THE DOUBT IS NOT GONE, ONLY MOVED: ANTI-CALIBRATED CONFIDENCE AND A FREE DENOISING-TRAJECTORY HALLUCINATION DETECTOR FOR DIFFUSION VLMS (SEISMO)

Anonymous authors
Paper under double-blind review

ABSTRACT

A model's confidence is usually treated as a property of its weights. In a masked-diffusion vision–language model (VLM), we show it is a property of how the model *decodes*. At each denoising step the sampler fills in the answer positions it is most sure about and re-masks the rest. The final distribution is therefore sharpened by the sampler itself, so it looks equally certain whether the answer is right or wrong. This creates a reliability trap. On LLaDA-V, output confidence—the score almost every abstention method uses—cannot find the model's own mistakes: it is at chance on POPE (0.50 AUROC) and *below* chance on GQA (0.37–0.39). A confident answer is no safer than a random one. So where did the doubt go? It is not gone, only moved. The step-by-step *denoising trajectory* still records which answer positions were contested, and for how long. Simply averaging the answer position's per-step entropy detects wrong answers at **0.65–0.79 AUROC** across four diffusion VLMs. It costs nothing, because it is read off the single decode the model already runs. A simple regularity says where it helps: the further a model's trajectory ranking drifts from its endpoint ranking, the larger the gain. **SEISMO** is our answer to that question—to our knowledge the first reliability framework built for diffusion VLMs. A calibration head cuts expected calibration error by $2.7\times$. A Learn-then-Test gate then adds the first *distribution-free* hallucination-risk guarantee for a diffusion VLM: it answers the most reliable half of POPE at **95%** accuracy, where trusting confidence gives 79%. The signal diagnoses but cannot repair: the visual evidence is lost before the decoder runs. It does *route*, though: sending flagged questions to an autoregressive fallback matches full replacement at a fifth of the cost ($0.830 \rightarrow 0.868$), while routing by confidence makes things worse. The lesson is general: a reliability method can silently *invert*, not merely weaken, when carried to a new kind of generation—and the check that catches this is free. All per-example records are released.[1]

1 INTRODUCTION

Vision–language models hallucinate: they describe things that are not in the image (Li et al., 2023; Rohrbach et al., 2018) (Figure 1). In medicine, documents, or assistive use, a wrong answer must be caught before it reaches a person (Geifman & El-Yaniv, 2017; Whitehead et al., 2022). The standard recipe is simple: treat the model's confidence as a reliability score, and abstain when it is low (El-Yaniv & Wiener, 2010; Guo et al., 2017). A large uncertainty-quantification (UQ) literature refines this score (Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016; Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023). All of it rests on one assumption: the model's own confidence carries usable information. For autoregressive (AR) VLMs it mostly does: AR models write their doubt into the output distribution, and a substantial line of work builds on exactly that (Leng et al., 2024; Wu et al., 2025). A parallel line reads AR *internal* states instead (Jiang et al., 2025); our own mirror finds those informative but secondary (Appendix B.1).

[1] `https://seismo-demo.pages.dev` — interactive demo, analysis code and all per-example records.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

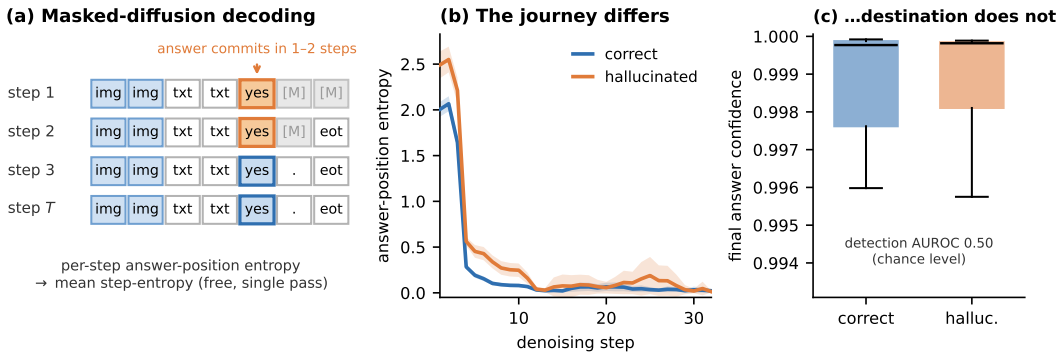

Figure 1: **Different journey, same destination (measured). (a)** A masked-diffusion VLM decodes by iteratively unmasking; the yes/no answer position commits within 1–2 steps, and we read out its per-step entropy—a free, single-pass signal. **(b)** Averaged over POPE (n=400, LLaDA-V), hallucinated answers travel through visibly higher-entropy early dynamics than correct ones. **(c)** Yet the *endpoint* is indistinguishable: final answer confidence sits just below 1.0 for correct and hallucinated answers alike (note the y-axis range), so output confidence detects errors at chance level. The uncertainty that the final distribution has squeezed out survives in the trajectory.

A new family of models breaks this assumption. Diffusion models moved from images (Ho et al., 2020) to text (Austin et al., 2021; Li et al., 2022; Lou et al., 2024; Gong et al., 2023), and recently to vision (You et al., 2026; Li et al., 2025; Dong et al., 2026). A *masked-diffusion* VLM (Nie et al., 2025) does not write left-to-right. The answer starts as a row of masked tokens, and over T denoising steps the model fills in the blanks it is most sure about. So before refining the confidence score, we had to ask a more basic question: *can this model's confidence be trusted at all?* We show the answer can be no—and that the decoding process itself offers something better: the whole *denoising trajectory* of intermediate predictions, not just the final distribution.

Both neighboring lines of work leave a gap (§2). Hallucination tools for VLMs are built for *AR* models; they lean on an output distribution that we show can be *anti-calibrated*—most confident exactly when wrong. Uncertainty work for diffusion LMs is *text-only*: zero-shot trajectory UQ in Vazhentsev et al. (2026), and, concurrently, a confidence-misalignment account for text reasoning in Kim & No (2026). No prior work brings a training-free trajectory signal to the *multimodal* setting, measures the anti-calibration mirror, or adds risk control. Ours does all three.

Our key insight is one sentence: in a diffusion VLM, the answer's *destination* hides the model's doubt, but the *journey* preserves it. The final distribution is sharpened by the sampler itself, so it looks confident whether the answer is right or wrong; the step-by-step trajectory still remembers which positions were contested, and for how long. One mechanism runs through everything that follows: the sampler sharpens the final distribution and erases its own doubt (tested directly, §6). This matters as diffusion (V)LMs spread. An uncertainty method carried over from AR models can *invert* here, not merely weaken, and nothing reveals that until someone measures it. The free trajectory signal avoids the trap.

We ask two plain questions: can we trust a diffusion VLM's confidence, and if not, is there a better signal hiding in the model? Across four diffusion VLMs (three model systems; LaViDa runs two backbones) the answers are "not always" and "yes" (Figure 2). Four contributions, in order of what we believe matters most:

1. **A reliability phenomenon, certified and explicitly scoped.** On LLaDA-family diffusion VLMs that decode *dynamically* (committing the positions they are most confident about first), output confidence detects errors at chance on POPE (0.50) and *below* chance on GQA (0.37–0.39 AUROC, with bootstrap intervals that exclude chance; 0.5 is chance). The same kind of signal works elsewhere: on an AR VLM (the mirror, Figure 2), and on diffusion VLMs that commit in roughly one step (LaViDa and MMaDA-POPE: 0.60–0.77). We therefore do not claim that this flip holds for every diffusion model. The general lesson is narrower, and we state it as such: *a model's confidence is a property of its decoding*

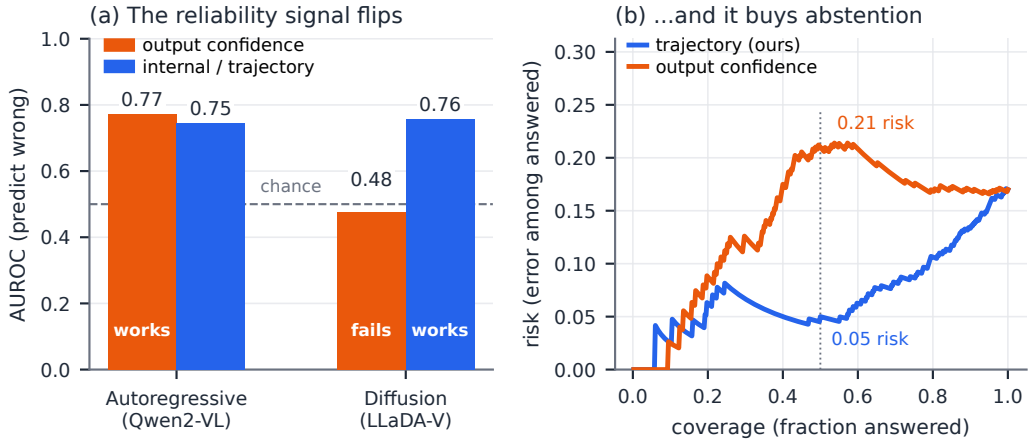

Figure 2: **The reliability signal flips across the generation paradigm—and the trajectory turns that into selective prediction. (a)** For an autoregressive VLM (Qwen2-VL) output confidence *works* as an error detector and the internal signal is slightly worse; for a masked-diffusion VLM (LLaDA-V) the order inverts: output confidence is at chance on POPE (0.48; certified *below* chance on GQA) while the denoising trajectory works. A calibration method tuned on one paradigm silently fails on the other. (POPE, matched AR/diffusion pools, n=600/300; values from Table 4.) **(b)** Abstaining by the trajectory (blue) holds risk far below abstaining by output confidence (orange) at every coverage: at 50% coverage it answers the most-reliable half at 95% accuracy (risk 0.05) versus 79% (risk 0.21), and roughly halves AURC (0.072 vs. 0.141; n=400).

procedure, not of its weights alone—so a reliability method must be re-checked for each training recipe, because it can silently invert rather than merely degrade.

2. **The mechanism, tested—and a predictive regularity.** The sampler's commit rule is what erases the doubt, and two tests show it. Switching the *same* model to random re-masking lifts confidence from 0.48 to 0.56, while the trajectory signal is unaffected by which re-masking schedule is used. Splitting the trajectory into early and late steps puts the signal in the earliest ones: step 1 alone reaches 0.70, whereas everything after the commit is at chance. A regularity follows. For each model–benchmark pair (a *cell*), the further its trajectory ranking drifts from its endpoint ranking, the larger the trajectory's advantage (r=0.89 over twelve cells; r=0.79 among the eight high-drift cells, so we report an empirical regularity, not a law). It also predicts *out-of-sample*: on a held-out fourth model it tells us which cells confidence will still win (§5.3).

3. **The deliverable: SEISMO.** The mean step-wise entropy of the trajectory is single-pass and training-free, read straight off the generation the model already runs. It detects wrong answers at **0.65–0.79 AUROC** on POPE/GQA across four diffusion VLMs, and at 0.59–0.79 across all thirteen in-domain settings we report. Built on top of it, a per-model calibration head lowers POPE ECE from 0.169 to 0.063. A Learn-then-Test gate then adds what is, to our knowledge, the first distribution-free hallucination-risk control for a diffusion VLM. On low-error, POPE-style queries it safely answers 2.2–10.6× more questions than confidence (§5.5).

4. **The boundaries, mapped—and the payoff.** Acting on the signal does not fix answers inside the model itself: not commit re-ordering, not step-voting, and not preference training, which shifts the model's default answer rather than what it perceives. Methods that sample several answers and look for disagreement fail here, and still fail at M=10 samples. *Supervised* probes over these same trace features *invert* across models (0.25–0.48) where the zero-shot statistic still works, so being training-free is *necessary*, not merely convenient. The diagnostic does pay off across systems, though: sending flagged queries to an AR fallback matches full replacement at a fifth of the cost (0.830→0.868 on POPE; §5.7).

2 Related Work

Table 15 positions us among three neighboring lines of work. **(i) Uncertainty for diffusion LMs** reads the decoding trace, but only for *text*: supervised trace and denoising-dynamics detectors (Chang et al., 2026; Qian et al., 2026), hidden-evidence verification (Zhao et al., 2026), zero-shot trajectory UQ (Vazhentsev et al., 2026), and further analyses of confidence and layer dynamics (Yang et al., 2026; Fang et al., 2026; Guo & Tan, 2026), with confidence misalignment identified concurrently for text reasoning (Kim & No, 2026). We bring the signal to vision, quantify *anti*-calibration and the AR mirror, and add risk control. **(ii) VLM hallucination work** (Leng et al., 2024; Wu et al., 2025; Jiang et al., 2025) is built for *autoregressive* models: it leans on an output distribution that carries signal there but, as we show, not always here. The closest analog is Jiang et al. (2025) (interpret–detect–act via attention); our instrument is the trajectory, and our act is abstention with guarantees. **(iii) Conformal abstention for VLMs** (Tayebati et al., 2025; Xu et al., 2026) relies on exactly the confidence scores that fail on diffusion VLMs. SEISMO supplies the missing score. Extended discussion in Appendix I.

3 The SEISMO Framework

SEISMO in three steps (base model frozen; no fine-tuning; one standard decode):
1. Score. While the model decodes, record the answer position's entropy at each step and average: high mean step-entropy = contested answer.
2. Calibrate. Fit a five-feature logistic head once on held-out data to turn the score into a probability the answer is wrong.
3. Guard. Pick a risk budget α. Learn-then-Test calibration returns a threshold such that, with probability $\geq 1-\delta$, answered questions err at rate $\leq \alpha$ (e.g., α=0.1: at most 10% of answered questions wrong). Answer, or route the rest to a human or a fallback model.

SEISMO works like a seismograph attached to a frozen model: the model decodes exactly as usual, and the three components—**SEISMO-Score** (§3.2), **SEISMO-Cal**, and **SEISMO-Guard** (§5.5)—score, calibrate, and act on its denoising trajectory. Nothing is trained or modified in the base model.

3.1 The Denoising Trajectory

What does a masked-diffusion VLM expose that an AR model does not? (Figure 5 in Appendix A diagrams the whole readout.) At every denoising step t, the model predicts a distribution p_t for every answer position, then commits only the positions it is most sure about. For each position we record what the model believed at all T steps, including after it locks an answer in. That history is the trajectory. It is computed anyway during decoding, and it keeps the doubt that the final distribution throws away.

3.2 SEISMO-Score: trajectory statistics

We want one number for how likely the final answer is *wrong* (higher = less reliable): two confidence baselines (`final_maxprob`, `final_entropy`) and five trajectory statistics, all read from one decode (Algorithm 1, Appendix A; every symbol and signal name is defined in Table 3). Our default **SEISMO-Score** is the strongest: `mean_step_entropy`, the answer position's entropy (how spread-out the guess is) averaged over the T steps. The intuition is simple: a correct answer settles early and stays flat; a wrong one keeps changing and stays uncertain.

We evaluate detection with rank-based AUROC (Mann–Whitney) and selective prediction with the risk–coverage curve and its area, AURC (lower is better).

3.3 Trajectory-Calibrated Remasking (TCR)

TCR is the corrective counterpart to detection: it replaces the sampler's confidence-based commit rule with a trajectory-stability score. The full definition and update rule are deferred to Appendix A.7 (results in §5.7).

4 EXPERIMENTAL SETUP

Models. Our main model is LLaDA-V (You et al., 2026), an 8B masked-diffusion VLM. We decode with T=32 reverse steps and generation length 32. The AR comparison is Qwen2-VL-2B (Wang et al., 2024), an instruction-tuned VLM in the LLaVA/InstructBLIP lineage (Liu et al., 2023; Dai et al., 2023). Note up front: LLaDA-V does not beat the AR model on accuracy in our matched check (GQA two-hop, counting the gold answer anywhere in the generation: 0.57 vs. 0.60; rule inventory: Appendix H). Hence our focus on *reliability*—when to trust an answer—not accuracy gains. **Benchmarks.** Two axes. For *reasoning* we use GQA (Hudson & Manning, 2019), compositional questions about everyday scenes; we take its harder two-hop subset (answers need two chained visual facts, e.g., "What color is the chair left of the table?"). For *perception* we use POPE (Li et al., 2023) (adversarial category; Appendix A), which asks yes/no object questions ("Is there a motorcycle in the image?"); a wrong "yes" is exactly an object hallucination. We use n=300–400 per benchmark. Scoring is dataset-aware (yes/no regex for POPE, first-word for GQA), and applied the same way to every signal, so the comparisons are fair (contains-match robustness: Appendix H). **Trajectory capture.** One forward hook on the output embedding records the answer-position logits at every step. It adds no compute. Full checkpoints, decoding settings, and scoring rules are in Appendix A.

5 RESULTS

5.1 OUTPUT CONFIDENCE IS AT-OR-BELOW CHANCE; THE TRAJECTORY IS STRONG

Table 1 reports AUROC for predicting a wrong answer; Figure 6 shows separation and ROC; Figure 10 walks through two examples. The headline: **the two kinds of signal sit on opposite sides of chance.** Output confidence (`final_entropy`) scores **0.385** on GQA and **0.500** on POPE (`final_maxprob`: 0.371/0.504); all at or below the 0.5 chance line: a confident LLaDA-V answer is no safer than random. The trajectory statistics, by contrast, sit well above chance on both benchmarks (the hybrid `conf_at_commit`, 0.561 on POPE, is the weakest). The best, `mean_step_entropy`, reaches **0.723** on GQA and **0.778** on POPE. That is a **+0.34** AUROC gain on GQA, and it holds on both tasks. It is not a small-sample artifact: the 95% intervals are *disjoint* (0.778 $[0.71, 0.84]$ vs. 0.500 $[0.43, 0.57]$; 10,000 resamples). **Takeaway: the model's stated confidence cannot be trusted as a warning signal here (at or below chance overall; App. H); the free trajectory statistic is a strong detector.**

Table 1: AUROC for predicting a *wrong* answer on LLaDA-V (higher is better; 0.5 = chance). Output confidence is at/below chance; the trajectory statistics sit well above it (the hybrid `conf_at_commit` is weakest on POPE). GQA n=300, POPE n=400; per-example records of every pool are released (Appendix H); ties midranked.

