STABLE CURVES, UNSTABLE ITEMS: ITEM-LEVEL SCALING HETEROGENEITY IN VIDEO LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

More visual input often improves average accuracy in video language models. However, aggregate gains do not reveal whether models preserve answers they already get right. We present a paired evaluation protocol that tracks the same questions across visual budgets within frozen models. Our study covers four models and four multiple-choice benchmark splits on a common budget grid, including text-only input. On these grids, 12.5–25.5% of questions exhibit at least one correct-to-incorrect transition as the budget increases. Choosing the best budget for each question in hindsight yields 8.8–18.9 percentage points higher accuracy than the best fixed budget. This gap reveals complementary successes that no single budget captures. We release over 120,000 per-item evaluation records and analysis code to support reproducible evaluation of visual scaling. Our findings show that larger visual budgets change which questions are answered correctly, not simply how many.

1 INTRODUCTION

Choosing a visual budget for a video language model requires judging what additional input buys. Average accuracy summarizes the net benefit, but does not reveal whether a higher budget preserves answers the model already gets right. This distinction matters when comparing frame counts or resolutions (Li et al., 2025; Zohar et al., 2025): an improving score can reflect both new successes and lost ones. To judge what is gained by a budget increase, we need to know both which questions become solvable and which earlier successes are lost.

Figure 1 makes this gap concrete. For Qwen2.5-VL-7B on Video-MME v1 short, moving from $128f$ to $256f$ changes correctness on 5.2% of questions despite a net accuracy loss of only 0.6 percentage points. Opposing changes largely cancel, so aggregate stability does not imply stable answers.

Prior work documents non-monotonic aggregate scaling (Brkic et al., 2025; Zohar et al., 2025) and example-level training dynamics (Swayamdipta et al., 2020). Our question instead concerns the same item's response to inference-time visual input changes within a frozen model. Knowing which questions are difficult at one budget does not tell us whether a success will persist at another, while an average curve does not identify which questions change. We therefore pair each model's responses to the same questions across budgets, making the gains and regressions directly observable even when the mean improves.

We evaluate four models on each of four multiple-choice splits from Video-MME and MLVU using the shared $\{\text{text}, 16f, 32f, 64f\}$ grid, where text denotes text-only input. An item oracle selects each question's best observed outcome in hindsight. A larger configuration set can increase this oracle simply by offering more choices; the shared grid separates this coverage effect from differences in model behavior. It matches requested frame budgets, not architecture-dependent visual tokens or computation.

In every model–split combination, 12.5–25.5% of all questions are correct at some lower budget but wrong at some higher one. Oracle accuracy is 8.8–18.9 percentage points above the best fixed accuracy on the same grid (Table 2). This retrospective upper bound quantifies complementary successes among the evaluated budgets. Thus the budget with the highest mean does not capture all the successes available on the grid, and its aggregate advantage can coexist with losses on individual questions. We then test whether this complementarity depends on answer options or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

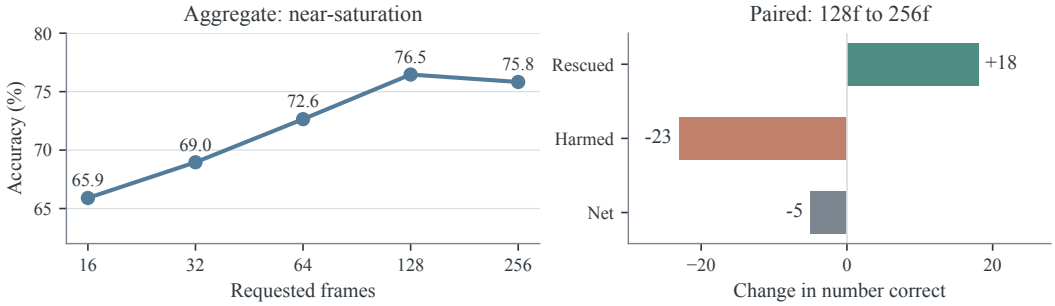

Figure 1: **Stable aggregate accuracy can conceal opposing item-level transitions.** Qwen2.5-VL-7B on Video-MME v1 short (n=786). Accuracy changes reflect rescues minus harms, divided by the item count. The final budget transition changes correctness on 41 items (5.2%), despite a net loss of only five correct answers (0.6 points).

input construction, and examine what it offers adaptive inference. The distinction is consequential: knowing that budgets have complementary successes does not itself identify which budget to select or establish a saving once all selection probes are counted.

Our contributions are:

- **A paired evaluation protocol for visual scaling.** We track the same questions across budgets within frozen models, separating gains from regressions and comparing complementarity on a common grid with explicit input-protocol provenance.
- **Systematic evidence of hidden item-level trade-offs.** Across all 16 model–split comparisons, we quantify how improved aggregate performance coexists with lost correct answers. Generation and input-control studies establish where the finding extends and how the affected items depend on the evaluation protocol.
- **A reproducible resource for budget-aware evaluation.** We release over 120,000 per-item records, input-protocol documentation, and analysis code to support audits of visual scaling and evaluation of adaptive inference strategies.

2 Evaluation Design and Paired Measurements

Our comparison holds the model and questions fixed, pairs outcomes across visual budgets, and controls the configurations available to the oracle. This separates three questions: whether the mean improves, whether earlier successes survive, and whether any fixed budget captures the successes available across the grid.

We evaluate Qwen2.5-VL-7B, Qwen3-VL-8B, InternVL3-8B, InternVL3.5-8B, and LLaVA-NeXT-Video-7B (Bai et al., 2025b;a; Zhu et al., 2025a; Wang et al., 2025b; Zhang et al., 2024) on Video-MME v1/v2, MLVU, and AVSD (Fu et al., 2025; 2026; Zhou et al., 2025; Alamri et al., 2019) (Table 1). The four main models share the $\{\text{text}, 16f, 32f, 64f\}$ MCQA grid. LLaVA uses $\{\text{text}, 8f, 16f, 32f\}$ under the recorded HF input setting and is analyzed separately. Matching the grid controls the configurations available to the oracle; model-default visual encoders are retained, so it does not imply equal visual tokens or FLOPs across architectures.

For a frozen model, let $y_{i,c} \in \{0, 1\}$ indicate correctness on item i at configuration $c \in C$. A configuration specifies frame count, spatial budget, and input protocol. The trajectory $\tau_i(C) = (y_{i,c})_{c \in C}$ pairs outcomes on the same n items. Frame-count grids are ordered by requested frames at fixed spatial budget. Equal-budget sampling policies are compared as alternatives, not assigned an artificial low-to-high order.

The best fixed accuracy and item-oracle accuracy are

$$A_{\text{fixed}} = \max_{c \in C} \frac{1}{n} \sum_i y_{i,c}, \qquad A_{\text{oracle}} = \frac{1}{n} \sum_i \max_{c \in C} y_{i,c}. \tag{1}$$

Table 1: Evaluation scope. Complete evaluated splits require exact item counts and no missing, duplicate, or error records. Per-model coverage and protocols are in Appendix A.1.

Regime	Datasets / items	Models	Trajectory signal
MCQA	V1 short 786; V1 medium 639; V2 medium 2,516; MLVU 2,170	5	option correctness
Open generation	MLVU: 201 QA + 217 summaries	3	Token-F1; ROUGE-L
Controlled dialogue	AVSD validation: 1,787 turn-10 targets	2	Token-F1; ROUGE-L
Input interventions	V1 short: 786 items / 262 videos	1	paired correctness

Their difference, *oracle headroom*, measures complementarity within the evaluated grid. It is a retrospective upper bound, not the expected gain of a deployable selector. An item is *visually confused* if it is correct at some lower budget and wrong at some higher budget. The reported rate includes the text-only configuration and uses all n items as its denominator. *Text overwrite* is the specific text-correct, visual-wrong transition at a stated target budget; conditional overwrite rates instead divide by the number of text-correct items and are explicitly identified.

For an ordered pair $c < c'$, the rescued and harmed fractions are

$$R_{c,c'} = \frac{1}{n}\sum_i (1 - y_{i,c})y_{i,c'}, \qquad H_{c,c'} = \frac{1}{n}\sum_i y_{i,c}(1 - y_{i,c'}). \tag{2}$$

The accuracy change is $A_{c'} - A_c = R_{c,c'} - H_{c,c'}$; pairwise churn is $R_{c,c'} + H_{c,c'}$. Thus a small net change does not bound either directional flow.

For generation, we replace $y_{i,c}$ with a deterministic per-item score $s_{i,c} \in [0, 1]$ in Equation 1 and call the difference a *score oracle gap*. A score regression occurs when some later budget reduces an item's score by more than δ. We use $\delta = 0.02$, with sensitivity checks at 0.01 and 0.05. Token-F1 is the primary metric and ROUGE-L a robustness check. Both measure reference overlap; we report score regressions separately from binary MCQA confusion.

Each cell records model, split, frame/pixel budgets, sampling, input size, and raw/cache path. The Qwen2.5 V1-short anchor uses raw-video decoding; cached cells use 1-fps JPEG frames. MCQA uses deterministic option scoring or answer generation per adapter; open generation is greedy. We use 5,000 paired item bootstraps, resampling item IDs jointly across configurations and reselecting best fixed; the resampling unit is an item, not a video cluster. McNemar tests compare binary policies. Overlap nulls preserve set sizes; exploratory tests use permutation and Benjamini–Hochberg false-discovery-rate (BH-FDR) correction. Appendix A.1 specifies model and generation settings.

3 Hidden Trade-offs Across Models and Budgets

Higher aggregate accuracy does not imply that a model retains the questions it could already answer. This gap recurs across models and benchmarks, but need not involve the same questions.

Qwen2.5-VL-7B accuracy on Video-MME v1 short rises from 65.9% at $16f$ to 76.5% at $128f$, then falls by 0.6 points at $256f$. Across adjacent visual budgets, 5–15% of items change correctness. The small net change at the final transition thus hides opposing flows, rather than indicating stable answers (Figures 1 and 2). The trajectory, rather than the mean alone, reveals both the successes acquired and those lost along this curve.

No fixed budget captures all observed successes in any of the 16 model–split comparisons. On the common text/$16f$/$32f$/$64f$ grid (Table 2), all retain 8.8–18.9 points of headroom and 12.5–25.5% visual confusion. Video-MME v2 expands the answer space from four to as many as eight options; MLVU changes the benchmark family. Both retain substantial complementarity. The matched results control frame-budget coverage, while preserving each model's documented input protocol.

LLaVA tests architectural breadth on its supported text/$8f$/$16f$/$32f$ grid. Across all 786 V1-short items, its best fixed accuracy is 45.93%, oracle accuracy 57.76%, and confusion rate 16.16%. Across all 2,170 MLVU MCQA items, these values are 48.02%, 60.46%, and 17.70%. The finding therefore extends beyond the two model families in the common-grid comparison.

