CERTIFIED SELECTIVE GATES FOR VIDEO-LANGUAGE MODELS EXCEED THE SELECTIVE-ERROR TARGET BEFORE GENRE DRIFT BEGINS

Anonymous authors
Paper under double-blind review

ABSTRACT

A selective gate lets a frozen video-language model answer only when its error rate on answered questions is certified below a target α, and the standard repair when the genre mix drifts is to detect the break and recalibrate on fresh labels. We find that on most drift streams there is nothing for the repair to restore. A split-conformal gate certifies the false-accept rate, and selective error equals that rate times the base error rate divided by coverage, so at the certified level the gate is over α wherever coverage falls below the base error rate, before any drift begins. We test this on 5 backbones (3B–32B) and 4 video-QA benchmarks with a per-genre certified gate, a label-free coverage-drift monitor, and drift streams built from benchmark metadata. The frozen gate is already over α before the break on 30 of 35 stream cells. Where α did hold before the break, recalibration on a deployment-sized label budget restores it on none, an oracle that recalibrates at the true break with no delay does no better, and a pre-registered sweep to a larger budget shows that more labels help but do not restore the target; the holds that do occur are thin under question-clustered upper bounds. A rule that certifies selective error directly is the honest alternative, and under a Bonferroni grid it is vacuous in most realizations at either budget, answering a third as often as the conformal rule. The quantity to certify, and the coverage regime in which certifying it is possible, are fixed by the identity before a label is collected.

1 INTRODUCTION

Selective prediction lets a model decline to answer. For a deployed video-language model (VLM) the quantity that matters to a user is the *selective error*: among the questions the system answered, how many were wrong. A gate that keeps it below a user-set level α with a finite-sample, distribution-free guarantee exists for static deployments: calibrating a threshold on an exchangeable held-out split bounds selective error at α with high probability while maximizing coverage (Wang et al., 2025b).

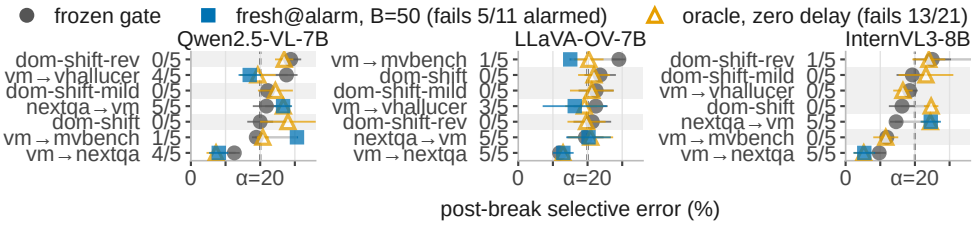


Figure 1: Detect-then-recalibrate does not restore the error target. One row per genre-drift stream cell, sorted by the frozen gate's post-break error (count: confirmed-alarm seeds of 5): pooled post-break selective error of the frozen conformal gate, of instant recalibration on $B = 50$ fresh labels at the alarm (post-recalibration segment), and of a zero-delay oracle at the true break; whiskers ± 1 s.d. across seeds. Gray: the monitor never alarms. Three of the five backbones are drawn (the other two: Table 12; all five: Figures 7 and 8); the legend counts are over all 35 cells, where the oracle sits right of $\alpha = 0.20$ in 21.

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.

Deployment is not a static split: a video assistant's question mix moves with its user base, and the certificate does not detect drift in it. That statically calibrated risk control degrades under drift is established for text streams (Opoku & Banahene, 2026b), with the video-side confidence analogue documented as a diagnostic (Ortiz, 2025); we take it as the premise. The natural repair, which the online-conformal line that adapts a level from observed errors (Gibbs & Candès, 2021; 2024) and budgeted evidence acquisition for VLMs (Xu et al., 2026c) both motivate, is to detect the break and recalibrate on a small batch of fresh labels. This paper tests that mechanism on video QA and finds, at the label budgets a deployment can spend, that it does not restore the target (Figure 1). On most of these streams there is no target to restore: the frozen certified gate is already above α before the break (30 of 35 stream cells, 6 of the 11 alarmed pre-break segments), because it certifies the false-accept rate, not selective error (Table 1, pre-break column).

CertSel is a training-free wrapper around a frozen VLM (Figure 2): a per-genre certified selective gate adopted from COIN; a label-free two-sided $|\Delta\mathrm{Cov}|$ monitor that watches the gate's own answered-coverage stream, confirms alarms, freezes the gate into abstain-all, and triggers recalibration on B fresh labels; and genre-drift streams with known break points built from benchmark metadata over 5 backbones (3B–32B) and 4 benchmarks; the anytime-valid e-process route (Khosravi & Huo, 2026; Hultberg et al., 2026) runs as an ablation arm on the same plumbing.

Contributions. **(1)** A negative result with its cause isolated: instant recalibration on $B = 50$ composition-matched fresh labels holds $\alpha = 0.20$ after recalibration in 10 of 18 alarmed stream cells and a zero-delay oracle in 9 of the same 18 and 14 of all 35 (5 seeds; 5/12 alarmed and 2/14 registered cells at 25, Table 2; question-clustered bounds certify 4 and 7 of those holds), and all 8 of the 8 detected-time failures persist under the oracle (RQ5, Figure 1). The recalibrated gate certifies the false-accept rate; selective error equals that rate times the post-break base error rate divided by coverage, so it exceeds α once $\mathrm{FAR} \cdot \pi_{\mathrm{wrong}} > \alpha \cdot \mathrm{Cov}$, at the certified level coverage below the base error rate, which a pre-registered test confirms on 47 of 50 such entries (Section 5, Figure 5). **(2)** A pre-registered budget sweep, $B \in \{10, 20, 30, 50, 75, 100, 150\}$ over 25 seeds and two recalibration rules: more labels help but do not restore the conformal rule, and a rule that certifies selective error directly is vacuous in 98.7% of realizations at $B = 50$, still in 63.7% at $B = 150$, where it answers a third as often (Figure 6). **(3)** The measurement machinery: a metadata-built genre-drift stream suite and a zero-GPU protocol comparing 9 static arms (seven baselines and our two) and 35 stream cells on one cached pool (Section 4). The $|\Delta\mathrm{Cov}|$ monitor is instrumentation, not a contribution (it alarms where ADWIN does, RQ3); the per-genre layer is an instantiation of the adopted gate, not a contribution: it coincides with COIN wherever genre is the benchmark and is vacuous on Video-MME in 25 and 25 of 25 splits (Section 5).

Scope. The static certificate is adopted from Wang et al. (2025b); that static gates break under drift is attributed to Opoku & Banahene (2026b) and Ortiz (2025); "selective conformal risk control" is a different framework (Xu et al., 2025); and nowhere do we claim that CertSel restores α under drift. On most of these streams there is no error target to restore, because a gate certified for the false-accept rate is already above α before the break.

2 Related Work

Distribution-free risk control. Risk-controlling prediction sets (Bates et al., 2021), their multiple-testing form (Angelopoulos et al., 2022), conformal risk control (Angelopoulos et al., 2024; Angelopoulos, 2026), anytime-valid statements over growing calibration sets (Hultberg et al., 2026) a map of what such certificates promise for generative models (Kotte, 2026), selective prediction under general risk control (Bai & Jin, 2026), and conformal adaptation after online shift detection (Leong, 2026) or configuration shift (Zhu et al., 2026) supply the object we adopt: a finite-sample bound that never touches the predictor, unlike checkers that certify by rule and stay silent elsewhere (Xiao et al., 2026a), and that assumes exchangeability. Weighted conformal prediction (Tibshirani et al., 2019), non-exchangeable coverage bounds (Barber et al., 2023) and robustness over shift balls (Aolaritei et al., 2025) bound a shift's effect without noticing a break or restoring the error among answered items; the anytime-valid branch (Hultberg et al., 2026) survives non-exchangeability by design and is compared against directly (RQ4).

Online and adaptive conformal inference. Adaptive conformal inference holds long-run coverage under shift by updating the working level (Gibbs & Candès, 2021), extended to adversarial shifts, sharpened to interval-wise regret, recast as a game, accelerated by revising past decisions and informed by the distribution (Gibbs & Candès, 2024; Bhatnagar et al., 2023; Principato & Stoltz, 2025; Jun & Ohn, 2025; Hu et al., 2026); we run this family's components as our drift layer. Its guarantee is a stream average a segment can violate, and no member of it certifies the error among answered items or abstains; monitors that fix an alarm point by tuning (Xiao et al., 2025; 2026c) give up the finite-sample statement.

Guarantees for foundation-model outputs. Conformal language modeling, factuality guarantees, conditional and boosted variants and feasibility-aware coverage calibrate sets (Quach et al., 2024; Mohri & Hashimoto, 2024; Cherian et al., 2024; Li et al., 2026; Campos et al., 2024); conformal alignment selects outputs under an FDR constraint (Gui et al., 2024). We adopt the static gate rather than extend it: selective conformal uncertainty screens queries with exchangeability outlier tests (Wang et al., 2025c), and COIN bounds the error among answered questions through Clopper–Pearson intervals while maximizing retention (Wang et al., 2025b), our static layer. A separate framework couples selective classification with risk control (Xu et al., 2025); online abstention under bandit feedback (Lee et al., 2025), anytime-valid selective acting (Khosravi & Huo, 2026), calibrated gates for LLM labelling (Xiao et al., 2026b), tuned agent gates without a bound (Xiao et al., 2026e) and minimal-change repair (Xiao & Li, 2026) complete the family.

Multimodal certification and drift-aware risk control. Sample-then-identify risk control on four video-QA datasets and eight backbones (Wang et al., 2025a) certifies risk on video QA, as we do, but emits sets and is silent on shift. Budgeted conformal evidence acquisition for VLMs shows that naive intervention voids the guarantee (Xu et al., 2026c), on static claims; learned abstention trains the policy (Tayebati et al., 2025), and confidence thresholding on video QA gives smooth in-distribution risk–coverage trade-offs that are not epistemic (Ortiz, 2025). Domain-shift-aware conformal prediction reweights calibration examples by prompt proximity (Lin et al., 2025); drift-aware risk control finds that static control fails once drift arrives while its adaptive variant keeps coverage (Opoku & Banahene, 2026b;a); we take that finding as our premise, with Ortiz (2025) as the video analogue. Evidence-grounded video verification (Xiao et al., 2026f;d) and training-free ensembling of a model's own candidates (Xu et al., 2026b), fused along a surveyed axis (Xu et al., 2026a), are our score's ingredients and issue no risk statement; CertSel measures whether recalibrating after a detected break restores the error target (RQ5).

3 METHOD

Setup. A frozen video-LLM answers a stream of questions $(x_1, g_1), (x_2, g_2), \ldots$, where $g_t \in \mathcal{G}$ is a genre tag observable at answer time (a Video-MME domain, or the originating benchmark). For each question the model emits 9 candidate answers (8 samples plus the greedy decode); a fixed nonconformity score $\hat{p} \in [0, 1]$ is computed from the candidate set: one minus the mean of answer self-consistency, frame-subset agreement and exponentiated confidence.

A *gate* maps $\hat{p}$ to ANSWER/ABSTAIN; the deployment quantities are the *selective error* $\mathrm{SelErr} = P(\text{wrong} \mid \text{ANSWER})$ and the *coverage* $P(\text{ANSWER})$. The user fixes a risk target $\alpha \in (0, 1)$ and a confidence $1 - \delta$; the gate must keep selective error at most α while answering as much as it can. A *segment* is a maximal interval over which the genre mix and the conditional error surface are fixed.

Static layer: adopted certified gates, made group-conditional. The static layer is prior art, adopted: COIN (Wang et al., 2025b) calibrates a threshold on exactly our static object, used here as a component and as our tightest static baseline. Two thresholds live in the layer, per genre g, from a calibration split of n_g labelled examples with n_g^+ wrong ones:

Selective (COIN-style) threshold. Fix a grid of $G{=}20$ candidate thresholds, the empirical quantiles of the calibration scores at levels $0.05, 0.10, \ldots, 1$; the grid reads scores only, never labels. Each candidate is tested with the exact Clopper–Pearson upper bound, at confidence $1 - \delta/G$, on the binomial error rate among the calibration points it answers. The gate keeps the largest-coverage candidate that passes, and abstains on everything when none does.