Score	GQA (reasoning)	POPE (perception)
Output-confidence baselines		
`final_maxprob`	0.371	0.504
`final_entropy`	0.385	0.500
Denoising-trajectory statistics (ours, training-free)		
`n_changes`	0.655	0.725
`commit_step`	0.672	0.735
`conf_at_commit`	0.676	0.561
`temporal_incons`	0.679	0.723
`mean_step_entropy`	**0.723**	**0.778**

5.2 A MIRROR IMAGE OF THE AUTOREGRESSIVE CASE

For AR VLMs the picture is *reversed*: on Qwen2-VL (same benchmarks and scoring) output confidence works (0.736 GQA / 0.771 POPE) while internal trajectory-style signals trail it (0.712 / 0.746). This is the mirror image of the diffusion case (Figure 2; Appendix B.1). The mirror is not a scale artifact: at matched scale, Qwen2-VL-*7B* confidence also works (0.69 POPE / 0.79 GQA, shared pools)—a cross-recipe contrast, not a controlled ablation. **Takeaway: each paradigm hides its uncertainty elsewhere—AR in the output, diffusion in the trajectory.**

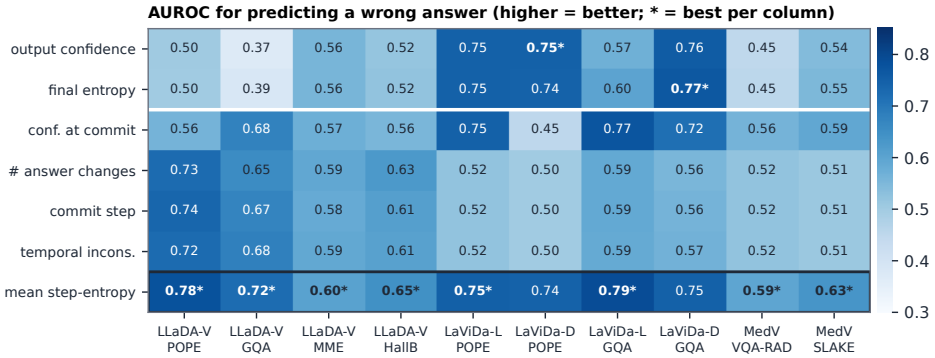

Figure 3: **Signal generality.** AUROC for predicting a wrong answer, for two confidence-family and five trajectory-family signals, across two general-domain diffusion VLM systems (three backbone instances; MMaDA appears in Table 5, Appendix B.2), four benchmarks, and the medical LLaDA-MedV (VQA-RAD, SLAKE). Mean step-entropy (boxed) is best or tied-best in every setting shown except the low-decoupling, near-one-step-commit cells (the Dream-backbone cells; also MMaDA-POPE in Table 5), where the answer commits early so confidence is as good or slightly better—matching our conditional claim.

5.3 Generality across diffusion VLMs

Is this specific to LLaDA-V? Repeating the experiment on LaViDa (LLaDA-8B and Dream-7B backbones) and MMaDA (a third recipe; both regimes reproduce within it) with 95% bootstrap CIs (Figure 3): **(1)** the trajectory detector generalizes (0.65–0.79 over the six general-domain cells; 0.59–0.79 with medical); **(2)** output confidence does not (at-or-below chance on LLaDA-V, works on LaViDa and MMaDA-POPE: 0.60–0.77)—its reliability is model-specific; **(3)** on the LLaDA-initialized *dynamic* cells the trajectory is never worse than confidence, and it beats it by $+0.34/+0.19/+0.11$ (LLaDA-V, LaViDa-LLaDA, MMaDA GQA); on near-one-step-commit cells the two coincide (App. B.9; on MMaDA-POPE confidence is slightly, significantly better, $\Delta = -0.057$—as the regularity predicts). On LLaDA-V it also beats confidence on HallusionBench (0.65 vs. 0.52, paired Δ $[+0.04, +0.21]$) and is positive but not significant on MME (0.60 vs. 0.56); Table 6, Appendix B.3. **Takeaway: on short-answer QA the detector travels across models; confidence's reliability is model-specific (and weaker on harder/open-ended tasks, §6).**

The medical domain inherits the pathology. Porting the identical capture to LLaDA-MedV (Dong et al., 2026), a biomedical fine-tune of LLaDA-V, on VQA-RAD (radiology) and SLAKE (clinical images): output confidence is again *at-or-below chance* on VQA-RAD (0.45 $[0.39, 0.50]$) while the trajectory leads in every cell (0.59/0.63; paired bootstrap deltas $+0.14$ $[0.08, 0.20]$ / $+0.09$ $[0.02, 0.15]$, both exclude zero). Absolute accuracies under our unified protocol differ from published constrained-scoring numbers (both in Appendix F); the calibration conclusions are protocol-robust. The *base* model under this shift (accuracy 0.42) has no working detector (traj 0.53/conf 0.56; a boundary—the advantage presupposes an operative signal, Appendix H). At medical error rates ($\approx$46%) risk control does not yet deliver useful coverage (§5.5), and plain selective prediction is weak in absolute terms: abstaining to 50% coverage by the trajectory lowers the answered half's error only from 0.458 to 0.405 on VQA-RAD. The medical value here is detection and triage signal, not guarantees or deployable accuracy.

5.4 Selective prediction: answer the reliable half at 95%

A good error score should let the model *abstain* well. At 50% coverage, abstaining by `mean_step_entropy` answers the most-reliable half of POPE at **95% accuracy**. Abstaining by confidence leaves **21%** error on the same half—*worse* than the unselected 17% base error. Over all coverages it roughly halves AURC (**0.072 vs. 0.141**; Table 11, Appendix C.2). Figure 2b shows the curves. **Takeaway: with SEISMO, a deployed diffusion VLM can answer half its questions at**

95% accuracy with no retraining. This assumes a labeled calibration set drawn from the same conditions as deployment (exchangeable).

Calibration (SEISMO-Cal). A five-feature logistic head over the trajectory features turns the ranking into a calibrated wrong-probability. Expected calibration error (ECE) measures the gap between the probability a model states and how often it is actually right. Ours is 0.063 against 0.169 for raw confidence, $2.7\times$ lower (details in Appendix C.1).

5.5 SEISMO-GUARD: DISTRIBUTION-FREE RISK CONTROL

Selective prediction shows the trade-off after the fact. Deployment needs the opposite order: fix a risk budget α *first* and guarantee—with probability at least $1-\delta$ over the draw of the calibration set—that answered questions err at most at rate α. We instantiate Learn-then-Test risk control (Angelopoulos et al., 2025; Bates et al., 2021): split off a calibration set, then scan keep-thresholds from most-reliable coverage upward while an exact binomial upper confidence bound on the kept-error (Clopper & Pearson, 1934) stays $\leq \alpha$ (δ=0.1). We report test coverage over 200 random splits. To our knowledge this is the **first distribution-free hallucination-risk guarantee for diffusion VLMs**. The guarantee is *marginal*—over the answered set, not per content class (Appendix H). Table 2 shows the cost of unreliable confidence: gating on it, LLaDA-V safely answers only 10% of POPE at α=0.15 and 24% at α=0.20; gating on the trajectory, 21% and 83%—a 2.2–3.5$\times$ gain at the same guaranteed risk. On LaViDa the trajectory matches confidence within a few points. Tripling the calibration set (n=1200) unlocks the strict α=0.10 budget: **47%** coverage versus **4%**. Figure 7 shows the gaps. The guarantee's two knobs trade off against each other in the way the theory predicts, and we map the whole (α, δ) surface in Figure 8. Empirical violation rates stay at or below δ except in three cells (see the table caption). Two limits, stated plainly: (i) α=0.05 delivers $\leq$1% coverage for every score at these sample sizes; (ii) on high-error benchmarks (base error 24–50%) no score achieves useful coverage—risk control is currently a tool for POPE-style existence queries.

Takeaway: SEISMO-Guard delivers what is, to our knowledge, the first finite-sample hallucination-risk guarantee for diffusion VLMs. A threshold calibrated on adversarial POPE transfers conservatively to the popular and random categories: kept-error stays $\leq \alpha$ at every budget, because those categories have lower base error. The opposite case is the unsafe one. If base error drifts *upward* after calibration, any marginal guarantee fails silently, so we add a sequential e-process monitor over realized kept-error. It alarms in a median of 47 answered examples at a 0.045 false-alarm rate (Table 12, Appendix C.5).

Table 2: Learn-then-Test risk control (δ=0.1, 200 splits): test coverage at risk budget α (higher is better); "conf" gates on output confidence, "traj" on mean step-entropy. On LLaDA-V the trajectory safely answers 2.2–3.5$\times$ more at n=400 and up to 10.6$\times$ at n=1200; on LaViDa it matches within a few points (one cell trails: Dream, α=0.15). Three cells' violation rates nominally exceed δ (max 0.125; all within MC error). Cells rounded; exact protocol, unrounded ratios, and derivations: Appendices A and H.

	cov. @ α=0.10		cov. @ α=0.15		cov. @ α=0.20	
POPE	conf	traj	conf	traj	conf	traj
LLaDA-V (n=400)	0.00	0.00	0.10	**0.21**	0.24	**0.83**
LLaDA-V (n=1200)	0.04	**0.47**	0.18	**0.80**	0.32	**0.96**
LaViDa-LLaDA (n=400)	0.00	0.00	0.36	0.36	**0.92**	0.90
LaViDa-Dream (n=400)	0.00	0.00	**0.74**	0.62	0.94	0.94

5.6 COMPARISON TO SAMPLING-BASED AND TRACE-BASED BASELINES

Sampling-based UQ. Semantic entropy and self-consistency *collapse* toward chance on LLaDA-V, because aggressive commitment removes the sample diversity they need. They stay there as samples grow (0.52 at M=3, 0.60 at M=10); P(True) reaches 0.65–0.71, matching but not exceeding the free single-pass trajectory (0.78) at an extra pass (Table 9, Appendix B.6). *Trace-based detectors.* Re-implementing the two learned trace-detector families for text diffusion LMs (a TraceDet-style supervised trace probe, a DynHD-style dynamics probe) on our LLaDA-V POPE captures over four

step budgets, they only *match* our zero-shot mean step-entropy in-domain ($\leq$0.04 either way, within cross-validation noise), while requiring training and *inverting* across models (0.25–0.48) where the zero-shot scalar still travels—so a trained probe buys nothing here and forfeits portability (Table 7, Appendix B.5).

5.7 Acting on the signal does not improve accuracy

Can the same signal *fix* answers? The natural test replaces the sampler's anti-calibrated confidence commit rule with our trajectory score (TCR, §3.3). **It does not help.** Accuracy is unchanged on both benchmarks (GQA 0.363→0.360; POPE 0.847→0.843; Table 13 in Appendix D). TCR only changes the *kind* of error (false "yes" 33→25 on POPE, traded for false "no"). The post-hoc route fails identically: four step-voting variants all tie the plain final-step readout (0.815 on POPE)—binary answers commit early and stably, even when wrong, so there is little flicker to exploit. **Takeaway: the signal is a diagnostic, not an in-stack cure—no trajectory-derived decode-time rule recovers the answer (§6 shows why); but it *routes*: gating a hand-off to an AR fallback matches full replacement at a fifth the cost (0.830→0.868 at 20% routed; confidence-gating *hurts*; App. D.2).**

We pushed further: the *first* hallucination-targeted preference training for a diffusion VLM (LoRA with a diffusion-correct DPO loss on detector-mined pairs; App. D.3). POPE rises 0.785 → 0.825 but the prediction distribution flips yes→no and GQA does not improve (0.544 → 0.528): a decision-threshold shift (Yin et al., 2025), not recovered grounding—at this budget, training-time optimization shifted the model's default answer, not what it perceives, exactly as the upstream account predicts (§6).

6 Analysis

The inversion has one cause: the sampler is greedy about confidence. Each step commits the most confident positions and re-masks the rest (Nie et al., 2025) (this confidence-misalignment is concurrently identified for text-diffusion *reasoning* by Kim & No, 2026; we show its consequence for *multimodal* reliability and turn it into a detector). An AR model's doubt lands in the output distribution; a diffusion endpoint is *sharpened by the sampler itself*, certain whether right or wrong. That one property explains the endpoint's *unreliability* (Table 1), the collapse of sampling-based UQ (re-samples agree regardless of truth; Appendix B.6), and the informative trajectory, whose record of contested positions is exactly the doubt the endpoint squeezed out. Decomposition confirms it: step-1 entropy alone detects at 0.70, the first 8 steps at 0.77 ($\approx$ full), post-commit readouts at *chance* (Table 10, Appendix B.10). Nor is it *merely* a difficulty meter. We hold difficulty fixed using an independent AR model's per-item correctness. Detection still holds in the AR-correct stratum (0.86/0.70), and is underpowered only in the smaller, error-saturated AR-wrong stratum (0.43/0.57). The detector therefore ranks idiosyncratic errors at a fixed difficulty (Appendix H). The *below-chance* GQA inversion has one further cause: the model's most confident answer classes are also its most error-prone (class-level corr. 0.63; Appendix H).

Why the signal detects but cannot correct. Both corrective tricks we tried—re-ordering the commits (TCR) and voting over the steps (§5.7)—leave accuracy unchanged. Instability says a position was *contested*, not which token is right. How much the endpoint hides is schedule-set: *random* remasking lifts confidence from 0.48 to 0.56 at a 2.7-point accuracy cost while the trajectory is schedule-robust (0.76 both; Appendix B.10), and LaViDa's one-step commit is the extreme (confidence fine, flicker empty).

The error is upstream of the decoder. Why does no trajectory-derived decode-time rule fix these errors? We localize where the evidence is lost, combining a *causal* image ablation with a *correlational* logit lens (protocol: Appendix E.1). The ablation shows the model uses vision (POPE 0.835→0.467 without the image, n=600; 61% of hallucinations flip: misreadings, not priors); and a validated logit lens (nostalgebraist, 2020) shows that in a *majority* of hallucinations the correct answer is never favored above a noise floor at *any* of the 33 lens read-out points, which are the embedding output plus each of the 32 blocks (Figure 4b). That is **85%** (82/97, n=600) at our default floor, and it stays the majority at any floor (53–98%; Table 14). Meanwhile the image still drives the wrong decision and visual attention looks normal (Figure 9; Appendix E.2). We flag one alternative reading of the same

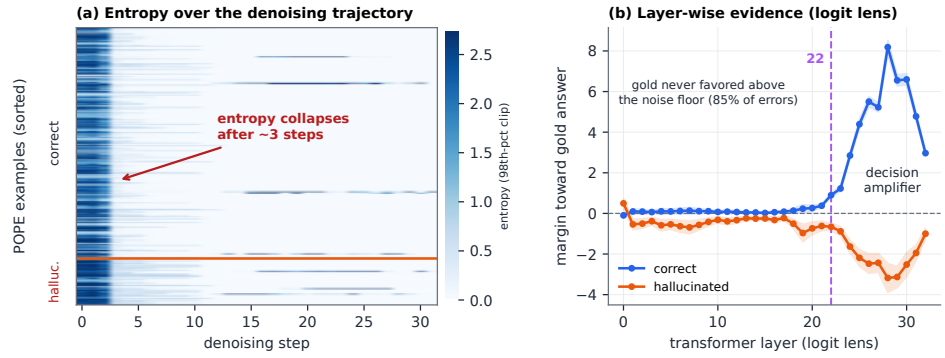

Figure 4: **Where the signal lives, and where the error lives. (a)** Answer-position entropy across all 32 denoising steps for 400 POPE examples (sorted; the divider separates correct from hallucinated). Entropy concentrates in the first $\sim$3 steps and then collapses; the correct/hallucinated difference is a shift in the mean (1.91 vs. 2.42 over steps 0–2), not a per-example contrast visible at this colour scale (Figure 1(b) and Table 1 quantify it). **(b)** Layer-wise logit lens (faithfulness 0.93): the gold margin surges in the late "decision-amplifier" layers ($\ell \geq 22$) for correct answers but is negative from mid-network on for hallucinations—in 85% of errors at the default floor (53–98% across floors) the correct answer is never favored above a measured noise floor at *any* depth.

margins: a per-layer *causal* cut (full-vs-blank lift toward the gold answer) shows the image pushes toward gold nearly as hard on hallucinations as on correct answers ($+3.64$ vs. $+3.92$ late-layer lift), under which the strictly under-encoded fraction is ~ 0.57 $[0.47, 0.67]$ and part of the mass reflects failed *arbitration* between competing evidence rather than absent evidence (Appendix E.1). The load-bearing conclusion is invariant to the partition: either way the failure sits *upstream of any decode-time rule*, which no re-ordering of the denoise can reach—detection succeeds, correction cannot.

Boundary: open-ended generation. In free-running captioning (LLaDA-V, 500 COCO images, 64-token blocks, CHAIR (Rohrbach et al., 2018)) the advantage *vanishes*: mention-level step-entropy 0.62 vs. confidence 0.63; at caption level confidence leads (0.62 vs. 0.54) and is *not* anti-calibrated here—outside the mirror. In a long block step-entropy mostly encodes *when* a position committed (rank-corr. 0.84 with commit vs. 0.27 with confidence; residualized 0.55): a decode-order clock where decoding is not contested; we scope SEISMO to prompted short-answer reliability (Appendix G).

7 Conclusion

We asked whether a diffusion VLM's confidence can be trusted. Not always: on LLaDA-V it is at or below chance. But the doubt survives in the denoising trajectory, and SEISMO reads it for free (0.65–0.79 across four models), calibrates it (ECE 2.7$\times$ lower), certifies it (up to 10.6$\times$ more questions at the same risk), and routes on it (0.830$\rightarrow$0.868 on POPE). The lesson travels: confidence is a property of how a model decodes, not just its weights—tooling ported across generation paradigms must be re-checked, and here the check is free.

Limitations

Our evidence spans four diffusion VLMs (three LLaDA-initialized recipes plus a Dream-7B backbone), six QA benchmarks plus COCO captioning, English only, n=225–1506 per cell, one seed, greedy decoding (robustness: Appendix B.4). **The advantage is conditional**: every dynamic cell to date is LLaDA-initialized, and forcing the independent Dream backbone into dynamic decoding did *not* transfer it. **The guarantee is not free**: Cal and Guard need 400–1200 labeled examples and per-step logits, and certify useful coverage only on low-error queries. **We diagnose, we do not correct**; the under-encoding/arbitration partition is floor-sensitive (53–98%). Appendix J gives all six boundaries in full.

Ethics Statement

This work studies *reliability* of vision–language models: detecting when a model's answer is likely wrong so it can be withheld or routed to a human. The intended impact is positive—reducing the risk of confidently-wrong (hallucinated) outputs in high-stakes settings such as medical or document understanding. We use only public benchmarks with no personally identifying content beyond what those benchmarks already contain. A residual risk is over-reliance: a reject-option signal that is good but not perfect could induce misplaced trust in the non-abstained answers; we therefore report full risk–coverage curves rather than a single operating point, and we are explicit that the signal is diagnostic, not corrective. Deployments should pair the abstention gate with a sequential risk monitor on realized kept-error (e.g., an anytime-valid e-process) that alarms if risk drifts above the certified level; the released per-example records suffice to implement one.

