INSTINCT VS. REFLECTION: UNIFYING TOKEN AND VERBALIZED CONFIDENCE IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have demonstrated exceptional capabilities in various perception and reasoning tasks. Despite this success, ensuring their reliability in practical deployment necessitates robust confidence estimation. Prior works have predominantly focused on text-only LLMs, often relying on computationally expensive self-consistency sampling based on agreement across sampled responses. In this paper, we extend this to multimodal settings and conduct a comprehensive evaluation of MLLMs' response confidence estimation. Our analysis reveals a significant *instinct-reflection misalignment*: the model's implicit token-level support frequently diverges from its verbal self-assessment confidence. To address this misalignment, we propose a monotone confidence fusion framework to merge dual-channel signals and cross-channel consistency to estimate correctness. Subsequently, an order-preserving mean alignment step is applied to correct global bias, which improves calibration while preserving the risk–coverage trade-off for selective prediction. Experiments on diverse open-source and closed-source MLLMs show that our method consistently yields more reliable confidence estimates and improves both calibration and failure prediction.

1 INTRODUCTION

A fundamental aspect of human intelligence is the ability to correctly express confidence in our responses (Koriat, 1993; Fleming & Daw, 2017; Ernst & Banks, 2002; Fleming, 2024). Whether based on immediate instinct or deep reflection, humans can generally discern the boundaries of their own knowledge (Pouget et al., 2016; Yeung & Summerfield, 2012). In recent years, Multimodal Large Language Models (MLLMs) (Achiam et al., 2023; Wang et al., 2025; Bai et al., 2025c; Abdin et al., 2024; Dang et al., 2024) have demonstrated remarkable capabilities across diverse scenarios, ranging from fine-grained visual perception to complex logical reasoning. However, high benchmark accuracy alone does not guarantee reliability. For models to approach human-like intelligence, they should not only provide correct answers but also faithfully quantify their certainty (Zhao et al., 2024; Tian et al., 2023; Huang et al., 2024a). Precise confidence estimation is critical for determining system reliability, enabling effective risk assessment, and mitigating errors in real-world deployment (Fort et al., 2019; Ovadia et al., 2019; Liu et al., 2020; Farquhar et al., 2024).

However, assessing **how well a model knows what it knows** remains an open challenge. MLLMs commonly suffer from poor confidence calibration such as overconfident hallucinations or underconfident hesitations (Xiong et al., 2023; Groot & Valdenegro-Toro, 2024). Prior works (Taubenfeld et al., 2025; Kumar et al., 2024; Geng et al., 2024) have primarily explored confidence estimation for text-only Large Language Models (LLMs) from a machine learning calibration perspective. These approaches generally fall into two categories: logits-based methods (Duan et al., 2024; Kumar et al., 2024; Xie et al., 2024; Kull et al., 2019; Shen et al., 2024), which utilize the internal probability scores of open-source models, and self-consistency methods (Xiong et al., 2023; Prato et al., 2024; Li et al., 2024; Xie et al., 2024; Li et al., 2025; Du et al., 2025) which measure the agreement among multiple reasoning paths for closed-source models. As depicted in Figure 1, self-consistency methods operate by sampling multiple responses (*e.g.*, via different prompts or parameters) and aggregating them to derive a final confidence score. While effective for LLMs, this iterative process is cost-prohibitive due to repeated API calls (Xiong et al., 2023; Dang et al., 2025). As a result, existing text-centric

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

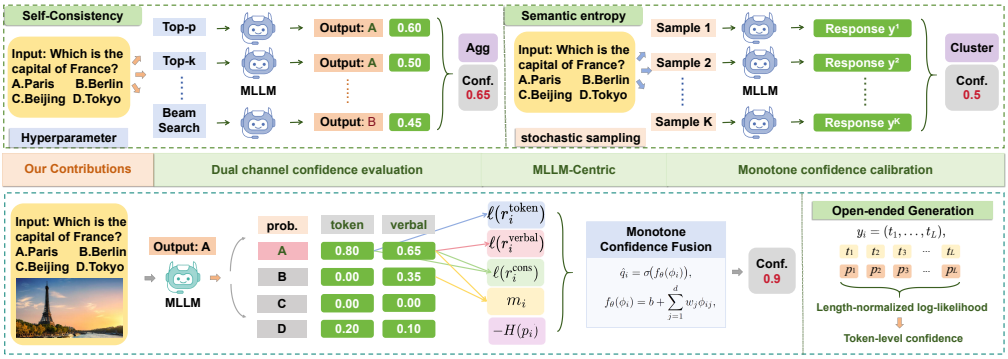


Figure 1: Comparison of confidence estimation paradigms. Top: sampling-based approaches, including self-consistency and semantic entropy, estimate confidence by aggregating multiple sampled responses. Bottom: Our framework extracts token confidence (instinct) and verbalized confidence (reflection) in a single pass and performs monotone confidence fusion.

paradigms do not seamlessly transfer to MLLMs, and establishing a systematic confidence evaluation framework in the multimodal setting remains an underexplored problem.

Addressing these gaps, this paper focuses on evaluating and calibrating the confidence estimation of MLLMs. We conduct a comprehensive evaluation of confidence estimation across diverse multimodal benchmarks. Specifically, we extract two confidence signals: **implicit token confidence** (instinct) and **explicit verbalized confidence** (reflection). To ensure generalizability, we adopt a unified black-box protocol for both open-source and proprietary MLLMs. We focus on a practical scenario where only single-token probabilities are available, without access to model weights or gradients. Our empirical analysis reveals a critical phenomenon: *instinct–reflection confidence misalignment*. For instance, a model may demonstrate high verbalized confidence while simultaneously exhibiting low token confidence. This decoupling exposes the inherent unreliability of single-source metrics, necessitating a more robust calibration approach.

Leveraging this insight, we propose a post-hoc, interpretable monotone confidence fusion framework to calibrate MLLMs' confidence estimates. Specifically, we integrate implicit token confidence, explicit verbalized confidence, and cross-channel consistency into a unified reliability descriptor. These features are mapped to a precise probability of correctness via a monotone fusion head, incorporating the essential inductive bias that stronger signals imply higher correctness. Subsequently, we employ an order-preserving mean alignment to refine the estimate and correct systematic bias. This pipeline effectively minimizes calibration error while strictly maintaining the relative sample ranking, thereby preserving the model's utility for risk-sensitive selective prediction. We evaluate both token and verbalized confidence across six multi-modal benchmarks, assessing performance via calibration (ECE) and correctness ranking metrics (AUROC and AUPRC). Experimental results demonstrate that our post-hoc calibration method significantly reduces ECE compared to single-source baselines.

Overall, our contributions can be summarized: **(1)** We combine implicit token probabilities (instinct) and explicit verbalized self-assessments (reflection) to evaluate MLLMs' confidence. **(2)** We propose a post-hoc and interpretable monotone confidence fusion mechanism to robustly calibrate the frequently misaligned dual-channel confidence. **(3)** We provide an evaluation across multiple benchmarks and MLLMs, comprehensively analyzing calibration error, failure prediction, and risk–coverage.

2 Related Work

Multimodal Large Language Models. Multimodal large language models (MLLMs) integrate visual representations with language modeling to support visually grounded understanding and reasoning. Recent models extend these capabilities through improved vision–language alignment, multimodal training, and reasoning-oriented post-training (Bai et al., 2025a; Wang et al., 2025; Hong et al., 2025). Qwen3-VL supports long-context understanding and reasoning over images and videos (Bai et al., 2025a), while InternVL3.5 improves reasoning and inference efficiency through cascade reinforcement learning and adaptive visual processing (Wang et al., 2025). GLM-4.1V-Thinking

further explores reinforcement learning for multimodal reasoning and releases an open-source model (Hong et al., 2025). More recently, Qwen3.5-Omni expands multimodal modeling to integrated text, vision, and audio understanding and interaction (Qwen Team, 2026). These developments broaden model capabilities, but task accuracy alone does not establish whether confidence estimates reflect actual correctness. Our work complements capability-oriented development by evaluating and calibrating response confidence across open-source and proprietary MLLMs.

Confidence Evaluation. Confidence estimation (Geng et al., 2024) has been studied most extensively in text-only LLMs. Existing methods broadly fall into two paradigms: logit-based (Duan et al., 2024; Kumar et al., 2024) and consistency-based estimation (Geng et al., 2024; Prato et al., 2024; Xiong et al., 2023; Li et al., 2024; Dang et al., 2025). Logit-based approaches derive confidence from the model's generative distribution, *e.g.*, aggregating token-level likelihoods (Kumar et al., 2024) or entropy into a sequence-level uncertainty signal (Farquhar et al., 2024). In contrast, self-consistency based approaches estimate confidence by probing the stability of model behavior across multiple samples (Khanmohammadi et al., 2025), prompts (Wang et al., 2024), or model variants (Xiong et al., 2023). Compared to the rapidly growing LLM literature, confidence evaluation in MLLMs is less standardized (Tu et al., 2024). Recent work Xuan et al. (2025) analyzes the verbalized confidence in MLLMs, but limits systematic analysis across model families and downstream scenarios. Overall, existing studies either focus on a single confidence channel (logits, sampling consistency, or verbalized confidence) or evaluate a limited slice of the MLLM settings.

Confidence Calibration. Previous studies (Huang et al., 2024b; Shen et al., 2024; Geng et al., 2024) have observed that LLMs often express overconfidence, which motivates post-hoc calibration to align predicted confidence with empirical accuracy. Classic post-hoc calibration methods provide tools for improving the reliability of predicted probabilities in discriminative models (Naeini et al., 2015; Kull et al., 2017; 2019). Representative approaches include Platt scaling (Platt et al., 1999; Zadrozny & Elkan, 2002) and temperature scaling (Guo et al., 2017; Xie et al., 2024). Their effectiveness can be evaluated using metrics such as Brier score and ECE. Following studies of MLLM overconfidence (Dang et al., 2025; Xuan et al., 2025; Groot & Valdenegro-Toro, 2024), recent work explores general calibration (Chen et al., 2025), robust verbalized uncertainty (Borszukovszki et al., 2025), reasoning-step confidence (He et al., 2025), decoupled visual–reasoning confidence (Xiao et al., 2026), and object-level calibration via semantic perturbations (Zhao et al., 2025). Existing work does not jointly calibrate implicit token and explicit verbalized confidence, and several methods require prompt optimization, or reinforcement learning. In contrast, our lightweight post-hoc framework fuses these complementary signals while keeping the MLLM backbone frozen.

3 METHOD

This section details the three stages of our confidence calibration framework. Section 3.1 extracts dual-channel confidence and consistency; Section 3.2 performs monotone confidence fusion; Section 3.3 corrects global bias through order-preserving mean alignment.

Task formulation. We consider post-hoc confidence estimation for MLLMs on K-way multiple-choice VQA tasks. The dataset is denoted as $\mathcal{D} = \{(x_i, q_i, \mathcal{A}_i, y_i)\}_{i=1}^{N}$, where N is the number of samples in the dataset, x_i represents the image, q_i denotes the question together with its candidate answers, $\mathcal{A}_i = \{1, \ldots, K\}$ is their index set, and $y_i \in \mathcal{A}_i$ is the ground-truth answer index. To obtain reproducible option-level scores, we use a structured prompt that enforces a deterministic answer format. The MLLM returns a predicted option label (*e.g.*, "1"... "K" or "A"... "D") together with a verbalized confidence score for each candidate. Using these outputs, our goal is to estimate a calibrated probability that the MLLM's top-1 prediction is correct.

3.1 DUAL-CHANNEL CONFIDENCE EXTRACTION

As shown in Figure 2, we extract two complementary signals from the MLLM: implicit token confidence (instinct), reflecting aleatoric uncertainty in the probability distribution, and explicit verbalized assessment (reflection). These signals often disagree, with high verbalized confidence sometimes accompanied by low token confidence. We leverage this informative misalignment to enhance reliability estimation in multimodal settings.

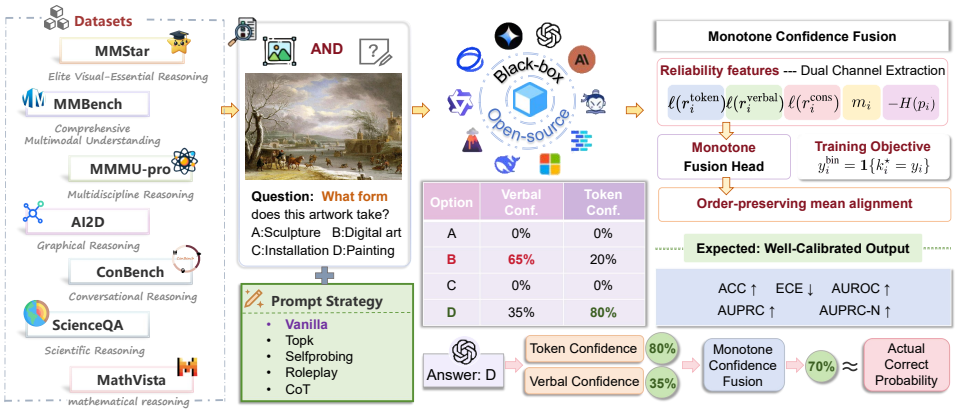


Figure 2: Overview of our method. We evaluate open-source and proprietary MLLMs across diverse benchmarks, employing varying prompt strategies to extract dual-channel uncertainty signals: *instinct* (implicit token-based probability confidence) and *reflection* (explicit verbal self-assessment confidence). These complementary cues are integrated via a monotone head followed by order-preserving mean alignment to correct distribution shifts, yielding well-calibrated reliability estimates.

Token confidence (implicit). Let z_{ik} denote the MLLM's log-probability of the option-label token corresponding to choice k for the i-th sample at the answer position, conditioned on (x_i, q_i) under the fixed answer-format instruction. We compute the option-token confidence vector $p_i = (p_{i1}, \ldots, p_{iK}) \in \Delta^{K-1}$, where each component p_{ik} is the normalized probability assigned to the k-th candidate among the K available options, with normalization restricted to these options:

$$p_{ik} = \frac{\exp(z_{ik})}{\sum_{k'=1}^{K} \exp(z_{ik'})}. \tag{1}$$

This token signal captures the model's local token-level support distribution over the discrete set of K candidates, rather than over the model's full vocabulary.

Verbalized confidence (explicit). In parallel, we prompt the MLLM to explicitly report its estimated probability for each option under a structured template that is easy to parse (Appendix B.2). We then parse the resulting text into a verbalized confidence vector consistent with the option set:

$$s_i = (s_{i1}, \ldots, s_{iK}) \in [0, 1]^K, \tag{2}$$

where s_{ik} is the model's stated probability (*e.g.*, 80%) for option k. Unlike $p_i \in \Delta^{K-1}$, s_i lies in the hypercube and may not sum to 1, allowing the model to express ambiguity (multiple plausible options) or ignorance (all options low).

Cross-channel confidence consistency. A trustworthy model should exhibit synchronization between instinct (token confidence) and reflection (verbalized confidence). This cross-channel coherence signifies an internal consensus across different cognitive inference stages. Specifically, for eac sample i and candidate k, we map their discrepancy into a similarity score using an RBF kernel:

$$\kappa_{ik} = \exp\left(-\frac{|p_{ik} - s_{ik}|^{\gamma}}{\tau}\right), \tag{3}$$

where $\kappa_{ik} \in (0, 1]$, $\gamma > 0$ controls the shape (we use $\gamma = 2$), and $\tau > 0$ is a bandwidth parameter. This metric assigns high agreement to both calibrated uncertainty (where both channels are low) and calibrated confidence (where both channels are high), while penalizing severe misalignment that is often indicative of hallucinated certainty. The non-linear curvature of the RBF kernel amplifies subtle discrepancies, providing better discriminative power than linear metrics.

For the i-th instance, the model first identifies the predicted option at inference time as $k_i^\star = \arg\max_{k \in \mathcal{A}_i} p_{ik}$ by maximizing the token probability over the candidate set $\mathcal{A}_i$. We define the token, verbalized, and consistency signals and the five-dimensional reliability descriptor:

$$(r_i^{\text{token}}, r_i^{\text{verbal}}, r_i^{\text{cons}}) = (p_{i,k_i^\star}, s_{i,k_i^\star}, \kappa_{i,k_i^\star}). \tag{4}$$

$$\phi_i = \big[\ell(r_i^{\text{token}}), \ell(r_i^{\text{verbal}}), \ell(r_i^{\text{cons}}), m_i, -H(p_i)\big] \in \mathbb{R}^5. \tag{5}$$

The three scalar terms respectively capture instinctive support, reflective self-assessment, and cross-channel agreement, while m_i and $-H(p_i)$ characterize local and global distributional sharpness.

Design rationale for MLLMs. Our fusion design follows four principles: (i) Monotone priors. Each reliability cue is constructed so that larger values imply higher correctness likelihood, motivating nonnegative fusion weights. (ii) Comparable geometry. Log-odds place heterogeneous confidence scales onto a shared linear domain, while margin and entropy capture local token sharpness. (iii) Shift-robust bias correction. A post-hoc bias adjustment matches mean confidence to an estimated correctness rate without changing ranking, enabling stable coverage control. (iv) Generalizes classical calibration. The linear head subsumes platt/temperature scaling as a special case and extends it to multi-signal fusion with more interpretable feature attribution.

3.2 Monotone Confidence Fusion

Reliability features. The descriptor ϕ_i in Eq.(5) combines both confidence channels using the numerically safeguarded log-odds transform to map probability-valued signals onto a shared scale:

$$\ell(z) = \log \frac{\text{clip}(z, \varepsilon, 1-\varepsilon)}{1 - \text{clip}(z, \varepsilon, 1-\varepsilon)}, \quad \varepsilon > 0. \tag{6}$$

This mapping mitigates saturation near probability boundaries, increasing the fusion head's sensitivity to confidence variations. We incorporate the top-2 margin $m_i = p_{i,(1)} - p_{i,(2)}$ and entropy, where $p_{i,(j)}$ denotes the j-th largest option probability. A larger margin indicates sharper decisions and reduced ambiguity. Shannon entropy $H(p_i) = -\sum_{k=1}^{K} p_{ik} \log p_{ik}$ characterizes global belief dispersion. We use negative entropy $-H(p_i)$ in ϕ_i so that greater concentration corresponds to higher reliability.

Monotone fusion head. Each component in Eq.(5) is tailored to capture a distinct dimension of uncertainty. Notably, these features are monotonically aligned by construction. The fusion stage in Figure 2 estimates the probability that the MLLM's top-1 prediction is correct using a coordinatewise monotone logistic head that combines five reliability features with nonnegative weights and a bias:

$$\hat{q}_i = \sigma\big(f_\theta(\phi_i)\big), \qquad f_\theta(\phi_i) = b + \sum_{j=1}^{d} w_j \phi_{ij}, \tag{7}$$

where $\sigma(t) = 1/(1+e^{-t})$ and $\phi_i \in \mathbb{R}^d$ denotes the reliability descriptor defined in Eq. (5), in our implementation, $d = 5$. We set $w_j = \text{softplus}(\tilde{w}_j) = \ln(1 + e^{\tilde{w}_j}) > 0$, ensuring that increasing any reliability feature cannot decrease predicted correctness.

Training objective. Let $\mathcal{D}_{\text{cal}}$ denote the calibration split with N_{cal} samples. Define the correctness indicator as $y_i^{\text{bin}} = \mathbf{1}\{k_i^\star = y_i\}$. We fit the model parameters $\theta = (b, \{\tilde{w}_j\})$ by minimizing the negative log-likelihood objective function:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\frac{1}{N_{\text{cal}}} \sum_{i=1}^{N_{\text{cal}}} \Big[y_i^{\text{bin}} \log(\hat{q}_i) + (1 - y_i^{\text{bin}}) \log(1 - \hat{q}_i) \Big]. \tag{8}$$

When d=1 and $\phi_i = \ell(r_i^{\text{token}})$, Eq.(7) reduces to Platt scaling, showing that our head generalizes standard post-hoc calibration. With the backbone MLLM frozen, we optimize θ on the calibration split. Mean alignment are performed on a disjoint validation split, while the test split is reserved exclusively for final evaluation to ensure out-of-sample generalization.

3.3 Order-preserving Mean Alignment

In the final stage of Figure 2, we apply the post-hoc bias shift $b \leftarrow b+\delta$, with δ chosen to satisfy Eq.(9). This step is intentionally order-preserving (Appendix B.6) and thus does not change any ranking-based evaluation (*e.g.*, risk–coverage curves), while correcting systematic base-rate misalignment.

$$\frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \sigma\big(f_\theta(\phi_i) + \delta\big) = \text{clip}(\text{ACC}, \varepsilon, 1-\varepsilon). \tag{9}$$

Here $\text{ACC}_{\text{val}} = N_{\text{val}}^{-1} \sum_{i=1}^{N_{\text{val}}} y_i^{\text{bin}}$ denotes the empirical accuracy on the validation split, and $\varepsilon = 10^{-6}$. We solve for δ using data-dependent bisection (Appendix B.7), prove the solution's existence and uniqueness in Appendix B.8, and separately evaluate the variant in Appendix C.2.

4 EXPERIMENTS

Table 1: Calibration and confidence-quality metrics of open-source and closed-source MLLMs on six benchmarks, comparing token confidence and verbalized confidence. Higher values indicate better performance for **AUROC**↑, **AUPRC**↑, and **AUPRC-N**↑, while lower values are better for **ECE**↓.

