VL-Calibration: Decoupled Confidence Calibration for Large Vision-Language Models Reasoning

Wenyi Xiao[*1], **Xinchi Xu**[*1], **Leilei Gan**[†1]
[1]Zhejiang University
{wenyixiao, leileigan}@zju.edu.cn

Abstract

Large Vision Language Models (LVLMs) achieve strong multimodal reasoning but frequently exhibit hallucinations and incorrect responses with high certainty, which hinders their usage in high-stakes domains. Existing verbalized confidence calibration methods, largely developed for text-only LLMs, typically optimize a single holistic confidence score using binary answer-level correctness. This design is mismatched to LVLMs: an incorrect prediction may arise from perceptual failures or from reasoning errors given correct perception, and a single confidence conflates these sources while visual uncertainty is often dominated by language priors. To address these issues, we propose **VL-Calibration**[1], a reinforcement learning framework that explicitly decouples confidence into visual and reasoning confidence. To supervise visual confidence without ground-truth perception labels, we introduce an intrinsic visual certainty estimation that combines (i) visual grounding measured by KL-divergence under image perturbations and (ii) internal certainty measured by token entropy. We further propose token-level advantage reweighting to focus optimization on tokens based on visual certainty, suppressing ungrounded hallucinations while preserving valid perception. Experiments on thirteen benchmarks show that VL-Calibration effectively improves calibration while boosting visual reasoning accuracy, and it generalizes to out-of-distribution benchmarks across model scales and architectures.

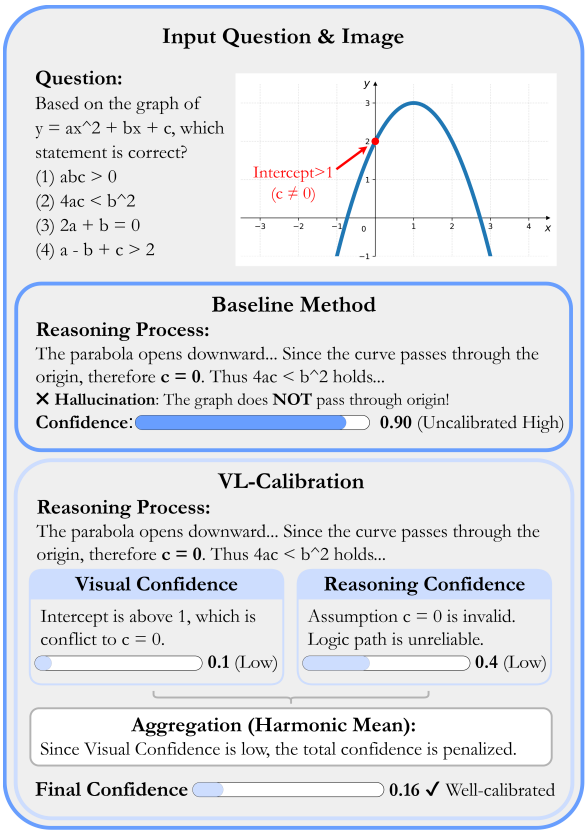


Figure 1: The baseline method (upper) makes overconfident assumptions. Instead, our method (lower) decouples confidence into visual and reasoning confidence, with clear identification of uncertainty sources and improved calibration.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated impressive capabilities in bridging visual perception and logical reasoning (Bai et al., 2025; Wang et al., 2025b; Zhang et al., 2026). Despite their success, these models often exhibit severe hallucinations that typically generate factually incorrect responses (Kadavath et al., 2022; Xiong et al., 2024), limiting their usage in high-stakes domains such as healthcare or law (Xiao et al., 2025; Hu et al., 2025; Shi et al., 2025).

One solution to overcome the aforementioned challenge is to teach models to verbalize their confidence (Kadavath et al., 2022; Xiong et al., 2024) (*e.g.*, "my confidence is 8/10") alongside the

[*]Equal Contribution
[†]Corresponding Author
[1]github.com/Mr-Loevan/VL-Calibration

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 44791–44815
July 2-7, 2026

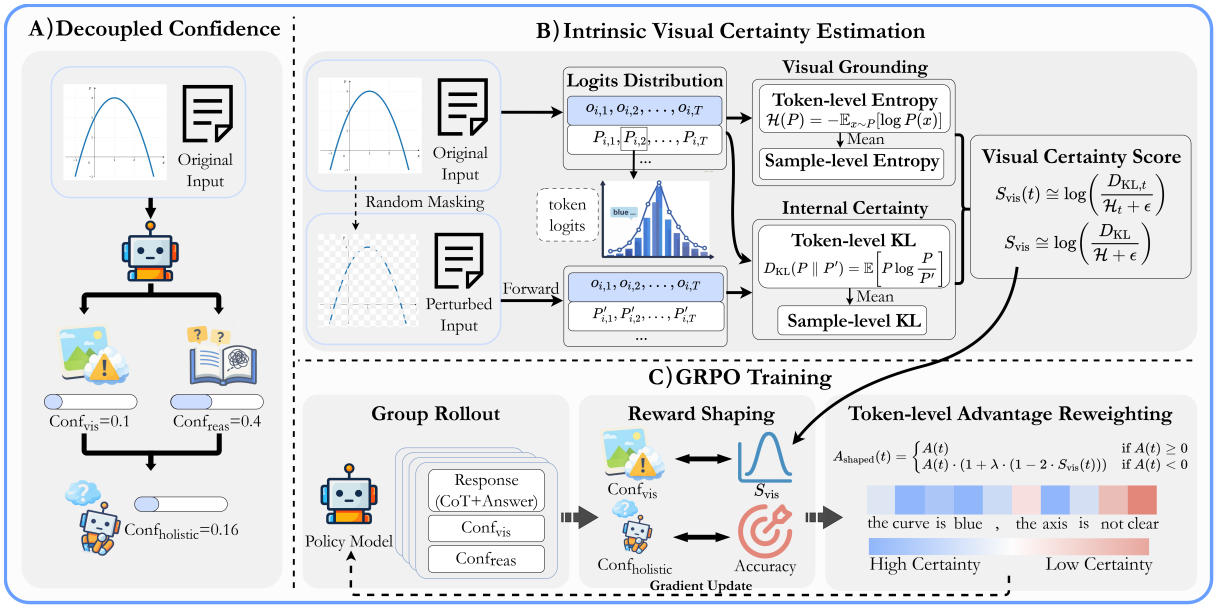


Figure 2: **Overview of our framework. A) Decoupled Confidence Inference.** The LVLM explicitly outputs separate visual and reasoning confidence to derive a holistic confidence. **B) Intrinsic Visual Certainty Estimation.** We quantify visual certainty by measuring visual grounding and internal certainty. **C) GRPO Training.** We align visual confidence with visual certainty score , and the holistic confidence with the answer accuracy. Additionally, we apply token-level advantage reweighting based on token visual certainty.

answers. Recent work has explored **verbalized confidence calibration** in large language models (LLMs)[2] (Stangel et al., 2025; Leng et al., 2025). For example, SaySelf (Xu et al., 2024) trains models to output a verbalized confidence alongside an answer via distilled rationale from GPT-4 and then uses reinforcement learning to align confidence with accuracy. While effective, the performance is limited by the capability of GPT-4. To solve this, RLCR (Damani et al., 2026) trains models to output verbalized confidence via RL using the Brier Score (Glenn et al., 1950), i.e., L2 distance of predicted confidence and binary accuracy, and simultaneously incentivizes reasoning capability through accuracy reward.

However, directly extending verbalized confidence calibration to LVLM reasoning faces several fundamental challenges. First, in LVLMs, an incorrect answer may arise from *perceptual hallucinations* (misreading or ignoring the image) or from *reasoning errors* given a correct perception. Consequently, a single confidence score conflates these error sources in LVLMs, thereby hindering precise error localization. Second, recent studies indicate LVLM reasoning is often dominated by language priors (Ramakrishnan et al., 2018; Jing et al., 2020; Zhang et al., 2025). Consequently, the intrinsic visual uncertainty is likely overshadowed by these language priors, leading to incorrect overall confidence calibration.

To address these limitations, we introduce **VL-Calibration**, a verbalized confidence calibration framework that decouples a single confidence regarding visual perception from that of logical reasoning. As shown in Figure 2 (a), we structure the calibration into two phases and elicit separate confidence tokens for the visual rationale and the reasoning chain. By doing so, VL-Calibration enables confidence expression for both perception and reasoning, and also identifies clearly the uncertainty source.

We employ reinforcement learning to train VL-Calibration. However, during training, this decoupling framework faces the lack of ground-truth labels for visual perception confidence. Existing uncertainty estimation methods for LLMs and LVLMs largely fall into two categories. **Sampling-based methods**, such as Self-Consistency (Wang et al., 2023) and VL-Uncertainty (Zhang et al., 2024), infer uncertainty by aggregating multiple generations, but incur substantial computational overhead. **Internal-state methods** leverage logits (Kadavath et al., 2022) or hidden representations (Vashurin et al., 2025) to predict correctness.

[2]For brevity, we use verbalized confidence calibration and calibration in this paper alternatively.

For example, Self-Certainty (Kang et al., 2025) measures KL-divergence of logits distribution from uniform distribution. However, they overlook the visual grounding characteristic of LVLMs.

To address the absence of ground-truth labels for visual confidence, we propose estimating the certainty of visual perception by simultaneously considering **Visual Grounding** and **Internal Certainty**. Specifically, to measure visual grounding, we compute the KL-divergence between the model's output distributions given the original image versus a perturbed image. A higher KL-divergence indicates that the model is sensitive to visual content, implying strong grounding. Second, to quantify internal uncertainty, we calculate the token entropy of the visual description, where lower entropy reflects higher model confidence. We integrate the two estimations via a log-scale formulation to derive the **Visual Certainty Reward**, which enjoys the merits of (i) rewarding responses that are both visually responsive and internally confident; and (ii) optimization stability by compressing the dynamic numeric range to facilitate stable RL training. Finally, to overcome the weakness that binary outcome reward treats each token equally, we propose **Token-level Advantage Reweighting**, which leverages the aforementioned visual certainty estimation to reweight advantage on high visual uncertainty tokens with negative advantage, thereby discouraging ungrounded hallucinations while preserving valid visual perception.

Extensive experiments across thirteen benchmarks demonstrate the effectiveness of our approach. VL-Calibration reduces the Expected Calibration Error (ECE) on Qwen3-VL models from 0.421 to 0.098 and simultaneously improves average accuracy by 2.3%–3.0% over the strongest baselines. We further observe consistent gains across model scales, model architectures, and out-of-distribution benchmarks, validating the effectiveness of our proposed paradigm.

2 Related Work

LLM and LVLM Uncertainty Estimation Sampling-based methods derive uncertainty estimation by aggregating multiple outputs. Self-Consistency (Wang et al., 2023) selects the most frequent answer via majority voting, while Semantic Entropy (Farquhar et al., 2024; Aichberger et al., 2025) measures meaning-level consistency across response clusters. For LVLMs, VL-Uncertainty (Zhang et al., 2024)) estimate uncertainty across multiple samples with semantically equivalent but perturbed inputs. Internal-state approaches leverage internal signals to quantify certainty. This includes both statistical metrics derived from logits, such as log-probabilities (Vashurin et al., 2025), perplexity (Zhao et al., 2025), Self-Certainty (i.e., logits divergence against uniform distribution) (Kang et al., 2025), and the probability assigned to a designated `true` token (Kadavath et al., 2022).

