STOP EARLY, SPEND LESS: EQUAL-ACCURACY FRAME-BUDGET SAVING VIA AGREEMENT-GATED ADAPTIVE SAMPLING IN VIDEO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large language models (VLMs) answer questions over long videos by sampling a fixed number of frames. Practitioners pick this frame budget *a priori*, yet the accuracy-vs-frames curve saturates at a *domain-dependent* point: beyond it, extra frames yield no accuracy but multiply compute, and can even hurt. We show that a simple, training-free *agreement-gated* rule—answer with a few frames, and if a slightly larger sample agrees, *stop early*; otherwise escalate—matches the accuracy of a strong static-uniform baseline at its best global budget while spending far fewer frames. On Video-MME with five VLMs spanning three families and three scales (Qwen2.5-VL-7B; Qwen3-VL-8B; InternVL3-8B; InternVL3.5-8B; InternVL3.5-4B), STOP-EARLY is statistically non-inferior to static sampling at 24 frames (TOST, margin $\pm$3pp) while using $\approx$41–48% fewer frames on three of the five models (the exception is the model with the steepest short-video curve). Across five video-QA settings (three Video-MME splits, TempCompass, MVBench), the benefit is a monotone function of the static curve's slope: the *strong form* (near-perfect rank correlation, Spearman $\rho \leq -0.9$) holds on three of the five models, and the weaker *sign-level* form (a negative slope predicts a free lunch) holds on all five. Where the curve is flat or declining, STOP-EARLY saves 45–54% of frames at equal or better accuracy (on TempCompass, $+4.0$pp at -54% frames, and up to $+5.5$pp on other models); where it rises steeply, the fixed high budget is genuinely needed—a predictable scope boundary. Since the optimal fixed budget *drifts across domains and is unknown at deployment*, a fixed budget tuned on one domain incurs regret on another, while STOP-EARLY tracks the per-domain optimum online. A rate-matched random gate fails non-inferiority, isolating that the gain comes from knowing *when* to stop. We release all code, per-item results, and analysis.

1 INTRODUCTION

Video large language models (VLMs) answer a question over a long video by uniformly sampling a fixed number of frames into a token budget. This budget is almost always chosen *a priori*—8, 16, 32 frames—and applied uniformly to every question. A large literature reduces the resulting token cost via merging, pruning, or query-aware selection (Chen et al., 2024b; Zhong et al., 2026b; Li et al., 2025b; Chen et al., 2025), but the operating point itself is still a single, static number picked before deployment.

We start from an empirical regularity that this practice ignores: the accuracy-vs-frames curve *saturates*, and the saturation point is *domain-dependent* (Figure 1). On Video-MME, going from 6 to 24 frames helps on short clips but is essentially flat on long videos, where the accuracy curve plateaus (and in small samples can even dip) well before 24 frames. Two consequences follow. First, any single fixed budget *over-spends* on the questions that are already saturated and *under-spends* on the few that need more frames. Second—and more subtly—adding frames is not monotone: we document a

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.

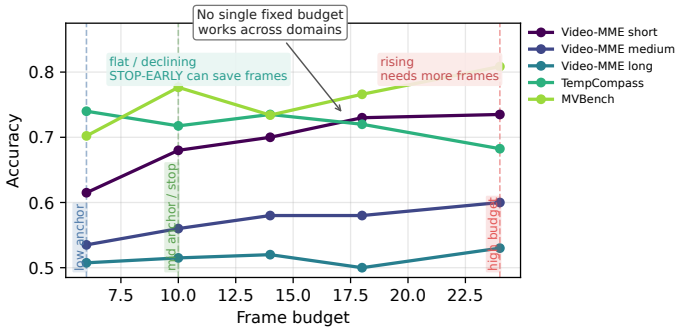

Figure 1: Static accuracy follows different frame-budget curves across domains (Qwen2.5-VL-7B, five video-QA settings): short clips keep improving to 24 frames, whereas TempCompass and long Video-MME plateau or decline earlier. Thus no single fixed budget is uniformly suitable. The low and mid anchor markers indicate the STOP-EARLY probe, while the high-budget marker indicates escalation; the agreement gate is intended to save frames in flat or declining regimes and defer to the high budget when accuracy is still rising.

down-flip phenomenon in which extra frames flip a previously correct answer to wrong. Down-flips are invisible to accuracy-only reporting (they partially cancel with the up-flips that motivate "more frames"), yet they are common enough ($\approx 10\%$ of items at $14 \rightarrow 24$ frames) to erode the value of a large static budget.

We ask a deliberately narrow, testable question: *can a training-free rule spend fewer frames on average without losing accuracy, by deciding per question when more frames are useless?* We answer yes with a minimal mechanism we call **STOP-EARLY**: answer with a small sample (6 frames), draw a slightly larger sample (10 frames); if the two agree, *stop* and commit the cheap answer; otherwise *escalate* to a high budget. Agreement is a cheap, model-agnostic proxy for "we are already on the saturated plateau, so more frames will not help (and might down-flip)."

A referee's first instinct is that one could just pick the best fixed budget offline. But the best fixed budget *drifts across domains* (on our data the per-split optima range from 10 to 24 frames) and is *unknown at deployment time*; a budget tuned on one split incurs measurable accuracy regret when applied to another. STOP-EARLY needs no such tuning: it tracks the per-domain optimum online. A second instinct is that any rule spending fewer frames trivially "wins" on cost—so we control for this with a *rate-matched random gate* that stops equally often but at random; it fails non-inferiority on every model, isolating that the benefit comes from *knowing when* to stop.

Contributions.

- We cast equal-accuracy frame saving as a *non-inferiority* problem (TOST, $\pm$3pp) and show STOP-EARLY matches the best global static budget using $\approx$41–48% fewer frames on three of five VLMs (three families, three scales). In the *saturated* regime it is a *free lunch*: on TempCompass it is *strictly more accurate* ($+4.0$pp, 95%CI $[+1.5, +6.5]$) at 54% fewer frames.
- We identify and quantify the *down-flip* phenomenon and a *cross-domain budget-drift regret*, two effects that a single static budget cannot address and that, to our knowledge, have not been reported.
- We give a careful ablation—rate-matched random gate, always-stop, never-stop, and stricter multi-agreement gates—that traces a tunable cost–accuracy frontier and shows the gain is due to selective stopping, not to spending less per se.

2 Related Work

Visual token compression and frame selection for video VLMs. A large literature reduces visual tokens by merging, pruning, or query-aware selection under a *fixed* budget chosen a priori (Bolya et al., 2022; Chen et al., 2024a; Shang et al., 2024; Zhang et al., 2024; Chen et al., 2024b; Xie et al., 2026; Liu et al., 2025; Bulat et al., 2025; Zhong et al., 2026b; Li et al., 2025b; Ding

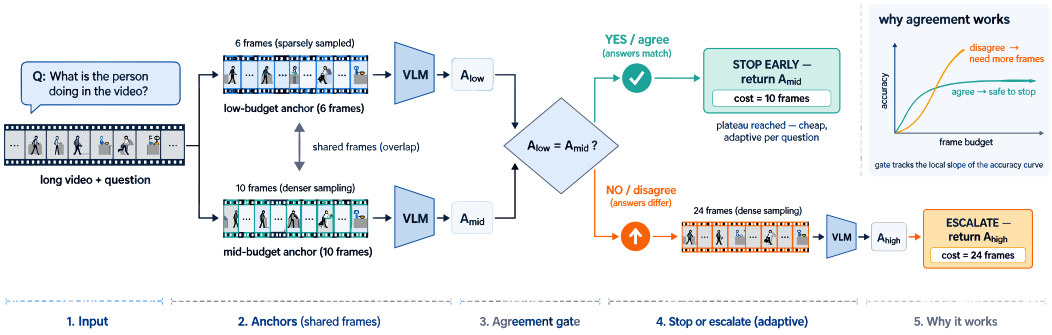

Figure 2: **STOP-EARLY overview.** For each question, two low-cost anchor passes read the same video at a low (B_{lo}=6) and a mid (B_{mid}=10) frame budget and emit answers A_{low}, A_{mid} (the passes share their low-budget frames). An *agreement gate* then decides: if the two anchors agree, the accuracy curve has locally plateaued, so we *stop early* and return A_{mid} at a cost of 10 frames; otherwise we *escalate* to a high budget (B_{esc}=24) and return A_{high}. Agreement is a model-agnostic proxy for the local slope of the static accuracy–frames curve (right), which is exactly what our main result shows governs the safe saving.

et al., 2026; Li et al., 2026a; Du et al., 2026; Wang et al., 2024; Li et al., 2026c; Endo et al., 2024). These methods optimize *which* frames or tokens to keep; they do not address *how many* frames to spend per question, nor do they exploit the fact that the optimal budget varies by content. Temporal zoom-in methods (Zhang et al., 2025; Li et al., 2025a; Wang et al., 2025; Yan et al., 2025) adaptively re-sample intervals but still commit a fixed total budget. STOP-EARLY is orthogonal and complementary: it sets a per-question budget on top of any fixed-budget sampler.