Conf.	Model	MMBench				MMStar				ConBench			
		ECE↓	AUROC↑	AUPRC↑	AUPRC-N↑	ECE↓	AUROC↑	AUPRC↑	AUPRC-N↑	ECE↓	AUROC↑	AUPRC↑	AUPRC-N↑
Token Confidence	**Open-source MLLMs**												
	MiniCPM-V-2.6 (Yao et al., 2024)	5.14	89.97	98.40	88.01	17.11	70.69	74.35	49.38	14.08	70.19	89.19	56.10
	InternVL2.5-4B (Chen et al., 2024b)	5.12	89.22	97.98	86.24	18.74	66.88	77.37	43.90	12.26	81.84	95.11	74.24
	InternVL3.5-8B (Wang et al., 2025)	5.63	88.86	98.66	86.92	20.12	70.43	84.69	56.25	15.98	83.33	94.08	73.11
	Qwen2.5-VL-7B (Bai et al., 2025c)	4.44	90.40	98.53	87.93	12.30	68.32	79.46	42.94	9.32	85.72	95.29	78.86
	Qwen3-VL-8B (Bai et al., 2025a)	5.43	83.56	96.40	62.91	22.61	66.91	72.08	30.21	10.32	76.23	91.78	58.91
	Phi-3.5-Vision (Abdin et al., 2024)	9.16	86.22	96.50	82.20	24.60	61.10	66.57	34.46	10.58	80.78	89.28	64.84
	Closed-source MLLMs												
	GPT-4o (Hurst et al., 2024)	7.61	91.33	98.81	89.51	28.37	74.01	77.46	49.50	12.33	90.52	98.22	88.99
	GPT-4o-mini (Hurst et al., 2024)	13.39	88.71	97.56	85.64	43.66	65.82	62.80	33.97	19.71	80.46	93.01	69.62
	GPT-5 (Singh et al., 2025)	4.99	79.32	96.01	67.74	32.02	70.66	72.91	39.30	12.28	73.67	85.42	44.02
Verbal Confidence	**Open-source MLLMs**												
	MiniCPM-V-2.6 (Yao et al., 2024)	7.45	74.71	93.62	52.27	26.87	54.68	53.22	7.69	10.98	62.96	82.27	28.00
	InternVL2.5-4B (Chen et al., 2024b)	10.99	56.67	86.50	7.94	26.97	52.06	62.94	8.11	13.87	61.26	87.27	33.02
	InternVL3.5-8B (Wang et al., 2025)	9.35	72.93	95.18	53.12	20.44	64.45	75.22	29.21	13.12	65.12	85.89	35.86
	Qwen2.5-VL-7B (Bai et al., 2025c)	37.64	45.92	87.54	-2.19	29.32	43.90	59.58	-12.27	38.65	35.91	74.90	-12.64
	Qwen3-VL-8B (Bai et al., 2025a)	24.19	71.55	95.47	53.25	41.14	55.95	68.04	20.11	28.24	65.90	89.02	45.11
	Phi-3.5-Vision (Abdin et al., 2024)	12.21	63.54	86.35	30.48	46.55	44.98	48.52	-0.95	20.33	63.14	72.65	10.32
	Closed-source MLLMs												
	GPT-4o (Hurst et al., 2024)	6.04	69.67	92.68	35.63	21.68	58.98	58.92	7.95	8.41	77.71	94.35	64.98
	GPT-4o-mini (Hurst et al., 2024)	8.08	77.61	92.66	56.78	19.06	61.85	54.55	19.32	14.10	64.82	83.41	27.87
	GPT-5 (Singh et al., 2025)	10.79	83.26	96.54	72.04	24.28	68.06	72.48	38.33	12.14	82.69	92.98	73.02

Conf.	Model	ScienceQA				AI2D				MMMU-Pro			
		ECE↓	AUROC↑	AUPRC↑	AUPRC-N↑	ECE↓	AUROC↑	AUPRC↑	AUPRC-N↑	ECE↓	AUROC↑	AUPRC↑	AUPRC-N↑
Token Confidence	**Open-source MLLMs**												
	MiniCPM-V-2.6 (Yao et al., 2024)	1.88	98.69	99.99	98.69	4.21	85.61	95.87	80.38	26.48	44.23	33.89	-5.05
	InternVL2.5-4B (Chen et al., 2024b)	0.88	92.81	99.95	92.57	8.13	86.12	95.98	81.31	39.49	31.64	43.00	-22.66
	InternVL3.5-8B (Wang et al., 2025)	4.53	98.67	99.88	98.68	9.23	86.03	96.66	82.52	51.36	50.29	28.58	-0.08
	Qwen2.5-VL-7B (Bai et al., 2025c)	4.76	99.25	99.92	99.20	5.07	87.93	97.01	83.13	35.52	56.65	32.14	11.26
	Qwen3-VL-8B (Bai et al., 2025a)	4.62	75.70	97.36	18.75	9.34	80.95	94.20	65.87	25.76	55.02	20.10	1.76
	Phi-3.5-Vision (Abdin et al., 2024)	2.69	92.43	99.46	91.75	12.24	84.54	94.97	79.00	21.01	55.51	53.64	9.16
	Closed-source MLLMs												
	GPT-4o (Hurst et al., 2024)	7.82	93.48	99.25	92.79	54.85	52.80	23.68	3.71	14.98	70.53	64.77	43.97
	GPT-4o-mini (Hurst et al., 2024)	7.93	92.55	99.26	91.88	49.43	49.98	29.02	2.08	18.56	67.41	59.88	38.80
	GPT-5 (Singh et al., 2025)	3.99	99.64	99.98	99.63	3.77	91.27	97.29	87.88	73.83	59.16	16.98	3.85
Verbal Confidence	**Open-source MLLMs**												
	MiniCPM-V-2.6 (Yao et al., 2024)	14.90	52.61	99.64	45.05	14.30	65.34	87.83	42.22	37.19	58.01	49.51	19.77
	InternVL2.5-4B (Chen et al., 2024b)	7.64	79.41	99.87	79.87	10.96	61.31	84.04	25.82	28.44	77.71	74.17	44.43
	InternVL3.5-8B (Wang et al., 2025)	9.09	87.73	98.62	84.79	19.23	66.12	89.85	46.85	41.91	70.40	46.50	25.04
	Qwen2.5-VL-7B (Bai et al., 2025c)	12.91	69.17	95.60	55.37	11.82	56.97	86.50	23.72	47.68	71.99	33.76	13.38
	Qwen3-VL-8B (Bai et al., 2025a)	10.71	62.55	98.10	41.50	28.74	62.40	87.94	29.02	40.81	73.74	36.02	21.34
	Phi-3.5-Vision (Abdin et al., 2024)	36.86	66.91	96.76	50.14	33.83	59.79	83.17	29.73	28.20	67.42	69.02	39.31
	Closed-source MLLMs												
	GPT-4o (Hurst et al., 2024)	4.25	91.71	99.07	91.02	23.74	84.67	57.72	46.65	30.91	70.78	57.48	32.36
	GPT-4o-mini (Hurst et al., 2024)	16.10	90.03	98.98	88.75	14.28	76.73	52.17	34.02	19.19	65.43	49.78	23.40
	GPT-5 (Singh et al., 2025)	18.47	82.32	97.48	53.47	14.04	80.90	93.88	72.60	46.02	72.56	34.32	23.93

Datasets and Models. We evaluate six open-source and three closed-source MLLMs on six multiple-choice VQA benchmarks. These benchmarks span diverse domains and difficulty levels, covering general-purpose visual understanding (MMBench (Liu et al., 2024), MMStar (Chen et al., 2024a), and ConBench (Zhang et al., 2024)), scientific reasoning (ScienceQA (Lu et al., 2022)), diagrammatic reasoning (AI2D (Kembhavi et al., 2016)), and multi-discipline professional reasoning (MMMU-Pro (Yue et al., 2024)). For each question, we assess response reliability using token-level probabilities assigned to the predicted option and verbalized confidence. Detailed benchmark descriptions, the complete model list, and confidence settings are provided in Appendix A.

Evaluation Metrics. We evaluate confidence along two complementary axes: calibration and failure prediction, while ACC (top-1 accuracy) is reported in Appendix Table 4. For calibration, ECE (Expected Calibration Error) (Guo et al., 2017) measures the sample-weighted absolute gap between mean predicted confidence and empirical accuracy across confidence bins. This definition applies to token, verbalized, and fused confidence, with lower ECE indicating better calibration. For failure prediction, AUROC measures how well confidence separates correct from incorrect predictions across decision thresholds. We also report AUPRC-N, which normalizes the area under the precision–recall curve relative to task-specific random baselines to facilitate comparisons across benchmarks with different class priors. Higher ACC, AUROC, and AUPRC-N indicate better predictive accuracy, discrimination, and confidence ranking, respectively. All metrics are reported on the $\times 10^2$ scale.

Results on Token Confidence. Table 1 shows that token confidence varies across models and benchmarks. On ScienceQA, most evaluated models achieve ECE below 5.0, with InternVL2.5-4B reaching 0.88 and AUROC 92.81. However, the same model obtains ECE 39.49, AUROC 31.64, and AUPRC-N -22.66 on MMMU-Pro, indicating poor calibration and ranking below the random baseline. The newer models struggle on this benchmark: Qwen3-VL-8B and GPT-5 yield ECE/AUROC of $25.76/55.02$ and $73.83/59.16$, respectively. Calibration can also deteriorate on diagram understanding, where GPT-4o reaches ECE 54.85 on AI2D. Thus, token confidence is neither uniformly reliable across benchmarks nor consistently superior to verbalized confidence.

Table 2: Performance comparison between raw confidence and calibrated confidence. Subscripts indicate Δ relative gains, calculated against the stronger baseline: for ECE $\downarrow$, Δ = Calibrated $-$ min(Token, Verbalized), Δ = Calibrated $-$ max(Token, Verbalized). All values are scaled by 10^2.

	MMBench				MMStar				ConBench			
Model	**ECE↓**	**AUROC↑**	**AUPRC↑**	**AUPRC-N↑**	**ECE↓**	**AUROC↑**	**AUPRC↑**	**AUPRC-N↑**	**ECE↓**	**AUROC↑**	**AUPRC↑**	**AUPRC-N↑**
Open-source MLLMs												
MiniCPM-V-2.6 (Yao et al., 2024)	$3.58_{(-1.56)}$	$90.34_{(+0.37)}$	$98.46_{(+0.06)}$	$88.45_{(+0.44)}$	$6.82_{(-10.29)}$	$72.44_{(+1.75)}$	$73.88_{(-0.47)}$	$48.46_{(-0.92)}$	$8.98_{(-2.00)}$	$74.60_{(+4.41)}$	$90.38_{(+1.19)}$	$60.92_{(+4.82)}$
InternVL2.5-4B (Chen et al., 2024b)	$3.59_{(-1.53)}$	$88.99_{(-0.23)}$	$97.92_{(-0.06)}$	$85.82_{(-0.42)}$	$10.66_{(-8.08)}$	$64.67_{(-2.21)}$	$76.08_{(-1.29)}$	$40.68_{(-3.22)}$	$7.02_{(-5.24)}$	$82.37_{(+0.53)}$	$95.19_{(+0.08)}$	$74.69_{(+0.45)}$
InternVL3.5-8B (Wang et al., 2025)	$3.35_{(-2.28)}$	$89.12_{(+0.26)}$	$98.68_{(+0.02)}$	$87.19_{(+0.27)}$	$7.05_{(-13.07)}$	$72.91_{(+2.48)}$	$85.70_{(+1.01)}$	$59.14_{(+2.89)}$	$5.94_{(-7.18)}$	$82.85_{(-0.48)}$	$93.41_{(-0.67)}$	$70.05_{(-3.06)}$
Qwen2.5-VL-7B (Bai et al., 2025c)	$2.24_{(-2.20)}$	$90.42_{(+0.02)}$	$98.45_{(-0.08)}$	$87.28_{(-0.65)}$	$8.68_{(-3.62)}$	$67.36_{(-0.96)}$	$77.30_{(-2.16)}$	$36.93_{(-6.01)}$	$5.91_{(-3.41)}$	$84.40_{(-1.32)}$	$93.31_{(-1.98)}$	$69.99_{(-8.87)}$
Qwen3-VL-8B (Bai et al., 2025a)	$3.00_{(-2.43)}$	$88.77_{(+5.21)}$	$98.61_{(+2.21)}$	$85.69_{(+22.78)}$	$8.77_{(-13.84)}$	$68.66_{(+1.75)}$	$76.76_{(+4.68)}$	$41.90_{(+11.69)}$	$5.74_{(-4.58)}$	$81.65_{(+5.42)}$	$94.59_{(+2.81)}$	$72.94_{(+14.04)}$
Phi-3.5-Vision (Abdin et al., 2024)	$4.27_{(-4.89)}$	$86.43_{(+0.21)}$	$96.54_{(+0.04)}$	$82.37_{(+0.17)}$	$13.99_{(-10.61)}$	$56.64_{(-4.46)}$	$63.88_{(-2.69)}$	$29.18_{(-5.28)}$	$7.17_{(-3.41)}$	$80.38_{(-0.40)}$	$89.32_{(+0.04)}$	$65.00_{(+0.16)}$
Closed-source MLLMs												
GPT-4o (Hurst et al., 2024)	$1.92_{(-4.12)}$	$90.84_{(-0.49)}$	$98.75_{(-0.06)}$	$88.96_{(-0.55)}$	$7.74_{(-13.94)}$	$72.89_{(-1.12)}$	$76.48_{(-0.98)}$	$47.30_{(-2.20)}$	$7.56_{(-0.85)}$	$90.66_{(+0.14)}$	$98.24_{(+0.02)}$	$89.08_{(+0.09)}$
GPT-4o-mini (Hurst et al., 2024)	$5.16_{(-2.92)}$	$86.08_{(-2.63)}$	$95.99_{(-1.57)}$	$76.40_{(-9.24)}$	$3.18_{(-15.88)}$	$65.11_{(-0.71)}$	$60.47_{(-2.33)}$	$29.82_{(-4.15)}$	$7.07_{(-7.03)}$	$77.65_{(-2.81)}$	$91.79_{(-1.22)}$	$64.31_{(-5.31)}$
GPT-5 (Singh et al., 2025)	$4.29_{(-0.70)}$	$90.74_{(+7.48)}$	$98.01_{(+1.47)}$	$83.91_{(+11.88)}$	$8.25_{(-16.03)}$	$72.01_{(+1.35)}$	$80.33_{(+7.43)}$	$55.94_{(+16.64)}$	$7.30_{(-4.84)}$	$84.86_{(+2.17)}$	$93.32_{(+0.35)}$	$74.36_{(+1.33)}$

	ScienceQA				AI2D				MMMU-Pro			
Model	**ECE↓**	**AUROC↑**	**AUPRC↑**	**AUPRC-N↑**	**ECE↓**	**AUROC↑**	**AUPRC↑**	**AUPRC-N↑**	**ECE↓**	**AUROC↑**	**AUPRC↑**	**AUPRC-N↑**
Open-source MLLMs												
MiniCPM-V-2.6 (Yao et al., 2024)	$1.58_{(-0.30)}$	$98.69_{(0.00)}$	$99.99_{(0.00)}$	$98.69_{(0.00)}$	$3.58_{(-0.63)}$	$86.38_{(+0.77)}$	$96.07_{(+0.20)}$	$81.33_{(+0.95)}$	$8.48_{(-18.00)}$	$46.46_{(-11.55)}$	$38.66_{(-10.85)}$	$2.53_{(-17.24)}$
InternVL2.5-4B (Chen et al., 2024b)	$1.13_{(-0.25)}$	$92.81_{(0.00)}$	$99.95_{(0.00)}$	$92.57_{(0.00)}$	$4.52_{(-3.61)}$	$86.19_{(+0.07)}$	$95.98_{(0.00)}$	$81.33_{(+0.02)}$	$21.87_{(-6.57)}$	$75.04_{(-2.67)}$	$72.36_{(-1.81)}$	$40.51_{(-3.92)}$
InternVL3.5-8B (Wang et al., 2025)	$3.19_{(-1.34)}$	$99.44_{(+0.77)}$	$99.95_{(+0.07)}$	$99.41_{(+0.73)}$	$3.80_{(-5.43)}$	$85.97_{(-0.06)}$	$96.58_{(-0.08)}$	$82.09_{(-0.43)}$	$2.41_{(-39.50)}$	$60.48_{(-9.92)}$	$40.73_{(-5.77)}$	$16.95_{(-8.09)}$
Qwen2.5-VL-7B (Bai et al., 2025c)	$3.57_{(-1.19)}$	$99.12_{(-0.13)}$	$99.90_{(-0.02)}$	$99.03_{(-0.17)}$	$3.54_{(-1.53)}$	$88.10_{(+0.17)}$	$97.24_{(+0.23)}$	$84.41_{(+1.28)}$	$10.56_{(-24.96)}$	$72.12_{(+0.13)}$	$38.64_{(+4.88)}$	$19.76_{(+6.38)}$
Qwen3-VL-8B (Bai et al., 2025a)	$3.06_{(-1.56)}$	$91.68_{(+15.98)}$	$99.72_{(+1.62)}$	$91.26_{(+49.77)}$	$3.65_{(-5.69)}$	$82.78_{(+1.83)}$	$95.63_{(+1.43)}$	$74.30_{(+8.43)}$	$9.61_{(-16.15)}$	$66.36_{(-7.38)}$	$28.06_{(-7.97)}$	$11.54_{(-9.80)}$
Phi-3.5-Vision (Abdin et al., 2024)	$4.48_{(+1.79)}$	$92.50_{(+0.07)}$	$99.46_{(0.00)}$	$91.73_{(-0.02)}$	$4.71_{(-7.53)}$	$84.07_{(-0.47)}$	$94.67_{(-0.30)}$	$77.74_{(-1.26)}$	$10.47_{(-10.54)}$	$63.28_{(-4.14)}$	$62.66_{(-6.36)}$	$26.85_{(-12.46)}$
Closed-source MLLMs												
GPT-4o (Hurst et al., 2024)	$6.35_{(+2.10)}$	$93.89_{(+0.41)}$	$99.30_{(+0.05)}$	$93.26_{(+0.47)}$	$11.34_{(-12.40)}$	$82.66_{(-2.01)}$	$48.21_{(-9.51)}$	$34.66_{(-11.99)}$	$7.88_{(-7.10)}$	$75.20_{(+4.42)}$	$70.43_{(+5.66)}$	$52.96_{(+8.99)}$
GPT-4o-mini (Hurst et al., 2024)	$9.57_{(+1.64)}$	$93.01_{(+0.46)}$	$99.31_{(+0.05)}$	$92.40_{(+0.52)}$	$9.04_{(-5.24)}$	$76.11_{(-0.62)}$	$49.56_{(-2.61)}$	$30.42_{(-3.60)}$	$3.74_{(-14.82)}$	$69.36_{(+1.95)}$	$60.17_{(+0.29)}$	$39.25_{(+0.45)}$
GPT-5 (Singh et al., 2025)	$4.60_{(+0.61)}$	$96.34_{(-3.30)}$	$99.79_{(-0.19)}$	$96.15_{(-3.48)}$	$4.35_{(+0.58)}$	$93.97_{(+2.70)}$	$97.99_{(+0.69)}$	$90.99_{(+3.11)}$	$7.19_{(-38.83)}$	$66.15_{(-6.41)}$	$29.61_{(-4.71)}$	$18.47_{(-5.46)}$

Results on Verbalized Confidence. Table 1 also shows substantial variation in verbalized confidence across models and benchmarks. GPT-4o-mini maintains ECE below 20.0 across all six datasets, whereas Qwen2.5-VL-7B reaches 47.68 on MMMU-Pro. On ConBench, Qwen2.5-VL-7B obtains AUROC 35.91 and AUPRC-N -12.64, indicating ranking below the random baseline. Nevertheless, poor calibration does not always imply weak discrimination: on MMMU-Pro, Qwen3-VL-8B and GPT-5 retain AUROC of 73.74 and 72.56, despite ECE of 40.81 and 46.02, respectively. Their verbalized scores remain useful for ranking despite miscalibration. On MMMU-Pro, Qwen2.5-VL-7B also yields higher AUROC with verbalized than token confidence (71.99 vs. 56.65), despite worse calibration. These differences across models, benchmarks, and metrics motivate the dual-channel fusion evaluated below, which combines complementary information from both confidence signals.

Results on Calibrated Confidence. Results in Table 2 indicate that our fused confidence primarily functions as a robust post-hoc recalibration tool, yielding substantial ECE reductions across most benchmarks. This gain is particularly significant in high-complexity reasoning tasks like MMMU-Pro and AI2D. For instance, InternVL3.5-8B's ECE on MMMU-Pro decreases by 39.50 (dropping to 2.41). This underscores that our proposed multi-channel fusion framework effectively mitigates single-channel miscalibration by leveraging complementary information. Conversely, foundational tasks like ScienceQA exhibit a marginal cost, where noise from the secondary channel causes slight ECE inflation (*e.g.*,+2.10 for GPT-4o) for models with near-perfect single-channel baselines. Furthermore, improvements in discriminative power (AUROC and AUPRC-N) are inconsistent and highly model-sensitive. For Qwen3-VL-8B, calibration lowers ECE relative to the stronger single channel on all six benchmarks, reaching 3.00 on MMBench and 9.61 on MMMU-Pro. Additional comparisons with generic calibrators under matched label budgets are reported in Appendix C.4.