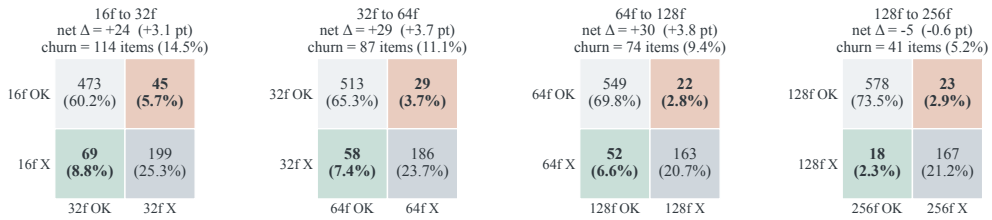

Figure 2: **The mean records net change, not total churn.** Adjacent-budget transition matrices for the Qwen2.5 V1-short raw-video grid. Teal and orange cells count rescued and harmed items. At $128f \rightarrow 256f$, 18 rescues and 23 harms produce a -0.6-point net change but 41 correctness changes.

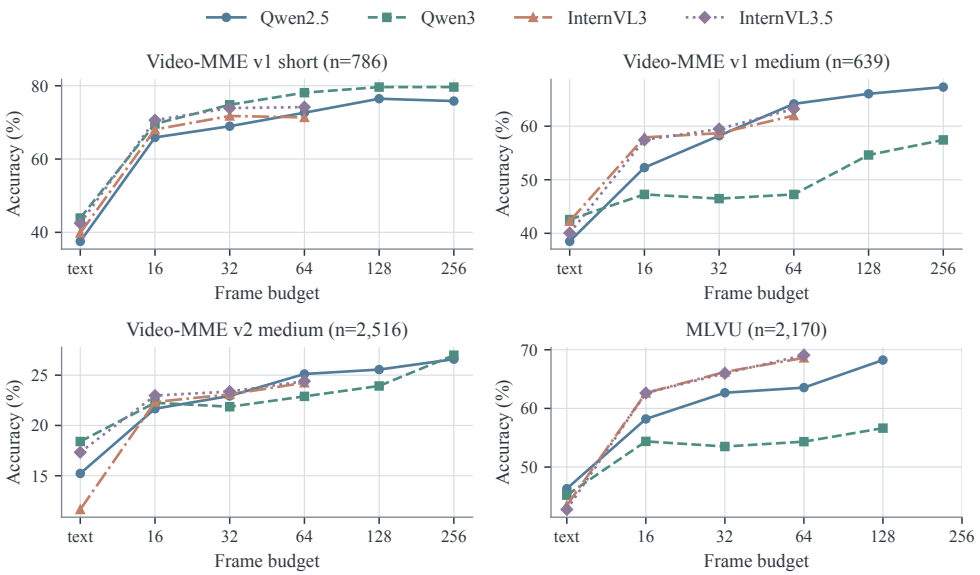

Figure 3: **Aggregate scaling across models and benchmark splits.** The curves mostly improve or saturate. Table 2 pairs outcomes within each model and restricts the comparison to the shared four-budget grid, revealing both headroom and harmful transitions beneath these trends. Reduced-input-size InternVL fallbacks are omitted.

The recurrence of confusion does not identify a common set of vulnerable questions. On the matched V1-short grid, Qwen2.5 and InternVL3 share 33 confused items (Jaccard 17.4%), above the 7.5% matched-size random expectation (2.30×, $p = 2.2 \times 10^{-6}$). Yet 157 of the 190 items in their union occur in only one confusion set. Shared susceptibility therefore coexists with substantial model-specific variation.

Text overwrite occurs in all four main models (Figure 4). At the stated target budgets, it affects 5.0–7.3% of all V1-short items, or 13.2–17.1% of text-correct items. Among the six largest task types, Counting has the highest conditional overwrite rate for every model. For Qwen2.5, 39/786 items are text-correct but wrong at $128f$, and the conditional Counting overwrite rate is 37.1%. Thus even successes obtained without visual evidence need not survive its addition.

4 Beyond Multiple-Choice Evaluation

Budget complementarity also appears in reference-overlap scores on open-generation tasks. Across two complete generation splits, increasing the visual budget usually improves mean reference overlap, yet individual scores can fall. We measure these changes with deterministic Token-F1 and check robustness with ROUGE-L; scores use a 0–100 scale.

Table 2: Frame-budget-matched MCQA results on text/$16f$/$32f$/$64f$. Right-hand entries give oracle headroom in accuracy points / visual-confusion rate. V1s, V1m, and V2 denote Video-MME v1 short, v1 medium, and v2 medium.

	Best-fixed accuracy (%)				**Headroom (pts) / confusion (%)**			
Model	V1s	V1m	V2	MLVU	V1s	V1m	V2	MLVU
Qwen2.5-VL-7B	72.6	64.2	25.1	63.5	+9.8/15.9	+10.6/19.7	+9.6/13.0	+14.5/20.3
Qwen3-VL-8B	78.1	47.3	22.9	54.4	+9.5/16.3	+18.9/23.3	+10.6/13.2	+18.6/25.5
InternVL3-8B	71.8	62.0	24.2	68.6	+8.8/12.5	+14.1/20.0	+13.6/17.5	+11.6/18.8
InternVL3.5-8B	74.2	63.2	24.4	69.1	+10.6/14.1	+13.3/20.3	+10.7/13.8	+13.1/20.7

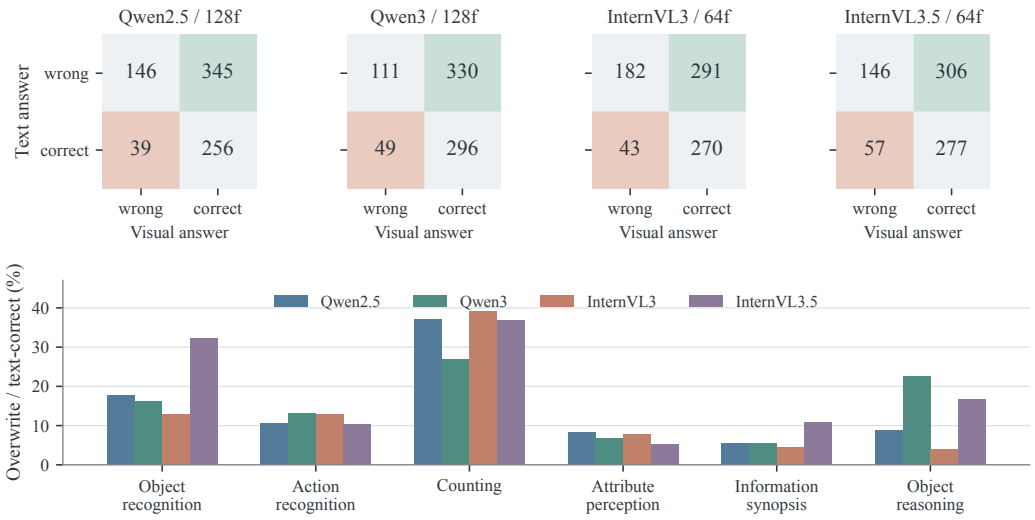

Figure 4: **Text overwrite across four models on Video-MME v1 short.** Top: paired text-only to visual correctness counts, using $128f$ for Qwen and standard-448 $64f$ for InternVL. Bottom: overwrite as a percentage of text-correct items within each of the six largest task types.

Complementarity persists within task groups, rather than arising only from pooling questions and summaries. On MLVU's 201 sub-scene questions and 217 video summaries, best-fixed Token-F1 ranges from 33.0 to 34.9, with item-oracle gaps of 2.7–3.7 points (Table 3). Gaps remain positive when QA and summarization are analyzed separately and under ROUGE-L (Table 4; Appendix B.2). Although the largest visual budget usually improves the mean, some items receive lower scores. Even with a stricter five-point drop threshold, 29.7–34.7% of items regress.

The finding also holds within longer-video and longer-summary strata. The $\geq$ 30-minute QA stratum ($n = 40$) and $\geq$ 100-reference-word summary stratum ($n = 130$) retain positive gaps for all three models, with all six paired-bootstrap lower bounds above zero. Videos reach 117.1 minutes and references 242 words (Appendix B.2).

Holding dialogue history fixed likewise leaves an oracle gap. For all 1,787 AVSD validation dialogues, turns 1–9 use the same reference history at every budget; only turn 10 is generated, without audio or the optional questioner summary. This isolates changes in current-turn visual input from changes in preceding responses. Both models improve in mean Token-F1 from text to $64f$, yet the item oracle remains 3.8–4.8 points higher. Gaps persist in both yes/no and other question groups and under ROUGE-L (Appendix B.2).

Complementarity thus extends beyond option correctness to continuous reference overlap, across task groups, length strata, and both metrics.

Table 3: Full-split generation results under Token-F1. Drop is the percentage of items with any later-budget decrease greater than 2 score points. Oracle gaps include 95% paired-bootstrap confidence intervals.

Dataset	Model / grid	n	Best fixed	Oracle	Gap [95% CI]	Drop
MLVU generation	Qwen2.5, text/16/64f	418	33.0	35.7	$+2.7\,[2.3, 3.1]$	43.1%
	Qwen3, text/16/64f	418	33.3	37.0	$+3.7\,[3.2, 4.2]$	49.0%
	LLaVA, text/8/32f	418	34.9	38.0	$+3.1\,[2.7, 3.6]$	53.6%
AVSD turn 10	Qwen2.5, text/16/64f	1,787	30.2	35.0	$+4.8\,[4.3, 5.2]$	36.3%
	Qwen3, text/16/64f	1,787	33.5	37.3	$+3.8\,[3.4, 4.2]$	29.2%

Table 4: Token-F1 oracle gaps by task group (0–100 score points). Long QA uses MLVU videos of at least 30 minutes; long summaries have at least 100 reference words. Dashes indicate model–task combinations not evaluated.

Model	MLVU QA	Summary	Long QA	Long summary	AVSD yes/no	AVSD other
Qwen2.5	+3.9	+1.6	+3.1	+1.7	+4.9	+4.7
Qwen3	+5.5	+1.8	+4.4	+1.7	+3.7	+4.0
LLaVA	+3.9	+2.3	+3.7	+1.9	–	–

5 Input Controls and Failure Analysis

Complementarity could reflect changes in evidence selection rather than visual budget alone. We examine this sensitivity through fixed-budget interventions, raw/cache replication, and sampling-based recovery on all 786 Qwen2.5 V1-short items (Table 5). These controls test the scope of the finding and which individual regressions alternative inputs can resolve.

