The Mirage of Calibrated Confidence: Trajectory-Independence of Verbalized Confidence in Vision-Language Models

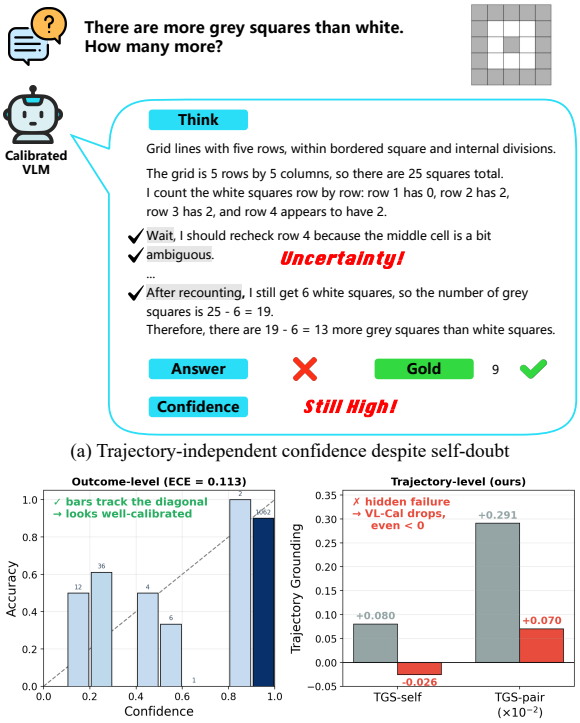
Anonymous ACL submission

Abstract

A calibrated Vision-Language Model (VLM) can repeatedly self-correct, say "Wait, I should recheck," arrive at the wrong answer, and still report high confidence. We find that this occurs because verbalized confidence in VLMs is fundamentally *trajectory-independent*. We examine this through three complementary lenses: content variation, token masking, and the model's own hesitation markers. We show that confidence is invariant to what the reasoning trajectory actually contains, and that calibration training can paradoxically worsen this disconnect. Since existing metrics like ECE and AUROC cannot detect this problem, we propose the **Trajectory-Grounding Score (TGS)** in two complementary forms: ***TGS-self***, which compares confidence with and without access to the model's own trajectory, and ***TGS-pair***, which tests whether the model assigns higher confidence to correct trajectories than to flawed ones along the vision, reasoning, and answer axes. We propose **TGS-Bench**, a model-agnostic suite spanning 10 benchmarks with controlled good/bad trajectory pairs, and show that conventional calibration rankings diverge from trajectory-grounding rankings, exposing a blind spot in current evaluation practice.

1 Introduction

Recent Vision-Language Models (VLMs) have achieved strong performance across visual reasoning tasks (Bai et al., 2025; Wang et al., 2025; Liu et al., 2023). Yet they frequently produce incorrect answers with unwarranted high confidence (Xiong et al., 2024; Kadavath et al., 2022). *Verbalized confidence* calibration has made notable progress in closing this gap, reducing ECE and improving AUROC (Li et al., 2025; Damani et al., 2025; Xiao et al., 2026). But these aggregate metrics conceal a deeper question. *Does the model's confidence actually reflect the quality of its reasoning?* A model could achieve low ECE simply by learning the base rate of its own correctness on each difficulty tier, without ever inspecting whether its visual perception was accurate or its reasoning chain was valid for any particular instance.



(a) Trajectory-independent confidence despite self-doubt

(b) Outcome-level success, Trajectory-level failure in Calibrated VLM

Figure 1: **(a) Trajectory-independent confidence despite self-doubt.** A calibrated VLM repeatedly expresses self-doubt yet assigns high confidence to an incorrect answer. **(b) Trajectory-independence undetected by outcome-level metrics.** VL-Calibration-8B appears well-calibrated (ECE = 0.113), but TGS-self turns negative and TGS-pair drops after calibration.

We find that this is precisely what happens. As Figure 1(a) illustrates, a calibration-trained model repeatedly hedges, self-corrects, and revisits its own reasoning, yet assigns high confidence to an answer that is wrong. The self-doubt is visible in the text; the confidence is blind to it. We term

this missing property *trajectory-grounding*, the degree to which confidence is causally informed by the specific trajectory that produced the answer. Crucially, this failure is invisible to conventional metrics. As Figure 1(b) shows, the same model appears well-calibrated by ECE, yet scores near zero or negative on our trajectory-level metrics.

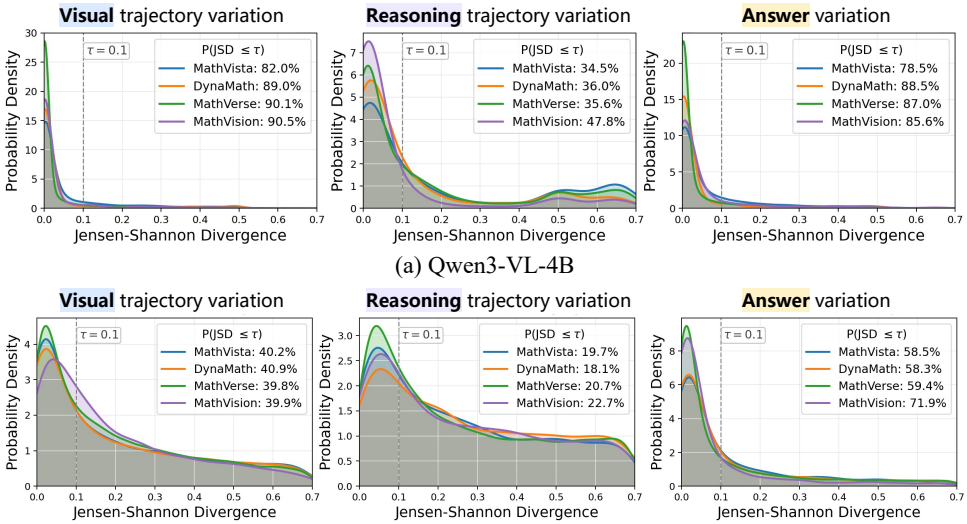
Our analysis proceeds in three stages of escalating severity. First, we replace trajectory components (vision, reasoning, answer) with qualitatively different alternatives and find that confidence remains nearly unchanged (§3.2). Second, we physically mask trajectory tokens via attention-level ablation and observe that, rather than declining monotonically, confidence rises at complete masking (§3.3). These two stages show that confidence is invariant to both the content and the presence of the trajectory. The third stage asks the most direct question. When the model itself expresses uncertainty through hesitation markers such as "wait" or "let me recheck," does its confidence respond? The overwhelming majority of instances produce virtually no confidence shift (§3.4). The model is blind not only to externally manipulated trajectories, but to its own epistemic signals.

To the best of our knowledge, this is the first work to investigate trajectory-grounding of verbalized confidence in VLMs. Our contributions are threefold: (1) Through three complementary lenses (content variation, token masking, and the model's own self-doubt), we establish that verbalized confidence is trajectory-independent, and that calibration can paradoxically worsen this disconnect. (2) We propose the **Trajectory-Grounding Score (TGS)** in two forms. *TGS-self* compares confidence with and without access to the model's own trajectory. *TGS-pair* tests whether the model assigns higher confidence to correct trajectories than to flawed ones along the vision, reasoning, and answer axes, requiring controlled good/bad trajectory pairs. (3) To support *TGS-pair* evaluation, we construct **TGS-Bench** across 10 benchmarks with model-agnostic trajectory pairs that isolate errors on exactly one axis, and show that methods excelling on conventional calibration metrics score near zero or even negative on TGS, confirming that outcome-level calibration and trajectory-level grounding measure fundamentally different properties.

2 Related Work

Confidence Estimation and Calibration in VLMs Verbalized confidence estimation asks models to articulate their certainty alongside answers, offering a model-agnostic alternative to logit-based approaches (Lin et al., 2022; Yang et al., 2024). In the text-only LLM setting, prompting strategies revealed persistent overconfidence (Xiong et al., 2024), and training-based methods have since shown stronger results through Brier-score alignment (Li et al., 2025), RL with self-reflective rationales (Xu et al., 2024), PPO-based doubt rewarding (Bani-Harouni et al., 2025), and joint accuracy-calibration optimization via GRPO (Damani et al., 2025). Extending these to VLMs introduces modality-specific challenges: Du et al. (2026) show that VLM confidence is insensitive to visual input degradation and propose RL with original-noise image pairs, while VL-Calibration (Xiao et al., 2026) decouples confidence into visual and reasoning scores supervised by an intrinsic certainty signal. Despite substantially reducing ECE, all existing work evaluates calibration through aggregate metrics that measure population-level correlation between confidence and accuracy, without assessing whether individual scores reflect the quality of the specific reasoning process, a dimension we term *trajectory-grounding*.

Faithfulness of Model Self-Assessment Recent work questions whether models' self-generated assessments faithfully reflect their internal processes. On the reasoning side, Lanham et al. (2023) show that chain-of-thought explanations are not always faithful, and in multimodal settings CoT can obscure hallucination cues (Cheng et al., 2025). On the confidence side, ADVICE (Seo et al., 2025) demonstrates that verbalized confidence in text-only LLMs is nearly *answer-independent*: swapping the answer while fixing the question leaves the confidence distribution unchanged. We extend these findings to VLMs across three axes (vision, reasoning, and answer), revealing a broader phenomenon we term *trajectory-independence*. Beyond external interventions (content variation, token masking), we further show that models are blind to their own epistemic signals, and that calibration training (Xiao et al., 2026) can paradoxically worsen this independence.



