FREE GENERATION-TIME SIGNALS GATE A VIDEO LLM BETTER THAN A 32B JUDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large language models sample many candidate answers, and a large share of those candidates are wrong, so a deployed system has to decide for each answer whether to act on it. We ask which verification signal should make that decision. Two signals the sampling protocol already produces, the model's own confidence and its self-consistency across samples, separate right answers from wrong ones better than a 32B vision-language judge run once per question: they beat the judge in 6 of 8 model-dataset cells with question-level bootstrap intervals excluding zero, at no inference cost beyond the samples already drawn. The evidence is a scored pool of 36,000 labelled candidates from two open backbones on four benchmarks, run through five families of verification signal (confidence, self-consistency, the judge, frame-evidence agreement, and temporal metamorphic probes), each with a finite-sample false-accept gate and its inference cost recorded. The two video-specific families are the weakest, and combining all five adds little over the best single signal. Gating on any of these signals needs one repair: on tied multiple-choice scores a naive split-conformal gate exceeds its false-accept bound in nearly every single-signal setting, and standard randomized tie-breaking restores validity within seed noise. We release the scored pool so that a new signal or gate can be compared against the free baselines without a GPU.

1 INTRODUCTION

On the four benchmarks in this study, when two widely deployed open backbones sample 9 answers per question, only 52.1–74.1% of the 36,000 sampled candidates are correct, so a system that acts on every sampled answer inherits a substantial error rate. Selective prediction formalizes the choice: accept or abstain, subject to a bound on the rate at which wrong answers slip through. It needs a signal that separates right from wrong answers and a gate with a citable guarantee.

Confidence, sampling self-consistency, judge models and perturbation probes have each been studied as uncertainty estimators, almost always one signal, model and dataset at a time, and mostly for text (Xiong et al., 2024; Wang et al., 2023; Farquhar et al., 2024; Liu & Zhang, 2025), and the closest

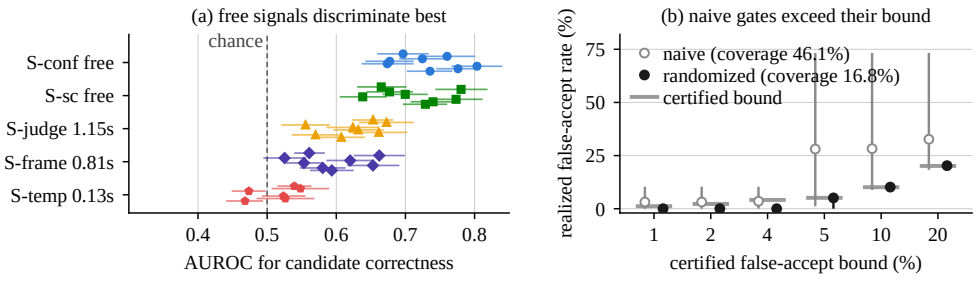

Figure 1: **Free signals lead; the randomized gate holds its bound.** *(a)* AUROC per cell with its 95% question-level bootstrap interval, families ordered by measured marginal cost. *(b)* Per certified bound, the range and median over the settings sharing it of the rate realized over 40 splits; naive hollow, randomized filled, the grey tick the bound.

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.

video-side selective-prediction study we found sweeps a confidence threshold on one closed model and one dataset (Ortiz, 2026). Conformal risk control supplies finite-sample false-accept bounds, instantiated for multimodal and foundation models by TRON and COIN (Wang et al., 2025a;b), but each certifies the one nonconformity score it proposes, with validity argued for continuous scores. Video question answering is the opposite regime: multiple-choice and yes/no answers induce heavily tied discrete scores, and sampling 9 candidates per question clusters test items within questions. Analyses of when such certificates hold (Kotte, 2026; Hu et al., 2026) do not cover this regime.

For each of four benchmarks (Video-MME, MVBench, NExT-QA, VideoHallucer) we fix a seeded 500-question pool, draw 9 candidates per question from two open backbones (Qwen2.5-VL-7B, LLaVA-OneVision-7B), and label every candidate against ground truth. Five signal families score each candidate: model-side confidence S_{conf}, sampling self-consistency S_{sc}, a 32B vision-language judge S_{judge}, frame-evidence agreement S_{frame} across disjoint frame subsets, and temporal metamorphic probes S_{temp} from frame shuffling and reversal, signed by each benchmark's temporal flag. We measure discriminative power per family (RQ1), combination structure (RQ2), certificate validity when each signal drives a wrong-side split-conformal gate under discreteness and clustering (RQ3), and per-item cost (RQ4), releasing the 36,000 labeled, fully scored candidates as a benchmark for gating research.

The signals that cost nothing at generation time are the strongest: S_{conf} is the numerically best single family in 6 of 8 cells, and the two free families beat the 32B judge in all 8 cells, in 6 of 8 with question-level bootstrap intervals excluding zero (Figures 1a and 3); a control that judges every candidate rather than one per question, on a 20% stratified subset, narrows that gap without reversing it (Section 4.1). The video-native families that intuition favors are the weakest: S_{frame} clears chance in 7 of 8 cells, and S_{temp} falls on either side of it, above chance in 2 of the 6 cells where it applies and below chance in 2. Combining all five adds at most +0.053 AUROC over the best single, and the judge's leave-out share of it clears zero in 4 of 8 cells.

On the certificate side the transfer from text fails where it can be tested: naive gates on tied scores exceed their nominal bounds in 95.6% of single-signal settings, by up to +68.1pp, a failure mode that survives certifying per question rather than per candidate. Randomized tie-breaking repairs validity to the resolution of our 40-seed design, but single-signal seed-means still scatter around the bound by a median 10.7pp, and a pre-registered sweep shows that scatter is set by the labelled test half, not the calibration half.

Two pre-registered analyses say why the video-native families fail: the temporal probes are broken by the flag that signs their score, not by the probing (0.337–0.549 on flagged questions, 0.521–0.602 on the rest), and the sparsity explanation for frame-evidence agreement holds in 1 of 2 backbones and so is not supported. More generally, this is a study of which verifier signal should drive the gate: the choice between a model's own signals and an external judge, and the repair a discrete score forces on the certificate, arise for any selective gate over a discretely scored candidate; per-step gates on an agent's actions are not tested here.

The contributions are the matrix itself (Table 1; all 36,000 candidates released), the ordering it inverts (Figures 1a and 3), the certificate stress test and calibration sweep that separate a testable naive failure from residual uncertainty more calibration cannot remove (Table 2, Figure 4), the two mechanism tests (Figure 5) and the cost axis (Table 3).

2 Related Work

What video-LLM benchmarks measure. Accuracy suites such as Video-MME and MVBench report a correctness gap that widens with clip length and temporal demand (Fu et al., 2024; Li et al., 2024b); a parallel line isolates hallucination with adversarial yes/no probes, captioning scores, long-video aggregation, synthetic stress tests, a cross-modal taxonomy and grounded-evidence benchmarks (Wang et al., 2024; Rawal et al., 2025; Lu et al., 2025; Li et al., 2025; Sahoo et al., 2024; Huang et al., 2026). These score model quality, not a per-answer signal a deployer could threshold, and none bounds the rate at which a wrong answer is accepted; we use them as label sources.

The signal families, studied one at a time. Confidence: post-hoc calibration (Guo et al., 2017), verbalized estimates (Xiong et al., 2024; Tian et al., 2023), and multimodal extensions (Xuan et al.,

2025; Dang et al., 2026; Zhao et al., 2025; Ding et al., 2025; Liu et al., 2025), with asymptotic risk calibration for selective QA (Lin & Dong, 2026). Sampling consistency: majority voting (Wang et al., 2023), semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024; Lau et al., 2026), video variants over frame subsets (Miyazawa & Okuno, 2026; Chung et al., 2025), among the strongest black-box scores in text (Xia et al., 2025). Second-model judges: reliability analyses of video-LLM judges (Liu & Zhang, 2025), conformal and selective treatments of a judge's own uncertainty (Sheng et al., 2025; Badshah et al., 2026), disagreement as a routing signal (Saluru, 2026), same-model self-verification as a conditional confidence signal (Phalod, 2026), target-checked reliability refinement on video QA (Ren & Chandra, 2026), and checkers that reject by rule and stay silent where they do not apply (Xiao et al., 2026a). Temporal probes: instability under reordering and rephrasing, stress tests of grounding, and channel decompositions of what a temporal score measures (Jung et al., 2024; V et al., 2026; Rahman, 2026), alongside long-video trustworthiness benchmarks and perturbation-plus-entropy detectors (Yu et al., 2026; Gautam et al., 2026). Each instruments one signal on one model and dataset; family comparisons exist in text (Richardson, 2026; Bouchard & Chauhan, 2025), and we found none on video.

The certificate. Conformal risk control wraps a score in a finite-sample guarantee: factuality bounds for text generation (Mohri & Hashimoto, 2024), selective-QA and sample-then-filter guarantees for foundation models (Wang et al., 2025b; 2026), two-level control extended to multimodal and VideoQA outputs (Wang et al., 2025a), and analyses of when such certificates hold (Kotte, 2026; Hu et al., 2026); the same calibrated gate risk-controls labeling outside text (Xiao et al., 2026b). A set-valued branch returns guaranteed-coverage answer sets for choice outputs, meeting the same tie arithmetic while optimizing set size, not false-accept rate (Li et al., 2026; Ye, 2025; Yang & Liu, 2025; Ye & Wen, 2025). Each certifies one score with validity argued for continuous scores; whether that survives video's tied scores and the clustering of many candidates per question is what RQ3 tests.

The two priors this matrix inverts. The closest precedent is a selective-prediction study for text-to-SQL that scores black-box, log-probability, judge and judge-ensemble families on one pool and reports judge ensembles reaching the highest AUROC (Richardson, 2026), with multi-scorer suites in text agreeing that a second model helps (Bouchard & Chauhan, 2025; Wang & Zhang, 2026); perturbation diagnostics are read as revealing genuine fragility (Jung et al., 2024; Gautam et al., 2026). The closest video-side abstention study we found sweeps a confidence threshold on a single closed model and dataset (Ortiz, 2026); the same one-knob pattern recurs over sensor streams (Xiao et al., 2025), and channel fusion is a surveyed design space (Xu et al., 2026) that RQ2 measures.

3 Study design

3.1 Task, pools, and labels

We study answer-level selective prediction for video question answering: a generator emits an answer to a question about a video, and a gate accepts it or abstains. Four public benchmarks supply complementary regimes: Video-MME (Fu et al., 2024) for multi-domain multiple choice, MVBench (Li et al., 2024b) for temporally heavy multiple choice, NExT-QA (Xiao et al., 2021) for causal and temporal reasoning, and VideoHallucer (Wang et al., 2024) for adversarial yes/no probing. From each we draw a seed-fixed pool of 500 questions. Every pool is multiple choice or yes/no, so labels are exact match on the canonicalized option or yes/no token; no human or model judges them.