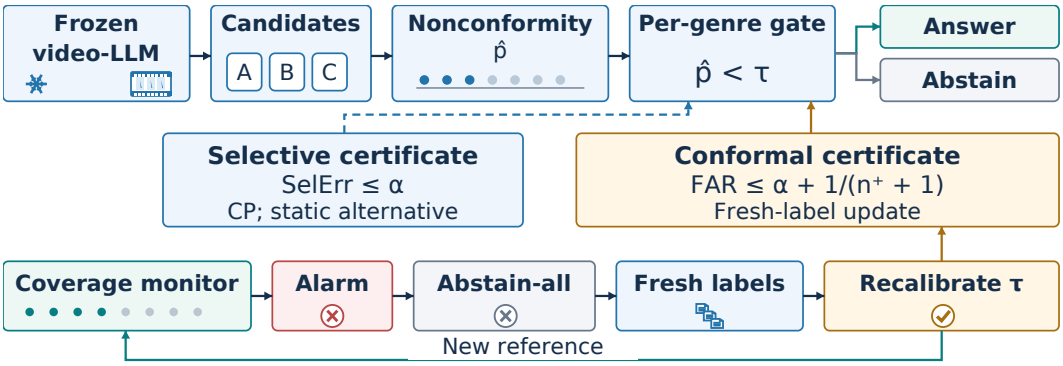


Figure 2: CertSel's gate and drift layer. A frozen video-LLM's candidates give a nonconformity score $\hat{p}$; the per-genre gate answers when $\hat{p} < \tau$. The selective (Clopper–Pearson) bound is static; only the conformal false-accept bound $\alpha + 1/(n^+ + 1)$ is renewed from the B fresh labels after a coverage alarm, during which the gate abstains on everything.

Theorem 1 (Per-genre selective-risk bound; adopted)**.** *Under exchangeability of the calibration and test points within genre g, with probability at least $1-\delta$ over the calibration draw no grid candidate whose selective error exceeds α is certified, so the selected threshold satisfies $P(\mathrm{SelErr}_g \le \alpha) \ge 1-\delta$ for any test point of genre g the gate answers.*

Conformal threshold. The split-conformal gate accepts when $\hat{p} < \tau(\alpha)$, with $\tau(\alpha)$ the *low* quantile $Q_{\alpha+1/(n_g^+ +1)}$ of the wrong-stratum scores, certifying the marginal false-accept rate $P(\textsc{answer} \mid \text{wrong}, g) \le \alpha + 1/(n_g^+ + 1)$; it is the cheaper certificate (no binomial tails, no δ) and the gate whose answered-coverage stream the drift monitor watches.

Corollary 1 (Monotone operating point)**.** *For both thresholds, certified coverage is monotone non-decreasing in α: relaxing the risk target never shrinks the answered set on a fixed calibration split.*

Both are standard machinery (exact binomial tails in the Learn-then-Test style (Angelopoulos et al., 2022); split-conformal quantiles) restated per genre, so easy genres cannot compensate for hard ones. The cost is granularity: a genre with few calibration points certifies nothing and abstains, the vacuous regime reported in RQ1.

Drift layer: detect, protect, recalibrate. Theorem 1 assumes a stationary world; the drift layer is built to restore its premise per segment, and whether the third state restores the target is measured at $B = 50$, $\alpha = 0.20$ (RQ5); the answer there is no.

Monitor (label-free). The gate's own answer decisions form a binary stream. A two-sided statistic $|\Delta\mathrm{Cov}| = |\overline{C}_w - \hat{c}_{\mathrm{ref}}|$ compares the coverage of a sliding window of width $w = 40$ against the calibration coverage $\hat{c}_{\mathrm{ref}}$, standardized by a two-sample binomial σ that accounts for the finite reference split; a coverage *rise* is as much evidence as a drop (RQ3). An alarm requires $z = 3.0$ standard deviations sustained for 5 consecutive windows. **Protect.** A confirmed alarm flips the gate to a conservative abstain-all fallback: while the system does not know which segment it is in, it does not answer. **Recalibrate.** The protocol spends a fresh-label budget of $B = 50$ items from the new segment, recomputes the *conformal* threshold as the low quantile of the fresh wrong-stratum scores, and re-enters monitoring with the recalibrated coverage as the new reference. This is exactly what the released code runs (Figure 2): the selective (Clopper–Pearson) threshold is not recomputed, so the certificate the recalibrated gate carries forward is the false-accept bound $\alpha + 1/(n^+ + 1)$, n^+ the wrong items among the B fresh labels, not Theorem 1; selective error after recalibration is an uncertified quantity, and RQ5 and Section 5 measure both it and the certified false-accept rate. Section 5 also evaluates offline a rule that recomputes the selective threshold instead, and prints the delay window's own error; the oracle arm of RQ5 removes that window entirely. **Baseline arm: the anytime-valid alternative.** The anytime-valid alternative to monitor-and-recalibrate is an e-process over the answered stream: a nonnegative supermartingale under $H_0 : \mathrm{SelErr} \le \alpha$ whose threshold crossing is an anytime-valid refutation, immune to drift by construction (Khosravi & Huo, 2026; Hultberg et al., 2026). We run a betting e-process on our own alarm and recalibration pipeline ($\delta_e = 0.05$), isolating the one difference: never invalid, at a cost in power (RQ4).

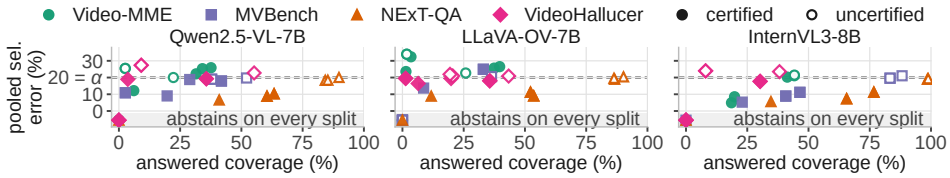


Figure 3: The static protocol at $\alpha = 0.20$: one point per backbone $\times$ benchmark $\times$ arm, mean answered coverage over 25 splits against pooled selective error; certified arms filled, uncertified hollow. The lower strip holds arms that abstain on every split and so have no selective error. Per cell, with seed spreads: Table 6.

4 Experiments

Pool. All experiments replay one cached candidate pool on CPU: 5 backbones (Qwen2.5-VL at 3B, 7B and 32B-AWQ, LLaVA-OneVision-7B, InternVL3-8B) $\times$ 4 video-QA benchmarks (Video-MME, MVBench, NExT-QA, VideoHallucer) $\times$ 500 questions $\times$ 9 candidates, generated once on a single 80 GB data-centre GPU and consumed read-only. Every number below is a deterministic function of this pool and a pinned 25-seed list; the matrix runs in 150.7 s of CPU wall clock.

Protocols. The *static* protocol splits each benchmark cell 0.5/0.5 into calibration and evaluation per seed, calibrates every arm at $\alpha \in \{0.05, 0.10, 0.20\}$, $\delta = 0.02$, and reports selective error against coverage. The *drift* protocol builds 7 streams per backbone with known break points: 3 Video-MME domain-composition shifts (both directions, plus a diluted post_mix $= 0.5$ strength) and 4 cross-benchmark concatenations. Stream arms run end to end: the CertSel protocol, the e-process arm, ACI/DtACI-style online adapters, a windowed recalibrating CRC-style arm, a SConU-style screen, and the two frozen static gates, with the reference detectors scored against the anchored break points. Main tables attach ± 1 sample s.d. across the 25 seeds in small type (over answering seeds only); alarm rates are seed counts with Wilson 95% intervals.

Baselines and deferred arms. 9 static arms run on the same pool: our two, the adopted COIN gate (Wang et al., 2025b), a TRON-style gate on the sampling-frequency score (Wang et al., 2025a), Conformal-Alignment-style FDR selection (Gui et al., 2024), SConU-style screening (Wang et al., 2025c), a learned abstention contrast (Tayebati et al., 2025), and two uncertified heuristic thresholds. Five design-list baselines are deferred with explicit blockers (BCEA, DS-CP, VL-Uncertainty, distribution-informed online CP, boosted/conditional conformal; Appendix K): each needs an input the cached pool lacks or a paper spec we refuse to guess.

RQ1: static certification, per benchmark cell. Figure 3 is the static matrix at $\alpha = 0.20$ (per-cell values: Table 6, Appendix D). The adopted COIN gate and our per-genre selective layer, the same static object on three of the four cells, hold α wherever they answer (at least 35 pooled accepts; no bold entry in those rows); on LLaVA-OV Video-MME both abstain in all 25 splits. Certification also has a coverage cost: the selective gate abstains on Video-MME in 25 and 25 of 25 splits, and where COIN answers it keeps near-zero coverage. **Static certification holds wherever the gate answers at scale (≥ 35 pooled accepts), and is vacuous on Video-MME.**

RQ2: genre drift. Who keeps α per segment? Table 1 pools each arm's post-break segment at $\alpha = 0.20$ over 25 stream seeds and 35 cells. The static gates bracket the failure modes: the frozen conformal gate keeps its coverage and violates α post-break in 17 of 35 cells, while the frozen COIN gate violates in 4 of the 5 nextqa$\rightarrow$vm cells, all still answering at scale, and elsewhere stays valid by going silent. The online adapters (ACI/DtACI-style, windowed CRC-style) sit between the brackets, as their average-coverage guarantee predicts.

CertSel itself violates the pooled post-segment target in 19 of 35 cells: 10 on the gentle domain shifts, where the label-free monitor almost never fires (at most 3 of 25 seeds per cell), a detector-power gap; the other 9 on abrupt cells it does alarm on, where the pooled number mixes the delay window with the recalibrated tail. **Post-break, CertSel violates α in 19 of 35 cells, no fewer than the 17 of the frozen gate it wraps (the per-cell differences are inside the cross-seed s.d., Table 3): the protocol does not restore validity.**

Table 1: Drift matrix at $\alpha = 0.20$: pooled post-break selective error (%) per arm over 25 stream seeds; CertSel columns carry ± 1 s.d. First column: the frozen conformal gate *before* the break, over α on 30 of 35 cells because it certifies the false-accept rate (the three vm→ streams of a backbone share one pre-break segment, so these are 20 of 25 distinct segments). **Bold**: certified arm over target (unrounded); –: nothing answered; last column: CertSel post-break coverage. Accept counts behind every entry are in Table 9. The omitted SConU-style arm equals conf. frozen everywhere; segment lengths in Table 5. Three of the 5 backbones are printed; Qwen2.5-VL-3B and 32B-AWQ are in Table 12, and every count in the text is over all 35 cells.

		pre-break	post-break selective error (%)							
Backbone	Stream	conf. frozen	COIN frozen	conf. frozen	ACI -style	DtACI -style	wCRC -style	e-proc (C3)	CertSel	CertSel cov
Qwen2.5-VL	dom-shift	24.6	–	20.0	18.5	15.7	22.8	**20.1**	20.0±7.2	22.1±4.9
	dom-shift-rev	25.5	–	**28.5**	26.3	24.8	27.3	**28.5**	**28.3**±4.6	45.3±7.8
	dom-shift-mild	25.6	–	**22.2**	20.6	19.1	25.2	**22.2**	**22.5**±5.4	27.8±4.3
	vm→nextqa	25.3	–	11.0	13.8	12.7	10.2	11.3	9.9±3.9	49.7±6.2
	vm→mvbench	25.3	–	19.0	17.9	17.7	18.6	19.6	19.1±3.0	43.1±5.7
	vm→vhallucer	25.3	–	**25.7**	22.1	21.3	21.2	**25.5**	**21.0**±4.3	33.9±8.9
	nextqa→vm	8.8	**23.7**	**21.4**	27.2	25.2	25.8	**21.4**	**24.4**±5.3	28.2±7.5
LLaVA-OV	dom-shift	25.5	–	**28.3**	25.3	21.4	27.5	**28.5**	**28.6**±5.0	38.7±5.8
	dom-shift-rev	26.7	–	**23.6**	21.6	20.6	25.1	**23.1**	**23.6**±4.4	35.0±5.3
	dom-shift-mild	25.5	–	**26.8**	23.4	21.6	25.7	**26.6**	**26.8**±5.4	38.7±5.1
	vm→nextqa	25.6	–	11.0	13.0	12.4	10.3	11.0	9.7±3.0	45.3±9.4
	vm→mvbench	25.6	–	**29.2**	23.7	21.4	27.3	**29.4**	**29.8**±3.1	43.6±6.3
	vm→vhallucer	25.6	–	**23.6**	20.2	19.2	19.8	**23.2**	**22.9**±3.3	42.5±8.1
	nextqa→vm	9.0	**24.8**	17.3	28.5	24.1	25.7	17.3	**23.9**±5.7	27.3±6.4
InternVL3	dom-shift	20.3	0.0	15.6	16.4	15.8	17.9	15.6	15.6±3.6	36.8±4.4
	dom-shift-rev	20.4	–	**24.2**	22.7	21.2	21.4	**24.0**	**24.1**±3.8	48.8±3.9
	dom-shift-mild	21.0	0.5	17.8	18.5	17.7	19.7	17.6	17.8±5.1	38.8±4.9
	vm→nextqa	20.2	3.7	8.9	12.2	10.3	7.5	8.8	7.3±2.7	52.2±5.9
	vm→mvbench	20.2	3.8	10.5	13.7	11.7	11.5	10.5	10.8±2.6	45.2±3.1
	vm→vhallucer	20.2	6.4	17.9	18.7	18.3	18.5	17.8	18.0±4.3	27.5±9.8
	nextqa→vm	7.3	**24.9**	13.5	25.8	20.3	19.2	13.5	**20.6**±5.6	33.1±5.4

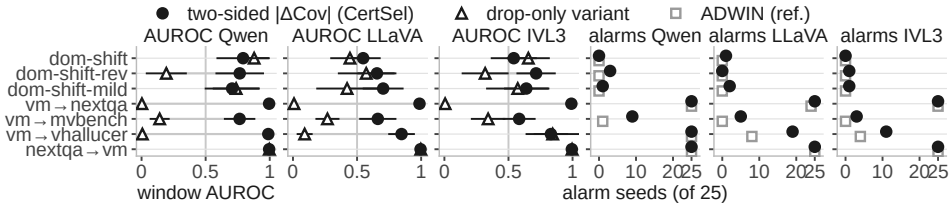


Figure 4: Two-sidedness detects coverage-*rise* breaks: the drop-only $|\Delta\mathrm{Cov}|$ variant collapses where the two-sided score separates (left pair; mean AUROC over 25 seeds, whiskers ± 1 s.d., gray line = chance); the confirmed-alarm rule reaches the caught bar on 7 of the 12 plotted abrupt cells (no vm→mvbench cell), every gentle shift near zero (right pair; CertSel vs ADWIN). Per-cell delays and false alarms: Table 10.