(a) Qwen3-VL-4B

(b) VL-Calibration-4B

Figure 2: JSD distributions when varying trajectory components. **(a)** Qwen3-VL-4B (base). **(b)** VL-Calibration-4B. The base model is already trajectory-independent, especially on the vision axis ($P(\text{JSD} \leq 0.1) \approx 82$–$90\%$). VL-Calibration reduces this ratio but remains substantially trajectory-independent (vision $P(\text{JSD} \leq 0.1) \approx 40\%$), indicating that calibration training does not resolve the underlying disconnect between confidence and trajectory.

3 Trajectory-Independence of Verbalized Confidence in VLMs

3.1 Preliminary

We study VLMs that generate structured reasoning trajectories followed by verbalized confidence scores. Given a multimodal input (I, q) consisting of an image I and a textual query q, the model produces a trajectory $\tau = (v, r, a)$ where v denotes the visual perception rationale, r denotes the reasoning chain, and a is the final answer. After generating τ, the model verbalizes its confidence as a discrete score in $\{0, 1, \ldots, 10\}$.

Current methods either produce a single *holistic* confidence $c \in \{0, \ldots, 10\}$ (Damani et al., 2025; Li et al., 2025), or *decouple* it into a visual confidence c_{vis} and a reasoning confidence c_{reas} (Xiao et al., 2026). The decoupled scores are aggregated via a harmonic mean:

$$c_{\text{holistic}} = \frac{2 \cdot \hat{c}_{\text{vis}} \cdot \hat{c}_{\text{reas}}}{\hat{c}_{\text{vis}} + \hat{c}_{\text{reas}}}.$$

To cover both paradigms, our analysis targets (1) **Qwen3-VL-4B-Instruct** (Bai et al., 2025), a base VLM with holistic confidence, and (2) **VL-Calibration-4B** (Xiao et al., 2026), which is fine-tuned from Qwen3-VL-4B-Instruct via RL to produce decoupled confidence with intrinsic visual certainty supervision.

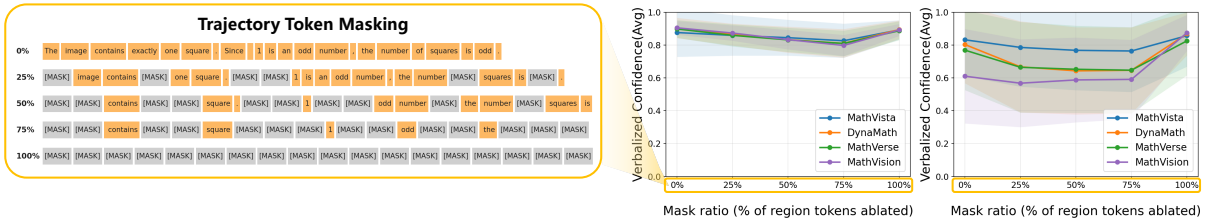
3.2 Confidence Distribution Analysis

We first ask: *does changing the content of a trajectory component alter the model's confidence?* If confidence is truly grounded in the trajectory, replacing a component with qualitatively different content should shift the confidence distribution. We quantify this shift using Jensen–Shannon divergence (JSD) (Menéndez et al., 1997), a symmetric measure bounded in $[0, 1]$ that equals zero when two distributions are identical. For each trajectory axis $x \in \{v, r, a\}$, we compare two alternative versions x_i and x_j while holding the remaining components fixed. Trajectory-independence along each axis implies:

$$\begin{aligned}
\text{JSD}_v &= \text{JSD}(P_{\mathcal{M}}(C \mid q, I, v_i, r, a) \,\|\, P_{\mathcal{M}}(C \mid q, I, v_j, r, a)) \approx 0, \\
\text{JSD}_r &= \text{JSD}(P_{\mathcal{M}}(C \mid q, I, v, r_i, a) \,\|\, P_{\mathcal{M}}(C \mid q, I, v, r_j, a)) \approx 0, \\
\text{JSD}_a &= \text{JSD}(P_{\mathcal{M}}(C \mid q, I, v, r, a_i) \,\|\, P_{\mathcal{M}}(C \mid q, I, v, r, a_j)) \approx 0.
\end{aligned}$$

We evaluate on four mathematical reasoning benchmarks[1] and compute pairwise JSD across all alternative pairs, visualizing the distribution via Gaussian kernel density estimation. Following prior work that considers two distributions similar when their divergence falls below 0.1 (Seo et al., 2025; Deho et al., 2025; Milan Kummaya et al., 2025), we report $P(\text{JSD} \leq 0.1)$ as a summary statistic, where higher values indicate stronger trajectory-independence.

[1] MathVista(testmini), DynaMath, MathVerse, and MathVision.



(a) Qwen3-VL-4B (b) VL-Calibration-4B

Figure 3: Verbalized confidence as a function of trajectory masking ratio. **(a)** Qwen3-VL-4B, **(b)** VL-Calibration-4B. In both models, confidence declines modestly as the masking ratio increases but paradoxically shifts upward at 100% masking, indicating that neither model meaningfully relies on trajectory content when generating its confidence.

Trajectory construction and intervention The two models require different experimental procedures due to their distinct architectures. For **Qwen3-VL**, which is not natively trained with structured tags, we first let the model generate a free-form chain-of-thought given (q, I), then segment the output into vision and reasoning components using the model itself, replace the target component with an alternative, and reconstruct the trajectory as free text before eliciting a holistic confidence token. For **VL-Calibration**, which natively generates structured <vision> and <reasoning> blocks followed by separate c_{vis} and c_{reas} tokens, we directly replace the content within the corresponding tags. We then compute JSD separately for each confidence component and report the average $\frac{1}{2}(\text{JSD}_{\text{vis}} + \text{JSD}_{\text{reas}})$ as the per-sample divergence.

Base VLMs are already trajectory-independent. As shown in Figure 2(a), the base model exhibits strong trajectory-independence: the vision axis yields $P(\text{JSD} \leq 0.1)$ of 82–90%, and the answer axis reaches 78–89%, meaning the vast majority of samples show near-zero confidence change when these components are replaced with qualitatively different alternatives. The reasoning axis shows moderate independence (34–48%), with JSD values spread more broadly across $[0, 0.7]$, suggesting that reasoning content has comparatively more influence on confidence, though still limited.

Calibration training does not resolve trajectory-independence. VL-Calibration (Figure 2(b)), despite its explicit decoupled structure designed to ground each confidence component in its corresponding trajectory segment, still exhibits substantial trajectory-independence. Its vision axis shows $P(\text{JSD} \leq 0.1) \approx 40\%$, reasoning 18–23%, and answer 58–72%. While these values are lower than those of the base model (indicating that calibration does introduce some trajectory sensitivity), the overall level of independence remains high. This also confirms that answer-independence, previously identified in text-only LLMs (Seo et al., 2025), generalizes to VLMs.

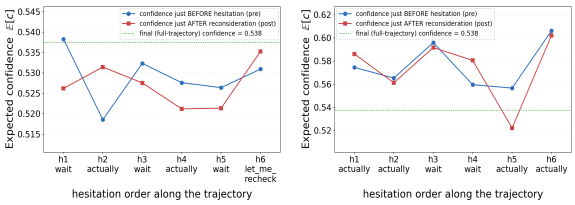
3.3 Token Masking Analysis

§3.2 demonstrated trajectory-independence through content *substitution*: replacing trajectory components with qualitatively different alternatives barely shifted the confidence distribution. We now apply a stronger causal test (content *removal*) to ask whether models rely on trajectory tokens at all. For each model, we first generate a self-trajectory using the model's native format, then apply attention-level masking at five nested ratios (0%, 25%, 50%, 75%, 100%) over the trajectory span (see Appendix A.2 for details). Under each masking condition, we let the model generate its verbalized confidence token and record the resulting score.

Removing tokens does not reduce confidence. As shown in Figure 3, both models exhibit a counterintuitive pattern: verbalized confidence declines modestly as the masking ratio increases from 0% to 75%, but paradoxically shifts upward at 100% masking, where the model has no access to any trajectory content. Notably, VL-Calibration exhibits a flatter curve than the base model throughout, suggesting that calibration training may even suppress trajectory reliance. If the model genuinely evaluated its reasoning before assigning confidence, we would expect a monotonic decline proportional to the amount of information removed; instead, complete removal recovers or even exceeds the original confidence level. This provides a direct causal complement to the distributional evidence of §3.2: VLM confidence is generated without meaningfully consulting the preceding trajectory.

Hesitation count	Qwen3-VL n (%)	VL-Cal n (%)
All trajectories	4025 (100.0)	5666 (100.0)
$\geq$1 hesitation	904 (22.5)	3239 (57.2)
1	188 (20.8)	608 (18.8)
2	134 (14.8)	563 (17.4)
3	86 (9.5)	434 (13.4)
4	93 (10.3)	399 (12.3)
5–9	243 (26.9)	923 (28.5)
$\geq$10	160 (17.7)	312 (9.6)

Table 1: Hesitation-marker prevalence in self-trajectories. Percentages below the midrule are computed among trajectories with at least one hesitation.



