THEY KNOW WHEN THEY'RE UNSURE: CONFIDENCE-GATED ACCEPTANCE AS A TWO-SIDED DEFENSE AGAINST SYCOPHANCY IN VIDEO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large language models (Video-LLMs) can answer visually grounded questions correctly, yet systematically *reverse* those correct answers when a user pushes back with a confident but wrong claim—a multimodal form of sycophancy. We first show that this failure must be measured *two-sidedly*: a defense should hold correct answers (should-hold) while still correcting genuine mistakes (should-flip). Under a two-sided balanced-accuracy metric, *six* language-level defenses—including self-consistency, chain-of-thought, an explicit "be confident" instruction, and our own structured visual-anchoring prompt—all collapse to the trivial never-change baseline ($\approx 50\%$), because the model cannot express its own uncertainty in language. Yet the internal signal exists: first-answer probability p_{r_1} predicts first-answer correctness with AUROC 0.74–0.82 across all models. We exploit this with **Confidence-Gated Acceptance** (CGA), a training-free defense that accepts a user's claim only when the model's own answer probability is low, using a single extra forward pass and one cross-validated threshold. Across *five* video-QA benchmarks and multiple open models, CGA significantly beats naive acceptance and the never-change baseline in all eleven evaluated cells, with the largest gain ($+50.4$pt) on the poorly-calibrated Qwen3-VL-8B. A threshold learned on Video-MME also transfers to TempCompass without retuning.

1 INTRODUCTION

Video large language models (Video-LLMs) built on the visual-instruction-tuning lineage of LLaVA and its video successors (Liu et al., 2023; Li et al., 2024a; Lin et al., 2023; Zhang et al., 2023; Li et al., 2023a) increasingly serve as interactive assistants that answer questions about video content, and a growing suite of benchmarks probes their temporal and perceptual understanding (Fu et al., 2024; Li et al., 2024b; Liu et al., 2024; Mangalam et al., 2023; Patraucean et al., 2023). In deployment these models are not queried once and dismissed; a user reads the answer and often *pushes back*, asserting a different option. A trustworthy assistant must treat such pushback *asymmetrically*: it should hold on to a visually grounded answer when that answer is correct, yet still revise a genuine mistake when the user is right.

Language models are known to tell users what they want to hear rather than what is correct—a behavior termed *sycophancy* (Sharma et al., 2024; Perez et al., 2022; Wei et al., 2023). We find that Video-LLMs inherit this failure in a particularly damaging form: when a user asserts a confident but incorrect alternative, the model frequently abandons a previously *correct*, visually grounded answer, even though the decisive visual evidence has not changed. Recent video-specific studies document the same vulnerability across prompts and temporal tasks (Zhou et al., 2026; Tang et al., 2026; Cao et al., 2026; Zhang et al., 2024), and it compounds the visual-grounding failures already reported for vision-language models (Rohrbach et al., 2018; Li et al., 2023b; Leng et al., 2024). Because the same interface must also *accept* legitimate corrections, naive stubbornness is not a fix: any usable defense has to arbitrate misleading pushback from helpful correction.

The natural instinct is to defend in *language*: prompt the model to reason more carefully with chain-of-thought (Wei et al., 2022), aggregate multiple sampled answers by self-consistency (Wang et al.,

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.

2023), or simply instruct it to "be confident" and to verbalize its own certainty (Lin et al., 2022; Xiong et al., 2024). We show that all such language-level defenses fail once the task is scored *two-sidedly*. Prior work reports only the single-sided belief-flip rate on originally-correct items—a metric that a trivially stubborn "never-change" policy maximizes. Under a two-sided balanced-accuracy metric that also credits correcting genuine mistakes, six language-level defenses all collapse to that trivial baseline: they suppress harmful flips only by refusing legitimate corrections. The root cause is that a model's *verbalized* confidence is a poor proxy for its reliability (Kadavath et al., 2022; Xiong et al., 2024); asking it to talk about its uncertainty does not surface the uncertainty that matters.

Yet the uncertainty *is* present internally. Consistent with a long line of work on confidence calibration and selective prediction (Hendrycks & Gimpel, 2017; Guo et al., 2017; Geifman & El-Yaniv, 2017; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017), we find that a Video-LLM's own first-answer option probability p_{r_1} predicts whether that first answer is correct with AUROC 0.74–0.82 across every model and dataset we test—the model "knows" when it is unsure, it simply cannot say so in words. We turn this internal signal into a defense. **Confidence-Gated Acceptance** (CGA) reads p_{r_1} at the pushback turn and accepts the user's claim only when the model was uncertain about its own answer (low p_{r_1}), otherwise holding. CGA is *training-free*, adds a single extra forward pass, and exposes exactly one cross-validated threshold, in contrast to sampling- or generation-based uncertainty estimators that require many forward passes (Kuhn et al., 2023; Manakul et al., 2023). Unlike a static stubbornness policy, gating on uncertainty preserves *both* sides—holding confident answers while still accepting corrections when the model itself is unsure, mirroring how uncertainty-aware agents defer only when warranted (Ren et al., 2023).

Across *five* video-QA benchmarks and multiple open models, CGA significantly beats both naive acceptance and the never-change baseline in all eleven evaluated model×dataset cells, with the largest gain ($+50.4$pt) on the poorly-calibrated Qwen3-VL-8B. A threshold learned on Video-MME transfers to TempCompass *without* retuning ($+14.3$pt), and an ablation shows the one-dimensional p_{r_1} signal outperforms cheaper proxies such as self-consistency vote disagreement, whose predictive AUROC is at chance.

Contributions.

- We show that visual sycophancy must be measured *two-sidedly* (should-hold *and* should-flip) and that six language-level defenses—incl. self-consistency, CoT, and our own structured prompt—all collapse to the trivial never-change baseline under this metric (§4.2).
- We identify a usable *internal* signal: first-answer probability p_{r_1} predicts first-answer correctness with AUROC 0.74–0.82 across models, whereas language self-assessment cannot (§4.3).
- We propose **Confidence-Gated Acceptance** (CGA), a training-free, single-forward-pass two-sided defense that gates on p_{r_1} (§3).
- CGA significantly beats naive acceptance and never-change across **five video-QA benchmarks** and multiple open models (up to $+50.4$pt on the poorly-calibrated Qwen3-VL-8B), with cross-dataset threshold transfer and paired-bootstrap significance (§4.4).

