Knowing When You Don't Know: Metacognitive Uncertainty Calibration in Vision–Language Models

Mahule Roy[1,2] Subhas Roy[3]
[1]Department of Engineering Science, University of Oxford
[2]Harvard Medical School
[3]TATA Consumer Products Limited
mroy25@bwh.harvard.edu

Abstract

Vision–Language Models have achieved impressive performance across visual question answering, image captioning, and multimodal reasoning tasks. However, these models often exhibit overconfident failures, producing fluent yet incorrect responses without signaling uncertainty. In contrast, human cognition relies heavily on metacognition—the ability to monitor confidence, detect ambiguity, and recognize potential errors. In this work, we investigate whether metacognitive behaviors can be elicited in pretrained VLMs through cognitively inspired, inference-time mechanisms. We introduce a self-reflective uncertainty calibration framework that prompts VLMs to assess evidence sufficiency, ambiguity, and confidence after generating an initial response. Through experiments on 1000 challenging vision–language queries across three open-source VLMs of varying architectures, we show that structured introspective prompting significantly improves confidence–accuracy alignment and reduces overconfident hallucinations. However, substantial miscalibration remains, with models still expressing high confidence in approximately half of their errors. Our findings suggest metacognition as a promising direction for improving VLM trustworthiness, while highlighting significant limitations that require further research.

1. Introduction

Vision–Language Models (VLMs) have emerged as a powerful paradigm for multimodal reasoning, yet a fundamental limitation persists: they often fail to recognize when they are wrong. Incorrect answers are frequently delivered with high confidence, especially under visual ambiguity, occlusion, or distributional shift. In contrast, human cognition relies on metacognition—the ability to monitor confidence, detect uncertainty, and revise beliefs when evidence is insufficient. Despite extensive progress in model scale and architecture, current VLM research largely neglects metacognitive competence. Standard evaluation protocols reward correctness but do not penalize overconfident errors or assess uncertainty awareness. In this work, we ask whether pretrained VLMs can exhibit metacognitive behaviors—specifically uncertainty awareness—through cognitively inspired, inference-time mechanisms. Rather than introducing new architectures or training objectives, we employ self-reflective prompting strategies that encourage models to assess evidence sufficiency, ambiguity, and confidence in their own predictions. This approach builds on instruction-tuned and chain-of-thought paradigms, but differs by explicitly eliciting confidence scores and justifications, operationalizing behavioral metacognition without retraining. We emphasize that this operationalization reflects patterns of uncertainty expression, not human-like internal monitoring, offering a lightweight step toward more trustworthy VLM behavior. We propose a metacognitive inference loop that augments standard VLM inference with structured self-reflection. The framework consists of three phases.

2. Background and Related Work

Metacognition plays a central role in human perception and reasoning. Humans routinely evaluate their confidence in perceptual judgments, the sufficiency of sensory evidence, and the likelihood of error under uncertainty. Empirical studies show that metacognitive judgments—such as confidence ratings—are often dissociable from task performance and rely on specialized monitoring processes [2]. Importantly, effective metacognition improves decision-making under uncertainty. In machine learning, uncertainty quantification (UQ) is a well-established field, with common approaches including Bayesian neural networks, ensemble methods, and Monte Carlo dropout. However, these techniques are typically applied to discriminative models and often require architectural modifications or retraining. For