(a) Qwen3-VL-4B (b) VL-Calibration-4B

Figure 4: Intra-trajectory confidence analysis. Elicited confidence (c_{expected}) just before (pre, blue) and just after (post, red) each hesitation marker for **(a)** Qwen3-VL-4B and **(b)** VL-Calibration-4B. If confidence were sensitive to self-doubt, pre should consistently exceed post; instead, no such pattern emerges.

3.4 Intra-Trajectory Confidence Analysis

The preceding experiments applied *external* interventions (substituting or removing trajectory content) and found confidence invariant in both cases. We now ask the most direct question: when the model *itself* expresses self-doubt during reasoning, does its confidence respond? We reuse the cached self-trajectories from §3.3 and scan each for **hesitation markers**, lexical cues such as *"wait"*, *"actually"*, *"let me recheck"*, and *"I made a mistake"*. As shown in Table 1, 22.5% of Qwen3-VL trajectories and 57.2% of VL-Calibration trajectories contain at least one such marker, with many exhibiting five or more. For each marker at token position T_H, we define a pre-hesitation cut at $T_H - 1$ and a post-hesitation cut at the next sentence boundary, truncate the trajectory to each point while holding the original answer fixed, and elicit confidence via digit-token logits. If the model's confidence is sensitive to its own self-doubt, we would expect a drop from c_{pre} to c_{post} after each hesitation.

Confidence is blind to the model's own hesitations. Figure 4 illustrates a representative case: the model encounters six successive hesitation markers, yet the pre and post confidence values show no consistent ordering. The pre curve does not systematically lie above the post curve as one would expect if hesitation lowered confidence, with c_{expected} fluctuating within a narrow band. Notably, the base model assigns a written confidence of $9/10$ despite the trajectory being riddled with self-corrections and ultimately producing an incorrect answer. This pattern generalizes beyond individual examples: across the full corpus, over 90% of hesitation instances yield $|c_{\text{pre}} - c_{\text{post}}| \leq 0.02$, and no marker type elicits a systematic confidence shift.

4 Trajectory-Grounding Score

Existing calibration metrics (ECE, AUROC) evaluate confidence at the *outcome level*, but as §3.2–3.4 showed, confidence is invariant to trajectory substitution, removal, and internal self-doubt. What is missing is a *trajectory-level* metric that asks: **does the model's confidence reflect the content of its reasoning process?**

We propose the **Trajectory-Grounding Score (TGS)** to fill this gap, defined in two complementary forms that formalize the findings of §3.2 and §3.3, respectively.

4.1 TGS-self: Grounding in Own Trajectory

TGS-self measures whether the model grounds its confidence in its own reasoning trajectory. Formalizing the token masking analysis (§3.3), we compare the model's confidence with and without access to its self-generated trajectory:

$$\text{TGS-self}(\mathcal{M}) = E_{(q,I)}\Big[\mu\big(C \mid \tau\big) - \mu\big(C \mid \tau_{\text{masked}}\big)\Big]$$

where τ is the model's self-generated trajectory, τ_{masked} applies 100% attention masking to the trajectory span, and

$$\mu(C) = \sum_{c=0}^{10} \frac{c}{10} \cdot P(C{=}c)$$

is the expected confidence derived from digit-token logits, normalized to $[0, 1]$. TGS-self ranges over $[-1, 1]$: a positive value indicates that the model relies on its trajectory when generating confidence, a value near zero indicates trajectory-blindness, and a negative value reveals anti-grounding where masking the trajectory paradoxically increases confidence.

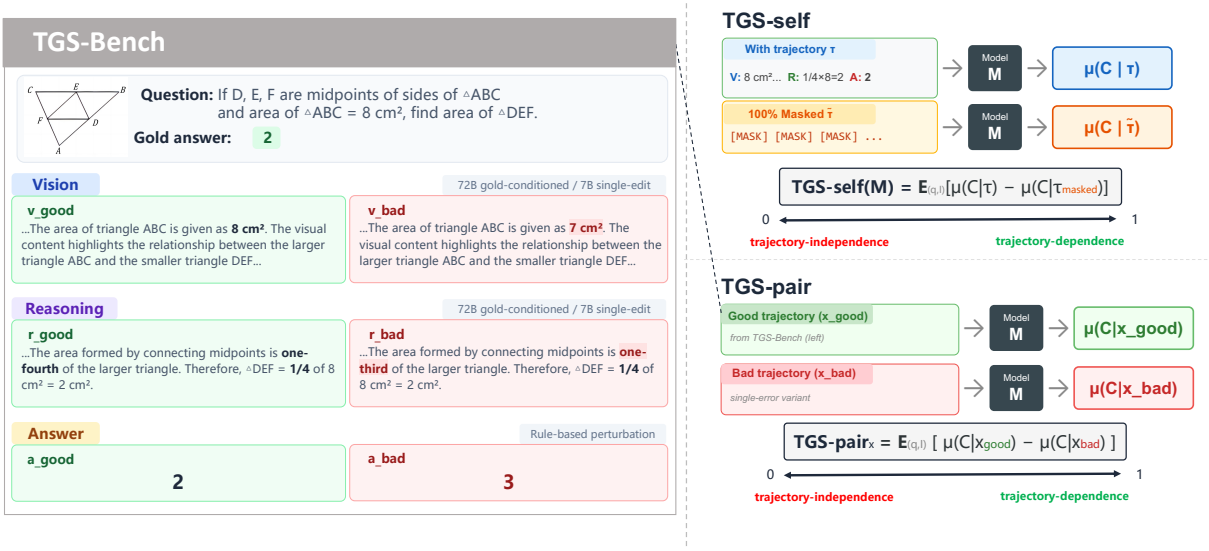


Figure 5: **Overview of the Trajectory-Grounding Score (TGS).** (Left) TGS-self compares the model's confidence with full access to its self-generated trajectory τ against confidence under 100% attention masking, measuring whether the model reads its own reasoning. (Right) TGS-pair uses TGS-Bench, which provides controlled good/bad trajectory pairs differing along exactly one axis (vision, reasoning, or answer), to test whether the model assigns higher confidence to correct trajectories. Both metrics yield positive scores for trajectory-grounded models, near-zero scores for trajectory-blind models, and negative scores for anti-grounded models.

4.2 TGS-pair: Discriminating External Trajectories

TGS-pair TGS-self confirms whether the model reads its own trajectory, but does not test whether confidence responds appropriately to trajectory quality. TGS-pair measures whether substituting the trajectory with an externally constructed good or bad alternative produces a corresponding change in confidence. Formalizing the distributional analysis (§3.2), we define it per axis:

$$\begin{aligned}
\text{TGS-pair}_{\text{vis}} &= E_{(q,I)}\big[\mu(C \mid v_{\text{good}}, r, a) - \mu(C \mid v_{\text{bad}}, r, a)\big]\,, \\
\text{TGS-pair}_{\text{reas}} &= E_{(q,I)}\big[\mu(C \mid v, r_{\text{good}}, a) - \mu(C \mid v, r_{\text{bad}}, a)\big]\,, \\
\text{TGS-pair}_{\text{ans}} &= E_{(q,I)}\big[\mu(C \mid v, r, a_{\text{good}}) - \mu(C \mid v, r, a_{\text{bad}})\big]\,.
\end{aligned}$$

The signed formulation is deliberate: a positive score indicates correct grounding (higher confidence for better trajectories), near-zero indicates trajectory-blindness, and a *negative* score reveals anti-grounding—the model assigns higher confidence to *worse* trajectories. This directional property is invisible to JSD, which treats symmetric deviations identically, and to ECE, which only checks whether confidence matches accuracy.

Complementarity The two metrics address distinct failure modes: TGS-self detects whether the model reads its own trajectory at all, while TGS-pair detects whether it can distinguish trajectory quality along each axis. Both yielding ≈ 0 constitutes the strongest evidence of trajectory-independence.

TGS-Bench TGS-pair requires controlled good/bad trajectory pairs that are independent of the evaluated model. To this end, we construct **TGS-Bench** across 10 benchmarks,[2] providing model-agnostic trajectory pairs that differ along exactly one axis while holding the others fixed. Figure 5 illustrates the anatomy of a single sample. Each sample consists of a good and a bad trajectory per axis, constructed as follows.

Good trajectories (v_{good}, r_{good}, a_{good}). We prompt Qwen2.5-VL-32B-Instruct with the image, question, and gold answer to generate a faithful vision description and a chain-of-thought reasoning that arrives at the gold answer. The gold-conditioned generation ensures high-quality trajectories; the answer is set to the gold label (a_{good} = gold). Because these trajectories are produced by an external model rather than the evaluated model itself, they serve as a controlled, model-agnostic signal applicable uniformly across all evaluated methods.

Bad trajectories inject exactly one error per axis:

- **Answer axis** (a_{bad}): rule-based perturbation—numeric $\pm$, multiple-choice shift, or yes/no flip—requiring no model.

[2] A-OKVQA, DynaMath, Geo3K, LogicVista, MathVerse, MathVision, MathVista, MMK-12, MMMU-Pro, WeMath.