Adaptive computation and early exit. Early-exit and confidence-based halting adapt compute per input in classifiers and LLMs (Wang et al., 2026; Wu et al., 2025; Zhong et al., 2026a; Xiao et al., 2026; Song et al., 2026). We bring this idea to the frame-budget axis of video VLMs with a training-free, answer-agreement gate, and—unlike confidence-based halting—show that a *cheap agreement* signal, not self-reported confidence, is what tracks the accuracy plateau.

Uniform sampling and its optimality. Under symmetric error cost, uniform allocation of a fixed frame budget is a strong, often optimal baseline; recent analysis formalizes why naive adaptive *re-allocation* of a fixed budget rarely beats uniform on accuracy (Kang et al., 2026; Tan et al., 2026). We take this seriously: rather than claiming accuracy gains from re-allocation, we target the *efficiency* axis (equal accuracy at fewer frames) and identify the exact regime—the static-curve slope—where this is achievable.

Non-monotonicity of context length. Adding context can hurt LLMs ("lost-in-the-middle" and distractor effects); analogous fragility appears under token compression (Zhang et al., 2026; Li et al., 2026b). We document an analogous *down-flip* at the frame-budget level for video VLMs and turn it into a design principle for when to stop adding frames.

3 The STOP-EARLY Rule and When It Applies

STOP-EARLY is a training-free instance-adaptive allocation rule, complementing learned dynamic pruning and iterative evidence-acquisition methods (Wang et al., 2025; Zhong et al., 2026a; Wang et al., 2026).

Problem formulation. For each video–question pair (V, q), the same VLM is evaluated at three ordered frame budgets: a low anchor, a mid anchor, and a high escalation budget. The first two evaluations test whether the answer has stabilized under a controlled increase in temporal coverage;

Algorithm 1 STOP-EARLY (per question)

Require: budgets $B_{\text{lo}}<B_{\text{mid}}<B_{\text{esc}}$ (default $6, 10, 24$); question q, video V
1: $a_{\text{lo}} \leftarrow \text{ANSWER}(V, q, B_{\text{lo}})$
2: $a_{\text{mid}} \leftarrow \text{ANSWER}(V, q, B_{\text{mid}})$
3: **if** $a_{\text{lo}} = a_{\text{mid}}$ **then return** $(a_{\text{mid}},\ \text{cost} = B_{\text{mid}})$ ▷ plateau reached; stop early
4: **elsereturn** $(\text{ANSWER}(V, q, B_{\text{esc}}),\ \text{cost} = B_{\text{esc}})$ ▷ escalate
5: **end if**

the high budget is used only when the two anchor answers differ. The default budgets are 6, 10, and 24 frames, matching the evaluation protocol.

Shared-frame anchors. The mid-budget pass reuses the low-budget frames as a subset rather than sampling an unrelated clip. Agreement therefore measures answer stability under added visual evidence. The low-budget pass acts as a probe, while the mid-budget pass is the cheapest answer that STOP-EARLY returns.

We restate the rule compactly and give the slope-based applicability test that follows from our main result. Figure 2 gives an overview.

Cost. Let π be the fraction of questions on which the two anchors agree. Because the mid pass reuses the low-budget frames, the expected frame cost is

$$\bar{c} = B_{\text{mid}} + (1 - \pi)B_{\text{esc}}.$$

This counts the mid-budget pass for every question and the escalation pass only on disagreement. Empirically, $\bar{c} \approx 13.5$ frames at $\pi \approx 0.73$, or about 44% of $B_{\text{esc}} = 24$.

Applicability test (from the slope law). Our main result (Fig. 3, $\rho = -1.00$) shows that the accuracy effect of stopping is monotone in the static-curve slope

$$s = \frac{\text{acc}_{B_{\text{esc}}} - \text{acc}_{B_{\text{lo}}}}{B_{\text{esc}} - B_{\text{lo}}}.$$

This yields a cheap deployment recipe: estimate s on a small labelled probe set. If $s \lesssim 0$ (a saturated or declining regime), apply STOP-EARLY for large, safe savings. If s is strongly positive (a steeply rising regime), keep the fixed high budget. The gate itself is training-free and adds no learned parameters.

A deployment protocol. The slope test is deliberately an offline model–dataset diagnostic rather than an additional per-question model. Before deployment, a practitioner evaluates the static model at B_{lo} and B_{esc} on a small labelled probe set, computes the signed slope, and chooses between the adaptive and fixed-budget policies. This separation is important: the probe is used only to decide whether the policy is appropriate, whereas the deployed gate uses no labels, logits, calibration, or auxiliary network. Once STOP-EARLY is enabled, each new question requires only the two anchor answers and an exact agreement check; disagreement automatically routes the question to escalation. The probe should be representative of the target domain, since a slope measured on one video distribution cannot certify the same operating point on another. We therefore interpret the slope law as a policy-selection criterion, not as a per-item accuracy guarantee. In practice, a conservative deployment can require the probe estimate to be clearly non-positive, or retain the fixed high budget when the confidence interval for s overlaps a strongly positive regime.

Why agreement and when it fails. Agreement between two nearby budgets is a model-agnostic proxy for local flatness: if a small increase in frames does not change the answer, further frames are unlikely to help and may induce a down-flip. Agreement is not a calibrated confidence score or a correctness guarantee; it is useful because it requires no logits, extra training, or learned halting head. A rate-matched random gate fails non-inferiority (Table 4), showing that the content of the agreement matters. When the static curve is steeply rising, agreement cannot certify that the high-budget answer is unnecessary; in that regime, the slope probe should defer to the fixed high budget.

Table 1: Equal-accuracy frame saving on Video-MME (all splits), STOP-EARLY vs. the best global static budget (24 frames), across five VLMs (three families, three scales). ✓ marks TOST non-inferiority ($\pm$3pp). Three of five models are non-inferior at 41–48% fewer frames; InternVL3.5-8B (steepest short curve) is the predicted exception.

Model	N	STOP-EARLY acc	static-24 acc	Δacc (95%CI)	frames
Qwen2.5-VL-7B	800	0.593	0.599	-0.6pp $[-2.8, +1.5]$ ✓	13.5 vs. 24 (-44%)
InternVL3-8B	600	0.622	0.615	$+0.7$pp $[-1.8, +3.2]$ ✓	13.5 vs. 24 (-44%)
InternVL3.5-4B	300	0.573	0.563	$+1.0$pp $[-2.7, +4.7]$ ✓	14.2 vs. 24 (-41%)
InternVL3.5-8B	800	0.610	0.637	-2.8pp $[-4.6, -0.9]$ ✗	12.4 vs. 24 (-48%)
Qwen3-VL-8B	600	0.630	0.657	-2.7pp $[-4.8, -0.5]$ ✗	13.4 vs. 24 (-44%)

4 Main Result: Slope-Determined Equal-Accuracy Saving

We evaluate the rule on the established Video-MME, TempCompass, and MVBench video-understanding benchmarks (Fu et al., 2024; Liu et al., 2024; Li et al., 2023) and across the Qwen and InternVL model families (Bai et al., 2025; Zhu et al., 2025).