5.1 Evidence Selection at Fixed Budgets

At $64f$ and 151K pixels per frame, uniform and random sampling are nearly tied in accuracy (73.92% versus 74.05%), yet exchange 18 correct-to-wrong and 19 wrong-to-correct items. Thus cancellation also occurs when frame count and spatial budget do not change. At a nominal 9.7M total-pixel budget, reallocating pixels between frames and resolution leaves 6.23 points of headroom (Figure 5). For example, $128f \times 76$K rescues 71 items relative to $32f \times 302$K, but loses 38 items that the latter answers correctly. Equal total pixels is an input-budget control, not a guarantee of equal tokens, FLOPs, or runtime. Expanding to the full 11-configuration frame–resolution grid yields an 87.9% oracle against a 76.5% best fixed accuracy (11.5 points of headroom). This expanded-grid gap describes a larger choice set than the matched frame-count comparison.

5.2 Repeatability and Protocol Dependence

Complementarity replicates across input protocols, whereas the affected items change. Repeating raw-video inference reproduces predictions and correctness for all 786 items at each visual budget. An independently executed 1-fps cache grid yields the same 8.78-point headroom; the paired difference is 0.00 points with 95% CI $[-1.78, +1.91]$. This numerical agreement coexists with different confusion rates (19.21% versus 15.90%). Across protocols, confusion labels agree on 87.0% of items, text-to-$128f$ overwrite on 96.3%, and the five trajectory groups on 81.4%.

Uniform sampling at different frame budgets need not produce nested frame sets; changing the budget can alter both the amount and temporal placement of evidence. Protocol records therefore identify the full input change associated with each measured transition.

5.3 Recoverable Regressions and Failure Hypotheses

Alternative sampling recovers part, but not all, of the observed regressions. For the independently executed cached text/$16f$/$32f$/$64f$ grid, 113 items are confused. Of these, $30/113 = 26.5\%$ change correctness across the three $64f$ sampling policies, compared with $58/673 = 8.6\%$ of other items

Table 5: Input interventions on Qwen2.5-VL-7B, V1 short ($n = 786$). Accuracy and confusion are percentages; gaps are accuracy points. Pixel budgets are per frame. Raw and cache are independent six-point frame-count grids.

Intervention grid	Best fixed	Oracle	Gap	Confusion
Sampling: uniform/random/dense-start, 64f–151K	74.05	77.74	+3.69	–
Resolution: 64f, 76K/151K/235K/360K	74.55	80.03	+5.47	–
Matched pixels: 32f–302K/64f–151K/128f–76K	75.32	81.55	+6.23	–
Raw video: text/16/32/64/128/256f	76.46	85.24	+8.78	19.21%
1-fps cache: text/16/32/64/128/256f	74.81	83.59	+8.78	15.90%

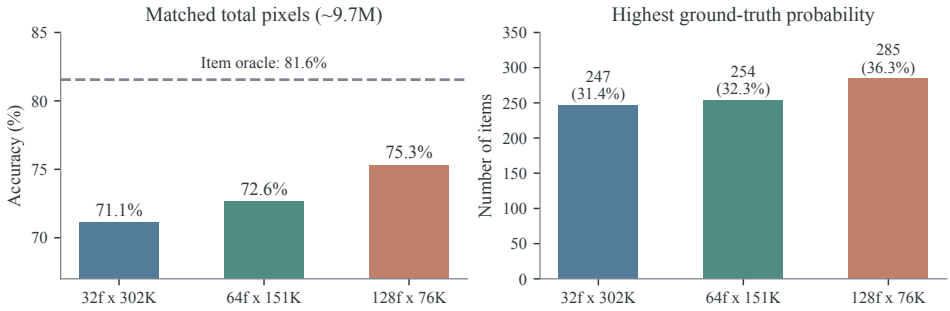

Figure 5: **Temporal–spatial complementarity at matched total pixels.** Left: accuracy at a nominal 9.7M total-pixel budget; the oracle exceeds the best fixed allocation by 6.2 points. Right: allocation maximizing the ground-truth option's probability for each item (247/254/285 items). These probability-based preferences are distinct from the binary correctness oracle and do not require a unique correct allocation.

(odds ratio 3.83, Fisher $p = 4.75 \times 10^{-7}$). These are cache-specific labels, distinct from the raw-anchor matched-grid labels in Table 2. Among 69 terminal-$64f$ regressions, alternative sampling recovers 20 (29.0%); the remaining 49 persist under all three policies.

A structured 12-case author audit covers sampling-recovered, sampling-persistent, counting, and OCR regressions (Appendix C.3). Undercounting of separated events persists across policies, suggesting temporal integration difficulty. An earlier visible clock value competing with a later queried event, and a brief target amid dominant scene content, suggest evidence competition. The intervention identifies a sampling-recoverable subset; persistent cases motivate distinct integration and referent-resolution hypotheses.

Across 786 questions and 262 videos, none of eight duration, frame-count, and visual proxies passes BH-FDR correction (maximum $|\rho| = 0.078$ and $|\delta_{\text{Cliff}}| = 0.087$, rounded). A multivariable question-side audit likewise detects no independent counting or numeric-option risk effect after permutation correction. These tests identify neither a reliable single-feature explanation nor a causal mechanism; the qualitative cases instead motivate targeted tests of temporal integration and competing evidence.

6 Implications for Adaptive Inference

Complementary successes create an opportunity for budget selection, not an automatic accuracy or efficiency gain. Realizing that opportunity requires both a useful decision signal and an accounting of the probes used to obtain it. Reference policies and retrospective diagnostics expose these two distinct constraints.

A margin-based $16f \to 32f \to 128f$ cascade matches fixed-$128f$ aggregate accuracy over all 786 Qwen2.5 V1-short items. Its mean terminal budget is 87.4 requested frames; charging every visited stage gives 118.9 total requested frames. The reduction relative to fixed $128f$ changes from 31.7% to 7.1%. Table 7 reports both. The margin threshold is selected and evaluated on this split, making the reported operating point an in-sample reference despite its label-free decision rule.

Table 6: Sampling intervention and diagnostic evidence on Qwen2.5 V1 short.

Diagnostic	Estimate	Interpretation
Policy sensitivity	88/786 items; oracle gap $+3.69$ pts	Selection changes which items succeed at fixed frame and pixel budgets.
Confusion enrichment	26.5% vs. 8.6%; OR 3.83	Confused items are more sensitive to sampling policy.
Terminal-$64f$ recovery	20/69 recovered; 49/69 persistent	Alternative policies recover a subset of regressions.
Video-level proxies	0/8 pass BH-FDR; maximum $\lvert\rho\rvert = 0.078$	No corrected association is detected for the tested low-level features.

Table 7: Reference policies and label-informed diagnostics on Qwen2.5-VL-7B, V1 short. Costs are mean requested frames: terminal counts the selected budget; total sums all visited stages. The selector and item oracle use different targets and configuration sets; their costs are not reported.

Policy	Accuracy	Terminal	Total	Label-free gate
Fixed $128f$	76.5%	128.0	128.0	✓
Margin cascade	76.5%	87.4	118.9	✓
Ground-truth-score cascade	78.8%	58.0	64.0	✗
Full-trajectory selector	77.9%	–	–	✗
11-config item oracle	87.9%	–	–	✗

Probe-inclusive accounting reverses the ranking of two accuracy-matched designs. The two-stage $16f \to 128f$ cascade uses 96.2 terminal frames but only 107.7 total frames, less than the three-stage design. Thus the smaller terminal budget requires more total requested frames when probes are included (Appendix D.3).

Measured stage times reveal a further distinction between frame savings and latency. On a separate 100-item stratified short-video subset, fixed $128f$ takes 2.308 seconds per item. Replaying the $16f \to 32f \to 128f$ cascade at margin threshold 0.91 by summing the measured times of all visited stages gives 2.624 seconds, or 13.7% more, despite using 118.1 rather than 128 requested frames. Both achieve 74.0% accuracy on this subset. This no-reuse estimate is an offline replay of separately timed stages, not an end-to-end cascade measurement (Appendix D.3).

For the decision signal, prediction agreement provides a label-free reference across models. Without per-model tuning, an agreement cascade evaluates $16f$ and $32f$, returns the latter prediction if they agree, and otherwise evaluates $64f$. Across three additional V1-short models, it stops early on 82.2–87.3% of items and stays within 0.0–1.0 accuracy points of the matched-grid best fixed policy (Table 8).

Confidence, accuracy, and calibration need not improve together as the visual budget increases. On Qwen2.5 short videos, increasing from 128 to 256 frames raises mean confidence while accuracy falls; on medium videos, accuracy improves from 64 to 128 frames while ECE increases (Table 9). Thus neither increasing confidence nor improving aggregate accuracy implies better calibration at the next budget. Appendix D.2 gives results for the full grid and the score definition.

Even label-informed early features leave a gap to full-trajectory information. On the anchor split, early and late changes in the ground-truth option's score correlate $r = -0.16$ ($r^2 \approx 2.6\%$). Under item-level cross-validation, a diagnostic using text/$16f$/$32f$ features yields 72.9% answer accuracy, versus 77.9% with full trajectories. Both predict the budget tier containing the highest ground-truth-option score on an eight-configuration grid, then select one of three budgets. Their label-derived features include probe correctness (Appendix D.2). The comparison diagnoses an information gap in the tested features, rather than the performance of a deployable selector.

The paired records therefore serve two roles: they reveal complementary successes that a fixed policy misses, and they expose the costs and signal limitations of attempts to select among them. Our release makes these comparisons recomputable from over 120,000 records with input-protocol provenance, rather than treating confused-item labels as universal across models or protocols (Appendix E.1).

Table 8: Cross-model reference cascade using one untuned agreement rule. Gap is relative to each model's matched-grid best fixed accuracy. Terminal and total are mean requested frame counts; total includes both the $16f$ and $32f$ probes.

Model	Accuracy	Gap	Terminal	Total
Qwen3-VL-8B	77.1%	−1.0 pts	37.7	59.4
InternVL3-8B	71.8%	0.0 pts	36.1	56.1
InternVL3.5-8B	73.8%	−0.4 pts	36.4	56.7

Table 9: Budget-dependent calibration of Qwen2.5-VL-7B on V1 short ($n = 786$) and medium ($n = 639$). Confidence is not renormalized over answer options; ECE is reported in percentage points (lower is better).