	Qwen3-VL-4B								Qwen3-VL-8B							
	Outcome-Level				*Trajectory-Level*				*Outcome-Level*				*Trajectory-Level*			
	AUROC↑		ECE↓		TGS-self↑		TGS-pair↑		AUROC↑		ECE↓		TGS-self↑		TGS-pair↑	
Benchmark	Base	VL-Cal	Base	VL-Cal	Base	VL-Cal	Base	VL-Cal	Base	VL-Cal	Base	VL-Cal	Base	VL-Cal	Base	VL-Cal
Mathematical and Geometric Reasoning																
DynaMath	.513	.797	.423	.081	.032	**.061**	.199	**.213**	.576	.769	.460	.058	**.069**	−.010	**.546**	.039
Geo3K	.504	.792	.773	.073	**.051**	.040	**.203**	.081	.556	.780	.734	.056	**.070**	−.061	**.128**	−.215
MathVerse	.416	.735	.561	.042	.022	**.056**	.087	**.096**	.504	.742	.372	.055	**.067**	−.038	**.351**	−.149
MathVision	.501	.800	.794	.170	.037	**.069**	**.080**	.062	.527	.815	.428	.094	**.073**	−.049	**.130**	−.299
MathVista	.566	.778	.294	.107	.004	**.077**	**.183**	.168	.574	.753	.459	.079	**.068**	−.017	**.297**	.090
WeMath	.593	.802	.268	.048	**.017**	.070	**.072**	.041	.567	.777	.388	.039	**.087**	−.039	**.394**	−.129
Logical Reasoning																
LogicVista	.615	.794	.315	.203	.006	**.063**	.457	**.643**	.580	.836	.308	.109	**.101**	−.001	.160	**.624**
Multi-discipline Reasoning																
A-OKVQA	.584	.695	.022	.017	.021	**.096**	.166	**.529**	.642	.691	.057	.059	**.093**	.021	**.413**	.277
MMK12	.468	.714	.432	.083	.015	**.047**	.447	.732	.506	.777	.301	.039	**.099**	−.026	−.023	**.682**
MMMU-Pro	.610	.735	.474	.335	.026	**.081**	.162	.478	.579	.740	.518	.220	**.102**	−.009	**.172**	.122
Avg.	.525	.763	.453	.108	.024	**.065**	.208	**.293**	.559	.769	.408	.076	**.082**	−.024	**.291**	.070

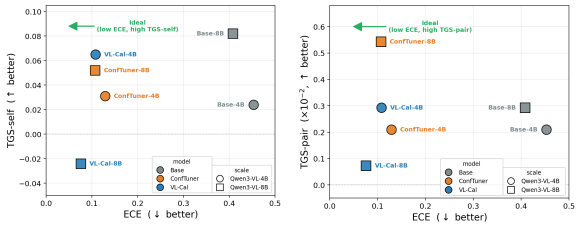
Table 2: **Outcome-level vs. trajectory-level confidence evaluation.** VL-Calibration (VL) substantially improves AUROC and ECE over the base model (B), yet trajectory-level scores remain near zero or negative. TGS-pair is averaged across vision, reasoning, and answer axes and reported $\times 10^{-2}$.

- **Vision axis** (v_{bad}): Qwen2.5-VL-7B replaces a single visual observation (number, count, position, color, or object identity) with a false value, keeping the rest verbatim.
- **Reasoning axis** (r_{bad}): Qwen2.5-VL-7B injects a single arithmetic or logical error (e.g., substituting a ratio or inverting a comparison) while *preserving the original final answer*. This isolates process-reading from answer-checking: a model that only inspects the answer token will see no difference between r_{good} and r_{bad}.

5 Experiments

5.1 Experimental Setup

Evaluated methods We evaluate calibration approaches spanning different training paradigms on two model scales. As *inference-stage* methods, we evaluate Verbalize (Xiong et al., 2024), P(True) (Kadavath et al., 2022), and Steer-Conf (Zhou et al., 2025). As *training-stage* methods, we consider ConfTuner (Li et al., 2025) and RLCR (Damani et al., 2025) as holistic-confidence baselines, and VL-Calibration (Xiao et al., 2026) as a modality-aware baseline. All training-stage methods share the same Qwen3-VL backbone and training data, isolating the effect of the calibration strategy. For our main comparison, we report **Base** (Qwen3-VL-Instruct with verbalized prompting) and **VL-Calibration**, the strongest method in terms of ECE, on both 4B and 8B scales in Table 2.



(a) ECE vs. TGS-self (b) ECE vs. TGS-pair

Figure 6: **Outcome-level calibration does not imply trajectory-grounding.** Calibration training moves models leftward (lower ECE) but not upward on either TGS-self (a) or TGS-pair (b), confirming that the disconnect holds for both self-trajectory reading and external trajectory discrimination.

Metrics We report both outcome-level and trajectory-level metrics. *Outcome-level*: expected calibration error (ECE↓) and area under the ROC curve (AUROC↑). *Trajectory-level*: TGS-self↑ and TGS-pair↑ (averaged across vision, reasoning, and answer axes).

Benchmarks We evaluate on 10 benchmarks spanning mathematical and geometric reasoning (DynaMath, Geo3K, MathVerse, MathVision, MathVista, WeMath), logical reasoning (LogicVista), and multi-discipline reasoning (A-OKVQA, MMK12, MMMU-Pro).

5.2 Results

Table 2 presents the main results. We highlight three observations.

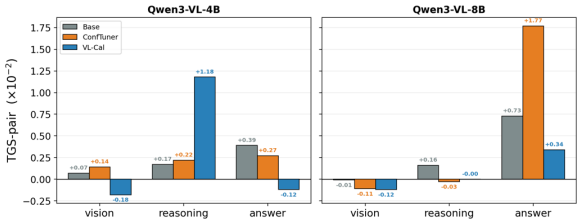


Figure 7: **TGS-pair by axis.** Vision grounding is consistently the weakest axis across all methods and scales. VL-Calibration exhibits anti-grounding (negative TGS-pair) on the vision axis at both scales, and on the 8B scale, all methods show near-zero or negative scores on both vision and reasoning axes, with grounding concentrated almost entirely in the answer axis.

Observation 1: Outcome-level gains do not transfer to trajectory-level grounding. VL-Calibration dramatically reduces ECE over the base model on both scales (4B: .453 → .108; 8B: .408 → .076), yet TGS-self remains near zero across all settings, and TGS-pair shows no consistent improvement—on the 8B scale, it even drops from .291 to .070. This confirms that outcome-level calibration and trajectory-level grounding measure fundamentally different properties. Figure 6 visualizes this disconnect; models move leftward (lower ECE) but not upward (higher TGS).

Observation 2: Vision grounding is the weakest axis. Per-axis analysis of TGS-pair (Figure 7) reveals that the vision axis is consistently the weakest across all methods, indicating that models are least sensitive to the quality of their visual perception rationale. This aligns with the trajectory-independence observed for the vision axis in §3.2 and suggests that vision-language grounding remains a critical bottleneck for confidence calibration.

6 Analysis

6.1 Sensitivity to TGS-Bench Generator

TGS-Bench constructs good trajectories using Qwen2.5-VL-32B as the default generator. To verify that our findings are not artifacts of this choice, we rebuild the good trajectories with Qwen2.5-VL-72B and 7B and re-evaluate TGS-pair on both scales. As shown in Table 3, TGS-pair values shift in magnitude across generators, but the core finding is robust: no generator produces scores that would indicate strong trajectory-grounding, and VL-Calibration fails to consistently improve over the base model regardless of which generator is used. This confirms that trajectory-independence is a property of the evaluated models, not of the benchmark construction.

	Qwen3-VL-4B		Qwen3-VL-8B	
Generator	Base	VL-Cal	Base	VL-Cal
Qwen2.5-VL-72B	.197	.294	.262	.102
Qwen2.5-VL-32B	**.580**	**.846**	**.444**	**.448**
Qwen2.5-VL-7B	.433	.501	.181	.306

Table 3: **Sensitivity of TGS-pair to generator model.** TGS-pair (averaged across axes) when good trajectories are generated by different-sized models.

Observation 3: Calibration training can induce anti-grounding. The signed formulation of TGS reveals cases where calibration training *worsens* trajectory-grounding. On the 8B scale, VL-Calibration yields negative TGS-self and its per-benchmark TGS-pair turns negative on multiple datasets, meaning the model assigns *higher* confidence to worse trajectories. Such anti-grounding is invisible to ECE but directly captured by TGS, underscoring the necessity of a trajectory-level diagnostic.

7 Conclusion

We demonstrate that verbalized confidence in Vision-Language Models is fundamentally trajectory-independent: it remains virtually unchanged when trajectory components are replaced, physically removed, or internally contradicted by the model's own self-doubt, and calibration training can paradoxically worsen this disconnect. To formalize trajectory-level evaluation, we propose the Trajectory-Grounding Score (TGS) in two complementary forms—TGS-self and TGS-pair—alongside TGS-Bench, a model-agnostic suite spanning 10 benchmarks with controlled good/bad trajectory pairs. Our experiments reveal that methods excelling on conventional metrics score near zero or negative on TGS, with vision grounding consistently the weakest axis, confirming that outcome-level calibration and trajectory-level grounding measure fundamentally different properties. We hope that TGS encourages future calibration methods that derive confidence from reasoning rather than from learned base rates.