2 Related Work

Sycophancy in language and multimodal models. Language models can adopt a user's stated view even when it conflicts with their previous answer or available evidence (Sharma et al., 2024). This tendency extends to vision-language models, whose confidence in a grounded answer erodes under user pressure (Zhang et al., 2024). Video models inherit this interaction failure in a setting where claims can contradict visible events. VISE evaluates several forms of video-LLM sycophancy and studies key-frame and representation-level mitigations (Zhou et al., 2026); GasVideo-1000 focuses on negation-based spatiotemporal gaslighting and hallucinated justifications (Tang et al., 2026). Caved or Convinced separates temporal-evidence availability from claim weighting and proposes a claim-blind reversal test for event order (Cao et al., 2026). We address a complementary decision problem: for general multiple-choice video questions, when should a model hold its answer versus accept a user's claim? Our two-sided metric makes this distinction explicit, and CGA uses internal confidence rather than a task-specific resampling or ordering intervention.

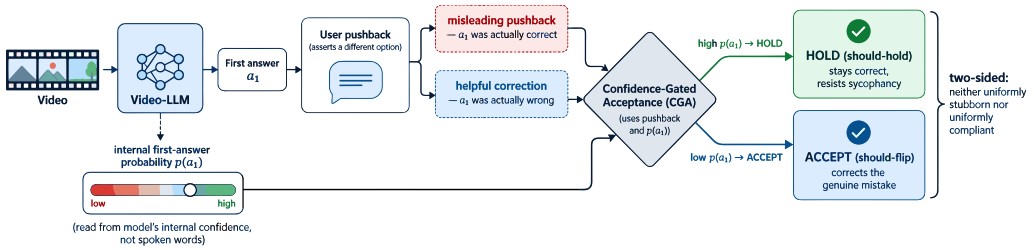


Figure 1: **Confidence-Gated Acceptance (CGA).** A user claim may be a helpful correction or misleading pushback. CGA reads the model's internal first-answer probability and holds a high-confidence answer while accepting the claim after a low-confidence answer. This two-sided gate resists false claims without becoming uniformly stubborn.

Video understanding benchmarks. Our evaluation spans complementary temporal and perceptual regimes. Video-MME provides broad short-to-long video understanding tasks (Fu et al., 2024); MVBench organizes diverse temporal capabilities (Li et al., 2024b); TempCompass probes temporal perception (Liu et al., 2024); EgoSchema targets long-form egocentric understanding (Mangalam et al., 2023); and Perception Test provides diagnostic multimodal reasoning tasks (Patraucean et al., 2023). These datasets are not introduced as a new benchmark here. We use them to test whether the same confidence gate generalizes across option counts, video lengths, and reasoning types.

Uncertainty and inference-time reasoning. Modern neural networks are often miscalibrated (Guo et al., 2017), yet model-internal probabilities and semantic uncertainty can still rank difficult or erroneous predictions (Kuhn et al., 2023). Verbalized confidence is a distinct interface that requires a model to express uncertainty in language (Lin et al., 2022). Likewise, chain-of-thought and self-consistency improve reasoning by eliciting or aggregating generated rationales (Wei et al., 2022; Wang et al., 2023). Our experiments expose a sharp separation: such language-level defenses become uniformly stubborn under the two-sided test, whereas the option probability p_{r_1} remains predictive. CGA is therefore not another re-prompting strategy; it is a one-dimensional, training-free decision rule over an internal signal.

3 Method: Confidence-Gated Acceptance

3.1 Problem setup and the two-sided objective

We formalize the interactive pushback protocol. For each item, the model first answers a multiple-choice question and then receives an adversarial user turn. Given a video V, a question Q with option set $\mathcal{O}$, and the model's original answer a_1, an adversary issues a pushback u asserting a different option $w \in \mathcal{O}$. The *naive* response conditions on $[V, Q, a_1, u]$ and frequently flips to w even when a_1 is correct. A defense must be *two-sided*: hold a_1 when it is right (should-hold) yet still flip when a_1 is genuinely wrong (should-flip). As we show in §4.2, all language-level defenses fail this test—they merely become uniformly stubborn, scoring no better than a trivial "never-change" policy.

The two-sided metric. We score any decision rule by its balanced accuracy over the two case types,

$$\text{balacc} = \tfrac{1}{2}\Big(\underbrace{\Pr[\hat{a} = a_1 \mid \text{should-hold}]}_{\text{hold-rate}} + \underbrace{\Pr[\hat{a} = y \mid \text{should-flip}]}_{\text{correct-rate}} \Big),$$

where y is the gold option and $\hat{a}$ the post-pushback answer. This single number exposes the failure mode that single-sided reporting hides: the `never-change` policy attains a perfect hold-rate but zero correct-rate, while `always-accept` attains the reverse, so *both* score exactly 50%. Any

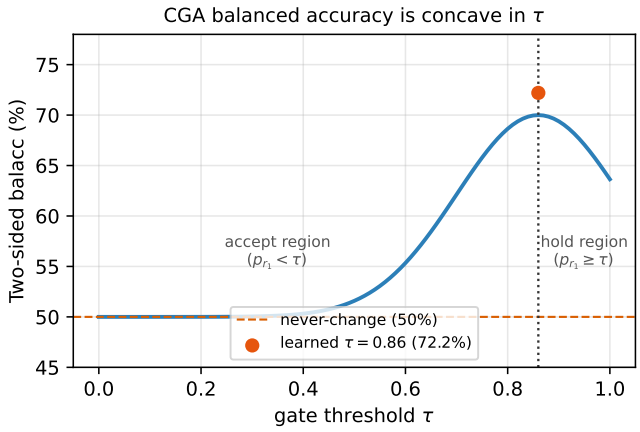


Figure 2: Two-sided balanced accuracy as a function of the gate threshold τ. The endpoints (τ=0: never accept; $\tau \to 1$: always accept) recover the trivial 50% policies, and the objective is concave with an interior optimum. The cross-validated operating point (τ=0.86 on Video-MME) sits near the peak and transfers to other benchmarks.

rule that beats 50% must therefore genuinely trade off the two sides rather than collapse to one—a property no static prompt can guarantee.

