CALIBRATED ANSWER-STABILITY HALTING FOR ACTIVE LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video question answering enables efficient access to lengthy recordings, but confidence from partial observations can be misleading when later evidence changes the answer. We seek to reduce video processing while controlling premature commitment to answers that would change under a fixed, expanded view. We introduce Stability-Based Halting (SBH), which uses a lightweight linear predictor combining current confidence, observation progress, and coarse cues from remaining video content to estimate the current answer's advantage over alternatives under that reference view. Calibration across videos yields a lower bound that triggers stopping when positive and controls early-stop reference disagreement across steps and questions under exchangeable video sampling. On held-out Video-MME videos with Qwen2.5-VL-7B, SBH achieves a $1.40\times$ speedup over scoring all 64 budgeted frames once, with 68.02% versus 67.65% accuracy and 5.19% observed video-level early-stop reference disagreement at a 10% target.

1 INTRODUCTION

Long-video question answering supports video search and instructional assistance by retrieving relevant details and reasoning across extended temporal contexts (Wu et al., 2024). Dense visual processing is costly in memory and computation. Active video understanding gathers evidence selectively, making both where to look and when to stop central to efficient inference.

Recent methods improve where computation is spent by selecting informative frames under limited budgets (Tang et al., 2025; Zhu et al., 2026). Deciding when to stop has also received explicit attention: AdaptToken-Lite uses response uncertainty to avoid processing further frame groups (Qi et al., 2026), while agentic systems assess whether more evidence is needed through reflection (Wang et al., 2026) or rubric-guided sufficiency checks (Yan et al., 2026). Stopping must also account for the risk that further evidence will change the answer.

Stopping is decided from partial evidence: a confident answer can change when a later scene resolves an ambiguity (Figure 1). Repeated stopping decisions also require reliability across the entire observation process. Calibrated early exit addresses prediction consistency as model computation grows (Schuster et al., 2021; 2022), and Thought Calibration extends calibrated stopping to reasoning length (Wu et al., 2025). Active video question answering adds a distinct challenge: observation changes the information available to the model. How can a stopping rule control premature commitment while retaining the computational benefit of selective observation?

We study *answer stability*: whether the same model would preserve its current answer under a fixed, expanded video view. Answer correctness remains a separate empirical question. Our *Stability-Based Halting* (SBH) rule combines current confidence with inexpensive remaining-video cues to predict the current answer's advantage over alternatives under that reference view. A calibrated lower bound determines when to stop and, under video exchangeability, controls early-stop reference disagreement jointly across observation steps and questions within a video.

Our contributions are threefold:

- **We reveal a confidence–stability gap.** A study of 5,040 observation states finds 10.3% endpoint disagreement among high-confidence states and an association between remaining-video rele-

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.

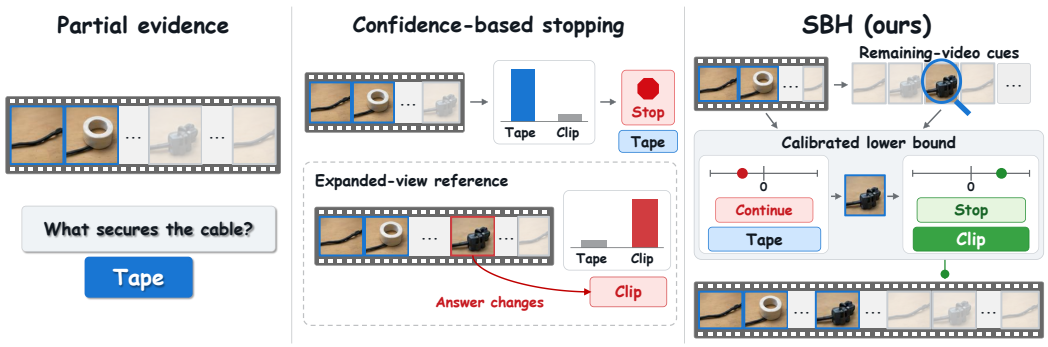

Figure 1: **High confidence does not imply a stable answer.** In this illustrative example, partial evidence favors "Tape," but an expanded video view changes the answer to "Clip." Confidence-based stopping commits too soon. SBH checks coarse remaining-video cues and stops early only when its calibrated lower bound is positive. SBH targets agreement with the same model under a fixed expanded view.

vance and answer instability (Section 3). This motivates assessing the effect of further evidence when deciding whether to stop.

- **We propose halting with explicit risk control.** SBH equips frozen video QA models with a calibrated stopping rule. We establish a finite-sample reference-consistency guarantee for adaptive stopping and characterize conditions under which the rule saves computation.
- **SBH achieves end-to-end inference savings.** On Video-MME, SBH provides 1.34–1.40× speedups across three backbones relative to scoring all budgeted frames once. With Qwen2.5-VL-7B, it attains 68.02% accuracy versus 67.65% for the budget endpoint, with 5.19% observed video-level early-stop reference disagreement at a prescribed risk level of 10%.

2 Related Work

Adaptive video computation. Query-aware frame selection balances relevance and temporal coverage in AKS (Tang et al., 2025) and adds bandit exploration in FOCUS (Zhu et al., 2026), while AdaReTaKe and FlexSelect reduce redundancy at the token level (Wang et al., 2025b; Zhang et al., 2025). More recent stopping rules use response entropy in AdaptToken-Lite (Qi et al., 2026) or coverage saturation in CSES (Zhang et al., 2026b). These criteria use confidence or representation coverage as stopping signals. SBH predicts the current winner's margin under an expanded reference view, isolating halting along fixed observation prefixes.

Evidence sufficiency. VideoAgent combines iterative retrieval with self-reflection (Wang et al., 2024). Recent agents structure this assessment through targeted observations in AVP (Wang et al., 2026), reusable sufficiency rubrics in REVEAL (Yan et al., 2026), and coverage, conflict, and hypothesis tracking in VESTA (Zhang et al., 2026a). In streaming video, StreamReady learns when supporting evidence permits a response (Azad et al., 2026). These mechanisms address semantic sufficiency or answer timing. SBH calibrates disagreement with a specified expanded-view answer using a remaining-video cue and a numerical margin bound.

Calibrated early stopping. Beyond prediction persistence in PABEE (Zhou et al., 2020), CATs and CALM calibrate early exits against full-computation predictions (Schuster et al., 2021; 2022). Thought Calibration extends this principle to reasoning length through hidden-state probes and calibrated stopping rules (Wu et al., 2025). These methods calibrate consistency as model computation increases. In active video understanding, further observation also changes the evidence available to the model. Building on split conformal prediction (Lei et al., 2018), SBH calibrates retained-margin lower bounds simultaneously across questions and observation prefixes within each video. Under video exchangeability, the resulting guarantee controls marginal early-stop disagreement with the reference.

3 PRELIMINARY STUDY

Does a confident answer remain stable as more video is observed? We analyze 5,040 intermediate states from 720 fixed eight-step video–question trajectories of a frozen QA model, without a learned stopping predictor or calibration. Let g_t be the current top-two probability gap and $F_t = \mathbf{1}\{a_t \neq a_T\}$ indicate disagreement with the trajectory's endpoint answer. This measures stability relative to the endpoint answer. Appendix B.6 details the protocol and statistical analysis.

Confidence does not ensure stability. Overall, 17.7% of intermediate answers disagree with the endpoint. Even when $g_t \geq 0.50$, disagreement is 10.3%, rising to 21.2% at early steps ($t/T \leq 3/8$; Figure 2a). Thus, confidence in observed evidence can coexist with sensitivity to further observation.

Unseen relevance provides additional information. Within 20 observation-step and confidence-gap strata, comparing the lowest and highest quartiles of recorded question-only unseen relevance yields weighted disagreement rates of 11.7% and 23.8%, a 12.1 percentage-point increase (Figure 2b). Eighteen strata show an increase, indicating an association between remaining-video relevance and endpoint disagreement within step–gap strata. Together, these findings motivate predicting whether the current winner will survive additional evidence, using remaining-video information and a calibrated stopping criterion.

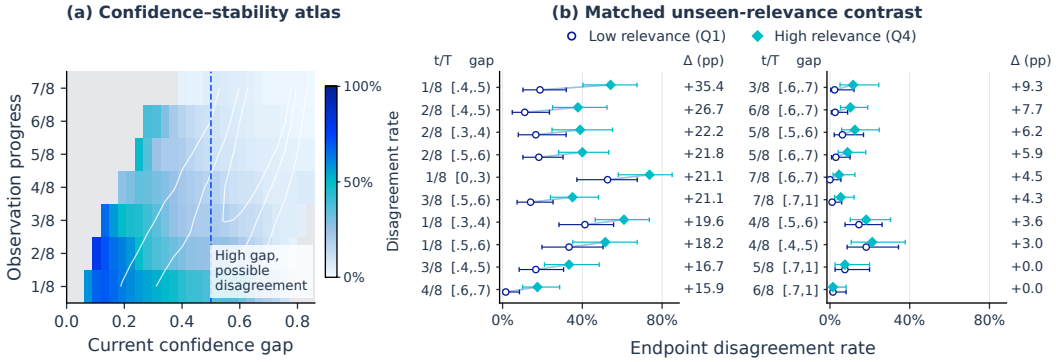

Figure 2: **Confidence and endpoint stability.** **(a)** Endpoint disagreement by confidence gap and observation progress. **(b)** Low- versus high-unseen-relevance quartiles within step–gap strata (open circles versus filled diamonds; 95% Wilson intervals). Δ denotes percentage-point differences. Appendix B.6.2 gives aggregation details.

4 METHOD

4.1 OVERVIEW

SBH predicts the *retained margin*: the current answer's advantage over its strongest competitor under an expanded reference view. A five-coefficient linear predictor combines confidence, observation progress, and a coarse remaining-video cue. Video-level calibration gives a simultaneous lower bound across questions and prefixes; a positive bound triggers stopping (Figure 3). The QA model stays frozen. Reference views supervise training offline; deployment follows a fixed observation path.

4.2 OBSERVATION PREFIXES AND THE RETAINED MARGIN

Given a video V, a question q, and $M \geq 2$ candidate texts $\mathcal{C} = \{c_1, \ldots, c_M\}$, divide V into K equal-duration blocks. A frozen SigLIP encoder (Zhai et al., 2023) maps block j's midpoint frame and the text "Question: {q}" to unit vectors $z_j \in \mathbb{R}^d$ and $v_q \in \mathbb{R}^d$, respectively. Rank blocks by decreasing $z_j^\top v_q$, breaking ties chronologically. Each selected block supplies b uniformly spaced fine-frame timestamps. Map them to frame identities and skip blocks that add no new frame. Each accepted block uses one step, yielding strictly nested evidence sets $E_1 \subset \cdots \subset E_T$. The accepted count T is capped at $T_{\max}$; the path and T are fixed before scoring. Appendix A.3 specifies timestamp mapping and short-video handling.

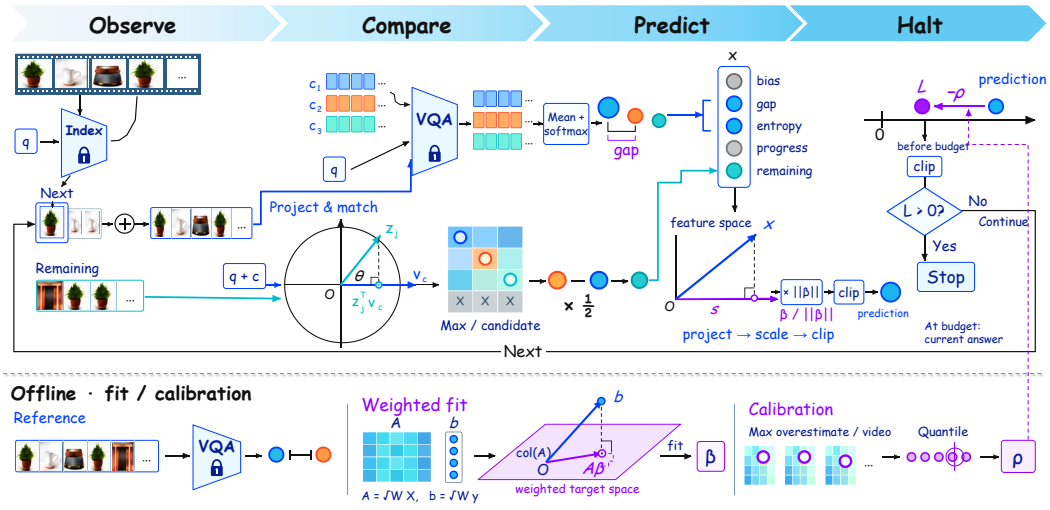

Figure 3: **Stability-Based Halting (SBH). Online (top):** confidence, progress, and remaining-video matches form x_t. A clipped linear predictor estimates the retained margin; SBH returns the current answer when $L_t > 0$ or the budget ends. **Offline (bottom):** reference margins supervise a weighted least-squares fit of β; disjoint calibration videos determine ρ_α from maximum overestimates. Geometric panels show dot-product matching, linear prediction, and weighted least-squares projection.

At each prefix, a frozen model P_θ scores candidates using the accumulated frames in temporal order and the same question and candidate list, without previous model answers:

$$s_t(c) = \frac{1}{\ell_c} \sum_{u=1}^{\ell_c} \log P_\theta(w_{c,u} \mid q, \mathcal{C}, E_t, w_{c,<u}),$$
$$p_t = \mathrm{softmax}(s_t), \qquad a_t = \arg\max_{c \in \mathcal{C}} p_t(c). \tag{1}$$

Here $w_{c,u}$ is candidate c's uth token and ℓ_c its token count; only candidate-text tokens are scored. Softmax uses temperature one, with ties broken by original candidate index.

For offline supervision, a deterministic expanded view E_H retains E_T and adds frames from across the video up to budget $B_H \geq bT_{\max}$ (Appendix A.3). The same scorer produces the reference distribution p_H and winner a_H. Define the current and retained margins as

$$m_t = p_t(a_t) - \max_{c \neq a_t} p_t(c), \qquad \gamma_t = p_H(a_t) - \max_{c \neq a_t} p_H(c). \tag{2}$$

The current margin $m_t \in [0, 1]$ measures confidence in observed evidence; the retained margin $\gamma_t \in [-1, 1]$ is positive exactly when a_t is the unique reference winner. This sign test captures stability even if losing candidates change order. Appendix A.1 explains its geometry and compares it with total-variation certificates.

4.3 A LINEAR PREDICTOR WITH REMAINING-VIDEO INFORMATION

Let $v_c \in \mathbb{R}^d$ be the unit SigLIP embedding of "Question: {q} Answer: {c}", and $\mathcal{U}_t$ the remaining-index blocks with at least one fine frame outside E_t. For $t < T$, define

$$r_t(c) = \max_{j \in \mathcal{U}_t} z_j^\top v_c, \qquad u_t = \frac{1}{2}\left[\max_{c \neq a_t} r_t(c) - r_t(a_t)\right],$$
$$x_t = [1, m_t, h_t, t/T, u_t]^\top, \qquad h_t = -\frac{\sum_c p_t(c) \log p_t(c)}{\log M}. \tag{3}$$

Here $r_t(c)$ is candidate c's strongest remaining coarse match; $u_t \in [-1, 1]$ compares the best competitor with the current winner. Normalized entropy $h_t \in [0, 1]$ uses $0 \log 0 = 0$. These, the current margin, an intercept, and observation fraction t/T form $x_t \in \mathbb{R}^5$. Every eligible early-stop state has nonempty $\mathcal{U}_t$.

A positive u_t favors a competitor in the remaining video, distinguishing prefixes with similar confidence and progress but different remaining content. Maxima may come from different blocks, so this frame-level cue ignores temporal order. Separately calibrated removal and randomization variants isolate its contribution (Appendix B.5).

Across all valid training prefixes $t < T$, stack feature rows $x_t^\top$ into $X \in \mathbb{R}^{N_s \times 5}$ and retained margins into $y \in \mathbb{R}^{N_s}$. Weighted least squares gives

$$\beta = (W^{1/2}X)^\dagger W^{1/2} y, \qquad \widehat{\gamma}_t = \mathrm{clip}_{[-1,1]}(\beta^\top x_t). \tag{4}$$

The dagger denotes the Moore–Penrose inverse. Diagonal W weights videos equally, then eligible questions within each video and states within each question equally. A thin SVD computes the minimum-norm fit $\beta \in \mathbb{R}^5$ (Appendix A.4). Clipping enforces the target range. Features, model settings, and coefficients are fixed before calibration.

4.4 Calibrated stopping

To account for adaptive stopping and dependent questions, SBH uses split conformal calibration (Lei et al., 2018; Angelopoulos & Bates, 2023) on each video's maximum overestimate across questions and prefixes. For n calibration videos with prespecified question sets $\mathcal{Q}_i$, define

$$R_i = \max_{q \in \mathcal{Q}_i} \max_{1 \le t < T_{i,q}} (\widehat{\gamma}_{i,q,t} - \gamma_{i,q,t}), \qquad k = \lceil (n+1)(1-\alpha) \rceil, \qquad \rho_\alpha = \max\{0, R_{(k)}\}. \tag{5}$$

Here $T_{i,q}$ is the prefix budget for question q in video i, $\alpha \in (0,1)$ is the prescribed risk level, and $R_{(k)}$ is the kth smallest calibration score. Set $R_{(n+1)} = +\infty$ and any empty maximum to $-\infty$. The shared correction ρ_α gives the lower bound and stopping rule

$$L_t = \mathrm{clip}_{[-1,1]}(\widehat{\gamma}_t - \rho_\alpha), \qquad \tau = \min(\{t < T : L_t > 0\} \cup \{T\}), \qquad \widehat{c} = a_\tau. \tag{6}$$

SBH returns the current winner at the first prefix with a positive bound, or a_T if no early stop occurs. For $\rho_\alpha = +\infty$, set $L_t = -1$.

Theorem 1 (Reference consistency)**.** *Condition on the specification and predictor fixed before calibration. Suppose the calibration videos and a new video, each including its question set, are exchangeable. All follow the same observation, scoring, and reference rules, with well-defined scores and features. Then*

$$\begin{aligned} &\Pr(\forall q \in \mathcal{Q}_{\mathrm{new}},\ \forall t < T_q : \gamma_{q,t} \ge L_{q,t}) \ge 1 - \alpha, \\ &\Pr\big(\exists q \in \mathcal{Q}_{\mathrm{new}} : \tau_q < T_q,\ a_{q,\tau_q} \ne a_{H,q}\big) \le \alpha. \end{aligned} \tag{7}$$

Proof sketch. Exchangeability gives $\Pr(R_{\mathrm{new}} > R_{(k)}) \le \alpha$. On the complementary event, every retained margin is at least its lower bound, including at an adaptively selected stop. Thus $L_{q,\tau_q} > 0$ implies $\gamma_{q,\tau_q} > 0$, so the returned answer is the unique reference winner. Appendix D.2 gives the full rank argument and treatment of ties.

When early stopping is possible. Since $\widehat{\gamma}_t \le 1$, any $\rho_\alpha \ge 1$ disables early stopping, even when finite. A correction $\rho_\alpha < 1$ permits stopping only if some prediction exceeds it. A difficult calibration question can raise the video's maximum overestimate and suppress stopping for other questions. Appendix A.2 gives a sufficient condition for early stopping. Correction size, positive-bound frequency, and early-stop rate measure conservatism (Appendix B.5). Computation savings must also outweigh repeated scoring and indexing costs, relative to a single call on E_T (Appendix A.6).

The guarantee averages over calibration and new videos, bounding any early-stop disagreement with the fixed reference within a new video. It does not cover budget-endpoint answers, task accuracy, disagreement conditional on stopping, or latency. SBH extends consistency-calibrated early exit (Schuster et al., 2021; 2022) to video observation prefixes through retained margins and remaining-video cues. Appendix A specifies the deterministic protocol.

4.5 Risk-constrained stopping efficiency

Theorem 1 controls early-stop disagreement. We now bound the additional computation needed to stop. Along the fixed path, let $\tau_\star = \min(\{t < T : \gamma_t > 0\} \cup \{T\})$ be the first prefix with a positive retained margin, or the budget endpoint. Let $C(t)$ be the complete cumulative cost, with nonnegative increments, and write $D = C(T) - C(\tau_\star)$ and $\bar{D} = \mathbb{E}D < \infty$.

For the clipped population least-squares predictor $f_\star$, let $\sigma(s)$ be the first reference-consistent prefix with $f_\star(x_t) > s$, or T if none exists. The profile $\Phi(s) = \mathbb{E}[C(\sigma(s)) - C(\tau_\star)]$ measures the expected extra cost of waiting for this threshold. Answers may change repeatedly; the comparator $\tau_\star$ uses a reference unavailable at deployment.