Split	Frames	Accuracy	Mean confidence	ECE (pts)
Short	128	76.5%	80.3%	4.7
Short	256	75.8%	81.0%	5.2
Medium	64	64.2%	69.8%	6.8
Medium	128	66.0%	72.5%	7.1

7 Related Work

Video-MME, MLVU, MVBench, TempCompass, and EgoSchema evaluate video understanding (Fu et al., 2025; Zhou et al., 2025; Li et al., 2024; Liu et al., 2024b; Mangalam et al., 2023). Studies of sampling and context allocation show that more visual input need not improve aggregate performance (Chung et al., 2025; Wang et al., 2025a; Brkic et al., 2025; Zohar et al., 2025). High frame rates introduce temporal-coverage and compression trade-offs (Li et al., 2025). We instead track item-level rescues and harms within each frozen model to measure how much opposing movements cancel, comparing shared budgets while retaining protocol-specific labels.

Frame selection, clip routing, and token compression retain useful evidence under constrained budgets (Zhu et al., 2025b; Yu et al., 2025; Zhong et al., 2025; Hu et al., 2025; Chen et al., 2024). Chain-of-thought prompting and self-consistency use intermediate reasoning and multiple sampled paths, respectively (Wei et al., 2022; Wang et al., 2023). Adaptive test-time scaling and cascades allocate computation using intermediate signals (Snell et al., 2025; Yue et al., 2024a; Wang et al., 2024; Narasimhan et al., 2025; Bajpai & Hanawal, 2025; Yue et al., 2024b; Yuan et al., 2025). We characterize the outcomes these policies seek to exploit: oracle complementarity, early versus full-trajectory predictability, and probe-inclusive accuracy–budget trade-offs.

Dataset cartography and example-difficulty analyses characterize training dynamics or the effective depth at which predictions stabilize (Swayamdipta et al., 2020; Baldock et al., 2021). Video-QA audits expose language priors (Feng et al., 2025); image benchmarks also screen for text-answerable questions (Liu et al., 2024a). We vary visual input at fixed model and item, tracking whether earlier successes survive; cross-model and raw/cache comparisons test how consistently the same items exhibit these transitions.

8 Conclusion

Visual-budget scaling changes which items a Video LLM answers correctly, not just how many. Paired trajectories expose opposing correctness changes in MCQA and reference-overlap changes in generation. Across the evaluated models and input protocols, recurring population-level complementarity coexists with variation in the affected items. Reference cascades further show that including probe costs can change policy rankings at the same accuracy. Together, these findings motivate reporting paired transitions, matched-grid comparisons, and probe-inclusive budgets alongside aggregate scaling curves. Our released records and code support this item-aware evaluation practice.

AI USE STATEMENT

Generative AI tools assisted with experiment planning, code implementation and debugging, statistical analysis, result interpretation, literature organization, and manuscript drafting and editing. Verification included inspecting execution artifacts, testing analysis code, and enforcing per-cell completeness checks. Primary generation results use deterministic reference-overlap metrics; a frozen local language-model judge is used only in secondary analyses (Appendix B.2). The authors are responsible for the final text, claims, code, and analyses, including all AI-assisted content.

ETHICS STATEMENT

This work evaluates frozen models on existing public benchmarks and releases derived model outputs and annotations rather than source media. We collected no new human-participant data and redistribute no videos, frames, audio, subtitles, question text, or model checkpoints. Users must obtain source assets from their original providers and follow the corresponding terms. The released tags describe model behavior and do not encode personal attributes of individuals appearing in videos. Appendix E.2 documents source governance, licensing boundaries, intended uses, and limitations.

REPRODUCIBILITY STATEMENT

The supplementary artifact contains per-item records, derived labels, protocol provenance, analysis scripts, and cached/raw inference adapters. Appendix E.1 specifies the schema and release contract; Appendix A.1 records inference settings; and Appendix C.1 reports protocol-replication checks. The main MCQA matrix and cross-model curves use complete per-item outputs with exact dataset counts and zero missing, duplicate, or execution-error records.

REFERENCES

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7558–7567, 2019.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, 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, 2025b. URL https://arxiv.org/abs/2502.13923.

Divya Jyoti Bajpai and Manjesh Kumar Hanawal. FREE: Fast and robust vision language models with early exits. In *Findings of the Association for Computational Linguistics (ACL)*, 2025.

Robert J. N. Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth $1/2$ tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Hyungjin Chung, Hyelin Nam, Jiyeon Kim, Hyojun Go, Byeongjun Park, Junho Kim, Joonseok Lee, Seongsu Ha, and Byung-Hoon Kim. Video parallel scaling: Aggregating diverse frame subsets for video LLMs. *arXiv preprint arXiv:2509.08016*, 2025.

Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video LLM benchmarks: Knowledge, spatial perception, or true temporal understanding? *arXiv preprint arXiv:2505.14321*, 2025.

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

Chaoyou Fu et al. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

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.

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

Yixuan Li, Changli Tang, Jimin Zhuang, Yudong Yang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Improving LLM video understanding with 16 frames per second. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024a.

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

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

Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Seungyeon Kim, Neha Gupta, Aditya Krishna Menon, and Sanjiv Kumar. Faster cascades via speculative decoding. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://arxiv.org/abs/2405.19261`.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Congchao Wang, Sean Augenstein, Keith Rush, Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Aditya Krishna Menon, and Alec Go. Cascade-aware training of language models. *arXiv preprint arXiv:2406.00060*, 2024.

Shaoguang Wang, Weiyu Guo, Ziyang Chen, Yijie Xu, Xuming Hu, and Hui Xiong. Less is more: Token-efficient video-qa via adaptive frame-pruning and semantic graph integration. *arXiv preprint arXiv:2508.03337*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL `https://arxiv.org/abs/2508.18265`.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Bowen Yuan, Sisi You, and Bing-Kun Bao. DToMA: Training-free dynamic token manipulation for long video understanding. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *International Conference on Learning Representations (ICLR)*, 2024a.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. DeeR-VLA: Dynamic inference of multimodal large language models for efficient robot execution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model. *LLaVA Blog*, 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. AIM: Adaptive inference of multi-modal LLMs via token merging and pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2412.03248.

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)*, pp. 13691–13701, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. *arXiv preprint arXiv:2510.27280*, 2025b.

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

APPENDIX CONTENTS

Supplementary material organized by topic.

A Experimental Details **13**

B Extended Evaluation Results **15**

C Robustness and Failure Analysis **17**

D Adaptive Inference **21**

E Artifact and Responsible Use **24**

Select a topic to jump to its section.

A EXPERIMENTAL DETAILS

A.1 MODELS, INPUTS, AND COVERAGE

MCQA experiments use BF16 precision and Flash Attention 2 where supported. Qwen and LLaVA base MCQA adapters select the option with the highest retained next-token score: for each letter, the score is the maximum vocabulary-softmax probability among its single-token spelling variants. The scores are not renormalized over answer options. InternVL base MCQA uses chat generation with `do_sample=False` and a 64-token cap, followed by answer-letter parsing. Responses without an extracted answer letter are scored incorrect and retained in the denominator. Open-ended generation uses greedy decoding with up to 256 new tokens. Visual inputs follow the recorded per-cell protocol: either uniform decoding from raw video or uniform subsampling from the 1-fps JPEG cache; the sampling intervention additionally uses random-seed-42 and dense-start policies. For Qwen2.5-VL-7B and Qwen3-VL-8B, spatial resolution is controlled via the `max_pixels` parameter. For InternVL3-8B and InternVL3.5-8B, standard cells use `input_size=448` and one square tile per video frame (`max_num=1`); reduced-input fallbacks are separate cells. For LLaVA-NeXT-Video-7B-hf, we use the HF-converted checkpoint (`llava-hf/LLaVA-NeXT-Video-7B-hf`) and the `LlavaNextVideoProcessor` with its default resolution, evaluating the context-safe text/$8f$/$16f$/$32f$ grid.

"V2 medium" denotes our evaluated 2,516-item, 629-video medium-duration subset, not an official full-benchmark split. The construction code defines medium as $120 \leq$ duration (s) ≤ 900 and retains locally available videos with a usable frame cache. The retained input snapshot spans 180.652–891.322 seconds; 2,466 questions have eight options and 50 have two to seven. Its exact item IDs are retained with the released labels. Results are per-question accuracies, not the benchmark's group-level score. The provenance audit did not recover the upstream revision or complete exclusion ledger; completeness checks apply to this fixed evaluated subset.

Table 10 summarizes the matched grid used in Table 2. Here and throughout, cross-model "matched" means frame-budget-matched: encoder behavior, FLOPs, and raw/cache execution are not assumed to be identical across models. The accepted run summaries record the expected dataset count and zero missing, duplicate, and error rows for every cell. Higher-budget and fallback configurations are reported separately from matched-grid comparisons.

Qwen2.5 and Qwen3 also have complete, error-free $128f$ cells and selected $256f$ cells. InternVL $128f$ results use a `input_size=224` fallback and are not substituted for the standard-448 matched cells. LLaVA uses its context-safe text-only/$8f$/$16f$/$32f$ grid and provides a third-family replication outside Table 2.

Table 10: Frame-budget-matched coverage for the four main models. All 16 model–split rows contain text/$16f$/$32f$/$64f$.

Model	Split	T	16	32	64
Qwen2.5-VL-7B	V1 short	✓	✓	✓	✓
	V1 medium	✓	✓	✓	✓
	V2 medium	✓	✓	✓	✓
	MLVU	✓	✓	✓	✓
Qwen3-VL-8B	V1 short	✓	✓	✓	✓
	V1 medium	✓	✓	✓	✓
	V2 medium	✓	✓	✓	✓
	MLVU	✓	✓	✓	✓
InternVL3-8B	V1 short	✓	✓	✓	✓
	V1 medium	✓	✓	✓	✓
	V2 medium	✓	✓	✓	✓
	MLVU	✓	✓	✓	✓
InternVL3.5-8B	V1 short	✓	✓	✓	✓
	V1 medium	✓	✓	✓	✓
	V2 medium	✓	✓	✓	✓
	MLVU	✓	✓	✓	✓

All Qwen3 V1-short statistics use the complete per-item outputs summarized in Table 13. On V1 medium, V2 medium, and MLVU, stratified-subset replications check consistency of the full-grid outputs. Minimum per-configuration correctness agreement is 98.0%, 100%, and 100%, respectively.

InternVL3's standard-448 V1-short $64f$ cell has $561/786$ correct answers (71.37%). On V1 medium, its text/$16f$/$32f$/$64f$ correct counts are $270/370/375/396$ out of 639. Their union contains 486 correct items, giving 14.1 points of matched-grid headroom, while 128 items (20.0%) have a downward correctness transition. Every configuration covers the full split without execution errors.

