GridProbe: Posterior-Probing for Adaptive Test-Time Compute in Long-Video VLMs

Mohamed Eltahir[1,2] **Lama Ayash**[1] **Ali Habibullah**[1] **Tanveer Hussain**[3*] **Naeemullah Khan**[1‡]
[1] King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[2] Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia
[3] Department of Computer Science, Edge Hill University, Ormskirk, England
{mohamed.hamid, lama.ayash, ali.habibullah}@kaust.edu.sa
hussaint@edgehill.ac.uk, naeemullah.khan@kaust.edu.sa

Abstract

Long-video understanding in VLMs is bottlenecked by a single monolithic forward pass over thousands of frames at quadratic attention cost. A common mitigation is to first *select* a small subset of informative frames before the forward pass, which most training-free selectors do via auxiliary *encoder-space* similarities. Such signals are capped by contrastive pretraining, which usually fails on reasoning-heavy queries (negation, cross-frame counting, holistic summarization). We propose ***GridProbe***, an efficient training-free *posterior-probing* inference paradigm that scores evidence in *answer space* using a frozen VLM's own reasoning and then selects question-relevant frames adaptively, resulting in attention cost sub-quadratic in the number of sampled frames with little to no accuracy loss. We arrange frames on a $K \times K$ grid and run lightweight row *R* and column *C* probes, where each probe reads its peak posterior as a query-conditioned confidence. The outer product of *R* and *C* yields an interpretable *importance map* whose skewness and kurtosis drive *Shape-Adaptive Selection*, a closed-form rule that reliably replaces the fixed frame budget M with a per-question M_{eff}. We show empirically that M_{eff}, surprisingly, tracks intrinsic question difficulty without ever seeing the answer, a sign of test-time adaptive compute. On Video-MME-v2, **GridProbe** matches the monolithic baseline within 1.6 pp Avg Acc at $3.36\times$ TFLOPs reduction, while on LongVideoBench it Pareto-dominates the baseline. Because the selector and QA models can be decoupled, pairing a small 2B selector with a stronger 4B or 8B QA is also strictly Pareto-dominant over the 2B monolithic baseline, with no retraining. The procedure transfers unchanged to InternVL3.5 and Qwen2.5-VL backbones. Finally, the interpretability of the importance maps opens future avenues for behavioral diagnostics, grounding, and frame-selection distillation.

1 Introduction

Modern video VLMs process long videos by compressing many frames into one forward pass. Qwen3-VL-2B [1], for example, uses an adaptive per-frame resolution that crushes individual frames to ≈ 240 visual tokens when 2048 frames are passed, an order of magnitude below the ≈ 2960 tokens per frame at the 64-frame setting. This trade-off exchanges per-frame fidelity for temporal coverage and reflects a structural limit: per-token cost is dominated by the (linear-in-tokens) FFN at current scales while attention adds asymptotically quadratic-in-sequence-length growth on top, so reducing the number of input tokens delivers the strongest compute savings, and even models trained with 256K-token contexts cannot afford dense sampling *and* dense attention at scale.

*Corresponding Author. ‡Principal Investigator (PI).
Code: https://www.github.com/mohammad2012191/GridProbe

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

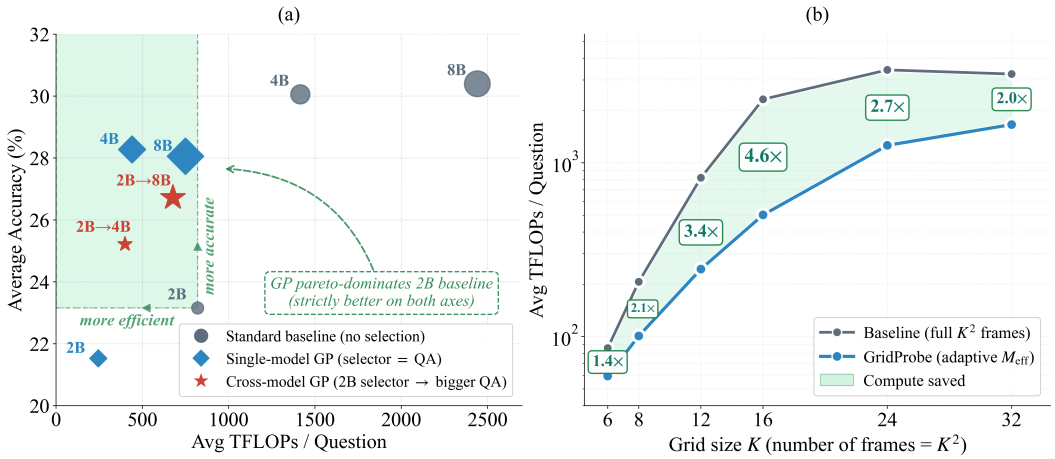

Figure 1: **(a) VMME-V2 Pareto across QA model sizes.** Single-model and cross-model **GridProbe** variants in the green region Pareto-dominate the 2B baseline (higher accuracy at lower compute). **(b) Compute reduction across K at fixed 2B QA.** The compute-reduction ratio is hump-shaped in K, peaking at $4.6\times$ near $K{=}16$.

An orthogonal response is *frame selection*: pick the $M \ll N$ most informative frames and run the VLM on only those. Recent training-free selectors (MDP3 [2], CLIP-matching, SigLIP-based scoring) and learned variants (Frame-Voyager [3], Focus [4], HFS [5]) share a common structure: frames and the query are embedded by separate vision and text encoders, and a similarity function in that shared space scores each frame. We call this paradigm **encoder-space selection**. Its weakness is documented: MDP3's own qualitative analysis shows SigLIP-matching failing on negation, cross-frame counting, and summarization queries, because these queries typically require reasoning outside the encoder's representational capacity.

We argue for a stronger move than swapping in a better selector. The VLM already knows which frames matter, it just needs to be asked. If we feed the VLM a subset of frames with the query, its posterior over the answer space encodes how confidently it can answer *given that subset* (Figure 3). High confidence on a small subset implies those frames carry the answer. This observation motivates a different inference paradigm rather than a different selector.

To this end, we introduce **GridProbe** (Figure 2), a training-free *posterior-probing* inference paradigm that replaces the standard one-shot forward pass with a self-probing recipe. We factorize the candidate frame pool into a $K \times K$ grid and run lightweight, axis-aligned probe passes over the rows and columns through a frozen VLM. The outer product of the row and column peak-posterior confidences yields a question-conditioned importance map. By default, the same frozen VLM serves as both the selector and the answerer. We further show that the two roles can be decoupled for a strict Pareto improvement.

This single design shift has three structural consequences. First, the selection signal is *reasoning-grounded*, it inherits the VLM's full reasoning capacity, so negation, cross-frame counting, and compositional queries are handled natively rather than being lost in contrastive embedding. Second, the signal needs no retraining and transfers across model families. Third, the maps are mechanically interpretable, rendering the model's evidence-gathering legible at the frame level. Notably, the current formulation reads a peak posterior over a finite answer space.

