TRACECONF: VISUALLY TRACEABLE AND CALIBRATED CONFIDENCE IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As multimodal large language models are increasingly used, confidence estimates are becoming central to assessing the reliability of their predictions. Existing calibration objectives, however, primarily ask whether an answer is likely to be correct, without revealing which visual evidence warrants that confidence. This limitation arises because binary correctness assigns the same calibration target to predictions with different evidential bases, allowing confidence to be statistically calibrated while remaining visually untraceable. We introduce TRACECONF, a post-training approach that jointly generates declared visual evidence, an answer, and a confidence choice. TRACECONF counterfactually assesses the necessity and sufficiency of the declared regions by removing and isolating them. The resulting evidence gate rewards evidence-grounded response generation and weights a local proper confidence loss, while the confidence-choice token is excluded from the shared sequence-level GRPO advantage to prevent response-level rewards from misassigning credit to confidence. Because the gate is determined before the confidence report, evidence weighting preserves the proper conditional target wherever the gate is positive. Across six multimodal benchmarks, TRACECONF improves confidence calibration on average and error discrimination across all benchmarks while maintaining competitive answer accuracy. Further experiments establish a direct link between model confidence and its declared visual evidence: removing the declared regions causes larger confidence drops, while retaining only these regions better preserves confidence than retaining their complementary regions. Together, these results demonstrate that visually traceable confidence can enhance multimodal reliability without compromising answer quality.

1 INTRODUCTION

As multimodal large language models (MLLMs) are increasingly used for complex visual reasoning, determining when their predictions can be trusted has become a central reliability challenge. Confidence estimates provide a natural signal for this purpose, yet MLLMs often remain overconfident and poorly calibrated across visual reasoning tasks (Groot & Valdenegro-Toro, 2024; Xuan et al., 2025). Confidence calibration addresses this problem by aligning reported confidence with the empirical likelihood of answer correctness (Guo et al., 2017). Most existing formulations, however, use binary answer correctness as the primary calibration target. Although this target indicates whether an answer is correct, it does not reveal which visual evidence supports the associated confidence.

This limitation matters in multimodal reasoning because MLLMs can arrive at correct answers through language priors without relying on task-relevant image evidence (Lin et al., 2024; Luo et al., 2025). Consequently, two predictions may receive the same correctness target despite relying on fundamentally different evidential bases. Calibration against correctness alone may therefore produce confidence that is statistically reliable at the population level while leaving its visual basis untraceable.

Figure 1 makes this distinction concrete. An answer-only calibration objective can regard a high-confidence correct prediction as well calibrated without revealing whether it relies on relevant visual information. Once the model is required to declare its evidence, however, a supported answer can be

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.

distinguished from an equally correct answer whose declared region does not justify the prediction. This observation motivates our central question: *Can MLLMs learn confidence that is not only calibrated with respect to answer correctness, but also warranted by their declared visual evidence?* To address this question, we introduce TRACECONF, a post-training approach to learning visually traceable multimodal confidence. TRACECONF first establishes a structured output protocol in which the model generates a reasoning trace, declared visual-evidence regions, an answer, and a binary confidence choice. It then evaluates the declared regions using counterfactual necessity and sufficiency tests. Specifically, removing the declared regions measures whether they are necessary for supporting the answer, while retaining only those regions measures whether they are sufficient. These signals are combined into a continuous evidence gate that rewards responses supported by their declared evidence and determines the strength of local confidence supervision.

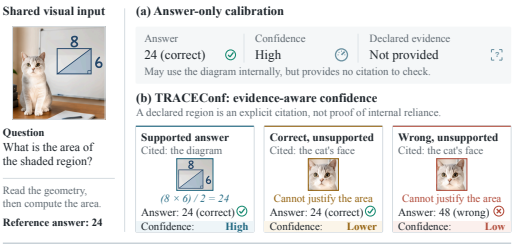


Figure 1: Answer-only calibration reports confidence without exposing its visual basis. TRACECONF additionally declares visual evidence, distinguishing supported answers from predictions whose declared regions do not justify the answer.

A second challenge arises from credit assignment during reinforcement learning. Standard Group Relative Policy Optimization (GRPO) assigns a shared completion-level advantage to all generated tokens (Shao et al., 2024). As a result, rewards for answer accuracy, evidence quality, or output format may reinforce the confidence choice even when that choice is itself incorrect. TRACECONF therefore excludes the confidence-choice token from the sequence-level GRPO objective and supervises it directly against answer correctness using a local binary logarithmic loss. The evidence gate modulates the strength of this supervision rather than changing its correctness target: weak evidence reduces the confidence-learning signal but is not treated as evidence that the answer is incorrect. Because the gate is determined before the confidence report and does not depend on that report, positive evidence weighting preserves the proper conditional target.

We evaluate TRACECONF on six multimodal benchmarks spanning mathematical, grounded, dynamic, and general visual reasoning. Compared with the strongest baseline, TRACECONF reduces macro-average ECE from 0.0951 to 0.0840 and improves macro-average AUROC from 0.7489 to 0.7982, while maintaining competitive answer accuracy. Further experiments establish a direct connection between confidence and model-declared visual evidence: removing the declared regions causes larger confidence drops, whereas retaining only these regions better preserves confidence than retaining the complementary image regions. These results show that TRACECONF improves both the statistical reliability and visual traceability of multimodal confidence.

Our main contributions are:

- We identify a limitation of correctness-based multimodal confidence calibration: binary correctness cannot distinguish predictions with different visual evidence bases, leaving confidence potentially calibrated but visually untraceable.
- We propose TRACECONF, which counterfactually assesses the necessity and sufficiency of model-declared evidence and uses the resulting gate to couple evidence-grounded response optimization with local proper confidence learning, while isolating the confidence choice from the shared sequence-level GRPO advantage.
- Across six multimodal benchmarks, we show that TRACECONF improves confidence calibration and error discrimination while maintaining competitive accuracy, and that its confidence is specifically associated with the visual evidence declared by the model.

2 RELATED WORK

2.1 TRAINING-FREE CONFIDENCE ESTIMATION

Without confidence-specific training, confidence can be estimated from token or sequence probabilities (Kadavath et al., 2022), agreement across sampled reasoning paths (Wang et al., 2023), or

semantic entropy over meaning-equivalent outputs (Farquhar et al., 2024). Multimodal work further studies verbalized uncertainty and develops evidential or perturbation-based uncertainty signals (Xuan et al., 2025; Huang et al., 2026). These approaches produce scalar confidence or uncertainty estimates while generally leaving their answer-specific visual basis implicit and unavailable for direct region-level inspection.

2.2 Post-Training Confidence Calibration

A complementary line of work explicitly trains models to express calibrated confidence. ConfTuner optimizes tokenized Brier scores (Li et al., 2025), while Rewarding Doubt, LoVeC, and RLCR combine task learning with proper confidence objectives (Bani-Harouni et al., 2026; Zhang et al., 2026; Damani et al., 2026). For multimodal models, VL-Calibration separates visual and reasoning confidence, MMBoundary supervises confidence at individual reasoning steps, and CDRL calibrates confidence through reinforcement learning over visually perturbed inputs (Xiao et al., 2026; He et al., 2025; Du et al., 2026). These methods improve how confidence is learned or expressed, but do not test whether visual regions declared for the generated answer provide counterfactual support. TraceConf couples confidence learning with necessity and sufficiency tests of declared evidence, while isolating the confidence choice from the shared sequence-level GRPO advantage.

A broader discussion of classical calibration, confidence elicitation, learning-based calibration, and visual grounding is provided in Appendix E.

3 Preliminaries

3.1 Problem Setup

Let $\pi_\theta(\cdot \mid x, q)$ denote the autoregressive distribution over output token sequences induced by a multimodal large language model conditioned on an image x and a question q. We represent the structured completion as $o = (r, e, a, s)$, where r is the reasoning trace, e is a set of declared visual-evidence regions, a is the final answer, and $s \in \{\texttt{yes}, \texttt{no}\}$ is the binary confidence choice.