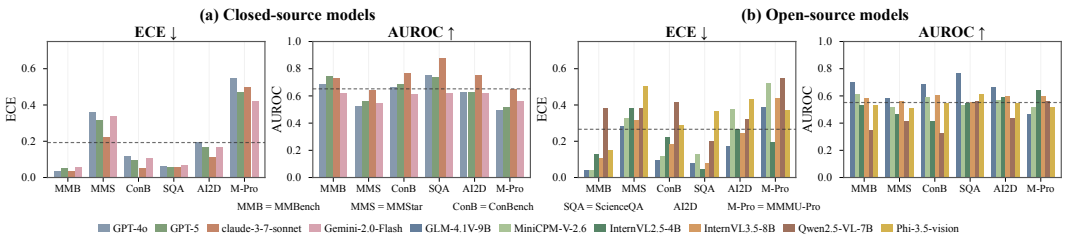
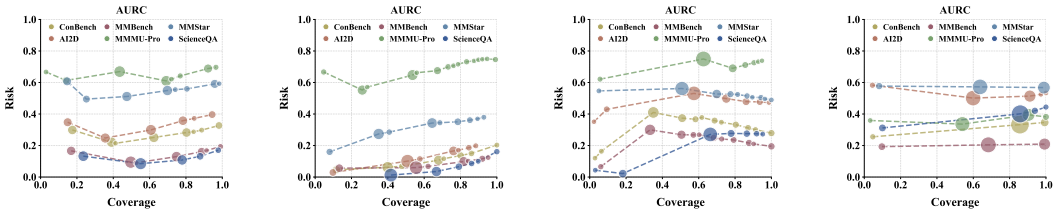


Figure 3: Comparison of ECE and AUROC using verbalized confidence across six datasets. Left: closed-source models. Right: open-source models. Metrics use the $[0, 1]$ scale.

Results on Calibration and Correctness Ranking. To compare probability calibration and correctness ranking, we visualize verbalized-confidence ECE and AUROC on six benchmarks in Figure 3. The left and right groups show closed-source and open-source models, with lower ECE and higher

AUROC indicating better performance. On MMBench, all displayed closed-source models achieve ECE below 0.10, whereas several open-source models exceed this level. Claude-3.7-Sonnet also combines low ECE (approximately 0.06) with high AUROC (approximately 0.88) on ScienceQA. However, these advantages are benchmark-dependent: on MMMU-Pro, closed-source ECE rises to approximately 0.42–0.55, exceeding the open-source results. The open-source group is also heterogeneous: Qwen2.5-VL-7B shows low ECE and relatively high AUROC on ScienceQA but substantially worse values on MMBench. These comparisons show that model access type alone does not determine reliability, motivating evaluation of both calibration and ranking across benchmarks.



(a) GPT-4o-mini (b) Claude-3.7-sonnet (c) Qwen2.5-VL-7B (d) Phi-3.5-Vision

Figure 4: Risk–coverage curves for four MLLMs across six benchmarks. Lower risk indicates better selective reliability. ScienceQA and MMBench generally show lower risk than MMMU-Pro.

Results of risk–coverage analysis. To assess selective reliability, we plot risk–coverage (RC) curves for GPT-4o-mini (Hurst et al., 2024), Claude-3.7-Sonnet, Qwen2.5-VL-7B, and Phi-3.5-Vision in Figure 4. For each benchmark in the single-image setting, we sort the n test instances by the verbal confidence c_i in descending order. At coverage $\alpha = k/n$, we accept the top-k instances and compute accuracy $\hat{a}(\alpha) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}[\hat{y}_i = y_i]$, with risk defined as $r(\alpha) = 1 - \hat{a}(\alpha)$. We summarize the trade-off by $\mathrm{AURC} = \int_0^1 r(\alpha)\, d\alpha$, approximated via the trapezoidal rule, where lower values indicate stronger selective reliability. In the GPT-4o-mini panel, ScienceQA and MMBench have lower risk than the other benchmarks over most of the displayed coverage range. For the same model, MMMU-Pro risk remains approximately 0.60–0.70 and varies non-monotonically with coverage, indicating that retaining fewer high-confidence answers does not consistently reduce error on this benchmark. Overall, GPT-4o-mini's verbalized confidence supports lower-risk selection on ScienceQA and MMBench, while MMStar and MMMU-Pro retain substantially higher risk.

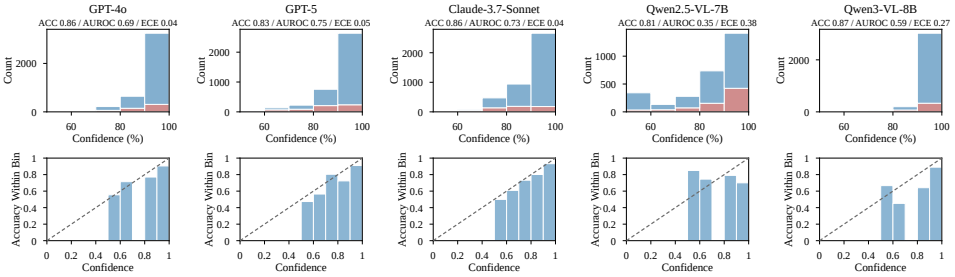


Figure 5: Verbalized-confidence calibration on MMBench. Histograms compare correct and incorrect predictions, while the dashed diagonal in each reliability diagram denotes perfect calibration.

Calibration and Reliability Analysis. To examine how verbalized confidence aligns with correctness, we analyze confidence histograms (top) and reliability diagrams (bottom) on MMBench in Figure 5. We observe a significant disparity between closed-source models (GPT-4o, GPT-5, Claude-3.7-Sonnet) and the open-source model (Qwen2.5-VL-7B). The closed-source models demonstrate higher reliability: correct predictions (blue) are concentrated in high-confidence regions (> 0.8), while incorrect predictions (red) are properly assigned lower confidence. This alignment results in reliability curves that closely follow the ideal diagonal, achieving minimal ECE (≈ 0.04–0.05) and comparatively higher AUROC. In contrast, the open-source model exhibits severe overconfidence alongside a notable lack of predictive orderliness. The ECE of Qwen2.5-VL-7B (0.38) is nearly an order of magnitude higher than closed-source counterparts. Its histogram reveals that a large

portion of wrong answers are incorrectly assigned the highest confidence. Results suggest that while closed-source models are better-calibrated, current open-source models still struggle to accurately estimate predictive uncertainty. Overall, these plots highlight the need to assess both confidence separation and probability calibration before using verbalized scores for selective prediction.

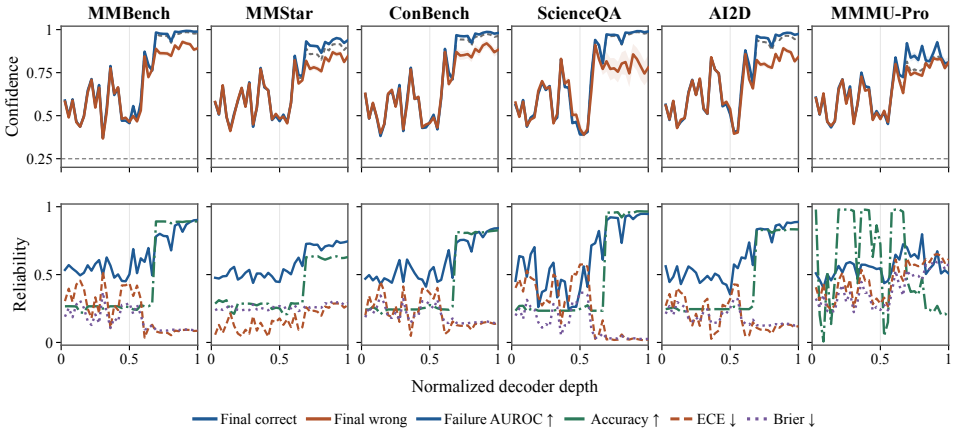


Figure 6: Layerwise confidence in Qwen3-VL-8B. Top: mean confidence for correct and incorrect final predictions. Bottom: failure AUROC, layer accuracy, and ECE/Brier evaluated against final correctness. The horizontal axis denotes normalized decoder depth, with one column per benchmark.

Results on Layerwise Confidence Formation. To examine when token confidence reflects correctness, we analyze Qwen3VL-8B across decoder depth in Figure 6. In the top row, confidence fluctuates through early and middle layers before rising sharply after roughly half the decoder blocks. Examples with correct final predictions generally attain higher late-layer confidence, but confidence also increases for incorrect answers. The bottom row shows that late-layer AUROC and accuracy improve on MMBench, ConBench, ScienceQA, and AI2D, accompanied by lower ECE and Brier scores. MMStar exhibits weaker improvements, while MMMU-Pro retains unstable layer accuracy and comparatively high calibration errors. These results indicate that confidence becomes more concentrated late in decoding, but its alignment with correctness remains benchmark-dependent. Thus, higher confidence does not establish reliability, motivating calibration beyond raw token confidence.

Results on Standard Calibration Baselines. To assess calibration and correctness ranking against established alternatives, we compare our method with standard calibration and uncertainty baselines on MMBench. As shown in Table 3, Platt scaling achieves the lowest ECE (2.68), while our method attains a comparable ECE of 2.83 and ties temperature scaling for the highest AUROC (89.56). Temperature scaling obtains the best AUPRC-N (86.91), with our method remaining close at 86.57. Our method also achieves the lowest NLL (0.261) and a Brier score of 0.083, comparable to the other post-hoc calibrators. In contrast, self-consistency and option-level entropy require five stochastic answer samples per question, whereas the post-hoc calibration methods require no repeated answer sampling.

Table 3: Comparison with standard calibration and uncertainty baselines on MMBench. ECE, AUROC, and AUPRC-N are reported as $\times 10^2$, while NLL and Brier scores are reported on their original scales.

Method	ECE↓	AUROC↑	AUPRC-N↑	NLL↓	Brier↓
Platt Scaling	2.68	89.44	86.51	0.262	0.083
Temperature Scaling	2.85	89.56	86.91	0.263	0.083
Self-Consistency	10.17	71.22	39.68	1.112	0.115
Option-level Entropy	8.71	71.23	39.80	1.101	0.110
Ours	2.83	89.56	86.57	0.261	0.083

5 Conclusion

This work unifies token and verbalized confidence through monotone fusion and mean alignment. It improves calibration across a broader range of benchmarks, with model-dependent ranking gains. Building on our open-ended VQA experiments, future work will investigate calibration for long-form multimodal responses. We will explore consistency-aware training and calibration under distribution shifts across more diverse evaluation settings. These directions could support reliable abstention, human oversight, and clearer uncertainty communication for more trustworthy MLLM deployment in practice, helping users decide when additional verification is needed.

Reproducibility Statement

We have taken steps to support the reproducibility of this work by describing the proposed method, experimental protocol, evaluated models and datasets, and evaluation metrics in the main paper. The appendix provides further implementation details, including data partitioning, prompting and parsing procedures, optimization settings, hyperparameters, and additional analyses. All experiments are evaluated on held-out data, and the datasets and models are identified through the corresponding citations or public interfaces. Although outputs from proprietary models may vary as their APIs evolve, the overall evaluation and calibration procedure is specified to facilitate replication under comparable settings.

AI Use Statement

Generative AI tools were used during manuscript preparation to assist with English-language editing, organization, LaTeX formatting, and drafting or revising descriptions based on author-provided experimental results and instructions. They were not used as independent sources of scientific evidence or to generate the reported experimental measurements. The authors manually reviewed the AI-assisted text, calculations, and formatting against the experimental records and cited sources, made all final methodological and interpretive decisions, and take full responsibility for the content of this work.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025c.

Mirko Borszukovszki, Ivo Pascal De Jong, and Matias Valdenegro-Toro. Know what you do not know: Verbalized uncertainty estimation robustness on corrupted images in vision-language models. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pp. 247–265, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zijun Chen, Wenbo Hu, Guande He, Zhijie Deng, Zheng Zhang, and Richang Hong. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3095–3109, 2025.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, et al. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*, 2024.

Yunkai Dang, Mengxi Gao, Yibo Yan, Xin Zou, Yanggan Gu, Jungang Li, Jingyu Wang, Peijie Jiang, Aiwei Liu, Jia Liu, et al. Exploring response uncertainty in mllms: An empirical evaluation under misleading scenarios. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18143–18184, 2025.

Weihua Du, Yiming Yang, and Sean Welleck. Optimizing temperature for language models with multi-sample inference. *arXiv preprint arXiv:2502.05234*, 2025.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5050–5063, 2024.

Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

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

Stephen M Fleming. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75(1):241–268, 2024.

Stephen M Fleming and Nathaniel D Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1):91, 2017.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. *arXiv preprint arXiv:2405.02917*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R Fung. Mmboundary: 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, 2025.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. A survey of uncertainty estimation in llms: Theory meets practice. *arXiv preprint arXiv:2410.15326*, 2024a.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. Uncertainty in language models: Assessment through rank-calibration. *arXiv preprint arXiv:2404.03163*, 2024b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese Smiley, Kundan S Thind, and Mohammad M Ghassemi. Calibrating llm confidence by probing perturbed representation stability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10459–10525, 2025.

Asher Koriat. How do we know that we know? the accessibility model of the feeling of knowing. *Psychological review*, 100(4):609, 1993.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pp. 623–631. PMLR, 2017.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. *arXiv preprint arXiv:2405.16282*, 2024.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. *arXiv preprint arXiv:2403.09972*, 2024.

Yiwei Li, Ji Zhang, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Jiayi Shi, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, et al. Revisiting self-consistency from dynamic distributional alignment perspective on answer aggregation. *arXiv preprint arXiv:2502.19830*, 2025.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512, 2020.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

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. *arXiv preprint arXiv:2310.02255*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, 2015.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Alexandre Pouget, Jan Drugowitsch, and Adam Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, 19(3):366–374, 2016.

Gabriele Prato, Jerry Huang, Prasanna Parthasarathi, Shagun Sodhani, and Sarath Chandar. Do large language models know how much they know? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6054–6070, 2024.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. Thermometer: Towards universal calibration for large language models. *arXiv preprint arXiv:2403.08819*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*, 2025.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. *arXiv preprint arXiv:2402.07417*, 2024.

Pei Wang, Yejie Wang, Muxi Diao, Keqing He, Guanting Dong, and Weiran Xu. Multi-perspective consistency enhances confidence estimation in large language models. *arXiv preprint arXiv:2402.11279*, 2024.

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

Wenyi Xiao, XU Xinchi, 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, 2026.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. *arXiv preprint arXiv:2409.19817*, 2024.

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. *arXiv preprint arXiv:2306.13063*, 2023.

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. *arXiv preprint arXiv:2505.20236*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Nick Yeung and Christopher Summerfield. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367 (1594):1310–1321, 2012.

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. *arXiv preprint arXiv:2409.02813*, 2024.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Yuan Zhang, Fei Xiao, Tao Huang, Chun-Kai Fan, Hongyuan Dong, Jiawen Li, Jiacong Wang, Kuan Cheng, Shanghang Zhang, and Haoyuan Guo. Unveiling the tapestry of consistency in large vision-language models. *Advances in Neural Information Processing Systems*, 37:118632–118653, 2024.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. Knowing what llms do not know: A simple yet effective self-detection method. In *Proceedings of the 2024 conference of the north American chapter of the Association for Computational Linguistics: Human language technologies (Volume 1: Long Papers)*, pp. 7051–7063, 2024.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Ruibo Hou, Jiaming Guo, Zihao Zhang, Yifan Hao, and Yunji Chen. Object-level verbalized confidence calibration in vision-language models via semantic perturbation. *arXiv preprint arXiv:2504.14848*, 2025.

APPENDIX CONTENTS

A Detailed Experiment Settings **16**

A.1 Benchmarks 16

A.2 Models 17

A.3 Evaluation Metrics 17

B Additional Implementation and Theoretical Details **18**

B.1 End-to-end procedure and dataset splits 18

B.2 Prompting and parsing verbalized confidence 18

B.3 Extracting token-option confidence 19

B.4 Reliability features: construction, bounds, and standardization 19

B.5 Monotone fusion head: optimization, interpretability, and special cases 20

B.6 Order-preserving mean alignment: theory and numerics 20

B.7 Solving mean alignment by bisection 20

B.8 Theoretical Guarantees 21

B.9 Complexity and practical considerations 22

B.10 Hyperparameters and recommended ranges 22

C Additional Experiments **22**

C.1 Confidence Diagnostics and Selective Prediction 22

C.2 Mean-alignment Robustness 26

C.3 Calibration and Prompting Analyses 28

C.4 Transferability and Data Efficiency 31

C.5 Broader Models, Tasks, and Perturbations 32

C.6 Component Analysis and Robustness Checks 35

C.7 Detailed Prompting Results and Templates 36

C.8 Confidence Case Studies 38

A Detailed Experiment Settings

A.1 Benchmarks

To comprehensively assess multimodal models in terms of general visual understanding, reasoning over diverse visual evidence (natural images, diagrams, and mixed text-image content), and knowledge-intensive problem solving, we conduct experiments on six established benchmarks, including MMBench (Liu et al., 2024), MMStar (Chen et al., 2024a), ConBench (Zhang et al., 2024), ScienceQA (Lu et al., 2022), AI2D (Kembhavi et al., 2016), and MMMU-Pro (Yue et al., 2024). These six datasets constitute our primary experimental suite. We focus on the multiple-choice VQA setting and extract per-question confidence scores for each instance, consisting of a question and its candidate options, to assess the reliability of model responses. Furthermore, we provide supplementary results on MathVista (Lu et al., 2023) in the appendix. These benchmarks cover (i) broad, real-world visual perception and instruction following, (ii) challenging multimodal reasoning with strong distractors, (iii) compositional/consistency-sensitive evaluation settings, (iv) science-oriented multimodal QA requiring domain knowledge, (v) diagram understanding with structured visual relations, and (vi) expert-level multi-discipline multimodal understanding.

MMBench (Liu et al., 2024) is a large-scale multi-choice benchmark designed to evaluate general-purpose visual understanding and reasoning across diverse image domains. It probes a model's comprehensive ability to seamlessly integrate complex visual cues with linguistic instructions through a fine-grained taxonomy of 20 ability dimensions. By incorporating rigorous evaluation protocols like CircularEval, it serves as a robust testbed for examining not only prediction accuracy but also the alignment between model confidence and actual performance across diverse visual tasks.

MMStar (Chen et al., 2024a) is a benchmark designed for high-difficulty multimodal reasoning, emphasizing the critical visual indispensability across all test questions. It targets settings with harder distractors and increased compositional complexity, effectively stressing fine-grained visual grounding and robust reasoning beyond superficial cues.

ConBench (Zhang et al., 2024) focuses on compositional and/or consistency-related aspects of multimodal reasoning, emphasizing whether a model can reliably handle controlled variations and multi-step reasoning patterns. It is particularly useful for stress-testing robustness under compositional changes and fine-grained content-level perturbations, while remaining critical for analyzing confidence behavior under harder conditions.

ScienceQA (Lu et al., 2022) evaluates science-centric multimodal question answering where solutions often require combining rich visual evidence (e.g., figures, diagrams, or illustrative images) with domain-specific scientific knowledge and reasoning. It therefore serves as a proxy for evaluating knowledge-intensive multimodal problem solving.

AI2D (Kembhavi et al., 2016) is a diagram understanding benchmark that assesses a model's inherent capability to interpret structured visual elements such as labels, arrows, layout, and relational cues. It emphasizes logical reasoning over diagram semantics rather than purely natural-image perception. Crucially, it requires the model to parse the complex topology between textual annotations and visual icons to resolve challenging multi-step scientific queries.

MMMU-Pro (Yue et al., 2024) measures expert-level multimodal understanding spanning multiple disciplines, specifically emphasizing visual indispensability. It requires deeper reasoning and broader knowledge coverage. Its difficulty makes it a strong indicator of advanced multimodal competence under challenging evaluation settings.

MathVista (Lu et al., 2023) is a comprehensive benchmark designed to evaluate mathematical reasoning within diverse visual contexts. It consolidates thirty-one distinct datasets, including three newly proposed challenging subsets—IQTest, FunctionQA, and PaperQA—to push the boundaries of visual logic. The benchmark organizes its problems into five task types and seven mathematical reasoning skills, while covering a wide range of visual contexts such as diagrams, charts, plots, tables, natural images, and scientific figures. Unlike conventional perception tasks, MathVista requires fine-grained visual grounding and complex, multi-step logical deduction.

A.2 MODELS

To obtain a representative view of confidence behavior across MLLMs, our primary evaluation includes six open-source and three closed-source models. The open-source models are MiniCPM-V-2.6 (Yao et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025c), Qwen3-VL-8B, InternVL2.5-4B (Chen et al., 2024b), InternVL3.5-8B (Wang et al., 2025), and Phi-3.5-Vision. The closed-source models are GPT-4o, GPT-4o-mini (Hurst et al., 2024), and GPT-5. For confidence estimation, we use token-level probabilities assigned to the predicted option tokens and verbalized confidence elicited from model responses. These signals do not require access to model gradients; token probabilities can be obtained through model interfaces or APIs that expose them, allowing the approach to be applied to both open-source and closed-source models with the requisite outputs.