Once frames are scored, selectors must determine how many to pass to the final model. Existing methods enforce a static budget M, creating an unavoidable trade-off: they waste compute on highly localized questions and bottleneck accuracy on holistic ones by discarding necessary context. Agentic methods adapt the budget by re-sampling frames over several reasoning rounds [6, 7, 8, 9], which multiplies the generative calls per question. Crucially, the **GridProbe** importance map resolves this natively. We demonstrate empirically that the shape of this importance distribution strongly correlates with question difficulty (Figure 4, right). Rather than using a static frame budget, we utilize this

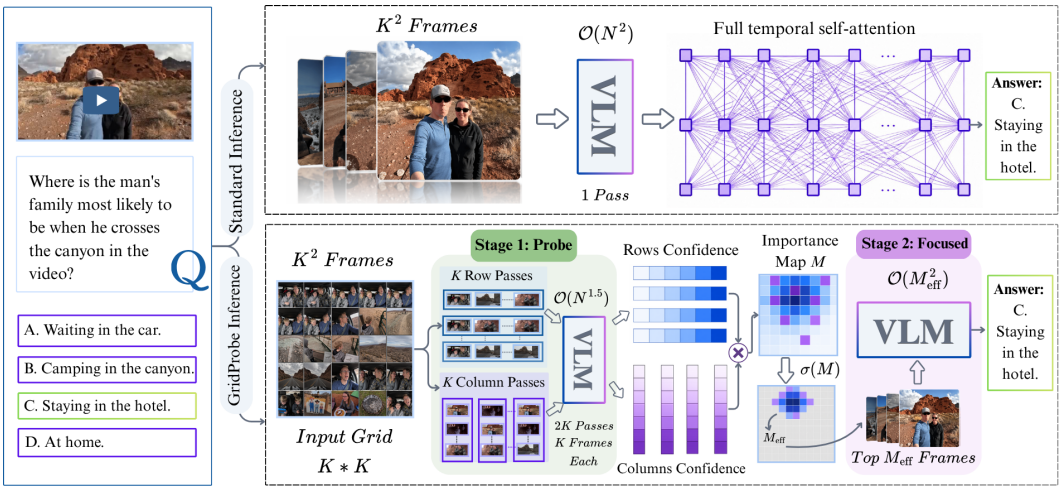

Figure 2: **GridProbe** pipeline. Stage 1: $2K$ row/column probes on K^2 candidate frames yield an importance map. Stage 2: one focused pass on the top-M_{eff} cells, sized adaptively from the map's distribution shape.

insight to introduce **shape-driven adaptive test-time compute**, which sets the per-question size M_{eff} via a closed-form rule on the map's skewness and kurtosis.

Coupling answer-space probing with shape-driven adaptive selection yields **GridProbe**, a training-free posterior-probing inference paradigm for long-video VLMs. Three findings anchor our empirical claims: (a) Pareto-dominant cross-model composition without retraining, (b) Pareto-efficient single-model operation, and (c) Adaptive test-time compute mirrors intrinsic difficulty.

Contributions:

Posterior-probing inference paradigm. We formalize **GridProbe**, a training-free inference method for long-video VLMs with attention cost sub-quadratic in the number of sampled frames that operates in answer space rather than encoder space, replacing the standard one-shot forward pass.

Question-conditioned importance map. A per-question, frame-level importance map exposes the VLM's evidence-gathering for each query, making long-video understanding interpretable.

Shape-driven adaptive test-time compute. A closed-form statistic on the importance map distribution replaces the fixed frame budget M with a per-question M_{eff} that adapts to the question difficulty.

The Redundancy Principle. Positive-skew (sparse peaks) and negative-skew (redundant high-importance) maps are different distribution shapes that share the same selection answer.

2 Related Work

Long-video VLMs and the cost of monolithic inference. Recent video VLMs such as Qwen3-VL [1], InternVL3.5 [10], and LLaVA-Video [11] scale to thousands of frames via extended context windows combined with adaptive per-frame visual-token budgets. Despite design differences, all share a structural commitment to a single monolithic forward pass with quadratic attention in input length $O(N^2)$. Even 256K-token contexts cannot afford dense attention over dense sampling, so reducing the cost of this single forward at inference time, without retraining the backbone or compromising visual fidelity, has become a practical priority.

Encoder-space frame selection. A dominant mitigation is to score and select a subset of informative frames before the forward pass. Training-free methods rely on similarities in vision-language encoder space (CLIP [12], SigLIP [13]). FOCUS [4] adds adaptive exploration over this signal, while MDP3 [2] generalizes ranking into a list-wise subset optimization that captures query relevance, diversity, and sequential structure. Learned variants (Frame-Voyager [3], HFS [5]) train auxiliary scoring heads or fine-tune the backbone to emit selection signals, trading training complexity for

accuracy. We collectively call this family *encoder-space selection*: the selection signal is computed in a representation space structurally separate from the QA model's reasoning, and its quality is therefore bounded by what that space was trained to encode. Reasoning-heavy queries (negation, cross-frame counting, holistic summarization) routinely defeat encoder-space signals that the QA model itself could resolve natively.

Multimodal frame scoring and the static-budget assumption. Recent work pushes scoring closer to the QA model. FRAG [14] evaluates each frame with a multimodal model and selects the top-M, which moves the signal from encoder space to model space but remains frame-wise (no temporal context, no reasoning about evidence sufficiency). Independently of the scoring axis, prior frame-selection methods share a second assumption: the selection size M is fixed a priori, wasting compute on localized queries (where $M \ll K^2$ suffices) and starving holistic queries (where the answer is genuinely dispersed). A scoring signal that captures sub-frame reasoning and a per-question budget that adapts to the shape of the evidence both remain open.

Test-time compute and agentic video inference. A growing body of work allocates *test-time compute* adaptively to improve answer quality. Text-domain efforts include longer chain-of-thought, self-consistency, and search-based decoding [15]. In the video domain, the closest prior work uses LLM-based agents to route compute per question. VideoAtlas [9] represents a video as a hierarchical grid explored by a Master-Worker agent loop, achieving logarithmic compute growth with video duration. VideoAgent [16] and AVUA [17] similarly use LLM agents that recursively re-sample frames based on their own intermediate reasoning. AVP [6] seeks evidence iteratively with a frozen planner, and VideoZoomer [7] and FrameThinker [8] learn temporal focusing with reinforcement learning. Streaming methods such as HERMES [18] compress the video into a KV cache before any question arrives. These systems achieve adaptive per-question compute by orchestrating multi-step agent loops. They inherit the orchestration overhead, control-flow complexity, and per-question planning costs of multi-step inference. A non-iterative, fixed-schedule mechanism that delivers comparable adaptive-compute behavior without agent orchestration is absent from this line of work.

Three threads converge on the same problem from different angles, each leaving a complementary gap. Encoder-space frame selection decreases input volume but operates in a representation space disconnected from the QA model's reasoning. Multimodal frame scoring bridges to model space but stays frame-wise and locks M a priori. Agentic adaptive inference routes per-question compute through multi-step agent orchestration. What is missing across all three threads is a *fixed-schedule training-free* mechanism that scores in the QA model's own answer space, captures cross-frame reasoning rather than per-frame similarity, and sizes the per-question budget in closed form. We describe how **GridProbe** fills all these gaps in the next section.

3 Methodology: GridProbe

3.1 Setup and Notation

Let $V = \{f_0, \ldots, f_{n-1}\}$ be an ordered sequence of video frames and q a natural-language query. The answer space $\mathcal{Y}$ depends on the task (for multiple-choice, $|\mathcal{Y}| = 4$). A frozen VLM θ defines a conditional probability distribution $p_\theta(\cdot \mid S, q)$ over $\mathcal{Y}$ for any frame subset $S \subseteq V$ paired with q. We define the *probe confidence* (c) as the peak of this posterior:

$$c(S, q) \;=\; \max_{y \in \mathcal{Y}} \; p_\theta(y \mid S, q). \tag{1}$$

Intuitively, $c(S, q)$ measures how confidently the model can commit to a single answer given S. We use this as a proxy for relevance: high confidence implies S contains frames needed to answer q, while a flat posterior signals that the subset lacks the evidence to discriminate among the candidates. Figure 3 contrasts this answer-space signal with encoder-space selection, where the score is a similarity computed by independent vision and text encoders.

