CoFi: Compress First, Focus Later for Long Video Understanding

Anonymous authors
Paper under double-blind review

Abstract

Long-video understanding requires allocating a limited visual input budget to observations that are both temporally informative and relevant to the question. Existing training-free methods typically address this bottleneck through query-aware frame selection, visual compression, or adaptive combinations of the two. However, adaptive representation methods generally determine where to allocate visual capacity before explicitly accounting for what information a compressed view of the video is estimated to preserve. We introduce **CoFi (Compress First, Focus Later)**, a training-free visual input construction framework that makes compression and focus acquisition sequentially dependent. CoFi first constructs a query-independent compressed temporal context and conditions a state–trajectory relational prior on this context to estimate the residual visual–temporal structure. It then introduces the question and sequentially acquires standalone focus observations that reduce query-relevant residual uncertainty, updating the residual structure after each acquisition. Under a fixed visual-unit budget, CoFi achieves the best performance on MLVU and LongVideoBench while remaining competitive with the best-performing frame-selection baseline on Video-MME. Ablation studies further show that compressed context and context-conditioned sequential focus acquisition play complementary roles, supporting the effectiveness of the proposed compress-first, focus-later design.

1 Introduction

Recent multimodal large language models (MLLMs) have made substantial progress in video understanding (Jin et al., 2024; Lin et al., 2024; Tang et al., 2026b; An et al., 2026). However, extending these models to long videos remains challenging, as dense visual inputs rapidly increase token and memory requirements, while current Video-LLMs can efficiently process only a limited number of visual observations (He et al., 2024; Tan et al., 2024; Shen et al., 2025; Zou et al., 2024; Liu et al., 2025a; Zohar et al., 2025; Jin et al., 2025). A common strategy is therefore to uniformly sample a fixed number of frames, but under a limited budget this provides increasingly sparse coverage as video duration grows and may miss important visual evidence. Long-video question answering (VideoQA) therefore becomes a problem of **visual budget allocation**: under a constrained input budget, which visual observations should be preserved, and at what fidelity?

Methods that improve upon this uniform-sampling baseline can be viewed along two complementary design dimensions, as illustrated in Fig. 1: *which visual observations to retain* and *how those observations should be represented* under a limited visual budget.

Along the first dimension, query-aware frame-selection methods prioritize observations based on query relevance together with temporal coverage, redundancy, diversity, or sequential structure (Liu et al., 2025b; Tang et al., 2025; Sun et al., 2025b; Wang et al., 2025b). These methods are effective when answer-relevant evidence occupies only a small portion of a long video.

Along the second dimension, efficient visual-representation methods reduce visual cost through temporal compression, composite inputs, or adaptive visual fidelity (Shen et al., 2025; Doorenbos et al., 2026; Tang et al., 2026a). These methods improve temporal coverage by representing more video content within a limited visual budget.

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.

These complementary perspectives suggest that representation and selection need not be treated independently. A compact representation can first expose broad, temporally distributed information, providing a basis for determining where additional visual evidence is still needed.

However, recent adaptive representation methods typically allocate visual fidelity directly from query relevance (Zhang et al., 2025a; Tang et al., 2026a), without explicitly accounting for what a compressed representation is estimated to preserve. This motivates conditioning subsequent focus allocation on the compressed context.

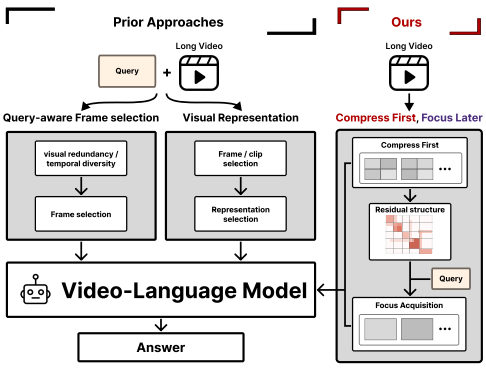


Figure 1: Conceptual comparison of existing approaches and CoFi.

We therefore introduce **CoFi (Compress First, Focus Later)**, a training-free framework for long-video visual input construction. CoFi first constructs a query-independent compressed context and conditions a state–trajectory relational prior on this context to obtain a residual covariance over the candidate sequence. It then sequentially allocates the remaining visual budget to observations that reduce query-relevant residual uncertainty, updating the residual structure after each acquisition. In this way, focus selection is explicitly conditioned on what the compressed context is estimated to already represent.

Experiments on Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), and LongVideoBench (Wu et al., 2024) demonstrate strong performance under a fixed visual-unit budget. CoFi achieves the best results on MLVU and LongVideoBench while remaining competitive on Video-MME.

Our contributions are threefold:

- We identify a limitation in existing long-video input construction: subsequent selection is typically not explicitly conditioned on what the compressed temporal context is estimated to already represent.
- We propose **CoFi**, a training-free framework that realizes this formulation by conditioning a state–trajectory relational prior on fidelity-aware compressed context and then sequentially acquiring query-relevant focus observations according to the resulting residual structure.
- We validate CoFi on three long-video benchmarks and multiple Video-LLM backbones, with extensive analyses of component roles, visual-budget allocation, representation robustness, and selection efficiency.

2 Method

2.1 Problem Setup and Overview

Given a long video, we construct a candidate frame sequence $\mathcal{I} = \{I_1, \ldots, I_N\}$ with a corresponding question q. Our goal is to construct a visual input for a frozen Video-LLM under a fixed visual-unit budget B.

We consider two types of visual units. A *context unit* is a rendered 2×2 panel containing four source frames, whereas a *focus unit* is a standalone frame. Each unit is provided to the downstream Video-LLM as a single image input under the same backbone-specific image preprocessing. Let B_C and B_F denote the numbers of context and focus units, respectively, with $B = B_C + B_F$.

Figure 2 illustrates the two-stage CoFi pipeline. *Compress First* constructs compressed temporal context and the context-conditioned residual covariance, while *Focus Later* allocates the remaining budget to query-relevant standalone observations.

The residual structure is a pre-inference estimate defined in the frozen visual representation space rather than a measure of what the downstream Video-LLM has actually understood from the compressed context.

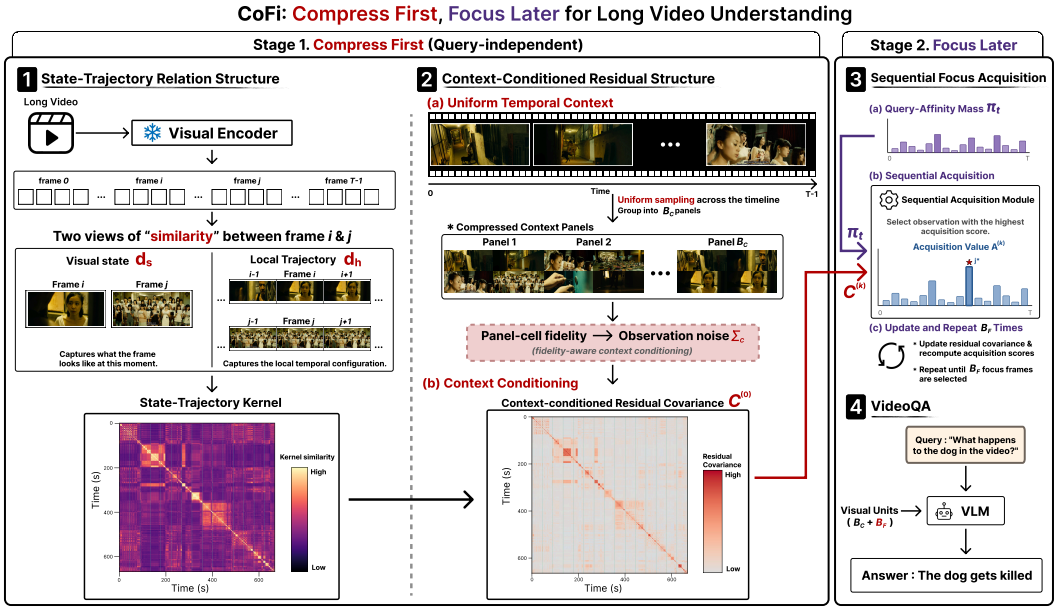


Figure 2: Overview of **CoFi**. CoFi first constructs a query-independent state–trajectory relation structure and compressed temporal context, then conditions the former on the latter to estimate the residual visual–temporal structure. The question is subsequently introduced to sequentially acquire standalone focus observations that reduce query-relevant residual uncertainty.

2.2 Compress First

The first stage defines all quantities independently of the query. This stage establishes a relational prior over the video, constructs a compact temporal context, and estimates the residual visual–temporal structure after conditioning on that context.

2.2.1 State–Trajectory Relation Structure

Video-centered state and local trajectory. Let E_V denote a frozen visual encoder, and let $x_t = E_V(I_t)/\|E_V(I_t)\|_2$ be the normalized representation of frame I_t. We characterize each candidate using both its instantaneous visual state and its ordered local temporal neighborhood:

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t, \quad s_t = \frac{x_t - \bar{x}}{\max\{\|x_t - \bar{x}\|_2, \epsilon_{\text{norm}}\}}, \quad h_t = \frac{[s_{t-r}; \ldots; s_t; \ldots; s_{t+r}]}{\sqrt{2r+1}} \tag{1}$$

Out-of-range indices are replaced by the nearest valid frame. The centered state s_t emphasizes frame-specific deviations from the video-wide mean representation, whereas the ordered neighborhood h_t describes the local semantic configuration surrounding frame t. This helps distinguish frames with similar instantaneous appearance when they occur under different local temporal contexts.

State–trajectory kernel. We define the state and trajectory distances as $d_s(i,j) = \|s_i - s_j\|_2^2$ and $d_h(i,j) = \|h_i - h_j\|_2^2$. Each distance is normalized using the video-specific median of the positive pairwise distances, denoted by m_s and m_h, respectively. For long candidate sequences, these bandwidth statistics are estimated from a uniformly sampled subset, as detailed in Appendix A.2.7. The resulting state–trajectory RBF kernel is

$$K(i,j) = \exp\left[-\alpha\frac{d_s(i,j)}{\gamma_{\text{bw}} m_s} - (1-\alpha)\frac{d_h(i,j)}{\gamma_{\text{bw}} m_h}\right], \tag{2}$$

where α balances instantaneous state similarity and local trajectory similarity, and γ_{bw} controls the overall kernel bandwidth.

The kernel is defined over all candidate locations and induces a conceptual kernel matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$. We use this kernel as a covariance prior over the candidate sequence. Its off-diagonal entries encode pairwise visual–temporal relations, allowing an observation at one location to modify residual structure at other related locations.

The full $N \times N$ matrix need not be explicitly materialized. The efficient implementation evaluates only the quantities required for context conditioning and subsequent focus acquisition; implementation details are provided in Appendix A.2.7.

The kernel does not determine which frames form the compressed context. It instead provides the relational prior through which the effect of those context observations is propagated across the candidate sequence.

2.2.2 Query-Independent Context

CoFi next constructs a compact temporal view of the full video by uniformly sampling $4B_C$ source frames across the video timeline. We denote their chronologically ordered indices by $\mathcal{C} = \{c_1, \ldots, c_{4B_C}\}$.

Uniform sampling is used only to construct a query-independent compressed context rather than to identify answer evidence. The sampled frames are grouped chronologically into sets of four and rendered as 2×2 panels, with temporal order preserved in raster order, from top-left to bottom-right.

2.2.3 Context-Conditioned Residual Structure

The source frames in $\mathcal{C}$ are presented inside compressed panel cells rather than as standalone observations. We therefore model the representation degradation induced by panel-cell downsampling as observation noise when conditioning the state–trajectory kernel prior.

Representation fidelity of compressed context. For each context source frame $c \in \mathcal{C}$, let ℓ_c denote the normalized representation obtained by re-encoding the same frame after panel-cell downsampling. We measure its representation-level fidelity as $f_c = \text{clip}(x_c^\top \ell_c, 0, 1)$ and map this fidelity to an effective observation-noise variance $\nu_c = \psi(f_c)$, where $\psi(\cdot)$ is monotonically decreasing. The corresponding context noise matrix is $\Sigma_C = \text{diag}(\nu_{c_1}, \ldots, \nu_{c_{4B_C}})$.

We therefore treat the context source frames as noisy observations under the kernel prior, with their conditioning strengths determined by representation fidelity. This independent-cell approximation is used only by the selection-time model; the downstream Video-LLM receives each completed 2×2 panel as a jointly rendered visual unit. The fidelity-to-noise mapping and further implementation details are provided in Appendix A.2.2.

