FULCRUM: OPTION-CONTRASTIVE FRAME SELECTION FOR LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video-capable vision-language models (VLMs) answer questions about long videos under a limited frame budget, making frame selection critical to the evidence available to the answering model. However, question relevance alone does not indicate whether a frame supports one candidate answer over the others. We introduce FULCRUM, a training-free frame selector for multiple-choice video question answering built on option-contrastive relevance. Using a frozen image–text encoder, FULCRUM scores each frame against each question–option pair and combines the strongest match with its margin over the mean match across options. This captures both relevance to an answer hypothesis and its support relative to competing hypotheses. The resulting scores guide budget allocation across temporal segments, while a coverage floor preserves broad temporal coverage and diversity-aware selection reduces redundancy. FULCRUM requires a single encoding pass over the candidate frames and option hypotheses, without querying the answering model during selection. Under a shared evaluation protocol on LongVideoBench and Video-MME, with Qwen2.5-VL-7B and LLaVA-Video-7B, FULCRUM outperforms uniform sampling and reimplementations of ten published selectors at all evaluated budgets of 4, 8, 16, and 32 frames. With eight frames on LongVideoBench, it improves Qwen2.5-VL-7B accuracy from 44.2% to 62.5%, exceeding the strongest competing selector by 3.1 percentage points and surpassing Qwen2.5-VL-72B with Adaptive Keyframe Sampling at the same frame budget. Across both models and benchmarks, four FULCRUM-selected frames outperform 32 uniformly sampled frames.

1 INTRODUCTION

Video large language models (Video-LLMs) answer a question about a video from a fixed number of frames, because every admitted frame enters the language model's context as hundreds of visual tokens (Bai et al., 2025; Zhang et al., 2024b; Li et al., 2024). For long videos this frame budget k covers a small fraction of the footage: an hour sampled at one frame per second yields 3,600 candidate frames, and selection methods are evaluated at budgets of 8 to 64 (Sun et al., 2025b; Zhang et al., 2025b). More frames do not lift the constraint: beyond a modest number, accuracy on long-video benchmarks flattens or degrades as redundant frames dilute the informative ones (Fu et al., 2024; Wu et al., 2024; Zhou et al., 2024). Long-video question answering therefore turns on a selection problem: which k frames the answerer receives.

Uniform sampling is the default in deployed Video-LLMs (Li et al., 2024; Bai et al., 2025). It costs nothing and is blind to content. Query-conditioned selectors improve on it by scoring each frame against the question with a vision–language encoder and spending the budget on high-scoring frames, combined with temporal coverage (Tang et al., 2025), list-wise diversity (Sun et al., 2025b; Zhang et al., 2025b), multi-resolution input (Zhang et al., 2025a), bandit search over clips (Zhu et al., 2026), or iterative verification by a vision–language model (VLM) (Zou et al., 2026b). Answer-aware selectors add the options to the text that frames are matched against, fused into one string (Wu et al., 2026) or appended to the question one at a time, keeping the best match (Sarkar et al., 2026). Across this family, relevance is one scalar per frame. A frame scored against the question, the fused options, or its best-matching option carries no measure of how far it separates that option from the others.

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.

We argue that this scalar measures the wrong quantity for multiple-choice inference. The answerer's output is a choice among options, and a frame informs that choice to the extent that it favors one option over the others. Scoring each frame against every option hypothesis exposes this directly: the gap between the best-supported option and the average option is the log-odds of the favored option against the geometric mean of all options, and the partition function cancels by construction (§3.2). The two scores select different frames: fusing the options into the query keeps 81% of the frames selected by the question alone, while contrasting them keeps 52% (App. F.2). Image–text similarity still locates evidence poorly: existing temporal search methods reach 2.1% temporal F1 on the LongVideoBench subset of LV-Haystack (Ye et al., 2025). A selector that commits its budget to the top-ranked frames inherits every ranking error, so the budget should follow the evidence in proportion, above a floor that guarantees temporal coverage.

We propose FULCRUM, a training-free frame selector for multiple-choice video question answering built from three components. **Option-contrastive relevance** scores each frame by the best-matching option hypothesis and its log-odds over the alternatives (§3.2). **Proportional allocation** places one frame in every part of the video and divides the rest of the budget in proportion to smoothed evidence (§3.3). **Diversity-aware assignment** fills each segment with non-redundant frames by the relevance-weighted log-determinant objective of AdaRD-Key (Zhang et al., 2025b) (§3.4). FULCRUM needs only a contrastive image–text encoder and leaves the answerer frozen. Our contributions are as follows:

- **Option-contrastive relevance**, a frame score for multiple-choice inference that measures how far a frame separates the answer options, with a log-odds interpretation that is independent of the encoder's temperature (§3.2).
- **Proportional budget allocation** under a coverage floor, solved exactly by greedy assignment, whose single parameter is expressed in within-video standard deviations and carries the same meaning in every video (§3.3).
- **Matched-budget gains**: at equal frame count, FULCRUM outperforms uniform sampling and ten published selectors run in one harness, at every budget from 4 to 32 frames, on two benchmarks and with two answerers (§5.1).

2 RELATED WORK

Query-conditioned frame selection. Uniform sampling remains the default input policy of open Video-LLMs (Li et al., 2024; Bai et al., 2025). Training-free selectors score each frame against the question with a vision–language encoder and differ in how they spend the budget on that score: by recursive splitting against temporal coverage (Tang et al., 2025), inverse-transform sampling (Liu et al., 2025), top-ranked frames at several resolutions (Zhang et al., 2025a), or a bandit over clips (Zhu et al., 2026). Others search the video iteratively with an object detector, a large language model (LLM) over captions, or a VLM in the loop (Ye et al., 2025; Guo et al., 2025; Wang et al., 2025b; Zou et al., 2026b; Li et al., 2026a), and learned selectors train the scorer from Video-LLM rankings, relevance labels, or reinforcement learning (Yu et al., 2025; Hu et al., 2025; Yao et al., 2025a; Tang et al., 2026; Qin et al., 2026; Yao et al., 2025b). In all of these methods, the relevance of a frame is computed against the question. App. A compares 24 selectors along these axes.

Answer options, allocation, and diversity. Two selectors read the answer options. VTR-VLM fuses the question with candidate options generated by the answerer into one text (Wu et al., 2026), and QAaF appends each option to the question and keeps the frames most aligned with any of the resulting texts (Sarkar et al., 2026). Both reduce the options to one score per frame, the fused and best-match rows of Table 2; FULCRUM adds the contrast disc, the log-odds of the favored option against the geometric mean of all options (Eq. (3)). Other selectors spread the budget across the video, by dynamic programming over segments (Sun et al., 2025b), importance-weighted clip budgets (Chen et al., 2026), or relevance-dependent sampling density (Wu et al., 2026), and control redundancy by volume, determinantal, or irreplaceability objectives (Li et al., 2026b; Zhang et al., 2025b; Ma et al., 2026; Sun et al., 2025a). FULCRUM allocates by a concave program above an explicit coverage floor (§3.3) and adopts the log-determinant assignment of AdaRD-Key within segments (§3.4).

Beyond single-pass selection. Token compression keeps the frame set and reduces the tokens each frame contributes, by merging, pruning, or adaptive ratios (Hyun et al., 2025; Shao et al., 2025; Fu et al., 2025; Shen et al., 2025; Tao et al., 2025; Wang et al., 2025a); agentic and retrieval-based systems answer through several model calls (Zhang et al., 2025c; Luo et al., 2025; Xie et al., 2026). FULCRUM selects the frames once, before the frozen answerer runs, and its output can be passed to any token compressor.

3 METHODOLOGY

FULCRUM selects k frames in three stages (Fig. 1). §3.2 scores every frame by how far it separates the answer options, §3.3 divides the budget across the video in proportion to that score above a coverage floor, and §3.4 chooses non-redundant frames within each part of the video. Implementation details, complexity, and pseudocode are given in App. B.

3.1 PROBLEM SETUP

Inputs. A video is represented by T frames $V = (v_0, \dots, v_{T-1})$ indexed by $[T] = \{0, \dots, T-1\}$. A multiple-choice question q has J options $\mathcal{O} = (o_1, \dots, o_J)$, labeled by the letters A, B, …, and a gold index $y^\star \in \{1, \dots, J\}$. A selector returns a set $S \subset [T]$ of exactly k distinct indices, which the answerer receives in increasing temporal order.

Answerer. The frozen answerer is prompted, in a single user turn, with the k frames followed by the question, the lettered options, and an instruction to reply with one option letter. Let Λ_j be the set of vocabulary tokens that spell letter j, written with and without a leading space, and let $p(u \mid S)$ be the answerer's next-token probability at the first answer position. The option posterior is

$$\hat{\pi}_j(S) = \frac{\sum_{u \in \Lambda_j} p(u \mid S)}{\sum_{l=1}^{J} \sum_{u \in \Lambda_l} p(u \mid S)}, \tag{1}$$

the prediction is $\arg\max_j \hat{\pi}_j(S)$, and a question is answered correctly when the prediction equals $y^\star$. The selector chooses S from the video, the question, and the options alone, without calling the answerer.

Segments. The video is divided into m contiguous segments $\mathcal{S}_1, \dots, \mathcal{S}_m$ of equal length (App. B.2). An allocation is a vector $\mathbf{n} = (n_1, \dots, n_m)$ of non-negative integers with $\sum_i n_i = k$, specifying how many selected frames each segment receives.

3.2 OPTION-CONTRASTIVE RELEVANCE

Let ϕ and ψ be the image and text towers of a contrastive encoder (Radford et al., 2021) with embedding dimension D, and let $z_t = \phi(v_t)/\|\phi(v_t)\|_2$.

Option-conditioned similarity. For each option we form the hypothesis string h_j, the question and the option text joined by a single space and truncated to the context length of the text encoder, and embed it as $e_j = \psi(h_j)/\|\psi(h_j)\|_2$. The similarity matrix is $a_{t,j} = \langle z_t, e_j \rangle$ for $t \in [T]$ and $j \in \{1, \dots, J\}$. Two frame-level statistics follow:

$$\mathrm{rel}(t) = \max_j a_{t,j}, \qquad \mathrm{disc}(t) = \max_j a_{t,j} - \frac{1}{J} \sum_{j=1}^{J} a_{t,j}. \tag{2}$$

$\mathrm{rel}(t)$ measures support for the best-matching answer hypothesis, and $\mathrm{disc}(t)$ measures how far that hypothesis stands above the average hypothesis. A frame that matches the question but no option in particular scores high on rel and near zero on disc; a frame that favors one option scores high on both (Fig. 1, step 2).

Log-odds interpretation. Let frame t induce a posterior over options $\pi_j(t) = \exp(a_{t,j}/\tau)/\sum_l \exp(a_{t,l}/\tau)$ at any temperature $\tau > 0$, and let $j^\star = \arg\max_j a_{t,j}$, which is also the mode of $\pi(t)$. Then

$$\log \pi_{j^\star}(t) - \frac{1}{J} \sum_{j=1}^{J} \log \pi_j(t) = \frac{\mathrm{disc}(t)}{\tau}. \tag{3}$$

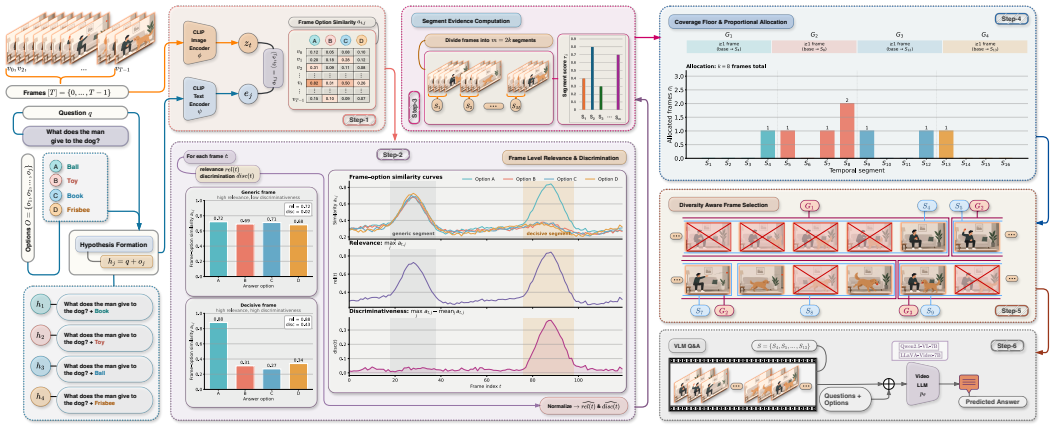

Figure 1: **FULCRUM pipeline.** (1) Frames v_t and the option hypotheses $h_j = q + o_j$ are embedded by a contrastive encoder into a frame–option similarity matrix. (2) Each frame yields two statistics, the support for the best-matching hypothesis (rel) and its log-odds over the alternatives (disc), standardized within the video. (3) They are averaged into segment scores r_i. (4) A coverage floor places one frame in each macro-cell, and concave allocation distributes the remaining budget with $1 + n_i \propto r_i - \min_j r_j + c$. (5) A greedy relevance-weighted log-determinant assignment chooses n_i non-redundant frames in each segment. (6) The k selected frames are passed to the frozen answerer in temporal order.

The partition function cancels by construction, so $\mathrm{disc}(t)$ is, up to the scale $1/\tau$, the log-odds of the favored option against the geometric mean of all options.

Within-video standardization. For a frame statistic $x : [T] \to \mathbb{R}$, let $\bar{x} = \frac{1}{T}\sum_t x(t)$ and $\sigma_x = \big(\frac{1}{T}\sum_t (x(t) - \bar{x})^2\big)^{1/2}$. We set $\tilde{x}(t) = (x(t) - \bar{x})/\sigma_x$ when $\sigma_x > 10^{-12}$, and $\tilde{x} \equiv 0$ otherwise. Standardization removes the video- and question-specific baseline of cosine similarity and leaves the within-video deviations that rank frames.

Segment score. Segment i is scored by the mean standardized statistic over its frames,

$$r_i = \frac{1}{|\mathcal{S}_i|} \sum_{t \in \mathcal{S}_i} \big(\widetilde{\mathrm{rel}}(t) + \widetilde{\mathrm{disc}}(t)\big), \qquad i = 1, \dots, m. \tag{4}$$

The scores r_i are expressed in within-video standard deviations, which fixes the units of the allocation parameter c in §3.3.

3.3 Proportional Budget Allocation

The allocation decides how many frames each segment receives, so several frames can go to a short stretch of video. Because the segment ranking of Eq. (4) locates evidence imperfectly, FULCRUM spends the budget in proportion to it, above a floor that guarantees temporal coverage.

Coverage floor. Let $k_0 = \mathrm{round}(fk)$ for a floor fraction $f \in [0, 1]$. We group the segment indices into k_0 contiguous macro-cells

$$\mathcal{G}_g = \big\{\, i : \lfloor (g-1)\, m/k_0 \rfloor < i \le \lfloor g\, m/k_0 \rfloor \,\big\}, \qquad g = 1, \dots, k_0. \tag{5}$$

Starting from $\mathbf{n} = \mathbf{0}$, each macro-cell gives one frame to its highest-scoring segment. The result is the floor allocation $\mathbf{n}^0$, under which every macro-cell, about $1/k_0$ of the video, holds at least one selected frame for every question and every score profile.

Concave allocation. The remaining $k - k_0$ frames are allocated by

$$\max_{\mathbf{n}} \sum_{i=1}^{m} w_i \log(1 + n_i) \quad \text{s.t.} \quad \sum_i n_i = k, \;\; n_i^0 \le n_i \le |\mathcal{S}_i|, \qquad w_i = r_i - \min_l r_l + c, \tag{6}$$

with $c > 0$. The objective is separable and concave, so assigning frames one at a time, starting from $\mathbf{n}^0$, to the segment with spare capacity and the largest marginal gain $w_i \log \frac{n_i+2}{n_i+1}$ attains the

exact integer optimum (Ibaraki & Katoh, 1988). The logarithm encodes diminishing returns: each additional frame in a segment adds less than the one before it.