3.2 Grid Formulation and Importance Map

We sample K^2 frames uniformly from V and index them as a conceptual $K \times K$ grid. For each row $r \in \{0, \ldots, K-1\}$ and column $c \in \{0, \ldots, K-1\}$, we define

$$S_r^{\text{row}} \;=\; \{\, f_{rK+j} \,\}_{j=0}^{K-1}, \qquad S_c^{\text{col}} \;=\; \{\, f_{c+jK} \,\}_{j=0}^{K-1}. \tag{2}$$

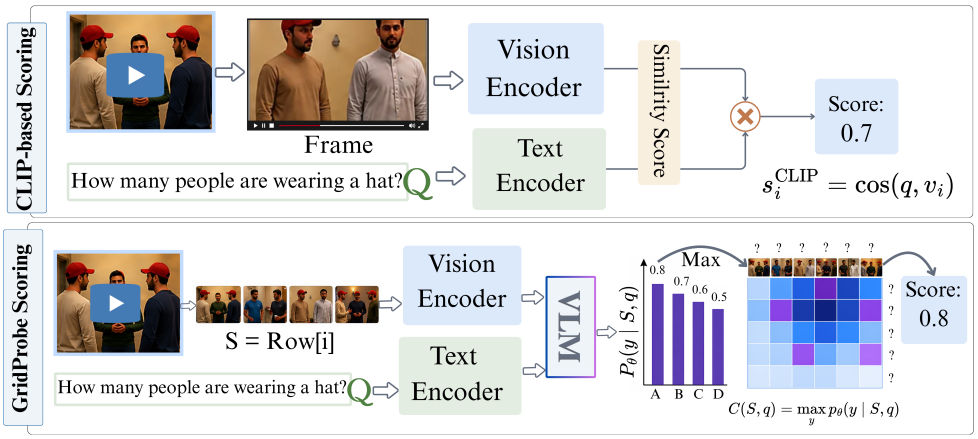

Figure 3: Encoder-space (top) vs answer-space (bottom) selection signals. Encoder-space scoring computes scalar similarity from independent vision and text encoders, while answer-space scoring reads the probe confidence directly from the QA VLM's posterior over answer candidates.

giving K row subsets (local temporal coverage) and K column subsets (strided, periodic coverage). In total, $2K$ probe passes are required, each seeing only K frames.

The K row subsets provide **local temporal coverage**: each row groups K contiguous frames from a localized segment of the video timeline, exposing fine-grained event-local evidence. The K column subsets provide **strided periodic coverage**: each column groups K frames at stride K, sampling the full timeline at uniform intervals and exposing distributed or recurring evidence. The two axes are complementary: any grid cell (r, c) is uniquely indexed by the intersection of one local row and one global column, so the same frame is scored once from a local-context view and once from a global-context view. Prior multimodal frame scoring [14] computes per-frame evidence one frame at a time, requiring K^2 forward passes to score all K^2 candidates. Our row+column factorization recovers a cell-level importance map at only $2K$ axis-level forward passes, each seeing only K frames.

For each axis subset we compute the probe confidence via Eq. 1: $c_r^{\text{row}} = c(S_r^{\text{row}}, q)$ and $c_c^{\text{col}} = c(S_c^{\text{col}}, q)$. The joint importance ($M$) of the grid cell (r, c), corresponding to frame f_{rK+c}, is the product

$$M[r, c] \;=\; c_r^{\text{row}} \cdot c_c^{\text{col}}. \tag{3}$$

Intuitively, a cell is important only if *both* the row and the column containing it produce confident answers (regardless of whether they are correct or not, as high confidence indicates relevance, not correctness). If only one marginal is confident, the cell is assigned moderate weight (partial evidence) and downweighted if neither.

In summary, the grid factorization combines local and strided periodic coverage in a single $O(K)$-pass scoring stage and produces a cell-level question-conditioned importance map without the per-frame scoring overhead of prior multimodal-scoring approaches.

3.3 Adaptive Selection Size via Distribution-Shape Statistics

Given the $K \times K$ importance map M, we need to pick how many cells to keep. A static M_{eff} is suboptimal: holistic questions benefit from many frames while localization queries need only a few. Our central observation is that these question types leave distinct fingerprints on M itself. A localization query concentrates evidence in a few cells, producing a sharply peaked, right-skewed map. A redundancy-heavy query spreads high importance across many overlapping cells, producing a left-skewed map. A holistic query distributes evidence broadly but sparsely, producing a near-uniform map. Across question types the shape of M co-varies with how hard the question is to answer from few frames, so we hypothesize that *distribution shape itself is an indirect signal for the optimal selection size* M_{eff}.

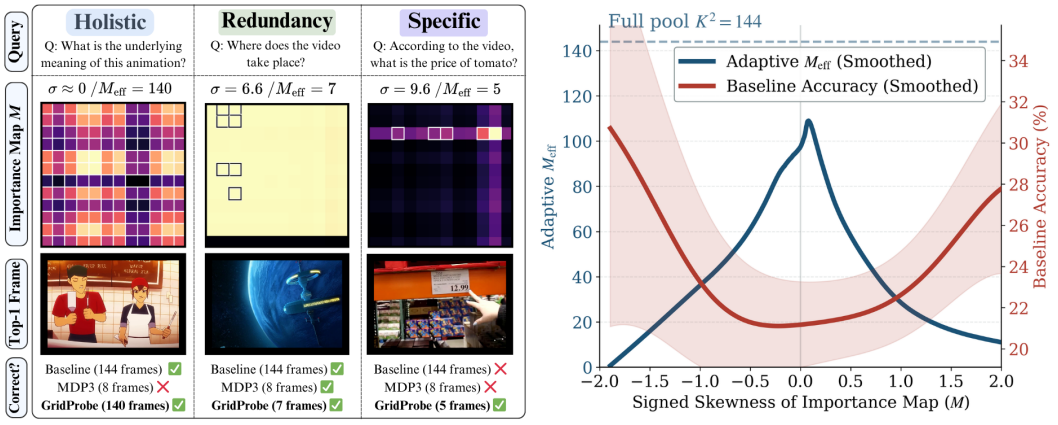

Figure 4: **Left:** three V2 queries exercise three distribution-shape regimes (Holistic, Redundancy, Specific). For each we show the question, importance map M, the top-1 selected frame by **GridProbe**, and per-method correctness. **GridProbe** answers all three correctly. MDP3 (paper-default fixed M=8) misses Holistic and Specific. The K^2=144 baseline misses Specific. **Right: GridProbe**'s adaptive M_{eff} (blue) and the 2B baseline accuracy (red), smoothed across signed $\text{skew}(M)$ on V2 (K=12, n=3,200). The two curves mirror each other: both signed extremes route to small M_{eff} on intrinsically easier questions, while the near-uniform middle gets near-K^2 coverage on intrinsically harder ones, an empirical realization of the redundancy principle (§3.3).