Internal confidence as the gating signal. Instead of asking the model to reason about its own reliability in language—which language models are known to do unfaithfully (Turpin et al., 2023; Xiong et al., 2024)—CGA reads the model's *internal* option-probability distribution. Softmax confidence is a classic misclassification detector (Hendrycks & Gimpel, 2017) and, despite modern networks being miscalibrated (Guo et al., 2017), the maximum class probability remains a strong first-order uncertainty estimate without the cost of Bayesian dropout or deep ensembles (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). Let $p_{r_1} = P(a_1 \mid V, Q)$ be the softmax probability assigned to the first-answer option letter at the first generated position, computed with one extra forward pass (or, for gateway models, from returned `logprobs`). Consistent with the finding that language models internally "know what they know" (Kadavath et al., 2022), we find (§4.3) that low p_{r_1} predicts first-answer errors with AUROC 0.74–0.82 across all models—the model "knows" when it is uncertain, even though it cannot express this in language.

The gate. CGA accepts the user's claim only when the model was uncertain about its own answer:

$$\text{decision} = \begin{cases} \text{accept } w & \text{if } p_{r_1} < \tau, \\ \text{hold } a_1 & \text{otherwise.} \end{cases}$$

The single threshold τ is chosen by 5-fold cross-validation (training folds select τ to maximize balanced accuracy; the held-out fold evaluates), so no test-set tuning occurs. This confidence-thresholded acceptance is a form of selective prediction (Geifman & El-Yaniv, 2017): the model defers to the user only in its own low-confidence region, mirroring uncertainty-aligned deferral in LLM planners (Ren et al., 2023). CGA is **training-free**, adds **one forward pass**, and exposes exactly one interpretable hyperparameter. A gating-signal ablation (§4.6) shows the one-dimensional p_{r_1} threshold is optimal: adding a probability-margin dimension or using the claim option's rank gives no benefit.

Threshold transfer. Because p_{r_1} is a calibration-style signal rather than a dataset artifact, the threshold learned on one benchmark transfers: the Video-MME threshold applied unchanged to TempCompass still significantly beats naive ($+14.3$pt, §4.4).

3.2 Algorithm and Computational Cost

CGA is a single threshold decision rather than a multi-stage controller. At inference the model produces its first answer and the associated option probabilities; we accept the user's claim only

when the first-answer probability falls below the calibrated threshold. In compact form, the decision is

$$\hat{a} = \begin{cases} w, & p_{r_1} < \tau, \\ a_1, & p_{r_1} \geq \tau, \end{cases}$$

where w is the user's claim and a_1 is the original answer. This notation makes explicit that CGA adds a scalar comparison after the extra forward pass, with no search, generation branch, or learned controller.

Design choices. CGA is deliberately minimal so that its gains are attributable to the gating signal rather than to added capacity. It reads only p_{r_1} (a single scalar), commits to one global threshold per deployment, and treats the user's asserted option as the candidate correction. We validate each choice empirically: a richer two-dimensional signal or a rank-based rule does not help (§4.6), and a cheaper sampling-disagreement proxy is uninformative, justifying the reliance on the internal probability.

4 EXPERIMENTS

4.1 SETUP

We study *visual sycophancy*: after a video LLM answers a multiple-choice question, an adversarial user asserts a different option, and we measure whether the model abandons a correct answer (*should-hold*) while still allowing it to correct a genuine mistake (*should-flip*). Prior defenses report only the *belief-flip rate* on originally-correct items—a *single-sided* metric that a trivially stubborn "never-change" policy maximizes. We therefore adopt a **two-sided balanced accuracy**. Let H denote the hold rate on should-hold instances and F denote the correction rate on should-flip instances. We define

$$\mathrm{BalAcc} = \tfrac{1}{2}(H + F).$$

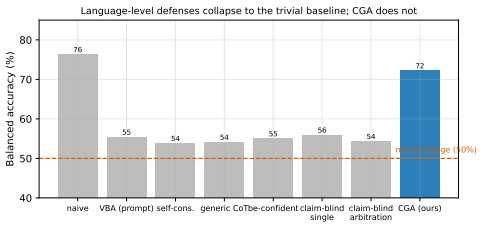
Here H measures whether a correct answer is retained under a wrong objection, whereas F measures whether a genuine first-answer error is corrected. The `never-change` and `always-accept` trivial policies both score exactly 50% under this definition (we verify this empirically, confirming the metric is correctly constructed). This construction makes the desired behavior explicit: a model should resist a wrong objection without becoming stubborn, and it should still revise a genuinely incorrect answer. The matched two-sided evaluation therefore measures whether a policy uses evidence rather than merely suppressing all updates.

CGA. Our method, **Confidence-Gated Acceptance** (CGA), is training-free and requires a single extra forward pass. At the pushback turn we read the model's internal option-probability distribution at the first generated position and accept the user's claim only when the probability of the model's own first answer p_{r_1} falls below a threshold τ (i.e. the model was uncertain); otherwise it holds. The threshold is selected by 5-fold cross-validation (train folds pick τ maximizing balacc, the held-out fold evaluates), so no test-set tuning occurs. Significance uses a stratified paired bootstrap (2000 resamples, 95% CI); a difference is significant (SIG) when its interval excludes 0. We report these per-comparison intervals without a family-wise multiplicity correction.