Closed-form interpretation. Relaxing integrality, the Karush–Kuhn–Tucker conditions of Eq. (6) give, for every segment strictly between its bounds,

$$1 + n_i = w_i/\mu, \tag{7}$$

with the multiplier $\mu > 0$ set by the budget. The allocation is proportional to smoothed evidence. The constant c is the weight of the lowest-scoring segment, and since r_i is measured in within-video standard deviations, c has the same meaning in every video: $c \to \infty$ spreads the budget evenly across segments, and $c \to 0$ concentrates it on high scorers.

Relevance gate. On a flat score profile, proportional allocation would distribute frames by noise. Let $\delta_i = r_i - \min_l r_l$ and $p_i = \delta_i / \sum_l \delta_l$, with $p_i = 1/m$ when every δ_i vanishes. Flatness is measured by the normalized entropy, with $0 \log 0 = 0$,

$$H = -\frac{1}{\log m} \sum_{i=1}^{m} p_i \log p_i \ \in [0, 1]. \tag{8}$$

If $H > \theta$, the segment scores r_i and the frame relevance ρ_t of §3.4 are set to zero, so the budget is spread evenly and the assignment uses diversity alone (App. B.2).

3.4 Diversity-Aware Frame Assignment

The allocation fixes how many frames each segment receives. Scoring frames independently within a segment would spend a concentrated allocation on near-duplicate frames, so FULCRUM selects the frames jointly.

Frame relevance. The frame relevance is the standardized support, rescaled to the unit interval in magnitude,

$$\rho_t = \frac{\widetilde{\mathrm{rel}}(t)}{\max\big(10^{-9},\, \max_{s \in [T]} |\widetilde{\mathrm{rel}}(s)|\big)} \ \in [-1, 1]. \tag{9}$$

The segment score of Eq. (4) uses both statistics of Eq. (2); the frame relevance uses rel alone.

Objective. Let $\mathcal{A}$ be the frames of the allocated segments, $\kappa_i(S)$ the number of frames of S attributed to segment i, and $\hat{z}_t$ the frame embeddings projected onto their leading d principal components over $\mathcal{A}$ and rescaled to unit norm (App. B.2). We maximize

$$G(S) = \sum_{t \in S} \rho_t + \lambda \log\det\big(I + \varepsilon^{-1} K_S\big) \quad \text{s.t.} \quad S \subseteq \mathcal{A}, \ \ \kappa_i(S) = n_i \ \ \forall i, \tag{10}$$

where K_S is the Gram matrix of $\{\hat{z}_t\}_{t \in S}$, $\lambda > 0$ weights diversity, and $\varepsilon > 0$ regularizes the volume. The log-determinant is the regularized log-volume spanned by the selected features; it grows when a frame adds a direction the set does not yet span. For unit-norm features, the first frame's diversity gain is $\lambda \log(1 + \varepsilon^{-1})$, which is 0.84 in the configuration of §4 and comparable to $\rho_t \in [-1, 1]$; each later gain is smaller.

Greedy assignment. S is built greedily across all allocated segments at once. Each step admits, among the frames whose segment has not yet met its quota, the frame with the largest marginal gain in G, so every candidate's redundancy is measured against the frames already chosen in all segments. The log-determinant gain is computed exactly by rank-one updates (App. B.2).

Approximation guarantee. Every feasible set of Eq. (10) has $|S| = k$, so adding a constant to every ρ_t changes neither the maximizer nor any greedy step. With the shift $\rho_t + 1 \geq 0$, the objective is monotone submodular, and the quota constraint $\kappa_i(S) \leq n_i$ is a partition matroid, so the greedy attains at least half of the shifted optimum (Fisher et al., 1978).

4 Experimental Setup

Benchmarks and answerers. We evaluate on the full LongVideoBench validation set (Wu et al., 2024) (1,337 questions over 752 videos) and on Video-MME (Fu et al., 2024), both without subtitles. The answerers are Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and LLaVA-Video-7B (Zhang

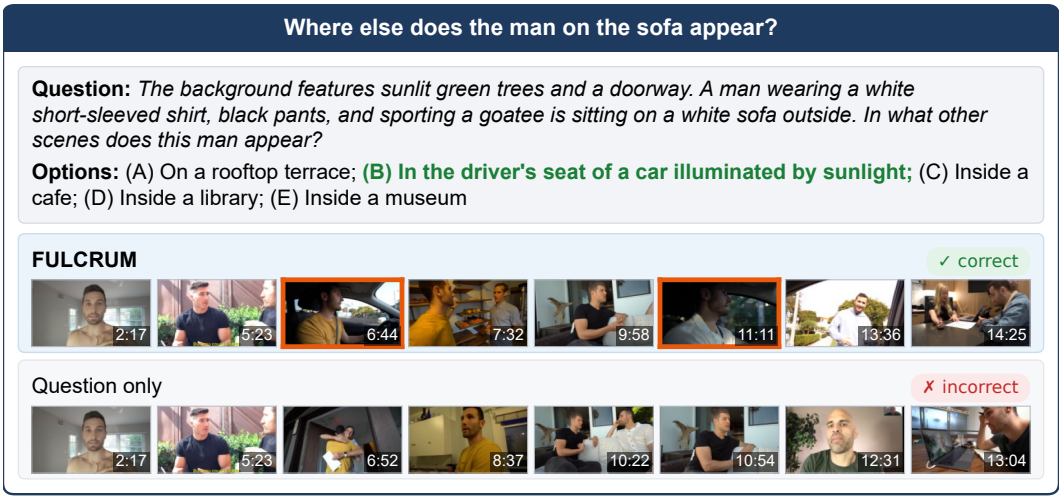

Figure 2: **Frames selected with and without the option contrast**, LongVideoBench, Qwen2.5-VL-7B, k=8. Outlined: the two frames that show the scene of the correct option (B), which question-only selection misses. More examples in App. F.1.

et al., 2024b). Answerer scale is studied separately with Qwen2.5-VL at 3B, 32B, and 72B on LongVideoBench at k=8. Benchmark statistics, frame extraction, and compute are given in App. C.

Candidate frames and answer scoring. Each video is decoded once into at most 128 candidate frames, evenly spaced in time and resized to 384×384, and every method selects from this same candidate set. The answerer receives the k selected frames in temporal order with the question and the lettered options, and is scored by the letter probabilities of Eq. (1), the log-likelihood protocol for multiple-choice evaluation (Biderman et al., 2024). A single prefill suffices, so no answer is lost to output parsing. Scored instead by generated answers with the parser of LMMs-Eval (Zhang et al., 2024a), the same frames lose 1.0 to 1.4 points, the methods keep their order, and the margin of FULCRUM over AKS changes from 6.9 to 6.8 points (App. I.1). Prompts and scored tokens are given in App. C.3.

Baselines. Uniform sampling at the same k accompanies every comparison; it is the default input policy of the answerers and a strong baseline for frame selectors (Brkic et al., 2025). The ten competing selectors of Table 1 are reimplemented from their papers (Zhang et al., 2025b; Tang et al., 2025; Liu et al., 2025; Li et al., 2026a; Zhu et al., 2026; Li et al., 2026b; Sun et al., 2025b; Zhang et al., 2025a; Wu et al., 2026; Chen et al., 2026) and run under the protocol above, with the same answerer, prompt, candidate frames, and frame budget k. Each keeps its own relevance model and hyperparameters (App. C), so the methods differ only in the frames they select.

FULCRUM configuration. FULCRUM uses one configuration on every benchmark, answerer, and budget: m=2k segments, floor f=0.5, smoothing c=0.5, gate θ=0.985, diversity weight λ=0.35, ridge ε=0.1, and projection rank d=32, with CLIP ViT-L/14 (Radford et al., 2021) as ϕ and ψ (D=768, 77 text tokens, half precision). Nothing is tuned per dataset; App. G reports the sensitivity of each parameter and of the encoder.

5 FINDINGS

5.1 MATCHED FRAME BUDGET

FULCRUM is the most accurate method in all sixteen settings (Table 1; App. D breaks the results down by question category, video duration, and question type). On LongVideoBench, AdaRD-Key is the strongest competitor in all eight settings, and FULCRUM leads it by 2.0 to 5.6 points. On Video-MME no single competitor is strongest (WFS-SB leads in three settings, MaxInfo in one, and AdaRD-Key in four), and FULCRUM leads each setting by 1.3 to 5.7 points. Over uniform sampling,

Table 1: **Accuracy (%) across frame budgets.** Each method selects k frames from the same candidate pool and feeds them to the same answerer under the same evaluation protocol. Best in each column in bold, second best underlined; subscripts give FULCRUM's margin over the strongest competitor. Competitors are grouped by how they select (Table 6): in a single pass, or iteratively, running the encoder or the answerer in rounds.

	Method	Venue	LongVideoBench				Video-MME			
	frames k		4	8	16	32	4	8	16	32
	Qwen2.5-VL-7B									
	Uniform	–	42.7	44.2	48.8	52.9	39.6	43.3	46.5	51.2
Single pass	AdaRD-Key (Zhang et al., 2025b)	–	55.2	59.4	60.8	60.1	49.3	54.5	57.9	58.2
Single pass	AKS (Tang et al., 2025)	CVPR'25	53.2	55.6	56.5	56.5	51.2	54.6	55.3	57.2
Single pass	BOLT (Liu et al., 2025)	CVPR'25	49.8	52.1	56.2	58.5	48.9	51.7	55.8	58.1
Single pass	MaxInfo (Li et al., 2026b)	WACV'26	52.5	56.8	58.0	57.3	51.3	55.0	56.6	57.1
Single pass	MDP^3 (Sun et al., 2025b)	ICCV'25	51.8	55.9	57.1	57.2	50.9	54.6	56.0	56.5
Single pass	Q-Frame (Zhang et al., 2025a)	ICCV'25	52.2	54.6	56.8	55.9	50.5	53.9	55.8	56.2
Single pass	VTR-VLM (Wu et al., 2026)	ICLR'26	49.6	50.4	51.9	52.3	48.4	50.0	52.1	52.8
Single pass	WFS-SB (Chen et al., 2026)	CVPR'26	50.9	54.8	56.8	58.1	51.8	54.7	57.1	58.3
Iterative	DIG (Li et al., 2026a)	CVPR'26	50.8	54.9	56.1	55.2	50.1	53.4	55.3	55.9
Iterative	FOCUS (Zhu et al., 2026)	ICLR'26	50.5	52.9	56.5	57.8	49.8	53.1	55.7	56.8
Iterative	FULCRUM *(Ours)*	–	**60.8**$_{+5.6}$	**62.5**$_{+3.1}$	**63.2**$_{+2.4}$	**63.2**$_{+3.1}$	**57.5**$_{+5.7}$	**59.8**$_{+4.8}$	**61.2**$_{+3.3}$	**60.8**$_{+2.5}$
	LLaVA-Video-7B									
	Uniform	–	44.4	47.3	51.2	53.8	41.6	45.1	48.3	52.6
Single pass	AdaRD-Key (Zhang et al., 2025b)	–	57.8	59.5	61.1	62.2	55.2	58.3	59.6	60.1
Single pass	AKS (Tang et al., 2025)	CVPR'25	51.5	55.2	56.8	56.1	49.5	53.1	55.6	55.6
Single pass	BOLT (Liu et al., 2025)	CVPR'25	48.3	51.8	55.2	55.4	47.6	51.2	54.9	55.1
Single pass	MaxInfo (Li et al., 2026b)	WACV'26	54.1	57.9	58.6	58.7	53.2	56.7	57.7	58.2
Single pass	MDP^3 (Sun et al., 2025b)	ICCV'25	54.9	57.8	59.1	59.2	52.8	56.4	57.8	58.3
Single pass	Q-Frame (Zhang et al., 2025a)	ICCV'25	53.8	56.8	57.7	57.8	52.6	55.8	57.0	57.5
Single pass	VTR-VLM (Wu et al., 2026)	ICLR'26	50.1	52.6	54.1	54.2	49.5	52.6	54.3	55.0
Single pass	WFS-SB (Chen et al., 2026)	CVPR'26	52.1	58.9	59.2	60.3	51.7	58.6	58.9	60.0
Iterative	DIG (Li et al., 2026a)	CVPR'26	52.2	55.4	58.6	59.6	51.2	54.8	57.1	58.2
Iterative	FOCUS (Zhu et al., 2026)	ICLR'26	51.2	55.8	57.0	59.5	50.9	55.3	56.9	58.4
Iterative	FULCRUM *(Ours)*	–	**62.5**$_{+4.7}$	**62.5**$_{+3.0}$	**64.2**$_{+3.1}$	**64.2**$_{+2.0}$	**60.4**$_{+5.2}$	**61.5**$_{+2.9}$	**61.5**$_{+1.9}$	**61.4**$_{+1.3}$

FULCRUM gains 10.3 to 18.3 points on LongVideoBench and 8.8 to 18.8 on Video-MME. In every answerer–benchmark pair, the margin over the strongest competitor is largest at k=4 (§5.3).

5.2 WHAT THE OPTIONS CONTRIBUTE

The contrast term carries most of the gain. Fusing the options into one query string gains 0.8 points on LongVideoBench and 1.1 on Video-MME over the question alone (Table 2, App. F). Scoring each option separately and keeping the best match (rel) gains 1.1 and 1.4. Adding the contrast term brings FULCRUM to 62.8 and 61.1, 3.5 and 3.3 points above the question alone, of which disc contributes 2.4 and 1.9 (Fig. 2 shows an example).

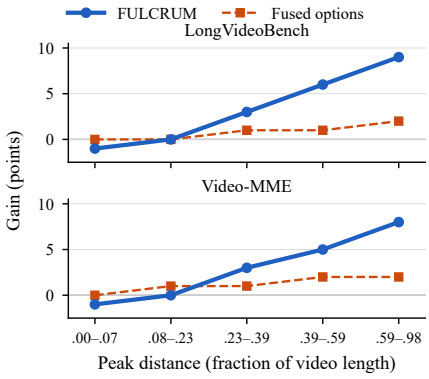

Figure 3: **Where the options help.** Gain over question-only relevance, for FULCRUM and for fused options, binned by peak distance.

The gain comes from the question's own options. Options taken from a question about another video lower accuracy below the question alone, by 1.6 points on LongVideoBench and 2.9 on Video-MME. A paraphrase of the question, which rewords it without naming any option, also falls below it, by 1.3 and 0.3. Shuffling the option order leaves the selection unchanged, since rel and disc are invariant to it. Options that the answerer writes from the question alone, without the video, recover part of the gain: 1.1 to 1.7 points over the question alone across both answerers and benchmarks, against 2.6 to 3.5 with the benchmark's options (Table 23).

The gain grows with peak distance. Binned by peak distance, the distance between the frame most similar to the question and the frame with the largest option contrast, the gain of FULCRUM over question-only selection grows from within one point where the two peaks coincide to 9 points on

Table 2: **What text the relevance score reads.** Only the text embedded for relevance changes; the rest of §3, k=8, and the answerer are fixed. Subscripts: change in points over question-only relevance.

Text embedded	LongVideoBench	Video-MME
Uniform (no relevance score)	44.2	43.3
Relevance text		
Question only	59.3	57.8
Question + all options, fused into one string	60.1 +0.8	58.9 +1.1
Question + each option, best match (rel)	60.4 +1.1	59.2 +1.4
rel + disc, own options (FULCRUM)	**62.8** +3.5	**61.1** +3.3
Controls		
Question + another video's options	57.7 −1.6	54.9 −2.9
Paraphrase of the question	58.0 −1.3	57.5 −0.3
Own options, order shuffled	**62.8** +3.5	**61.1** +3.3

Table 3: **Language prior.** Split by answerer correctness without frames, on LongVideoBench with Qwen2.5-VL-7B at k=8; subscripts: margin over AKS.