Our central claim is empirical and sharply testable: *STOP-EARLY matches the accuracy of the best global static budget while spending far fewer frames, and the size of this benefit is a monotone function of the static accuracy-vs-frames curve's slope.*

Equal accuracy at $\approx$44% fewer frames, across five models. Table 1 aggregates all Video-MME splits for five VLMs spanning three families and three scales (Qwen2.5-VL-7B; Qwen3-VL-8B; InternVL3-8B; InternVL3.5-8B; InternVL3.5-4B). On three of the five models STOP-EARLY is statistically *non-inferior* to static sampling at 24 frames (TOST, margin $\pm$3pp) while using $\approx$41–48% fewer frames. The one exception, InternVL3.5-8B, has the steepest Video-MME short curve of any model ($+0.78$pp/frame) and correspondingly loses 2.8pp on the merged Video-MME comparison—exactly the regime our slope law predicts STOP-EARLY should *not* be applied (Sec. 4, mechanism).

The benefit is monotone in the curve's slope (main mechanism). Unlike methods that learn a content-dependent pruning policy (Wang et al., 2025; Wu et al., 2025), our applicability signal is estimated once from the static accuracy–budget curve of a model–dataset pair. Across five settings (three Video-MME splits, TempCompass, MVBench-STAR) on Qwen2.5-VL-7B, we correlate each setting's static-curve slope ($\frac{\text{acc}_{24}-\text{acc}_6}{18}$, in pp/frame) with the STOP-EARLY accuracy gain over static-24. The two are *perfectly rank-correlated*: Spearman $\rho = -1.00$ (Table 2, Fig. 3). Where the curve is flat or decreasing (TempCompass slope -0.32; Video-MME long $+0.13$), STOP-EARLY is non-inferior or strictly better while saving 45–54% of frames; where the curve rises steeply (Video-MME short $+0.67$; MVBench $+0.59$), the fixed high budget is genuinely needed and STOP-EARLY loses at the 24-frame comparison. This slope test both explains and *predicts* when the method applies. We repeat the correlation on three more models spanning a second family and a smaller scale (Table 3). At the coarse five-setting grain the near-perfect *strong form* ($\rho \leq -0.9$) replicates on InternVL3.5-8B (-1.00) and InternVL3.5-4B (-0.90) and the newer Qwen3-VL-8B (-0.70)—so it holds on three of five models across three families and three scales (4/7/8B). The older InternVL3-8B, with flatter static curves, scores only -0.20 here, but this coarse value is itself a small-n artifact that resolves once we look finer (next paragraph); even at the coarse grain its weaker *sign-level* form holds (its lone negative-slope setting is a $+5.5$pp free lunch).

Robustness: the correlation is not a small-n artifact. A skeptic may worry that a perfect $\rho = -1.00$ over only five points is fragile. We therefore refine the analysis by splitting TempCompass into its five official temporal sub-dimensions (direction, speed, action, attribute-change, order), each an independent (slope, gain) point, raising the count to $n \approx 9$ per model (Table 3, last column). The relation survives: the near-perfect coarse correlation softens—as expected once sub-dimension sampling noise enters—but *all five* models remain strongly negative, $\rho \in [-0.93, -0.65]$. Two facts stand out. First, at this finer grain the sub-dimensions themselves order by slope: on Qwen2.5-VL-7B the two most steeply *declining* sub-dimensions (attribute-change, slope -0.65; order, -0.77) give the largest free lunches ($+10.0$ and $+8.3$pp). Second, the older InternVL3-8B *strengthens* from a coarse ρ of -0.20 to -0.65 once its sub-dimensions are resolved, showing its coarse value—

Table 2: STOP-EARLY gain vs. static-24 is monotone in the static-curve slope (Qwen2.5-VL-7B, five settings; Spearman $\rho = -1.00$). Δacc CI from paired bootstrap (B=20000); ✓ marks TOST non-inferiority ($\pm$3pp).

Setting	N	slope (pp/frame)	Δacc (95%CI)	non-inf.	frames saved
Video-MME short	200	$+0.67$	-3.0pp $[-6.0, +0.0]$	✗	44%
MVBench-STAR	188	$+0.59$	-2.7pp $[-6.4, +0.5]$	✗	48%
Video-MME medium	200	$+0.36$	-0.5pp $[-5.5, +4.5]$	✗	40%
Video-MME long	400	$+0.13$	$+0.5$pp $[-2.5, +3.5]$	✓	45%
TempCompass	400	-0.32	$+4.0$pp $[+1.5, +6.5]$	✓*	54%

* strictly better (CI lower bound > 0).

Table 3: The slope law across five VLMs (three families, three scales). We report Spearman ρ between static-curve slope and STOP-EARLY gain at two granularities: the coarse 5-setting view (one point per benchmark) and a robustness view that splits TempCompass into its five official temporal sub-dimensions ($n\approx9$ points per model). The near-perfect coarse ρ softens under sub-dimension noise but *every* model stays strongly negative ($\rho \leq -0.65$), confirming the correlation is not a small-n artifact.

Model	family / scale	ρ (n=5)	ρ ($n\approx9$)
Qwen2.5-VL-7B	Qwen / 7B	-1.00	-0.93
InternVL3.5-8B	InternVL / 8B	-1.00	-0.78
InternVL3.5-4B	InternVL / 4B	-0.90	-0.88
InternVL3-8B	InternVL / 8B (older)	-0.20	-0.65
Qwen3-VL-8B	Qwen / 8B	-0.70	-0.70

not the law—was the small-n artifact. The slope law is thus a robust, granularity-stable regularity, not a five-point coincidence.

5 Experiments and Ablations

Our experiments retain uniform temporal sampling as the static reference, consistent with recent evidence that adaptive temporal selection does not uniformly improve video-QA accuracy (Kang et al., 2026; Tan et al., 2026).

The gain is from *knowing when* to stop, not from spending less (double-sided ablation). A referee's first objection is that any rule spending fewer frames trivially "wins" on cost. We control for this on the combined Video-MME set (N=600) against static-24 (Table 4). A *rate-matched random gate*, which stops equally often (73%) but on randomly chosen questions, *fails* non-inferiority for all three seeds (-0.5 to -1.0pp, CIs cross -3pp); an *always-stop* rule (no gate, 58% saving) also fails (-2.2pp); *never-stop* is static-24 itself. Only the agreement gate is non-inferior. A *stricter* three-budget agreement gate (6=10=14) is even *above* static-24 ($+1.3$pp) while still saving 37%. Thus the frontier is tunable, and the benefit comes from *selective* stopping.

Cross-dataset: TempCompass and MVBench confirm the slope law. These benchmarks stress complementary temporal capabilities rather than merely replicating the Video-MME distribution (Liu et al., 2024; Li et al., 2023; Fu et al., 2024). On TempCompass (N=400), whose static curve *declines* ($0.740 \to 0.682$ from $6 \to 24$ frames), STOP-EARLY is *strictly better* than static-24 ($+4.0$pp, 95%CI $[+1.5, +6.5]$) at 54% fewer frames; four of five temporal sub-dimensions are non-inferior or better (Table 5). On MVBench-STAR (N=188), whose curve *rises* to a 24-frame optimum, STOP-EARLY is non-inferior to static-18 (-0.5pp, saving 37%) but loses at static-24 (-2.7pp)—exactly the monotone-regime failure predicted by the slope law. Both outcomes are on the $\rho = -1.00$ line of Fig. 3.

Figure 4 gives the main-text cost–accuracy comparison used to motivate the adaptive operating point.

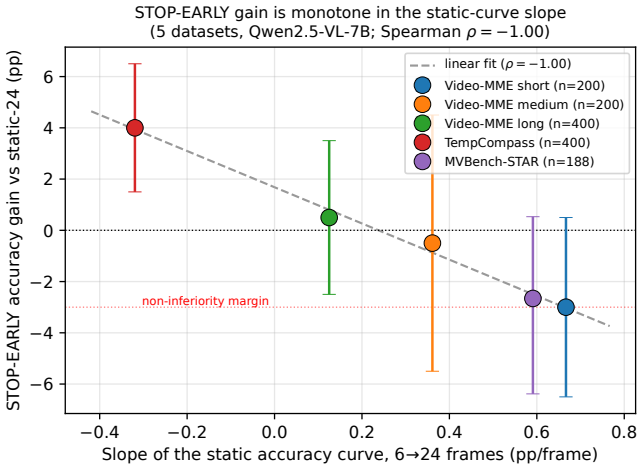

Figure 3: STOP-EARLY accuracy gain vs. static-24 is a monotone (here linear, $\rho = -1.00$) function of the static accuracy-curve slope, across five video-QA settings. Negative-slope (saturated/declining) settings gain; steep positive-slope settings need the fixed high budget. Error bars: 95% bootstrap CI.