Context-conditioned residual covariance. Conditioning the kernel prior on the compressed context observations gives

$$\mathbf{C}^{(0)} = \mathbf{K} - \mathbf{K}_{:\mathcal{C}} \left(\mathbf{K}_{\mathcal{C}\mathcal{C}} + \Sigma_C + \epsilon_{\text{cov}} I \right)^{-1} \mathbf{K}_{\mathcal{C}:}. \tag{3}$$

We refer to $\mathbf{C}^{(0)}$ as the *context-conditioned residual covariance*. Context conditioning modifies the full covariance structure rather than only its diagonal. Accordingly, $\mathbf{C}^{(0)}_{ij}$ represents the residual covariance between candidate locations i and j under the kernel model after accounting for the compressed context. The diagonal $\mathbf{C}^{(0)}_{tt}$ represents residual variance at candidate t, while the off-diagonal terms encode residual dependencies between candidate locations.

2.3 Focus Later

After the query-independent relation structure has been conditioned on the compressed context, CoFi introduces the question and allocates the remaining visual budget according to the resulting residual structure.

2.3.1 QUERY RELEVANCE

Let E_T denote the frozen text encoder associated with E_V, and let $z_q = E_T(q)/\|E_T(q)\|_2$ be the normalized query representation. We compute the frame–query affinity as $a_t = x_t^\top z_q$ and collect the scores into $\mathbf{a} = [a_1, \ldots, a_N]$. After robust normalization, $\widehat{a}_t = \mathrm{RobustNorm}(\mathbf{a})_t$, we transform the scores into a positive query-affinity mass:

$$\pi_t = \frac{\mathrm{softplus}(\widehat{a}_t)}{\sum_{u=1}^{N} \mathrm{softplus}(\widehat{a}_u)}. \tag{4}$$

The resulting weights satisfy $\sum_t \pi_t = 1$. We interpret π_t as relative query-affinity mass rather than a calibrated probability that frame t contains the answer.

Writing $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_N]$, we restrict focus acquisition to the higher-affinity portion of the candidate pool using $\mathcal{E}_\rho = \mathrm{TopMass}_\rho(\boldsymbol{\pi}, B_F)$, the smallest top-ranked set whose cumulative query-affinity mass reaches ρ while containing at least B_F candidates.

2.3.2 QUERY-WEIGHTED FOCUS ACQUISITION

Starting from $\mathbf{C}^{(0)}$, CoFi sequentially selects B_F standalone focus frames. Let $\mathcal{F}_k$ denote the set selected after k acquisitions, with $\mathcal{F}_0 = \varnothing$, and let $\mathbf{C}^{(k)}$ denote the corresponding residual covariance. A standalone focus observation is modeled with a small observation-noise variance ν_F. For each admissible candidate j, CoFi evaluates

$$A^{(k)}(j) = \frac{\pi_j \sum_{t=1}^{N} \pi_t \left(\mathbf{C}^{(k)}_{tj}\right)^2}{\mathbf{C}^{(k)}_{jj} + \nu_F}. \tag{5}$$

The acquisition score combines candidate-local query relevance with query-weighted residual reduction. Excluding the candidate-local factor π_j, the remaining term equals the one-step decrease in the query-weighted residual-variance objective $\sum_t \pi_t C^{(k)}_{tt}$ under the current kernel-space model. The factor π_j additionally favors candidates that are themselves query relevant. Through the off-diagonal terms $\mathbf{C}^{(k)}_{tj}$, this reduction can extend beyond candidate j to other query-relevant locations.

At step k, CoFi greedily selects

$$j^\star_{k+1} = \arg\max_{j \in \mathcal{E}_\rho \setminus \mathcal{F}_k} A^{(k)}(j). \tag{6}$$

2.3.3 SEQUENTIAL RESIDUAL UPDATE

After selecting $j^\star_{k+1}$, the residual covariance is updated as

$$\mathbf{C}^{(k+1)} = \mathbf{C}^{(k)} - \frac{\mathbf{C}^{(k)}_{:j^\star_{k+1}} \mathbf{C}^{(k)}_{j^\star_{k+1}:}}{\mathbf{C}^{(k)}_{j^\star_{k+1} j^\star_{k+1}} + \nu_F}. \tag{7}$$

The selected set is updated as $\mathcal{F}_{k+1} = \mathcal{F}_k \cup \{j^\star_{k+1}\}$, and all acquisition scores are recomputed under the updated residual covariance. Because the covariance changes after each acquisition, observations carrying overlapping residual information tend to receive smaller subsequent marginal gains, providing implicit redundancy control. The process repeats until B_F focus frames have been selected. Additional mathematical derivations and implementation details are provided in Appendix A.2.

2.4 FINAL VISUAL INPUT

The final Video-LLM input consists of the B_C compressed context panels, the B_F selected standalone focus frames, and the question. The mixed visual units are ordered approximately by video

time using representative timestamps, while the source frames within each panel preserve their internal chronological order.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate CoFi on Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), and LongVideoBench (Wu et al., 2024). We report benchmark-level accuracy on the multiple-choice evaluation sets, using the public validation split for LongVideoBench. For Video-MME, the Short, Medium, and Long duration breakdown is reported in Appendix D.1. All benchmarks are evaluated without subtitles, using the official question annotations and answer keys.

Video-LLM. We use LLaVA-Video-7B (Zhang et al., 2025b) for the main comparison and additionally evaluate cross-backbone transfer with Qwen2.5-VL-7B (Bai et al., 2025) and InternVL2.5-8B (Chen et al., 2025) under the same visual-input construction setting.

Baselines. We compare against Uniform Sampling; training-free keyframe-selection methods AKS† (Tang et al., 2025), BOLT† (Liu et al., 2025b), MDP^{3†} (Sun et al., 2025b), Q-Frame (QFS)† (Zhang et al., 2025a), FOCUS† (Zhu et al., 2026), GIFT* (Ma et al., 2026), and WFS-SB† (Chen et al., 2026); and visual-representation methods Video Panels† (Doorenbos et al., 2026) and Q-Fold* (Tang et al., 2026a). † denotes methods based on the authors' official implementations, whereas $*$ denotes our reimplementations from the paper when official code was unavailable. All methods are evaluated under the same 1-fps candidate pool, 32-VU budget, and LLaVA-Video-7B backbone, using the same frozen SigLIP encoder where applicable. Q-Frame† uses only its QFS component, and MDP^{3†} is evaluated under this unified 1-fps setting.

Visual Budget. Candidate frames are extracted at 1 fps, and all methods use a 32-visual-unit budget unless otherwise specified. We count one standalone frame or one rendered 2×2 panel as one visual unit (VU). CoFi allocates four VUs to context panels and 28 VUs to standalone focus frames. Since each context panel contains four source-frame observations, CoFi represents 44 source-frame observations using 32 VUs. We report #VU and the number of represented source-frame observations (#Src); Table 1 reports Avg. #Src, the macro-average across the three benchmarks.

Implementation Details. Downstream Video-LLM inference is conducted on NVIDIA RTX A6000 GPUs. We use SigLIP-SO400M-Patch14-384 (Zhai et al., 2023) as the default frozen vision-language encoder for visual input construction. We set $\alpha = 0.5$, $r = 1$, $\gamma_{\mathrm{bw}} = 1$, $\rho = 0.8$, $\nu_F = 10^{-6}$, $\epsilon_{\mathrm{norm}} = 10^{-12}$, and $\epsilon_{\mathrm{cov}} = 10^{-6}$ across all benchmarks without dataset-specific tuning. For long candidate sequences, bandwidth estimation uses at most $L_{\max} = 2048$ uniformly spaced candidate frames; further details are provided in Appendix A.2.7. Additional sensitivity, robustness, scalability, and efficiency analyses are provided in Appendix D.

3.2 MAIN RESULTS

Overall comparison. Under LLaVA-Video-7B, CoFi achieves the strongest performance on MLVU and LongVideoBench while remaining competitive on Video-MME. It also outperforms the visual-representation baselines despite representing fewer source-frame observations than Video Panels and Q-Fold, indicating that the gains are not explained simply by exposing the model to more source frames. CoFi further improves over Uniform Sampling across all three benchmarks, showing consistent gains from the overall visual input construction under the same 32-VU budget.

Transfer across Video-LLM backbones. CoFi consistently improves over Uniform Sampling across all three benchmarks with both Qwen2.5-VL-7B and InternVL2.5-8B under the same 32-VU budget. This indicates that the benefit of CoFi's visual input construction is not specific to the LLaVA-Video backbone and transfers across different downstream Video-LLM architectures.

Selection efficiency. Restricting the admissible candidate pool with the default $\rho = 0.8$ reduces CoFi's median selection latency by approximately 68–70% compared with using the full candidate

Table 1: Main comparison with training-free visual input construction methods under LLaVA-Video-7B, together with cross-backbone transfer results for Uniform Sampling and CoFi. All experiments use the same 32-VU budget within each backbone. #VU denotes visual units, and Avg. #Src denotes the macro-average number of source-frame observations represented across Video-MME, MLVU, and LongVideoBench. For the LLaVA-Video-7B comparison, best and second-best results are shown in **bold** and underline. Within each cross-backbone block, the better result is shown in **bold**. † denotes methods based on the authors' official implementations, whereas $*$ denotes our reimplementations from the paper when official code was unavailable.

Method	#VU	Avg. #Src	Video-MME (w/o sub)	MLVU	LongVideoBench
LLaVA-Video-7B					
Uniform Sampling	32	32.0	62.22	68.26	55.94
Keyframe Selection					
AKS†	32	32.0	63.70	71.34	61.11
BOLT†	32	32.0	62.74	72.36	60.00
MDP^3†	32	32.0	**65.11**	74.33	61.26
Q-Frame (QFS)†	32	32.0	62.00	67.07	59.69
FOCUS†	32	32.0	62.48	71.39	61.33
GIFT*	32	32.0	63.30	<u>74.56</u>	61.32
WFS-SB†	32	32.0	63.11	74.15	<u>61.85</u>
Visual Representation					
Video Panels†	32	108.7	63.78	70.70	57.82
Q-Fold*	32	69.5	62.78	72.95	61.48
CoFi (Ours)	32	44.0	<u>64.52</u>	**74.70**	**62.68**
Qwen2.5-VL-7B					
Uniform Sampling	32	32.0	59.81	64.03	58.56
CoFi (Ours)	32	44.0	**61.59**	**68.58**	**60.51**
InternVL2.5-8B					
Uniform Sampling	32	32.0	61.07	66.42	59.46
CoFi (Ours)	32	44.0	**62.93**	**69.83**	**64.70**

Table 2: Median CPU selection latency after candidate visual representations have been obtained. All methods are profiled under the same evaluation environment. For CoFi, $\rho = 1.0$ uses the full candidate pool, whereas the default $\rho = 0.8$ restricts the candidate pool using $\mathrm{TopMass}_\rho$. MDP^3 is measured under the same unified 1-fps candidate setting used in the main comparison.

Method	Selection Latency (s)		
	Video-MME	MLVU	LongVideoBench
GIFT*	0.006	0.006	0.006
WFS-SB†	0.021	0.022	0.022
MDP^3†	20.61	26.62	24.96
CoFi ($\rho = 1.0$)	0.106	0.106	0.108
CoFi ($\rho = 0.8$, default)	0.034	0.032	0.035

pool. CoFi remains slower than lightweight selectors such as GIFT and WFS-SB, but substantially faster than MDP^3. Detailed runtime and memory analyses are provided in Appendix D.

3.3 Ablation Study

Component ablation. We first examine the roles of compressed context and standalone focus observations, then isolate context conditioning and sequential residual updating during focus acquisition. Panel Only follows the Video Panels input, while Focus Only performs sequential acquisition from the unconditioned state–trajectory prior.

All $C4 + F28$ variants use the same context panels and focus budget. Uniform Focus selects focus frames uniformly. Unconditioned Sequential starts from $C^{(0)} = K$, whereas Static Acquisition selects all focus frames from the initial context-conditioned scores without intermediate updates.

Table 3: Ablation of CoFi on Video-MME, MLVU, and LongVideoBench. The upper group compares the roles of compressed context and standalone focus observations. The lower group evaluates context conditioning and sequential residual updating under the same $C4 + F28$ final visual input. Uniform Focus selects the focus observations uniformly, while the remaining focus-acquisition variants use the same query-affinity weighting. Cond. and Seq. denote context-conditioned residual modeling and sequential residual updating, respectively.