To act on this hypothesis we capture shape with two complementary moments combined into a single statistic $\sigma(M)$ that drives the adaptive size:

$$\sigma(M) \;=\; \big|\,\text{skew}(M)\,\big| \;+\; 0.5 \cdot \max\big(0,\, \text{kurt}_{\text{ex}}(M)\big), \qquad M_{\text{eff}} \;=\; \left\lceil \frac{K^2}{1 + \gamma_0 \cdot K \cdot \sigma(M)} \right\rceil. \quad (4)$$

Here $\text{skew}(\cdot)$ is the third standardized moment (asymmetric concentration of evidence) and $\text{kurt}_{\text{ex}}(\cdot)$ is the excess fourth standardized moment (peakedness). Each captures a complementary departure from uniformity. Skewness detects evidence biased toward a small subset of cells. Excess kurtosis detects sharp peaks even in symmetric distributions. On a perfectly uniform map the sample variance is zero and the standardized moments are formally undefined; we set σ=0 in this degenerate case (implemented numerically via a variance threshold), so $M_{\text{eff}} = K^2$ and the method falls back to the full pool, equivalent to the monolithic baseline. On a one-hot map ($\sigma \to \infty$ formally), $M_{\text{eff}} \to 1$. In practice M_{eff} varies smoothly between these extremes per question. The half-weight on kurtosis downweights its larger absolute scale relative to skewness. The factor of K in the denominator (rather than just $\gamma_0\sigma$) keeps M_{eff} growing linearly with K on peaked maps instead of quadratically. Without it, doubling K to gain finer probe resolution would also quadruple M_{eff} on the same map, undoing the focused-pass savings.

§4.5 validates this distribution shape hypothesis empirically and shows how $\sigma(M)$ helps to allocate more compute exactly to the questions the QA model finds intrinsically hardest.

Why $|\text{skew}|$? The redundancy principle. The absolute value collapses two regimes that have opposite distribution geometries but *identical* selection requirements. A right-skewed map (positive skew, mass at low importance) is the sparse-peak regime, where a few decisive frames carry the answer and the rest can be discarded. A left-skewed map (negative skew, mass at high importance) is the redundancy regime, where most frames are individually informative for the query but show overlapping content, so a small representative subset suffices. The truly compute-hungry case is the low $|\text{skew}|$ near-uniform map, where evidence is sparse-and-dispersed across the timeline and full coverage is warranted.

Figure 4 (right) makes the inverted-U pattern in M_{eff} explicit: both signed extremes of $\text{skew}(M)$ route to small M_{eff} while only the near-uniform middle draws near K^2 coverage, confirming that $|\text{skew}|$ correctly groups the two "few-needed" regimes together. Figure 4 (left) realizes the three regimes qualitatively on Video-MME-v2 clips, where questions produce M_{eff} from 140 (holistic) to 5 (specific).

3.4 Two-Stage Inference Pipeline

GridProbe (Figure 2) combines the probe and a focused pass:

1. **Stage 1 (probe):** run K row-passes and K column-passes on K-frame subsets through the frozen VLM. Record $2K$ probe confidences and build M via Eq. 3.
2. **Compute** M_{eff} from the shape statistic in Eq. 4.
3. **Stage 2 (focused pass):** select the M_{eff} frames corresponding to the top entries of M, denoted $S^\star$. Run the VLM once on $S^\star$ at full resolution and read off the final answer as $\arg\max_y p_\theta(y \mid S^\star, q)$.

3.5 Complexity

A monolithic pass on $N{=}K^2$ frames has attention cost $O(N^2)$ in the attention-dominated regime. All costs below are in N, the number of sampled frames, which tracks video length. **GridProbe** runs $2K$ probe passes of K frames each (cost $O(K \cdot K^2) = O(N^{1.5})$) plus one focused pass of M_{eff} frames. For non-uniform importance maps $M_{\text{eff}} \ll N$ and the total attention cost is $O(N^{1.5} + M_{\text{eff}}^2)$, sub-quadratic. In the worst case (perfectly uniform maps) $M_{\text{eff}} \to N$ and the focused pass falls back to the monolithic baseline. Empirically, FFN cost (linear in tokens) dominates at current model scales. The probe stage runs at reduced spatial resolution ($224{\times}224$ in our experiments), making each probe forward markedly cheaper than a full-resolution pass. So the $2K$ probe passes plus a focused pass on M_{eff} full-resolution frames remain net-cheaper than a single full-resolution pass on all N frames (Fig. 1).

4 Experiments

4.1 Experimental Setup

We evaluated on Video-MME-v2 [19] (8-option MCQ, 3,200 questions across 800 videos with three-level cognitive hierarchy and grouped non-linear scoring, reported visual-only with no subtitles) and LongVideoBench [20] (with subtitles). We further report Video-MME [21] (no subtitles) and EgoSchema [22]. Backbones are Qwen3-VL-Instruct [1] (2B, 4B, 8B), InternVL3.5 [10] and Qwen2.5-VL [23], all frozen at inference. Unless stated otherwise, $K{=}12$ (a $12{\times}12$ grid yielding a 144-frame candidate pool), $\gamma_0{=}0.25$ in Eq. 4, probe resolution $224{\times}224$ pixels, and uncapped focused-pass resolution. Frame sampling draws K^2 frames uniformly from the video timeline. We reported Average Accuracy, the official Non-Linear grouped score (VMME-V2 only), per-question TFLOPs, and time to first token (TTFT) on one A100. Video-MME and EgoSchema setups are in App. A.

4.2 Main Results: Video-MME-v2 and LongVideoBench

Same-model results. On Qwen3-VL-2B at $K{=}12$ (Block 2 of Table 1), **GridProbe**($M{=}$AUTO) trades 1.63 pp Avg Acc on V2 for a $3.36\times$ TFLOPs reduction, and reaches a Pareto-dominant point on LVB ($+0.9$ pp at $0.35\times$ compute). The trade-off is broadly invariant across QA size: comparing each GP-X to its same-size monolithic baseline, the accuracy cost is -1.63/-1.78/-2.88 pp on V2 at 2B/4B/8B and -1.7/-2.0 pp on LVB at 4B/8B, shifting modestly toward higher accuracy cost at larger backbones because stronger QAs extract more from the full-frame baseline. Video-MME shows the same pattern at about $3.3\times$ less compute. EgoSchema stays close to the baseline at a smaller saving, because its median evidence span covers about 100 s of each 180 s clip [22], so no small subset carries the answer. The gain of **GridProbe** therefore acts as a soft indicator of a benchmark's temporal certificate. As a side benefit, GP-8B reaches 28.06% Avg Acc on V2 at 842 TFLOPs, matching the 2B baseline's compute (820 TFLOPs) within 3% at $+4.9$ pp accuracy, a near-matched-compute upgrade for users willing to deploy the 8B answerer.

Cross-model: a free Pareto move. Pairing the 2B selector with a stronger QA (Block 3 of Table 1, visualized in Fig. 1(a)) Pareto-dominates the 2B-monolithic baseline on both benchmarks: $+3.56$ pp Avg Acc at $0.83\times$ compute on V2 and $+3.30$ pp at $0.85\times$ on LVB for GP-2B$\to$8B, with GP-2B$\to$4B delivering an even larger LVB gain ($+4.0$ pp at $0.52\times$, widening to $+4.5$ pp on the 3600-sec bin). The mechanism is straightforward: attention on M_{eff} frames at 8B is cheaper than on K^2 frames at

Table 1: Main results on Video-MME-v2 (no subtitles), LongVideoBench, Video-MME, and EgoSchema at $K{=}12$. `GP-X` denotes single-model **GridProbe** (selector = QA = Qwen3-VL-X). `GP-X→Y` denotes cross-model pipelines. Non-Lin is V2's official grouped score. Long is LVB's 3600-second bin. The Qwen3-VL-2B baseline is the comparison anchor for the lower blocks. numbers in gray are for reference. **Bold** marks the best **GridProbe** operating point per block on V2 and LVB.