Two open-weight backbones generate answers: Qwen2.5-VL-7B (Bai et al., 2025) and LLaVA-OneVision-7B (Li et al., 2024a), both sampling 8 frames per video. For each question we generate one greedy answer plus 8 temperature samples, giving 9 candidates per question and $2 \times 4 \times 500 \times 9 = 36,000$ labeled candidates overall (Figure 2). The two backbones $\times$ four benchmarks yield 8 model-dataset cells; all measurements below are reported per cell.

3.2 Five verification-signal families

Model-side confidence S_{conf}. Length-normalized log-probability of the candidate under its own generator, read directly from generation. Zero added inference.

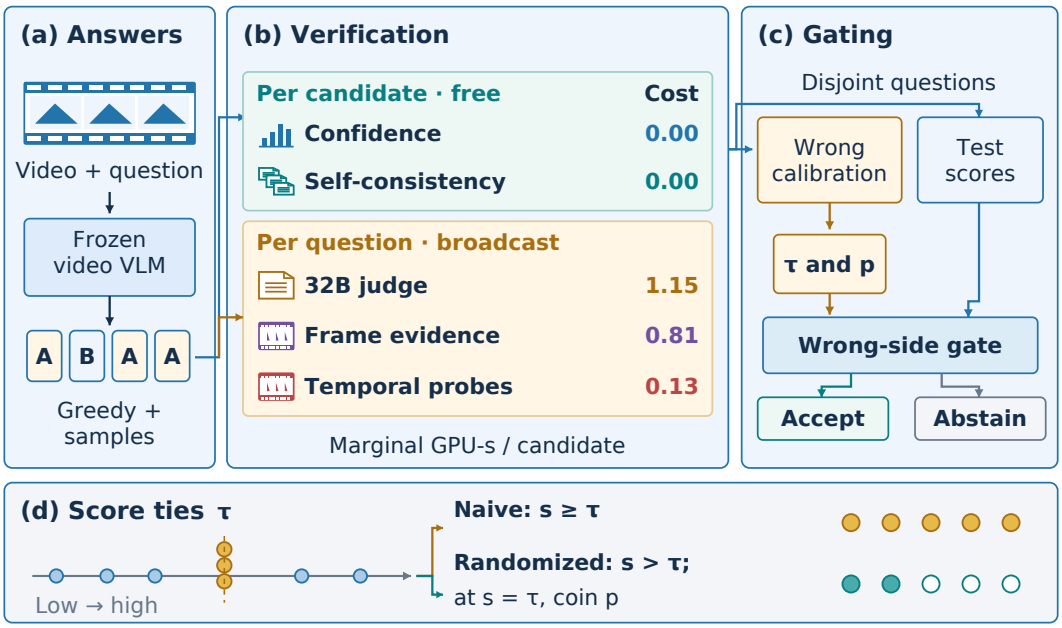

Figure 2: **The study, and the gate inside it.** *(a)* One seeded pool of 500 questions per benchmark, 9 candidates per question per backbone (one greedy, the rest sampled), exact-match labels. *(b)* Confidence and self-consistency cost no extra inference; the judge, frame-evidence and temporal-probe families score the greedy answer once per question and broadcast it to every candidate. The costs shown are mean marginal GPU-seconds per scored candidate over the cells (Table 3); the judge's figure includes a superseded first pass. *(c)* Calibration on the wrong-only half, test on a question-disjoint half; the wrong-side gate is read as false-accept rate and coverage. *(d)* The threshold lands inside a tie block of coarse scores; naive acceptance takes the block whole, the calibrated coin takes a fraction. Panels are schematic.

Self-consistency S_{sc}**.** The share of the question's other sampled candidates whose canonical answer agrees with this candidate (Wang et al., 2023). No inference beyond the existing samples.

VLM judge S_{judge}**.** Qwen2.5-VL-32B-Instruct, run unquantized in bf16, scores the greedy answer 0–10 given the question and the same 8 frames, following the LLM-as-judge pattern (Liu & Zhang, 2025). The score is per question and broadcast (per-candidate control: Appendix E); a strict single-`SCORE` format is required and unparseable outputs are excluded, never imputed, so S_{judge} metrics use the parseable subset and paired comparisons the common subset (Appendices A and C).

Frame-evidence agreement S_{frame}**.** The answer is regenerated from disjoint sparse frame subsets, and the signal is the agreement rate between the *greedy* answer and the subset answers, scored once per question and broadcast like S_{judge}.

Temporal metamorphic probes S_{temp}**.** The greedy answer is regenerated on temporally shuffled and on reversed frames, and the agreement between the probe answers and the original is *signed by the benchmark's own temporal flag*: on a flagged question the answer should change once frame order is destroyed, so the score is one minus the agreement, and on a non-flagged question it is the agreement itself. The flag is benchmark metadata, not ours (every MVBench question, the NExT-QA types labelled T, the VideoHallucer temporal category); Video-MME publishes none, so S_{temp} is not applicable there and covers 6 of 8 cells. Three of the five families are thus per-question; the two that vary per candidate are the two that win (Section 4.1).

As a discriminative upper envelope, a logistic regression over all family scores is fit per cell with group-aware 5-fold cross-validation (RQ2); in this stage only, missing signal values enter as midpoint-imputed features with explicit missingness indicators (Appendix C). TRON's sampling-frequency nonconformity (Wang et al., 2025a) is additionally implemented from the banked samples as a tested gate variant, not a signal family; within each pool it is a monotone map of S_{sc}, so its gate rows coincide with S_{sc}'s at $\alpha \leq 0.1$; Appendix B records the mapping and where it diverges.

Table 1: **RQ1/RQ2/RQ4: the free families top every cell.** AUROC for candidate correctness over each cell's full pool, with the all-family logistic combination and marginal cost. Bold marks the best single family; intervals in Table 17 and Figure 3. S_{judge} uses its parseable subset; S_{temp} is n/a on Video-MME.

Signal	Cost (s/cand.)	Qwen2.5-VL-7B				LLaVA-OneVision-7B			
		V-MME	MVB	NExT	VHall	V-MME	MVB	NExT	VHall
Candidate accuracy (%)	–	52.1	59.6	74.1	67.1	52.7	55.9	73.4	63.7
S_{conf} (confidence)	0.00	**0.736**	**0.776**	**0.804**	**0.674**	0.677	**0.725**	0.760	**0.697**
S_{sc} (self-consistency)	0.00	0.729	0.740	0.773	0.638	**0.700**	0.677	**0.780**	0.665
S_{judge} (32B judge)	1.15	0.607	0.570	0.662	0.632	0.624	0.556	0.673	0.653
S_{frame} (frame evidence)	0.81	0.593	0.580	0.653	0.553	0.620	0.525	0.662	0.561
S_{temp} (temporal probes)	0.13	–	0.468	0.526	0.524	–	0.474	0.548	0.539
All families (logistic)	–	0.741	0.770	0.835	0.711	0.734	0.717	0.833	0.730

3.3 Certified false-accept gates

Each signal is fed into a wrong-side split-conformal gate (Wang et al., 2025b; Mohri & Hashimoto, 2024). Candidates from a calibration split that are labeled wrong contribute their scores; the gate accepts a test candidate when its score exceeds the upper α order statistic of these wrong-side scores. Its nominal bound is $\Pr[\text{accept} \mid \text{wrong}] \leq \alpha + \epsilon_n$, with $\epsilon_n = 1/(n^- + 1)$ over the n^- wrong-side calibration candidates. The bound is exact only when calibration and test scores are exchangeable, and candidates of one question are not exchangeable with those of another (a per-question signal repeats one value across all its candidates), so at the candidate unit the bound is nominal, not certified. The exchangeable unit is the question; Appendix G recertifies every setting with one anchor candidate per calibration question, and Section 4.3 reports both units. Main-text macros use the candidate unit. The bounded quantity is the false-accept rate, not the error rate among accepted answers; the two differ by a base-rate factor, and we report coverage alongside. A composed *all* gate Bonferroni-splits α across the families; for a conjunction the full-α bound already holds, so the split is conservative and the gate abstains almost everywhere (Table 2).

Two properties of video scores stress this guarantee. First, *discreteness*: every family takes few values per cell (Table 3); even S_{conf} puts 12.3–50.7% of a cell's candidates on one value, so the order statistic sits inside a tie block (Figure 2d). Second, *within-question clustering*: a question's 9 candidates share the video, so test items are not independent. We therefore evaluate each gate in two modes: *naive* thresholding, and *randomized* tie-breaking, which accepts score-equal-to-threshold items with a plug-in probability that sets the expected count of accepted wrong calibration candidates to $\alpha(n^- + 1)$. Calibration and test splits are group-aware (250 question groups each side, questions never straddle the split), and every number is the mean over 40 split seeds. Gates are evaluated at $\alpha \in \{0.05, 0.1, 0.2\}$ per cell, giving 138 (cell, signal, α) settings per mode once TRON's rows, which duplicate S_{sc}'s, are set aside (Appendix B).

3.4 Cost accounting

Every family's marginal GPU cost is measured on the same hardware as GPU-seconds per scored candidate, with the GPU-hours its stage consumed (Table 3); the judge figure includes a superseded first pass (Appendix A). The main matrix and its judge controls consumed 25.7 of 56 GPU-hours.

4 Results

4.1 RQ1: which families discriminate?

Table 1 presents the matrix, from a single generation run, so paired claims below carry question-level bootstrap intervals (Figure 3). First, the free families lead: S_{conf} is the numerically best single signal in 6 of 8 cells and S_{sc} in 2 more, with AUROC between 0.638 and 0.804 (the S_{conf}-vs-S_{sc} separation excludes zero in 5 of 8 cells). Second, the 32B judge, the most expensive signal, is dominated by

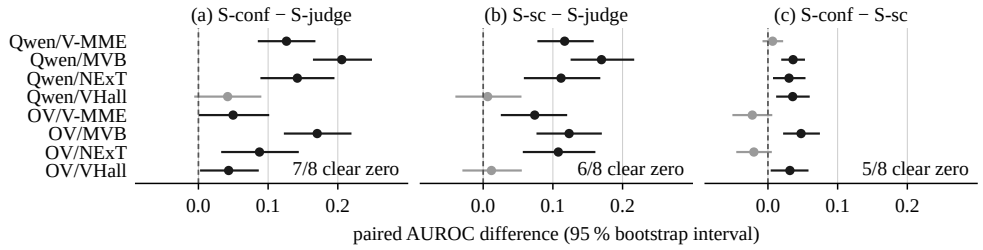

Figure 3: **RQ1: both free families beat the judge.** One row per cell, one 95% question-level cluster bootstrap per contrast; dark rows exclude zero. Values in Table 4.