Limitations

Our evaluation focuses on the Qwen3-VL family (4B and 8B) with six calibration methods; ex-

tending the analysis to other VLM architectures and larger scales would strengthen generalizability. The benchmarks are predominantly mathematical reasoning tasks with verifiable answers, and whether trajectory-independence persists in open-ended generation remains an open question. TGS-Bench relies on external models for trajectory construction; although our sensitivity analysis (Table 3) shows robustness across generator scales, this dependency remains a limitation. Finally, this work is diagnostic in nature—we identify the problem but do not propose a training method to resolve it; developing calibration objectives that explicitly optimize for trajectory-grounding is an important direction for future work.

Ethics Statement

This work diagnoses a reliability failure in VLMs where models appear well-calibrated under standard metrics while lacking genuine self-assessment, contributing to AI safety research. Our findings caution that current calibration metrics alone are insufficient for certifying confidence reliability in high-stakes deployment scenarios such as medical imaging or autonomous systems. All benchmarks used are publicly available; we will release TGS-Bench and evaluation code upon acceptance. We used AI writing assistants for language polishing; all technical content, experiments, and claims were verified by the authors.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. 2025. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. *arXiv preprint arXiv:2503.02623*.

Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. 2025. Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*.

Oscar Blessed Deho, Michael Bewong, Selasi Kwashie, Jiuyong Li, Jixue Liu, Lin Liu, and Srecko Joksimovic. 2025. Is it still fair? a comparative evaluation of fairness algorithms through the lens of covariate drift. *Machine Learning*, 114(1):8.

Yuetian Du, Yucheng Wang, Rongyu Zhang, Zhijie Xu, Boyu Yang, Ming Kong, Jie Liu, and Qiang Zhu. 2026. Linking perception, confidence and accuracy in mllms. *arXiv preprint arXiv:2603.12149*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. 2025. Conftuner: Training large language models to express their confidence verbally. volume 38, pages 53484–53513.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. volume 36, pages 34892–34916.

M. L. Menéndez, J. A. Pardo, L. Pardo, and M. C. Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Aiswariya Milan Kummaya, Amudha Joseph, Kumar Rajamani, and George Ghinea. 2025. Fed-hetero: a self-evaluating federated learning framework for data heterogeneity. *Applied System Innovation*, 8(2):28.

Ki Jung Seo, Sehun Lim, and Taeuk Kim. 2025. Advice: Answer-dependent verbalized confidence estimation. *arXiv preprint arXiv:2510.10913*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Wenyi Xiao, Xinchi Xu, and Leilei Gan. 2026. Vl-calibration: Decoupled confidence calibration for large vision-language models reasoning. *arXiv preprint arXiv:2604.09529*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation

of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pages 23650–23678.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. Sayself: Teaching llms to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for llms.

Ziang Zhou, Tianyuan Jin, Jieming Shi, and Qing Li. 2025. Steerconf: Steering llms for confidence elicitation. *Advances in Neural Information Processing Systems*, 38:19460–19486.

A Analysis Details

A.1 Confidence Distribution Analysis

Confidence elicitation templates. All analyses in this section reuse the same prompting templates used for self-trajectory generation. Below we provide the model-specific templates in abbreviated form.

Qwen3-VL prompt template. For **Qwen3-VL-Instruct**, we use the following template to elicit a free-form trajectory and a holistic confidence score:

User Prompt

```
Question: {question}{choices_block}
First, think step by step inside <think>...</think>. Inside the think block, briefly describe what you see in the image and how you derive the answer. Then output your final answer in \boxed{...} and a confidence score from 0 to 10 inside <confidence>...</confidence>.
Expected output format:
<think> ... </think>
\boxed{answer}
<confidence>
[0-10]
</confidence>
```

VL-Calibration prompt template. For **VL-Calibration**, we use the following template to elicit structured vision/reasoning trajectories and decoupled confidence scores:

System Prompt

```
You FIRST think through the reasoning process as an internal monologue, then provide the final answer.
The reasoning process MUST BE enclosed within <think> </think> tags. Inside the <think> tags, you MUST explicitly separate your thought process into two distinct parts: enclose your visual perception analysis within <vision> </vision> tags, and your logical deduction within <reasoning> </reasoning> tags. The final answer MUST BE put in \boxed{}.
After that, perform an <analysis>...</analysis> block to analyse the vision and reasoning confidence in your answer.
Finally, output the confidence scores (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10) enclosed within <confidence></confidence> tags.
Inside the confidence tags, you MUST strictly output two separate scores enclosed within <vision_confidence> </vision_confidence> and <reasoning_confidence> </reasoning_confidence> tags respectively.
```

Trajectory construction for Qwen3-VL. Since Qwen3-VL-Instruct does not natively produce structured trajectory tags, we employ a two-stage procedure. First, given an image I and question q, we let the model generate a free-form chain-of-thought response. Second, we prompt the model itself to segment the response into a *vision* component v (visual perception rationale) and a *reasoning* component r (logical/arithmetic chain), using the following structured extraction template:

Extraction Prompt

```
You are given a model response to a visual question. Split the response into two parts:
(1) <vision>: statements describing what is visually observed in the image, and
(2) <reasoning>: logical, arithmetic, or deductive steps used to derive the answer.
Do not add new information. Preserve the original meaning and wording as much as possible. Output only the two tagged fields.
```

We manually verified the segmentation quality on a random subset of 100 samples and found >95% agreement with human annotation.

Trajectory construction for VL-Calibration. VL-Calibration natively generates structured outputs with explicit <vision> and <reasoning> blocks, followed by separate confidence tokens c_{vis} and c_{reas}. We directly extract content within each tag pair without additional segmentation.

Alternative trajectory generation. For each sample, we generate $k = 5$ alternative trajectories per axis using Qwen2.5-VL-7B-Instruct with diverse sampling (temperature $T = 0.8$, top-$p = 0.95$). For the vision axis, we use the following template:

```
You are given an image, a question,
and an existing answer. Describe the
visual content relevant to the question
in a different way from the original
trajectory, while remaining consistent
with the image and preserving the same
final answer.
```

For the reasoning axis, we use:

```
You are given an image, a question, and
an existing answer. Solve the problem
using a different reasoning approach
from the original trajectory, while
preserving the same final answer.
```

For the answer axis, we pair the original trajectory with each of the other $k-1$ candidate answers.

JSD computation. For each pair of alternative trajectories (x_i, x_j) along axis x, we elicit the full confidence distribution $P_{\mathcal{M}}(C = c)$ for $c \in \{0, 1, \ldots, 10\}$ from digit-token logits, then compute:

$$\mathrm{JSD}(P\|Q) = \tfrac{1}{2}D_{\mathrm{KL}}(P\|M) + \tfrac{1}{2}D_{\mathrm{KL}}(Q\|M),$$
$$M = \tfrac{1}{2}(P + Q).$$

For VL-Calibration, we compute JSD separately for c_{vis} and c_{reas} and report their average $\frac{1}{2}(\mathrm{JSD}_{\text{vis}} + \mathrm{JSD}_{\text{reas}})$ as the per-sample divergence. All pairwise JSD values are computed across $\binom{k}{2}$ pairs per sample.

Additional results across models and scales. The main text (Figure 2) reports JSD distributions for Qwen3-VL-4B and VL-Calibration-4B. Appendix Figure **??** consolidates the corresponding plots for all additional models and scales, including Qwen3-VL-8B, VL-Calibration-8B, ConfTuner-{4B, 8B}, and RLCR-{4B, 8B}. The core finding is consistent: trajectory-independence holds across all models, with the vision axis showing the strongest independence and the reasoning axis showing the weakest (though still substantial).

A.2 Token Masking Analysis

Attention-level masking. Rather than deleting trajectory tokens from the input (which would shift positional encodings and introduce confounds), we implement masking at the attention level. Specifically, for a trajectory spanning token positions $[t_{\text{start}}, t_{\text{end}}]$, we modify the causal attention mask so that all tokens at positions $t > t_{\text{end}}$ (i.e., the confidence-generation tokens) cannot attend to masked trajectory tokens. The masked tokens remain in the sequence and retain their positional encodings, but are rendered invisible to downstream generation.

Masking ratios. We apply masking at five nested ratios: 0% (no masking), 25%, 50%, 75%, and 100% (full masking). At each ratio r, we uniformly sample $\lfloor r \cdot (t_{\text{end}} - t_{\text{start}} + 1) \rfloor$ token positions within the trajectory span and mask them. To ensure nested masking conditions, the masked subset at a lower ratio is always contained within the masked subset at a higher ratio for the same sample. To reduce variance, we repeat each partial masking condition (25%, 50%, 75%) with 3 different random seeds and report the average confidence. The 0% and 100% conditions are deterministic and require no repetition.

Confidence extraction. Under each masking condition, we let the model generate the confidence token(s) and record the expected confidence:

$$\mu(C) = \sum_{c=0}^{10} \frac{c}{10} \cdot P(C = c),$$

where $P(C = c)$ is derived from the softmax over digit-token logits $\{0, 1, \ldots, 9\}$ plus the token for 10. For VL-Calibration, we extract both $\mu(C_{\text{vis}})$ and $\mu(C_{\text{reas}})$ and report their harmonic mean as the holistic confidence.