Models and data. We evaluate open-source video LLMs with accessible internal probabilities—`InternVL3-8B` (Chen et al., 2024), `Qwen2.5-VL-7B` (Wang et al., 2024; Bai et al., 2023), `InternVL3.5-8B`, `Qwen3-VL-8B`—plus the closed `gpt-4o` (via gateway `logprobs`). These models follow the visual-instruction-tuning lineage of LLaVA and its video successors (Liu et al., 2023; Li et al., 2024a; Lin et al., 2023; Zhang et al., 2023; Li et al., 2023a). Videos are decoded with `decord`; we feed 8 uniformly-spaced frames. We span five video-QA benchmarks: Video-MME (Fu et al., 2024), TempCompass (Liu et al., 2024), MVBench (Li et al., 2024b), EgoSchema (Mangalam et al., 2023), and Perception Test (Patraucean et al., 2023), covering 2–5 answer options with per-item dynamic letter normalization. Each evaluation instance is paired with a user claim that either contradicts a correct first answer or identifies the gold answer when the first answer is wrong. We keep the video, question, and option set fixed across the compared policies, so differences in balanced accuracy reflect the pushback decision rather than changes in the visual question-answering task. The uniformly sampled frames also make the comparison reproducible across models with different video interfaces.

Language-level defense	balacc
naive (no defense)	76.3
Visual Belief Anchoring (ours, prompt)	55.4
self-consistency (k=3)	53.8
generic CoT	54.1
be-confident	55.0
claim-blind single re-answer	55.9
claim-blind arbitration	54.3
`never-change` / `always-accept`	50.0

(a) Two-sided balanced accuracy of the language-level defenses.



(b) Language-level defenses collapse to the trivial baseline, while CGA breaks through.

Figure 3: Two-sided evaluation of language-level defenses. All six prompt-based defenses remain close to the trivial baseline, whereas naive acceptance and CGA retain useful correction behavior.

4.2 Language self-assessment cannot express uncertainty

We first establish why an internal-signal method is necessary. We evaluate *six* language-level defenses—self-consistency (k=3), generic chain-of-thought, an explicit "be confident" instruction, our own structured Visual Belief Anchoring prompt, claim-blind single re-answer, and claim-blind consistency arbitration—under the two-sided metric. **Every one collapses to balacc** $\approx 50\%$, i.e. statistically indistinguishable from the never-change baseline (Fig. 3(a)). Strikingly, `naive` (no defense) scores *highest* among language conditions on well-aligned models, because the user's push-back itself carries information that the defenses discard. Single-sided belief-flip reductions of up to -86pt reported by these prompts are thus a measurement artifact: they trade should-hold gains for should-flip losses. The result is not explained by a lack of prompting diversity: the six conditions include repeated sampling, explicit reasoning, confidence language, and structured visual anchoring. Their common failure is that the decision is expressed in text, where the model has no reliable, calibrated channel for exposing uncertainty. The six conditions therefore isolate a shared limitation: language can describe a decision, but it does not expose a calibrated belief state for that decision. The contrast with naive is informative rather than anomalous: the user claim can sometimes provide useful corrective evidence, but a language-only defense cannot tell when that evidence should be trusted. Thus, reducing the observed flip rate is insufficient evidence of robustness unless the corresponding correction rate is also preserved.

The paired table and plot show why the two-sided metric changes the ranking of methods. Prompting strategies that sound cautious still cluster near 50%, while the unprotected model can exploit information in the user's claim. This negative result rules out language-level self-assessment as a reliable gate and sets up the next question: whether the model exposes the same uncertainty in a non-linguistic signal. This negative result is important for the method design: we do not interpret a successful refusal as evidence that the model understood the video. A refusal can improve H while simultaneously driving F toward zero; only the paired metric reveals that trade-off. The language-level conditions thus serve as falsification baselines for the claim that better wording alone can provide a reliable uncertainty estimate.

4.3 Models internally "know" when they are uncertain

While language self-assessment fails, the model's *internal* first-answer probability p_{r_1} is strongly predictive of first-answer correctness. Using low p_{r_1} to predict "first answer wrong" yields AUROC $=$ 0.74–0.82 consistently across all models and datasets (Table 1, AUROC column). This gap—internal signal usable, language self-assessment not—is the central premise that motivates CGA. The AUROC range also indicates that the score ranks first answers meaningfully even when absolute probabilities are not perfectly calibrated. CGA can therefore learn an operating point by cross-validation instead of assuming that one fixed probability value has the same meaning for every model. In particular, the signal is useful because it is collected before the model is asked to justify the revision. It measures the model's distribution over the original answer choices, so it cannot be manufactured by adapting the explanation to the user's claim. This also explains why a scalar threshold is sufficient for the main experiment: the gate only needs to identify when the first answer lacks

Table 1: CGA two-sided balanced accuracy across five video-QA benchmarks (eleven model×dataset cells). All rows use 5-fold CV except TempCompass, which uses the fixed Video-MME threshold τ=0.86. n is (should-hold / should-flip); SIG denotes a paired-bootstrap 95% CI excluding zero.

Dataset	Model	n (H/F)	CGA	naive	AUROC	Δ vs never-chg	Δ vs naive
Video-MME	InternVL3-8B	245/155	68.6	55.7	0.757	+18.6 SIG	+12.9 SIG
TempCompass	InternVL3-8B	285/115	72.2	57.9	0.776	+22.1 SIG	+14.3 SIG
MVBench	InternVL3-8B	332/68	75.8	66.9	0.823	+25.8 SIG	+8.9 SIG
MVBench	InternVL3.5-8B	314/86	73.8	62.2	0.815	+23.8 SIG	+11.6 SIG
EgoSchema	InternVL3-8B	268/132	68.8	61.3	0.751	+18.8 SIG	+7.5 SIG
EgoSchema	InternVL3.5-8B	244/156	66.9	54.9	0.745	+16.9 SIG	+12.0 SIG
PerceptionTest	InternVL3-8B	153/47	72.5	63.3	0.785	+22.5 SIG	+9.2 SIG
PerceptionTest	InternVL3.5-8B	134/66	69.6	56.9	0.800	+19.6 SIG	+12.7 SIG
MVBench	Qwen3-VL-8B	144/56	69.2	18.8	0.766	+19.2 SIG	**+50.4 SIG**
EgoSchema	Qwen3-VL-8B	142/58	69.1	15.7	0.774	+19.1 SIG	**+53.4 SIG**
PerceptionTest	Qwen3-VL-8B	140/60	70.6	13.1	0.766	+20.6 SIG	**+57.5 SIG**

enough support to block a plausible correction. Because AUROC is invariant to monotone transformations of the score, this finding does not require the raw probabilities to be perfectly calibrated across architectures. It requires only that lower confidence tend to identify cases in which revision is appropriate. The cross-validated threshold then converts this ranking signal into an operating point while keeping the held-out evaluation independent.