Our InternVL runner sets `tokenizer.model_max_length` to 32,768. InternVL3.5 reports a 32K supervised fine-tuning sequence length (Wang et al., 2025b), while the released 8B model configuration specifies `max_position_embeddings`= 40,960.[1] These are distinct settings, not a single hard architectural frame limit. Under standard video inference with `max_num`=1 (one tile per frame), pixel-unshuffle produces exactly 256 visual tokens per frame: $(448/14)^2/4 = 256$ for a 448×448 input, ViT patch size 14, and downsample ratio 0.5. Thus $64f$ contributes 16,384 visual tokens, whereas $128f$ contributes 32,768 before question and frame-marker overhead. The latter exceeds the runner's configured 32K budget, but token accounting alone does not establish whether a particular failed attempt was caused by truncation, a placeholder mismatch, or GPU memory exhaustion.

The standard $128f$@448 MLVU attempt did not yield a complete, error-free cell. We exclude failed attempts and use the clean standard-input text-only/$16f$/$32f$/$64f$ grid for matched comparisons. This is an evaluated-grid choice, not a claim that the architecture cannot process more than 64 frames. The runner aligns prompt-side image placeholders with the encoder's tokens per frame when input size changes.

We report $128f$ at `input_size`=224 as a context-safe fallback. Each frame then contributes $(224/14)^2/4 = 64$ tokens, giving $128f \times 64 = 8{,}192$ visual tokens within the 32K budget. The MLVU $128f$@224 run yields 65.71% (1,426/2,170, 0 errors in the accepted run summary). This reduced-input result extends evaluation to 128 frames but does not substitute for the standard-input protocol. Its accuracy is below the InternVL3.5 $64f$@448 result (69.08%): among these evaluated settings, standard-input $64f$ performs better than reduced-input $128f$.

A.2 Correctness Labels and Trajectories

We order configurations by an input-budget proxy, not by measured FLOPs. At fixed resolution, the ordering follows frame count; at fixed frame count, it follows per-frame pixel count. For mixed

[1]Released configuration: `https://huggingface.co/OpenGVLab/InternVL3_5-8B/blob/main/config.json`.

comparisons, we use total pixel count (frames $\times$ resolution). An MCQA item is visually confused if it is correct at *any* lower-budget configuration and incorrect at *any* higher-budget configuration. Visual-confusion and marginal text-overwrite rates use all evaluated items as their denominator. A conditional text-overwrite rate instead uses only items answered correctly in the text-only configuration. These binary measures are distinct from the continuous score regressions used for generation.

Each item on the short split yields a binary correctness vector of length 6 (text-only, $16f$, $32f$, $64f$, $128f$, $256f$, with all visual configurations at 151K). For a consistent summary across grids, we use five mutually exclusive groups:

- **Always correct / Always wrong**: all entries are 1 / 0.
- **Monotonic increasing**: at least one $0\rightarrow1$ transition, no $1\rightarrow0$ transition.
- **Monotonic decreasing**: at least one $1\rightarrow0$ transition, no $0\rightarrow1$.
- **Non-monotonic**: at least one transition in each direction.

The medium split uses a 4-point vector (text-only, $16f$, $64f$, $128f$). These summaries apply the same rules but retain their respective grid sizes; they are not frame-budget-matched comparisons. Finer short-split labels distinguish the locations of peaks, dips, recovery, and oscillation within the non-monotonic group.

Table 11: Trajectory groups on V1 short and V1 medium for Qwen2.5-VL-7B, recomputed from per-item correctness. Entries are counts (percentages). The final row combines monotonic decreases and non-monotonic trajectories.

Pattern	**V1 short** (n=786)	**V1 medium** (n=639)
Always correct	209 (26.6%)	174 (27.2%)
Always wrong	116 (14.8%)	140 (21.9%)
Monotonic increasing	310 (39.4%)	210 (32.9%)
Monotonic decreasing	28 (3.6%)	28 (4.4%)
Non-monotonic	123 (15.6%)	87 (13.6%)
Any downward transition	151 (19.2%)	115 (18.0%)

Table 12: Per-task visual-confusion rate on Qwen2.5-VL-7B $\times$ V1 short across the six-point text/$16f$/$32f$/$64f$/$128f$/$256f$ grid at 151K (eight highest-rate task types among twelve). Rates use all items of each task type, not only text-correct items.

Task type	**Confusion rate (%)**	**Confused items**
Counting Problem	36.1	39
OCR Problems	24.5	12
Action Recognition	22.0	24
Object Reasoning	19.7	13
Object Recognition	19.5	29
Spatial Reasoning	15.4	4
Spatial Perception	14.8	4
Attribute Perception	14.0	15

B Extended Evaluation Results

B.1 Cross-Model Multiple-Choice Evaluation

Table 13 reports accuracy at every evaluated configuration for Qwen3-VL-8B and InternVL3.5-8B on Video-MME short.

The curves show three patterns:

- Qwen3-VL-8B has identical accuracy and predictions at $128f$ and $256f$ ($626/786 = 79.64\%$), plateauing in the short-split cached 1-FPS regime. Qwen2.5-VL-7B instead loses five net correct answers at the final transition of its raw-video grid.

Table 13: Per-configuration accuracy on Video-MME short for Qwen3-VL-8B (6 configurations, up to $256f$) and InternVL3.5-8B (4 configurations, up to $64f$). Oracle headroom and visual-confusion rates use each model's full evaluated grid, not a common grid.

Config	Qwen3-VL-8B (%)	InternVL3.5-8B (%)
text-only	43.9	42.5
$16f$	69.6	70.6
$32f$	74.8	73.9
$64f$	78.1	74.2
$128f$	**79.6**	—
$256f$	**79.6**	—
Item-level oracle	88.7	84.7
Oracle headroom (pts)	+9.0	+10.6
Confusion rate (%)	16.7	14.1

- InternVL3.5-8B's accuracy increases monotonically through $64f$, with diminishing gains within the evaluated range.
- Qwen3-VL-8B's confusion rate is 16.7% over its six-configuration full grid and 16.3% on the four-configuration frame-budget-matched grid. The matched-grid value is comparable to InternVL3.5-8B (14.1%) and the raw-video Qwen2.5-VL-7B anchor (15.9%). Its larger matched-grid headroom on V1 medium and MLVU is visible in Table 2.

On medium, Qwen3-VL-8B yields text-only 42.6%, $16f$ 47.3%, $64f$ 47.3%, $128f$ 54.6%, while InternVL3.5-8B yields text-only 40.1% and $64f$ 63.2%. The Qwen3-VL-8B medium curve shows $16f \approx 64f$ followed by a gain at $128f$, a pattern absent from its short-split curve.

We include LLaVA-NeXT-Video-7B-hf (Zhang et al., 2024) as an older-generation architectural reference (trained in 2024-04 on Vicuna-7B-v1.5). We evaluate $\{\text{text-only}, 8f, 16f, 32f\}$ on the short split at the recorded HF input setting: text-only 28.4%, $8f$ 42.4%, $16f$ 43.6%, $32f$ 45.9%; best-fixed 45.9%, item-level oracle 57.8%, oracle headroom $+11.8$ points, visual confusion rate 16.2%. Despite lower absolute accuracy than the four main models, LLaVA-NeXT-Video exhibits both oracle headroom and visual confusion. This qualitatively replicates the two phenomena in an older model family under its own context-limited grid.

A full 2,170-item MLVU run under the same $\{\text{text-only}, 8f, 16f, 32f\}$ grid yields text-only 36.59%, $8f$ 45.30%, $16f$ 47.14%, $32f$ 48.02%; best-fixed 48.02% ($32f$), item-level oracle 60.46%, oracle headroom $+\mathbf{12.4}$ **points**, and visual confusion rate 17.7%. LLaVA also has the lowest absolute MLVU accuracy among these models (Qwen2.5/Qwen3/InternVL3/InternVL3.5 reach 56–69% best-fixed over their respective full supported grids), while showing comparable oracle headroom and visual-confusion rates on a benchmark with a different task composition from Video-MME. Its evaluated MLVU grid ends at $32f$, below the $64f$/$128f$ budgets at which the Qwen and InternVL families peak, so we report it separately from the main frame-budget-matched matrix.

Over the shared short configurations (text-only, $16f$, $32f$, $64f$), Qwen2.5-VL-7B has 125 visually confused items and InternVL3-8B has 98. Their intersection contains 33 items and their union 190, giving a Jaccard overlap of $\mathbf{17.4}\%$. Our null preserves the two set sizes and randomizes item membership over the shared 786-item universe. The exact matched-size random-set expectation is 7.5% Jaccard with a 95% interval of $[4.2\%, 10.9\%]$. The observed overlap is $\mathbf{2.30}\times$ this expectation ($p = 2.2 \times 10^{-6}$, exact upper-tail probability).

Consistent with §3, the above-chance overlap suggests a shared benchmark or task component: some items are scaling-sensitive in both models. Nevertheless, $157/190 = 82.6\%$ of items in the union belong to only one model's confused set, indicating substantial model-specific membership.

B.2 Open-Ended Generation and Dialogue

We evaluate all 418 MLVU generation items: 201 sub-scene questions and 217 video summaries. Qwen2.5 and Qwen3 use text-only/$16f$/$64f$; LLaVA uses text-only/$8f$/$32f$. All models use greedy decoding with at most 256 new tokens. Table 14 reports score oracle gaps measured by deterministic Token-F1; ROUGE-L also yields a positive gap in every row. These continuous scores follow the

generation-score definitions in §2. A frozen Qwen2.5-VL-32B rubric judge provides a secondary analysis following exact-answer and direct-contradiction controls. Primary results use the deterministic metrics; the secondary rubric scores are reported separately from the official MLVU GPT-4 evaluation protocol.

Table 14: MLVU Token-F1 score oracle gaps by task group, in points on a 0–100 scale.

Model	**Task**	n	**Oracle gap**
Qwen2.5	sub-scene QA	201	+3.9
	summary	217	+1.6
Qwen3	sub-scene QA	201	+5.5
	summary	217	+1.8
LLaVA	sub-scene QA	201	+3.9
	summary	217	+2.3

The gaps use the same complete set of 418 outputs as Table 3. Machine-readable means, per-item scores, and confidence-interval summaries accompany the analysis scripts. Token-F1 uses lowercased alphanumeric word tokens, retaining repeated words, and computes multiset-overlap F1 against one reference. ROUGE-L uses the same tokens and longest-common-subsequence F1. Neither metric removes articles or stems words. Empty predictions receive zero. Scores are averaged equally over items.

The $\geq$ 30-minute sub-scene subset contains 40 questions (median 63.1 minutes; maximum 117.1), and the $\geq$ 100-reference-word summary subset contains 130 items (median 126; maximum 242 words). Table 15 reports Token-F1 score oracle gaps for these strata. Because each stratum also selects a task type, these analyses assess coverage rather than isolate the effects of video duration or reference-answer length.

Table 15: Long-video and longer-summary Token-F1 score oracle gaps with 95% paired-bootstrap CIs.