Method	Correct (37%)	Incorrect (63%)
Uniform	65.9	31.4
AKS	78.0	42.3
FULCRUM	**81.0** +3.0	**51.5** +9.2

Table 4: **Answerer scale.** Qwen2.5-VL from 3B to 72B, set up as in Table 3.

Method	3B	7B	32B	72B
No frames	34.8	37.1	41.5	44.0
Uniform	40.1	44.2	50.3	53.6
AKS	50.2	55.6	59.0	61.2
FULCRUM	**56.0** +5.8	**62.5** +6.9	**65.2** +6.2	**67.0** +5.8

LongVideoBench and 8 on Video-MME in the farthest bin, while fused options stay within 2 points of the question alone throughout (Fig. 3, App. F.3).

The gain does not come from the language prior. Without frames, Qwen2.5-VL-7B answers 37% of the LongVideoBench questions correctly. On the other 63%, where the frames must supply the answer, FULCRUM reaches 51.5%, 9.2 points above AKS and 20.1 above uniform sampling. These questions carry about 84% of the margin of FULCRUM over AKS (Table 3, App. I.2).

5.3 FRAME BUDGET

Selection matters most when frames are scarce. At k=4, FULCRUM exceeds uniform sampling by 17.9 to 18.8 points across the four answerer–benchmark pairs, and at k=32 by 8.8 to 10.4 (Fig. 5). From 4 to 32 frames, uniform accuracy rises by 9.4 to 11.6 points and FULCRUM by 1.0 to 3.3, and the best competitor's margin over uniform sampling shrinks from 12.2–13.6 to 7.1–8.4 points.

Selection matters more as videos get longer. On LongVideoBench at k=4 with Qwen2.5-VL-7B, uniform accuracy falls by 27.6 points from the shortest duration group (8–15 s) to the longest (15–60 min), and FULCRUM by 6.3. Its margin over uniform sampling grows from 2.2 to 23.5 points (App. D.3).

Four frames chosen by FULCRUM exceed 32 uniform frames and rival 32 frames from the best competitor. At 32 frames, uniform sampling remains 6.3 to 8.7 points below FULCRUM at 4 frames. FULCRUM at 4 frames is ahead of the best of the ten competitors at 32 frames in three of the four pairs and 0.8 points behind in the fourth, the dotted line of Fig. 5 (App. E).

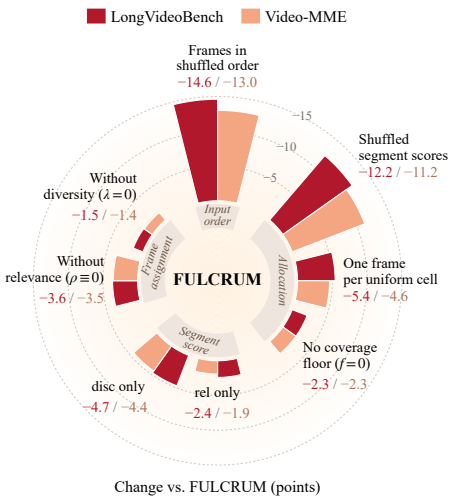

Figure 4: **Ablations** at k=8 with Qwen2.5-VL-7B: the change in accuracy, in points, when one part of FULCRUM is removed or perturbed. Accuracies are in Table 24 (App. G).

5.4 COMPONENTS

Every component contributes. Fig. 4 removes or perturbs one part of FULCRUM at a time, at k=8 with Qwen2.5-VL-7B, and every change costs accuracy on both benchmarks. The segment scores carry the allocation: permuting them across segments keeps the shape of the allocation and costs 12.2 and 11.2 points, while one frame per uniform cell gives even coverage and costs 5.4 and 4.6.

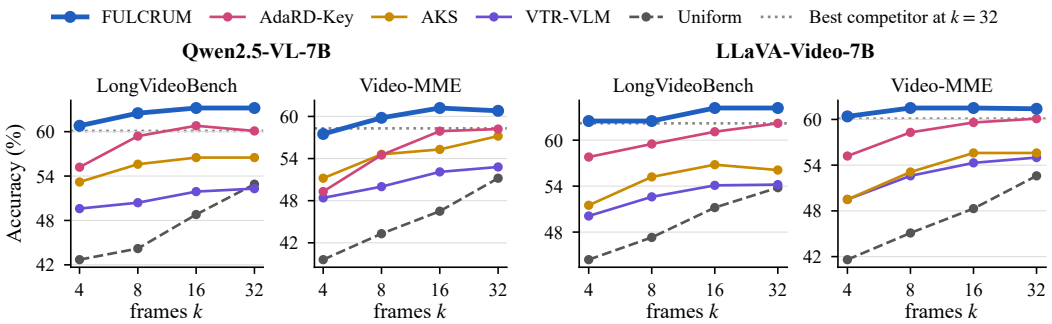

Figure 5: **Selected frames against uniform frames.** Accuracy against frame budget k for uniform sampling, AKS, AdaRD-Key, VTR-VLM, and FULCRUM, per answerer and benchmark. The dotted line is the best of all ten competitors at k=32 (Table 18, App. E). Values are those of Table 1; iso-frame comparisons for every method are in Tables 19 and 20.

Both segment statistics are needed: rel alone costs 2.4 and 1.9 points, and disc alone 4.7 and 4.4. Removing the coverage floor costs 2.3 points on both benchmarks, relevance in the fill 3.6 and 3.5, and diversity 1.5 and 1.4. The answerer also relies on temporal order: the same frames in shuffled order cost 14.6 and 13.0, and every result in this paper presents frames in order. One configuration holds across parameters and encoders: every parameter keeps accuracy within 1.5 points of the locked configuration except at three extremes, and with each of six encoders the option contrast adds 2.7 to 3.2 points on LongVideoBench and 1.6 to 2.1 on Video-MME (App. G).

6 DISCUSSION

Scope. FULCRUM requires answer options, so open-ended questions fall outside it. Options generated by the answerer from the question alone recover a third to a half of the gain (Table 23), which suggests a route to open-ended settings. Both answerers in the main comparison have 7B parameters; with Qwen2.5-VL from 3B to 72B, FULCRUM remains the most accurate method and its margin over AKS stays between 5.8 and 6.9 points (Table 4, App. H). The choice of frames outweighs a tenfold larger answerer: the 7B answerer with FULCRUM (62.5) is more accurate than the 72B answerer with AKS (61.2).

Limitations. FULCRUM separates only options that its encoder can tell apart in a single frame. It fails when the options differ in direction of motion or in left against right, when the question locates its moment by a spoken line, when every option matches the video equally, and when a wrong option names an object visible throughout the video (App. J). The advantage over uniform sampling narrows from 18.3 points at 7B to 13.4 at 72B (Table 4), so stronger answerers absorb part of the benefit of frame selection.

7 CONCLUSION

We introduced FULCRUM, a training-free frame selector for multiple-choice video question answering that scores each frame by the margin of the best-supported option over the average option and spends the frame budget in proportion to that evidence, using one pass of a frozen image–text encoder and no query to the answerer. Against uniform sampling and ten published selectors, FULCRUM is the most accurate method at every budget from 4 to 32 frames on two benchmarks with two answerers; four frames chosen by FULCRUM outperform thirty-two uniform frames, and with it a 7B answerer exceeds a 72B answerer with AKS. The gain comes from contrasting the question's own options: adding them without the contrast gains at most 1.4 points over the question alone, against 3.3 to 3.5 with it, and most of the margin falls on questions the answerer misses without frames.

AI USE STATEMENT

AI tools were used to aid and polish the writing of this manuscript, including grammatical correction, clarity improvements, and final proofreading. AI tools were not used to propose the method, design the experiments, or interpret the results. The authors reviewed and verified all AI-assisted text and take full responsibility for the content of this paper.

ETHICS STATEMENT

We adhere to the ICLR Code of Ethics. This work evaluates frame selection on two public benchmarks, LongVideoBench and Video-MME, used under their research licenses and as released by their authors. We collect no new data, involve no human subjects or annotators, and train no model. FULCRUM scores frames with a pretrained image–text encoder and may inherit that encoder's biases; it only chooses which frames a video-language model sees, and adds no capability beyond that of the answerer it feeds.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe every component needed to rerun our experiments. FULCRUM is deterministic: it has no sampling step and breaks every tie toward the lowest index, so a video, a question, and its options always yield the same frames. The single configuration used throughout is given in §4, the pseudocode in App. B.1, and the implementation details in App. B.2. Candidate frames, prompts, scored tokens, and competitor reimplementations are documented in App. C, and model checkpoints and library versions in Table 10. All benchmarks and models are publicly available, and anonymized code is provided in the supplementary material.

REFERENCES

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. *arXiv preprint arXiv:2502.13923*, 2025.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

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

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)*, 2026. arXiv:2603.00512.

Marshall L. Fisher, George L. Nemhauser, and Laurence A. Wolsey. An analysis of approximations for maximizing submodular set functions—II. *Mathematical Programming Study*, 8:73–87, 1978.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. arXiv:2405.21075.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. FrameFusion: Combining similarity and importance for video token reduction on large vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b. arXiv:2501.01986.

Weiyu Guo, Ziyang Chen, Shaoguang Wang, Jianxiang He, Yijie Xu, Jinhui Ye, Ying Sun, and Hui Xiong. Logic-in-Frames: Dynamic keyframe search via visual Semantic-Logical verification for long video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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

Jeongseok Hyun, Sukjun Hwang, Su Ho Han, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Joon-Young Lee, Seon Joo Kim, and Minho Shim. Multi-Granular Spatio-Temporal token merging for Training-Free acceleration of video LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2507.07990.

Toshihide Ibaraki and Naoki Katoh. *Resource Allocation Problems: Algorithmic Approaches*. MIT Press, 1988.

Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a VLM. *IEEE Access*, 12, 2024. arXiv:2403.18406.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2408.03326.

Jialuo Li, Bin Li, Jiahao Li, and Yan Lu. Divide, then ground: Adapting frame selection to query types for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a. arXiv:2512.04000.

Pengyi Li, Irina Abdullaeva, Alexander Gambashidze, Andrey Kuznetsov, and Ivan Oseledets. MaxInfo: A Training-Free Key-Frame selection method using maximum volume for enhanced video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026b.

Zongyao Li, Kengo Ishida, Satoshi Yamazaki, Xiaotong Ji, and Jianquan Liu. KFS-Bench: Comprehensive evaluation of key frame sampling in long video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026c. arXiv:2512.14017.

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)*, 2025. arXiv:2503.21483.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned Retrieval-Augmented long video comprehension. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2411.13093.

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. *arXiv preprint arXiv:2603.25072*, 2026.

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient frame selection for long video understanding via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16944–16953, 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 *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Dilip Sarkar, Md. Safayet Islam, and Liang Liang. Evaluation of MLLM-agnostic plug-and-play keyframe selection methods for long video understanding. *arXiv preprint arXiv:2609.13250*, 2026.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. HoliTom: Holistic token merging for fast video large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. FastVID: Dynamic density pruning for fast video large language models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2503.11187.

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. *arXiv preprint arXiv:2510.02262*, 2025a.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP3: A training-free approach for list-wise frame selection in Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b. arXiv:2501.02885.

Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. TSPO: Temporal sampling policy optimization for long-form video language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

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)*, 2025. arXiv:2502.21271.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. DyCoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. AdaReTaKe: Adaptive redundancy reduction to perceive longer for video-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. arXiv:2405.19209.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved Video-Language understanding. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024. arXiv:2407.15754.

Zhirong Wu, Xiaodong Wang, Langling Huang, Teng Xu, and Peixi Peng. A training-free framework for long video understanding via Video-Query-Options similarity. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=hfMfYMoRLk`.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-MTR: Reinforced multi-turn reasoning for long video understanding. In *International Conference on Machine Learning (ICML)*, 2026.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding. In *Proceedings of the Findings of the Association for Computational Linguistics (ACL Findings)*, 2025a. arXiv:2503.09146.

Yifeng Yao, Yike Yun, Jing Wang, Huishuai Zhang, Dongyan Zhao, Ke Tian, Zhihao Wang, Minghui Qiu, and Tao Wang. K-frames: Scene-Driven any-k keyframe selection for long video understanding. *arXiv preprint arXiv:2510.13891*, 2025b.

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. T*: Re-thinking temporal search for Long-Form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2504.02259.

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. Frame-Voyager: Learning to query frames for video large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.03226.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025a.

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)*, 2025b. arXiv:2506.22139.

Xian Zhang, Zexi Wu, Zinuo Li, Hongming Xu, Luqi Gong, Farid Boussaid, Naoufel Werghi, and Mohammed Bennamoun. AdaRD-key: Adaptive Relevance-Diversity keyframe sampling for long-form video understanding. *arXiv preprint arXiv:2510.02778*, 2025c.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025d. arXiv:2505.18079.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *Transactions on Machine Learning Research (TMLR)*, 2025e. arXiv:2410.02713.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2406.04264.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.27280.

Junbo Zou, Ziheng Huang, Shengjie Zhang, Liwen Zhang, and Weining Shen. VideoBrain: Learning adaptive frame sampling for long video understanding. *arXiv preprint arXiv:2602.04094*, 2026a.

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 *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2510.04428.

A Extended Related Work **15**

B Method Details **16**

B.1 Pseudocode 16

B.2 Implementation Details 17

B.3 Complexity 18

C Protocol and Reproducibility **18**

C.1 Benchmarks 18

C.2 Candidate Frames and Answerer Input 18

C.3 Prompt and Answer Scoring 18

C.4 Competitor Implementations 19

C.5 Compute and Software 19

D Full Results **20**

D.1 Every Method, Answerer, and Budget 20

D.2 By Question Category 20

D.3 By Video Duration 21

D.4 By Question Type across Budgets 21

E Frame Budget **23**

F What the Options Contribute **27**

F.1 Lexical Controls 27

F.2 Which Frames Each Text Selects 27

F.3 Peak Distance 27

F.4 Options Generated by the Answerer 27

G Components and Sensitivity **30**

G.1 Ablations 30

G.2 One Parameter at a Time 30

G.3 Two Parameters at a Time 30

G.4 Entropy Gate 30

G.5 Encoders 30

H Answerer Scale **32**

I Evaluation Checks **33**

I.1 Answer Scoring 33

I.2 Language Prior 33

J Failure Analysis **34**

A Extended Related Work

Table 6 places 24 published frame selectors and FULCRUM along six axes: training, the text each frame is scored against, the selection unit, how the budget is spread over the video, whether an explicit diversity term is used, and what runs during selection. Table 5 restricts the comparison to the ten selectors evaluated in Table 1. FULCRUM is the only entry that scores frames by the contrast between answer options.

Table 5: **Positioning against the ten evaluated selectors.** ✓ present, ∼ partial, ✗ absent.

Method	Reads options	Option contrast	Allocation	Coverage floor	Set diversity
AKS	✗	✗	✓	∼	✗
BOLT	✗	✗	∼	✗	✗
Q-Frame	✗	✗	✗	✗	✗
DIG	✗	✗	✗	∼	✗
FOCUS	✗	✗	✓	✗	✗
MaxInfo	✗	✗	✗	✗	✓
MDP^3	✗	✗	✓	✗	✓
AdaRD-Key	✗	✗	✗	✗	✓
WFS-SB	✗	✗	✓	✗	✓
VTR-VLM	∼	✗	✓	✗	✗
FULCRUM	✓	✓	✓	✓	✓

Table 6: **Frame selectors by design axis.** FULCRUM is the only selector that scores frames by the contrast between options. *Scored against*: the text each frame's relevance is computed against. *Budget*: how frames are spread over the video. *Div.*: an explicit diversity or redundancy term. *Selection*: what runs during selection beyond one encoder pass over the candidates. SFT: supervised fine-tuning; RL: reinforcement learning.