	Video-MME-v2			LongVideoBench			Video-MME		EgoSchema	
Pipeline	Non-Lin	Avg Acc	TFLOPs	Long Acc	Overall	TFLOPs	Acc	TFLOPs	Acc	TFLOPs
Monolithic Baselines (full $K^2{=}144$ pool). 2B is the comparison anchor.										
Qwen3-VL-2B	9.45	23.16	820	49.8	56.4	868	61.9	757	65.8	107
Qwen3-VL-4B	14.11	30.06	1415	57.3	64.1	1493	69.9	1304	73.2	199
Qwen3-VL-8B	14.91	30.94	2441	55.1	62.7	2574	72.3	2238	73.6	370
*Single-Model **GridProbe**: efficient trade-off at fixed model size (vs same-size baseline).*										
GP-2B	8.39	21.53	245	51.4	57.3	301	59.3	214	64.8	76
GP-4B	**12.86**	**28.28**	**440**	**55.9**	**62.4**	**575**	66.1	391	69.4	154
GP-8B	12.60	28.06	842	53.4	60.7	1068	**67.7**	**711**	**72.0**	**299**
*Cross-Model **GridProbe**: Pareto-dominance over the 2B baseline (2B probe → larger QA).*										
Uniform-M_{eff} → 8B	10.83	26.22	677	49.8	58.5	735	–	–	71.0	161
GP-2B→4B	10.76	25.22	399	**54.3**	**60.4**	**452**	–	–	–	–
GP-2B→8B	**11.70**	**26.72**	**677**	52.0	59.7	735	–	–	**71.4**	**161**

Table 2: Selector quality at fixed $M{=}8$, 2B QA. Both use the exact same per-question frame budget. Only the scoring selector differs.

2B QA, $M{=}8$	V2 NL / Acc	LVB Long / Ov
+ MDP3	7.38 / 20.09	44.9 / 51.0
+ **GridProbe**	7.49 / 20.12	47.3 / 54.6
Δ	**+0.11 / +0.03**	**+2.4 / +3.6**

Table 3: Matched-budget decomposition on LVB, same model, about 53 of 144 frames. Size is M_{eff} minus fixed, Rank is GP minus M_{eff}, both before rounding.

	Uniform				
Backbone	Fixed	M_{eff}	GP	Size	Rank
Qwen3-VL-2B	54.6	53.7	**57.3**	−0.8	+3.6
InternVL3.5-2B	54.9	52.9	**56.8**	−2.0	+3.9
Qwen2.5-VL-7B	54.2	52.4	**59.2**	−1.8	+6.8
Qwen3-VL-8B	59.3	59.5	**60.7**	+0.2	+1.3
InternVL3.5-8B	61.9	59.7	**62.2**	−2.3	+2.6

2B because sequence length dominates parameter count in the multi-frame regime, and the larger QA produces sharper answer posteriors on the focused subset. The Uniform-M_{eff}→8B control (gray row in Table 1) uses GridProbe's per-question M_{eff} but draws those frames *uniformly* from the K^2 pool (no importance ranking). Against this matched-compute baseline, the importance ranking adds $+0.50$ pp on V2 and $+1.20$ pp on LVB, close to the $+1.3$ pp it adds to a same-model 8B answerer (Table 3). A stronger answerer recovers more from weaker frames, so the ranking matters most at small answerers.

Decomposing the same-model gap. Tables 2 and 3 factor the gap into selector quality (vs MDP3 at matched $M{=}8$) and the split between ranking and sizing. At fixed $M{=}8$, the two selectors are matched on V2 but diverge sharply on LVB, where **GridProbe** wins by $+3.6/+2.4$ pp Overall / 3600s while MDP3 lands -4.9 pp *below* the no-selection baseline. Encoder-space scoring is not just suboptimal there but actively harmful. Figure 4 (left) illustrates the failure mode qualitatively: MDP3 misses both the holistic and specific cases that **GridProbe** answers correctly. Since MDP3 itself dominates a broad suite of training-free selectors (CLIP-based scoring, scene-change detection, optical flow, Frame-Voyager [3], and others) [2], the gap extends transitively over that family. Table 3 then holds the mean budget at about 53 of 144 frames on LVB and changes one component at a time. Adaptive sizing with uniform frames never beats fixed uniform sampling by more than 0.2 pp, so the accuracy comes from the ranking. The ranking adds $+3.6$ to $+6.8$ pp at the 2B and 7B operating points that a compute-saving method targets, and $+1.3$ to $+2.6$ pp at 8B. The role of adaptive sizing is compute allocation. It sets each question's budget from the map with no tuning, while the fixed control is only definable in hindsight as the mean of M_{eff}.

Table 4: Positioning on a shared backbone (InternVL3.5-2B, Video-MME Long, n=900, one A100). Uniform and **GridProbe** run under vLLM [24] with prefix caching, and the other methods use their released code at default settings. TTFT is per question, with the per-video ingest of HERMES amortised.

Method	How frames reach the answerer	Acc	TTFT
Uniform, 144 frames	no selection	44.6	1.9 s
HERMES [18]	KV cache compressed before the question	42.4	1.4 s
AVP [6]	multi-round agentic loop	43.6	81.3 s
GridProbe	$2K+1$ scored forward passes	**47.0**	5.4 s

4.3 The Ranking Transfers Across Families and Paradigms

Model families. The procedure runs unchanged on InternVL3.5 and Qwen2.5-VL, and the ranking beats both matched uniform controls on every same-model backbone (Table 3). Against the full 144-frame pass, two factors set how much the ranking can change the outcome. First, the 144-frame pass of InternVL3.5 fills 38K of its 41K-token context, and **GridProbe** lifts it by $+10.0$ pp at 2B and $+5.5$ pp at 8B on LVB. Second, inside LVB the ranking gains on 600 s and 3600 s videos and loses on 15 to 60 s clips, where uniform frames already cover the timeline (Table 8).

Agentic and streaming methods. Table 4 compares them on a shared backbone. **GridProbe** is the only method above uniform sampling. HERMES answers faster by compressing each video once before questions arrive, and AVP spends over a minute per question. Its latency depends on the backbone, since the $2K$ probes see $2K^2$ frame instances and pay off only when a probe frame costs under half the tokens of a full-resolution frame. The adaptive per-frame budget of Qwen3-VL gives a $4.2\times$ gap, and GP-2B answers in 2.0 s against 4.6 s for the monolithic pass on LVB. InternVL3.5 uses a fixed per-frame budget, so there **GridProbe** trades latency for accuracy.

4.4 Probe Confidence Tracks Evidence

LongVideoBench annotates the frames that carry each question's evidence, so the relevance proxy of Eq. 1 can be scored directly. The map ranks an annotated frame in its top 5 of 144 at twice the chance rate, and the lift over chance is positive in every duration bin and at every selector size (Tables 5 and 9). The signal depends on the question. When we inject the evidence frame at a random grid slot and swap only the question text for one from another video, its top-1 recovery falls from 6.0% to 0% (App. C). A confidently wrong probe also cannot corrupt which frames are chosen. Selection reads only the ranking of M, and any monotone miscalibration $c \mapsto c^p$ maps M to M^p. The selected set is identical under $p \in \{0.5, 2, 3\}$ on all 3,200 V2 questions. Miscalibration changes how many frames are taken, never which ones.