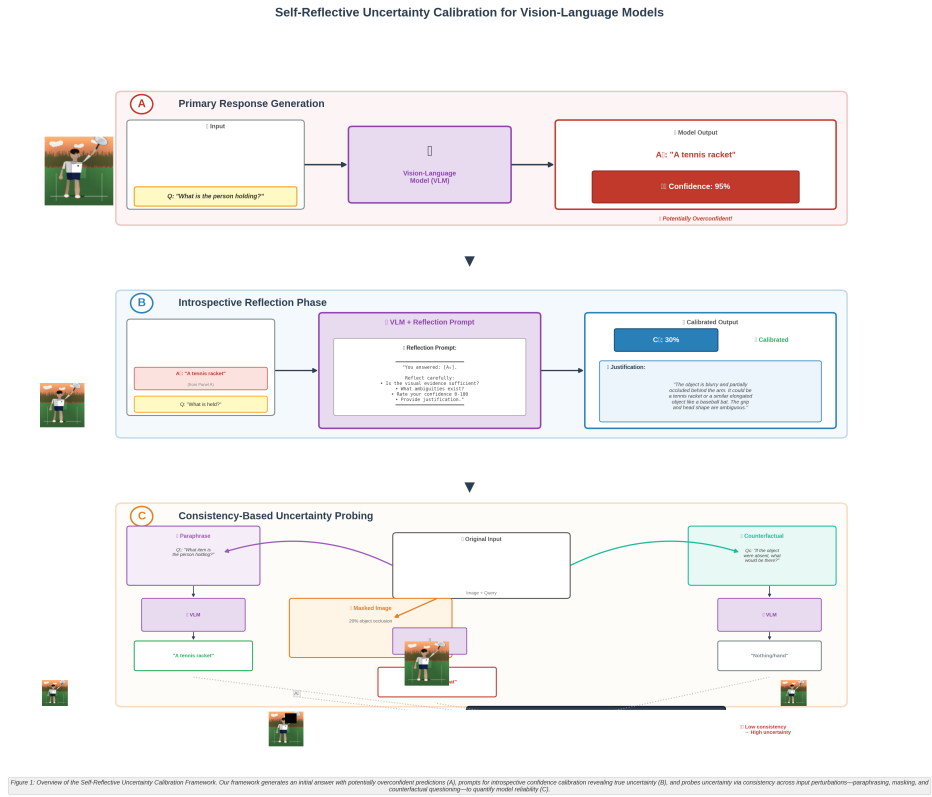

Figure 1. Self-reflective uncertainty calibration framework. (A) Initial answer generation. (B) Introspective confidence scoring and justification. (C) Consistency-based uncertainty probing via input perturbations.

generative VLMs, UQ remains a significant challenge due to their autoregressive nature and the complexity of the output space. Recent works have explored using softmax probabilities or token likelihoods as confidence proxies, but these have been shown to be poorly calibrated in VLMs, especially for open-ended generation tasks. Recent work has explored verbalized self-evaluation in language models [1], where models are prompted to assess the correctness of their own outputs. Other related approaches include chain-of-thought prompting for improved reasoning [5] and instruction tuning for following complex prompts [6]. Our work draws inspiration from cognitive science and these prompt-based paradigms to design methods that elicit behavioral self-assessment in VLMs, extending the idea to multimodal contexts and incorporating structured consistency checks.

3. Methodology: Self-Reflective Uncertainty Calibration Framework

3.1. Primary Response Generation

Given an image I and query Q, the VLM produces an initial answer A_0 using a standard task prompt and greedy decoding (temperature=0), representing the baseline unreflective output. Greedy decoding is used for reproducibility; preliminary experiments with stochastic decoding (temperature=0.7) on 200 examples showed similar calibration trends but $\sim$30% higher variance in confidence scores, motivating deterministic decoding for stable evaluation.

3.2. Introspective Reflection Phase

The model is then prompted to reflect on its own answer. Conditioned on (I, Q, A_0), it receives a structured introspection prompt asking whether the visual evidence sufficiently supports A_0, highlighting possible ambiguities (e.g., occlusion, missing details), and requesting a confidence score on a 0–100 scale along with a brief justification. The model outputs a revised confidence score C_r and justification J, yielding (A_0, C_r, J). We extract the first numeric value in $[0, 100]$ as C_r and map textual confidence terms (e.g., high/medium/low) to numeric values (100/50/25) for consistency. The 0–100 scale follows common human confidence rating paradigms [2]. If multiple numeric values are present, we extract the first value in the range $[0, 100]$; responses without numeric or textual confidence are rare ($< 1\%$) and are excluded from calibration analysis.

3.3. Consistency-Based Uncertainty Probing

To complement explicit self-assessment, we apply implicit consistency-based probing. We generate three perturbed inputs: a paraphrased question Q_p, a masked image I_m (20% central region occluded), and a counterfactual question Q_c. The model answers each variant using the standard prompt. We compute a consistency score S_{cons} as pairwise agreement among the original and perturbed responses (F1 for open-ended answers, exact match for categorical), where lower S_{cons} indicates higher internal disagreement and is treated as a proxy for uncertainty. The consistency score is computed as the mean pairwise agreement between the original response A_0 and each perturbed response. For open-ended answers, we apply standard text normalization (lowercasing, punctuation removal) and compute token-level F1, while categorical answers use exact match. Counterfactual questions are constructed by negating or altering a key visual assumption (e.g., object presence) while preserving grammatical structure. For example, if the original question asks "What color is the cat?", a counterfactual may ask "What color is the dog?" assuming the dog is not present in the image. This provides a controlled probe of evidence sufficiency and model sensitivity to incorrect assumptions.