Variant	Allocation	Cond.	Seq.	Video-MME	MLVU	LongVideoBench
Representation Roles						
Panel Only	$C32 + F0$	–	–	63.78	70.70	57.82
Focus Only	$C0 + F32$	–	✓	63.93	73.28	61.85
Focus Acquisition under $C4 + F28$						
Uniform Focus	$C4 + F28$	–	–	62.22	67.53	58.49
Unconditioned Sequential	$C4 + F28$	–	✓	64.56	73.92	62.23
Static Acquisition	$C4 + F28$	✓	–	61.74	72.08	61.26
CoFi (Ours)	$C4 + F28$	✓	✓	64.52	74.70	62.68

Panel Only performs reasonably on Video-MME but is substantially weaker on MLVU and LongVideoBench, whereas Focus Only performs more strongly across the three benchmarks. CoFi outperforms both single-type inputs, showing that the full context–focus design is more effective than relying exclusively on compressed context or standalone focus observations.

Under the same $C4 + F28$ final visual input, recomputing $A^{(k)}$ after each residual update consistently improves over one-shot Static Acquisition across all three benchmarks. This shows that the acquisition criterion benefits from the evolving residual structure, allowing each selected observation to reshape the value of subsequent candidates.

Context conditioning complements this process by initializing the residual structure from the compressed temporal context, improving MLVU and LongVideoBench while leaving Video-MME essentially unchanged. Together, these results support CoFi's two-stage design: compressed context initializes what remains unexplained, while sequential acquisition progressively refines that residual structure.

Context–focus allocation. We next examine how the fixed 32-VU budget should be divided between compressed context and standalone focus observations.

Table 4: Effect of context–focus allocation under a fixed 32-VU budget.

Allocation	Video-MME				MLVU	LongVideoBench
	Short	Medium	Long	Overall		
$C0 + F32$	74.89	62.44	54.44	63.93	73.28	61.85
$C2 + F30$	74.56	61.78	54.00	63.44	73.78	**62.75**
$C4 + F28$ (default)	**75.89**	62.89	54.78	64.52	**74.70**	<u>62.68</u>
$C8 + F24$	75.44	<u>63.11</u>	<u>56.22</u>	<u>64.93</u>	73.23	61.63
$C16 + F16$	<u>75.67</u>	**63.22**	**56.56**	**65.15**	71.80	61.41

Table 4 shows different allocation preferences across benchmarks. Video-MME generally benefits from larger context allocations, whereas MLVU and LongVideoBench peak with smaller context allocations and degrade as context consumes more of the focus budget. The default $C4 + F28$ provides a strong cross-benchmark balance rather than a per-benchmark optimum.

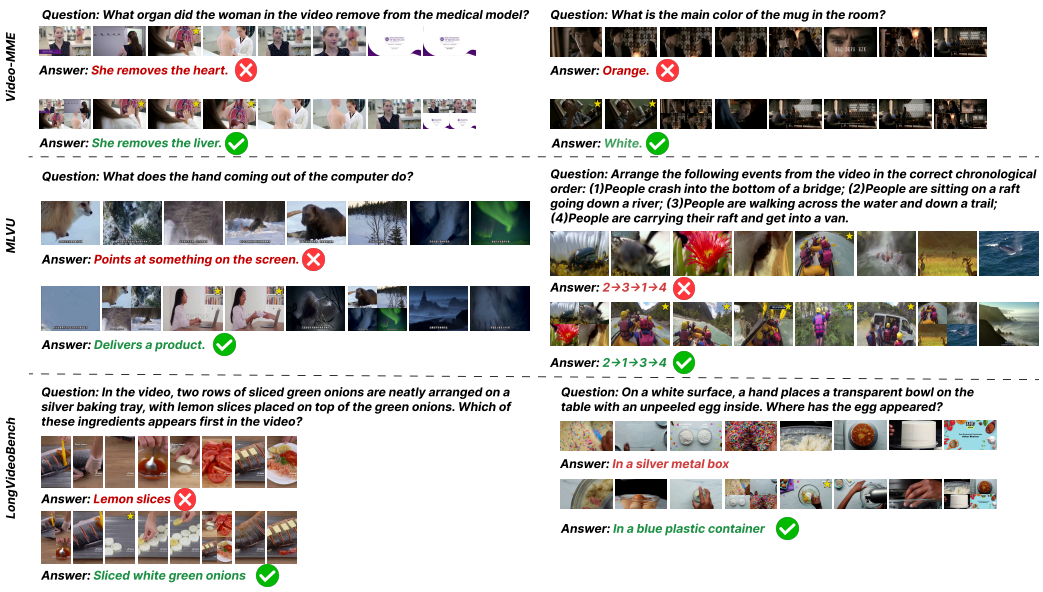


Figure 3: Qualitative comparison of Uniform Sampling and CoFi across Video-MME, MLVU, and LongVideoBench under the same 8-VU budget. The reduced budget enables direct inspection of the constructed visual inputs. Red and green denote incorrect and correct predictions, respectively.

3.4 Qualitative Analysis

Figure 3 compares Uniform Sampling and CoFi ($C4 + F4$) across the three benchmarks under the same reduced 8-VU budget for visual clarity. Across diverse evidence types, CoFi combines compressed temporal context with query-conditioned focus observations that expose evidence absent or less explicit in the corresponding uniform inputs.

The examples illustrate complementary roles of the two visual-unit types under a tight budget. For questions requiring localized visual evidence, such as identifying the removed organ or the mug color, CoFi allocates standalone focus observations near the relevant moment while retaining compressed coverage of the broader video. For temporally structured questions, such as event ordering, the combination of distributed context and selected focus observations provides evidence from multiple parts of the sequence that is less visible in the corresponding uniform input. These examples are consistent with the intended compress-first, focus-later design: context preserves broad temporal coverage, while focus acquisition uses the remaining capacity for query-relevant evidence.

4 Conclusion

We introduced **CoFi (Compress First, Focus Later)**, a training-free framework for long-video visual input construction under a fixed visual budget. CoFi first constructs a query-independent compressed temporal context, conditions a state–trajectory relational prior on it to estimate residual visual–temporal structure, and then sequentially allocates the remaining budget to query-relevant focus observations.

Across Video-MME, MLVU, and LongVideoBench, CoFi achieves strong performance under the same visual-unit budget, attaining the best results on MLVU and LongVideoBench while remaining competitive on Video-MME. Component and allocation studies further show that broad temporal context and context-conditioned sequential acquisition play complementary roles.

Taken together, these results support a broader design principle for long-video visual input construction: *representation should inform subsequent selection*. By first estimating what is already represented by the compressed context, the remaining visual capacity can be directed toward complementary, query-relevant evidence rather than allocated independently of the existing context.

Reproducibility Statement

We provide the full formulation of CoFi and its visual-input construction procedure in Sec. 2, together with the experimental setup, visual budgets, and default hyperparameters in Sec. 3.1. Additional implementation details, including representation fidelity, candidate-pool construction, sequential acquisition, computational complexity, and benchmark evaluation protocols, are provided in Appendices A and C. Anonymous source code for CoFi is available at `https://github.com/Anonymous-CoFi/CoFi/tree/main`. These details and resources are intended to support reproduction of the reported experiments.

LLM Usage Statement

In the preparation of this work, we used large language models (LLMs) for grammar correction and language polishing. We also used LLMs to assist with data visualization. All AI-assisted work was reviewed by the authors.

References

Zhaochong An, Zirui Li, Mingqiao Ye, Feng Qiao, Jiaang Li, Zongwei Wu, Vishal Thengane, Chengzu Li, Lei Li, Luc Van Gool, Guolei Sun, and Serge Belongie. Video understanding: From geometry and semantics to unified models. *Machine Intelligence Research*, 2026. ISSN 2731-5398. doi: 10.1007/s11633-026-1656-7. URL http://dx.doi.org/10.1007/s11633-026-1656-7.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Lorenzo Bottarelli and Marco Loog. Gaussian process variance reduction by location selection. *Pattern Recognition Letters*, 125:727–734, 2019.

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. Wavelet-based frame selection by detecting semantic boundary for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24052–24061, June 2026.

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, 2025. URL https://arxiv.org/abs/2412.05271.

David Cohn, Zoubin Ghahramani, and Michael Jordan. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994. URL https://proceedings.neurips.cc/paper_files/paper/1994/file/7f975a56c761db6506eca0b37ce6ec87-Paper.pdf.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. VideoZoomer: Reinforcement-learned temporal focusing for long video reasoning. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=ARHCFvgx6G.

Lars Doorenbos, Federico Spurio, and Juergen Gall. Video panels for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31293–31303, June 2026.

Junjie Fei, Jun Chen, Zechun Liu, Yunyang Xiong, Chong Zhou, Wei Wen, Junlin Han, Mingchen Zhuge, Saksham Suri, Qi Qian, Shuming Liu, Lemeng Wu, Raghuraman Krishnamoorthi, Vikas Chandra, Mohamed Elhoseiny, and Chenchen Zhu. Small vision-language models are smart compressors for long video understanding, 2026. URL https://arxiv.org/abs/2604.08120.

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. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 1986.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514. IEEE, 2024.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. FrameThinker: Learning to think with long videos via multi-turn frame spotlighting. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=nsNpsCpVG1.

Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, Sungjin Ahn, Jan Kautz, Hongxu Yin, Yao Lu, Song Han, and Wonmin Byeon. Storm: Token-efficient long video understanding for multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 5889–5900, October 2025.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13700–13710. IEEE, 2024.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, and Lizhuang Ma. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3:27, 2025.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008. URL http://jmlr.org/papers/v9/krause08a.html.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. LensWalk: Agentic video understanding by planning how you see in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19518–19528, 2026.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Jingren Liu, Yun Wang, Long Zhang, Yiheng Wang, Shuning Xu, Ling Wang, Jiaqi Yan, Dell Zhang, and Xiangyu Chen. Towards training-free long video understanding: methods, benchmarks, and open challenges. *Vicinagearth*, 2(1):6, 2025a.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3327, June 2025b.

Junpeng Ma, Sashuai Zhou, Guanghao Li, Xin Gao, Yue Cao, Hengyu Zeng, Yuxiang Yan, Zhibin Wang, Jun Song, Bo Zheng, Shanghang Zhang, and Jian Pu. Gift: Global irreplaceability frame targeting for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25610–25620, June 2026.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Aidean Sharghi, Boqing Gong, and Mubarak Shah. Query-focused extractive video summarization. *CoRR*, abs/1607.05177, 2016. URL http://arxiv.org/abs/1607.05177.

Aidean Sharghi, Jacob S Laurel, and Boqing Gong. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2127–2136. IEEE, 2017.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/shen25j.html.

Guangyu Sun, Archit Singhal, Burak Uzkent, Mubarak Shah, Chen Chen, and Garin Kessler. From frames to clips: Training-free adaptive key clip selection for long-form video understanding, 2025a. URL https://arxiv.org/abs/2510.02262.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp 3: a training-free approach for list-wise frame selection in video-llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24090–24101. IEEE, 2025b.

Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13581–13591. IEEE, 2024.

Biao Tang, Xu Chen, Shuxiang Gou, Jingyi Yuan, Yuhan Zhang, Deyu Meng, and Chenqiang Gao. Reunit: Multi-granularity visual unitization for long video understanding, 2026a. URL https://arxiv.org/abs/2606.12125.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128, June 2025.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376, 2026b. doi: 10.1109/TCSVT.2025.3566695.

Mengyue Wang, Shuo Chen, Kristian Kersting, Volker Tresp, and Yunpu Ma. METok: Multi-stage event-based token compression for efficient long video understanding. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings*

of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 18870–18884, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.954. URL https://aclanthology.org/2025.emnlp-main.954/.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Computer Vision – ECCV 2024*, pp. 58–76. Springer, 2024. doi: 10.1007/978-3-031-72989-8_4.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025b.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9088–9099, 2026.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857. Curran Associates, Inc., 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/329ad516cf7a6ac306f29882e9c77558-Paper-Datasets_and_Benchmarks_Track.pdf.

Chenwei Xu, Zhen Ye, Shang Wu, Weijian Li, Zihan Wang, Zhuofan Xia, Lie Lu, Pranav Maneriker, Fan Du, Manling Li, and Han Liu. Towards sparse video understanding and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11357–11368, 2026.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. VCA: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20168–20179, 2025.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591, June 2025.

