TEMPCORE: Are Video QA Benchmarks Temporally Grounded? A Frame Selection Sensitivity Analysis and Benchmark

Anonymous ACL submission

Abstract

Vision-language models (VLMs) can ingest only a limited number of video frames, making frame selection a practical necessity. But do current Video QA benchmarks genuinely require temporal frame selection, or can most questions be answered regardless of which frames are shown? We introduce **Frame Selection Sensitivity** (FSS), a per-sample diagnostic that measures how much VLM accuracy changes when the most relevant frames are replaced with the least relevant ones. Across six benchmarks and eight VLMs, we find that a large majority of samples are frame-agnostic: only a minority are genuinely sensitive to frame choice. Combining FSS with a Language Independence Score (LIS) reveals that merely 8–33% of samples are *Temporally Sensitive*. We construct TEMPCORE, compact evaluation subsets that isolate these temporal samples from existing benchmarks, and will release code and per-sample annotations upon publication.

1 Introduction

Large language models (LLMs) have transformed natural language processing, and their extension to vision-language models (VLMs) has brought comparable advances to image understanding (Li et al., 2025; Zhu et al., 2025; Bai et al., 2025b). These developments have spurred growing interest in applying VLMs to *video* understanding, where temporal dynamics—event ordering, causal transitions, and changes over time—are fundamental.

Extending VLMs to video processing, however, is fundamentally constrained by input capacity: videos may contain thousands of frames, but current models can process only around 32 at a time (Lin et al., 2024; Tang et al., 2025). The standard approach is to rank frames by relevance using encoders such as SigLIP (Zhai et al., 2023; Tschannen et al., 2025) or CLIP (Radford et al., 2021), and retain the top-K frames (Tang et al., 2025; Himakunthala et al., 2023; Shen et al., 2025). Accordingly, a growing line of work explores more sophisticated selection strategies—including combinatorial ranking, spatiotemporal saliency, hierarchical clustering, and generative index prediction (Yu et al., 2025; Hu et al., 2025; Wang et al., 2025; Yao et al., 2025)—all based on the premise that selecting the right frames is critical.

While these methods report gains on standard benchmarks, a closer look complicates the picture. On some benchmarks, simple uniform sampling performs comparably to learned selection (Brkic et al., 2025), and a growing literature suggests that VLMs are largely "temporally blind," exhibiting limited sensitivity to event ordering and duration (Upadhyay et al., 2025; Jung et al., 2025; Cai et al., 2024). If temporally insensitive models can still achieve high scores, the benchmarks themselves may not genuinely require temporal reasoning—and if most questions can be answered from a single frame or from text alone, then frame selection is irrelevant regardless of the method. This tension raises two intertwined questions:

1. *Does key-frame selection actually improve Video QA performance?*
2. *Do current benchmarks genuinely require temporal reasoning?*

To address these questions, we examine whether encoder-based frame relevance translates into meaningful differences in downstream VLM accuracy. Concretely, we score every video frame with a SigLIP encoder and compare VLM accuracy when the model receives the most relevant (MAXPROB) versus least relevant (MINPROB) frames. This yields **Frame Selection Sensitivity** (FSS), a per-sample diagnostic that partitions each benchmark sample into *Key* (temporally sensitive), *Non-key* (frame-agnostic), or *Anti-key* (hurt by "relevant" frames). Applying FSS to six benchmarks with 8 VLMs ranging from 256M to 8B parameters—

the largest systematic study of frame selection to date—we answer:

Answer to Q1: Frame selection works, but only on long videos. On long-video benchmarks, MaxProb consistently outperforms MinProb, while on short-video benchmarks, the effect is near zero. Our approach uses a general-purpose image-text encoder (SigLIP) in question-answer (QA-mode) scoring, which leverages full question-answer context for more accurate temporal classification; encoder improvements—such as question-aware models—could narrow this gap further. The large oracle headroom confirms that substantial temporal signal exists in the video but remains unexploited by current selection methods.

Answer to Q2: Most existing benchmarks are not temporally grounded. Only 6–30% of samples are Key; the vast majority are Non-key or Anti-key. A complementary **Language Independence Score** (LIS) reveals that only 8.3–32.8% of samples are *Temporally Sensitive*—both vision-dependent and frame-sensitive (§7.1). No-frame baselines reach 20–50% accuracy on language priors alone, and aggregate scores are dominated by Non-key samples, allowing models with strong language priors to appear competitive even when they fail on frame-sensitive temporal questions.

Based on this analysis, we construct TempCore, temporally grounded evaluation subsets that retain only these samples (§7), providing a focused measure of temporal reasoning ability.

Contributions.

1. We propose FSS, a per-sample diagnostic that quantifies temporal sensitivity, and conduct a systematic frame selection study (§3, §5).
2. We introduce a two-dimensional FSS$\times$LIS taxonomy that reveals only 8.3–32.8% of samples are *Temporally Sensitive* (§6, §7).
3. We construct TempCore, temporally grounded evaluation subsets extracted from existing benchmarks without manual re-annotation (§7).

2 Related Work

Large language models and multimodal LLMs. The expansion of large language models (LLMs) has enabled strong performance across various general-purpose language tasks (Achiam et al., 2023; Grattafiori et al., 2024; Yang et al., 2024; Gemma Team et al., 2024). Building upon this foundation, research has extended such capabilities to other modalities by leveraging the knowledge of LLMs (Alayrac et al., 2022; Liu et al., 2023; Chu et al., 2024; Wang et al., 2024; Zhu et al., 2025). Video MLLMs, in particular, have achieved strong performance in video question answering, captioning, and summarization (Zhang et al., 2023; Lin et al., 2024; Zhu et al., 2024; Maaz et al., 2024; Shen et al., 2025; Zhang et al., 2025).

Key frame selection. Key frame selection reduces the computational cost of long-video understanding by identifying the most relevant frames for a given query. Encoder-based methods score individual frames via vision-language similarity: Tang et al. (2025) select frames that are both query-relevant and visually diverse to avoid redundancy, Yu et al. (2025) rank frame subsets by how well they complement each other rather than scoring frames independently, and Hu et al. (2025) leverage multimodal LLM prompting to derive spatial importance and temporal selection signals for frame ranking. A second line of work explores structured selection strategies: Wang et al. (2025) build a hierarchical tree of visual segments that enables query-adaptive retrieval from long videos, Buch et al. (2025) learn to dynamically adjust how many frames to select per query instead of using a fixed budget, and Sun et al. (2025) optimize frame orderings using a list-wise ranking objective that considers the full sequence jointly. These methods *improve* selection; we *diagnose* when selection matters and construct evaluation subsets that isolate temporal ability.

Temporal diagnostics & benchmarks. Several studies document that VLMs struggle with temporal reasoning: near-zero accuracy on ordering tasks (Upadhyay et al., 2025), inconsistency across equivalent temporal queries (Jung et al., 2025), and failures in fine-grained temporal understanding (Cai et al., 2024). Complementary work shows that many Video QA questions can be solved from a single frame or without temporal reasoning: Cores et al. (2025) show that most MVBench tasks can be solved without temporal reasoning, with single-frame and text-only baselines remaining competitive; Lei et al. (2023) demonstrate similar single-frame biases across multiple video-language benchmarks. On the diagnostic side, counterfactual interventions (Niu et al., 2021), text-bias reduction (Kulgod et al., 2026), confidence intervals (Du,

2025), and sampling strategy comparisons (Brkic et al., 2025) have been proposed to isolate specific biases. Dedicated temporal benchmarks take complementary approaches. TempCompass (Liu et al., 2024) constructs synthetic test cases by applying video-level manipulations—reversal, spatial concatenation, and temporal concatenation—to probe whether models truly perceive temporal order rather than relying on static appearance. TemporalBench (Cai et al., 2024) instead enlists expert annotators to label fine-grained temporal events within each clip, producing densely annotated samples that test ordering, duration, and causal sequencing. TEMPCORE differs on two fronts: (i) it provides a fully automated, model-based diagnostic pipeline that requires no manual annotation or synthetic video generation, and (ii) it draws from six widely adopted benchmarks covering egocentric, perceptual, and causal-reasoning domains, yielding a compact yet diverse evaluation set.

3 Diagnostic Framework

3.1 Frame Scoring

Given a video $V = (f_1, f_2, \ldots, f_L)$ of L frames uniformly extracted at 1–3 FPS, and a text query t (question, or concatenation with answer), we compute a per-frame relevance score using a pretrained SigLIP vision-language encoder (Zhai et al., 2023):

$$s(f_i, t) = \sigma\big(\cos\big(E_V(f_i),\ E_T(t)\big)\big)\,, \qquad (1)$$

where E_V and E_T are the vision and text encoders, $\cos(\cdot,\cdot)$ denotes cosine similarity, and σ is the sigmoid function.

Text modes. We employ two modes for t:

- Q-mode: t is the question text.
- QA-mode: t is the concatenation of the question and the correct answer.

QA-mode requires oracle answer labels and serves as an upper bound on encoder-based selection.

3.2 Frame Selection Strategies

Using the frame scores $\{s(f_i, t)\}_{i=1}^{L}$, we define five different strategies:

MAXPROB. Select the top-K scoring frames:

$$\mathcal{F}^{+} = \text{top-K}\big(f_1, \ldots, f_L;\ s(\cdot, t)\big). \qquad (2)$$

MINPROB. Select the bottom-K scoring frames:

$$\mathcal{F}^{-} = \text{bot-K}\big(f_1, \ldots, f_L;\ s(\cdot, t)\big). \qquad (3)$$

This adversarial baseline serves as the diagnostic counterpart to MAXPROB.

Uniform. Select K frames uniformly distributed over the video timeline. This is the standard baseline in most deployed systems.

Window oracle. Slide a window of K frames with stride S across the video, run the VLM on each window, and report the best accuracy. When the total number of windows exceeds a budget $W_{\max}$ (default 50), we uniformly subsample $W_{\max}$ windows to bound computation. This measures the theoretical upper bound achievable by any temporal selection strategy given window size K.