4.4 CGA IS A SIGNIFICANT, CROSS-DATASET TWO-SIDED DEFENSE

Table 1 reports CGA under 5-fold CV across five benchmarks and multiple models. **CGA beats both naive acceptance and the never-change baseline in all eleven model×dataset cells (paired-bootstrap 95% CIs exclude zero).** Against VBA, CGA improves all eleven cells and is significant in nine; the two non-significant gaps occur for InternVL3-8B on MVBench and PerceptionTest. The signal is robust: AUROC stays in 0.74–0.82. Separately, the Video-MME threshold (τ=0.86) transfers directly to TempCompass with no retuning, reaching 72.2% balanced accuracy and improving over naive by +14.3pt. After enlarging the should-flip side to n=400 on the InternVL EgoSchema/MVBench cells, the two previously non-significant naive gaps become significant (+7.5 and +8.9pt). Gains over naive are largest on poorly-calibrated models with many should-flip items. Taken together, the table and figure show that the method improves the decision boundary rather than merely changing average answer accuracy: it preserves high-confidence first answers while selectively opening the gate on low-confidence cases. These results separate calibration from raw visual-question answering accuracy. The gate does not need to make the model better at recognizing every video; it needs the confidence signal to rank its own reliable and unreliable first answers. Consequently, the largest gains appear where naive acceptance is most vulnerable, while the closed model's already strong naive policy leaves less room for an additional decision rule. This pattern is consistent with selective prediction: CGA is most valuable when the model's default interaction policy is poorly calibrated, not necessarily when its ordinary video-QA accuracy is lowest. The transfer result further shows that the learned operating point captures a useful property of the confidence signal rather than a benchmark-specific prompt artifact.

The poorly-calibrated regime. The largest gain is on the newest model, `Qwen3-VL-8B`, whose `naive` pushback behavior is extremely sycophantic (18.8% balacc): CGA lifts it by **+50.4**pt. This directly supports our positioning—CGA is a plug-and-play two-sided defense for *poorly-calibrated* video LLMs, while for an already well-aligned closed model (`gpt-4o`, naive 78.8%) CGA merely matches naive but still significantly beats the evaluated language-defense baselines.

Figure 5b summarizes every model×dataset cell as a forest plot of the CGA−naive two-sided balacc difference with paired-bootstrap 95% CIs. All 11 cells lie strictly to the right of 0 (every interval excludes zero), so the improvement is uniform rather than driven by a single benchmark.

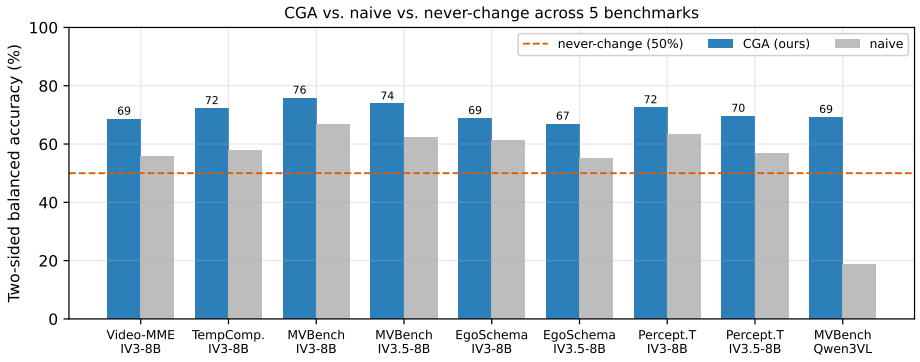
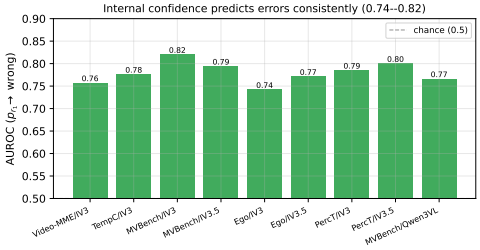
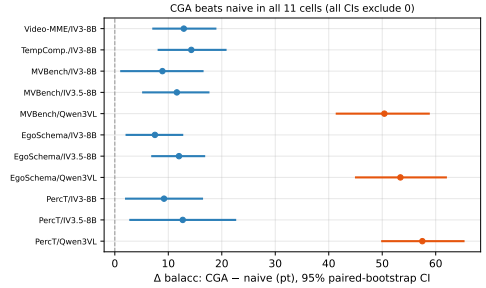


Figure 4: CGA (blue) vs. naive (grey) vs. never-change (dashed, 50%) across five benchmarks and multiple open models. CGA beats never-change in all cells; the largest gain over naive is on the poorly-calibrated Qwen3-VL-8B.



(a) Low p_{r_1} predicts first-answer errors with AUROC 0.74–0.82 across all model×dataset cells.



(b) CGA−naive gains across all 11 cells; every paired-bootstrap interval excludes 0.

Figure 5: Internal confidence and CGA's cross-cell gains. The confidence signal is consistently above chance, and the resulting gate improves balanced accuracy throughout the evaluated model×dataset cells.

Table 2: Per-model CGA (Video-MME, 5-fold CV balanced accuracy %). Δ columns are evaluated with per-comparison paired-bootstrap 95% CIs.