Verbalized Confidence Calibration Recent studies have explored various alignment strategies to calibrate verbalized confidence (Stengel-Eskin et al., 2024; Leng et al., 2025; Xu et al., 2024; Stangel et al., 2025; Damani et al., 2026). One line of work utilizes **Supervised Fine-Tuning (SFT)** with consistency-based labels (Xu et al., 2024) or token-level logit supervision (Li et al., 2025). Another line leverages **Reinforcement Learning (RL)** to incentivize calibrated uncertainty. For instance, LACIE (Stengel-Eskin et al., 2024) employs Direct Preference Optimization (DPO) (Rafailov et al., 2023; Xiao et al., 2024a) within a speaker-listener framework, while PPO-C (Leng et al., 2025), Sky-Self (Xu et al., 2024), and Rewarding Doubt (Stangel et al., 2025) use Proximal Policy Optimization (PPO) (Schulman et al., 2017) via tailored reward functions (*e.g.*, Brier score or log-penalties) to reward accurate confidence. More recently, RLCR (Damani et al., 2026) uses GRPO (Shao et al., 2024; Xiao and Gan, 2025) to jointly improve task accuracy and calibration. While effective for text-only tasks, these approaches remain largely unexplored for LVLM calibration.

3 Methodology

3.1 Preliminary

Let π_θ denote an LVLM, $q = (I, x)$ a multimodal input consisting of an image I and a textual query x, $\tau = (z, y)$ a generation trajectory sampled from the policy π_θ conditioned on the multimodal input (I, x), where z denotes the reasoning and y the answer. Reinforcement Learning with Verifiable Rewards (RLVR; (Shao et al., 2024)) optimizes the model using a binary correctness reward: $R_{\text{acc}}(y, y^*) = \mathbb{1}_{y \equiv y^*}$, where $\mathbb{1}_{(\cdot)}$ is indicator function for correctness against the ground-truth y^*.

To teach the model to verbalize their confidence alongside the answers, recent work (Damani et al., 2026) extends the trajectory to include a holistic

confidence score $c \in [0, 1]$:

$$\tau = (z, y, c) \tag{1}$$

In addition to correctness reward, it jointly optimizes accuracy and confidence calibration with an additional Brier-based calibration term (Glenn et al., 1950):

$$R_{\text{acc\_conf}}(y, c, y^*) = \mathbb{1}_{y \equiv y^*} - (c - \mathbb{1}_{y \equiv y^*})^2 \tag{2}$$

Optimization Objective We employ GRPO to optimize π_θ. For each query $q = (I, x)$, we sample a group of G outputs $\{o_1, \ldots, o_G\}$ from the old policy $\pi_{\theta_{old}}$. First, we calculate the composite reward R_i for each output o_i and the advantage $\hat{A}_i$ by normalizing rewards within the group. The overall objective maximizes the surrogate of $\hat{A}_i$ with a KL constraint. Let $\rho_{i,t}$ denote the token probability ratio between π_θ and $\pi_{\theta_{old}}$, and clip$(\cdot)$ denote clip$(\rho_{i,t}, 1-\epsilon, 1+\epsilon)$. The objective is defined as:

$$\begin{aligned}\mathcal{J}(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\} \sim \pi_{\theta_{\text{old}}}} \\ \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left(\rho_{i,t} \hat{A}_{i,t}, \text{clip}(\cdot) \hat{A}_{i,t} \right) \right. \right. \\ \left. \left. - \beta \mathbb{D}_{KL} \left(\pi_\theta || \pi_{\text{ref}} \right) \right\} \right]\end{aligned} \tag{3}$$

where ϵ and β are the clipping hyperparameter and the coefficient controlling the KL regularization.

3.2 Decoupling Visual and Reasoning Verbalized Confidence

To realize verbalized confidence calibration for LVLM reasoning and address the limitations of Eq. 2 discussed in §1, we reformulate LVLM reasoning as two phases that explicitly decouples visual perception from reasoning. Specifically, the policy π_θ is instructed to generate structured visual and reasoning rationales, each followed by a decoupled verbalized confidence:

$$\tau = (\underbrace{z_{vis}, c_{vis}}_{\text{Visual Phase}}, \underbrace{z_{reas}, c_{reas}}_{\text{Reasoning Phase}}, y) \tag{4}$$

where z_{vis} denotes the visual rationale (*e.g.*, image dense caption) and z_{reas} denotes the reasoning chain. The confidence tokens c_{vis}, c_{reas} represent the model's certainty regarding z_{vis} and z_{reas}, respectively, normalized to $\hat{c}_{vis}, \hat{c}_{reas} \in [0, 1]$.

To derive the final holistic confidence Φ for the answer y, we employ the harmonic mean of the decoupled scores:

$$\Phi(\hat{c}_{vis}, \hat{c}_{reas}) = \frac{2 \cdot \hat{c}_{vis} \cdot \hat{c}_{reas}}{\hat{c}_{vis} + \hat{c}_{reas}} \tag{5}$$

We select the harmonic mean for its conservative property: unlike the arithmetic mean, Φ is dominated by the minimum of the two scores.

To achieve supervision on decoupled visual confidence, we further introduce a Brier-based vision calibration term. Finally, in our decoupled verbalized confidence framework, the reward is denoted as follows:

$$\begin{aligned}R(\tau, y^*, z^*_{vis}) = \mathbb{1}_{y \equiv y^*} - \left(\Phi(\hat{c}_{vis}, \hat{c}_{reas}) - \mathbb{1}_{y \equiv y^*} \right)^2 \\ - \left(\hat{c}_{vis} - \mathbb{1}_{z_{vis} \equiv z^*_{vis}} \right)^2\end{aligned} \tag{6}$$

3.3 Visual Certainty Estimation

Supervising the decoupled c_{vis} presents a challenge due to the absence of ground-truth labels (z_{vis^*} and $\mathbb{1}_{z_{vis} \equiv z^*_{vis}}$). To achieve effective visual certainty estimation as a reliable pseudo label, we propose a composite metric that evaluates visual certainty from two complementary dimensions: **Visual Grounding** and **Internal Certainty**.

Visual Grounding. Visual grounding measures the extent to which the model's generation relies on the input image rather than hallucinating from language priors (Zhang et al., 2025; Jing et al., 2020). We quantify this via sensitivity analysis: if an LVLM effectively grounds its rationale in the image, perturbing visual features should significantly alter the output distribution. Conversely, insensitivity implies possible hallucination. We calculate the KL-divergence between the logits of the original image I and a perturbed version I' via random patch masking with a ratio = 0.8:

$$D_{KL} = \frac{1}{T} \sum_{t=1}^{T} \text{KL} \left(\pi(\cdot | z_{vis,<t}, I) \parallel \pi(\cdot | z_{vis,<t}, I') \right) \tag{7}$$

Here, a high D_{KL} indicates strong grounding, confirming that the generation is actively conditioned on visual tokens.

Internal Certainty. However, visual grounding alone is insufficient. A model may rely on the image yet remain conflicted among multiple plausible interpretations (*e.g.*, due to visual ambiguity). To capture this internal state, we measure the average token entropy $\mathcal{H}$ over the visual rationale z_{vis}:

$$\mathcal{H} = -\frac{1}{T} \sum_{t=1}^{T} \sum_{v \in \mathcal{V}} \pi(v | z_{vis,<t}, I) \log \pi(v | z_{vis,<t}, I) \tag{8}$$

A low $\mathcal{H}$ indicates a sharp probability distribution, reflecting that the model is internally certain about its generated token.

Visual Certainty Score. Intuitively, a high visual certainty score S_{vis} should indicate both *well-grounded* (high D_{KL}) and *internally certain* (low $\mathcal{H}$). Therefore, we integrate two metrics to formulate the S_{vis}. Considering the diffirent scale of two metrics, we define S_{vis} as a log-ratio signal:

$$S_{vis} = \log(D_{KL} + \epsilon) - \log(\mathcal{H} + \epsilon) \quad (9)$$

Compared with sampling-based and internal-state estimation methods, including the KL (visual grounding), Entropy (internal certainty), Self-Consistency, Semantic Entropy, VL-Uncertainty, and Self-Certainty, our proposed visual certainty estimation shows superior correlation with `Gemini-3-pro-preview` perception judgement. Detailed discussion refers to §5.

3.4 Reinforcement Learning with Certainty-Aware Calibration

Building upon the decoupled confidence and visual certainty estimation, we introduce **Visual Certainty Reward** to provide visual confidence supervision and **Token-Level Advantage Reweighting** to penalize ungrounded hallucinations while preserving valid visual perception.

Reward Shaping. To address the lack of a visual confidence label, we integrate visual certainty estimation via reward shaping. We reformulate the reward function in Eq. 6 as

$$R(\tau, y^*) = \lambda_{acc} R_{acc} + \lambda_{cal} R_{cal} + \lambda_{vis} R_{vis} \quad (10)$$

where λ_{acc}, λ_{cal}, and λ_{vis} are coefficients. Beyond binary accuracy reward: $R_{acc} = \mathbb{1}_{y \equiv y^*}$ where $\equiv$ denotes semantical equivalence, and holistic confidence calibration reward: $R_{cal} = -(\Phi(\hat{c}_{vis}, \hat{c}_{reas}) - \mathbb{1}_{y=y^*})^2$, we propose the **Visual Certainty Reward** (R_{vis}). Specifically, we use S_{vis} as a proxy of $\mathbb{1}_{z_{vis} \equiv z^*_{vis}}$. Since the raw visual certainty S_{vis} (Eq. 9) varies in scale across batches, we first normalize the raw visual certainty score S_{vis} by applying batch-wise z-score standardization $z = \frac{S_{\text{vis}} - \mu_B}{\sigma_B + \epsilon}$ and then mapping it to $[0, 1]$ via a sigmoid transform $\tilde{S}_{\text{vis}} = \sigma(z)$. Therefore we derive R_{vis} as :

$$R_{vis} = -\left(\hat{c}_{vis} - \text{sg}(\tilde{S}_{vis})\right)^2 \quad (11)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. This term explicitly anchors the model's verbalized visual confidence to its actual perceptual certainty, preventing ungrounded hallucinations.

Token-Level Advantage Reweighting Standard GRPO employs uniform credit assignment, treating all errors equally. However, we posit that *ungrounded hallucinations*, where errors arising from high visual uncertainty, indicate a severe perception failure and warrant stricter penalties than other visual errors. To address this, we propose Token-Level Advantage Reweighting (TAR) to dynamically reweight the advantage $\hat{A}_t$ based on the error source and visual certainty. Specifically, we calculate the visual certainty $S_{vis}(t)$ for each specific token. Following the aforementioned Z-score standardization and sigmoid transformation, we normalize these token-wise scores within a single sample to $[0, 1]$ to obtain $\tilde{S}_{vis}(t)$. We then introduce a reweighting mechanism derived from this token-wise $\tilde{S}_{vis}(t)$ to reweight the advantage of tokens within z_{vis} with $\lambda_{\text{TAR}} = 0.1$:

$$\hat{A}_t^{\text{TAR}} = \begin{cases} \hat{A}_t \cdot \left(1 + \lambda_{\text{TAR}}(1 - 2\tilde{S}_{vis}(t))\right) & \text{if } t \in z_{vis} \wedge \hat{A}_t < 0 \\ \hat{A}_t & \text{otherwise} \end{cases} \quad (12)$$

Intuitively, when the model errs ($\hat{A}_t < 0$) under low visual certainty ($\tilde{S}_{vis}(t) \to 0$), the penalty is amplified to discourage blind guessing. Conversely, if the model is well-grounded ($\tilde{S}_{vis}(t) \to 1$), the penalty is softened to preserve valid perception.