RQ3: detection quality, measured. Figure 4 shows the detection layer against the anchored break points (per-cell table: Table 10, Appendix G). The two-sided $|\Delta\mathrm{Cov}|$ score reaches AUROC 0.541–0.995 across cells; the drop-only variant collapses to near zero on the coverage-*rise* streams (e.g. vm→nextqa). Two bars recur: a cell is *caught* when the confirmed-alarm rule fires post-break in at least 20 of 25 seeds (80%), *alarmed* when it fires in one seed of the isolation tier's 5-seed list, the weaker bar that admits cells to RQ5. CertSel catches 11 of the 20 abrupt cross-benchmark cells: both nextqa-direction shifts on every backbone but Qwen2.5-VL-32B (vm→nextqa only) and vm→vhallucer on Qwen2.5-VL-7B and 32B. Qwen2.5-VL-32B nextqa→vm and LLaVA-OV and InternVL3 vm→vhallucer confirm in 19/25, 19/25 and 11/25 seeds, under the bar (Wilson 95% CIs in Table 10), and every vm→mvbench cell is missed (at most 9/25 seeds). Mean per-cell delays over the caught cells span 30.3–95.7 steps; the 35 cells carry 12 pre-boundary false alarms (Table 10). No gentle cell alarms in over 3 of 25 seeds (upper Wilson bounds $\leq 30\%$): the statistic is underpowered at $w = 40$, and gentle cells account for 10 of CertSel's 19 post-break violations. **Two-sidedness detects coverage-rise breaks: 11 of 20 abrupt cells are caught, no gentle shift.**

The rule's one parameter, z (pinned at 3.0), is walked over $z \in \{2.5/4.0\}$ in Appendix H: sensitivity and false alarms trade off, but the pooled post-break selective error spreads at most 5.5 points in any cell, so we report the curve rather than a tuning.

RQ4: the anytime-valid arm, head-to-head (C3). The e-process arm shares CertSel's alarm and recalibration pipeline, so Table 1's e-proc column against the CertSel column is the power-versus-safety comparison, and the detection counts decide it. The e-process fires in no seed on the four registered nextqa-direction cells (once in 25 on the 3B twin), where the coverage monitor fires in 25/25, 25/25, 25/25 and 25/25; on the other four abrupt cells it fires in 1/25, 9/25, 13/25 and 4/25 seeds with conditional mean delays of 16.0, 130.3, 117.2 and 131.0 steps, against 30.3–95.7 for CertSel on its caught cells. On the gentle shifts the ranking reverses: the e-process alarms more often than the coverage monitor in 7 of 15 cells (up to 17/25 against 0/25), and those cells carry 10 of CertSel's 19 violations. Where the null is already false pre-break (the conformal threshold certifies the false-accept rate, so $H_0 : \mathrm{SelErr} \leq \alpha$ can be false on calibration-like data) a pre-break e-process alarm is a true refutation, not a false alarm (Table 4). **Guaranteed safety costs nearly all of the detection power on the abrupt cells (at most 13/25, and at most one seed on the nextqa-direction ones) and buys back part of it on the gentle cells, out-alarming our own monitor on 7 of 15.**

RQ5: recalibration isolated from detection delay (GPU tier). Table 1's pooled numbers mix detection speed, the stale gate's cost during the delay window, and what the B-label fresh recalibration does on the new segment. A 9.4-GPU-hour fresh-label sweep (31050 release rows on held-out questions; Appendix M) replays every stream cell under 5 recalibration policies (Appendix I), with fresh draws composition-matched to the realized post segment (Table 12).

The isolated verdict at $\alpha = 0.20$ is the paper's central finding. Of 35 stream cells, 18 alarm (the at-least-one-seed bar of RQ3 at this tier's 5 seeds): 17 of the 20 abrupt cells, no gentle one, 5 of them in a single seed (at the three-seed bar, 8 of 13 hold below). Fresh@alarm, instant fresh recalibration, holds the pooled post-recalibration target in 10 of those 18 (7 under the stricter every-seed reading). The oracle arm isolates the cause: in all 8 of the 8 failing alarmed cells, recalibrating at the true boundary with zero delay also fails; on the same 18 alarmed cells the oracle holds α in 9 against fresh@alarm's 10, and over all 35 cells in 14. Against the frozen control on matched post segments, recalibration improves the error in 10 of 18 alarmed cells and worsens it in 8 (all but two differences inside the cell's cross-seed s.d.); 4 of the 10 holding cells were brought back under α, the other 6 already held frozen. On 13 of the 18 the frozen gate is already over α *before* the break (20.2–31.0% on 30 of 35 cells, Table 1), so there recalibration is asked to reach α, not to restore it; on the 5 where it held (7.2–11.4%) it fails on all 5. Every verdict here is a pooled point estimate over seeds that re-ask the same questions, so the 95% upper bounds of Table 12 resample questions, not accepts. They certify 4 of the 10 post-recal holds and 7 of the 14 oracle holds; the thinnest hold rests on 20 accepts from one alarming seed (bound 30.0%), the thinnest alarmed pool of all, 18 accepts, fails, and the tightest failure misses by 1 wrong accept in 291. The 17 silent cells are RQ2's detector-power gap and count for neither side. **On the 5 alarmed cells where α held before the break, recalibration restores it on none, and zero-delay oracle recalibration holds on 9 of the same 18 alarmed cells and fails in 21 of all 35: the bottleneck is not the detector or its delay but what $B = 50$ fresh labels can certify.**

5 Analysis: why budgeted recalibration does not restore α

3 of the 5 predictions written before these analyses ran (`PREREG.md`, Appendix J) held.

The certified quantity is met; the reported one fails. Recalibration recomputes the conformal threshold (Section 3), so the certificate it carries forward is a false-accept bound $\mathrm{FAR} \leq \alpha + 1/(n^+ + 1)$ on the fresh draw. Selective error is tied to it by an identity that holds on every segment, $\mathrm{SelErr} = \mathrm{FAR} \cdot \pi_{\mathrm{wrong}}/\mathrm{Cov}$, with π_{wrong} the post-break base error rate and Cov the coverage. The identity is not the finding; where $\mathrm{FAR} \cdot \pi_{\mathrm{wrong}} > \alpha \cdot \mathrm{Cov}$ holds after a break is. On Qwen nextqa→vm after instant recalibration the false-accept rate is 23.0%, inside the recorded certificate of 24.3%; the base error rate is 45.3% against a coverage of 39.1%, so the selective error is 26.6%, the target missed with the certificate intact. Over alarmed cells $\times$ three α the pooled post-recalibration false-accept rate is within the recorded bound in 62 of 68 cell-arms, and in 50 of the 54 whose selective error exceeds α (Figure 5, right); per realization it is marginal, 61% landing inside their own bound at $B = 50$ over 25 seeds. P1 held.

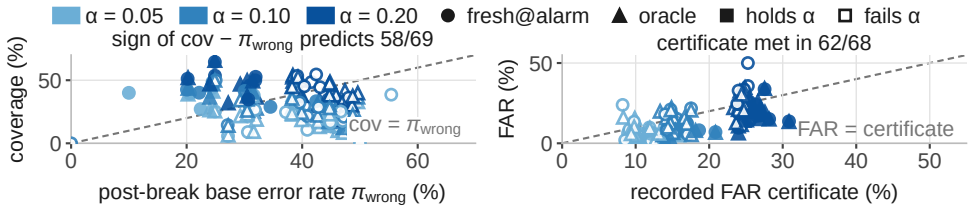


Figure 5: Mechanism, one point per stream cell $\times$ α $\times$ recalibration arm (circle: fresh@alarm; triangle: oracle) on the isolation table's post-recalibration segment (filled: holds α; open: fails). Left: post-recalibration coverage against the post-break base error rate; the diagonal is the pre-registered decision line. Right: realized post-recalibration false-accept rate (FAR) against the recorded certificate $\alpha + 1/(n^+ + 1)$; points below the diagonal meet it, including most open ones.

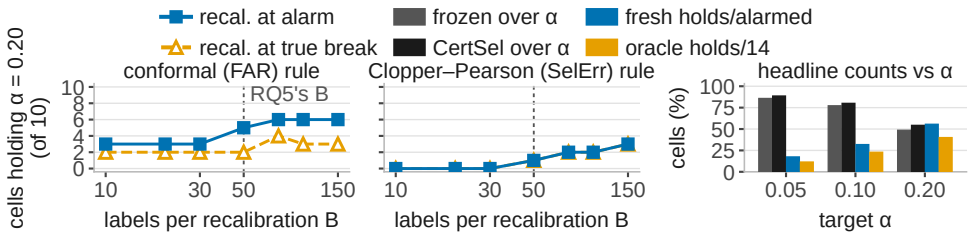


Figure 6: Recalibration budget, $\alpha = 0.20$, 25 seeds. Left, middle: cells (of the 10 reaching $B = 150$) holding α after recalibration at the alarm and at the true break, per rule; right: headline counts at every α over their own denominators.

Coverage against the base error rate. The exact condition is $\mathrm{FAR} \cdot \pi_{\mathrm{wrong}} \leq \alpha \cdot \mathrm{Cov}$, which at the certified level $\mathrm{FAR} = \alpha$ is coverage at least the base error rate. We pre-registered that the sign of $\mathrm{Cov} - \pi_{\mathrm{wrong}}$ predicts the measured verdict in at least 80% of (cell, α, arm) entries; it does in 58 of 69 (84%; Figure 5, left), but the two directions differ: 47 of the 50 entries with coverage below the base error rate fail, while only 11 of the 19 above it hold, so the diagonal is an approximate discriminant, not the condition: an entry above it fails only when its false-accept rate sits above α, which the identity forces. The failure is structural: the oracle fails with the detected-time arm (P2 held).