Let z_{yes} and z_{no} denote the logits at the confidence-choice position; we define the reported confidence as

$$\hat{p} := \frac{\exp(z_{\text{yes}})}{\exp(z_{\text{yes}}) + \exp(z_{\text{no}})}. \tag{1}$$

The correctness of a generated answer a against the reference answer $a^\star$ is encoded by $y := \mathbf{1}\{a \equiv a^\star\} \in \{0, 1\}$, where $\equiv$ denotes the task-specific answer-matching criterion.

Given the conditioning context $h = (x, q, r, e, a)$ preceding the confidence choice, the probabilistic target of $\hat{p}$ is

$$p^\star(h) := \Pr(Y = 1 \mid H = h), \tag{2}$$

where H and Y are the random counterparts of h and y, respectively.

Across inputs and sampled completions, the reported confidence $\hat{p}$ induces a random variable $\hat{P}$. The confidence reports are calibrated if $\mathbb{E}[Y \mid \hat{P}] = \hat{P}$.

3.2 Proper Losses for Confidence Estimation

A binary loss $\ell(\hat{p}, y)$ is strictly proper if, for every conditioning context h, its conditional risk is uniquely minimized by the true conditional correctness probability $p^\star(h)$:

$$\arg\min_{\hat{p} \in [0,1]} \mathbb{E}\left[\ell(\hat{p}, Y) \mid H = h\right] = p^\star(h). \tag{3}$$

We use the binary logarithmic loss $\ell_{\log}(\hat{p}, y) = -y \log \hat{p} - (1 - y) \log(1 - \hat{p})$, which is strictly proper for binary outcomes.

4 METHODOLOGY

4.1 OVERVIEW

As illustrated in Figure 2, TRACECONF is trained in two stages. Supervised fine-tuning first establishes the structured output protocol, under which the model generates a reasoning trace, declared visual evidence, an answer, and a binary confidence choice in a single autoregressive sequence. In the second stage, structured response generation and confidence estimation are optimized through distinct learning objectives. All completion tokens except the confidence-choice token are optimized by sequence-level GRPO using accuracy, format, and evidence rewards, whereas the confidence-choice token is excluded from the policy-gradient loss and directly supervised against answer correctness using a local binary logarithmic loss. A shared evidence gate, computed by a frozen vision–language scorer through counterfactual interventions, serves as the sequence-level evidence reward and weights the local confidence loss according to how strongly the declared evidence supports the generated answer. We refer to this design as evidence-gated local proper confidence learning; the evidence gate, role-specific objectives, and properness of the gated confidence objective are detailed below.

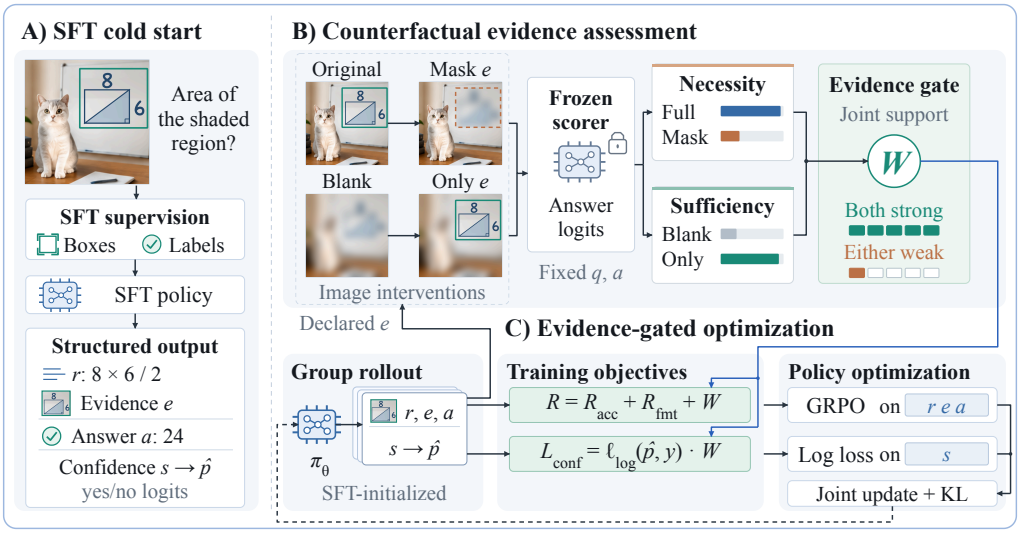


Figure 2: TRACECONF overview. Supervised fine-tuning first establishes a structured output interface with evidence boxes, answers, and confidence tokens. Evidence-gated GRPO then uses counterfactual necessity and sufficiency checks from a frozen scorer to gate calibration credit during policy optimization.

4.2 STRUCTURED SUPERVISED FINE-TUNING

Before reinforcement learning, we perform structured supervised fine-tuning to initialize the model's ability to produce a reasoning trace, declared visual-evidence regions, an answer, and a binary confidence choice within a unified response. For each image–question pair (x_i, q_i) with reference answer $a_i^\star$, the base model generates a reasoning trace r_i and a candidate answer a_i, while a teacher vision–language model conditioned on $(x_i, q_i, a_i^\star)$ annotates one or more visual-evidence regions e_i. We retain examples with nonempty evidence annotations for which a frozen scorer detects measurable visual dependence; the same scorer scores how well each annotated region supports the reference answer. The correctness label $y_i = \mathbf{1}\{a_i \equiv a_i^\star\}$ determines the target confidence choice $s_i^\star = \texttt{yes}$ if $y_i = 1$ and $s_i^\star = \texttt{no}$ otherwise. The resulting SFT target serializes $(r_i, e_i, a_i, s_i^\star)$ according to the structured output protocol defined in Section 3.

To learn from incorrect candidates without reinforcing their generated content, we apply a correctness-dependent token mask. Let $o_i^{\mathrm{SFT}} = (o_{i,1}^{\mathrm{SFT}}, \ldots, o_{i,T_i}^{\mathrm{SFT}})$ denote an SFT target, let t_i^{conf} denote the position of its confidence-choice token, and let $\mathcal{T}_i^{\mathrm{gen}}$ be all completion positions except the confidence choice. The masked SFT objective first averages generation-token cross-entropy within

each example, gates it by correctness, adds a weighted binary loss on the final confidence token, and then averages across the batch:

$$\mathcal{L}_{\mathrm{SFT}} = \frac{1}{B}\sum_{i=1}^{B}\left[\lambda_{\mathrm{sol}}\cdot y_i\cdot\frac{1}{|\mathcal{T}_i^{\mathrm{gen}}|}\sum_{t\in\mathcal{T}_i^{\mathrm{gen}}}\ell_{\mathrm{CE}}\big(o_{i,t}^{\mathrm{SFT}}\mid x_i,q_i,o_{i,<t}^{\mathrm{SFT}}\big)+\lambda_{\mathrm{conf}}\cdot w_{y_i}\cdot\ell_{\mathrm{CE}}^{\mathrm{bin}}\Big(o_{i,t_i^{\mathrm{conf}}}^{\mathrm{SFT}}\Big)\right], \tag{4}$$

where $\ell_{\mathrm{CE}}^{\mathrm{bin}}$ is the binary cross-entropy over the yes/no logits at the confidence position, w_{y_i} balances the yes/no classes, and $\lambda_{\mathrm{sol}}, \lambda_{\mathrm{conf}}$ weight the two terms. When $y_i = 0$, the generation term vanishes and only the confidence token is supervised; its reasoning, evidence, and answer remain in the conditioning prefix but are not used as token-level imitation targets. The resulting policy π_{SFT} initializes reinforcement learning and serves as the fixed reference policy for KL regularization.

4.3 Counterfactual Assessment of Declared Evidence

Given a generated answer a and its declared evidence e, we assess whether e counterfactually supports a using a frozen vision–language model f. Let $U(e)$ denote the union of the declared regions. We construct four image conditions: the original image x_{full}, an image x_{blur} with $U(e)$ blurred, an evidence-only image x_{only} with its complement blurred, and a fully blurred reference x_{blank}. The corresponding answer logits are

$$c_v = f(x_v, q, a), \qquad v \in \{\mathrm{full}, \mathrm{blur}, \mathrm{only}, \mathrm{blank}\}, \tag{5}$$

where $f(x_v, q, a)$ denotes the frozen model's output logit for candidate answer a under image condition x_v.