Theorem 2 (Finite-sample stopping efficiency)**.** *Assume the independent-video sampling, moment, and numerical conditions in Appendix A.7. Use N eligible training videos and n calibration videos with $k \le n$, and choose parameters satisfying Eq.* (13)*. Fix a deterministic population-level threshold a. Let e_N and $\bar{\delta}$ be the uniform-error radius relative to $f_\star$ and the joint fitting/calibration failure bound in Eqs.* (13) *and* (52)*. With $s_N = [a + e_N]_+ + e_N$, where $[\cdot]_+$ denotes the positive part, SBH retains the risk guarantee of Theorem 1 and satisfies*

$$\mathbb{E}\big[(C(\tau) - C(\tau_\star))_+\big] \le \Phi(s_N) + \bar{\delta}\big(\bar{D} - \Phi(s_N)\big). \tag{8}$$

Proof sketch. Video-level regression concentration and an exact binomial calibration count imply that, with probability at least $1 - \bar{\delta}$, prediction error is uniformly at most e_N and $\rho_\alpha \le [a + e_N]_+$. On this event, SBH stops by the first prefix qualifying for $\sigma(s_N)$, so $\tau \le \sigma(s_N)$. Otherwise, excess cost is at most D. Independence from the test video yields the weighted bound above. Appendix D.9 gives the full proof.

The bound separates finite-sample uncertainty from the population cost profile Φ. Appendix D.10 gives explicit training and calibration sample requirements. Corollary 1 relates Φ to reference margins and states when SBH saves computation relative to a single full-budget call.

5 Experiments

We organize the evaluation around four questions, moving from overall utility to the information and calibration needed for stopping, then to decision behavior:

Q1: Does SBH offer a useful accuracy–cost tradeoff across models and datasets? (Section 5.2)
Q2: Does remaining-video information improve stopping? (Section 5.3)
Q3: Why calibrate across an entire video? (Section 5.4)
Q4: When should a persistent answer be reconsidered? (Section 5.5)

5.1 Experimental setup

Benchmarks and metrics. Evaluation covers LongVideoBench, Video-MME, MLVU-MCQ, LVBench (Wu et al., 2024; Fu et al., 2025; Zhou et al., 2025; Wang et al., 2025a), and Video-MME-v2 (Fu et al., 2026) with visual-only input. Metrics include official accuracy, MLVU M-Avg, Video-MME-v2 group scores, video-level reference disagreement and simultaneous coverage, early-stop rate, and end-to-end latency. Dataset splits, supplementary benchmarks, and metric definitions are in Appendix B.1.

Comparison protocol. Stopping rules share observation prefixes, backbones, and answer scoring, with operating points selected on development videos. Native frame/token-selection results, including FOCUS (Zhu et al., 2026) and AdaptToken (Qi et al., 2026), provide separate system-level context. Baseline definitions and calibration, component, and cost analyses appear in Appendices B.2 and C.

Backbones and implementation. The backbones are frozen Qwen2.5-VL-7B-Instruct, LLaVA-OneVision-7B, and InternVL2.5-8B (Bai et al., 2025; Li et al., 2024; Chen et al., 2024). The hardware configuration comprises eight NVIDIA A100 GPUs with inference batch size one. SBH uses $\alpha = 0.10$, with separate fitting and calibration per backbone. Hyperparameters, preprocessing, and cost accounting appear in Appendix B.3.

5.2 Q1: Does SBH offer a useful accuracy–cost tradeoff?

Full-benchmark accuracy. Table 1 groups native systems by backbone and benchmark. SBH has the highest reported score in all 15 backbone–metric combinations, with descriptive gains of 0.5–1.1 percentage points over the strongest baseline in each column. Native frame, token, and context

budgets differ; these comparisons provide system-level context, while the shared-path experiments below isolate stopping (Appendix B.8).

Table 1: **Full-benchmark accuracy (%).** Evaluation details and notation appear in Appendix C.1.

			Video-MME		LongVB	MLVU	LVBench
Backbone	Method	Budget	Overall	Long	Val	M-Avg	Acc.
Qwen2.5-VL 7B-Instruct	AdaReTaKe (Wang et al., 2025b)	16K c	67.7	58.3	60.8	75.0	51.2
	FlexSelect (Zhang et al., 2025)	7010 t	68.2	59.3	62.4	72.5	51.2
	FlexSelect-Lite (Zhang et al., 2025)	7010 t	67.4	58.6	61.9	70.3	50.0
	AdaptToken (Qi et al., 2026)	7010 t	68.0	59.1	62.6	73.6	54.8
	AdaptToken-Lite (Qi et al., 2026)	7010 t	67.2	58.4	61.7	72.8	53.3
	SBH (ours)	≤ 64 f	**69.0**$_{+0.8}$	**60.4**$_{+1.1}$	**63.1**$_{+0.5}$	**75.6**$_{+0.6}$	**55.3**$_{+0.5}$
InternVL2.5 8B	FlexSelect (Zhang et al., 2025)	8256 t	67.0	58.1	59.8	71.9	49.7
	FlexSelect-Lite (Zhang et al., 2025)	8256 t	67.2	57.9	59.4	71.9	49.9
	AdaptToken (Qi et al., 2026)	8256 t	67.5	58.4	60.4	72.1	52.1
	AdaptToken-Lite (Qi et al., 2026)	8256 t	66.8	57.6	59.6	71.3	51.3
	SBH (ours)	≤ 64 f	**68.0**$_{+0.5}$	**59.0**$_{+0.6}$	**61.0**$_{+0.6}$	**72.7**$_{+0.6}$	**52.8**$_{+0.7}$
LLaVA-OneVision-7B	Uniform (Zhu et al., 2026)	32 f	56.5	47.3	54.8	64.3	46.8
	AKS (Tang et al., 2025)	32 f	57.7	48.6	57.4	65.4	48.1
	FOCUS (Zhu et al., 2026)	32 f	58.3	49.8	60.7	66.5	49.7
	SBH (ours)	≤ 64 f	**59.0**$_{+0.7}$	**50.7**$_{+0.9}$	**61.4**$_{+0.7}$	**67.2**$_{+0.7}$	**50.6**$_{+0.9}$

Controlled stopping and executed cost. With identical prefixes and scoring on Qwen2.5-VL/Video-MME (Table 2), SBH achieves 68.02% accuracy versus the endpoint's 67.65%, while latency falls from 10.84 to 7.74 seconds (1.40× speedup). At $\alpha = 0.10$, it stops early on 92.35% of questions with 5.19% observed video-level reference risk. Confidence and persistence rules incur higher risk and latency; calibrated TV reduces risk to 2.96% at a latency above the endpoint. Cost accounting includes indexing and all scoring calls (Table 8): sequential inference without early exit takes 34.96 seconds, exposing the cost of repeatedly revisiting evidence. Paired 95% intervals for latency differences against the four primary controls lie below zero. Accuracy differences remain uncertain, with all four intervals spanning zero (Table 9).

Table 2: **Controlled stopping on Video-MME.** Qwen2.5-VL-7B; see Appendix B.7.

	Answer quality		Efficiency		
Stopping rule	Acc.↑	$\widehat{J}_{\text{video}}$↓	Early↑	Time (s)↓	Speedup↑
Reference controls					
Endpoint E_T, scored once	67.65$_{-0.37}$	**0.00**$_{-5.19}$	0.00$_{-92.35}$	10.84$_{+3.10}$	1.00×$_{-0.40}$
Sequential, no early exit	67.65$_{-0.37}$	**0.00**$_{-5.19}$	0.00$_{-92.35}$	34.96$_{+27.22}$	0.31×$_{-1.09}$
Confidence and persistence					
Candidate entropy (Qi et al., 2026)	66.91$_{-1.11}$	11.85$_{+6.66}$	85.93$_{-6.42}$	8.82$_{+1.08}$	1.23×$_{-0.17}$
Candidate margin (Schuster et al., 2022)	67.16$_{-0.86}$	10.74$_{+5.55}$	84.94$_{-7.41}$	8.66$_{+0.92}$	1.25×$_{-0.15}$
Persistent answer (Zhou et al., 2020)	66.79$_{-1.23}$	12.96$_{+7.77}$	82.72$_{-9.63}$	9.05$_{+1.31}$	1.20×$_{-0.20}$
Confidence count (Qi et al., 2026)	67.04$_{-0.98}$	11.48$_{+6.29}$	86.67$_{-5.68}$	8.57$_{+0.83}$	1.26×$_{-0.14}$
Classification and calibration					
Equal-feature classifier (Schuster et al., 2021)	67.41$_{-0.61}$	8.52$_{+3.33}$	89.38$_{-2.97}$	8.21$_{+0.47}$	1.32×$_{-0.08}$
Calibrated TV bound (Lei et al., 2018)	67.53$_{-0.49}$	2.96$_{-2.23}$	56.54$_{-35.81}$	12.47$_{+4.73}$	0.87×$_{-0.53}$
Verifier adaptations					
AVP-style verifier (Wang et al., 2026)	67.28$_{-0.74}$	7.78$_{+2.59}$	87.78$_{-4.57}$	9.36$_{+1.62}$	1.16×$_{-0.24}$
REVEAL-style verifier (Yan et al., 2026)	67.53$_{-0.49}$	6.67$_{+1.48}$	88.89$_{-3.46}$	8.91$_{+1.17}$	1.22×$_{-0.18}$
SBH (ours)	**68.02**	5.19	**92.35**	**7.74**	**1.40×**

Breadth across backbones and datasets. Controlled speedups are 1.34–1.40× across three backbones on Video-MME and 1.34–1.37× on three further datasets (Tables 3 and 7). Source-only trans-

fer also gives the highest point estimates among the compared rules on Video-MME-v2, EgoSchema, and MINERVA (Table 4). Coverage under transfer is assessed empirically. Detailed scores and uncertainty appear in Appendices C.1 and C.

Takeaway (Q1). Early stopping pays off only when avoided computation outweighs repeated inspection. Its value depends on net savings at a stated reference risk, with accuracy assessed separately.

5.3 Q2: Does remaining-video information improve stopping?

Figure 4 tests the cue u_t on Qwen2.5-VL/Video-MME with a shared observation path, scorer, and reference (Table 6A).

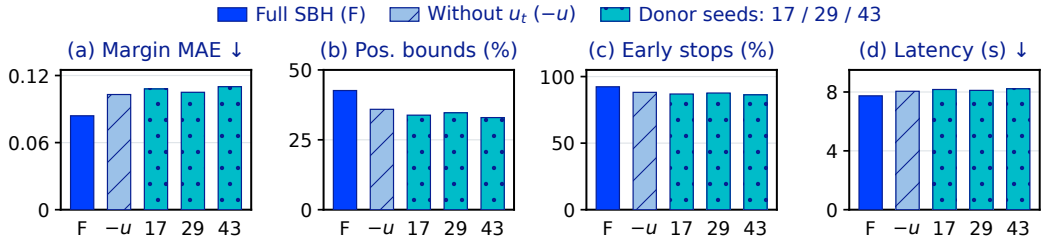

Figure 4: **The remaining-video cue improves the observed stopping tradeoff.** Each variant is refitted and separately video-calibrated at $\alpha = 0.10$. F denotes full SBH; 17/29/43 identify training-only donor seeds. Positive-bound frequency uses all eligible held-out prefixes. Bars show point estimates from Table 6A.

Prediction quality and stopping utility. With u_t, retained-margin MAE falls from 0.103 to 0.084 (18.4% relative), positive bounds rise by 6.72 percentage points, early stopping rises from 88.15% to 92.35%, and latency falls from 8.05 to 7.74 seconds. All three donor replacements increase MAE (0.105–0.110) and latency (8.11–8.22 seconds), supporting the cue's incremental value under separate calibration.

Takeaway (Q2). Coarse retrieval can inform whether to keep looking, as well as where to look. Remaining-video information can thus guide computation before detailed inspection.

5.4 Q3: Why calibrate across an entire video?

We next hold the full predictor and zero stopping threshold fixed, changing only the calibration unit (Figure 5; Table 6B). This separates the cue's predictive value in Q2 from the reliability of repeated stopping decisions.

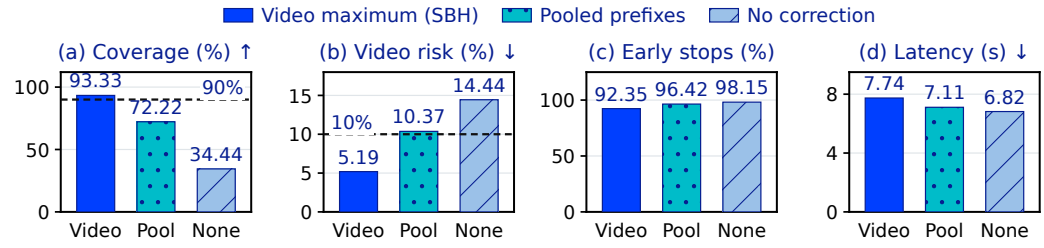

Figure 5: **Video-level calibration trades some speed for reference-risk control.** The fitted full predictor and zero stopping threshold are fixed. Dashed lines mark nominal 90% simultaneous coverage and 10% video-level reference risk at $\alpha = 0.10$. Bars show point estimates from Table 6B.

Simultaneous reliability versus faster exits. Video-maximum calibration yields 93.33% simultaneous coverage and 5.19% observed reference risk. Pooled-prefix calibration and no correction

reduce latency to 7.11 and 6.82 seconds. Their coverage falls to 72.22% and 34.44%, with risk rising to 10.37% and 14.44%. These faster exits come with lower reliability under the same predictor.

Risk levels and evaluation uncertainty. Across the settings in Table 5, observed coverage exceeds $1-\alpha$ and reference risk remains below α. On Video-MME, tightening α to 0.05 lowers observed risk to 1.85–2.96% and early-stop frequency to 80.86–84.07% across backbones. These held-out rates describe the fitted rules. The guarantee controls reference disagreement marginally over calibration and new videos. Evaluation uncertainty is larger on LVBench, whose 32 videos yield a risk interval of [0.00, 15.63]% (Table 9).

Takeaway (Q3). Calibration should match the unit at which failure matters: any premature commitment within a video. Pooling dependent prefixes can make a rule appear more efficient by leaving repeated opportunities for error unaccounted for.

5.5 Q4: When should a persistent answer be reconsidered?

Figure 6 illustrates the confidence–stability gap (Section 3) through a constructed mango-preparation example. Scenes are AI-generated; responses and bound signs are manually specified (Appendix C.5). The initial answer C (serving fruit pieces) persists through three prefixes, yet $L_1, L_2, L_3 \leq 0$ keeps observation active. At the fourth prefix, the blender resolves the intended use: the answer changes to B (blending a smoothie), and $L_4 > 0$ triggers stopping.

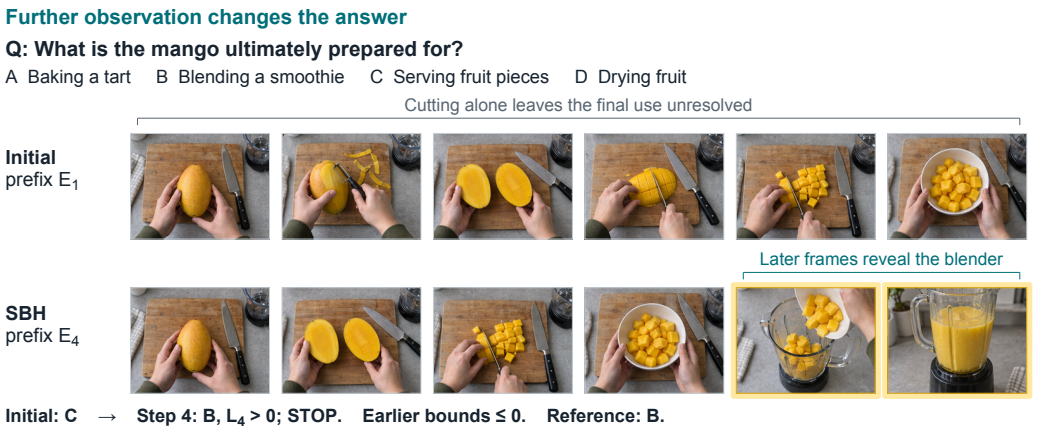

Figure 6: **Illustration of answer correction before early stopping.** Representative frames from the initial and stopping prefixes; yellow borders highlight the later evidence. Case provenance and additional cases appear in Appendix C.5.

Decision boundaries. The additional cases in Appendix C.5 contrast early stopping, budget termination without a positive bound, and agreement with an erroneous reference, separating ambiguity resolution from answer correctness.

Takeaway (Q4). Repeated answers may reflect repeated exposure to the same ambiguity, so persistence alone is a weak reason to stop. Remaining-video cues help a stability criterion respond to unresolved evidence; answer correctness is evaluated against ground truth.

6 Conclusion

SBH combines confidence and remaining-video cues to predict answer stability under an expanded reference view, addressing premature stopping in active long-video understanding. Video-level calibration controls early-stop reference disagreement in finite samples under video exchangeability; our cost analysis characterizes when stopping saves computation. On Video-MME, SBH reduces end-to-end latency across three backbones with comparable observed accuracy, supporting stability as a practical stopping criterion. Jointly optimizing evidence acquisition and calibrated stopping while preserving reference-disagreement control remains an open question.

AI USE STATEMENT

Generative AI tools were used to assist with organizing and interpreting experimental results, language polishing, and manuscript formatting. The authors take full responsibility for all text and results reported in this paper.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://arxiv.org/abs/2107.07511.

Shehreen Azad, Vibhav Vineet, and Yogesh S. Rawat. StreamReady: Learning what to answer and when in long streaming videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40494–40504, 2026. URL https://arxiv.org/abs/2603.08620.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Zhe Chen, Weiyun Wang, Yue Cao, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. URL https://arxiv.org/abs/2412.05271.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, 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*, pp. 24108–24118, 2025. URL https://arxiv.org/abs/2405.21075.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026. URL https://arxiv.org/abs/2604.05015.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://arxiv.org/abs/1604.04173.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

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*, volume 36, 2023. URL https://arxiv.org/abs/2308.09126.

Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, Cordelia Schmid, and Tobias Weyand. MINERVA: Evaluating complex video reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23968–23978, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Nagrani_MINERVA_Evaluating_Complex_Video_Reasoning_ICCV_2025_paper.html.

Haozhe Qi, Kevin Qu, Mahdi Rad, Rui Wang, Alexander Mathis, and Marc Pollefeys. AdaptToken: Entropy-based adaptive token selection for MLLM long video understanding. *arXiv preprint arXiv:2603.28696*, 2026. URL https://arxiv.org/abs/2603.28696.

Tal Schuster, Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Consistent accelerated inference via confident adaptive transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4962–4979, 2021. doi: 10.18653/v1/2021.emnlp-main.406. URL https://aclanthology.org/2021.emnlp-main.406/.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6fac9e316a4ae75ea244ddcef1982c71-Abstract-Conference.html.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29118–29128, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tang_Adaptive_Keyframe_Sampling_for_Long_Video_Understanding_CVPR_2025_paper.html.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z. URL https://tropp.caltech.edu/papers/Tro11-User-Friendly-preprint.pdf.

Weihan Wang, Zehai He, Wenyi Hong, et al. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a. URL https://arxiv.org/abs/2406.08035.

Xiao Wang, Qingyi Si, Shiyu Zhu, Jianlong Wu, Li Cao, and Liqiang Nie. AdaReTaKe: Adaptive redundancy reduction to perceive longer for video-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5417–5432, 2025b. doi: 10.18653/v1/2025.findings-acl.283. URL https://aclanthology.org/2025.findings-acl.283/.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/10325_ECCV_2024_paper.php.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*, 2026. URL https://arxiv.org/abs/2512.05774. Version 2.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, 2024. URL https://arxiv.org/abs/2407.15754.

Menghua Wu, Cai Zhou, Stephen Bates, and Tommi Jaakkola. Thought calibration: Efficient and confident test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14291–14305, 2025. doi: 10.18653/v1/2025.emnlp-main.722. URL https://aclanthology.org/2025.emnlp-main.722/.

Caijun Yan, Yang Zhou, Meixing Shi, Haoran Sun, Yichen Li, Yuxiang Cai, and Yankai Jiang. REVEAL: A rubric-guided agent for explicit evidence sufficiency verification in long-video question answering. *arXiv preprint arXiv:2608.08612*, 2026. URL https://arxiv.org/abs/2608.08612.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. URL https://arxiv.org/abs/2303.15343.

Can Zhang, Baofeng Zhang, Xiaotian Han, Junyuan Shang, Yuchen Ding, Shuohuan Wang, Dianhai Yu, and Ruirui Li. From intent to evidence: Policy-steered multi-strategy retrieval for long-video agents. *arXiv preprint arXiv:2608.31005*, 2026a. URL https://arxiv.org/abs/2608.31005.

Junyang Zhang, Puhan Luo, Chen Tang, Yuxi Shi, and Xiang-Yang Li. Coverage-driven adaptive keyframe selection for video understanding. *arXiv preprint arXiv:2608.00714*, 2026b. URL https://arxiv.org/abs/2608.00714.

Yunzhu Zhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. FlexSelect: Flexible token selection for efficient long video understanding. *arXiv preprint arXiv:2506.00993*, 2025. URL https://arxiv.org/abs/2506.00993.

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*, pp. 13691–13701, 2025. URL https://arxiv.org/abs/2406.04264.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. BERT loses patience: Fast and robust inference with early exit. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d4dd111a4fd973394238aca5c05bebe3-Abstract.html.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a010577f93a8301dada61f0b97cda923-Abstract-Conference.html.