Budget and rule. Figure 6 and Table 2 report a sweep over $B \in \{10, 20, 30, 50, 75, 100, 150\}$ on the CPU matrix's 25 seeds with the isolation table's fresh draws and monitor, for the shipped conformal rule and for a Clopper–Pearson rule that recomputes the selective threshold of Theorem 1 ($\delta = 0.02$) on the fresh draw and abstains on everything when no threshold certifies, each at the alarm and at the true break; both are online policies on the same monitor and fresh draw, identical up to the first confirmed alarm. On 4 cells the matched draw exhausts a fresh Video-MME domain at $B = 100$, so the curves run over the 10 cells that reach $B = 150$ (eight capped cells were predicted; P5 failed). For the conformal rule the budget helps but does not restore: instant recalibration holds $\alpha = 0.20$ in 5 of 10 cells at $B = 50$ and 6 at $B = 150$, the oracle in 2 and 3, and at $\alpha = 0.05$ the oracle fails 13 of the 14 registered cells at the largest budget every cell reaches (P3 held; 32 of 35 over all backbones; Qwen2.5-VL at 3B and 32B-AWQ changes none of this, and nextqa→vm fails α on every backbone, Appendix J); at 25 seeds and $B = 50$ the oracle holds on 2 of the 14 cells (5 in the five-seed table) and instant recalibration on 5 of 12 alarmed cells; the alarmed set grows with the seed count (8 at 5) because its bar is a resampled statistic, so both readings are reported. The Clopper–Pearson rule certifies the right quantity at two costs, measured with the 20-candidate Bonferroni grid at $\delta = 0.02$ (a fixed-sequence scan or a larger δ would be less conservative): at $B = 50$ its certificate is vacuous in 98.7% of the alarmed (cell, seed) realizations on these 10 cells at $\alpha = 0.20$ and 100.0% at $\alpha = 0.05$, and at no budget in the grid does it hold α in more cells than the conformal rule (3 against 6 at $B = 150$, alarm timing) at a coverage of 13.4% against 41.4% (P4: vacuity held, the lead clause failed). **Timing and budget help the conformal rule but do not restore it; the rule that certifies selective error on the 20-candidate grid is still vacuous in 63.7% of realizations at $B = 150$ labels per event and answers a third as often (coverage ratio 0.324 at the alarm).**

Table 2: Budget sweep, $\alpha = 0.20$, 25 seeds: cells holding α after recalibration (of the 10 reaching every budget) per rule and timing; vacuous CP certificates (% of alarmed cell–seed realizations); mean coverage at the alarm; last two columns: the 14 registered cells (twins: Appendix J).

	conformal rule		CP rule		CP	coverage (%)		all 14, conformal	
B	alarm	oracle	alarm	oracle	vacuous (%)	conformal	CP	alarm	oracle
10	3	2	0	0	100.0	51.1	0.0	4/12	2
20	3	2	0	0	100.0	48.4	0.0	4/12	2
30	3	2	0	0	100.0	46.7	0.0	4/12	2
50	5	2	1	1	98.7	42.7	0.5	5/12	2
75	6	4	2	2	86.5	41.6	4.1	6/12	4
100	6	3	2	2	70.7	41.3	8.5	–	–
150	6	3	3	3	63.7	41.4	13.4	–	–

Table 3: Cost of protection on the caught cells, $\alpha = 0.20$, 25 seeds: post-break coverage of the frozen gate and of CertSel (%), their difference (points), post-break selective error (%), mean delay (steps; the abstain window adds $B = 50$) and pre-break false alarms, each spending B labels.

		coverage (%)			sel-err (%)			
Backbone	Stream	frozen	CertSel	Δ	frozen	CertSel	delay	FA
Qwen2.5-VL	vm→nextqa	67.1	49.7	+17.4	11.0	9.9	43.1	2
Qwen2.5-VL	vm→vhallucer	64.7	33.9	+30.8	25.7	21.0	43.7	2
Qwen2.5-VL	nextqa→vm	25.6	28.2	−2.6	21.4	24.4	30.3	0
LLaVA-OV	vm→nextqa	64.4	45.3	+19.0	11.0	9.7	45.9	1
LLaVA-OV	nextqa→vm	22.5	27.3	−4.8	17.3	23.9	32.6	0
InternVL3	vm→nextqa	71.2	52.2	+18.9	8.9	7.3	36.9	0
InternVL3	nextqa→vm	34.6	33.1	+1.4	13.5	20.6	43.5	1
Qwen2.5-VL-3B	vm→nextqa	62.8	43.6	+19.2	16.5	14.2	48.2	0
Qwen2.5-VL-3B	nextqa→vm	24.7	26.6	−1.9	22.4	28.3	67.6	0
Qwen2.5-VL-32B	vm→nextqa	57.3	46.0	+11.4	12.4	11.5	75.1	0
Qwen2.5-VL-32B	vm→vhallucer	54.8	43.1	+11.7	20.0	18.5	95.7	0

The delay window. The selective error between the break and the confirmed alarm, decided under the stale threshold, exceeds the post-recalibration error in 5 of 8 registered alarmed cells (Qwen nextqa→vm: 35.0% against **26.6**%); removing it (the oracle arm) does not suffice.

What protection costs. Table 3 gives the cost of the drift layer on the 11 caught cells. On the 8 where CertSel answers less, the frozen gate is inside α on all but Qwen2.5-VL-7B vm→vhallucer, yet the protocol abstains for the delay window plus B steps and recalibrates to a coverage 1.4–30.8 points lower; on the other 3, all nextqa→vm, it covers more (up to 4.8 points) and errs more. The noise floor is the same-arm cross-seed s.d. of post-break selective error, 2.5–7.3 points; 31 of the 35 CertSel-versus-frozen differences in Table 1 are inside it, so we do not rank the two.

Negative results. The SConU-style screen equals the frozen conformal gate item for item on 35 of 35 cells at every α. The per-genre selective layer coincides with the adopted COIN gate on 59 of 60 static entries, and the 1 entry where they differ is vacuous under both (Appendices D and E).

Limitations. The Clopper–Pearson arm is not re-armed after a vacuous certificate (Appendix J). The pool is 5 backbones (3B–32B) and 4 benchmarks at 500 questions per cell; genre tags are assumed observable; streams are metadata-built; the monitor misses gentle shifts at $w = 40$.

6 Conclusion

CertSel measures whether detect-then-recalibrate restores the target under genre drift. It does not: at $B = 50$ fresh labels, 10 of 18 alarmed cells hold $\alpha = 0.20$ after recalibration, a zero-delay oracle 9 of the same 18 and 14 of all 35. On 13 alarmed cells the frozen gate was over α before the break; elsewhere recalibration restored it on none. The gate certifies a false-accept rate, and selective error exceeds α once that rate times the base error rate exceeds α times coverage. More labels help but do not restore: a rule certifying selective error on a 20-candidate Bonferroni grid at $\delta = 0.02$ is still vacuous in 63.7% of realizations at 150 labels and answers a third as often. The lesson carries to process supervision: check what a certificate bounds and whether it survives drift.

REPRODUCIBILITY STATEMENT

Every certificate, drift replay, and detector statistic in this paper is deterministic post-processing over two provenanced inputs: the cached candidate pool (release sha `5fd8f57c`), replayed by both tiers, and the fresh-label release (sha `10f5c506`, 9.4 GPU-hours) additionally consumed by the recalibration-isolation tier; the macro generator refuses to build if the two tiers record different pool shas. Each contract file has exactly one seeded writer (`experiments/certsel/run_matrix.py`, `recal_isolation.py`, `recal_budget_sweep.py`, and the read-only `analysis_recal_mechanism.py`, whose predictions were registered in `PREREG.md` before its first run and whose failures are reported as failures); `cluster_bounds.py` writes a fifth, derived file of question-clustered bounds. Every numeral in the body and tables is generated from those five files by one macro generator (`gen_macros.py`), which refuses to build if any of their recorded shas has moved; no number is transcribed by hand. The scripts, the five result files, the candidate pool (`experiments/certsel/data/pool/benchmark_release.jsonl`), the fresh-label release and unit tests are in the supplementary zip; start at `supp/README_supp.md`. Appendix M details the pool schemas, the protocol knobs, and the compute ledger.

AI USE STATEMENT

We disclose our use of generative AI by the categories the ICLR 2027 policy lists. Within the categories that require disclosure, these tools helped us settle the conceptual frame, put forward and sharpen the hypotheses under test, write down the formal claims and the statistical treatment behind them, comment on the experimental design and the research methodology, build the released code that runs the experiments, analyses them and draws the figures, tidy dataset files into a common format, and help interpret the results. Four further categories do not arise: no synthetic dataset was built, nothing was translated, no thematic or qualitative analysis of data took place, and the paper proves nothing new: its one theorem and corollary are adopted from cited prior work. The rest of the required list is out of scope: no human subjects are involved, and nothing was gathered beyond published benchmarks and what open-weight checkpoints produced on them. Among the categories the policy only recommends disclosing, they also served for brainstorming, literature search and reading, software code, preparing figures, and editing this paper's prose. They were kept away from the results: no measurement here was produced, chosen or discarded by such a tool. Every experimental quantity comes out of one committed program, the only writer of its result file, and reaches the source as a generated macro, so no such value can be typed in by hand; what we do type are protocol settings, design constants, and values quoted from earlier work with a citation. The questions this paper asks, its claims and the verdict it reaches belong to the authors, who checked every AI-assisted output and answer for all of it, each number and each released artifact included.

ETHICS STATEMENT

All five backbones are open-weight checkpoints their developers published, and the four evaluation sets (Video-MME, MVBench, NExT-QA, VideoHallucer) are distributed publicly; we ran no data collection and recruited no one. The videos show people who have no connection to us, so the artifact ships no footage at all (no frame, no clip, no still) and no candidate answers. Both the cached pool and the fresh-label release are tables of numbers: each row carries the model and the evaluation set, the question identifier that set assigns, a candidate index, a single bit for whether the answer was right, and the signal values. Anyone repeating the experiment gets the video files from the benchmark authors on the conditions those authors set, which is how we got them. On the method, the misreading worth avoiding is treating a satisfied certificate as a safety property. What is certified is a bounded rate of the marker the selector was calibrated against, answer error, under the exchangeability the calibration split assumes, and the paper's own drift experiment is a study of what happens when that assumption breaks. It bounds nothing about harm, and a deployment would need a loss marker that tracks the harm of concern and a calibration sample from the deployment distribution; neither is supplied here.

REFERENCES

Anastasios N. Angelopoulos. Conformal risk control for non-monotonic losses. arXiv:2602.20151, 2026.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. arXiv:2110.01052, 2022.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2208.02814. arXiv:2208.02814.

Liviu Aolaritei, Zheyu Oliver Wang, Julie Zhu, et al. Conformal prediction under levy-prokhorov distribution shifts: Robustness to local and global perturbations. arXiv:2502.14105, 2025.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. arXiv:2603.24704, 2026.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 2023. doi: 10.1214/23-aos2276. arXiv:2202.13415.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, et al. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 2021. doi: 10.1145/3478535. arXiv:2101.02703.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning (ICML)*, 2023. doi: 10.48550/arXiv.2302.07869. arXiv:2302.07869.

Margarida M. Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal prediction for natural language processing: A survey. arXiv:2405.01976, 2024.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.09714.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. doi: 10.48550/arXiv.2106.00170. arXiv:2106.00170.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024. arXiv:2208.08401.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. arXiv:2405.10301, 2024.

Dongjian Hu, Junxi Wu, Shu-Tao Xia, et al. Distribution-informed online conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2026. doi: 10.48550/arXiv.2512.07770. arXiv:2512.07770.

Bror Hultberg, Dave Zachariah, and Antônio H. Ribeiro. Anytime-valid conformal risk control. arXiv:2602.04364, 2026.

Jungbin Jun and Ilsang Ohn. Online conformal inference with retrospective adjustment for faster adaptation to distribution shift. arXiv:2511.04275, 2025.

Hamed Khosravi and Xiaoming Huo. Conformal selective acting: Anytime-valid risk control for RLVR-trained LLMs. arXiv:2605.20270, 2026.

Varun Kotte. When can conformal risk control certify LLM outputs? bounds, impossibility, and adaptation for structured generation. arXiv:2606.29054, 2026.

Minjae Lee, Yoonjae Jung, and Sangdon Park. Online conformal abstention for factuality control under adversarial bandit feedback. arXiv:2506.14067, 2025.

Jun Wen Leong. Online shift detection and conformal adaptation for deployed safety classifiers. arXiv:2606.11949, 2026.

Ye Li, Anqi Hu, Yuanchang Ye, et al. Set-valued prediction for large language models with feasibility-aware coverage guarantees. arXiv:2603.22966, 2026.

Zhexiao Lin, Yuanyuan Li, Neeraj Sarna, et al. Domain-shift-aware conformal prediction for large language models. arXiv:2510.05566, 2025.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. arXiv:2402.10978, 2024.

Jeffery Opoku and David Banahene. Drift-aware spectral conformal prediction for non-exchangeable streaming data. arXiv:2606.15953, 2026a.

Jeffery Opoku and David Banahene. PromptShift-CRC: Drift-aware conformal risk control for foundation models under prompt and domain shift. arXiv:2606.15964, 2026b.

Jorge Ortiz. Explicit abstention knobs for predictable reliability in video question answering. arXiv:2601.00138, 2025.

Guillaume Principato and Gilles Stoltz. Adaptive conformal inference through the lens of blackwell approachability. arXiv:2510.15824, 2025.

Victor Quach, Adam Fisch, Tal Schuster, et al. Conformal language modeling. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.10193.

Sina Tayebati, Divake Kumar, Nastaran Darabi, et al. Learning conformal abstention policies for adaptive risk management in large language and vision-language models. arXiv:2502.06884, 2025.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 2526–2536, 2019. arXiv:1904.06019.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, et al. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2410.08174.

Zhiyuan Wang, Jinhao Duan, Qingni Wang, et al. COIN: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. arXiv:2506.20178, 2025b.

Zhiyuan Wang, Qingni Wang, Yue Zhang, et al. SConU: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Main Conference*, 2025c. doi: 10.18653/v1/2025.acl-long.934. arXiv:2504.14154.