Overall answer accuracy. Table 4 reports the accuracy of the verbalized responses, providing the task-performance context for the confidence-quality results in the main text. Performance varies substantially across both models and benchmarks: Qwen3-VL-8B achieves the highest accuracy on MMBench and ConBench, InternVL3.5-8B leads on MMStar and ScienceQA, and GPT-4o performs best on AI2D. No single model dominates all five benchmarks, and strong answer accuracy does not by itself imply well-calibrated or discriminative confidence. The table should therefore be read as a reference for underlying task difficulty, while ECE, AUROC, and AUPRC-N separately assess the reliability of the associated confidence estimates.

Table 4: Answer accuracy (%) of verbalized responses across five benchmarks.

Model	MMBench	MMStar	ConBench	ScienceQA	AI2D
Open-source MLLMs					
MiniCPM-V-2.6 (Yao et al., 2024)	82.79	50.84	71.13	83.64	50.44
InternVL2.5-4B (Chen et al., 2024b)	79.35	49.53	71.50	89.48	62.34
InternVL3.5-8B (Wang et al., 2025)	84.04	60.20	76.20	90.13	67.91
Qwen2.5-VL-7B (Bai et al., 2025c)	80.53	51.05	72.11	73.80	53.24
Qwen3-VL-8B (Bai et al., 2025a)	87.03	59.44	79.06	85.85	59.47
Phi-3.5-Vision (Abdin et al., 2024)	78.89	43.27	65.20	55.58	46.63
Closed-source MLLMs					
GPT-4o (Hurst et al., 2024)	85.87	51.41	76.18	89.06	72.67
GPT-4o-mini (Hurst et al., 2024)	80.41	40.27	66.87	81.95	59.75
GPT-5 (Singh et al., 2025)	82.89	50.18	77.32	84.31	68.56

A.3 EVALUATION METRICS

We report both task performance and reliability-oriented metrics to provide a holistic evaluation. Specifically, **ACC** measures end-task correctness, while **ECE** quantifies probabilistic calibration (how well predicted confidence matches empirical accuracy). In addition, we evaluate the discriminative ability of model confidence using threshold-free ranking metrics: AUROC, AUPRC, and its normalized variant AUPRC-N. Finally, we employ AURC to assess the potential for risk-sensitive selective prediction. Unless otherwise specified, higher values are better for ACC, AUROC, AUPRC, and AUPRC-N, while lower values are preferred for ECE and AURC.

ACC (↑). Accuracy is the fraction of total examples for which the model produces the correct answer (e.g., the correct option in multiple-choice settings). It is widely considered as the primary metric for evaluating overall end-task performance.

ECE (↓). Expected Calibration Error measures the gap between predicted confidence and empirical accuracy, typically computed by partitioning predictions into confidence bins and aggregating the absolute difference between average confidence and accuracy per bin. Lower ECE indicates better-calibrated confidence estimates.

AUROC (↑). Area Under the ROC Curve evaluates how well a scalar confidence score ranks correct predictions above incorrect ones across all possible thresholds. AUROC is insensitive to the choice of a specific operating point and reflects overall separability. Specifically, a score of 1.0 signifies perfect discrimination between the two classes.

Algorithm A.1 (Monotone Confidence Fusion with mean alignment). For each instance $(x_i, q_i, \mathcal{A}_i)$:

1. **Token confidence channel (instinct).** Compute option log-probabilities $\{z_{ik}\}_{k=1}^{K}$ at the answer position and form $p_i = \mathrm{softmax}(z_i)$ as in Eq.(1). Let $k_i^\star = \arg\max_k p_{ik}$ and $r_i^{\text{token}} = \max_k p_{ik}$.
2. **Verbalized confidence channel (reflection).** Prompt the MLLM to output per-option confidence (Appendix B.2); parse into $s_i \in [0,1]^K$ and set $r_i^{\text{verbal}} = s_{i,k_i^\star}$.
3. **Cross-channel consistency.** Compute κ_{ik} by Eq.(3) and set $r_i^{\text{cons}} = \kappa_{i,k_i^\star}$.
4. **Features.** Form ϕ_i by Eq.(5) using $\ell(\cdot)$ in Eq.(6), the top-2 margin $m_i = p_{i,(1)} - p_{i,(2)}$, and Shannon entropy $H(p_i) = -\sum_{k=1}^{K} p_{ik} \log p_{ik}$.

Fitting: Standardize features using calibration-split statistics; fit θ by minimizing Eq.(8) with $w_j = \mathrm{softplus}(\tilde{w}_j)$. **Mean alignment:** Estimate a_{ref} on validation split using the empirical target, or the Laplace target for the explicitly designated few-shot variant. Solve $g(\delta) = a_{\mathrm{ref}}$ with the data-dependent bracket in Eq.(12), checking endpoint conditions and the final residual (Appendix B.7). At zero target-label budget, retain the complete source calibrator. **Inference:** Output merged confidence $\hat{q}_i = \sigma(f_\theta(\phi_i) + \delta)$.

AUPRC (↑). Area Under the Precision–Recall Curve emphasizes performance under class imbalance and is often more informative than AUROC when the positive class is relatively rare. Here, we treat correct predictions as positives and compute AUPRC from the continuous confidence-based rankings to assess reliability thoroughly.

AUPRC-N (↑). AUPRC-N normalizes the Precision-Recall area against task-specific random baselines. Unlike standard AUPRC, it mitigates sensitivity to class priors, enabling fair reliability comparisons across diverse benchmarks by explicitly decoupling ranking quality from the task's intrinsic accuracy to measure the performance gain over random chance.

AURC (↓). Area Under the Risk-Coverage Curve serves as a pivotal metric for evaluating a model's uncertainty quantification capabilities. By reflecting the model's self-awareness, it exhibits high sensitivity to probability calibration, making it indispensable for assessing performance in safety-critical scenarios. Specifically, it measures the cumulative trade-off between coverage (fraction of the samples predicted) and selective risk (the error rate on those samples).

B Additional Implementation and Theoretical Details

This appendix details the end-to-end procedure for reproducing our reliability estimation pipeline, which encompasses four cardinal stages: (i) extracting and quantifying dual-channel confidence metrics, (ii) constructing monotonic reliability features, (iii) fitting the constrained monotonic fusion head, and (iv) executing order-preserving mean alignment.

B.1 End-to-end procedure and dataset splits

Data splits and leakage prevention. We assume the original benchmark provides a test set that is never used for any fitting or tuning. From the remaining data (or a held-out calibration corpus), we partition the samples into: (i) a calibration split to fit the fusion head parameters $\theta = (b, \{\tilde{w}_j\})$ in Eq.(8); (ii) a validation split to solve mean-alignment Eq.(9).

Deterministic answer protocol. To obtain comparable option-level scores and multi-dimensional confidence, we enforce a structured response format. Specifically, the MLLM is constrained to output the predicted option label token (e.g., "A"..."D" or "1"..."4") and its verbalized confidence score within a standardized and parseable template (e.g., a JSON object). We treat the model as a black box and directly extract the token probabilities for the option label at its specific generation index. This makes the token confidence p_i in Eq.(1) well-defined and consistently reproducible across multiple inference runs.

B.2 Prompting and parsing verbalized confidence

We obtain per-option verbalized confidence via a structured instruction that constrains the output format. The protocol in this subsection applies to multiple-choice questions. The scalar prompt,

parsing rules, and complete-case evaluation for open-ended answers are specified alongside Table 10. A representative template is: **Instruction:** You are given an image and a multiple-choice question with K options. First, output exactly one option label as your final answer. Then, output a JSON object only with keys {1,...,K} and numeric values in {0,...,100} indicating your confidence percentage for each option. Do not include any additional text.

Decoding settings for robustness. We use deterministic decoding (e.g., temperature 0) to reduce formatting variance. We constrain the maximum generation length so that the model has sufficient budget to output a complete JSON object but not enough to drift into free-form text. If supported, we additionally enforce stop sequences after the closing brace `}`.

Parsing and normalization. We parse the JSON object into $s_i \in [0,1]^K$ by: (i) extracting numeric values (allowing integers or floats), (ii) dividing by 100, and (iii) clipping each entry to $[0,1]$. If the output is malformed, we fall back to a regex-based extractor that matches `(option_id, number)` pairs and applies the same normalization.

Range handling and inclusion criteria. Parsed confidence percentages are divided by 100 and clipped to $[0,1]$. The inclusion criteria for the supplementary paired evaluations are specified with their respective experimental protocols.

Why we do not enforce $\sum_k s_{ik} = 1$. Unlike $p_i \in \Delta^{K-1}$, verbalized confidence can legitimately violate normalization: the model may express global uncertainty (all low) or multi-modality (several high). For this reason, the main text treats s_i as a hypercube vector and only uses the predicted-option component $r_i^{\text{verbal}} = s_{i,k_i^\star}$ together with cross-channel agreement.

B.3 Extracting token-option confidence

The token confidence defined in Eq.(1) requires model-retrieved log-probabilities for each option-label token at the answer position.

Single-token option labels. We choose option labels that are single tokens under the model tokenizer (e.g., `A/B/C/D` or `1/2/3/4` depending on the model). Let the prompt end with a fixed prefix such as `Final answer:` and let the next generated token be strictly restricted (implicitly or explicitly) to one of the K predefined labels. We obtain the log-probability vector $z_i = (z_{i1}, \ldots, z_{iK})$ for these label tokens and compute token confidence $p_i = \mathrm{softmax}(z_i)$ as in Eq.(1).

Numerical Stability and Reproducibility. Since the token confidence score p_i is derived directly from a single-step next-token distribution, it avoids the potential cascading error accumulation inherent in multi-step sequence generation. This approach exhibits inherent robustness to long-horizon decoding noise and sampling stochasticity. Consequently, we enforce the standardized single-label token protocol to ensure both determinism and experimental reproducibility.

B.4 Reliability features: construction, bounds, and standardization

We construct a compact reliability descriptor $\phi_i \in \mathbb{R}^d$ combining both channels' confidence features as in Eq.(5):

$$\phi_i = \Big[\ell(r_i^{\text{token}}),\ \ell(r_i^{\text{verbal}}),\ \ell(r_i^{\text{cons}}),\ m_i,\ -H(p_i)\Big]. \tag{10}$$

Log-odds transform with clipping. The map $\ell(\cdot)$ in Eq.(6) turns heterogeneous $(0,1)$ confidence values into an approximately linear domain for better regression stability. We clip inputs to $[\varepsilon, 1-\varepsilon]$ to avoid infinities. In practice, ε can be 10^{-6}.

Margin and entropy are complementary sharpness signals. The margin $m_i = p_{i,(1)} - p_{i,(2)} \in [0,1]$ captures top-1 separability, while $H(p_i) \in [0, \log K]$ captures global dispersion. We include both because they capture different aspects of distributional sharpness. For example, $p_i = [0.5, 0.5, 0, 0]$ has zero top-2 margin but entropy $H(p_i) = \log 2$, below the four-option maximum of $\log 4$, using the convention $0 \log 0 = 0$.

Sign conventions and monotone intent. We specifically use $-H(p_i)$ so that larger values indicate lower entropy and thus higher reliability, thereby matching the monotone prior established in the main text analysis.

Standardization preserves monotonicity. We standardize each feature dimension using calibration-split mean μ_j and standard deviation $\sigma_j > 0$: $\tilde{\phi}_{ij} = (\phi_{ij} - \mu_j)/\sigma_j$. Since this is an affine transform with positive scaling, and assuming the non-negativity constraint $w_j \geq 0$, the fused predictor remains coordinatewise nondecreasing in the original features:

$$\frac{\partial \hat{q}_i}{\partial \phi_{ij}} = \frac{\partial \hat{q}_i}{\partial \tilde{\phi}_{ij}} \cdot \frac{\partial \tilde{\phi}_{ij}}{\partial \phi_{ij}} = \sigma'\big(f_\theta(\tilde{\phi}_i)\big)\, w_j \cdot \frac{1}{\sigma_j} \geq 0. \tag{11}$$

If a feature has near-zero variance on the calibration split, we drop it or set $\sigma_j = 1$ and report this preprocessing.

B.5 Monotone fusion head: optimization, interpretability, and special cases

We predict correctness probability with $\hat{q}_i = \sigma\big(f_\theta(\phi_i)\big)$, $f_\theta(\phi_i) = b + \sum_{j=1}^{d} w_j \phi_{ij}$, and $w_j = \mathrm{softplus}(\tilde{w}_j) \geq 0$.

Coordinate-wise Monotonicity & Interpretability. The constraint $w_j \geq 0$ enforces a monotonic inductive bias, ensuring that predicted correctness is non-decreasing with respect to each reliability cue. In low-data regimes, this prevents overfitting to counter-intuitive negative weights. Since standardized features ensure higher values signify greater reliability, the magnitude of w_j directly quantifies each cue's marginal contribution. In practice, we enforce this by reparameterizing the weights through a Softplus function, ensuring coordinate-wise monotonicity of the calibration function.

Theoretical Connections & Optimization. Our formulation generalizes standard post-hoc calibration; specifically, when $d = 1$ and $\phi_i = \ell(r_i^{\text{token}})$, the framework reduces to a classic one-dimensional logistic calibrator (e.g., Platt scaling). This allows our approach to subsume common baselines while enabling multi-signal fusion within a single interpretable framework. We learn the parameters $\{\tilde{w}, b\}$ by minimizing the negative log-likelihood Eq.(8) on the calibration split. Given the low dimensionality of ϕ_i, the optimization is inherently stable. We employ the Adam optimizer with a small learning rate and mild weight decay, applying early stopping based on validation NLL to ensure robust generalization.

B.6 Order-preserving mean alignment: theory and numerics

Corollary 1 (Ranking invariance). Updating the bias as $b \leftarrow b + \delta$ adds the same constant to all output logits $\{f_\theta(\phi_i)\}_i$, preserving their relative ordering; thus the risk–coverage curves and any selection rule based on ranking are unchanged. This guarantee holds in real arithmetic within each setting that uses a shared finite shift. Ranking metrics can be computed from logits to avoid numerical ties caused by sigmoid saturation.

We apply a post-hoc bias shift $b \leftarrow b + \delta$, with δ to satisfy Eq.(9) using validation split statistics. This step is intentionally order-preserving (Corollary 1) and thus does not change any ranking-based evaluation (e.g., risk–coverage curves), while effectively correcting systematic base-rate misalignment arising from the initial calibration-split optimization.

Why a one-dimensional correction is often sufficient. In many deployment shifts, the primary change is the base accuracy (e.g., questions become slightly harder), while the relative ordering of easy vs. hard instances remains informative. A pure bias shift exactly targets this scenario without refitting weights and without altering selection ordering.

B.7 Solving mean alignment by bisection

Let $u_i = f_\theta(\phi_i)$ be the finite, pre-alignment logits on a nonempty alignment set $\mathcal{D}_{\text{align}}$ of size m, and define $g(\delta) = m^{-1} \sum_{i=1}^{m} \sigma(u_i + \delta)$. For a reference rate $a_{\text{ref}} \in (0, 1)$, mean alignment solves $g(\delta) = a_{\text{ref}}$. Here a_{ref} denotes the target mean confidence on the labeled alignment set, which is disjoint from the test set. The empirical method uses $a_{\text{ref}} = \mathrm{clip}(k/m, \varepsilon, 1 - \varepsilon)$, where $k = \sum_{i=1}^{m} y_i^{\text{bin}}$. For the main experimental setting of Eq. (9), the alignment set corresponds to the validation split, and thus $m = N_{\text{val}}$. The separately evaluated few-shot variant uses the Laplace target defined below. The original results retain the empirical target.

Data-dependent bracketing. Set $t = \mathrm{logit}(a_{\mathrm{ref}})$ and construct

$$\delta_{\min} = t - \max_i u_i, \qquad \delta_{\max} = t - \min_i u_i. \tag{12}$$

For every i, $u_i + \delta_{\min} \le t \le u_i + \delta_{\max}$. Monotonicity of the sigmoid therefore guarantees

$$g(\delta_{\min}) \le a_{\mathrm{ref}} \le g(\delta_{\max}). \tag{13}$$

The interval contains the solution in real arithmetic, without assuming that the logits are centered near zero or requiring sigmoid saturation. If all logits are equal, the endpoints coincide and give the exact solution.

Numerical convergence. The revised solver uses float64 arithmetic, a stable sigmoid, and $t = \log(a_{\mathrm{ref}}) - \log(1 - a_{\mathrm{ref}})$ with the second term computed using `log1p`. It verifies both endpoint inequalities. A violated endpoint is moved outward by one representable float64 step using `nextafter`, with at most 64 repair rounds. After this check, an endpoint is accepted if its absolute mean-probability residual is at most $\eta = 10^{-10}$. Otherwise, bisection uses the midpoint $\delta_{\min}/2 + \delta_{\max}/2$, updates the lower endpoint when $g(\delta) < a_{\mathrm{ref}}$ and the upper endpoint otherwise, and stops only when $|g(\delta) - a_{\mathrm{ref}}| \le \eta$. Invalid inputs, unrepresentable bounds, failed bracket repair, floating-point stagnation, or nonconvergence within 256 iterations raise an explicit error rather than silently returning a boundary value. Appendix C.2 audits this solver against the historical fixed-bracket implementation.

Few-shot reference-rate smoothing. Clipping the empirical rate to $[10^{-6}, 1 - 10^{-6}]$ prevents an infinite solution but does not resolve small-sample estimation error. Our separately reported few-shot variant uses $a_{\mathrm{ref}} = (k + 1)/(m + 2)$ for $m \ge 1$ binary-labeled target examples, where k is the number of correct predictions. This is the posterior mean of the marginal correctness rate under a fixed $\mathrm{Beta}(1, 1)$–Bernoulli model, used as a plug-in target rather than a prior on the fusion head. The prior is fixed independently of target test labels. For $m = 8$, all-correct and all-incorrect samples yield targets 0.9 and 0.1, respectively. Smoothing moderates the alignment target. It neither bounds every test confidence to this interval nor guarantees improvement in every calibration metric. At $m = 0$, we retain the source-calibrated predictor without target alignment. This binary-label variant is separate from the open-ended soft-target protocol in Table 10.

B.8 Theoretical Guarantees

Proposition 1 (Coordinatewise monotonicity). For any j, $\partial \hat{q}_i / \partial \phi_{ij} = \sigma'\big(f_\theta(\phi_i)\big)\, w_j \ge 0$. Hence the fused score Eq.(7) is nondecreasing in each feature, ensuring interpretable alignment with input reliability signals (Appendix B.8).

Theorem 1 (Existence and uniqueness of δ). For $m \ge 1$ finite logits $u_1, \ldots, u_m$, define

$$g(\delta) = \frac{1}{m} \sum_{i=1}^{m} \sigma(u_i + \delta). \tag{14}$$

For every fixed $a_{\mathrm{ref}} \in (0, 1)$, the equation $g(\delta) = a_{\mathrm{ref}}$ admits a unique finite solution in the interval of Eq.(12).

Proof of Proposition 1.

Recall $\hat{q}_i = \sigma(f_\theta(\phi_i))$ with $f_\theta(\phi_i) = b + \sum_{j=1}^{d} w_j \phi_{ij}$ and $w_j \ge 0$. For any coordinate j,

$$\frac{\partial \hat{q}_i}{\partial \phi_{ij}} = \sigma'\big(f_\theta(\phi_i)\big) \cdot \frac{\partial f_\theta(\phi_i)}{\partial \phi_{ij}} = \sigma'\big(f_\theta(\phi_i)\big)\, w_j. \tag{15}$$

Since $\sigma'(t) = \sigma(t)(1 - \sigma(t)) > 0$ for all t and $w_j \ge 0$, we obtain $\partial \hat{q}_i / \partial \phi_{ij} \ge 0$, proving coordinatewise monotonicity.

Proof of Theorem 1. Each term $\sigma(u_i + \delta)$ is continuous and strictly increasing in δ. Finite sums preserve these properties, so g is continuous and strictly increasing. Moreover, $\lim_{\delta \to -\infty} g(\delta) = 0$ and $\lim_{\delta \to +\infty} g(\delta) = 1$. Therefore, for any $a_{\mathrm{ref}} \in (0, 1)$ there exists a unique finite δ such that $g(\delta) = a_{\mathrm{ref}}$. The endpoint inequalities in Appendix B.7 establish containment in the data-dependent bracket. This proof applies to both the clipped empirical target and the Laplace target in real arithmetic. Finite-precision convergence is checked separately by the solver.

B.9 Complexity and practical considerations

Per-instance overhead. Computing p_i requires obtaining log-probabilities for K label tokens at one answer position; this is $O(K)$ additional work after the forward pass. Parsing s_i is linear in the output length and negligible in practice. For m alignment examples, each bisection evaluation costs $O(m)$. Let $W = \max_i u_i - \min_i u_i$ be the initial bracket width. Since $0 < g'(\delta) \leq 1/4$, a probability-residual tolerance η is reached in $O(\log(2 + W/\eta))$ evaluations in real arithmetic, giving $O(m \log(2 + W/\eta))$ total time, including the coincident-endpoint case $W = 0$. Alignment is performed once per evaluation setting, subject to the numerical checks in Appendix B.7.