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 (ICCV)*, pp. 11975–11986, October 2023.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pp. 310–325. Springer, 2024.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22056–22065, October 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856. URL https://openreview.net/forum?id=EElFGvt39K.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, 2025. doi: 10.1109/CVPR52734.2025.01278.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1OQKqLFcbB.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. Apollo: An exploration of video understanding in large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18891–18901, 2025. doi: 10.1109/CVPR52734.2025.01760.

Heqing Zou, Tianze Luo, Guiyang Xie, Victor Zhang, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, and Huaijian Zhang. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding, 2024. URL https://arxiv.org/abs/2409.18938.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=SZVpOKw0YD.

A Appendix

A.1 Related Work

Query-aware frame and clip selection. Recent long-video methods reduce visual input by selecting query-relevant observations using relevance, temporal coverage, redundancy, diversity, or sequential structure (Tang et al., 2025; Liu et al., 2025b; Zhu et al., 2026; Ma et al., 2026; Chen et al., 2026; Sun et al., 2025b). Earlier query-focused video summarization also combined query relevance with contextual importance or diversity when selecting representative video segments (Sharghi et al., 2016; 2017). Other approaches progressively refine relevant temporal regions through hierarchical or coarse-to-fine search (Wang et al., 2025b; Ye et al., 2025). Q-Frame (Zhang et al., 2025a) jointly performs query-aware frame selection and adaptive resolution, while F2C (Sun et al., 2025a) extends selection from isolated frames to temporally coherent clips and adjusts spatial resolution under a constrained visual budget. These approaches primarily determine which temporal observations should receive limited visual capacity, optionally adapting their representation during selection.

Efficient and adaptive visual representation. A complementary line of work reduces visual cost by changing how video content is represented. LongVU (Shen et al., 2025) removes temporally redundant content and applies text-guided compression, while Video Panels (Doorenbos et al., 2026) combines temporally distributed frames into composite images for broader temporal coverage. Q-Fold (Tang et al., 2026a) constructs heterogeneous visual representations by preserving query-relevant observations at higher fidelity and compressing less relevant content. Beyond frame-level input construction, recent methods also reduce visual tokens within the Video-LLM pipeline through event-aware pruning, temporal token compression, or learned query-aware compression (Wang et al., 2025a; Jiang et al., 2025; Fei et al., 2026). Together, these approaches show that limited visual capacity can be allocated not only by selecting observations, but also by adapting how those observations are represented.

Agentic and adaptive video reasoning. Recent work increasingly moves beyond fixed visual inputs by allowing additional video evidence to be acquired during reasoning. VideoAgent (Wang et al., 2024) iteratively retrieves visual information using foundation models as tools, while VCA (Yang et al., 2025) actively explores video segments through reward-guided search. A.I.R. (Zou et al., 2026) performs adaptive iterative frame selection using VLM-based semantic analysis, and ReViSe (Xu et al., 2026) maintains a compact reasoning state across multiple rounds to guide subsequent sparse frame acquisition. Other recent approaches explicitly learn or execute observation actions during reasoning: FrameThinker (He et al., 2026) and VideoZoomer (Ding et al., 2026) use reinforcement learning to optimize multi-turn frame selection or temporal focusing, while Active

Video Perception (Wang et al., 2026) and LensWalk (Li et al., 2026) couple planning, observation, and evidence assessment through iterative visual interaction. These methods demonstrate the benefit of adapting visual observation to information accumulated during reasoning, but typically realize this adaptivity through multiple rounds of downstream model inference or tool interaction. CoFi shares the principle that the value of an additional observation depends on information already available, but realizes it before downstream reasoning by constructing a context-conditioned visual input from frozen representations.

Uncertainty-based observation selection. Uncertainty reduction and information gain have long been used to select informative observations in active learning and Gaussian-process-based sensor placement (Cohn et al., 1994; Krause et al., 2008; Bottarelli & Loog, 2019). Such methods evaluate candidate observations by how much they reduce uncertainty over unobserved variables, with sequential conditioning naturally accounting for previously acquired information. CoFi applies this general principle to long-video visual input construction, where the relational structure is first conditioned on compressed visual context and subsequent acquisition is weighted by query relevance.

A.2 Method Details

This section provides additional mathematical and implementation details for the CoFi pipeline, following the order of Sec. 2.

A.2.1 State–Trajectory Kernel

The main paper defines the video-centered state s_t, local trajectory h_t, and the resulting state–trajectory kernel in Eqs. 1–2. Here we provide additional details on bandwidth normalization and the validity of the resulting kernel.

Video-adaptive distance normalization. In the full formulation, the state and trajectory distances are normalized independently for each video using their median positive pairwise distances:

$$m_s = \mathrm{median}_{i<j,\ d_s(i,j)>0}\, d_s(i,j), \qquad m_h = \mathrm{median}_{i<j,\ d_h(i,j)>0}\, d_h(i,j). \tag{8}$$

If no valid positive pairwise distance exists, we use a unit fallback statistic for numerical stability.

This normalization prevents the kernel scale from depending directly on the magnitude of representation variation within a particular video. At temporal boundaries, the local trajectory representation uses the nearest valid state for out-of-range indices, as described in the main paper.

Positive semidefiniteness. The state–trajectory kernel can be interpreted as a Gaussian RBF kernel over a scaled joint representation. Define

$$\phi_t = \left[\sqrt{\frac{\alpha}{\gamma_{\mathrm{bw}} m_s}}\, s_t \,;\, \sqrt{\frac{1-\alpha}{\gamma_{\mathrm{bw}} m_h}}\, h_t \right]. \tag{9}$$

Then,

$$\|\phi_i - \phi_j\|_2^2 = \frac{\alpha}{\gamma_{\mathrm{bw}} m_s} d_s(i,j) + \frac{1-\alpha}{\gamma_{\mathrm{bw}} m_h} d_h(i,j), \tag{10}$$

and therefore

$$K(i,j) = \exp\left(-\|\phi_i - \phi_j\|_2^2\right). \tag{11}$$

For $0 \le \alpha \le 1$ and positive bandwidth terms, this is a Gaussian RBF kernel and is positive semidefinite. This justifies using $\mathbf{K}$ as a kernel-induced covariance prior in the subsequent conditioning procedure.

A.2.2 REPRESENTATION FIDELITY

The main paper models representation degradation induced by panel-cell downsampling as observation noise during context conditioning. For each context source frame $c \in \mathcal{C}$, let x_c denote its normalized original frame representation.

To estimate the representation change induced by panel-cell downsampling, we render the same frame at the spatial resolution of a single 2×2 panel cell and re-encode it using the same frozen visual encoder. Let ℓ_c denote the resulting normalized representation. The representation-level fidelity is then

$$f_c = \mathrm{clip}\left(x_c^\top \ell_c, 0, 1\right). \tag{12}$$

Since both representations are normalized, f_c corresponds to their cosine similarity and measures how well the frozen visual representation is preserved under panel-cell downsampling.

For numerical stability, we define $\bar{f}_c = \mathrm{clip}(f_c, 10^{-4}, 1)$ and convert fidelity to observation noise as

$$\nu_c = \frac{1 - \bar{f}_c}{\bar{f}_c}. \tag{13}$$

Thus, context observations whose representations change more under panel-cell downsampling exert weaker influence during kernel conditioning. For the $M_C = 4B_C$ context source observations, we construct

$$\Sigma_C = \mathrm{diag}\left(\nu_{c_1}, \ldots, \nu_{c_{M_C}}\right). \tag{14}$$

Independent-cell approximation. The selection-time model evaluates each context source frame independently at panel-cell resolution rather than encoding the completed 2×2 panel. Cross-cell interactions are therefore not modeled during context conditioning, although the downstream Video-LLM receives the completed panel as a jointly rendered visual unit.

A.2.3 CONTEXT CONDITIONING

The context-conditioned residual covariance is defined in Eq. 3. Here we clarify the interpretation and bounds of its diagonal values.

Since $\mathbf{K} \succeq 0$ and $\Sigma_C + \epsilon_{\mathrm{cov}} I \succ 0$, standard Gaussian conditioning gives (Rasmussen & Williams, 2006)

$$\mathbf{C}^{(0)} \succeq 0, \qquad \mathbf{C}^{(0)} \preceq \mathbf{K}. \tag{15}$$

Since the RBF kernel satisfies $K(t,t) = 1$, $0 \le C_{tt}^{(0)} \le 1$. These diagonal values characterize residual variance within the frozen representation space rather than downstream Video-LLM uncertainty.

A.2.4 QUERY-AFFINITY NORMALIZATION

Given the raw frame–query affinities $a_t = x_t^\top z_q$, collected in $\mathbf{a} = [a_1, \ldots, a_N]$, we first apply a median–MAD robust normalization (Hampel et al., 1986):

$$\widehat{a}_t = \frac{a_t - \mathrm{median}(\mathbf{a})}{1.4826\ \mathrm{MAD}(\mathbf{a})}. \tag{16}$$

For degenerate scales, we fall back to the standard deviation and then to a unit scale, using 10^{-8} as the degeneracy threshold.

We then map the normalized scores to positive values using $\mathrm{softplus}$ and normalize them across the video timeline:

$$\pi_t = \frac{\mathrm{softplus}(\widehat{a}_t)}{\sum_{u=1}^{N} \mathrm{softplus}(\widehat{a}_u)}. \tag{17}$$

The resulting π_t represents relative query-affinity mass over the candidate frames rather than a calibrated probability of containing the answer.

A.2.5 Candidate-Pool Construction

The $\mathrm{TopMass}_\rho$ operator in the main paper constructs the admissible candidate set used by sequential focus acquisition.

Let $\tau(1), \ldots, \tau(N)$ denote candidate indices sorted by decreasing query-affinity mass:

$$\pi_{\tau(1)} \geq \pi_{\tau(2)} \geq \cdots \geq \pi_{\tau(N)}. \tag{18}$$

We retain

$$M = \max\left[B_F,\, \min\left\{m \in \{1, \ldots, N\} : \sum_{k=1}^{m} \pi_{\tau(k)} \geq \rho\right\}\right], \tag{19}$$

and define

$$\mathcal{E}_\rho = \{\tau(1), \ldots, \tau(M)\}. \tag{20}$$

Thus, $\mathcal{E}_\rho$ contains the smallest set of top-ranked candidates covering at least a fraction ρ of the query-affinity mass, while always retaining at least B_F candidates.

The query-affinity mass is used both to construct the admissible candidate set and to weight acquisition within that set. Specifically, π_j weights candidate j directly, while $\{\pi_t\}_{t=1}^{N}$ weights its residual-variance reduction across video locations.

A.2.6 Sequential Focus Acquisition

The main paper defines the acquisition score in Eq. 5. Here we decompose the score to clarify the distinct roles of query-weighted residual-variance reduction and candidate-local query relevance.

Suppose that after k focus observations the current residual covariance is $\mathbf{C}^{(k)}$. If candidate j is observed with focus observation noise ν_F, applying the same Gaussian conditioning rule gives the rank-one covariance update

$$\mathbf{C}^{(k+1;j)} = \mathbf{C}^{(k)} - \frac{\mathbf{C}^{(k)}_{:j}\mathbf{C}^{(k)}_{j:}}{C^{(k)}_{jj} + \nu_F}. \tag{21}$$

The corresponding reduction in residual variance at location t is

$$\begin{aligned}\Delta^{(k)}_j(t) &= C^{(k)}_{tt} - C^{(k+1;j)}_{tt} \\ &= \frac{\left(C^{(k)}_{tj}\right)^2}{C^{(k)}_{jj} + \nu_F}.\end{aligned} \tag{22}$$

This quantity measures how much observing candidate j reduces the remaining variance at a particular video location t. Importantly, the reduction depends on the residual covariance $C^{(k)}_{tj}$ rather than only on the candidate's own variance $C^{(k)}_{jj}$. Thus, observing j can reduce residual variance not

only at j itself but also at other locations that remain strongly related to it after conditioning on the compressed context and previously selected focus observations.

Weighting this reduction across all candidate locations by the query-affinity mass gives the integrated query-weighted reduction

$$G^{(k)}(j) = \sum_{t=1}^{N} \pi_t \Delta_j^{(k)}(t) = \frac{\sum_{t=1}^{N} \pi_t \left(C_{tj}^{(k)}\right)^2}{C_{jj}^{(k)} + \nu_F}. \tag{23}$$

Thus, $G^{(k)}(j)$ measures the one-step reduction in query-weighted residual variance induced by observing candidate j. The weights $\{\pi_t\}_{t=1}^{N}$ determine where residual reduction is most valuable: reductions at locations with higher query affinity contribute more strongly to the integrated gain. Through the off-diagonal covariance terms, $G^{(k)}(j)$ also rewards candidates whose observation can explain residual structure shared across multiple query-relevant locations.