4.5 Adaptive Compute Mirrors Intrinsic Question Difficulty

Figure 4 (right) provides direct empirical evidence that **GridProbe**'s adaptive selection responds to question content. Firstly, the M_{eff} curve is symmetric in skew sign: both extremes of the importance-map skewness axis route to small M_{eff}, while the near-uniform middle gets near-full coverage. This is the redundancy principle of §3.3 made empirical: positive-skew (sparse peaks) and negative-skew (redundant high-importance frames) are different distribution shapes with the *same* selection answer. Secondly, baseline accuracy mirrors the M_{eff} curve: questions in the near-uniform regime are intrinsically harder ($\sim 21\%$ baseline accuracy) while questions at either tail are easier (~ 28 to 32%), so the two curves visibly mirror each other across $\text{skew}(M)$. The selector's compute allocation tracks intrinsic difficulty without ever observing the answer. Thirdly, the adaptive variability is quantitatively large: the cross-question coefficient of variation (CV, standard deviation over mean) in compute is 0.78 for **GridProbe** versus 0.018 for the fixed-input baseline ($44\times$ higher), at $0.30\times$ the per-question average compute. This level of input-dependent compute variability is a signature of adaptive test-time compute that no fixed-input baseline can produce.

Table 5: Selector-size ablation. VMME-V2 stratified n=400. $\bar{M}_{\text{eff}}$ is the per-question average. Gold@8 is the recall lift over random for annotated LVB evidence frames at a top-8 budget.

Selector	QA	Avg Acc	$\bar{M}_{\text{eff}}$	TFLOPs	Gold@8
2B	2B	**26.25%**	**52.9**	**233.3**	+7.9
4B	2B	26.00%	56.2	302.0	+7.3
8B	2B	24.75%	58.5	415.1	+4.1

Table 6: Frame-order ablation. V2 stratified n=400. Same frames, different input order.

Frame order	Avg Acc	Δ
Temporal (default)	**26.25%**	(ref.)
Importance (descending)	25.00%	−1.25 pp

5 Ablation Study

We run three ablations: selector size in cross-model pairings, the temporal vs. importance order of the focused pass, and the constants of the shape rule. App. C adds an image-collation efficiency variant. All ablations use Qwen3-VL-2B at K=12 on Video-MME-v2 with M=AUTO unless stated.

Selector size. Holding the QA fixed at 2B, scaling the selector from 2B to 8B *decreases* Avg Acc while increasing TFLOPs (Table 5), the opposite of the naive capability-scaling expectation.The per-question average $\bar{M}_{\text{eff}}$ is roughly stable across selector sizes ($52.9 \to 56.2 \to 58.5$ for 2B/4B/8B), which suggests that the accuracy degradation is not driven by more aggressive selection. Instead, localization saturates at the small selector. The gold-frame recall lift of the 8B selector is about half that of the 2B (Table 5), so a larger selector only adds probe compute.

Frame ordering. The focused pass receives the top-M_{eff} frames in temporal order. Passing them in descending-importance order (same frames, different positional encoding) costs -1.25 pp Avg Acc (Table 6). The effect is small but consistent: temporal ordering preserves the positional encoding the VLM was trained to read.

6 Conclusion and Limitations

We introduced **GridProbe**, a training-free posterior-probing paradigm in which row and column probes over a $K \times K$ grid produce a question-conditioned importance map, and a closed-form shape statistic sets the per-question budget M_{eff}. **GridProbe** delivers Pareto-efficient single-model operation on Video-MME-v2 and Pareto-dominant operation on LongVideoBench, the cross-model variant (2B selector with a stronger QA) Pareto-dominates the 2B-monolithic baseline on both benchmarks without retraining, the importance ranking carries the gain and transfers across model families, and the per-question M_{eff} tracks intrinsic question difficulty without ever observing the answer.

A few caveats and natural refinements remain. The TFLOPs reduction is most pronounced when the focused pass dominates inference, and more modest with large prompts (e.g., LVB's ∼700–1,000 subtitle tokens that every probe re-processes) or small grid sizes (K<10, where the $2K$ probe passes themselves become a non-trivial fraction of total cost). The cross-model pipeline shifts cost from compute toward host memory, since both selector and QA are loaded simultaneously. Two refinements are natural follow-ups: allowing γ_0 to adapt to video length or pool density, and extending the shape statistic to other per-cell importance signals such as attention magnitudes or retrieval scores.

Acknowledgments and Disclosure of Funding

We are grateful to the KAUST Academy for its generous support, and especially to Prof. Sultan Albarakati, who made this work possible. Funding in direct support of this work: the first author was a visiting student at the KAUST Academy, which provided access to KAUST research resources. For computer time, this research used Ibex managed by the Supercomputing Core Laboratory at King Abdullah University of Science & Technology (KAUST) in Thuwal, Saudi Arabia. Competing interests: the authors declare no financial relationships outside the submitted work that could be perceived as influencing it.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

[3] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, volume 2025, pages 84154–84179, 2025.

[4] Zirui Zhu, Hailun Xu, Yang Luo, LIU YONG, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, volume 2026, pages 98463–98486, 2026.

[5] Yiqing Yang and Kin-Man Lam. Hfs: Holistic query-aware frame selection for efficient video reasoning. *arXiv preprint arXiv:2512.11534*, 2025.

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

[7] Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pages 20087–20111, 2026.

[8] Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. In *International Conference on Learning Representations*, volume 2026, pages 74904–74933, 2026.

[9] Mohamed Eltahir, Ali Habibullah, Yazan Alshoibi, Lama Ayash, Tanveer Hussain, and Naeemullah Khan. Videoatlas: Navigating long-form video in logarithmic compute. *arXiv preprint arXiv:2603.17948*, 2026.

[10] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[11] Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8592–8603, 2025.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[13] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952. IEEE, 2023.

[14] De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. Frag: Frame selection augmented generation for long video and long document understanding. *arXiv preprint arXiv:2504.17447*, 2025.

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[16] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024.

[17] Sullam Jeoung, Goeric Huybrechts, Bhavana Ganesh, Aram Galstyan, and Sravan Bodapati. Adaptive video understanding agent: Enhancing efficiency with dynamic frame sampling and feedback-driven reasoning. *arXiv preprint arXiv:2410.20252*, 2024.

[18] Haowei Zhang, Shudong Yang, Jinlan Fu, See-Kiong Ng, and Xipeng Qiu. HERMES: KV cache as hierarchical memory for efficient streaming video understanding. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8411–8430, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[19] Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

[20] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[21] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118. IEEE, 2025.

[22] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

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

[24] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

[25] Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhu Chen. Videoeval-pro: Robust and realistic long video understanding evaluation. *arXiv preprint arXiv:2505.14640*, 2025.

A Implementation Details

Models. All experiments use Qwen3-VL-Instruct backbones at three sizes (2B, 4B, 8B parameters) loaded from HuggingFace Transformers. Transfer experiments (§4.3) use InternVL3.5 (2B, 8B) and Qwen2.5-VL-7B with the same procedure and constants. We perform zero-shot inference: no fine-tuning, no chain-of-thought prompting, no tool-use. Final answers are read off via letter-token scoring, taking $\arg\max_y p_\theta(y \mid \cdot)$ over $y \in \{A, B, C, D, E, F, G, H\}$ (or the letter set for LongVideoBench's 5-way questions).

Hyperparameters. The defaults used throughout the paper are $K{=}12$ (so $K^2{=}144$ frames in the pool), $\gamma_0{=}0.25$ for the M_{eff} closed-form rule, and a small variance threshold to set $\sigma{=}0$ in the degenerate near-uniform case (§3.3). The probe stage runs at reduced spatial resolution $224{\times}224$, and the focused stage runs at the model's native resolution with Qwen3-VL's adaptive per-frame token allocation.