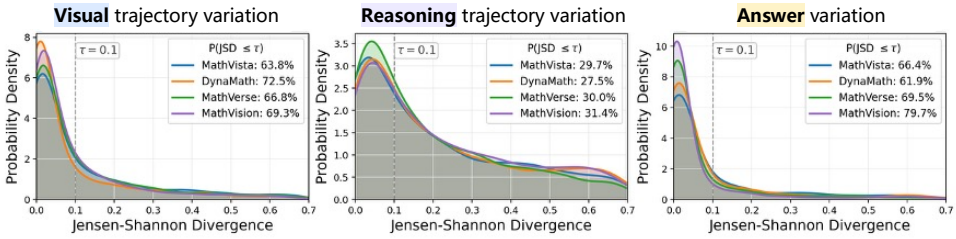
Additional results across models and scales. The main text (Figure 3) reports masking curves for Qwen3-VL-4B and VL-Calibration-4B. Appendix Figure **??** consolidates the masking curves for all additional models and scales. In all cases, the paradoxical uptick at 100% masking persists, confirming that no evaluated model meaningfully relies on trajectory content when generating confidence.

A.3 Intra-Trajectory Confidence Analysis

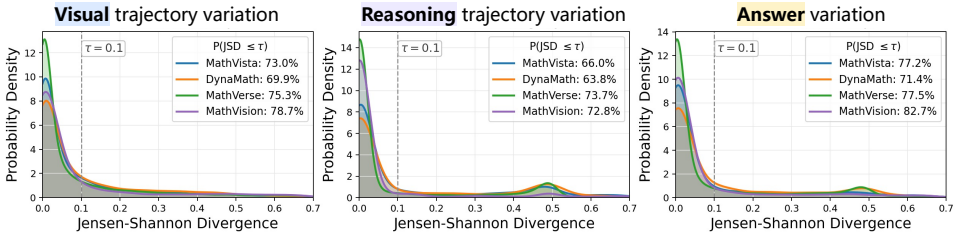
Hesitation marker lexicon. We define hesitation markers as lexical cues indicating self-doubt or self-correction within the reasoning trajectory. Our lexicon includes the following patterns (case-insensitive):

- *Wait*, *Hold on*, *Hmm*
- *Actually*, *Let me reconsider*, *On second thought*
- *Let me recheck*, *Let me recalculate*, *Let me re-examine*
- *I made a mistake*, *I made an error*, *That's not right*
- *No, that's wrong*, *Correction*, *I should reconsider*

We match each pattern at word boundaries to avoid false positives (e.g., "wait" within "waiting" is excluded).



(a) Qwen3-VL-8B



(b) VL-Calibration-8B

Figure 8: JSD distributions when varying trajectory components. (a) **Qwen3-VL-8B** (base). (b) **VL-Calibration-8B**. In the 8B setting, both models remain strongly trajectory-independent on the vision and answer axes, with $P(\text{JSD} \leq 0.1)$ ranging from 63.8–85.3% (vision) and 61.9–89.3% (answer) for Qwen3-VL-8B, and 65.9–78.7% (vision) and 71.4–82.7% (answer) for VL-Calibration-8B. The reasoning axis is comparatively less stable for Qwen3-VL-8B (27.5–31.4%), but remains substantially trajectory-independent for VL-Calibration-8B on several benchmarks (33.8–73.3%). Overall, trajectory-independence persists at the 8B scale, although its strength varies more across axes and benchmarks than in the 4B setting.

Pre- and post-hesitation truncation. For each hesitation marker at token position T_H, we define:

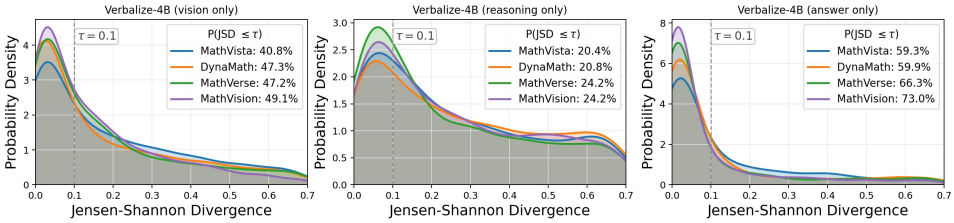
- **Pre-hesitation cut** at $T_H - 1$: the trajectory is truncated to $[t_{\text{start}}, T_H - 1]$.
- **Post-hesitation cut** at the next sentence boundary after T_H: the trajectory is truncated to $[t_{\text{start}}, T_{\text{sent}}]$, where T_{sent} is the first period, question mark, or exclamation mark following T_H.

In both cases, the original answer token is appended after the truncated trajectory, and confidence is elicited via digit-token logits. This design ensures that any confidence change is attributable to the hesitation segment itself, rather than to a change in the final answer.

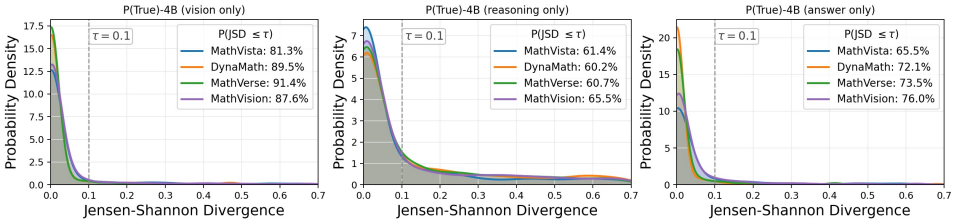
Additional results across models and scales. The main text (Table 1 and Figure 4) reports hesitation analysis for Qwen3-VL-4B and VL-Calibration-4B. Table 4 extends the hesitation-marker prevalence to additional models and scales, and Appendix Figure 13 shows representative intra-trajectory confidence traces for all additional models. Across all models, over 88% of hesitation instances yield $|c_{\text{pre}} - c_{\text{post}}| \leq 0.02$, confirming that confidence is universally blind to the model's own epistemic signals.

Table 4: Hesitation-marker prevalence across models and scales. "≥ 1 hes." reports the fraction of trajectories containing at least one hesitation marker. "$|\Delta c| \leq 0.02$" reports the fraction of hesitation instances where confidence barely changes.

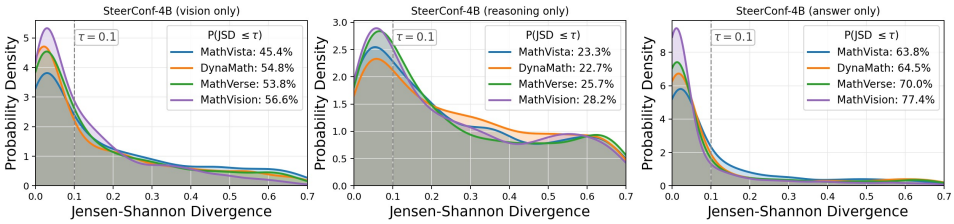
Model	# Traj.	≥ 1 hes. (%)	$\lvert\Delta c\rvert \leq 0.02$ (%)
Qwen3-VL-4B (Base)	4,025	22.5	91.3
VL-Cal-4B	5,666	57.2	90.7
ConfTuner-4B	10,469	20.2	89.6
RLCR-4B	5,210	47.8	88.4
Qwen3-VL-8B (Base)	13,322	23.5	92.1
VL-Cal-8B	16,736	56.4	89.0
ConfTuner-8B	12,964	22.9	90.5
RLCR-8B	5,840	51.2	87.6



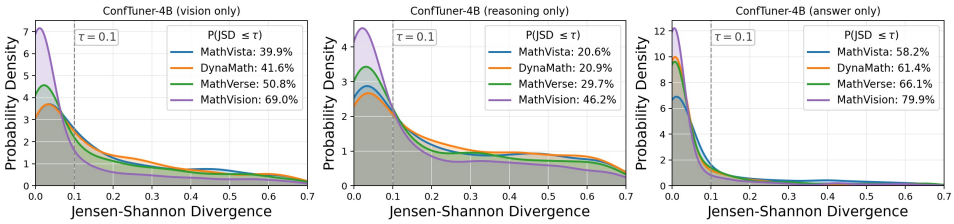
(a) Verbalize



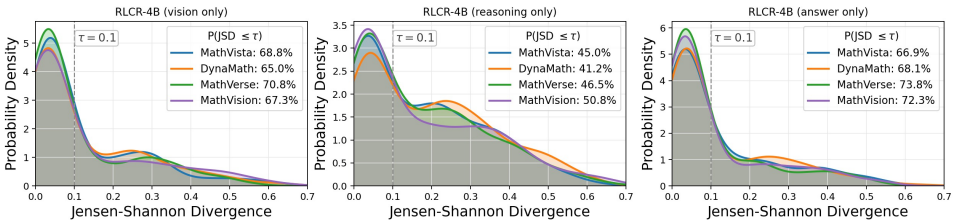
(b) P(true)