signals that cost nothing: both S_{conf} and S_{sc} exceed it numerically in all 8 cells, with paired intervals excluding zero for both in 6 of 8 (S_{conf} alone: 7); the two S_{sc} margins on VideoHallucer are within noise. How the judge is deployed does not change the ordering: rescoring every candidate with the same judge on a 20% stratified subset lifts it by +0.021 AUROC (across-cell 95% interval [+0.010, +0.031]) and still leaves it under S_{conf} in 6 and under S_{sc} in 6 of 8 cells (Appendix E); the suspected family confound finds no support (Appendix D). Third, the video-native families are weaker than their motivation predicts: S_{frame} ranges 0.525–0.662 with an interval clear of chance in 7 of 8 cells, while S_{temp} ranges 0.468–0.548 over the 6 cells where it applies and excludes chance in 2 of them from above and 2 from below, so it is not uninformative so much as inconsistently signed (Section 5). A second sampling seed leaves all 16 interval verdicts standing (largest AUROC shift 0.040), and the free signals stay above chance at 3B and 32B (Appendix F). **The signals a deployer already has discriminate better than the ones bought by perturbing the video or by asking a larger model.**

4.2 RQ2: IS THERE COMPLEMENTARY STRUCTURE?

If the families measured different failure modes, combining them should raise AUROC. It does, by little: the all-family logistic combination improves on the best single signal by between -0.008 and +0.053 AUROC per cell, of the order of the fold SD (0.019–0.064); with both sides out of fold on the same question groups and a paired question-level bootstrap (2,000 resamples, Appendix H) the gain's interval clears zero in 4 of 8 cells and falls below it in 0. Removing S_{judge} from the combination costs an interval that clears zero in 4 cells, at most +0.044 AUROC (Appendix H). **The verification information is concentrated in the generation logs; the remaining four signals together add a second, small source of it, at most +0.053 AUROC; the judge's leave-out cost is the largest of the four in 7 of 8 cells and clears zero in 4, for 10.4–17.8 GPU-seconds per question.**

4.3 RQ3: DO THE CERTIFICATES HOLD?

Table 2 reports realized false-accept rates at target $\alpha = 0.1$ (other budgets: Table 7; coverage: Table 8). Shares below are over the 114 single-signal (cell, signal, α) settings, the composed *all* gate being reported apart and TRON's rows, which duplicate S_{sc}'s (Appendix B), not counted; significance counts are over the 138 settings of a mode, the composed gate's included. The naive gate, the direct port of the text procedure, is not usable on video scores: its seed-mean exceeds the nominal bound in 95.6% of them (106 of 138 significant under one-sided seed-level t-tests with Bonferroni correction), by +24.8pp on average among the violating single-signal settings and up to +68.1pp (Figure 1b), reaching 73.0% against a bound of 5.1% where the score is coarsest.

Randomized tie-breaking is the textbook repair, and the evidence is consistent with it: the mean excess drops to +0.1pp over the single-signal settings (-0.3pp with the composed gate included) and no setting's excess is distinguishable from zero (0 of 138 Bonferroni-corrected rejections, against 106 naive). 59 of the 114 single-signal settings (51.8%) sit above their bound in seed-mean, about the share an exactly valid gate produces, and the *all* gate is 0 of 24. The largest excess, +2.2pp, is smaller than its own setting's 40-seed range of 25.7pp, and the single-signal residue of +0.1pp sits an order below the median seed range (10.7pp): uncertainty, not failure. Certifying per question rather than per candidate, one anchor drawn per calibration question (Appendix G), widens the bound by

Table 2: **RQ3: naive gates break the bound on ties; randomized holds it** (target $\alpha = 0.1$; other α in Appendix C). Realized false-accept rate (%), mean over 40 group-aware splits. The nominal bound $\alpha + \epsilon_n$ (candidate unit; question unit in Appendix G) is 10.1–10.2% for the signal gates, the composed *all* gate certifies against its Bonferroni-split bound, and bold marks seed-means above their bound. TRON's rows equal the S_{sc} rows at this α (Appendix B).

		Qwen2.5-VL-7B				LLaVA-OneVision-7B			
Gate signal	Mode	V-MME	MVB	NExT	VHall	V-MME	MVB	NExT	VHall
S_{conf}	naive	**10.9**	**10.8**	**17.2**	**21.1**	8.8	10.0	**15.9**	9.7
	rand.	**10.8**	**10.5**	**12.0**	9.9	8.8	9.9	**12.2**	9.4
S_{sc}	naive	**15.4**	**20.0**	**28.3**	**38.5**	**22.6**	**27.6**	**27.9**	**16.0**
	rand.	**10.4**	**11.2**	**11.0**	10.0	9.6	**10.5**	**10.9**	9.7
S_{judge}	naive	**35.0**	**56.8**	**39.4**	**47.1**	**31.1**	**56.7**	**34.3**	**39.0**
	rand.	9.9	10.0	**10.9**	**10.3**	9.7	9.6	**11.2**	10.1
S_{frame}	naive	**51.4**	**58.1**	**56.8**	**73.3**	**45.3**	**66.3**	**55.2**	**72.7**
	rand.	**10.4**	10.0	**10.4**	9.9	**10.2**	9.5	**10.5**	10.0
All (Bonferroni)	naive	**3.2**	0.0	**3.9**	**10.4**	1.1	0.0	**4.4**	**3.1**
	rand.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 3: **RQ4: the strongest families are free.** Marginal GPU-seconds per scored candidate, the GPU-hours the family's stage consumed over the 8 cells, and the distinct values the score takes in a cell, the quantity RQ3 turns on. The judge's measured cost includes a superseded first pass; the declared protocol alone projects to 1.98s.

Family	GPU-s per candidate	stage GPU-hours	distinct score values per cell
S_{conf} (confidence)	0.00	—	599–1476
S_{sc} (self-consistency)	0.00	—	14–15
S_{judge} (32B judge)	1.15	11.51	4–7
S_{frame} (frame evidence)	0.81	8.11	3–4
S_{temp} (temporal probes)	0.13	0.98	3
Generation 2.05; judge controls 0.28 and 2.76		25.7 of 56	

0.74–1.36pp and moves realized risk by at most 2.61pp; the naive gate still exceeds it in 101 of 114 settings, the randomized gate in 10. Both halves replicate on the 3B and 32B pools: the naive gate exceeds α (not $\alpha+\epsilon_n$) in 15 of 16 entries by up to +27.4pp, the randomized gate by at most +1.5pp (Appendix F). Validity costs coverage: averaged over all gate settings, the naive gate accepts 46.1% of candidates and the randomized gate 16.8% (Figure 1b), so selective video QA at these risk levels means abstaining on most inputs. **Discreteness breaks the ported bound testably, the known repair restores it to the resolution of a 40-seed design, and what remains is a cost in coverage rather than a broken bound.**

4.4 RQ4: WHAT DOES DISCRIMINATION COST?

Table 3 and Figure 1a give the cost side. The two strongest families are free given the protocol's nine candidates per question (S_{conf} unconditionally, S_{sc} only once several candidates are drawn; a single-answer deployment pays 1.64 GPU-seconds per question for its eight extra samples); S_{judge} adds 1.15 GPU-seconds per candidate and 11.51 GPU-hours for numerically less discrimination than either free family in every cell, and S_{frame} adds 0.81s and 8.11 GPU-hours for no discrimination gain over the free signals, so the cost-discrimination frontier is degenerate. We release all 36,000 scored candidates as JSONL (Appendix I), on which new scores and gates can be evaluated without GPU inference. **No paid family reached the discrimination the generation logs already held.**

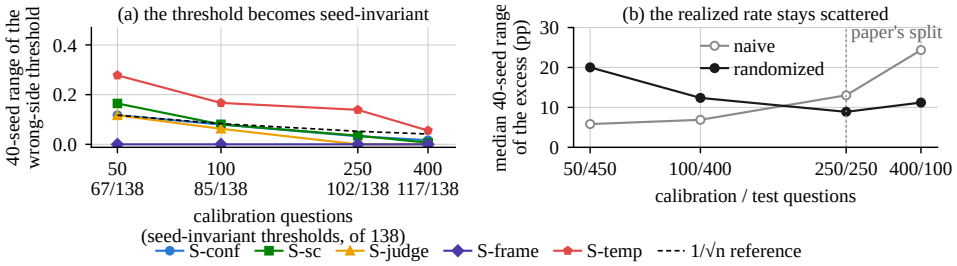

Figure 4: **Calibration size does not deliver per-setting assurance.** *(a)* 40-seed range of the wrong-side threshold, averaged over cells and targets, against a $1/\sqrt{n_{\text{cal}}}$ reference; counts are the settings whose threshold is seed-invariant. *(b)* Median 40-seed range of the excess over the bound, over the 162 settings of each mode, TRON's and the composed gate's included.

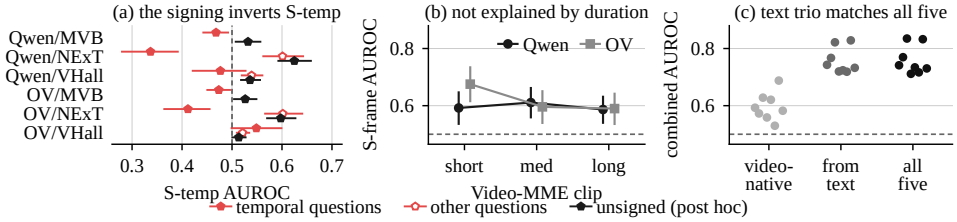

Figure 5: **The temporal flag inverts S_{temp}, duration does not explain S_{frame}, and the text trio matches all five.** *(a)* S_{temp} by temporal flag, 95% intervals; MVBench is entirely flagged, and the black points undo the signing and are post hoc. *(b)* S_{frame} by Video-MME clip duration. *(c)* Video-native, text-transplanted and all five combined.

5 Why the ordering holds

Section 4 reports what the matrix measures; this section reports why. The first account re-reads reported measurements; the three that follow were registered before they were run (Appendix K), and every outcome is reported, including the two that went against this paper's own explanations.