Reproducibility Statement

All results are computed from a frozen, publicly released base model with no training except one negative-control LoRA run (Appendix D.3). The trajectory capture (one forward hook, no extra forward passes), the five-feature SEISMO-Cal head, and the Learn-then-Test SEISMO-Guard procedure are specified in Section 3 and Appendix A; models, checkpoints, decoding settings (T=32, generation length 32), example-pool sizes (n=225–1506), and dataset-aware scoring rules are in Appendix A and H. We report 95% bootstrap/Wilson intervals and, for risk control, realized test coverage and risk over 200 random calibration/test splits. Per-example records (trajectory statistics plus benchmark example IDs) for every pool are released, together with the analysis code, at an anonymous companion site, `https://seismo-demo.pages.dev`, so that every table and figure can be reproduced from the released logs on CPU. The site presents the same tables and figures alongside a step-by-step viewer of the denoising trajectory on real examples, and serves the analysis code and every per-example record file directly, so the headline AUROCs of Table 1 can be recomputed on CPU by a fifteen-line script that downloads nothing else. Appendix A.9 expands this statement.

Use of Large Language Models

Generative AI tools were used as general-purpose assistants in preparing this paper, in three ways. *Writing*: drafting and editing prose, including sections of the main text and the appendix, which the authors then reviewed and revised. *Coding*: drafting and refactoring the capture, analysis and plotting scripts that produce every table and figure. *Retrieval and discovery*: locating related work and checking bibliographic records against primary sources. No experimental result was generated or altered by an AI tool: every number reported here was produced by the released scripts from the released per-example records, and re-verified against those records by the authors, who also reviewed and tested all AI-assisted code. Two further categories do not arise here: no data were generated by an AI tool (every evaluation pool is a public benchmark), and the paper develops no theory and contains no proofs. The authors take full responsibility for the content of this paper, including all claims, numbers, and released artifacts.

References

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics*, 19(2):1641–1662, 2025. First released as arXiv:2110.01052, 2021.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

Shenxu Chang, Junchi Yu, Weixing Wang, Yongqiang Chen, Jialin Yu, Philip Torr, and Jindong Gu. TraceDet: Hallucination detection from the decoding trace of diffusion large language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Charles J. Clopper and Egon S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alessio Devoto. Peeking inside diffusion language models with the logit lens. https://alessiodevoto.github.io/Diffusion-Language-Models-Inner-Workings/, 2026.

Xuanzhao Dong et al. LLaDA-MedV: Exploring large language diffusion models for biomedical image understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. First released as arXiv:2508.01617, 2025.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 2010.

Liancheng Fang, Aiwei Liu, Henry Peng Zou, Yankai Chen, Enze Ma, Leyi Pan, Chunyu Miao, Wei-Chieh Huang, Xue Liu, and Philip S. Yu. Locally confident, globally stuck: The quality-exploration dilemma in diffusion language models. *arXiv preprint arXiv:2604.00375*, 2026.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, 2019.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Zhengnan Guo and Fei Tan. Lost in diffusion: Uncovering hallucination patterns and failure modes in diffusion large language models. *arXiv preprint arXiv:2604.10556*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. First released as arXiv:2311.17911, 2023.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Metod Jazbec, Theo X. Olausson, Louis Béthune, Pierre Ablin, Michael Kirchhof, João Monteiro, Victor Turrisi, Jason Ramapuram, and Marco Cuturi. Learning unmasking policies for diffusion language models. *arXiv preprint arXiv:2512.09106*, 2025.

Yatai Ji, Teng Wang, Yuying Ge, Zhiheng Liu, Sidi Yang, Ying Shan, and Ping Luo. From denoising to refining: A corrective framework for vision-language diffusion model. *arXiv preprint arXiv:2510.19871*, 2025.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Dueun Kim and Albert No. The confidence shortcut: A reasoning failure mode of masked diffusion models. *arXiv preprint arXiv:2605.29123*, 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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Sanghyun Lee, Seungryong Kim, Jongho Park, and Dongmin Park. Lookahead unmasking elicits accurate decoding in diffusion language models. *arXiv preprint arXiv:2511.05563*, 2025.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shufan Li, Konstantinos Kallidromitis, et al. LaViDa: A large diffusion language model for multimodal understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning (ICML)*, 2024.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Vishal Narnaware, Animesh Gupta, Kevin Zhai, Zhenyi Wang, and Mubarak Shah. Seeing to ground: Visual attention for hallucination-resilient MDLLMs. *arXiv preprint arXiv:2603.25711*, 2026.

Shen Nie et al. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

nostalgebraist. Interpreting gpt: the logit lens. LessWrong blog post, 2020.

Seongheon Park et al. VAUQ: Vision-aware uncertainty quantification for LVLM self-evaluation. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2026. First released as arXiv:2602.21054, 2026.

Yanyu Qian, Yue Tan, Yixin Liu, Wang Yu, and Shirui Pan. DynHD: Hallucination detection for diffusion large language models via denoising dynamics deviation learning. *arXiv preprint arXiv:2603.16459*, 2026.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Shreshth Saini, Avinab Saha, Balu Adsumilli, Neil Birkbeck, Yilin Wang, and Alan C. Bovik. TABES: Trajectory-aware backward-on-entropy steering for masked diffusion models. *arXiv preprint arXiv:2602.00250*, 2026.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2024. Introduces the MMHal-Bench benchmark; first released as arXiv:2309.14525, 2023.

Sina Tayebati, Divake Kumar, Nastaran Darabi, Dinithi Jayasuriya, Theja Tulabandhula, Ranganath Krishnan, and Amit Ranjan Trivedi. CAP: Conformalized abstention policies for context-adaptive risk management for LLMs and VLMs. In *Proceedings of the 17th Asian Conference on Machine Learning (ACML)*, volume 304 of *Proceedings of Machine Learning Research*, pp. 926–941, 2025. First released as arXiv:2502.06884, 2025, under the title "Learning Conformal Abstention Policies for Adaptive Risk Management in Large Language and Vision-Language Models".

Artem Vazhentsev, Vladislav Smirnov, David Li, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Uncertainty quantification for large language diffusion models. *arXiv preprint arXiv:2605.14570*, 2026.

Chengcheng Wang, Tingzhang Luo, Wenhao Li, Jianyuan Guo, and Chang Xu. TACG: Trajectory-aware commit gating for diffusion language model decoding. *arXiv preprint arXiv:2607.03236*, 2026.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023a.

Peng Wang et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Spencer Whitehead et al. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision (ECCV)*, 2022.

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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

Bohan Yang, Yijun Gong, Zhi Zhang, Ge Zhang, Wenpeng Xing, and Meng Han. TriLens: Per-layer logit-lens entropy for white-box hallucination detection. *arXiv preprint arXiv:2606.01033*, 2026.

Ling Yang et al. Mmada: Multimodal large diffusion language models. arXiv:2505.15809, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in mllms? arXiv:2504.10020, 2025.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. LLaDA-V: Large language diffusion models with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Longxuan Yu, Yu Fu, Shaorong Zhang, Hui Liu, Mukund Varma T, Greg Ver Steeg, and Yue Dong. Thinking out of order: When output order stops reflecting reasoning order in diffusion language models. *arXiv preprint arXiv:2601.22035*, 2026.

Yushan Zhang et al. DiffSF: Diffusion models for scene flow estimation. *arXiv preprint arXiv:2403.05327*, 2024.

Guoshenghui Zhao, Tan Yu, and Weijie Zhao. HIVE: Hidden-evidence verification for hallucination detection in diffusion large language models. *arXiv preprint arXiv:2604.26139*, 2026.

SUPPLEMENTARY MATERIAL

This supplement is organized by topic. Appendix A specifies the experimental setup and the signal definitions, and Appendix B extends the detection results. Appendix C covers calibration, selective prediction and risk control; Appendix D reports what happens when the signal is acted on. Appendix E localizes the underlying failure, and Appendices F and G give the two boundary settings (medical domain shift and open-ended generation). Appendix H lists every capture pool with its cross-section numbers. Finally, Appendices I and J give the extended related work and the unabridged limitations. Each table and figure appears in the section that discusses it.

A EXPERIMENTAL DETAILS

This appendix specifies every model, checkpoint, decoding setting, data filter, scoring rule, signal definition, and metric used in the paper, so that all results can be reproduced from public artifacts.

A.1 NOTATION AND ABBREVIATIONS

Table 3 collects every symbol, captured-signal name, and abbreviation used in the paper. The signal names are the identifiers that label the rows of every detection table; their formal definitions are Algorithm 1. Models and benchmarks are specified in Appendices A.2 and A.4 rather than repeated here.

Table 3: Notation, captured signals, and abbreviations. "Error score" means the convention used throughout: *higher* indicates *more likely wrong*, so every signal is oriented for predict-wrong AUROC.

Trajectory quantities	
T	Number of reverse (denoising) steps; T=32 unless stated (App. A.3).
p_t	The model's distribution over the vocabulary at the answer position at step t.
a_t	$\arg\max_v p_t(v)$: the token favored at step t.
$H(p)$	Shannon entropy $-\sum_{v\in V} p(v)\ln p(v)$ (natural log); V is the vocabulary.
position 0	The first answer token. All signals are read at this position (Limitation 5).
Captured signals (Algorithm 1; all are error scores)	
`mean_step_entropy`	$\frac{1}{T}\sum_t H(p_t)$ — the **SEISMO-Score**: free, single-pass, training-free.
`final_entropy`	$H(p_T)$ — output-confidence baseline (the endpoint's entropy).
`final_maxprob`	$1-\max_v p_T(v)$ — output-confidence baseline. Stored *negated* in the released records, i.e. already an error score.
`n_changes`	$\lvert\{t\geq 2 : a_t \neq a_{t-1}\}\rvert$: how often the favored token changes.
`commit_step`	$\max(\{t : a_t \neq a_T\}\cup\{0\})$: last step disagreeing with the final answer (0 if none).
`temporal_incons`	$1-\frac{1}{T}\sum_t \mathbf{1}[a_t = a_T]$: fraction of steps that disagree with the final answer.
`conf_at_commit`	$1-\max_v p_{\min(\texttt{commit_step}+1,T)}(v)$: uncertainty at the step the answer locks in.
Framework components	
SEISMO	The framework as a whole: Score, Cal, and Guard (§3).
SEISMO-Score	The zero-shot trajectory statistic; the only component that transfers across models (§3.2).
SEISMO-Cal	Five-feature logistic head turning the score into a calibrated wrong-probability (§5.5).
SEISMO-Guard	Learn-then-Test abstention gate giving a distribution-free risk guarantee (§5.5).
TCR	Trajectory-Calibrated Remasking: replacing the sampler's commit rule with the trajectory score (§3.3; a reported *null*).
Risk-control parameters	
LTT	Learn-then-Test: the fixed-sequence testing procedure behind Guard (§5.5).
α	Risk budget: the target error rate *among answered* examples.
δ	Confidence level of the guarantee: $\mathbb{P}(\text{risk} > \alpha) \leq \delta$.

(Table 3 continued)

coverage	Fraction of examples answered (not abstained).
risk	Error rate among the answered examples ("kept error").
Metrics	
AUROC	Area under the ROC curve for predicting a *wrong* answer; 0.5=chance, ties midranked.
AURC	Area under the risk–coverage curve (mean risk over answered prefixes); lower is better.
ECE	Expected calibration error, 15 equal-width bins.
CI	95% bootstrap confidence interval (10,000 resamples unless stated).
ρ	Spearman rank correlation; *decoupling* $= 1-\rho$ between trajectory and endpoint (§5.3).
Other abbreviations	
VLM / LM	Vision–language model / language model.
AR	Autoregressive (left-to-right) generation.
dLLM	Diffusion large language model (text-only).
UQ	Uncertainty quantification.
M	Number of samples drawn by sampling-based UQ baselines (§5.6).
P(True)	Self-evaluation baseline: the model is re-asked whether its own answer is correct (scored both as a binary verdict and as a continuous probability).
τ	Decoding temperature; τ=0.7 in the sampled-decoding check (App. B.4).
ℓ	Transformer layer index in the logit-lens analysis (§6).
DPO / LoRA	Direct preference optimization / low-rank adapter (App. D.3).
VCD	Visual contrastive decoding, the image-ablation baseline family (App. B.6).

A.2 Models and checkpoints

Diffusion VLM (primary). LLaDA-V (You et al., 2026) (HuggingFace `GSAI-ML/LLaDA-V`), an 8B masked-diffusion VLM whose language backbone is the LLaDA-8B diffusion LM (Nie et al., 2025), equipped with a SigLIP-2 vision tower and a two-layer `mlp2x_gelu` projector. We load in half precision (`fp16`) with `sdpa` attention on a single 48 GB GPU. **Autoregressive comparison.** Qwen2-VL-2B-Instruct (Wang et al., 2024) (`Qwen/Qwen2-VL-2B-Instruct`), `fp16`, `sdpa`, image cap `max_pixels`=512×28×28; the matched-scale check uses Qwen2-VL-7B-Instruct (`Qwen/Qwen2-VL-7B-Instruct`), same settings, greedy decoding, max 16 new tokens. **Diffusion VLMs (cross-model check).** LaViDa (Li et al., 2025) in both released variants: LaViDa-LLaDA (`jacklishufan/lavida-llada-v1.0-instruct`, LLaDA-8B backbone) and LaViDa-Dream (`jacklishufan/lavida-dream-v1.0-instruct`), whose backbone is the *Dream-7B* diffusion LM. That backbone is a different diffusion family from LLaDA, and both variants carry a SigLIP-SO400M tower. We include them to test whether the phenomenon is a class property rather than a single-model artifact. Both decode with the same T=32, generation length 32 settings as LLaDA-V. **Third recipe.** MMaDA (`Gen-Verse/MMaDA-8B-MixCoT`), a unified discrete-diffusion multimodal LM. Images enter as MAGVITv2 tokens (`showlab/magvitv2`, resolution 512), not through a vision tower, and the model is LLaDA-8B-initialized (a shared init, distinct recipe—disclosed accordingly). It is decoded with its own `mmu_generate` at T=32, generation length 32, single block, default low-confidence remasking. **Medical.** LLaDA-MedV (`XZDong123/LLaDA-MedV`), a biomedical fine-tune of LLaDA-V, loaded with the vision tower overridden to `google/siglip2-so400m-patch14-384` (the released config hardcodes a local path); identical capture and decoding.

A.3 Decoding settings

Unless stated otherwise, LLaDA-V generates with T=32 reverse (denoising) steps, generation length 32, block length 32, `prefix_refresh_interval`=32 and `threshold`=1. It uses the model's default *low-confidence remasking* schedule, in which each step commits the highest-confidence masked positions and re-masks the rest. Token selection is greedy (`temperature`=0), stopping at `<|eot_id|>`. The denoising loop always executes all T steps; `<|eot_id|>` truncates the decoded string only

(the capture check confirms T=32 hook firings per decode). `threshold`=1 leaves the default committed-count schedule in place (values below 1 enable confidence-thresholded parallel commits); the multi-step trajectories verified in Appendix B.9 confirm decoding was not degenerate. The yes/no scorer matches "yes" with priority when both tokens appear. The GQA parser keeps the text *before* the first newline/rationale marker, so "The answer is red" scores by its first word. That is one reason we also report contains-matching, under which all conclusions strengthen. The prompt template is the model's own `llava_llada` conversation format with the instruction "Answer with a single word or phrase." The AR model decodes greedily (`do_sample`=False) for `max_new_tokens`=16 with `output_scores` enabled to recover the first-token distribution. The denoising-steps ablation (Table 10) sweeps $T \in \{4, 8, 16, 32, 64\}$ with n=200 per point except T=32 (n=400). **The AR "internal signal" in Table 4** is training-free intermediate-layer/final-answer disagreement read through the logit lens. It is the strongest of four internal statistics we measured on Qwen2-VL (internal disagreement 0.712/0.746 on GQA/POPE; late-change 0.51/0.50; resolved-not-final 0.48/0.44; mid–final KL 0.37/0.34). All four are in the released records, and none beats AR output confidence. When a table reports a single "output confidence" column, it shows the *stronger* of the two output-only statistics (max-probability / final entropy) for that cell; both variants are in the released records for every cell.

A.4 Benchmarks, example pools, and scoring

POPE (Li et al., 2023): the object-presence test split (HuggingFace `lmms-lab/POPE`), a balanced yes/no hallucination benchmark. Pools take the first n examples in dataset order. We verified that the first 400 rows of this split are all from the *adversarial* category—the hardest variant, where negatives are frequently co-occurring objects. Our POPE numbers are therefore POPE-adversarial numbers. MME and HallusionBench pools are likewise the first n valid yes/no examples in split order, scored by the same word-boundary yes/no rule. MME is grouped by task category in split order, so a prefix samples the earliest categories rather than the full benchmark; its cell is a within-category probe, not a breadth claim. **GQA** (Hudson & Manning, 2019): the balanced test-dev split (HuggingFace `lmms-lab/GQA`), taking questions from the instruction config and images from the image config. We restrict to the *compositional/two-hop* subset, defined programmatically as examples whose functional program either contains a `relate` operation or has a semantic chain of length $\geq$3; this isolates multi-hop reasoning. The AR and diffusion runs construct their example pools with *identical* code and ordering; the AR mirror extends the same ordered pools to larger n (first-n rule), so every diffusion example is contained in the AR pool. Sample sizes: detection n=300 (GQA two-hop) and n=300–400 (POPE); selective prediction n=400 (POPE); the mirror comparison (Table 4) aggregates the matched GQA and POPE pools. **Scoring is dataset-aware.** Because LLaDA-V often answers verbosely, POPE is scored by a case-insensitive word-boundary regular expression matching `yes` or `no` in the decoded text. GQA is scored by first-word match after truncating any rationale, splitting off text after a newline or the tokens "rationale", "because", "the answer is". The same rules are applied identically to both models.

A.5 Trajectory capture and signal definitions

We register a forward hook on the language model's output embedding (`get_output_embeddings`, i.e. the LM head). Each reverse step performs one forward pass, so the hook fires once per step; it stores the logits of the last 32 (answer-span) positions, cast to `fp32` on CPU. Position 0 is the first answer token. From its per-step distributions $p_1, \ldots, p_T$ we take the argmax a_t, max-probability $\max_v p_t(v)$, and entropy $H(p_t)$, from which Algorithm 1 computes the five trajectory statistics and the two output-confidence baselines. The capture adds *no* extra forward passes—only the memory to hold $T \times 32 \times |V|$ logits on CPU. The AR "output confidence" uses the first generated token's distribution (`final_maxprob`, `final_entropy`); its "best internal" number (Table 4) is the strongest hidden-state disagreement signal from a parallel AR error-detection probe.