Appendix

CONTENTS Page

A Technical Details **14**
A.1 Contrast geometry and the total-variation comparison . . . 14
A.2 Nonvacuity of the stopping rule . . . 14
A.3 Deterministic observation and reference views . . . 14
A.4 Weighted fit and numerical conventions . . . 15
A.5 Offline and online computation . . . 15
A.6 Computation accounting . . . 15
A.7 Sampling units, population projection, and cost conventions . . . 16

B Experimental Details **18**
B.1 Benchmarks, splits, and metrics . . . 18
B.2 Comparison families and alignment . . . 19
B.3 Backbones, implementation, and resource accounting . . . 19
B.4 Encoders, prompts, and image processing . . . 20
B.5 Quantities for assessing the stopping mechanism . . . 21
B.6 Preliminary study: protocol and statistical details . . . 22
B.7 Reading the main comparisons . . . 23
B.8 Score provenance and resource conditions . . . 24

C Additional Experimental Results **25**
C.1 Additional backbones and transfer evaluation . . . 25
C.2 Calibration and component controls . . . 26
C.3 Controlled generalization across datasets . . . 27
C.4 Executed cost and statistical uncertainty . . . 28
C.5 Qualitative illustrations . . . 29
C.6 Budget termination and reference errors . . . 30

D Proofs **31**
D.1 Proof of the exact margin comparison . . . 31
D.2 Proof of Theorem 1 . . . 31
D.3 Population projection and cluster moments . . . 31
D.4 A matrix Laplace argument with scalar conditional bounds . . . 32
D.5 A variance-sensitive bound for the video residual vector . . . 34
D.6 Uniform estimation of the weighted least-squares target . . . 35
D.7 Exact finite-video calibration, including atoms and empty trajectories . . . 35
D.8 Nonmonotone answers and the exact excess-cost profile . . . 37
D.9 Proof of the main theorem and comparison with single-call inference . . . 38
D.10 Explicit training and calibration sample requirements . . . 39
D.11 Optional sharper score radii and deterministic numerical checks . . . 41
D.12 Sharpness and the independent-video sampling unit . . . 42

A TECHNICAL DETAILS

A.1 CONTRAST GEOMETRY AND THE TOTAL-VARIATION COMPARISON

For candidate a, define $A_a \in \mathbb{R}^{(M-1)\times M}$ with rows $(e_a - e_c)^\top$ for $c \neq a$, where e_c is the cth standard basis vector. Then $(A_a p)_c = p(a) - p(c)$ and $A_a \mathbf{1} = 0$. Within the probability simplex, $A_a p > 0$ describes exactly the region where a is the unique winner. Equation (2) takes the smallest coordinate of this contrast vector at the current and reference distributions. This representation explains why common score offsets cancel; evaluating SBH requires only the two scalar margins.

Proposition 1 (Exact margin comparison)**.** *For any candidate distributions p_t, p_H, the retained margin satisfies $\gamma_t \geq m_t - 2D_t$, where $D_t = \frac{1}{2}\|p_t - p_H\|_1$. Thus $m_t > 2D_t$ implies $\gamma_t > 0$, and the converse can fail.*

A proof is given in Appendix D.1.

Ties and relabeling. If the reference maximum is tied, then $\gamma_t \leq 0$, with equality exactly when a_t belongs to the tied maximum. For example, $p_H = (0.4, 0.4, 0.2)^\top$ and $a_t = c_3$ give $\gamma_t = -0.2$. Such a reference cannot produce a strictly positive certified bound on the coverage event. A tie at the current prefix can still yield $\gamma_t > 0$ for its tie-broken answer. Under *simultaneous relabeling of fixed scores and embeddings*, m_t, h_t, u_t, γ_t are invariant when the current winner is unique. With a current tie, this also requires preserving the identity selected by the tie breaker. Changing the candidate order in the model prompt and rescoring is a different operation: the scorer need not be permutation equivariant.

A.2 NONVACUITY OF THE STOPPING RULE

For any fixed trajectory with $T > 1$, an early stop occurs exactly when $\max_{t<T} \widehat{\gamma}_t > \rho_\alpha$. Consequently, $\rho_\alpha < 1$ is necessary but not sufficient. If $k = n + 1$, no early stop is possible; a finite corrected quantile requires $n \geq \lceil 1/\alpha \rceil - 1$. Finite corrections in $[1, 2]$ can also disable the rule even when every score is available.

A deterministic sufficient condition clarifies the role of prediction quality. Suppose $k \leq n$ and $|\widehat{\gamma} - \gamma| \leq \epsilon$ at every calibration state and every state of a new trajectory, for some $\epsilon \geq 0$. Then each calibration score is at most ϵ, so $\rho_\alpha \leq \epsilon$. If a new prefix has $\gamma_t > 2\epsilon$, then $\widehat{\gamma}_t \geq \gamma_t - \epsilon > \epsilon \geq \rho_\alpha$, and SBH stops no later than that prefix. This sufficient condition relates approximation quality to early stopping. Theorem 1 applies to any predictor frozen before calibration.

For fixed predictions and trajectories, enlarging a calibration video's question set cannot decrease its maximum score. Enlarging every calibration set therefore cannot decrease the order statistic and hence cannot decrease its nonnegative correction. This monotonicity exposes the cost of simultaneous protection: rare hard questions can determine the threshold for all deployment questions. Useful early stopping depends on both small upper-tail prediction errors and positive reference margins.

A.3 DETERMINISTIC OBSERVATION AND REFERENCE VIEWS

Let the video have duration $D > 0$ and decodable frame timestamps in $[0, D]$, ordered by timestamp and then source-frame index. A requested timestamp maps to the nearest frame, with ties resolved by this order. Block j spans $[(j-1)D/K, jD/K]$ and its representative timestamp is $(j - \frac{1}{2})D/K$. Its fine-frame set G_j contains the frames requested at $(j - 1 + (l - \frac{1}{2})/b)D/K$ for $l = 1, \ldots, b$, with duplicate frame identities removed.

Sort blocks by decreasing $z_j^\top v_q$, breaking ties by increasing j. Start with $E_0 = \varnothing$ and scan that order. When $G_j \setminus E_t$ is empty, skip the block without consuming a step. Otherwise append the prefix $E_{t+1} = E_t \cup G_j$. Continue until $T_{\max}$ nonempty additions have been accepted; if the order is exhausted sooner, its accepted count is T. Frame identities and coarse similarities determine the schedule before any prefix is scored. It fixes T before scoring and guarantees strict nesting. Define $\mathcal{U}_t = \{j : G_j \setminus E_t \neq \varnothing\}$, including blocks beyond the observation cap. Each prefix is presented in timestamp order, regardless of retrieval order. When $T = 1$, there is no early-stop state.

For the reference, initialize $E_H = E_T$ and let $r = B_H - |E_T|$. If $r > 0$, request the midpoints of r equal-duration bins spanning the whole video and add their distinct frame identities. If duplicates leave spare capacity, repeatedly add the unused frame maximizing its minimum timestamp distance to E_H, breaking ties toward the earlier frame. Stop at B_H frames; if fewer frames exist, use all of them. This gives $E_T \subseteq E_H$ without changing the online trajectory. The reference configuration fixes $K = 32$, $b = 8$, $T_{\max} = 8$, $B_H = 128$, and $\alpha = 0.10$.

A.4 Weighted fit and numerical conventions

For training video i, let $\mathcal{Q}_i^+$ contain its questions with at least one valid early state, and let $\mathcal{T}_{i,q}$ contain those states. Let N_v be the number of videos with nonempty $\mathcal{Q}_i^+$. Then

$$w_{i,q,t} = \frac{1}{N_v |\mathcal{Q}_i^+| |\mathcal{T}_{i,q}|}, \qquad \beta \in \arg\min_{\vartheta \in \mathbb{R}^5} \sum_{i,q,t} w_{i,q,t} \big(\gamma_{i,q,t} - \vartheta^\top x_{i,q,t}\big)^2. \tag{9}$$

All weights sum to one. Among minimizers, Eq. (4) selects the one with smallest Euclidean norm. Compute the thin SVD $W^{1/2}X = U\Sigma Q^\top$ in float64 and evaluate $\beta = Q\Sigma^\dagger U^\top W^{1/2} y$. A singular value is retained exactly when it exceeds $\epsilon_{\text{mach}} \max(N_s, 5)\sigma_{\max}$, with float64 machine precision ϵ_{mach} and largest singular value $\sigma_{\max}$. This fixes numerical rank without a regularization search. Features use their stated bounded scales. Record the retained rank and condition number on the retained subspace. If $N_v = 0$, set $\beta = 0$; the resulting nonpositive lower bounds disable early stopping.

A.5 Offline and online computation

Training, development, calibration, and evaluation splits are disjoint at the source-video level, including known duplicate videos and derived clips. Question sets follow a prespecified collection rule shared by calibration and evaluation. Ground-truth answers are unnecessary for fitting reference margins; task accuracy uses them at evaluation.

1. Fit. Construct the complete observation path for each training question. Score every prefix, construct its reference view, and compute γ_t for each early state. Form X, y, W and fit β by Eq. (4). Freeze the full specification before calibration.

2. Calibrate. Collect complete calibration trajectories, including prefixes that the fitted predictor would skip after stopping. Compute one maximum score R_i per video and the common correction ρ_α. Store (β, ρ_α) with the configuration.

3. Deploy. On a video's first question, construct its coarse image index; reuse that index for subsequent questions. Encode the question and candidate templates, precompute their similarities, and fix the observation path. At each accepted block, score E_t and save its winner. At $t < T$, compute x_t, $\widehat{\gamma}_t$, and L_t; return the winner immediately if $L_t > 0$. At T, return the current winner with a budget-termination flag. Reference views and reference scores are confined to offline computation.

Failure handling. If any required computation at a prefix fails, that prefix cannot trigger early stopping; processing advances along the same fixed path. At budget termination, return the most recent valid winner; if none exists, abstain with a processing-failure flag. Empty videos, fewer than two candidates, empty candidate tokenizations, and invalid embeddings (nonfinite entries, zero norm) are input failures. Failed training states are excluded with counts recorded. A failed calibration prefix contributes $-\infty$ to the maximum, matching its exclusion from early stopping. A missing reference for a video with eligible states gives that video score $+\infty$, while a video with no eligible states has score $-\infty$. All calibration videos are retained. Theorem 1 concerns well-defined reference scores; evaluations retain reference failures in their denominators and count them as uncertifiable outcomes.

A.6 Computation accounting

Let $C_{\text{index}}(V)$ denote video-index construction time, $C_{\text{text}}(q, \mathcal{C})$ question and candidate encoding time, and $C_{\text{sim}}(q, \mathcal{C})$ similarity computation and block-ranking time. Let $C_\theta(E_t, M)$ denote the cost of the *single scoring invocation at step* t, including all M candidate branches at that step.

Let $C_{\mathrm{pool}}(t)$ and $C_{\mathrm{lin}}(t)$ denote feature construction and the affine-prediction/sign-test times. The total online time is $C_{\mathrm{index}} + C_{\mathrm{text}} + C_{\mathrm{sim}} + \sum_{t=1}^{\tau} C_\theta(E_t, M) + \sum_{t=1}^{\min(\tau,T-1)} [C_{\mathrm{pool}}(t) + C_{\mathrm{lin}}(t)]$. Every term is measured in the same time unit and includes its decoding, preprocessing, transfer, and execution overheads. End-to-end timing also captures synchronization and orchestration. The timing comparison uses first-use latency, charging fresh video-index construction to every query.

The fixed-budget baseline uses the same retrieved E_T, coarse index, candidate scoring, and within-call cache policy, with one scorer invocation. Its cost is $C_{\mathrm{index}} + C_{\mathrm{text}}(q) + C_{\mathrm{sim}}(q) + C_\theta(E_T, M)$, where the two question-only terms cover retrieval and exclude candidate embeddings used exclusively by SBH. SBH's cost depends on cumulative scoring across prefixes. Under the illustrative assumptions $|E_t| = bt$, no cross-prefix reuse, and scoring cost proportional to frame count, stopping at τ processes $b\tau(\tau+1)/2$ frame instances, versus bT for the single call. With $b = 8$ and $T = 8$, a full trajectory processes 288 frame instances versus 64; stopping at step four still processes 80. Thus even in this simplified model, savings require $\tau(\tau+1)/2 < T$ before accounting for halting overhead. These illustrative counts describe frame-processing work; end-to-end latency is measured separately.

Arithmetic complexity is distinct from measured time. After embedding extraction, similarities take $O(K(M+1)d)$ operations and ranking takes $O(K \log K)$. Direct remaining-block pooling takes $O(KM)$ per eligible prefix; prediction uses five coefficients. The thin-SVD fit takes $O(25N_s)$ operations. Offline model computation for training and calibration, including the expanded references, is accounted for separately from deployment cost.

A.7 Sampling units, population projection, and cost conventions

Throughout, logarithms are natural, $\|\cdot\|$ is the Euclidean or spectral norm as appropriate, and $A \preceq B$ means that $B - A$ is positive semidefinite.

A complete video object V includes its prespecified question set, all potential prefix features and scores, the reference view, and potential execution costs. All sets are finite, and all analyzed features, scoring calls, and reference scores are well defined. Thus every $t < T_q$ is an eligible state in this analysis; processing-failure rules remain as specified in Appendix A.5. Let $\mathcal{E}$ be the event that at least one eligible early state exists. Assume $\Pr(\mathcal{E}) > 0$ and sample $N \geq 1$ independent training videos from $P(\cdot \mid \mathcal{E})$. This is the law obtained by retaining eligible videos from an independent source-video sample. Conditional on the retained count and eligibility pattern, the retained objects are independent with this law. The training sample is independent of the n calibration videos and the test video, which are independent draws from the unconditional law P; videos without eligible states remain in these splits. They have maximum score $-\infty$, cannot stop early, and contribute zero excess cost.

For $V \in \mathcal{E}$, the exact within-video weighting is

$$\langle u \rangle_V = \frac{1}{|\mathcal{Q}_V^+|} \sum_{q \in \mathcal{Q}_V^+} \frac{1}{|\mathcal{T}_{V,q}|} \sum_{t \in \mathcal{T}_{V,q}} u_{q,t}, \qquad \mathcal{T}_{V,q} = \{1, \ldots, T_q - 1\}. \tag{10}$$

All its state weights are positive and sum to one. In population regression expressions only, $\mathbb{E}_+$ denotes expectation under $P(\cdot \mid \mathcal{E})$. Define

$$\begin{aligned} &\Sigma = \mathbb{E}_+ \langle x x^\top \rangle_V, \quad b = \mathbb{E}_+ \langle x\gamma \rangle_V, \qquad \beta_\star = \Sigma^\dagger b, \\ &\Sigma = U H U^\top, \qquad\quad z = H^{-1/2} U^\top x, \quad \theta_\star = H^{1/2} U^\top \beta_\star. \end{aligned} \tag{11}$$

Here $U \in \mathbb{R}^{5 \times r}$ has orthonormal columns spanning $\mathrm{range}(\Sigma)$ and $H \in \mathbb{R}^{r \times r}$ is positive definite. The intercept ensures $1 \leq r \leq 5$. The feature ranges in Eq. (3) imply $\|x\|^2 \leq 5$, while $\gamma \in [-1, 1]$.

Let κ be the essential supremum of $\|z\|^2$ over all eligible states. For each training video define

$$G_V = \langle z z^\top \rangle_V, \quad g_V = \langle z(\gamma - \beta_\star^\top x) \rangle_V, \quad G_V \preceq \Lambda I_r, \quad \|g_V\| \leq B, \quad \nu = \mathbb{E}_+ \|g_V\|^2. \tag{12}$$

States within a video may be arbitrarily dependent. The constants can satisfy $1 \leq \Lambda \leq \kappa$, $B \leq \sqrt{\Lambda} + \Lambda$, and $\nu \leq \Lambda$; smaller valid bounds sharpen the result. The clipped population projection is $f_\star(x) = \mathrm{clip}_{[-1,1]}(\beta_\star^\top x)$; the true reference margin may vary nonlinearly with the features.

For N independent eligible training videos, choose $\omega \in (0,1)$ and failure probabilities $\delta_G, \delta_g > 0$ with $\delta_f = \delta_G + \delta_g < 1$. Write $h(u) = (1+u)\log(1+u) - u$, and let kl denote binary relative entropy. Define

$$d_\Lambda(\omega) = \mathrm{kl}\left(\frac{1-\omega}{\Lambda} \,\middle\|\, \frac{1}{\Lambda}\right), \qquad b_N = \frac{\nu}{B}\, h^{-1}\left(\frac{B^2 \log((r+1)/\delta_g)}{N\nu}\right), \\ e_N = \frac{\sqrt{\kappa}\, b_N}{1-\omega} + \zeta, \qquad N d_\Lambda(\omega) \geq \log(r/\delta_G). \tag{13}$$

If $\Lambda = 1$, take $\omega = 0$ and $\delta_G = 0$, and omit the last condition; if $\nu = 0$, take $b_N = 0$. Exact least squares has $\zeta = 0$. A numerical implementation is covered when its clipped predictions differ from the exact fit by at most a deterministic ζ on the eligible feature support; Appendix D.11 states a sufficient condition for retaining the correct rank. Under these conditions, the essential supremum of $|\widehat{\gamma}(x) - f_\star(x)|$ over eligible states is at most e_N with probability at least $1 - \delta_f$. Appendix D.11 gives optional sharper score radii.

Test-question expectations use $V \sim P$ followed by a question drawn from a prespecified distribution on $\mathcal{Q}_V$, independent of model outputs. Equivalently, results hold for a within-video average with fixed nonnegative question weights summing to one. A video with no questions contributes zero. Index cost can be allocated across its questions by any fixed accounting convention, used identically in the compared costs. Videos are the independent sampling units for both regression and test averaging.

Reference-informed comparator and costs. For one question, let Δ_H be the difference between the largest and second-largest reference probabilities, and define

$$\tau_\star = \min\big(\{t < T : \gamma_t > 0\} \cup \{T\}\big), \qquad C(t) = C_0 + \sum_{j=1}^{t} c_j, \qquad D = C(T) - C(\tau_\star). \tag{14}$$

Here $c_j \geq 0$ is the complete incremental cost of executing prefix j, including its scoring call and, when $j < T$, its stopping test; C_0 includes the initial indexing and encoding costs. These potential costs may depend on the video and question. They are fixed along the prescribed trajectory, without depending on the chosen stopping rule. The comparator $\tau_\star$ is an offline diagnostic benchmark along the prescribed path, using the reference unavailable at deployment. Whenever $\gamma_t > 0$, the current winner equals the unique reference winner and $\gamma_t = \Delta_H$.

For each question $C(t) = C_0 + \sum_{j \leq t} c_j$ with finite, nonnegative costs. The choice $c_j = C_\theta(E_j, M) + \mathbf{1}\{j < T\}[C_{\mathrm{pool}}(j) + C_{\mathrm{lin}}(j)]$ matches Appendix A.6. The theorem requires only $\mathbb{E}D < \infty$ for excess cost. Comparisons involving expected total costs additionally assume those costs are integrable. These are additive potential-cost statements for the fixed protocol; a change in caching or execution policy requires its corresponding cost model.

B EXPERIMENTAL DETAILS

This protocol complements Appendix A. The preliminary study (Appendix B.6) uses the trajectory endpoint a_T, whereas this evaluation uses the expanded reference a_H.

B.1 BENCHMARKS, SPLITS, AND METRICS

Benchmarks. The core benchmarks are LongVideoBench's public validation split (Wu et al., 2024), Video-MME (Fu et al., 2025), the seven publicly labeled MLVU multiple-choice development tasks (Zhou et al., 2025), and LVBench (Wang et al., 2025a). All use visual-only input. Video-MME contains 2,700 questions on 900 videos; its long score uses the official 900-question duration group. MLVU uses task-macro M-Avg, excluding open-ended tasks and hidden-test scores. LVBench contains 1,549 questions over 103 videos, making video-level sample counts particularly important. Complementary transfer results use Video-MME-v2's complete four-question groups and official average, non-linear, and L1–L3 scores (Fu et al., 2026), EgoSchema's public 500-question subset (Mangalam et al., 2023), and MINERVA's 1,515 questions (Nagrani et al., 2025). Accuracy is the common metric for the latter two.

Data isolation. Each core dataset uses source-video partitions for fitting, development, calibration, and evaluation in proportions $40/10/20/30$. Seed 2026 permutes canonical source-video identifiers; the first three counts are rounded down and evaluation receives the remainder. All questions, groups, clips, and identifiable duplicates stay together, including across datasets. The split manifest records release hashes, identifiers, and video/question counts. Source-video identifiers are used to exclude pilot and preliminary-study videos from calibration and evaluation. Fitting estimates predictors, development freezes implementation choices and baseline thresholds, calibration computes the correction, and evaluation remains held out. Custom held-out scores are distinct from published full-benchmark scores.

For full-benchmark transfer, each target is excluded from fitting, development, and calibration. A separate predictor is fitted and calibrated on the remaining core datasets after removing source-side overlap with the target, then evaluated on the target's entire designated public split. Learned stopping baselines follow the same source-only selection. Target identifiers, input modality, and evaluator must match for a score comparison. Coverage under transfer is evaluated empirically. The additional controlled comparisons use within-dataset held-out videos on LongVideoBench, MLVU, and LVBench, with Qwen2.5-VL and $\alpha = 0.10$ (Table 7).