Why the naive certificate breaks. The wrong-side threshold is an order statistic of coarse calibration scores (Table 3), so it lands inside a tie block and naive thresholding accepts the whole block (Figure 2d). The same coarseness shows from the calibration side: at the study's calibration size the threshold is identical across all 40 seeds in 102 of 138 single-signal settings, TRON's included (Figure 4a). Decoding at T=1.0 lowers the tie share, and the naive excess falls with it in all 4 cells tried (Appendix F). **Discreteness is the failure mode of the naive gate, not a second-order effect.**

What a larger calibration set changes. The registered quantity is the threshold's 40-seed range, predicted to fall as $1/\sqrt{n_{\text{cal}}}$ (log-log slope in $[-0.75, -0.25]$) against a competing prediction that discreteness holds it above -0.25. Both fail: the range falls faster (slope -0.93 for S_{conf}) and reaches exactly zero for the coarse families, so by $n_{\text{cal}} = 400$ the threshold is seed-invariant in 117 of 138 settings. The quantity a deployer has to check, the excess over the bound, is descriptive here and not monotone in calibration size: its median 40-seed range is 20.0pp at $n_{\text{cal}} = 50$, 12.4pp at 100, 8.9pp at 250 (the split this study uses) and 11.2pp at 400 (Figure 4b, Table 18): the test half is the complement, so whichever half is smaller sets the dispersion. **Per-setting verification is bounded by the smaller half of a fixed labelled budget, so no split of that budget provides it.**

Why the temporal family is weak. S_{temp} is signed by the benchmark's temporal flag (Section 3.2), so the registered contrast splits each benchmark by that flag. The registered explanation (instability measures the model's fragility, not this answer's correctness) predicted both strata at chance; the competing reading (real signal on genuinely temporal questions) predicted the flagged stratum above chance. The measurement rejects both. Over the 4 registered flagged strata the AU-

ROC runs 0.337–0.549: the interval falls entirely *below* chance in 2 of them, above it in 0 and contains chance in 2. The 2 all-flagged MVBench cells, descriptive under the registration, also fall below chance. Over the 4 non-flagged strata the AUROC runs 0.521–0.602 and lies entirely above chance in 4, and the difference excludes zero in 3 of 4 contrasts (Figure 5a, Table 19). The probes carry information; the signing inverts it on just those questions the family was designed for, because a stable answer is the likelier to be correct whether or not the question is temporal. Scoring raw agreement instead, an analysis chosen after seeing this result and reported as exploratory, gives 0.513–0.625, above chance in 5 of 6 cells. **The failure is in the signing rule, not the probe, and the repaired score still sits under the free signals' 0.638 floor: Table 1 stands.**

Why the frame-evidence family is weak. The registered prediction was that a fixed 8-frame budget grows sparser with clip length, so S_{frame} AUROC should fall with duration on Video-MME. The ordering holds in 1 of 2 backbones (short minus long +0.086 [+0.002, +0.171] against +0.006 [−0.072, +0.082]), and the one interval that clears zero clears it by 0.002 AUROC (Figure 5b, Table 19). **Sparsity is not supported at this frame budget.** The budget is: 16 frames lift S_{frame} by +0.048 and +0.040 AUROC on Qwen2.5-VL, within interval (Appendix F).

Why combination adds little, and why the judge carries what it adds. The two video-native families combined reach 0.530–0.688 AUROC, the three transplanted from text 0.719–0.829, and all five 0.711–0.835, the text trio being at least as good as the full set in 4 of 8 cells (Figure 5c). Redundancy explains the flat free-pair marginals (S_{conf} and S_{sc} correlate at 0.69–0.87) and why the largest marginal belongs to S_{judge} (+0.044, the largest in 5 of 8 cells, clearing zero in 4, Section 4.2): the judge is the least correlated feature, not the most informative one. Its disadvantage is what it scores, the greedy answer rather than the candidate ranked (judging every candidate, at 9 times the cost, closes only 23–25% of the gap, Appendix E), and its coarse integer output.

6 Discussion, limitations and conclusion

What should a deployer actually run? Threshold S_{conf}, which is free, or S_{sc}, free once the deployment already samples several candidates and otherwise 1.64 GPU-seconds per question (Table 3). Adding all other signals, the judge among them, gains at most +0.053 AUROC, and the judge's own leave-one-out contribution is smaller, for 10.4–17.8 GPU-seconds more per question. If a finite-sample certificate is required, calibrate per question (Appendix G) and use the randomized wrong-side gate on a free signal, read the guarantee as aggregate (Section 5), and expect abstention (Table 2).

Limitations and threats to validity. Labels are exact-match on multiple-choice and yes/no pools; free-form video QA is out of scope. Both backbones are open 7B models; larger or closed ones may reorder the families. Unparseable judge outputs are excluded, not imputed (Appendix A).

What the measurement says. Across five signal families, two backbones and four benchmarks, the free signals are the strongest gates, the paid families add little, and certificates designed for continuous scores break on video's tied scores until ties are randomized.

Reproducibility statement

All signal scores, correctness labels, and gate outcomes derive from a single seeded pipeline over public benchmarks and open-weight models. We release the full 36,000-candidate scored pool, the harness, and per-stage artifacts; during review they are provided as an anonymized artifact with the supplementary material. Every numeral in this paper is generated from the released result file by a macro generator, with no hand-transcribed numbers. Appendix J details prompts, decoding parameters, and compute.

AI use statement

In this work we disclose generative AI use under the categories of the ICLR 2027 AI policy. Within the required-disclosure categories, we used generative AI tools to help develop the conceptual framework, propose and refine hypotheses, formulate the paper's formal claims and statistical analyses,

give feedback on research methodology and experiment design, implement methods (the experiment, analysis and figure-generating code released with this work), clean and reformat datasets, and interpret results. We have *not* used generative AI tools to generate synthetic data sets, assist with translation, support qualitative or thematic data analysis, or write proofs (this paper states no formal proof); the remaining required-disclosure categories are not applicable to this work, which has no human-subject component and no data collection beyond public benchmarks and the outputs of open-weight models. Under the recommended categories, generative AI was additionally used for software code, literature search and analysis, brainstorming, figure preparation, and editing the text of this paper. No experimental result was produced, selected or filtered by a generative AI tool: every experimental numeral in this paper is emitted by a committed single-writer artifact and reaches the text through an auto-generated macro bridge that admits no hand-typed value; design constants, protocol settings and figures quoted from prior work are of course written by hand and cited. The research questions, the claims and the conclusions are the authors'. We have reviewed all AI-assisted work and take responsibility for the final content, including every claim, number and artifact.

Ethics statement

The four benchmarks used here (Video-MME, MVBench, NExT-QA and VideoHallucer) are public, and both backbones are public open-weight checkpoints. We collected no data and recruited no participants, and the videos contain people we have no relationship with, so we release no frames, clips or stills. The scored pool is numeric: per row, the model and dataset names, the benchmark's own question id and a candidate index, one correctness bit, and the five signal values. Nothing in it describes what is in a video. The review package also contains the per-stage generation dumps (Appendix I), which carry each benchmark question's text and options and each candidate's decoded answer text, but no frames, clips or stills. Anyone reproducing this work obtains the videos from the benchmark authors under their terms, as we did. On the method itself, one misreading is worth naming: the gate scores whether an answer is likely to be correct, not whether it is safe, and a deployment that abstained on the strength of this signal would be abstaining on predicted error alone. Nothing here is calibrated against harm.

References

Sher Badshah, Ali Emami, and Hassan Sajjad. Scope: Selective conformal optimized pairwise llm judging. arXiv:2602.13110, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. arXiv:2502.13923, 2025.

Dylan Bouchard and Mohit Singh Chauhan. Uncertainty quantification for language models: A suite of black-box, white-box, llm judge, and ensemble scorers. *Transactions on Machine Learning Research (TMLR)*, 2025.

Hyungjin Chung, Hyelin Nam, Jiyeon Kim, et al. Video parallel scaling: Aggregating diverse frame subsets for videollms. arXiv:2509.08016, 2025. Accepted CVPR Findings 2026.

Yunkai Dang, Yifan Jiang, Yizhu Jiang, Anqi Chen, Wenbin Li, and Yang Gao. Instinct vs. reflection: Unifying token and verbalized confidence in multimodal large models. arXiv:2604.17274, 2026.

Zhikai Ding, Shiyu Ni, and Keping Bi. Do lvlms know what they know? a systematic study of knowledge boundary perception in lvlms. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024. doi: 10.1038/s41586-024-07421-0.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52734.2025.02245.

Sushant Gautam, Cise Midoglu, Vajira Thambawita, Michael A. Riegler, and Pål Halvorsen. Videohedge: Entropy-based hallucination detection for video-vlms via semantic clustering and spatiotemporal perturbations. arXiv:2601.08557, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Anqi Hu, Zhiyuan Wang, Zijun Jia, et al. Mird: Reliable set-valued prediction for open-ended question answering via miscoverage risk decomposition. arXiv:2605.27091, 2026.

Linpeng Huang, Weixing Chen, Zexin Chen, et al. Eg-vqa: Benchmarking verifiable video question answering with grounded temporal evidence. arXiv:2606.24797, 2026.

Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. On the consistency of video large language models in temporal comprehension. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52734.2025.01280.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Gregory Kang Ruey Lau, Hieu Dao, Nicole Kan Hui Lin, et al. Uncertainty quantification for multimodal large language models with incoherence-adjusted semantic volume. arXiv:2602.24195, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, et al. Llava-onevision: Easy visual task transfer. arXiv:2408.03326, 2024a.

Kunchang Li, Yali Wang, Yinan He, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. doi: 10.1109/CVPR52733.2024.02095.

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

Zongxia Li, Xiyang Wu, Guangyao Shi, et al. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Shufan Lin and Sijin Dong. Asymptotic risk calibration for selective question answering. arXiv:2608.12008, 2026.

Ming Liu and Wensheng Zhang. Is your video language model a reliable judge? In *International Conference on Learning Representations (ICLR)*, 2025.

Xiaoou Liu, Tiejin Chen, Longchao Da, et al. Uncertainty quantification and confidence calibration in large language models: A survey. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025. doi: 10.1145/3711896.3736569.

Hao Lu, Jiahao Wang, Yaolun Zhang, et al. Elv-halluc: Benchmarking semantic aggregation hallucinations in long video understanding. arXiv:2508.21496, 2025.

Tomoya Miyazawa and Hiroyasu Okuno. Answer self-consistency with margin-triggered question re-arbitration for the cvpr 2026 vidllms challenge. arXiv:2606.04323, 2026. CVPR 2026 VidLLMs Challenge (workshop).

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *International Conference on Machine Learning (ICML)*, 2024.

Jorge Ortiz. Explicit abstention knobs for predictable reliability in video question answering. arXiv:2601.00138, 2026.

Aditya Ajay Phalod. When should a language model trust itself? same-model self-verification as a conditional confidence signal. arXiv:2605.02915, 2026.

Farrukh Rahman. What does a temporal benchmark score measure? decomposing channel use in video vlm evaluation. arXiv:2607.12304, 2026.

Ruchit Rawal, Reza Shirkavand, Heng Huang, et al. Argus: Hallucination and omission evaluation in video-llms. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/ICCV51701.2025.01886.

Guoxiang Ren and Rohitash Chandra. Target-checked reliability score refinement for video question answering. arXiv:2609.13288, 2026.

Robert Richardson. What predicts correctness in text-to-sql? a selective-prediction study. arXiv:2607.06799, 2026.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, et al. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. doi: 10.18653/v1/2024.findings-emnlp.685.

Durga Sandeep Saluru. Disagreement-based cross-model routing for implicit video question answering. arXiv:2606.14723, 2026.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. Analyzing uncertainty of llm-as-a-judge: Interval evaluations with conformal prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Katherine Tian, Eric Mitchell, Allan Zhou, et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Sethuraman T V, Savya Khosla, Aditi Tiwari, et al. Stress tests reveal fragile temporal and visual grounding in video-language models. arXiv:2602.11244, 2026.

Jiayi Wang and Xu-Yao Zhang. A systematic evaluation of black-box uncertainty estimation methods for large language models. arXiv:2606.19868, 2026.