We quantify necessity by the logit drop when the declared regions are removed, and sufficiency by the logit recovery when only those regions are retained:

$$\Delta_{\mathrm{nec}} = c_{\mathrm{full}} - c_{\mathrm{blur}}, \qquad \Delta_{\mathrm{suf}} = c_{\mathrm{only}} - c_{\mathrm{blank}}. \tag{6}$$

Each signal is mapped to $[0, 1)$ by

$$w(\Delta;\tau) = \frac{[\Delta]_+}{[\Delta]_+ + \tau}, \qquad [\Delta]_+ = \max(0, \Delta), \tag{7}$$

and the final evidence gate is

$$W = w(\Delta_{\mathrm{nec}};\tau_{\mathrm{nec}})\, w(\Delta_{\mathrm{suf}};\tau_{\mathrm{suf}}). \tag{8}$$

The product assigns substantial weight only when the declared regions are both necessary for and sufficient to support the generated answer.

4.4 Evidence-Gated Local Proper Confidence Learning

Standard GRPO assigns a single completion-level advantage to every generated token. This misattributes credit when the confidence choice is miscalibrated: a correct answer yields a positive advantage that also reinforces the confidence token, even if its reported confidence is unjustified. We address this by excluding the confidence choice from the sequence objective and supervising it with a local, evidence-gated loss.

Sequence-level response optimization. For each input, we sample G completions and assign each completion o_i the reward and centered advantage

$$R_i = R_{\mathrm{format},i} + R_{\mathrm{accuracy},i} + W_i, \qquad A_i = R_i - \frac{1}{G}\sum_{j=1}^{G} R_j, \tag{9}$$

where $R_{\mathrm{accuracy},i} = y_i$, $R_{\mathrm{format},i}$ measures compliance with the structured output protocol, and W_i is the evidence gate from Section 4.3. Unlike standard GRPO (Shao et al., 2024), we omit standard-deviation scaling and apply the clipped surrogate only to non-confidence tokens, using $m_{i,t}^{\mathrm{seq}} = 0$ at the confidence-choice position and $m_{i,t}^{\mathrm{seq}} = 1$ elsewhere. We denote the resulting masked loss by $\mathcal{L}_{\mathrm{seq}}$. Thus, sequence-level rewards optimize the structured response without directly reinforcing its confidence choice.

Local confidence optimization. For completions that reach the confidence-choice position, we directly supervise that choice using

$$\mathcal{L}_{\text{local-conf}} = \frac{1}{|\mathcal{I}_{\text{conf}}|} \sum_{i \in \mathcal{I}_{\text{conf}}} W_i \, \ell_{\log}(\hat{p}_i, y_i), \tag{10}$$

where $\mathcal{I}_{\text{conf}}$ indexes such completions. The proper loss targets answer correctness directly, while W_i modulates supervision strength without changing the target; unlike $\mathcal{L}_{\text{seq}}$, it is independent of the group-relative advantage A_i. The overall objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seq}} + \lambda \mathcal{L}_{\text{local-conf}} + \beta \mathcal{L}_{\text{KL}}$, where $\mathcal{L}_{\text{KL}}$ regularizes the policy against the fixed SFT reference at every generated position, including the confidence choice.

Properness under evidence gating. For a fixed conditioning context $h = (x, q, r, e, a)$, the gate $W(h)$ is determined before the confidence choice and is independent of the reported confidence. Therefore, whenever $W(h) > 0$,

$$\arg\min_{\hat{p} \in [0,1]} \mathbb{E}[W(H)\ell_{\log}(\hat{p}, Y) \mid H = h] = p^{\star}(h). \tag{11}$$

Conditioned on $H = h$, the positive factor $W(h)$ only rescales the strictly proper conditional risk and therefore leaves its unique minimizer unchanged. When $W(h) = 0$, the local objective provides no confidence supervision for that context.

5 Experiments

In this section, we first describe the experimental setup, then investigate whether TRACECONF produces reliable confidence estimates and whether these estimates can be traced to explicit visual evidence. Finally, we conduct ablation studies to examine how the proposed training objective affects confidence calibration and visual traceability.

5.1 Experimental Setup

Datasets. Following VL-Calibration (Xiao et al., 2026), we construct our training data from the ViRL-39K (Wang et al., 2025). We first perform structured supervised fine-tuning to establish the joint generation of visual evidence, answers, and confidence estimates, and then conduct GRPO training on prompts with verifiable answers.

For evaluation, we use six benchmarks covering complementary multimodal reasoning settings. MATH-Vision (Wang et al., 2024) and MathVista (Lu et al., 2024) evaluate mathematical reasoning across diverse visual contexts, while MathVerse-V (Zhang et al., 2024) focuses on problems whose solutions depend explicitly on information contained in the image. DynaMath-2V (Zou et al., 2025) further evaluates the robustness of mathematical reasoning over dynamically generated visual problems. Beyond visual mathematics, MMStar (Chen et al., 2024) assesses a broad range of vision-indispensable multimodal capabilities, and MMMU-Pro-v2 (Yue et al., 2025) evaluates expert-level multimodal understanding and reasoning across multiple disciplines.

Baselines. We compare TRACECONF with five representative confidence elicitation and calibration methods: CoT and TopK (Xiong et al., 2024), ConfTuner (Li et al., 2025), Rewarding Doubt (Bani-Harouni et al., 2026), and VL-Calibration (Xiao et al., 2026). CoT and TopK are training-free confidence elicitation methods, while the remaining methods explicitly optimize confidence through training. Each method uses its native confidence readout and is evaluated under the same judging protocol.

Evaluation Metrics. Following prior work on confidence calibration and uncertainty estimation (Naeini et al., 2015; Guo et al., 2017; Xiong et al., 2024; Xiao et al., 2026), we report accuracy (ACC), Expected Calibration Error (ECE), and Area Under the Receiver Operating Characteristic Curve (AUROC). ACC measures task performance, ECE measures the discrepancy between predicted confidence and empirical accuracy, and AUROC measures how well confidence ranks correct predictions above incorrect ones.

Table 1: Main results under the TRACECONF evaluation protocol. Bold marks the best result for each dataset and metric, and underlining marks the second best. The average row reports macro averages across all six benchmarks.

Dataset	Metric	CoT	TopK	ConfTuner	RewardingDoubt	VL-Calib.	TRACECONF
MATH-Vision	ACC ↑	0.5600	**0.6130**	0.6010	0.6020	0.5120	0.6075
	ECE ↓	0.1577	0.1765	0.1982	0.2680	0.0962	**0.0436**
	AUROC ↑	0.6308	0.7159	0.7935	0.7699	0.6631	**0.8678**
MathVerse-V	ACC ↑	0.7262	0.7391	0.7365	0.7095	0.6915	**0.7545**
	ECE ↓	0.1299	0.1374	0.1538	0.2090	**0.1037**	0.1203
	AUROC ↑	0.5643	0.7040	0.6897	0.6617	0.5926	**0.7804**
DynaMath-2V	ACC ↑	0.8211	0.8244	**0.8448**	0.8308	0.8050	0.8160
	ECE ↓	0.1164	0.1030	0.0841	0.1129	**0.0638**	0.1209
	AUROC ↑	0.6615	0.6756	0.7541	0.7379	0.6533	**0.7760**
MMStar	ACC ↑	0.7380	0.7400	0.7467	0.7100	0.7127	**0.7540**
	ECE ↓	0.1893	0.1577	0.1588	0.1890	0.1299	**0.0943**
	AUROC ↑	0.7205	0.7010	0.7242	0.6489	0.6331	**0.7282**
MathVista	ACC ↑	0.8150	0.8230	**0.8300**	0.7890	0.7990	0.7910
	ECE ↓	0.1170	0.1060	0.0710	0.1350	**0.0470**	0.0680
	AUROC ↑	0.7320	0.7780	0.7950	0.7830	0.6520	**0.8700**
MMMU-Pro-v2	ACC ↑	0.6790	0.6930	**0.7100**	0.6850	0.6430	0.6800
	ECE ↓	0.1720	0.1700	0.1480	0.1780	0.1300	**0.0570**
	AUROC ↑	0.7230	0.7370	0.7370	0.6890	0.6730	**0.7670**
Average	ACC ↑	0.7232	0.7388	**0.7448**	0.7211	0.6939	0.7338
	ECE ↓	0.1471	0.1418	0.1357	0.1820	0.0951	**0.0840**
	AUROC ↑	0.6720	0.7186	0.7489	0.7151	0.6445	**0.7982**

To assess whether confidence tracks visual support, we report Support Gap, Pairwise Preference, and High-confidence Masked Rate. To test whether this dependence is specific to the model-declared evidence, we further report Necessity Drop and Sufficiency Gap, formally defined in Section 5.3.