4 Experiments

4.1 Experimental Setup

Implementation Details To control training overhead, we randomly pick 12,000 data points from ViRL-39K (Wang et al., 2025a), a diverse categories visual reasoning dataset, as the training dataset, namely VL-Calibration-12K. We apply VL-Calibration on Qwen3-VL-4B-Instruct (Bai et al., 2025), Qwen3-VL-8B-Instruct, and InternVL3.5-4B-MPO (Wang et al., 2025b) to confirm efficacy on different model sizes and base models. For more details, refer to Appendix A.1.

Baselines Our comparison involves inference-stage methods including Verbalize (Xiong et al., 2024), P(`True`) (Kadavath et al., 2022), SteerConf (Zhou et al., 2025), and training-stage methods including RLVR (Guo et al., 2025), LACIE (Stengel-Eskin et al., 2024), ConfTuner (Li et al., 2025), PPO-C (Leng et al., 2025), SaySelf (Xu et al., 2024), Rewarding Doubt (Stangel et al., 2025), and RLCR (Damani et al., 2026). To ensure fair comparisons, we re-implemented these methods on LVLMs with same settings of ours.

Benchmark	Qwen3-VL-4B									Qwen3-VL-8B								
	Accuracy ↑			AUROC ↑			ECE ↓			Accuracy ↑			AUROC ↑			ECE ↓		
	Base	Best	**Ours**	Base	Best	**Ours**	Base	Best	**Ours**	Base	Best	**Ours**	Base	Best	**Ours**	Base	Best	**Ours**
Mathematical and Geometric Reasoning																		
DynaMath	.486	.718	**.753**	.513	.716	**.797**	.423	.165	**.081**	.680	.766	**.784**	.576	.667	**.769**	.460	.160	**.058**
Geo3K	.514	.616	**.671**	.504	**.801**	.792	.773	.159	**.073**	.514	.621	**.729**	.556	.761	**.780**	.734	.192	**.056**
MathVerse	.426	.796	**.807**	.416	.659	**.735**	.561	.142	**.042**	.622	.813	**.838**	.504	.656	**.742**	.372	.129	**.055**
MathVision	.171	.440	**.483**	.501	**.814**	.800	.794	.207	**.170**	.266	.473	**.540**	.527	.771	**.815**	.428	.249	**.094**
MathVista	.679	**.772**	.730	.566	.710	**.778**	.254	.132	**.107**	.678	.733	**.771**	.574	.644	**.753**	.459	.198	**.079**
WeMath	.580	.771	**.820**	.593	.647	**.802**	.268	.164	**.048**	.699	.801	**.836**	.567	.730	**.777**	.388	.110	**.039**
Logical Reasoning																		
LogicVista	.456	.519	**.570**	.615	.757	**.794**	.315	.232	**.203**	.508	.600	**.611**	.580	.688	**.836**	.308	.253	**.109**
Vision-Dominant Reasoning																		
CLEVR	.920	**.935**	**.935**	.517	.577	**.797**	**.025**	.058	.035	.910	.935	**.940**	.545	.495	**.723**	.332	.069	**.029**
MathVerse_V	.283	.748	**.781**	.519	.669	**.721**	.508	.171	**.056**	.573	.776	**.804**	.502	.660	**.743**	.398	.162	**.052**
Multi-discipline Reasoning																		
A-OKVQA	.836	.861	**.875**	.584	.592	**.695**	.022	.112	**.017**	.829	.872	**.875**	.642	.593	**.691**	.057	.107	**.059**
MMK12	.489	.741	**.747**	.468	.651	**.714**	.432	.182	**.083**	.585	.780	**.809**	.506	.691	**.777**	.301	.131	**.039**
MMMU-Pro	.249	.436	**.458**	.610	.694	**.735**	.474	.340	**.335**	.383	.518	**.522**	.579	.634	**.740**	.518	.357	**.220**
ViRL-39K	.620	.796	**.816**	.406	.729	**.753**	.622	.113	**.026**	.689	.811	**.835**	.537	.723	**.783**	.460	.109	**.033**
Avg.	.516	.704	**.727**	.524	.694	**.763**	.421	.167	**.098**	.610	.731	**.761**	.553	.670	**.764**	.401	.171	**.071**

Table 1: **Main Results.** Comparison between the base model using verbalized confidence (Xiong et al., 2024) (**Base**), the strongest re-implemented baseline (**Best**), and our method (**Ours**) across Qwen3-VL 4B and 8B scales. **Bold** indicates the best result. Full results of all baselines are reported in Appendix § D.1.

Model	Method	ACC	AUROC	ECE
Qwen3-VL-30B	P(True)	0.652	0.569	0.388
	Ours	**0.803**	**0.767**	**0.082**
InternVL3.5-4B	RLCR	0.656	0.649	0.209
	Ours	**0.689**	**0.701**	**0.103**

Table 2: **Generalization Analysis on Different Model Scales and Architectures.** We report the average Accuracy, AUROC, and ECE across all 12 evaluation benchmarks. **Bold** indicates the best results.

Evaluation To systematically evaluate the efficacy of VL-Calibration, we evaluate on thirteen benchmarks that cover diverse visual reasoning topics, and multi-disciplinary reasoning. Our evaluation metrics include Accuracy (ACC), Expected Calibration Error (ECE), and Area Under the Receiver Operating Characteristic Curve (AUROC) to evaluate models' both reasoning and calibration capability. Details of evaluation benchmarks and metrics are provided in Appendix A.2.

4.2 Main Results

Table 1 presents the main results regarding reasoning task performance and calibration. Across both Qwen3-VL-4B and Qwen3-VL-8B models, our method consistently outperforms strong baselines on all metrics. **First**, our method achieves a significant improvement in ECE, lowering it from 0.421 to 0.098 on the 4B model and from 0.204 to 0.071 on the 8B model. **Second**, while existing calibration methods often struggle to maintain the original reasoning accuracy, our method improves the average accuracy by a remarkable margin (**+2.3%** over the best baseline on 4B and **+3.0%** on 8B), particularly on complex visual reasoning benchmarks like DynaMath and MathVerse. **Third**, our method demonstrates out-of-distribution generalization, extending its efficacy to multi-disciplinary tasks. On benchmarks requiring broad knowledge integration like MMMU-Pro, our method consistently outperforms baselines, achieving accuracy gains of +2.2%. On the commonsense reasoning benchmark A-OKVQA, it reduces ECE from 0.112 to 0.017. We provide significance analysis in Appendix C. **Lastly**, to confirm that our method can generalize across various model scales and architectures, we further evaluate its performance on Qwen3-VL-30B and InternVL3.5-4B-MPO. As shown in Table 2, on the larger 30B scale, it continues to effectively achieve calibration, significantly improving the AUROC to 0.767 and reducing the ECE to 0.082, while simultaneously boosting reasoning accuracy from 0.652 to 0.803. Similarly, when applied to the InternVL architecture, our method outperforms the strong RLCR baseline, achieving

a superior task performance (ACC=0.689) and calibration (ECE=0.103).

4.3 Ablations

Model	VCE		TAR	Metrics		
	Ent.	KL		ACC	AUROC	ECE
Qwen3-VL-4B	–	–	–	0.516	0.763	0.421
RLCR	–	–	–	0.704	0.694	0.167
+ Decoupled	–	–	–	0.701	0.682	0.164
+ VCE	✓	–	–	0.688	0.723	0.119
	-	✓	–	0.709	0.721	0.124
	✓	✓	–	0.715	0.751	0.121
Ours	✓	✓	✓	**0.727**	**0.763**	**0.098**

Table 3: **Ablation Study** of Visual Certainty Estimation (VCE) and Token Advantage Reweighting (TAR). Decoupled denotes that we decouple the verbalized confidence in training without VCE and TAR.

We conduct ablation studies to validate the effectiveness of each design component, as presented in Table 3. **First, RLCR with decoupling alone does not help calibration**, a variant that only changes the output to (c_{vis}, c_{reas}) but optimizes the same holistic Brier score performs nearly identically to RLCR, suggesting that explicit visual confidence supervision is necessary. **Second, on Visual Certainty Estimation (VCE)**, incorporating either entropy or KL-divergence significantly reduces ECE compared to the base model. Notably, the combination of both metrics yields best performance. We empirically observe that relying on a single metric leads to training instability: using only entropy tends to cause *entropy collapse*, while using only KL-divergence risks *entropy explosion*. We provide a detailed visualization and analysis of this phenomenon in § E.1. **Third, on Token Advantage Reweighting (TAR)**, applying TAR on top of VCE achieves the best overall performance. By uncertainty-aware reweighting the optimization of high-uncertainty tokens advantage, TAR further boosts accuracy to **0.727** and minimizes ECE to **0.098**, confirming that fine-grained advantage reweighting is essential for effective calibration.

5 Analyses

Validation of Visual Certainty Estimation To validate the effect of our visual certainty estimation, we evaluate its correlation with the quality of image captions. Specifically, we use the base model (Qwen3-VL-4B) to generate 1,500 dense

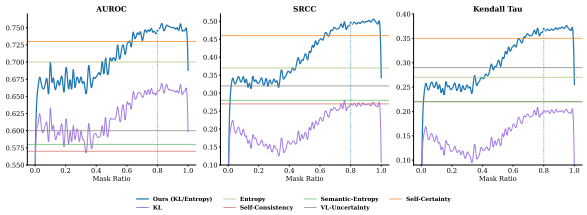


Figure 3: **Effectiveness of Visual Certainty Estimation.** Our estimation outperforms the strongest baseline, Self-Certainty, at mask ratios > 0.65. The vertical dashed line marks the mask ratio (0.8) adopted in the following experiments.

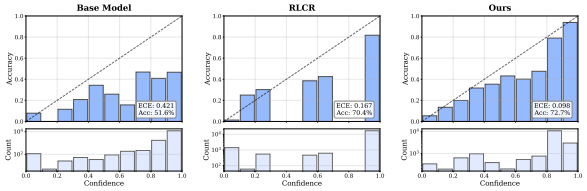


Figure 4: **Comparison with Holistic Confidence Calibration.** Reliability diagrams comparison: Base Model (Qwen3-VL-4B, Left), RLCR (Middle), and Ours (Right) across all evaluation datasets.

captions and employ `Gemini-3-pro-preview` as a judge to assess the image captions: (i) whether the caption is correct, and (ii) a quality scoring scalar $S_{quality} \in [0, 10]$. Next, we derive uncertainty estimation and utilize AUROC to measure the hallucination detection capability, as well as Spearman's rank correlation coefficient (SRCC) and Kendall's Tau to evaluate the correlation with quality scores. As illustrated in Figure 3, our proposed estimation surpasses strong baselines, including VL-Uncertainty (Zhang et al., 2024), Self-Consistency (Wang et al., 2023), and Self-Certainty (Kang et al., 2025), with AUROC=0.746, SRCC=0.496, and Kendall's Tau=0.370. Beyond the above validation, we also provide **more analyses of proposed estimation** regarding computation overhead, combination, and perturbation in Appendix § E.1.

Training Dynamics We present the training dynamics of Qwen3-VL-4B and Qwen3-VL-8B in Figure 14. As illustrated, training with visual certainty estimation exhibits fast initial convergence with ECE reducing to 0.1 in less than 100 steps, achieving higher calibration performance more efficiently. These results indicate that the proposed supervision signal not only improves calibration and reasoning accuracy, but also provides a strong and effective training signal.