Yifan Xiao and Shijie Li. ARGUS: Argumentation-based minimal-change repair for verifiable LLM self-explanations. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2026.

Yifan Xiao, Zhongjun Wang, Tiantian Yan, Weiwei Li, Dazhong Wu, and Jie Ma. Time series anomaly detection: A hybrid deep learning framework with adaptive thresholding for electricity industry data. In *IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)*, 2025. doi: 10.1109/CSCLOUD66326.2025.00046.

Yifan Xiao, Shijie Li, and Yuhao Ge. RunbookFX: Type- and effect-safe LLM synthesis for executable incident diagnosis and mitigation. *Proceedings of the ACM on Programming Languages (ICFP)*, 2026a. doi: 10.1145/3828675.

Yifan Xiao, Shijie Li, and Yu Huang. CALM-TS: Risk-controlled LLM labeling for time-series via calibrated selective gating. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026b. doi: 10.1145/3770855.3817859.

Yifan Xiao, Shijie Li, and Yu Huang. CodeFault: Predicting fault risks from code changes via multi-modal learning. In *Proceedings of the IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2026c.

Yifan Xiao, Shijie Li, and Yu Huang. ProgV-Trace: An LLM-driven verification framework for evidence-grounded video misinformation attribution. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2026d. doi: 10.1145/3767308.3836284.

Yifan Xiao, Dexuan Xu, Shijie Li, Weiping Ding, and Yu Huang. Fuzzy-gated verification of LLM agents for industrial fault diagnosis: An evolutionary-optimized closed-loop framework. *Applied Soft Computing*, 2026e. In press.

Yifan Xiao, Dexuan Xu, Shijie Li, and Yu Huang. Plan-retrieve-verify: An LLM-driven framework for long-video audio-visual evidence reasoning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2026f. doi: 10.1145/3767308.3836283.

Dexuan Xu, Yanyuan Chen, Zhongyan Chai, Yifan Xiao, Yandong Yan, Weiping Ding, Hanpin Wang, Zhi Jin, Wenpin Jiao, Weihua Yue, Hang Li, and Yu Huang. Knowledge fusion in deep learning-based medical vision-language models: A review. *Information Fusion*, 125:103455, 2026a.

Dexuan Xu, Yifan Xiao, Shijie Li, Xingyu Yang, Siqi Hou, Hanpin Wang, and Yu Huang. Beyond one granularity: Spectral ensemble and consensus expansion for training-free composed image retrieval. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2026b. doi: 10.1145/3767308.3835648.

Jian Xu, Delu Zeng, John Paisley, and Qibin Zhao. Look again before you abstain: Budgeted conformal evidence acquisition for reliable vision-language models. arXiv:2606.16667, 2026c.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. arXiv:2512.12844, 2025.

Yuqicheng Zhu, Jialin Yu, Lin Li, Gengyuan Zhang, Zhen Yang, Steffen Staab, Puneet Dokania, and Philip Torr. Conformalized large language models under configuration shift. arXiv:2608.01460, 2026.

A READING KEY: WHAT EACH ARM CERTIFIES

Arms differ in the quantity they certify, in their label model, and several are "-style" mappings with documented deviations from the named paper. This table is the contract every results row must be read against; the machine-readable version ships as `method_registry` and `stream_method_registry` in the released result file.

B THEORY NOTES

Theorem 1 is the exact-binomial certificate with Bonferroni control over a fixed grid: for genre g, the $G{=}20$ candidates are score quantiles chosen without reading labels, so conditioning on the scores leaves G fixed hypotheses; each is tested with a Clopper–Pearson upper bound at level δ/G, and a union bound over the grid gives the stated $1-\delta$. This is the fixed-grid variant of Learn-then-Test (Angelopoulos et al., 2022) applied per genre. Testing every distinct calibration score at level δ instead, with no multiplicity control, does not carry the bound; the implementation and its null test (`tests/test_selective_static.py`) use the grid. Corollary 1 follows because feasibility is monotone in α. The conformal threshold's false-accept statement is the usual split-conformal quantile argument with the finite-sample $1/(n^+ + 1)$ slack; the quantile used is the *low* one and the gate answers only when $\hat{p} < \tau(\alpha)$ (with $\leq$, a block of wrong-stratum scores tied at $\tau(\alpha)$ would all be accepted), both encoded as permanent regression tests (`tests/test_gate_quantile_direction.py`). The detection-delay accounting

Table 4: Arm registry (static protocol, top; drift protocol, bottom).

Arm	Certified?	Certified quantity / role
CertSel selective	yes	per-genre $P(\text{wrong} \mid \text{answered}) \leq \alpha$ w.p. $1-\delta$ (exact Clopper–Pearson); ours, group-conditional layer
CertSel conformal	yes	per-genre $P(\text{answered} \mid \text{wrong}) \leq \alpha + 1/(n^+ + 1)$, marginal; ours, the monitored gate
COIN (adopted)	yes	pooled $P(\text{wrong} \mid \text{answered}) \leq \alpha$ w.p. $1-\delta$; prior art (Wang et al., 2025b), run through our grid implementation (`gate.selective_threshold`), static baseline
TRON-style	yes	pooled false-accept on the sampling-frequency score (Wang et al., 2025a); head-to-head
Conformal-Align-style	yes	FDR among selected $\leq \alpha$ (BH) (Gui et al., 2024); static certified arm
SConU-style	yes	wrong-stratum false-accept $\leq \alpha + 1/(n^+ + 1)$ on non-outliers (Wang et al., 2025c); on the drift streams identical to the frozen conformal gate on 35/35 cells at every α (its outlier test never rejects)
learned abstention	no	not training-free contrast arm (Tayebati et al., 2025)
self-consistency thr.	no	heuristic abstention arm
semantic-entropy thr.	no	heuristic abstention arm
CertSel (drift)	yes	label-free $\lvert\Delta\mathrm{Cov}\rvert$ confirmed alarm $\rightarrow$ abstain-all $\rightarrow$ B-fresh-label recalibration
e-process (C3)	yes	betting e-process detector, same recalibration protocol; pre-break alarms are true refutations, not false alarms (RQ4)
ACI/DtACI-style	yes[†]	online miscoverage adaptation (Gibbs & Candès, 2021; 2024); average marginal coverage, no abstention
wCRC-style	yes[†]	windowed recalibrating conformal risk control on a rolling window
static frozen gates	yes	the two static thresholds held fixed across the break

[†]Certifies an asymptotic average quantity along the stream, not a per-segment selective error.

of Section 3 is deliberately not a proposition: its constants (a bounded-shift modulus, the alarm's confirmation-delay distribution) are measured rather than assumed, the isolation table (Appendix I) prints the delay-window selective error per cell, and Section 5 shows that the dominant post-break failure term is what the B-item recalibration certifies, which any delay bound takes as given. We deliberately do not restate the adopted proofs: Theorem 1 is the Clopper–Pearson selection argument of Wang et al. (2025b), the binomial-tail machinery of Angelopoulos et al. (2022), applied per genre over the fixed G-candidate grid, each candidate tested at level δ/G (the union bound above); Corollary 1 is immediate because the certified feasible set only grows with α; and the conformal statement is the standard split-conformal quantile bound with its $1/(n^+ + 1)$ slack. The mapping from each statement to the implementation it describes (`selective_threshold`, `conformal_threshold`) is pinned by the unit suite in the supplementary zip (`supp/tests/`).

C Stream construction and meta

Streams are built from metadata only. *Domain shifts* reorder Video-MME by its 6 domains: the pre-break segment draws from one fixed half of the taxonomy (large-volume domains), the post-break segment from the other half (reversed in dom-shift-rev; diluted to post_mix $= 0.5$ in dom-shift-mild). *Cross-benchmark* streams concatenate two benchmark cells at an anchored break. Table 5 gives the pinned segment lengths.

D Static matrix

Table 6 is the static matrix of RQ1 at $\alpha = 0.20$, all 9 arms on all 5×4 cells.

Table 5: Stream meta at $\alpha = 0.20$ (reference seed): pre/post segment lengths per backbone.

	Qwen2.5-VL		LLaVA-OV		InternVL3		Qwen2.5-VL-3B		Qwen2.5-VL-32B-AWQ	
Stream	n_{pre}	n_{post}	n_{pre}	n_{post}	n_{pre}	n_{post}	n_{pre}	n_{post}	n_{pre}	n_{post}
dom-shift	162	87	162	87	162	87	162	87	162	87
dom-shift-rev	87	162	87	162	87	162	87	162	87	162
dom-shift-mild	75	174	75	174	75	174	75	174	75	174
vm→nextqa	249	250	249	250	249	250	249	250	249	250
vm→mvbench	249	250	249	250	249	250	249	250	249	250
vm→vhallucer	249	250	249	250	249	250	249	250	249	250
nextqa→vm	250	249	250	249	250	249	250	249	250	249

E PER-GENRE STATIC CERTIFICATION

Tables 7 and 8 tabulate the static layer on both genre axes at $\alpha = 0.20$: the cross-benchmark axis (genre = originating benchmark) and the Video-MME domain axis (genre = domain). The domain axis is the granularity-honesty exhibit: with per-domain calibration splits of only tens of items, the selective gate certifies nothing on any of the 30 (backbone, domain) cells in any split and abstains (the certificate refuses to promise what its sample size cannot pay for), while the conformal gate still answers (its certificate is on the false-accept rate, not the selective error, whose uncertified pooled value is shown for reference).

F ACCEPT COUNTS BEHIND THE DRIFT MATRIX

Table 1 reports pooled selective error per arm; a 0.0% entry for COIN can rest on a handful of accepts. Table 9 gives, per cell and segment, the number of stream items and the pooled accept count of the three certified arms over the same 25 seeds, so every entry of Table 1 can be read with its denominator.

G DETECTION LAYER: FULL PER-CELL TABLE

Table 10 tabulates the detection layer of RQ3 per stream cell: window AUROC (two-sided vs drop-only), the ADWIN reference detector, and CertSel's confirmed-alarm rate, mean detection delay, and pre-boundary false alarms.

H DETECTION OPERATING CURVE (z SENSITIVITY)

The confirmed-alarm rule has one knob a deployment must set: the standard-deviation threshold z, pinned at 3.0 in the main text and walked over $z \in \{2.5/4.0\}$ in the same replay matrix (Table 11). Loosening to $z = 2.5$ buys sensitivity: 15 of 35 cells caught at $\geq 20/25$ seeds (versus 11 at the pinned point), and more alarming seeds on the gentle in-benchmark domain shifts (up to 9 of 25 per cell, versus at most 3 at the pinned point, still below the confirmation bar) at the price of 55 total pre-boundary false alarms against 12. Its delays are *not* uniformly shorter: each per-cell mean conditions on that arm's own alarming seeds, and the extra marginal seeds $z = 2.5$ admits tend to alarm late, so while it fires earlier where the signal is strongest (e.g. Qwen vm→nextqa 32.7 versus 43.1 steps, nextqa→vm 25.6 versus 30.3), its conditional mean is *higher* in 6 of the 25 cells where both operating points fire; over each point's own caught cells the ranges are 25.6–125.7 steps ($z = 2.5$; the maximum is the hard vm→mvbench cell only it catches) versus 30.3–95.7 pinned. Tightening to $z = 4.0$ drives the false-alarm total to 0, but every gentle domain shift goes fully silent and the 8 cells it still catches are caught later (58.4–96.5 steps over those cells, versus 30.3–95.7 at the pinned point). The pooled post-break selective error itself barely tracks the knob (at most 5.5 points of spread across the three operating points in any cell; summarized in RQ3).

I RECALIBRATION ISOLATION: THE GPU-TIER TABLE

Table 12 is the per-cell bridge for the recalibration-isolation experiment (RQ5): five arms replay the *same* conformal gate on the *same* streams at $\alpha = 0.20$, differing only in recalibration policy.

Table 6: Static matrix at $\alpha = 0.20$: answered coverage (%, left of each pair) and pooled selective error (%, right) over 25 calibration splits, each ± 1 sample s.d. across seeds (sel-err is pooled over splits; its $\pm$ is the s.d. of the per-seed realizations over the *answering* splits only, as few as 1 and 1 on Qwen2.5-VL-32B-AWQ VideoHallucer for CertSel selective and COIN; LLaVA-OV Video-MME abstains in all 25). **Bold**: certified arm over α; ABSTAIN: every seed certified nothing (vacuous). Certified quantities differ per arm; see Table 4. Coverage carries ± 1 s.d. over seeds; sel-err is the pooled point (its seed s.d. is in the released results). The two heuristic rows are functions of the same answer clusters and coincide on 50 of 60 entries over the α grid.