5.2 CAN TRACECONF LEARN EFFECTIVE VERBALIZED CONFIDENCE ESTIMATION CAPABILITIES?

Table 1 summarizes task accuracy, confidence calibration, and correct–incorrect discrimination across six multimodal benchmarks. TRACECONF achieves the highest AUROC on all six benchmarks. Averaged across the six benchmarks, it attains an ECE of 0.0840 and an AUROC of 0.7982, outperforming the corresponding best baseline averages of 0.0951 and 0.7489, respectively. Its average accuracy is 0.7338, compared with the best baseline accuracy of 0.7448.

The most substantial improvements are observed on MATH-Vision, where TRACECONF reduces ECE from 0.0962, the best baseline value, to 0.0436 and improves AUROC from 0.7935 to 0.8678. It also achieves the lowest ECE on MMStar and MMMU-Pro-v2. Overall, these results show that TRACECONF improves confidence calibration and correct–incorrect discrimination while maintaining task accuracy comparable to existing baselines. Figure 3 provides a complementary view of these results: its confidence more closely tracks empirical accuracy across bins, while correct and incorrect predictions are more clearly separated across the confidence range.

5.3 DOES TRACECONF CONFIDENCE TRACK VISUAL SUPPORT?

The result of confidence calibration does not establish whether confidence is grounded in visual evidence. We examine this question at two levels: whether confidence responds when relevant visual information is weakened, and whether this response is specifically tied to the evidence locations declared by the model.

Confidence responds to reduced visual support. To isolate the effect of visual support from answer correctness, we pair each original input x_i^+ with an evidence-blurred input x_i^-. For each

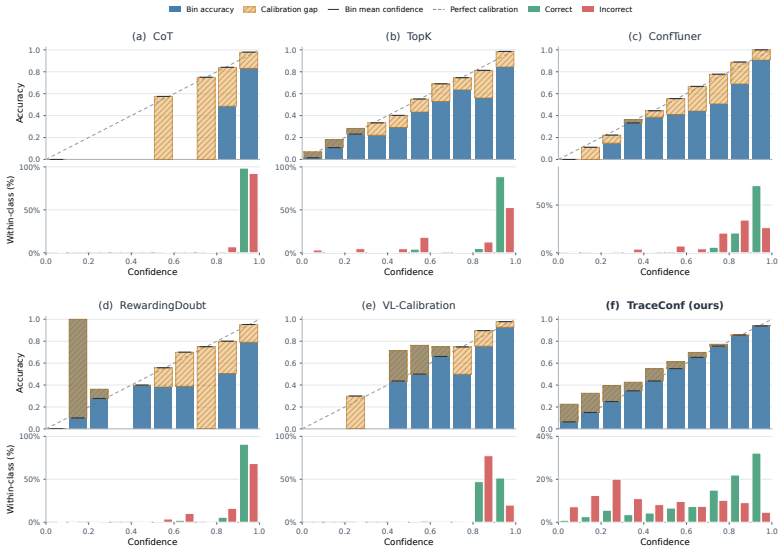


Figure 3: **Reliability diagrams and confidence distributions across all six benchmarks.** In each upper panel, blue bars show empirical accuracy within confidence bins, the black line shows mean predicted confidence, and the dashed diagonal indicates perfect calibration. Hatched regions visualize the bin-wise gap between accuracy and confidence. The lower panels show the confidence distributions of correct (green) and incorrect (red) predictions.

example, an independent teacher model uses the image, question, and reference answer to identify regions supporting the reference answer, which are then blurred to construct x_i^-. All methods are evaluated on the same paired inputs and generate new answers and confidence estimates under both conditions. For each evaluated method m, we retain the pairs answered correctly under both conditions:

$$\mathcal{C}_m = \left\{ i : y_{m,i}^+ = y_{m,i}^- = 1 \right\}.$$

We define the paired confidence change and Support Gap as

$$\Delta_{\text{support},i}^{(m)} = \hat{p}_m(x_i^+) - \hat{p}_m(x_i^-), \qquad \text{Gap}_m = \frac{1}{|\mathcal{C}_m|} \sum_{i \in \mathcal{C}_m} \Delta_{\text{support},i}^{(m)}. \tag{12}$$

We further report Pairwise Preference and High-confidence Masked Rate:

$$\text{Pref}_m = \frac{1}{|\mathcal{C}_m|} \sum_{i \in \mathcal{C}_m} \mathbb{1}\left\{ \Delta_{\text{support},i}^{(m)} > 0 \right\}, \qquad \text{HCM}_m = \frac{1}{|\mathcal{C}_m|} \sum_{i \in \mathcal{C}_m} \mathbb{1}\left\{ \hat{p}_m(x_i^-) > 0.8 \right\}. \tag{13}$$

We evaluate 378 paired examples drawn from the six benchmarks. Restricting the analysis to predictions that remain correct under both conditions isolates the confidence response to reduced visual support from changes in answer correctness.

Table 2: **Confidence response to reduced visual support.** Metrics are computed over correct/correct pairs. Support Gap is reported in percentage points, with 95% CIs in brackets. Pref. and HCM denote Pairwise Preference and High-confidence Masked Rate, respectively.

Method	**Original**	**blurred**	**Gap (↑)**	**Pref. (↑)**	**HCM (↓)**
CoT	0.9817	0.9344	4.72 [3.26, 6.42]	0.5229	0.9450
TopK	0.9840	0.8463	13.78 [10.32, 17.40]	0.5194	0.6589
ConfTuner	0.9569	0.8285	12.84 [9.10, 16.67]	0.5517	0.6121
Rewarding Doubt	0.9385	0.7962	14.23 [10.00, 18.94]	0.5000	0.6250
VL-Calib.	0.9332	0.7812	15.20 [11.14, 19.53]	0.6606	0.7156
TRACECONF	0.7768	0.5701	**20.67** [15.93, 25.59]	**0.8174**	**0.3217**

Table 2 shows that TRACECONF is the most responsive to reduced visual support among the evaluated methods, achieving a Support Gap of 20.67 percentage points (95% CI: [15.93, 25.59]) compared with 4.72–15.20 points for the baselines. Its higher Pairwise Preference and lower HCM

further indicate that confidence decreases when relevant visual evidence is removed, even when answer correctness is preserved.

Confidence depends on the declared evidence locations. The preceding experiment shows that confidence changes when independently identified visual support is weakened, but does not establish whether such changes are specifically associated with the evidence declared by TRACECONF. We therefore extend the paired confidence difference defined in Eq. 12 to the model-declared regions while keeping the predicted answer unchanged, and denote the resulting quantity as Necessity Drop (NecDrop). To measure whether the declared evidence alone can preserve model confidence, we additionally define $\mathrm{SufGap}(e) = \left|\hat{p}(x) - \hat{p}(x_{\mathrm{only}(e)})\right|$. A smaller SufGap indicates better preservation of the full-image confidence. As shown in Figure 4, the declared regions exhibit a larger NecDrop and a smaller SufGap than their complements, indicating that the model's confidence depends more strongly on its declared visual evidence.