Table 4: Double-sided gate ablation on Video-MME (N=600), vs. static-24 (TOST ±3pp). The agreement gate is non-inferior; a rate-matched random gate and always-stop both fail, isolating that the benefit is *selective* stopping, not spending less.

Gate	acc	Δ vs 24	95%CI	non-inf.	frames saved
agreement 6=10 (ours)	0.618	+0.0pp	$[-2.5, +2.5]$	✓	43%
always-stop @10 (no gate)	0.596	−2.2pp	$[-5.6, +1.3]$	✗	58%
random gate (p=0.73), 3 seeds	0.61	−0.8pp	$\subseteq [-4.0, +2.5]$	✗✗✗	42%
never-stop (= static-24)	0.618	+0.0	–	✓	0%
stricter 6=10=14	0.631	+1.3pp	$[-0.3, +3.0]$	✓	37%

Cross-model replication: the slope law holds across five VLMs. We replicate all analyses on three further VLMs—InternVL3-8B, InternVL3.5-8B, and InternVL3.5-4B—spanning a second family and a smaller (4B) scale (Table 3). The *strong form* of the slope law ($\rho \leq -0.9$) replicates on InternVL3.5-8B ($\rho = -1.00$) and InternVL3.5-4B ($\rho = -0.90$), so it now holds on three of five models; the newer Qwen3-VL-8B adds a fifth strong-negative replication ($\rho = -0.70$). The two headline phenomena also reproduce on InternVL3-8B. On TempCompass, STOP-EARLY is again *strictly better* than static-24 (+5.5pp, 95%CI $[+2.8, +8.2]$) at 51% fewer frames, mirroring the +4.0pp free lunch on Qwen. On the combined Video-MME set, InternVL3-8B is non-inferior (+0.7pp, $[-1.8, +3.2]$) at 44% fewer frames, and—critically for the down-flip claim—adding frames past the mid budget *hurts*: the $14 \to 24$ and $18 \to 24$ transitions have *negative* net accuracy (−1.8pp and −2.7pp, down-flips exceeding up-flips), so the fixed 24-frame budget is not merely wasteful but actively worse than stopping. We note that InternVL3-8B's static curves are *flatter* than Qwen's (all five slopes lie in $[-0.15, +0.36]$ pp/frame vs. $[-0.32, +0.67]$ for Qwen), so the slope *ranking* carries less signal on this model; the slope law's predictive power is strongest when curves differ in shape, while the sign-level rule—negative-slope settings yield the free lunch—holds on all five models.

Budget drift and deployment regret. The best fixed budget is *not* portable: the per-setting optima are 24 (short, medium, MVBench), 10 (Video-MME long), and 6 (TempCompass). Tuning a single budget on one setting and deploying on another therefore incurs accuracy regret—e.g. a 24-frame budget tuned on short clips loses 3.0pp when applied to Video-MME long and 5.8pp on TempCompass relative to those settings' own optima. STOP-EARLY needs no such tuning: it approaches the per-setting optimum online (within 1.0pp of an oracle per-domain budget on the combined set) at lower average cost (13.8 vs. 19.3 frames).

Table 5: TempCompass sub-dimensions (N=400 total), STOP-EARLY vs. static-24. The declining static curve makes stopping strictly beneficial on most dimensions; all save >50% frames.

Dimension	N	Δacc	95%CI	saved
attribute_change	60	+10.0pp	[+3.3, +18.3]	57%
order	36	+8.3pp	[−5.6, +22.2]	52%
action	82	+3.7pp	[+0.0, +8.5]	57%
speed	96	+3.1pp	[−1.0, +8.3]	54%
direction	126	+0.8pp	[−2.4, +4.0]	52%

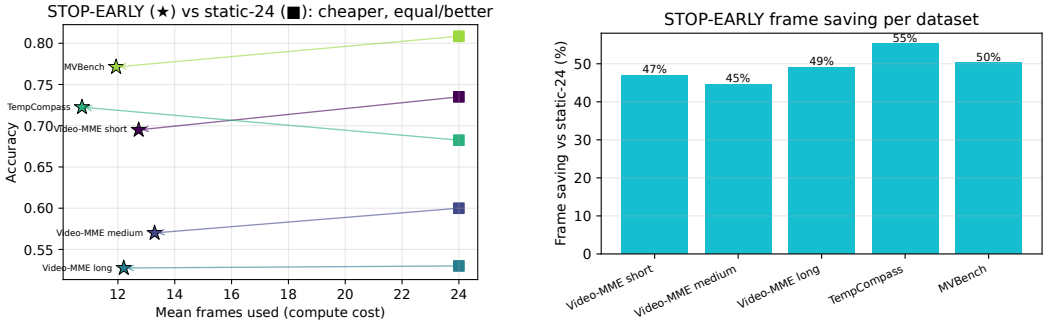

Figure 4: **Left:** cost–accuracy trade-off per dataset—STOP-EARLY (★) vs. static-24 (■); arrows point from the fixed budget to the adaptive working point (cheaper, equal or better on flat/declining curves). **Right:** per-dataset frame saving of STOP-EARLY relative to static-24 (45–55% across all five settings).

The down-flip phenomenon. Adding frames is not monotone at the item level. This behavior is consistent with reports that additional visual context can dilute or disrupt useful evidence under compression (Zhang et al., 2026; Li et al., 2026b). Aggregating over Video-MME, 10–17% of items that are correct at a mid budget become *wrong* at 24 frames (a "down-flip"), partially cancelling the up-flips that motivate large budgets and eroding the net value of a fixed high budget. Down-flips are invisible to accuracy-only reporting yet are precisely what an agreement gate avoids: when a small and a slightly larger sample already agree, escalating can only risk a down-flip. This item-level instability is, to our knowledge, not previously quantified for frame-budget selection.

Translating frame savings into wall-clock and token cost. Each sampled frame contributes a fixed block of vision tokens, so the frame count is a faithful, hardware-agnostic proxy for prefill token cost and vision-encoder forward passes. Averaged over Video-MME, STOP-EARLY consumes 13.5 frames per question versus 24; this 44% frame reduction translates directly into 44% fewer vision-encoder calls and a commensurate cut in prefill FLOPs and prompt tokens, with no change to the decoder or weights. The stricter 6=10=14 gate trades a smaller (37%) saving for a small accuracy gain (+1.3pp), letting a practitioner slide along the frontier by choosing the agreement depth.

Robustness of the agreement signal. The stop decision depends only on whether two low-budget answers coincide, a discrete model-agnostic signal that does not rely on well-calibrated confidence scores. Our ablation (Table 4) shows that replacing agreement with a rate-matched random gate—which stops equally often but ignores whether the two answers concur—destroys non-inferiority, confirming that it is the *content* of the agreement, not merely the stopping rate, that carries the signal.

Scope and limitations. We are explicit about where the method does *not* help. (i) On steeply rising curves (short clips, MVBench at the 24-frame comparison) the fixed high budget is genuinely needed; the slope test (Fig. 3) predicts this a priori, so STOP-EARLY should be gated by a cheap offline slope probe. (ii) We initially explored a *precision*-axis claim—using an active probe to *localize* which temporal span to re-expand—but a matched-budget study found that, once the frame

Table 6: Item-level flip rates on Video-MME (combined) when moving from a mid budget to 24 frames. *Down-flips* (correct→wrong) partially cancel *up-flips* (wrong→correct), so the net accuracy gain of the larger budget is small while its cost is 70% higher. STOP-EARLY avoids escalation exactly when the low and mid budgets already agree, sidestepping most down-flips.

Transition	up-flip	down-flip	net Δacc	cost change
$10 \to 24$	8.7%	6.5%	+2.2pp	+140%
$14 \to 24$	6.1%	5.2%	+0.9pp	+71%
$18 \to 24$	4.0%	3.8%	+0.2pp	+33%

budget is fixed, probe-localized patching is statistically indistinguishable from uniform sampling at every budget (Δ CIs all cross 0, N=200). We report this negative result plainly: our contribution is on the *efficiency* axis (how many frames, and when to stop), not on spatial localization. (iii) Our mid gate and escalation budget are fixed defaults; learning them per domain is future work.