CoFi additionally weights this integrated reduction by the candidate's own query-affinity mass π_j, yielding

$$A^{(k)}(j) = \pi_j G^{(k)}(j) = \frac{\pi_j \sum_{t=1}^{N} \pi_t \left(C_{tj}^{(k)}\right)^2}{C_{jj}^{(k)} + \nu_F}, \tag{24}$$

which recovers Eq. 5.

The two uses of query affinity play distinct roles. The weights $\{\pi_t\}_{t=1}^{N}$ determine how strongly residual reduction at each video location contributes to the integrated gain, whereas π_j favors candidates that are themselves directly relevant to the query. Consequently, a candidate receives a high acquisition value when it is itself query relevant and its observation also reduces residual variance across other query-relevant locations.

These weighting roles are also distinct from the TopMass restriction $\mathcal{E}_\rho$. The set $\mathcal{E}_\rho$ determines which candidates are admissible for focus acquisition, whereas $A^{(k)}(j)$ ranks those admissible candidates by combining candidate-local relevance with query-weighted residual reduction.

At iteration k, the selected candidate is

$$j_{k+1}^{\star} = \arg \max_{j \in \mathcal{E}_\rho \setminus \mathcal{F}_k} A^{(k)}(j), \tag{25}$$

after which Eq. 7 is applied and all acquisition scores are recomputed.

Implicit redundancy control. After an observation is acquired, conditioning updates the residual covariance, which tends to lower the acquisition value of observations carrying overlapping residual information. Thus, the evolving residual covariance provides implicit redundancy control without an explicit diversity penalty.

A.2.7 Long-Video Approximation and Efficient Implementation

The formulation above defines the bandwidth statistics, kernel, and residual covariance over the full candidate sequence. For long videos, exhaustive bandwidth estimation and explicit materialization of the full $N \times N$ covariance become increasingly expensive. CoFi therefore bounds the computation used for bandwidth estimation and directly evaluates only the kernel quantities needed for context conditioning and the posterior columns needed for subsequent focus acquisition.

Bandwidth approximation. The exact bandwidth statistics in Eq. 8 are defined using positive pairwise distances over the full candidate sequence. Computing these statistics exactly requires quadratic pairwise computation in N.

For efficient processing of long videos, we use $L = \min(N, L_{\max})$ frames for bandwidth estimation.

When $N \leq L_{\max}$, all candidate frames are used for bandwidth estimation. For longer videos, we uniformly sample $L_{\max}$ frames across the candidate sequence. Let $\mathcal{S}_L$ denote the resulting set of indices. We then estimate

$$\widehat{m}_s = \mathrm{median}_{i<j,\, i,j\in\mathcal{S}_L,\, d_s(i,j)>0}\, d_s(i,j), \qquad \widehat{m}_h = \mathrm{median}_{i<j,\, i,j\in\mathcal{S}_L,\, d_h(i,j)>0}\, d_h(i,j). \tag{26}$$

For $N > L_{\max}$, the kernel in Eq. 2 uses $\widehat{m}_s$ and $\widehat{m}_h$ in place of the full-sequence statistics m_s and m_h. This bounds the pairwise computation required for bandwidth estimation as video duration increases. The subset $\mathcal{S}_L$ is used only for bandwidth estimation; subsequent kernel evaluation retains all N candidate locations as targets, while admissible focus candidates are determined independently by $\mathcal{E}_\rho$.

Posterior-column context conditioning. Once the bandwidth statistics are fixed, the full $N \times N$ kernel matrix need not be explicitly materialized.

Although $\mathbf{C}^{(0)}$ is defined independently of the query, the implementation evaluates only the columns indexed by $\mathcal{E}_\rho$ after the admissible set is known. This is a lazy evaluation of the same query-independent context-conditioned covariance rather than query-dependent conditioning.

Let $M_C = |\mathcal{C}| = 4B_C$ denote the number of context source observations.

The required computation uses

$$\mathbf{K}_{:\mathcal{C}} \in \mathbb{R}^{N\times M_C}, \qquad \mathbf{K}_{\mathcal{C}\mathcal{C}} \in \mathbb{R}^{M_C\times M_C}, \qquad \mathbf{K}_{\mathcal{C}\mathcal{E}_\rho} \in \mathbb{R}^{M_C\times M}, \qquad \mathbf{K}_{:\mathcal{E}_\rho} \in \mathbb{R}^{N\times M}. \tag{27}$$

Rather than forming the inverse explicitly, we solve

$$(\mathbf{K}_{\mathcal{C}\mathcal{C}} + \Sigma_C + \epsilon_{\mathrm{cov}} I)\, \mathbf{X} = \mathbf{K}_{\mathcal{C}\mathcal{E}_\rho}, \tag{28}$$

where $\mathbf{X} \in \mathbb{R}^{M_C\times M}$. The context-conditioned covariance columns required for focus acquisition are then obtained directly as

$$\mathbf{C}^{(0)}_{:\mathcal{E}_\rho} = \mathbf{K}_{:\mathcal{E}_\rho} - \mathbf{K}_{:\mathcal{C}}\mathbf{X}, \tag{29}$$

with $\mathbf{C}^{(0)}_{:\mathcal{E}_\rho} \in \mathbb{R}^{N\times M}$.

This is exactly the corresponding set of columns of the full context-conditioned covariance in Eq. 3. Thus, all N candidate locations remain represented as targets, while only the covariance columns associated with admissible focus candidates are evaluated and stored.

Posterior-column sequential update. The same posterior-column representation is maintained during sequential focus acquisition. Although Eq. 21 defines the update over the full residual covariance, acquisition scores require only the columns associated with $\mathcal{E}_\rho$. For a selected candidate $j^\star \in \mathcal{E}_\rho$, we update

$$\mathbf{C}^{(k+1)}_{:\mathcal{E}_\rho} = \mathbf{C}^{(k)}_{:\mathcal{E}_\rho} - \frac{\mathbf{C}^{(k)}_{:j^\star}\mathbf{C}^{(k)}_{j^\star\mathcal{E}_\rho}}{C^{(k)}_{j^\star j^\star} + \nu_F}. \tag{30}$$

Because $j^\star \in \mathcal{E}_\rho$, all quantities in this update are available from the maintained $\mathbf{C}^{(k)}_{:\mathcal{E}_\rho}$ representation. The update therefore preserves the effect of each acquired observation on all N candidate locations without constructing covariance columns for locations outside the admissible set $\mathcal{E}_\rho$.

The only approximation introduced by the long-video implementation is therefore the bounded estimation of the bandwidth statistics. Given these statistics, the context-conditioning and sequential-update steps above evaluate exactly the posterior quantities required for focus acquisition without materializing the full $N \times N$ kernel or residual covariance.

A.2.8 COMPUTATIONAL COMPLEXITY

Let N denote the number of candidate frames, $M_C = 4B_C$ the number of context source observations, $M = |\mathcal{E}_\rho|$ the number of admissible focus candidates, B_F the number of acquired focus observations, $L_{\max}$ the bandwidth-estimation cap, and $L = \min(N, L_{\max})$ the number of candidate frames used for bandwidth estimation.

Bandwidth estimation. Computing the exact median pairwise distances in Eq. 8 over the full candidate sequence requires $O(N^2)$ pairwise evaluations. For long videos, CoFi instead uses $L = \min(N, L_{\max})$ uniformly spaced candidate frames, requiring $O(L^2)$ pairwise evaluations. Thus, for fixed $L_{\max}$, the bandwidth-estimation cost remains bounded as N increases beyond the cap.

Context conditioning. CoFi directly constructs the posterior columns associated with the M admissible focus candidates rather than the full $N \times N$ context-conditioned covariance. Factorizing the $M_C \times M_C$ context-conditioning matrix $\mathbf{K}_{\mathcal{C}\mathcal{C}} + \Sigma_C + \epsilon_{\text{cov}} I$ requires $O(M_C^3)$ operations. Solving Eq. 28 for M right-hand sides requires $O(M_C^2 M)$ operations. The subsequent multiplication $\mathbf{K}_{:\mathcal{C}}\mathbf{X}$ in Eq. 29 requires $O(NM_C M)$ operations.

The required kernel quantities involve $O(NM + NM_C + M_C M + M_C^2)$ pairwise kernel evaluations, which is dominated by $O(NM)$ when M_C is fixed. Excluding kernel evaluation, the linear-algebra cost of context conditioning is therefore

$$O(M_C^3 + M_C^2 M + NM_C M).$$

Because M_C is fixed by the visual budget and is small in our experiments, the dominant scaling is governed by N and M.

Sequential focus acquisition. At each acquisition step, the score in Eq. 5 is evaluated for the M admissible focus candidates, with the integrated residual reduction summed over all N candidate locations. This requires $O(NM)$ operations. The posterior-column rank-one update in Eq. 30 likewise requires $O(NM)$ operations. For a focus budget of B_F, sequential acquisition therefore requires

$$O(B_F NM)$$

operations.

The maintained posterior representation $\mathbf{C}^{(k)}_{:\mathcal{E}_\rho}$ requires $O(NM)$ storage. The context kernel quantity $\mathbf{K}_{:\mathcal{C}}$ requires an additional $O(NM_C)$ storage. Thus, with fixed M_C, the overall maintained storage is $O(NM)$.

The bounded bandwidth estimator further prevents full-sequence pairwise median estimation from becoming an initial quadratic bottleneck for very long videos. Nevertheless, M may increase with N, so the $O(NM)$ posterior representation and sequential acquisition can still approach quadratic scaling when the retained candidate set grows proportionally with video length. Given cached frozen representations, the complete selection procedure requires no downstream Video-LLM inference.

B LIMITATIONS

Residual covariance interpretation. CoFi models residual visual–temporal structure in a frozen representation space rather than the internal representation of the downstream Video-LLM. The resulting covariance should therefore be interpreted as a representation-space proxy for residual structure, not as calibrated downstream Video-LLM uncertainty.

Temporal resolution and scalability. CoFi constructs the candidate sequence at 1 fps, which may miss short-lived evidence between sampled frames. Although the implementation avoids materializing the full $N \times N$ residual covariance, focus acquisition maintains an $N \times M$ posterior cross-covariance block between all candidate locations and the M admissible focus candidates. Its computation and memory can therefore still approach quadratic scaling when M grows proportionally with video length. More scalable residual representations may be beneficial for denser sampling or substantially longer videos.

Context and temporal-ordering approximations. The mapping from representation fidelity to observation noise is heuristic, and context source frames are modeled independently during selection-time conditioning even though the downstream Video-LLM processes each rendered panel jointly. Moreover, mixed context and focus units are ordered by representative timestamps, which does not preserve exact frame-level temporal interleaving. Thus, cross-cell interactions are not explicitly modeled during context conditioning, and fine-grained temporal ordering is only approximately represented in the final mixed visual input.

C Benchmark and Evaluation Details

C.1 Benchmarks

C.1.1 Video-MME

Video-MME is a comprehensive multiple-choice benchmark for video understanding across diverse content and temporal scales (Fu et al., 2025). It contains 900 videos and 2,700 human-annotated question–answer pairs across six visual domains and 30 fine-grained categories, with an average video duration of approximately 17.0 minutes. We evaluate the visual-only setting without subtitles and report accuracy for the Short, Medium, and Long subsets together with the aggregate Overall score.

C.1.2 MLVU

MLVU is a multi-task benchmark designed for long-video understanding (Zhou et al., 2025), covering diverse video genres and reasoning tasks. Our experiments evaluate its multiple-choice portion, including Topic Reasoning, Anomaly Recognition, Needle QA, Ego Reasoning, Plot QA, Action Order, and Action Count. The evaluated videos have an average duration of approximately 12.6 minutes. We report aggregate multiple-choice accuracy.

C.1.3 LongVideoBench

LongVideoBench evaluates long-context video-language understanding through referring-reasoning questions over long videos (Wu et al., 2024). Following prior frame-selection evaluations, we use the public validation split without subtitles, containing 1,337 question–answer pairs. The evaluated videos have an average duration of approximately 8.0 minutes, and we report multiple-choice accuracy on this split.