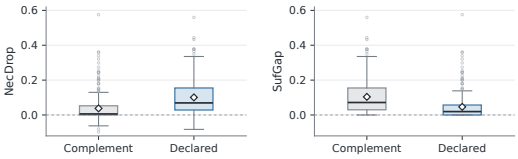


Figure 4: **Declared evidence versus its complement.** Declared regions yield higher NecDrop and lower SufGap than their complements, indicating stronger confidence dependence on the declared evidence.

5.4 How Does Training Shape Calibrated and Traceable Confidence?

We conduct an ablation study to examine the contributions of structured initialization and the full TRACECONF training objective. We compare three training variants: structured SFT alone, the role-separated objective without evidence gating ($W = 1$), and the full TRACECONF objective.

Table 3: **Ablation study of the TRACECONF training pipeline.** Answer-level metrics are averaged over six benchmarks. Visual metrics are reported in percentage points. Δ_{gain} denotes the direction-normalized change from the no-gate variant to the full TRACECONF objective, where positive values indicate improvement.

Answer-level confidence				Visual-evidence dependence			
Variant	**ACC ↑**	**ECE ↓**	**AUROC ↑**	**Variant**	**Gap ↑**	**NecDrop ↑**	**SufGap ↓**
SFT only	0.7018	0.1486	0.7653	SFT only	17.92	10.35	4.64
w/o gate	0.7126	0.1243	0.7960	w/o gate	20.59	8.96	4.65
TRACECONF	**0.7338**	**0.0840**	**0.7982**	TRACECONF	**20.67**	**11.94**	**3.28**
Δ_{gain} vs. w/o gate	+0.0212	+0.0403	+0.0022	Δ_{gain} vs. w/o gate	+0.08	+2.98	+1.37

As shown in Table 3, structured SFT already yields confidence that responds to localized visual evidence, while role-separated optimization without evidence gating further improves answer-level performance and the Support Gap. The full TRACECONF objective performs best across all six metrics, substantially reducing ECE from 0.1243 to 0.0840 and strengthening confidence dependence on model-declared evidence, as reflected by higher NecDrop and lower SufGap.

6 Conclusion

We introduced TRACECONF, a framework for evidence-warranted multimodal confidence that makes confidence traceable to model-declared visual evidence. TRACECONF jointly generates declared evidence, an answer, and a confidence choice; evaluates the necessity and sufficiency of the declared regions through counterfactual interventions; and separates sequence-level response optimization from local proper confidence learning. The resulting evidence gate modulates confidence supervision while preserving its proper conditional target. Across six benchmarks, TRACECONF improves calibration and error discrimination while maintaining competitive accuracy. Further experiments show that confidence tracks visual support and depends more strongly on the declared regions than on their complements, demonstrating that calibrated multimodal confidence can also be visually traceable.

LLM USAGE STATEMENT

Large language models were used to assist with the organization and refinement of the manuscript, including improving English expression, condensing overly verbose passages, checking notation and cross-section consistency, and suggesting LaTeX formulations and presentation improvements. They were also used to provide editorial feedback on the clarity of the argument and the presentation of figures and tables. The research questions, methodological design, implementation, experiments, data analysis, and scientific conclusions were developed and verified by the authors. All LLM-assisted content was reviewed and revised by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

REFERENCES

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yResLmrVO1. Also available as arXiv:2503.02623.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2f8ee6a3d766b426d2618e555b5aeb39-Paper-Conference.pdf.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ASQ649zdHm. Also available as arXiv:2507.16806.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.408/.

Yuetian Du, Yucheng Wang, Rongyu Zhang, Zhijie Xu, Boyu Yang, Ming Kong, Jie Liu, and Qiang Zhu. Linking perception, confidence and accuracy in MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.12149. Introduces Confidence-Driven Reinforcement Learning (CDRL).

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pp. 145–171, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.trustnlp-1.13. URL https://aclanthology.org/2024.trustnlp-1.13/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R. Fung. MM-Boundary: Advancing MLLM knowledge boundary awareness through reasoning step confidence calibration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16427–16444, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.802. URL https://aclanthology.org/2025.acl-long.802/.

Tao Huang, Rui Wang, Xiaofei Liu, Yi Qin, Li Duan, and Liping Jing. Detecting misbehaviors of large vision-language models by evidential uncertainty quantification. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xJT4fXJr1Q. Also available as arXiv:2602.05535.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VZQ04Ojhu5. Also available as arXiv:2508.18847.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29914–29934. PMLR, 2024. URL https://proceedings.mlr.press/v235/lin24c.html.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156. PMLR, 2025. URL https://proceedings.mlr.press/v267/luo25b.html.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2901–2907. AAAI Press, 2015. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4oYxzssbVg.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Wenyi Xiao, Xinchi Xu, and Leilei Gan. VL-Calibration: Decoupled confidence calibration for large vision-language models reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44791–44815, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.2074. URL https://aclanthology.org/2026.acl-long.2074/.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1408–1450, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.74. URL https://aclanthology.org/2025.emnlp-main.74/.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for LLMs. *arXiv preprint arXiv:2412.14737*, 2024. URL https://arxiv.org/abs/2412.14737.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. URL https://aclanthology.org/2025.acl-long.736/.

Caiqi Zhang, Xiaochen Zhu, Chengzu Li, Nigel Collier, and Andreas Vlachos. LoVeC: Reinforcement learning for better verbalized confidence in long-form generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33336–33363, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1539. URL https://aclanthology.org/2026.acl-long.1539/.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/01270.pdf.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3623–3643, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.198. URL https://aclanthology.org/2024.acl-long.198/.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VOAMTA8jKu.

A Implementation Details

This section documents the data construction, optimization, visual interventions, and evaluation protocol used for the 4B results in the main paper. We distinguish the scorer used to construct the SFT data from the visual-support verifier used during reinforcement learning: the former is a local Qwen3-VL-4B-Instruct model evaluated by teacher-forced answer likelihood, whereas the latter is a frozen Doubao-Seed-2.0-Lite model queried as a binary visual-support verifier.

A.1 Data Construction

Evidence annotation and counterfactual filtering. We construct the SFT source from the training portion of the VL-Calibration-12K subset of ViRL-39K (Xiao et al., 2026; Wang et al., 2025). For each image–question pair, a frozen VLM annotator is given the reference answer and asked to identify the smallest set of visible regions sufficient for answering the question. It returns up to eight bounding boxes in a normalized 0–1000 coordinate system. Multiple boxes are retained as a set: we neither snap coordinates to a grid nor replace a disconnected set by its enclosing rectangle. This is important for questions whose evidence is distributed across a plot, an axis, a legend, or several diagram elements.

The initial annotation pool contains 11,961 examples, of which 11,958 were successfully scored. To audit whether an annotation contains measurable visual evidence, we use a frozen Qwen3-VL-4B-Instruct model and the teacher-forced mean log probability of the reference-answer tokens. Let L_{full}, L_{blank}, L_{mask}, and L_{only} denote this score on the original image, the fully blurred image, the image with the proposed regions blurred, and the image with only the proposed regions retained, respectively. We classify a sample as *local* when

$$L_{\text{full}} - L_{\text{blank}} \geq 0.05, \qquad L_{\text{full}} - L_{\text{mask}} \geq 0.05, \qquad L_{\text{only}} - L_{\text{blank}} \geq 0.05,$$

and $L_{\text{full}} - L_{\text{only}} \leq 1.5$. A sample is marked *nonvisual* when the full-versus-blank difference is below 0.05, and *unresolved visual* when the full image has measurable visual signal but the proposed region set does not satisfy all local-evidence criteria. All interventions use Gaussian blur with radius 14 on the original-resolution image. The resulting counts are summarized in Table 4.

The final 5,368-example source consists of 3,744 local and 1,624 unresolved-visual examples. We retain the latter because they remain visually dependent examples with nonempty teacher-proposed boxes, while excluding the 6,485 nonvisual examples. The original construction removes records without usable boxes and images assigned to the held-out sets available at data-construction time. A subsequent audit against the final six evaluation sets found no normalized-question overlap. It found nine shared image hashes affecting 11 evaluation questions, but the questions paired with those images are different from the SFT questions. The independently constructed RL split is disjoint from all six evaluation sets at both the image-hash and normalized-question levels.