Pipeline. For each (video, question) pair, we extract $K^2{=}144$ frames uniformly from the video duration. Stage 1 runs $2K{=}24$ probe passes through the frozen VLM (K row passes and K column passes), where each pass collates $K{=}12$ frames into a single $\lceil\sqrt{K^2}\rceil{\times}\lceil\sqrt{K^2}\rceil$ tiled image at $2048{\times}2048$. Each pass produces a softmax confidence over the answer letters; we take the top-1 confidence as the probe score per row/column. The importance map is $M[r, c] = c_r \cdot c_c$ (Eq. 3). We then compute the shape statistic $\sigma = |\text{skew}(M)| + \frac{1}{2}\max(0, \text{kurt}_{\text{ex}}(M))$ and the per-question budget $M_{\text{eff}} = K^2/(1 + \gamma_0 K\sigma)$. Stage 2 runs a single forward pass on the top-M_{eff} frames (by M score) at the model's native resolution.

Benchmarks. **Video-MME-v2** (visual-only): $n{=}3{,}200$ questions across 800 videos, grouped into three cognitive levels (App. B.1). We report Non-Lin (V2's official grouped score) and Avg Acc. **LongVideoBench validation set** (with subtitles): subtitles are concatenated to the prompt as text in both the probe and focused stages. We report Long (3600-second bin) and Overall accuracy. **EgoSchema** uses its 500-question labeled subset of 180 s clips.

Hardware. All experiments ran on a single node with $8{\times}$ NVIDIA A100 GPUs. We shard each evaluation across the 8 GPUs by sample id (interleaved partitioning to balance the duration-bin distribution across shards) and merge per-shard JSON outputs. TTFT for Qwen3-VL-2B is measured on one A100 over 20 LVB questions with HuggingFace Transformers, summing the probe and focused passes and excluding frame decoding, which both methods share.

B Detailed Breakdown Tables

This section provides the per-bin breakdown referenced from Table 1.

B.1 Video-MME-v2: Per-Level and per-Group-Type

Video-MME-v2 organizes questions into three cognitive levels (L1: Information Aggregation, L2: Temporal Understanding, L3: Complex Reasoning) and two group types (*Consistency* for capability-consistency groups, *Coherence* for reasoning-coherence groups). The Consistency block contains 519 of 800 groups and the Coherence block contains 281 of 800 groups. Table 7 reports Non-Lin score for each method at each breakdown axis.

Against the same-model 2B baseline, **GridProbe**($M{=}$AUTO) *wins* on Level 2 temporal understanding ($\Delta{=}{+}0.37$ Non-Lin), where evidence is concentrated at a transition-bearing event, and *loses* most on Level 1 multi-point aggregation ($\Delta{=}{-}1.92$ Non-Lin), where evidence is genuinely distributed across many timestamps and no small subset suffices. The σ-driven adaptive M degrades gracefully toward K^2 on near-uniform maps, so the method does not overcommit to selection when evidence is dispersed but has nothing to gain when it is already fully covered.

B.2 LongVideoBench: Full Per-Duration Breakdown

The main table reports only LongVideoBench's *Long* bin and *Overall*. Table 8 adds the four duration buckets. **GridProbe**($M{=}$AUTO) *wins* on the 600-sec (+2.2 pp) and 3600-sec (+1.6 pp) bins, where uniform K^2 undersamples the timeline and question-conditioned selection picks the relevant moments, and *loses* on the 15-sec and 60-sec bins (-3.1 pp each), where 144 uniform frames already saturate the timeline and selection is unnecessary. The same pattern appears on V2's cognitive levels (§B.1): the method earns its keep where evidence is sparse and degrades gracefully when it is already fully covered.

Table 7: V2 Non-Lin breakdown by cognitive level (L1/L2/L3) and group type (Consistency/Coherence). Visual-only, K=12, n=3,200.

Method	Level 1	Level 2	Level 3	Consistency	Coherence	Overall
Single-model: Qwen3-VL-2B-Instruct						
Baseline (no selection)	13.83	7.50	8.30	10.18	8.12	9.45
+ **GridProbe** (M=AUTO)	11.91	7.87	6.72	9.13	7.03	8.39
Same-model: Qwen3-VL-4B-Instruct						
Baseline (no selection)	19.43	12.56	12.12	15.04	12.40	14.11
+ **GridProbe** (M=AUTO)	18.98	10.73	10.82	14.20	10.40	12.86
Same-model: Qwen3-VL-8B-Instruct						
Baseline (no selection)	18.98	15.19	12.32	15.53	13.75	14.91
+ **GridProbe** (M=AUTO)	16.18	11.78	11.11	12.95	11.97	12.60
Cross-model: 2B selector → 4B QA						
4B baseline (no selection)	19.43	12.56	12.12	15.04	12.40	14.11
+ **GridProbe** (2B selector, ours)	14.42	9.30	9.68	11.37	9.64	10.76
Cross-model: 2B selector → 8B QA						
8B baseline (no selection)	18.98	15.19	12.32	15.53	13.75	14.91
+ **GridProbe** (2B selector, ours)	15.28	11.26	9.93	12.02	11.12	11.70

Table 8: LongVideoBench Avg Acc % by source-video duration (15s / 60s / 600s / 3600s / Overall), K=12, with subtitles. Δ rows: **GridProbe**'s gain over the corresponding no-selection baseline, except the fixed-uniform row, which samples the mean M_{eff} of about 53 frames uniformly.

Method	15-sec	60-sec	600-sec	3600-sec	Overall
Single-model: Qwen3-VL-2B-Instruct					
2B baseline (no selection)	72.9	77.3	55.3	49.8	56.4
+ **GridProbe** (M=8, fixed)	66.7	69.1	57.3	47.3	54.6
+ **GridProbe** (M=AUTO)	69.8	74.2	**57.5**	**51.4**	**57.3**
Δ **GridProbe** (auto) vs. 2B base.	−3.1	−3.1	**+2.2**	**+1.6**	**+0.9**
Δ **GridProbe** (auto) vs. fixed uniform	−2.3	−4.1	**+3.6**	**+4.4**	**+2.7**
Same-model: Qwen3-VL-4B-Instruct (selector = QA)					
4B baseline (no selection)	81.4	79.4	64.6	57.3	64.1
+ **GridProbe** (M=AUTO, same-model)	79.8	**80.4**	61.7	55.9	62.4
Same-model: Qwen3-VL-8B-Instruct (selector = QA)					
8B baseline (no selection)	80.6	74.2	64.8	55.1	62.7
+ Uniform sampling	77.5	76.3	60.9	51.1	59.3
+ **GridProbe** (M=AUTO, same-model)	76.7	74.2	62.6	53.4	60.7
Cross-model: 2B selector → 4B QA					
+ Uniform sampling	77.5	78.4	60.0	50.4	58.8
+ **GridProbe** (2B selector, ours)	76.7	75.3	60.2	54.3	**60.4**
Δ **GridProbe** (2B→4B) vs. 2B base.	**+3.8**	−2.0	**+4.9**	**+4.5**	**+4.0**
Cross-model: 2B selector → 8B QA					
+ Uniform sampling	79.8	75.3	59.7	49.8	58.5
+ **GridProbe** (2B selector, ours)	76.0	74.2	61.7	52.0	59.7
Δ **GridProbe** (2B→8B) vs. 2B base.	**+3.1**	−3.1	**+6.4**	**+2.2**	**+3.3**

C Additional Analyses

Gold-frame recall. LongVideoBench marks the frames that carry each question's evidence. We map them to the K^2 pool and test whether the importance map ranks them highly (Qwen3-VL-2B selector, K=12, n=1,200, random baseline averaged over 20 draws). Table 9 reports recall at small budgets and the lift at the per-question M_{eff} budget by subset.