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. Spotlight.

Qingni Wang, Yue Fan, and Xin Eric Wang. Safer: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, et al. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. arXiv:2406.16338, 2024.

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.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL*, 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations (ICLR)*, 2024.

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, 2026. doi: 10.1016/j.inffus.2025.103455.

Weihao Xuan, Qingcheng Zeng, Heli Qi, et al. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Guang Yang and Xinyang Liu. Conformal sets in multiple-choice question answering under black-box settings with provable coverage guarantees. arXiv:2508.05544, 2025.

Yuanchang Ye. Conformal p-value in multiple-choice question answering tasks with provable risk control. arXiv:2508.10022, 2025.

Yuanchang Ye and Weiyan Wen. Data-driven calibration of prediction sets in large vision-language models based on inductive conformal prediction. arXiv:2504.17671, 2025. Accepted ICIPCA 2025.

Xueqing Yu, Bohan Li, Yan Li, and Zhenheng Yang. Virtuebench: Evaluating trustworthiness under uncertainty in long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yunpu Zhao, Rui Zhang, Junbin Xiao, et al. Object-level verbalized confidence calibration in vision-language models via semantic perturbation. arXiv:2504.14848, 2025.

A JUDGE PROTOCOL AND REMEDIATION

The judge is the unquantized bf16 Qwen2.5-VL-32B-Instruct checkpoint. An AWQ-quantized variant was tried first and abandoned before any reported score was produced, because under forced fp16 it emitted overflow garbage on a large share of calls; the run's configuration key kept the `-awq` suffix from that plan, so the key recorded in the artifacts names a configuration and not the checkpoint's precision. The judge stage prompts it with the question, the answer options when present, the greedy candidate answer, and the same 8 frames the generator saw, and requests an integer score with the exact instruction "`Output ONLY the final score in the exact format: SCORE: <integer from 0 to 10>`". The declared protocol decodes greedily with a 512-token budget; what the retained scores were actually decoded under is the subject of the next paragraph. An output is parsed only when it contains exactly one in-range `SCORE:` match; anything else is excluded from S_{judge} and never imputed.

An initial run with a 32-token budget truncated the judge's unrequested reasoning before the score line on a large share of items, with truncation correlated with item difficulty. The remediation re-ran only the items that had failed to parse, so the S_{judge} column mixes two decoding budgets: 1,765 of 4,000 retained scores were produced at 512 tokens and 2,235 (55.9%) were kept from the 32-token pass. Two consequences, one benign and one not. The parsed *integer* is protocol-invariant: greedy decoding is prefix-deterministic, and each re-run generation is byte-prefixed by its 32-token predecessor in all 1,765 cases, so the score a longer budget would have produced is the score already recorded. *Parseability* is not invariant, because a continuation may emit a second `SCORE:` and our parser accepts only outputs with exactly one. 485 of the retained rows were cut mid-generation at

the 32-token cap, and 10 of the 1,765 re-run generations (0.57%) carry a second `SCORE:`; applying that rate bounds the reported judge coverage (Table 5) at 3 items of 4,000 above what the declared protocol alone would give. Coverage is therefore a union over two budgets, and the AUROCs, which depend only on the parsed integers, are not affected.

B TRON GATE-VARIANT MAPPING

TRON (Wang et al., 2025a) defines its nonconformity from the sampling frequency of a response's semantic cluster among K sampled responses, with clusters formed by an entailment model over free text. In multiple-choice and yes/no pools, semantic equivalence degenerates to exact match on the canonicalized answer, so a candidate's cluster is its answer bucket and no entailment model is defensible. The variant score is the fraction of the question's temperature samples sharing the candidate's canonical answer, counting the candidate itself, which follows TRON's cluster-size definition. This deliberately differs from S_{sc}, which excludes the candidate's own vote: the TRON column is a faithful port of the gate for RQ3, while S_{sc} is the family measurement for RQ1 and RQ2. The TRON score then receives the identical wrong-side calibration and tie-break protocol as every family gate. It is excluded from the composed gate and from RQ1/RQ2, since it is a gate variant rather than a sixth family. The two scores of one gated candidate differ only by whether its own vote is counted: with k temperature samples and a of them agreeing with the candidate, S_{sc} is $a/(k-1)$ for a sampled candidate and a/k for the greedy one, TRON is $(a+1)/k$ and a/k respectively, a fixed strictly increasing map within each pool. The wrong-side threshold is an order statistic of calibration scores, so within one pool the two gates accept the same test candidates. The release's TRON rows match the S_{sc} rows seed for seed at $\alpha = 0.05$ and $\alpha = 0.1$ (0 of the 32 (cell, mode) seed-mean rates differ); at $\alpha = 0.2$ the threshold can fall where the greedy and sampled pools interleave, and since the two maps differ across pools, 11 of the 16 seed-mean rates differ there, by up to 7.1pp. The body therefore reports the S_{sc} block once and counts TRON's settings out of the gate family.

C PER-CELL DETAIL

Combination-stage protocol. Single-family AUROCs (Table 1) are full-sample; combined, best-pair, and leave-one-family-out AUROCs are group-aware 5-fold cross-validation estimates, so gains near zero, including small negative ones, are protocol artifacts rather than evidence that a family subtracts information. In the combination stage a candidate's missing signal values (unparseable judge scores, S_{temp} on Video-MME) are midpoint-imputed and paired with explicit missingness-indicator features; per-family analyses exclude those candidates instead (Section 3.2).

Uncertainty for paired signal comparisons. Table 4 reports question-level cluster-bootstrap intervals for the paired AUROC differences behind the Section 4.1 claims: the 500 questions of a cell are resampled with replacement 20,000 times (fixed seed, all candidates of a drawn question retained, tie-corrected Mann-Whitney AUROC), and 95% percentile intervals are formed. Differences against S_{judge} are computed on the judge-parseable common subset, so both signals in a pair score an identical item set; S_{conf}-vs-S_{sc} uses all candidates.

D CROSS-FAMILY JUDGE CONTROL

The primary judge shares the Qwen2.5-VL family with one of the two backbones, so per-backbone S_{judge} comparisons carry a potential family confound. As the pre-registered control, a cross-family judge (InternVL3-8B) rescored a stratified subsample of greedy candidates in every cell under the identical protocol, and Table 9 reports the AUROC of both judges for greedy-answer correctness on the same paired items; questions where either judge's output was unparseable are excluded pairwise, so both rows of each column score an identical item set ($n = 51$–54 per cell).

The control supports two statements, both at the evidential strength its small subsamples allow. First, it shows no evidence of a same-family advantage: on the paired subsample the primary judge is no better on its own family's answers than on the other backbone's (mean AUROC 0.625 over the Qwen cells vs. 0.653 over the LLaVA-OneVision cells), and under the out-of-family judge the two backbones are indistinguishable (0.573 vs. 0.574). At $n = 51$–54 paired questions per cell the

Table 4: Paired AUROC differences with 95% question-level cluster-bootstrap percentile intervals (20,000 resamples), per cell. Judge comparisons use the judge-parseable common subset. Intervals excluding zero back the CI-qualified cell counts in the main text; the remaining margins are within noise. An endpoint that would round to zero at three decimals is printed with the extra digit its sign needs, so no non-zero bound reads as zero.

Cell	$S_{\text{conf}} - S_{\text{judge}}$	$S_{\text{sc}} - S_{\text{judge}}$	$S_{\text{conf}} - S_{\text{sc}}$
Qwen2.5-VL V-MME	+0.126 [+0.086, +0.167]	+0.117 [+0.076, +0.158]	+0.007 [−0.009, +0.022]
Qwen2.5-VL MVB	+0.206 [+0.164, +0.249]	+0.170 [+0.126, +0.213]	+0.036 [+0.018, +0.054]
Qwen2.5-VL NExT	+0.142 [+0.088, +0.197]	+0.112 [+0.056, +0.168]	+0.030 [+0.008, +0.054]
Qwen2.5-VL VHall	+0.042 [−0.006, +0.090]	+0.006 [−0.040, +0.055]	+0.035 [+0.011, +0.061]
LLaVA-OV V-MME	+0.050 [+0.001, +0.099]	+0.074 [+0.027, +0.120]	−0.022 [−0.051, +0.006]
LLaVA-OV MVB	+0.170 [+0.122, +0.219]	+0.123 [+0.075, +0.172]	+0.048 [+0.021, +0.074]
LLaVA-OV NExT	+0.088 [+0.033, +0.144]	+0.108 [+0.055, +0.163]	−0.020 [−0.046, +0.007]
LLaVA-OV VHall	+0.043 [+0.003, +0.085]	+0.012 [−0.030, +0.056]	+0.032 [+0.004, +0.058]

Table 5: Per-cell AUPRC for each signal family, the judge coverage share, and leave-one-family-out marginal AUROC contributions of the all-family logistic combination. AUPRC collapses each tie block into one threshold (the average-precision convention), which these heavily tied columns require: walking tied items one at a time makes the value depend on their arbitrary input order. Judge coverage is a union over the two decoding budgets of Appendix A. The best two-family combination per cell is in Table 6.

	Qwen2.5-VL-7B				LLaVA-OneVision-7B			
	V-MME	MVB	NExT	VHall	V-MME	MVB	NExT	VHall
AUPRC								
S_{conf}	0.731	0.821	0.897	0.778	0.679	0.769	0.873	0.797
S_{sc}	0.708	0.773	0.875	0.750	0.685	0.686	0.873	0.755
S_{judge}	0.589	0.633	0.813	0.743	0.608	0.587	0.816	0.728
S_{frame}	0.575	0.639	0.807	0.696	0.600	0.572	0.806	0.668
S_{temp}	–	0.587	0.760	0.683	–	0.548	0.762	0.657
Judge coverage (%)	97.4	99.8	100.0	100.0	98.2	99.8	100.0	100.0
Leave-one-family-out marginal AUROC contribution								
S_{conf}	+0.004	+0.017	+0.004	+0.011	+0.002	+0.026	+0.003	+0.008
S_{sc}	+0.004	−0.001	+0.001	−0.001	+0.014	−0.002	+0.019	+0.000
S_{judge}	+0.006	−0.003	+0.031	+0.041	+0.012	+0.000	+0.028	+0.044
S_{frame}	−0.002	+0.004	+0.016	−0.003	+0.010	−0.002	+0.006	+0.000
S_{temp}	+0.000	+0.003	+0.001	−0.006	+0.000	−0.001	−0.002	−0.002

per-cell differences carry wide uncertainty: a question-level bootstrap of the primary-minus-cross-family AUROC difference (Table 10, 20,000 resamples) puts 0 of 8 per-cell intervals clear of zero, so the table reports plain values with no winner marking and these means bound the family effect rather than exclude it. Second, switching families shows no sign of rescuing judge-based gating: the cross-family judge is the weaker scorer in 7 of 8 cells, a sign count and not a per-cell significance claim. The control judge is also smaller than the primary (8B vs. 32B), so family and capacity are confounded in the second statement; the conclusion the paper relies on – we find no evidence that S_{judge}'s weakness is an artifact of judging in-family – needs only the first.