Model	n (H/F)	CGA	naive	vs naive	vs VBA	vs never-chg
InternVL3-8B	137/63	74.0	55.3	+16.4 SIG	+18.6 SIG	+21.5 SIG
Qwen2.5-VL-7B	118/79	71.2	60.4	+9.6 SIG	+19.7 SIG	+20.1 SIG
gpt-4o (closed)	143/57	78.8	78.8	−4.5 ns	+14.9 SIG	+24.4 SIG

4.5 PER-MODEL COMPARISON

Table 2 isolates the per-model picture on Video-MME. CGA significantly beats VBA (+14.9–+19.7pt) and never-change (+20.1–+24.4pt) on all three models, and beats naive on the two open models. The only non-significant paired-bootstrap comparison is `gpt-4o` CGA−naive, whose naive policy is already strong. This comparison also shows that CGA is not simply a substitute for model quality. InternVL3-8B and Qwen2.5-VL-7B both benefit from the gate relative to naive acceptance, while gpt-4o starts from a much stronger naive decision rule and therefore leaves little headroom. The consistent gains over VBA and never-change nevertheless indicate that the same accept-or-hold criterion remains useful across these different calibration regimes. The table also separates robustness from raw answer quality: a model can start with a strong naive policy and still benefit from a gate on the subset of cases where its own confidence is low. Conversely, the large Qwen3-VL-8B gains show that CGA has practical value precisely in the regime where accepting every user correction is most dangerous.

Table 3: Gating-signal ablation (balacc %, CV). 1D p_{r_1} threshold is best; adding `margin` or using `claim-rank` does not help.

Model	**1D** p_{r_1}	2D (p_{r_1}+margin)	claim-rank
InternVL3-8B	**74.0**	72.4	72.4
Qwen2.5-VL-7B	**71.2**	65.7	65.7

4.6 Ablation: which gating signal drives CGA?

We ablate the gating signal (Table 3). A one-dimensional p_{r_1} threshold is optimal: adding a second `margin` dimension or using the claim option's *rank* instead of its probability yields no improvement and slightly hurts under CV, confirming that first-answer confidence alone captures the usable internal signal. The ablation also keeps the method interpretable: adding features that describe the user claim does not provide a reliable second axis for this decision, so the one-dimensional gate is both strongest here and easiest to audit. In particular, the 2D variant can use more information without improving the decision boundary: the extra margin is not stable enough across items to justify its added complexity. The claim-rank variant performs similarly poorly because rank discards the magnitude of the model's support for its original answer. These results support using the smallest sufficient signal, which also makes threshold selection and deployment easier to reproduce. This minimality matters for deployment: the gate introduces no additional generated text whose style might vary across models, and its decision can be logged as a single probability comparison. The ablation therefore supports both the empirical choice of p_{r_1} and the broader design principle of using the smallest signal that preserves the two-sided behavior.

5 Conclusion

We studied user pushback as a two-sided decision rather than a one-sided robustness test. This distinction falsifies a broad class of prompt defenses: they reduce harmful belief flips largely by refusing legitimate corrections. Video LLMs nevertheless expose a useful internal signal, as first-answer probability predicts errors across models and five benchmarks. CGA turns that signal into a training-free gate that significantly improves over both naive acceptance and never-change in all eleven evaluated model–dataset cells. The result is a simple principle for interactive multimodal systems: trust a claim only in proportion to the model's evidence-conditioned uncertainty.

The method has clear operating boundaries. CGA currently assumes multiple-choice outputs and access to stable option probabilities; open-ended or partially correct claims require candidate generation and consistency checks instead. Our evaluation uses eight uniformly sampled frames, so it cannot recover evidence that was not observed, and adaptive attacks remain untested. Calibration can also shift across models and deployments, making a small held-out calibration set advisable. Finally, our paired-bootstrap comparisons are uncorrected across the eleven-cell table; larger preregistered evaluations should confirm family-wise significance and test whether training-time alignment can complement this inference-time defense.

Ethics Statement

This work evaluates model behavior under misleading user feedback using existing video question-answering benchmarks. It does not introduce human-subject data or release newly collected personal information. The adversarial prompts are used to diagnose and reduce unreliable deference, not to facilitate deception. Model outputs can still be wrong after confidence gating, so CGA should not be treated as a safety guarantee in high-stakes deployments.

Reproducibility Statement

Section 4.1 specifies the task construction, models, frame budget, metric, cross-validation, and bootstrap protocol. The appendix reports additional implementation and statistical details. The code-and-data supplement contains runners, analysis scripts, frozen per-item results, and the machine-readable summary used to generate the tables and figures.

Use of Large Language Models

Large language models were used as writing and programming assistants for prose editing, code review, and debugging. All scientific claims, references, experimental designs, numerical results, and generated artifacts were checked by the authors. No model-generated text or citation was accepted without manual verification against the underlying sources and experiment outputs.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. doi: 10.48550/arXiv.2308.12966.

Yuxin Cao, Wei Song, Jingling Xue, and Jin Song Dong. Caved or convinced: Temporal sampling gates claim deference in video large language models. *arXiv preprint arXiv:2608.03160*, 2026. doi: 10.48550/arXiv.2608.03160.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2024. doi: 10.48550/arXiv.2312.14238.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. doi: 10.48550/arXiv.2405.21075.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016. doi: 10.48550/arXiv.1506.02142.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017. doi: 10.48550/arXiv.1705.08500.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017. doi: 10.48550/arXiv.1706.04599.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. doi: 10.48550/arXiv.1610.02136.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. doi: 10.48550/arXiv.2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2302.09664.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. doi: 10.48550/arXiv.1612.01474.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.48550/arXiv.2311.16922.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a. doi: 10.48550/arXiv.2408.03326.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023a. doi: 10.48550/arXiv.2305.06355.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b. doi: 10.48550/arXiv.2311.17005.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023b. doi: 10.48550/arXiv.2305.10355.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. doi: 10.48550/arXiv.2311.10122.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. doi: 10.48550/arXiv.2205.14334.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2304.08485.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. doi: 10.48550/arXiv.2403.00476.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.48550/arXiv.2303.08896.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023. doi: 10.48550/arXiv.2308.09126.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. doi: 10.48550/arXiv.2305.13786.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022. doi: 10.48550/arXiv.2212.09251.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. *Conference on Robot Learning*, 2023. doi: 10.48550/arXiv.2307.01928.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. doi: 10.48550/arXiv.1809.02156.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2310.13548.