Failure modes and mitigations. (i) Malformed JSON: use the regex fallback to recover labeled option–confidence pairs (Appendix B.2). (ii) Extremal probabilities (0 or 1): managed via numerical clipping in Eq.(6). (iii) Tokenizer mismatch for option labels: verify that each label is single-token before running experiments.

B.10 Hyperparameters and recommended ranges

We summarize the key hyperparameters and their recommended ranges to facilitate reproducibility. We retain $\varepsilon = 10^{-6}$ for the logit-based feature clipping in Eq.(6) and for the empirical reference rate. The few-shot Laplace target uses the fixed $\mathrm{Beta}(1, 1)$ prior and is already strictly inside $(0, 1)$ for every nonempty binary-labeled adaptation sample. Regarding the consistency coefficient in Eq.(3), we set the shape parameter γ to 2 (selecting from $\{1, 2\}$) and determine the bandwidth τ via a coarse grid search on the validation set. The revised bisection solver computes its bounds from Eq.(12), with no search-width hyperparameter. We use an absolute mean-probability residual tolerance $\eta = 10^{-10}$, at most 256 bisection iterations, and the endpoint checks specified in Appendix B.7.

C Additional Experiments

C.1 Confidence Diagnostics and Selective Prediction

Scope of the layerwise diagnostics. The layerwise analyses in Figures 7, 8, 9, and 10 use a separate answer-only run on six models with accessible hidden states. They probe how candidate confidence varies across decoder depth and are not an additional inference requirement of the output-based fusion method. When a layerwise score is compared with verbalized confidence, the latter is obtained through an independent structured prompt and mapped to the answer-only run's final predicted option. Such comparisons measure cross-prompt association rather than consistency between two signals from the same forward pass.

Layerwise evolution of token confidence. Figure 7 reveals that candidate confidence is formed non-monotonically across decoder depth. For each sample, the maximum softmax probability (MSP) is the largest candidate-normalized probability over the four answer options and the y-axis reports its mean across samples. Several models, including MiniCPM, InternVL2.5, and Qwen2.5-VL, exhibit pronounced U-shaped trajectories. The maximum candidate probability decreases or fluctuates through the middle blocks before rising sharply near the output layer. InternVL3.5, Qwen3-VL, and Phi-3.5 show similarly strong late-stage transitions, although their intermediate-layer dynamics are more step-like and model-specific. Within each model, the overall trajectory is broadly consistent across benchmarks, suggesting that confidence formation is governed primarily by the model architecture, while task difficulty modulates the final concentration level. In particular, MMMU-Pro frequently produces less saturated terminal confidence than easier benchmarks such as MMBench and ScienceQA. These results indicate that token confidence is not accumulated uniformly layer by layer. Instead, intermediate representations repeatedly reorganize candidate evidence before the final decoder blocks commit to a concentrated prediction. The layerwise view therefore complements our terminal token-confidence analysis by exposing when the model's instinctive preference emerges and stabilizes.

Scope of the paired-channel experiments. The selective-prediction experiment in Figure 11 use 69,936 paired complete cases with independently elicited verbal confidence. Both channels are evaluated against the token-predicted option, even when the verbal-confidence run selects a different

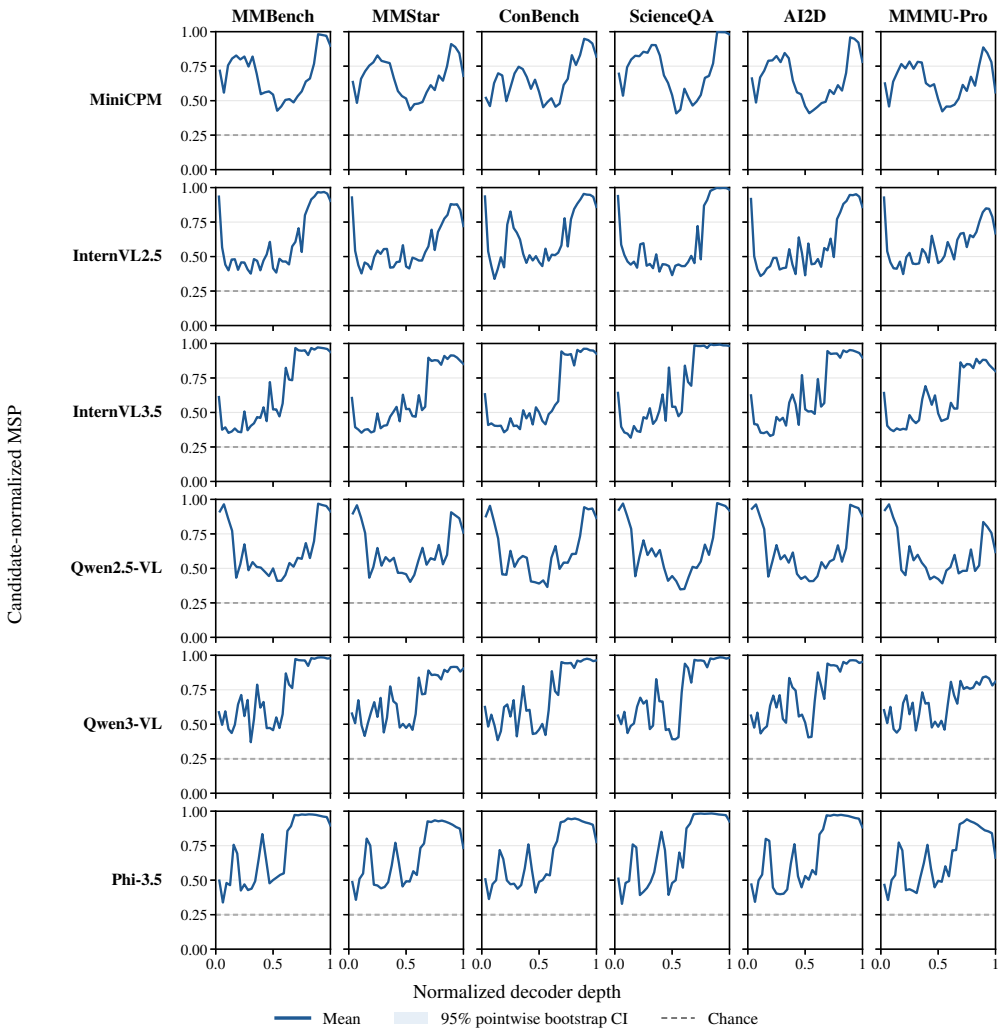


Figure 7: Layerwise candidate-normalized logit-lens confidence across six vision-language models and six benchmarks.

answer. This independent-call protocol is specific to these supplementary experiments and should not be interpreted as a measurement of same-forward-pass agreement or single-pass inference cost. The layerwise correlation analysis in Figure 10 separately retains 70,276 successfully parsed and mapped verbal scores. The two reported sample counts refer to different inclusion criteria and are not interchangeable.

Layerwise confidence separation. We collect 71,334 confidence trajectories from six vision-language models and six multiple-choice benchmarks. At decoder layer ℓ, we pass the hidden state at the answer position through the model's terminal normalization and original language-model head, aggregate tokenizer aliases of the candidate labels, and renormalize the resulting scores over the four options. Their maximum defines the layerwise MSP at depth ℓ/L. The models contain 28, 36, 36, 28, 36, and 32 decoder layers, respectively. Examples are grouped according to the correctness of the final prediction without fitting an additional classifier. As shown in Figure 8, the macro correct-minus-wrong MSP gap at the final layer is $+0.178$ and is positive in 34 of 36 settings. The only negative gaps occur for MiniCPM on MMMU-Pro (-0.030) and InternVL2.5 on MMMU-Pro (-0.036). Thus, correctness-related separation is a robust late-layer phenomenon on most benchmarks, while difficult professional reasoning settings can remain confidently ambiguous.

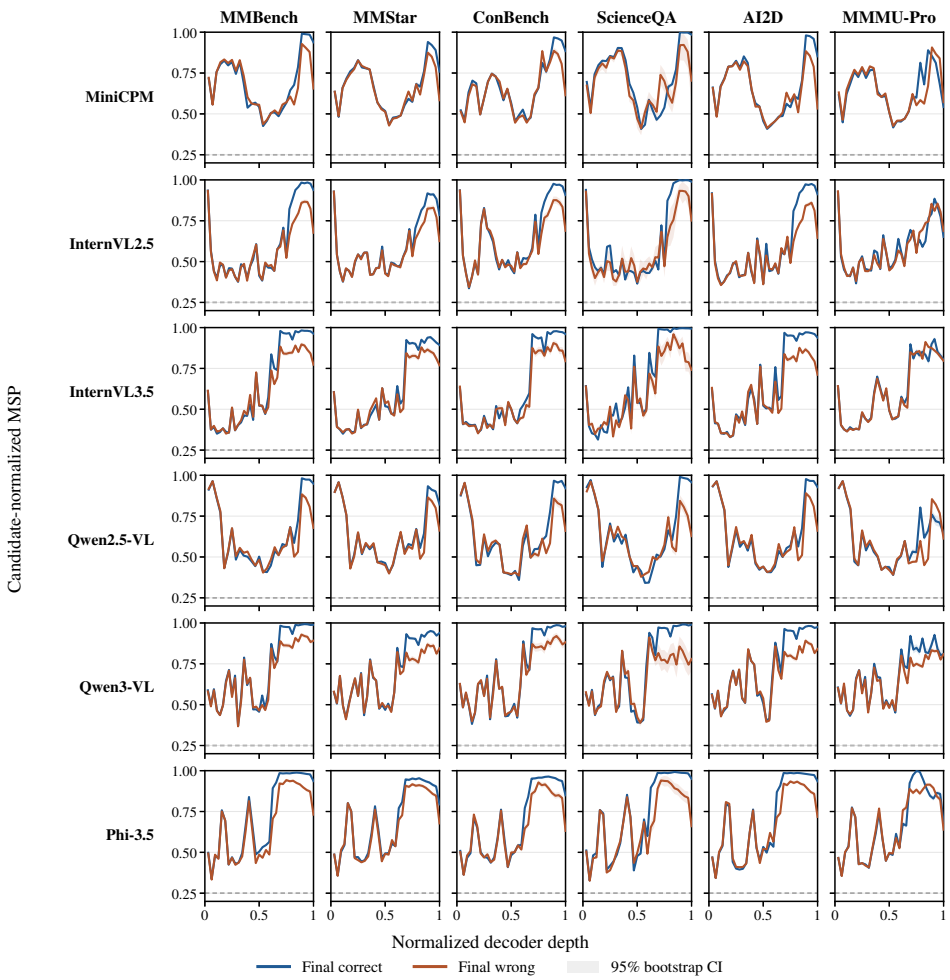


Figure 8: Confidence separation between correct and incorrect predictions emerges primarily in the late decoder.

Layerwise reliability evolution. Using the same 71,334 trajectories, we evaluate whether the confidence exposed by each intermediate decoder layer can predict the eventual model outcome. Failure AUROC measures the ranking of incorrect final answers using $1 - \mathrm{MSP}$, whereas layer accuracy evaluates the top-1 option directly readable from that layer. Final-outcome ECE and Brier score treat layerwise MSP as a probabilistic forecast of final correctness. No layer-specific calibrator is fitted. Averaged over the 36 model–dataset settings, final-layer failure AUROC is 0.801, while selecting the best recorded layer within each setting increases it to 0.847; the mean depth of this best layer is 0.955. Final-layer ECE and Brier score average 0.134 and 0.175, respectively. Figure 9 therefore indicates that reliability is formed late but not always monotonically: some intermediate representations retain failure-discriminative information that is weakened at the output layer.

Cross-prompt layerwise association. For every layer and model–dataset pair, we correlate the candidate-normalized MSP with the verbalized confidence assigned to the answer-only run's final predicted option. The layerwise trajectory is collected using an answer-only prompt, whereas verbal confidence is elicited independently using a structured per-option prompt. Of the 71,334 examples, 70,276 provide a valid mapped verbal score and are retained for this analysis. Final-layer correlations are positive in all 36 settings, with a macro Spearman correlation of $\rho = 0.386$ and a range of 0.171–0.622. The macro best-layer correlation is 0.424 and occurs at an average normalized depth of 0.924. However, the two prompts produce the same answer on only 73.84% of examples with a valid

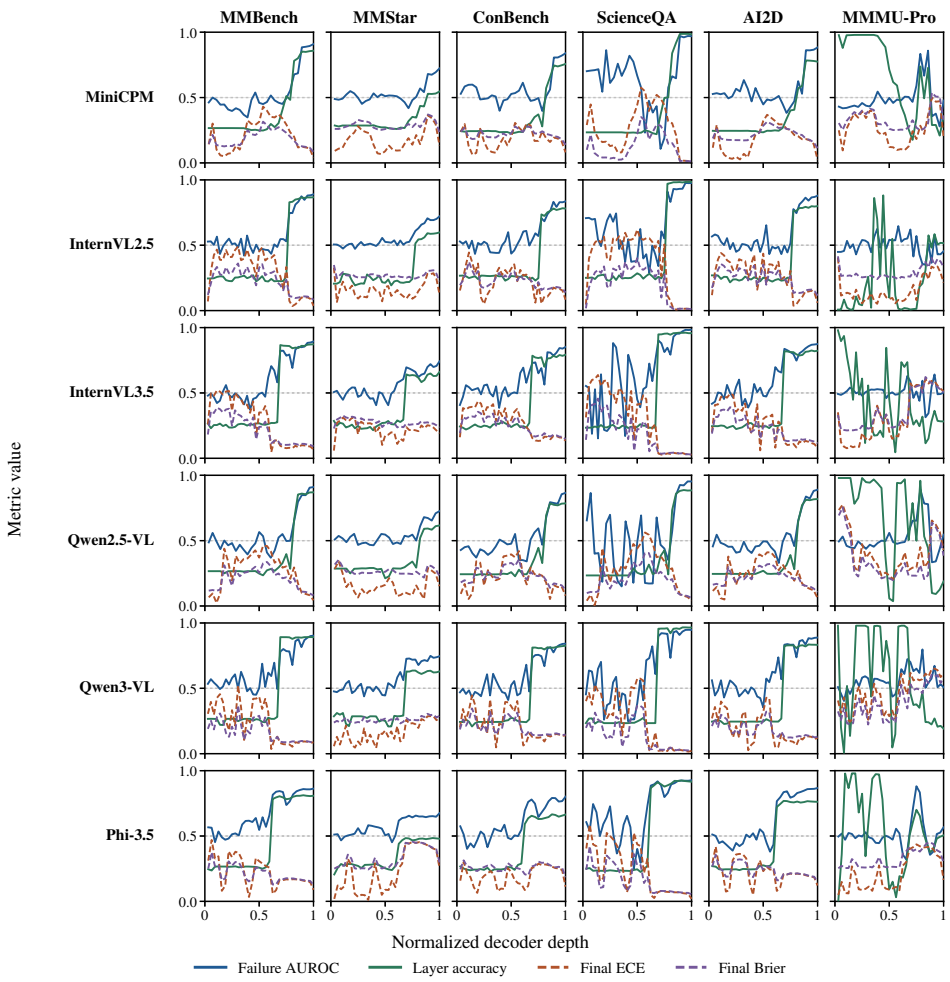


Figure 9: Failure detection, answer accuracy, and calibration evolve non-monotonically across decoder depth.

verbal answer. We therefore interpret Figure 10 as evidence of cross-prompt association, rather than same-forward internal–verbal consistency or a causal relationship.

Selective prediction. We use the same 69,936 paired complete cases and perform ten stratified 60/20/20 train–validation–test splits for each model–dataset setting. The fusion features comprise token confidence, verbalized confidence, RBF cross-channel consistency, the top-two candidate margin, and candidate entropy. Feature scaling and fusion weights are learned only on the training split, the RBF bandwidth is selected using validation NLL, and the scalar mean-alignment offset is estimated on the validation split. The curves in Figure 11 use the held-out test set from the first split, while numerical summaries are averaged over all ten splits. Macro E-AURC decreases from 0.0986 for token confidence and 0.1382 for verbal confidence to 0.0806 for our method, while the unconstrained head obtains 0.0770. Our method also improves macro AUROC from 0.796 and 0.664 for the two individual channels to 0.823. Because mean alignment is order preserving, it changes calibration but not the risk–coverage ordering. The ranking gains are therefore meaningful but setting dependent rather than universal.

Cross-dataset transfer and target-label budget. For each vision-language model, we treat every ordered pair of distinct datasets as a transfer task, yielding $6 \times 6 \times 5 = 180$ directed transfers. The five-feature monotone calibrator is fitted exclusively on the source domain: feature scaling and weights are learned from a stratified 60% training split, the RBF bandwidth is selected from

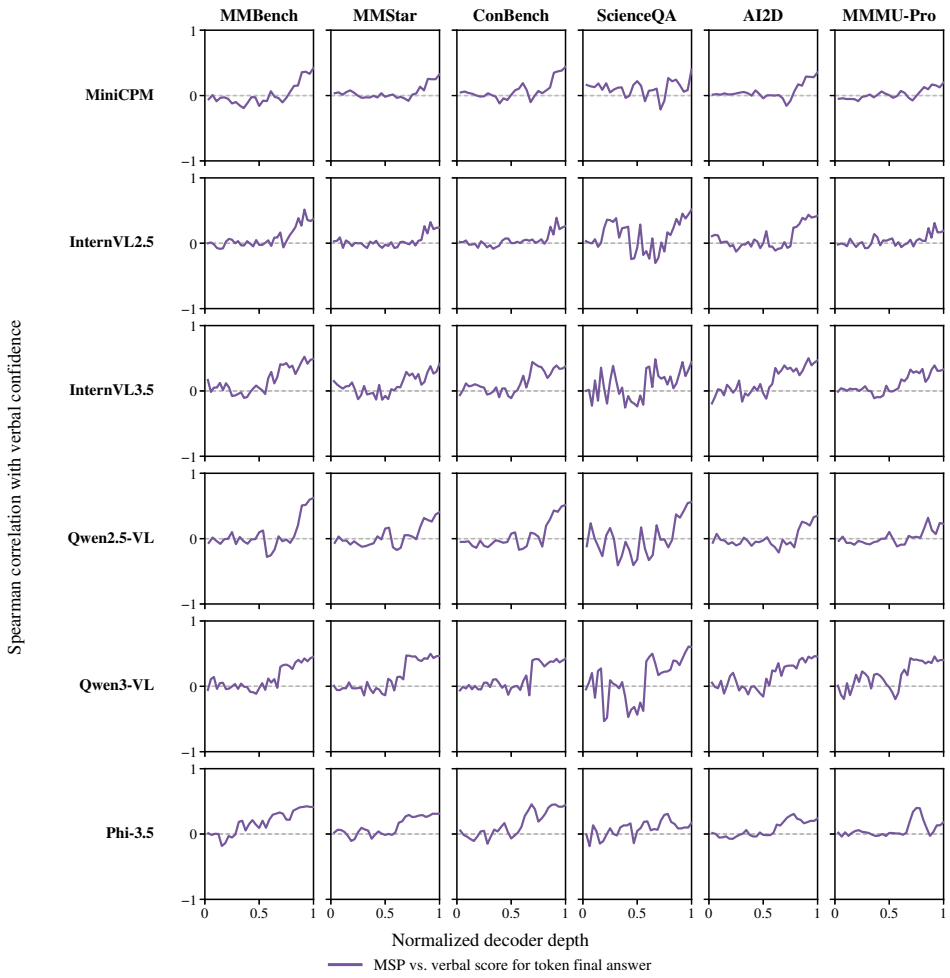


Figure 10: Layerwise logit confidence becomes increasingly associated with verbalized confidence near the output.

$\{0.02, 0.05, 0.1, 0.2, 0.5, 1.0\}$ using NLL on a disjoint 20% validation split, and the source alignment offset is estimated on that validation split. For a target budget $m > 0$, we sample m labeled target examples using correctness-stratified sampling and re-estimate only the scalar offset; all remaining target examples are reserved for evaluation. Sampling is repeated 20 times, and 95% confidence intervals use 2,000 bootstrap resamples of pair-level means. Figure 12 shows that macro ECE and NLL decrease from 0.1645 and 0.600 in the source-frozen setting to 0.0960 and 0.508 with eight selected target labels, and to 0.0758 and 0.468 with 128 labels. AUROC remains essentially unchanged (0.7774 versus 0.7773), confirming that the target offset mainly repairs base-rate calibration rather than ranking. Because this sampler accesses correctness labels for the full target pool, these results are a controlled-composition diagnostic. The label-independent protocol below tests adaptation using only the m sampled target labels. At $m = 0$, no target alignment is performed and the complete source calibrator is retained.