Accuracy, reference risk, and coverage. Official task accuracy scores the returned answer against ground truth, including budget terminations. Reference risk compares the returned answer with the model's high-budget prediction. For evaluation video i with prespecified question set $\mathcal{Q}_i$, define

$$S_{i,q} = \mathbf{1}[\tau_{i,q} < T_{i,q}], \qquad D_{i,q} = \mathbf{1}[a_{i,q,\tau_{i,q}} \neq a_{H,i,q}], \tag{15}$$

$$\widehat{J}_{\text{video}} = \frac{1}{N_V} \sum_{i=1}^{N_V} \mathbf{1}[\exists q \in \mathcal{Q}_i : S_{i,q} D_{i,q} = 1], \tag{16}$$

$$\widehat{C}_{\text{sim}} = \frac{1}{N_V} \sum_{i=1}^{N_V} \mathbf{1}[\forall q \in \mathcal{Q}_i,\ \forall t < T_{i,q} : \gamma_{i,q,t} \geq L_{i,q,t}]. \tag{17}$$

These video-level quantities align with Theorem 1. Early-stop frequency is $\widehat{Q} = N_Q^{-1} \sum_{i,q} S_{i,q}$, with $N_Q = \sum_i |\mathcal{Q}_i|$. Disagreement conditional on stopping follows the J/Q relation in Appendix D.2. Coverage uses complete held-out trajectories, including prefixes after the first stop. Table 5 reports calibration/evaluation video counts, correction size, coverage, risk, and early-stop frequency. Reference failures remain in the denominators as uncertifiable outcomes under Appendix A.

Components and uncertainty. The component study measures video-balanced retained-margin MAE, ρ_α, and positive-bound frequency, together with accuracy, reference risk, early stopping, and latency (Table 6). State-level summaries average within questions and then videos. The positive-bound frequency uses every eligible held-out prefix. Primary SBH accuracy, reference-risk, and

speedup intervals cover the six held-out dataset/backbone settings in Table 9. Paired accuracy and latency differences compare SBH with the endpoint, entropy, margin, and equal-feature controls on Qwen2.5-VL/Video-MME. Both analyses use 2,000 source-video bootstrap replicates and percentile 95% intervals, keeping all questions and paired method predictions together and recomputing official macro metrics and ratios within each replicate. Risk intervals use the binary video event as the sampling unit. All intervals describe evaluation uncertainty conditional on the frozen fitted and calibrated rule.

B.2 Comparison families and alignment

Stopping rules on a shared path. Every controlled method receives the same nested E_t, timestamps, candidate ordering, and distributions p_t. The endpoint scores E_T once; the sequential no-halting control exposes repeated-scoring overhead. Candidate entropy and confidence count adapt confidence criteria from AdaptToken (Qi et al., 2026); margin follows CALM's softmax-response criterion (Schuster et al., 2022); persistent answer applies PABEE's unchanged-prediction criterion (Zhou et al., 2020). The count and persistence lengths are chosen from $\{1, 2, 3\}$ on development videos. Scalar thresholds use 101 development empirical quantiles plus never-stop. A single operating point maximizes development accuracy under a mean-latency cap of 0.75 times the single-call endpoint, with ties favoring lower cost and then a fixed grid index. An unattainable cap is marked infeasible; actual held-out latency is reported even when the cap is missed. SBH uses $\alpha = 0.10$; the calibration study reports both this setting and 0.05 on Video-MME across all three backbones.

Equal-information controls and component ablations. The consistency classifier, motivated by CATS (Schuster et al., 2021), uses the same five features, video-balanced weights, and labels $\mathbf{1}[a_t = a_H]$. Logistic regularization is chosen from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ on development videos, with an unpenalized intercept. The TV control predicts $\mathrm{TV}(p_t, p_H)$ from the same features, clips to $[0, 1]$, and uses video-maximum split-conformal calibration (Lei et al., 2018) before applying $m_t > 2U_t$. Both targets use the same offline scores. On Qwen2.5-VL/Video-MME, removing u_t and substituting a training-only donor cue isolate its information contribution; each predictor is refitted and separately video-calibrated. Donor seeds 17, 29, and 43 are reported individually. Calibration controls retain the full fitted predictor and compare no correction with a pooled prefix-level correction. Their exact definitions and scope appear in Appendix C.

Verifier adaptations. AVP-style reflection (Wang et al., 2026) and a REVEAL-style rubric verifier (Yan et al., 2026) inspect only the shared prefix and common coarse index, using the same frozen backbone. Prompts, rubric-construction data, score scales, and threshold grids are frozen on fit/development videos; every additional model call is charged. Both verifiers are adapted to the shared observation path, with access limited to the observed prefix and common coarse index.

Native frame/token-selection comparisons. Table 1 retains native resource and scoring conditions for Uniform, AKS (Tang et al., 2025), FOCUS (Zhu et al., 2026), AdaReTaKe (Wang et al., 2025b), FlexSelect and FlexSelect-Lite (Zhang et al., 2025), and AdaptToken and AdaptToken-Lite (Qi et al., 2026). Published scores and supplementary measurements are distinguished in Appendix B.8. Final frames, retained tokens, context/cache caps, selector input costs, and additional selector training are different resource quantities. Consequently, native accuracy differences provide system-level context; the shared-path tables support controlled stopping comparisons.

B.3 Backbones, implementation, and resource accounting

Model and processor configurations. The frozen checkpoints are `Qwen/Qwen2.5-VL-7B-Instruct`, `llava-hf/llava-onevision-qwen2-7b-ov-hf`, and `OpenGVLab/InternVL2_5-8B`. The Qwen/SigLIP configuration follows Appendix B.4, including Transformers 4.51.3, recorded revisions, 448-pixel fine frames, and 224-pixel coarse frames. LLaVA-OneVision uses its native video pathway and pooling; InternVL2.5 uses one 448-pixel tile per frame without an extra thumbnail. Chronological frames, timestamps, processor resolution, pooling, chat template, tokenizer, and context settings remain fixed within each backbone's controlled comparisons. The complete 128-frame reference, question, options, and longest candidate must fit without truncation. Dependency versions, native positional scaling, and device maps are fixed on fit/development data

before fitting/calibration; configurations are feasible only if the complete reference fits for every example.

SBH settings. Defaults are $K = 32$ coarse blocks, $b = 8$ fine frames per accepted block, $T_{\max} = 8$, and reference budget $B_H = 128$. Timestamp mapping, duplicate handling, deterministic ties, and reference completion follow Appendix A.3. Candidate scoring uses length-normalized teacher-forced answer-text log-likelihood with temperature one, excluding label and end-of-sequence tokens. VLM inference uses bfloat16; likelihood accumulation and SigLIP use float32; weighted least squares uses float64 and the thin-SVD rank tolerance in Appendix A.4. SBH has no ridge penalty, optimizer, learning rate, or training epochs. Fits and corrections are separate for each backbone and dataset protocol. The stopping threshold is strictly zero. An out-of-range conformal index yields infinite correction and disables early stopping, as specified in Eq. (5).

Hardware and timing. The resource pool comprises eight NVIDIA A100 GPUs with inference batch size one; the reported online comparisons allocate one GPU per run. The run manifest records GPU memory capacity, interconnect, CPU, software, and device allocation. Timed methods share a fixed allocation within each backbone, without unrelated concurrent jobs. After 30 fit/development warm-up queries, three synchronized executions per evaluation query measure latency from a locally available compressed video to the returned answer, excluding weight loading and network download. Query means are aggregated for end-to-end latency; speedup divides the single-call endpoint mean by the method mean on identical questions. All reported cost comparisons use first-use indexing: every timed query is charged fresh video-index construction.

Cost accounting. Table 8 separates decoding/indexing, cumulative candidate scoring, and all remaining online work (question/candidate encoding, retrieval, feature/stopper computation, transfers, and orchestration), with nonoverlapping wall-clock attribution and synchronized totals. It also records processed fine-frame instances, scoring invocations, device counts, and offline trajectory/reference scoring and fit/calibration cost. Expanded references remain offline. Eight 8-frame increments without cross-prefix reuse process $8\sum_{t=1}^{8} t = 288$ fine-frame instances before candidate-scoring multiplicity, versus 64 for a single endpoint call. Cumulative scoring work therefore determines the cost of repeated observation. Statistical analysis uses prefix replay; latency measurements execute the stopping policy. Invalid inputs and scoring failures follow Appendix A and remain in the reported denominators.

B.4 Encoders, prompts, and image processing

The reference configuration fixes Transformers 4.51.3, the SigLIP checkpoint `siglip-base-patch16-224` at revision `7fd15f0`,[1] and the frozen scorer `Qwen2.5-VL-7B-Instruct` at revision `cc59489`.[2] Both checkpoints use evaluation mode. SigLIP runs in float32; the answer model uses bfloat16 weights and activations, with float32 log-softmax and score accumulation. No sampling is involved in candidate scoring.

Coarse embeddings. Convert each representative frame to RGB, resize bicubically to 224×224, rescale to $[0, 1]$, and normalize each channel with mean 0.5 and standard deviation 0.5. Use the checkpoint's pooled image and text features, normalized to unit Euclidean norm; similarities are their dot products without learned logit scaling. The ranking template is exactly `Question: {q}` and the candidate template is exactly `Question: {q} Answer: {c}`. Use the checkpoint tokenizer and its 64-token input length, terminal token, and padding conventions. To preserve both payloads in long candidate templates, tokenize the fixed labels and payloads separately without special tokens, with one leading space on each payload and on `Answer:`. While their concatenation exceeds 63 tokens, remove the last token of the longer payload, choosing the question on equal lengths. Append the terminal token and pad to 64. For the question-only template, right-truncate its payload to the same limit. The embeddings therefore use a deterministic truncation rule without removing either fixed label.

[1] https://huggingface.co/google/siglip-base-patch16-224/tree/7fd15f0

[2] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct/tree/cc59489

Fine-frame inputs. Represent each observed frame as an image with its actual timestamp. For an RGB frame of width w and height h, set $s = 448/\max(w, h)$, resize bicubically to $(\lfloor sw + \frac{1}{2} \rfloor, \lfloor sh + \frac{1}{2} \rfloor)$, clamping each dimension to at least one pixel, and center the result on a 448×448 RGB canvas with fill value $(127, 127, 127)$. Assign any odd padding pixel to the right and bottom. Disable further resizing in the checkpoint processor. Rescale by $1/255$ and use channel means $(0.48145466, 0.45782750, 0.40821073)$ and standard deviations $(0.26862954, 0.26130258, 0.27577711)$. The processor retains patch size 14, temporal patch size 2, and spatial merge size 2. These settings apply identically to all prefixes and reference views.

Complete scoring prompt. Use the checkpoint's chat template with one system message and one user message, followed by its assistant generation prefix. The system message is:

```
You answer multiple-choice questions about video frames.
Select the best-supported candidate.  Output its answer text
only.
```

The user message interleaves one timestamp text item and one image item for every frame, then appends the question and candidates:

```
Time:  {timestamp in seconds, three decimal places} s
{image item}
...
Question:  {q}
Candidates:
1.  {c_1}
...
M. {c_M}
Answer with the candidate text.
```

The ellipses denote repetition of the indicated items and are absent from the actual prompt. Candidate numbering follows the input order. Tokenize each complete candidate text without additional special tokens, append these tokens to the shared assistant prefix, and teacher-force their likelihoods. Exclude candidate labels, the end-of-sequence token, and the prompt tokens from Eq. (1). Cache the shared prompt within a scoring call across candidate branches; rebuild the prompt for every observation prefix, with no cross-prefix cache reuse. Previous answers never enter subsequent prompts.

B.5 Quantities for assessing the stopping mechanism

Calibration validity and computational utility have different observables. Table 5 relates ρ_α, simultaneous coverage, early-stop disagreement, and early-stop frequency for the listed settings. The component study (Table 6) additionally measures the fraction of eligible prefixes with $L_t > 0$ using complete held-out trajectories, including states after the first stop. Questions without eligible prefixes have early-stop indicator zero. Disagreement conditional on stopping has the J/Q interpretation given in Appendix D.2. Task accuracy and executed cost are evaluated in their corresponding controlled comparisons.

Two controls isolate the remaining-video cue: removing u_t and replacing it with a random cross-video donor value. The donor pool contains only training-video cues, grouped by candidate count and observation fraction; each replacement is sampled uniformly from its matching group, excluding the recipient video. An empty donor group supplies zero and is counted. The pool is frozen before calibration, and donor draws are independent across videos and independent of targets, so calibration and evaluation videos never exchange features. Each control refits its predictor and computes its own correction using the same video partitions and risk level. Held-out retained-margin error, correction size, positive-bound frequency, and latency quantify the cue's incremental value.

A complementary target control fits D_t from the same features, clips its prediction to $[0, 1]$, and calibrates the maximum of $D_t - \widehat{D}_t$ within each video using Eq. (5). Its upper bound is $U_t^{\mathrm{TV}} = \mathrm{clip}_{[0,1]}(\widehat{D}_t + \rho_\alpha^{\mathrm{TV}})$, with value one for an infinite correction, and its stopping rule is $m_t > 2U_t^{\mathrm{TV}}$.

B.6 PRELIMINARY STUDY: PROTOCOL AND STATISTICAL DETAILS

B.6.1 OBSERVATION TRAJECTORIES AND RECORDED QUANTITIES

The diagnostic uses a frozen video QA model along a fixed observation trajectory. Each video is partitioned into 32 coarse blocks ranked by question-only image–text similarity. One block is revealed per step, with eight fine frames per selected block and a cap of eight steps. Every trajectory is evaluated to its endpoint, with no learned stopping predictor, calibration, or candidate-conditioned remaining-video feature. The analyzed records contain 720 video–question trajectories, all with $T = 8$ and exactly one record for each intermediate step $t = 1, \ldots, 7$. Thus, the analysis includes all $720 \times 7 = 5{,}040$ intermediate states; the endpoint itself is excluded because its disagreement is zero by definition. The sampling units here are video–question trajectories, which may share a source video.

Let p_t be the candidate probability distribution and $a_t = \arg\max_c p_t(c)$ the selected answer. The recorded confidence gap is

$$g_t = p_t(a_t) - \max_{c \neq a_t} p_t(c), \tag{18}$$

and the recorded outcome is

$$F_t = \mathbf{1}\{a_t \neq a_T\}. \tag{19}$$

The comparator is the endpoint a_T of the prescribed eight-step path; SBH uses the expanded-reference answer a_H. Endpoint disagreement measures answer stability. An answer that changes and then returns to a_t by step T has $F_t = 0$.

Each state record contains a trajectory identifier, t, T, progress t/T, $g_t \in [0, 1]$, the question-only unseen-relevance score $r_t^Q \in [0, 1]$, and $F_t \in \{0, 1\}$. The recorded candidate-agnostic covariate r_t^Q defines the relevance groups in the diagnostic. Observation follows the fixed path, and every trajectory runs to its endpoint. SBH's candidate-aware stopping feature u_t is defined separately in Equation (3).

B.6.2 AGGREGATION AND FIGURE CONSTRUCTION

Unsmoothed disagreement rates. All percentages in the main text are computed from the binary outcomes, without heatmap smoothing. Overall, 891 of 5,040 records have $F_t = 1$, yielding 17.7% disagreement. Restricting to $g_t \geq 0.50$ gives $361/3{,}503 \approx 10.3\%$; further restricting to $t \leq 3$, equivalently $t/T \leq 3/8$, gives $202/955 \approx 21.2\%$. Every eligible state receives equal weight. Because each trajectory contributes seven states, the overall rate also equals the mean of the 720 trajectory-level disagreement fractions. Conditioning on the gap can change the number of states contributed by each trajectory.

Confidence–progress atlas. Figure 2a uses a two-dimensional histogram with 29 equal-width gap bins from zero to the greater of 0.86 and the largest gap across all states, and seven step bins with edges $0.5, 1.5, \ldots, 7.5$. The vertical labels report progress $t/8$. Let N_{bt} and D_{bt} denote the state count and sum of disagreement indicators in histogram cell (b, t). We separately smooth these arrays with a Gaussian kernel G whose standard deviations are 1.15 gap bins and 0.55 step bins, using reflective boundaries and a four-standard-deviation truncation. The displayed rate is

$$\widehat{R}_{bt} = \frac{(G * D)_{bt}}{(G * N)_{bt}} \quad \text{when } (G * N)_{bt} \geq 1. \tag{20}$$

Cells below this smoothed-count threshold are gray. White contours mark the 55th, 75th, and 90th percentiles of positive smoothed counts, retaining distinct levels strictly within the observed count range. Color uses a fixed 0–100% scale; the dashed line marks $g_t = 0.50$. The surface visualizes the smoothed empirical disagreement rate.

Step–gap strata and relevance quartiles. Figure 2b uses separate, coarser gap bins: $[0, 0.30)$, $[0.30, 0.40)$, $[0.40, 0.50)$, $[0.50, 0.60)$, $[0.60, 0.70)$, and $[0.70, 1]$. A stratum s consists of states at the same step in one such interval. Within each stratum, the 25th and 75th percentiles of r_t^Q use linear interpolation between adjacent ordered values at fractional index $(n_s - 1)p$, for $p \in \{0.25, 0.75\}$. The low group contains scores at or below the 25th percentile; the high group contains scores at or

above the 75th percentile. Boundary ties are included, allowing group sizes to exceed one quarter of n_s. We retain a stratum only if $n_s \geq 120$, its two percentile thresholds are distinct, and both groups contain at least 20 states. No criterion uses the sign or magnitude of the outcome difference.

For retained strata $\mathcal{S}$, let $\widehat{p}_{s,L}$ and $\widehat{p}_{s,H}$ be the mean outcomes in the low and high groups, and define $\delta_s = \widehat{p}_{s,H} - \widehat{p}_{s,L}$. The aggregate comparison uses the full stratum sizes, including its middle-quartile states, as weights:

$$w_s = \frac{n_s}{\sum_{u \in \mathcal{S}} n_u}, \qquad \overline{p}_L = \sum_{s \in \mathcal{S}} w_s \widehat{p}_{s,L}, \qquad \overline{p}_H = \sum_{s \in \mathcal{S}} w_s \widehat{p}_{s,H}. \tag{21}$$

This standardized average weights each within-stratum rate by the full stratum size. The 20 eligible strata give $\overline{p}_L \approx 11.7\%$ and $\overline{p}_H \approx 23.8\%$, with a weighted difference of 12.1 percentage points. Eighteen δ_s are positive and two are zero; the unweighted median is 12.6 points. Rates and differences are computed before rounding and displayed to one decimal place. Rows are ordered by decreasing δ_s, breaking ties by step and then gap-bin index, and read down the left column before the right column of panel (b). Open circles and filled diamonds mark the low and high groups, respectively; the numeric columns show $100\delta_s$ in percentage points.

Intervals and interpretation. For a group with n states, empirical disagreement $\widehat{p}$, and $z = 1.96$, the plotted Wilson interval has center c and half-width h:

$$c = \frac{\widehat{p} + z^2/(2n)}{1 + z^2/n}, \qquad h = \frac{z\sqrt{\widehat{p}(1-\widehat{p})/n + z^2/(4n^2)}}{1 + z^2/n}, \tag{22}$$
$$I = [\max(0, c-h), \min(1, c+h)].$$

These nominal 95% binomial intervals describe individual group proportions. They use state counts with pointwise coverage and no source-video clustering adjustment. Each trajectory contributes at most one state to a given step–gap stratum; it can contribute to multiple rows, and several trajectories may share a source video. Stratification conditions on the step and a confidence-gap interval, allowing exact gaps and other covariates to vary within each stratum. The contrasts therefore describe an association with recorded unseen relevance. The contribution of SBH's candidate-aware feature is evaluated in Section 5.3.

B.7 Reading the main comparisons

Table 1 groups native pipelines by backbone. Baseline rows combine 32 published scores and 28 supplementary measurements, detailed below. Scores are reported separately for each dataset to retain their weighting and temporal coverage. Within each backbone and metric, bold and underlined values denote the best and second-best scores, respectively (ties receive the same formatting). Green subscripts beside SBH show its absolute gain in percentage points over the strongest baseline in that column. These differences describe performance under each system's native resource and scoring conditions.

Controlled stopping: settings and notation. Table 2 uses Qwen2.5-VL-7B on held-out Video-MME videos, with shared evidence prefixes E_t and common candidate scoring. Accuracy, video-level reference risk $\widehat{J}_{\text{video}}$, and early-stop frequency are percentages; time is mean end-to-end latency in seconds. Speedup divides the single-call endpoint's mean latency by each method's mean latency on the same questions and device allocation. Threshold baselines use the development operating point at the $0.75\times$ endpoint-latency cap; held-out costs determine the reported speedups. SBH retains $\alpha = 0.10$. Operating-point selection follows Appendix B.2.