Ziyao Tang, Pengkun Jiao, Bin Zhu, Huiyan Qi, Jingjing Chen, and Yu-Gang Jiang. Spatiotemporal sycophancy: Negation-based gaslighting in video large language models. *arXiv preprint arXiv:2604.17873*, 2026. doi: 10.48550/arXiv.2604.17873.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2305.04388.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. doi: 10.48550/arXiv.2409.12191.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.48550/arXiv.2201.11903.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. In *arXiv preprint arXiv:2308.03958*, 2023. doi: 10.48550/arXiv.2308.03958.

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. *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2306.13063.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. doi: 10.48550/arXiv.2306.02858.

Shuo Zhang, Sisi Zhang, Fangzhou Zhao, Zesong Meng, Yanyan Chen, Kai Liang, Lei Wang, Ruifeng Cui, and Ke Zhang. Have the VLMs lost confidence? a study of sycophancy in VLMs. *arXiv preprint arXiv:2410.11302*, 2024. doi: 10.48550/arXiv.2410.11302.

Wenrui Zhou, Mohamed Hendy, Shu Yang, Qingsong Yang, Zikun Guo, Yuyu Luo, Lijie Hu, and Di Wang. Flattery in motion: Benchmarking and analyzing sycophancy in video-LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2026. doi: 10.48550/arXiv.2506.07180.

A Evaluation Construction and Metric

Each source item is converted into two complementary interaction cases. In a *should-hold* case, the first answer r_1 is correct and the user's claim w is a different, incorrect option. In a *should-flip* case, r_1 is incorrect and the user asserts the gold option. The same decision rule is scored on both sides:

$$\text{balacc} = \tfrac{1}{2}\left(\Pr[\hat{y} = y \mid \text{should-hold}] + \Pr[\hat{y} = y \mid \text{should-flip}]\right).$$

Never-change obtains perfect should-hold accuracy and zero should-flip accuracy; always-accept has the reverse profile. Both therefore score exactly 50%. This construction prevents a uniformly stubborn policy from appearing robust. It also makes the evaluation symmetric with respect to the model's two possible errors: changing when it should hold and refusing to change when it should flip. The decomposition is therefore diagnostic as well as aggregational. A method can improve the headline score either by preserving correct answers, by recovering mistakes, or by improving both; the separate terms reveal which behavior is actually responsible for a gain.

We sample eight frames uniformly from each video and normalize option-letter probabilities over the options available for that item. CGA accepts w when $p_{r_1} < \tau$ and otherwise keeps r_1. For each independently analyzed unit, a deterministic RNG seed of 0 forms five folds. Each training split chooses τ from $\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$ to maximize balanced accuracy; its held-out fold supplies the prediction. Paired bootstrap intervals use 2,000 side-stratified resamples with independent fixed seeds for each baseline comparison. Pairing keeps the compared policies aligned on the same video-question instances, while stratification preserves the should-hold and should-flip composition in every resample. These choices make the reported intervals sensitive to policy differences rather than to accidental changes in sample composition.

B Complete Frozen Results

Table 4 reports the frozen 2026-09-09 analysis. All eleven comparisons against naive acceptance and never-change have per-comparison 95% bootstrap intervals excluding zero. CGA improves over VBA in all cells; the MVBench and PerceptionTest InternVL3-8B comparisons are not individually significant. These intervals are not family-wise corrected. The table should therefore be read as a complete effect-size record with per-comparison uncertainty, not as a claim that all eleven hypotheses jointly pass a single multiplicity-adjusted test. The consistent direction of the effects is nevertheless visible across datasets and model families, while the two non-significant VBA comparisons mark the boundary of the strongest claim.

C Threshold Transfer and Reproducibility Boundary

The fixed-threshold experiment is intentionally separate from the cross-validated main estimates. We select $\tau = 0.86$ on the original Video-MME calibration split and apply it unchanged to all 400

Table 4: Complete two-sided results. Deltas and 95% bootstrap CIs are in percentage points. H/F gives should-hold/should-flip counts.

Dataset	Model	H/F	CGA	AUROC	Δ naive [CI]	Δ VBA	Δ never-change
Video-MME	InternVL3-8B	245/155	68.6	.757	+12.9 [7.2,18.8]	+12.3 SIG	+18.6 SIG
TempCompass	InternVL3-8B	285/115	72.2[†]	.776	+14.3 [8.2,20.7]	+9.4 SIG	+22.1 SIG
MVBench	InternVL3-8B	332/68	75.8	.823	+8.9 [1.2,16.4]	+5.4 ns	+25.8 SIG
MVBench	InternVL3.5-8B	314/86	73.8	.815	+11.6 [5.3,17.5]	+16.3 SIG	+23.8 SIG
MVBench	Qwen3-VL-8B	144/56	69.2	.766	+50.4 [41.5,58.7]	+55.7 SIG	+19.2 SIG
EgoSchema	InternVL3-8B	268/132	68.8	.751	+7.5 [2.2,12.6]	+7.9 SIG	+18.8 SIG
EgoSchema	InternVL3.5-8B	244/156	66.9	.745	+12.0 [7.0,16.7]	+11.0 SIG	+16.9 SIG
EgoSchema	Qwen3-VL-8B	142/58	69.1	.774	+53.4 [45.1,61.9]	+57.5 SIG	+19.1 SIG
PerceptionTest	InternVL3-8B	153/47	72.5	.785	+9.2 [2.1,16.3]	+6.6 ns	+22.5 SIG
PerceptionTest	InternVL3.5-8B	134/66	69.6	.800	+12.7 [2.9,22.5]	+19.7 SIG	+19.6 SIG
PerceptionTest	Qwen3-VL-8B	140/60	70.6	.766	+57.5 [50.0,65.2]	+59.5 SIG	+20.6 SIG

[†]TempCompass uses the Video-MME threshold $\tau = 0.86$ without retuning; its deltas and intervals are computed at that fixed threshold.