	Unanswerable	Answerable	Δ
Qwen3-VL-4B	0.698	0.926	0.228
RLCR	0.532	0.937	0.405
Ours	0.218	0.834	**0.616**

Table 4: Comparison of Qwen3-VL-4B, RLCR, and Ours of Confidence Gap (Δ) in visual unanswerable and answerable problems of DynaMath.

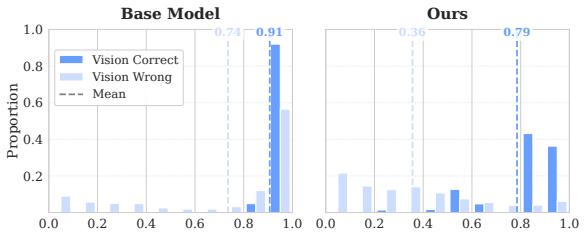


Figure 5: Visual confidence distribution comparison of visually correct and incorrect responses. Base Model: Qwen3-VL-4B.

Comparison with Holistic Confidence Calibration As shown in Figure 4, we construct reliability diagrams by binning predicted holistic confidence into $M{=}10$ equal-width bins in $[0, 1]$. For each bin, we plot the average confidence against the empirical accuracy (fraction of correct predictions), with the diagonal line indicating perfect calibration. The reliability diagrams of the base model reveal a severe ***overconfidence*** phenomenon, where predicted confidence consistently exceeds accuracy, particularly in high-confidence intervals, with a poor ECE of 0.421, while ours reduce ECE by over $4\times$ to 0.098. Visually, the confidence bins of our model align closely with the diagonal identity line, indicating that the predicted confidence scores serve as a reliable proxy for correctness. Detailed reliability diagrams of each benchmark are provided in Appendix E.2.

Effect of Proposed Visual Confidence To further investigate the effect of decoupled visual confidence, **first**, we randomly sample 1,000 problems across all benchmarks to evaluate Qwen3-VL-4B and our model, then manually label responses as visually correct or incorrect. As shown in Figure 5, the base model assigns high visual confidence to both visually correct and incorrect responses. However, decoupled visual confidence substantially lowers confidence on visually incorrect ones while maintaining relatively high confidence on visually correct ones, effectively distinguishing visual errors. **Second**, we further evaluate our model's performance on curated ***visually unanswerable*** problems. Specifically, we preserve the text input while removing images of `DynaMath` benchmark, which are generated dynamically to alleviate data contamination. We report the average confidence on the answerable (original) and visual unanswerable (lack of images) problems in Table 4. We observe that, compared to RLCR and the base model, our method achieves the largest **Confidence Gap (Δ)**, lowering confidence for unanswerable problems while maintaining high certainty for answerable ones.

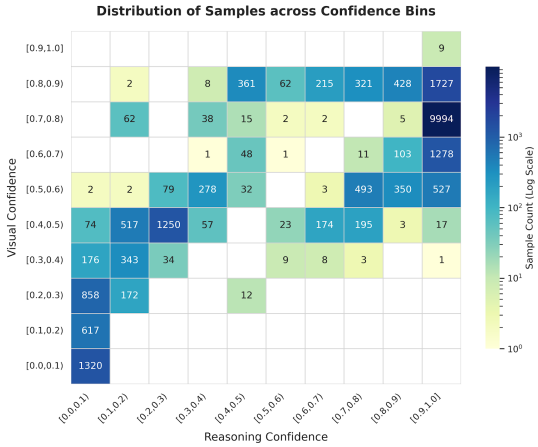


Figure 6: Heatmap of visual vs. reasoning confidence. The off-diagonal distribution indicates that the model can be visually certain but logically uncertain, and vice versa.

Decoupled Distribution of Visual and Reasoning Confidence To further verify that visual and reasoning confidence measure different things, we visualize their distribution across all samples in a 2D heatmap (Figure 6). The results show that the two scores are clearly separated. For instance, when visual confidence is high (the $[0.8, 0.9)$ bin), reasoning confidence still varies widely from 0.1 to 1.0. Similarly, even when reasoning confidence is at its highest ($[0.9, 1.0]$), visual confidence ranges from 0.3 to 1.0. This indicates that the model can be certain about what it sees but unsure about its logic, or vice versa. This clear separation confirms that a single overall confidence score is insufficient, as it mixes two different sources of uncertainty.

Qualitative Analysis of Token Advantage Reweighting To better understand token advantage reweighting, we visualize the most visually uncertain tokens in Figure 7. High uncertainty appears not only on visually grounded content to-

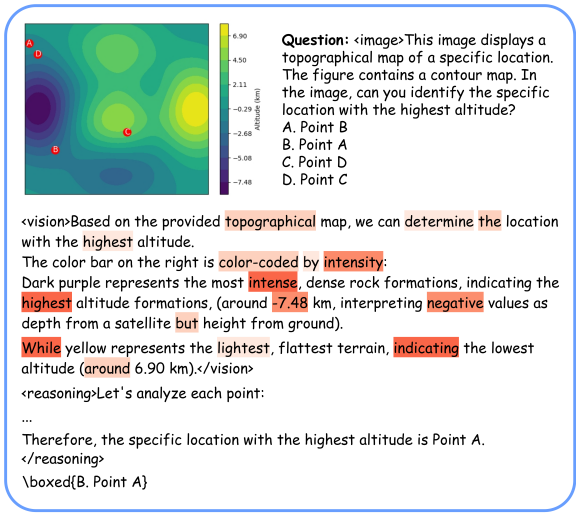


Figure 7: **Visualization of most visual-uncertain tokens.** Darker red represents higher uncertainty.

kens (*e.g.*, `highest`, `negative`), but also on logical connectives (*e.g.*, `while` and `indicating`) that help guide the reasoning process. This observation motivates our use of the visual certainty score for advantage reweighting, where tokens with lower visual certainty receive stronger penalties under negative advantage.

6 Conclusion

We present VL-Calibration, an RL-based calibration framework for LVLMs that decouples verbalized confidence into visual and reasoning confidence. We further propose an intrinsic visual certainty estimation signal based on KL-divergence under image perturbation and token entropy, and a token-level advantage reweighting strategy to better suppress ungrounded hallucinations. Experiments on thirteen benchmarks show that VL-Calibration consistently reduces calibration error while improving visual reasoning accuracy, and generalizes across model scales and architectures.

Limitations

While VL-Calibration demonstrates improved calibration and reasoning across diverse model families and scales, our current evaluation is constrained by computational resources. We present results on Qwen3-VL (4B to 30B) and InternVL3.5-4B, observing consistent gains across these settings. However, the efficacy of our method on larger-scale vision-language models (*e.g.*, 70B+) remains to be empirically verified, as reweighting behaviors may present distinct challenges.

Ethics Statement

While improved calibration enhances the reliability of VLM systems, it does not guarantee safety, particularly in high-stakes applications. Moreover, explicit confidence signals introduce potential risks of overreliance. In domains such as healthcare, law, or infrastructure monitoring, high confidence scores may induce automation bias, where users accept model outputs without sufficient independent verification. We emphasize that confidence estimates are intended to support, not replace, human judgment. Deployment in high-stakes settings should incorporate additional safeguards, including human-in-the-loop verification.

Acknowledgment

This work was supported in part by the Ningbo Youth Science and Technology Innovation Leading Talent Program (No. 2025QL059), the "Pioneer and Leading Goose" R&D Program of Zhejiang (No. 2025C02037), the Science and Technology Project of State Grid Beijing Electric Power Company (Project Title: Research on Urban Cable Network Operation Status Detection and Risk Identification Technology Based on Soft Robots and Artificial Intelligence, Project Number: 520246250003), and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2025. Improving uncertainty estimation through semantically diverse language generation. In *The Thirteenth International Conference on Learning Representations*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2026. Beyond binary rewards: Training LMs to reason about their uncertainty. In *The Fourteenth International Conference on Learning Representations*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

W Brier Glenn and 1 others. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yinghao Hu, Leilei Gan, Wenyi Xiao, Kun Kuang, and Fei Wu. 2025. Fine-tuning large language models for improving factuality in legal question answering. In *Proceedings of the 31st international conference on computational linguistics*, pages 4410–4427.

Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11181–11188.

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*.

Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2025. Taming overconfidence in llms: Reward calibration in rlhf. In *The Thirteenth International Conference on Learning Representations*.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. 2025. Conftuner: Training large language models to express their confidence verbally. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. 2023. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, and 1 others. 2025. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, and 1 others. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual

question answering with adversarial regularization. *Advances in neural information processing systems*, 31.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Fulin Shi, Wenyi Xiao, Bin Chen, Liang Din, and Leilei Gan. 2025. Revealer: Reinforcement-guided visual reasoning for element-level text-image alignment evaluation. *arXiv preprint arXiv:2512.23169*.

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

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. Lacie: Listener-aware finetuning for calibration in large language models. *Advances in Neural Information Processing Systems*, 37:43080–43106.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. Cocoa: A minimum bayes risk framework bridging confidence and consistency for uncertainty quantification in LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

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

Wenyi Xiao and Leilei Gan. 2025. Fast-slow thinking GRPO for large vision-language model reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2025. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25543–25551.

Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Zongrui Li, Ruirui Lei, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, and 1 others. 2024a. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications. *arXiv preprint arXiv:2410.15595*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024b. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

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 *The Twelfth International Conference on Learning Representations*.

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.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. 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)*, pages 15134–15186.

Peng Zhang, Wanggui He, Mushui Liu, Wenyi Xiao, Siyu Zou, Yuan Li, Xingjian Wang, Guanghao Zhang, Yanpeng Liu, Weilong Dai, and 1 others. 2026. Fuse: Fine-grained and semantic-aware learning for unified image understanding and generation. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 40, pages 28355–28363.

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. 2025. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, pages 169–186, Cham. Springer Nature Switzerland.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. 2024. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*.

Yang Zhao, Kai Xiong, Xiao Ding, Li Du, YangouOuyang, Zhouhao Sun, Jiannan Guan, Wenbin Zhang, Bin Liu, Dong Hu, Bing Qin, and Ting Liu. 2025. UFO-RL: Uncertainty-focused optimization for efficient reinforcement learning data selection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ziang Zhou, Tianyuan Jin, Jieming Shi, and Li Qing. 2025. Steerconf: Steering LLMs for confidence elicitation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models.

This Appendix for *"VL-Calibration: Decoupled Verbalized Confidence for Large Vision-Language Models Reasoning"* is organized as follows:

- **Experimental Setup and Reproducibility.** In §A, we describe the experimental setup in detail, including training details (§A.1) and evaluation details (§A.2). The evaluation section further covers the adopted metrics (§A.2.1) and benchmark datasets (§A.2.2).
- **Prompts.** In §B, we provide the complete *system prompt* used for VL-Calibration training and inference.
- **Statistical Significance Analysis.** In §C, we report statistical significance analyses to verify the robustness and reliability of the observed performance improvements.
- **Additional Results.** In §D, we present detailed quantitative results, including the complete main results (§D.1).
- **Additional Analyses.** In §E, we provide more in-depth analyses of VL-Calibration, including visual certainty estimation (§E.1), reliability diagrams (§E.2), behavior on unanswerable visual problems (§E.3), training dynamics (§E.4), failure mode analysis (§E.5), and case study in Figure 17.