No-frame. Replace all input frames with a uniform gray dummy image. The VLM processes the question but receives no visual content.

3.3 Frame Selection Sensitivity (FSS)

To measure how much each sample depends on receiving the right frames, we define *Frame Selection Sensitivity* (FSS). Let $\mathcal{M}$ denote the set of models, and let $\text{acc}^{+}_{m,i}$ and $\text{acc}^{-}_{m,i}$ be the binary correctness (1 or 0) of model m on sample i when given MAXPROB and MINPROB frames, respectively.

$$\text{FSS}_i = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \big(\text{acc}^{+}_{m,i} - \text{acc}^{-}_{m,i}\big). \qquad (4)$$

$\text{FSS}_i > 0$ indicates that the sample benefits from relevant frames; $\text{FSS}_i < 0$ indicates that relevant frames *hurt*—a visual appearance bias.

Sample categories. Given a threshold τ (default $\tau = 0.15$), we classify each sample as:

- *Key* ($\text{FSS}_i > \tau$): temporally sensitive; benefits from correct frame selection.
- *Non-key* ($|\text{FSS}_i| \leq \tau$): frame-agnostic; answered regardless of frame choice.
- *Anti-key* ($\text{FSS}_i < -\tau$): appearance-biased; "relevant" frames are detrimental.

This classification is an *operational* definition of temporal sensitivity: it captures whether frame selection affects model behavior, which is a necessary—but not necessarily sufficient—condition for a question to require temporal reasoning. FSS-based pruning retains samples with $|\text{FSS}| > \tau$, yielding an evaluation subset in which frame selection is consequential.

Choice of τ. We select $\tau = 0.15$ via a systematic search validated against human temporal judgments (F1 $= 0.845$). Details and ablations over τ are in Section B.2.

Default configuration. We adopt QA-mode as the primary FSS scoring mode, as it more accurately identifies temporally sensitive questions by leveraging full question-answer context. Since FSS is an analytical diagnostic, maximizing accuracy takes priority over label-free operation.

3.4 Language Independence Score (LIS)

Let $\text{correct}^{\text{nf}}_{m,i}$ be the binary correctness of model m on sample i under the no-frame strategy.

$$\text{LIS}_i = 1 - \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \text{correct}^{\text{nf}}_{m,i}. \quad (5)$$

$\text{LIS}_i = 1$ when no model answers correctly without meaningful visual input, indicating maximal vision dependence; $\text{LIS}_i = 0$ when all models answer correctly, indicating that the item is solvable from language priors alone.

4 Experimental Setup

4.1 Models

We evaluate eight VLMs from five families: LLaVA-OneVision (Li et al., 2025), VideoLLaMA3 (Zhang et al., 2025), SmolVLM2 (Marafioti et al., 2025), Qwen3-VL (Bai et al., 2025a), and InternVL3 (Zhu et al., 2025), spanning 256M to 8B parameters. All models receive $K = 32$ frames in a zero-shot multiple-choice setting; ablations over $K \in \{16, 32, 48\}$ appear in Section C.1. We compute frame relevance scores (Equation (1)) with SigLIP-400M (Zhai et al., 2023) as the default; ablations with SigLIP2-400M and CLIP ViT-L/14 appear in Section C.2. See Table 5 for full details.

4.2 Benchmarks

We use six Video QA benchmarks grouped by video length. *Short* ($\leq$3 min): MVBench (Li et al., 2024), NExTQA (Xiao et al., 2021), and EgoSchema (Mangalam et al., 2023). *Long* (>5 min): MLVU (Zhou et al., 2025), LongVideoBench (Wu et al., 2024), and Video-MME (Fu et al., 2025). Detailed information and full statistics appear in Table 6.

Roadmap. We first examine whether frame selection affects VLM accuracy (§5), then characterize how temporally grounded current benchmarks are (§6), and finally construct TEMPCORE, temporally grounded evaluation subsets (§7).

Benchmark	Q-mode	QA-mode	Δ
Short-video			
MVBench	47.1	49.4	+2.3
NExTQA	57.8	59.9	+2.0
EgoSchema	32.8	38.1	+5.4
Long-video			
MLVU	53.9	57.0	+3.2
LongVideoBench	49.2	45.6	-3.6
Video-MME	44.7	48.4	+3.6
Average	47.6	49.7	+2.2

Table 1: **QA-mode consistently outperforms Q-mode.** MaxProb accuracy (%) averaged across 8 VLMs. Δ = QA-mode − Q-mode. LongVideoBench is the sole exception, where descriptive questions already provide sufficient visual grounding.

5 Does Frame Selection Work?

5.1 RQ1: Are Vision Encoders Robust to Interrogative Cues?

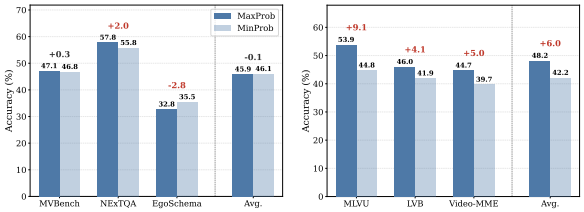
Vision-language similarity is an intuitive basis for frame scoring, yet VQA queries are inherently interrogative: they contain words like "who," "what," and "how many" that carry little visual grounding. Since encoders such as SigLIP are trained on declarative image–caption pairs, their similarity estimates may be poorly calibrated for question-form inputs. We test this by comparing Q-mode (question only) with QA-mode (question + ground-truth answer), where the latter supplies a declarative phrase that better matches the encoder's training distribution. Table 1 summarizes benchmark-averaged accuracy.

QA-mode outperforms Q-mode on five of six benchmarks, with an average gain of +2.2% (Table 1), indicating that encoders are weakened by interrogative cues. The exception is LongVideoBench, whose questions are three to four times longer *in word count* than those of other benchmarks and already contain rich visual descriptions, making Q-mode sufficient on its own (see Section C.3 for statistics and examples).

Figure 1 further illustrates this vulnerability via Grad-CAM (Selvaraju et al., 2017) attention maps: when scoring with question text, attention is diffuse across the frame, whereas answer text concentrates attention on the task-relevant region. This confirms that ambiguous, interrogative queries yield noisier similarity estimates.

Takeaway. These results suggest that text-based frame selection methods must be robust to interrogative and ambiguous input cues. Developing encoders that handle question-form queries as effectively as declarative captions remains an impor-

Figure 1: **Grad-CAM attention maps for question-text vs. answer-text scoring.** Question-text attention is diffuse, while answer-text attention concentrates on the relevant region, illustrating the noise introduced by interrogative cues.



(a) Short-video (b) Long-video

Figure 2: **MaxProb vs. MinProb accuracy (8-model average).** Frame selection yields negligible effect on short-video benchmarks but consistent gains on long-video benchmarks. LVB denotes LongVideoBench.

tant direction for future work.

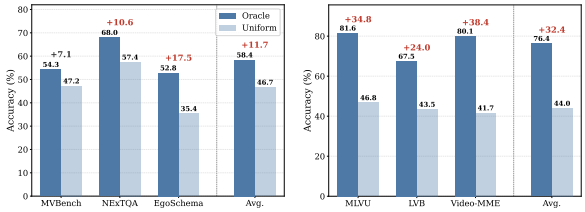
5.2 RQ2: MaxProb vs. MinProb

Despite the interrogative-cue limitation identified above, does vision-language similarity still provide an effective frame selection signal? Figure 2 compares MaxProb and MinProb accuracy (8-model average) across six benchmarks (full per-model results in Table 16).

On long-video benchmarks, MaxProb consistently outperforms MinProb, with an average gain of $+6.1\%$ across models. On short-video benchmarks (MVBench, NExTQA), differences are small and sometimes reversed: in short clips, most $K = 32$ windows already contain the relevant content, so the scorer's ranking provides little additional benefit.

EgoSchema shows a reversed pattern: MaxProb *underperforms* MinProb for 6 of 8 models. This is the only benchmark where Anti-key samples outnumber Key samples (Key/AK ratio = 0.48, Table 3), suggesting that SigLIP's top-scoring frames are frequently misleading on this benchmark.

Takeaway. Vision-language similarity is an effective frame selection signal for long videos, where relevant content is temporally sparse. On short videos, the benefit vanishes—temporal information demand is inherently low, and any $K = 32$ window



(a) Short-video (b) Long-video

Figure 3: **Oracle gaps reveal large untapped headroom for frame selection.** Uniform-sampling vs. window-oracle accuracy (model average). Long-video gaps far exceed short-video gaps, confirming that temporal frame selection is most consequential for extended videos. LVB denotes LongVideoBench.

already covers most of the relevant content.

5.3 RQ3: Oracle Gap

How much room remains for better frame selection? Figure 3 compares window-oracle accuracy—the best accuracy achievable across all K-frame sliding windows—against a uniform-sampling baseline (full per-model results in Table 17).

The gaps are substantial: long-video benchmarks average $+29.4\%$ while short-video benchmarks average $+11.7\%$. However, the oracle selects the single best window per sample, and different windows may elicit correct answers through distinct reasoning paths, potentially inflating the gap beyond pure temporal content.

To control for this concern, Table 2 decomposes oracle gaps by FSS category. Key samples exhibit consistently larger gaps than Non-key samples on five of six benchmarks, confirming that the more temporal a sample is, the more headroom remains for improved frame selection.

Takeaway. Large oracle gaps indicate substantial room for better frame selection, and this gap is concentrated in temporally sensitive samples. Longer benchmarks with higher temporal purity stand to benefit the most from advances in frame retrieval.

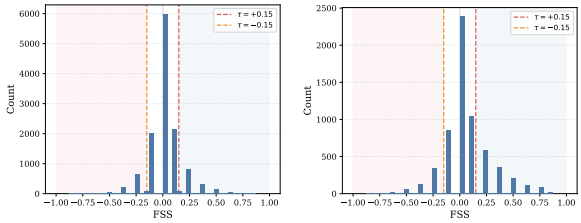
6 How Temporal Are Video QA Benchmarks?