E Per-candidate judge control

The deployed S_{judge} scores the greedy answer once per question and broadcasts that score to the question's 9 candidates. It therefore differs from a per-candidate judge in two ways at once: the score is constant inside a question, and – the substantive one – it is a judgement of a *different object*, the greedy answer rather than the candidate being ranked. Whether the free families beat judging as such, or only beat this deployment of it, the matrix alone cannot say. Scoring every candidate separately settles it at 9 times the cost; the control runs on a stratified subset of every cell.

Table 6: Best two-family combination per cell and its AUROC (group-aware 5-fold cross-validation), the data behind the best-pair claim of Section 4.2. S_{judge} enters the best pair in 6 of 8 cells; the gain over the best single family never exceeds +0.048 AUROC. The pair is selected by the same cross-validation estimate this column reports, so the AUROC is a maximum over 10 correlated estimates and is upward-biased; in 4 of 8 cells it exceeds the all-family combination, which a feature superset cannot be beaten by in population and which therefore bounds the selection noise.

Cell	Best pair	AUROC
Qwen2.5-VL V-MME	conf+judge	0.740
Qwen2.5-VL MVB	conf+temporal	0.776
Qwen2.5-VL NExT	conf+judge	0.822
Qwen2.5-VL VHall	conf+judge	0.721
LLaVA-OV V-MME	selfcon+judge	0.722
LLaVA-OV MVB	conf+temporal	0.723
LLaVA-OV NExT	selfcon+judge	0.825
LLaVA-OV VHall	conf+judge	0.737

Table 7: Certificate validity at $\alpha = 0.05$ (top) and $\alpha = 0.2$ (bottom): realized false-accept rate (%), seed-mean, randomized tie-breaking. Bold marks seed-means above the nominal bound – a descriptive flag; no randomized setting's excess is significant after Bonferroni correction (Section 4.3). The corresponding naive-mode numbers violate almost uniformly and are omitted for space; they are in the released result file.

	Qwen2.5-VL-7B				LLaVA-OneVision-7B			
Gate signal	V-MME	MVB	NExT	VHall	V-MME	MVB	NExT	VHall
$\alpha = 0.05$, *randomized*								
S_{conf}	**5.5**	**5.6**	**6.3**	**5.2**	4.8	**5.1**	**6.2**	4.6
S_{sc}	**5.2**	**5.6**	**5.3**	5.0	4.8	**5.3**	**5.3**	4.9
S_{judge}	4.9	5.0	**5.6**	5.1	4.8	4.8	**5.6**	5.1
S_{frame}	**5.2**	5.0	**5.3**	5.1	5.0	4.8	**5.3**	5.0
S_{temp}	–	5.1	5.0	5.1	–	**5.5**	5.1	**5.2**
TRON	**5.2**	**5.6**	**5.3**	5.0	4.8	**5.3**	**5.3**	4.9
All (Bonferroni)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\alpha = 0.2$, *randomized*								
S_{conf}	**20.3**	**20.8**	**21.6**	19.6	18.3	**20.2**	**21.7**	19.1
S_{sc}	19.7	**21.4**	**21.6**	19.4	19.2	**21.0**	**22.1**	18.9
S_{judge}	19.7	**20.1**	**21.6**	**20.9**	19.3	19.3	**22.4**	**20.5**
S_{frame}	**20.5**	**20.3**	**20.8**	20.0	**20.4**	19.3	**21.0**	19.9
S_{temp}	–	19.4	19.9	**20.6**	–	**20.2**	**20.7**	**20.5**
TRON	19.7	**21.4**	**21.6**	19.4	19.2	**21.0**	**22.1**	18.9
All (Bonferroni)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Protocol. Within each cell a deterministic seeded draw (nominal fraction 20%, realized 20.4–20.6% after rounding up per stratum; seed 20260716, a fixed-seed `random.Random` shuffle over sorted item ids, never Python's process-salted `hash()`, keeping at least one question per metadata stratum) selects the questions, and the identical primary judge (configuration key `qwen2.5-vl-32b-awq`, the bf16 checkpoint of Appendix A) rescores each of their 9 candidates under the unchanged prompt, frame budget, and decoding parameters: only *what* is scored changes. The draw uses the same machinery and seed as the cross-family control of Appendix D at twice its fraction; because a wider fraction at a fixed seed only adds questions, that control's subset is contained in this one, cell for cell. The same nesting is why the slate could be scored in 2 passes (10% then 20%) without rejudging anything: widening the draw re-used every question already scored and paid only for the increment. In total 818 questions were rescored in 7,362 judge calls costing 2.8 GPU-hours, of which 52 (0.71%) were unparseable and dropped – concentrated rather than spread, at up to 2.72% in one cell and exactly zero in 3 of 8. The 7,285 candidates carrying both a per-candidate and a broadcast judge score form the paired slate of Table 11, on which all four columns are computed.

Table 8: Coverage at $\alpha = 0.1$: share of test candidates accepted (%), seed-mean over 40 splits, randomized tie-breaking. S_{temp} is n/a on Video-MME.

	Qwen2.5-VL-7B				LLaVA-OneVision-7B			
Gate signal	V-MME	MVB	NExT	VHall	V-MME	MVB	NExT	VHall
S_{conf}	24.0	33.8	33.9	16.8	17.9	26.8	29.3	24.9
S_{sc}	21.4	24.3	24.4	13.9	17.5	16.6	24.8	20.1
S_{judge}	12.4	11.3	16.6	14.1	13.2	10.4	18.6	15.1
S_{frame}	11.8	11.4	14.0	10.8	12.7	10.0	14.4	11.0
S_{temp}	–	9.0	11.2	10.5	–	9.2	11.9	10.7
TRON	21.4	24.3	24.4	13.9	17.5	16.6	24.8	20.1
All (Bonferroni)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 9: **Same-family vs. cross-family judge control:** AUROC for greedy-answer correctness on the stratified paired subsample, per cell. The primary judge (Qwen2.5-VL-32B) is same-family for the Qwen backbone only; InternVL3-8B is out-of-family for both backbones. Both judges are scored on the identical pairwise-comparable item set of size n. No per-cell winner is marked: at these n the per-cell differences are within noise.

	Qwen2.5-VL-7B				LLaVA-OneVision-7B			
Judge	V-MME	MVB	NExT	VHall	V-MME	MVB	NExT	VHall
Qwen2.5-VL-32B (primary)	0.652	0.581	0.717	0.549	0.642	0.615	0.692	0.662
InternVL3-8B (cross-family)	0.533	0.595	0.660	0.505	0.539	0.558	0.629	0.568
Paired questions n	52	51	53	54	51	51	53	54

What it shows. Judging the candidate itself helps the judge and does not close the gap. Per-candidate scoring gains +0.021 AUROC over the broadcast score on average, positive in 8 of 8 cells and ranging +0.005 to +0.045 (Table 12: 0 of 8 per-cell intervals exclude zero, across-cell 95% interval [+0.010, +0.031], sign test $p = 0.008$). The gain recovers 23–25% of the broadcast judge's distance to the free families in the cell mean (0.644 to 0.665, against 0.728 for S_{conf} and 0.736 for S_{sc}), and the per-candidate judge outscores S_{conf} in 2 of 8 cells and S_{sc} in 2 – and not the same cell, so both free families lead it in 6 of 8. The class-level reading of Section 4.1 – free generation-time signals outrank the paid judge – survives the repair that would most plausibly have overturned it.

Doubling the subset. The first pass drew 10% of each cell, and its per-cell estimates did not agree on the sign of the effect. The draw was therefore widened to 20% at the same seed – nested, so it added questions without invalidating any already judged – and the identical estimator was recomputed on both slates from the same artifacts. On the first pass's 3,755 candidates the gain was +0.026 AUROC, positive in 7 of 8 cells and ranging -0.016 to +0.073; on the 7,285 of the full slate it is +0.021, positive in 8 of 8 and ranging +0.005 to +0.045. The spread across cells more than halved and the cell that had been negative turned positive, which is what a subset-noise diagnosis predicts and why the scope note below attaches the noise warning to the free-family columns rather than to the contrast this table is for. The point estimate fell slightly, and the class-level reading did not move: it now rests on twice the evidence, and on a sign that no longer depends on which cell one looks at.

Where the gain comes from. Not from score granularity. AUROC averages over (correct, incorrect) candidate pairs, and those pairs split exactly into pairs inside one question and pairs across questions, so the gain splits with them. Only 0.35% of the slate's pairs are within-question – a question's candidates are almost always unanimously right or unanimously wrong – and the broadcast score, being constant there, scores 0.5 on them by definition. Of the +0.021 mean gain, +0.0002 comes from that within-question channel and +0.0205 (99%) from re-ordering questions against each other. The control therefore measures what judging the right object is worth, not what per-candidate resolution is worth: on this slate the tightest per-question ceiling in any cell is 0.9975 AUROC, and on the full matrix, where the pool is 500 questions per cell and within-question pairs

Table 10: Primary minus cross-family judge AUROC per cell with 95% question-level bootstrap percentile intervals (20,000 resamples) on the paired subsample of Table 9. 0 of 8 intervals exclude zero: at these subsample sizes the control bounds the family effect rather than isolating it.

Cell	n	Primary − cross-family
Qwen2.5-VL V-MME	52	+0.119 [−0.055, +0.289]
Qwen2.5-VL MVB	51	−0.014 [−0.159, +0.125]
Qwen2.5-VL NExT	53	+0.058 [−0.164, +0.276]
Qwen2.5-VL VHall	54	+0.045 [−0.099, +0.191]
LLaVA-OV V-MME	51	+0.102 [−0.082, +0.283]
LLaVA-OV MVB	51	+0.057 [−0.106, +0.213]
LLaVA-OV NExT	53	+0.063 [−0.069, +0.212]
LLaVA-OV VHall	54	+0.093 [−0.037, +0.231]

Table 11: **Per-candidate judge control** on the nominal 20% stratified subset (seed 20260716): AUROC for candidate correctness of the same 32B judge run per candidate and, as deployed, per question and broadcast, with the two free families, all on the identical paired slate of n candidates. Δ is per-candidate minus broadcast computed before rounding, so it can differ by one in the last digit from the printed columns; the last row averages the eight cells, except †, which totals them. The subset columns are not Table 1 re-estimated; see *Scope*. Intervals for Δ are in Table 12.