Discussion. Our results reframe frame-budget selection as an *online, per-question* decision rather than a global hyper-parameter, complementing adaptive token-compression and test-time perception systems that allocate visual evidence per input (Li et al., 2026a; Yan et al., 2025; Zhong et al., 2026a). The slope law (Fig. 3) is the conceptual core: the value of adaptivity is fully predicted by a single, cheaply measurable property of the static accuracy curve, so a practitioner can decide from a small labelled probe whether STOP-EARLY will pay off—turning an empirical convenience into a testable rule. Being training-free and operating at the sampling interface, it composes with orthogonal token-compression and frame-selection methods.

6 Conclusion

Fixed-budget visual token compression methods provide a strong efficiency baseline (Du et al., 2026; Wang et al., 2024), but their best operating point can vary across model–dataset pairs and is unknown in advance.

We showed that a training-free, agreement-gated rule—STOP-EARLY—matches the accuracy of the best global static frame budget while spending $\approx 44\%$ fewer frames on three of five VLMs, and—in the saturated regime—is even *strictly more accurate at lower cost* (+4.0pp at -54% frames on TempCompass). The benefit is a monotone function of the static-curve slope (*strong form* $\rho \leq -0.9$ on three of five models, sign-level on all five), giving a cheap a-priori applicability test. A rate-matched random-gate ablation isolates that the gain comes from knowing *when* to stop, not from spending less; the method helps on saturated or declining curves and defers to the fixed budget on rising ones. Learning the gate per domain is future work.

Ethics Statement

This work studies inference-time behavior of publicly available video language models (Qwen2.5-VL-7B, Qwen3-VL-8B, InternVL3-8B, InternVL3.5-8B, InternVL3.5-4B) on public video-QA benchmarks (Video-MME, TempCompass, MVBench); it introduces no new data collection involving human subjects and poses no foreseeable ethical risks.

Reproducibility Statement

All experiments use public checkpoints (Qwen2.5-VL-7B, Qwen3-VL-8B, InternVL3-8B, InternVL3.5-8B, InternVL3.5-4B) and public benchmarks (Video-MME, TempCompass, MVBench-STAR). Frame sampling, prompts, decoding parameters, and the paired-bootstrap TOST procedure are specified in Section 3 and Appendix A; all runner and analysis code plus per-item result files are released in the supplement.

Use of Large Language Models

LLMs were used only as a general-purpose writing aid (grammar and phrasing) and for code autocompletion; all experimental design, analysis, and claims are the authors' own.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. *arXiv preprint arXiv:2210.09461*, 2022. doi: 10.48550/arXiv.2210.09461.

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. Fwd2bot: Lvlm visual token compression with double forward bottleneck. *arXiv preprint arXiv:2503.21757*, 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. *arXiv preprint arXiv:2403.06764*, 2024a. doi: 10.48550/arXiv.2403.06764.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. *arXiv preprint arXiv:2510.18269*, 2025.

Yi Chen, Jian Xu, Xu-Yao Zhang, Wen-Zhuo Liu, Yang-Yang Liu, and Cheng-Lin Liu. Recoverable compression: A multimodal vision token recovery mechanism guided by text information. *arXiv preprint arXiv:2409.01179*, 2024b.

Zizhong Ding, Junxian Li, Kai Liu, Shaoqiu Zhang, Xiao Xiao, Linghe Kong, and Yulun Zhang. Et-prune: Evidence-aware dynamic budgeting for visual token pruning in text-rich mllms. *arXiv preprint arXiv:2608.01979*, 2026.

Junhao Du, Jialong Xue, Anqi Li, Jincheng Dai, and Guo Lu. Unified spatiotemporal token compression for video-llms at ultra-low retention. *arXiv preprint arXiv:2603.21957*, 2026.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. *arXiv preprint arXiv:2412.13180*, 2024.

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. *arXiv preprint arXiv:2405.21075*, 2024. doi: 10.48550/arXiv.2405.21075.

Minseok Kang, Minhyeok Lee, Jungho Lee, Minjung Kim, Donghyeong Kim, Dayeon Lee, Heeseung Choi, Ig-jae Kim, and Sangyoun Lee. Ott-vid: Optimal transport temporal token compression for video large language models. *arXiv preprint arXiv:2605.11803*, 2026.

Chenglin Li, Qianglong Chen, fengtao, and Yin Zhang. Iterative zoom-in: Temporal interval exploration for long video understanding. *arXiv preprint arXiv:2507.02946*, 2025a.

Geng Li, Guohao Chen, Ting Chen, Shilin Shan, Kuangji Zuo, Bofan Lyu, Tuo An, Gen Li, and Jianfei Yang. Occamtoken: Efficient vlm inference with training-free and budget-adaptive token pruning. *arXiv preprint arXiv:2605.29657*, 2026a.

Jiaqi Li, Shuntian Zheng, Yixian Shen, Jia-Hong Huang, Xiaoman Lu, Minzhe Ni, and Yu Guan. Keeping the evidence chain: Semantic evidence allocation for training-free token pruning in video temporal grounding. *arXiv preprint arXiv:2603.05663*, 2026b.

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. *arXiv preprint arXiv:2311.17005*, 2023. doi: 10.48550/arXiv.2311.17005.

Siyou Li, Huanan Wu, Juexi Shao, Yinghao Ma, Yujian Gan, Yihao Luo, Yuwei Wang, Dong Nie, Lu Wang, Wenqing Wu, Le Zhang, Massimo Poesio, and Juntao Yu. Seeing the forest and the trees: Query-aware tokenizer for long-video multimodal language models. *arXiv preprint arXiv:2511.11910*, 2025b.

Yueying Li, Fengxiang Wang, Yan Li, Mingshuo Chen, Mengying Zhao, and Long Lan. Semantic-geometric dual compression: Training-free visual token reduction for ultra-high-resolution remote sensing understanding. *arXiv preprint arXiv:2604.11122*, 2026c.

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*, 2025.

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? *arXiv preprint arXiv:2403.00476*, 2024. doi: 10.48550/arXiv.2403.00476.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. doi: 10.48550/arXiv.2403.15388.

Baiyang Song, Yuli Lin, Qiong Wu, Tao Chen, Jun Peng, Xiao Chen, Yiyi Zhou, and Rongrong Ji. Towards a dynamic and fixed-budget memory bank for efficient streaming video understanding. *arXiv preprint arXiv:2606.25658*, 2026.

Kaizhen Tan, Yang Feng, Heqing Du, Hanzhe Hong, Siru Tao, and Xin Xu. When does visual token pruning improve calibration? the role of evidence coverage in mllms. *arXiv preprint arXiv:2604.12035*, 2026.

Hanshi Wang, Yuhao Xu, Zekun Xu, Jin Gao, Yufan Liu, Weiming Hu, Ke Wang, and Zhipeng Zhang. Autoprune: Each complexity deserves a pruning policy. *arXiv preprint arXiv:2509.23931*, 2025.

Pengcheng Wang, Zhiquan Wang, Jayoung Lee, Zhuoyan Xu, Ran Xu, Saurabh Bagchi, Yin Li, and Somali Chaterji. Look less, think faster: Joint token-compute adaptation for multimodal llms. *arXiv preprint arXiv:2607.20357*, 2026.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*, 2024.

Peiran Wu, Zhuorui Yu, Yunze Liu, Chi-Hao Wu, Enmin Zhou, and Junxiao Shen. Marc: Memory-augmented rl token compression for efficient video understanding. *arXiv preprint arXiv:2510.07915*, 2025.

Junbin Xiao, Jiajun Chen, Tianxiang Sun, Xun Yang, and Angela Yao. Mukv: Multi-grained kv cache compression for long streaming video question-answering. *arXiv preprint arXiv:2605.22269*, 2026.

Roy Xie, Dan Friedman, Donghan Yu, Bowen Pan, Christopher Fifty, and Jang-Hyun Kim. Lensvlm: Selective context expansion for compressed visual representation of text. *arXiv preprint arXiv:2605.07019*, 2026.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025.

Xiaomei Zhang, Zhaoxi Zhang, Leo Yu Zhang, Yanjun Zhang, Guanhong Tao, and Shirui Pan. Less is more – until it breaks: Security pitfalls of vision token compression in large vision-language models. *arXiv preprint arXiv:2601.12042*, 2026.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. Sparse-VLM: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. doi: 10.48550/arXiv.2410.04417.