	Method	Training	Scored against	Unit	Budget	Div.	Selection
Training-free, single pass	AKS	none	question	frame	recursive split, relevance vs. coverage	✗	one pass
	BOLT	none	question	frame	inverse-transform sampling	✗	one pass
	Q-Frame	none	question	frame	top-ranked, resolution by rank	✗	one pass
	MaxInfo	none	no text	frame	maximum embedding volume	✓	one pass
	MDP^3	none	question	frame	dynamic program over segments	✓	one pass
	AdaRD-Key	none	question	frame	relevance plus diversity, gated	✓	one pass
	WFS-SB	none	question	frame	importance-weighted clip budgets	✓	one pass
	F^{2}C	none	question	clip	anchored clips, adaptive length	✓	one pass
	QAaF	none	question + each option	frame	top-ranked	✗	one pass
	VTR-VLM	none	question + generated options, fused	frame	density by segment relevance	✗	answerer writes options
	FULCRUM	none	question + options, contrast	frame	concave allocation above a floor	✓	one pass
Training-free, iterative	FOCUS	none	question	clip, frame	bandit over clips	✗	encoder, in rounds
	GIFT	none	question	frame	batches, re-scored	✓	encoder, in rounds
	T*	none	objects in the question	frame	iterative search	✗	detector, in rounds
	Logic-in-Frames	none	objects and relations	frame	iterative search	✗	detector, in rounds
	VideoTree	none	question	cluster	adaptive tree	✗	LLM over captions
	A.I.R.	none	question	frame	high-potential batches	✗	VLM, in rounds
	DIG	none	question	segment	uniform for global queries	✗	answerer, in rounds
Learned	Frame-Voyager	ranking	question	frame set	learned combination ranking	✗	one pass
	M-LLM Select	pseudo-labels	question	frame	learned relevance	✗	one pass
	GenS	fine-tuning	question	span	scored time spans	✗	one pass
	K-frames	SFT + RL	question	clip	any-k within key clips	✗	one pass
	TSPO	RL	question	frame	probabilistic selection	✗	one pass
	Qin et al.	pre-train + RL	question	frame	top-ranked	✗	one pass
	VideoBrain	RL	question	frame	retrieval and uniform agents	✗	VLM, in rounds

Training-free, single pass. These selectors share the setting of FULCRUM: frozen models score the candidate frames once, and the answerer is not called during selection, except by VTR-VLM to write candidate options. They differ in how they spend the budget. AKS splits the video recursively and trades relevance against temporal coverage (Tang et al., 2025); BOLT samples frames by inverse transform over query similarity (Liu et al., 2025); Q-Frame keeps the top-ranked frames and assigns their resolution by rank (Zhang et al., 2025a). MDP^3 selects within segments by a query-conditioned determinantal point process and sizes the segments by dynamic programming (Sun et al., 2025b). WFS-SB segments the video at semantic boundaries found by a wavelet transform, gives each clip a budget by an importance score, and selects within clips by maximal marginal relevance (Chen et al.,

2026). F^2C selects clips instead of frames and sets each clip's length by trading relevance against redundancy (Sun et al., 2025a). MaxInfo ignores the query and maximizes the volume spanned by the selected embeddings (Li et al., 2026b); AdaRD-Key adds a log-determinant diversity term to relevance and falls back to diversity alone when relevance is weak (Zhang et al., 2025b). Two selectors read the answer options. VTR-VLM fuses the question with candidate options generated by the answerer into one text, and raises sampling density and resolution in the segments that match it (Wu et al., 2026). QAaF appends each option to the question, computes the alignment of every frame with every resulting text, and keeps the frames with the highest alignment (Sarkar et al., 2026). Their scoring rules correspond to the fused and best-match rows of Table 2, which gain 0.8–1.4 points over the question alone, against 3.3–3.5 for the contrast of FULCRUM.

Training-free, iterative. A second group spends extra computation during selection. FOCUS treats clips as bandit arms and scores only the most promising ones (Zhu et al., 2026). GIFT scores each frame's irreplaceability, its relevance conditioned on its difference from other frames, and re-scores the remaining frames after each selected batch (Ma et al., 2026). T* and Logic-in-Frames search the video for the objects, and for the spatial, temporal, and causal relations, that the question names, using an object detector (Ye et al., 2025; Guo et al., 2025). VideoTree clusters the video and lets an LLM rate cluster captions to decide where to expand (Wang et al., 2025b). A.I.R. passes small batches of high-potential frames to a VLM for analysis until the budget is spent (Zou et al., 2026b). DIG lets an LLM classify the query as global or localized, samples uniformly for global queries, and for localized ones has the answerer score one frame per segment, iterating until the set is stable (Li et al., 2026a).

Learned selectors. Frame-Voyager learns to rank frame combinations by the prediction loss of a reference Video-LLM (Yu et al., 2025). M-LLM Select trains a lightweight selector on two pseudo-labels, single-frame importance from an M-LLM and multi-frame selections from an LLM over frame captions (Hu et al., 2025). GenS fine-tunes a VLM to output relevant time spans with scores (Yao et al., 2025a), and K-frames trains a model by supervised fine-tuning and reinforcement learning to predict key clips, from which any number of frames is drawn (Yao et al., 2025b). TSPO trains an event-aware temporal agent jointly with the answer by group-relative policy optimization (Tang et al., 2026). Qin et al. (2026) pre-train a frame–query relevance model and refine it by reinforcement learning with frame-level and combination-level rewards. VideoBrain trains a VLM to call a CLIP-based retrieval agent and a uniform sampling agent over several rounds and to decide when the gathered frames suffice (Zou et al., 2026a).

How selection is evaluated. Most selectors are compared by answer accuracy at a fixed frame budget, the protocol of §4. Three studies examine the protocol itself. Frame-Sampling-Strategies finds uniform sampling at a fixed frame rate the best strategy for four small VLMs on Video-MME (Brkic et al., 2025). KFS-Bench annotates the scenes each question requires, identifies sampling precision, scene coverage, and sampling balance as the factors that drive accuracy, and proposes a sampler that balances question–frame similarity against diversity (Li et al., 2026c). A comparison of five plug-and-play selectors across three answerers and three benchmarks ranks QAaF first in 13 of 15 settings and FOCUS second (Sarkar et al., 2026). LV-Haystack measures temporal search directly, by the overlap of selected frames with annotated evidence (Ye et al., 2025). Uniform frames remain a strong input even without selection: six of them arranged in one grid image support zero-shot video question answering (Kim et al., 2024).

B METHOD DETAILS

B.1 PSEUDOCODE

Algorithm 1 gives FULCRUM end to end for one question, and marks the section of §3 that defines each stage. Every $\arg\max$, in the floor, the concave allocation, and the greedy assignment, breaks ties toward the lowest index. FULCRUM contains no sampling step, so the selection is a deterministic function of the video, the question, and the options.

Algorithm 1 FULCRUM. All $\arg\max$ operations break ties toward the lowest index.

Require: Frames $v_0, \dots, v_{T-1}$; question q; options $o_1, \dots, o_J$; budget k; segments m; floor fraction f; smoothing c; gate θ; diversity weight λ; ridge ε; rank d
Ensure: Frame indices $S \subset [T]$ with $|S| = k$, in increasing order

Relevance (§3.2)
$z_t \leftarrow \phi(v_t)/\|\phi(v_t)\|_2$ for $t \in [T]$; $e_j \leftarrow \psi(h_j)/\|\psi(h_j)\|_2$ for $j = 1, \dots, J$ ▷ one encoder pass
$a_{t,j} \leftarrow \langle z_t, e_j \rangle$; $\mathrm{rel}(t) \leftarrow \max_j a_{t,j}$; $\mathrm{disc}(t) \leftarrow \mathrm{rel}(t) - \frac{1}{J}\sum_j a_{t,j}$
$\widetilde{\mathrm{rel}}, \widetilde{\mathrm{disc}} \leftarrow$ within-video standardization, zero where $\sigma \le 10^{-12}$
$\mathcal{S}_i \leftarrow \mathcal{B}_i^{(m)}$; $r_i \leftarrow \mathrm{mean}_{t \in \mathcal{S}_i}\big(\widetilde{\mathrm{rel}}(t) + \widetilde{\mathrm{disc}}(t)\big)$ for $i = 1, \dots, m$
$\rho_t \leftarrow \widetilde{\mathrm{rel}}(t)/\max\big(10^{-9}, \max_s |\widetilde{\mathrm{rel}}(s)|\big)$ for $t \in [T]$
Gate (§3.3)
$\delta_i \leftarrow r_i - \min_l r_l$; $p_i \leftarrow \delta_i / \sum_l \delta_l$ if $\sum_l \delta_l > 10^{-12}$, else $p_i \leftarrow 1/m$
$H \leftarrow -\sum_i p_i \log p_i / \log m$
if $H > \theta$ **then**
  $r \leftarrow \mathbf{0}$; $\rho \leftarrow \mathbf{0}$ ▷ flat profile: coverage and diversity only
end if
Allocation (§3.3)
$k_0 \leftarrow \mathrm{round}(fk)$; $\mathbf{n} \leftarrow \mathbf{0}$
for $g = 1, \dots, k_0$ **do**
  $i^\star \leftarrow \arg\max\{r_i : i \in \mathcal{G}_g,\ n_i < |\mathcal{S}_i|\}$; $n_{i^\star} \leftarrow n_{i^\star} + 1$ ▷ coverage floor
end for
$w_i \leftarrow r_i - \min_l r_l + c$ for $i = 1, \dots, m$
while $\sum_i n_i < k$ **do**
  $i^\star \leftarrow \arg\max_{i:\, n_i < |\mathcal{S}_i|}\ w_i \log \frac{n_i+2}{n_i+1}$; $n_{i^\star} \leftarrow n_{i^\star} + 1$ ▷ exact concave allocation
end while
Assignment (§3.4)
$\mathcal{A} \leftarrow \bigcup_{i:\, n_i > 0} \mathcal{S}_i$; $\iota(t) \leftarrow \max\{i : t \in \mathcal{S}_i,\ n_i > 0\}$ for $t \in \mathcal{A}$
if $D > d$ **then**
  $\hat{z}_t \leftarrow$ scores on the leading $\min(d, |\mathcal{A}|)$ principal components of $\{z_t\}_{t \in \mathcal{A}}$, centered on their mean
else
  $\hat{z}_t \leftarrow z_t$
end if
$\hat{z}_t \leftarrow \hat{z}_t / \max\big(10^{-12}, \|\hat{z}_t\|_2\big)$; $M^{-1} \leftarrow I$; $S \leftarrow \varnothing$
while $|S| < k$ **and** an admissible frame exists **do**
  $t^\star \leftarrow \arg\max_{t \in \mathcal{A} \setminus S:\ \kappa_{\iota(t)}(S) < n_{\iota(t)}}\ \rho_t + \lambda \log\big(1 + \varepsilon^{-1} \hat{z}_t^\top M^{-1} \hat{z}_t\big)$
  $u \leftarrow M^{-1}\hat{z}_{t^\star}$; $M^{-1} \leftarrow M^{-1} - uu^\top/(\varepsilon + \hat{z}_{t^\star}^\top u)$; $S \leftarrow S \cup \{t^\star\}$ ▷ Sherman–Morrison
end while
while $|S| < k$ **do**
  $S \leftarrow S \cup \{\min([T] \setminus S)\}$ ▷ only under shared boundary frames
end while
return S in increasing order

B.2 IMPLEMENTATION DETAILS

Segments and uniform cells. A partition of $[T]$ into P parts has parts

$$\mathcal{B}_i^{(P)} = \Big\{ t \in [T] :\ \big\lfloor (i-1)T/P \big\rfloor \ \le\ t\ \le\ \min\big(\lceil iT/P \rceil - 1,\ T-1\big) \Big\}, \qquad i = 1, \dots, P. \tag{11}$$

The segments are $\mathcal{S}_i = \mathcal{B}_i^{(m)}$, and the k uniform cells are $\mathcal{C}_c = \mathcal{B}_c^{(k)}$. When P divides T, the parts are disjoint and contain T/P frames each; otherwise consecutive parts share their boundary frame, which then enters the score of both segments. With $m = 2k$, each segment lies inside one uniform cell. The uniform set of size $n \le T$ is $U^{(n)} = \{\lfloor (c - \frac{1}{2})\, T/n \rfloor : c = 1, \dots, n\}$, whose n elements are distinct because consecutive values differ by at least one; $U^{(k)}$ is the frame set of the uniform baseline.

Gate fallback. When the gate fires ($H > \theta$), the floor selects the lowest-index segment of each macro-cell, Eq. (6) spreads the remaining frames under equal weights $w_i = c$, and the assignment uses diversity alone. At $\theta = 0.985$ the gate fires only when H is within 1.5% of its maximum, that is, when the segment scores are close to uniform and carry no usable ranking.

Candidates and attribution. The candidate set is $\mathcal{A} = \bigcup_{i:\, n_i>0} \mathcal{S}_i$. Each candidate is attributed to one segment, $\iota(t) = \max\{i : t \in \mathcal{S}_i,\ n_i > 0\}$, which matters only for boundary frames shared by two allocated segments. For a set S, the number of its frames attributed to segment i is $\kappa_i(S) = |\{s \in S : \iota(s) = i\}|$.

Feature space. When $D > d$, the candidate embeddings are centered on their mean over $\mathcal{A}$, a thin singular value decomposition of the resulting $|\mathcal{A}| \times D$ matrix is computed, and the scores on the leading $d' = \min(d, |\mathcal{A}|)$ principal components are kept. When $D \le d$, the embeddings are used directly. Each resulting vector is rescaled to unit norm, with norms floored at 10^{-12}, giving $\hat{z}_t \in \mathbb{R}^{d'}$. Centering removes the appearance component shared by the candidate frames, so the volume measures variation within the video.

Greedy with rank-one updates. Starting from $S = \varnothing$ and $M^{-1} = I_{d'}$, a frame t is admissible when $t \in \mathcal{A} \setminus S$ and $\kappa_{\iota(t)}(S) < n_{\iota(t)}$. Each step admits the admissible frame with the largest marginal gain $\rho_t + \lambda\Delta_t$, where the marginal log-determinant gain is exact,

$$\Delta_t = \log\bigl(1 + \varepsilon^{-1}\hat{z}_t^\top M_S^{-1}\hat{z}_t\bigr), \qquad M_S = I_{d'} + \varepsilon^{-1}\sum_{s\in S}\hat{z}_s\hat{z}_s^\top, \tag{12}$$

by the matrix determinant lemma. After admitting $t^\star$, M_S^{-1} is updated by Sherman–Morrison,

$$M^{-1} \leftarrow M^{-1} - \frac{M^{-1}\hat{z}_{t^\star}\hat{z}_{t^\star}^\top M^{-1}}{\varepsilon + \hat{z}_{t^\star}^\top M^{-1}\hat{z}_{t^\star}}. \tag{13}$$

The greedy stops when $|S| = k$ or no admissible frame remains. The second case occurs only when shared boundary frames leave a segment with fewer attributed candidates than its quota; S is then completed with the lowest-index frames of $[T] \setminus S$. The output is S in increasing order.

B.3 Complexity

After the encoder pass over the T frames and J hypothesis strings, the similarity matrix costs $O(TJD)$, the segment scores and the gate cost $O(T + m)$, and the allocation costs $O(km)$. The projection is a thin singular value decomposition (SVD) of the $|\mathcal{A}| \times D$ candidate matrix, costing $O(|\mathcal{A}|\, D\, \min(|\mathcal{A}|, D))$, and the greedy costs $O(k\,|\mathcal{A}|\,d'^2)$ through Eqs. (12) and (13). Selection uses no answerer forward pass and no fitted parameters.