		Video-MME		MVBench		NExT-QA		VideoHallucer	
Backbone	Arm	cov	sel-err	cov	sel-err	cov	sel-err	cov	sel-err
Qwen2.5-VL	CertSel selective	ABSTAIN	–	2.6±7.4	10.9	63.3±6.2	10.3	ABSTAIN	–
	CertSel conformal	37.6±5.8	**25.9**	41.9±4.4	17.9	60.4±3.6	8.8	35.7±5.2	19.3
	COIN (adopted)	ABSTAIN	–	2.6±7.4	10.9	63.3±6.2	10.3	ABSTAIN	–
	TRON-style	31.5±6.9	**22.2**	28.9±22.2	18.8	ABSTAIN	–	ABSTAIN	–
	Conformal-Align-style	6.1±7.4	12.2	19.7±12.2	9.0	40.9±18.2	6.6	3.6±8.9	18.9
	SConU-style	34.1±5.4	**25.3**	41.9±4.4	17.9	60.4±3.6	8.8	35.7±5.2	19.3
	learned abstention	22.2±5.5	20.0	52.1±4.9	19.7	89.8±4.5	20.0	55.2±13.1	22.8
	self-consistency thr.	2.6±9.0	25.5	37.9±17.2	19.0	84.2±4.6	18.4	9.1±21.3	27.4
	semantic-entropy thr.	2.6±9.0	25.5	37.9±17.2	19.0	85.3±5.3	18.4	9.1±21.3	27.4
LLaVA-OV	CertSel selective	ABSTAIN	–	ABSTAIN	–	52.3±21.1	11.2	1.3±4.6	19.5
	CertSel conformal	39.7±4.6	**26.5**	33.0±4.0	**25.0**	53.6±5.7	9.0	35.6±4.3	18.1
	COIN (adopted)	ABSTAIN	–	ABSTAIN	–	52.3±21.1	11.2	1.3±4.6	19.5
	TRON-style	3.5±12.0	**32.4**	ABSTAIN	–	ABSTAIN	–	19.8±16.6	19.4
	Conformal-Align-style	1.4±5.5	**23.6**	8.6±11.3	13.7	11.7±15.8	8.9	6.4±10.5	16.5
	SConU-style	37.4±4.5	**25.6**	33.0±4.0	**25.0**	53.6±5.7	9.0	35.6±4.3	18.1
	learned abstention	25.8±8.2	22.7	36.4±6.1	20.6	90.2±4.1	20.4	43.3±10.0	20.9
	self-consistency thr.	1.8±8.8	33.9	ABSTAIN	–	86.6±5.9	18.9	20.5±17.8	21.0
	semantic-entropy thr.	1.8±8.8	33.9	ABSTAIN	–	86.3±6.1	19.6	19.4±18.2	21.9
InternVL3	CertSel selective	ABSTAIN	–	40.7±10.3	8.9	76.7±6.5	11.2	ABSTAIN	–
	CertSel conformal	44.2±3.0	**21.1**	46.5±4.7	11.1	65.6±4.2	7.3	30.3±5.1	17.7
	COIN (adopted)	18.6±8.1	4.9	40.7±10.3	8.9	76.7±6.5	11.2	ABSTAIN	–
	TRON-style	ABSTAIN	–	ABSTAIN	–	ABSTAIN	–	ABSTAIN	–
	Conformal-Align-style	19.8±11.3	8.5	22.9±17.6	5.3	34.7±30.1	5.6	ABSTAIN	–
	SConU-style	41.3±4.0	**20.2**	46.5±4.7	11.1	65.6±4.2	7.3	30.3±5.1	17.7
	learned abstention	44.2±4.3	21.5	88.2±6.3	21.1	98.5±1.8	19.0	38.2±25.1	23.5
	self-consistency thr.	ABSTAIN	–	83.3±8.2	19.6	98.8±1.9	19.1	8.0±22.1	24.0
	semantic-entropy thr.	ABSTAIN	–	83.2±8.2	19.6	98.8±1.4	19.3	8.0±22.1	24.0
Qwen2.5-VL-3B	CertSel selective	ABSTAIN	–	2.6±7.2	11.7	37.3±8.4	7.8	ABSTAIN	–
	CertSel conformal	38.5±4.1	**29.2**	46.9±3.9	18.1	49.6±5.2	11.3	37.0±3.7	15.1
	COIN (adopted)	ABSTAIN	–	2.6±7.2	11.7	37.3±8.4	7.8	ABSTAIN	–
	TRON-style	28.8±4.0	**22.4**	38.8±4.5	15.8	24.1±25.7	13.7	22.7±2.1	10.6
	Conformal-Align-style	7.3±9.4	12.0	8.8±11.6	12.0	24.0±11.7	4.5	10.7±10.6	8.9
	SConU-style	36.1±3.2	**26.8**	46.9±3.9	18.1	49.6±5.2	11.3	37.0±3.7	15.1
	learned abstention	28.9±4.6	20.3	50.4±7.5	19.4	77.1±4.8	20.2	60.0±9.9	21.4
	self-consistency thr.	1.2±5.9	28.4	41.1±5.4	17.1	63.2±9.3	18.6	41.9±10.1	17.8
	semantic-entropy thr.	1.2±5.9	28.4	41.7±6.8	17.5	63.3±10.3	18.9	41.5±8.2	17.4
Qwen2.5-VL-32B-AWQ	CertSel selective	ABSTAIN	–	1.1±3.8	15.9	49.4±12.5	10.0	0.6±2.8	14.3
	CertSel conformal	43.7±4.0	**25.0**	36.9±4.3	**20.7**	55.3±3.2	10.7	39.4±5.5	15.6
	COIN (adopted)	ABSTAIN	–	1.1±3.8	15.9	49.4±12.5	10.0	0.6±2.8	14.3
	TRON-style	1.8±9.2	**30.4**	ABSTAIN	–	5.1±17.8	15.0	ABSTAIN	–
	Conformal-Align-style	8.7±10.1	11.8	8.2±12.4	16.5	21.9±21.4	5.8	8.9±8.9	9.6
	SConU-style	41.0±3.0	**25.1**	36.9±4.3	**20.7**	55.3±3.2	10.7	39.4±5.5	15.6
	learned abstention	37.1±3.8	22.4	43.3±7.8	20.5	80.9±6.0	20.5	57.7±9.7	21.1
	self-consistency thr.	ABSTAIN	–	7.3±17.2	25.9	75.2±6.8	18.5	29.4±28.8	22.5
	semantic-entropy thr.	ABSTAIN	–	7.3±17.2	25.9	76.3±6.8	18.9	29.4±28.8	22.5

no-recal freezes the gate across the break (control); *in-stream* is the CPU-tier CertSel protocol (the B-label window is consumed from the stream while the gate abstains); *fresh@alarm* recalibrates instantly at the confirmed alarm on $B = 50$ fresh post-composition labels drawn from a parallel annotation channel, composition-matched to the realized post segment; *oracle* recalibrates on the same fresh draw at the true break with zero delay (not a deployable policy; the boundary is oracle

Table 7: Per-genre static certification, cross-benchmark axis, $\alpha = 0.20$: coverage (%) and pooled selective error (%).

		CertSel selective		CertSel conformal	
Backbone	Genre	cov	sel-err	cov	sel-err[‡]
Qwen2.5-VL	Video-MME	ABSTAIN	–	34.1	25.3
	MVBench	2.6	10.9	41.9	17.9
	NExT-QA	63.3	10.3	60.4	8.8
	VideoHallucer	ABSTAIN	–	35.7	19.3
LLaVA-OV	Video-MME	ABSTAIN	–	37.4	25.6
	MVBench	ABSTAIN	–	33.0	25.0
	NExT-QA	52.3	11.2	53.6	9.0
	VideoHallucer	1.3	19.5	35.6	18.1
InternVL3	Video-MME	18.6	4.9	41.3	20.2
	MVBench	40.7	8.9	46.5	11.1
	NExT-QA	76.7	11.2	65.6	7.3
	VideoHallucer	ABSTAIN	–	30.3	17.7
Qwen2.5-VL-3B	Video-MME	ABSTAIN	–	36.1	26.8
	MVBench	2.6	11.7	46.9	18.1
	NExT-QA	37.3	7.8	49.6	11.3
	VideoHallucer	ABSTAIN	–	37.0	15.1
Qwen2.5-VL-32B-AWQ	Video-MME	ABSTAIN	–	41.0	25.1
	MVBench	1.1	15.9	36.9	20.7
	NExT-QA	49.4	10.0	55.3	10.7
	VideoHallucer	0.6	14.3	39.4	15.6

[‡]Uncertified for the conformal gate (it certifies the false-accept rate); shown for reference.

information); *e-proc* is the C3 anytime-valid detector with the same instant fresh recalibration. The oracle-vs-detected contrast isolates detection delay; the fresh@alarm-vs-no-recal contrast isolates the value of the recalibration itself.

Silent cells, the 17 cells where the monitor never fires at this tier's seeds (all 14 gentle domain-shift cells and 3 abrupt ones: InternVL3 vm→mvbench and vm→vhallucer, Qwen2.5-VL-3B vm→vhallucer, Qwen2.5-VL-32B-AWQ vm→mvbench) (at the 25-seed CPU tier they reach at most 3 alarming seeds, Figure 4), are listed with alarm $0/5$ and no post-recal column. Following the contract note shipped inside the result file, they are a *detector-power* gap, reported in the CPU tier (RQ2); they are outside what recalibration can fix and are counted neither for nor against the recalibration verdict here. The oracle column is still populated on them and carries its own information: even boundary-time recalibration fails α in 12 of the 17 silent cells.

Two small-sample caveats belong next to the table rather than in a footnote. First, the post-recal pools on the mvbench-target cells are thin (49 and 20 accepted items for Qwen and LLaVA-OV respectively, against 398 on vm→nextqa), so those two cells' verdicts (one over, one under α) carry the widest bounds in the table (43.2% and 34.4%); the LLaVA-OV vm→mvbench cell holds α post-recal while its *oracle* arm does not, which is the pattern small-n noise produces. The tightest failure in the table is the other direction of the same problem: LLaVA-OV nextqa→vm misses by 1 wrong accept in 291, and its bound (24.6%) does not separate it from α either. Second, the per-seed strict criterion (no single seed's realization over α, adjudicated as appendix-grade) is harsher: 7 of 18 alarmed cells pass it, versus 10 under seed pooling.

J RECALIBRATION-BUDGET SWEEP

The sweep of Section 5 is a third single-writer contract file (`results_recal_budget.json`, 67500 replays over the CPU matrix's 25 seeds, $B \in \{10, 20, 30, 50, 75, 100, 150\}$, two rules $\times$ two timings $\times$ the α grid). Its predictions were written to `PREREG.md` before the first run: P1 (the certified false-accept rate is met where selective error fails) held; P2 (the sign of coverage minus base error rate predicts the verdict in $\geq 80\%$ of entries) held; P3 (budget does not restore the conformal rule; the oracle still fails at the largest budget, and the hold count is non-decreasing in B in aggregate: it rises from 3 to 6 of 10 at the alarm, with an interior dip at the oracle timing

Table 8: Per-genre static certification, Video-MME domain axis, $\alpha = 0.20$: coverage (%) and pooled selective error (%).

		CertSel selective		CertSel conformal	
Backbone	Domain	cov	sel-err	cov	sel-err[‡]
Qwen2.5-VL	Artistic Performance	ABSTAIN	–	37.7	18.8
	Film & Television	ABSTAIN	–	30.6	20.0
	Knowledge	ABSTAIN	–	37.9	25.7
	Life Record	ABSTAIN	–	37.0	28.4
	Multilingual	ABSTAIN	–	64.4	22.1
	Sports Competition	ABSTAIN	–	37.4	32.8
LLaVA-OV	Artistic Performance	ABSTAIN	–	49.7	20.3
	Film & Television	ABSTAIN	–	42.6	20.5
	Knowledge	ABSTAIN	–	36.2	29.5
	Life Record	ABSTAIN	–	41.6	27.9
	Multilingual	ABSTAIN	–	52.0	14.5
	Sports Competition	ABSTAIN	–	32.4	34.4
InternVL3	Artistic Performance	ABSTAIN	–	47.7	22.4
	Film & Television	ABSTAIN	–	54.6	17.6
	Knowledge	ABSTAIN	–	41.4	21.4
	Life Record	ABSTAIN	–	39.5	24.1
	Multilingual	ABSTAIN	–	59.1	20.3
	Sports Competition	ABSTAIN	–	41.9	19.7
Qwen2.5-VL-3B	Artistic Performance	ABSTAIN	–	43.6	21.9
	Film & Television	ABSTAIN	–	39.1	24.0
	Knowledge	ABSTAIN	–	35.0	31.9
	Life Record	ABSTAIN	–	36.4	32.1
	Multilingual	ABSTAIN	–	30.2	52.9
	Sports Competition	ABSTAIN	–	44.5	27.9
Qwen2.5-VL-32B-AWQ	Artistic Performance	ABSTAIN	–	42.6	22.1
	Film & Television	ABSTAIN	–	47.6	18.6
	Knowledge	ABSTAIN	–	44.7	27.2
	Life Record	ABSTAIN	–	38.5	28.9
	Multilingual	ABSTAIN	–	63.1	14.1
	Sports Competition	ABSTAIN	–	42.3	27.2

[‡]Uncertified for the conformal gate; shown for reference.