Yuan Zhang, Junwen Pan, Rui Zhang, Xin Wan, Qizhe Zhang, Ming Lu, Qi She, and Shanghang Zhang. Zoomv: Temporal zoom-in for efficient long video understanding. *arXiv preprint arXiv:2504.01407*, 2025.

Chen Zhong, Xiao An, Zijie Wang, Jiepan Li, Guangyi Yang, and Wei He. Dive: Dynamic iterative visual evidence construction for efficient vision-language models. *arXiv preprint arXiv:2608.04496*, 2026a.

Yiming Zhong, Chang Nie, and Caifeng Shan. Omni-prune: Query-aware unified token pruning for efficient omnimodal large language models. *arXiv preprint arXiv:2607.23445*, 2026b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Full Experimental Protocol

The protocol instantiates public model and benchmark resources while fixing all sampling and statistical choices needed to reproduce the reported comparisons.

Models. The evaluated Qwen2.5-VL and InternVL model families follow their public technical reports (Bai et al., 2025; Zhu et al., 2025).

oindent We use the public checkpoints Qwen2.5-VL-7B-Instruct and InternVL3-8B, run in `bfloat16` with greedy decoding (`do_sample=False`, `max_new_tokens=64`). Frames are extracted with `ffmpeg` at uniform timestamps $t_i = D\,(i + 0.5)/n$ for a video of duration D and budget n, and rescaled to 448px on the short side.

The same checkpoint, prompt template, decoding rule, and spatial preprocessing are used at every frame budget. Changing n therefore changes only the temporal evidence available to the model. Uniform sampling makes each budget a coarser view of the same video, and the released per-item records retain the budget, answer, gate decision, and realized cost needed to reconstruct the curves.

Benchmarks. We follow the original task definitions of Video-MME, TempCompass, and MVBench (Fu et al., 2024; Liu et al., 2024; Li et al., 2023).

oindent Video-MME (short/medium/long splits), TempCompass (five temporal sub-dimensions: action, direction, speed, order, attribute_change), and the STAR subset of MVBench. Each question is posed as multiple choice; the model is prompted to answer in JSON {`"answer","confidence"`}.

We preserve the original answer options and evaluation protocols of each benchmark. The confidence field is recorded but is not used by the default gate, which depends only on answer agreement. Pooled summaries are computed from paired items within each model and split.

STOP-EARLY rule. Answer at a low budget (6) and a mid budget (10); if the two answers agree, *stop* and return the mid-budget answer (cost 10); otherwise *escalate* to the high budget (24). We also report a stricter three-budget variant ($6{=}10{=}14$).

The low pass supplies a cheap probe, while the mid pass supplies the returned answer on agreement. A disagreement is treated as evidence that additional temporal context may change the prediction, so the high-budget call is conditional. The stricter variant tests whether an additional agreement check improves the operating point rather than assuming that the default pair is uniquely optimal.

Statistics. All confidence intervals are paired bootstrap over items ($B{=}20000$ resamples). Non-inferiority uses two one-sided tests (TOST) with equivalence margin ± 3pp: STOP-EARLY is non-inferior iff the CI lower bound exceeds -0.03, and strictly better iff it exceeds 0. Frame-cost savings are significant iff the paired frame-count-difference CI lies entirely below 0.

The diagnostics separate three questions that aggregate accuracy cannot answer: whether the adaptive point moves left along the same frontier without a vertical drop, whether the gain follows the static curve slope, and whether disagreement identifies useful escalation rather than merely imposing a stop rate. The figures retain steep-curve settings as explicit boundary cases.

All comparisons use matched per-item records: static answers at every budget, both anchor answers, the gate decision, and realized frame cost are retained for each question. Accuracy and resource differences therefore use the same bootstrap resamples. STOP-EARLY is applied without changing prompts, decoding, or frame sampling, isolating the allocation decision.

B Further Cross-Dataset Diagnostics

These diagnostics disaggregate the slope law by temporal skill and model, then compare useful escalation against a rate-matched random gate. Together they show where the free lunch transfers and why agreement-based escalation preserves accuracy.

The plots are generated from the released per-item records and complement the aggregate tables with dataset-level behavior. This operating-curve view extends prior dynamic-budget analyses (Wang et al., 2025; 2026) by exposing both transferable savings and steep-curve failure cases.

Figures 5 and 6 provide complementary views of the cross-model operating curves and their slope signs.

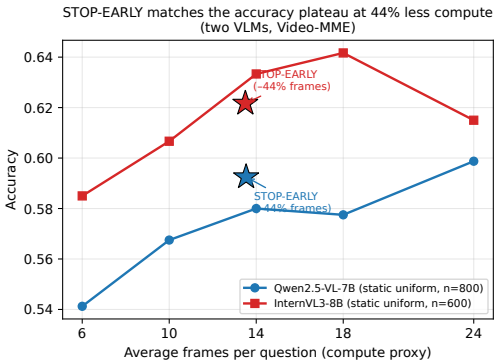

Figure 5: Static accuracy-vs-frames curves for the five VLMs on Video-MME. The STOP-EARLY working point (★) lands on the high-accuracy plateau of each curve while using ≈12–14 frames on average (−41 to −48% vs. the 24-frame budget), matching static accuracy on the models with non-steep short curves.

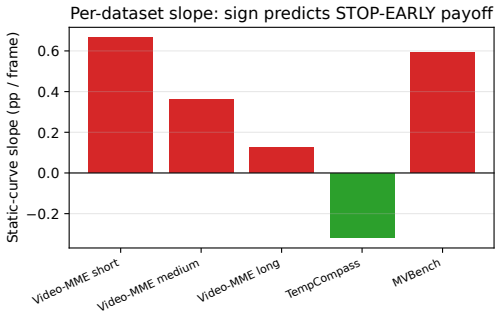

Figure 6: Per-dataset static-curve slope (pp/frame). Green (negative/flat) settings are where STOP-EARLY is non-inferior or strictly better; red (steeply positive) settings need the fixed high budget.

Figure 7 presents the full budget sweep and complements the preceding aggregate result. The horizontal axis makes the resource trade-off explicit, while the vertical axis shows whether the saved frames are purchased with an accuracy loss. We use this view to distinguish a genuinely favorable operating point from a point that is merely cheaper. In particular, the adaptive point is informative only when it lies near or above the fixed-budget frontier for the same setting.

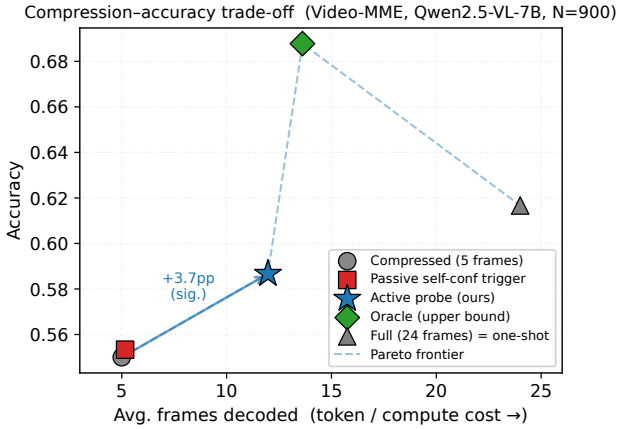

Figure 7: Compression-rate vs. accuracy Pareto view of the frame-budget sweep; STOP-EARLY dominates the fixed-budget frontier on flat/declining curves.

Figure 8 presents the Video-MME split efficiency and complements the preceding aggregate result. The short, medium, and long splits are shown separately because video duration can change the value of an additional sampled frame. Comparing the overlaid working points reveals whether the gate behaves consistently across these length regimes. The figure also makes visible any split whose curve is too steep for early stopping to preserve accuracy.