6.1 FSS Distributions

FSS is computed from MaxProb and MinProb accuracy (defined in §3). Figure 4 shows the distribution of FSS scores across samples in each benchmark. The overwhelming majority of samples cluster near zero, indicating that most Video QA items are insensitive to frame selection.

Benchmark	Split	Uniform	Oracle	Gap
Short-video				
MVBench	All	47.2	54.3	+7.1
	Key	49.5	66.8	+17.4
	Non-key	46.8	51.1	+4.3
	Anti-key	45.1	55.5	+10.4
NExTQA	All	57.4	68.0	+10.6
	Key	51.3	70.1	+18.7
	Non-key	59.4	67.8	+8.3
	Anti-key	49.3	64.4	+15.1
EgoSchema	All	35.4	52.8	+17.5
	Key	39.1	63.2	+24.1
	Non-key	33.9	49.5	+15.6
	Anti-key	44.6	60.7	+16.1
Long-video				
MLVU	All	46.3	77.2	+30.9
	Key	41.1	80.9	+39.8
	Non-key	49.5	74.0	+24.4
	Anti-key	50.0	79.5	+29.5
LongVideoBench	All	43.5	67.5	+24.0
	Key	42.9	66.1	+23.2
	Non-key	42.8	66.8	+24.0
	Anti-key	66.0	75.3	+9.3
Video-MME	All	40.8	74.1	+33.3
	Key	40.3	82.2	+41.9
	Non-key	41.1	70.9	+29.9
	Anti-key	42.3	73.7	+31.4

Table 2: **Key samples exhibit the largest oracle gaps, confirming FSS validity.** Oracle gap (%) decomposed by FSS category. Key samples consistently show larger gaps than Non-key on five of six benchmarks, demonstrating that FSS-classified Key samples are precisely those where temporal frame selection is consequential.



(a) Short-video (b) Long-video

Figure 4: **FSS scores concentrate near zero, confirming non-temporal dominance.** Aggregated FSS distributions for short-video and long-video benchmarks. Dashed lines mark classification thresholds ($\tau = \pm 0.15$). The heavy mass near zero shows that frame selection has negligible effect on the majority of samples. Per-benchmark breakdowns are in Section D.1.

Benchmark	Key (%)	Anti-key (%)	Key/AK
Short-video			
MVBench	9.2	8.5	1.08
NExTQA	12.7	8.0	1.59
EgoSchema	5.6	11.6	0.48
Long-video			
MLVU	**30.3**	11.1	2.73
LongVideoBench	16.3	8.1	2.01
Video-MME	18.1	7.6	2.38

Table 3: **Key fractions vary widely across benchmarks ($\tau = 0.15$).** "Key/AK" is the ratio of Key to Anti-key samples.

Quantitative summary. Table 3 reports Key and Anti-key fractions for each benchmark. EgoSchema has the lowest Key fraction (5.6%), while MLVU reaches a Key fraction of 30.3%. Other short-video benchmarks fall in between (9–13%), reflecting that even short clips contain some frame-sensitive samples, but relatively few.

6.2 Implications for Benchmark Design

Our analysis calls for a reinterpretation of current Video QA benchmarks. Only a small fraction of samples in any benchmark are temporally sensitive to frame selection; the remainder are resolved by language priors, visual appearance, or other non-temporal cues. This does not invalidate existing benchmarks—appearance and semantic comprehension are important capabilities—but it underscores that *temporal* reasoning performance should be reported separately. FSS-based pruning (retaining $|\text{FSS}| > \tau$) provides a lightweight, model-agnostic mechanism to do so. We address this by combining FSS with the Language Independence Score (LIS) to construct TEMPCORE, temporally grounded evaluation subsets.

7 TEMPCORE

FSS captures sensitivity to frame selection but cannot distinguish genuinely temporal samples from those solvable by language priors alone—a high-FSS question may simply contain answer-leaking text cues. To filter out such cases, we combine FSS with the Language Independence Score (LIS, introduced in §3.4) and construct TEMPCORE, temporally grounded evaluation subsets extracted from existing benchmarks without manual re-annotation.

7.1 Two-Dimensional Taxonomy

Combining LIS and FSS yields a three-tier taxonomy (with Temporally Sensitive further subdivided into two subcategories). Given thresholds τ (FSS, default 0.15) and λ (LIS, default 0.50):

- **Temporally Sensitive** (LIS $> \lambda$, $|\text{FSS}| > \tau$): vision-dependent *and* frame-sensitive. Subdivided into:
  - *Temporally Grounded* (FSS $> \tau$): benefits from relevant frames; the question genuinely requires the temporal segment.
  - *Visual Bias* (FSS $< -\tau$): harmed by

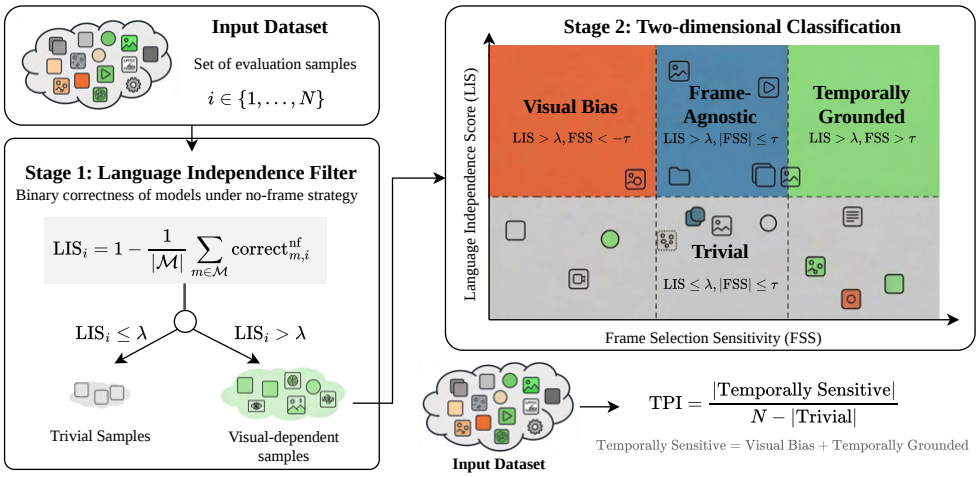


Figure 5: **Overview of the TempCore construction pipeline.** Stage 1 filters out Trivial samples (solvable from language priors, LIS $\leq \lambda$). Stage 2 classifies the remaining vision-dependent samples along the FSS axis into three categories: Temporally Grounded (FSS $> \tau$), Frame-Agnostic ($|\mathrm{FSS}| \leq \tau$), and Visual Bias (FSS $< -\tau$). The Temporal Purity Index (TPI) summarizes the fraction of vision-dependent samples that are temporally sensitive.

selected frames, which capture visually matching but temporally wrong content.

- **Frame-Agnostic** (LIS $> \lambda$, $|\mathrm{FSS}| \leq \tau$): requires vision but is frame-choice insensitive.
- **Trivial** (LIS $\leq \lambda$): solvable from text alone.

Including Anti-key samples in the Temporally Sensitive is motivated by human evaluation: $|\mathrm{FSS}|$ correlates more strongly with human temporal scores than signed FSS ($\rho = 0.528$ vs. 0.495), and Anti-key samples receive higher temporal ratings than Non-key samples (details in Section 7.3).

LIS–FSS orthogonality. A key design requirement is that LIS and FSS capture distinct properties. Spearman rank correlations between the two scores are near zero across all six benchmarks ($|\rho| < 0.10$), establishing that language independence and frame sensitivity are empirically orthogonal dimensions (see Figure 13).

7.2 Benchmark Quality Profiles

We summarize each benchmark with the **Temporal Purity Index** (TPI):

$$\mathrm{TPI} = \frac{|\text{Temporally Sensitive}|}{N - |\text{Trivial}|}, \tag{6}$$

where Temporally Sensitive $=$ Key $\cup$ Anti-key (i.e., $|\mathrm{FSS}| > \tau$ and LIS $> \lambda$), and Trivial $= \{i : \mathrm{LIS}_i \leq \lambda\}$ encompasses all samples solvable from language priors. TPI thus measures the fraction of *vision-dependent* samples for which temporal frame selection is consequential.

Table 4 reports the category breakdown. Several patterns emerge: (i) **Temporally Sensitive** fractions range from 8.3% (LongVideoBench) to 32.8% (MLVU), indicating that even long-video benchmarks contain $<33\%$ temporally sensitive, vision-dependent samples. (ii) **Frame-Agnostic** samples dominate EgoSchema (58.8%) and LongVideoBench (55.1%), suggesting these benchmarks heavily test semantic scene understanding rather than temporal reasoning. (iii) **Trivial** fractions (20–37%) are substantial. (iv) The Temporal Purity Index ranges from 0.131 to 0.510, providing a single-number summary of each benchmark's temporal grounding quality.

Table 4: **Benchmark quality profiles under two-stage (LIS, FSS) classification.** *Trivial*: solvable from text alone (LIS $\leq \lambda$). *Fr.Agn*: vision-dependent but frame-insensitive. *Vis.Bias*: harmed by encoder-selected frames (FSS $< -\tau$). *Temp.Grnd*: benefits from relevant frames (FSS $> \tau$). TempSens = Vis.Bias + Temp.Grnd. TPI = TempSens / (N − Trivial).

Benchmark	Trivial (%)	Fr.Agn (%)	Vis.Bias (%)	Temp.Grnd (%)	**TempSens (%)**	TPI
Short-video						
MVBench	33.3	49.6	3.7	13.4	**17.1**	0.256
NExTQA	37.0	46.3	3.4	13.3	**16.7**	0.264
EgoSchema	20.2	58.8	2.4	18.6	**21.0**	0.263
Long-video						
MLVU	35.6	31.6	5.5	27.3	**32.8**	0.510
LongVideoBench	36.6	55.1	1.3	7.0	**8.3**	0.131
Video-MME	31.3	46.6	3.0	19.1	**22.1**	0.321

7.3 Validation

Fully validating temporal necessity is inherently difficult, as it is a latent property that cannot be observed directly from annotations alone. We therefore treat validation as converging evidence across complementary signals: human judgments, oracle-gap patterns, and internal and external correlations.