A.6 Selective prediction and calibration

Risk–coverage / AURC. Examples are ordered most-reliable-first (ascending wrong-score); at coverage c we answer the top c fraction and report the error rate among them (risk). AURC is the mean risk over all answered-prefix coverages (equivalently, the area under the full risk–coverage

Algorithm 1 Training-free trajectory error signals (per answer position)

Require: step distributions $p_1, \ldots, p_T$ at the answer position; argmaxes $a_t = \arg\max_v p_t(v)$; vocabulary V. $H(p) = -\sum_{v \in V} p(v) \ln p(v)$ is Shannon entropy (natural log); $\mathbf{1}[\cdot]$ is the indicator.

1: `n_changes` $\leftarrow |\{t \geq 2 : a_t \neq a_{t-1}\}|$
2: `commit_step` $\leftarrow \max(\{t : a_t \neq a_T\} \cup \{0\})$ ▷ last step disagreeing with the final answer; 0 if none
3: `temporal_incons` $\leftarrow 1 - \frac{1}{T}\sum_t \mathbf{1}[a_t = a_T]$
4: `mean_step_entropy` $\leftarrow \frac{1}{T}\sum_t H(p_t)$
5: `conf_at_commit` $\leftarrow 1 - \max_v p_{\min(\texttt{commit_step}+1,\, T)}(v)$ ▷ uncertainty when the final answer locks in
6: **return** scores ▷ higher ⇒ more likely wrong

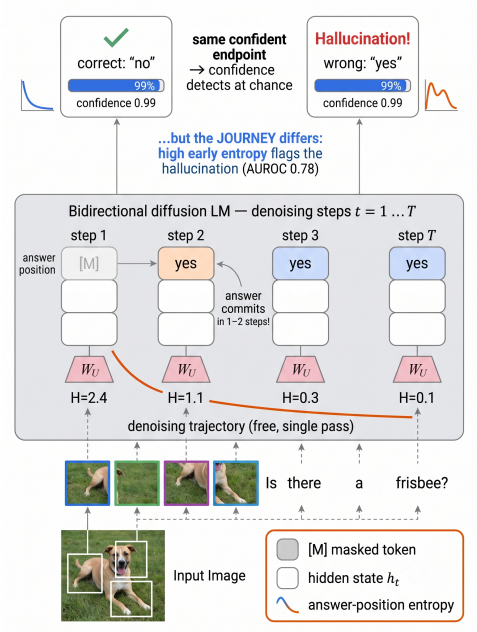

Figure 5: **Overview.** A masked-diffusion VLM answers by iteratively unmasking; the answer position commits within 1–2 steps. Reading out the answer-position entropy at every denoising step gives a free, single-pass *trajectory* signal. Correct and hallucinated answers end at the same saturated confidence (right cards)—output confidence detects errors at chance—but their *journeys* differ, and the trajectory flags the hallucination (0.78 AUROC). Figure 1 shows the corresponding measured data. The trapezoid marked W_U is the model's output embedding (LM head), the single tensor our forward hook reads; no attention weights are involved.

curve); lower is better (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019). **Calibration.** The learned head is a logistic regression (standardized inputs, C=1.0) over the five trajectory features, evaluated with 5-fold stratified cross-validation; we take out-of-fold predicted p(wrong) and compute a 15-bin equal-width ECE (Guo et al., 2017) and the Brier score. The output-confidence baseline uses p(wrong)=$1 - \max_v p_T$. On POPE (n=400) the head attains ECE=0.063 / Brier 0.134 versus 0.169 / 0.170 for confidence; the single `mean_step_entropy` remains the best pure ranker (0.778 AUROC, vs. 0.724 for the head).

A.7 Trajectory-Calibrated Remasking (TCR)

TCR (§3.3) replaces the sampler's confidence-based commit rule. It maintains an exponential moving average of step-entropy, $\overline{H}_t = (1 - \frac{1}{w})\overline{H}_{t-1} + \frac{1}{w}H(p_t)$ with w=4 and $\overline{H}_0$=0. Each masked position is then scored by $\mathbf{1}[a_t{=}a_{t-1}] \cdot (1 - \overline{H}_t / \ln|V|)$, i.e. argmax stable *and* recently low-entropy, with

step 1 falling back to confidence. The sampler commits the highest-scoring positions each step instead of the highest-confidence ones. It is training-free and adds no forward passes; baseline and TCR run on identical inputs via a global toggle.

The anti-calibration result has a direct algorithmic consequence. Masked-diffusion decoding commits the tokens it is *most confident* about and remasks the least confident, step by step (Nie et al., 2025). But if confidence is anti-calibrated, the sampler commits exactly the tokens it is confidently *wrong* about, then locks them in. **TCR** replaces this commit rule: instead of committing by $\max_v p_t^{(i)}(v)$, it commits a position i only when its argmax has been *temporally stable* and its recent step-entropy is low. Concretely, with an exponential moving average of the (normalized) step-entropy,

$$\overline{H}_t^{(i)} = \left(1 - \tfrac{1}{w}\right) \overline{H}_{t-1}^{(i)} + \tfrac{1}{w}\, H\big(p_t^{(i)}\big), \qquad \overline{H}_0^{(i)} = 0,$$

the commit score is

$$\mathrm{stab}_t^{(i)} \;=\; \mathbf{1}\Big[a_t^{(i)} = a_{t-1}^{(i)}\Big] \cdot \Big(1 - \overline{H}_t^{(i)} / \ln|V|\Big),$$

with window $w{=}4$, and $\mathrm{stab}_1^{(i)}$ defined as the confidence score, since $a_0^{(i)}$ does not exist. Two properties of this rule are worth stating, because they bear on the null result it produces. The moving average is *zero-initialized and not bias-corrected*, so $\overline{H}_t^{(i)}$ reaches only $1-(1-1/w)^t$ of its steady-state value—0.25 at $t{=}1$, 0.68 at $t{=}4$, 0.90 at $t{=}8$—which *inflates* stab early and therefore biases TCR toward committing in the first few steps. Since commitment is already early under the default sampler, this works against TCR rather than for it, and a bias-corrected variant (dividing by $1-(1-1/w)^t$) is untested; we flag it as a confound on the null rather than claiming the null is robust to it. Here $p_t^{(i)}$ and $a_t^{(i)}$ are the step-t distribution and argmax at masked position i. The sampler thus commits the highest-scoring positions each step, keeping flickering or high-entropy positions masked for further refinement. TCR is training-free and adds no forward passes. Unlike post-hoc consistency voting (which only re-reads the produced trajectory), TCR changes *which tokens are committed during sampling*, so it can in principle change the output. Whether this helps accuracy is an empirical question we answer in §5.7.

How does the free trajectory signal compare to the uncertainty methods a reviewer would reach for first? On the same POPE examples and the same LLaDA-V model, we run three standard baselines. The first is *semantic entropy* (Kuhn et al., 2023; Farquhar et al., 2024), which clusters M sampled answers by meaning and takes the entropy over clusters. The second is *self-consistency* (Wang et al., 2023b), one minus the majority-answer fraction over the M samples. The third is *P(True)* self-evaluation (Kadavath et al., 2022), in which the model is re-asked whether its own answer is correct. We report both scorings. The *binary* greedy self-verdict is a 0/1 indicator—one ROC operating point, and on this pool it ties 98% of example pairs, so it cannot rank. Scoring the same self-evaluation *continuously*, by $p(\text{no})/(p(\text{yes})+p(\text{no}))$ at the answer slot, removes that handicap and is worth $+0.076$ AUROC ($0.640 \to 0.716$, paired re-run on the same 125 examples; the binary value reproduces the earlier capture's 0.645). The conclusion is unchanged: against the continuous variant the trajectory remains ahead but statistically tied (paired $[-0.18, +0.05]$). Samples for these baselines are drawn with per-step Gumbel noise at temperature 1.0, since token-level temperature is weak in masked-diffusion decoding. Sampling runs over $T{=}16$ steps, which is $2\times$ cheaper per sample than the main decode. Even so, $M{=}3$ samples add 48 denoising steps on top of the decode the model already runs—$1.5\times$ the cost of that decode—whereas our signal adds none. Sampling runs under the model's default confidence-greedy schedule—the deployment-relevant setting; a randomized schedule changes answers (accuracy $0.847 \to 0.820$, Appendix B.10) and could yield more inter-sample diversity, an untested variant. We quantify the handicap rather than leave it implicit: at $M{=}3$ the disagreement scores take only *two* distinct values and tie 82.5% of all example pairs; at $M{=}10$ they take seven and tie 79.1%. Midranked ties pull a rank statistic toward 0.5 mechanically, and indeed more samples buy resolution and accuracy together (semantic entropy $0.524 \to 0.597$, P(True) $0.645 \to 0.706$). Our trajectory statistic is continuous on the same pool (no ties) at no extra passes; the comparison is therefore between a score whose resolution is bought with compute and one whose resolution is free, not between a good and a bad idea. P(True) re-asks the model "Question: … Proposed answer: … Is the proposed answer correct? Answer yes or no." at temperature 0. One property to note: on binary answers semantic entropy and self-consistency are monotone transforms of the same majority count, so they coincide at *any* M (Table 9). We report $M{=}3$ in the main table as a cost decision and then verify below that the collapse is *not* a low-sample artifact with an $M{=}10$ rerun (the semantic-entropy literature's regime). Table 9 reports the results.

Two things stand out. First, our free single-pass signal **decisively beats the disagreement-based baselines**—by 0.26 AUROC at $M{=}3$ and, in the $M{=}10$ rerun, by 0.18 (paired bootstrap $[0.08, 0.29]$, excluding zero)—at a fraction of the cost ($M{+}1$ passes). It *matches* the strongest baseline, P(True): ahead by 0.14 at $M{=}3$ but statistically *tied* at $M{=}10$ (0.78 vs. 0.71; paired $[-0.05, 0.19]$). P(True) also costs an extra forward pass, and it is itself a confidence-family self-report whose reliability we show is model-specific (§5.3). We therefore claim only that the trajectory *equals or exceeds* P(True) at zero extra cost, while dominating the sampling baselines. Second, and more diagnostically, **the disagreement-based methods collapse toward chance and stay there as M grows** (0.525 at $M{=}3$; 0.60 $[0.51, 0.70]$ at $M{=}10$, so this is not a low-sample artifact). Semantic entropy and self-consistency quantify uncertainty through *disagreement across stochastic samples*, but masked-diffusion decoding yields near-identical samples. Their inter-sample entropy is therefore ≈ 0 regardless of correctness; we trace this to the sampler's aggressive commitment in §6. The message is practical: *the disagreement-based UQ toolkit built for AR LLMs does not transfer to diffusion VLMs*, whereas the trajectory—an artifact these baselines never observe—does. Bootstrap 95% CIs ($n{=}125$): ours 0.780 [0.685, 0.868]; semantic entropy 0.525 [0.454, 0.609] ($M{=}3$) rising only to 0.597 [0.505, 0.700] ($M{=}10$). Records: `baselines_uq_m10_analysis.json`.

A.8 Metrics and compute

Detection uses rank-based AUROC (equivalent to the Mann–Whitney statistic), computed by an exact rank-sum formula rather than trapezoidal integration for numerical stability. Uncertainty on every AUROC is reported as a percentile bootstrap 95% confidence interval (10,000 resamples over examples; draws with a single class are dropped). Table-wide CIs are per-cell and uncorrected; the two headline paired contrasts survive a Bonferroni check over all 84 model×benchmark×signal cells ($z{=}6.2/8.2$), and the risk-control grid uses fixed-sequence testing, which needs no multiplicity correction. All detector captures decode greedily in a single block (gen-length 32; 64 for captioning), and block-length effects remain untested. Sampled (non-greedy) decoding of the detector is tested at $\tau{=}0.7$ (Appendix B.4: 0.736 vs. 0.756 greedy, confidence still at chance). **SEISMO-Guard protocol, exactly:** each of the 200 runs draws a random 50/50 calibration/test split. Candidate keep-fractions form a coarse grid (10%, 15%, …, 100% of the calibration set, most-reliable first). Thresholds are tested in fixed-sequence order (no multiplicity correction needed), accepting while the Clopper–Pearson upper bound ($\delta{=}0.1$) on kept-error stays $\leq \alpha$ and stopping at the first failure. The last accepted threshold is applied to the test half. Table 2's cells reproduce under exactly this specification from the released records. For reference, at POPE's 17% base error the *unselected* model is itself certifiable at $\alpha{\approx}0.20$ (exact Clopper–Pearson at full coverage: 0.197). Selection therefore earns its keep at the stricter budgets ($\alpha{\leq}0.15$)—exactly where the trajectory's coverage advantage concentrates and confidence's collapses.

A.9 Reproducibility details

This appendix expands the Reproducibility Statement in the main text. All experiments use publicly released checkpoints (LLaDA-V; Qwen2-VL-2B/7B; LaViDa-LLaDA and LaViDa-Dream; MMaDA-8B-MixCoT; LLaDA-MedV—exact IDs in Appendix A) and public benchmarks (GQA, POPE, MME, HallusionBench, VQA-RAD, SLAKE, and COCO `val2014` for CHAIR captioning). Our signals are computed by a single forward hook on the output head and, for the grounding analysis, a hook on the attention weights; no training or fine-tuning of the base model is required. Algorithm 1 fully specifies the trajectory statistics; the exact prompts, scoring rules (dataset-aware yes/no and first-word matching), denoising settings ($T{=}32$, gen-length 32), and per-benchmark sample sizes are given in the appendix. All checkpoints and benchmarks are used for non-commercial academic research, consistent with their public licenses and intended use. Experiments are single-pass inference: one forward hook, no base-model training. The sole exception is one diffusion-DPO LoRA run reported as a negative result. All runs use a single 48 GB GPU (RTX A6000/A40/L40S), totaling on the order of tens of GPU-hours. Package versions (PyTorch 2.4/CUDA 12.4, `transformers`, `numpy`/`scipy`/`scikit-learn`) are pinned in the released script headers. All analysis code and the raw per-example trajectory records (the per-example signal values and correctness labels behind every table and figure) are served at the anonymous companion site (`https://seismo-demo.pages.dev`). Every number in the paper can therefore be recomputed at review time.

Code and data availability. The anonymous companion site (`https://seismo-demo.pages.dev`) serves, in addition to the figures and tables of this paper, the complete analysis code and the complete set of per-example record files. The code is the 35 experiment and analysis scripts, exactly as submitted, grouped by what they produce: trajectory capture and the seven signal definitions; detection and the sampling-based, self-evaluation and learned-probe baselines; calibration, selective prediction and the Learn-then-Test gate; the acting experiments (TCR, remasking controls, preference training), all of which are reported as negatives; and the logit-lens and attention analyses behind the root-cause split. The plotting scripts that only redraw the figures from these records are not included. The records are the 99 released files (97 JSON, one JSONL and one NPZ) holding, for every pool, the per-example signal values and the correctness label behind each number we report. No GPU, no model download and no credentials are needed to recompute the detection results: the site shows a fifteen-line script that fetches one record file and reproduces both AUROCs of Table 1 (0.778 for mean step-entropy, 0.504 for output confidence) in a few seconds. Two conventions matter when recomputing: `final_maxprob` is stored *negated* in the records, since it is already an error score, and AUROC is computed in the predict-wrong orientation with ties midranked.

B Detection: Extended Results

B.1 The autoregressive mirror

Why is LLaDA-V's anti-calibration striking? Because for AR VLMs the picture is reversed. On Qwen2-VL (Table 4, same benchmarks and scoring), **output confidence works** (`final_entropy` 0.736 GQA / 0.771 POPE) and the best internal signal is slightly *worse* (0.712/0.746). That is the mirror image of LLaDA-V, where confidence is at/below chance (0.385/0.476) and the trajectory is strong (0.723/0.756). A calibration method tuned on AR VLMs can thus silently fail on a diffusion VLM like LLaDA-V. The contrast survives a matched-scale check. Qwen2-VL-*7B* on the shared first-n pools (n=400 POPE / 300 GQA, greedy decoding, same scoring) again has working output confidence (`final_entropy` 0.690 POPE / 0.792 GQA). The mirror is thus paradigm-linked, not a 2B-vs-8B scale artifact, though it remains a cross-recipe comparison rather than a controlled ablation. As §5.3 shows, however, this inversion is model-specific—not every diffusion VLM has broken confidence—which is why a signal that works *regardless* matters.

Table 4: Mirror-image calibration: AUROC for predicting a wrong answer. AR confidence works (internal slightly worse); diffusion confidence fails (trajectory works). AR: Qwen2-VL-2B (n=600) and Qwen2-VL-7B (n=700, matched-scale check); diffusion: LLaDA-V (n=300); same benchmarks and scoring. The AR-2B row uses the n=600 mirror pool (summary released); on the shared first-n pools whose per-example records we release the AR-2B confidence AUROCs are 0.72 (POPE, n=400) / 0.76 (GQA, n=300)—the same qualitative mirror.