C Protocol and Reproducibility

C.1 Benchmarks

Table 7 summarizes the two benchmarks. The LongVideoBench validation set holds 1,337 questions over 752 videos, with four or five options per question; Video-MME holds 2,700 questions over 900 videos, with four options each, split evenly into short, medium, and long videos. The median video lasts about 20 minutes on LongVideoBench and 17 on Video-MME. Both benchmarks are evaluated without subtitles, so the answerer receives only the selected frames, the question, and its options.

C.2 Candidate Frames and Answerer Input

Each video is decoded once into at most 128 candidate frames, evenly spaced in time and resized to 384×384, for both answerers. Every method, uniform sampling included, selects its k frames from this pool, so the methods differ only in which candidates they pick. The selected frames reach the answerer in temporal order.

C.3 Prompt and Answer Scoring

Table 8 gives the prompt for each answerer, and Table 9 the letter tokens whose probabilities are read. Both answerers use the Qwen2 vocabulary, so the token ids coincide. A letter can be spelled with or without a leading space; the probabilities of the two spellings are summed, and the option with the highest total is predicted. The prompt, the letters, and the scoring are the same for every method. App. I compares this protocol with generated answers.

C.4 Competitor Implementations

All ten competitors are reimplemented from their papers (Zhang et al., 2025b; Tang et al., 2025; Liu et al., 2025; Li et al., 2026a; Zhu et al., 2026; Li et al., 2026b; Sun et al., 2025b; Zhang et al., 2025a; Wu et al., 2026; Chen et al., 2026). Each keeps its own relevance model and hyperparameters, and each selects from the same candidate pool and feeds the same answerer as FULCRUM.

C.5 Compute and Software

Table 10 lists the model checkpoints and the library versions. FULCRUM runs its encoder in half precision and involves no training.

Table 7: **Benchmark statistics.**

Benchmark	Videos	Questions	Median dur. (s)	Options J
LongVideoBench val	752	1,337	1,224	4–5
Video-MME	900	2,700	1,020	4

Table 8: **Prompt template per answerer.** Each line is one line of the prompt. Grey: chat-template tokens; blue: fields filled per question. The Qwen2.5-VL vision block is repeated once per selected frame, and the LLaVA-Video `<image>` token stands for the k selected frames. The `E.` line is present only for five-option questions.

Qwen2.5-VL-7B-Instruct : Prompt for the Answerer

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|> × k {question}

A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
E. {option_E}

Answer the multiple-choice question by selecting the single best option.  Respond with
ONLY the letter of the correct option and nothing else.<|im_end|>
<|im_start|>assistant
```

LLaVA-Video-7B : Prompt for the Answerer

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<image>
{question}

A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
E. {option_E}

Answer the multiple-choice question by selecting the single best option.  Respond with
ONLY the letter of the correct option and nothing else.<|im_end|>
<|im_start|>assistant
```

Table 9: **Scored letter tokens, shared by both answerers.** Qwen2.5-VL and LLaVA-Video use the Qwen2 vocabulary, so the ids coincide; each letter is scored with and without a leading space, and the two ids are summed.

Letter	A	B	C	D	E
Token ids summed	32, 362	33, 425	34, 356	35, 422	36, 468

Table 10: **Model checkpoints and software.**

Item	Value
Answerer weights	`Qwen/Qwen2.5-VL-7B-Instruct`; `lmms-lab/LLaVA-Video-7B-Qwen2`
Encoder weights	`openai/clip-vit-large-patch14`
Library versions	torch 2.10.0+cu128, transformers 4.57.6, accelerate 1.13.0, Python 3.12.13

D Full Results

This section gives every accuracy behind Table 1 and breaks the Qwen2.5-VL-7B results down by question category, video duration, and question type.

D.1 Every Method, Answerer, and Budget

Tables 11 and 12 repeat Table 1 with one benchmark per table, so that the two answerers can be read side by side. The last doubling of the budget separates uniform sampling from the selectors. From k=16 to k=32, uniform sampling gains 2.6 to 4.7 points in each of the four answerer–benchmark pairs, the ten competitors change by -0.9 to $+2.5$, and FULCRUM by -0.4 to 0.0. Frames placed on the evidence at 16 frames leave little for 16 more to add, while uniform sampling, which places its frames without regard to the question, keeps gaining.

Table 11: **LongVideoBench val: every method, answerer and budget.** Accuracy (%); best in each column in bold, second best underlined; subscripts give FULCRUM's margin over the strongest competitor.

Method	**Qwen2.5-VL-7B**				**LLaVA-Video-7B**			
frames k	4	8	16	32	4	8	16	32
Uniform	42.7	44.2	48.8	52.9	44.4	47.3	51.2	53.8
AdaRD-Key	<u>55.2</u>	<u>59.4</u>	<u>60.8</u>	<u>60.1</u>	<u>57.8</u>	<u>59.5</u>	<u>61.1</u>	<u>62.2</u>
AKS	53.2	55.6	56.5	56.5	51.5	55.2	56.8	56.1
BOLT	49.8	52.1	56.2	58.5	48.3	51.8	55.2	55.4
DIG	50.8	54.9	56.1	55.2	52.2	55.4	58.6	59.6
FOCUS	50.5	52.9	56.5	57.8	51.2	55.8	57.0	59.5
MaxInfo	52.5	56.8	58.0	57.3	54.1	57.9	58.6	58.7
MDP^3	51.8	55.9	57.1	57.2	54.9	57.8	59.1	59.2
Q-Frame	52.2	54.6	56.8	55.9	53.8	56.8	57.7	57.8
VTR-VLM	49.6	50.4	51.9	52.3	50.1	52.6	54.1	54.2
WFS-SB	50.9	54.8	56.8	58.1	52.1	58.9	59.2	60.3
FULCRUM	**60.8**+5.6	**62.5**+3.1	**63.2**+2.4	**63.2**+3.1	**62.5**+4.7	**62.5**+3.0	**64.2**+3.1	**64.2**+2.0

Table 12: **Video-MME: every method, answerer and budget.** Accuracy (%), marked as in Table 11.

Method	**Qwen2.5-VL-7B**				**LLaVA-Video-7B**			
frames k	4	8	16	32	4	8	16	32
Uniform	39.6	43.3	46.5	51.2	41.6	45.1	48.3	52.6
AdaRD-Key	49.3	54.5	<u>57.9</u>	58.2	<u>55.2</u>	58.3	<u>59.6</u>	<u>60.1</u>
AKS	51.2	54.6	55.3	57.2	49.5	53.1	55.6	55.6
BOLT	48.9	51.7	55.8	58.1	47.6	51.2	54.9	55.1
DIG	50.1	53.4	55.3	55.9	51.2	54.8	57.1	58.2
FOCUS	49.8	53.1	55.7	56.8	50.9	55.3	56.9	58.4
MaxInfo	51.3	<u>55.0</u>	56.6	57.1	53.2	56.7	57.7	58.2
MDP^3	50.9	54.6	56.0	56.5	52.8	56.4	57.8	58.3
Q-Frame	50.5	53.9	55.8	56.2	52.6	55.8	57.0	57.5
VTR-VLM	48.4	50.0	52.1	52.8	49.5	52.6	54.3	55.0
WFS-SB	<u>51.8</u>	54.7	57.1	<u>58.3</u>	51.7	<u>58.6</u>	58.9	60.0
FULCRUM	**57.5**+5.7	**59.8**+4.8	**61.2**+3.3	**60.8**+2.5	**60.4**+5.2	**61.5**+2.9	**61.5**+1.9	**61.4**+1.3

D.2 By Question Category

Table 13 splits LongVideoBench at k=8 into its 17 question categories, and Fig. 6 plots them. FULCRUM is the most accurate method, alone or tied, in 14 of the 17 categories, and exceeds uniform sampling in all but one, where the two tie at 70.1. The largest gains over uniform sampling are on

Sequence of Scenes (13.4 to 46.4) and Text-referred Object Existence (28.9 to 61.8). FULCRUM falls below AdaRD-Key in three categories: Event-referred Object (-4.6), Object before/after Object (-9.1), and Event before/after Text (-6.9). Two of the three ask for the order of two moments; of the other two ordering categories, FULCRUM ties AdaRD-Key on one and leads it on the other (Event before/after Event, $+4.3$). An ordering question is answered by a pair of frames in the right order, while the option contrast scores each frame on its own, the first failure mode of Table 32. Table 14 gives the same breakdown for the twelve task types of Video-MME.

Table 13: **LongVideoBench by question category.** Qwen2.5-VL-7B, k=8. Accuracy (%); best in each row in bold, second best underlined; subscripts give FULCRUM's margin over the stronger of AdaRD-Key and AKS.

Question category	n	Uniform	AdaRD-Key	AKS	FULCRUM
Perception					
Scene-referred Event	93	57.0	71.0	67.7	**74.2**+3.2
Scene-referred Object Existence	72	52.8	58.3	52.8	**61.1**+2.8
Scene-referred Object Attribute	88	59.1	81.8	78.4	**86.4**+4.6
Event-referred Object	65	69.2	**89.2**	86.2	84.6−4.6
Object-referred Event	87	**70.1**	66.7	63.2	**70.1**+3.4
Text-referred Event	65	49.2	50.8	55.4	**63.1**+7.7
Text-referred Object Existence	76	28.9	56.6	56.6	**61.8**+5.2
Text-referred Object Attribute	79	41.8	58.2	55.7	**65.8**+7.6
Relation					
Event before/after Event	94	52.1	67.0	63.8	**71.3**+4.3
Object before/after Object	66	27.3	**53.0**	40.9	43.9−9.1
Sequence of Scenes	97	13.4	43.3	39.2	**46.4**+3.1
Scene-referred Object Tracking	81	24.7	50.6	42.0	**54.3**+3.7
Scene-referred Object Attribute Change	72	45.8	58.3	54.2	**65.3**+7.0
Event before/after Text	73	31.5	**52.1**	49.3	45.2−6.9
Object before/after Text	74	41.9	**45.9**	41.9	**45.9**+0.0
Text-referred Object Tracking	73	42.5	46.6	42.5	**50.7**+4.1
Text-referred Object Attribute Change	82	45.1	57.3	52.4	**67.1**+9.8
Overall	1337	44.2	59.4	55.6	**62.5**+3.1

Table 14: **Video-MME by task type.** Qwen2.5-VL-7B, k=8. Accuracy (%), marked as in Table 13; subscripts give FULCRUM's margin over the strongest of AdaRD-Key, AKS and VTR-VLM.

Task type	n	Uniform	AdaRD-Key	AKS	VTR-VLM	FULCRUM
Perception						
Temporal Perception	55	43.6	58.2	58.2	45.5	**61.8**+3.6
Spatial Perception	54	46.3	**57.4**	51.9	53.7	55.6−1.8
Attribute Perception	222	47.3	55.9	58.6	46.4	**60.4**+1.8
Action Recognition	313	40.9	**56.9**	53.4	49.8	55.3−1.6
Object Recognition	354	43.2	52.5	56.2	50.8	**56.5**+0.3
OCR Problems	139	42.4	51.8	52.5	49.6	**57.6**+5.1
Counting Problem	268	41.8	57.5	51.5	53.4	**63.8**+6.3
Reasoning						
Temporal Reasoning	177	43.5	54.2	55.4	52.0	**61.6**+6.2
Spatial Reasoning	56	41.1	53.6	55.4	46.4	**60.7**+5.3
Action Reasoning	285	43.2	52.3	50.2	51.9	**58.2**+5.9
Object Reasoning	454	41.4	55.5	53.7	47.4	**63.7**+8.2
Synopsis						
Information Synopsis	323	47.1	52.0	59.4	51.4	**60.4**+1.0
Overall	2700	43.3	54.5	54.6	50.1	**59.8**+5.2

D.3 BY VIDEO DURATION

On LongVideoBench at k=4 (Table 15, Fig. 7), accuracy falls from the shortest duration group to the longest by 27.6 points for uniform sampling (62.4 to 34.8), 13.1 for AdaRD-Key (64.0 to 50.9), and 6.3 for FULCRUM (64.6 to 58.3). The margin of FULCRUM over uniform sampling therefore grows from 2.2 points on videos of 8–15 s to 23.5 on videos of 15–60 min, and its margin over AdaRD-Key from 0.6 to 7.4. On the shortest videos, four evenly spaced frames are at most a few seconds apart, and selection has little to add. Table 16 gives Video-MME at k=8 by duration split.

D.4 BY QUESTION TYPE ACROSS BUDGETS

Table 17 follows the two question types of LongVideoBench, perception (625 questions) and relation (712), across budgets. FULCRUM at k=4 exceeds uniform sampling at k=32 on both, by 8.0 points

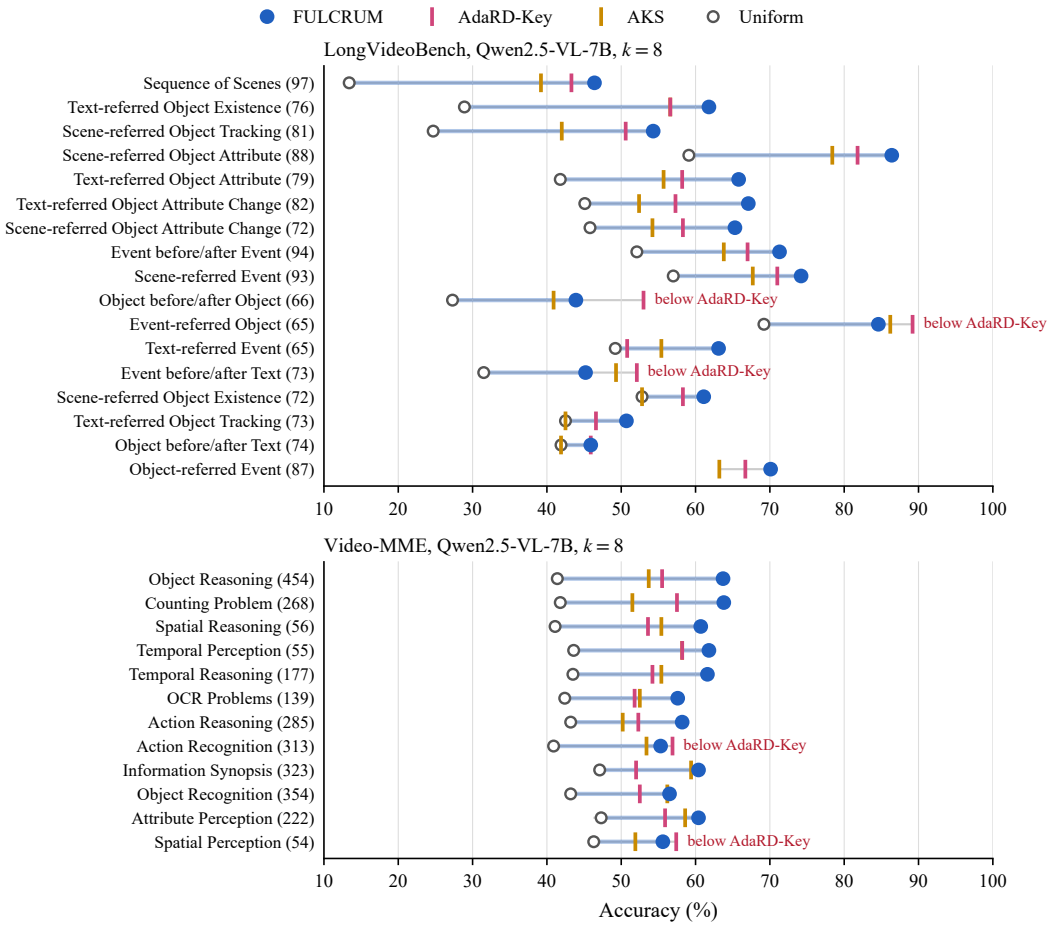

Figure 6: **Accuracy by question category** (LongVideoBench) and task type (Video-MME), Qwen2.5-VL-7B, k=8, sorted by the gain of FULCRUM over uniform sampling; question counts in parentheses. The shaded bar spans uniform sampling to FULCRUM. FULCRUM falls below AdaRD-Key on three LongVideoBench categories and two Video-MME task types (marked). Values are those of Tables 13 and 14; VTR-VLM, reported only for Video-MME, is omitted.