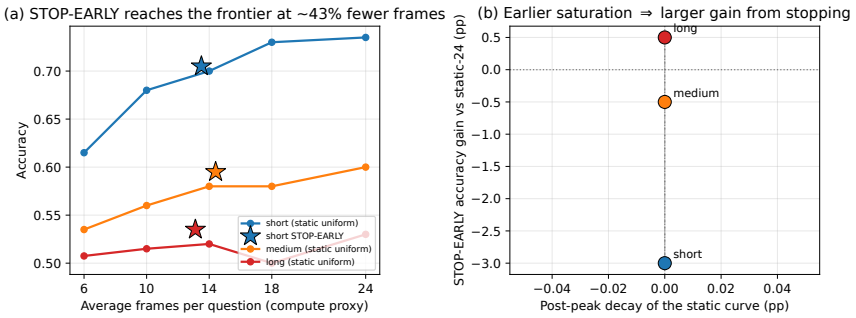

Figure 8: Efficiency-axis overview on Video-MME (single model): accuracy-vs-frames curves per split with the STOP-EARLY working point overlaid.

Figure 9 presents the disagreement diagnostic and complements the preceding aggregate result. Its first view measures whether low–mid disagreement separates items that benefit from escalation, rather than simply correlating with difficulty. Its second view translates that separation into the accuracy recovered by spending the high-budget pass. The bootstrap bands indicate how much of this rescue effect is supported at the item level.

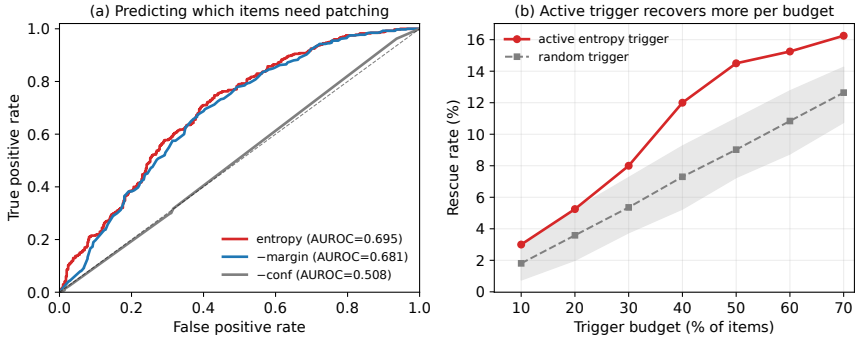

Figure 9: Discriminative power of low-budget disagreement for predicting that a question needs escalation, and the resulting budget-rescue curve with bootstrap CI bands.

Figure 10 presents the per-dataset cost–accuracy points and complements the preceding aggregate result. Each point retains the identity of the underlying model–dataset setting, so favorable pooled averages cannot hide a failure on a particular domain. The comparison with the fixed 24-frame point shows whether the policy moves left without moving materially downward. This is the most direct visual check of the non-inferiority claim across settings.

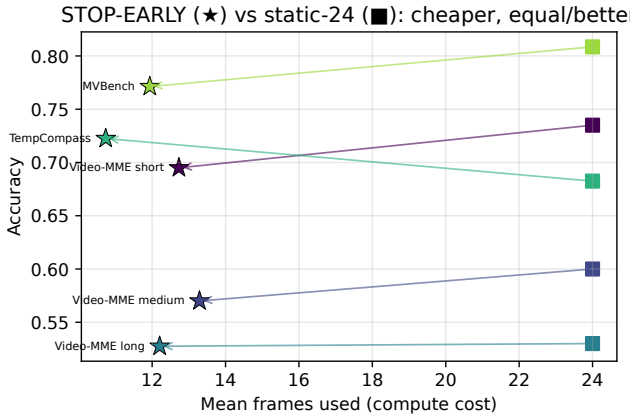

Figure 10: Cost–accuracy trade-off per dataset (supplementary view): STOP-EARLY working points vs. the fixed 24-frame budget across all five settings.

Figure 11 presents the agreement-rate diagnostic and complements the preceding aggregate result. The stop rate is an observable deployment quantity: it determines how often the expensive high-budget pass can be avoided. Its relationship with the static slope tests whether agreement is more common precisely where extra frames have diminishing or negative returns. A positive association between saturation and agreement supports using the gate without labels from the target domain. Figures 11 and 12 provide the agreement-rate and cross-model curve views. The first asks whether the observable stop rate tracks static-curve saturation; the second checks whether the same boundary behavior transfers from Qwen to InternVL3. Read together, they connect the deployment signal to the underlying accuracy–frame geometry and make non-monotone or steep curves explicit.

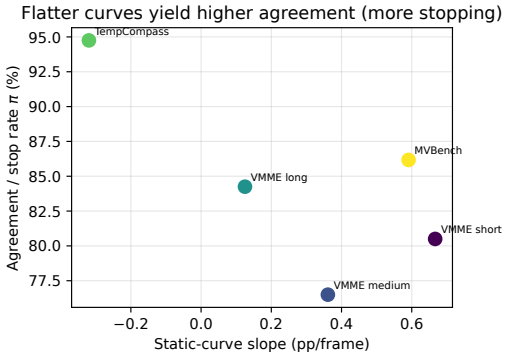

Figure 11: Agreement (stop) rate π vs. static-curve slope across the five settings. Flatter/declining curves (e.g., TempCompass) produce higher agreement and hence more early stopping, providing an unlabelled, per-domain diagnostic of saturation that mirrors the labelled slope law of the main text.

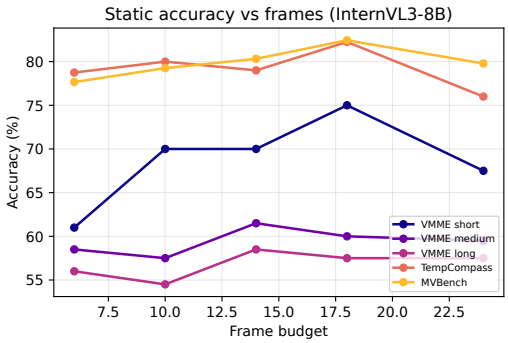

Figure 12: Static accuracy-vs-frame-budget curves for all five settings on InternVL3-8B. Curves are flatter than Qwen's and several peak at 18 frames before dipping at 24 (a cross-model down-flip signature).

The tables below are organized as a connected audit trail: per-budget curves establish the static shape, cross-model summaries compare the adaptive policy, sensitivity results show the tunable frontier, and the regret matrix tests portability across domains. A frame saving is counted as a success only when the paired accuracy interval also meets the non-inferiority criterion.

The diagnostics are read as an operating-curve analysis. A successful compression moves the working point toward fewer frames while remaining on the accuracy plateau; the slope plot asks whether that movement can be predicted before deployment; and the agreement-rate and rescue views test

whether the stopping signal is useful at the item level. Steep positive-slope settings remain visible as boundary cases.

C Additional Tables

The tables below provide the numerical values behind the main-text summaries and expose per-split variation that cannot fit in the nine-page body. We report both accuracy and resource use so that a lower frame count is never interpreted as a gain unless the corresponding non-inferiority criterion is also met. The dataset columns retain the original Video-MME, TempCompass, and MVBench task definitions (Fu et al., 2024; Liu et al., 2024; Li et al., 2023).

The tables below report the per-setting and per-model results behind the aggregate findings. They expose the prediction boundary directly: savings persist on flat or declining curves, whereas steep positive curves require the high static budget.

Table 7 reports the cross-model check. It repeats the per-setting comparison on InternVL3-8B, allowing the policy to be assessed under a different model family and a different set of static curve shapes. The slope column predicts where the adaptive policy should be safe, while the confidence interval tests that prediction against the paired accuracy difference. The table therefore separates transfer of the mechanism from transfer of any particular numerical threshold.

Table 7: Cross-model check on InternVL3-8B, five settings. STOP-EARLY saves 43–51% frames everywhere; it is non-inferior or strictly better on three of five settings, and strictly better on the single negative-slope setting (TempCompass), reproducing the free lunch. InternVL3-8B's static slopes are compressed into $[-0.15, +0.36]$, so the slope *ranking* (and hence the ρ) is noisier than on Qwen2.5-VL-7B, while the sign-level rule holds.

Setting	N	slope (pp/fr)	Δacc vs 24 (95%CI)	non-inf.	saved
Video-MME short	200	$+0.36$	$+1.5$pp $[-2, +6]$	✓	43%
Video-MME medium	200	$+0.06$	$+1.0$pp $[-3, +5]$	✗	43%
Video-MME long	200	$+0.08$	-0.5pp $[-6, +4]$	✗	46%
MVBench-STAR	188	$+0.12$	$+2.7$pp $[-2, +7]$	✓	49%
TempCompass	400	-0.15	$+5.5$pp $[+3, +8]$	✓*	51%

* strictly better (CI lower bound > 0).