Paradigm	Signal	GQA	POPE
AR (Qwen2-VL-2B)	output confidence (entropy)	**0.736**	**0.771**
AR (Qwen2-VL-2B)	best internal (disagreement)	0.712	0.746
AR (Qwen2-VL-7B)	output confidence (entropy)	**0.792**	**0.690**
Diffusion (LLaDA-V)	output confidence (entropy)	0.385	0.476
Diffusion (LLaDA-V)	trajectory (mean step-entropy)	**0.723**	**0.756**

B.2 Cross-model generality

Is this specific to LLaDA-V? We repeat the experiment on three further diffusion VLMs, with 95% bootstrap CIs (Table 5). These are LaViDa (Li et al., 2025) with LLaDA-8B and *Dream-7B* backbones (a distinct diffusion family), and MMaDA (Yang et al., 2025), a third recipe: discrete MAGVIT image *tokens*, no vision tower; LLaDA-8B-initialized. MMaDA reproduces *both* regimes in one model: near-one-step on POPE (decoupling 0.03; confidence slightly ahead, paired Δ=-0.057 $[-0.088, -0.026]$), dynamic on GQA (0.46, $+0.11$). Two things generalize and one does not. **(1) The trajectory detector generalizes**: mean step-entropy predicts errors above chance in every cell (0.65–0.79 over the six general-domain cross-model cells; 0.59–0.79 including medical). **(2) Output confidence does not**: it is at-or-below chance on LLaDA-V (0.37–0.50; below chance

on GQA) but well above chance on LaViDa (0.60–0.77)—its reliability is model-specific. **(3) On the LLaDA-initialized *dynamic* cells the trajectory is never worse than confidence, and it beats it wherever confidence is broken.** On the near-one-step-commit cells the trajectory instead *coincides* with confidence (LaViDa-Dream GQA is a statistical tie, 0.75 vs. 0.77) or trails it slightly (MMaDA-POPE 0.65 vs. 0.71, paired $\Delta{=}-0.057$ $[-0.088, -0.026]$). MMaDA-POPE is the one general-domain cell where plain confidence wins outright, and it is exactly the low-decoupling regime the regularity predicts. We verified (Appendix B.9) that when the answer position commits *early and near-deterministically* the step-entropy statistic mathematically *coincides* with output confidence—hence the near-identical CIs where confidence works (comparisons via paired bootstrap on shared items throughout). Where decoding stays dynamic, the trajectory instead *diverges from and beats* confidence. On LLaDA-V (both tasks, confidence broken) it leads by up to $+0.34$, and on LaViDa-LLaDA GQA by $+0.19$ (0.79 vs. 0.60). That GQA cell's longer answers keep *position 0 itself* uncommitted for more steps; the signal is read at the first answer position throughout (Appendix A). So the trajectory *equals* confidence when the answer is decided in one step, and *improves* on it whenever there is a denoising trajectory to read. Confidence alone, by contrast, is a model-specific gamble (anti-calibrated on LLaDA-V, usable on LaViDa). This is quantitative, not anecdotal. Across all *twelve* model×benchmark cells, the trajectory–endpoint rank-decoupling ($1 - \rho_{\text{Spearman}}$) predicts the trajectory's AUROC advantage at $r{=}0.89$ (Spearman 0.84), spanning $0.004/{+}0.001$ (LaViDa POPE) to $0.94/{+}0.34$ (LLaDA-V GQA). The fit holds *out-of-sample*: MMaDA's two cells, measured after the fit, land on the line. All cells are listed in Appendix H. The striking anti-calibration "mirror" is thus specific to *some* diffusion VLMs; the robust default is model-independent. One caveat: only the entropy statistic generalizes—the discrete flicker signals (`n_changes`, `commit_step`) require argmax flicker and so are informative only under dynamic decoding (§6). **Benchmark breadth.** On LLaDA-V the detector also beats confidence on MME (0.60 vs. 0.56) and HallusionBench (0.65 vs. 0.52)—five benchmarks in all (Appendix B.3).

Table 5: Cross-model generality: AUROC for predicting a wrong answer (95% bootstrap CI; POPE $n{=}400$, GQA $n{=}300$; higher of the two scores per row in **bold**, ties unbolded). The trajectory detector (mean step-entropy) is strong on all three diffusion VLMs and both tasks; output confidence is anti-calibrated on LLaDA-V but works on LaViDa. The trajectory ties confidence where confidence works and beats it where it fails. LLaDA-V GQA CIs come from an exact replication of that capture (Appendix H). On one-step-commit models `conf_at_commit` is numerically degenerate (saturated near zero, e.g. mean $4{\times}10^{-4}$ on Dream), so its at-chance values there are noise, not signal.

Model (diffusion backbone)	Bench	Output confidence	Trajectory (mean step-ent.)
LLaDA-V (LLaDA-8B)	POPE	0.50 [0.43, 0.57]	**0.78 [0.71, 0.84]**
LLaDA-V (LLaDA-8B)	GQA	0.39 [0.32, 0.45]	**0.72 [0.66, 0.78]**
LaViDa (LLaDA-8B)	POPE	0.75 [0.68, 0.81]	0.75 [0.68, 0.81]
LaViDa (LLaDA-8B)	GQA	0.60 [0.53, 0.67]	**0.79 [0.73, 0.84]**
LaViDa (Dream-7B)	POPE	**0.75 [0.69, 0.80]**	0.74 [0.69, 0.80]
LaViDa (Dream-7B)	GQA	**0.77 [0.71, 0.82]**	0.75 [0.69, 0.80]
MMaDA (LLaDA-8B init)	POPE	**0.71 [0.65, 0.77]**	0.65 [0.59, 0.71]
MMaDA (LLaDA-8B init)	GQA	0.60 [0.54, 0.67]	**0.71 [0.66, 0.77]**

B.3 Benchmark breadth

Beyond POPE and GQA we run the trajectory detector on three further hallucination benchmarks with the same trajectory capture and dataset-aware scoring (Table 6). On all five, mean step-entropy predicts errors above chance. Paired bootstrap deltas over the same examples certify the lead on HallusionBench ($+0.12$ $[+0.04, +0.21]$), while MME is positive but not significant ($+0.05$ $[-0.01, +0.11]$). On the harder MME and HallusionBench, output confidence is near chance while the trajectory retains signal; on AMBER it falls *below* chance (0.47), reproducing the anti-calibration result outside POPE and GQA. AMBER also settles whether the signal is specific to object *existence*, which is all POPE tests: across its three discriminative types the detector reaches 0.787 on *relation* questions, 0.722 on *existence* and 0.612 on *attribute* (accuracies 0.669/0.865/0.609), so it is strongest on exactly the relational questions POPE never asks. We confirmed on each that the capture is a genuine multi-step trajectory (per-example correlation between mean step-entropy and final-step entropy ≈ 0.5), not a collapse to confidence.

Table 6: Benchmark breadth on LLaDA-V: AUROC for predicting a wrong answer (95% bootstrap CI, 10,000 resamples). AMBER (Wang et al., 2023a) is sampled *stratified* over its three discriminative types (133 each: existence, attribute, relation), not as a split prefix. It is the hardest cell (accuracy 0.714 vs. 0.830 on POPE) and the only one where output confidence falls *below* chance, replicating the anti-calibration result on a benchmark used nowhere else in the paper.

Benchmark	n	Output confidence	Trajectory (mean step-ent.)
POPE (object presence)	400	0.50 [0.43, 0.57]	**0.78 [0.71, 0.84]**
GQA (compositional/two-hop)	300	0.39 [0.32, 0.45]	**0.72 [0.66, 0.78]**
MME (perception + cognition)	400	0.56 [0.49, 0.62]	**0.60 [0.55, 0.66]**
HallusionBench	225	0.52 [0.43, 0.61]	**0.65 [0.56, 0.73]**
AMBER (exist./attr./relation)	399	0.47 [0.40, 0.54]	**0.66 [0.60, 0.71]**

B.4 Robustness of the headline detector

Three checks probe whether the LLaDA-V detection gap is an artifact of a single evaluation pool, a single prompt, or greedy decoding. **(i) Disjoint second pool.** On an entirely disjoint POPE-adversarial pool (items 1200–1799, n=600, base error 0.173) the gap reproduces: mean step-entropy **0.756** vs. output confidence **0.530** (cf. the original n=1200 pool, 0.779 vs. 0.509). **(ii) Prompt phrasing.** Across three prompt templates (n=300 each) output confidence stays broken while the trajectory holds: bare prompt (confidence 0.418 / trajectory 0.743), the paper's prompt (0.482 / 0.756), and an explicit "answer yes or no" suffix (0.503 / 0.773). Anti-calibration is worst with no instruction; it is not a prompt artifact. **(iii) Sampled decoding.** Under temperature-sampled remasking (τ=0.7, n=288) the detector degrades only mildly and does not collapse: trajectory **0.736** (vs. 0.756 greedy on matched items), while output confidence remains at chance (0.515). Consistent with the mechanism—step-1 entropy alone detects at 0.70 before any commit—the trajectory signal does not depend on greedy commitment.

B.5 Trace-detector head-to-head (learned trace baselines)

We re-implement the two learned trace-detector families proposed for text diffusion LMs on our released LLaDA-V POPE captures. The first is a *supervised trace probe* (TraceDet-style: 5-fold cross-validated logistic regression on the length-normalized per-step answer-position entropy trace). The second is a *supervised dynamics probe* (DynHD-style: cross-validated logistic regression over hand-crafted denoising-dynamics scalars). We evaluate four n=400 captures spanning three step budgets (the two T=32 captures are independent runs over the same items). Only entropy-trace features are reproducible from the released captures; features requiring per-step token identities or top-k logits are out of scope and noted as such. Table 7 shows the result. In-domain, the learned probes sit within cross-validation noise of our zero-shot mean step-entropy. They nonetheless (i) require a labeled training set and (ii) *invert* across models (0.25–0.48; §5.3, transfer analysis) where the zero-shot scalar still travels. The trained probe therefore buys nothing here and forfeits the portability that is the point of a training-free detector.

B.6 Sampling-based and white-box baselines

A training-free internal-signal baseline (VAUQ-style). Sampling and self-evaluation are not the only training-free options: Park et al. (2026) scores an answer by how much the *image* reduces predictive uncertainty (an image-information score, IS). We implement that backbone—run each question with the real image and with a blank one, and take the uncertainty drop—on two pools (Table 8). We do not implement VAUQ's unsupervised core-region masking, so this is VAUQ-*style*, not VAUQ. The result sharpens our thesis rather than competing with it: the image-information idea works when the uncertainty is read over the *trajectory* (0.76/0.65) and collapses to chance when the same quantity is read at the *endpoint* (0.49/0.53)—on both datasets. A competing method's own score is informative only in the place this paper argues the information lives. Note also that the trajectory variant contains our statistic by construction (it is mean step-entropy minus a blank-image term), so the tie with our signal is expected; the operative difference is cost, since the IS needs a second forward pass and ours needs none.

Table 7: Detection AUROC (predict wrong; POPE, LLaDA-V, n=400 per cell) for our zero-shot mean step-entropy vs. learned trace-detector baselines, across denoising-step budgets T. Best per column in **bold**. In-domain the learned probes only match the free zero-shot statistic ($\leq$0.04 either way, within CV noise); unlike it, they must be trained and do not transfer across models. The T=8 and T=16 columns are a fresh n=400 capture that independently reproduces the schedule trend of Table 10 (there n=200). The two T=32 columns are two n=400 captures of the *same* 400 items (the second is the main capture used in Table 1); they bound the run-to-run spread—0.768 vs. 0.778 zero-shot and 0.805 vs. 0.797 for the trained probe—which is the scale against which the probe-vs-scalar gap should be read.

Detector	T=8	T=16	T=32	T=32 (repl.)
Base error rate	0.265	0.210	0.168	0.170
`mean_step_entropy` (ours, zero-shot)	0.837	**0.815**	0.768	0.778
Supervised trace probe (TraceDet-style)	**0.856**	0.799	**0.805**	**0.797**
Supervised dynamics probe (DynHD-style)	0.817	0.785	0.752	0.778

Table 8: VAUQ-style image-information scoring vs. our trajectory statistic. AUROC for predicting a wrong answer (LLaDA-V). The same image-information quantity is informative read over the trajectory and at chance read at the endpoint. POPE here is a 130-example paired capture (accuracy 0.823, matching the 0.824 main pool); AMBER is the stratified 399-example pool of Table 6.

Score	POPE (n=130)	AMBER (n=399)
output confidence (stronger variant)	0.519	0.471
VAUQ-style IS, endpoint	0.493	0.530
VAUQ-style IS, trajectory	0.761	0.649
Ours (mean step-entropy)	**0.792**	**0.656**

A white-box comparator: the step-0 internal margin (and its VCD-style contrast). For completeness we also evaluate the strongest *white-box* alternative our own captures contain: the answer-position final-layer logit-lens margin from the *clean step-0 forward*, before any commit. We also evaluate its image-ablation contrast $|1.5\, m_{\text{full}} - 0.5\, m_{\text{blank}}|$ in the style of contrastive decoding (Leng et al., 2024). On the root-cause pool (n=600, POPE) these detect at 0.818 $[0.776, 0.857]$ and 0.821 $[0.781, 0.858]$—numerically above the trajectory's 0.778 on the main pool (CIs overlap). Three qualifications keep this a comparator rather than a replacement. It requires *intermediate-layer access* (white-box), its contrastive form requires a *second, image-ablated forward pass*, and the signed margin exists only for *binary* answers. A top-1$-$top-2 lens-margin analog on open-vocabulary GQA is definable but unmeasured, and the trajectory's largest advantage lives there. Notably, the contrast term adds essentially nothing over the plain internal margin (0.821 vs. 0.818). The comparison is itself evidence for our thesis: a *clean single-forward* internal read-out stays discriminative (0.82) while the same weights' *sampler-sharpened output* confidence sits at chance (0.50). The information is present before iterative decoding erases it, and the trajectory recovers most of it (0.78) for free, task-generally, and black-box (`vcd_baseline.json`).

B.7 Why training-free is necessary, not merely convenient

It is reasonable to expect that a *supervised* trace detector, trained on labelled errors, should beat our zero-shot statistic. We test this directly, and the answer is not that supervision fails in-domain—it is that it buys nothing and costs portability. *In-domain*, a TraceDet-style logistic probe on the full entropy trace tracks the zero-shot scalar to within $\pm$0.04 across four step budgets (0.805 vs. 0.768 at T=32; 0.799 vs. 0.815 at T=16; Table 7). That is inside cross-validation noise—for the price of a labeled training set per model. *Out of domain* it is worse than useless: a seven-signal probe trained on LLaDA-V POPE **inverts** when transferred across models. It reaches 0.25–0.48 AUROC over the four LaViDa cells, below chance in three of four, versus 0.74–0.79 for the zero-shot statistic on the same targets. The individual shape statistics (variance, slope, curvature, monotonicity of the entropy trace)—the feature family on which text-dLLM trace detectors build—are each $\leq$0.52 zero-shot on our captures. Supervised trajectory probes overfit one decoder's signal polarity—which we showed

Table 9: Uncertainty baselines on LLaDA-V (POPE, n=125, M=3 samples). Our single forward-pass trajectory statistic dominates the disagreement-based baselines, which collapse toward chance because masked-diffusion sampling produces near-identical answers—the answer *diversity* these methods depend on is largely absent—and stay collapsed at M=10 (text below). P(True) is reported under both scorings, since a 0/1 verdict cannot rank; against its stronger continuous form the trajectory is ahead but tied. AUROC for predicting a wrong answer.

Method	AUROC (predict wrong)	Cost
Semantic entropy (Kuhn et al., 2023)	0.525	M samples + clustering
Self-consistency (Wang et al., 2023b)	0.525	M samples
P(True) self-eval, binary (Kadavath et al., 2022)	0.640	+1 pass
same, continuous scoring	0.716	+1 pass
Ours (mean step-entropy)	**0.780**	1 pass (free)

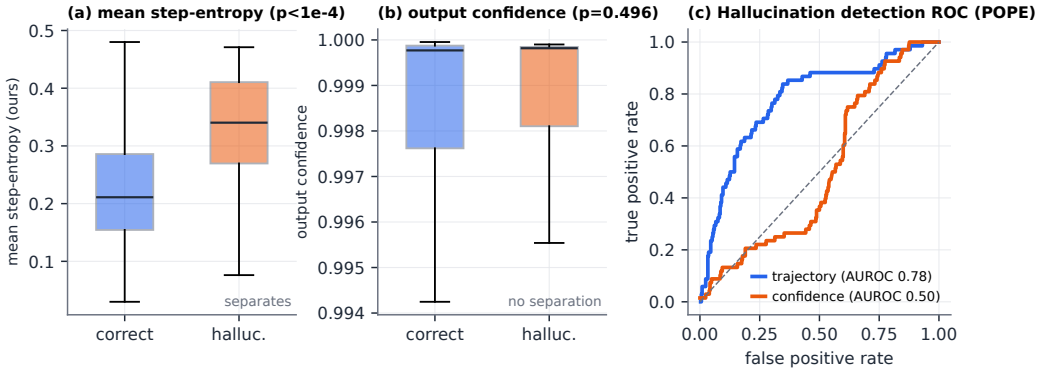

Figure 6: **Detection on POPE (LLaDA-V, n=400). (a)** Mean step-entropy separates correct from hallucinated answers (permutation $p<10^{-4}$); **(b)** output confidence does not (p=0.50; note the y-range, 0.994–1.000, and that outliers are suppressed—confidence is saturated, not merely uninformative); **(c)** ROC: the trajectory statistic reaches 0.78 AUROC; confidence sits at chance.

is decoding-dependent (§5.3)—and actively mislead on the next model, while the un-trained mean step-entropy transfers everywhere. This mirrors the supervised text-dLLM detectors (Chang et al., 2026; Qian et al., 2026), which are trained and evaluated per-model; in the multimodal setting we find the training-free statistic is not a cheap approximation to a trained probe but the more robust object.

B.8 Distribution of the signal and ROC

B.9 Trajectory-capture verification (cross-model)

To rule out a measurement artifact behind the cross-model results, we logged the full per-step, per-position answer trajectory on LaViDa. The LM-head hook fires once per denoising step (T=32 confirmed), and the number of still-masked answer positions decreases monotonically (32 $\rightarrow$ 2), so the trajectory is fully observed—not truncated. However, the per-position predictive entropy is near-zero throughout (0.0–0.24) and the answer's argmax never flips: *LaViDa commits the answer early and near-deterministically*. Consequently the mean step-entropy *coincides* with the final-step (output-confidence) entropy on LaViDa *on this POPE capture* (per-example correlation 0.99, mean $|\Delta| < 0.013$; the correlation is benchmark-dependent, see below). On LLaDA-V, by contrast, the two differ substantially (correlation 0.34, mean $|\Delta|$=0.25), because its low-confidence remasking keeps the answer in flux. The same check on LaViDa-LLaDA *GQA*—the one LaViDa cell where the trajectory beats confidence by +0.19—shows correlation 0.46 (mean $|\Delta|$=0.65): decoding stays dynamic exactly where the advantage appears, completing the pattern. This confirms that the trajectory statistic *reduces to* output confidence exactly when decoding is deterministic, and carries additional information only when it is not. That is precisely the pattern in Table 5, and evidence that those numbers are real rather than a capture artifact.

B.10 DENOISING STEPS: WHERE IN THE TRAJECTORY THE SIGNAL LIVES

Where in the trajectory the signal lives (pre/post-commit decomposition). Averaging over all T steps mixes genuine pre-commit predictions with post-commit readouts (the model re-reading its own committed token). Decomposing on the POPE capture (per-step entropy matrix released): **post-commit readouts alone are at chance** (0.48)—they neither drive nor hurt the signal—while **step-1 entropy alone reaches** 0.70 (before any commit exists), the first-2-step mean 0.70, and the first-8-step mean 0.77, indistinguishable from the full mean (0.768). (A window-conditioned "pre-commit mean" scores only 0.53, an artifact of conditioning the averaging window on commit timing; fixed windows are the clean analysis.) The doubt is in the *early* steps, exactly as the mechanism account says. This also answers "is the mean just commit timing repackaged?": within the zero-flicker stratum (`commit_step`$=0$, n=199, where all timing signals are identically zero), mean step-entropy still detects at 0.60 (0.73 in the dynamic stratum). Truncating the T=32 capture to its first 8 steps preserves the detector (0.772)—so deployments can read early and decide early—though the T=8-*schedule* peak of 0.841 (Table 10) reflects genuinely coarser denoising, not mere truncation.