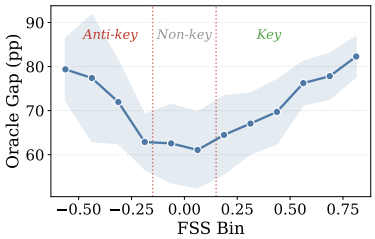
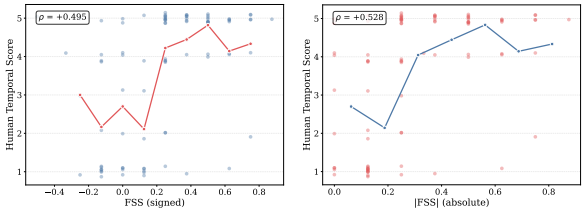


Figure 6: **Oracle gap exhibits a U-shape across FSS bins (cross-benchmark average).** Both Key and Anti-key tails show higher oracle gaps than the Non-key center, confirming that Anti-key samples are temporally sensitive. Shaded region: ± 1 std across benchmarks.



(a) FSS ($\rho = 0.495$) (b) $|\text{FSS}|$ ($\rho = 0.528$)

Figure 7: **FSS vs. human score.** (a) FSS shows positive correlation. (b) $|\text{FSS}|$ yields stronger correlation, capturing temporal dependence at both extremes.

Anti-key temporal dependence. A crucial design choice in TEMPCORE is including Anti-key samples (FSS $< -\tau$) alongside Key samples as temporally sensitive. Figure 6 aggregates oracle gap across all six benchmarks by FSS bin, revealing a clear U-shape: both positive (Key) and negative (Anti-key) tails exhibit elevated oracle gaps compared to the near-zero center, confirming that Anti-key samples—like Key samples—benefit substantially from oracle frame selection.

Human evaluation corroborates this finding. Figure 7 shows that $|\text{FSS}|$ achieves a stronger correlation with human temporal scores than FSS ($\rho = +0.528$ vs. $+0.495$), and Anti-key samples receive markedly higher temporal ratings than Non-key samples (3.60 vs. 2.32). Anti-key samples are visually misled but temporally dependent: SigLIP's highest-confidence frames capture appearance rather than temporal content, and removing them paradoxically helps because the model falls back on more temporally informative frames. See additional analysis and details in Section B.

Internal consistency. Oracle gaps on Key samples exceed those on Non-key samples on five of six benchmarks (mean $+27.5$ vs. $+17.8$; Table 2), and human annotators rate Key samples highest,

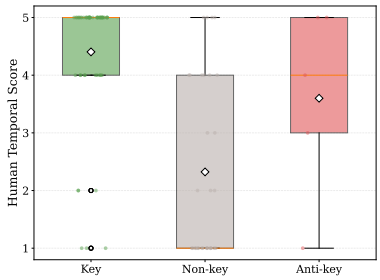


Figure 8: **Human temporal scores by FSS category ($\tau = 0.15$).** Key samples receive the highest ratings; Anti-key exceeds Non-key, supporting their inclusion in Temporally Sensitive.

followed by Anti-key and Non-key (see Figure 8), confirming that TEMPCORE retains those samples where temporal frame selection is consequential.

External correlation. Non-key model rankings correlate more strongly with six image VLM benchmarks than Key rankings (mean $\rho = 0.85$ vs. 0.67 across benchmarks; Section B.3), confirming that Key performance captures a distinct temporal skill rather than general VLM capability.

Practical usage. TEMPCORE subsets contain samples with $|\text{FSS}| > \tau$, comprising 8.3–32.8% of the original evaluation set. They can be used as a complementary metric: reporting TEMPCORE-subset accuracy alongside overall benchmark scores separates temporal from non-temporal performance. The pruning removes 67–92% of non-temporal samples, yielding a focused evaluation of temporal reasoning ability.

8 Conclusion

We presented three contributions toward understanding temporal grounding in Video QA evaluation. First, we proposed FSS, a per-sample diagnostic that quantifies dependence on temporal frame selection, and conducted a large-scale systematic study across multiple models, benchmarks, and strategies. Encoder-based frame selection yields substantial gains on long-video benchmarks, while oracle ceilings confirm large untapped headroom. Second, combining FSS with a LIS reveals that only a minority of benchmark samples are temporally sensitive—both vision-dependent and frame-sensitive. Third, we constructed TEMPCORE, temporally grounded evaluation subsets, providing a focused measure of temporal reasoning ability.

Limitations

While FSS effectively identifies frame-sensitive samples, it captures only one facet of temporal reasoning—dependence on which frames are selected. Other temporal requirements, such as causal ordering, event duration estimation, or state-change detection, may not manifest as frame-selection sensitivity and are therefore invisible to FSS. Incorporating complementary temporal diagnostics (*e.g.*, temporal ordering probes, action phase segmentation metrics) could yield a more comprehensive characterization of benchmark temporality. We leave the development of such multi-signal temporal taxonomies to future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, and 1 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, and 1 others. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

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, and 1 others. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Marija Brkic, Anas Filali Razzouki, Yannis Tevissen, Khalil Guetari, and Mounim A. El Yacoubi. 2025. Frame sampling strategies matter: A benchmark for small vision language models. *arXiv preprint arXiv:2509.14769*.

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2025. Flexible frame selection for efficient video reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. TemporalBench: Benchmarking fine-grained temporal understanding for multimodal video models. In *NeurIPS Workshop*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. 2025. Lost in time: A new temporal benchmark for VideoLLMs. In *British Machine Vision Conference (BMVC)*.

Wenzhang Du. 2025. When +1% is not enough: A paired bootstrap protocol for evaluating small improvements. *arXiv preprint arXiv:2511.19794*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Sihan Chen, Timin Gao, Peixian Chen, Mengdan Zhang, Xiawu Zheng, Rongrong Ji, Yunhang Shen, Caifeng Shan, and Xing Sun. 2025. Video-MME: The first-ever comprehensive evaluation benchmark of multimodal LLMs in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Yang Wang. 2023. Let's think frame by frame with VIP: A video infilling and prediction dataset for evaluating video chain-of-thought. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. 2025. M-LLM based video frame selection for efficient video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. 2025. On the consistency of video large

language models in temporal comprehension. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sutej Kulgod, Sean Ye, Sanchit Tanwar, and Christoffer Heckman. 2026. Reducing text bias in synthetically generated MCQAs for VLMs in autonomous driving. *arXiv preprint arXiv:2602.17677*.

Jie Lei, Tamara Berg, and Mohit Bansal. 2023. Revealing single frame bias for video-and-language learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, and 1 others. 2025. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A cause-effect look at language bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

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, and 1 others. 2025. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *International Conference on Machine Learning (ICML)*.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. 2025. MDP3: A training-free approach for list-wise frame selection in video-LLMs. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

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

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

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. 2025. Time blindness: Why video-language models can't see what humans can? *arXiv preprint arXiv:2505.24867*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, and 1 others. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. 2025. GenS: Generative frame sampler for long video understanding. In *Findings of the Association for Computational Linguistics*.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. 2025. Frame-Voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations (ICLR)*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025. VideoLLaMA3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024. LMMs-Eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. MLVU: Benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations (ICLR)*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, and 1 others. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

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. 2025. Apollo: An exploration of video understanding in large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Model	Architecture	Params
LLaVA-OV-7B	LLaVA-OneVision (Li et al., 2025)	7B
LLaVA-OV-0.5B	LLaVA-OneVision (Li et al., 2025)	0.5B
VideoLLaMA3-2B	VideoLLaMA3 (Zhang et al., 2025)	2B
SmolVLM2-2.2B	SmolVLM2 (Marafioti et al., 2025)	2.2B
SmolVLM2-500M	SmolVLM2 (Marafioti et al., 2025)	500M
SmolVLM2-256M	SmolVLM2 (Marafioti et al., 2025)	256M
Qwen3-VL-8B	Qwen3-VL (Bai et al., 2025a)	8B
InternVL3.5-8B	InternVL3 (Zhu et al., 2025)	8B

Table 5: **Eight open-source VLMs spanning three size tiers (256M–8B).** Five model families evaluated in zero-shot, multiple-choice setting with 32 input frames.

Benchmark	# QA	Median Len.	Type
EgoSchema (Mangalam et al., 2023)	500	∼3 min	Short
MVBench (Li et al., 2024)	4,000	<1 min	Short
NExTQA (Xiao et al., 2021)	8,564	∼1 min	Short
MLVU (Zhou et al., 2025)	2,175	∼12 min	Long
LongVideoBench (Wu et al., 2024)	1,337	∼10 min	Long
Video-MME (Fu et al., 2025)	2,700	various	Mixed

Table 6: **Six Video QA benchmarks spanning seconds to tens of minutes.** "Length" is the median video duration; "# QA" is the full evaluation set size.

Appendix overview.

- Section A – Experimental setup
- Section B – FSS and TempCore validation
- Section C – Method ablations
- Section D – Detailed results

A Experimental Setup

A.1 Model Details

We evaluate eight open-source VLMs spanning five model families and three size tiers (256M–8B parameters). The selection covers both established architectures (LLaVA-OneVision, InternVL) and recent lightweight models (SmolVLM2, VideoLLaMA3), ensuring diversity in design choices such as vision encoder integration, temporal token handling, and training data composition. All models are evaluated in a zero-shot, multiple-choice setting with 32 input frames.