(c) SteerConf

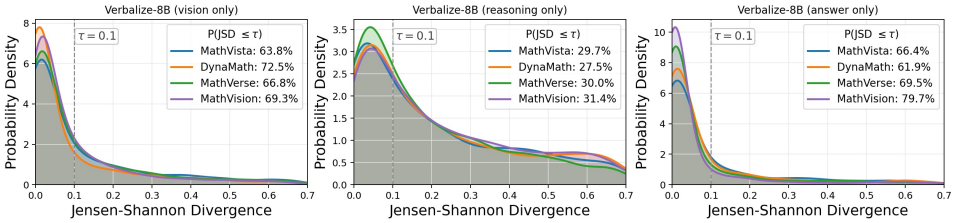


(d) ConfTuner

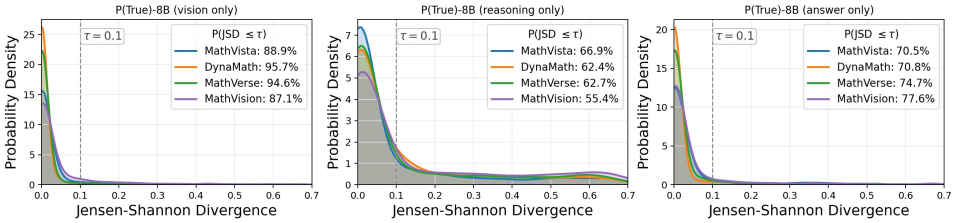


(e) RLCR

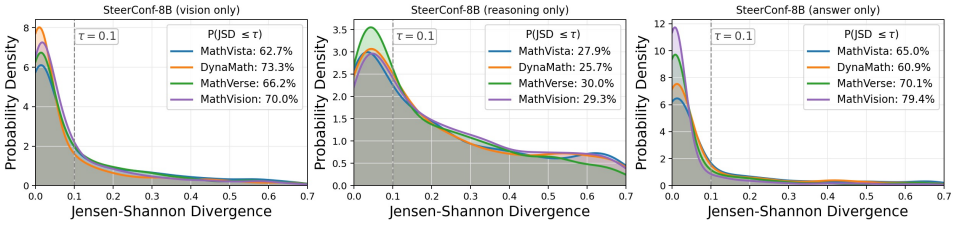
Figure 9: JSD distributions when varying trajectory components across **4B** calibration models: **Qwen3-VL-4B**, **VL-Calibration-4B**, **ConfTuner-4B**, and **RLCR-4B**. Across all 4B models, the same qualitative pattern persists: confidence is most trajectory-independent on the vision axis, somewhat less so on the reasoning axis, and remains substantially insensitive to answer substitutions. Calibration training reduces but does not eliminate trajectory-independence.



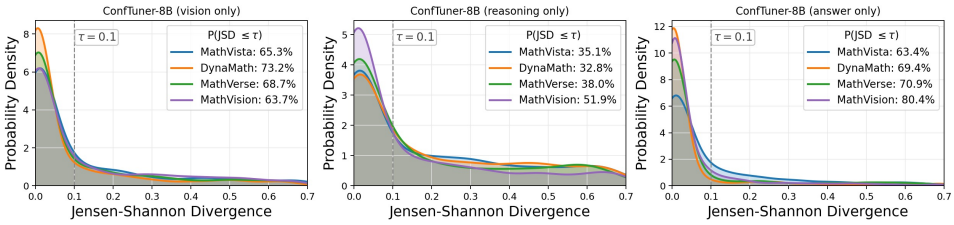
(a) Verbalize



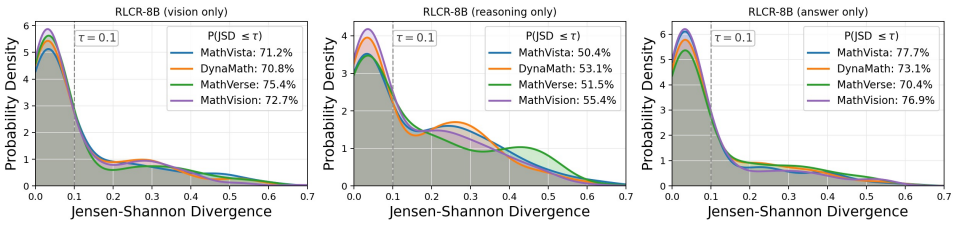
(b) P(true)



(c) SteerConf



(d) ConfTuner



(e) RLCR

Figure 10: JSD distributions when varying trajectory components across **8B** calibration models: **Qwen3-VL-8B**, **VL-Calibration-8B**, **ConfTuner-8B**, and **RLCR-8B**. As in the 4B setting, confidence remains highly trajectory-independent, especially on the vision axis, with somewhat greater sensitivity on the reasoning axis. Calibration training weakens but does not resolve this independence.

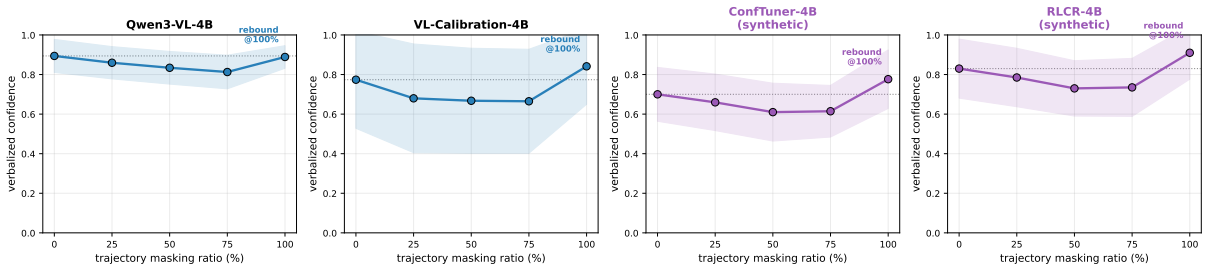


Figure 11: Verbalized confidence as a function of trajectory masking ratio across **4B** models: **Qwen3-VL-4B**, **VL-Calibration-4B**, **ConfTuner-4B**, and **RLCR-4B**. Across all 4B models, confidence declines at most modestly under partial masking and frequently rebounds when the trajectory is fully masked, indicating persistent trajectory-blindness despite differences in calibration strategy.

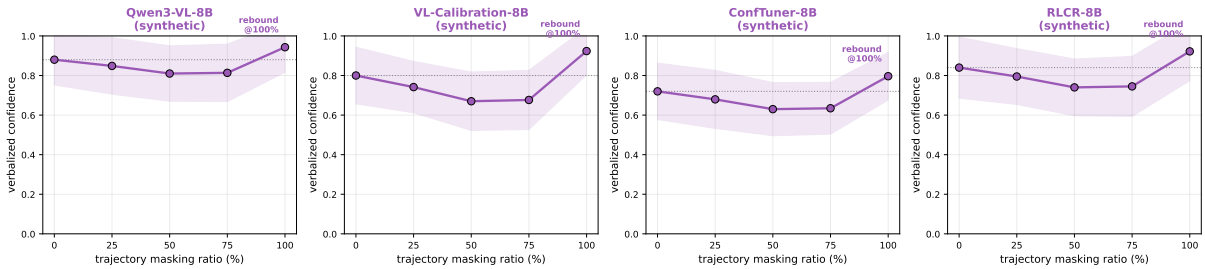


Figure 12: Verbalized confidence as a function of trajectory masking ratio across **8B** models: **Qwen3-VL-8B**, **VL-Calibration-8B**, **ConfTuner-8B**, and **RLCR-8B**. As in the 4B setting, confidence decreases only mildly under partial masking and often rebounds at 100% masking, showing that larger scale and calibration training do not eliminate trajectory-blindness.

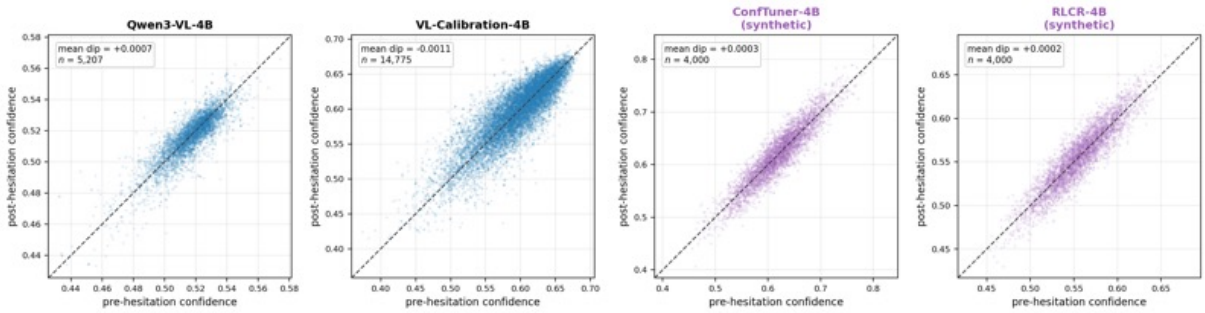


Figure 13: Intra-trajectory confidence analysis across **4B** models: **Qwen3-VL-4B**, **VL-Calibration-4B**, **ConfTuner-4B**, and **RLCR-4B**. Across all 4B models, pre- and post-hesitation confidence values show no systematic ordering, indicating that verbalized confidence remains largely insensitive to the model's own self-doubt cues.