Table 15: **LongVideoBench accuracy by video duration group.** k=4. Best in each column in bold, second best underlined; subscripts give FULCRUM's margin over the stronger of AKS and AdaRD-Key.

Method	8–15 s	15–60 s	180–600 s	900–3600 s
Uniform	62.4	52.3	40.5	34.8
AKS	63.5	60.5	52.2	48.2
AdaRD-Key	<u>64.0</u>	<u>61.6</u>	<u>54.4</u>	<u>50.9</u>
FULCRUM	**64.6**+0.6	**65.7**+4.1	**60.4**+6.0	**58.3**+7.4

Table 16: **Video-MME by duration split.** k=8. Accuracy (%); subscripts give FULCRUM's margin over uniform sampling.

Method	Short	Medium	Long
Uniform	52.0	42.0	35.9
FULCRUM	**62.0**+10.0	**59.8**+17.8	**57.7**+21.8

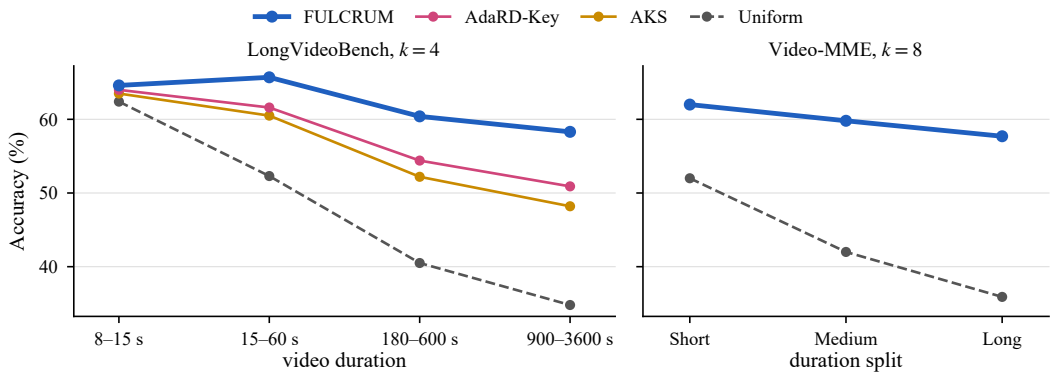

Figure 7: **Accuracy by video duration**: LongVideoBench at k=4 (Table 15) and Video-MME at k=8 (Table 16). On LongVideoBench the gap to uniform sampling grows from 2.2 points on the shortest videos to 23.5 on the longest; on Video-MME it grows from 10.0 to 21.8 points.

on perception (68.0 against 60.0) and 7.9 on relation (54.5 against 46.6). Beyond that the two types diverge. On relation questions, FULCRUM improves up to k=16 (54.5 to 60.8) and its margin over uniform sampling stays between 14.2 and 19.1 points at every budget. On perception questions, FULCRUM peaks at k=8 (71.2) and falls to 65.9 at 16 and 32 frames, while uniform sampling rises to 60.0, so the margin narrows from 18.1 to 5.9 points. A perception question typically rests on one moment, which a small budget can hold; a relation question rests on several, and more frames keep helping.

Table 17: **Budget sweep by question type.** LongVideoBench, Qwen2.5-VL-7B, accuracy (%). The benchmark has 625 perception and 712 relation questions, 1,337 in total. Subscripts give FULCRUM's margin over uniform sampling.

Question type	**k=4**	**k=8**	**k=16**	**k=32**
Uniform				
Perception	49.9	53.8	55.0	60.0
Relation	36.4	35.8	43.3	46.6
Overall	*42.7*	*44.2*	*48.8*	*52.9*
FULCRUM				
Perception	68.0+18.1	71.2+17.4	65.9+10.9	65.9+5.9
Relation	54.5+18.1	54.9+19.1	60.8+17.5	60.8+14.2
Overall	*60.8*+18.1	*62.5*+18.3	*63.2*+14.4	*63.2*+10.3

E FRAME BUDGET

Tables 19 and 20 compare every method at k frames with uniform sampling at every budget k', for both answerers, and Fig. 8 draws the same numbers as heatmaps. Each cell is a method's accuracy at k frames (row) minus the accuracy of uniform sampling at k' frames (column). The diagonal, k'=k, is the matched-budget comparison of Table 1; the cells to its right ask how many uniform frames a selector's k frames are worth.

Of the 704 cells, 38 are red, and every one of them compares a competitor with 16 or 32 uniform frames. Every competitor has at least one. At k=4, 6 to 8 of the ten competitors fall below uniform sampling at 32 frames in each answerer–benchmark pair, and each competitor does so in at least one pair. Depending on the method and the pair, a competitor needs 4 to 16 frames to match 32 uniform frames, where it matches them at all; BOLT needs 16 in three of the four pairs. VTR-VLM with Qwen2.5-VL-7B on LongVideoBench matches it at no budget: at 32 frames it is 0.6 points below uniform sampling at the same budget, the only matched-budget loss in either table. FULCRUM has no red cell. At k=4 it is 6.3 to 8.7 points above uniform sampling at 32 frames, the comparison of §5.3. Table 18 compares FULCRUM at k=4 with the best of the ten competitors at k=32: FULCRUM is ahead in three of the four pairs and 0.8 points behind with Qwen2.5-VL-7B on Video-MME, where WFS-SB is the best competitor.

Table 18: **FULCRUM at the smallest budget against the best competitor at the largest.** Accuracy of FULCRUM at k=4 and of the best of the ten competitors at k=32, with the gap in points. The full comparison for every method and budget is in Tables 19 and 20.

		FULCRUM	Best competitor		
Answerer	**Benchmark**	k=4	Method	k=32	**Gap**
Qwen2.5-VL-7B	LongVideoBench	60.8	AdaRD-Key	60.1	+0.7
	Video-MME	57.5	WFS-SB	58.3	−0.8
LLaVA-Video-7B	LongVideoBench	62.5	AdaRD-Key	62.2	+0.3
	Video-MME	60.4	AdaRD-Key	60.1	+0.3

Table 19: **Iso-frame gaps, LongVideoBench.** Accuracy of each method at k frames minus the accuracy of uniform sampling at k' frames (U@k'), in points, for both answerers. Cells are shaded by the gap; red cells fall below uniform sampling.

Method		**Qwen2.5-VL-7B**				**LLaVA-Video-7B**			
	k	U@4	U@8	U@16	U@32	U@4	U@8	U@16	U@32
AdaRD-Key	4	12.5	11.0	6.4	2.3	13.4	10.5	6.6	4.0
	8	16.7	15.2	10.6	6.5	15.1	12.2	8.3	5.7
	16	18.1	16.6	12.0	7.9	16.7	13.8	9.9	7.3
	32	17.4	15.9	11.3	7.2	17.8	14.9	11.0	8.4
AKS	4	10.5	9.0	4.4	0.3	7.1	4.2	0.3	−2.3
	8	12.9	11.4	6.8	2.7	10.8	7.9	4.0	1.4
	16	13.8	12.3	7.7	3.6	12.4	9.5	5.6	3.0
	32	13.8	12.3	7.7	3.6	11.7	8.8	4.9	2.3
BOLT	4	7.1	5.6	1.0	−3.1	3.9	1.0	−2.9	−5.5
	8	9.4	7.9	3.3	−0.8	7.4	4.5	0.6	−2.0
	16	13.5	12.0	7.4	3.3	10.8	7.9	4.0	1.4
	32	15.8	14.3	9.7	5.6	11.0	8.1	4.2	1.6
DIG	4	8.1	6.6	2.0	−2.1	7.8	4.9	1.0	−1.6
	8	12.2	10.7	6.1	2.0	11.0	8.1	4.2	1.6
	16	13.4	11.9	7.3	3.2	14.2	11.3	7.4	4.8
	32	12.5	11.0	6.4	2.3	15.2	12.3	8.4	5.8
FOCUS	4	7.8	6.3	1.7	−2.4	6.8	3.9	0.0	−2.6
	8	10.2	8.7	4.1	0.0	11.4	8.5	4.6	2.0
	16	13.8	12.3	7.7	3.6	12.6	9.7	5.8	3.2
	32	15.1	13.6	9.0	4.9	15.1	12.2	8.3	5.7
MaxInfo	4	9.8	8.3	3.7	−0.4	9.7	6.8	2.9	0.3
	8	14.1	12.6	8.0	3.9	13.5	10.6	6.7	4.1
	16	15.3	13.8	9.2	5.1	14.2	11.3	7.4	4.8
	32	14.6	13.1	8.5	4.4	14.3	11.4	7.5	4.9
MDP^3	4	9.1	7.6	3.0	−1.1	10.5	7.6	3.7	1.1
	8	13.2	11.7	7.1	3.0	13.4	10.5	6.6	4.0
	16	14.4	12.9	8.3	4.2	14.7	11.8	7.9	5.3
	32	14.5	13.0	8.4	4.3	14.8	11.9	8.0	5.4
Q-Frame	4	9.5	8.0	3.4	−0.7	9.4	6.5	2.6	0.0
	8	11.9	10.4	5.8	1.7	12.4	9.5	5.6	3.0
	16	14.1	12.6	8.0	3.9	13.3	10.4	6.5	3.9
	32	13.2	11.7	7.1	3.0	13.4	10.5	6.6	4.0
VTR-VLM	4	6.9	5.4	0.8	−3.3	5.7	2.8	−1.1	−3.7
	8	7.7	6.2	1.6	−2.5	8.2	5.3	1.4	−1.2
	16	9.2	7.7	3.1	−1.0	9.7	6.8	2.9	0.3
	32	9.6	8.1	3.5	−0.6	9.8	6.9	3.0	0.4
WFS-SB	4	8.2	6.7	2.1	−2.0	7.7	4.8	0.9	−1.7
	8	12.1	10.6	6.0	1.9	14.5	11.6	7.7	5.1
	16	14.1	12.6	8.0	3.9	14.8	11.9	8.0	5.4
	32	15.4	13.9	9.3	5.2	15.9	13.0	9.1	6.5
FULCRUM	4	18.1	16.6	12.0	7.9	18.1	15.2	11.3	8.7
	8	19.8	18.3	13.7	9.6	18.1	15.2	11.3	8.7
	16	20.5	19.0	14.4	10.3	19.8	16.9	13.0	10.4
	32	20.5	19.0	14.4	10.3	19.8	16.9	13.0	10.4

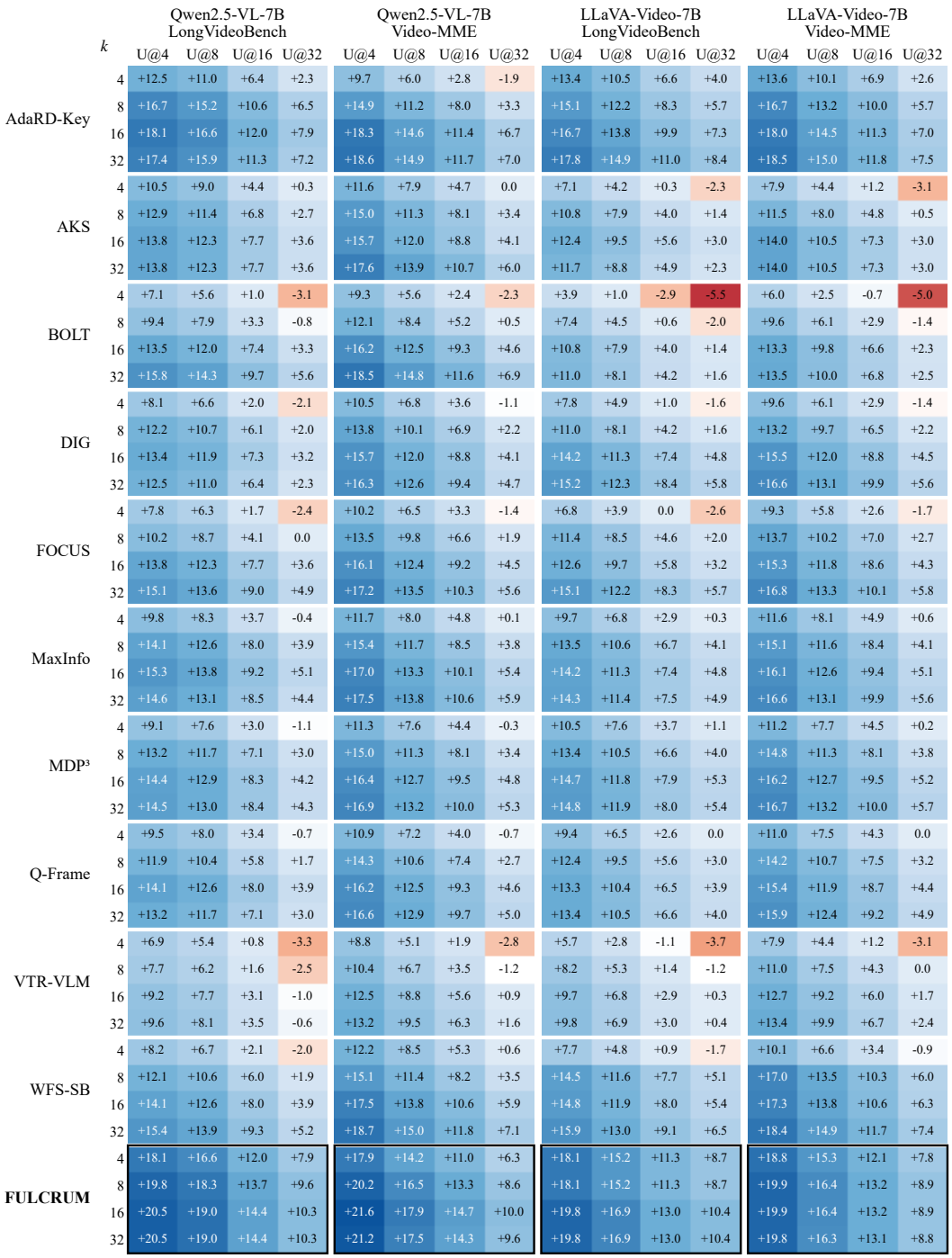

Figure 8: **Iso-frame gaps**, both answerers and benchmarks. Each cell is a method's accuracy at k frames (rows) minus uniform accuracy at k' frames (columns, U@k'), in points; red cells fall below uniform sampling. At k=4, every competitor falls below uniform sampling at 32 frames in at least one of the four settings; FULCRUM (framed) stays at least 6.3 points above it. The numbers are those of Tables 19 and 20.

Table 20: **Iso-frame gaps, Video-MME.** Laid out as in Table 19.