A.2 Benchmark Details

We select six Video QA benchmarks that span a range of video lengths, question types, and evaluation protocols. Short-video benchmarks (EgoSchema, MVBench, NExTQA) feature clips under three minutes and test fine-grained actions, scene transitions, and causal reasoning. Long-video benchmarks (MLVU, LongVideoBench, Video-MME) contain videos of ten minutes or longer and require multi-hop temporal reasoning over extended narratives. All benchmarks use multiple-choice format, enabling deterministic accuracy evaluation without LLM-based judging.

B FSS and TempCore Validation

B.1 Anti-key Samples as Temporally Sensitive

Our main analysis includes Anti-key samples in TEMPCORE alongside Key samples (§7). The following evidence motivates this design decision: *Anti-key samples are also temporally sensitive*. The key insight is that SigLIP's max-prob frames can be *misleading*—selecting visually matching but temporally wrong content (*e.g.*, for "Did the yellow object leave the scene?", SigLIP selects frames *with* the object rather than frames showing its disappearance).

We hypothesize two mechanisms underlying Anti-key behavior: (i) for *static-appearance tasks* (*e.g.*, Object Existence, Fine-grained Action), SigLIP preferentially selects close-up shots of salient objects, but the question asks about *absence* or subtle state changes that are better captured by contextual background frames; (ii) for *temporal-sequence tasks* (*e.g.*, Action Sequence), the highest-similarity frames depict the endpoint of an action, depriving the model of the preceding context needed to infer order. Anti-key fractions are highest in temporal sub-tasks, and in MLVU the Anti-key fraction (11.1%) is substantial despite MLVU having the highest Key fraction (30.3%), indicating that visual appearance bias affects even temporally demanding benchmarks.

Oracle gap vs. FSS bin. Figure 9 bins samples by quantized FSS values and plots mean oracle gap per bin. Both positive (Key) and negative (Anti-key) tails show elevated oracle gaps compared to the near-zero center (Non-key), producing a U-shaped curve. This confirms that *both* Key and Anti-key samples benefit substantially from oracle frame selection, while Non-key samples—which respond similarly regardless of frame selection strategy—show lower oracle gaps.

Oracle gap vs. |FSS|. Figure 10 plots per-sample oracle gap against |FSS|. The binned mean (red line) rises monotonically: higher |FSS| correlates with larger oracle gaps regardless of sign, reinforcing that both Key and Anti-key samples are temporally localized.

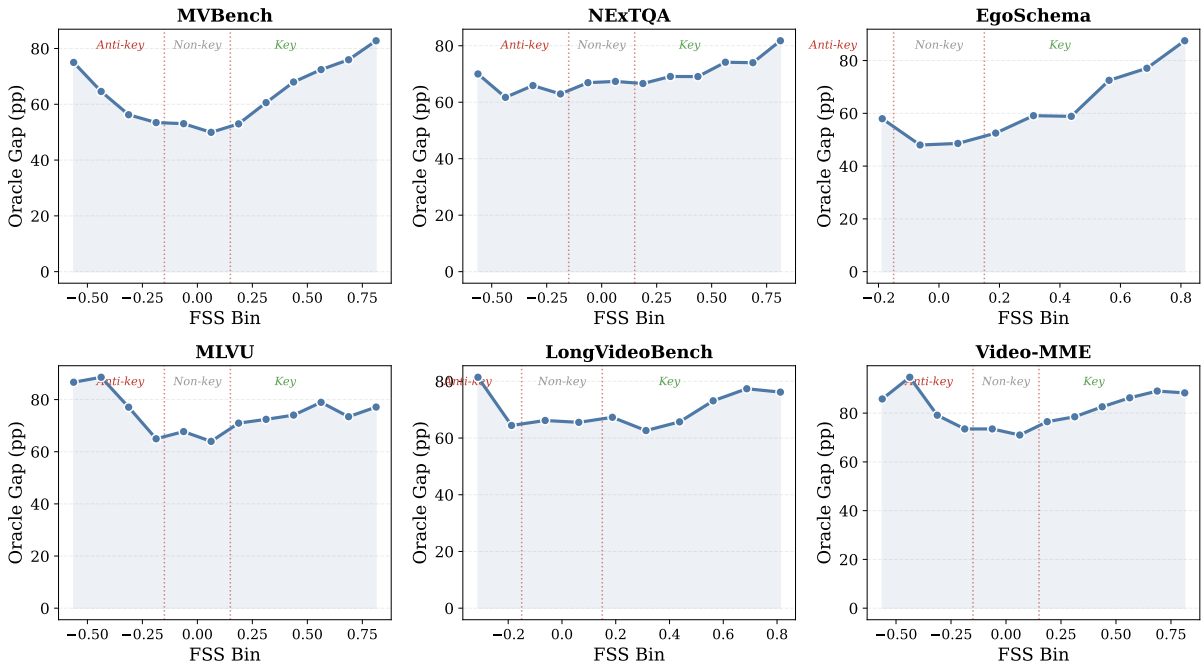


Figure 9: **Oracle gap exhibits a U-shape across FSS bins.** Both positive (Key) and negative (Anti-key) FSS tails show higher oracle gaps than the Non-key center, confirming that Anti-key samples are also temporally sensitive—they require specific temporal segments, but SigLIP selects misleading frames. Bins with fewer than 5 samples are omitted.

Correct window centroid position. Table 7 analyzes where in the video the correct-answer windows are located. Key and Anti-key samples tend to have higher centroid standard deviation than Non-key, confirming that their correct windows are distributed across specific temporal regions rather than clustering uniformly in the middle.
Together with the oracle gap subset analysis (Table 2), these results establish that Anti-key samples are temporally sensitive but misled by visually plausible-but-temporally-wrong frames. The true dichotomy is Temporal (Key + Anti-key) vs. Non-temporal (Non-key).

B.2 Human Temporal Judgment Validation

To validate FSS against human perception, we conducted a human evaluation study. An annotator assessed 120 stratified samples (20 per benchmark $\times$ 6 benchmarks, balanced across three FSS categories: Key, Non-key, Anti-key at $\tau = 0.15$) on a 1–5 temporal necessity scale. For each sample, annotators were shown the video, the question, and the answer options, and asked to assign a score indicating the extent to which temporal information was necessary to answer the question.

Metric correlations. Table 8 reports individual metric correlations with human temporal scores. Crucially, $|\text{FSS}|$ (absolute value) achieves higher Spearman ρ than signed FSS ($+0.528$ vs. $+0.495$), confirming that *both* Key and Anti-key samples exhibit temporal dependence.

FSS category analysis. Table 9 shows human temporal scores grouped by FSS category. Key samples receive the highest mean score (4.40) with 88.5% classified as temporal (≥ 4). Anti-key samples score markedly higher than Non-key (3.60 vs. 2.32, 60.0% vs. 32.3% temporal), supporting the U-shape hypothesis—Anti-key frames do signal temporal dependence, though less strongly than Key frames.

τ-sweep against human ground truth. Table 10 compares two temporal classification methods at varying FSS thresholds, using human scores ≥ 4 as ground truth. Method A (Key-only: FSS $> \tau$) achieves higher precision, while Method B (Key+Anti-key: $|\text{FSS}| > \tau$) gains in recall. At $\tau = 0.15$, Method B achieves F1 = 0.845 ($\geq$ Method A's 0.829), confirming the inclusion of Anti-key in TempCore. The advantage of Method B grows at higher τ values. We adopt Method B ($|\text{FSS}| > \tau$) as the default classification throughout this work.

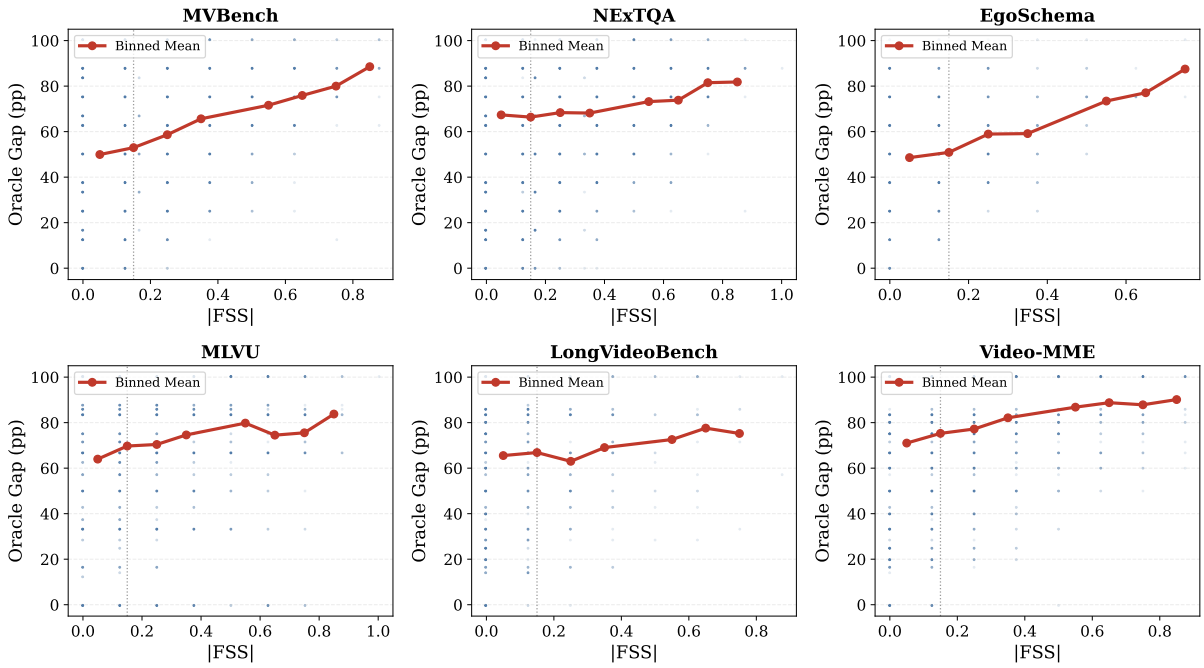


Figure 10: **Absolute FSS predicts oracle gap magnitude.** Per-sample scatter with binned means (red). The positive trend confirms that both high-FSS (Key) and low-FSS (Anti-key) samples benefit from temporal frame selection.