(4 at $B = 75$ against 3 at $B = 150$), which the pre-registration rules on as a non-monotone curve, so the "budget is the bottleneck" reading is not available) held; P4 (the Clopper–Pearson certificate is vacuous in most realizations at $B \leq 50$: held; and holds at least as many cells as the conformal rule at $B \leq 50$, failed: it holds fewer at the alarm timing and no more at the oracle timing at every budget in the grid) failed; P5 (eight cells hit the fresh-pool cap before $B = 150$) failed: 4 did, at $B = 100$. Table 2 (main text) gives the counts behind Figure 6. At $B = 50$ on the first 5 seeds the sweep's conformal arms reproduce Table 12 field for field (1974 fields compared, a refusal-to-write check in the writer). The Clopper–Pearson arm is the same online policy as the conformal arm with the threshold rule swapped: it monitors, confirms the alarm, buys the same B fresh labels, and abstains on everything while no threshold certifies. Its one limitation is registered here: after a vacuous certificate the arm keeps abstaining for the rest of the stream, since a coverage of zero gives the monitor nothing to alarm on; a re-arming rule was not run.

K Deferred baselines

Five design-list baselines are computable only with inputs the cached pool does not contain, and are deferred to the GPU tier with explicit blockers (recorded as `deferred_baselines` in the result file): *BCEA* (Xu et al., 2026c) needs budgeted staged evidence-acquisition re-inference passes; *DS-CP* needs prompt/question embeddings to reweight calibration samples; *VL-Uncertainty* needs semantic-perturbation re-inference passes; *distribution-informed online CP* (Hu et al., 2026) and *boosted/conditional conformal* (Cherian et al., 2024) need update rules read off their papers (delib-

Table 9: Accept counts, drift layer, $\alpha = 0.20$, 25 seeds pooled: stream items (n) and accepts of the frozen conformal gate, frozen COIN and CertSel (frozen / COIN / CertSel) per segment.

		pre-break				post-break			
Backbone	Stream	n	frozen	COIN	CertSel	n	frozen	COIN	CertSel
Qwen2.5-VL	dom-shift	4050	1392	0	1392	2175	481	0	481
	dom-shift-rev	2175	833	0	833	4050	1939	0	1836
	dom-shift-mild	1875	659	0	659	4350	1214	0	1211
	vm→nextqa	6225	2122	0	2048	6250	4196	0	3107
	vm→mvbench	6225	2122	0	2048	6250	2864	0	2692
	vm→vhallucer	6225	2122	0	2048	6250	4043	0	2117
	nextqa→vm	6250	3778	3959	3778	6225	1594	1830	1757
LLaVA-OV	dom-shift	4050	1595	0	1568	2175	863	0	842
	dom-shift-rev	2175	804	0	804	4050	1419	0	1419
	dom-shift-mild	1875	721	0	721	4350	1737	0	1684
	vm→nextqa	6225	2329	0	2318	6250	4022	0	2834
	vm→mvbench	6225	2329	0	2318	6250	2785	0	2726
	vm→vhallucer	6225	2329	0	2318	6250	3312	0	2654
	nextqa→vm	6250	3349	3267	3349	6225	1399	1765	1697
InternVL3	dom-shift	4050	1662	181	1662	2175	801	84	801
	dom-shift-rev	2175	961	0	961	4050	1993	0	1976
	dom-shift-mild	1875	763	83	763	4350	1700	182	1688
	vm→nextqa	6225	2573	1156	2573	6250	4448	2792	3264
	vm→mvbench	6225	2573	1156	2573	6250	2882	1717	2822
	vm→vhallucer	6225	2573	1156	2573	6250	1803	204	1716
	nextqa→vm	6250	4098	4792	4079	6225	2152	3163	2063
Qwen2.5-VL-3B	dom-shift	4050	1446	0	1446	2175	565	0	565
	dom-shift-rev	2175	838	0	838	4050	1855	0	1855
	dom-shift-mild	1875	675	0	675	4350	1336	0	1336
	vm→nextqa	6225	2247	0	2247	6250	3924	0	2723
	vm→mvbench	6225	2247	0	2247	6250	2819	0	2722
	vm→vhallucer	6225	2247	0	2247	6250	2259	0	2259
	nextqa→vm	6250	3099	2331	3099	6225	1538	1100	1656
Qwen2.5-VL-32B-AWQ	dom-shift	4050	1662	0	1662	2175	687	0	687
	dom-shift-rev	2175	975	0	975	4050	1985	0	1965
	dom-shift-mild	1875	789	0	789	4350	1560	0	1560
	vm→nextqa	6225	2551	0	2551	6250	3584	0	2872
	vm→mvbench	6225	2551	0	2551	6250	2632	0	2614
	vm→vhallucer	6225	2551	0	2551	6250	3426	0	2693
	nextqa→vm	6250	3454	3088	3442	6225	2275	1892	2053

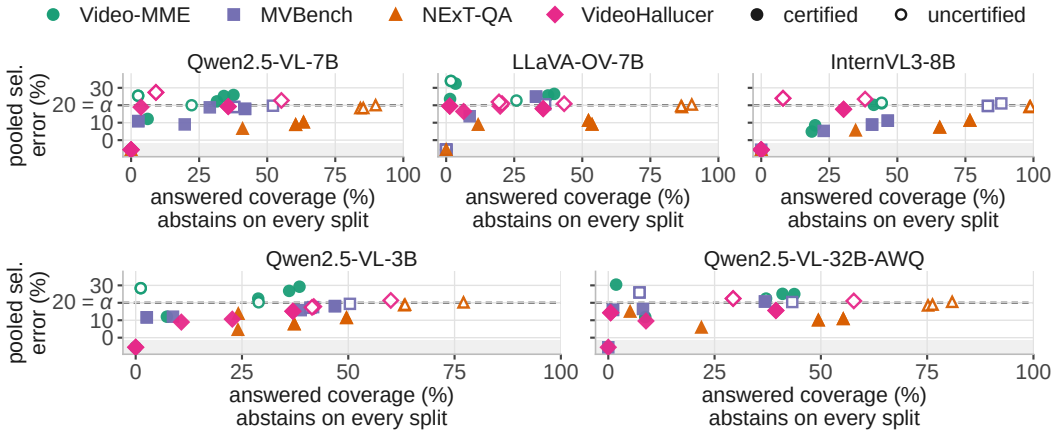


Figure 7: The static matrix (Figure 3) over all five backbones: the three reported backbones on the first row and the Qwen2.5-VL scale twins (3B, 32B-AWQ) on the second. Same axes, arms and marker conventions.

erately not guessed, because running an invented update under a real paper's name would violate the same honesty discipline this table serves).

Table 10: Detection at $\alpha = 0.20$: $|\Delta\mathrm{Cov}|$ window AUROC (two-sided vs drop-only) on the frozen gate's answer stream; ADWIN reference alarms; CertSel's confirmed-alarm rate (of 25 seeds), mean detection delay (stream steps; each delay averages only that cell's alarming seeds; the alarm count beside it is its n), and total pre-boundary false alarms. Alarm/delay follow the in-stream protocol, whose recalibration consumes stream entries; on cells with a pre-boundary false alarm it can therefore diverge post-break from the isolation tier's parallel-channel fresh arm (Table 12, which pins the original 5-seed list, a prefix of this table's 25).

		$\lvert\Delta\mathrm{Cov}\rvert$ AUROC		ADWIN (ref.)		CertSel confirmed alarm		
Backbone	Stream	two-sided	drop-only	alarm	delay	alarm	delay	FA
Qwen2.5-VL	dom-shift	0.792	0.878	0/25	–	0/25	–	0
	dom-shift-rev	0.766	0.193	0/25	–	3/25	85.0	0
	dom-shift-mild	0.706	0.738	0/25	–	1/25	73.0	0
	vm→nextqa	0.995	0.003	25/25	81.5	25/25	43.1	2
	vm→mvbench	0.765	0.141	1/25	198.0	9/25	132.2	2
	vm→vhallucer	0.989	0.006	25/25	89.2	25/25	43.7	2
	nextqa→vm	0.995	0.997	25/25	58.8	25/25	30.3	0
LLaVA-OV	dom-shift	0.547	0.445	0/25	–	1/25	15.0	1
	dom-shift-rev	0.655	0.571	0/25	–	0/25	–	0
	dom-shift-mild	0.704	0.422	0/25	–	2/25	20.5	0
	vm→nextqa	0.987	0.008	24/25	99.3	25/25	45.9	1
	vm→mvbench	0.662	0.269	0/25	–	5/25	184.6	1
	vm→vhallucer	0.847	0.091	8/25	154.0	19/25	121.7	1
	nextqa→vm	0.995	0.997	25/25	80.5	25/25	32.6	0
InternVL3	dom-shift	0.541	0.655	0/25	–	0/25	–	0
	dom-shift-rev	0.715	0.318	0/25	–	1/25	128.0	0
	dom-shift-mild	0.639	0.572	0/25	–	1/25	139.0	0
	vm→nextqa	0.991	0.006	25/25	93.0	25/25	36.9	0
	vm→mvbench	0.583	0.341	0/25	–	3/25	116.3	0
	vm→vhallucer	0.831	0.845	4/25	174.0	11/25	145.0	0
	nextqa→vm	0.994	0.996	25/25	75.4	25/25	43.5	1
Qwen2.5-VL-3B	dom-shift	0.690	0.794	0/25	–	0/25	–	0
	dom-shift-rev	0.689	0.259	0/25	–	0/25	–	0
	dom-shift-mild	0.688	0.700	0/25	–	0/25	–	0
	vm→nextqa	0.976	0.013	25/25	108.4	25/25	48.2	0
	vm→mvbench	0.700	0.196	0/25	–	9/25	123.0	0
	vm→vhallucer	0.544	0.499	0/25	–	0/25	–	0
	nextqa→vm	0.979	0.988	25/25	112.5	25/25	67.6	0
Qwen2.5-VL-32B-AWQ	dom-shift	0.738	0.825	0/25	–	0/25	–	0
	dom-shift-rev	0.705	0.313	0/25	–	1/25	107.0	0
	dom-shift-mild	0.765	0.741	0/25	–	0/25	–	0
	vm→nextqa	0.887	0.072	13/25	153.7	24/25	75.1	0
	vm→mvbench	0.546	0.451	0/25	–	1/25	83.0	0
	vm→vhallucer	0.826	0.106	3/25	123.3	20/25	95.7	0
	nextqa→vm	0.901	0.944	15/25	175.7	19/25	95.7	1

L BROADER IMPACT

A gate that can abstain makes errors visible, countable, and contractible. That cuts both ways: abstention shifts work back to the user, and a miscalibrated promise of safety is worse than none, which is why this paper reports its own silent-violation regime in the main text.

M REPRODUCIBILITY

The CPU-tier result contract file (`experiments/certsel/results/results.json`, stage cpu-matrix, generated 2026-09-22T09:23:38) carries a provenance block: input pool sha `5fd8f57c`, genre sidecar sha `64fb236f`, full config and version pins. The single writer is `run_matrix.py`; `gen_macros.py` is the only path by which a number enters this PDF; `run_ledger.jsonl` is an append-only history of full runs (GPU hours 0.0, CPU wall

Table 11: z operating curve at $\alpha = 0.20$: confirmed-alarm rate (of 25 seeds), mean detection delay (steps), and total pre-boundary false alarms, per stream cell. Middle group repeats the pinned $z = 3.0$ of Table 10 for side-by-side reading.