Model	$\geq$ 30 min (n=40)	$\geq$ 100 words (n=130)
Qwen2.5	$+3.1\,[1.8, 3.7]$	$+1.7\,[1.2, 2.3]$
Qwen3	$+4.4\,[2.7, 5.8]$	$+1.7\,[1.3, 2.1]$
LLaVA	$+3.7\,[2.0, 5.6]$	$+1.9\,[1.4, 2.4]$

At stricter $\geq$ 60-minute (n=22) and $\geq$ 150-word (n=34) thresholds, all six Token-F1 point estimates remain positive ($+1.1$ to $+3.9$ points), with all six paired-bootstrap lower bounds above zero. Each resample jointly selects item IDs across configurations and reselects the best fixed configuration within that stratum; the interval is the 2.5th–97.5th percentile of 5,000 resulting gaps.

We use all 1,787 ten-turn dialogues in AVSD@DSTC7 validation. Turns 1–9 use the same ground-truth history at every budget, and only turn 10 is generated. We omit audio and the optional questioner summary to isolate changes in current-turn visual evidence. All six inference cells and six secondary-judge cells contain 1,787 complete, error-free rows. Frame counts are requested caps: each model's $64f$ run uses 32–54 effective frames on eight items, comprising six shorter sources and two decoder fallbacks. Table 16 reports Token-F1; the ROUGE-L score oracle gaps are $+4.6/+3.7$ points for Qwen2.5/Qwen3.

At later-budget score-drop thresholds of 1/2/5 points, score-regression rates are 40.1/36.3/30.4% for Qwen2.5 and 32.3/29.2/23.5% for Qwen3. Regressions persist across thresholds, although fewer items qualify as the threshold increases.

C Robustness and Failure Analysis

C.1 Raw-Video and Cached-Frame Protocols

For cached-frame cells, we extract frames at 1 FPS with `ffmpeg`, store JPEGs at quality q=2, and uniformly subsample the requested budget. Raw cells decode directly from source video. The

Table 16: AVSD Token-F1 by question-type group. "Means" lists text-only/$16f$/$64f$; gaps are score oracle gaps. Yes/no questions are identified by an initial auxiliary verb, not by the polarity of a generated or reference answer.

Model	Group	n	Means	Best fixed	Oracle	Gap
Qwen2.5	all	1,787	29.2/30.2/30.2	30.2	35.0	+4.8
	yes/no	1,127	30.8/31.6/31.6	31.6	36.5	+4.9
	other	660	26.4/27.6/27.8	27.8	32.5	+4.7
Qwen3	all	1,787	31.6/32.7/33.5	33.5	37.3	+3.8
	yes/no	1,127	33.2/34.6/35.1	35.1	38.8	+3.7
	other	660	28.9/29.5/30.7	30.7	34.7	+4.0

Qwen2.5 anchor uses raw video; cached-frame cells carry a distinct protocol tag. Item labels from these two paths remain separate. Each requested budget is sampled uniformly and independently, so the selected frame sets need not be nested. Increasing the budget is therefore a configuration intervention, not necessarily the addition of frames to an otherwise unchanged evidence set.

The initial Qwen2.5-VL-7B single-cell check compared cached and raw execution at $64f$: raw accuracy is 72.65% versus cached 73.92%, a difference of 1.27 points. Correctness agreement is 91.86% and answer-prediction agreement is 89.69% over the same 786 items.

We extend this check to the full six-point replication below. Both full grids have $+8.78$ points of headroom, although oracle accuracy, visual-confusion rate, and item membership differ. Replication of aggregate findings therefore does not imply invariance of individual labels.

We compare all 786 items under both six-point protocols rather than extrapolate from a selected subset. Table 17 reports the agreement of the derived labels. High overwrite agreement partly reflects the many negative items; Cohen's κ makes that class imbalance explicit.

Table 17: Raw-vs-cache label agreement on all 786 Video-MME short items. Confusion and trajectory groups use the six-point grid; overwrite compares text-only with $128f$.

Per-item label	Agreement (%)	Cohen's κ
Final correctness @ $64f$	91.86	0.792
Visual-confusion tag (binary)	87.02	0.553
Text-overwrite tag (binary)	96.31	0.623
Five trajectory groups	81.42	0.743

The Qwen2.5 V1-short anchor uses raw-video decoding. Repeating inference under this protocol reproduces the prediction and correctness of every item at each of the five visual budgets. We also evaluate an independent cached-frame grid. Table 18 reports both full grids.

Table 18: Full six-point raw/cache replication. Accuracy values are percentages.

	Accuracy by configuration						Best	Oracle	Gap	Confusion
Protocol	T	16	32	64	128	256	fixed			
Raw	37.53	65.90	68.96	72.65	76.46	75.83	76.46	85.24	+8.78	19.21
Cache	37.40	64.50	70.74	73.92	74.81	74.81	74.81	83.59	+8.78	15.90

The difference in full-grid headroom is 0.00 points (95% CI $[-1.78, +1.91]$). Across the two protocols, exact trajectory vectors agree on 68.83% of items, visual-confusion labels on 87.02%, text-to-128f overwrite labels on 96.31%, and the five trajectory groups defined in Appendix A.2 on 81.42%. We retain protocol-specific labels: this comparison tests end-to-end replication, not the causal effect of caching alone. Grid size also matters. On the four-point frame-budget-matched grid, the cached protocol has 113 visually confused items, whereas the raw-video protocol has 125. These are distinct from the full six-point counts. On the raw-video anchor, the six-point frame grid has 151 confused items; the full 11-configuration frame–resolution grid has 227/786. These counts describe different sets of configuration comparisons.

C.2 Resolution and Sampling

The following analyses vary spatial resolution, temporal–spatial allocation, and sampling policy, complementing the frame-count comparison in §3.

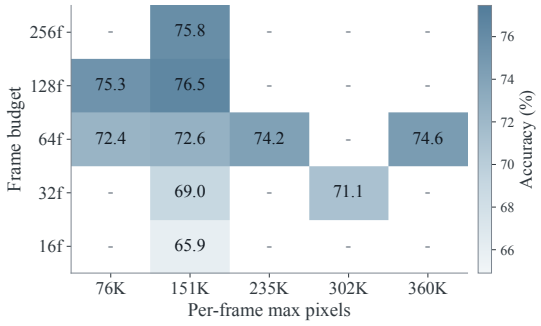

Figure 6: Aggregate accuracy on the partially evaluated frame×resolution grid (Qwen2.5-VL-7B on short). The highest aggregate accuracies occur at high frame counts ($128f$), while item-level preferences vary. Figure 5 shows the matched-total-pixel slice; equal pixel budgets do not imply equal FLOPs.

At a fixed 64 frames, we evaluate four resolutions on Qwen2.5-VL-7B × V1 short. While aggregate accuracy increases monotonically ($72.4\% \rightarrow 74.6\%$), the per-item resolution oracle reaches $\mathbf{80.0}\%$, a $+5.5$-point gap over the best fixed resolution. The configuration with the highest ground-truth-option probability varies across items: 194, 205, 226, 161 items select 76K, 151K, 235K, 360K, respectively. These counts are probability argmax assignments, not counts of unique binary-correct optima or lowest-cost correct configurations; ties follow configuration order. We observe *resolution confusion* paralleling frame confusion: 29 items are correct at 76K but wrong at 360K, and 23 at 151K but wrong at 360K.

We also evaluate Qwen2.5-VL-7B on V1 short with a fixed requested $64f$ cap across three pixel budgets, each in a complete 786-item run: 50K pixels yields 70.10%, 151K pixels yields 73.92%, and 200K pixels yields 74.43%. The 50K $\rightarrow$ 151K increase adds $+3.82$ points, whereas 151K $\rightarrow$ 200K adds $+0.51$ points, indicating diminishing returns above the $\sim$151K default in this sweep. The 151K result (73.92%) differs from the original 11-configuration grid's $64f \times 151K$ entry (72.6%; §2). The independent sweep uses cached frames, whereas the anchor grid uses raw-video decoding. Each within-grid comparison retains its own execution protocol. Actual cached frame counts are identical across the three pixel settings for each item; 234 of 786 items have fewer than 64 frames (overall range: 14–64).

We test sampling-policy sensitivity on Qwen2.5-VL-7B × V1 short at fixed $64f \times 151K$ pixels by replacing uniform sampling with two alternatives, each as a full 786-item run: *dense-at-start*, a prefix-biased policy, and *random* (seed 42). On these caches, dense-at-start selects the first 64 available frames, or all frames when the cache contains at most 64. All policies use the same requested cap and the same number of available frames per item. Aggregate accuracies are 73.92% for uniform, 70.48% for dense-at-start, and 74.05% for random sampling. Random and uniform sampling do not differ significantly ($+0.13$ pt), whereas dense-at-start loses 3.44 points while restricting temporal coverage. Per-item correctness sets agree strongly between random and uniform (Jaccard 0.94, 4.7% of items flip), and only modestly less between dense-at-start and uniform (Jaccard 0.88, 9.0% flip). Random sampling gives similar aggregate accuracy and substantial item-level agreement with uniform sampling. The dense-at-start result shows the sensitivity of this evaluation to concentrating frames near the beginning of a video.

We compare temporal–spatial allocations at matched total-pixel budgets (Figure 5), without assuming equal computational cost. At a fixed $\sim$9.7M total-pixel budget, $32f \times 302K$, $64f \times 151K$, and $128f \times 76K$ achieve aggregate accuracies of 71.1%, 72.6%, and 75.3%, respectively. The corresponding ground-truth-option probability argmax counts are 247, 254, and 285 items, with ties resolved by configuration order. These assignments differ from binary-correctness optima and from the cheapest-

correct labels in the artifact schema. The binary item-level oracle reaches $\mathbf{81.6}\%$, a $+6.2$-pt gap over the best fixed allocation, showing complementary correctness across temporal–spatial allocations.

Table 19: Resolution confusion at fixed 64 frames (Qwen2.5-VL-7B on short): number of items correct at row resolution but wrong at column resolution.

	76K	**151K**	**235K**	**360K**
76K	—	29	27	29
151K	—	—	26	23
235K	—	—	—	13

At a matched $\sim$9.7M total pixel budget on Qwen2.5-VL-7B short:

- $128f\times76K$ correct but $32f\times302K$ wrong: 71 items
- $32f\times302K$ correct but $128f\times76K$ wrong: 38 items
- $128f\times76K$ correct but $64f\times151K$ wrong: 51 items
- $64f\times151K$ correct but $128f\times76K$ wrong: 30 items
- $64f\times151K$ correct but $32f\times302K$ wrong: 52 items
- $32f\times302K$ correct but $64f\times151K$ wrong: 40 items

The nonzero counts in both directions show that no allocation dominates item-level correctness, even though their aggregate accuracies differ. These comparisons match total pixels, not FLOPs.