B.3 External Correlation with Image VLM Benchmarks

To validate that FSS correctly separates temporal from non-temporal performance, we correlate Key-subset and Non-key-subset model rankings with six image VLM benchmarks (MMMU, MMMU-Pro, MME, AI2D, TextVQA, OCRBench) evaluated using `lmms-eval` (Zhang et al., 2024). If FSS correctly identifies frame-sensitive samples, Non-key accuracy should correlate more strongly with image benchmarks (measuring general VLM capability), while Key accuracy should diverge.

Table 11 confirms this pattern on all six benchmarks: the Non-key–Key gap is largest on LongVideoBench (mean $\rho = 0.89$ vs. 0.42) and Video-MME (0.84 vs. 0.65), with a moderate gap on MLVU (0.82 vs. 0.65) and NExTQA (0.84 vs. 0.72). On MVBench and EgoSchema the gap narrows ($+0.04$ and $+0.09$), consistent with these benchmarks' lower temporal purity (Table 4).

The benchmarks where Key–Non-key divergence is largest are precisely those with high temporal content (LongVideoBench, Video-MME, MLVU), while the short-video benchmarks (MVBench, EgoSchema) show smaller gaps. This provides external evidence that FSS captures distinct, frame-sensitive capabilities rather than general VLM strength.

C Method Ablations

C.1 Frame Count (K) Ablation

We ablate the number of selected frames $K \in \{16, 32, 48\}$ on EgoSchema with LLaVA-OV-7B and LLaVA-OV-0.5B (Table 12). Accuracy differences across K values are modest ($\leq 6\%$), suggesting that MaxProb is robust to the frame budget. We use $K = 32$ throughout as a practical default that balances temporal coverage with computational cost.

C.2 Per-Encoder Comparison

We compare MaxProb accuracy with three vision-language encoders: SigLIP-400M (default), SigLIP2-400M, and CLIP ViT-L/14. Table 13 reports Q-mode accuracy on the Perception benchmark (19,140 samples) for two LLaVA models. All three encoders produce nearly identical accuracy, with differences within 0.8 pp. This confirms that frame selection quality is robust to the choice of scoring encoder, consistent with the finding of Zohar et al. (2025) that modern vision-language encoders converge in frame-level ranking ability.

C.3 Question and Answer Text Statistics

Table 14 reports the mean question and answer option lengths across six benchmarks. LongVideoBench questions average 43.4 words,

Benchmark	Category	Begin (%)	Middle (%)	End (%)	Std
Short-video					
MVBench	Key	9.1	73.1	17.8	0.154
	Non-key	3.4	89.9	6.6	0.103
	Anti-key	8.7	84.3	7.0	0.134
NExTQA	Key	5.9	85.8	8.3	0.116
	Non-key	2.2	94.9	2.9	0.078
	Anti-key	3.5	90.6	5.9	0.102
EgoSchema	Key	6.4	84.4	9.2	0.111
	Non-key	5.5	89.6	4.9	0.101
	Anti-key	7.1	78.6	14.3	0.113
Long-video					
MLVU	Key	6.9	86.9	6.2	0.116
	Non-key	3.3	93.2	3.6	0.085
	Anti-key	3.5	91.8	4.7	0.105
LongVideoBench	Key	12.0	78.0	10.0	0.150
	Non-key	4.8	90.6	4.5	0.107
	Anti-key	0.0	81.0	19.0	0.117
Video-MME	Key	8.9	83.1	8.0	0.126
	Non-key	5.6	88.7	5.7	0.107
	Anti-key	5.7	88.6	5.7	0.099

Table 7: **Key and Anti-key samples show concentrated correct-window positions.** Distribution of correct-window centroids across video thirds (Begin: $[0, 1/3)$, Middle: $[1/3, 2/3)$, End: $[2/3, 1]$). Lower standard deviation indicates more concentrated temporal localization. Both Key and Anti-key tend to have lower Std than Non-key.

Metric	Spearman ρ	PB r	AUC
FSS	$+0.495$	$+0.414$	0.777
|FSS|	$+0.528$	$+0.447$	0.799

Table 8: **Individual metric correlations with human temporal scores.** |FSS| captures both Key and Anti-key temporal dependence, yielding higher correlation than signed FSS (PB r: Point-Biserial correlation). Threshold for binarization: score ≥ 4.

three to four times longer than those of other benchmarks (11–24 words), because they embed detailed scene descriptions —clothing, spatial layout, character actions—before posing a query. This explains why Q-mode outperforms QA-mode on LongVideoBench (§5.1): the question text alone provides sufficient visual grounding for the encoder, and appending answer text adds noise rather than signal.

LongVideoBench case study. The following examples illustrate the descriptive nature of LongVideoBench questions:

> **LongVideoBench (77 words):** "In a room decorated with various paintings, there are two men standing in the middle. On the left is a man with short hair wearing a red short-sleeve shirt, holding a beer. On the right is a man with a black beard, wearing a gray short-sleeve shirt. Behind the man in red is a white door. When the subtitle mentions 'over 2600 meters yeah there you go we', what did the man in the red shirt do?"
>
> **LongVideoBench (79 words):** "In the video, there are two men in a room. The man on the left is wearing a yellow shirt, a hat backwards, and a watch on his left hand. The man on the right is wearing a shirt with red flowers and green leaves and has a bracelet on his left hand. When the subtitle mentions 'interested enough to join us but first,' what change happens to the man wearing the shirt with red flowers and green leaves?"
>
> **MLVU (4 words):** "Did I drink water?"

The contrast is striking: LongVideoBench questions specify clothing colors, accessories, and spatial positions in enough detail for a vision-language encoder to localize the relevant frames from the question alone. In contrast, short questions like the MLVU example provide almost no visual cues, making the answer text a valuable source of grounding signal.

FSS Bin	Mean	Temporal%
Key	4.40	88.5%
Non-key	2.32	32.3%
Anti-key	3.60	60.0%

Table 9: **Human temporal scores by FSS category ($\tau = 0.15$).** Key and Anti-key samples both exhibit elevated temporal scores compared to Non-key, supporting the U-shape hypothesis. Temporal%: score ≥ 4.

τ	Key-Only (FSS $> \tau$) Prec.	Rec.	F1	Key+Anti-Key (|FSS| $> \tau$) Prec.	Rec.	F1
0.050	0.787	0.814	0.800	0.705	0.932	0.803
0.150	0.885	0.780	0.829	0.860	0.831	**0.845**
0.250	0.912	0.525	0.667	0.917	0.559	0.695
0.375	0.920	0.390	0.548	0.923	0.407	0.565
0.500	0.857	0.203	0.329	0.857	0.203	0.329

Table 10: **τ-sweep: Key-only vs. Key+Anti-key temporal classification against human ground truth.** Including Anti-key samples (right) improves recall and F1 across most τ values, confirming that anti-key frames also indicate temporal dependence.

D Detailed Results

D.1 Per-Benchmark Figures

Figure 11 presents the per-benchmark FSS distributions corresponding to the aggregated view in Figure 4. Figure 12 shows the cross-benchmark model ranking correlation matrices; Figure 13 shows the per-benchmark LIS–FSS scatter plots.

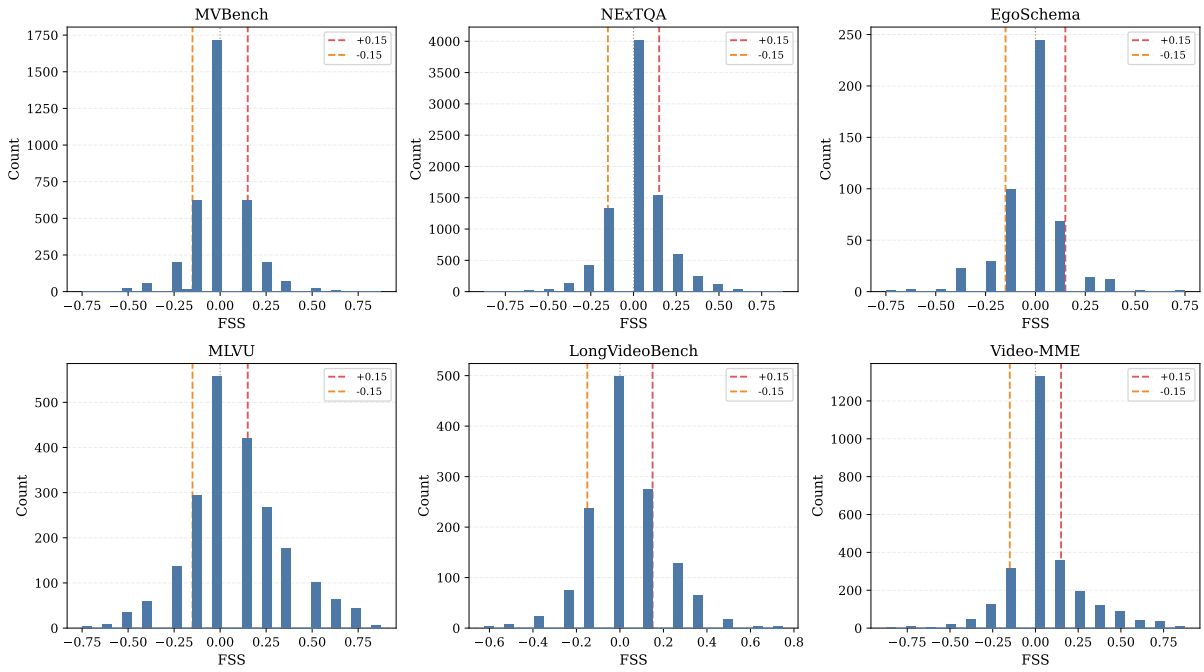


Figure 11: **Per-benchmark FSS distributions** ($\tau = 0.15$)**.** All six benchmarks show a dominant near-zero peak, with MLVU exhibiting the broadest spread.