A Experimental Setup

A.1 Training Details

Table 5: Training Hyperparameters

Hyperparameter	Value
Model	Qwen3-VL
Epochs	15
Learning Rate	1e-6
Train Batch Size	256
Temperature	1.0
Rollout per Prompt	8
Prompt Max Length	4096
Generation Max Length	4096
Precision	BF16
Max Pixels	1000000
λ_{acc}	1.0
λ_{cal}	2.0
λ_{vis}	0.4
λ_{TAR}	0.1

We implement VL-Calibration using Qwen3-VL-4B, 8B, and 30B-A3B as our base models. Below, we detail our training setup and hyperparameters.

General Training Hyperparameters. For VL-Calibration training, we use our 12K dataset with a learning rate of 1e-6, a batch size of 256. We set the maximum sequence length to 4096 for both prompts and generation, and apply BF16 precision throughout training. The training process runs for 15 epochs, requiring approximately 240 H200 GPU hours for Qwen3-VL-4B model, and 450 H200 GPU hours for Qwen3-VL-8B model, 1900 H200 GPU hours for Qwen3-VL-30B-A3B model.

Method-specific Training Hyperparameters. For our reinforcement learning approach, we employ a temperature of 1.0, 8 rollouts per prompt. For the reward weights, we set accuracy reward weight $\lambda_a = 1.0$, calibration reward weight $\lambda_c = 2.0$ and visual certainty reward weight $\lambda_v = 0.4$. The mask ratio is 0.8.

Computation Environment. All training experiments were conducted using H200 GPUs. Model inference in evaluations is performed using the vLLM framework (Kwon et al., 2023), and our training implementation extends the VeRL codebase (Sheng et al., 2024).

The complete set of hyperparameters is provided in Table 5. We commit to releasing all the code, data, and model checkpoints for experimental results reproducibility.

A.2 Evaluation Details

A.2.1 Evaluation Metrics

We use the following evaluation metrics:

1. **Accuracy:** A measure of reasoning performance.
2. **Area Under the Receiver Operating Characteristic Curve (AUROC):** Measures calibration ability of classifier to distinguish between positive/negative classes across thresholds.

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))\, dt \quad (13)$$

   where TPR is the True Positive Rate and FPR is the False Positive Rate.
3. **Expected Calibration Error (ECE):** Calibration metric that groups confidences into bins

and computes difference between the average correctness and confidence.

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{14}$$

where M is the number of bins , B_m is the set of samples in bin m, and N is the number of samples. We use M=10.

A.2.2 Evaluation Datasets

This section provides a brief analysis of the eight benchmarks used in our main evaluation. We deliberately selected this suite to cover a wide spectrum of challenges, from domain-specific mathematical skills to general logical cognition, ensuring a holistic assessment of our model's capabilities.

Mathematical and Geometric Reasoning.

- **DynaMath** (Zou et al., 2024) is a unique benchmark designed to test the *robustness* of visual mathematical reasoning. Instead of using a static set of questions, it employs program-based generation to create numerous variants of seed problems, systematically altering numerical values and function graphs to challenge a model's ability to generalize rather than memorize.
- **Geo3k** (Lu et al., 2021) is a large-scale benchmark focused on high-school level *geometry*. Its key feature is the dense annotation of problems in a formal language, making it particularly well-suited for evaluating interpretable, symbolic reasoning approaches.
- **MathVerse** (Zhang et al., 2025) is specifically designed to answer the question: "Do MLLMs truly see the diagrams?" It tackles the problem of textual redundancy by providing six distinct versions of each problem, systematically shifting information from the text to the diagram. This allows for a fine-grained analysis of a model's reliance on visual versus textual cues.
- **MathVista** (Lu et al., 2024) is a benchmark designed to combine challenges from diverse mathematical and visual tasks.
- **MATH-Vision** (Wang et al., 2024) elevates the difficulty by sourcing its problems from *real math competitions* (e.g., AMC, Math Kangaroo). Spanning 16 mathematical disciplines and 5 difficulty levels, it provides a challenging testbed for evaluating advanced, competition-level multimodal reasoning.
- **We-Math** (Qiao et al., 2025) introduces a novel, human-centric evaluation paradigm. It assesses reasoning by *decomposing composite problems into sub-problems* based on a hierarchy of 67 knowledge concepts. This allows for a fine-grained diagnosis of a model's specific strengths and weaknesses, distinguishing insufficient knowledge from failures in generalization.

Logical Reasoning.

- **LogicVista** (Xiao et al., 2024b) is designed to fill a critical gap by evaluating *general logical cognition* beyond the mathematical domain. It covers five core reasoning skills (inductive, deductive, numerical, spatial, and mechanical) across a variety of visual formats, testing the fundamental reasoning capabilities that underlie many complex tasks.

Visual-Dominant Reasoning.

- **SuperClevr** (Li et al., 2023) is a counting benchmark for testing the perception capability.
- **MathVerse**$_V$ (Zhang et al., 2025) We also report MathVerse's vision-dependent subset result, where the problem cannot be solved without its visual input.

Multi-discipline Reasoning.

- **A-OKVQA** (Schwenk et al., 2022) is a benchmark requiring a broad base of commonsense and world knowledge to answer. The questions generally cannot be answered by simply querying a knowledge base, and instead require some form of commonsense reasoning about the scene depicted in the image.
- **MMK12** (Meng et al., 2025) is a benchmark focused on K-12 level multimodal STEM problems. It provides a strong test of foundational scientific reasoning skills that are essential for more advanced applications.
- **MMMU-Pro** (Yue et al., 2025) is a hardened version of the popular MMMU benchmark.

It was specifically created to be unsolvable by text-only models by filtering out questions with textual shortcuts, augmenting the number of choices to reduce guessing, and introducing a vision-only format. It serves as a strong test of a model's ability to seamlessly integrate visual and textual information in a high-stakes, academic context.

- **ViRL-39K-Test** (Wang et al., 2025a) is a holdout dataset containing 1,800 problems from ViRL-39K excluding VL-Calibration-12K, covering comprehensive topics and categories: from grade school problems to broader STEM and Social topics; reasoning with charts, diagrams, tables, documents, spatial relationships, etc.

B Prompts

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 visual 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 <visual confidence> </visual confidence> and <reasoning confidence> </reasoning confidence> tags respectively.

C Statistical Significance Analysis

To evaluate the statistical significance of our method on the qwen3-VL-4B model, we conducted 5 independent inference runs with different random seeds. For each run, we recorded the following metrics on the evaluation set: Accuracy, AUROC, and ECE. The average results across the 5 runs are reported in the main paper (ACC = 0.727, AUROC = 0.763, ECE = 0.098). To assess the stability and significance of these results, we computed the mean and standard deviation for each metric as follows:

Table 6: Stability Analysis across 5 Random Seeds. We report the Mean ± Standard Deviation.

Metric	Acc (↑)	AUROC (↑)	ECE (↓)
Ours	0.727 ± 0.008	0.763 ± 0.009	0.098 ± 0.005
RLCR	0.704 ± 0.007	0.694 ± 0.010	0.167 ± 0.006

To confirm whether the improvements over the baseline are statistically significant, we performed paired t-tests on the metrics collected from the 5 independent runs. The significance level was set to $\alpha = 0.05$. The resulting p-values for Accuracy, AUROC, and ECE were $p = 0.012$, $p = 0.008$, and $p = 0.004$, respectively, indicating statistically significant improvements in all three metrics. Overall, these results demonstrate that our proposed decoupled confidence calibration method not only achieves stable performance across different random seeds but also significantly outperforms the baseline method in terms of calibration and accuracy on the Qwen3-VL-4B model.

D Results

D.1 Detailed Main Results

In this section, we report more detailed main results. We report detailed baseline results on all benchmarks in Figure 15 and Figure 16. We observe that, RLCR is the strongest baseline, outperforming other methods in both reasoning and calibration performance.

E Analyses

E.1 Visual Certainty Estimation

Analysis of Certainty Estimation Combination Complementing the effectiveness validation in Section 5, we further examine the training dynamics in Figure 8. We observe that single-metric supervision is prone to optimization pathology: optimizing Entropy only leads to *entropy collapse*, while KL only results in *entropy explosion*. In contrast, our estimation leverages the two metrics as mutual regularizers, effectively preventing both extremes to ensure a robust and stable training trajectory.

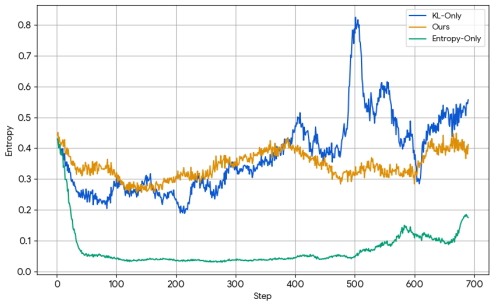


Figure 8: Entropy curves of different estimation: KL-Only, Entropy-Only, Ours (Combination of KL and Entropy). The entropy curve of ours shows training stability compared to others.

Estimation Computation Overhead Analysis While visual certainty estimation is effective, it introduces additional computational overhead due to the second forward pass required to compute the KL divergence. We analyze this overhead in terms of both runtime and monetary cost, comparing our approach with self-consistency and external annotator baselines.

In terms of runtime, relative to GRPO, the additional forward pass incurs a 11% time overhead: it adds 15 seconds to the 140-second step time for the 8B model, and 12 seconds to the 100-second step time for the 4B model. By contrast, although self-consistency in GRPO avoids rollout latency, it relies on semantic clustering, which can require up to $O(N^2)$ inferences from an external NLI model in the worst case, leading to substantial latency.

In terms of cost, external annotators are expensive. For example, using `Gemini-3-pro-preview` costs $0.03 per judgment, amounting to $43,200 per training cycle.

Perturbation Design Analysis To investigate the effect of perturbations on the visual certainty estimation, we evaluate Qwen3-VL-4B across multiple perturbation types and mask ratios. As shown in Table 7, global perturbations (Gaussian Blur and Noise) achieve comparable effectiveness to our default Random Masking (80%). In contrast, Center Crop underperforms because preserving central objects fails to sufficiently disrupt visual cues.

Regarding mask ratio, performance improves as the ratio increases from 20% to 80%, since weaker masks leave excessive visual information that trivializes the grounding evaluation. However, complete masking (100%) slightly degrades performance, aligning with Figure 3 where optimal certainty estimation peaks at a 0.8 ratio. Overall, these results confirm that our metric is robust across perturbation designs.

Table 7: Robustness evaluation of Qwen3-VL-4B across different perturbation types and severities.

Perturbation Type	Settings	Accuracy (↑)	AUROC (↑)	ECE (↓)
Random Masking (Ours)	**Ratio = 80%**	**0.727**	**0.763**	**0.098**
Random Masking	Ratio = 20%	0.650	0.688	0.188
Random Masking	Ratio = 50%	0.682	0.711	0.151
Random Masking	Ratio = 100%	0.691	0.722	0.142
Gaussian Blur	$\sigma = 5.0$	0.708	0.758	0.105
Gaussian Noise	$\sigma = 0.5$	0.714	0.749	0.110
Center Crop	Crop 50%	0.699	0.706	0.148

Aggregation Function Analysis To justify using the Harmonic Mean for aggregating visual (c_{vis}) and reasoning (c_{reas}) confidences, we compare it against alternative aggregation functions.

Table 8: Comparison of different aggregation functions for decoupled confidence on Qwen3-VL-4B.