C.3 Failure Cases and Interpretation

At fixed $64f$/151K, uniform, random-seed-42, and dense-start sampling achieve 73.92%, 74.05%, and 70.48% accuracy, respectively. Correctness varies across policies for 88/786 items. The policy oracle reaches 77.74%, $+3.69$ points above the best fixed policy. Relative to non-confused items, visually confused items have 3.83 times the odds of being sampling-sensitive; alternative policies recover 20/69 terminal regressions. This enrichment analysis uses the cached, frame-budget-matched confusion labels, whereas the question-side model below uses the raw six-point frame-grid labels.

The structured analysis selects 12 distinct cases using prespecified behavioral signatures: four sampling-recovered, four sampling-persistent, two counting regressions, and two OCR regressions. Representative cases include (i) separated flag or toast occurrences that all policies undercount; (ii) a visible 5:59 clock that may compete with the queried later 6:00 event; and (iii) a rarely visible camera-holder amid dominant rugby/referee frames. The artifact includes case identifiers and the selection manifest, but not source frames or frame-contact sheets.

Across all 786 questions and 262 videos, none of eight duration, frame-count, and visual features is significant after BH-FDR correction. A six-feature question-side logistic analysis likewise finds no positive risk association that survives permutation-based BH-FDR correction after accounting for correlated counting and numeric-option indicators. These analyses provide no evidence for a simple explanation based on the tested features; they do not exclude semantic integration or attention-based mechanisms.

The evidence does not identify a single mechanism. Changing only the sampling policy alters correctness and recovers 29.0% of terminal regressions, implicating evidence composition for a subset of items. Persistent counting errors are consistent with limitations in temporal accumulation or occurrence tracking. OCR and dominant-scene cases suggest competition between a salient but incomplete cue and the queried event or referent. Attention dilution and distractor competition remain hypotheses, not causally measured mechanisms. The question-side analysis includes counting/OCR task indicators, temporal-language and negation indicators, an indicator that every option contains a numeral, and mean option length in words; no positive risk association survives its permutation-based multiple-testing correction.

Three practices follow from these findings. First, aggregate accuracy should be accompanied by per-item correctness traces to expose changes that cancel in the mean. Second, cross-model comparisons should use frame-budget-matched grids while retaining protocol provenance. On V1 short, raw-video Qwen2.5 has 19.2% confusion over its six-point frame grid and 15.9% over the shared four-point grid; neither rate describes its 11-configuration frame–resolution surface. Qwen3 changes from 16.7% over

six configurations to 16.3% over four. Grid differences otherwise confound scaling coverage with scaling behavior. Third, option-coverage diagnostics should accompany accuracy. Predicted-letter frequencies should be compared with the available options and ground-truth distribution before attributing an answer imbalance to a model's letter prior.

D Adaptive Inference

D.1 Cascade Policies

A single agreement-based cascade rule, without per-model tuning, gives similar qualitative results across models (Table 8). These reference cascades remain within 0.0–1.0 points of the corresponding matched-grid best-fixed accuracy. Charging both probes gives mean totals of 56.1–59.4 requested frames. They illustrate frame-budget–accuracy trade-offs across architectures rather than optimized adaptive-inference policies: early agreement allows some items to stop, while the rest receive additional frames.

Table 20: Operating points from the retrospective cascade sweep. For each margin design, we choose the smallest terminal cost among thresholds matching or exceeding fixed-$128f$ accuracy. Cost is terminal equivalent frames at fixed resolution; the mixed-resolution row instead charges both stages, with pixels normalized to $151{,}200$ per frame.

Cascade Design	Accuracy (%)	Cost	Stop Rate (%)	Signal
Label-free decisions (margin-gated)				
$16f \to 128f$	76.5	96.2	28.4	margin
$\mathbf{16f \to 32f \to 128f}$	**76.5**	**87.4**	**37.5**	margin
$32f \to 128f$	76.5	103.0	26.1	margin
$64f \to 128f$	76.5	99.6	44.4	margin
Ground-truth-score gates (diagnostics)				
$16f \to 128f$	78.8	58.0	62.5	p_{correct}
$32f_{302K} \to 128f_{76K}$	78.9	84.4	68.2	p_{correct}

Each design sweeps 101 equally spaced thresholds between its observed minimum and maximum signal. A stage stops when its signal strictly exceeds the threshold. The shared threshold for the three-stage margin cascade is 0.9099005; the corresponding two-stage thresholds are 0.9098862, 0.9698940, and 0.8799048 for $16f$, $32f$, and $64f$ probes. Selection and reporting use the same split. The ground-truth-score gate uses the known correct option's score, whereas the margin gate uses the top-two option-score difference.

The label-informed oracle gate protects 21 of the 151 visually confused items by stopping early, while incorrectly stopping 3 items, for a net benefit of $+18$ items. The label-free margin gate is more conservative: it matches static-$128f$ accuracy while reducing total requested frames by 7.1%, with a 37.5% early-stop rate.

We additionally evaluate cascades along the resolution dimension (*e.g.*, $64f{\times}76K \to 64f{\times}235K$) and mixed-dimension designs (*e.g.*, $32f{\times}302K \to 128f{\times}76K$). The mixed-resolution ground-truth-score gate reaches 78.9% at a total pixel-normalized equivalent-frame cost of 84.4. The frame-only $16f \to 32f \to 128f$ margin cascade has the smallest terminal-frame cost among the reported accuracy-matched margin designs. Charging all stages instead favors the two-stage $16f \to 128f$ design (Appendix D.3).

The cross-model cascade reference baselines are summarized in Table 8. An additional Qwen3 variant, *agree(*$16f$*,*$32f$*)*$\to 128f$, reaches 77.0% with an 82.2% stop rate. It uses 49.1 terminal and 70.8 total requested frames, versus 37.7 and 59.4 for the $\to 64f$ variant at 77.1%. The same untuned agreement rule is used for every model.

D.2 Calibration and Predictive Signals

On the short split, ECE improves relative to text-only but does not decrease monotonically with frame count. It is lowest at $128f$. On medium, ECE is lowest at $64f$, while $128f$ improves accuracy but has

Table 21: Calibration on Video-MME v1 short ($n = 786$) and medium ($n = 639$), Qwen2.5-VL-7B. Confidence is the maximum recorded option-token score. ECE uses ten equal-width bins on $[0, 1]$, weighted by bin frequency. All visual configurations use 151K pixels per frame.

Split	Config	Accuracy (%)	Avg. Confidence	ECE
short	text-only	37.5	0.549	0.174
short	$16f$	65.9	0.728	0.069
short	$32f$	69.0	0.759	0.070
short	$64f$	72.6	0.791	0.069
short	$128f$	76.5	0.803	**0.047**
short	$256f$	75.8	0.810	0.052
medium	text-only	38.5	0.568	0.183
medium	$16f$	52.3	0.651	0.128
medium	$64f$	64.2	0.698	**0.068**
medium	$128f$	66.0	0.725	0.071

slightly higher ECE. The runner takes the largest vocabulary-softmax probability among single-token letter variants for each option; it does not renormalize these scores over the four options. Among the 65 items that are confused on the six-point frame-count grid and incorrect at $128f$, mean confidence is 0.5473. Increasing confidence and improving average calibration therefore need not eliminate item-level regressions.

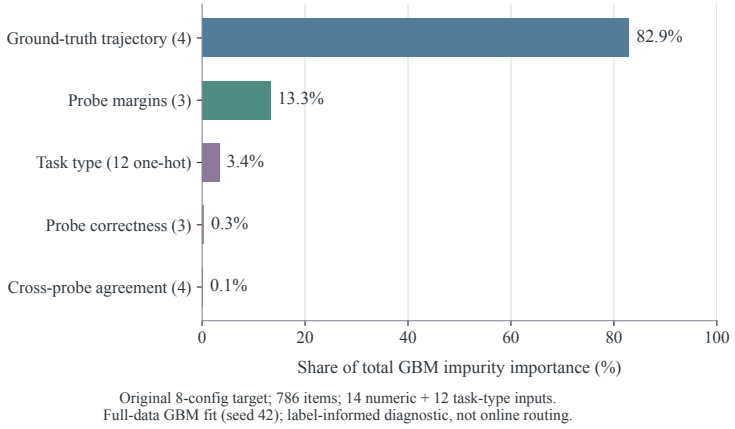

Figure 7: Normalized impurity importance from the label-informed GBM fitted to all 786 Qwen2.5 V1-short items. Four ground-truth-trajectory features contribute 82.9%. Groups are disjoint; importance is a fitted-model diagnostic, not a causal contribution.

The full selector uses 14 numeric features and twelve task-type indicators. Its target is the budget tier containing the maximum ground-truth-option score on the original eight configurations: text/$16f$/$32f$, three $64f$ resolutions (76K/151K/235K), $128f$, and $256f$. Predicted low/medium/high tiers select $32f$/$64f$/$128f$ at 151K. This target differs from the eleven-configuration binary item oracle. Importance by feature group is:

- **Ground-truth trajectory** (82.9%): score changes $64\rightarrow256f$ (35.6%) and $16\rightarrow64f$ (24.1%), regression severity (17.8%), and range (5.5%).
- **Probe margins** (13.3%): top-two option-score differences at text/$16f$/$32f$.
- **Task type** (3.4%): twelve one-hot indicators.
- **Probe correctness** (0.3%): text/$16f$/$32f$ correctness labels.
- **Cross-probe agreement** (0.1%): three pairwise indicators and all-probes agreement.

Removing trajectory features reduces accuracy from 77.9% to 72.3% (-5.6 points). In the original per-probe ablation groups, removing text, $16f$, $32f$, agreement, or task features individually changes accuracy by less than 0.4 points. The early selector uses 38 numeric features from text/$16f$/$32f$,

including score dynamics and correctness labels. Both selectors therefore use ground-truth information at evaluation time: they diagnose the information available in a trajectory rather than provide executable routing policies.

Five-fold stratified item cross-validation (seed 42) gives 77.86% for full features and 72.90% for early features. Holding out whole videos with stratified group folds gives 77.85% and 72.86%, respectively; both schemes happen to yield the same pooled counts of 612 and 573 correct answers out of 786. The feature set and GBM are held fixed for this grouping check. These are fold-mean accuracies, not confidence intervals. The early-to-late correlation compares $p_{32f} - p_{\text{text}}$ with $p_{256f} - p_{64f}$, where p is the correct option's score; $r = -0.1612$. The features called slopes in the implementation are score differences, not changes per frame. Selected-target budgets exclude feature acquisition and are not total inference costs.