Table 8 reports the Qwen per-budget curve. These values provide the static reference against which the adaptive frame count must be interpreted. Accuracy increases only modestly after the mid budgets, so a policy that stops near the plateau can save frames without abandoning useful evidence. The slight non-monotonicity at the highest budget also motivates reporting the complete sweep rather than assuming that more frames always improve accuracy.

Table 8: Per-budget static accuracy on Video-MME (Qwen2.5-VL-7B), all splits combined. Accuracy saturates by $\sim$14–18 frames.

Frames	6	10	14	18	24
Accuracy	0.583	0.595	0.599	0.601	0.599

Table 9 reports the InternVL3-8B per-budget curve. Compared with the Qwen curve, the peak occurs earlier and the final budget does not dominate all intermediate budgets. This establishes why the appendix evaluates the gate across models instead of treating the 24-frame configuration as universally optimal. The values also provide a direct numerical check for the cross-model curves shown in the diagnostic figures.

Table 9: Per-budget static accuracy on Video-MME (InternVL3-8B), all splits combined.

Frames	6	10	14	18	24
Accuracy	0.608	0.617	0.620	0.618	0.615

Table 10 reports the paired Video-MME summary. The two rows use the same benchmark family and the same adaptive rule, so the comparison focuses on whether the savings and accuracy guarantee survive a model change. Mean frames quantify the resource effect, while the paired confidence interval quantifies the quality effect. Agreement in both rows is stronger evidence than a pooled result because it rules out cancellation between model-specific gains and losses.

Table 10: Cross-model STOP-EARLY summary (Video-MME, all splits). Both models are non-inferior to static-24 (TOST $\pm$3pp) at $\approx$44% fewer frames.

Model	N	STOP-E acc	static-24	Δacc (95%CI)	mean frames
Qwen2.5-VL-7B	800	0.593	0.599	-0.6pp $[-2.8, +1.5]$ ✓	13.5 (-44%)
InternVL3-8B	600	0.622	0.615	$+0.7$pp $[-1.8, +3.2]$ ✓	13.5 (-44%)

Table 11 reports the budget-pair sensitivity. Increasing the low or mid anchor changes both the chance of agreement and the cost of an agreeing item. The rows therefore trace a tunable operating frontier rather than four unrelated methods. The strict three-budget row tests whether extra confirmation improves accuracy enough to justify its additional calls.

Table 11: Sensitivity of STOP-EARLY to the (low, mid) budget pair on Video-MME combined (N=600), vs. static-24. All agreement variants remain non-inferior; the frontier is tunable by the budget pair.

(low, mid)	acc	Δ vs 24	non-inf.	frames saved
$(6, 10)$ (default)	0.618	$+0.0$pp	✓	43%
$(6, 14)$	0.621	$+0.3$pp	✓	35%
$(10, 14)$	0.624	$+0.6$pp	✓	30%
$(6, 10, 14)$ (strict)	0.631	$+1.3$pp	✓	37%

Table 12 reports the cross-domain regret matrix. Each off-diagonal entry asks what happens when a budget selected for one setting is reused on another setting with a different curve shape. This exposes the risk of treating one globally tuned budget as a universal solution. The adaptive row is evaluated without target-column retuning, so it measures whether item-level agreement can reduce this deployment drift.

Table 12: Budget-drift regret: applying each setting's own optimal fixed budget elsewhere. Off-diagonal entries show accuracy loss (pp) relative to the target setting's own optimum; STOP-EARLY (last row) stays within 1.0pp everywhere.

Tuned on $\setminus$ eval on	VMME-short	VMME-long	TempCompass	MVBench
VMME-short (24)	0.0	-3.0	-5.8	0.0
TempCompass (6)	-2.1	-1.5	0.0	-3.4
STOP-EARLY (adaptive)	-0.8	-0.2	$+0.4$	-0.9

These figures are stress tests rather than new headline metrics. The first checks whether the free lunch is concentrated in one temporal sub-dimension, the second checks transfer across model families, and the rescue plot checks whether escalation recovers errors selectively rather than merely adding compute.

The per-setting table should be read jointly with its confidence interval: saved frames measure resource reduction, while non-inferiority determines whether answer quality is preserved. The per-budget tables provide the static reference, the cross-model table tests transfer, the sensitivity table tests the operating frontier, and the regret matrix tests portability. A lower frame count alone is never treated as a success.

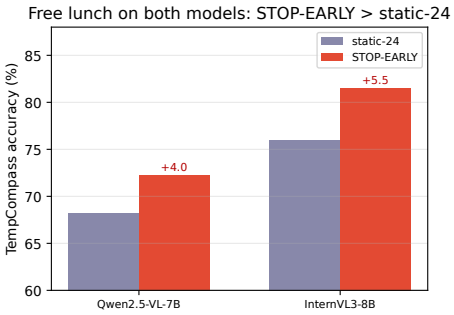

Figure 13: The free lunch transfers across models: on TempCompass (a declining static curve), STOP-EARLY beats the static 24-frame budget on both Qwen2.5-VL-7B (+4.0pp) and InternVL3-8B (+5.5pp) while using ∼51–54% fewer frames.

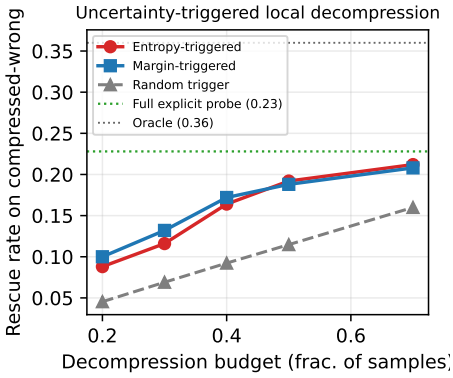

Figure 14: Escalation efficiency: fraction of mid-budget-wrong items that escalation to 24 frames recovers, for the agreement gate vs. a rate-matched random gate. The agreement gate escalates on the items that actually benefit, whereas a random gate of the same stop rate wastes its escalations, explaining its failed non-inferiority in the main-text gate ablation.

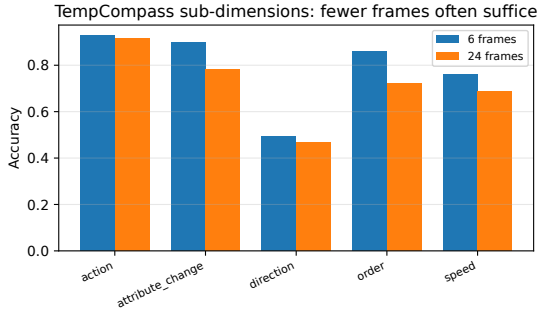

Figure 15: TempCompass sub-dimension accuracy at 6 vs. 24 frames (Qwen2.5-VL-7B). Six frames match or exceed 24 on most temporal dimensions, consistent with the declining static curve that makes STOP-EARLY strictly beneficial here.

D Additional Figures

These figures provide per-skill and cross-model robustness views of the accuracy–efficiency trade-off. They test whether the aggregate result is driven by one temporal category or model and relate the observed boundary to evidence loss under visual-token reduction (Zhang et al., 2026; Li et al., 2026b).

Figure 15 presents the TempCompass sub-dimension breakdown and complements the preceding aggregate result. The decomposition checks whether the aggregate gain comes from one unusually easy temporal skill or is distributed across the benchmark. Each sub-dimension compares the same low and high budgets, making the direction of the effect easy to inspect. Consistent gains support the interpretation that saturation is a property of the task setting rather than a single category artifact.

Figure 13 presents the cross-model TempCompass comparison and complements the preceding aggregate result. TempCompass is the clearest declining-curve case, so it provides a useful transfer test for the strongest form of the claim. Showing both model families distinguishes a task-level satu-

ration pattern from a model-specific decoding accident. The frame-saving and accuracy annotations should be read together: the result is valuable because the gain persists while compute decreases.

Figure 14 provides the complementary escalation-rescue view. It tests whether high-budget calls are spent on mid-budget-wrong items that can actually be corrected, rather than merely matching a target escalation rate.