Schedule control: random remasking. The direct test of the sampler-sharpening mechanism is to change nothing but the commit rule. On the same POPE n=300 pool, the default confidence-greedy schedule gives accuracy 0.847 with confidence AUROC 0.482/0.476 (max-probability/entropy) and trajectory 0.756—an exact replication of the original capture. Switching to *random* remasking gives accuracy 0.820 (a 2.7-point cost) with confidence AUROC 0.560/0.560 and trajectory 0.764: removing confidence-greedy commitment *partially restores* the endpoint's informativeness while leaving the trajectory unchanged. The restoration is partial (0.56, not 0.76), consistent with the residual between-answer-class anti-calibration documented in Appendix H; the control ran on POPE, and its GQA analog is untested. Per-example records released.

Table 10: Denoising-steps ablation (LLaDA-V, POPE; n=200 per cell except T=32, the main n=400 capture). As steps increase, answer accuracy rises but the trajectory error-detector weakens—the error signal is strongest at coarse/early denoising. An independent n=400 re-capture at $T \in \{8, 16\}$ (Table 7) reproduces the trend: 0.837/0.815 there vs. 0.841/0.809 here.

Denoising steps T	4	8	16	32	64
Answer accuracy	0.650	0.745	0.790	0.833	**0.840**
Trajectory AUROC (predict wrong)	0.816	**0.841**	0.809	0.768	0.753

The same control on an independent backbone (Dream). We next ask whether the trajectory *advantage* is a property of dynamic decoding in general or of the LLaDA-family's uncertainty encoding specifically. To test this, we ran the identical random-remasking control on LaViDa's *Dream-7B* backbone, a different diffusion family (POPE n=400, GQA n=300; `dreamdyn_generality_result.json` released). Random remasking raised Dream's trajectory–endpoint decoupling on POPE from 0.06 (native) to 0.51. Yet the trajectory did *not* become a detector: POPE trajectory 0.53 vs. confidence 0.58 (gap -0.05, paired $[-0.13, +0.04]$); GQA 0.67 vs. 0.68 (gap -0.01). Both are at or near chance on POPE, and neither signal leads on either. The contrast with LLaDA-V is the point: under the *same* random schedule, LLaDA-V's trajectory stays robust (0.76) while Dream's collapses to chance. So the advantage is not produced by decode-order decoupling *per se*. It requires the model to *encode* its uncertainty in the evolving answer-position representation across steps, a property of the LLaDA-initialized recipe that the Dream backbone does not share. That is precisely why forcing decoupling through a random schedule—rather than the native confidence-ordered dynamics the regularity measures—does not carry the signal. This scopes the trajectory *advantage* to LLaDA-family diffusion VLMs—the *absolute* detector still generalizes (§5.3)—and explains, mechanistically, why every dynamic-advantage cell in our study is LLaDA-initialized.

Is the anti-calibration a property of the learned distribution or of the *iterative sampling*? We vary the number of denoising steps T and measure both answer accuracy and the trajectory detector's AUROC (LLaDA-V, POPE; n=200 per point except T=32 at n=400). Two clean trends (Table 10): as T grows, accuracy rises monotonically while the trajectory error-signal peaks at a coarse T=8 and then

weakens (the T=8 vs. T=4 gap is within noise at n=200). The reliability information therefore lives in the early, coarse denoising dynamics. Extra refinement improves the answer but smooths away the trajectory's discriminativeness, localizing the phenomenon to the iterative remasking process. The practical corollary is that for *error detection* fewer steps are better, even though more steps give better *answers*.

Where the miscalibration comes from. A denoising-steps ablation (Appendix B.10, Table 10) localizes the effect to the *iterative sampling*: as steps grow, accuracy rises ($0.65 \rightarrow 0.84$) while the trajectory detector weakens, peaking at a coarse T=8 (0.84) and falling to 0.75 at T=64. The reliability signal therefore lives in the early, coarse dynamics. For detection fewer steps are better, even though answering wants more.

C Calibration, Selective Prediction, and Risk Control

C.1 Calibration

Calibration. The trajectory not only *ranks* errors but can be *calibrated* into a trustworthy probability. A lightweight logistic head over the five trajectory features (5-fold cross-validated, POPE n=400) yields an expected calibration error of **ECE** $= 0.063$ (15-bin) and a Brier score of 0.134. Raw output confidence ($p_{\text{wrong}}=1-\max_v p_T$) gives **ECE** $= 0.169$ and Brier 0.170. That is a 2.7× reduction in calibration error. Two controls pin down where the gain comes from. Giving confidence the *same* supervision (Platt/temperature scaling, 5-fold) yields Brier 0.144—*worse than the constant base-rate predictor* (0.141)—with resolution ≈0.001: perfectly calibrated ignorance, because confidence carries no ranking information to recalibrate. The trajectory head's Murphy decomposition shows genuine resolution (0.017 of a maximum 0.141; reliability 0.010), so its Brier advantage is information, not supervision. The single `mean_step_entropy` statistic remains the strongest pure *detector* (0.778 AUROC, vs. 0.724 for the multi-feature head); the head's contribution is not sharper ranking but a usable, well-calibrated wrong-*probability*. A diffusion VLM can therefore be made to report trustworthy hallucination probabilities at the cost of a five-feature logistic head—no model retraining. Cal and Guard are *per-model* components fit on the target model's own captures—our transfer result shows why (trained heads over these features invert across models; Appendix B.7); only the zero-shot Score travels.

C.2 Selective prediction and risk–coverage

Table 11: Selective prediction on POPE (n=400, base error 17%). AURC and error rate among answered examples at 50%/80% coverage. Lower is better.

Reject-option score	AURC	risk@50%cov	risk@80%cov
output confidence (entropy)	0.141	0.210	0.169
`commit_step` (trajectory)	0.088	0.070	0.109
`n_changes` (trajectory)	0.089	0.070	0.113
`mean_step_entropy` (trajectory)	**0.072**	**0.050**	**0.106**

C.3 Learn-then-Test risk control

C.4 Sensitivity to the two knobs (α, δ)

Sensitivity to the two knobs. The guarantee has exactly two parameters—the risk budget α and the confidence level δ. Figure 8 sweeps both with the *same* fixed-sequence protocol as Table 2 (LLaDA-V POPE n=400, 200 splits per cell). **(a)** The surface has the shape finite-sample certification must have. There is a *certification floor* at the most stringent settings ($\alpha \leq 0.10$ at small δ). There, the first grid point's Clopper–Pearson bound already exceeds α on an n=200 calibration half, so the system certifies nothing and answers nothing—coverage 0, never a violation. (α=0.10 becomes usable at n=1200, Table 2.) Above that floor, coverage rises monotonically in both knobs, with the trajectory (*ours*) dominating output confidence at every certified point. **(b)** Wherever anything is certified, the

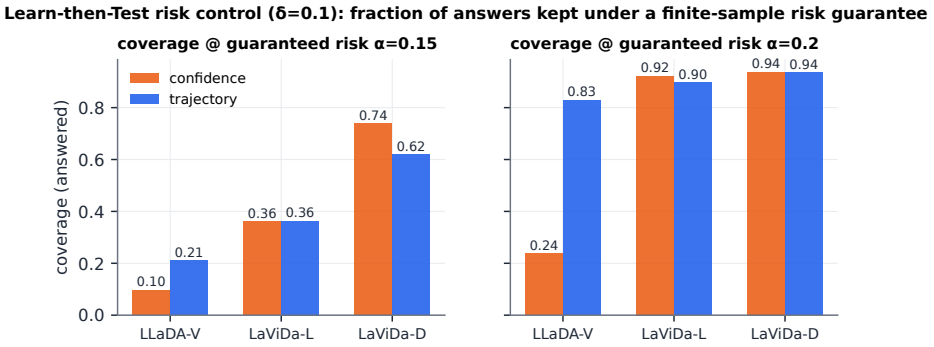

Figure 7: **Coverage under Learn-then-Test risk control** (δ=0.1; POPE). Bars show the fraction of questions each score can answer *with a finite-sample risk guarantee*. On the anti-calibrated LLaDA-V the trajectory covers 2.2–3.5$\times$ more than standard confidence; on LaViDa the two match within a few points, except the Dream cell at α=0.15, where confidence leads (0.74 vs. 0.62).

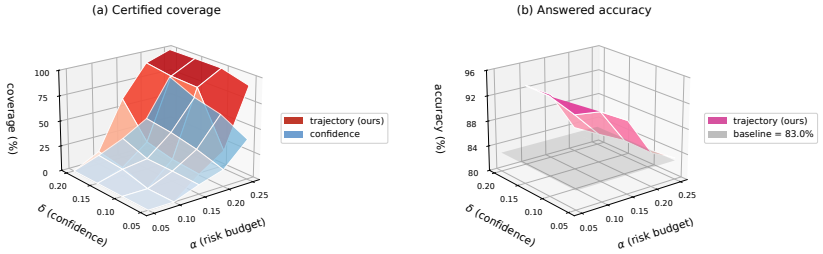

Figure 8: **SEISMO-Guard across its two knobs** (α risk budget, δ confidence level; LLaDA-V POPE n=400; 200 Learn-then-Test splits per cell, fixed-sequence protocol identical to Table 2; cells are independent Monte-Carlo draws, within sampling error). **(a)** Certified coverage: zero at the stringent-(α, δ) floor (nothing certifiable at this n—abstain-all, not a violation), then monotone in both knobs; the trajectory (*ours*) dominates *confidence* at every certified point. **(b)** Answered-set accuracy (shown where a set is certified) tracks at or above the unselected baseline (83.0%; the two α=0.25 cells sit 0.1 pp below it), converging to the baseline as coverage approaches 100%. Note the truncated z-axis in (b) (80–96%), which magnifies the accuracy gain, and that these are independent Monte-Carlo draws of the protocol in Table 2—the α=0.20, δ=0.10 cell reads $\approx$86% here against the table's 82.9%, a 1.6-SE difference at the coverage sd of 0.29 we report.

answered set's accuracy is at or above the unselected baseline (83.0%; converging to it as coverage approaches 100%), highest at stringent budgets. No cell violates its budget; the reported operating points sit on this monotone surface rather than at a cherry-picked spike. Surfaces recompute on CPU from the released records (`guard_sensitivity_grid.json`).

Cross-host threshold transfer: a threshold calibrated on the main n=400 capture and applied to the independent same-protocol rerun on a different GPU host keeps test error at 0.115 (α=0.15) and 0.168 (α=0.20). The guarantee therefore survives the hardware-drift scenario. We still recommend recalibrating on the serving stack, and prefer δ=0.05 with $n \geq 1200$ at deployment. Selective prediction uses risk–coverage and AURC; calibration uses ECE and Brier. All runs fit on one 48 GB GPU; a full n=400 detection pass takes roughly two hours. For every run we additionally dump the raw per-example signals and labels (`*_raw.json`), so all AUROC/AURC/ECE numbers can be recomputed offline without re-running the model.

C.5 RISK CONTROL UNDER DISTRIBUTION SHIFT, AND A DRIFT MONITOR

Two properties bound SEISMO-Guard's guarantee in deployment. **(1) Conservative transfer under category shift.** We calibrate the fixed-sequence Learn–Then–Test threshold on the released adversarial POPE pool (base error 0.183). We then deploy it *verbatim* on the popular and random POPE categories (base error 0.138 and 0.100) and on a disjoint adversarial pool (0.173). The certificate holds at every budget (Table 12): because the shifted categories are easier, the transferred threshold keeps kept-error at or below α with a safety margin. This is the benign direction of shift. **(2) A monitor for the unsafe direction.** A marginal, exchangeable guarantee fails *silently* if base error drifts *upward* after calibration. We therefore run a sequential e-process (a mixture betting supermartingale on the realized kept-error indicators, Ville threshold $1/\delta$) over the answered stream. On the in-distribution stream it does not alarm (final $\log_{10}$ e-value -5.3; kept-error 0.154 at α=0.15). Under simulated upward drift it alarms at rate 1.0 with a median of 47 answered examples to alarm when base error rises to 0.35, and at 0.465 (median 194) at the near-boundary 0.25. The false-alarm rate under in-distribution resampling is $0.045 \leq \delta$=0.10. Guard thus ships with both a conservative-transfer property and a drift detector; the monitor is ∼50 lines and runs on the released records.

Table 12: SEISMO-Guard threshold calibrated on adversarial POPE (n=1200) and deployed verbatim under category shift. Kept-error stays $\leq \alpha$ in every cell (certificate holds); coverage is the answered fraction.

α	Deploy category	base error	coverage	kept-error
0.10	popular	0.138	0.678	0.054
0.10	random	0.100	0.732	0.034
0.10	adversarial (disjoint)	0.173	0.665	0.093
0.15	popular	0.138	0.813	0.088
0.15	random	0.100	0.855	0.059
0.15	adversarial (disjoint)	0.173	0.820	0.120
0.20	popular	0.138	1.000	0.138
0.20	random	0.100	1.000	0.100
0.20	adversarial (disjoint)	0.173	1.000	0.173

Threshold transfer across step budgets. A deployment may change the denoising budget T after Guard is calibrated. We calibrate the fixed-sequence LTT threshold at T=32 (first-400 POPE pool) and deploy it *verbatim* at $T \in \{8, 16\}$. At α=0.10 and α=0.15 the transferred threshold keeps kept-error $\leq \alpha$ at *every* budget (kept-error 0.033–0.118), with coverage degrading gracefully toward small T (0.58 at T=8 vs. 0.85 matched at α=0.15). At α=0.20 it fails at T=8 (kept-error 0.263) for a transparent reason: the model's own base error at T=8 is $0.265 > \alpha$, so no transferred threshold could certify near-full coverage—changing the step budget changes the *model*. Deployment rule: recalibrate Guard at the deployment T (the same per-model calibration cost as §5.5).

Class-conditional certification. The marginal guarantee says nothing per answer class, and POPE's yes/no strata differ in base error (0.105 vs. 0.260 on the n=1200 pool). We run the identical LTT protocol *within* each stratum (n=600 each, α=0.15, δ=0.10). The *yes* stratum certifies coverage 0.992 ± 0.014 at kept-error 0.100, while the *no* stratum (base error 0.260) certifies essentially nothing—for the trajectory *and* for confidence. That stratum sits above the operating envelope: kept-error below α<0.26 is not certifiable at meaningful coverage for any score at this accuracy and n. The marginal certificate's coverage is therefore carried by the yes class. The abstention audit (kept-error $5.4\%/4.5\%$ per class at the marginal threshold) shows the kept sets remain class-balanced in *error*. Per-class *certification* of the hard stratum, however, requires a larger risk budget or more per-class calibration data. Records: `rebuttal_t_mismatch_guard.json`, `rebuttal_class_ltt.json`.

Table 13: TCR vs. the standard low-confidence sampler (LLaDA-V, matched budget). Accuracy ties on both benchmarks; on POPE, false-positive "yes" hallucinations fall but accuracy is unchanged. The POPE row ships per-example records (n=300); for the GQA row only the run summary was retained, so it is reported as a summary-level replication of the same null.

	GQA acc	POPE acc	POPE false-yes
low-confidence (baseline)	0.363	0.847	33
TCR (trajectory-stability)	0.360	0.843	25
Δ	-0.003	-0.003	-8

D Acting on the Signal

D.1 Trajectory-Calibrated Remasking: results

D.2 Detector-triggered routing

Detector-triggered routing: the diagnostic acts across systems. In-stack correction fails (§5.7), but the detector can *route*. On the aligned POPE pool (n=400; both models decoded greedily), sending the flagged fraction (highest mean step-entropy) from LLaDA-V to the Qwen2-VL fallback lifts system accuracy monotonically to a peak *above both endpoints*. Accuracy goes from 0.830 (LLaDA-V alone) $\rightarrow$ 0.860/0.868/0.868 at 10/20/30% routed, vs. 0.850 for always-Qwen2-VL (0.870–0.873 with the 7B fallback). The 20%-routed gain over LLaDA-V alone is +3.8 points (paired bootstrap $[+0.8, +7.0]$, excluding zero). Against *always* routing it is a statistical tie ($+1.7$ $[-1.3, +4.8]$). That is, the router matches full replacement while invoking the AR model on a fifth of queries. Two controls make this the detector's achievement. Routing the same fractions by output *confidence* is no better than never routing (0.828 at 20%, point-estimate *below* baseline—anti-calibration routes the wrong items). Random routing merely interpolates between the endpoints. On GQA, where the AR model dominates outright, trajectory routing recovers 36% of the full replacement gain at 30% routed (0.367$\rightarrow$0.443; ceiling 0.577). The operating fraction is a post-hoc choice on the same pool—we report the full sweep, and the rank ordering (no tuned threshold) does the work. Records: `routing_demo.json`; computed entirely from the released aligned per-example records.

D.3 Preference training (diffusion-DPO)

Full diffusion-DPO experiment. We pushed one step further and tested whether the failure is specific to *decode-time* intervention by moving to *training time*: to our knowledge the first hallucination-targeted preference training for a diffusion VLM. We mine on-policy preference pairs with the detector itself: we sample LLaDA-V on GQA-train yes/no questions, and when the answer is wrong we prefer the gold yes/no over the model's own. We then fine-tune with a diffusion-correct DPO, as a LoRA (rank 32, α=64, fp32 adapter on a bf16 base; AdamW, learning rate 5×10^{-5}, β_{DPO}=0.1, 2 epochs, gradient accumulation 8; 274 balanced pairs; loss decreases 0.69$\rightarrow$0.42). For single-token yes/no responses the masked-forward log-probability is *exact*, so no ELBO estimate is needed. Because the base model is yes-biased, the preferred side of the mined pairs skews toward "no". A threshold shift is therefore the *expected* failure mode. Observing it confirms—at this budget—that the adapter learned the prior rather than the perception. We scope the conclusion accordingly: it is a statement about detector-mined preference training at LoRA scale, not an impossibility claim for larger-budget training. The result is instructive: POPE accuracy rises 0.785$\rightarrow$0.825, but the prediction distribution flips from yes-biased (245:152; 3/400 unparseable) to no-biased (142:258), and **GQA does not improve** (0.544$\rightarrow$0.528)—the gain is a decision-threshold shift, not recovered grounding (Yin et al., 2025). Even training-time preference optimization, aimed directly at the model's own mined hallucinations, moves the answer *prior* rather than the visual *evidence*—exactly what the upstream under-encoding account predicts (§6).