Method	k	Qwen2.5-VL-7B				LLaVA-Video-7B			
		U@4	U@8	U@16	U@32	U@4	U@8	U@16	U@32
AdaRD-Key	4	9.7	6.0	2.8	−1.9	13.6	10.1	6.9	2.6
	8	14.9	11.2	8.0	3.3	16.7	13.2	10.0	5.7
	16	18.3	14.6	11.4	6.7	18.0	14.5	11.3	7.0
	32	18.6	14.9	11.7	7.0	18.5	15.0	11.8	7.5
AKS	4	11.6	7.9	4.7	0.0	7.9	4.4	1.2	−3.1
	8	15.0	11.3	8.1	3.4	11.5	8.0	4.8	0.5
	16	15.7	12.0	8.8	4.1	14.0	10.5	7.3	3.0
	32	17.6	13.9	10.7	6.0	14.0	10.5	7.3	3.0
BOLT	4	9.3	5.6	2.4	−2.3	6.0	2.5	−0.7	−5.0
	8	12.1	8.4	5.2	0.5	9.6	6.1	2.9	−1.4
	16	16.2	12.5	9.3	4.6	13.3	9.8	6.6	2.3
	32	18.5	14.8	11.6	6.9	13.5	10.0	6.8	2.5
DIG	4	10.5	6.8	3.6	−1.1	9.6	6.1	2.9	−1.4
	8	13.8	10.1	6.9	2.2	13.2	9.7	6.5	2.2
	16	15.7	12.0	8.8	4.1	15.5	12.0	8.8	4.5
	32	16.3	12.6	9.4	4.7	16.6	13.1	9.9	5.6
FOCUS	4	10.2	6.5	3.3	−1.4	9.3	5.8	2.6	−1.7
	8	13.5	9.8	6.6	1.9	13.7	10.2	7.0	2.7
	16	16.1	12.4	9.2	4.5	15.3	11.8	8.6	4.3
	32	17.2	13.5	10.3	5.6	16.8	13.3	10.1	5.8
MaxInfo	4	11.7	8.0	4.8	0.1	11.6	8.1	4.9	0.6
	8	15.4	11.7	8.5	3.8	15.1	11.6	8.4	4.1
	16	17.0	13.3	10.1	5.4	16.1	12.6	9.4	5.1
	32	17.5	13.8	10.6	5.9	16.6	13.1	9.9	5.6
MDP^3	4	11.3	7.6	4.4	−0.3	11.2	7.7	4.5	0.2
	8	15.0	11.3	8.1	3.4	14.8	11.3	8.1	3.8
	16	16.4	12.7	9.5	4.8	16.2	12.7	9.5	5.2
	32	16.9	13.2	10.0	5.3	16.7	13.2	10.0	5.7
Q-Frame	4	10.9	7.2	4.0	−0.7	11.0	7.5	4.3	0.0
	8	14.3	10.6	7.4	2.7	14.2	10.7	7.5	3.2
	16	16.2	12.5	9.3	4.6	15.4	11.9	8.7	4.4
	32	16.6	12.9	9.7	5.0	15.9	12.4	9.2	4.9
VTR-VLM	4	8.8	5.1	1.9	−2.8	7.9	4.4	1.2	−3.1
	8	10.4	6.7	3.5	−1.2	11.0	7.5	4.3	0.0
	16	12.5	8.8	5.6	0.9	12.7	9.2	6.0	1.7
	32	13.2	9.5	6.3	1.6	13.4	9.9	6.7	2.4
WFS-SB	4	12.2	8.5	5.3	0.6	10.1	6.6	3.4	−0.9
	8	15.1	11.4	8.2	3.5	17.0	13.5	10.3	6.0
	16	17.5	13.8	10.6	5.9	17.3	13.8	10.6	6.3
	32	18.7	15.0	11.8	7.1	18.4	14.9	11.7	7.4
FULCRUM	4	17.9	14.2	11.0	6.3	18.8	15.3	12.1	7.8
	8	20.2	16.5	13.3	8.6	19.9	16.4	13.2	8.9
	16	21.6	17.9	14.7	10.0	19.9	16.4	13.2	8.9
	32	21.2	17.5	14.3	9.6	19.8	16.3	13.1	8.8

F What the Options Contribute

This section documents the controls of Table 2 (§5.2) and adds four measurements: how far each relevance text moves the selection, whether the option order matters, how the gain depends on peak distance, and how much of the gain survives when the answerer writes the options itself.

F.1 Lexical Controls

The two controls of Table 2 ask whether the gain needs the question's own answer hypotheses or only more text. *Foreign options* replace the options with the option set of a question about another video that has the same number of options. The donor is drawn at random (seed 0) from the ten candidates closest in mean option length, so the two sets match in count and length: the median gap in mean option length is 0.6 characters (90th percentile 7.1). One question with no same-count donor takes a donor with more options, truncated to its count. The foreign options are scored exactly as the own options are. *Paraphrases* replace the question and its options with a single rewrite of the question, produced by Qwen2.5-VL-7B with greedy decoding and instructed to keep the meaning, change the wording and sentence structure, and neither name nor imply any option. Figs. 9 and 10 show the four texts for two questions and the frames each one selects.

F.2 Which Frames Each Text Selects

The texts that do not contrast the question's own options stay close to the question: fused options share 0.81 of their selected frames with question-only selection, and foreign options 0.74 (Table 22). FULCRUM shares 0.52, so the question's own options move about half of the selection. Foreign options and FULCRUM share 0.49, the lowest overlap in the table: the two texts differ only in which options they carry, and they select mostly different frames. Shuffling the option order changes none of the selected frames (Table 21), since rel and disc are invariant to it, and accuracy is 62.5 in the original order and in both permutations.

F.3 Peak Distance

Fig. 3 bins questions by peak distance, the distance between the candidate frame most similar to the question and the frame with the largest option contrast, as a fraction of video length (§5.2). It is computed from the encoder alone, with no answer and no answerer. The options help most where the question alone points away from the frames that separate the answers.

F.4 Options Generated by the Answerer

When the selector reads options that the answerer generates from the question alone, instead of the benchmark's options, it recovers a third to a half of the gain over question-only selection (Table 23): on LongVideoBench with Qwen2.5-VL-7B, 61.0 against 59.3 for the question alone and 62.8 for FULCRUM. Across the four answerer–benchmark settings, generated options gain 1.1 to 1.7 points over the question alone, and the benchmark's options 2.6 to 3.5. The answerer still answers with the benchmark's options, so only the text the selector reads changes.

Table 21: **Option-order permutations.** LongVideoBench, k=8, locked configuration. The option order enters neither rel nor disc, so no selected frame changes.

Permutation	Acc.	Frames changed (%)	Δ vs. original
Original order	62.5	0.00	0.00
Permutation 1	62.5	0.00	0.00
Permutation 2	62.5	0.00	0.00

Table 22: **Selected-frame overlap between relevance-text variants.** Mean fraction of selected frames two variants share, at k=8. Darker cells share more frames.

	Question only	Fused	FULCRUM	Foreign
Question only	–			
Fused	0.81	–		
FULCRUM	0.52	0.55	–	
Foreign	0.74	0.70	0.49	–

Table 23: **Generated options against given options.** k=8. Only the text FULCRUM selects with changes; the answerer always answers with the benchmark's options. Qwen2.5-VL-7B columns use m=32, as in Table 2. Subscripts give the change in points over question-only relevance. LVB: LongVideoBench.

Selector reads	Qwen2.5-VL-7B		LLaVA-Video-7B	
	LVB	Video-MME	LVB	Video-MME
Question only	59.3	57.8	59.6	58.9
Options generated by the answerer	61.0 $_{+1.7}$	58.9 $_{+1.1}$	61.1 $_{+1.5}$	60.2 $_{+1.3}$
Benchmark options (FULCRUM)	**62.8** $_{+3.5}$	**61.1** $_{+3.3}$	**62.5** $_{+2.9}$	**61.5** $_{+2.6}$

Example (a) Where else does the man on the sofa appear?

Question: *The background features sunlit green trees and a doorway. A man wearing a white short-sleeved shirt, black pants, and sporting a goatee is sitting on a white sofa outside. In what other scenes does this man appear?*

Own options (read by FULCRUM)
(A) On a rooftop terrace
(B) In the driver's seat of a car illuminated by sunlight
(C) Inside a cafe
(D) Inside a library
(E) Inside a museum

Foreign options (another video's question)
(A) Silver bracelet
(B) Watch
(C) Vegetable knife without letters on the blade
(D) Golden ring
(E) Vegetable knife with letters on the blade

Paraphrase: *In which additional settings can we find the man who is seated on a white sofa outside, dressed in a white short-sleeved shirt, black pants, and with a goatee?*

Gold answer: (B) In the driver's seat of a car illuminated by sunlight

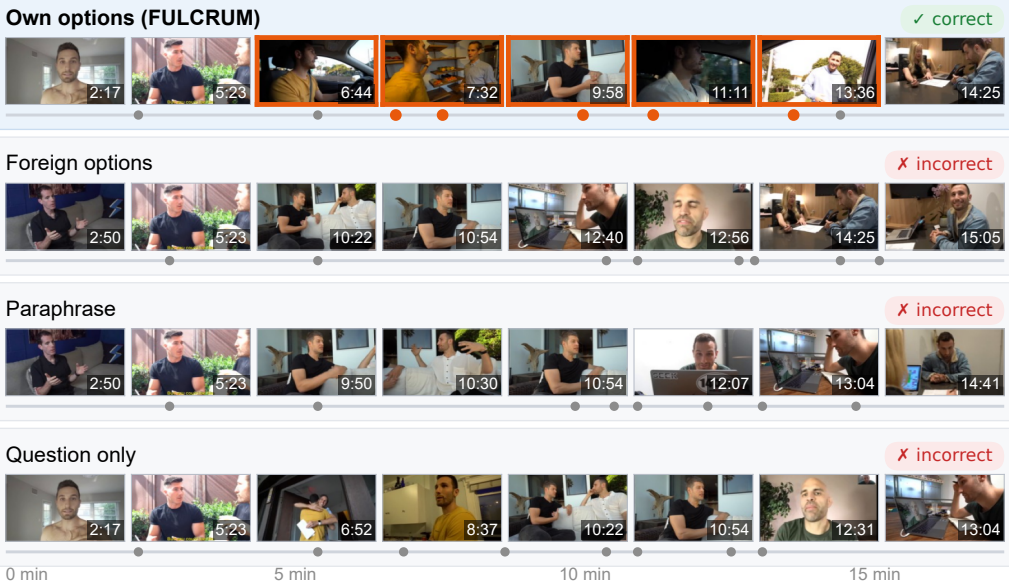

Figure 9: **Frames selected under each relevance text**, k=8, on a question where FULCRUM answers correctly and both the foreign options and the question alone do not. Each row shows the frames one text selects, their timestamps, and where they fall in the video. Outlined frames are selected by FULCRUM and by neither of those two controls. Only FULCRUM selects the man in the driver's seat of a car (6:44, 11:11), the scene the correct option names.

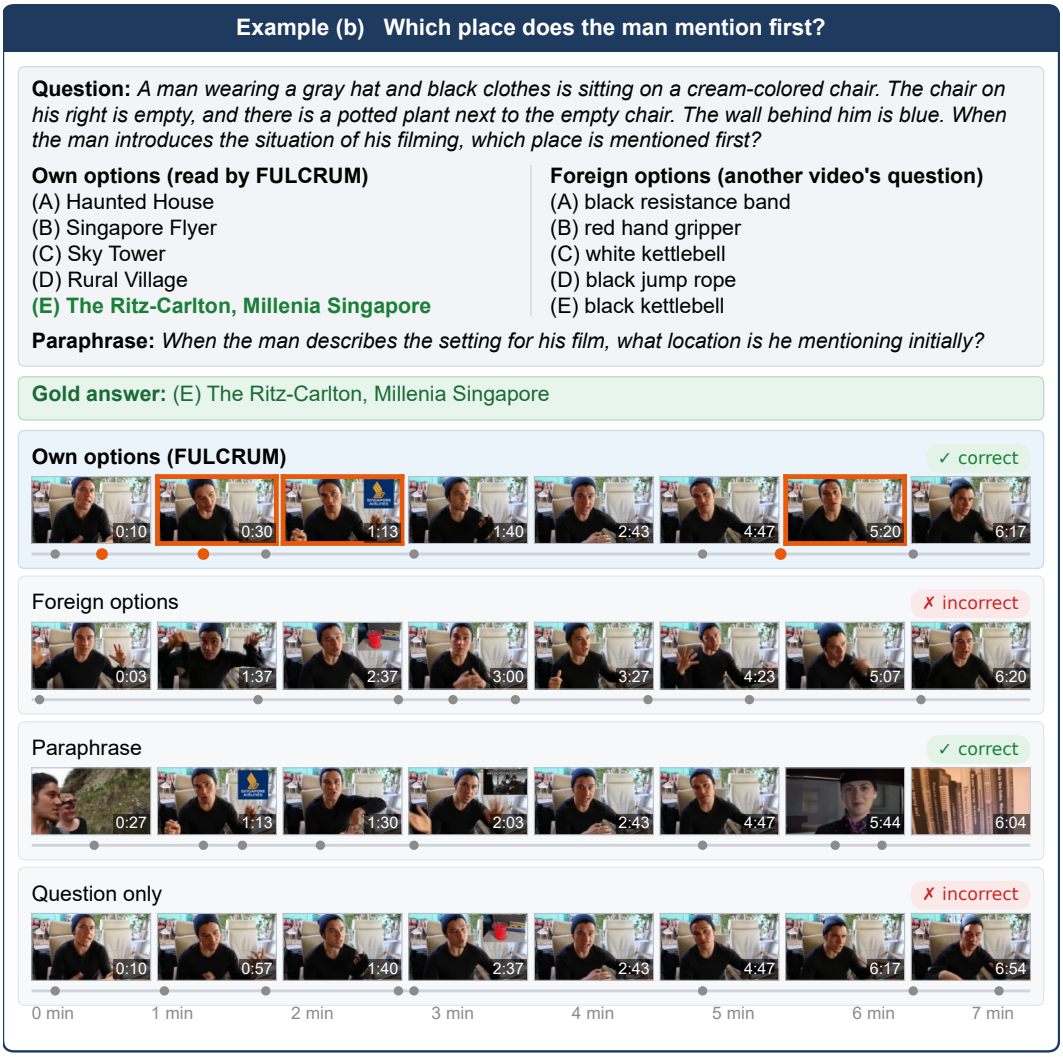

Figure 10: **Frames selected under each relevance text**, a second question, laid out as in Fig. 9. FULCRUM and the paraphrase select the frame showing a Singapore Airlines logo (1:13), and both answer correctly; the foreign options and the question alone miss it and do not.

G COMPONENTS AND SENSITIVITY

This section varies the parameters of §4, the entropy gate, and the encoder, on LongVideoBench with Qwen2.5-VL-7B at k=8, where the locked configuration reaches 62.5.

G.1 ABLATIONS

Table 24 lists the accuracies behind Fig. 4: the locked configuration, and each variant that removes or perturbs one part of FULCRUM, on both benchmarks.

Table 24: **Ablations.** k=8, Qwen2.5-VL-7B. Accuracy (%); subscripts give the change in points relative to FULCRUM.

Variant	LongVideoBench	Video-MME
FULCRUM	**62.5**	**59.8**
Input order		
Frames in shuffled order	47.9 $_{-14.6}$	46.8 $_{-13.0}$
Allocation		
Shuffled segment scores	50.3 $_{-12.2}$	48.6 $_{-11.2}$
One frame per uniform cell	57.1 $_{-5.4}$	55.2 $_{-4.6}$
No coverage floor (f=0)	60.2 $_{-2.3}$	57.5 $_{-2.3}$
Segment score		
rel only	60.1 $_{-2.4}$	57.9 $_{-1.9}$
disc only	57.8 $_{-4.7}$	55.4 $_{-4.4}$
Frame assignment		
Without relevance ($\rho \equiv 0$)	58.9 $_{-3.6}$	56.3 $_{-3.5}$
Without diversity (λ=0)	61.0 $_{-1.5}$	58.4 $_{-1.4}$

G.2 ONE PARAMETER AT A TIME

Table 25 moves each parameter over five values with all others locked, and the top of Fig. 11 plots the sweeps. Accuracy stays within 1.5 points of 62.5 except at three extremes. Removing the coverage floor (f=0) costs 2.3 points, the ablation of Table 24. At f=1 the floor takes the whole budget, the allocation has no free frames to distribute, and accuracy falls to 59.1. A gate threshold of θ=0.95 lets the gate fire on profiles that still carry a ranking, and costs 1.7 points. The locked configuration is not the best value of every sweep: 32 segments reach 62.8, the setting of Table 2. It is within 0.3 points of the best value in every sweep and grid.