C.2 Implementation and Evaluation

C.2.1 Prompt and Temporal Ordering

CoFi provides the Video-LLM with a mixed visual sequence consisting of temporal context panels and standalone focus frames. To indicate the temporal structure of this mixed representation, we prepend the following instruction to each benchmark question:

> *Some video images are temporal panels. Within every panel, time proceeds in raster order: left-to-right and then top-to-bottom. The visual units are ordered approximately by video time using their representative timestamps.*

Within each 2×2 temporal panel, the four source frames are arranged chronologically in raster order. Each panel is assigned the mean timestamp of its four constituent frames, which is used to order the panel together with the standalone focus frames. Because each panel spans multiple temporal moments, this unit-level ordering does not necessarily preserve the exact temporal interleaving between individual panel cells and neighboring focus frames.

C.2.2 Evaluation Protocol

For all benchmarks, we use the official question annotations, answer choices, and ground-truth labels. Candidate video frames are extracted at 1 fps before applying each visual-input construction

method. All benchmarks are evaluated using visual inputs only; subtitles and other auxiliary video-text streams are not provided to the downstream Video-LLM.

All methods compared in the main experiment are evaluated using the same downstream Video-LLM and the same visual-unit budget. For CoFi, the final visual input consists of the compressed context panels and selected focus frames, temporally ordered as described above. The Video-LLM receives the constructed visual sequence together with the original benchmark question and answer choices.

D Additional Experimental Analysis

D.1 Video-MME Duration-wise Results

Video-MME organizes its evaluation into Short, Medium, and Long duration subsets. Table 5 reports the corresponding breakdown for the methods in the main comparison.

Table 5: Duration-wise results on Video-MME under LLaVA-Video-7B and the same 32-VU budget as the main comparison. Best and second-best results are shown in **bold** and underline. † and $*$ follow the implementation-source notation in Table 1.

Method	Short	Medium	Long	Overall
Uniform Sampling	74.22	60.11	52.33	62.22
Keyframe Selection				
AKS†	**76.78**	61.11	53.22	63.70
BOLT†	74.44	60.22	53.56	62.74
MDP^3†	<u>76.67</u>	**63.67**	**55.00**	**65.11**
Q-Frame (QFS)†	73.56	59.56	52.89	62.00
FOCUS†	74.22	59.22	54.00	62.48
GIFT*	76.00	61.11	52.78	63.30
WFS-SB†	75.78	60.44	53.11	63.11
Visual Representation				
Video Panels†	75.22	62.44	53.67	63.78
Q-Fold*	74.78	59.44	54.11	62.78
CoFi (Ours)	75.89	<u>62.89</u>	<u>54.78</u>	<u>64.52</u>

CoFi remains strong across video durations, achieving the second-best Medium, Long, and Overall scores. In particular, its Medium and Long results show that CoFi remains competitive on the longer-duration subsets.

D.2 Hyperparameter Sensitivity

We evaluate the sensitivity of CoFi to its state–trajectory kernel, query-affinity threshold, and RBF bandwidth. Unless otherwise specified, all experiments use the default $C4+F28$ configuration with 32 visual units.

State–trajectory kernel. We examine different combinations of state and trajectory information and vary the temporal radius used to construct the local trajectory representation.

The default combined kernel with $\alpha = 0.5$ and $r = 1$ performs consistently across benchmarks, achieving the best performance on MLVU and LongVideoBench while remaining within 0.07 points of the best Video-MME Overall result. These results support jointly modeling instantaneous visual state and local temporal configuration using a compact $r = 1$ neighborhood.

Query-affinity mass threshold. We vary ρ, which controls the admissible candidate pool for focus acquisition.

Table 6: Sensitivity to the state–trajectory kernel configuration.

Kernel	Video-MME				MLVU	LongVideoBench
	Short	Medium	Long	Overall		
Combined ($\alpha = .5, r = 1$)	75.89	**62.89**	54.78	64.52	**74.70**	**62.68**
State only ($\alpha = 1, r = 1$)	**76.33**	62.78	54.56	64.56	74.24	61.71
Trajectory only ($\alpha = 0, r = 1$)	76.00	62.22	**55.56**	**64.59**	74.10	61.85
Combined ($\alpha = .5, r = 2$)	76.00	62.56	54.11	64.22	74.24	61.78
Combined ($\alpha = .5, r = 4$)	75.78	62.67	54.78	64.41	73.74	61.85

Table 7: Sensitivity to the query-affinity mass threshold ρ.

ρ	Video-MME				MLVU	LongVideoBench
	Short	Medium	Long	Overall		
0.6	75.33	**63.11**	54.00	64.15	74.52	62.23
0.8	**75.89**	62.89	**54.78**	**64.52**	**74.70**	**62.68**
1.0	75.56	63.00	54.67	64.41	74.61	62.45

The default $\rho = 0.8$ achieves the best benchmark-level performance on all three datasets. A moderate query-affinity mass threshold therefore provides a useful balance between candidate filtering and downstream accuracy.

RBF bandwidth. We further evaluate the RBF bandwidth multiplier γ_{bw}.

Table 8: Sensitivity to the RBF bandwidth multiplier γ_{bw}.

γ_{bw}	Video-MME				MLVU	LongVideoBench
	Short	Medium	Long	Overall		
0.5	75.78	**63.11**	**55.11**	**64.67**	74.01	61.71
1.0	**75.89**	62.89	54.78	64.52	**74.70**	**62.68**
2.0	**75.89**	62.56	54.56	64.33	74.47	61.71

The default $\gamma_{\text{bw}} = 1$ achieves the best performance on MLVU and LongVideoBench while remaining close to the best Video-MME configuration. Accordingly, the median-normalized bandwidth provides a consistent, dataset-independent default without benchmark-specific tuning.

D.3 Panel Compression Granularity

We examine the granularity of the compressed context representation while keeping the number of context units and focus units fixed at $C4 + F28$. A 1×1, 2×2, and 3×3 context unit therefore represents 1, 4, and 9 source frames, respectively.

The 2×2 configuration achieves the best benchmark-level performance on all three datasets. The 1×1 setting provides less temporal context, whereas the 3×3 setting covers more source frames at the cost of stronger per-frame compression. The superior performance of the 2×2 setting is consistent with a favorable balance between temporal coverage and preserved spatial detail.

D.4 Performance across Visual Budgets

We examine whether the benefit of context–focus decomposition persists under different total visual-unit budgets. For each budget, we compare uniform sampling with different context–focus allocations. Since each context panel compresses four source frames into one visual unit, we additionally report the number of source frames used to construct each input.

Table 9: Effect of panel compression granularity under the fixed $C4 + F28$ allocation. #CtxSrc denotes the number of source frames represented by the four context units, and #Src denotes the total number of source-frame observations represented by the final visual input.

Panel	#CtxSrc	#Src	Video-MME	MLVU	LongVideoBench
1×1	4	32	64.04	73.28	62.00
$\mathbf{2 \times 2}$	16	44	**64.52**	**74.70**	**62.68**
3×3	36	64	64.26	73.32	62.30

Table 10: Performance under different visual-unit budgets. #Src denotes the number of source frames used to construct each input.

#VU	#Src	Configuration	Video-MME				MLVU	LongVideoBench
			Short	Medium	Long	Overall		
8	8	Uniform Sampling	66.33	53.56	46.44	55.44	61.22	55.42
8	14	$C2 + F6$	71.44	56.56	51.22	59.74	70.52	**59.91**
8	20	$C4 + F4$	**73.00**	**57.22**	**52.00**	**60.74**	**70.88**	59.54
16	16	Uniform Sampling	71.78	56.44	50.33	59.52	65.09	55.03
16	22	$C2 + F14$	**73.33**	60.78	54.33	62.81	72.36	60.28
16	28	$C4 + F12$	72.56	**61.89**	**54.89**	**63.11**	**72.86**	62.23
16	40	$C8 + F8$	72.89	59.89	53.00	61.93	71.07	**62.60**
32	32	Uniform Sampling	74.22	60.11	52.33	62.22	68.26	55.94
32	44	$C4 + F28$	**75.89**	**62.89**	**54.78**	**64.52**	**74.70**	**62.68**

Context–focus configurations outperform Uniform Sampling at every tested visual-unit budget across all three benchmarks. The gains are particularly pronounced at lower budgets on Video-MME and MLVU, while remaining substantial on LongVideoBench across all budgets. These results show that the context–focus decomposition is effective beyond the default 32-VU setting and remains useful under tighter visual budgets.

D.5 Compressed Context with Existing Selectors

We further examine whether the benefit of CoFi can be explained simply by combining compressed temporal context with an existing frame selector. For each selector, we compare its original $F32$ configuration with a $C4 + F28$ variant that uses the same four uniformly constructed context panels as CoFi and allocates the remaining 28 visual units using the corresponding selector. This comparison separates the effect of introducing compressed context from the use of CoFi's context-conditioned focus acquisition.

Replacing four selector-chosen frames with compressed context preserves or improves performance in most selector–benchmark combinations. In particular, all four selectors improve on Video-MME and LongVideoBench, indicating that compressed context can complement standalone frame selection rather than merely consuming part of the focus budget. The effect is not universal, however: GIFT and WFS-SB decrease on MLVU, showing that compressed context alone does not guarantee an improvement over the original 32-frame selector.

Under the same $C4 + F28$ allocation, CoFi achieves the strongest result on all three benchmarks. This suggests that CoFi's performance is not explained simply by adding the same compressed context to the tested selectors. Together with the component ablation in Table 3, these results support coupling compressed context with subsequent context-conditioned sequential focus acquisition.

D.6 Context Observation Modeling

We further examine two design choices in context conditioning: how the reliability of compressed context observations should be modeled, and whether observations from the same rendered panel

Table 11: Effect of combining compressed context with existing frame selectors. For each selector, $F32$ denotes its original 32-frame configuration, while $C4 + F28$ replaces four focus units with the same four compressed context panels used by CoFi. Values in parentheses indicate the change from the corresponding $F32$ configuration; green and red denote improvements and decreases, respectively. † and $*$ follow the implementation-source notation in Table 1.

Selector	Allocation	Video-MME	MLVU	LongVideoBench
FOCUS†	$F32$	62.48	71.39	61.33
	$C4 + F28$	63.22 (+0.74)	72.17 (+0.78)	62.23 (+0.90)
GIFT*	$F32$	63.30	74.56	61.32
	$C4 + F28$	63.67 (+0.37)	74.06 (-0.50)	62.15 (+0.83)
BOLT†	$F32$	62.74	72.36	60.00
	$C4 + F28$	62.78 (+0.04)	72.45 (+0.09)	60.51 (+0.51)
WFS-SB†	$F32$	63.11	74.15	61.85
	$C4 + F28$	64.37 (+1.26)	73.51 (-0.64)	62.45 (+0.60)
CoFi	$C4 + F28$	**64.52**	**74.70**	**62.68**

should be modeled independently or jointly. All variants use the same $C4 + F28$ final visual input and differ only in the observation-noise model used during selection-time conditioning.

Fidelity-aware observation noise. We first examine how compression-induced degradation should be reflected in the reliability of context observations. No Fidelity treats all context source frames as noiseless observations ($\nu_c = 0$). Constant Noise preserves a video-level estimate of compression noise but removes cell-specific variation: for each video, we compute the median positive fidelity-derived noise across its context cells and assign this shared value to all observations. The default Fidelity-aware model instead retains the cell-specific noise values derived from representation fidelity.

Table 12: Effect of observation-noise modeling for compressed context under the default $C4 + F28$ configuration.

Observation Noise	Video-MME	MLVU	LongVideoBench
No Fidelity ($\nu_c = 0$)	64.22	74.47	62.23
Constant Noise	64.48	74.66	62.53
Fidelity-aware	**64.52**	**74.70**	**62.68**

As shown in Table 12, introducing a shared compression-aware noise level consistently improves over noiseless conditioning, indicating that compressed context is better modeled as an imperfect observation. Retaining cell-specific fidelity differences provides a further, though modest, improvement across all three benchmarks.

Independent-cell and joint-panel observations. The default CoFi estimates fidelity for each source frame independently at panel-cell resolution and uses a diagonal observation-noise matrix. We compare this with two alternatives that derive observation statistics from the actually rendered 2×2 panel. Joint panel, diagonal obtains cell-wise fidelities from the corresponding regions of the rendered panel while retaining diagonal observation noise. Joint panel, correlated additionally models correlations between the representation residuals of cells within the same panel, resulting in a block-diagonal observation covariance across panels.