D.4 What does not work

What does *not* work (negative results). Two negatives sharpen the contribution. (i) *Trajectory shape*: shape statistics of the entropy trajectory (monotonicity, bumps, slope, curvature)—which help for AR chain-of-thought—do not transfer to denoising (all $\leq$0.52 AUROC vs. 0.77 for the scalar). (ii) *Visual grounding*: the entropy of the answer's cross-attention over image patches is at chance (0.48–0.59); only a weak early-layer image-attention-fraction cue emerges (layer 8, AUROC 0.66), still below the token signal. The diffusion error signal is carried by *token-level* temporal dynamics, not visual attention (Figure 9 shows this per-head on the case-study pair).

E Mechanism and Root Cause

E.1 Root-cause localization: the error is upstream of the decoder

This appendix details the two experiments behind the "error is upstream of the decoder" claim (§6). Both run on LLaDA-V, POPE, greedy decoding (T=32, gen-length 32); the pilot used n=150 and the scaled primary run n=600 (numbers below).

Experiment 1: causal condition ablation. For each question we decode the answer under three visual conditions with an otherwise identical prompt and decode: the *real* image, a content-free *blank* (black) image, and the real image plus heavy Gaussian noise (σ=110). We then classify each hallucination (wrong under the real image) as *language-prior-driven* if the yes/no answer is unchanged when the image is removed (the prior overrides vision), or *visual-misreading* otherwise (the answer depends on the image). Results at n=600 (pilot n=150 values in parentheses, stable across checkpoints): $\mathrm{acc}_{\mathrm{real}}$=0.835 (0.833), $\mathrm{acc}_{\mathrm{blank}}$=0.467 (0.453), $\mathrm{acc}_{\mathrm{noise}}$=0.640 (0.660); image-dependence rate 0.57 (0.62); of the 99 hallucinations, 61% are visual-misreadings and 39% language-prior (72%/28% in the pilot). As a control, *correct* answers also flip often under the blank image (58% on the stored subset)—as they must, since vision is load-bearing for correctness (the 0.835→0.467 collapse). The taxonomy's claim is directional: a hallucination that flips when the image is removed was *driven* by the image (a misreading), exactly as a correct answer that flips was driven by it. The collapse from 0.835 to 0.467 (near the binary prior) confirms the model genuinely uses vision.

Table 14: Root-cause split of LLaDA-V hallucinations (POPE, n=600; 97 classified hallucinations; pilot n=150 replicated). Noise floor 2.39; lens faithfulness 0.93. The correct answer is almost never represented above noise—the fault is visual under-encoding (upstream of the decoder), not a late-layer read-out that overwrites a correct intermediate answer.

Quantity	Value
Hallucinations where gold is never favored above noise (under-encoding)	**85%** (82/97)
Ambiguous (peak above floor only outside the mid/late band)	3% (3/97)
Hallucinations with correct answer mid-network then overwritten	12% (12/97)
Mean peak gold-margin, mid/late band (hallucinations)	0.72 (floor 2.39)
Full-vs-blank margin swing: hallucinations / correct	7.4 / 10.1

Experiment 2: layer-wise logit lens. We reconstruct the exact step-0 decision—the multimodal `inputs_embeds` the sampler builds, followed by `gen_length` masked answer slots—and run one bidirectional forward with `output_hidden_states`. At the first answer position we project every layer's hidden state through the final RMSNorm and the (tied) LM head (a logit lens) and read the yes-minus-no margin per layer, under the real and the blank image. Three controls make the read-out trustworthy. **(i) Faithfulness:** the lens's final-layer argmax matches the real diffusion sampler on 93% of examples (and the lens accuracy, 0.833, equals the ablation's $\mathrm{acc}_{\mathrm{real}}$), so the intermediate-layer reads reflect the real decision. **(ii) Noise floor:** because a binary answer across 33 layers is favored *somewhere* almost surely, the naive "correct-at-some-layer" rate is 1.0 even for correct predictions. We therefore define the noise floor as the 90th percentile of the spurious anti-gold margin peak within correct examples' mid band. That band is layers 10–29 of 33, excluding the embedding/early layers and the final three readout layers. The floor is 2.39 at n=600 and 2.42 in the pilot. Only a mid/late-band gold peak that exceeds this floor is counted. **(iii) Floor sensitivity:** the headline

85% is at this default floor; sweeping the floor moves the under-encoded rate over 53–98% (floor 0→3) and 78–93% (75th–95th percentile). The *qualitative ordering* nevertheless holds at *every* floor: under-encoding is the majority mode and dominates late-layer overwrite. The direction is therefore robust even though the exact split is floor-dependent. Orienting margins toward the gold answer and averaging per layer gives a clean picture. On hallucinations the gold margin is negative from mid-network onward (-0.4 at layers 10–20), and the late "decision-amplifier" layers sharpen it to -3.1 at layer 28. Correct examples are the mirror image ($+8.2$ at layer 28).

The large full-vs-blank swing on hallucinations (7.4, close to the 10.1 on correct answers) shows the image strongly drives the wrong decision. The failure is visual *fidelity*—the model reads the image and forms a confidently-wrong representation—not the decoder ignoring vision. Since the discriminative evidence is absent from the representation before the decoder runs, no rule operating on the denoising dynamics can recover it—the mechanistic explanation for the null corrective results in §5.7.

Two further probes triangulate the same conclusion. First, a *feature-decodability probe*. We extract the connector's image tokens exactly as the LM receives them (mean-, max-, and object-conditioned attention-pooled, concatenated with the question-object embedding) and train a regularized linear probe to predict the POPE answer (PCA-64; standardization and PCA fit on the training fold only). POPE asks several questions about each image, so these 260 rows come from only 44 distinct images, each carrying both *yes* and *no* labels, and the mean-/max-pooled vision blocks are constant across an image's rows. Folds are therefore *grouped by image* (5-fold, averaged over 20 shuffles of the group assignment); without grouping, an image's paired questions straddle the split and the probe is scored on vectors it saw under the opposite label. The joint probe reaches 0.690 accuracy (0.715 AUROC) against a 0.50 base rate—well below the model's own 0.84. An image-blind control decomposes it. The question-object text embedding *with no image* already reaches 0.721 AUROC (the POPE-adversarial lexical prior), the vision blocks *alone* reach 0.616, and the mean-/max-pooled blocks sit exactly at chance because they do not vary within an image. Adding vision to the text prior does not improve on it (0.715 vs. 0.721): the visual representation carries little decodable object evidence beyond an image-blind prior, which is the under-encoding reading. The LM already extracts *more* than is linearly present in its visual input. Under these probes there is no *decodable* reservoir of unused visual signal for a better read-out: nonlinear MLP probes (1–2 hidden layers, 64–256 units, same image-grouped protocol) reach 0.62–0.64, no better than the linear probe's 0.690. The absence is therefore not an artifact of probe linearity on these pooled features, though token-level nonlinear readouts remain untested. Second, a *fidelity-ceiling probe*. Re-answering each question under six views of the image—native, center-zoom, and four overlapping corner crops at $\sim$1.7$\times$ effective resolution—yields only a $+0.025$ net gain. That is under the best realistic aggregator (per-view majority; n=40 pilot, so indicative rather than tight). The naive "any view says yes" rule is strongly asymmetric—perfect recovery on *yes*-questions but collapsing *no*-question accuracy to 0.35 by importing false positives. (Per-view "oracle recovery" rates are inflated by the same binary-chance effect controlled for above, and should not be read as headroom.) Three independent measurements therefore converge: the null decode-time interventions, the logit-lens result, and this probe. The n=40 multi-view pilot is consistent with them but too small to carry weight. The missing signal must be added *upstream*—in the encoder, data, or training—and the trajectory's proper role is detection and abstention. The scripts (`localize_pain.py`, `root_cause_lens.py`, `analyze_lens.py`, `feature_probe.py`, `reperception_ceiling.py`) and the raw per-layer trajectories and per-view records are served at the companion site.

The error is upstream of the decoder. A deeper question is *why* no trajectory-derived decode-time rule can recover the right answer. We localize where the evidence is lost on POPE, combining a *causal* image ablation with a *correlational* logit lens (LLaDA-V, n=150; Appendix E.1). First, we ablate the visual input across three conditions: the real image, a blank one, and a heavily-noised one. The model genuinely uses vision—accuracy collapses from 0.835 (real) to 0.467 (blank, near the yes/no prior; n=600). Of the hallucinations, 61% are *visual misreadings*, meaning the answer flips once the image is removed; the remaining 39% are language-prior overrides. Second, a layer-wise logit lens at the answer position asks whether the correct answer is ever computed internally. We validate the lens by checking that its final-layer read-out matches the real sampler on 93% of examples. We also control for chance by measuring a noise floor from correct examples. A binary answer across 33 layers is favored *in some layer* almost surely, so the naive "present-at-some-layer" metric reads

1.0 even for correct predictions and has to be discarded. Against that floor, in 85% of hallucinations the correct answer is *never* favored above noise at any depth (mean peak gold-margin 0.72 against a floor of 2.39). Only 12% show the correct answer computed mid-network and then overwritten by the final layers. Moreover the image still moves the wrong decision strongly on hallucinations: the full-vs-blank margin swing is 7.4, close to the 10.1 on correct answers. The failure is therefore one of visual *fidelity*—the model reads the image and forms a confidently-wrong representation—rather than the decoder ignoring vision. One caveat on the partition itself. A per-layer *causal* cut of the same records (full-vs-blank lift toward *gold*) finds the image pushing toward the correct answer nearly as strongly on hallucinations as on correct items (+3.64 vs. +3.92). Under that cut the strictly under-encoded share drops to $\sim$0.57 $[0.47, 0.67]$, and the remainder is better described as failed *arbitration* between competing internal evidence. We therefore treat the floor-robust statement as the finding: for the majority of hallucinations the correct answer is never the internally *favored* option at any depth, so a decode-time selection rule has nothing to select. We leave the under-encoding/arbitration split as an open question—a follow-up causal-patching pilot leaned toward under-encoding, but its step-0 logit-diff metric failed its own controls, so we do not adjudicate here. The discriminative evidence is thus never the *favored* option before the decoder runs—whether because it is absent or because it loses an internal arbitration. A corrective rule operating on the denoising dynamics has nothing to recover, which is precisely why detection succeeds while every decode-time correction we tried is null. This scopes the finding to LLaDA-V on object-presence questions, but it makes the detect-but-cannot-correct result a statement about *where* the error lives, not merely about the trajectory signal.

E.2 Grounding analysis

We locate the image-token block from the `<image>` placeholder position j in the original ids: after multimodal merging, image features occupy positions $[j, j{+}N)$ where $N = L - (L_{\text{orig}} - 1)$ and L is the merged sequence length. We compute, per denoising step and per layer, the mean over answer rows of the attention over these image columns, then its entropy (grounding-entropy) and its sum (image-attention fraction). Grounding-entropy is at chance for predicting hallucination across layers (0.48–0.59 AUROC). The image-attention *fraction* shows a weak but reproducible signal at layer 8 (mean over steps, AUROC 0.657 at $n{=}400$ and 0.673 at $n{=}200$) and chance at layers 16/24, suggesting early-layer grounding carries a minor complementary cue; it remains well below the token-level trajectory (0.77), which is the informative signal.

E.3 Case study: a hallucination that is as committed as a correct answer

F Domain Shift: Medical VQA

The medical domain inherits the pathology. We port the identical capture to *LLaDA-MedV* (Dong et al., 2026), a biomedical fine-tune of LLaDA-V, on VQA-RAD and SLAKE (the English-only distribution `SLAKE-vqa-english`; closed yes/no; $n{\approx}400$ each, unified single-word protocol[2]). The anti-calibration finding *transfers*: output confidence predicts MedV's errors at **0.45 AUROC** $[0.39, 0.50]$ **(at-or-below chance)** on VQA-RAD, while the trajectory statistic reaches 0.59 (VQA-RAD) and 0.63 (SLAKE)—the best signal in every medical cell (Figure 3, last two columns). Two further observations sharpen the story. First, the base LLaDA-V collapses to 0.42 accuracy (below chance) under this domain shift, so reliability signals matter most exactly here. Second, our capture check shows MedV's trajectory and final entropy correlate at 0.98 in Pearson terms but only 0.34 in *rank* terms: the fine-tune commits quickly (so the discrete flicker signals carry nothing), yet within the resulting near-zero entropy mass the two statistics order examples differently—which is exactly why their AUROCs differ (0.59 vs. 0.45). MedV is therefore an intermediate regime between LLaDA-V's dynamic decoding and LaViDa's strict one-step coincidence (where rank correlation is also high; Appendix B.9); a per-step verification like Appendix B.9's on the medical captures is released with the records.

[2]Under our unified protocol LLaDA-MedV reaches 0.54 accuracy on VQA-RAD closed; model-specific protocols report higher absolute numbers. Our claims concern within-protocol *reliability*, not absolute accuracy: prompt variants (0.60–0.65) and constrained yes/no scoring (0.50–0.57) do not change the calibration picture.



Figure 9: **Per-head answer→image attention for the two case-study examples** (the four most focused heads by a computed concentration criterion—top-5% attention mass on the base view, averaged over both cases—shown as the *same* columns for both rows; border ring masked as attention-sink artifact; red ellipses mark the two largest computed peaks; each heat-map is normalized within its own panel, so the comparison is of attention *shape*, not of absolute magnitude—the quantitative claim rests on the AUROCs below, not on the images). The hallucinated "yes" (bottom) attends to the vehicles as sharply and plausibly as the correct answer (top) attends to the person: visual attention does not separate right from wrong, consistent with the at-chance grounding-*entropy* AUROC (a weak layer-8 attention-fraction cue exists, 0.66; see text) in our detection analysis. The failure lies in what the visual pathway *encodes*, not where it looks.

G Open-Ended Generation: Protocol and Results

Protocol. 500 unique COCO `val2014` images (the distinct image sources of the POPE-adversarial split), captioned by LLaDA-V ("Describe this image in detail."), generation length 64, T=64, one block, default low-confidence remasking. Object mentions are extracted with the standard 80-class CHAIR synonym table (Rohrbach et al., 2018) via greedy longest-match with token-span alignment; ground truth is the official `instances_val2014` annotation set. **Hallucination rates.** CHAIR_s $= 0.234$ ($117/500$ captions with $\geq$1 hallucinated object); $\text{CHAIR}_i = 0.141$ ($213/1506$ mentions). **Detection (per mention, signal aggregated over the mention's token span; AUROC for predicting a hallucinated object, 95% bootstrap CI).** Trajectory (mean step-entropy) 0.616 $[0.572, 0.661]$; confidence (final entropy) 0.632 $[0.590, 0.671]$; confidence ($1-$max-prob) 0.627 $[0.586, 0.667]$; commit step 0.613 $[0.566, 0.660]$. Caption-level (predicting any-hallucination, n=500): trajectory 0.54 $[0.48, 0.60]$, confidence 0.62 $[0.57, 0.68]$. **Mechanism.** Within a caption, per-position mean step-entropy is a decode-order proxy: median within-caption Spearman correlation with the position's commit step is 0.836 (0.925 at mention level), versus 0.265 with the position's content confidence. Commit step alone detects at 0.613; step-entropy residualized on commit step falls to 0.550. In one long free-running block, low-confidence remasking makes late-committed positions carry high step-entropy by construction, so the statistic collapses onto decode order and loses the content-specific component that drives its short-form advantage. This is consistent with concurrent work showing that in diffusion LMs the commit (reasoning) order diverges from output-position order and that harder positions stabilize later (Yu et al., 2026). It is a boundary of the *statistic* under this decoding regime, not a contradiction of the QA results: the advantage appears where the readout position is prompted and contested (all QA settings above). Records: `openended_raw.json`, `openended_analysis.json`, `openended_decodeorder.json`.

H Capture Pools and Cross-Section Numbers

Each analysis in the paper runs on its own captured pool, so quantities that sound alike can differ slightly across sections; we list them here so every number is traceable. **Base POPE accuracy:** 0.847 on the n=300 signal-capture pool (used by the decode-time interventions, Table 13); 0.833

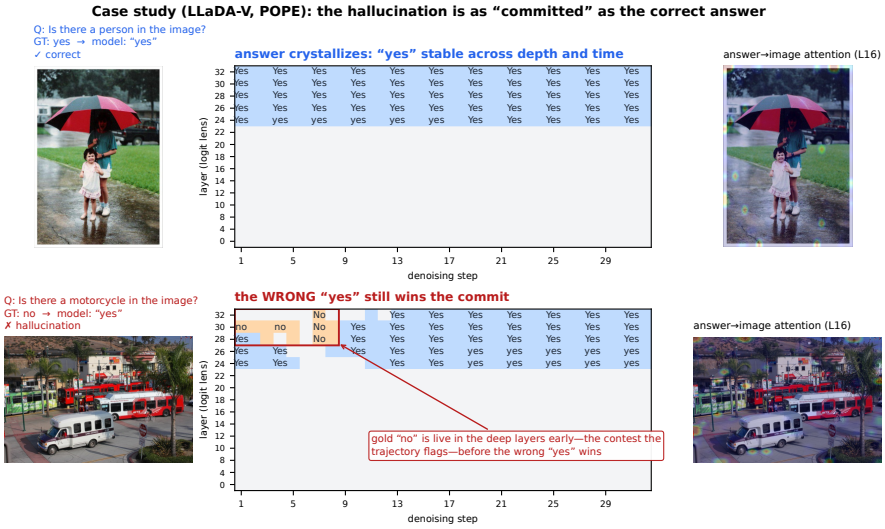

Figure 10: **Case study on two POPE images** (real data; examples chosen for grid clarity—the hallucinated case is from the 12% overwrite class, the one mode with mid-network gold to display; the dominant 85% under-encoded mode shows none at any depth, Fig. 4). Center: the logit-lens-decoded token at the answer position over a subsampled layer set (rows; 15 of the 33 read-out points, every second block above ℓ=8) and every denoising step (columns), printed where the decoded token is a decision token (cell shading: grey = other, blue = "yes", orange = "no")—an axis no autoregressive analysis can draw (the logit lens of nostalgebraist, 2020, adapted to the layer×step surface of *text* dLLMs by Devoto, 2026; we bring the layer×step lens to a diffusion *VLM* and tie it to hallucination). *Top:* for the correct answer, "yes" crystallizes in the late layers at step 1 and never moves. *Bottom:* the wrong "yes" crystallizes with identical stability—and the gold "no" that is live in the deep layers ($\ell \geq 24$) during the earliest steps is overwritten (red box), a live instance of the 12% overwrite class from our root-cause analysis. Right: answer-to-image attention (layer 16, base view) is comparable in both cases—the failure is what gets *encoded*, not where the model looks. Replay decodes reproduce the real sampler's answers for both cases.

on the n=150 causal-ablation pilot and 0.835 on its n=600 scaled run (§6); 0.785 on the n=400 preference-training evaluation pool (§5.7); 0.830 on the n=400 main capture (and its replication); 0.824 on the n=125 UQ-baseline pool; 0.815 is the plain final-step readout in the separate n=400 decode-variants (voting) run; and the AR-matched mirror pools (n=300/600, Table 4 and Figure 2) give confidence 0.476 and trajectory 0.756, slightly different from the n=400 capture's 0.504/0.778. **GQA accuracies:** 0.3667 under first-word scoring on the n=300 two-hop capture (replication identical); 0.544 under the same rule on the separate n=250 preference-training evaluation pool; and two *different* verbosity-robust rules—0.57 counts the gold string anywhere in the full generation (the matched check of §4, on the original stored generations), while 0.49 counts it within the parsed answer span (the replication's robustness analysis below). The two contains-rules are distinct by design; both are released. **Headline detection AUROC:** the main capture gives 0.778 (POPE n=400). The denoising-steps study (Table 10) is a separate run of the same first-n protocol on a different host. Half-precision kernel nondeterminism across hosts flips a handful of marginal examples, so its T=32 value is 0.768, within the main capture's 95% CI $[0.713, 0.838]$. All pools are first-n prefixes of their split; the preference-training evaluation additionally runs the model in `bf16` (matching its training arm) rather than `fp16`, which accounts for its lower base accuracy (0.785), and the decode-variants run reads the final-step argmax without the commit rule (0.815). LaViDa base accuracies: 0.850/0.850 (POPE, LLaDA/Dream backbones) and 0.510/0.500 (GQA); Qwen2-VL-2B: 0.85 (POPE), 0.605 (GQA, n=600). **Rounding:** table cells are rounded to two decimals; ratios and multipliers quoted in the text are computed from the unrounded values in the released records.