Table 9: Gold-frame recall on LongVideoBench. **Left:** recall at small budgets and the mean normalised rank of the best gold frame. **Right:** recall lift over random at the M_{eff} budget, by number of gold frames and by duration.

Budget	**GridProbe**	Random
Top-5	11.6	5.7
Top-8	16.9	9.0
Mean rank	0.36	0.50

Gold frames	1	2+	3+	
Lift (pp)	+8.9	+3.6	+1.0	
Duration	15 s	60 s	600 s	3600 s
Lift (pp)	+8.1	+14.5	+6.4	+4.2

Question-swap injection. On 100 LVB questions at $K{=}6$, we insert the annotated evidence frame at a random grid slot and probe the identical frames twice, once with the matched question and once with a question from another video. Top-1 recovery is 6.0% against 0.0% (chance 2.8%), and the mean normalised rank moves from 0.44 to 0.58. We use $K{=}6$ because each probe shows K frames at once, so at $K{=}12$ a single injected frame moves a 12-frame verdict and its effect dilutes.

Shape-rule sensitivity. Table 10 sweeps γ_0 and the kurtosis weight of Eq. 4. Using signed skew in place of $|\text{skew}|$ would hand the 23.8% of V2 questions with negative-skew maps 122 frames on average, against 70 with $|\text{skew}|$.

Table 10: Shape-rule sensitivity on V2 with Qwen3-VL-2B.

γ_0	0.10	**0.25**	0.50	1.00
Avg Acc	26.5	26.0	22.5	23.3
Mean frames	78	53	37	25
Kurtosis weight	0.25	**0.5**	1.0	2.0
Mean M_{eff}	55.2	53.4	51.0	48.3

Image collation. As an efficiency variant, the focused pass can composite the top-M_{eff} frames into a single $\lceil\sqrt{M_{\text{eff}}}\rceil{\times}\lceil\sqrt{M_{\text{eff}}}\rceil$ tiled image at $2048{\times}2048$ (with empty cells when M_{eff} is not a perfect square), reducing compute to one image's worth of tokens at the cost of temporal positional encoding and per-frame pixel budget. Collation reaches $0.29\times$ the standard two-stage's compute at a -1.16 pp Avg Acc cost (Table 11), making it a credible operating point when extreme compute is the constraint.

Table 11: Collated single-image variant vs the standard M_{eff}-frame two-stage, V2 $n{=}1{,}800$.

Method	Avg Acc	TFLOPs
Baseline (no selection)	23.10	819.6
Two-stage (M_{eff} frames as video)	**21.27**	232.2
Collated (one $\sqrt{M_{\text{eff}}}{\times}\sqrt{M_{\text{eff}}}$ image)	20.11	**66.6**

Open-ended pilot. For open-ended QA the probe asks "Can this question be answered from these frames? Answer Yes or No." and reads $p(\text{Yes})$. The outer product, the shape rule, M_{eff}, and the focused pass are unchanged, and the final pass generates freely. On 80 VideoEval-Pro [25] questions with Qwen3-VL-2B, graded by Qwen3-VL-30B-A3B-Instruct against the reference answer, **GridProbe** scores 22.5% against 21.2% for the 144-frame pass, using $3.5\times$ fewer tokens and $3.9\times$ less wall clock. VideoEval-Pro re-poses questions from existing benchmarks without options, so this pilot isolates the change of answer format.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 list three findings (Pareto-dominant cross-model composition, Pareto-efficient single-model operation and adaptive test-time compute that mirrors intrinsic question difficulty); each is supported by results in §4.2, §4.3, §4.4, §4.5, §5, and Table 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §6 discusses the regimes where the TFLOPs reduction is most modest (large prompts, small grid sizes), the host-memory cost of cross-model deployment, the backbone dependence of latency gains, and the weaker row and column factorization when evidence spans several frames.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include formal theorems with proofs. The closed-form M_{eff} rule (§3.3) and complexity analysis (§3) are derivations and informal arguments, not theorem–proof statements.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The method is fully specified in §3 (grid formulation, M_{eff} rule, two-stage pipeline, complexity); experimental setup is documented in §4 and §A (models, hyperparameters, frame extraction, benchmarks, hardware).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use the publicly available Video-MME-v2 [19] and LongVideoBench [20] validation sets, Video-MME [21], EgoSchema [22], and VideoEval-Pro [25], and the publicly available Qwen3-VL-Instruct [1], InternVL3.5 [10], and Qwen2.5-VL [23] models. Source code (probe + focused-pass pipeline, evaluation scripts, plotting and analysis utilities) is publicly available at `https://www.github.com/mohammad2012191/GridProbe`.

   Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The work is training-free (no fine-tuning, no optimizer). Test details (model checkpoints, K=12, γ_0=0.25, probe vs. focused resolutions, benchmark splits, evaluation protocol) are documented in §4 and §A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Inference is fully deterministic (greedy letter-token scoring, fixed uniform frame extraction). We evaluate on the complete test sets (V2, LVB, VMME). EgoSchema uses its labeled subset.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Hardware (1 node, 8× NVIDIA A100 GPUs) is documented in §A. Per-question wall time and TFLOPs are reported alongside accuracy in Table 1 and Tables 5–11, with TTFT in Table 4 and §A, both overall and per duration bin.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. The work uses only publicly available pre-trained models and benchmarks, conducts no human-subjects research, and releases inference code without high-risk capabilities.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: **GridProbe** is an inference-time frame-selection algorithm applied to existing pre-trained video VLMs and public benchmarks; it does not introduce new generative capabilities, new datasets, or new model artifacts. Its societal-impact profile is inherited from the underlying VLM and benchmark assets, which are addressed by their respective creators.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases inference code only; no new pre-trained models, generative artifacts, or scraped datasets are released. The underlying assets (Qwen3-VL, Video-MME-v2, LongVideoBench, InternVL3.5, Qwen2.5-VL, Video-MME, EgoSchema, VideoEval-Pro) are existing public resources released by their respective creators.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All underlying assets are credited via citations: Qwen3-VL-Instruct [1] (probe and answerer backbone), Video-MME-v2 [19] (benchmark), and LongVideoBench [20] (benchmark), InternVL3.5 [10] and Qwen2.5-VL [23] (transfer backbones), Video-MME [21], EgoSchema [22], and VideoEval-Pro [25] (benchmarks), HERMES [18] and AVP [6] (released baseline code), and vLLM [24] (serving). Each asset is used per its respective public license; exact license terms are documented in the original sources.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release inference code with documentation at `https://www.github.com/mohammad2012191/GridProbe`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All evaluation uses publicly released benchmarks whose human annotation was performed by the benchmark authors.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subjects research; IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The core method uses pre-trained Vision-Language Models (Qwen3-VL-Instruct at 2B, 4B, and 8B sizes [1]) as both the probe and the answerer in a two-stage inference pipeline (§3, §A). Transfer experiments use InternVL3.5 and Qwen2.5-VL in the same roles, and the open-ended pilot is graded by Qwen3-VL-30B-A3B-Instruct (App. C). The method is training-free: we do not fine-tune or modify the underlying VLM.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.