These variants modify only the selection-time posterior used to choose focus frames. The downstream Video-LLM receives the same four rendered context panels and 28 selected focus frames in all settings.

As shown in Table 13, explicitly modeling the jointly rendered panel does not provide a consistent improvement over the independent-cell approximation. Independent-cell modeling performs best on

Table 13: Comparison of context observation models under the default $C4 + F28$ configuration.

Observation Model	Video-MME	MLVU	LongVideoBench
Independent-cell	**64.52**	**74.70**	62.68
Joint panel, diagonal	63.30	72.54	**62.90**
Joint panel, correlated	63.30	72.54	62.75

Video-MME and MLVU while remaining competitive on LongVideoBench, and introducing cross-cell correlations provides little additional benefit.

D.7 Representation and Backbone Robustness

We examine robustness to the representation space used for CoFi and to the downstream Video-LLM backbone.

Representation encoder. We compare CLIP-B/32 (Radford et al., 2021), SigLIP-SO400M (Zhai et al., 2023), and LongCLIP-L/14 (Zhang et al., 2024), while keeping the downstream Video-LLM, visual-unit budget, and all other CoFi components unchanged.

Table 14: Performance of CoFi with different frozen representation encoders.

Representation Encoder	Video-MME				MLVU	LongVideoBench
	Short	Medium	Long	Overall		
CLIP-B/32	74.67	61.11	54.33	63.37	73.09	61.85
SigLIP-SO400M	**75.89**	62.89	54.78	64.52	74.70	**62.68**
LongCLIP-L/14	75.67	**63.11**	**55.22**	**64.67**	**75.07**	61.18

CoFi maintains strong performance across all three representation encoders. LongCLIP provides small improvements on Video-MME and MLVU, whereas the default SigLIP performs best on LongVideoBench. We retain SigLIP as a single shared encoder across benchmarks, avoiding benchmark-specific encoder selection.

Representation-space alignment. CoFi constructs its residual covariance in a frozen representation space, which need not be identical to the visual representation used internally by the downstream Video-LLM. We therefore examine whether explicitly aligning the covariance representation with the downstream model improves performance.

Specifically, we replace the features from the standalone frozen encoder used to construct the state–trajectory relation structure and residual covariance with features from the fine-tuned vision tower of the evaluated LLaVA-Video checkpoint, while keeping the remaining CoFi pipeline and visual-unit budget unchanged. Table 15 compares the two settings.

Table 15: Effect of aligning the residual-covariance representation with the downstream Video-LLM.

Covariance Representation	Video-MME	MLVU	LongVideoBench
SigLIP-SO400M	**64.52**	**74.70**	62.68
LLaVA-tuned vision tower	63.74	73.78	**62.83**

Using the LLaVA-tuned vision tower changes approximately 20% of the selected focus observations, showing that the two representation spaces produce meaningfully different acquisition decisions. Nevertheless, explicit downstream alignment does not improve overall cross-benchmark performance: it lowers Video-MME and MLVU while providing only a small gain on LongVideoBench. These results support the use of a standalone frozen representation space without requiring explicit downstream alignment in the evaluated setting.

Downstream Video-LLM. We next examine whether the benefit of CoFi transfers across downstream Video-LLM architectures. In addition to the primary LLaVA-Video-7B backbone, we evaluate CoFi with Qwen2.5-VL-7B and InternVL2.5-8B while keeping the visual-unit budget and visual input construction setting unchanged. For each backbone, CoFi is compared with Uniform Sampling under the same 32-VU budget.

Table 16: Detailed results across downstream Video-LLM backbones. Δ denotes the improvement of CoFi over Uniform Sampling.

Video-LLM	Strategy	Video-MME				MLVU	LongVideoBench
		Short	Medium	Long	Overall		
Qwen2.5-VL-7B	Uniform	71.89	56.56	51.00	59.81	64.03	58.56
	CoFi	71.22	60.22	53.33	61.59	68.58	60.51
	Δ	-0.67	+3.66	+2.33	+1.78	+4.55	+1.95
InternVL2.5-8B	Uniform	73.11	58.22	51.89	61.07	66.42	59.46
	CoFi	74.00	60.56	54.22	62.93	69.83	64.70
	Δ	+0.89	+2.34	+2.33	+1.86	+3.41	+5.24

CoFi improves benchmark-level performance over Uniform Sampling on all three benchmarks for both Qwen2.5-VL-7B and InternVL2.5-8B. Together with the primary LLaVA-Video-7B results, these results support the transferability of CoFi across different downstream Video-LLM backbones.

D.8 Validation of Bounded Bandwidth Estimation

The long-video implementation in Appendix A.2.7 bounds the pairwise computation used to estimate the kernel bandwidths. We therefore evaluate whether bounded bandwidth estimation preserves the behavior obtained with exhaustive full-pair bandwidth estimation, from the resulting bandwidth statistics and kernel values to focus selection and downstream predictions.

Bandwidth and kernel approximation accuracy. We evaluate the $L_{\max} = 2048$ bandwidth approximation against exhaustive full-pair estimation on all 290 benchmark videos whose candidate sequences exceed 2048 frames. As shown in Table 17, the approximation yields mean absolute relative errors of only 0.0247% and 0.0175% for the state and trajectory bandwidths, respectively, with a kernel MAE of 6.31×10^{-5}.

Table 17: Comparison of bounded and exhaustive bandwidth estimation on videos with more than 2048 candidate frames. State Err. and Traj. Err. denote the mean absolute relative errors of the approximate state and trajectory bandwidths, respectively, relative to exhaustive full-pair estimates, averaged equally over videos. Kernel MAE denotes the mean absolute error between the kernel values obtained using exhaustive and approximated bandwidth statistics, evaluated over the full candidate sequence.

Dataset	# Videos	State Err. (%)	Traj. Err. (%)	Kernel MAE
Video-MME	219	0.0237	0.0163	6.17×10^{-5}
MLVU	43	0.0336	0.0260	8.18×10^{-5}
LongVideoBench	28	0.0187	0.0138	4.54×10^{-5}
Overall	290	0.0247	0.0175	6.31×10^{-5}

For the longest evaluated video with 32,551 candidate frames, the state and trajectory bandwidth errors remain 0.1032% and 0.0961%, respectively, with a kernel MAE of 3.27×10^{-4}. These results show that bounded estimation closely approximates exhaustive full-pair bandwidth estimation across datasets and video lengths.

Selection and prediction stability. We next examine whether the small numerical differences introduced by bounded bandwidth estimation propagate to the actual focus-selection behavior. For all 788 questions associated with videos whose candidate sequences exceed 2048 frames, we rerun

CoFi using exhaustive full-pair bandwidth estimation while keeping the remainder of the selection pipeline unchanged.

Table 18: Stability of CoFi focus selection when replacing the default bounded bandwidth estimation ($L_{\max} = 2048$) with exhaustive full-pair bandwidth estimation. Exact Set denotes the fraction of questions for which all 28 selected focus frames are identical, and Frame Overlap denotes the total number of shared selected focus frames divided by the total number of selected focus frames across all questions. Pred. Changes counts questions whose final Video-LLM prediction changes.

Dataset	# Questions	Exact Set	Frame Overlap	Pred. Changes
Video-MME	657	99.54%	99.93%	1
MLVU	47	97.87%	99.09%	0
LongVideoBench	84	100.00%	100.00%	0
Overall	788	99.49%	99.89%	1

As shown in Table 18, bounded and exhaustive bandwidth estimation produce exactly the same 28-frame focus set for 784 of 788 questions (99.49%). Across the 22,064 focus-frame selections aggregated over all questions, 99.89% are shared between the two settings. Only four questions exhibit any difference in the selected set, with three occurring on Video-MME, one on MLVU, and none on LongVideoBench.

Re-evaluating the four questions with different focus sets changes the final Video-LLM prediction in only one case. In that case, the bounded estimator produces the correct answer, whereas exhaustive bandwidth estimation produces an incorrect one. Consequently, exhaustive estimation leaves MLVU and LongVideoBench accuracy unchanged and changes Video-MME from 64.52 to 64.48 (-0.04 points).

Together, these results show that the $L_{\max} = 2048$ approximation closely preserves the bandwidth statistics, resulting kernel values, and downstream focus-selection behavior obtained with exhaustive full-pair bandwidth estimation, with no observed loss in downstream accuracy on the evaluated benchmarks.

D.9 Computational Efficiency and Scalability

Having established that bounded bandwidth estimation closely matches the behavior obtained with exhaustive full-pair estimation, we next quantify its computational benefit and examine the runtime and memory scalability of CoFi.

Efficiency of bounded bandwidth estimation. In GPU profiling on an NVIDIA RTX A6000, bounded bandwidth estimation provides median speedups of $3.02\times$, $10.34\times$, and $3.15\times$ on Video-MME, MLVU, and LongVideoBench, respectively. For the longest evaluated video with 32,551 candidate frames, bounded estimation reduces the bandwidth-estimation time from approximately 70 seconds to less than 0.4 seconds. Thus, the bounded estimator substantially reduces the cost of video-specific bandwidth estimation while the approximation analysis above shows that its effect on the resulting selection behavior is minimal.

Effect of candidate-pool restriction. We next examine the computational effect of restricting the admissible candidate pool using the threshold ρ. Both settings use cached 1-fps SigLIP representations, FP32 computation, one CPU thread, and the same sequential focus budget.

With $\rho = 0.8$, CoFi retains only 54–64% of the candidate frames while achieving slightly higher accuracy than $\rho = 1.0$ on all three benchmarks. Restricting the candidate pool also substantially reduces observed selection latency in our FP32 measurements. Together with the accuracy sensitivity results in Table 7, these results show that the TopMass-based candidate-pool restriction reduces computation without degrading downstream performance.

Scaling with video duration. Figure 4 shows that selection latency increases with video duration as the 1-fps candidate sequence grows. The separation between $\rho = 0.8$ and $\rho = 1.0$ becomes

Table 19: Effect of candidate-pool restriction on the retained candidate fraction and selection latency. Candidate Ret. denotes the average fraction of candidates retained.

Dataset	ρ	Candidate Ret.	Mean (ms)	Median (ms)
Video-MME	0.8	56.4%	213.84	33.53
	1.0	100%	1099.92	106.44
MLVU	0.8	54.0%	189.21	31.68
	1.0	100%	920.47	106.04
LongVideoBench	0.8	64.1%	109.68	34.82
	1.0	100%	411.13	107.91

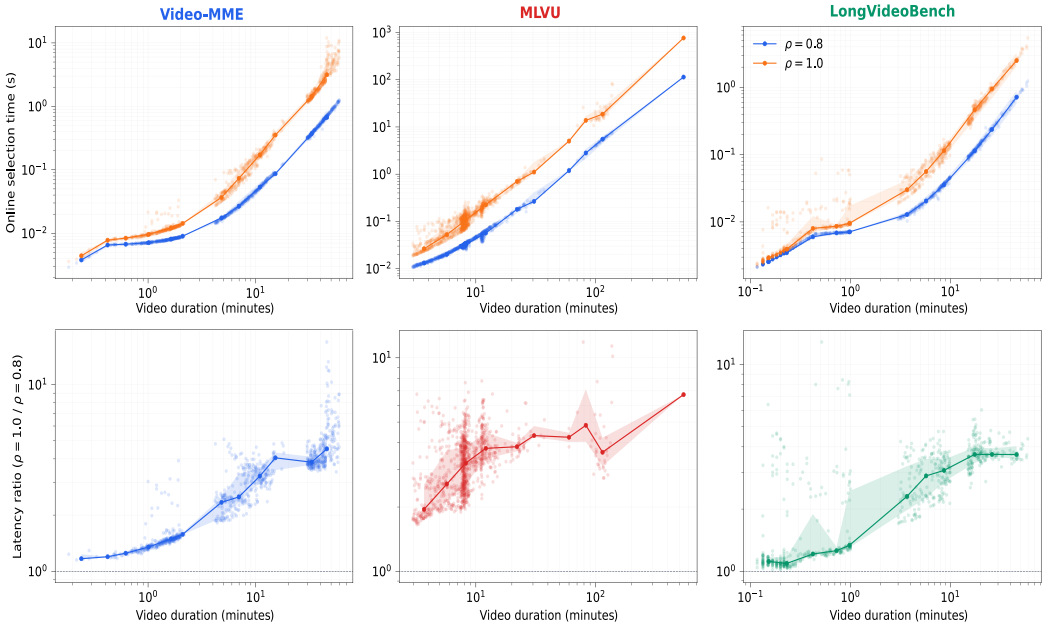


Figure 4: FP32 CoFi selection latency as a function of video duration. Each point represents the per-video median across its questions using cached 1-fps SigLIP features and one CPU thread. Lines and shaded regions show binned medians and interquartile ranges. The bottom row reports the latency ratio between $\rho = 1.0$ and $\rho = 0.8$.

larger for longer videos, consistent with candidate-pool restriction reducing repeated acquisition-score evaluation as the search space grows.