Exact replication of the LLaDA-V GQA capture. The original LLaDA-V *GQA* capture stored summary statistics only, so we re-ran the identical protocol end-to-end (same pools, prompts, and

decoding) with full per-example records. The replication agrees to within 10^{-4} on every statistic: GQA accuracy 0.3667 vs. 0.3667, confidence AUROC 0.3714 vs. 0.3713, entropy 0.3853 vs. 0.3852, and trajectory 0.7228 vs. 0.7228. The released POPE per-example records likewise reproduce the main-text POPE capture exactly (trajectory 0.7781, confidence 0.5040). The GQA CIs in Table 5 are computed from this replication's released records. Confidence's GQA interval $[0.32, 0.45]$ lies entirely below chance: the anti-calibration is CI-certified, not a point estimate.

Root-cause split: uncertainty and threshold sensitivity. The causal-ablation pool contains 97 lens-classifiable hallucinations at the scaled n=600 (2 of the 99 ablation hallucinations were not lens-classifiable; the pilot had 25): 85% under-encoded is $82/97$, Wilson 95% interval $[76\%, 90\%]$ (pilot: 88%, $[70\%, 96\%]$—replicated); 61% visual-misread is $60/99$, Wilson $[51\%, 70\%]$ (pilot 72% $[52\%, 86\%]$). The split is robust to the noise-floor choice: at the 75th/90th/95th-percentile floors the n=600 under-encoded fraction is $78\%/85\%/93\%$ (pilot: $80\%/88\%/92\%$)—the dominant class never changes. We therefore read the percentages as a *direction* supported by four converging measurements, not as precise magnitudes.

Robustness to the scoring rule (GQA). First-word matching on open-vocabulary GQA answers has label noise, and a skeptic might worry the detector exploits it. Rescoring the replication with *contains*-matching (gold string contained in the parsed answer; accuracy rises $0.37 \rightarrow 0.49$) makes the result *stronger*, not weaker: trajectory AUROC 0.747 $[0.689, 0.802]$ and confidence 0.346 $[0.285, 0.408]$—still entirely below chance. The finding is scoring-rule-robust.

Detection within predicted-answer strata (POPE). On binary POPE one might worry the detector merely senses a mismatch between the answer and the model's yes/no prior rather than genuine error. Stratifying the replication by the *predicted* answer refutes this: among predicted-*yes* answers only (n=229), the trajectory detects errors at 0.79 $[0.72, 0.86]$; among predicted-*no* (n=169), at 0.73 $[0.60, 0.86]$ (the remaining 2 of 400 answers were unparseable). Confidence, for comparison, recovers *some* within-class ranking (0.66 predicted-yes, 0.69 predicted-no) even though it is at chance marginally (0.50)—the anti-calibration on POPE is thus partly a between-class (Simpson-type) structure, exactly what a deployed threshold on *marginal* confidence inherits. The trajectory needs no such conditioning: it leads confidence *within both strata* (0.79 vs. 0.66; 0.73 vs. 0.69) and marginally (0.78 vs. 0.50). Both stratified baselines are in `recapture_analysis.json`.

The GQA inversion is between answer classes, not a frequency artifact. Two follow-up analyses on the replication records. (i) Final confidence is essentially uncorrelated with the predicted answer's frequency in the pool (Pearson 0.06)—ruling out a "frequent tokens get high confidence" artifact. (ii) Grouping examples by *predicted answer*: across classes with $n \geq 5$, the class's mean confidence correlates $+0.63$ with its error rate—the model is most confident on its most error-prone answer types (e.g., person-gender answers: max-probability 0.996–0.999 at 57–100% error). Within the two large binary classes confidence retains some ranking signal (within-class AUROC 0.58 yes / 0.89 no); within open-vocabulary classes it is at or below chance. The trajectory retains within-class signal throughout (0.55 yes, 0.77 no, 0.60 pooled over open-vocabulary classes at small n=24), so part—not all—of its GQA aggregate reflects between-class structure. The below-chance aggregate is thus genuine between-class anti-calibration—confidently-wrong answer types—not a scoring or frequency artifact (small per-class n; read as structure, not precise magnitudes). **Abstention audit:** at the 50%-coverage operating point on POPE, both answer classes are answered at similar rates (yes 48%, no 53%) with kept-error $5.4\%/4.5\%$—selection tracks instance correctness, not class identity, on this benchmark.

The difficulty control (rival-theory test). A skeptic's most parsimonious rival reading is that early-step entropy is a question-*difficulty* meter (hard items have diffuse posteriors and high error, so any difficulty proxy predicts errors). We test it by holding difficulty constant with an *external* covariate: the matched AR model's per-item correctness on the shared pools. Within the majority stratum that Qwen2-VL answers *correctly* (the easy items: n=340 POPE, n=173 GQA), the trajectory still detects LLaDA-V's errors at **0.86** $[0.80, 0.91]$ and **0.70** $[0.62, 0.77]$ respectively—if the signal were item difficulty, these within-stratum AUROCs would sit at chance. (The residual Qwen-wrong strata are small and error-saturated—n=60 at 42% and n=127 at 86% LLaDA-V error—with wide CIs spanning chance.) The trajectory ranks *idiosyncratic model errors within a difficulty*

level, not the difficulty of the question (`difficulty_partial.json` released). The control is rater-robust: stratifying instead by Qwen2-VL-*7B* correctness gives 0.85 $[0.77, 0.92]$ (n=341) and 0.71 $[0.63, 0.79]$ (n=168) (`difficulty_partial_7b.json`).

The dynamism regularity (all twelve cells). Trajectory–endpoint rank-decoupling ($1 - \rho_S$ between mean step-entropy and final entropy) against the trajectory's AUROC advantage over the stronger confidence variant: LLaDA-V POPE $0.74/{+}0.27$, GQA $0.94/{+}0.34$, MME $0.51/{+}0.05$, HallB $0.40/{+}0.12$; LaViDa-L POPE $0.004/{+}0.001$, LaViDa-D POPE $0.06/{-}0.003$, LaViDa-L GQA $0.69/{+}0.19$, LaViDa-D GQA $0.12/{-}0.02$; MedV VQA-RAD $0.66/{+}0.14$, SLAKE $0.69/{+}0.08$; **out-of-sample** (measured after the regularity was fit): MMaDA POPE $0.03/{-}0.06$, MMaDA GQA $0.46/{+}0.11$. Correlation over all twelve: r=0.89 (Spearman 0.84); leave-one-out refits stay at r=0.86–0.94, so no single cell carries the fit. (Limitation 1's *thirteen* in-domain settings are these twelve model×benchmark cells plus the POPE n=1200 confirmation capture.) Where decoding sharpens the endpoint away from the trajectory, the trajectory wins by the amount of the divergence; where they coincide, neither wins. Two scope conditions, made explicit: the regularity is fit on cells decoded with their *native* confidence-ordered schedules (artificially randomized schedules decouple without transferring the advantage; Appendix B.10), and it presupposes an *operative* detector—under severe domain shift (base LLaDA-V on VQA-RAD: accuracy 0.42, trajectory 0.53, confidence 0.56) decoupling is high (0.65) yet there is no signal for it to release, so that boundary cell sits off the line (`med_vqarad_base.json`). The conditional claim is thus a predictor within its stated scope, not a post-hoc taxonomy (`dynamism_law.json` in the released records).

Span-aggregation ablation (readout-position control). Recapturing both benchmarks with per-position final-step statistics over the whole 32-token span (mean span length 31; per-example records released): position-0 values replicate the main captures to the third decimal (GQA 0.372/0.385/0.723; POPE 0.504/0.500/0.778). Aggregating the *final-step* confidence over the span yields 0.55/0.55/0.56 on GQA and 0.52–0.54 on POPE—uninformative, not inverted—and the span-aggregated trajectory drops to 0.60 (GQA) / 0.63 (POPE). Both signals are strongest at the answer position, where the scored answer lives and where their comparison is readout-matched; positions 2–32 are verbose continuation whose statistics dilute both signals symmetrically. The answer-position inversion itself is not a verbosity-scoring artifact (contains-match rescoring strengthens it) and has the between-class structure documented above. The signal separates right from wrong *within* each answer class, where the prior is constant.

I Extended Related Work

Two neighbours that differ in kind. A reviewer may reasonably ask why we do not benchmark against training-free AR methods that read *internal* state rather than the output distribution. They divide into two kinds, and only one is comparable. **OPERA** (Huang et al., 2024) does read internals—the attention paid to previously generated tokens—but it is a decoding *intervention*: a beam-search logit penalty plus a rollback rule. Its output is better text, not a per-example score, so there is no quantity to rank by and no detection AUROC to compare against; and masked diffusion has neither beam search nor a left-to-right attention history to penalize, so the mechanism does not port. Reporting a "diffusion OPERA" would be reporting our own re-invention under someone else's name. **VAUQ** (Park et al., 2026) is the apt comparison: a training-free *score* that asks how much the image reduces predictive uncertainty. We implement its image-information backbone and report it alongside our own signal in Appendix B.6.

Extended related-work discussion. Table 15 positions our work against the three neighboring lines of work it sits between. **(i) Uncertainty for diffusion LMs is text-only.** Decoding-trace and denoising-dynamics hallucination detectors (Chang et al., 2026; Qian et al., 2026), hidden-evidence verification from the denoising process (Zhao et al., 2026), and the first systematic UQ study for diffusion LMs (Vazhentsev et al., 2026) all evaluate *text* dLLMs (LLaDA-1.5, Dream) on QA/summarization. Our work differs on three axes: it is *multimodal* (diffusion VLMs, not text), it *quantifies* output confidence as unreliable-to-*anti-calibrated* (AUROC 0.37–0.50: at chance on POPE, below chance on GQA), and it contrasts autoregressive with diffusion as a mirror—none of these text-only detectors does. Improving or gating the unmasking order is a parallel text-only line (Lee et al., 2025; Jazbec et al., 2025; Saini et al., 2026). The closest is concurrent *trajectory-aware*

Table 15: Positioning against prior work. Ours is the only entry at the diffusion × multimodal × trajectory intersection, and the only one to report the anti-calibration diagnosis and the AR-vs-diffusion mirror.

	Diffusion	Multimodal	Trajectory signal	Anti-calib.+mirror
Text-dLLM UQ (Chang et al., 2026; Qian et al., 2026; Vazhentsev et al., 2026) (ICLR '26; arXiv '26)	✓	✗	✓	✗
AR-VLM UQ (Leng et al., 2024; Wu et al., 2025; Manakul et al., 2023) (CVPR '24; NeurIPS '25; EMNLP '23)	✗	✓	✗	✗
Sampling UQ (Kuhn et al., 2023; Lakshminarayanan et al., 2017) (ICLR '23; NeurIPS '17)	✗	Both	✗ (multi-pass)	✗
Calibration (temp. scaling, (Guo et al., 2017)) (ICML '17)	✗	✗	✗	✗
Sample-variance UQ (Zhang et al., 2024) (arXiv '24)	✓	✗	✗ (multi-pass)	✗
Ours	✓	✓	✓	✓

commit gating (Wang et al., 2026), which uses a step-entropy/commit signal to *accelerate* decoding by committing early once confident. That is the same trajectory-commit primitive as our TCR (§5.7), but optimized for the opposite objective (speed rather than reliability), text-only, and with no detection or calibration analysis. **(ii) Uncertainty for VLMs is autoregressive.** Hallucination detection/mitigation for AR VLMs leans on the output distribution—semantic entropy over perturbed inputs and contrastive/retrospective decoding (Leng et al., 2024; Wu et al., 2025)—which work *because* AR output confidence carries signal. Closest in spirit, Jiang et al. (2025) interpret, detect, and mitigate AR-VLM hallucinations through an internal *attention* lens; we pursue the same interpret–detect–act program for diffusion VLMs, where the natural internal signal is the *denoising trajectory* and where the acting step must be abstention with guarantees rather than decoding intervention (§5.7). Sample-variance uncertainty in dense generative models (e.g. diffusion-based scene flow, Zhang et al., 2024) needs extra forward passes; our trajectory signal is single-pass and free. A separate line *corrects* diffusion-VLM errors during generation (Ji et al., 2025); we instead *detect* and abstain, and show (§5.7) that acting on the trajectory does not improve accuracy *in-stack*—so the two are complementary. Cross-system acting, however, works: see the routing result below. **(iii) Conformal abstention for VLMs is also autoregressive.** Conformal abstention policies (Tayebati et al., 2025) and budgeted evidence acquisition (Xu et al., 2026) deliver risk guarantees for AR VLMs, but both consume softmax-confidence-family scores as input—the very quantity that is anti-calibrated on diffusion VLMs; VISAGE (Narnaware et al., 2026) gates unmasking commits by attention entropy on MMaDA; to our knowledge it reports no detection-AUROC or calibration analysis. Our risk-control results (§5.5) supply the missing ingredient for this family on diffusion VLMs: a score that sustains risk-controlled coverage where confidence cannot. Our contribution occupies the empty intersection of these literatures.

J Extended Limitations

This appendix states all six boundaries in full; the main text summarizes the first four.

(1) Scope of evidence. Four diffusion VLMs (three LLaDA-initialized recipes plus a Dream-7B backbone), six QA benchmarks plus open-ended COCO captioning, English only, n=225–1506 per cell, one seed per cell with one full replication (exact to 10^{-4}; Appendix H). Greedy decoding is deterministic, so the only randomness is the evaluation sample (quantified by our bootstrap CIs); the headline gap also reproduces on a disjoint POPE pool, across three prompt phrasings, and under temperature-sampled decoding (Appendix B.4). The abstract's 0.65–0.79 detector range covers POPE/GQA. Over the thirteen reported in-domain settings the range is 0.59–0.79. Under severe domain shift, meaning the base general-domain model on medical imagery, detection itself degrades to 0.53; we report that as a boundary in §5.3 and Appendix H. At the lower end of the range, risk-controlled coverage is not yet useful. More model families and seeds, and generative hallucination

suites such as AMBER (Wang et al., 2023a) and MMHal-Bench (Sun et al., 2024), would strengthen generality.

(2) What the regularity does not cover. The trajectory's *advantage* appears only under dynamic decoding, and every dynamic cell to date is LLaDA-initialized; Dream-initialized cells tie, exactly as the one-step-commit analysis predicts (Appendix B.9). We tested this directly: forcing the independent Dream backbone into high-decoupling dynamic decoding did *not* transfer the advantage (Appendix B.10), so it is a representational property of the LLaDA-family recipe, not of decoding dynamics alone. Only the entropy statistic travels; the discrete flicker signals are decoding-specific. In long free-running blocks (open-ended captioning) the statistic collapses onto decode order and the advantage vanishes (§6; Appendix G). The AR mirror replicates at matched scale (Qwen2-VL-7B) but remains a cross-recipe contrast, not a controlled ablation.

(3) Deployment costs. SEISMO-Cal and SEISMO-Guard are per-model components consuming a *labeled* calibration set (400–1200 examples): the score is free, the guarantee is not. The capture itself needs per-step logits (one forward hook)—available with code access but not through serving APIs that expose only final outputs. The guarantee is marginal—over the answered set, under exchangeability—and currently delivers useful certified coverage only on low-error, POPE-style queries (§5.5).

(4) Diagnosis, not correction. SEISMO detects and contains; it does not correct, and we show why correction fails in-stack (the evidence is lost upstream). The root-cause analysis covers LLaDA-V on object-presence questions (piloted on 25 hallucinations, replicated at n=600: 85% $[76\%, 90\%]$ at the default floor; 53–98% across floors, and a causal cut of the same margins lowers the strictly under-encoded fraction to $\sim$0.57—the upstream localization, not the partition, is the claim); other question types and models, and any fix at the vision–language interface, remain open.

(5) Measurement choices. All signals are read at the first answer position; the span-aggregation ablation (Appendix H) shows the below-chance result is specific to that readout—the quantity the scoring evaluates—while whole-span confidence is merely uninformative. LLaDA-MedV's released config hardcodes a local vision-tower path, so its cells assume our documented reconstruction (Appendix A).

(6) In-stack framing. Our comparison is *within* the diffusion stack. If one is free to choose the stack, a well-calibrated AR model can dominate end-to-end; our matched 2B AR baseline is more accurate with usable native confidence. This paper's tooling addresses the case where a diffusion VLM is chosen for its own reasons—parallel decoding, infilling, editability—and reliability must be recovered in-stack.