Aggregation Function	Accuracy (↑)	AUROC (↑)	ECE (↓)
Arithmetic Mean	0.725	0.741	0.145
Geometric Mean	0.724	0.752	0.107
Minimum	0.718	0.749	0.121
Harmonic Mean	**0.727**	**0.763**	**0.098**

Technically, a reliable aggregated confidence should only be high when both decoupled confidences are high. The *Arithmetic Mean* is overly optimistic; for example, a severe visual hallucination ($c_{vis} \approx 0$) coupled with strong language priors ($c_{reas} \approx 1$) still yields an aggregated score of 0.5. The *Minimum* function is strictly conservative but suffers from vanishing reward signals for the non-minimum term, which hinders effective joint optimization. Compared to the *Geometric Mean*, the *Harmonic Mean* serves as a more conservative penalty.

Empirically, we evaluate these functions during the RL training of Qwen3-VL-4B. As shown in Table 8, the Harmonic Mean achieves the optimal balance, yielding the highest accuracy and AUROC, along with the lowest ECE.

E.2 Reliability Diagrams

We illustrate the reliability diagrams of VL-Calibration-4B and VL-Calibration-8B on each benchmark in Figure 11 and Figure 12.

E.3 Visually Unanswerable Problems

In Figure 13, we provide concrete confidence distributions of Base Model, RLCR, and Ours across visually answerable and unanswerable problems.

E.4 Training Dynamics

We present the training dynamics of Qwen3-VL-4B and Qwen3-VL-8B in Figure 14.

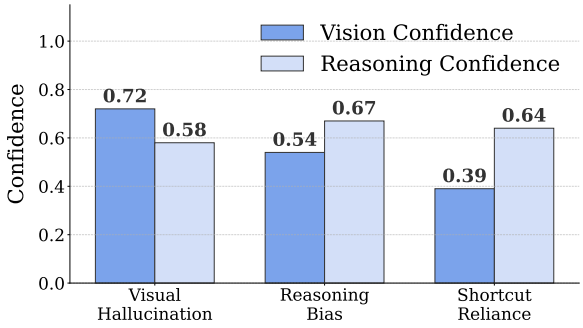


Figure 9: **Confidence of overconfident wrong answers.**

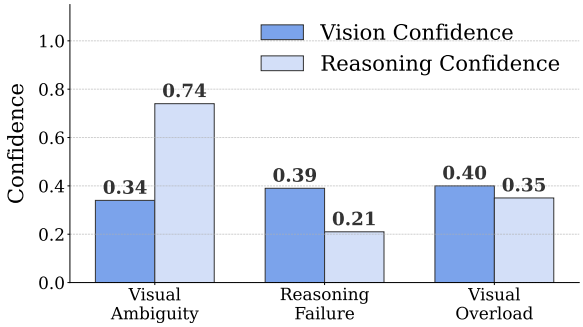


Figure 10: **Confidence of underconfident correct answers.**

E.5 Failure Mode Analysis

All analyses are conducted on a manually curated subset of 1,000 samples drawn from metric dataset. We systematically reviewed these samples to categorize failure modes based on their holistic confidence scores and prediction correctness.

Wrong predictions with high holistic confidence. We analyze wrong predictions with high holistic confidence (> 0.40), as shown in Figure 9. These cases are categorized into three types. *Visual Hallucination* is the dominant source, where the model confidently infers nonexistent or misperceived visual attributes, leading to incorrect answers and indicating overconfidence. *Reasoning Bias* arises when the model relies on flawed logical shortcuts, producing confident yet incorrect conclusions. The remaining cases involve *Shortcut Reliance*, where prior-driven decision making overrides image-specific evidence, resulting in higher confidence in reasoning rationale.

Correct predictions with low holistic confidence. We analyze correct predictions with low holistic confidence (< 0.50), as shown in Figure 10, and categorize them into three types. The majority fall into *Visual Uncertainty*, where critical visual cues are ambiguous or hard to distinguish, leading to lower visual confidence than reasoning confidence. This highlights perception-driven uncertainty captured by our decoupled confidence modeling. The remaining cases include *Reasoning Failure*, where the task surpasses the model's reasoning capacity causing uncertain guesses, and *Visual Overload*, where extremely dense visual inputs make perception difficult. Both latter types reflect increased task difficulty, resulting in uniformly low confidence across visual and reasoning.

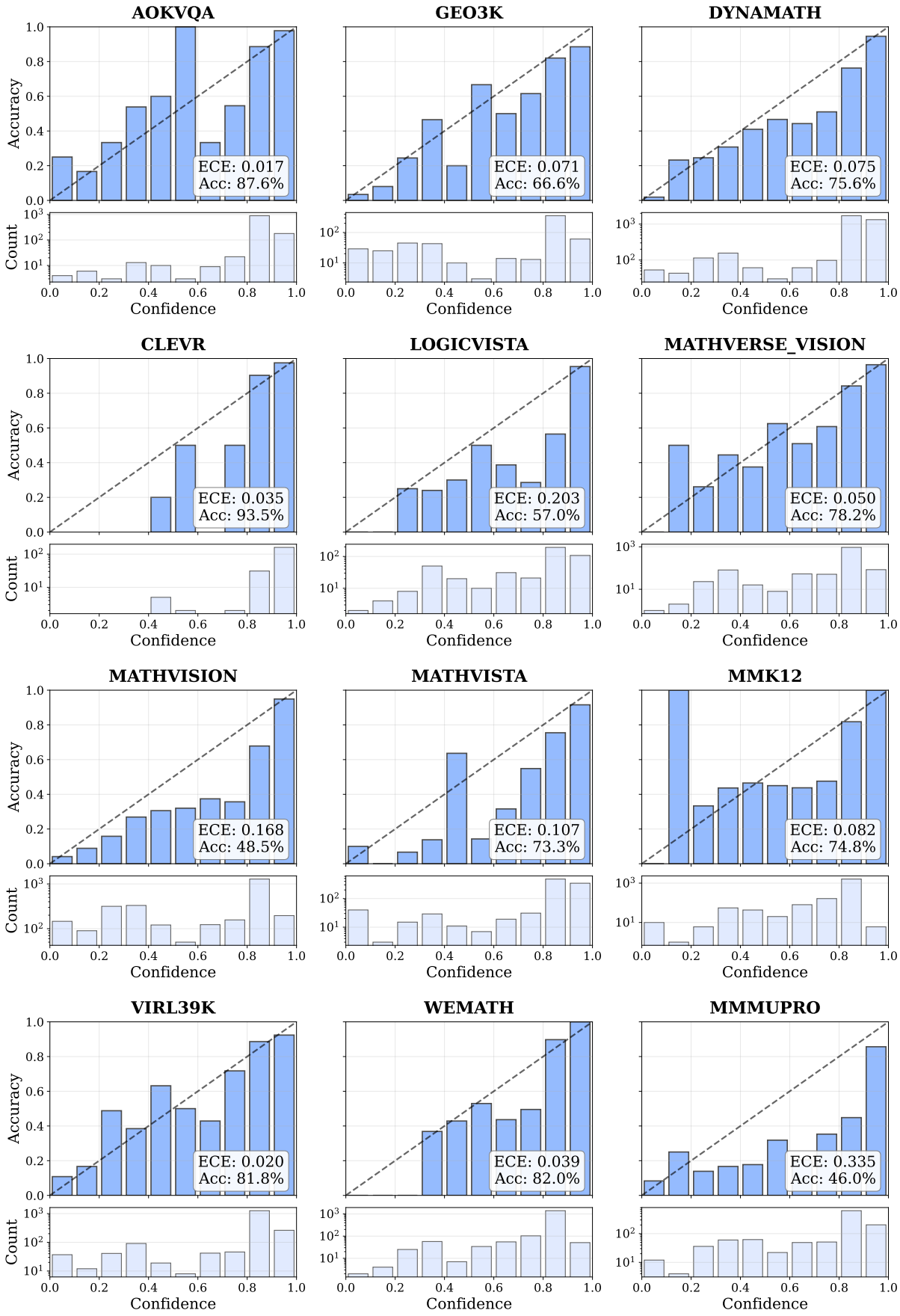


Figure 11: Detailed reliability diagrams of VL-Calibration-4B across the evaluation datasets. While Table 9 reports metrics averaged over 8 rollouts (avg@8), this figure illustrates the results of a single representative rollout.

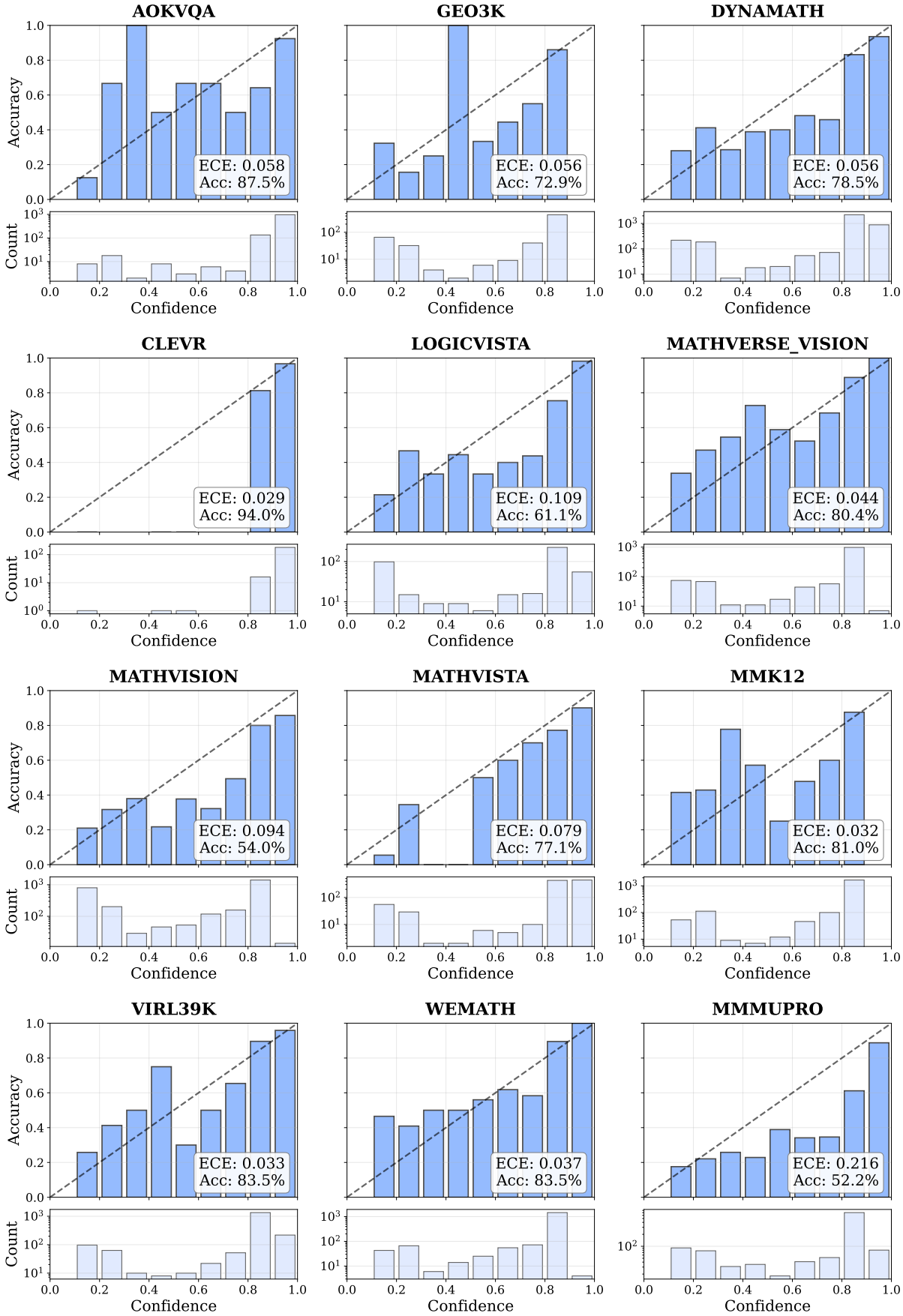


Figure 12: Detailed reliability diagrams of VL-Calibration-8B across the evaluation datasets.