Bold and underlined values mark the best and second-best distinct values in each column, with ties sharing the same formatting. Each baseline subscript is the signed difference *baseline minus SBH*, in percentage points for the first three columns, seconds for time, and multiplicative-factor units for speedup. Red indicates a less favorable value in the indicated direction; green indicates a more favorable value. Early-stop frequency is descriptive and must be read jointly with accuracy and risk. The endpoint and sequential controls never stop early, so their zero reference risk is structural.

Baseline families and citation scope. The endpoint scores E_T once; the sequential control additionally exposes repeated-scoring overhead. Both are internal reference controls. Candidate entropy and confidence count adapt AdaptToken's confidence criteria (Qi et al., 2026); candidate margin follows the softmax-response criterion in CALM (Schuster et al., 2022). Persistent answer applies PABEE's unchanged-prediction criterion across evidence steps (Zhou et al., 2020). The equal-feature classifier is motivated by consistency prediction in CATS (Schuster et al., 2021), while the TV control applies split-conformal regression calibration (Lei et al., 2018). These baselines adapt the cited stopping principles to the shared observation path and scorer. AVP/REVEAL rows are shared-prefix verifier adaptations (Wang et al., 2026; Yan et al., 2026); all additional calls are charged. Their scores come from the shared-prefix evaluation. The offline expanded reference, video-level risk denominator, and source-video bootstrap units follow Appendix B.1.

B.8 Score provenance and resource conditions

Table conventions. All columns in Table 1 use visual-only input. Video-MME Long uses the official 900-question subset; MLVU uses seven-task development M-Avg. Native baselines retain their original scorer and budgets. Budget units are f (final fine frames), t (retained visual tokens), and c (context/cache cap); each describes a different resource, with selector input costs additional. SBH also uses a 32-block coarse index. The paragraphs below distinguish published scores from supplementary measurements. Uniform and the published AKS values follow FOCUS's reproduction; AKS's method citation refers to Tang et al. (2025).

AdaReTaKe. The published Qwen2.5-VL-7B values in Table 1 come from Table 1 of Wang et al. (2025b): Video-MME overall/long, MLVU development M-Avg, and LVBench. The public configuration uses a 2,048-frame input cap and a 16K context/cache budget. Its Video-MME configuration selects the annotation file without subtitles, and the evaluator groups the official duration labels. The MLVU evaluator separately computes task-macro and pooled accuracy; the table uses the former. The LongVideoBench entry (60.8) is a supplementary visual-only measurement; the released builder includes subtitles. The paper states 2-fps sampling, and the released Video-MME configuration uses 4 fps. Published scores retain these native preprocessing conditions.

FlexSelect. Table 1 of Zhang et al. (2025) supplies all scores except InternVL LongVideoBench. The native maximum input is 1,024 frames for Qwen2.5-VL and 512 for InternVL2.5, with 64-frame selection groups. Appendix A.3 specifies retained budgets of 7,010 and 8,256 visual tokens, respectively; the released Qwen script specifies 7,040. Table 1 uses the paper's budget label. Qwen's LongVideoBench script uses visual-only input; InternVL's includes subtitles. The InternVL entries (59.8/59.4) are supplementary visual-only measurements. Both use the development MLVU task and the visual-only Video-MME task. FlexSelect-Lite trains a lightweight selector on video–text data while keeping its answering backbone frozen.

AdaptToken. The LVBench values are transcribed from Table 1 of Qi et al. (2026). Appendix B extends context limits by $16\times$, allowing up to 4,096 input frames for Qwen2.5-VL and 1,024 for InternVL2.5, with 7,010/8,256 retained visual tokens. The reported LMMS-Eval results leave the Video-MME/LongVideoBench subtitle variants and MLVU split and aggregation unspecified, and omit Video-MME long scores. All Video-MME, LongVideoBench, and MLVU entries are supplementary measurements under the conditions in Table 1. Native AdaptToken computes entropy from generated tokens and reports H100 timings. Our stopping adaptation uses candidate entropy, with latency measured separately on A100.

FOCUS and AKS. Tables 1 and 5 of Zhu et al. (2026) provide the visual-only, 32-final-frame LLaVA-OneVision-7B results for Uniform, FOCUS, and their AKS reproduction on Video-MME overall and LongVideoBench validation. The Video-MME long, MLVU, and LVBench entries are supplementary measurements. Native retrieval uses BLIP-based relevance; AKS's candidate prefilter and FOCUS's timeline search incur different selection costs despite equal final frame counts. The reported LongVideoBench score covers the full validation split.

C Additional Experimental Results

C.1 Additional backbones and transfer evaluation

On LLaVA-OneVision and InternVL2.5, SBH reaches 58.64% and 67.28% accuracy with $1.34\times$ and $1.38\times$ speedups, respectively (Table 3). Their observed video-level reference risks are 6.30% and 4.44%. Table 4 reports Video-MME-v2 average/non-linear scores of 57.5/49.6 for SBH, alongside 55.2% on EgoSchema and 46.9% on MINERVA. Each transfer target is excluded from fitting, development, and calibration; target groups remain intact and target coverage is empirical. The tables report point estimates; metric definitions appear in Appendix B.1, and the supporting calibration, component, and cost tables appear in Appendix C.

Table 3: **Controlled stopping across backbones.** The Video-MME held-out partition and stopping rules match Table 2. Accuracy and $\widehat{J}_{\text{video}}$ are percentages; speedup uses each backbone's own single-call endpoint baseline.

	LLaVA-OneVision-7B			InternVL2.5-8B		
Stopping rule	Acc.↑	$\widehat{J}_{\text{video}}$↓	Speedup	Acc.↑	$\widehat{J}_{\text{video}}$↓	Speedup
Endpoint, scored once	57.90	0.00	1.00×	66.67	0.00	1.00×
Sequential, no early exit	57.90	0.00	0.29×	66.67	0.00	0.32×
Candidate entropy	57.28	13.33	1.16×	66.05	10.74	1.23×
Candidate margin	57.53	11.85	1.20×	66.30	10.00	1.26×
Persistent answer	57.04	14.44	1.14×	65.93	12.22	1.19×
Confidence count	57.41	12.59	1.21×	66.17	10.37	1.28×
Equal-feature classifier	58.15	9.26	1.27×	66.91	7.41	1.31×
Calibrated TV bound	58.02	3.33	0.84×	66.79	2.59	0.90×
AVP-style verifier	57.78	8.89	1.08×	66.67	6.67	1.16×
REVEAL-style verifier	58.27	7.41	1.14×	67.04	5.93	1.22×
SBH (ours)	**58.64**	6.30	**1.34**×	**67.28**	4.44	**1.38**×

Table 4: **Transfer to recent and complementary benchmarks.** Qwen2.5-VL-7B, visual-only input, source-only fitting/calibration, and the common candidate scorer. Video-MME-v2 uses its complete public question groups and official evaluator; EgoSchema uses the public 500-question subset. All scores are percentages.

	Video-MME-v2					Ego	MINERVA
Method	Avg.	Nonlin.	L1	L2	L3	Acc.	Acc.
Endpoint, scored once	56.4	47.9	65.9	56.2	46.7	53.4	45.5
Candidate entropy	55.9	46.8	65.0	55.8	46.3	52.8	44.9
Equal-feature classifier	56.8	48.5	66.2	56.8	47.1	54.2	46.1
Confidence count	56.2	47.2	65.2	56.0	46.7	53.0	45.3
AVP-style verifier	56.6	48.1	65.8	56.7	47.2	53.8	45.9
REVEAL-style verifier	56.9	48.7	66.4	56.9	47.5	54.4	46.3
SBH (ours)	**57.5**	**49.6**	**67.1**	**57.6**	**47.9**	**55.2**	**46.9**

Avg./Nonlin. denote official average/non-linear scores; L1–L3 are official level scores. All methods are evaluated under the adaptations defined in Table 2. Values are point estimates under the stated transfer protocol.

C.2 CALIBRATION AND COMPONENT CONTROLS

All results follow Appendix B; rates are percentages. Bold/underlining mark the best/second-best distinct error, coverage, risk, accuracy, and cost in each comparison, including ties. Correction size and positive-bound/early-stop frequencies are descriptive.

Table 5 reports calibration/evaluation video counts (n_{cal}/N_V), coverage, and risk. Observed coverage exceeds $1-\alpha$ and risk remains below α throughout. On Video-MME, $\alpha = 0.05$ lowers risk to 1.85–2.96% and early-stop frequency to 80.86–84.07%. The table reports held-out rates for each fitted rule; Theorem 1 gives marginal coverage over calibration and new videos.

Table 5: **Calibration, simultaneous coverage, and early stopping.**

Setting / SBH backbone	α	n_{cal}/N_V	ρ_α	$\widehat{C}_{\mathrm{sim}}$	$\widehat{J}_{\mathrm{video}}$	Early
Video-MME: separate fit and calibration per backbone						
Qwen2.5-VL-7B	0.10	180/270	0.214	93.33	5.19	92.35
	0.05	180/270	0.271	96.30	2.22	83.21
LLaVA-OneVision-7B	0.10	180/270	0.236	92.22	6.30	90.74
	0.05	180/270	0.298	95.56	2.96	80.86
InternVL2.5-8B	0.10	180/270	0.207	94.07	4.44	91.73
	0.05	180/270	0.263	97.04	1.85	84.07
Other core datasets: SBH with Qwen2.5-VL-7B						
LongVideoBench	0.10	150/227	0.226	93.39	5.29	90.72
MLVU-MCQ	0.10	224/338	0.198	94.08	4.73	91.28
LVBench	0.10	20/32	0.284	90.63	6.25	86.37

Table 6 retains the complete numerical results underlying Figures 4 and 5. It fixes Qwen2.5-VL, Video-MME partitions, observation path, scorer, and reference. Panel A separately refits/calibrates each variant at $\alpha = 0.10$; donor replacement follows Appendix B.5. Removing u_t raises video-balanced retained-margin MAE from 0.084 to 0.103 and latency from 7.74 to 8.05 seconds; all donor seeds further increase both. Pos. uses every eligible held-out prefix.

Table 6: **Component ablations on Qwen2.5-VL / Video-MME.**

A. Remaining-video information: separate video-level calibration

Variant	MAE↓	ρ_α	Pos.	$\widehat{J}_{\mathrm{video}}$	Early	Acc.	Time (s)
SBH (full)	**0.084**	0.214	42.63	**5.19**	92.35	**68.02**	**7.74**
SBH without u_t	<u>0.103</u>	0.251	35.91	<u>5.93</u>	88.15	<u>67.78</u>	<u>8.05</u>
SBH with donor (17)	0.108	0.264	33.82	6.67	86.91	67.53	8.17
SBH with donor (29)	0.105	0.257	34.67	6.30	87.65	67.65	8.11
SBH with donor (43)	0.110	0.269	32.95	7.04	86.30	67.41	8.22

B. Calibration unit: same full predictor and zero stopping threshold

Variant	ρ	$\widehat{C}_{\mathrm{sim}}$	Pos.	$\widehat{J}_{\mathrm{video}}$	Early	Acc.	Time (s)
SBH (video maximum)	0.214	**93.33**	42.63	**5.19**	92.35	**68.02**	7.74
SBH without correction	0.000	34.44	62.47	14.44	98.15	66.91	**6.82**
SBH with pooled prefixes	0.069	<u>72.22</u>	54.18	<u>10.37</u>	96.42	<u>67.41</u>	<u>7.11</u>

Panel B freezes the full predictor and tests $L_t > 0$. Without correction, $\rho = 0$. Pooled-prefix calibration applies the nonnegative finite-sample quantile rule to individual residuals $\widehat{\gamma} - \gamma$, using their pooled count and an infinite correction for an out-of-range index. Pooling treats each prefix residual as a calibration unit, while video-maximum calibration accounts jointly for dependent prefixes. Without correction and with pooling, observed coverage falls to 34.44% and 72.22%, respectively, while risk rises to 14.44% and 10.37%, versus SBH's 5.19%.

C.3 CONTROLLED GENERALIZATION ACROSS DATASETS

Table 7 extends the shared-path comparison to three within-dataset held-out partitions. All rows use Qwen2.5-VL, visual-only input, common candidate scoring, a 64-frame maximum observation budget, and a 128-frame reference. Each dataset has disjoint fitting, development, calibration, and evaluation videos. SBH and calibrated TV use $\alpha = 0.10$; threshold selection follows Appendix B.2. Score is official accuracy except for MLVU's seven-task M-Avg. The endpoint has zero early-stop risk because it always reaches the budget.

Table 7: **Controlled stopping beyond Video-MME.**

Stopping rule	Score↑	$\widehat{J}_{\text{video}}$↓	Early	Time (s)↓	Speedup↑
LongVideoBench (accuracy)					
Endpoint, scored once	59.74	**0.00**	0.00	12.18	1.00×
Candidate entropy (Qi et al., 2026)	59.18	11.01	86.84	10.20	1.19×
Candidate margin (Schuster et al., 2022)	59.42	9.69	85.09	9.91	1.23×
Equal-feature classifier (Schuster et al., 2021)	59.89	7.05	89.03	9.42	1.29×
Calibrated TV (Lei et al., 2018)	59.82	3.08	54.11	13.84	0.88×
SBH (ours)	**60.31**	5.29	90.72	**8.89**	**1.37**×
MLVU-MCQ (M-Avg)					
Endpoint, scored once	74.91	**0.00**	0.00	9.76	1.00×
Candidate entropy (Qi et al., 2026)	74.32	10.36	87.96	8.14	1.20×
Candidate margin (Schuster et al., 2022)	74.56	9.17	86.55	7.94	1.23×
Equal-feature classifier (Schuster et al., 2021)	75.04	6.51	89.87	7.57	1.29×
Calibrated TV (Lei et al., 2018)	75.13	2.96	58.04	11.48	0.85×
SBH (ours)	**75.48**	4.73	91.28	**7.12**	**1.37**×
LVBench (accuracy)					
Endpoint, scored once	54.62	**0.00**	0.00	13.46	1.00×
Candidate entropy (Qi et al., 2026)	53.91	12.50	82.06	11.51	1.17×
Candidate margin (Schuster et al., 2022)	54.11	9.38	80.90	11.22	1.20×
Equal-feature classifier (Schuster et al., 2021)	54.83	6.25	84.95	10.51	1.28×
Calibrated TV (Lei et al., 2018)	54.76	3.13	46.71	15.47	0.87×
SBH (ours)	**55.21**	6.25	86.37	**10.04**	**1.34**×

SBH has the highest score in each dataset block and gives 1.34–1.37× endpoint speedup, with observed risk of 4.73–6.25%. Calibrated TV combines lower risk among early-stopping rules with latency above the endpoint. These within-dataset comparisons remain distinct from native full-benchmark results (Table 1) and source-only transfer (Table 4).

With 32 evaluation videos, LVBench has a risk interval of [0.00, 15.63]% (Table 9), spanning values on both sides of the 10% target.

C.4 EXECUTED COST AND STATISTICAL UNCERTAINTY

Table 8 uses one A100 per online run. Fine inst. counts fine-frame inputs across scoring calls before candidate-branch multiplicity; Score includes every branch. Index, Score, and Other partition latency. Including all online work, SBH gives 1.34–$1.40\times$ endpoint speedup.

Table 8: **Executed latency and computation accounting.**

A. First-use online cost per query

Method	Fine inst.	Calls	Index (s)	Score (s)	Other (s)	Total (s)
Qwen2.5-VL-7B						
Endpoint, scored once	64.0	1.00	0.83	9.48	0.53	10.84
Sequential, no early exit	288.0	8.00	0.83	33.35	0.78	34.96
SBH (ours)	52.4	2.61	0.83	6.13	0.78	7.74
LLaVA-OneVision-7B						
Endpoint, scored once	64.0	1.00	0.76	9.91	0.63	11.30
Sequential, no early exit	288.0	8.00	0.76	37.24	0.97	38.97
SBH (ours)	57.8	2.72	0.76	6.78	0.89	8.43
InternVL2.5-8B						
Endpoint, scored once	64.0	1.00	0.89	10.57	0.67	12.13
Sequential, no early exit	288.0	8.00	0.89	36.04	0.98	37.91
SBH (ours)	54.1	2.66	0.89	6.93	0.97	8.79

B. SBH allocation and offline cost for the Video-MME fit

Backbone	Online GPUs	Offline scoring (GPU-h)	Fit + calibration (s)
Qwen2.5-VL-7B	1	26.7	1.36
LLaVA-OneVision-7B	1	29.8	1.44
InternVL2.5-8B	1	28.5	1.39

Table 9 uses 2,000 source-video bootstrap replicates at $\alpha = 0.10$. Paired differences (SBH minus comparator) support latency savings: all latency intervals are below zero. Accuracy intervals span zero, leaving the direction of the accuracy differences uncertain.

Table 9: **Uncertainty in SBH performance and paired differences.**

A. SBH estimates with video-bootstrap 95% intervals

Setting	Score [95% CI]	$\widehat{J}_{\text{video}}$ [95% CI]	Speedup [95% CI]
Video-MME / Qwen	68.02 [64.94, 71.11]	5.19 [2.59, 8.15]	1.40 [1.35, 1.45]
Video-MME / LLaVA	58.64 [55.31, 61.98]	6.30 [3.33, 9.63]	1.34 [1.29, 1.39]
Video-MME / InternVL	67.28 [64.20, 70.37]	4.44 [1.85, 7.41]	1.38 [1.33, 1.43]
LongVB / Qwen	60.31 [55.61, 64.89]	5.29 [2.20, 8.81]	1.37 [1.31, 1.44]
MLVU / Qwen	75.48 [72.42, 78.47]	4.73 [2.37, 7.10]	1.37 [1.32, 1.42]
LVBench / Qwen	55.21 [49.92, 60.43]	6.25 [0.00, 15.63]	1.34 [1.23, 1.47]

B. Paired differences: Qwen2.5-VL / Video-MME

Comparison	ΔAcc. (pp) [95% CI]	ΔTime (s) [95% CI]
SBH vs. endpoint	+0.37 [-0.74, 1.48]	-3.10 [-3.48, -2.74]
SBH vs. entropy	+1.11 [-0.12, 2.35]	-1.08 [-1.32, -0.84]
SBH vs. margin	+0.86 [-0.25, 1.98]	-0.92 [-1.17, -0.68]
SBH vs. equal-feature classifier	+0.61 [-0.49, 1.73]	-0.47 [-0.68, -0.26]

C.5 QUALITATIVE ILLUSTRATIONS

Figures 7 and 8 present four cases of SBH's decision logic. The scenes are AI-generated, and the questions, candidate answers, responses, and signs of L_t are manually specified.

Reading the cases. Each pair contrasts an initial prefix with a later accumulated prefix on a fixed observation path. Only representative frames are displayed; all earlier evidence remains in the later prefix. Yellow borders mark evidence discussed in each case. The reference answer is an offline comparator and is unavailable to the online stopping rule. The budget is $T = 8$.

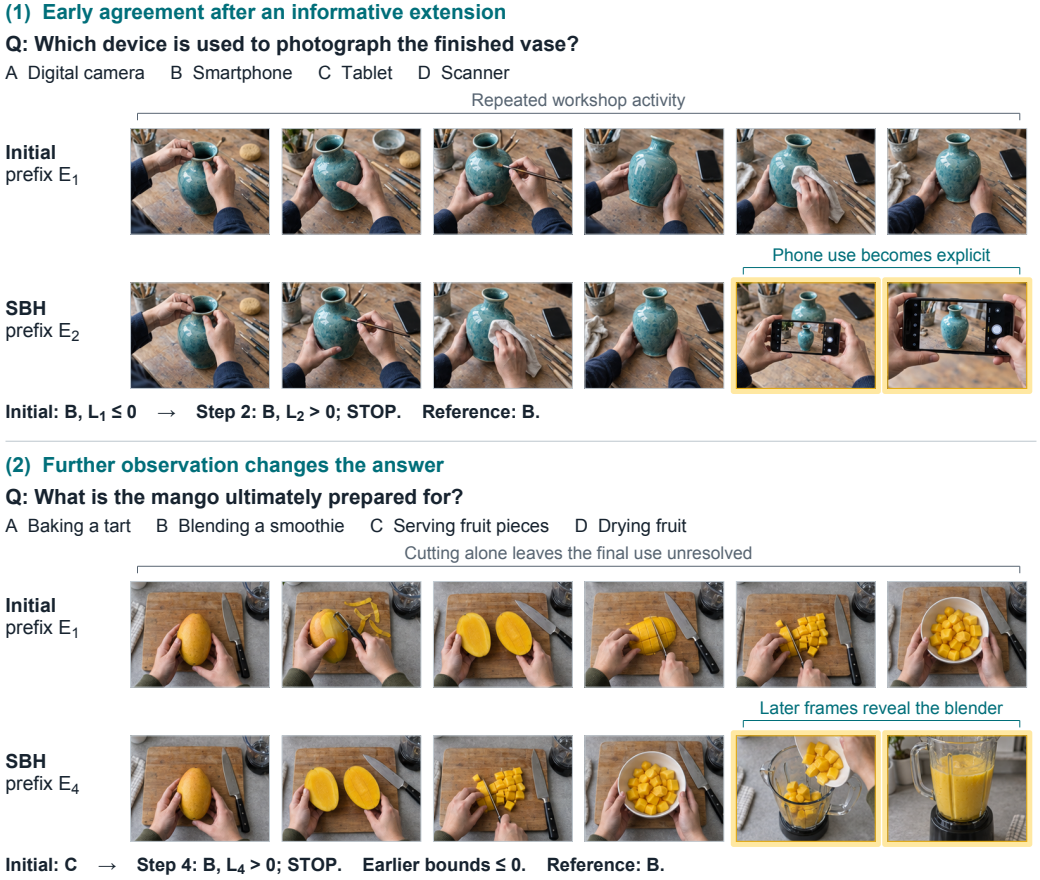

Figure 7: **Early stopping.** (1) Explicit phone use supports answer B after an informative extension. (2) Cutting mango initially supports several possible purposes; observing the blender changes the answer from C to B. Both traces stop at their first positive lower bound according to Eq. (6).