Reasoning and confidence targets. We generate candidate reasoning traces with Qwen3-VL-4B-Thinking conditioned only on the image and question; the reference answer is not included in this generation prompt. Sampling uses temperature 0.6, top-$p = 0.95$, top-$k = 20$, and a maximum of 8,192 new tokens. Closed reasoning traces containing at most 5,000 tokens are retained. For every retained completion, the model's own answer is preserved and its correctness label is obtained by task-specific matching against the reference answer. The confidence target is `yes` exactly for a correct candidate and `no` otherwise. Unclosed or excessively long traces are replaced by a direct-answer fallback before this correctness test; completions with a valid reasoning trace but a malformed tail are repaired in the same manner. The final data contain 3,728 examples with a retained reasoning trace and 1,640 direct fallbacks, with 2,771 `yes` and 2,597 `no` targets. Every assistant target ends with the same ordered fields, `<evidence>`, `<answer>`, and `<conf>`.

RL prompts. The RL data are drawn independently from MMFineReason-Visual-Verifiable-v2-10k. We first remove image overlap with the SFT and evaluation sets, deduplicate by image SHA-256, and enforce a fail-closed answer-matching contract. The resulting training split contains 3,000 unique images and prompts: 2,336 integer, 556 decimal, and 108 fraction answers. Of the original 3,000-prompt split, 2,646 pass the answer contract directly and 354 are replaced deterministically from the unused, disjoint candidate pool while preserving the answer-type counts. The policy receives no evidence annotation in RL; its evidence boxes are generated on policy and scored online.

Table 4: Construction of the structured SFT data. The three scorer categories are mutually exclusive. The final source retains examples with nonempty evidence from the local and unresolved-visual categories after held-out and empty-box exclusions.

Stage or category	Count	Criterion or role
Initial annotation pool	11,961	Candidate image–question pairs
Successfully scored	11,958	Complete counterfactual scorer output
Local	3,766	Proposed regions pass visual-dependence, necessity, and sufficiency checks
Nonvisual	6,485	Insufficient full-versus-blank visual signal
Unresolved visual	1,707	Visual signal is present, but the proposed set fails at least one local-evidence check
Final SFT source	5,368	Nonempty-box local and unresolved-visual examples after exclusions

A.2 Training and Evaluation Protocols

Structured SFT. The policy is initialized from Qwen3.5-4B and trained for three epochs, or 504 optimizer steps, on eight GPUs. We update all language-model parameters while freezing the vision tower and multimodal aligner. The effective batch size is 32 (per-device batch size 1 and four gradient-accumulation steps), the learning rate is 5×10^{-6} with cosine decay and a 5% warm-up ratio, and the maximum sequence length is 8,192. Training uses BF16, gradient checkpointing, and ZeRO-3, with model and data seeds set to 42. The SFT objective follows the correctness-masked formulation in the main text. Let $\mathcal{T}_i^{\mathrm{gen}}$ contain the active assistant positions before the final confidence choice and let t_i^{conf} denote that choice position. The implemented per-example loss is

$$\mathcal{L}_{\mathrm{SFT}} = \frac{1}{B}\sum_{i=1}^{B}\left[\lambda_{\mathrm{sol}} y_i \frac{1}{|\mathcal{T}_i^{\mathrm{gen}}|}\sum_{t\in\mathcal{T}_i^{\mathrm{gen}}} \ell_{\mathrm{CE}}(o_{i,t}^{\mathrm{SFT}}) + \lambda_{\mathrm{conf}} w_{y_i} \ell_{\mathrm{CE}}^{\mathrm{bin}}\big(o_{i,t_i^{\mathrm{conf}}}^{\mathrm{SFT}}\big)\right]. \tag{14}$$

The generation-token loss is normalized within each example before the batch average is taken. For $y_i = 1$, the reasoning, evidence, and answer tokens are supervised together with the confidence choice. For $y_i = 0$, the factor y_i masks all generation-token targets, while the final confidence choice remains supervised toward `no`. Thus incorrect candidates teach the model to express low confidence without being imitated as solutions.

Reinforcement learning. We initialize RL from the final SFT checkpoint and train on all 3,000 prompts with eight rollouts per prompt, giving 24,000 trajectories and 125 optimizer steps. Each update contains 24 prompt groups. Training runs on 32 GPUs with per-device batch size 1, six gradient-accumulation steps, and six updates per generation batch. We use a learning rate of 10^{-6}, five warm-up steps, BF16, gradient checkpointing, an 8,192-token completion limit, sampling temperature 1.1, and no explicit thinking-token budget. Rewards are centered within each group without standard-deviation scaling.

The sequence reward follows the role-separated objective in the main text:

$$R_i = R_{\mathrm{format},i} + R_{\mathrm{accuracy},i} + W_i. \tag{15}$$

Here $R_{\mathrm{format},i}$ includes the format-component coefficient used in optimization, $R_{\mathrm{accuracy},i} = y_i$, and W_i is the evidence gate. This scalar reward is used only for the reasoning, evidence, and answer tokens; the final confidence-choice token is excluded from the shared GRPO advantage. At that position, the implementation instead applies the unclipped local binary log loss from Eq. 10, weighted by $\lambda = 0.01$. Consequently, the confidence probability does not appear in the sequence-level reward. KL regularization to the fixed SFT policy is applied at every generated position with $\beta = 0.01$. An unfinished rollout receives zero task-reward components and a terminal penalty of -3.

Online visual-support verification. During RL, the frozen verifier is given the current image, question, and the answer generated in the current rollout, and decides whether the visible image supports that answer. The dataset reference answer is used only to compute y_i and $R_{\mathrm{accuracy},i}$; it is not provided to the visual-support verifier. Consequently, the gate is a function of the pre-confidence context $H = (X, Q, R, E, A)$ and does not use the realized correctness label or the confidence

Table 5: Optimization settings for the reported 4B model.

Setting	Structured SFT	TraceConf RL
Training data	5,368 examples	3,000 prompts, 8 rollouts each
Optimization length	3 epochs / 504 steps	125 steps
Effective prompt batch	32	24
Learning rate	5×10^{-6}	10^{-6}
Scheduler	cosine, 5% warm-up	constant, 5 warm-up steps
Maximum length	8,192 total tokens	4,608 prompt; 8,192 completion
Precision and parallelism	BF16, ZeRO-3, 8 GPUs	BF16, ZeRO-3, 32 GPUs
Frozen components	vision tower and aligner	vision tower; fixed SFT KL reference
Confidence coefficient	λ_{conf} in the SFT objective	$\lambda = 0.01$
KL coefficient	—	$\beta = 0.01$

logits. The two output labels, `A` and `B`, are counterbalanced deterministically by sample identifier. We orient the same-position token log-probability difference so that a positive margin always means stronger visual support. Applying the verifier to the full, evidence- blurred, evidence-only, and fully blurred images gives margins c_{full}, c_{blur}, c_{only}, and c_{blank}, from which the gate in the main text is computed. We use Gaussian blur with radius 14, $\tau_{\mathrm{nec}} = \tau_{\mathrm{suf}} = 5$, and the smooth-ratio mapping. A rollout may contain at most 16 boxes, and a box union covering more than 95% of the image is assigned zero evidence weight. Invalid or empty evidence also receives no gated term. Exact A/B log probabilities are used when available; a missing top-20 alternative is retried up to three times and then handled by a bounded, explicitly recorded top-k censoring rule. The verifier is required only during training and mechanistic evaluation, not during ordinary deployment.

Inference and answer-level metrics. For all six reported benchmarks, each model produces one response per example. TraceConf uses greedy decoding with temperature 0, top-$p = 1$, seed 0, a 12,288-token thinking budget, a 16,384-token output limit, and a 32,768-token context limit. Its confidence is read directly from the final `yes`/`no` logits according to Eq. 1. The correctness judge receives only the question, reference answer, answer choices when available, and the parsed final answer; it does not receive the model's reasoning, evidence boxes, or confidence. ACC is computed over all evaluated examples. ECE uses ten equal-width bins, and ECE and AUROC are computed on examples with both a valid confidence and a successful correctness judgment. The evaluated benchmark sizes are 1,000 for MATH-Vision, 778 for MathVerse-V, 928 for DynaMath-2V, 1,500 for MMStar, 1,000 for MathVista, and 1,726 for MMMU-Pro-v2, totaling 6,932 examples. The Average rows in the main paper are unweighted arithmetic means of the six per-benchmark values, rather than metrics recomputed after pooling all examples.