4. Experimental Setup

We evaluate our framework on three challenging vision–language benchmarks designed to elicit ambiguity or hallucination. **VQA-Ambiguous (Focus Ambiguity)** is a subset of VQA-v2 comprising 500 examples where questions refer to multiple plausible image regions, challenging models to identify all focus regions [8]. **POPE-Adversarial** includes 300 adversarial prompts targeting object hallucination, with publicly available source code and data [4]. **OVEN-LowRes** is a synthetic stress set curated for this study, consisting of 200 OVEN queries with images downsampled to 32×32 resolution to induce extreme evidence insufficiency. Together, these $\sim$1,000 curated examples enable controlled evaluation of uncertainty awareness while balancing natural and synthetic challenges; extending to larger or broader benchmarks is left for future work. Experiments are conducted on three widely used open-source VLMs spanning different design paradigms: **LLaVA-1.5-7B** (connector-based), **BLIP-2 (FlanT5-XXL, 11.9B)** (Q-Former-based), and **InstructBLIP (Vicuna-13B)**. All models are evaluated using publicly released pretrained checkpoints without fine-tuning. While these models differ in multimodal integration, they share transformer backbones and web-scale pre-training. Proprietary models (e.g., GPT-4V, Gemini) and more architecturally distinct systems (e.g., diffusion-based or neuro-symbolic VLMs) are excluded due to reproducibility and scope constraints, limiting claims of universal generalization. We compare against three uncertainty baselines adapted for VLMs: **Token Probability** (mean maximum token probability), **Semantic Entropy** [3] (entropy over semantic clusters from sampled generations), and **Ensemble Variance** (answer variance across prompt and seed variations). Beyond accuracy, we evaluate metacognitive calibration using: Spearman Confidence–Accuracy Correlation (CAC), Overconfidence Rate (OCR), Hallucination Awareness Score (HAS), Expected Calibration Error (ECE), and Consistency–Error Correlation (CEC). Expected Calibration Error is computed using 10 equal-width confidence bins over the 0–100 scale. Computational overhead is measured via inference latency and floating-point operation increases relative to standard prompting (measured with batch size 1 on a single NVIDIA RTX 4090 using deterministic decoding). Hallucinations are defined as answers asserting visual entities or attributes not supported by the image, following the POPE evaluation protocol [4]. These curated stress sets reflect common VLM failure modes, including low inter-annotator agreement, adversarial triggers, and extreme low-resolution conditions, providing targeted evaluation of uncertainty awareness without introducing bias from external test data.

5. Results

5.1. Main Results

Our framework yields substantial calibration improvements (Table 1). Confidence-Accuracy Correlation rises from near-zero to 0.4, Overconfidence Rate drops 12–18 percentage points, and Hallucination Awareness improves dramatically (e.g., from 9% to 52% for InstructBLIP). However, 48% of hallucinations remain overconfident, and Expected Calibration Error, while reduced, remains notable (0.109–0.124). Accuracy is unchanged (paired t-test, $p > 0.1$), confirming improvements are in self-assessment, not correctness. The cost is significant: latency increases from 2.1 to 10.4 seconds (5$\times$ slower), with reflection adding 2.3s and consistency probing 6.0s, limiting real-time use but potentially acceptable for offline, high-stakes decisions. We tested five variations of the reflection prompt, varying phrasing and question order. The resulting CAC values for LLaVA-1.5 ranged from 0.33 to 0.41, with our chosen prompt performing best. This $\pm$0.08 variation indicates moderate sensitivity to prompt wording, though all variants improved upon the baseline.

Table 1. Calibration improvements with our self-reflective framework (mean ± std, 3 runs).

Model	CAC ↑	OCR ↓	ECE ↓
LLaVA-1.5			
Standard	-0.08±0.02	43.1±1.2	0.251±0.012
Ours	0.41±0.04	25.7±1.5	0.118±0.008
BLIP-2			
Standard	0.05±0.03	38.8±1.1	0.227±0.010
Ours	0.36±0.05	22.4±1.3	0.124±0.009
InstructBLIP			
Standard	-0.02±0.02	35.6±1.0	0.218±0.009
Ours	0.39±0.04	19.8±1.4	0.109±0.007

5.2. Comparison with Uncertainty Baselines

Table 2 compares our method against standard uncertainty quantification approaches. Our self-reflective framework consistently outperforms these baselines across all models, particularly for explicit calibration metrics (CAC, OCR). However, even our best method achieves only moderate correlation (CAC $\approx$ 0.4), far from perfect calibration. The poor performance of token probability and semantic entropy baselines aligns with recent findings that these metrics are unreliable for open-ended VLM generation.