Cell	per-cand.	broadcast	Δ	S_{conf}	S_{sc}	n
Qwen2.5-VL V-MME	0.641	0.616	+0.024	0.759	0.739	884
Qwen2.5-VL MVB	0.628	0.612	+0.016	0.810	0.773	914
Qwen2.5-VL NExT	0.684	0.679	+0.005	0.812	0.834	927
Qwen2.5-VL VHall	0.652	0.639	+0.013	0.612	0.613	914
LLaVA-OV V-MME	0.668	0.650	+0.018	0.670	0.722	885
LLaVA-OV MVB	0.627	0.613	+0.013	0.728	0.741	916
LLaVA-OV NExT	0.732	0.702	+0.030	0.771	0.794	927
LLaVA-OV VHall	0.688	0.643	+0.045	0.659	0.672	918
Mean over cells	0.665	0.644	+0.021	0.728	0.736	7,285†

scale only linearly with it, a per-question score is capped at most 0.0005 below 1. "The broadcast score cannot rank inside a question" is true and nearly free.

Scope. This is a nominal 20% stratified subset of each cell, one judge model, 8 cells, one run: it bounds what judging the candidate rather than the greedy answer is worth, and it is not a re-measurement of Table 1 at per-candidate granularity. Note also that the within-question share, and so the broadcast score's structural ceiling, is 4.6–5.6 times looser here than on the full matrix, because within-question pairs scale with the number of questions and total pairs with its square: the subset over-represents the very channel it is best placed to observe. At $n = 884$–927 candidates per cell the free-family columns of Table 11 carry subset sampling noise and differ from their full-sample values in Table 1 in both directions – S_{conf} on the Qwen2.5-VL VideoHallucer cell reads 0.612 here against 0.674 over all candidates, the largest downward deviation, while S_{sc} on the LLaVA-OneVision MVBench cell reads 0.741 against 0.677, the largest upward one. Those deviations do not push the judge's gap one way, but they are large enough to reorder cells: on this subset even the deployed broadcast judge falls below both free families in 7 of 8 cells rather than all 8 as in Table 1. The control is therefore read as the paired within-subset contrast between two judge granularities and not as a second estimate of the matrix.

F Second seed, decoding temperature, frame budget and model scale

Four checks registered after the main results (PREREG items B0/B1 and amendments 4 and 5), each run on the A800 with the shipped code. Logistic combination numbers in the main text carry three decimals, an order above their cross-environment reproduction drift. **Second seed.** The k temper-

Table 12: Per-candidate minus broadcast judge AUROC per cell with 95% question-level cluster-bootstrap percentile intervals (20,000 resamples) on the paired slate of Table 11, the same battery Table 10 applies to the cross-family control. 0 of 8 per-cell intervals exclude zero; across the 8 cell estimates the 95% t-interval is [+0.010, +0.031] and the sign test on 8 positive cells gives two-sided $p = 0.008$. The effect is real in the aggregate and per-cell-inconclusive.

Cell	n	Per-candidate − broadcast
Qwen2.5-VL V-MME	884	+0.024 [−0.023, +0.075]
Qwen2.5-VL MVB	914	+0.016 [−0.033, +0.065]
Qwen2.5-VL NExT	927	+0.005 [−0.070, +0.081]
Qwen2.5-VL VHall	914	+0.013 [−0.030, +0.059]
LLaVA-OV V-MME	885	+0.018 [−0.030, +0.067]
LLaVA-OV MVB	916	+0.013 [−0.029, +0.055]
LLaVA-OV NExT	927	+0.030 [−0.033, +0.092]
LLaVA-OV VHall	918	+0.045 [−0.017, +0.109]

ature candidates were re-drawn with sampling seed 20260913 (greedy candidates unchanged) and RQ1 recomputed: 0 of 16 interval verdicts change; 12 deltas are positive; the largest, 0.040 on S_{sc} at Qwen2.5-VL VHall, exceeds that cell's half-width 0.035; of 96 gate rows, 4 change the sign of their excess over the bound, all within ± 0.0025. Table 15 lists the pairs. **Temperature.** Regenerating the four cheapest cells at $T{=}1.0$ (fresh draw, seed 20260913) lowers S_{sc}'s largest tie share and the naive gate's excess at $\alpha{=}0.1$ together in every cell (Table 16). **Frame budget.** Scoring the shipped Qwen2.5-VL candidates with 16 frames: NExT-QA S_{frame} AUROC 0.653→0.701 ([0.615, 0.694] / [0.662, 0.742]), MVBench 0.580→0.620 ([0.546, 0.614] / [0.586, 0.655]); 1486 of 1500 MVBench subsets reached 16 frames (shortest clip 7). **Scale.** The two free signals were re-scored on Qwen2.5-VL at 3B and 32B (AWQ), candidates generated under the protocol of Section 3 on the same 500 questions per benchmark, with the 7B pool regenerated beside them so that the three points share seeds; candidate accuracy runs 48.0–53.6% on Video-MME and 66.6–74.1% on NExT-QA. Table 13 gives the 8 three-point curves. All 24 intervals sit above chance; 4 of the 8 curves are strictly monotone in scale, every one of them falling, and the rest are not ordered; the largest 3B–32B gap is 0.054 AUROC (S_{sc} on MVB), of the order of one interval half-width. Discrimination by a free signal is therefore a property of the 3B–32B range, not of the 7B point, and it does not improve with the backbone. The registered third test of RQ3 reads the same pools through the gate at $\alpha{=}0.1$ over the 40 split seeds (Table 14): the naive gate exceeds α in 15 of 16 (pool, signal) entries, by +6.6pp on average and up to +27.4pp, with 5 entries above their nominal bound on every seed; the randomized gate's excess is +0.5pp on average, at most +1.5pp, and its seed-mean sits above α in 12 entries, with 15–25 of 40 seeds above their nominal bound, the share an exactly valid gate produces. The 7B finding replicates at both scales in both directions: the port breaks, the repair holds.

Table 13: Free-signal AUROC by backbone scale (Qwen2.5-VL 3B / 7B / 32B-AWQ), with question-level bootstrap half-widths.

Benchmark	Signal	3B	7B	32B
MVB	S_{conf}	0.761± 0.031	0.776± 0.030	0.745± 0.031
MVB	S_{sc}	0.761± 0.031	0.740± 0.031	0.707± 0.031
NExT	S_{conf}	0.813± 0.032	0.804± 0.038	0.797± 0.034
NExT	S_{sc}	0.779± 0.033	0.773± 0.038	0.783± 0.035
V-MME	S_{conf}	0.744± 0.031	0.736± 0.032	0.732± 0.032
V-MME	S_{sc}	0.716± 0.031	0.729± 0.032	0.710± 0.031
VHall	S_{conf}	0.704± 0.036	0.674± 0.039	0.672± 0.037
VHall	S_{sc}	0.710± 0.037	0.638± 0.035	0.658± 0.036

G CERTIFYING PER QUESTION

The main-text certificate pools all wrong candidates of the calibration questions, so its effective calibration size is the number of wrong candidates (median $n^{-} = 818.1$) while the exchangeable unit is the question. The per-question certificate draws one anchor candidate per calibration question

Table 14: Realized risk excess over α=0.1 (pp) of the naive and randomized gates on the 3B and 32B pools, seed-mean over 40 split seeds; in parentheses, the number of seeds whose realized risk exceeds its nominal bound.

		3B		32B	
Benchmark	Signal	naive	randomized	naive	randomized
MVB	S_{conf}	+0.7(20)	+0.7(20)	+1.2(26)	−0.3(15)
MVB	S_{sc}	+3.0(37)	+0.5(23)	+14.0(40)	−0.5(15)
NExT	S_{conf}	+0.9(22)	+0.7(22)	+5.6(39)	+1.1(23)
NExT	S_{sc}	+8.8(40)	+0.9(18)	+14.9(40)	+0.3(20)
V-MME	S_{conf}	+1.1(25)	+1.0(25)	+3.1(34)	+1.5(23)
V-MME	S_{sc}	+3.6(33)	+1.4(25)	+15.9(40)	+0.8(20)
VHall	S_{conf}	−0.2(17)	−0.3(17)	+0.5(20)	+0.4(20)
VHall	S_{sc}	+5.4(26)	−1.0(17)	+27.4(40)	+0.1(20)

Table 15: Second-draw AUROC spread per (cell, signal).

Cell	Signal	AUROC (published draw)	AUROC (second draw)	Δ	CI half-width
LLaVA-OV MVB	S_{conf}	0.725	0.724	−0.001	0.033
LLaVA-OV MVB	S_{sc}	0.677	0.683	+0.006	0.032
LLaVA-OV NExT	S_{conf}	0.760	0.765	+0.005	0.038
LLaVA-OV NExT	S_{sc}	0.780	0.803	+0.023	0.037
LLaVA-OV V-MME	S_{conf}	0.677	0.689	+0.011	0.034
LLaVA-OV V-MME	S_{sc}	0.700	0.721	+0.021	0.033
LLaVA-OV VHall	S_{conf}	0.697	0.705	+0.009	0.037
LLaVA-OV VHall	S_{sc}	0.665	0.685	+0.020	0.036
Qwen2.5-VL MVB	S_{conf}	0.776	0.771	−0.005	0.030
Qwen2.5-VL MVB	S_{sc}	0.740	0.755	+0.015	0.031
Qwen2.5-VL NExT	S_{conf}	0.804	0.806	+0.002	0.038
Qwen2.5-VL NExT	S_{sc}	0.773	0.782	+0.008	0.038
Qwen2.5-VL V-MME	S_{conf}	0.736	0.744	+0.009	0.032
Qwen2.5-VL V-MME	S_{sc}	0.729	0.727	−0.001	0.032
Qwen2.5-VL VHall	S_{conf}	0.674	0.672	−0.001	0.039
Qwen2.5-VL VHall	S_{sc}	0.638	0.679	+0.040	0.035

by a deterministic hash of (anchor, seed, question id), which reads neither label nor score, and takes the same order statistic over the wrong anchors (median $n_q^- = 90.9$, from 250 calibration questions per cell); the test side is unchanged and realized risk is still the false-accept share over all test wrong candidates. Over the 228 (cell, signal, α, mode) settings, TRON's set aside as in Section 4.3, the bound widens by 0.74–1.36pp and the realized seed-mean risk moves by at most 2.61pp, so 57 verdicts change, every one from violating to holding: naive violations go from 109 to 101 of 114 single-signal settings, randomized from 59 to 10 of 114. The per-setting grid under both units ships as `gate_unit_readout.json`; every macro in the main text is computed from the candidate-unit rows.

H Out-of-fold paired comparison

The combination gains of Section 4.2 are read from the same five grouped folds as the combination itself (`combine.py`, 500 question groups, cv seed 20260716): the best single signal is scored on the held-out fold with the same 0.5 imputation the fusion feature matrix uses, so both sides see the same candidates, and the statistic is the difference of fold-mean AUROCs. The interval is a question-level paired bootstrap of 2,000 resamples that moves whole questions with folds fixed. The same procedure gives the leave-S_{judge}-out cost. Per cell (gain over best single; judge left out): Qwen Video-MME +0.005 [−0.010, +0.020]; +0.006 [−0.004, +0.017]. Qwen MVBench −0.006 [−0.021, +0.008]; −0.003 [−0.013, +0.007]. Qwen NExT-QA +0.031 [+0.001, +0.063]; +0.031 [+0.001, +0.061]. Qwen VideoHallucer +0.034 [−0.004, +0.072]; +0.041 [+0.008, +0.076]. LLaVA-OV Video-MME +0.035 [+0.010, +0.059]; +0.012 [−0.008, +0.031]. LLaVA-OV MVBench −0.008 [−0.029, +0.014]; −0.0003 [−0.021, +0.020]. LLaVA-OV NExT-QA +0.053 [+0.026, +0.080]; +0.028 [+0.005, +0.052]. LLaVA-OV VideoHallucer +0.031 [+0.004,

Table 16: Tie share and naive excess at T=0.7 and 1.0 (S_{sc}, α=0.1).