C.6 BUDGET TERMINATION AND REFERENCE ERRORS

The remaining cases illustrate budget termination and reference errors. In the packing example, all eligible bounds are nonpositive, so the rule reaches its budget and returns the current answer. In the tea case, the early answer agrees with an erroneous reference. The white crystals in the unmarked jar leave the ingredient ambiguous between salt and sugar.

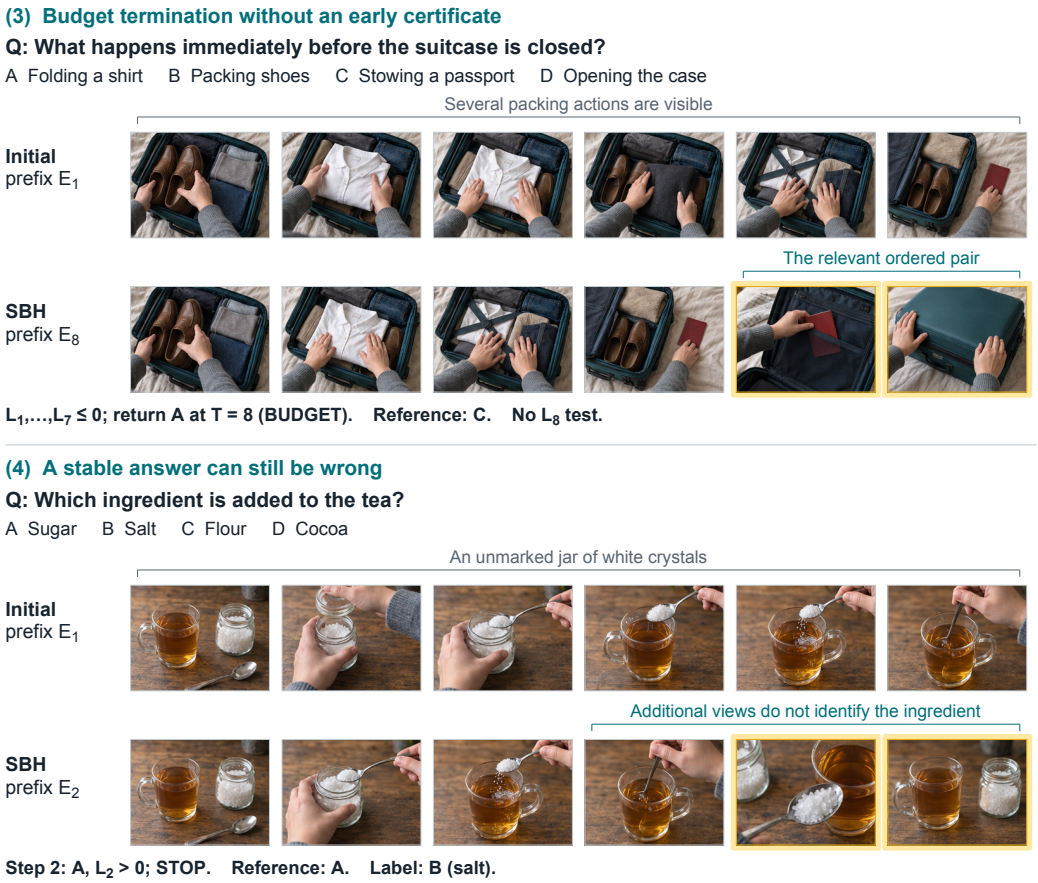

Figure 8: **Budget termination and reference errors.** (3) The budget answer disagrees with its reference even though the relevant frames are shown. There is no lower-bound test at T. (4) Agreement with an erroneous reference can coexist with a positive bound. The salt/sugar error separates reference consistency from correctness, as in Theorem 1.

D PROOFS

D.1 PROOF OF THE EXACT MARGIN COMPARISON

Proof. Let $\delta = p_H - p_t$. Since $\mathbf{1}^\top \delta = 0$, the positive coordinates of δ sum to D_t, as do the magnitudes of its negative coordinates. Hence $|\delta(c)| \le D_t$ for every candidate. For each $c \neq a_t$,

$$p_H(a_t) - p_H(c) = p_t(a_t) - p_t(c) + \delta(a_t) - \delta(c) \ge m_t - 2D_t. \tag{23}$$

Taking the minimum proves the bound. With $p_t = (0.6, 0.3, 0.1)^\top$ and $p_H = (0.6, 0.1, 0.3)^\top$, both margins equal 0.3 while $D_t = 0.2$, giving strict inclusion. Moving mass ϵ from the winner to its nearest competitor decreases their gap by 2ϵ at total-variation distance ϵ, so the factor two is sharp. The inclusion compares exact targets; prediction and calibration errors determine the behavior of learned rules. □

D.2 PROOF OF THEOREM 1

Let $\mathcal{F}$ contain the training data and every choice frozen before calibration: β, feature definitions, models, prompts, observation rules, budgets, and reference construction. Conditional on $\mathcal{F}$, applying the same maximum-residual map to exchangeable complete video objects $(V_i, \mathcal{Q}_i)$ yields exchangeable scores $R_1, \dots, R_{n+1}$. These scores may be extended real numbers; all dependence among a video's questions and prefixes is retained inside its score.

For $k \le n$, attach independent continuous tie breakers to the $n+1$ scores. The new video's lexicographic rank is uniform on $\{1, \dots, n+1\}$. Strict exceedance of the kth calibration score requires rank greater than k, giving the standard split conformal bound (Lei et al., 2018; Angelopoulos & Bates, 2023)

$$\Pr\big(R_{n+1} > R_{(k)} \mid \mathcal{F}\big) \le \frac{n+1-k}{n+1} \le \alpha. \tag{24}$$

Here $R_{(k)}$ is computed from the first n scores. If $k = n+1$, its prescribed value is $+\infty$ and strict exceedance is impossible. Since $\rho_\alpha \ge R_{(k)}$, with probability at least $1-\alpha$ every new-video state satisfies $\widehat{\gamma}_{q,t} - \gamma_{q,t} \le \rho_\alpha$. For finite ρ_α, this gives $\gamma_{q,t} \ge \widehat{\gamma}_{q,t} - \rho_\alpha$. Clipping the right-hand side to $[-1, 1]$ preserves the inequality because $\gamma_{q,t} \in [-1, 1]$. The infinite-correction convention $L_{q,t} = -1$ gives deterministic coverage. This proves the simultaneous lower-bound assertion.

On that coverage event, every early stop satisfies $\gamma_{q,\tau_q} \ge L_{q,\tau_q} > 0$, so the returned answer is the unique reference winner. An early disagreement on any question is therefore contained in the complement of the coverage event. Its probability is at most α, proving the second assertion. Integration over $\mathcal{F}$ gives the unconditional statement.

The probability weights videos equally; a question-weighted disagreement rate uses different weights when videos contain different numbers of questions. For a single question with early-stop probability $Q > 0$ and joint early-disagreement probability $J \le \alpha$, the conditional disagreement probability is $J/Q \le \min\{1, \alpha/Q\}$. Changing the question collection rule changes the object to which exchangeability applies. Recalibration after such a change requires the calibration and deployment collections to follow the same rule.

D.3 POPULATION PROJECTION AND CLUSTER MOMENTS

The proof of Theorem 2 combines estimation of a population linear projection, finite-video calibration, and the geometry of stopping. The concentration arguments below build on the classical two-matrix Golden–Thompson inequality. The definitions and sampling assumptions are given in Appendix A.7.

Lemma 1 (Support, normal equations, and cluster moments)**.** *All eligible states of almost every video lie in* $\mathrm{range}(\Sigma)$*. With* $\xi = \gamma - z^\top \theta_\star$*,* $G_V = \langle zz^\top \rangle_V$*, and* $g_V = \langle z\xi \rangle_V$*, one has*

$$\mathbb{E}_+ G_V = I_r, \quad \mathbb{E}_+ g_V = 0, \quad \|\theta_\star\|^2 \le 1, \quad \sigma_\xi^2 := \mathbb{E}_+ \langle \xi^2 \rangle_V = \mathbb{E}_+ \langle \gamma^2 \rangle_V - \|\theta_\star\|^2 \le 1. \tag{25}$$

If $G_V \preceq \Lambda I_r$*, valid deterministic bounds are*

$$\begin{aligned} \|g_V\| &\le \sqrt{\Lambda} + \Lambda \|\theta_\star\| \le \sqrt{\Lambda} + \Lambda, \\ \nu := \mathbb{E}_+ \|g_V\|^2 &\le \Lambda \sigma_\xi^2 \le \Lambda, \qquad 1 \le \Lambda \le \kappa, \qquad r \le \kappa \le \frac{5}{\lambda_{\min}^+(\Sigma)}. \end{aligned} \tag{26}$$

The inequalities concerning Λ use a choice no larger than the pointwise leverage bound κ.

Proof. For $v \in \ker(\Sigma)$, positivity gives

$$0 = v^\top \Sigma v = \mathbb{E}_+ \langle (v^\top x)^2 \rangle_V, \qquad \langle (v^\top x)^2 \rangle_V \geq 0. \tag{27}$$

Thus the within-video average is zero almost surely. Each of its finitely many summands has positive weight and is nonnegative, so each summand vanishes. Applying this argument to a finite basis of $\ker(\Sigma)$ establishes the asserted simultaneous support property. It also gives $v^\top b = 0$, hence $b \in \mathrm{range}(\Sigma)$ and $\Sigma\beta_\star = b$. Whitening these equations proves $\mathbb{E}_+ G_V = I_r$ and $\mathbb{E}_+ g_V = 0$. These population normal equations allow nonzero conditional residual means.

The population normal equation and Cauchy–Schwarz give the chain

$$\begin{aligned} \|\theta_\star\|^2 &= \theta_\star^\top \mathbb{E}_+ \langle z\gamma \rangle_V = \mathbb{E}_+ \langle (z^\top \theta_\star)\gamma \rangle_V \\ &\leq \big(\mathbb{E}_+ \langle (z^\top \theta_\star)^2 \rangle_V\big)^{1/2} \big(\mathbb{E}_+ \langle \gamma^2 \rangle_V\big)^{1/2} \\ &= \|\theta_\star\| \big(\mathbb{E}_+ \langle \gamma^2 \rangle_V\big)^{1/2} \leq \|\theta_\star\|. \end{aligned} \tag{28}$$

Division when $\|\theta_\star\| > 0$, and the trivial zero case otherwise, imply $\|\theta_\star\| \leq 1$. Expanding the residual square and substituting the normal equation yields

$$\begin{aligned} \sigma_\xi^2 &= \mathbb{E}_+ \langle \gamma^2 \rangle_V - 2\theta_\star^\top \mathbb{E}_+ \langle z\gamma \rangle_V + \theta_\star^\top \mathbb{E}_+ G_V \theta_\star \\ &= \mathbb{E}_+ \langle \gamma^2 \rangle_V - \|\theta_\star\|^2 \leq \mathbb{E}_+ \langle \gamma^2 \rangle_V \leq 1. \end{aligned} \tag{29}$$

For every scalar within-video quantity v_s, duality of the Euclidean norm followed by weighted Cauchy–Schwarz gives

$$\begin{aligned} \|\langle zv \rangle_V\|^2 &= \sup_{\|u\|=1} \langle (u^\top z) v \rangle_V^2 \\ &\leq \sup_{\|u\|=1} \langle (u^\top z)^2 \rangle_V \langle v^2 \rangle_V = \|G_V\| \langle v^2 \rangle_V \leq \Lambda \langle v^2 \rangle_V. \end{aligned} \tag{30}$$

Taking $v = \gamma$, and then $v = \xi$, proves

$$\begin{aligned} \|g_V\| &= \|\langle z\gamma \rangle_V - G_V \theta_\star\| \leq \|\langle z\gamma \rangle_V\| + \|G_V\| \|\theta_\star\| \\ &\leq \sqrt{\Lambda \langle \gamma^2 \rangle_V} + \Lambda \|\theta_\star\| \leq \sqrt{\Lambda} + \Lambda, \\ \mathbb{E}_+ \|g_V\|^2 &\leq \Lambda \mathbb{E}_+ \langle \xi^2 \rangle_V = \Lambda \sigma_\xi^2 \leq \Lambda. \end{aligned} \tag{31}$$

Finally, $G_V \preceq \kappa I_r$ and $\mathbb{E}_+ G_V = I_r$ allow $1 \leq \Lambda \leq \kappa$, while

$$r = \mathbb{E}_+ \langle \|z\|^2 \rangle_V \leq \kappa, \qquad \|z\|^2 = x^\top \Sigma^\dagger x \leq \frac{\|x\|^2}{\lambda_{\min}^+(\Sigma)} \leq \frac{5}{\lambda_{\min}^+(\Sigma)}. \tag{32}$$

Thus leverage is finite for this fixed finite-dimensional population and can be large for an ill-conditioned feature distribution. The moments in Eq. (31) preserve all within-video dependence. □

D.4 A MATRIX LAPLACE ARGUMENT WITH SCALAR CONDITIONAL BOUNDS

We use the classical inequality $\operatorname{tr} e^{A+B} \leq \operatorname{tr}(e^A e^B)$ for self-adjoint A, B; see the Golden–Thompson discussion in Tropp (2012). The recursion applies this two-matrix inequality at each conditioning step.

Lemma 2 (Scalar-isotropic Laplace recursion)**.** *Let $A_1, \ldots, A_N$ be independent self-adjoint $d \times d$ random matrices with $\mathbb{E} e^{A_i} \preceq m_i I_d$, where $m_i > 0$ are deterministic. Then*

$$\mathbb{E} \operatorname{tr} \exp\left(\sum_{i=1}^N A_i\right) \leq d \prod_{i=1}^N m_i. \tag{33}$$

Proof. Put $S_j = \sum_{i \le j} A_i$ and $S_0 = 0$. Conditional independence, Golden–Thompson, and positivity give, successively,

$$\begin{aligned} \mathbb{E}\,\mathrm{tr}\, e^{S_j} &\le \mathbb{E}\,\mathrm{tr}(e^{S_{j-1}} e^{A_j}) \\ &= \mathbb{E}\,\mathrm{tr}\big(e^{S_{j-1}} \mathbb{E}[e^{A_j} \mid A_1, \dots, A_{j-1}]\big) \\ &= \mathbb{E}\,\mathrm{tr}\big(e^{S_{j-1}} \mathbb{E} e^{A_j}\big) \le m_j \mathbb{E}\,\mathrm{tr}\, e^{S_{j-1}} \\ &\le m_j m_{j-1} \mathbb{E}\,\mathrm{tr}\, e^{S_{j-2}} \le \cdots \le \left(\prod_{i=1}^{j} m_i\right) \mathrm{tr}\, I_d. \end{aligned} \tag{34}$$

The trace-order step follows because, with $P = e^{S_{j-1}} \succeq 0$ and $Q = m_j I_d - \mathbb{E} e^{A_j} \succeq 0$, one has $\mathrm{tr}(PQ) = \mathrm{tr}(P^{1/2} Q P^{1/2}) \ge 0$. Taking $j = N$ proves the claim. □

Lemma 3 (Support-Gram lower tail)**.** *For independent copies $G_1, \dots, G_N$ of G_V, let $\overline{G} = N^{-1} \sum_i G_i$. If $\Lambda > 1$ and $0 < \omega < 1$, then*

$$\Pr\big(\lambda_{\min}(\overline{G}) \le 1 - \omega\big) \le r \exp\{-N d_\Lambda(\omega)\}, \quad \mathrm{kl}(a \| p) = a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p}. \tag{35}$$

If $\Lambda = 1$, then $G_i = I_r$ almost surely. The simpler exponent satisfies

$$d_\Lambda(\omega) \ge \frac{\omega + (1-\omega)\log(1-\omega)}{\Lambda} \ge \frac{\omega^2}{2\Lambda}. \tag{36}$$

Proof. For $\theta > 0$, convexity places $e^{-\theta t}$ below its chord on $[0, \Lambda]$. Functional calculus and $\mathbb{E} G_i = I_r$ imply

$$\begin{aligned} e^{-\theta G_i} &\preceq I_r + \frac{e^{-\theta\Lambda} - 1}{\Lambda} G_i, \\ \mathbb{E} e^{-\theta G_i} &\preceq \left(1 - \frac{1}{\Lambda} + \frac{e^{-\theta\Lambda}}{\Lambda}\right) I_r. \end{aligned} \tag{37}$$

Spectral mapping, exponential Markov, trace domination, and Lemma 2 then give the full bound

$$\begin{aligned} \Pr\left(\lambda_{\min}\left(\sum_i G_i\right) \le N(1-\omega)\right) &= \Pr\Big(e^{\theta \lambda_{\max}(-\sum_i G_i)} \ge e^{-\theta N(1-\omega)}\Big) \\ &\le e^{\theta N(1-\omega)} \mathbb{E} e^{\lambda_{\max}(-\theta \sum_i G_i)} \\ &\le e^{\theta N(1-\omega)} \mathbb{E}\,\mathrm{tr}\, e^{-\theta \sum_i G_i} \\ &\le r \exp\left\{N\left[\theta(1-\omega) + \log\left(1 - \frac{1}{\Lambda} + \frac{e^{-\theta\Lambda}}{\Lambda}\right)\right]\right\}. \end{aligned} \tag{38}$$

Let $p = 1/\Lambda$, $a = (1-\omega)/\Lambda$, and $u = \theta\Lambda$. Since $0 < a < p < 1$, differentiation gives the positive optimizer $u_\star = \log[p(1-a)/(a(1-p))]$. Substitution gives the binary relative-entropy exponent:

$$\begin{aligned} a u_\star + \log(1 - p + p e^{-u_\star}) &= a \log \frac{p(1-a)}{a(1-p)} + \log \frac{1-p}{1-a} \\ &= -a \log \frac{a}{p} - (1-a) \log \frac{1-a}{1-p} = -\,\mathrm{kl}(a \| p). \end{aligned} \tag{39}$$

This proves Eq. (35). If $\Lambda = 1$, then $I_r - G_i \succeq 0$ and $\mathbb{E}\,\mathrm{tr}(I_r - G_i) = 0$, so every nonnegative eigenvalue of $I_r - G_i$ is zero almost surely.

For the relaxed exponent, put $v = \omega/(\Lambda - 1)$ and use $(1+v)\log(1+v) \ge v$, which follows by differentiating their difference. Therefore

$$\begin{aligned} d_\Lambda(\omega) &= \frac{1-\omega}{\Lambda} \log(1-\omega) + \frac{\Lambda - 1}{\Lambda} (1+v) \log(1+v) \\ &\ge \frac{(1-\omega)\log(1-\omega) + \omega}{\Lambda} \\ &= \frac{1}{\Lambda} \int_0^\omega -\log(1-u)\, du \ge \frac{1}{\Lambda} \int_0^\omega u\, du = \frac{\omega^2}{2\Lambda}. \end{aligned} \tag{40}$$

The concentration argument uses independent videos, preserving dependence among their states. □

D.5 A VARIANCE-SENSITIVE BOUND FOR THE VIDEO RESIDUAL VECTOR

Lemma 4 (Video-score Bennett bound)**.** *Let $g_1, \dots, g_N \in \mathbb{R}^r$ be independent copies of g_V, with $\mathbb{E}g_i = 0$, $\|g_i\| \le B$, and $\mathbb{E}\|g_i\|^2 = \nu > 0$. For $t \ge 0$,*

$$\Pr\left(\left\|\frac{1}{N}\sum_i g_i\right\| \ge t\right) \le (r+1)\exp\left\{-\frac{N\nu}{B^2}h\left(\frac{Bt}{\nu}\right)\right\}. \tag{41}$$

Consequently, b_N in Eq. (13) *is a valid confidence radius. Writing $\ell = \log((r+1)/\delta_g)$, it obeys*

$$b_N \le \frac{B\ell}{3N} + \sqrt{\frac{2\nu\ell}{N} + \frac{B^2\ell^2}{9N^2}} \le \sqrt{\frac{2\nu\ell}{N}} + \frac{2B\ell}{3N}. \tag{42}$$

If $\nu = 0$, every $g_i = 0$ almost surely and $b_N = 0$.