C.2 Mean-alignment Robustness

Sampling and evaluation protocol. We reuse the 36 archived source-trained heads for the same six models and six benchmarks (180 directed transfers), reconstructing feature standardizers from the original source training splits without refitting the heads. We compare source-frozen inference, dynamic-bracket alignment to clipped empirical accuracy, and dynamic-bracket alignment to the fixed Laplace target $(k + 1)/(m + 2)$. For each model–target pair, 100 predetermined uniform

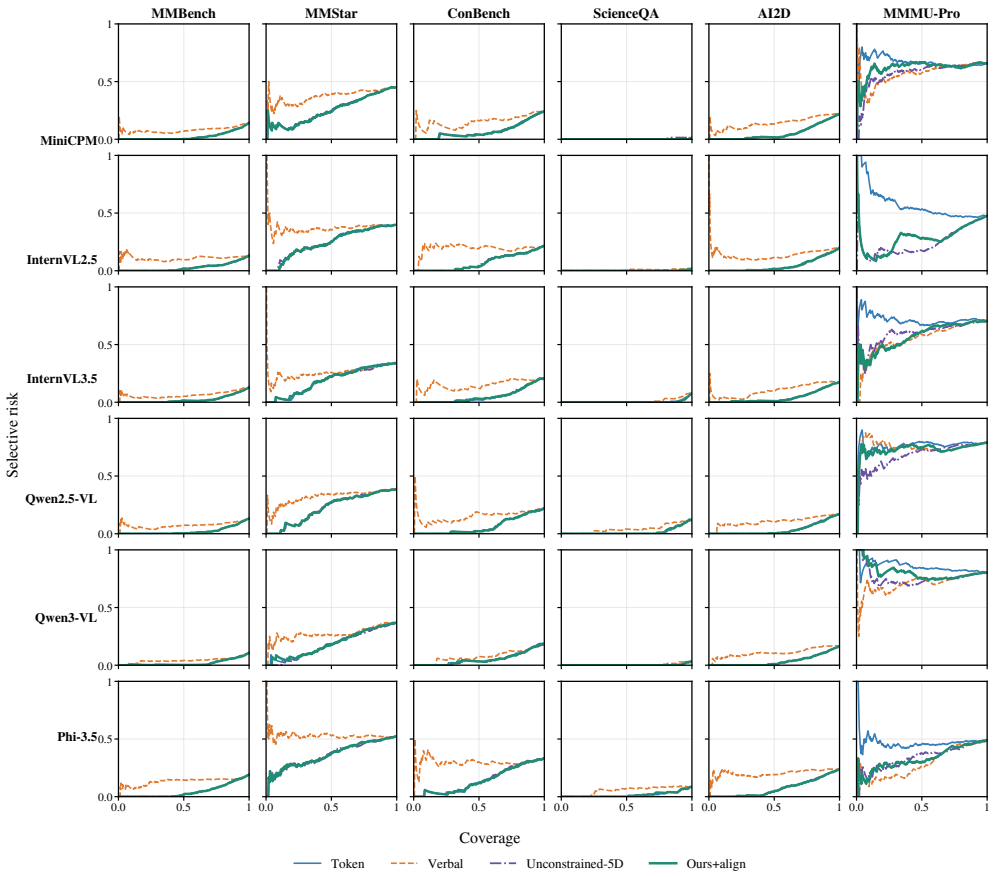


Figure 11: Multi-channel fusion generally improves selective prediction over either confidence channel alone.

random permutations, generated without accessing correctness labels, define nested adaptation sets of $m \in \{8, 16, 32, 64, 128\}$ by their prefixes. The same indices are shared across all five source heads and all estimators. The remaining target examples form the disjoint evaluation set. We keep source weights, feature scaling, and RBF bandwidth fixed. For $m > 0$, the target offset replaces the source offset. At $m = 0$, all estimators retain the complete source calibrator and are evaluated once on the full target set. ECE uses 15 equal-width bins, including confidence 1 in the last bin. ECE and Brier use sigmoid probabilities without additional output clipping, and NLL is evaluated directly from logits using a stable log-sum-exp expression. We average repeats within each transfer, then the five source heads within each model–target cell, and finally the 36 cells equally. Paired 95% percentile intervals use 2,000 bootstrap resamples of these cells and describe variation across this benchmark panel, rather than population-level significance for independent models or datasets. We also evaluate the original correctness-stratified sampler with its 20 repeats under these same metric definitions. The probability-clipped historical metrics in Figure 12 are retained separately.

Few-shot robustness and its limits. With $m = 8$, 815 of 3,600 distinct uniform adaptation draws (22.64%) are entirely correct or entirely incorrect. Laplace smoothing reduces macro NLL from 0.8018 to 0.5426, with a paired difference of -0.2592 and a descriptive 95% interval of $[-0.3117, -0.2085]$. NLL improves in 35 of the 36 model–target cells. Brier decreases from 0.1697 to 0.1673, whereas ECE increases from 13.30% to 13.56%. The paired difference intervals for both Brier and ECE include zero. At $m = 128$, NLL changes from 0.4734 to 0.4722. Under the original correctness-stratified sampler, smoothing instead worsens $m = 8$ ECE from 9.60% to 12.26% and NLL from 0.5081 to 0.5226. These results support improved log-loss robustness under limited, randomly acquired labels, with benefits that depend on the sampling regime and metric. We therefore

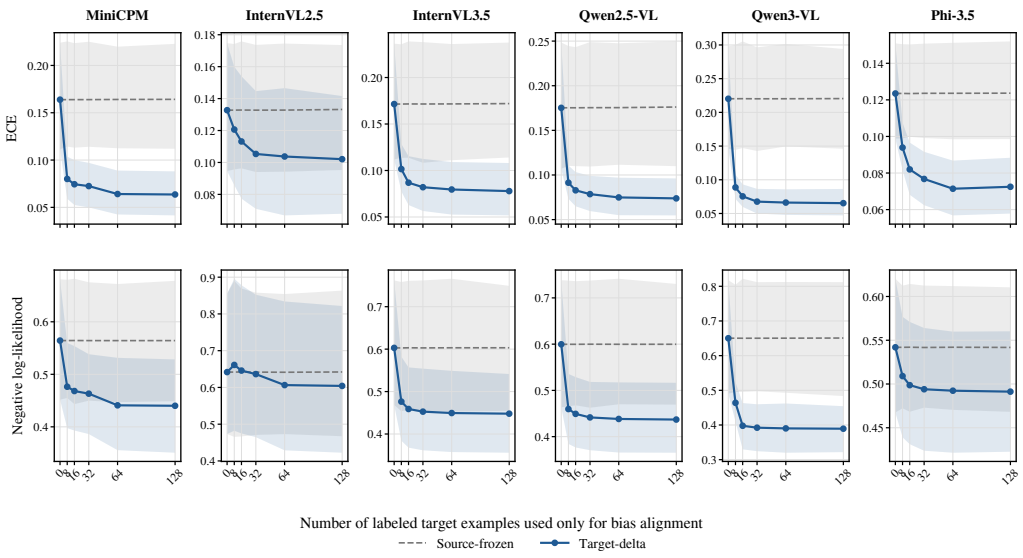


Figure 12: Cross-dataset mean alignment under correctness-stratified target sampling.

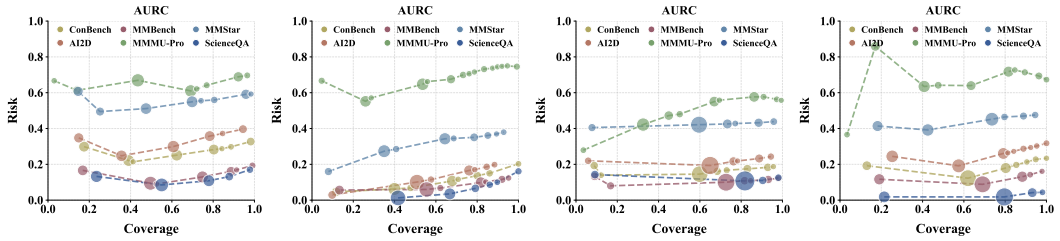
report Laplace as a separate few-shot variant and retain the empirical-target definition for the original results.

Numerical bracketing audit. Neither the previously specified interval $[-20, 20]$ nor the historical transfer implementation's interval $[-30, 30]$ fails to bracket the solution on the 18,000 original target-alignment fits. To test dependence on logit location, we translate both adaptation and evaluation logits by $c \in \{-100, 100\}$ for the first original repeat of every transfer and positive budget, giving 1,800 stress cases. Both fixed intervals fail to bracket the root in all 1,800 cases, whereas the dynamic solver succeeds in all cases with absolute mean-probability residual below 10^{-10}. Test probabilities are invariant to translation within 2×10^{-14}. Thus the revised interval removes a mathematical limitation of fixed bracketing. These stress-test failures do not imply failures in the original runs.

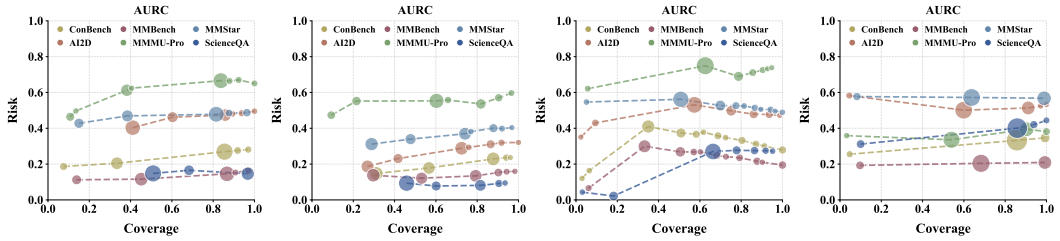
C.3 Calibration and Prompting Analyses

Results of risk–coverage analysis. We extend our analysis to eight additional Multi-modal Large Language Models (MLLMs), comparing both closed-source models (*e.g.*, Claude-3.7-sonnet, Gemini-2.0-Flash) and open-source models (*e.g.*, Qwen2.5-VL, InternVL3.5). Figure 13 presents the risk–coverage (RC) curves for these models. The results confirm the trends observed in the main text while highlighting performance differences between the two categories. First, on the challenging MMMU-Pro benchmark, both closed-source and open-source models exhibit persistently high risk across all coverage levels. This indicates that current confidence estimation methods struggle to reliably distinguish correct from incorrect answers on complex reasoning tasks, regardless of model access type. Second, on standard benchmarks like MMBench and ScienceQA, most models demonstrate effective selective prediction, where the risk decreases significantly as we reduce coverage. However, closed-source models generally achieve lower overall risk compared to open-source models, particularly in the high-coverage regions. Finally, the non-monotonic behavior on MMStar is visible in models across both groups, such as GPT-4o-mini and Qwen2.5-VL. This suggests that the misalignment between confidence and correctness on specific visual dependencies is a common issue shared by diverse model architectures.

Effect of Temperature on Calibration. To examine whether decoding temperature can improve verbalized-confidence calibration, we vary the sampling temperature while keeping the remaining decoding configuration fixed, motivated by work on confidence-aware answer aggregation (Li et al., 2025; Taubenfeld et al., 2025) and multi-sample temperature optimization (Du et al., 2025). We sweep $T \in \{0.1, 0.5, 0.7, 0.9\}$ and measure ECE across the six benchmarks. The temperature experiments show setting-dependent changes: higher temperatures improve calibration on ConBench in some



(a) GPT-4o-mini (b) Claude-3.7-sonnet (c) Gemini-2.0-Flash (d) GLM-4.1V-9B



(e) MiniCPM-V-2.6 (f) InternVL3.5-8B (g) Qwen2.5-VL-7B (h) Phi-3.5-Vision

Figure 13: Risk–coverage analysis for eight comparative MLLMs. We report the risk ($1 -$ accuracy) at varying coverage levels across six multi-modal benchmarks. The top row displays results for closed-source models (first three) and GLM-4.1V-9B, while the bottom row shows open-source models. Across both groups, ScienceQA and MMBench (bottom curves) show high reliability, whereas MMMU-Pro (top curves) remains challenging with high risk, indicating limited effectiveness of confidence-based selection on expert-level tasks.

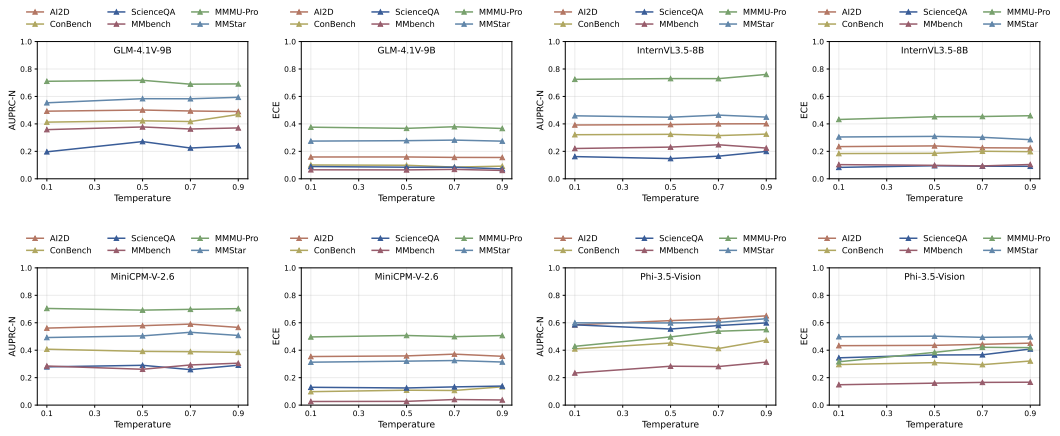
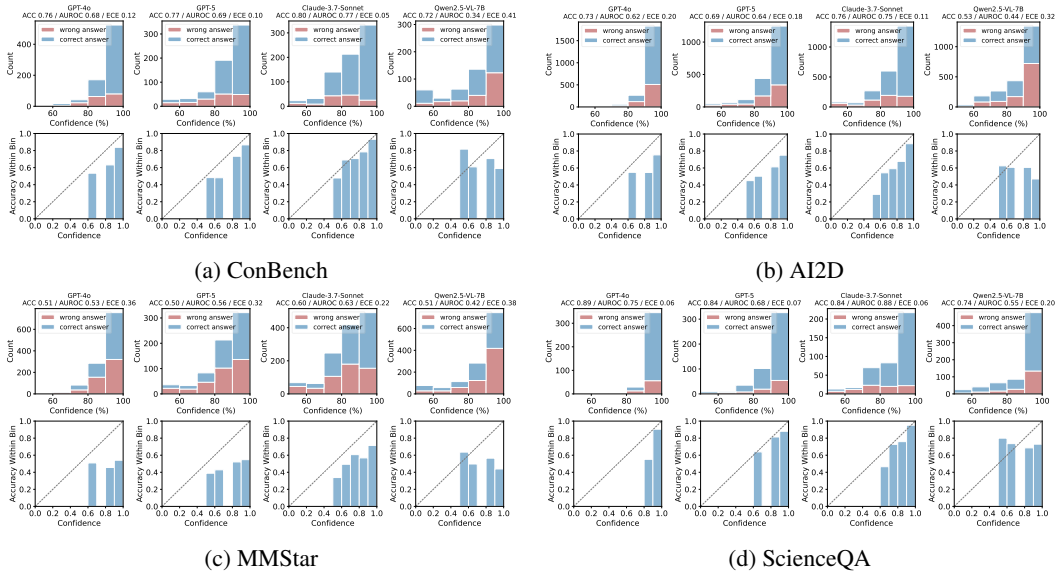


Figure 14: Effect of Decoding Temperature on Calibration and Ranking. We report the AUPRC-N and ECE metrics for four multimodal models: GLM-4.1V-9B, InternVL3.5-8B, MiniCPM-V-2.6, and Phi-3.5-Vision. We evaluate performance across six benchmarks by sweeping temperature $T \in \{0.1, 0.5, 0.7, 0.9\}$. Most curves vary little over the tested temperature range.

settings, while MMStar slightly degrades at the highest temperature; ECE remains relatively stable on most other benchmarks. To examine whether these trends generalize, Figure 14 reports both ECE and AUPRC-N for GLM-4.1V-9B, InternVL3.5-8B (Wang et al., 2025), MiniCPM-V-2.6 (Yao et al., 2024), and Phi-3.5-Vision (Abdin et al., 2024) under the same sweep. For these four models, most curves are nearly flat, with modest sensitivity in Phi-3.5-Vision, indicating limited changes in calibration and ranking over the tested range. Thus, any benefits from increasing temperature are model- and dataset-dependent rather than universal. Changing the generation temperature alone does not consistently resolve confidence miscalibration, and this experiment should be distinguished from fitting a post-hoc temperature calibrator to fixed predictions.



(a) ConBench

(b) AI2D

(c) MMStar

(d) ScienceQA

Figure 15: Extended calibration analysis across ConBench (top left), AI2D (top right), MMStar (bottom left), and ScienceQA (bottom right). Each panel displays the confidence histograms (top) for correct versus wrong predictions and the corresponding reliability diagrams (bottom).

Calibration and Reliability Analysis. Figure 15 extends the visualization to ConBench, AI2D, MMStar, and ScienceQA. The top-left panel shows ConBench, where GPT-4o, GPT-5, and Claude-3.7-Sonnet have ECE values of 0.12, 0.10, and 0.05, respectively, compared with 0.41 for Qwen2.5-VL-7B. The three closed-source models also have lower ECE than Qwen2.5-VL-7B on the other displayed benchmarks, but their calibration is not uniformly close to ideal: on MMStar, their ECE ranges from 0.22 to 0.36, compared with 0.06–0.07 on ScienceQA. These values use the $[0, 1]$ scale printed in the panels. Thus, the visualization supports a calibration advantage for these closed-source models over this open-source comparator, while also showing substantial task-dependent variation.

Prompting Strategies for Confidence Estimation. To comprehensively evaluate MLLM confidence, we design and compare six distinct prompting strategies to elicit model confidence. Specifically, the Vanilla (Verbalized) approach asks models to directly assign a 0-100 scalar score based on its holistic judgment. To encourage deeper reasoning, we employ Chain-of-Thought (CoT) and Multi-Step methods, which require the model to break down the process into step-by-step explanations or aggregate confidence from sub-tasks. We also explore introspection via Self-Probing and Roleplay, where the model critiques its own answers or adopts a cautious expert persona to avoid over-assertion. Finally, the Confidence Interval strategy captures uncertainty intervals by asking for lower and upper bounds rather than a single point estimate. Detailed illustrations of the specific prompt templates corresponding to these strategies are provided in Figure 16, 17, 18.

Calibration Analysis of Prompting Strategies The cross-sectional comparisons across Table 17, 18, 19, and 20 reveal the complex dynamics of prompting strategies in multimodal calibration. These findings suggest that relying solely on prompt engineering is not a panacea for improving calibration performance. Its effectiveness is significantly influenced by both the model and the task. Specifically, complex strategies exhibit a "double-edged sword" effect: for instance, the Chain-of-Thought (CoT) strategy significantly improves calibration for Qwen2-VL-7B (with ECE decreasing from 16.77 to 6.99 on MMBench), yet it causes performance degradation for MiniCPM-V-2.6. This suggests that for models with already well-calibrated confidence, forced reasoning may introduce noise. Additionally, a clear decoupling phenomenon is observed between accuracy (ACC) and calibration (ECE). For example, although the Roleplay strategy improves the accuracy of InternVL3.5-8B, it does not correspondingly reduce ECE on tasks such as AI2D. It indicates that increased problem-solving ability does not necessarily equate to improved self-calibration. More critically, in high-difficulty reasoning tasks such as MathVista, all models exhibit persistent overconfidence (with ECE mostly

above 20.0) under various prompting strategies. The Confidence Interval strategy even leads to a broad performance regression. These results demonstrate that relying solely on prompt engineering is insufficient to address multimodal hallucination, further validating the superiority of the proposed dual-channel confidence fusion method for achieving robust calibration.

C.4 Transferability and Data Efficiency

Table 5: Cross-dataset transfer on InternVL3.5-8B. Each row denotes the source dataset used to fit the fusion head, and columns denote target datasets. We report ECE ($\times 10^2$; lower is better). The first row is the stronger of token and verbalized confidence on each target dataset.

Source \ Target	MMBench	MMStar	ConBench	ScienceQA	AI2D	MMMU-Pro
Best single channel	5.63	20.12	13.12	4.53	9.23	41.91
MMBench	3.22	7.22	4.78	2.44	2.85	24.48
MMStar	4.79	10.01	4.75	5.05	6.04	17.17
ConBench	5.77	8.32	11.42	3.27	5.70	19.98
ScienceQA	5.25	14.19	8.30	3.05	5.87	29.68
AI2D	1.59	7.49	5.34	2.59	4.63	25.19
MMMU-Pro	6.48	7.79	8.99	8.38	9.84	5.38

Cross-dataset transfer. Table 5 shows that source-trained fusion heads, with target-validation mean alignment, improve ECE over the target's best single-channel baseline in 25 of the 30 cross-dataset transfers. On MMMU-Pro, all five transferred heads reduce ECE from 41.91 to 17.17–29.68, while the in-domain head reaches 5.38. All five transfers to MMStar also improve ECE, obtaining 7.22–14.19 compared with 20.12. The five exceptions are ConBench $\rightarrow$ MMBench (5.77 vs. 5.63), MMStar $\rightarrow$ ScienceQA (5.05 vs. 4.53), and MMMU-Pro $\rightarrow$ MMBench, ScienceQA, and AI2D (6.48, 8.38, and 9.84 vs. 5.63, 4.53, and 9.23, respectively). These results support the usefulness of transferred fusion weights in most settings, but show that target mean alignment does not guarantee an improvement over single-channel confidence for every source–target pair.

Table 6: Cross-prompt transfer on ConBench with InternVL3.5-8B. Rows and columns denote source and target prompting strategies, respectively. Values are ECE ($\times 10^2$; lower is better).