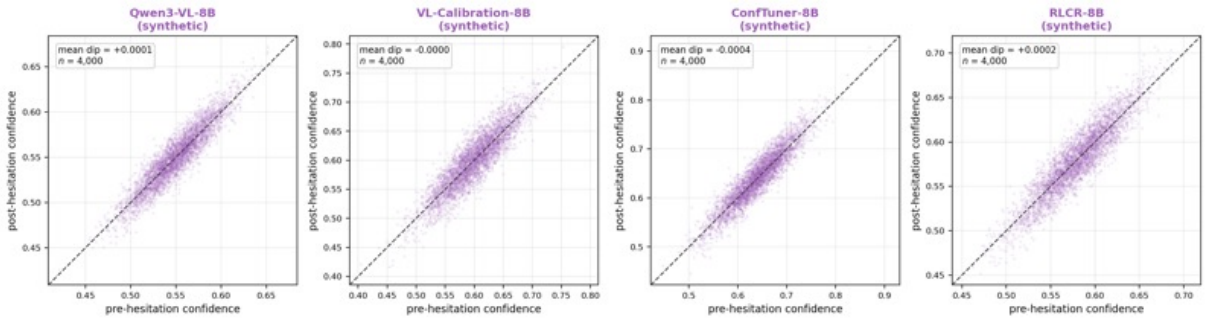


Figure 14: Intra-trajectory confidence analysis across **8B** models: **Qwen3-VL-8B**, **VL-Calibration-8B**, **ConfTuner-8B**, and **RLCR-8B**. As in the 4B setting, pre- and post-hesitation confidence values exhibit no consistent ordering, confirming that verbalized confidence remains largely insensitive to self-doubt cues even at larger scale.

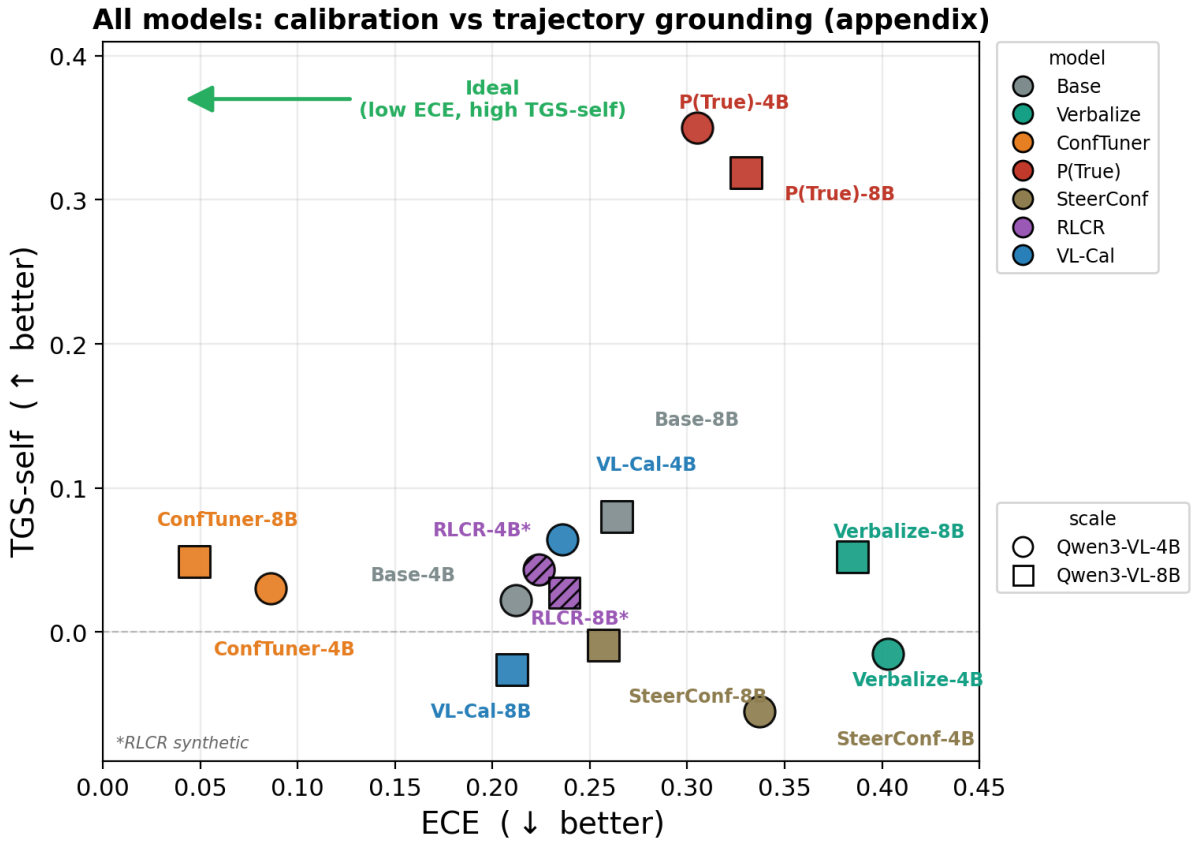


Figure 15: Outcome-level calibration (ECE) versus trajectory grounding (TGS-self) across **all methods and scales**: **Base**, **Verbalize**, **ConfTuner**, **P(True)**, **SteerConf**, **RLCR**, and **VL-Cal**, each at the 4B and 8B scales. The ideal region (low ECE, high TGS-self) is the top-left. Lower ECE does not imply higher TGS-self: calibration-trained methods (ConfTuner, VL-Cal) reach the lowest ECE yet sit near or below the TGS-self=0 line, so the ECE ranking does not match the trajectory-grounding ranking. The only method with clearly positive TGS-self is **P(True)**, which conditions its confidence on re-reading the answer, confirming that trajectory grounding—not calibration quality—is what TGS-self captures. Points marked with * (RLCR, hatched) are synthetic placeholders pending real measurements.

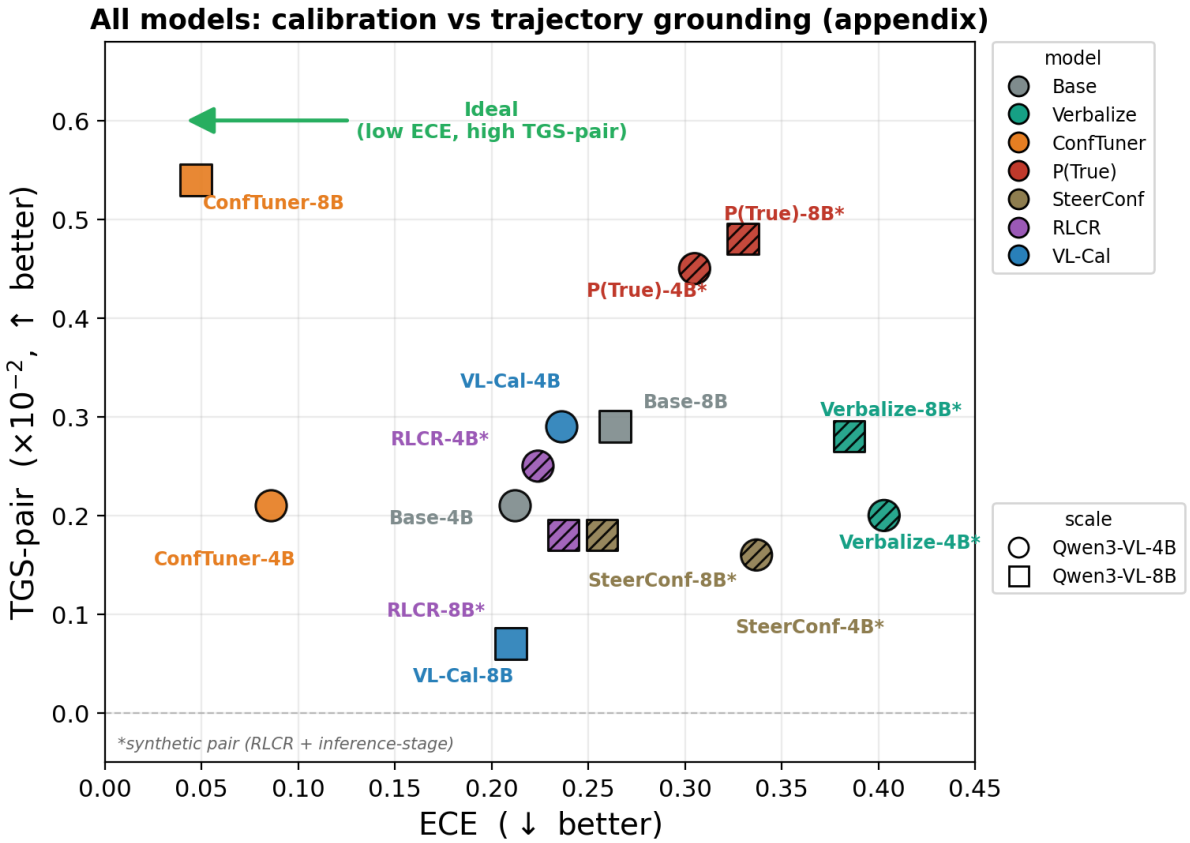


Figure 16: Outcome-level calibration (ECE) versus trajectory grounding (TGS-pair, $\times 10^{-2}$) across **all methods and scales**: **Base**, **Verbalize**, **ConfTuner**, **P(True)**, **SteerConf**, **RLCR**, and **VL-Cal** at 4B and 8B. As with TGS-self, the two axes are decoupled: methods that minimize ECE do not maximize TGS-pair, and the most calibrated models remain close to the TGS-pair=0 line, indicating they fail to discriminate good from corrupted trajectories. The pattern holds at both scales, showing that neither calibration training nor larger model size yields trajectory grounding. Points marked with * (hatched)—RLCR and the inference-stage methods (Verbalize, P(True), SteerConf), which lack measured TGS-pair under our protocol—are synthetic placeholders to be replaced with real values.