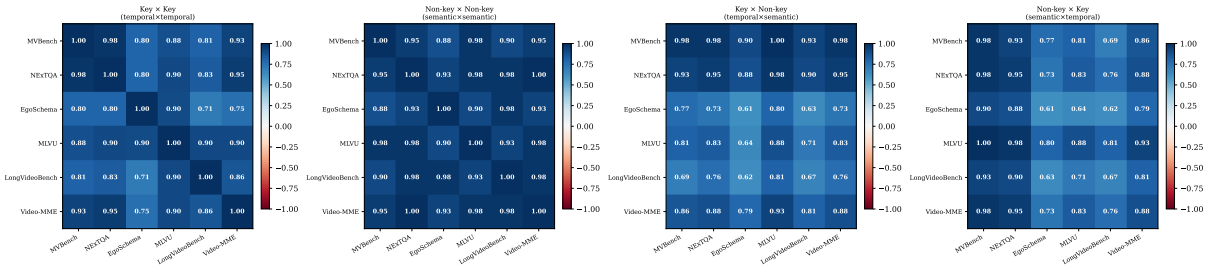


Figure 12: **Cross-benchmark model ranking correlations.** 1×4 Spearman ρ matrices for Key×Key, Non-key×Non-key, Key×Non-key, and Non-key×Key subsets.

D.2 Performance Tables

Table 15–Table 23 provide complete per-model results for each evaluation strategy and benchmark.

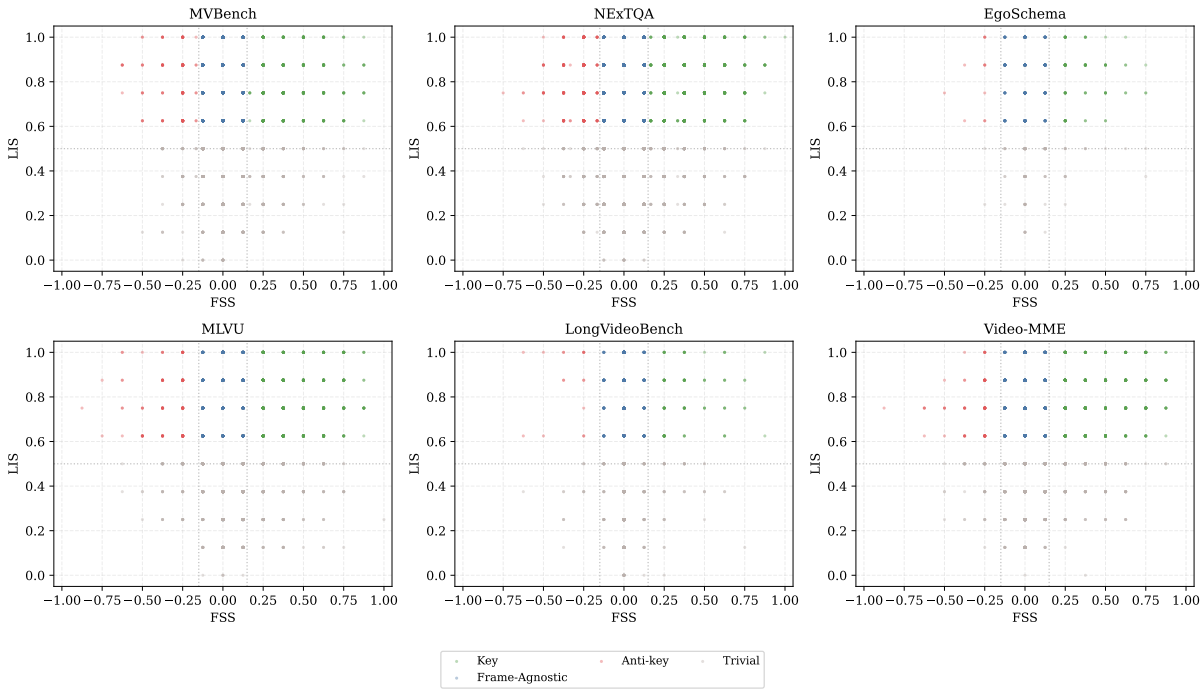


Figure 13: **Per-benchmark LIS–FSS scatter plots.** Each panel shows the two-dimensional (LIS, FSS) distribution with four-category coloring. Dashed lines mark thresholds ($\tau = 0.15$, $\lambda = 0.50$).

	Spearman ρ vs. Image VLM Benchmark						
Split	MMMU	MMMU-Pro	MME	AI2D	TextVQA	OCRBench	Mean
Short-video							
MVBench							
Key	.83	.83	.86	.79	.76	.60	.78
Non-key	.86	.86	.90	.83	.81	.67	.82
NExTQA							
Key	.76	.76	.79	.74	.76	.52	.72
Non-key	.90	.90	.90	.86	.83	.64	.84
EgoSchema							
Key	.83	.83	.86	.79	.76	.60	.78
Non-key	.93	.93	.90	.88	.81	.79	.87
Long-video							
MLVU							
Key	.69	.69	.76	.67	.64	.43	.65
Non-key	.86	.86	.90	.83	.81	.67	.82
LongVideoBench							
Key	.50	.50	.52	.43	.45	.14	.42
Non-key	.95	.95	.93	.90	.86	.74	.89
Video-MME							
Key	.69	.69	.76	.67	.64	.43	.65
Non-key	.90	.90	.90	.86	.83	.64	.84

Table 11: **Non-key accuracy correlates more strongly with image VLM benchmarks than Key accuracy.** Spearman ρ between Key and Non-key (including Anti-key) model rankings and six image benchmarks evaluated via `lmms-eval`. Key consistently shows lower correlation, with the gap largest on long-video benchmarks (LongVideoBench: .42 vs. .89; MLVU: .65 vs. .82), confirming that Key performance captures a distinct temporal skill rather than general VLM capability. $N = 8$ models.

Model	K=16	$\boldsymbol{K}$**=32**	K=48
LLaVA-OV-7B	55.4	**50.2**	56.2
LLaVA-OV-0.5B	22.6	**23.4**	25.6

Table 12: **MaxProb accuracy is robust to frame budget K.** Accuracy (%) on EgoSchema across $K \in \{16, 32, 48\}$ for two representative models. Differences are modest, supporting K=32 as a practical default. **Bold** marks the default K=32.

Model	SigLIP2-400M	SigLIP-400M	CLIP ViT-L
LLaVA-OV-7B	56.5	57.1	57.3
LLaVA-OV-0.5B	48.6	48.6	48.8
Δ (max$-$min)		≤ 0.8 pp	

Table 13: **Encoder choice has negligible impact on MaxProb accuracy.** Q-mode MaxProb accuracy (%) on Perception (19,140 samples) with three vision-language encoders. All differences are within 1 pp, confirming that frame selection quality is robust to encoder choice.

Benchmark	Q (words)	A (words)	#Opts
MVBench	13.0	3.9	3.7
NExTQA	11.5	2.8	5.0
EgoSchema	24.1	19.3	5.0
MLVU	15.1	2.5	4.0
LongVideoBench	**43.4**	**8.9**	**4.7**
Video-MME	13.4	5.9	4.0

Table 14: **LongVideoBench questions are 3–4× longer than other benchmarks.** Mean question length (Q), answer option length (A), and number of options (#Opts) across six benchmarks. The descriptive nature of LongVideoBench questions explains why Q-mode alone provides sufficient visual grounding.

	Short-video						*Long-video*					
	MVBench		NExTQA		EgoSchema		MLVU		LongVideoBench		Video-MME	
Model	Q	QA	Q	QA	Q	QA	Q	QA	Q	QA	Q	QA
InternVL3.5-8B	64.8	67.7	77.3	80.0	52.4	62.4	64.7	69.0	59.2	57.5	55.2	59.6
LLaVA-OV-7B	52.7	55.6	76.8	79.4	50.2	61.6	63.6	68.6	63.6	56.0	54.2	59.6
Qwen3-VL-8B	56.4	59.5	70.6	73.1	47.8	56.8	60.7	64.4	51.2	49.6	50.9	55.4
SmolVLM2-2.2B	45.5	48.5	60.2	62.5	25.8	30.8	57.2	60.1	50.0	47.1	47.3	50.6
LLaVA-OV-0.5B	44.8	47.8	57.3	60.3	23.4	26.8	55.8	59.5	49.2	43.7	42.8	47.8
SmolVLM2-500M	41.6	43.8	46.1	48.6	22.8	26.2	50.8	54.9	42.4	41.1	39.0	43.9
VidLLaMA3-2B	39.6	39.7	46.8	47.0	23.6	23.2	42.9	42.2	47.0	42.2	36.6	36.5
SmolVLM2-256M	31.7	32.9	27.6	28.1	16.0	17.2	35.3	37.6	31.2	27.6	31.9	33.7

Table 15: **Per-model Q-mode vs. QA-mode MaxProb accuracy (%).** QA-mode improves accuracy for all models except on LongVideoBench, where long, descriptive questions already provide sufficient visual grounding. The QA advantage is consistent across model sizes, with the largest gains on EgoSchema.