Source \ Target	Vanilla	CoT	Multi-step	Roleplay
Best single channel	13.12	10.20	8.62	16.32
Vanilla	5.94	8.53	8.13	7.04
CoT	6.42	5.12	4.74	6.49
Multi-step	6.10	5.86	6.15	6.87
Roleplay	5.69	5.46	6.67	7.39

Cross-prompt transfer. The results in Table 6 demonstrate that calibration is robust to changes in confidence-elicitation prompts. Every transferred fusion head improves over the corresponding best single-channel reference, including the challenging Roleplay target, for which ECE decreases from 16.32 to 6.49–7.39. A head trained with CoT prompting transfers particularly well, achieving the lowest ECE on CoT, Multi-step, and Roleplay targets. This behavior suggests that the fusion mechanism does not merely memorize the numerical distribution induced by a particular prompt; instead, it exploits stable relationships among the two confidence channels and their agreement.

Transfer across related MLLMs. Table 7 further transfers a fusion head between related models without refitting its weights on the target model. All eight InternVL transfer results and seven of the eight GPT-family results improve upon the target model's best single confidence channel. The largest reduction occurs for InternVL2.5-4B $\rightarrow$ InternVL3.5-8B on MMMU-Pro, where ECE decreases by 19.45 points. The sole degradation, GPT-4o-mini $\rightarrow$ GPT-4o on ConBench, increases ECE by 5.39 points, indicating that model transfer is most reliable within closely matched confidence distributions and should still be validated when model scale or API behavior changes.

Table 7: Transfer across related MLLMs. Values are target ECE ($\times 10^2$; lower is better), with changes relative to the target's best single-channel confidence in parentheses.

Source model	Target model	MMStar	ConBench	AI2D	MMMU-Pro
InternVL2.5-4B	InternVL3.5-8B	10.76 (-9.36)	10.09 (-3.03)	6.56 (-2.67)	22.46 (-19.45)
InternVL3.5-8B	InternVL2.5-4B	9.57 (-9.17)	9.87 (-2.39)	6.48 (-1.65)	26.67 (-1.77)
GPT-4o	GPT-4o-mini	14.91 (-4.15)	7.87 (-6.23)	10.03 (-4.25)	6.72 (-11.84)
GPT-4o-mini	GPT-4o	12.48 (-9.20)	13.80 ($+5.39$)	8.90 (-14.84)	5.44 (-9.54)

Table 8: Mean calibration budget: the minimum number of labeled samples required for the fused confidence to stably outperform token-only and verbal-only confidence in ECE.

Dataset	MMBench	MMStar	ConBench	AI2D	MMMU-Pro
Mean budget N	97	39	139	109	20

Labeled calibration data. To quantify supervision requirements, we use the same stratified 60/20/20 calibration–validation–test split for each model–dataset pair. We repeatedly subsample different numbers of labeled examples from the 60% calibration split to fit the fusion head, while keeping the 20% validation split fixed for hyperparameter selection and mean alignment. All calibration budgets are evaluated on the same held-out 20% test split. Table 8 reports the mean minimum budget at which our fused confidence stably achieves lower ECE than both single-channel baselines. The required budget is modest across the reported benchmarks, ranging from only 20 examples on MMMU-Pro to 139 on ConBench. This result is consistent with the low dimensionality of the fusion head and indicates that the method can provide calibration gains without requiring a large labeled adaptation set.

Table 9: Matched-budget calibration across seven models and six datasets ($[0, 1]$ scale).

Budget	Method	ECE↓	NLL↓	Brier↓
64	Ours	0.08	0.46	0.14
64	Best generic	0.07 (Iso/Platt)	0.44 (Platt-p)	≈ 0.14 (Platt-p)
128	Ours	0.07	**0.4186**	**0.1383**
128	Best generic	0.05 (Iso-s)	0.4204 (Platt-p)	0.1403 (LR-5D)
512	Ours	0.07	0.40	0.13
512	Best generic	0.04 (Iso)	0.40 (LR-5D)	≈ 0.13 (LR-5D)

Comparison with generic calibrators. Table 9 compares our method with Platt, isotonic, and logistic-regression calibrators under identical splits, seeds, and label budgets, averaged over seven models and six datasets. "Best generic" reports the best result separately for each metric, with the selected configuration in parentheses, rather than a single configuration that attains all reported values. At $N = 128$, our method obtains slightly lower mean NLL and Brier score (0.4186 and 0.1383) than the strongest generic alternatives (0.4204 and 0.1403), while isotonic calibration achieves lower ECE (0.05 versus 0.07). At $N = 64$ and $N = 512$, our Brier score is close to the best generic result at the reported precision. Thus, the proposed fusion is competitive on proper scoring rules but does not uniformly outperform generic calibrators in ECE.

C.5 Broader Models, Tasks, and Perturbations

Extension to open-ended generation. We evaluate whether the framework transfers beyond candidate-based multiple-choice QA on OK-VQA (Marino et al., 2019) and VQAv2 (Goyal et al., 2017). For a generated answer $y_i = (t_1, \ldots, t_L)$, token confidence is defined by the geometric mean of its autoregressive token probabilities, $c_i^{\mathrm{tok}} = \exp\Big(\frac{1}{L}\sum_{\ell=1}^{L} \log P(t_\ell \mid x_i, q_i, t_{<\ell})\Big)$. Because no candidate distribution is available, the top-2 margin and entropy features are omitted, and Kendall correlation is used to measure ranking against the soft VQA agreement score. Table 10 shows that

Table 10: Open-ended VQA results. Each cell reports token/verbalized/calibrated confidence (T/V/C), with all metrics shown on the $\times 10^2$ scale. Lower ECE and higher Kendall correlation are better. All three scores use the token-run answer's VQA agreement as the common target; V is elicited in a separate answer-and-confidence run.

Model	OK-VQA ECE	OK-VQA Kendall	VQAv2 ECE	VQAv2 Kendall
InternVL3.5-8B	44.41/53.66/**6.63**	28.06/28.96/**32.30**	25.64/34.33/**5.80**	18.67/**26.67**/22.84
MiniCPM-V-2.6	35.20/39.33/**3.89**	**41.74**/23.75/40.61	14.08/16.25/**3.46**	44.19/23.72/**44.96**
Phi-3.5-Vision	32.49/38.12/**3.68**	**38.55**/15.39/37.78	16.87/26.25/**3.32**	47.20/18.48/**47.65**
Qwen2.5-VL-7B	65.05/45.80/**5.46**	31.34/30.50/**33.01**	64.99/49.67/**6.17**	**30.28**/21.02/27.88

calibrated confidence consistently produces the lowest ECE, reducing it to 3.32–6.63 across all model–dataset pairs. Kendall correlation remains competitive and is best for four of eight pairs, supporting the use of paired token and question-level verbal confidence for predicting soft answer quality in open-ended VQA.

Open-ended confidence elicitation and parsing. For both datasets, a separate image–question inference run jointly generates a short answer and one scalar verbalized confidence. The image is supplied through the model's multimodal interface, and the following user prompt is used verbatim, with `{question}` replaced by the question:

```
You are a highly capable AI assistant.
Read the image and answer the question.

Return strictly in JSON with no extra text:
{
  "answer": "short answer",
  "confidence": 0
}

Rules:
1. "answer" must be concise (usually a short phrase).
2. "confidence" must be an integer from 0 to 100.
3. No explanation. No markdown. No code block.

Question: {question}
```

We unwrap the model response and extract a dictionary from a fenced block, the full response, or the substring between its outer braces, trying JSON decoding followed by Python-literal parsing. We read `answer` (or `chosen_answer`) and `confidence`, also accepting the supported nested `message` list. Numeric confidence values are converted by `int(float(value))`, clipped to $[0, 100]$, and divided by 100. If the dictionary does not supply a confidence, a case-insensitive, line-anchored `confidence: number` fallback accepts one to three digits and applies the same clipping and scaling. Answer extraction falls back to an `answer:` line and then the first nonempty line, unless an explicitly empty structured answer was supplied; fixed answer prefixes, surrounding quotes, and a terminal period are removed, and whitespace is collapsed. An unrecoverable confidence remains missing.

Pairing, target identity, and fitting. We join the token and verbal runs by `problem_id` and retain complete pairs with both confidence scores and a VQA agreement target. All three scores are evaluated against the agreement score of the token-run answer. The independently generated verbal response may contain a different answer: its confidence is therefore an auxiliary question-level signal, not necessarily self-confidence in that same token-run answer. For each model–dataset pair, we first sort the examples by their string-valued question IDs and then, using random seed 2025, randomly split the data into training, validation, and test sets. In the model order shown in Table 10, the retained/test counts are $5046/1010$, $4971/995$, $4967/994$, and $4573/915$ for OK-VQA, and $3000/600$, $2967/594$, $2920/584$, and $2785/557$ for VQAv2. The three fusion inputs are the clipped log-odds of token confidence, verbal confidence, and $\exp(-|c_i^{\mathrm{tok}} - c_i^{\mathrm{verb}}|^2/0.1)$, with logit clipping at 10^{-6} and $1 - 10^{-6}$ and no feature standardization. The monotone head is fitted on the calibration

split using soft-target binary cross-entropy, while mean alignment uses the validation split's mean agreement score as a_{ref}. The test split is used exclusively for final evaluation.

Soft VQA targets and expected calibration error. Let $\hat{y}_i^{\mathrm{tok}}$ be the token-run answer and g_{ij} the jth valid reference answer. Prediction and references undergo VQA answer normalization, including lowercasing, punctuation processing, number-word conversion, article removal, and contraction normalization. With J_i valid reference answers (normally ten), the leave-one-annotator-out agreement target is

$$a_i = \frac{1}{J_i} \sum_{j=1}^{J_i} \min \left\{ 1, \frac{1}{3} \sum_{\substack{k=1 \\ k \neq j}}^{J_i} \mathbf{1}\left[\mathrm{norm}(\hat{y}_i^{\mathrm{tok}}) = \mathrm{norm}(g_{ik})\right] \right\}. \tag{16}$$

We retain $a_i \in [0, 1]$ without thresholding it into correct/incorrect labels. For ten references, one, two, and three matching answers yield 0.3, 0.6, and 0.9, respectively. For each confidence channel, we partition $[0, 1]$ into $M = 15$ equal-width bins: $I_m = [(m-1)/M, m/M)$ for $m < M$, and $I_M = [(M-1)/M, 1]$. Let $B_m = \{i : c_i \in I_m\}$ on the same N held-out complete pairs. Soft-target ECE replaces binary bin accuracy by mean VQA agreement:

$$\mathrm{ECE}_{\mathrm{soft}} = \sum_{m:|B_m|>0} \frac{|B_m|}{N} \left| \frac{1}{|B_m|} \sum_{i \in B_m} c_i - \frac{1}{|B_m|} \sum_{i \in B_m} a_i \right|. \tag{17}$$

Empty bins contribute zero, and both boundary scores $c_i = 0$ and $c_i = 1$ are included. This measures calibration to expected VQA task credit and reduces to ordinary ECE when the targets are binary. Table 10 reports this ECE and tie-corrected Kendall's τ_b against a_i, both multiplied by 100.

Table 11: Instinct–reflection discrepancy under visual corruption. Corr-L2 and Corr-L5 denote severity levels 2 and 5, and Corr-All aggregates all tested corruption levels.

Split	Mean $\lvert t - v \rvert$ (%)	Wrong–correct gap (pp)
Clean	50.1	+6.4
Corr-L2	47.7	+5.4
Corr-L5	53.4	+6.8
Corr-All	50.5	+6.1

Robustness under visual corruption. Table 11 examines whether instinct–reflection misalignment is merely an artifact of clean benchmark images. The mean absolute difference between token and verbalized confidence remains close to 50% across clean, level-2, level-5, and pooled corruption settings. More importantly, the discrepancy is consistently larger for wrong answers than for correct ones by 5.4–6.8 percentage points. Thus, visual degradation does not eliminate the dual-channel structure; instead, disagreement remains associated with model failure, supporting its use as a reliability signal under perceptual perturbations.

Table 12: Effect of increasing self-consistency samples on 17,118 aligned examples from four benchmarks and twelve model–dataset pairs.

Method	Calls/query	Relative cost	ACC (%)	Non-unanimous (%)	Flip vs. SC-5 (%)
Self-Consistency-5	5	5×	55.69	47.87	0.00
Self-Consistency-10	10	10×	56.35	55.74	16.64

Cost and instability of multi-sample self-consistency. Table 12 compares five and ten multimodal inference calls per query. Doubling the sampling cost improves accuracy by only 0.66 percentage points, while the fraction of non-unanimous predictions increases from 47.87% to 55.74%; moreover, 16.64% of predictions flip relative to Self-Consistency-5. These results show that additional multimodal samples do not simply converge to a stable answer and motivate the fixed-cost, single-pass confidence estimator used in our framework.

Table 13: Ablation of the top-2 margin feature. Values are average ECE ($\times 10^2$; lower is better).

Model	w/o margin	Top-2	Δ
InternVL2.5-4B	8.21	8.13	-0.08
InternVL3.5-8B	4.65	4.29	-0.36
Qwen2.5-VL-7B	6.21	5.75	-0.45
GPT-4o	7.32	7.13	-0.19
GPT-4o-mini	6.43	6.29	-0.14

Dataset	w/o	Top-2	Δ
ConBench	7.43	7.09	-0.34
MMMU-Pro	9.70	9.34	-0.35
MMStar	8.56	8.31	-0.25
ScienceQA	4.44	4.27	-0.17

C.6 Component Analysis and Robustness Checks

Contribution of the margin feature. Table 13 compares the full descriptor with a variant that removes the top-2 margin. Adding the margin reduces average ECE for every evaluated model, with improvements ranging from 0.08 to 0.45 points, and for every reported dataset, with improvements of 0.17–0.35 points. Although the absolute changes are moderate, their consistency shows that the separation between the predicted option and its strongest competitor provides information not fully captured by top-1 confidence or global entropy.

Table 14: Sensitivity to the rank k used in the margin feature. Metrics are averaged over all evaluated model–dataset pairs and reported on the $[0, 1]$ scale.

k	ACC	ECE	AUROC	AUPRC-N
2	0.79	0.0735	0.8101	0.6437
3	0.79	0.0738	0.8101	0.6441
4	0.79	0.0736	0.8100	0.6441

k	AI2D	ConB.	MMB.	MMMU	MMStar	SciQA
2	0.0495	0.0659	0.0298	0.1454	0.1085	0.0420
3	0.0497	0.0687	0.0288	0.1434	0.1104	0.0419
4	0.0498	0.0661	0.0290	0.1430	0.1096	0.0420

Sensitivity to the top-k choice. Table 14 varies the rank used by the margin feature over $k \in \{2, 3, 4\}$. All aggregate metrics remain nearly unchanged: ECE varies by at most 0.0003, AUROC by 0.0001, and AUPRC-N by 0.0004. Dataset-level differences are similarly small, and no value of k dominates across all benchmarks. We therefore retain $k = 2$ as a parsimonious and interpretable default because it measures ambiguity against the strongest competing option, while entropy already summarizes the broader probability distribution.

Table 15: Ablation of the RBF cross-channel consistency feature. Values are mean ECE ($\times 10^2$) across ten models.

Dataset	With RBF	Without RBF
ConBench	**7.21**	7.77
MMBench	**3.63**	3.96
MMStar	**8.02**	8.53
ScienceQA	**4.23**	4.58

Contribution of cross-channel consistency. Removing the RBF agreement feature increases mean ECE on every dataset in Table 15, with degradations of 0.33–0.56 points. The consistent reduction from the full model indicates that the absolute confidence levels alone do not completely describe reliability: whether instinct and reflection agree supplies an additional signal. This ablation also supports the nonlinear mapping in Eq.(3), which converts raw channel disagreement into a bounded feature that can be combined with the remaining confidence descriptors.

Unconstrained-head comparison. Table 16 reports a fusion variant without the Softplus nonnegativity constraint. Its values exceed the better raw confidence channel on all three ranking metrics for four of five models, while MiniCPM-V-2.6 has lower values. The reference channel is selected separately for each model and metric. These differences do not measure gains over the monotone head and therefore do not isolate the effect of removing the nonnegativity constraint. Such a comparison requires matched data splits, training settings, and evaluation examples. We retain the monotone formulation in the main method for its coordinatewise monotonicity.

Table 16: Unconstrained fusion-head ranking performance on MMMU-Pro. Values are $\times 10^2$. Parentheses report differences from the better raw token or verbal confidence channel for each model and metric in Table 1.

Model	AUROC↑	AUPRC↑	AUPRC-N↑
InternVL2.5-4B	84.43 (+6.72)	85.02 (+10.85)	67.76 (+23.33)
Qwen2.5-VL-7B	75.46 (+3.47)	42.11 (+8.35)	24.29 (+10.91)
MiniCPM-V-2.6	56.79 (−1.22)	47.59 (−1.92)	16.72 (−3.05)
GPT-4o	77.09 (+6.31)	69.19 (+4.42)	50.99 (+7.02)
GPT-4o-mini	70.60 (+3.19)	61.84 (+1.96)	41.79 (+2.99)

C.7 Detailed Prompting Results and Templates

Detailed results across prompting strategies. Tables 17–20 report task accuracy and verbalized-confidence quality for four open-source MLLMs under Self-Probing, Role-Play, Confidence Interval, and Chain-of-Thought (CoT) prompting, respectively. These tables evaluate the raw verbalized confidence elicited by each strategy rather than the calibrated dual-channel score. The results are strongly model- and benchmark-dependent. Self-Probing yields low ECE for MiniCPM-V-2.6 on MMBench and ScienceQA, but Qwen2.5-VL-7B remains substantially miscalibrated on MMStar and AI2D. Role-Play performs well for InternVL3.5-8B on ScienceQA, reaching an ECE of 3.83 and an AUROC of 76.70, while Phi-3.5-Vision retains high ECE on the more difficult AI2D and MathVista benchmarks. Confidence Interval prompting can improve confidence ranking in some settings, such as Qwen2.5-VL-7B on MMBench, but often leaves large calibration errors on challenging tasks. CoT similarly produces isolated gains, including an ECE of 1.70 for InternVL3.5-8B on ScienceQA, yet it can substantially weaken correctness ranking, as illustrated by Qwen2.5-VL-7B's AUROC of 19.55 on the same benchmark. Overall, no prompting strategy consistently improves both calibration and discrimination across models and datasets, reinforcing the need for post-hoc dual-channel calibration.

Table 17: Calibration and confidence-quality metrics for open-source MLLMs across six benchmarks (MMBench, MMStar, ConBench, ScienceQA, AI2D, and MathVista) under the self-probing verbalized confidence setting. Higher values indicate better performance for **ACC**↑, **AUROC**↑, **AUPRC**↑, and **AUPRC-N**↑, while lower values are better for **ECE**↓. Metrics are reported as $\times 10^2$.

	MMBench					MMStar					ConBench				
Model	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N
Open-source MLLMs															
MiniCPM-V-2.6 (Yao et al., 2024)	81.81	2.47	59.38	84.89	26.29	50.87	21.78	52.88	53.02	51.37	71.80	8.81	58.62	75.93	37.64
Qwen2.5-VL-7B (Bai et al., 2025b)	68.61	20.45	55.17	71.98	36.94	35.80	51.06	48.28	35.74	65.18	62.56	25.99	56.82	68.43	43.40
InternVL3.5-8B (Wang et al., 2025)	78.16	11.17	56.60	80.82	26.28	48.27	30.35	54.42	52.01	54.49	68.60	16.74	51.14	69.29	32.93
Phi-3.5-Vision (Abdin et al., 2024)	77.79	11.60	58.29	80.83	28.03	46.73	41.31	53.12	48.65	55.86	64.60	25.03	58.96	69.64	41.48

	ScienceQA					AI2D					MathVista				
Model	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N
Open-source MLLMs															
MiniCPM-V-2.6 (Yao et al., 2024)	86.88	2.65	59.99	89.38	18.13	52.78	12.12	54.60	55.45	51.29	40.66	27.17	52.12	42.23	61.07
Qwen2.5-VL-7B (Bai et al., 2025b)	58.18	39.18	35.34	60.58	46.05	40.06	48.25	47.72	40.03	59.99	37.36	47.73	45.91	35.76	61.90
InternVL3.5-8B (Wang et al., 2025)	86.36	10.80	54.13	88.39	14.65	55.15	23.46	58.96	61.29	51.08	52.75	21.71	49.06	51.95	48.39
Phi-3.5-Vision (Abdin et al., 2024)	70.52	22.99	55.74	73.99	32.30	44.33	43.23	50.30	44.52	56.35	35.53	51.32	47.58	34.72	63.71

Table 18: Calibration and confidence-quality metrics for open-source MLLMs across six benchmarks (MMBench, MMStar, ConBench, ScienceQA, AI2D, and MathVista) under the role-play verbalized confidence setting. Higher values indicate better performance for **ACC**↑, **AUROC**↑, **AUPRC**↑, and **AUPRC-N**↑, while lower values are better for **ECE**↓. Metrics are reported as $\times 10^2$.