Proof. Introduce the self-adjoint dilation

$$A_i = \begin{pmatrix} 0_{r\times r} & g_i \\ g_i^\top & 0 \end{pmatrix}, \qquad A_i^2 = \begin{pmatrix} g_i g_i^\top & 0 \\ 0 & \|g_i\|^2 \end{pmatrix}. \tag{43}$$

Its nonzero eigenvalues are $\pm\|g_i\|$, so $\|A_i\| \le B$, $\mathbb{E}A_i = 0$, and $\lambda_{\max}(\sum_i A_i) = \|\sum_i g_i\|$. Moreover,

$$\mathbb{E}g_i g_i^\top \preceq \mathbb{E}\|g_i\|^2 I_r = \nu I_r, \qquad \mathbb{E}A_i^2 \preceq \nu I_{r+1}. \tag{44}$$

For each integer $m \ge 2$ and each $\lambda \in [-B, B]$, $\lambda^m \le B^{m-2}\lambda^2$. Spectral calculus gives $A_i^m \preceq B^{m-2}A_i^2$. Boundedness justifies termwise expectation in the exponential series. For $\theta > 0$,

$$\begin{aligned}
\mathbb{E}e^{\theta A_i} &= I_{r+1} + \theta\mathbb{E}A_i + \sum_{m=2}^{\infty}\frac{\theta^m}{m!}\mathbb{E}A_i^m \\
&\preceq I_{r+1} + \sum_{m=2}^{\infty}\frac{\theta^m B^{m-2}}{m!}\mathbb{E}A_i^2 \\
&= I_{r+1} + \frac{e^{\theta B} - 1 - \theta B}{B^2}\mathbb{E}A_i^2 \\
&\preceq \left[1 + \frac{\nu}{B^2}(e^{\theta B} - 1 - \theta B)\right] I_{r+1} \\
&\preceq \exp\left\{\frac{\nu}{B^2}(e^{\theta B} - 1 - \theta B)\right\} I_{r+1}.
\end{aligned} \tag{45}$$

Apply exponential Markov and Lemma 2:

$$\begin{aligned}
\Pr\left(\left\|\sum_i g_i\right\| \ge Nt\right) &= \Pr\left(\lambda_{\max}\left(\sum_i A_i\right) \ge Nt\right) \\
&\le e^{-\theta Nt}\mathbb{E}\,\mathrm{tr}\, e^{\theta\sum_i A_i} \\
&\le (r+1)\exp\left\{-\theta Nt + \frac{N\nu}{B^2}(e^{\theta B} - 1 - \theta B)\right\}.
\end{aligned} \tag{46}$$

At $\theta_\star = B^{-1}\log(1 + Bt/\nu)$, writing $u = Bt/\nu$, the exponent becomes

$$\begin{aligned}
-\theta_\star Nt + \frac{N\nu}{B^2}(e^{\theta_\star B} - 1 - \theta_\star B) &= -\frac{N\nu}{B^2}\big[u\log(1+u) - u + \log(1+u)\big] \\
&= -\frac{N\nu}{B^2}\big[(1+u)\log(1+u) - u\big].
\end{aligned} \tag{47}$$

This proves the tail and its inverse-radius statement since $h'(u) = \log(1+u) > 0$ for $u > 0$.

To derive the elementary relaxation, induction gives $m! \ge 2\cdot 3^{m-2}$ for $m \ge 2$. For $0 \le \theta B < 3$,

$$\frac{e^{\theta B} - 1 - \theta B}{B^2} = \sum_{m=2}^{\infty}\frac{\theta^m B^{m-2}}{m!} \le \frac{\theta^2}{2}\sum_{j=0}^{\infty}(\theta B/3)^j = \frac{\theta^2}{2(1 - \theta B/3)}. \tag{48}$$

Choosing $\theta = t/(\nu + Bt/3)$ in the relaxed exponent yields $-Nt^2/[2(\nu + Bt/3)]$. Solving $Nt^2 = 2\ell(\nu + Bt/3)$ gives the first root in Eq. (42). The exact Bennett optimization is no weaker than this feasible relaxed choice. The second inequality follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Finally, $\nu = 0$ implies $\|g_i\|^2 = 0$ almost surely by nonnegativity. □

D.6 Uniform estimation of the weighted least-squares target

Proposition 2 (Uniform estimation of the clipped projection)**.** *Under the population and sampling conditions above, suppose Eq.* (13) *holds. Let $\widehat{\beta}$ be the exact minimum-norm weighted fit, and let the implemented predictor have clipped prediction discrepancy at most ζ from this fit on all eligible states. There is a training-only event $\mathcal{A}_f$ of probability at least $1-\delta_f$ such that*

$$\operatorname*{ess\,sup}_{V,\,q,\,t<T_q} |\widehat{\gamma}_{q,t} - f_\star(x_{q,t})| \le e_N. \tag{49}$$

For $\Lambda = 1$, one may take $\omega = 0$ and $\delta_G = 0$.

Proof. On the Gram event from Lemma 3, $\overline{G} \succeq (1-\omega) I_r \succ 0$. The support lemma ensures that every training feature is in $\mathrm{range}(U)$, so the raw empirical Gram has exactly that rank-r support. The minimum-norm fit lies on the same support. Its whitened coefficient $\widehat{\theta} = H^{1/2} U^\top \widehat{\beta}$ satisfies the exact weighted normal equation

$$\overline{G}\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N} \langle z\gamma \rangle_{V_i}, \qquad \overline{G}(\widehat{\theta} - \theta_\star) = \frac{1}{N}\sum_{i=1}^{N} \langle z(\gamma - z^\top \theta_\star)\rangle_{V_i} =: \overline{g}. \tag{50}$$

Lemma 4 bounds $\|\overline{g}\| \le b_N$ outside an event of probability δ_g. A union bound gives a common event with probability at least $1 - \delta_G - \delta_g$. The union bound accommodates dependence between $\overline{G}$ and $\overline{g}$.

Clipping onto $[-1,1]$ is nonexpansive: for $u \le v$, its monotonicity and slope in $[0,1]$ give $0 \le \mathrm{clip}(v) - \mathrm{clip}(u) \le v - u$, including intervals crossing either endpoint. On the common event, every eligible state therefore obeys

$$\begin{aligned} |\widehat{\gamma}(x) - f_\star(x)| &\le \zeta + |\,\mathrm{clip}(x^\top \widehat{\beta}) - \mathrm{clip}(x^\top \beta_\star)| \\ &\le \zeta + |x^\top(\widehat{\beta} - \beta_\star)| = \zeta + |z^\top(\widehat{\theta} - \theta_\star)| \\ &= \zeta + |z^\top \overline{G}^{-1}\overline{g}| \le \zeta + \|z\|\,\|\overline{G}^{-1}\overline{g}\| \\ &\le \zeta + \|z\|\,\|\overline{G}^{-1}\|\,\|\overline{g}\| = \zeta + \frac{\|z\|\,\|\overline{g}\|}{\lambda_{\min}(\overline{G})} \\ &\le \zeta + \frac{\sqrt{\kappa}\, b_N}{1-\omega} = e_N. \end{aligned} \tag{51}$$

Taking the essential supremum gives uniform estimation of $f_\star$. The population approximation error between $f_\star$ and γ remains a separate term. □

D.7 Exact finite-video calibration, including atoms and empty trajectories

For an independent video let $R_\star = \max_{q,t<T_q}(f_\star(x_{q,t}) - \gamma_{q,t})$, with an empty maximum equal to $-\infty$. Fix a deterministic population-level $a \in \mathbb{R}$, and set $p_a = \Pr(R_\star \le a)$. For $k = \lceil (n+1)(1-\alpha) \rceil \le n$, define

$$\begin{aligned} \psi_{n,k}(p) &= \sum_{j=0}^{k-1} \binom{n}{j} p^j (1-p)^{n-j}, \qquad \bar{\delta} = 1 - (1-\delta_f)(1 - \psi_{n,k}(p_a)), \\ s_N &= [a + e_N]_+ + e_N, \qquad \sigma(s) = \min\big(\{t < T : \gamma_t > 0,\ f_\star(x_t) > s\} \cup \{T\}\big), \\ \Phi(s) &= \mathbb{E}\big[C(\sigma(s)) - C(\tau_\star)\big], \qquad \bar{D} = \mathbb{E}D < \infty. \end{aligned} \tag{52}$$

Here $[u]_+ = \max\{u, 0\}$. Test-question expectations follow the convention in Appendix A.7. Both $\tau_\star$ and $\sigma(s)$ equal T if the reference maximum is tied.

Lemma 5 (Exact calibration-count identity and transfer)**.** *Fix $a \in \mathbb{R}$ and $f_\star$ at the population level. For independent calibration videos define $R_{\star,i}$ as defined above and $p_a = \Pr(R_\star \le a)$. If $1 \le k \le n$, then*

$$\begin{aligned} \Pr(R_{\star,(k)} > a) &= \Pr(\mathrm{Bin}(n, p_a) \le k-1) = \psi_{n,k}(p_a) \\ &= \Pr(\mathrm{Bin}(n, 1-p_a) \ge m), \qquad m = n - k + 1 = \lfloor (n+1)\alpha \rfloor. \end{aligned} \tag{53}$$

On the intersection of $\mathcal{A}_f$ and $\mathcal{A}_c = \{R_{\star,(k)} \le a\}$, the actual correction obeys $\rho_\alpha \le [a + e_N]_+$. Moreover, $\mathcal{A}_f$ and $\mathcal{A}_c$ are independent and

$$\Pr(\mathcal{A}_f \cap \mathcal{A}_c) \ge (1 - \delta_f)(1 - \psi_{n,k}(p_a)) = 1 - \bar{\delta}. \tag{54}$$

Proof. The indicators $I_i = \mathbf{1}\{R_{\star,i} \le a\}$ are independent Bernoulli variables of mean p_a. There are at least k scores at most a if and only if the kth order statistic is at most a. Hence

$$\begin{aligned}\Pr(R_{\star,(k)} > a) &= \Pr\left(\sum_{i=1}^{n} I_i < k\right) = \sum_{j=0}^{k-1} \binom{n}{j} p_a^j (1 - p_a)^{n-j} \\ &= \Pr\left(\sum_{i=1}^{n} (1 - I_i) \ge n - k + 1\right).\end{aligned} \tag{55}$$

This uses the weak inequality inside I_i and the strict order-statistic exceedance. It is therefore exact for atoms, ties, and $R_\star = -\infty$. Since $k = \lceil (n+1)(1-\alpha) \rceil$, the elementary identity $j - \lceil j - u \rceil = \lfloor u \rfloor$ for integer j gives the stated m.

On $\mathcal{A}_f$, every calibration state satisfies $\widehat{\gamma} \le f_\star + e_N$. For nonempty videos, and also for empty videos using $-\infty + e_N = -\infty$,

$$\begin{aligned}R_i &= \max_{q,t<T_q} (\widehat{\gamma}_{q,t} - \gamma_{q,t}) \\ &\le \max_{q,t<T_q} (f_\star(x_{q,t}) + e_N - \gamma_{q,t}) = R_{\star,i} + e_N.\end{aligned} \tag{56}$$

Coordinatewise order implies order-statistic order: if at least k of the quantities $R_{\star,i}$ are at most a, their corresponding R_i are at most $a + e_N$. Thus, on $\mathcal{A}_f \cap \mathcal{A}_c$,

$$R_{(k)} \le R_{\star,(k)} + e_N \le a + e_N, \qquad \rho_\alpha = \max\{0, R_{(k)}\} \le \max\{0, a + e_N\}. \tag{57}$$

The event $\mathcal{A}_f$ depends only on training videos, whereas $\mathcal{A}_c$ uses only independent calibration videos and the fixed population function $f_\star$. Defining $\mathcal{A}_c$ through the population residual $R_\star$ makes it independent of training, giving Eq. (54). □

Lemma 6 (Rounding-aware exponential calibration bound)**.** *Let $u = (k-1)/n$. If $u < p < 1$, then*

$$\psi_{n,k}(p) \le \exp\{-n \operatorname{kl}(u \| p)\}. \tag{58}$$

At $u = 0$ this is the exact identity $\psi_{n,1}(p) = (1-p)^n$. At $p = 1$ the failure probability is zero. Equivalently, for $q = 1 - p < m/n$, the exponent is $n \operatorname{kl}(m/n \| q)$.

Proof. Let $S \sim \mathrm{Bin}(n, p)$. For any $\lambda < 0$, independence of the Bernoulli summands and exponential Markov imply

$$\begin{aligned}\Pr(S \le nu) &= \Pr(e^{\lambda S} \ge e^{\lambda n u}) \le e^{-\lambda n u} \mathbb{E} e^{\lambda S} \\ &= \exp\{n[\log(1 - p + p e^{\lambda}) - \lambda u]\}.\end{aligned} \tag{59}$$

For $0 < u < p$, the minimizer is $\lambda_\star = \log[u(1-p)/(p(1-u))] < 0$. Direct substitution gives

$$\begin{aligned}\log(1 - p + p e^{\lambda_\star}) - \lambda_\star u &= \log \frac{1-p}{1-u} - u \log \frac{u(1-p)}{p(1-u)} \\ &= -u \log \frac{u}{p} - (1-u) \log \frac{1-u}{1-p} = -\operatorname{kl}(u \| p).\end{aligned} \tag{60}$$

The case $u = 0$ follows by $\lambda \to -\infty$, or by observing that $S = 0$ exactly when all n indicators are zero. The case $p = 1$ is deterministic. Finally, binary relative entropy is invariant under simultaneously replacing its two arguments by their complements. This gives the equivalent upper-tail form. The strict order-statistic exceedance gives the integer threshold $k - 1$. □

D.8 NONMONOTONE ANSWERS AND THE EXACT EXCESS-COST PROFILE

Lemma 7 (First-consistent comparator and repeated opportunities)**.** *Define $\tau_\star$, $\sigma(s)$, and C as in Eqs.* (14) *and* (52)*. Then $\tau_\star \le \sigma(s) \le T$. For $s = [a + e_N]_+ + e_N$, on $\mathcal{A}_f \cap \mathcal{A}_c$,*

$$\tau \le \sigma(s), \qquad 0 \le (C(\tau) - C(\tau_\star))_+ \le C(\sigma(s)) - C(\tau_\star) \le D. \tag{61}$$

With $\max \varnothing = -\infty$, the population cost profile has the exact representation

$$\Phi(s) = \mathbb{E} \sum_{j=2}^{T} c_j \, \mathbf{1}\left\{ \tau_\star < j, \ \max_{t<j:\, \gamma_t>0} f_\star(x_t) \le s \right\}. \tag{62}$$

Proof. By definition, the qualifying early states for $\sigma(s)$ are a subset of those for $\tau_\star$. Hence $\tau_\star \le \sigma(s)$. If the reference has a unique winner, then $\gamma_t > 0$ implies $a_t = a_H$ and

$$\gamma_t = p_H(a_H) - \max_{c \ne a_H} p_H(c) = \Delta_H. \tag{63}$$

If the maximum is tied, no candidate can have a strictly positive retained margin; both comparators then equal T. The argument allows the answer to change repeatedly along the path.

On $\mathcal{A}_f \cap \mathcal{A}_c$, any state counted in $\sigma(s)$ obeys the strict chain

$$\widehat{\gamma}_t \ge f_\star(x_t) - e_N > s - e_N = [a + e_N]_+ \ge \rho_\alpha. \tag{64}$$

The clipping in the lower bound preserves the sign of $\widehat{\gamma}_t - \rho_\alpha$; therefore $L_t > 0$, and SBH must have stopped by this state. If no such state exists, $\sigma(s) = T$ and $\tau \le T$ is automatic. Since $c_j \ge 0$, C is nondecreasing, so

$$(C(\tau) - C(\tau_\star))_+ \le (C(\sigma(s)) - C(\tau_\star))_+ = C(\sigma(s)) - C(\tau_\star) \le C(T) - C(\tau_\star). \tag{65}$$

This cost comparison also holds when coverage fails: an incorrect premature stop can only decrease the cost on the left.

For every fixed trajectory, telescoping the cumulative costs yields

$$\begin{aligned} C(\sigma(s)) - C(\tau_\star) &= \sum_{j=2}^{T} c_j \mathbf{1}\{\tau_\star < j \le \sigma(s)\} \\ &= \sum_{j=2}^{T} c_j \mathbf{1}\left\{ \tau_\star < j, \ \max_{t<j:\, \gamma_t>0} f_\star(x_t) \le s \right\}. \end{aligned} \tag{66}$$

Indeed, $j \le \sigma(s)$ exactly means that no eligible state before j both has $\gamma_t > 0$ and satisfies $f_\star(x_t) > s$. Taking expectations proves Eq. (62). Nonnegative terms justify the interchange if the number of states is random; integrability of D ensures finiteness. This identity includes trajectories that leave and later return to the reference answer. □

Proposition 3 (A fitted-predictor refinement)**.** *For any fixed fitted predictor and correction $\rho_\alpha \le r_0 < \infty$, define $d_t = \gamma_t - \widehat{\gamma}_t$. Then, on each trajectory,*

$$\begin{aligned} (C(\tau) - C(\tau_\star))_+ &\le \sum_{j=2}^{T} c_j \mathbf{1}\left\{ \tau_\star < j, \ \Delta_H \le r_0 + \min_{t<j:\, \gamma_t>0} d_t \right\} \\ &\le D \, \mathbf{1}\{\tau_\star < T, \ \Delta_H \le r_0 + d_{\tau_\star}\}. \end{aligned} \tag{67}$$

The minima are used only when $\tau_\star < j$, so their index sets are nonempty; the last indicator is evaluated only when $\tau_\star < T$.

Proof. If $\tau_\star < j \le \tau$, the algorithm did not stop at any earlier state $t < j$. At every such reference-consistent state, the lower-bound sign test therefore gives $\widehat{\gamma}_t \le \rho_\alpha$. By Eq. (63),

$$\Delta_H = \gamma_t = \widehat{\gamma}_t + d_t \le \rho_\alpha + d_t \le r_0 + d_t, \qquad t < j, \ \gamma_t > 0. \tag{68}$$

Taking the minimum over these states, multiplying by $c_j \ge 0$, and summing the exact identity $(C(\tau) - C(\tau_\star))_+ = \sum_{j=2}^{T} c_j \mathbf{1}\{\tau_\star < j \le \tau\}$ proves the first bound. The minimum is at most $d_{\tau_\star}$, and the sum of all eligible remaining costs is D, proving the second. The bound holds pathwise for every fixed predictor and correction satisfying the stated envelope. □

D.9 Proof of the main theorem and comparison with single-call inference

The theorem also has the following conditional-efficiency refinement. With probability at least $1-\bar{\delta}$ over training and calibration, $\rho_\alpha \le [a + e_N]_+$ and the conditional expected excess cost is at most $\Phi(s_N)$. The video-level early-disagreement probability is at most α after averaging over calibration and new videos, as in Theorem 1.

Proof of Theorem 2. Condition on all training data and the frozen specification. The predictor, including its deterministic numerical implementation, is now fixed. Independent identically distributed calibration and new video objects are exchangeable, and applying the same maximum-residual map preserves exchangeability. The rank argument in Appendix D.2 gives simultaneous lower-bound coverage at least $1-\alpha$, and hence early disagreement probability at most α. Averaging over the training data establishes the retained risk guarantee. Exchangeability and a predictor fixed before calibration suffice for this risk guarantee.

For efficiency, let $\mathcal{A} = \mathcal{A}_f \cap \mathcal{A}_c$, $b_0 = \Pr(\mathcal{A}^c)$, and, for the independent test question, define

$$Z = (C(\tau) - C(\tau_\star))_+, \qquad A_s = C(\sigma(s_N)) - C(\tau_\star). \tag{69}$$

Proposition 2 and Lemma 5 give $b_0 \le \bar{\delta}$. Lemma 7 gives $Z \le A_s$ on $\mathcal{A}$, and the universal bound $Z \le D$ holds on its complement. Both A_s and D depend only on the independent test object and fixed population quantities. They are thus independent of $\mathcal{A}$, although Z itself need not be. Consequently,

$$\begin{aligned}
\mathbb{E}Z &= \mathbb{E}[Z\mathbf{1}_{\mathcal{A}}] + \mathbb{E}[Z\mathbf{1}_{\mathcal{A}^c}] \\
&\le \mathbb{E}[A_s\mathbf{1}_{\mathcal{A}}] + \mathbb{E}[D\mathbf{1}_{\mathcal{A}^c}] \\
&= \Pr(\mathcal{A})\mathbb{E}A_s + \Pr(\mathcal{A}^c)\mathbb{E}D \\
&= (1-b_0)\Phi(s_N) + b_0\bar{D} \\
&= \Phi(s_N) + b_0(\bar{D} - \Phi(s_N)) \\
&\le \Phi(s_N) + \bar{\delta}(\bar{D} - \Phi(s_N)).
\end{aligned} \tag{70}$$

The last step uses $0 \le \Phi(s_N) \le \bar{D}$. Retaining this mixture is sharper than adding an independent worst-case cost penalty, and requires no deterministic upper bound on cost.