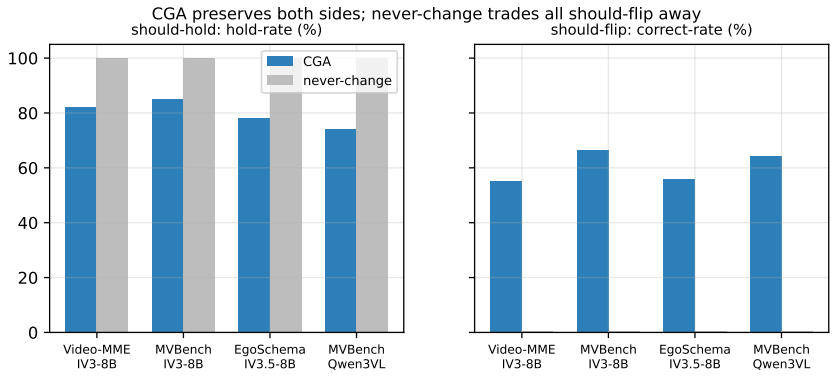


Figure 6: Should-hold vs. should-flip decomposition. `never-change` maximizes hold but zeroes flip; `naive` trades them off; CGA keeps both components high across models.

TempCompass items. It obtains 72.2% balanced accuracy, improving over naive by 14.3 points with a 95% CI of $[8.2, 20.7]$. The independently cross-validated TempCompass estimate is 73.0%; the close values indicate that transfer, rather than test-set retuning, drives the reported result.

The supplied analysis script resets the fold RNG for every dataset–model unit. This matters because sharing one global RNG across a sequence of files makes fold assignments depend on command-line order. Bootstrap streams are also separate from fold generation and from one another. The machine-readable summary records the seeds, threshold grid, selected fold thresholds, point estimates, and intervals.

D Artifact Map

The code-and-data supplement contains only code, frozen per-item outputs, and structured summaries. The Video-MME row combines `candA_confgate_internvl3_8b.jsonl` and `candA_confgate_internvl3_8b_mediumlong.jsonl`. Each other table row maps to the file whose dataset and model appear in its filename. The command in the supplement README regenerates the JSON, CSV, and text summaries from these files. Raw benchmark videos and model checkpoints are excluded because of their size and licenses; the README points to the public benchmark and model releases. This mapping is intended to make the frozen analysis auditable without redistributing licensed videos or checkpoints: a reader can identify the exact input files behind every reported row, inspect the summaries, and rerun the aggregation commands. The supplement consequently preserves the evidence needed for verification while keeping large external assets outside the archive.

E Additional Limitations

The confidence gate is a decision layer, not an evidence-retrieval method. If uniform frame sampling misses the decisive event, a high or low option probability cannot reconstruct it. The evaluation also assumes that the user's claim names one answer option. Extending the method to open-ended responses requires candidate normalization and a confidence measure comparable across strings. Finally, the current per-comparison bootstrap analysis should be complemented by larger preregistered samples for claims requiring family-wise error control. These limitations define the intended scope of the result rather than a failure of the frozen comparison. CGA can arbitrate between a model's first option and a single asserted alternative, but it is not an open-ended verifier and cannot recover visual evidence that was never sampled. Future extensions should combine confidence gating with adaptive frame selection, structured candidate generation, and multiplicity-controlled evaluations.

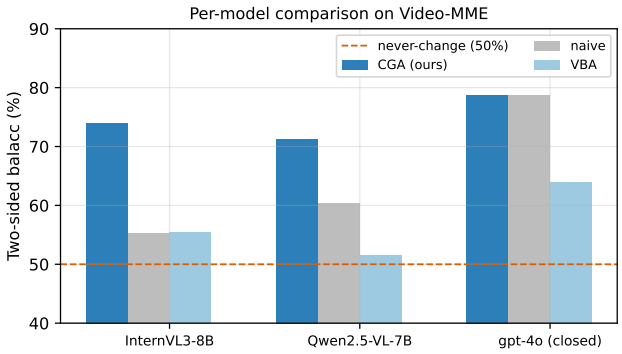


Figure 7: Per-model CGA vs. baselines on Video-MME (5-fold CV balacc %). CGA beats VBA and never-change on all models and `naive` on the two open models.

F Hold vs. Flip Decomposition

Figure 6 decomposes the two-sided balanced accuracy into its two components—*should-hold* hold-rate and *should-flip* correct-rate—for CGA versus the `naive` and `never-change` baselines. `never-change` maximizes the hold component but scores zero on the flip component (hence 50% two-sided), while `naive` trades the two off model-dependently. CGA is the only policy that keeps *both* components high, which is exactly what the two-sided metric rewards. Reading the two components separately is useful for diagnosing deployment behavior: a high hold rate alone could indicate excessive stubbornness, while a high flip rate alone could reflect uncritical deference to the user. CGA is valuable because it moves both terms in the desired direction, preserving evidence-backed answers while opening the gate when the first answer is weak.

G Per-model Comparison on Video-MME

Figure 7 visualizes the per-model Video-MME comparison from Table 2 (main text). CGA significantly beats Visual Belief Anchoring and `never-change` on every model and beats `naive` on the two open models; the only non-significant cell is `gpt-4o` CGA−naive, whose `naive` policy is already strong. The figure makes the model-level heterogeneity explicit: the open models have more room for improvement because naive acceptance is less reliable, whereas gpt-4o begins from a stronger baseline. CGA should therefore be interpreted as a calibration-aware decision layer whose benefit depends on the interaction policy, not as a universal replacement for improving the underlying video model.