		$z = 2.5$			$z = 3.0$ (pinned)			$z = 4.0$		
Backbone	Stream	alarm	delay	FA	alarm	delay	FA	alarm	delay	FA
Qwen2.5-VL	dom-shift	7/25	48.1	1	0/25	–	0	0/25	–	0
	dom-shift-rev	9/25	91.0	0	3/25	85.0	0	0/25	–	0
	dom-shift-mild	3/25	74.3	1	1/25	73.0	0	0/25	–	0
	vm→nextqa	25/25	32.7	2	25/25	43.1	2	25/25	58.4	0
	vm→mvbench	20/25	114.8	2	9/25	132.2	2	1/25	38.0	0
	vm→vhallucer	25/25	31.3	2	25/25	43.7	2	25/25	73.7	0
	nextqa→vm	25/25	25.6	3	25/25	30.3	0	25/25	59.6	0
LLaVA-OV	dom-shift	0/25	–	2	1/25	15.0	1	0/25	–	0
	dom-shift-rev	0/25	–	0	0/25	–	0	0/25	–	0
	dom-shift-mild	3/25	20.7	0	2/25	20.5	0	0/25	–	0
	vm→nextqa	25/25	34.0	3	25/25	45.9	1	23/25	76.6	0
	vm→mvbench	11/25	116.7	3	5/25	184.6	1	0/25	–	0
	vm→vhallucer	21/25	94.9	3	19/25	121.7	1	5/25	90.4	0
	nextqa→vm	24/25	35.0	3	25/25	32.6	0	24/25	71.2	0
InternVL3	dom-shift	0/25	–	2	0/25	–	0	0/25	–	0
	dom-shift-rev	5/25	69.0	0	1/25	128.0	0	0/25	–	0
	dom-shift-mild	3/25	103.7	0	1/25	139.0	0	0/25	–	0
	vm→nextqa	25/25	30.6	1	25/25	36.9	0	25/25	62.3	0
	vm→mvbench	6/25	81.2	1	3/25	116.3	0	0/25	–	0
	vm→vhallucer	21/25	125.7	1	11/25	145.0	0	1/25	131.0	0
	nextqa→vm	25/25	28.9	2	25/25	43.5	1	25/25	65.2	0
Qwen2.5-VL-3B	dom-shift	0/25	–	1	0/25	–	0	0/25	–	0
	dom-shift-rev	4/25	94.8	0	0/25	–	0	0/25	–	0
	dom-shift-mild	2/25	75.5	0	0/25	–	0	0/25	–	0
	vm→nextqa	25/25	46.2	4	25/25	48.2	0	24/25	96.5	0
	vm→mvbench	13/25	130.8	4	9/25	123.0	0	0/25	–	0
	vm→vhallucer	3/25	162.7	4	0/25	–	0	0/25	–	0
	nextqa→vm	23/25	41.5	4	25/25	67.6	0	14/25	117.0	0
Qwen2.5-VL-32B-AWQ	dom-shift	1/25	23.0	1	0/25	–	0	0/25	–	0
	dom-shift-rev	7/25	88.4	0	1/25	107.0	0	0/25	–	0
	dom-shift-mild	2/25	123.0	0	0/25	–	0	0/25	–	0
	vm→nextqa	25/25	63.8	0	24/25	75.1	0	6/25	108.0	0
	vm→mvbench	2/25	143.5	0	1/25	83.0	0	0/25	–	0
	vm→vhallucer	24/25	84.2	0	20/25	95.7	0	2/25	142.0	0
	nextqa→vm	21/25	71.0	5	19/25	95.7	1	5/25	143.8	0

clock 150.7 s for the pinned matrix). The GPU-tier isolation experiment has its own contract file (`results_recal_isolation.json`, generated 2026-09-22T09:23:54, single writer `recal_isolation.py`, 2625 replays), fed by a fresh-label sweep on held-out questions: 31050 release rows over fresh pools of 240 Video-MME and 150 per-benchmark questions per backbone (release sha `10f5c506`), at 9.4 GPU-hours total; the same `gen_macros.py` bridges both files into this PDF, together with the budget sweep (`results_recal_budget.json`, single writer `recal_budget_sweep.py`) and the read-only mechanism analysis (`analysis_recal_mechanism.json`, writer `analysis_recal_mechanism.py`) and the question-clustered bounds (`results_cluster_bounds.json`, writer `cluster_bounds.py`, derived from the isolation file), each carrying the sha256 of its inputs; the bridge refuses to build if any of the five files disagree on their inputs. The unit suite (per-seed determinism, the quantile-direction and AUROC tie-credit postmortem regressions, protocol guardrails) runs in under five seconds and is green at submission of this draft. Protocol knobs: window $w = 40$, $z = 3.0$ (operating curve $z \in \{2.5/4.0\}$ in the result file), persistence 5, recalibration budget $B = 50$, e-process significance $\delta_e = 0.05$, calibration fraction 0.5, $\delta = 0.02$.

Table 12: Recalibration isolation at $\alpha = 0.20$: post-break pooled selective error (%) per arm over this tier's 5 seeds (the original seed list, a prefix of the CPU matrix's 25); **bold** marks a value over α. "alarm"/"delay" are the fresh arm's $|\Delta\mathrm{Cov}|$ confirmed-alarm seed count and mean delay (steps); on cells with a pre-boundary false alarm they can diverge from Table 10, whose in-stream recalibration consumes stream entries while the parallel fresh channel does not (at the shared seeds, the two vm→nextqa cells). "e-al" is the e-process arm's own alarm count. "fresh@alarm" pools the full post segment (delay window included); "post-recal" pools only the steps after its recalibration, the isolated quantity; – marks silent cells. "acc / nq" is the pooled accept count the post-recal verdict is decided on and the number of distinct questions among those accepts (the 5 seeds re-ask the same questions; at most 5 repeats of one question); "UB" / "UB_o" are 95% question-clustered percentile-bootstrap upper bounds (%) on that verdict and on the oracle's (20000 resamples of questions, seed 20260921, rounded up). A hold is certified only where its bound is at or below α: 4 of the 10 post-recal holds and 7 of the 14 oracle holds (the exact pooled Clopper–Pearson bound, which treats the repeats as independent, certified 6 and 9). On 12 cell-arms with at most two repeats per question the clustered bound is narrower than the pooled one; taking the wider of the two there changes no certification.

					post-break selective error (%)					post-recal verdict		
Backbone	Stream	alarm	delay	e-al	no-recal	in-stream	fresh@alarm	oracle	e-proc	post-recal	acc / nq	UB / UB_o
Qwen2.5-VL	dom-shift	0/5	–	0/5	20.0	20.0	20.0	**28.0**	20.0	–	–/–	–/38.0
	dom-shift-rev	0/5	–	2/5	**28.8**	**28.8**	**28.8**	**26.9**	**28.5**	–	–/–	–/33.9
	dom-shift-mild	0/5	–	0/5	**22.1**	**22.1**	**22.1**	**24.4**	**22.1**	–	–/–	–/30.8
	vm→nextqa	4/5	63.3	0/5	12.5	11.4	8.1	7.3	10.8	8.0	398/230	11.6/10.3
	vm→mvbench	1/5	148.0	0/5	18.9	18.0	19.4	**20.8**	19.4	**30.6**	49/49	40.9/25.3
	vm→vhallucer	4/5	35.8	3/5	**27.6**	**21.2**	**20.9**	19.3	**25.4**	17.0	265/157	22.4/24.7
	nextqa→vm	5/5	24.8	0/5	**21.8**	**23.4**	**27.0**	**26.7**	**21.8**	**26.6**	436/198	32.5/32.5
LLaVA-OV	dom-shift	0/5	–	0/5	**23.6**	**23.6**	**23.6**	**21.9**	**23.6**	–	–/–	–/31.1
	dom-shift-rev	0/5	–	0/5	**21.4**	**21.4**	**21.4**	19.7	**21.4**	–	–/–	–/27.1
	dom-shift-mild	0/5	–	1/5	**22.5**	**22.5**	**22.5**	**21.2**	**22.5**	–	–/–	–/27.0
	vm→nextqa	5/5	36.8	0/5	11.9	9.6	12.5	12.9	11.9	13.0	684/326	16.4/16.2
	vm→mvbench	1/5	186.0	2/5	**29.0**	**31.2**	**27.4**	**20.4**	**27.9**	15.0	20/20	30.0/26.2
	vm→vhallucer	3/5	182.3	0/5	**22.4**	**22.6**	**20.8**	19.1	**22.4**	16.3	92/81	23.4/24.3
	nextqa→vm	5/5	29.4	0/5	19.4	**21.8**	20.0	**20.9**	19.4	**20.3**	291/159	26.1/26.6
InternVL3	dom-shift	0/5	–	0/5	16.3	16.3	16.3	**24.7**	16.3	–	–/–	–/32.8
	dom-shift-rev	1/5	128.0	0/5	**25.0**	**24.5**	**25.2**	**24.0**	**25.0**	**38.9**	18/18	55.6/29.9
	dom-shift-mild	0/5	–	0/5	19.3	19.3	19.3	**23.2**	18.9	–	–/–	–/28.4
	vm→nextqa	5/5	38.8	0/5	9.7	6.8	5.6	5.2	8.5	5.4	538/278	7.9/7.6
	vm→mvbench	0/5	–	0/5	11.9	11.9	11.9	11.6	11.9	–	–/–	–/15.2
	vm→vhallucer	0/5	–	0/5	18.5	18.5	18.5	16.5	17.7	–	–/–	–/21.4
	nextqa→vm	5/5	28.8	0/5	14.6	19.6	**23.9**	**24.5**	14.6	**24.6**	586/266	29.4/29.1
Qwen2.5-VL-3B	dom-shift	0/5	–	0/5	16.7	16.7	16.7	16.0	16.7	–	–/–	–/25.6
	dom-shift-rev	0/5	–	0/5	**30.3**	**30.3**	**30.3**	**27.6**	**30.3**	–	–/–	–/34.7
	dom-shift-mild	0/5	–	0/5	**20.6**	**20.6**	**20.6**	**20.5**	**20.6**	–	–/–	–/26.5
	vm→nextqa	5/5	86.2	0/5	16.5	15.6	14.4	13.5	16.5	15.9	429/241	20.1/17.3
	vm→mvbench	1/5	76.0	0/5	16.2	15.8	15.8	15.3	16.2	18.3	60/60	26.7/19.9
	vm→vhallucer	0/5	–	0/5	12.7	12.7	12.7	16.2	12.7	–	–/–	–/20.3
	nextqa→vm	5/5	63.6	0/5	**22.1**	**23.6**	**24.6**	**27.3**	**22.1**	**25.1**	334/177	31.0/32.9
Qwen2.5-VL-32B-AWQ	dom-shift	0/5	–	0/5	18.6	18.6	18.6	**23.8**	18.4	–	–/–	–/31.9
	dom-shift-rev	0/5	–	0/5	**29.9**	**29.9**	**29.9**	**29.9**	**29.9**	–	–/–	–/36.4
	dom-shift-mild	0/5	–	0/5	**24.3**	**24.3**	**24.3**	**24.1**	**22.3**	–	–/–	–/29.9
	vm→nextqa	5/5	74.6	0/5	13.1	10.5	10.9	9.6	10.5	9.7	435/234	13.0/12.9
	vm→mvbench	1/5	83.0	0/5	**25.5**	**25.4**	**25.2**	**25.0**	**25.3**	**29.3**	75/75	38.7/30.4
	vm→vhallucer	5/5	90.6	0/5	**21.3**	18.4	18.4	17.1	18.2	15.6	275/162	21.1/21.7
	nextqa→vm	3/5	47.0	0/5	**21.6**	**21.8**	**24.0**	**28.8**	**21.6**	**23.9**	268/179	29.6/33.8

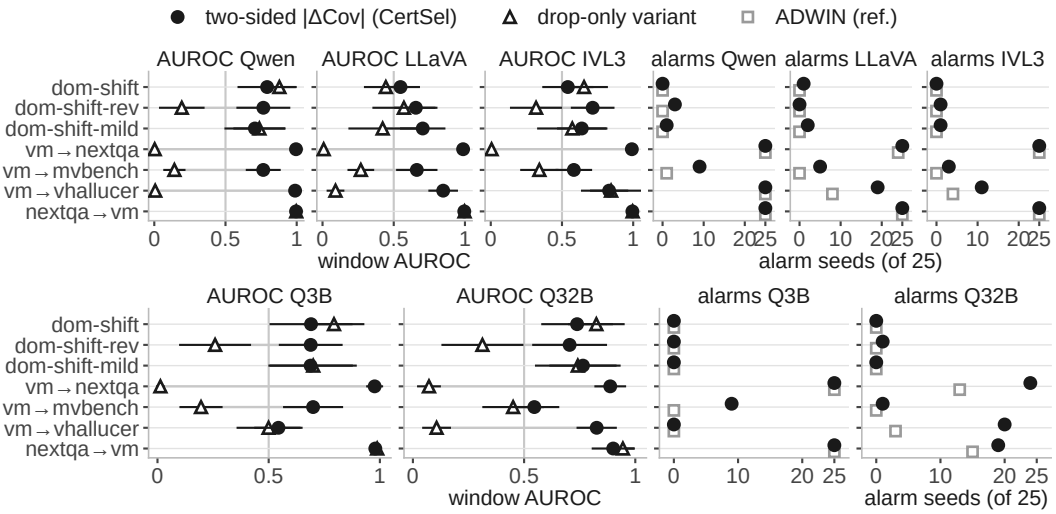


Figure 8: Detection (Figure 4) over all five backbones, same layout as Figure 7: window AUROC of the two-sided and drop-only statistics and ADWIN reference, and confirmed-alarm seeds of 25, per cell.