Visual-intervention statistics. The reference-region experiment begins with 64 deterministically selected examples from each benchmark. A separate annotation prompt yields usable reference regions for 378 of the resulting 384 examples. For each method, we then restrict to its own pairs that remain correct before and after the reference-region blur. Support Gap and its confidence interval are computed from the paired per-example differences; Pairwise Preference uses a strict inequality, and the High-confidence Masked Rate uses threshold 0.8. Confidence intervals use 100,000 paired percentile-bootstrap replicates with seed 20260920.

For the declared-region experiment, we use a fixed-prefix replay: the original reasoning, declared evidence, and answer are held constant, and only the image condition supplied before the final confidence choice is changed. The cohort contains 384 outputs, 64 per benchmark; the main box plot uses the 265 examples whose original answer is correct. The declared condition uses the union $U(e)$, and the complementary condition uses $\Omega \setminus U(e)$; the two regions are not area matched. All reported box-plot statistics are therefore paired at the example level and should be interpreted as confidence dependence on the declared region versus the remainder of the same image.

B Prompts and Qualitative Examples

We report the operative prompt formats below. Line breaks and capitalization are preserved because they determine parsing and the location of the final confidence choice. The annotation prompt uses a

reference answer during SFT data construction. During RL, the policy does not observe the reference answer, and the support-verification prompt evaluates the answer generated in the current rollout.

Evidence-box annotation prompt. The annotation model receives the following system instruction, followed by the image, question, and known reference answer.

```
You are a visual evidence box annotator for VQA.
Given one image, a question, and the known correct answer,
return evidence regions as normalized 0--1000 bounding
boxes.  Your goal is to propose the smallest visual region
set that is sufficient for a human to answer.
Output strict JSON only.  Coordinates are [x1,y1,x2,y2].
Multiple boxes are allowed and preferred to a large
enclosing box when the evidence is disconnected.  Include
the required axis, legend, label relation, complete formula
line, or execution path when applicable.  If the image is
unnecessary, return visual_required=false and an empty box
list.
```

The required JSON fields are `visual_required`, `question_type`, `evidence_boxes`, and `notes`; each box additionally carries a short role, description, granularity, and justification. These textual annotations are retained for auditing, but only the coordinates enter the policy target.

SFT generation prompt. The reasoning model used to construct the 4B SFT data receives:

```
Answer the question and output visual evidence as normalized
0-1000 bounding boxes.  Return exactly three lines:
<evidence>[[x1,y1,x2,y2],[...]]</evidence>
<answer>answer</answer>
<conf>yes|no</conf>
Question:  {question}
```

The model may produce a `<think>` block before these fields. The reference answer is withheld while the reasoning trace and candidate answer are generated.

RL and evaluation prompt. RL uses the following stricter version of the same interface:

```
Answer the question and output visual evidence as normalized
0-1000 bounding boxes.  Return exactly three lines.  In the
final <conf> tag, write lowercase yes or no exactly:
<evidence>[[x1,y1,x2,y2],[...]]</evidence>
<answer>answer</answer>
<conf>yes|no</conf>
Put only the final answer in <answer>.  For multiple-choice
questions, either the option label or its answer content is
accepted.
Question:  {question}
```

Evaluation uses the same output schema. The reported probability is obtained from the `yes`/`no` logits at the final confidence position, not by mapping the generated word itself to a hard confidence value.

Visual-support verification prompt. For each of the four counterfactual image conditions, the verifier receives the image and this prompt:

```
System:  You are a strict visual answer-support verifier.
Decide whether the visible image evidence supports the
proposed answer to the question.  Output exactly one
uppercase ASCII letter and nothing else:  {support label} if
the image supports the proposed answer; {not-support label}
```

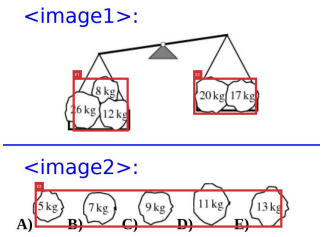
```
if it does not support the proposed answer or the needed
visual evidence is absent.

User:  Question:

{question}

Proposed answer:

{generated answer}

Output A or B.
```

The mapping of the two semantic outcomes to A/B is deterministically counterbalanced across sample identifiers and is held fixed across the four views of an example. The reference answer is not included in this verifier prompt; it is used separately to compute the rollout's correctness label.

Qualitative examples from the final policy. We report three outputs from the final RL-125 policy. Each example includes the question, the input image with policy-declared evidence boxes, the reasoning trace, the answer, the binary confidence choice, and the continuous confidence. Red boxes are rendered directly from the normalized coordinates generated in the `<evidence>` field. For readability, repetitive reasoning sentences are omitted, while the argument, declared coordinates, answer, and confidence choice are preserved.

Example 1: Multi-region visual arithmetic (MATH-Vision).

Question. Which stone should Mr Flintstone place on the right side of the scales so that both sides weigh the same? **Options:** A: 5 kg, B: 7 kg, C: 9 kg, D: 11 kg, E: 13 kg.



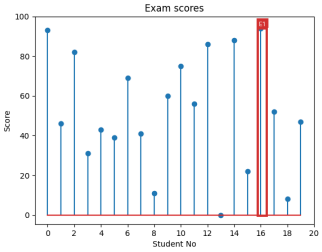
TRACECONF output $\hat{p} = 0.924$

```
<think> The left side has 26 + 8 + 12 = 46 kg.  The right side
has 20 + 17 = 37 kg.  The missing weight is 46 − 37 = 9 kg,
corresponding to option C. </think>
<evidence>[[220,330,395,555], [600,330,800,485],
[100,800,880,960]]</evidence>
<answer>C</answer>  <conf>yes</conf>
```

Verification. The declared regions cover both sides of the scale and the candidate weights. The generated answer C matches the reference answer.

Example 2: Dynamic chart reasoning (DynaMath-2V).

Question. The diagram shows the exam scores of several students. What is the student number of the student who obtained the highest score?



TRACECONF output $\hat{p} = 0.835$

```
<think> The x-axis indicates Student No and the y-axis
indicates Score.  The highest point is at Student No 16,
with a score of approximately 95, which is higher than the
other points.  </think>
<evidence>[[740,160,770,860]]</evidence>
<answer>16</answer>  <conf>yes</conf>
```

Verification. The declared region isolates the highest point and its corresponding student number. The generated answer 16 matches the reference answer.

Example 3: General visual recognition (MMStar).

Question. How many persons are playing guitar in the concert? **Options:** A: 1, B: 2, C: 3, D: 4.



TRACECONF output $\hat{p} = 0.953$

```
<think> There is one person in the background playing a guitar.
The main singer is holding a microphone rather than a guitar.
Therefore, only one person is playing guitar, corresponding to
option A. </think>
<evidence>[[560,345,800,768]]</evidence>
<answer>A</answer>  <conf>yes</conf>
```

Verification. The declared region encloses the guitarist while excluding the singer. The generated answer A matches the reference answer.

These examples illustrate the structured output format and the inspectability of the declared evidence regions. Population-level calibration and evidence dependence are established by the quantitative experiments rather than by selected qualitative examples.

C ADDITIONAL RESULTS ON THE 8B MODEL

We replicate the complete SFT-to-RL pipeline with **Qwen3-VL-8B-Thinking**, using the same 5,368-example SFT source, 3,000 RL prompts, eight rollouts per prompt, and TRACECONF objective as the reported Qwen3.5-4B model. This 8B experiment does *not* use a Qwen3.5-8B checkpoint: both the model architecture and scale change from Qwen3.5-4B to Qwen3-VL-8B-Thinking. It therefore evaluates whether the method transfers to a larger vision–language backbone rather than providing a controlled within-family scaling comparison.