Memory scaling with video duration. Table 20 reports the empirical memory footprint of CoFi selection as video duration, and consequently the number of 1-fps candidate frames, increases. Memory usage grows substantially with video length as both the number of target frames N and the retained candidate count M increase, consistent with storing the $N \times M$ posterior cross-covariance block used for focus acquisition.

Across the grouped benchmark videos, median incremental peak memory increases from 4.7 MiB for videos shorter than one minute to 370.1 MiB for videos between 40 and 138.9 minutes, with a corresponding P90 of 720.6 MiB. The extreme 542.5-minute MLVU video contains 32,551 candidate frames and requires 7.76 GiB of additional CPU memory. These measurements complement the runtime scaling analysis by quantifying practical memory cost of the maintained posterior cross-covariance block.

Selection efficiency against baselines. Table 21 compares CoFi with representative training-free methods in terms of downstream accuracy and selection latency. All methods are profiled on CPU

Table 20: Empirical memory consumption of CoFi selection across video-duration groups. Memory denotes incremental peak CPU RSS above the pre-selection process baseline using cached FP32 visual representations.

Video Duration	# Videos	Median (MiB)	P90 (MiB)
<1 min	427	4.7	7.7
1–3 min	234	11.9	15.0
3–10 min	1,231	64.4	76.2
10–20 min	412	99.3	150.9
20–40 min	258	299.4	337.9
40–138.9 min	212	370.1	720.6
542.5 min	1	7,946 MiB	

under the same evaluation environment. The reported latency measures the visual-input selection stage after candidate representations have been obtained.

Table 21: Accuracy and selection latency of representative training-free visual input construction methods. Acc. denotes downstream accuracy, and Lat. denotes the median selection latency after candidate representations are obtained, measured in seconds. All methods are profiled on CPU under the same evaluation environment. † and $*$ follow the implementation-source notation in Table 1.

Method	Video-MME		MLVU		LongVideoBench	
	Acc.	Lat. (s)	Acc.	Lat. (s)	Acc.	Lat. (s)
GIFT*	63.30	0.006	74.56	0.006	61.32	0.006
WFS-SB†	63.11	0.021	74.15	0.022	61.85	0.022
MDP^{3†}	65.11	20.61	74.33	26.62	61.26	24.96
CoFi Sequential	64.52	0.034	74.70	0.032	62.68	0.035
CoFi Block-2	64.41	0.024	74.66	0.022	62.68	0.024

CoFi incurs more selection overhead than lightweight selectors such as GIFT and WFS-SB, but remains in the tens-of-milliseconds range. In contrast, MDP^3 requires tens of seconds under the same CPU profiling setup. Compared with MDP^3, CoFi achieves higher MLVU and LongVideoBench accuracy with substantially lower selection latency, while remaining within 0.59 points on Video-MME.

Block-wise focus acquisition. The default CoFi performs fully sequential acquisition. Given the current residual covariance $\mathbf{C}^{(k)}$, it computes the acquisition scores, selects one focus observation, updates the residual structure, and recomputes the scores before selecting the next observation. Thus, each newly acquired observation can influence the following selection.

To reduce the number of sequential recomputation rounds, we additionally evaluate a block-wise variant that selects the top-b candidates under the current acquisition scores before refreshing the residual covariance.

Block-wise acquisition reduces the number of score-recomputation rounds from B_F to approximately $\lceil B_F/b \rceil$ using the same acquisition score before refreshing the residual covariance. Increasing the block size reduces selection latency, while accuracy varies across benchmarks; in particular, Block-2 lowers median latency by approximately 30% with at most a 0.11-point accuracy change across the three benchmarks. We retain $b = 1$ as the default to preserve fully sequential focus acquisition, while block-wise acquisition provides an optional efficiency–accuracy trade-off when lower latency is preferred.

Table 22: Effect of block size on CoFi accuracy and selection latency. Acc. denotes downstream accuracy, and Lat. denotes selection latency in milliseconds. Block size 1 corresponds to the default fully sequential acquisition.

Block Size	Video-MME		MLVU		LongVideoBench	
	Acc.	Lat. (ms)	Acc.	Lat. (ms)	Acc.	Lat. (ms)
1 (Sequential)	64.52	33.53	74.70	31.68	62.68	34.82
2	64.41	23.58	74.66	22.34	62.68	24.34
3	64.19	22.64	74.61	21.34	62.60	23.63
4	64.59	21.41	74.98	20.21	61.93	22.08
5	64.56	21.01	74.47	19.97	62.45	21.80
6	64.41	20.78	74.93	19.63	62.08	21.41
7	64.07	20.43	74.38	19.32	62.68	21.31

E Additional Task-Wise and Qualitative Analysis

E.1 Task-wise Analysis on MLVU

Figure 5 compares the task-wise performance of CoFi with Uniform Sampling, AKS, and Video Panels on the multiple-choice portion of MLVU.

AKS and Video Panels exhibit complementary strengths across different task types. AKS performs relatively well on *Action Count* but is weaker on *Action Order*, whereas Video Panels shows the opposite tendency, excelling on *Action Order* but performing less effectively on *Action Count*. CoFi performs well on both categories, achieving the best result on *Action Count* while remaining close to Video Panels on *Action Order*.

Beyond these two categories, CoFi achieves the best performance on *Topic Reasoning*, *Needle QA*, *Ego Reasoning*, and *Plot QA*, while remaining competitive on *Anomaly Recognition*. Overall, CoFi remains competitive across task types with different evidence requirements, indicating that its performance is not confined to a single type of long-video question.

E.2 Failure-Case Analysis

Figure 6 presents representative failure cases under the default 32-VU setting ($C4 + F28$). These examples illustrate that failures can occur even when relevant visual evidence is present in the constructed input.

In Fig. 6a, the Amazon stock-price example shows that CoFi retrieves relevant visual content containing the salient "$600/share" peak, but the downstream model predicts this value instead of the correct value for the dot-com bubble period. In the repeated-action counting example, CoFi includes observations of the shredding-paper action, yet the downstream model overestimates how many times the action occurs. Uniform Sampling, in contrast, produces the correct answer although the displayed Uniform samples do not clearly expose the corresponding action.

In Fig. 6b, CoFi again captures relevant observations but fails in their interpretation or temporal organization. For the park-yoga example, CoFi captures the target scene but incorrectly identifies the time of day as morning, whereas Uniform Sampling predicts the correct answer although the displayed Uniform samples do not clearly show the target scene. In the temporal-ordering example, CoFi identifies observations related to the queried events and correctly resolves the latter part of the sequence, but confuses the ordering of the first two events.

Overall, these cases separate *evidence acquisition* from *evidence utilization*. CoFi can expose task-relevant visual observations, yet the downstream Video-LLM may still misinterpret the evidence, miscount repeated events, or misresolve their temporal order. Conversely, a correct prediction from Uniform Sampling does not necessarily imply that the decisive visual evidence was explicitly captured. Future work could incorporate reasoning-aware feedback into the acquisition process, linking representation-space selection more directly to downstream evidence utilization.

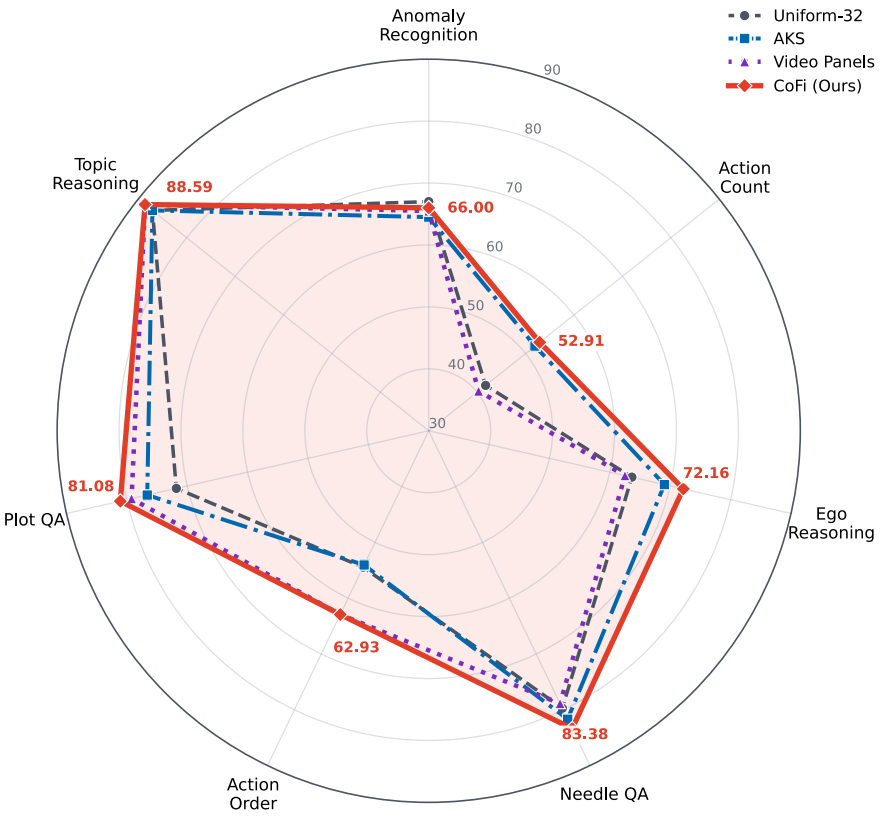


Figure 5: Task-wise accuracy on the multiple-choice portion of MLVU for CoFi, Uniform Sampling, AKS, and Video Panels.

Question: In this video, how many times does the scene of the 'shredding paper' action appear in total?

Uniform Sampling (32 VUs)



Answer: 2 ✓

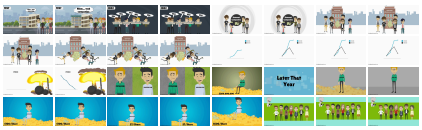
CoFi (C4+F28)



Answer: 3 ✗

Question: What was Amazon's stock price at the peak of the dot-com bubble?

Uniform Sampling (32 VUs)



Answer: $100/Share ✓

CoFi (C4+F28)



Answer: $600/Share ✗

(a) Failures in repeated-action counting and factual-value identification.

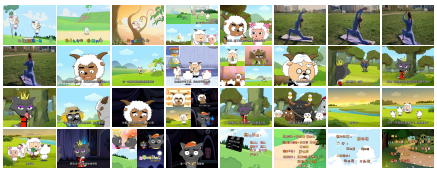
Question: What time of day is it when the young girl in a tracksuit is doing yoga in the park?

Uniform Sampling (32 VUs)



Answer: Sunset ✓

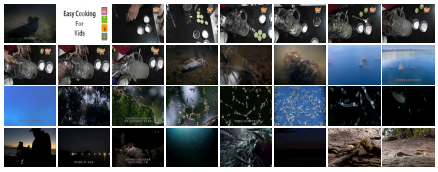
CoFi (C4+F28)



Answer: Morning ✗

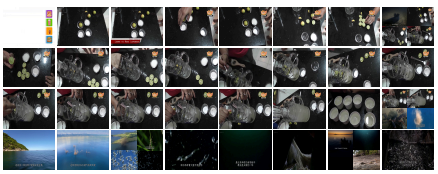
Question: Arrange the following events from the video in the correct chronological order:
(1)Water is added to the cup and more limes are squeezed in by hand;
(2)The small bowls of salt are arranged and limes are sliced in halves;
(3)The cup is stirred with more water and a set of cups filled with the refreshment are seen;
(4)The limes are juiced into a cup using a hand held press.

Uniform Sampling (32 VUs)



Answer: 2→4→1→3 ✓

CoFi (C4+F28)



Answer: 4→2→1→3 ✗

(b) Failures in visual-state interpretation and temporal ordering.

Figure 6: Representative failure cases under the default 32-VU setting ($C4 + F28$). Each example compares Uniform Sampling and CoFi under the same visual-unit budget. Green and red denote correct and incorrect predictions, respectively.