For each realization of training and calibration in $\mathcal{A}$, the same pathwise inequality holds for almost every new test video. Integration over that video gives conditional expected excess cost at most $\Phi(s_N)$. Equation (54) gives the claimed probability of these realizations. The risk bound follows from the rank argument, which averages over calibration and test objects. □

Corollary 1 (Cost-weighted margin and real computational savings)**.** *Suppose $\gamma_t - f_\star(x_t) \le b_-$ for all reference-consistent early states, almost surely, for a fixed $b_- \ge 0$. For $s \ge 0$,*

$$\Phi(s) \le \Psi(s + b_-), \qquad \Psi(u) = \mathbb{E}\big[D\,\mathbf{1}\{\tau_\star < T,\ 0 < \Delta_H \le u\}\big]. \tag{71}$$

This profile retains the association between difficult margins and expensive trajectories. In particular, if $\Psi(u) \le K_0 u^\eta$ for $u \ge 0$ and $\eta > 0$, then

$$\mathbb{E}[(C(\tau) - C(\tau_\star))_+] \le (1-\bar{\delta})\min\{\bar{D}, K_0(s_N + b_-)^\eta\} + \bar{\delta}\bar{D}. \tag{72}$$

For integrable compared costs, let $A_{\mathrm{one}} = \mathbb{E}[C_{\mathrm{one}} - C(\tau_\star)]$. Then

$$\mathbb{E}[C_{\mathrm{one}} - C(\tau)] \ge A_{\mathrm{one}} - \big[\Phi(s_N) + \bar{\delta}(\bar{D} - \Phi(s_N))\big]. \tag{73}$$

Proof. If $\tau_\star < T$ and $\Delta_H > s + b_-$, the first reference-consistent state satisfies

$$f_\star(x_{\tau_\star}) \ge \gamma_{\tau_\star} - b_- = \Delta_H - b_- > s, \qquad \tau_\star \le \sigma(s) \le \tau_\star. \tag{74}$$

Thus $\sigma(s) = \tau_\star$. If $\tau_\star = T$, both times equal T as well. In all remaining cases, the excess comparator cost is at most D, so

$$\begin{aligned}
0 \le C(\sigma(s)) - C(\tau_\star) &\le D\,\mathbf{1}\{\tau_\star < T,\ 0 < \Delta_H \le s + b_-\}, \\
\Phi(s) \le \Psi(s + b_-) &\le \min\{\bar{D}, K_0(s + b_-)^\eta\}.
\end{aligned} \tag{75}$$

The power-bound assertion follows because the main mixture is increasing in Φ. The single-call comparison uses the pointwise relation

$$\begin{aligned} C(\tau) - C(\tau_\star) &\le (C(\tau) - C(\tau_\star))_+, \\ \mathbb{E}[C_{\text{one}} - C(\tau)] &= A_{\text{one}} - \mathbb{E}[C(\tau) - C(\tau_\star)] \\ &\ge A_{\text{one}} - \mathbb{E}[(C(\tau) - C(\tau_\star))_+] \\ &\ge A_{\text{one}} - \big[\Phi(s_N) + \bar{\delta}(\bar{D} - \Phi(s_N))\big]. \end{aligned} \tag{76}$$

In particular, positivity requires a positive baseline advantage after counting all costs of the reference-informed sequential comparator. The single-call baseline includes question-based retrieval and one endpoint scoring call, as specified in Appendix A.6. □

For integrable costs, if C_{one} is the complete single-call E_T baseline cost from Appendix A.6, a sufficient condition for positive expected savings is

$$\mathbb{E}[C_{\text{one}} - C(\tau_\star)] > \Phi(s_N) + \bar{\delta}\big(\bar{D} - \Phi(s_N)\big). \tag{77}$$

Repeated scoring and candidate encoding can make the left side negative. Computational savings therefore depend on the complete execution cost.

Corollary 2 (Equality with the first-consistent time)**.** *Let $\pi_s = \Pr(\sigma(s_N) > \tau_\star)$ and $\pi_\star = \Pr(\tau_\star < T)$. Under the main theorem,*

$$\Pr(\tau \neq \tau_\star) \le \min\{1,\ \alpha + (1 - \bar{\delta})\pi_s + \bar{\delta}\pi_\star\}. \tag{78}$$

Proof. On simultaneous coverage, a positive lower bound implies $\gamma_t > 0$, so $\tau < \tau_\star$ is impossible. The coverage failure probability is at most α, including for any prespecified sampled question. On $\mathcal{A}$, $\tau > \tau_\star$ implies $\sigma(s_N) > \tau_\star$; on $\mathcal{A}^c$, it at least implies $\tau_\star < T$. Independence from the test object gives

$$\begin{aligned} \Pr(\tau \neq \tau_\star) &= \Pr(\tau < \tau_\star) + \Pr(\tau > \tau_\star) \\ &\le \alpha + (1 - b_0)\pi_s + b_0\pi_\star \\ &\le \alpha + (1 - \bar{\delta})\pi_s + \bar{\delta}\pi_\star, \end{aligned} \tag{79}$$

where $b_0 \le \bar{\delta}$ and $\pi_s \le \pi_\star$. Clamping by one proves the statement. Unlike excess computation, equality of stopping times does require the coverage term. □

D.10 Explicit training and calibration sample requirements

Proposition 4 (Exact-exponent training requirement)**.** *Fix a target estimation tolerance $\varepsilon > \zeta$ and let $v = (1 - \omega)(\varepsilon - \zeta)/\sqrt{\kappa}$. For $\Lambda > 1$ and $\nu > 0$, the sufficient sample requirement*

$$N \ge \max\left\{ \frac{\log(r/\delta_G)}{d_\Lambda(\omega)},\ \frac{B^2 \log((r+1)/\delta_g)}{\nu h(Bv/\nu)} \right\} \tag{80}$$

ensures $e_N \le \varepsilon$. A simpler sufficient requirement is

$$N \ge \max\left\{ \frac{2\Lambda}{\omega^2} \log \frac{r}{\delta_G},\ \left(\frac{2\nu\kappa}{(1-\omega)^2(\varepsilon - \zeta)^2} + \frac{2B\sqrt{\kappa}}{3(1-\omega)(\varepsilon - \zeta)} \right) \log \frac{r+1}{\delta_g} \right\}. \tag{81}$$

Omit the first term when $\Lambda = 1$ and use $\omega = 0, \delta_G = 0$; omit the second when $\nu = 0$. At least one eligible training video is still required.

Proof. The first term of Eq. (80) is exactly the Gram condition. Since h is increasing, the second term implies

$$\frac{B^2 \log((r+1)/\delta_g)}{N\nu} \le h(Bv/\nu), \qquad b_N = \frac{\nu}{B} h^{-1}\left(\frac{B^2 \log((r+1)/\delta_g)}{N\nu} \right) \le v. \tag{82}$$

Consequently, $e_N \le \sqrt{\kappa}v/(1-\omega) + \zeta = \varepsilon$. For the elementary version, Eq. (36) handles the Gram term. The second term of Eq. (81) is exactly $N \ge 2(\nu + Bv/3)\log((r+1)/\delta_g)/v^2$. The Bernstein exponent established in Lemma 4 then satisfies

$$\frac{Nv^2}{2(\nu + Bv/3)} \ge \log \frac{r+1}{\delta_g}, \qquad \Pr(\|\bar{g}\| > v) \le \delta_g. \tag{83}$$

Equivalently, the relaxed radius and hence b_N are at most v. The zero-variance and deterministic-Gram cases were proved in the corresponding lemmas. □

Proposition 5 (Finite calibration size with upper-tail slack)**.** *The finite-order-statistic convention requires exactly*

$$k \le n \quad \Longleftrightarrow \quad (n+1)\alpha \ge 1 \quad \Longleftrightarrow \quad n \ge \lceil 1/\alpha \rceil - 1. \tag{84}$$

For fixed a, the exact condition for $\Pr(\mathcal{A}_c) \ge 1 - \delta_c$ is $\psi_{n,k}(p_a) \le \delta_c$. Suppose further that $q_a := \Pr(R_\star > a) \le \alpha - \epsilon_c$, where $0 < \epsilon_c < \alpha$. Then it suffices that

$$n \ge \max\left\{ \lceil 1/\alpha \rceil - 1,\ \frac{2(1-\alpha)}{\epsilon_c},\ \frac{8\alpha + 4\epsilon_c/3}{\epsilon_c^2} \log \frac{1}{\delta_c} \right\}. \tag{85}$$

The exact binomial condition, or the rounded KL bound in Lemma 6, is preferable when numerically evaluating a sample requirement.

Proof. Since n is an integer, $\lceil (n+1)(1-\alpha) \rceil \le n$ if and only if $(n+1)(1-\alpha) \le n$, proving Eq. (84). When it fails, $R_{(n+1)} = +\infty$ and SBH cannot stop early, independently of predictor accuracy.

The exact condition follows directly from Lemma 5. For the explicit bound, set $q_0 = \alpha - \epsilon_c$ and couple Bernoulli variables by common independent uniforms. Since $q_a \le q_0$, the number of scores above a is stochastically bounded by $S_0 \sim \mathrm{Bin}(n, q_0)$. Writing $A = 1 - \alpha$ and $m = \lfloor (n+1)\alpha \rfloor$, one has

$$m \ge n\alpha - A, \qquad t_n := n\epsilon_c - A > 0, \qquad \psi_{n,k}(p_a) \le \Pr(S_0 - nq_0 \ge t_n). \tag{86}$$

The scalar Bernstein step follows from the same series argument. A centered Bernoulli variable Y has $\mathbb{E}Y = 0$, $|Y| \le 1$, and $\mathbb{E}Y^2 = v_0 = q_0(1-q_0)$. For $\lambda \in (0, 3)$, the same series argument as Eq. (48) gives

$$\begin{aligned} \mathbb{E}e^{\lambda Y} &= 1 + \sum_{j=2}^{\infty} \frac{\lambda^j \mathbb{E}Y^j}{j!} \le 1 + v_0 \sum_{j=2}^{\infty} \frac{\lambda^j}{j!} \\ &\le 1 + \frac{v_0\lambda^2}{2(1-\lambda/3)} \le \exp\left\{ \frac{v_0\lambda^2}{2(1-\lambda/3)} \right\}. \end{aligned} \tag{87}$$

Independence, exponential Markov, and the choice $\lambda = t_n/(nv_0 + t_n/3)$ therefore yield

$$\begin{aligned} \Pr(S_0 - nq_0 \ge t_n) &\le \inf_{0<\lambda<3} \exp\left\{ -\lambda t_n + \frac{nv_0\lambda^2}{2(1-\lambda/3)} \right\} \\ &\le \exp\left\{ -\frac{t_n^2}{2nv_0 + 2t_n/3} \right\}. \end{aligned} \tag{88}$$

The second requirement in Eq. (85) gives $t_n \ge n\epsilon_c/2$. Since $v_0 \le q_0 \le \alpha$ and $t \mapsto t^2/(2n\alpha + 2t/3)$ is increasing for $t > 0$,

$$\begin{aligned} \frac{t_n^2}{2nv_0 + 2t_n/3} &\ge \frac{t_n^2}{2n\alpha + 2t_n/3} \ge \frac{(n\epsilon_c/2)^2}{2n\alpha + n\epsilon_c/3} \\ &= \frac{n\epsilon_c^2}{8\alpha + 4\epsilon_c/3} \ge \log \frac{1}{\delta_c}. \end{aligned} \tag{89}$$

Combining the preceding inequalities proves the sufficient condition. Integer ceilings give a sufficient integer sample size. □

Corollary 3 (An explicit sample-to-computation guarantee)**.** *Suppose the one-sided approximation condition in Corollary 1 holds. Choose a target reference-margin scale $u > b_- + \max\{a, 0\}$ and set*

$$\varepsilon_u = \min\left\{ u - b_-,\ \frac{u - b_- - a}{2} \right\} > 0, \qquad \delta_{fc} = 1 - (1-\delta_f)(1-\delta_c). \tag{90}$$

If $\varepsilon_u > \zeta$, the training requirement of Proposition 4 with $\varepsilon = \varepsilon_u$, together with $\psi_{n,k}(p_a) \le \delta_c$, gives

$$\mathbb{E}[(C(\tau) - C(\tau_\star))_+] \le (1 - \delta_{fc})\Psi(u) + \delta_{fc}\bar{D}. \tag{91}$$

Under $\Psi(u) \le K_0 u^\eta$, replace $\Psi(u)$ by $\min\{\bar{D}, K_0 u^\eta\}$. All sample sizes count independent videos.

Proof. The positive-part correction can be inverted without loosening its negative-a branch:

$$\begin{aligned} [a+e]_+ + e &= \max\{e, a+2e\}, \\ [a+e]_+ + e + b_- \le u &\iff e \le u - b_- \text{ and } e \le \frac{u - b_- - a}{2}. \end{aligned} \tag{92}$$

The training requirement therefore gives $s_N + b_- \le u$. Since Ψ is nondecreasing, $\Phi(s_N) \le \Psi(u) \le \bar{D}$. Also $\bar{\delta} \le \delta_{fc}$. Applying these two monotonicities to the main mixture proves

$$\begin{aligned} \mathbb{E}[(C(\tau) - C(\tau_\star))_+] &\le (1-\bar{\delta})\Phi(s_N) + \bar{\delta}\bar{D} \\ &\le (1-\bar{\delta})\Psi(u) + \bar{\delta}\bar{D} \\ &\le (1-\delta_{fc})\Psi(u) + \delta_{fc}\bar{D}. \end{aligned} \tag{93}$$

The cost-weighted margin assumption gives the final specialization. □

For fixed conditioning and moment parameters, the elementary training requirement has leading order $\kappa\nu(\varepsilon-\zeta)^{-2}\log((r+1)/\delta_g)$, in addition to Gram conditioning. Calibration with tail slack ϵ_c has the sufficient order $\alpha\epsilon_c^{-2}\log(1/\delta_c) + \epsilon_c^{-1}$. Increasing N improves estimation of $f_\star$; increasing n controls uncertainty in a tail quantile. The population approximation term remains: as estimation error decreases, the margin threshold approaches the possibly positive value $[a]_+ + b_-$.

D.11 Optional sharper score radii and deterministic numerical checks

Proposition 6 (A computable refinement of the score bound)**.** *For $t > 0$ and $\nu > 0$, the same assumptions as Lemma 4 give*

$$\begin{aligned} &\Pr(\|\overline{g}\| > t) \le J_N(t), \\ &J_N(t) = \min\left\{1, \frac{\nu}{Nt^2},\ (r+1)\inf_{\theta>0} e^{-N\theta t}\left[1 + \frac{\nu}{B^2}(e^{\theta B} - 1 - \theta B)\right]^N\right\}. \end{aligned} \tag{94}$$

Any deterministic t with $J_N(t) \le \delta_g$ may replace b_N. The deterministic bound $\|\overline{g}\| \le B$ also permits replacing b_N by $\min\{b_N, B\}$.

Proof. Retain the penultimate scalar factor in Eq. (45), apply Lemma 2, and use Markov exactly as in Eq. (46). This gives the variational term without the relaxation $1 + x \le e^x$. Independently, the centered, independent video vectors satisfy

$$\begin{aligned} \mathbb{E}\|\overline{g}\|^2 &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\|g_i\|^2 + \frac{1}{N^2}\sum_{i\ne j}\mathbb{E}(g_i^\top g_j) \\ &= \frac{\nu}{N} + \frac{1}{N^2}\sum_{i\ne j}(\mathbb{E}g_i)^\top(\mathbb{E}g_j) = \frac{\nu}{N}, \\ \Pr(\|\overline{g}\| > t) &\le \frac{\mathbb{E}\|\overline{g}\|^2}{t^2} = \frac{\nu}{Nt^2}. \end{aligned} \tag{95}$$

Each of these bounds holds for the same event, so their minimum is valid without an additional union bound. Finally, $\|\overline{g}\| \le N^{-1}\sum_i \|g_i\| \le B$ deterministically. Any such replacement propagates monotonically through the uniform radius and the cost profile. This one-dimensional optimization tightens the closed-form Bennett expression. □

Proposition 7 (Sufficient retention of the numerical SVD support)**.** *Let $A = W^{1/2}X$, where the weights are exactly those of Eq.* (9)*, and let N_s be the number of training states. On the Gram event, all positive singular values exceed the prescribed numerical threshold provided*

$$\epsilon_{\mathrm{mach}}\max(N_s, 5) < \sqrt{\frac{(1-\omega)\lambda_{\min}^+(\Sigma)}{5}}. \tag{96}$$

This condition ensures rank retention under the stated cutoff. Floating-point perturbations are accounted for separately by ζ.

Proof. The sample covariance is $\widehat{\Sigma} = A^\top A$. For any $v \in \mathrm{range}(U)$, the whitening identity and Gram event give

$$\begin{aligned} v^\top \widehat{\Sigma} v &= (H^{1/2}U^\top v)^\top \overline{G}(H^{1/2}U^\top v) \\ &\geq (1-\omega)\|H^{1/2}U^\top v\|^2 = (1-\omega)v^\top \Sigma v \\ &\geq (1-\omega)\lambda^+_{\min}(\Sigma)\|v\|^2. \end{aligned} \tag{97}$$

All features lie in this support, so these are exactly its positive eigenvalues. The feature norm bound and normalized weights also imply

$$\begin{aligned} &\sigma^+_{\min}(A) \geq \sqrt{(1-\omega)\lambda^+_{\min}(\Sigma)}, \\ &\sigma^2_{\max}(A) = \|\widehat{\Sigma}\| \leq \mathrm{tr}(\widehat{\Sigma}) = \frac{1}{N}\sum_i \langle \|x\|^2 \rangle_{V_i} \leq 5. \end{aligned} \tag{98}$$

Combining Eq. (96) with these inequalities yields

$$\epsilon_{\mathrm{mach}} \max(N_s, 5)\sigma_{\max}(A) < \sqrt{(1-\omega)\lambda^+_{\min}(\Sigma)} \leq \sigma^+_{\min}(A), \tag{99}$$

which is the required strict retention criterion. True zero singular values are discarded in exact arithmetic. For a floating-point implementation, the clipped-prediction perturbation contributes ζ in Eq. (13) and requires a separate bound, including when using float64. □

D.12 Sharpness and the independent-video sampling unit

Proposition 8 (Sharpness of the deterministic transfer and the sampling unit)**.** *The following statements hold.*

(i) Given only (n, k, p_a), the calibration failure probability in Eq. (53) *cannot be reduced: it is an identity.*

(ii) The coefficient two in the envelope $a + b_- + 2e$ when $a \geq 0$ cannot be uniformly reduced using only the stated two-sided estimation error, calibration envelope, and safe-state underprediction envelope. Strict separation is necessary for a strictly positive stopping test.

(iii) Duplicating every eligible question within a training video any number of times leaves its regression contribution unchanged. Therefore a generic improvement based only on the total number of correlated states cannot replace the independent-video sample size N.

Proof. For (i), the distribution of each indicator $\mathbf{1}\{R_{\star,i} \leq a\}$ is completely determined by p_a. Its independent sum has exactly the binomial law used in Eq. (55). In particular, a two-point score law with probability p_a below or at a and probability $1 - p_a$ above a realizes the same expression. The identity holds for both discrete and continuous score distributions.

For (ii), fix $0 < e < 1/3$. Consider permissible scalar prediction envelopes with $a = b_- = 0$. Give every calibration video an eligible state with $\gamma = 0$, $f_\star = 0$, and $\widehat{\gamma} = e$. Then $R_\star = 0$, $R = e$, and, for $k \leq n$, $\rho_\alpha = e$. For a test trajectory with $T = 2$, at its first and only early state $t = 1$, take

$$\gamma = \Delta_H = 2e, \qquad f_\star = 2e, \qquad \widehat{\gamma} = e, \qquad L = \mathrm{clip}(e - e) = 0. \tag{100}$$

All scores and predictions lie in $[-1, 1]$, both estimation deviations have magnitude e, and the safe-state approximation error is zero. A binary reference distribution $(1/2 + e, 1/2 - e)$ realizes the positive margin. Nevertheless, this state does not trigger stopping. Any proposed sufficient separation $\Delta_H > ce$ with $c < 2$ would be satisfied and hence would be false for these envelopes. The example establishes sharpness of the deterministic error transfer under these envelopes. The two estimation errors can have opposite signs; replacing their sum by a single error would need an additional structural relation.

For (iii), suppose a video has Q eligible questions and duplicate each question m times, preserving its states. Every original state contribution to the average then becomes

$$\sum_{j=1}^{m} \frac{1}{mQ|\mathcal{T}_q|} u_{q,t} = \frac{1}{Q|\mathcal{T}_q|} u_{q,t}. \tag{101}$$

Thus G_V, g_V, the empirical Gram, the empirical response vector, and the exact weighted least-squares fit are all unchanged, while the number of recorded states is multiplied by m. In a non-degenerate regression population, the unchanged estimator retains its uncertainty as $m \to \infty$. A vanishing error rate based on state count alone would therefore require further within-video assumptions. The implemented numerical fit is also unchanged if its retained support stays the same; increasing N_s changes the specified SVD cutoff and need not preserve that support. □

The support normal equation, binomial calibration count, strict trigger implication, and cost-survival identity are exact. Matrix Laplace bounds, leverage control, and the margin envelope provide sufficient inequalities. The population quantities p_a, κ, ν, and Ψ characterize conditions for computational savings; using them quantitatively on data requires estimation. The threshold a is fixed independently of calibration. Selecting it from calibration residuals requires a simultaneous selection argument, and changing the question-collection rule, reference definition, or scoring protocol requires calibration under the corresponding law.