We evaluate the final 8B checkpoint on the same six held-out benchmarks and with the same inference and answer-judging protocol used for the main results. Table 6 reports accuracy over all examples, along with ECE (10 equal-width bins) and AUROC on examples with both a valid confidence and a successful correctness judgment. The average row is the unweighted arithmetic mean of the six per-benchmark values.

Table 6: Held-out results of the 8B TRACECONF model on the six benchmarks used in the main paper. Lower ECE and higher ACC/AUROC are better.

Benchmark	N	**ACC** ↑	**ECE** ↓	**AUROC** ↑
MATH-Vision	1,000	0.5330	0.0734	0.8231
MathVerse-V	778	0.6478	0.0591	0.7864
DynaMath-2V	928	0.7414	0.1064	0.7339
MMStar	1,500	0.7047	0.1164	0.7135
MathVista	1,000	0.7550	0.0356	0.8303
MMMU-Pro-v2	1,726	0.6483	0.0982	0.7536
Macro average	**6,932**	**0.6717**	**0.0815**	**0.7735**

The 8B model retains meaningful confidence discrimination on every benchmark: AUROC is above 0.71 throughout and reaches 0.8303 on MathVista and 0.8231 on MATH-Vision. Calibration varies across datasets, with the lowest ECE on MathVista (0.0356) and the highest on MMStar (0.1164). Overall, the 8B replication obtains a macro-average ACC of 0.6717, ECE of 0.0815, and AUROC of 0.7735. These results confirm that the evidence-conditioned confidence interface transfers to the larger backbone, while also showing that increased model size alone does not uniformly improve task accuracy or confidence ranking under the present training recipe.

Figure 5 shows the corresponding per-benchmark reliability diagrams. The confidence estimates retain a clear monotonic relationship with empirical correctness across all six datasets, while the magnitude and direction of the bin-wise calibration gap vary by benchmark.



Figure 5: **Per-benchmark calibration of the Qwen3-VL-8B-Thinking TRACECONF model.** The upper panels are reliability diagrams: blue bars show empirical accuracy, black marks show mean confidence, the dashed diagonal denotes perfect calibration, and hatched regions show the calibration gap. The lower panels show confidence distributions for correct and incorrect predictions, normalized within each class; matching dashed vertical lines indicate the corresponding class means. ECE uses 10 equal-width bins.

D PROOF OF PROPERNESS UNDER EVIDENCE GATING

We make explicit the conditional statement underlying Eq. 11. Let $H = (X, Q, R, E, A)$ denote the complete context before the confidence choice, let $Y \in \{0, 1\}$ denote whether the generated answer is correct, and let $W = w(H) \geq 0$ be the evidence gate. The gate is computed from the image, question, generated reasoning and answer, and declared evidence before the confidence choice is made. In particular, the verifier does not receive the reference answer, the realized correctness label, or the confidence logits. Thus W is fixed after conditioning on $H = h$ and is independent of the reported probability.

Proposition 1 (Properness under evidence gating)**.** *For a fixed context $H = h$ with $W(h) > 0$, the conditional risk of the gated binary logarithmic loss is uniquely minimized by the conditional probability of answer correctness:*

$$\arg\min_{p\in[0,1]} \mathbb{E}[W(H)\ell_{\log}(p, Y) \mid H = h] = \Pr(Y = 1 \mid H = h) = p^\star(h). \tag{16}$$

Proof. Write $p^\star(h) = \Pr(Y = 1 \mid H = h)$. Since $W = w(H)$, conditioning on $H = h$ makes the evidence weight the nonnegative constant $W(h)$. Therefore

$$\begin{aligned}\mathcal{R}_h(p) &= \mathbb{E}[W(H)\ell_{\log}(p, Y) \mid H = h] \\ &= W(h)\left[-p^\star(h)\log p - \big(1 - p^\star(h)\big)\log(1 - p)\right].\end{aligned} \tag{17}$$

The bracket is the cross-entropy between a Bernoulli variable with parameter $p^\star(h)$ and the report p. Equivalently, it is the entropy of $\mathrm{Bern}(p^\star(h))$ plus $\mathrm{KL}(\mathrm{Bern}(p^\star(h)) \,\|\, \mathrm{Bern}(p))$. Its unique minimizer is therefore $p = p^\star(h)$, including the limiting cases $p^\star(h) \in \{0, 1\}$. Multiplication by the positive constant $W(h)$ does not change that minimizer. □

The construction requires more than independence from the reported confidence: the gate must be determined by the pre-confidence context itself. TraceConf satisfies this condition by evaluating whether the declared regions support the generated answer, while reserving the reference answer solely for computing the correctness

target Y. Consequently, the same $W(H)$ may control how much local confidence supervision a context receives without changing which probability is optimal for that context.b

When $W(h) = 0$, the gated conditional risk is identically zero and every report p has the same risk. This is an abstention of supervision rather than a target of zero confidence: weak declared evidence does not by itself establish that the generated answer is wrong.

Finally, the proposition characterizes the population optimum of the local confidence objective for a fixed conditional distribution. It does not claim that a finite neural model trained jointly with sequence rewards and KL regularization reaches this optimum, nor that marginal calibration follows automatically. Those properties are evaluated empirically using ECE, AUROC, and the visual-intervention tests in the main paper.

E Extended Related Work

Classical calibration. Probabilistic calibration asks that events assigned probability p occur with frequency p. Modern neural classifiers are often overconfident, motivating post-hoc methods such as temperature scaling and nonparametric binning (Guo et al., 2017; Naeini et al., 2015). Proper scoring rules provide a complementary training principle: in expectation, a strictly proper score is optimized by reporting the underlying event probability. TraceConf builds on this view but changes which examples contribute confidence supervision by conditioning its strength on an evidence-dependent gate. Because calibration and ranking are distinct, we report both ECE and AUROC.

Confidence elicitation from generative models. For language models, confidence can be estimated from token or sequence probabilities, self-evaluation, agreement across sampled reasoning paths, or semantic uncertainty over meaning-equivalent generations (Kadavath et al., 2022; Wang et al., 2023; Farquhar et al., 2024). Models can also verbalize their uncertainty directly, although the resulting values depend on prompt design, scale, and task distribution (Xiong et al., 2024; Zhou et al., 2024; Yang et al., 2024). These approaches primarily ask whether confidence separates correct from incorrect predictions; they do not require the reported value to expose a spatially testable basis.

Learning calibrated confidence. Recent methods incorporate confidence into post-training. ConfTuner optimizes tokenized Brier objectives, while Rewarding Doubt and related RL approaches use proper scoring rewards to train models to express calibrated uncertainty (Li et al., 2025; Bani-Harouni et al., 2026; Zhang et al., 2026; Damani et al., 2026). In the multimodal setting, VL-Calibration separates visual and reasoning confidence, MMBoundary supervises uncertainty along reasoning steps, and CDRL uses visually perturbed inputs during reinforcement learning (Xiao et al., 2026; He et al., 2025; Du et al., 2026). These methods improve the estimation or expression of confidence, but their objectives do not test whether the particular image regions declared for a generated answer are necessary and sufficient for supporting that answer. VL-Calibration is the closest conceptual baseline, but its visual-confidence target does not require a declared region to pass paired removal and retention tests for the generated answer.

Visual grounding and faithfulness. Grounding and rationale evaluation ask whether a prediction can be associated with interpretable evidence, and perturbation-based evaluations test this association by deleting or retaining candidate evidence (DeYoung et al., 2020). Multimodal models nevertheless remain vulnerable to language priors and visually unsupported shortcuts (Lin et al., 2024; Luo et al., 2025). TraceConf connects these two lines of work: the model first declares answer-specific visual regions, those regions are tested through paired removal and isolation interventions, and the resulting evidence signal controls the credit assigned to confidence learning. The declared regions therefore define testable intervention targets rather than only explanatory visualizations. Localization agreement and confidence traceability are distinct, so intervention metrics are primary and pseudo-IoU is complementary; the claim is visual testability, not exhaustive human rationale recovery.