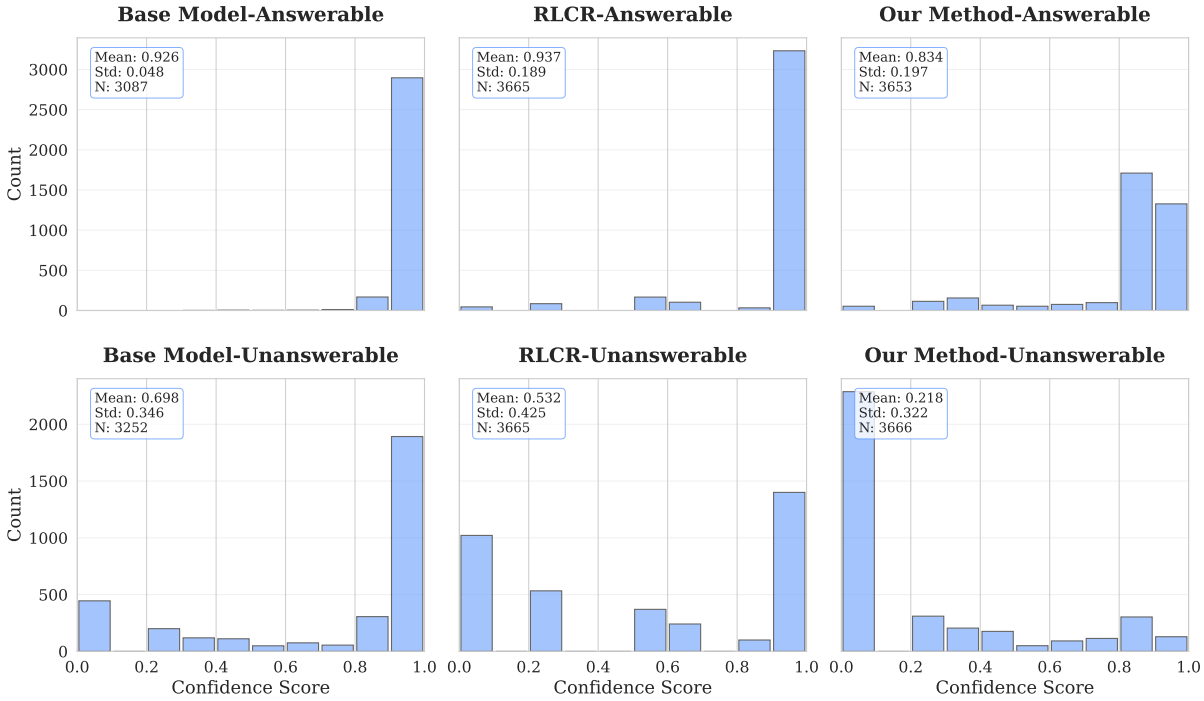


Figure 13: **Confidence Distribution across Visually Answerable and Unanswerable Problems.** We visualize the distribution of confidence scores for the Base Model, RLCR, and Our Method (columns) on both Answerable and Unanswerable datasets (rows). While baselines tend to remain overconfident even on unanswerable queries (bottom row), **Our Method** exhibits a significant distributional shift towards lower confidence, demonstrating superior capability in recognizing visual input lack.

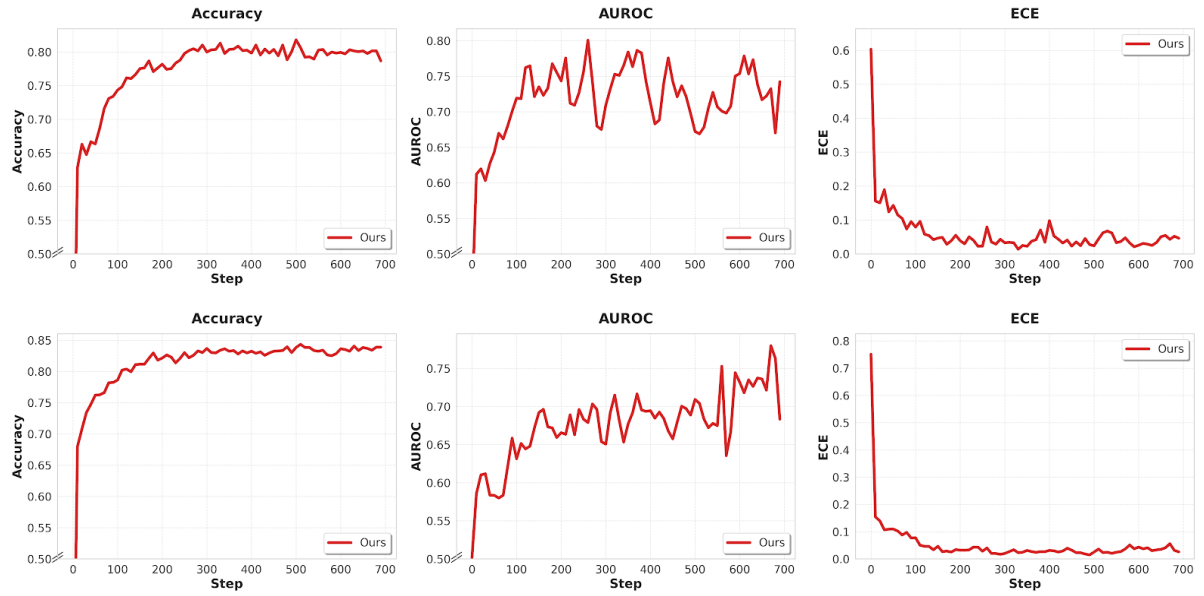


Figure 14: Training dynamics of Qwen3-VL-4B (upper) and Qwen3-VL-8B (bottom). We visualize the ACC, AUROC, ECE curves on the validation set.

Table 9: **Comparison of Base Model (Qwen3-VL-4B), Best, and Our Method across various benchmarks.** We report Accuracy, AUROC, and ECE. Best results are highlighted in **bold**.

Benchmark	Accuracy ↑			AUROC ↑			ECE ↓		
	Base	Best	Ours	Base	Best	Ours	Base	Best	Ours
Mathematical and Geometric Reasoning									
DynaMath	0.486	0.718	**0.753**	0.513	0.716	**0.797**	0.423	0.165	**0.081**
Geo3K	0.514	0.616	**0.671**	0.504	**0.801**	0.792	0.773	0.159	**0.073**
MathVerse	0.426	0.796	**0.807**	0.416	0.659	**0.735**	0.561	0.142	**0.042**
MathVision	0.171	0.440	**0.483**	0.501	**0.814**	0.800	0.794	0.207	**0.170**
MathVista	0.679	**0.772**	0.730	0.566	0.710	**0.778**	0.254	0.132	**0.107**
WeMath	0.580	0.771	**0.820**	0.593	0.647	**0.802**	0.268	0.164	**0.048**
Logical Reasoning									
LogicVista	0.456	0.519	**0.570**	0.615	0.757	**0.794**	0.315	0.232	**0.203**
Visual-Dominant Reasoning									
CLEVR	0.920	0.935	**0.935**	0.517	0.577	**0.797**	**0.025**	0.058	0.035
MathVerse$_V$	0.283	0.748	**0.781**	0.519	0.669	**0.721**	0.508	0.171	**0.056**
Multi-discipline Reasoning									
A-OKVQA	0.836	0.861	**0.875**	0.584	0.592	**0.695**	0.022	0.112	**0.017**
MMK12	0.489	0.741	**0.747**	0.468	0.651	**0.714**	0.432	0.182	**0.083**
MMMU-Pro	0.249	0.436	**0.458**	0.610	0.694	**0.735**	0.474	0.340	**0.335**
ViRL-39K-Test	0.620	0.796	**0.816**	0.406	0.729	**0.753**	0.622	0.113	**0.026**
Average	0.516	0.704	**0.727**	0.524	0.694	**0.763**	0.421	0.167	**0.098**

Table 10: **Comparison of Base Model (8B), Best, and Our Method across various benchmarks.** We report Accuracy, AUROC, and ECE. Best results are highlighted in **bold**.

Benchmark	Accuracy ↑			AUROC ↑			ECE ↓		
	Base	Best	Ours	Base	Best	Ours	Base	Best	Ours
Mathematical and Geometric Reasoning									
DynaMath	0.680	0.766	**0.784**	0.576	0.667	**0.769**	0.460	0.160	**0.058**
Geo3K	0.514	0.621	**0.729**	0.556	0.761	**0.780**	0.734	0.192	**0.056**
MathVerse	0.622	0.813	**0.838**	0.504	0.656	**0.742**	0.372	0.129	**0.055**
MathVision	0.266	0.473	**0.540**	0.527	0.771	**0.815**	0.428	0.249	**0.094**
MathVista	0.678	0.733	**0.771**	0.574	0.644	**0.753**	0.459	0.198	**0.079**
WeMath	0.699	0.801	**0.836**	0.567	0.730	**0.777**	0.388	0.110	**0.039**
Logical Reasoning									
LogicVista	0.508	0.600	**0.611**	0.580	0.688	**0.836**	0.308	0.253	**0.109**
Visual-Dominant Reasoning									
CLEVR	0.910	0.935	**0.940**	0.545	0.495	**0.723**	0.332	0.069	**0.029**
MathVerse$_V$	0.573	0.776	**0.804**	0.502	0.660	**0.743**	0.398	0.162	**0.052**
Multi-discipline Reasoning									
A-OKVQA	0.829	0.872	**0.875**	0.642	0.593	**0.691**	0.057	0.107	**0.059**
MMK12	0.585	0.780	**0.809**	0.506	0.691	**0.777**	0.301	0.131	**0.039**
MMMU-Pro	0.383	0.518	**0.522**	0.579	0.634	**0.740**	0.518	0.357	**0.220**
ViRL-39K-Test	0.689	0.811	**0.835**	0.537	0.723	**0.783**	0.460	0.109	**0.033**
Average	0.610	0.731	**0.761**	0.553	0.670	**0.764**	0.401	0.171	**0.071**