	Tie mass (%)		Naive-gate excess (pp)	
Cell	T=0.7	T=1.0	T=0.7	T=1.0
LLaVA-OV MVB	43.6	34.6	+17.5	+6.2
LLaVA-OV NExT	63.8	49.8	+17.7	+5.7
Qwen2.5-VL MVB	43.6	33.8	+9.9	+4.9
Qwen2.5-VL NExT	65.8	51.2	+18.1	+4.2

+0.054]; +0.044 [+0.022, +0.066]. Shuffling the fusion score within question removes the gain (negative control in `tests/test_oof_paired.py`); the read-out ships as `oof_paired.json`.

I RELEASE SCHEMA

Each row of the released JSONL has eleven fields: `model`, `dataset`, the benchmark's own `question_id`, `candidate_idx`, the binary correctness label `correct`, and the signal values `s_conf`, `s_selfcon`, `s_judge`, `s_frameev`, `s_temporal` and `s_tron`. Rows where a signal is undefined carry an explicit null rather than an imputed value: S_{judge} where the judge output was unparseable, and S_{temp} on Video-MME. The release contains 36,000 rows covering 8 model-dataset cells with 500 questions each. For review it is in the supplementary zip at `experiments/mmgating/results/benchmark_release.jsonl`, with `README_SUPPLEMENTARY.md` at the archive root, alongside the harness and the CPU-only statistics stage. The zip also holds the per-stage dumps under `results/raw*/`, which carry the benchmark question text and each candidate's decoded answer so that the scoring stages can be re-run. A permanent public link replaces the zip on publication. The archive is built by a script that redacts author, host and project identifiers, replaces absolute data roots with `$DATA_ROOT`, and then re-scans the staged tree, refusing to emit an archive in which any identifier survives – so anonymity of the code package is a checked property, not an intention. Porting-provenance comments are kept; only the identifiers in them are redacted.

J REPRODUCIBILITY AND COMPUTE

All pools are seed-fixed subsets of public benchmark splits; candidate generation, signal scoring, gate calibration, and aggregation run from one configuration file with pinned model revisions, and each stage writes append-only JSONL checkpoints. The gate stage was first run at 5 split seeds and re-run at 40 to separate a small single-signal excess from seed noise; the decision rule was not changed, the 40 seeds include the original 5, and the naive-mode failure became more significant, not less. Gate splits use seed-derived deterministic randomness, including a hash-based tie-break draw, so every number in this paper is exactly reproducible from the released artifacts. A dedicated CPU-only statistics stage recomputes all inferential quantities – the seed-level one-sided tests with Bonferroni correction and the question-level bootstrap intervals – deterministically from the released result file and candidate pool under a fixed seed, and the macro generator consumes only its output, so no inferential number is hand-derived either. The GPU ledger records 25.7 GPU-hours for the main matrix and its judge controls, on a single 80GB accelerator per stage against a 56-hour budget, of which the judge stage is the largest share (11.51 GPU-hours); the runs of Appendix F are not in the ledger. The aggregate layer is pure CPU post-processing and regenerates all derived statistics, including every macro in this paper, in minutes.

K PRE-REGISTRATION OF THE MECHANISM ANALYSES

The three analyses of Section 5 were registered in a file dated 2026-09-02 that names the sha256 of every input and states that none of the analyses had been run. It entered the repository on 2026-09-03 in the same commit as the analysis script and its output, so the commit history does not by itself order registration and analysis. A later registration file replaced it in the repository on 2026-09-13; the original ships in the supplementary archive as `PREREG_H1H3_20260902.md`. The registration

fixes the strata, the estimator, the predictions, the competing predictions and the decision rules; they are restated below so that the outcomes can be read against what was claimed in advance.

H1, the temporal family. S_{temp} is signed by the benchmark's temporal flag (Section 3.2). The strata are fixed by benchmark metadata and are identical for both backbones: in NExT-QA, the 174 questions whose type label begins with T against the other 326; in VideoHallucer, the 83 questions of the temporal category against the other 417. MVBench is entirely flagged and Video-MME publishes no flag, so neither admits a within-dataset contrast. The paper's own explanation, that perturbation instability measures model fragility rather than answer correctness, predicts both strata at chance, that is, an interval containing 0.5. The competing prediction, from the literature that reads perturbation probes as revealing genuine fragility, is that the flagged stratum lies strictly above chance and the flagged-minus-non-flagged difference excludes zero. Registered decision rule: the first is supported if no flagged interval lies entirely below 0.5, the second if at least one lies entirely above 0.5 and its difference excludes zero, and any other outcome is reported as it falls. Section 5 applies this rule.

H2, the frame-evidence family. The strata are the Video-MME duration labels, short, medium and long. The paper's explanation, that a fixed 8-frame budget grows sparser as clips lengthen, predicts AUROC decreasing with duration. Registered decision rule: supported if the ordering holds in both backbones and at least one short-minus-long interval excludes zero; falsified if the ordering reverses with an interval excluding zero; otherwise reported as an unresolved ordering.

H3, calibration size. The gate is re-run at $n_{\text{cal}} \in \{50, 100, 250, 400\}$ question groups with the complement as the test set, 40 seeds, both modes and all three targets. Because the test half shrinks as the calibration half grows, the realized-rate range mixes calibration noise with evaluation noise, so the calibration side is measured separately as the 40-seed range of the wrong-side threshold, which does not depend on the test split. Registered prediction: that threshold range shrinks as $1/\sqrt{n_{\text{cal}}}$, with a log-log slope in $[-0.75, -0.25]$. Competing prediction: discreteness holds it up, so the slope stays above -0.25 for the coarsely quantized signals. Extrapolated calibration sizes, had we reported any, were to be marked as extrapolations rather than measurements.

Estimator, fixed in advance for all three. The tie-corrected Mann-Whitney AUROC already used for the paired battery; question-level cluster bootstrap with 20,000 replicates and a fixed seed; 95% percentile intervals; stratum differences resampling the two strata independently within a replicate; strata reported with their n and never merged or dropped. The sweep's own correctness is checked by re-running it at the paper's split and comparing every gate row against the committed result file: 12,960 rows agree exactly.

What was not registered. Undoing the flag signing and scoring raw probe agreement (the green points of Figure 5a) was chosen after seeing the H1 outcome. It is reported as exploratory, and it changes no claim: the repaired score reaches 0.513–0.625 AUROC, still below the free families' 0.638 floor.

Table 17: Single-family AUROC with a 95% question-level cluster bootstrap interval, per cell. These are the intervals behind the chance-level statements of Section 4.1; the point estimates are those of Table 1. S_{temp} is not applicable to Video-MME.

Signal	V-MME	MVB	NExT	VHall
Qwen2.5-VL-7B				
S_{conf}	0.736 [0.702, 0.768]	0.776 [0.744, 0.807]	0.804 [0.766, 0.841]	0.674 [0.636, 0.713]
S_{sc}	0.729 [0.696, 0.761]	0.740 [0.707, 0.773]	0.773 [0.736, 0.812]	0.638 [0.604, 0.674]
S_{judge}	0.607 [0.571, 0.642]	0.570 [0.535, 0.605]	0.662 [0.618, 0.703]	0.632 [0.595, 0.669]
S_{frame}	0.593 [0.561, 0.626]	0.580 [0.547, 0.614]	0.653 [0.615, 0.692]	0.553 [0.526, 0.582]
S_{temp}	–	0.468 [0.440, 0.495]	0.526 [0.484, 0.569]	0.524 [0.492, 0.556]
LLaVA-OneVision-7B				
S_{conf}	0.677 [0.642, 0.712]	0.725 [0.690, 0.758]	0.760 [0.719, 0.801]	0.697 [0.659, 0.734]
S_{sc}	0.700 [0.666, 0.733]	0.677 [0.643, 0.711]	0.780 [0.743, 0.819]	0.665 [0.630, 0.702]
S_{judge}	0.624 [0.585, 0.662]	0.556 [0.520, 0.591]	0.673 [0.632, 0.712]	0.653 [0.623, 0.683]
S_{frame}	0.620 [0.585, 0.654]	0.525 [0.494, 0.557]	0.662 [0.625, 0.700]	0.561 [0.539, 0.584]
S_{temp}	–	0.474 [0.448, 0.498]	0.548 [0.506, 0.590]	0.539 [0.515, 0.565]

Table 18: The pre-registered calibration sweep (H3). Threshold settings whose wrong-side threshold is identical across all 40 seeds, out of 138 single-signal settings (TRON's included), and the median 40-seed range of the excess over the nominal bound, over the 162 settings of each mode (TRON's and the composed gate's included). The test half is the complement of the calibration half.

n_{cal} (questions)	50	100	250	400
Test questions	450	400	250	100
Seed-invariant thresholds	67	85	102	117
Median excess range, naive (pp)	5.8	6.9	13.0	24.4
Median excess range, rand. (pp)	20.0	12.4	8.9	11.2

Table 19: The pre-registered strata (H1, H2) with 95% question-level cluster-bootstrap intervals; n is questions, and every candidate of a drawn question travels with it. MVBench is entirely temporal-flagged, so its whole-cell value in Table 17 is its flagged stratum.

		Qwen2.5-VL-7B	LLaVA-OneVision-7B
H1: S_{temp} by the benchmark's temporal flag			
NExT-QA	flagged ($n = 174$)	0.337 [0.277, 0.394]	0.412 [0.362, 0.457]
	other ($n = 326$)	0.601 [0.560, 0.645]	0.602 [0.564, 0.643]
	difference	−0.264 [−0.336, −0.193]	−0.190 [−0.254, −0.131]
VideoHallucer	flagged ($n = 83$)	0.477 [0.419, 0.530]	0.549 [0.497, 0.601]
	other ($n = 417$)	0.539 [0.517, 0.563]	0.521 [0.507, 0.537]
	difference	−0.062 [−0.125, −0.006]	+0.027 [−0.027, +0.081]
H2: S_{frame} by Video-MME clip duration			
Video-MME	short	0.592 [0.532, 0.651]	0.675 [0.612, 0.738]
	medium	0.611 [0.555, 0.666]	0.596 [0.535, 0.655]
	long	0.587 [0.536, 0.635]	0.590 [0.532, 0.646]
	short − long	+0.006 [−0.072, +0.082]	+0.086 [+0.002, +0.171]