G.3 TWO PARAMETERS AT A TIME

Table 26 varies two pairs of parameters jointly (bottom of Fig. 11). The floor f and the concavity c both control how far the budget may concentrate. Without a floor, accuracy rises with c from 58.9 to 61.4: a flatter allocation partly supplies the coverage the floor would guarantee. With $f \geq 0.5$ the grid is flat, 61.7 to 62.5. The best cell, 62.6 at f=0.25 and c=1.0, is 0.1 points above the locked one. The diversity weight λ and the regularizer ε enter the fill through the first-frame diversity gain $\lambda \log(1 + \varepsilon^{-1})$ (§3.4). The two lowest cells of the grid are its smallest value, 0.07 at λ=0.1 and ε=1.0 (61.4), and its largest, 3.2 at λ=0.7 and ε=0.01 (61.2). In the first, diversity barely counts; in the second, it outweighs the relevance $\rho_t \in [-1, 1]$. Every other cell lies between 62.0 and 62.5.

G.4 ENTROPY GATE

When the segment scores are nearly flat, the gate (§3.3) discards them and leaves selection to coverage and diversity. At the locked m=16 it fires on 9.4% of the questions (Table 27), and on those it raises accuracy from 46.8 to 51.6, six more correct answers. Disabling it (θ=1.00) costs 0.4 points overall, 62.1 against 62.5 (Table 25). With more segments the profile is flat more often: the gate fires on 12.1% of the questions at m=32 and 16.3% at m=64, and gains 3.7 and 3.2 points on them.

G.5 ENCODERS

Table 28 runs FULCRUM with six encoders, each compared with question-only relevance scored by the same encoder. The option contrast adds 2.7 to 3.2 points on LongVideoBench and 1.6 to 2.1

Table 25: **One-at-a-time parameter sweeps.** LongVideoBench, k=8. For each parameter, the upper (gray) row lists the swept values and the lower row the accuracy (%); all other parameters are held at the locked configuration of §4, whose value is shaded.

Parameter	Swept value / Accuracy (%)				
Segments m	8	16	24	32	64
	61.4	62.5	62.3	62.8	61.9
Coverage floor f	0	0.25	0.5	0.75	1.0
	60.2	61.8	62.5	62.0	59.1
Concavity c	0.1	0.25	0.5	1.0	2.0
	62.1	62.4	62.5	62.2	61.7
Diversity weight λ	0	0.1	0.35	0.7	1.0
	61.0	62.0	62.5	62.1	61.3
Regularizer ε	0.01	0.05	0.1	0.5	1.0
	62.2	62.4	62.5	62.3	62.0
Projection rank d	8	16	32	64	128
	61.6	62.3	62.5	62.4	62.4
Gate θ	0.95	0.97	0.985	0.99	1.00
	60.8	62.1	62.5	62.5	62.1

Table 26: **Two-parameter grids.** LongVideoBench, k=8, accuracy (%). Each grid varies two parameters jointly while all others stay at the locked configuration of §4; the locked cell is shaded. (a) Coverage floor and concavity both control how far the budget may concentrate, so a one-at-a-time sweep could miss an interaction between them. (b) Diversity weight and regularizer.

(a) Floor f × concavity c

Floor f \ Concavity c	0.1	0.5	1.0	2.0
0	58.9	60.2	60.9	61.4
0.25	61.3	61.8	62.6	62.0
0.5	62.1	**62.5**	62.2	61.7
0.75	62.0	62.0	61.9	61.9

(b) Diversity λ × regularizer ε

Diversity λ \ Regularizer ε	0.01	0.1	1.0
0.1	62.3	62.0	61.4
0.35	62.2	**62.5**	62.0
0.7	61.2	62.1	62.4

Table 27: **Activation rate and effect of the entropy gate.** LongVideoBench, k=8. *Rate* is the percentage of questions on which the gate is activated (n = number of questions). Accuracy (%) is reported on the activated questions only, with and without the gate; Δ denotes the difference in points. The locked value of m is shaded.

	Gate activation		Accuracy on activated questions		
Segments m	Rate (%)	n	w/ gate	w/o gate	Δ
16	9.4	126	51.6	46.8	+4.8
32	12.1	162	50.6	46.9	+3.7
64	16.3	218	50.0	46.8	+3.2

on Video-MME with every encoder, so the gain does not depend on a particular encoder. Stronger encoders raise both variants: with SigLIP2, FULCRUM reaches 63.9 and 61.0, 1.4 and 1.2 points above the locked CLIP ViT-L/14. All results in the main text use CLIP ViT-L/14.

Table 28: **FULCRUM with different visual encoders.** k=8, Qwen2.5-VL-7B, locked configuration (m=16). Accuracy (%); subscripts give the improvement in points over the question-only variant scored by the same encoder. The locked encoder is shaded. The improvement stays within +2.7 to +3.2 on LongVideoBench and +1.6 to +2.1 on Video-MME across all six encoders. BLIP ITM: the image–text matching head of BLIP.

Encoder	LongVideoBench	Video-MME
CLIP ViT-B/16	60.1 $_{+2.7}$	57.6 $_{+1.7}$
CLIP ViT-L/14	62.5 $_{+3.0}$	59.8 $_{+1.9}$
CLIP ViT-L/14-336	62.8 $_{+3.2}$	60.1 $_{+2.1}$
SigLIP-so400m	63.4 $_{+2.9}$	60.6 $_{+1.9}$
SigLIP2	**63.9** $_{+3.0}$	**61.0** $_{+2.0}$
BLIP ITM	61.2 $_{+2.8}$	58.5 $_{+1.6}$

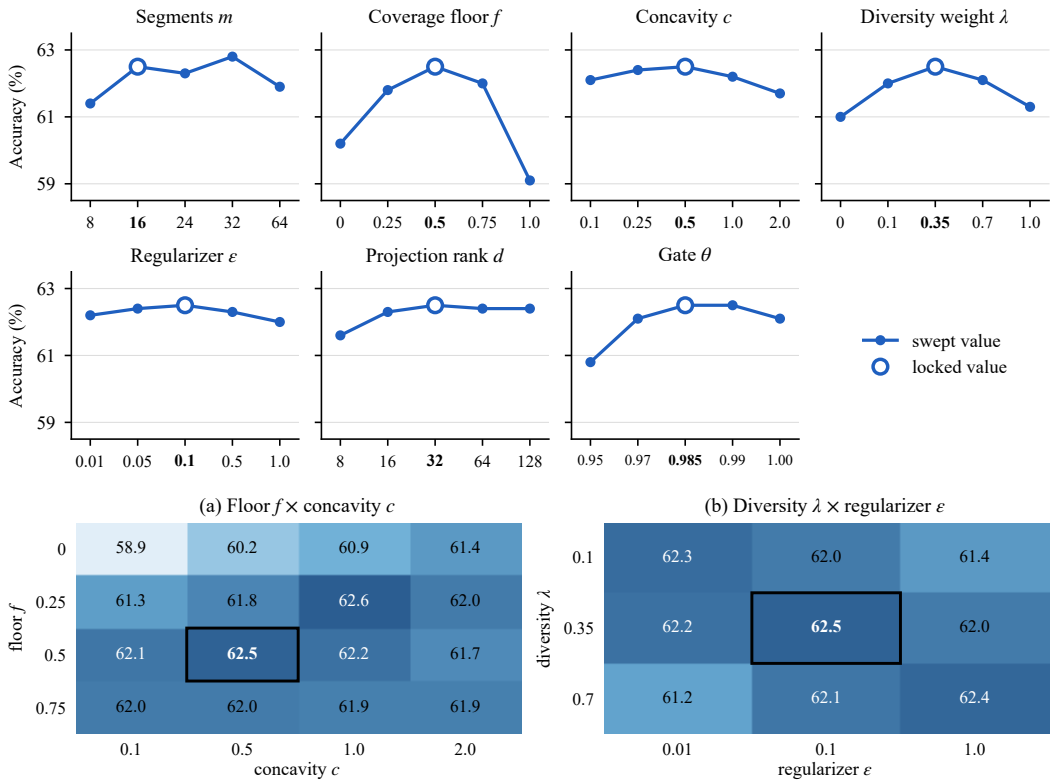

Figure 11: **Parameter sensitivity** on LongVideoBench (k=8). Top: the one-at-a-time sweeps of Table 25; the open marker is the locked value. Bottom: the two-parameter grids of Table 26; the framed cell is the locked configuration. Accuracy stays within 1.5 points of the locked configuration everywhere except at the extremes of the coverage floor (f=0, f=1) and the gate (θ=0.95).

H Answerer Scale

Table 29 and Fig. 12 vary the size of the Qwen2.5-VL answerer from 3B to 72B on LongVideoBench at k=8, with AKS as the reference selector. Of the two answerer families, Qwen2.5-VL is the one released at all four sizes. The 7B column repeats Table 1. FULCRUM is the most accurate method at every size. Its margin over AKS stays between 5.8 and 6.9 points, while its margin over uniform sampling peaks at 7B (18.3) and narrows to 13.4 at 72B. Larger answerers get more from the question alone, with accuracy without frames rising from 34.8 to 44.0, and more from any frames: uniform sampling adds 5.3 points over no frames at 3B and 9.6 at 72B. Within this range, the choice of frames matters more than the size of the answerer. With FULCRUM, the 7B answerer (62.5) is more accurate than the 72B answerer with AKS (61.2) or with uniform sampling (53.6), and the 3B answerer (56.0) is more accurate than the 72B answerer with uniform sampling.

Table 29: **Effect of answerer scale.** LongVideoBench, Qwen2.5-VL, k=8. Accuracy (%); AKS is the reference selector and is held fixed across sizes. Δ rows give the improvement of FULCRUM in points.

Method	**Answerer size**			
	3B	7B	32B	72B
No frames	34.8	37.1	41.5	44.0
Uniform	40.1	44.2	50.3	53.6
AKS	50.2	55.6	59.0	61.2
FULCRUM	**56.0**	**62.5**	**65.2**	**67.0**
Δ vs. Uniform	+15.9	+18.3	+14.9	+13.4
Δ vs. AKS	+5.8	+6.9	+6.2	+5.8

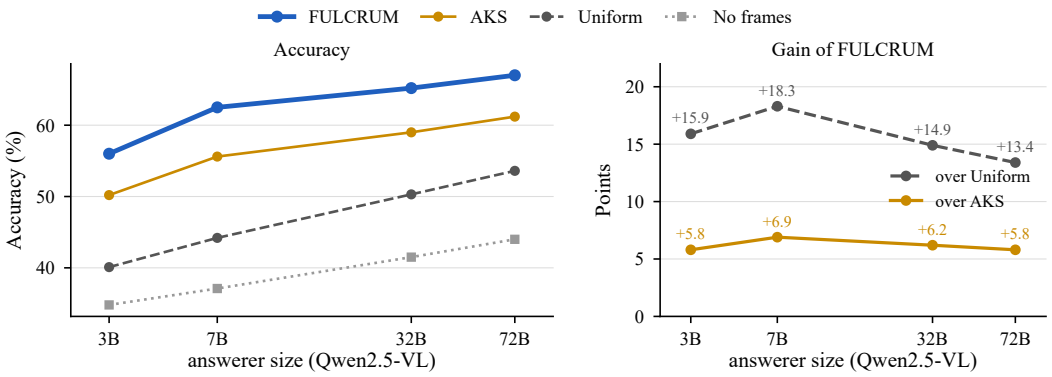

Figure 12: **Answerer scale**, LongVideoBench, k=8 (Table 29). Left: accuracy. Right: gain of FULCRUM over uniform sampling and over AKS. The margin over AKS stays between 5.8 and 6.9 points from 3B to 72B; the margin over uniform sampling peaks at 7B and narrows to 13.4 points at 72B, as accuracy without frames rises from 34.8 to 44.0.

I EVALUATION CHECKS

Two checks concern the evaluation rather than the selector: whether the answer-scoring protocol favors any method, and whether the gain comes from questions the answerer could answer without seeing the video.

I.1 ANSWER SCORING

Our harness reads the answerer's next-token probabilities over the option letters (§4), the multiple-choice protocol of Biderman et al. (2024). Published selectors are commonly evaluated with LMMs-Eval (Zhang et al., 2024a), which generates a reply and parses the letter from it. Table 30 scores the same selected frames both ways, with the prompt and parser of LMMs-Eval unchanged. Generated answers are 1.0 to 1.4 points less accurate, and 2.2 to 2.4% of the replies do not parse. The three methods keep their order, and the margin of FULCRUM over AKS changes from 6.9 to 6.8 points.

Table 30: **Answer scoring: letter probabilities against generated answers.** The same selected frames are scored both ways. LongVideoBench, Qwen2.5-VL-7B, k=8.

Method	Letter probabilities	Generated answer, LMMs-Eval	
	Accuracy (ours)	Accuracy	Parse failures (%)
Uniform	44.2	43.2	2.3
AKS	55.6	54.3	2.4
FULCRUM	**62.5**	**61.1**	2.2

I.2 LANGUAGE PRIOR

Table 31 splits the LongVideoBench questions by whether the answerer picks the correct option with no frames at all, which it does on 37% of them (the no-frames accuracy of Table 29). On these questions, frames can only keep or overturn the answer, and uniform frames keep it on 65.9%; part of this stratum is answered correctly by chance, and the frames change such guesses. FULCRUM keeps 81.0%, 3.0 points more than AKS. On the other 63%, the frames must supply the answer. There FULCRUM reaches 51.5, 9.2 points above AKS and 20.1 above uniform sampling. This stratum carries about 84% of the overall margin over AKS (0.63 of 9.2 points, against 6.9), so the gain is not explained by the language prior.

Table 31: **Accuracy stratified by the language prior.** LongVideoBench, Qwen2.5-VL-7B, k=8. Questions are split by whether the answerer is correct *without* any frames; the share of questions in each stratum is given in parentheses. Accuracy (%) with frames selected by each method; subscripts give FULCRUM's margin over AKS.

Method	Answerer without frames	
	Correct (37%)	Incorrect (63%)
Uniform	65.9	31.4
AKS	78.0	42.3
FULCRUM	**81.0**+3.0	**51.5**+9.2

J FAILURE ANALYSIS

Table 32 describes four ways in which FULCRUM fails, read from questions that uniform sampling answers correctly and FULCRUM does not. The first two modes are limits of scoring frames one at a time with an image–text encoder. Options that differ in the direction of a motion, or in left against right, are separated by the order of several frames or by fine spatial detail, which no single-frame similarity captures. The ordering categories of Table 13 on which FULCRUM trails AdaRD-Key show the same limit at the level of a category (App. D). A question located by a spoken line needs the audio or the subtitles, which none of the evaluated methods receives. In the third mode the contrast is flat, so the segment scores rest mostly on rel, or, when they are flat as well, the gate leaves the question to coverage and diversity (§3.3). The fourth is the contrast working as designed on a misleading option: a wrong option that names an object visible throughout the video draws frames toward that object. FULCRUM does not fall below uniform sampling on any category of Table 13, so these failures are spread across question types rather than concentrated in one.

Table 32: **Failure modes.** Read from questions uniform answers correctly and FULCRUM does not.

Mode	Description
Options differ in details the encoder cannot separate	The options differ only in direction of motion, left against right, or near-identical wording. A single frame carries none of these, so the contrast points at the wrong frames. *"Flying straight up" against "falling vertically down"; "touched his beard with his left hand" against "right hand".*
The question points by speech	The moment is marked by a spoken line, which no frame encoder can read. *"What characters appear after the subtitle 'happily as they walk towards their rooms'?"*
Flat contrast	Every option matches the video equally, disc is near zero across the video, and selection falls back on rel. *"What material is this boat made of?" Options: plastic, wood, ceramic, stainless steel, glass.*
A wrong option names what is everywhere	A wrong option mentions an object visible in most frames, so the contrast gathers evidence for it. *A man crouching in front of a camera;* FULCRUM *selects for "picked up the camera and took a photo".*