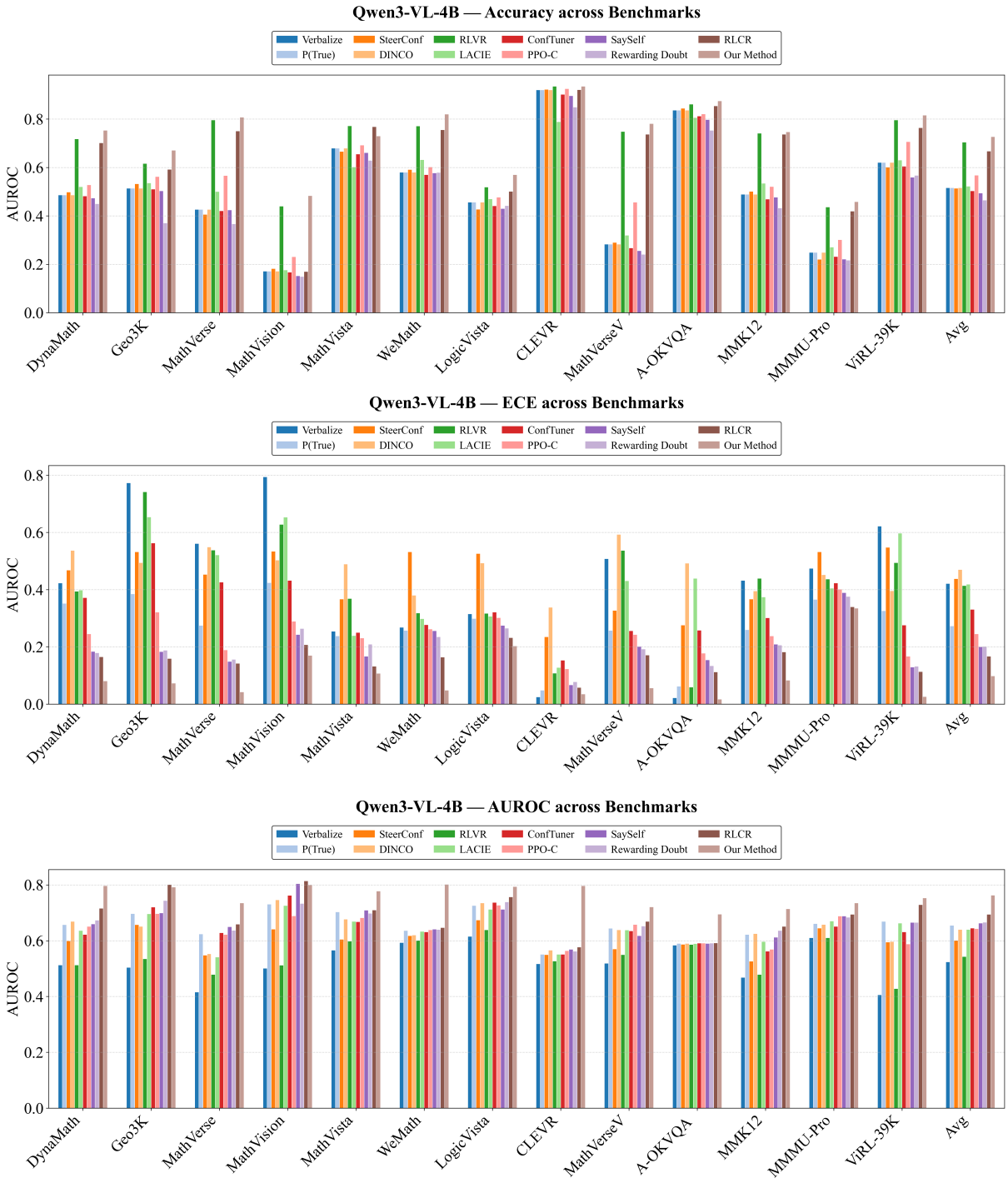


Figure 15: Baselines Performance comparison with Qwen3-VL-4B in terms of Accuracy, ECE, and AUROC.

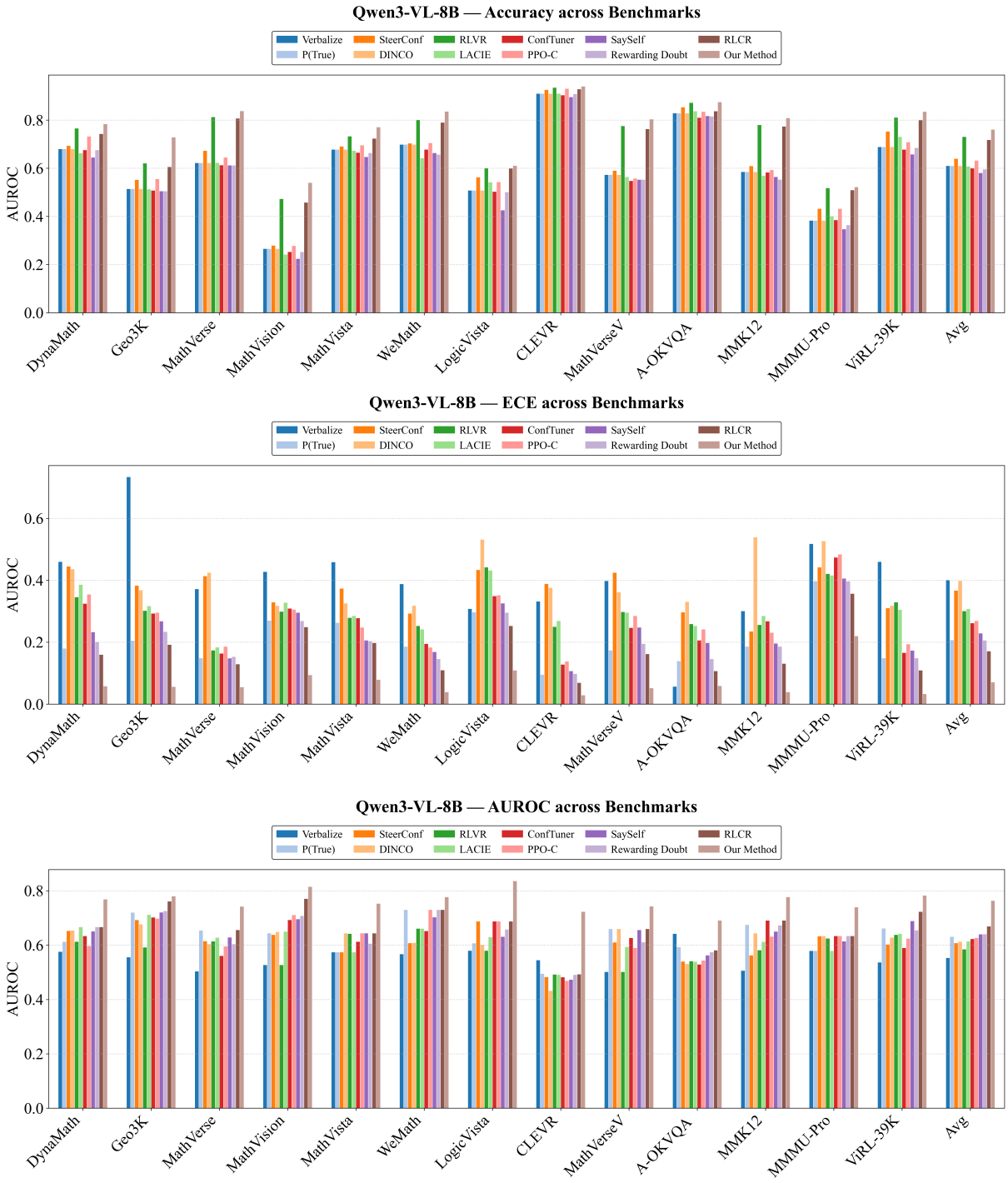
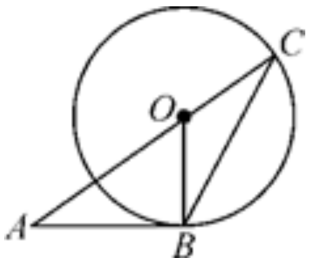


Figure 16: Baselines Performance comparison with Qwen3-VL-8B in terms of Accuracy, ECE, and AUROC.

Q: As shown in the figure, if angle A = 36.0, then angle C is equal to ( )?
A.36°
B.54°
C.60°
D.27°

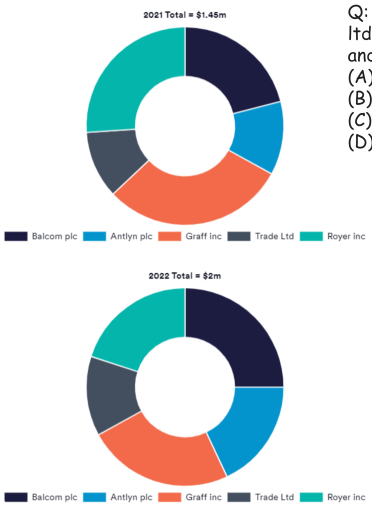
Dataset: MathVerse
Ground Truth: D



<vision>Arrow label under point and base with its name. Point B labeled under the base line. The circle has center O which lies on chord BC (since O is on segment AC and also inside circle). Since O is the center and also on AC, and OB is perpendicular to AB as shown, and B is on the circle, OB should be the radius. Key observation: O lies on chord AC and is the center.O is on AC and inside the circle, and OB ⊥ AB.</vision>

<reasoning> Therefore, ∠ABO = 90°. Given ∠A = 36° and ∠ABO = 90°, in triangle ABO: ∠AOB = 180° - 90° - 36° = 54°. Since O lies on AC, ∠AOB and ∠BOC are supplementary: ∠BOC = 180° - ∠AOB = 180° - 54° = 126°. In triangle OBC, since OB = OC, the base angles are equal: ∠OBC = ∠OCB = ∠C. Using the sum of angles in triangle OBC: ∠BOC + 2∠C = 180° 126° + 2∠C = 180° 2∠C = 54° ∠C = 27° This matches option D.</reasoning>

<confidence>
<vision_confidence>8</vision_confidence>
<reasoning_confidence>10</reasoning_confidence>
</confidence>
\boxed{D}



Q: What was the combined value of Balcom plc's and Trade ltd's share of the industry in 2022? Select from A, B, C, and D.
(A) $464k
(B) $551k
(C) $604k
(D) $760k

Dataset: Logic-Vista
Ground Truth: D

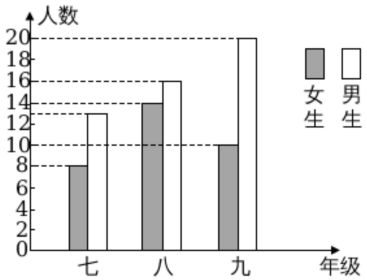
<vision>Color correct. Two donut charts, labeled. No share data. As the donut charts lack percentages or exact values, calculating the specific combined value for Balcom plc and Trade Ltd from the $2m total is impossible without additional data. </vision>

<reasoning>While the question asks for a specific dollar amount, the visual alone doesn't provide the necessary scale or proportions. Despite being a test question, the absence of measurable data means an accurate calculation cannot be performed based on the visual alone, leading to an uncertain selection.</reasoning>

<confidence>
<vision_confidence>2</vision_confidence>
<reasoning_confidence>0</reasoning_confidence>
</confidence>
\boxed{C}

Q: As shown in the figure, which of the following statements is correct?
A. The number of students in Grade 7 is the highest.
B. The number of boys in Grade 9 is twice the number of girls.
C. The number of girls in Grade 9 is more than the number of boys.
D. The number of students in Grade 8 is more than the number of students in Grade 9.

Dataset: MMK12
Ground Truth: B



<vision>Black background noted. Grey and white bars on dark base. The chart displays bars for different grades, likely with grey representing girls and white representing boys. Analyzing the groups: Grade 7 (First group): White bar is taller than the grey bar. Grade 8 (Second group): Both bars are tall and roughly equal. Grade 9 (Third group): The white bar is significantly taller than the grey bar.</vision>

<reasoning>Evaluating the options: Option A: Incorrect. Grade 9 appears to have a higher total count than Grade 7.
...
Option B: Visually, the white bar in Grade 9 is approximately twice the height of the grey bar. This suggests the number of boys is roughly double the number of girls, making this the most plausible answer.</reasoning>

<confidence>
<vision_confidence>5</vision_confidence>
<reasoning_confidence>4</reasoning_confidence>
</confidence>
\boxed{B}

Figure 17: Case Study of Qwen3-VL-4B model trained with our method. Figure (a) (upper) showcases a correct response with high confidence, and Figure (b) (middle) illustrates a incorrect response with low confidence. In Figure (c) demonstrate a correct response with modest confidence.