stage than any budget-matched baseline while still matching the same target visual-token budget.

Table 3: Component ablation on InternVL3-8B, averaged over $K \in \{4, 2, 1\}$. Raw-query selection corresponds to disabling both VerCoRe and CoMo.

VerCoRe	CoMo	VideoMME	MVBench	LongVB
×	×	54.00	56.36	54.15
✓	×	53.94	57.98	52.43
×	✓	56.28	62.63	53.78
✓	✓	**57.91**	**70.28**	**54.75**

Table 4: Retrieval strategy comparison on Qwen2.5-VL-7B, VideoMME, with CoMo and the answer-stage budget fixed. The *Raw query* row differs from Table 1 because CoMo remains active.

Retrieval strategy	50%	25%
Raw query	55.60	53.75
Content words	55.80	54.10
One-shot LLM concepts	55.95	53.81
VerCoRe (ours)	**56.11**	**54.19**

Table 5: End-to-end inference efficiency on VideoMME at K=2 using Qwen2.5-VL-7B, averaged over 100 samples. Latency includes the selection-stage verifier. Q-Frame at K=2 passes 5 encoder images (1 high + 4 mid, see Appendix H) targeting the same nominal frame-equivalent budget. Round-1 termination for ViCoS: 60.0%.

Method	Answer imgs	Ret. budget	Verif. calls	Lat. (s)
Uniform	K	K	0	4.88
Q-Frame	5	nominal equiv. K	0	9.15
SparseVLM	K=8	equiv. K	0	7.80
ViCoS (ours)	K	K	1.4	9.68

Table 6: Answer-stage budget realization across compressor families. All methods target the same K frame-equivalent budget but realize it at different compression stages. For Q-Frame the encoder-image count follows the fixed-token multi-resolution allocation of Appendix H; the resulting frame-equivalent budget matches K exactly by construction.

Method	50% (K=4)	25% (K=2)	12.5% (K=1)	Budget realization
Uniform	4 imgs	2 imgs	1 img	Pre-encoder
Q-Frame	7 imgs → 4 equiv.	5 imgs → 2 equiv.	1 img → 1 equiv.	Multi-resolution
Post-encoder baselines	8 imgs → 50%	8 imgs → 25%	8 imgs → 12.5%	Token reduction
ViCoS	4 imgs	2 imgs	1 img	Pre-encoder

VerCoRe rounds. We fix T_A=2 as the default. Combined with the 60% round-1 termination rate (Table 5), a second verifier call is avoided for 60% of the samples. The per-round accuracy breakdown is provided in Appendix F.

5 Conclusion

We diagnosed two distinct but related problems in aggressive video-token compression: *mis-selection*, where raw-query CLIP retrieval misranks answer-relevant evidence, and *discarded-context*, where unselected frames still retain substantial query-aligned visual signal. We introduced *ViCoS*, a training-free two-module framework that addresses both. *VerCoRe* refines the retrieval scoring by having the video-LLM self-verify the current top-P frames and emit corrective visual concepts. *CoMo* preserves the peripheral temporal context by packing the unselected frames into a single low-resolution mosaic image that takes the last of the K encoder slots, so ViCoS uses no more encoder images at the final answer stage than any budget-matched baseline. On two backbones and five diverse video QA benchmarks, ViCoS achieves the highest compressed accuracy among the compared methods in all 30 backbone–benchmark–budget settings, and on VideoVista reaches 78.80, only 0.04 points below the full-budget K=8 reference using a single encoder image (the CoMo mosaic).