	Short-video									*Long-video*									*Avg. Δ*	
	MVBench			NExTQA			EgoSchema			MLVU			LongVideoBench			Video-MME				
Model	Max	Min	Δ	Max	Min	Δ	Max	Min	Δ	Max	Min	Δ	Max	Min	Δ	Max	Min	Δ	Short	Long
InternVL3.5-8B	64.8	64.9	-0.1	77.3	74.6	+2.8	52.4	57.8	-5.4	64.7	52.3	+12.5	57.5	52.1	+5.5	55.2	48.0	+7.2	-0.9	+8.4
LLaVA-OV-7B	52.7	52.6	+0.1	76.8	73.4	+3.4	50.2	56.2	-6.0	63.6	51.2	+12.3	56.0	50.5	+5.5	54.2	47.3	+6.9	-0.8	+8.3
Qwen3-VL-8B	56.4	54.7	+1.7	70.6	69.0	+1.6	47.8	52.8	-5.0	60.7	49.0	+11.8	49.6	46.4	+3.1	50.9	45.6	+5.3	-0.6	+6.8
SmolVLM2-2.2B	45.5	45.5	-0.1	60.2	57.2	+3.0	25.8	29.2	-3.4	57.2	45.9	+11.2	48.5	42.0	+6.5	47.3	40.4	+6.9	-0.1	+8.2
LLaVA-OV-0.5B	44.8	44.5	+0.3	57.3	54.9	+2.4	23.4	24.6	-1.2	55.8	43.6	+12.2	44.9	38.1	+6.8	42.8	36.7	+6.1	+0.5	+8.4
SmolVLM2-500M	41.6	40.9	+0.6	46.1	44.2	+1.9	22.8	22.8	+0.0	50.8	42.6	+8.2	42.1	37.2	+5.0	39.0	34.2	+4.8	+0.9	+6.0
VidLLaMA3-2B	39.6	39.7	-0.2	46.8	46.6	+0.2	23.6	23.6	+0.0	42.9	43.1	-0.2	40.9	40.8	+0.2	36.6	36.4	+0.2	+0.0	+0.1
SmolVLM2-256M	31.7	31.6	+0.1	27.6	26.9	+0.7	16.0	17.0	-1.0	35.3	30.6	+4.6	27.2	27.2	+0.0	31.9	29.3	+2.6	-0.1	+2.4
Average	47.1	46.8	+0.3	57.8	55.8	+2.0	32.8	35.5	-2.8	53.9	44.8	+9.1	45.9	41.8	+4.1	44.7	39.7	+5.0	-0.1	+6.1

Table 16: **MaxProb consistently outperforms MinProb on long-video benchmarks.** Per-model accuracy (%) with Short and Long average Δ columns.

	Short-video									*Long-video*									*Avg. Gap*	
	MVBench			NExTQA			EgoSchema			MLVU			LongVideoBench			Video-MME				
Model	Uniform	Oracle	Gap	Uniform	Oracle	Gap	Uniform	Oracle	Gap	Uniform	Oracle	Gap	Uniform	Oracle	Gap	Uniform	Oracle	Gap	Short	Long
InternVL3.5-8B	65.4	77.1	+11.7	76.2	87.7	+11.5	55.8	83.8	+28.0	53.0	93.5	+40.4	48.9	82.7	+33.8	49.4	85.2	+35.9	+17.1	+36.7
LLaVA-OV-7B	52.6	64.4	+11.8	76.1	88.2	+12.1	54.6	83.4	+28.8	52.2	93.0	+40.8	58.6	81.8	+23.2	51.2	91.7	+40.5	+17.5	+34.8
Qwen3-VL-8B	57.1	66.4	+9.4	70.1	84.4	+14.3	53.6	81.0	+27.4	51.0	84.8	+33.8	45.6	80.9	+35.2	47.8	85.5	+37.8	+17.0	+35.6
LLaVA-OV-0.5B	44.6	52.1	+7.5	56.8	69.9	+13.2	24.6	40.2	+15.6	47.3	83.2	+35.9	44.6	63.8	+19.2	36.7	73.7	+37.0	+12.1	+30.7
SmolVLM2-2.2B	45.6	52.1	+6.5	59.2	70.4	+11.2	28.2	41.8	+13.6	48.7	76.8	+28.1	43.7	67.8	+24.2	40.4	73.8	+33.4	+10.4	+28.5
SmolVLM2-500M	41.7	47.7	+5.9	46.2	57.4	+11.2	26.0	44.4	+18.4	43.5	78.9	+35.4	38.9	69.2	+30.3	35.5	75.3	+39.8	+11.8	+35.2
SmolVLM2-256M	31.1	33.6	+2.5	27.6	37.2	+9.6	16.8	22.2	+5.4	32.0	61.0	+29.1	26.7	48.9	+22.2	29.1	67.5	+38.4	+5.8	+29.9
VidLLaMA3-2B	39.6	41.0	+1.4	47.2	48.9	+1.7	23.2	25.8	+2.6	42.5	46.1	+3.5	41.1	45.1	+4.0	36.3	40.3	+4.0	+1.9	+3.8
Average	47.2	54.3	+7.1	57.4	68.0	+10.6	35.3	52.8	+17.5	46.3	77.2	+30.9	43.5	67.5	+24.0	40.8	74.1	+33.3	+11.7	+29.4

Table 17: **Large oracle gaps reveal substantial headroom for frame selection.** Uniform-sampling vs. window-oracle accuracy (%). Long-video gaps far exceed short-video gaps, confirming that temporal frame selection is most consequential for extended videos.

Model	No Frame	Uniform	Oracle	MaxProb	MinProb
InternVL3.5-8B	38.9	69.9	77.1	64.8	64.9
Qwen3-VL-8B	37.5	57.4	66.4	56.4	54.7
LLaVA-OV-7B	36.5	57.2	64.4	52.7	52.6
SmolVLM2-2.2B	30.2	46.6	52.1	45.5	45.5
LLaVA-OV-0.5B	33.0	46.4	52.1	44.8	44.5
SmolVLM2-500M	31.2	42.2	47.7	41.6	40.9
VidLLaMA3-2B	31.9	39.4	41.0	39.6	39.7
SmolVLM2-256M	29.3	31.9	33.6	31.7	31.6

Table 18: **Window oracle dominates all strategies on MVBench.** Cross-strategy accuracy (%) across 8 VLMs. Oracle substantially outperforms all other strategies, while MaxProb and uniform sampling perform comparably on this short-video benchmark.

Model	No Frame	Uniform	Oracle	MaxProb	MinProb
InternVL3.5-8B	48.8	80.2	87.7	77.3	74.6
LLaVA-OV-7B	49.2	78.3	88.2	76.8	73.4
Qwen3-VL-8B	49.8	70.5	84.4	70.6	69.0
LLaVA-OV-0.5B	35.7	57.0	69.9	57.3	54.9
SmolVLM2-2.2B	22.6	61.1	70.4	60.2	57.2
VidLLaMA3-2B	30.0	46.8	48.9	46.8	46.6
SmolVLM2-500M	25.2	47.2	57.4	46.1	44.2
SmolVLM2-256M	21.1	27.9	37.2	27.6	26.9

Table 19: **Window oracle reveals large untapped potential on NExTQA.** Cross-strategy accuracy (%) across 8 VLMs. MaxProb and uniform sampling remain closely matched.

Model	No Frame	Uniform	Oracle	MaxProb	MinProb
InternVL3.5-8B	33.0	64.6	83.8	52.4	57.8
LLaVA-OV-7B	34.6	63.0	83.4	50.2	56.2
Qwen3-VL-8B	36.6	54.0	81.0	47.8	52.8
SmolVLM2-2.2B	20.2	29.4	41.8	25.8	29.2
SmolVLM2-500M	19.4	26.2	44.4	22.8	22.8
LLaVA-OV-0.5B	15.6	26.4	40.2	23.4	24.6
VidLLaMA3-2B	25.6	24.4	25.8	23.6	23.6
SmolVLM2-256M	20.2	17.0	22.2	16.0	17.0

Table 20: **MinProb outperforms MaxProb on EgoSchema, reflecting high Anti-key prevalence.** Cross-strategy accuracy (%) across 8 VLMs. The reversed MaxProb–MinProb pattern is consistent with EgoSchema's high Anti-key fraction.

Model	No Frame	Uniform	Oracle	MaxProb	MinProb
InternVL3.5-8B	44.9	65.6	93.5	64.7	52.3
LLaVA-OV-7B	45.1	63.9	93.0	63.6	51.2
Qwen3-VL-8B	41.4	52.0	84.8	60.7	49.0
LLaVA-OV-0.5B	35.8	51.3	83.2	55.8	43.6
SmolVLM2-2.2B	30.2	54.6	76.8	57.2	45.9
SmolVLM2-500M	31.1	48.6	78.9	50.8	42.6
VidLLaMA3-2B	34.2	42.9	46.1	42.9	43.1
SmolVLM2-256M	25.7	33.3	61.0	35.3	30.6

Table 21: **MLVU shows the strongest strategy differentiation among all benchmarks.** Cross-strategy accuracy (%) across 8 VLMs, reflecting MLVU's high temporal demand.

Model	No Frame	Uniform	Oracle	MaxProb	MinProb
LLaVA-OV-7B	39.7	56.5	81.8	56.0	50.5
InternVL3.5-8B	44.6	57.3	82.7	57.5	52.1
Qwen3-VL-8B	42.8	48.8	80.9	49.6	46.4
SmolVLM2-2.2B	29.2	47.3	67.8	48.5	42.0
LLaVA-OV-0.5B	33.7	43.0	63.8	44.9	38.1
SmolVLM2-500M	31.0	40.9	69.2	42.1	37.2
VidLLaMA3-2B	36.5	41.1	45.1	40.9	40.8
SmolVLM2-256M	25.1	26.8	48.9	27.2	27.2

Table 22: **MaxProb provides meaningful gains over MinProb on LongVideoBench.** Cross-strategy accuracy (%) across 8 VLMs, demonstrating effective frame retrieval on long videos.

Model	No Frame	Uniform	Oracle	MaxProb	MinProb
InternVL3.5-8B	44.5	61.3	85.2	55.2	48.0
LLaVA-OV-7B	38.2	57.7	91.7	54.2	47.3
Qwen3-VL-8B	40.6	49.7	85.5	50.9	45.6
SmolVLM2-2.2B	28.1	49.4	73.8	47.3	40.4
LLaVA-OV-0.5B	30.7	44.3	73.7	42.8	36.7
SmolVLM2-500M	29.5	40.3	75.3	39.0	34.2
SmolVLM2-256M	25.4	30.8	67.5	31.9	29.3
VidLLaMA3-2B	28.2	33.5	40.3	36.6	36.4

Table 23: **Video-MME exhibits the largest oracle gaps among all benchmarks.** Cross-strategy accuracy (%) across 8 VLMs.