	MMBench					MMStar					ConBench				
Model	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N
Open-source MLLMs															
MiniCPM-V-2.6 (Yao et al., 2024)	81.15	3.31	59.48	84.33	25.84	51.53	26.65	50.92	52.38	48.98	72.00	7.23	57.40	76.15	32.08
Qwen2.5-VL-7B (Bai et al., 2025b)	80.40	6.69	56.71	82.63	22.58	50.13	35.00	53.99	53.09	52.02	74.70	11.03	56.97	77.80	31.95
InternVL3.5-8B (Wang et al., 2025)	78.50	9.03	63.31	84.07	29.10	48.40	32.07	56.04	52.59	55.99	70.80	15.64	64.24	78.02	41.08
Phi-3.5-Vision (Abdin et al., 2024)	79.32	11.20	59.62	83.03	25.63	46.13	41.89	52.98	47.97	55.49	66.10	23.49	61.23	72.50	41.17

	ScienceQA					AI2D					MathVista				
Model	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N
Open-source MLLMs															
MiniCPM-V-2.6 (Yao et al., 2024)	79.87	11.00	46.18	79.20	26.35	45.95	32.02	48.16	45.11	55.88	41.03	34.08	51.36	41.21	62.68
Qwen2.5-VL-7B (Bai et al., 2025b)	70.00	18.53	44.84	68.15	29.05	53.21	29.28	53.74	55.35	51.32	56.04	26.90	49.59	56.19	44.72
InternVL3.5-8B (Wang et al., 2025)	87.66	3.83	76.70	94.20	40.30	58.68	24.16	61.79	66.11	51.57	49.08	29.09	58.35	53.31	60.34
Phi-3.5-Vision (Abdin et al., 2024)	55.19	32.06	69.39	70.44	58.27	43.04	39.71	55.63	47.19	60.35	38.46	47.32	55.12	44.17	64.39

Table 19: Calibration and confidence-quality metrics for open-source MLLMs across six benchmarks (MMBench, MMStar, ConBench, ScienceQA, AI2D, and MathVista) under the confidence interval verbalized confidence setting. Higher values indicate better performance for **ACC**↑, **AUROC**↑, **AUPRC**↑, and **AUPRC-N**↑, while lower values are better for **ECE**↓. Metrics are reported as $\times 10^2$.

	MMBench					MMStar					ConBench				
Model	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N
Open-source MLLMs															
MiniCPM-V-2.6 (Yao et al., 2024)	83.14	5.61	69.74	90.53	29.02	52.40	21.80	57.16	59.55	51.95	74.90	4.24	60.93	79.77	35.17
Qwen2.5-VL-7B (Bai et al., 2025b)	73.77	16.84	75.90	86.82	48.22	44.33	41.43	62.60	55.79	63.87	71.30	18.22	69.61	81.72	42.50
InternVL3.5-8B (Wang et al., 2025)	78.14	9.31	60.83	82.91	29.84	48.47	33.89	58.25	55.60	58.18	66.20	21.49	63.27	74.40	44.87
Phi-3.5-Vision (Abdin et al., 2024)	79.30	14.38	60.04	83.76	27.02	48.00	42.48	58.30	57.57	58.30	66.50	28.55	57.76	70.91	39.75

	ScienceQA					AI2D					MathVista				
Model	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N
Open-source MLLMs															
MiniCPM-V-2.6 (Yao et al., 2024)	72.86	12.61	53.24	73.32	36.91	47.44	23.98	53.25	49.61	56.68	60.81	15.08	55.08	65.63	42.92
Qwen2.5-VL-7B (Bai et al., 2025b)	66.23	23.94	72.47	83.61	50.46	56.77	29.00	61.37	67.95	50.93	37.00	48.55	53.32	40.04	65.72
InternVL3.5-8B (Wang et al., 2025)	87.40	5.16	58.44	89.20	22.76	58.87	25.75	59.87	66.28	49.88	47.99	32.41	50.76	47.34	57.41
Phi-3.5-Vision (Abdin et al., 2024)	48.57	43.65	61.82	56.54	60.41	39.38	50.70	55.89	46.49	64.03	49.08	37.32	49.29	51.09	52.47

Table 20: Calibration and confidence-quality metrics for open-source MLLMs across six benchmarks (MMBench, MMStar, ConBench, ScienceQA, AI2D, and MathVista) under the CoT verbalized confidence setting. Higher values indicate better performance for **ACC**↑, **AUROC**↑, **AUPRC**↑, and **AUPRC-N**↑, while lower values are better for **ECE**↓. Metrics are reported as $\times 10^2$.

	MMBench					MMStar					ConBench				
Model	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N
Open-source MLLMs															
MiniCPM-V-2.6 (Yao et al., 2024)	81.81	3.60	63.10	86.52	25.91	51.33	28.59	54.53	55.03	51.90	72.00	11.45	61.19	78.00	34.49
Qwen2.5-VL-7B (Bai et al., 2025b)	78.98	6.99	53.63	80.30	22.52	48.87	34.85	50.29	50.13	51.83	72.90	11.87	50.44	73.19	28.22
InternVL3.5-8B (Wang et al., 2025)	77.63	10.61	61.89	81.65	32.81	49.80	31.59	54.46	51.90	54.41	70.60	17.74	59.05	73.91	38.58
Phi-3.5-Vision (Abdin et al., 2024)	79.32	12.02	57.69	82.66	23.99	46.27	42.95	49.20	46.16	53.28	65.80	25.83	58.20	71.18	38.53

	ScienceQA					AI2D					MathVista				
Model	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N	ACC	ECE	AUROC	AUPRC	AUPRC-N
Open-source MLLMs															
MiniCPM-V-2.6 (Yao et al., 2024)	78.57	9.08	53.25	80.78	27.63	46.89	29.85	55.82	52.05	57.98	42.86	34.98	49.75	44.08	59.60
Qwen2.5-VL-7B (Bai et al., 2025b)	72.34	25.90	19.55	66.75	23.93	49.06	31.31	43.10	47.17	51.25	51.28	30.35	42.97	47.84	46.31
InternVL3.5-8B (Wang et al., 2025)	89.74	1.70	61.21	90.93	27.50	58.81	24.48	61.16	65.20	53.29	54.58	24.74	48.92	52.76	47.61
Phi-3.5-Vision (Abdin et al., 2024)	49.61	31.73	64.86	58.95	63.44	42.00	41.80	54.02	44.30	61.02	49.45	39.15	44.28	47.59	47.85

Prompt for vanilla

You are a highly capable AI assistant. Carefully read the question and the attached image. Then answer in a strictly **JSON format** with no additional text or explanation outside the JSON. Your response MUST follow exactly this structure (with no extra keys, no code fences, no markdown formatting):

```
{  "message": [
    { "answer": "A", "confidence": "XX" },
    { "answer": "B", "confidence": "XX" },
    { "answer": "C", "confidence": "XX" },
    { "answer": "D", "confidence": "XX" },
    { "chosen_answer": "X" }  ] }
```

Important rules (extended):

1. "confidence" must be an integer in the range 0 to 100 (no extra symbols or text).
2. "chosen_answer" must be one of "A", "B", "C", or "D".
3. Do not include any explanation or text outside of the JSON object.
4. Do not wrap your JSON inside code fences or add any prefixes like "Here is your answer:".
5. No additional keys are allowed. Only the exact structure shown above.
6. Please ensure that the "confidence" values are output in the exact order corresponding to the options provided in the question. For example, if the question has N options (such as A, B, C, etc.), then the first "confidence" value must correspond to the first option, the second to the second option, and so on. Do not change or shuffle this order.
7. When you assign the "confidence" value for each option, carefully consider:
   - The difficulty or complexity of the question,
   - The availability (or lack) of relevant knowledge,
   - Any ambiguity in the prompt or the attached image,
   - Other potential sources of uncertainty (e.g., incomplete reasoning).

   Your "confidence" should reflect how likely you believe that specific option is actually correct, from 0 (very unsure) to 100 (absolutely certain).

Remember: Output must **ONLY be the JSON**.
Now, here is the question you need to answer using the picture:
{query}
{choices}

Figure 16: Prompt templates for verbalized confidence strategy.

C.8 Confidence Case Studies

As illustrated in Figures 19 and 21, a distinct Verbal-Internal Disconnect (VID) pervades MLLMs, where implicit token confidence conceptually diverges from explicit verbalized confidence. To elucidate this dual-channel discrepancy, we dissect representative samples, systematizing them into four distinct groups defined by the intersection of prediction correctness and confidence alignment.

(1) Well-Calibrated Correct (WCC): This is the best scenario where the model answers correctly and is also very sure about it. As shown in Figure 25 and Figure 26, the model picks the right option. More importantly, both confidence channels favor the correct option, with high verbalized confidence. This shows that the model trusts its own answer, and users can trust it too.

(2) Under-Confident Correct (UCC): In these cases, the model implicitly identifies the correct answer but lacks sufficient verbal certainty. For example, in Figure 21 and Figure 22, the model selects the right choice. The correct option receives the highest token confidence among the candidates. However, the verbal confidence is surprisingly low (less than 60% or even 25%). This happens

Prompt for topk

Please read the question carefully and output your best 4 guesses for the correct choice (A/B/C/D), from the most likely (G1) to the least likely (G4). For each guess, give an integer probability (P1..P4)
indicating how likely it is to be correct (0 to 100).
Important:

- Only output 8 lines in the exact pattern:
  G1: X
  P1: N
  G2: X
  P2: N
  G3: X
  P3: N
  G4: X
  P4: N
- Where 'X' is one of [A, B, C, D], and 'N' is an integer [0..100].
- Do not include any explanation or any extra text besides those 8 lines.

Prompt for self-probing

Carefully read the question and any attached image. After reviewing the content, use self-probing to evaluate each of the four options (A, B, C, D) and determine how likely each is to be the correct answer. This process involves internally assessing the options, but the reasoning itself should not be included in your final output. For each option, assign a confidence level between 0 and 100 to represent how likely it is to be correct.

Once you've evaluated all the options, choose the one that you believe is most likely to be correct and mark it as the "chosen_answer."

Your response should only include the final confidence levels for each option and the selected answer. Make sure your confidence for each option is a number between 0 and 100, and that the "chosen_answer" is one of the options (A, B, C, or D). Avoid any additional explanations or text outside the required structure.

Figure 17: Prompt templates for top-K and self-probing prompting strategies.

Prompt for role-play

You are a cautious and humble expert.

Although you have considerable knowledge, carefully evaluate the difficulty, missing information, and potential ambiguities in the question. When assessing your confidence for each choice, avoid overestimating your certainty. If you are uncertain, reflect that in lower or moderate confidence values rather than being overly confident.

Now, carefully read the question and any attached image. After analyzing the content, assign a confidence level to each of the four options (A, B, C, D). Be mindful that, as a cautious expert, you should only provide confidence values that reflect your careful evaluation, with 100% confidence being rare.

Finally, choose the option that you believe is most likely correct, marking it as the "chosen_answer."

Prompt for confidence interval

Carefully read the question and any attached image. After evaluating the content, for each of the four options (A, B, C, D), provide a confidence range that reflects your belief in the correctness of each answer. The confidence range should include both a lower and an upper value, with the lower bound being less than or equal to the upper bound.

After assigning the confidence intervals for each option, choose the option that seems most likely to be correct and identify it as the "chosen_answer."

Ensure that your response includes only the confidence intervals for each option and the final chosen answer, with no extra explanations or text.

Prompt for chain of thought

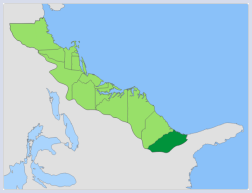
Carefully read the question and any attached image. Analyze the question step by step, considering each option (A, B, C, D), and evaluate your confidence in the answer for each one.

As you proceed through the reasoning, determine your confidence in each option, taking into account any difficulties or ambiguities. Once you have completed your analysis, choose the most likely answer and mark it as the "chosen_answer."

Your response should only include the confidence values for each option (A, B, C, D) and the "chosen_answer." Avoid any extra text, explanations, or symbols.

Figure 18: Prompt templates for different role-play, confidence interval, and COT confidence strategies.

Category: Geography
Data Source: ScienceQA



Question: What is the name of the colony shown?
Ground Truth: A
Chosen Answer: A

Confidence Comparison

Option	Verbal Conf.	Token Conf.
A: Georgia	1%	**48.65%**
B: New Hampshire	0%	5.41%
C: South Carolina	**99%**	43.24%
D: West Virginia	0%	2.70%

Analysis of Discrepancy:
This presents an instance of **Verbal-Internal Disconnect (VID)**, where Token Confidence accurately identifies the correct answer A (48.65%), whereas Verbal Confidence hallucinates the incorrect option C (99%).

Figure 19: An illustration of Verbal-Internal Disconnect (VID) in MLLMs during a geography-based QA task.

because the model gets distracted by other competing options when generating the final text, failing to properly express its true internal knowledge.

(3) Over-Confident Wrong (OCW): This is the most dangerous situation where the model makes a mistake but claims to be completely correct. As seen in Figure 23 and Figure 24, the model chooses the wrong answer. Of particular concern is that the verbal confidence is extremely high (often 100%). Even if the internal token confidence shows some doubt or supports another option, the final output sounds fully confident. This misleads users into thinking the wrong answer is true.

(4) Well-Calibrated Wrong (WCW): Finally, Figure 27 and Figure 28 show a safer type of failure. Here, the model gives a wrong answer, but it explicitly admits that it is not sure. The verbal confidence is appropriately low (below 50%), which matches the fact that the answer is incorrect. Although the model might still have some internal confusion, the low confidence score effectively warns the user to exercise caution, signaling that the generated answer might be unreliable.

Category: General Knowledge
Data Source: MMStar



Question: Which option describe the object relationship in the image correctly?
Ground Truth: A
Chosen Answer: A

Confidence Comparison

Option	Verbal Conf.	Token Conf.
A: The suitcase is on the book.	35%	**80%**
B: The suitcase is beneath the cat.	0%	0%
C: The suitcase is beneath the bed.	0%	0%
D: The suitcase is beneath the book.	**65%**	20%

Analysis of Discrepancy:

This presents an instance of **Verbal-Internal Disconnect**, where Token Confidence accurately identifies the correct answer A (**80%**), whereas Verbal Confidence hallucinates the incorrect option D (**65%**).

Figure 20: An illustration of Verbal-Internal Disconnect (VID) in MLLMs during a general knowledge QA task.

Category: General Knowledge
Data Source: MMStar



Question: What seems to be the theme of the image?
Ground Truth: D
Chosen Answer: D

Confidence Comparison

Option	Verbal Conf.	Token Conf.
A: Hanging Posters	0%	26.67%
B: Music	41%	23.33%
C: Home decoration	0%	20.0%
D: Playing guitar	**59%**	**30.0%**

Analysis of Discrepancy:

This case demonstrates **Under-Confident Correct (UCC)**. While the model answers correctly with high internal certainty (Token Conf: **30.0%**), the verbal output exhibits significant hesitation (Verbal Conf: **59%**) due to the distraction from option B (41%).

Figure 21: An illustration of Under-Confident Correct(UCC) in MLLMs during a general knowledge QA task.

Category: Public Health
Data Source: MMMU-Pro

Area	Number of imported cases	Number of susceptible contacts	Number of local cases
Fuzhou	74	409	13
Xiamen	25	138	1#
Quanzhou	5	28	1#
Putian	3	17	0
Total	107	592	15

Cases imported from outside the province

Question: In March 2009, an outbreak of Influenza A (H1N1) occurred... The following table showed the regional distribution... The secondary attack rate of H1N1 influenza in Fuzhou was...

Ground Truth: A

Chosen Answer: A

Confidence Comparison

Option	Verbal Conf.	Token Conf.
A: 3.18%	25%	**30.0%**
B: 0.00%	25%	20.0%
C: 2.20%	25%	26.6%
D: 1.50%	25%	23.33%

Analysis of Discrepancy:

This case represents **Under-Confident Correct (UCC)**. The model answers correctly with high internal certainty (Token Conf: **30.0%**). However, the verbal output collapses into a uniform distribution (25% each), failing to express the model's actual internal knowledge.

Figure 22: An illustration of Under-Confident Correct(UCC) in MLLMs during a public health QA task.

Category: Geography
Data Source: ScienceQA



Question: Which country is highlighted?

Ground Truth: A

Chosen Answer: C

Confidence Comparison

Option	Verbal Conf.	Token Conf.
A: Palau	0%	33.33%
B: Nauru	0%	25.00%
C: The Federated States of Micronesia	**100%**	**37.5%**
D: New Zealand	0%	4.17%

Analysis of Discrepancy:

This presents an instance of **Over-Confidence Wrong (OCW)**, where Token Confidence captures the competitive state between options A and C, whereas Verbal Confidence exhibits a complete failure.

Figure 23: An illustration of Over-Confident Wrong(OCW) in MLLMs during a geography QA task.

Category: General Knowledge
Data Source: MMStar



Question: What is the center of focus in the image?
Ground Truth: A
Chosen Answer: B

Confidence Comparison

Option	Verbal Conf.	Token Conf.
A: A man writing in a book.	0%	25.93%
B: A boy with his head in his hands surrounded by books.	**100%**	**33.33%**
C: A cluttered desk with books and a pen.	0%	22.22%
D: A stack of books covering a young man's face.	0%	18.52%

Analysis of Discrepancy:

This case represents **Over-Confident Wrong (OCW)**. The model provides an incorrect answer (B) with absolute verbal certainty (**100%**). Although the internal Token Confidence for the correct answer A was relatively high (25.93%), it was overshadowed by the incorrect option B (33.33%), leading to a high-confidence failure.

Figure 24: An illustration of Over-Confident Wrong(OCW) in MLLMs during a general knowledge QA task.

Category: Literature
Data Source: MMMU-Pro



Question: Refer to the figure, which term refers to the materials and technical means used by artists to create pictures?
Ground Truth: A
Chosen Answer: A

Confidence Comparison

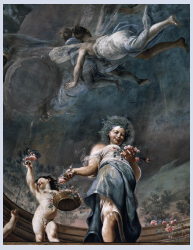
Option	Verbal Conf.	Token Conf.
A: Artistic media	**100%**	**34.62%**
B: Artistic elements	0%	23.09%
C: Functions of illustrations	0%	15.38%
D: Visual elements	0%	26.92%

Analysis of Discrepancy:

This case represents **Well-Calibrated Correct (WCC)**. The model correctly identifies the answer (A) and exhibits high certainty in both Verbal (**100%**) and Token (**34.62%**) metrics. The model is confident and accurate, showing ideal calibration.

Figure 25: An illustration of Well-Calibrated Correct(WCC) in MLLMs during a literary QA task.

Category: Arts
Data Source: ConBench



Question: In which form does this artwork exist?

Ground Truth: B
Chosen Answer: B

Confidence Comparison

Option	Verbal Conf.	Token Conf.
A: Sculpture	0%	16%
B: Painting	**100%**	**72%**
C: Digital Art	0%	8%
D: Performance Art	0%	4%

Analysis of Discrepancy:

This case represents **Well-Calibrated Correct (WCC)**. The model correctly identifies the artwork form (B) and maintains high confidence in both generative output (Verbal: **100%**) and internal representation (Token: **72%**), showing ideal alignment.

Figure 26: An illustration of Well-Calibrated Correct(WCC) in MLLMs during a artistic QA task.

Category: Arts
Data Source: ConBench



Question: Which museum displays this artwork?
A) Rijksmuseum, Amsterdam
B) Stedelijk Museum de Lakenhal, Leiden
C) Van Gogh Museum, Amsterdam
D) Mauritshuis, The Hague.

Ground Truth: B
Chosen Answer: D

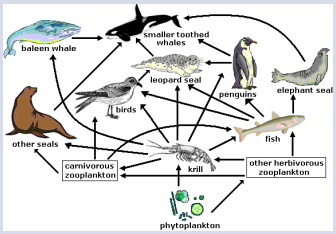
Confidence Comparison

Option	Verbal Conf.	Token Conf.
A: Rijksmuseum, Amsterdam	4%	20%
B: Stedelijk Museum de Lakenhal, Leiden	36%	30%
C: Van Gogh Museum, Amsterdam	0%	10%
D: Mauritshuis, The Hague	**59%**	**40%**

Analysis of Discrepancy:
This case represents **Well-Calibrated Wrong (WCW)**. The model incorrectly selects D, but exhibits significant uncertainty (Token Conf: **40%**, Verbal Conf: **59%**). Notably, it assigns a substantial probability to the correct answer B (Token: 30%, Verbal: 36%), reflecting genuine ambiguity rather than confident hallucination.

Figure 27: An illustration of Well-Calibrated Wrong(WCW) in MLLMs during a artistic QA task.

Category: Science / Ecology
Data Source: AI2D



Question: The diagram below shows a food chain. If the smaller toothed whales ate all the leopard seals, the population of krills would most likely
Ground Truth: C
Chosen Answer: B

Confidence Comparison

Option	Verbal Conf.	Token Conf.
A: remain the same	27%	20%
B: decrease	**44%**	**70%**
C: increase	27%	10%
D: NA	3%	0%

Analysis of Discrepancy:
This case represents **Well-Calibrated Wrong (WCW)**. Although the model selects the incorrect answer B, its Verbal Confidence is low (**44%**), indicating significant uncertainty. While the internal Token Confidence (70%) suggests a stronger latent misconception, the explicit verbal output successfully conveys hesitation, warning the user of potential error.

Figure 28: An illustration of Well-Calibrated Wrong(WCW) in MLLMs during a ecological QA task.