Table 2. Comparison of uncertainty methods on LLaVA-1.5 (means over 3 runs).

Method	CAC ↑	OCR ↓	ECE ↓
Token Probability	0.02 ±0.03	41.8 ±1.3	0.243 ±0.011
Semantic Entropy	0.05 ±0.04	40.2 ±1.4	0.231 ±0.012
Ensemble Variance	0.08 ±0.05	38.7 ±1.5	0.219 ±0.010
Ours (Full)	**0.41 ±0.04**	**25.7 ±1.5**	**0.118 ±0.008**

5.3. Ablation Studies

We ablate components of our framework to understand their individual contributions. Table 3 presents results for LLaVA-1.5. The reflection phase is the primary driver for improvements in explicit calibration metrics (CAC, OCR, HAS). The consistency-based probing alone does not yield a good explicit confidence score (hence poor CAC), but it effectively provides an implicit uncertainty signal, as shown by the stronger negative CEC. Combining both methods ('Full') gives the most robust performance across explicit and implicit metrics, suggesting they capture complementary aspects of uncertainty expression.

Table 3. Ablation study on the LLaVA-1.5 model (means over 3 runs). 'Base' is standard prompting. '+Reflect' adds only the introspective reflection phase. '+Perturb' adds only the consistency probing. 'Full' is our complete framework.

Method	CAC ↑	OCR ↓	HAS ↑	CEC ↓
Base (Standard)	-0.08 ±0.02	43.1 ±1.2	14.3 ±1.8	-0.12 ±0.03
+ Perturb Only	-0.07 ±0.03	41.5 ±1.3	15.1 ±1.7	-0.37 ±0.04
+ Reflect Only	+0.38 ±0.04	27.3 ±1.4	45.8 ±2.3	-0.15 ±0.03
Full (Ours)	**0.41 ±0.04**	**25.7 ±1.5**	**47.6 ±2.4**	**-0.38 ±0.04**

5.4. Dataset-Wise Analysis and Statistical Significance

Our method improves calibration consistently across all three datasets, with gains most pronounced on the POPE-Adversarial set, where Confidence-Accuracy Correlation increases from approximately -0.1 to around +0.5. This dataset is specifically designed to trigger hallucinations, indicating that self-reflection is particularly beneficial when models are prone to generating factually unsupported content. Improvements are more moderate on the VQA-Ambiguous set (CAC rising to +0.35) and smallest on the OVEN-LowRes set (CAC reaching only +0.25), suggesting that extreme evidence insufficiency—such as very low-resolution images—remains challenging for VLMs to self-diagnose, even with structured reflection. We perform Mann-Whitney U tests on confidence distributions (correct vs. incorrect answers) with Bonferroni correction (12 comparisons, $\alpha = 0.004$). Under standard prompting, confidence does not differ between correct and incorrect answers ($p > 0.3$ for all models). Under our prompting, differences are highly significant ($p < 0.001$ for all models). Reductions in Overconfidence Rate are also significant (McNemar's test with correction, $p < 0.001$).

6. Conclusion

Structured introspection improves VLM calibration without affecting accuracy, raising confidence–accuracy correlation to $\sim$0.4 and reducing overconfidence. Models provide plausible uncertainty justifications (e.g., occlusion, blur) but show systematic issues: 23% mismatches, anchoring, and mid-range compression. Consistency probing offers a complementary error-correlated signal. Gains reflect behavioral calibration, not true uncertainty awareness, and generalize across models. Limitations include persistent overconfidence ($\sim$50% of hallucinations), calibration error, $\sim$5$\times$ latency, small-scale evaluation, and reduced robustness under distribution shift. Overall, improvements likely arise from learned patterns, indicating the need for architectural advances for reliable uncertainty estimation.

References

[1] Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207. 2

[2] Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. Frontiers in human neuroscience, 8, 443. 1, 2

[3] Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. arXiv preprint arXiv:2302.09664. 3

[4] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J. R. (2023). Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355. 3

[5] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837. 2

[6] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744. 2

[7] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in neural information processing systems, 36, 34892-34916.

[8] Chen, C., Tseng, Y. Y., Li, Z., Venkatesh, A., & Gurari, D. (2025). Acknowledging focus ambiguity in visual questions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1228-1238). 3