D.3 Runtime Accounting

Let $b_1, \ldots, b_K$ be the requested frame budgets of a cascade and k_i its stopping stage on item i. We report two costs at fixed spatial resolution:

$$B_{\text{terminal}} = \frac{1}{n} \sum_i b_{k_i}, \qquad B_{\text{total}} = \frac{1}{n} \sum_i \sum_{j=1}^{k_i} b_j. \tag{3}$$

The first describes the selected budget; the second counts repeated processing when each visited stage runs independently. Independently sampled frame sets need not be nested, so terminal cost cannot be assumed to represent an implemented reuse mechanism. Table 22 derives both quantities from the stage decisions on all 786 items.

Table 22: Budget accounting for accuracy-matched margin cascades on all 786 Qwen2.5 V1-short items. Every row has 601 correct answers. All costs are requested frames, not measured latency, visual tokens, or FLOPs.

Policy	**Terminal**	**Total**	**Total vs.** $128f$
Fixed $128f$	128.0	128.0	0.0%
$16f \to 128f$	96.2	107.7	-15.9%
$32f \to 128f$	103.0	126.6	-1.1%
$64f \to 128f$	99.6	135.2	$+5.6\%$
$16f \to 32f \to 128f$	87.4	118.9	-7.1%

For the three-stage cascade, 223 items stop after $16f$, 72 after $32f$, and 491 after $128f$. Their total requested costs are 16, 48, and 176 frames, respectively. The two-stage $16f \to 128f$ cascade has the same 223 early stops, but avoids the intermediate probe for the remaining 563 items. Its larger terminal budget therefore coexists with a lower total cost. Conversely, the $64f \to 128f$ design exceeds the static budget once probes are charged. Actual latency also depends on decoding, preprocessing, visual tokens, attention kernels, and any verified reuse; these frame counts establish a budget trade-off, not a wall-clock speedup.

We separately time six static configurations on a 100-item Video-MME short subset. All six runs cover the same 100 items without missing, duplicate, or error records. The timed adapter performs a single forward pass and scores answer-option tokens. Wall time includes cached-JPEG loading amortized over questions sharing a video, preprocessing, forward execution, and option scoring; model initialization and offline cache construction are excluded. Each item is timed once, including the first item. Table 23 reports means from this recorded run.

The $16f \to 32f \to 128f$ cascade stops when the current option-score margin is at least 0.91, with 33, 4, and 63 items exiting at the respective stages. On this subset, both the cascade and fixed $128f$ answer the same 74 items correctly. Summing full visited-stage times gives 2.624 seconds per item, 13.7% above fixed $128f$. An incremental rule that charges the first stage plus positive consecutive timing differences gives 1.657 seconds, 28.2% below fixed $128f$; it assumes savings that would require an implemented reuse mechanism. The subset's mean terminal and total requested budgets are 87.2 and 118.1 frames, distinct from the 786-item results above.

Table 23: Static measurements and offline cascade replay on 100 aligned items. Cache denotes amortized cached-JPEG loading. Memory is the mean per-item forward-window peak allocated memory in GiB, not the maximum across items. Replay rows are estimates from independently timed stages, not end-to-end cascade measurements.

Configuration	Acc. (%)	Wall (s)	Cache (s)	Forward (s)	Memory (GiB)
Text-only	38	0.041	0.000	0.033	15.51
$16f$	66	0.542	0.088	0.417	16.03
$32f$	70	0.951	0.120	0.744	16.55
$64f$	73	1.795	0.204	1.426	17.45
$128f$	74	2.308	0.205	1.901	18.08
$256f$	74	2.273	0.195	1.874	18.08
Incremental replay	74	1.657	–	–	17.33
No-share replay	74	2.624	–	–	17.33

The $128f$ and $256f$ timing runs load the same number of frames on 99 of 100 items: their mean pre-processor frame counts are 78.55 and 78.58, respectively, because many caches contain fewer frames than the requested cap. Their similar timings therefore should not be interpreted as a general absence of extra cost when doubling the frame budget. The logs record loaded frames, not post-processor visual-token counts.

E Artifact and Responsible Use

E.1 Released Records and Code

The artifact combines per-item records, derived annotations, aggregate analyses, and evaluation code. A self-contained MCQA supplement freezes the records for Table 2, Figure 3, and the anchor frame–resolution grid.

For each released (model, benchmark split, configuration) triple, the core CSV contains one row per item with fields: `sample_id`, `model`, `split`, `config`, `frame_count`, `pixel_count`, `predicted_option`, `ground_truth_option`, `correct`, `source_tag`, `protocol_note`. Task and duration metadata are joined by sample identifier for the corresponding analyses; they are not mandatory columns in the core label CSV. The provenance fields and accompanying run manifests identify raw/cache execution, sampling policy, and input-size fallback. Numeric pixel/input-size metadata specify the actual spatial budget. Reduced-input InternVL configurations are identified separately: for example, `128f_input224` has `pixel_count`$= 50{,}176$, corresponding to 224^2 input pixels. These records support deterministic derivation of the MCQA annotations. The core release contains over 120,000 rows across the per-model coverage matrix (Appendix A.1).

Derived annotations are indexed by `grid_id`, model, split, and sample identifier. Each grid explicitly lists its configurations and `overwrite_target`; the matched four-point and full frame-count grids therefore retain distinct labels. The annotations include `is_visually_confused`, `is_text_overwritten`, `trajectory_class`, and `first_correct_config`. The latter is the lowest-frame correct configuration within an ordered frame grid, or empty if none is correct. For the eleven-configuration anchor surface, strict total-pixel ordering determines confusion; equal-budget configurations are incomparable. Its `lowest_budget_correct_configs` retains all minimum-budget ties, and an ordered trajectory class is inapplicable. Across the anchor grids, text overwrite always targets $128f$ at 151K, independently of the configurations used to calculate the oracle and confusion.

Per-grid JSON files report accuracy curves, oracle headroom, visual-confusion and overwrite counts, trajectory distributions for ordered grids, and adjacent-budget churn matrices. Task-wise analyses join benchmark metadata by sample identifier. The self-contained supplement reproduces the main matched-grid table, curve data, and anchor-surface metrics using Python's standard library; its scripts validate unique IDs, complete configuration coverage, and explicit grid definitions before computing the results.

The package contains raw-video and cached-frame inference adapters, a one-time 1-fps JPEG cache builder, completeness and merge checks, label-derivation and figure-data scripts, and the reference cascade implementation. Sampling policy and protocol provenance are explicit command-line arguments and output fields.

Per-item outputs use deterministic option scoring or greedy answer generation, with recorded precision and attention settings. The main matched-grid and cross-model curve cells have exact dataset counts with no missing, duplicate, or execution-error records. The raw-video frame-count anchor reproduces exactly under the same runner; independently evaluated cached-frame cells retain separate protocol records (Appendix C.1). Adding a new Video LLM requires an inference adapter and leaves existing records unchanged.

Code is released under the MIT License and newly created labels and annotations under CC-BY-4.0. These licenses do not override source-benchmark terms. We do *not* redistribute videos, extracted frames, audio, subtitles, question text, or checkpoints; users must obtain source assets from the original providers. A Croissant 1.1 record documents provenance, schema, licenses, intended uses, limitations, and sensitive-information boundaries.

The artifact supports reproduction of the reported empirical analyses, development of frame-budget adaptive-inference methods (cascades, routers, and early-exit predictors) with the per-item oracle as an upper bound, and benchmark-level studies of item-level scaling heterogeneity. The labels characterize the evaluated model–split–protocol combinations; they do not establish a universal aggregate decline with increasing frames. Model-agnostic difficult-item prediction requires renewed cross-model overlap checks, given the mixture of shared and model-specific structure. Cross-model comparisons of visual-confusion rates require matched grids.

Several limitations apply to the released labels. (i) The configuration grid is non-uniform across models and input protocols. (ii) Exact item labels are protocol-conditioned; raw and cache outputs must not be pooled. (iii) Alternative sampling policies are available only for the Qwen2.5 V1-short intervention. (iv) Confidence/margin fields are available only for cells used by the cascade analysis. (v) Continuous generation scores are metric-dependent and are not binary correctness labels.

E.2 Licensing and Broader Impact

Per-item records expose failures hidden by aggregate accuracy and support reproducible comparisons, adaptive-inference research, and benchmark audits of letter-prior bias and text overwrite. Deployment and energy savings require direct measurement.

Per-item correctness labels may enable fine-tuning or prompt tuning that overfits to the released Video-MME and MLVU items. Excluding source videos limits redistribution, but neither that restriction nor frame-budget matching prevents benchmark overfitting. The documented failure modes could also inform adversarial inputs that change Video-LLM answers, although this work does not develop such attacks. Finally, lower equivalent-frame costs do not eliminate the oracle gap or establish reliability: the reference cascades do not justify replacing more accurate models in high-stakes settings.

The study evaluates five open Video LLMs and provides over 120,000 core per-item records. Frame caches reduce redundant decoding; raw-video runs provide protocol comparisons. Appendix D.3 reports probe-inclusive frame costs and static timing measurements, without extrapolating them to project-wide energy savings.

We do not redistribute underlying benchmark media or question text. Code is released under MIT and newly created model outputs and annotations under CC-BY-4.0; benchmark-derived fields remain subject to their source terms. The artifact contains no newly collected human- participant data or person-descriptive attributes.

The artifact contains no surveillance-oriented fine-tuning recipes, facial-recognition or person-identification training data, or harmful-content classifiers. It characterizes scaling behavior rather than establishing readiness for downstream deployment.

Table 24 summarizes the source-provider terms checked for this release. The derived-artifact license covers only components created by this study; it does not relicense source material or supersede benchmark and media-owner rights.

Table 24: Source-provider terms and release boundaries. This summary refers to the linked provider pages; the current source terms govern use.

Source	Provider-stated terms	Handling in this artifact
Video-MME v1/v2	Academic research only; commercial use prohibited; video copyrights remain with owners; redistribution requires prior approval.	No videos, frames, subtitles, audio, or question text. Derived rows retain source provenance; v1/v2 labels are tagged separately.
MLVU	CC-BY-NC-SA-4.0 and research-only notice; benchmark authors state that they do not own raw-video copyrights and provide a removal channel.	No source media or question text; generation and MCQA records remain source-tagged.
AVSD / Charades	AVSD annotations and Charades videos are obtained through their official provider channels and remain governed by those providers' terms.	No source videos, audio, dialogue text, or reference answers are redistributed.

Provider documentation is linked for Video-MME v1, Video-MME v2, MLVU, and AVSD (Alamri et al., 2019). If a provider withdraws an item, we will remove or tombstone its derived identifier. Source providers remain responsible for original collection and media rights; we are responsible for minimizing exposure in the derived release. Its task and trajectory tags describe model behavior, not demographic or person-identifying attributes.