Assessing Perceptual Metacognition in Vision-Language Models

Ran Ji[*†1], **Maijunxian Wang**[*2], **Qingying Gao**[3], **Yijiang Li**[4], **Hokin Deng**[5] **& Dezhi Luo**[†6]
[1]Department of Psychology, University of California San Diego
[2]Cognitive Science Program, University of California, Berkeley
[3]Department of Computer Science & Wilmer Eye Institute, Johns Hopkins University
[4]Department of Electrical and Computer Engineering, University of California San Diego
[5]Robotics Institute, Carnegie Mellon University
[6]Weinberg Institute for Cognitive Science, University of Michigan

Abstract

Vision-language models (VLMs) have demonstrated unprecedented capabilities in perception and reasoning, yet their perceptual metacognitive abilities remain unexamined, a gap with implications for reliable deployment and welfare considerations. We evaluated six open-source VLMs on perceptual decision-making tasks, comparing implicit (logit-based) and explicit (self-reported) metacognition. Models exhibited robust implicit calibration and sensitivity on high-performance tasks but unreliable explicit confidence reports with poor calibration and task-dependent sensitivity. Process modeling further revealed that both confidence measures converged on classic Signal Detection Theory, suggesting VLMs lack the metacognitive monitoring mechanisms that differentiate decision and confidence processes in humans. Together, we showed that current VLMs thus limited perceptual metacognition with notable dissociations between implicit and explicit confidence.

Keywords: perceptual metacognition; vision-language models; signal detection theory; confidence; calibration

Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse cognitive tasks, from complex reasoning to knowledge synthesis (Bai et al., 2023; Brown et al., 2020; Jaech et al., 2024; Touvron et al., 2023). However, substantial doubts have been raised about the nature of their understanding (Bender et al., 2021; Mitchell & Krakauer, 2023): do they genuinely know what they are doing, or are they just *stochastic parrots*? To tackle such concerns, **metacognition**, the ability to assess and monitor one's own cognitive process, emerges as a critical index. This question has been argued to hold critical implications for model reliability (Tian et al., 2023), safe deployment (Bricken et al., 2023), and the prospect of machine consciousness (S. Fleming et al., 2025).

Recent work has begun examining metacognition in language-only models, revealing both promise and fundamental limitations. Studies show that LLMs can introspect about their internal representations (Binder et al., 2024; Chen et al., 2024; Ghandeharioun et al., 2024), exhibit awareness of their learned behaviors (Betley et al., 2025; Plunkett et al., 2025), and monitor their internal activations (Ji-An et al., 2025). Some evidence suggests models possess calibrated uncertainty about their knowledge (Kadavath et al., 2022). However, other work reveals systematic failures (Cheng et al., 2024; Song et al., 2025), indicating that metacognitive abilities in LLMs remain incomplete and highly context-dependent.

Critically, this emerging literature focuses almost exclusively on language-only models. Yet vision-language models (VLMs)—which integrate visual perception with linguistic reasoning (Alayrac et al., 2022; Liu et al., 2024; Radford et al., 2021)—present a fundamentally different metacognitive challenge. VLMs must assess confidence not only about linguistic knowledge but also about their *perceptual* judgments—a capacity extensively studied in humans but entirely unexamined in artificial systems. This gap is particularly significant given the rapid deployment of VLMs in perception-critical applications requiring reliable uncertainty estimation, from medical image analysis to autonomous navigation.

We address this gap by conducting the first systematic investigation of **perceptual metacognition** in vision-language models. We evaluate six contemporary open-source VLMs on three perceptual decision-making tasks varying in structure and noise. Crucially, we employ a dual-measurement approach comparing *implicit* confidence (derived from output logits) with *explicit* confidence (self-reported ratings) (Steyvers & Peters, 2025). This design is motivated by prior findings that self-reported confidence in LLMs can be notoriously unreliable; we thus compare whether explicit confidence judgments actually reflect models' internal uncertainty representations.

Our results reveal fundamental differences in metacognitive mechanisms depending on measurement approach: VLMs demonstrate robust implicit calibration on high-performance tasks but unreliable explicit confidence reports with poor calibration and task-dependent efficiency. Process modeling further reveals that both confidence measures converge on classic Signal Detection Theory (SDT)-like mechanisms—a notable finding given that human perceptual metacognition exhibits levels of computational complexity that transcend simple SDT accounts. In addition, to examine scaling effects, we evaluated self-reported confidence in three state-of-the-art closed-source VLMs, finding similar limitations. This convergence suggests VLMs may lack the differentiated metacognitive monitoring processes that distinguish decision and confidence formation in humans for their perceptual judgments.

Related Works

Perceptual Metacognition in Humans. Perceptual metacognition involves monitoring one's own perceptual decisions through confidence judgments (S. M. Fleming,

*Equal Contributions
†Correspondence: r9ji@ucsd.edu; ihzedoul@umich.edu

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

2024). In typical paradigms, participants make a perceptual decision (Type-1) and rate their confidence (Type-2) (S. M. Fleming & Lau, 2014). This structure enables measurement of three properties: *metacognitive calibration* (whether confidence reflects accuracy), *metacognitive sensitivity* (ability to discriminate correct from incorrect responses via confidence, quantified through meta-d'/d' (Maniscalco & Lau, 2012)), and *metacognitive efficiency* (sensitivity relative to task performance (Maniscalco et al., 2016)). Visual metacognition research reveals these properties dissociate across individuals and tasks (Rahnev, 2021), with computational models showing that human confidence formation involves complex evidence accumulation beyond simple decision readout (Shekhar & Rahnev, 2024).

Metacognition in Language Models. Research approaches LLM confidence from two perspectives. The first treats confidence as a *calibration problem*, examining whether output probabilities or self-reports align with correctness (Tian et al., 2023; Xiong et al., 2023). Work reveals systematic overconfidence (Kumaran et al., 2025; Wen et al., 2024) and debates whether models genuinely possess confidence states (Keeling & Street, 2025). The second examines *metacognitive introspection*—whether models monitor internal states. Evidence suggests LLMs can introspect representations (Binder et al., 2024; Chen et al., 2024) and exhibit awareness of learned behaviors (Betley et al., 2025), though systematic failures persist (Cheng et al., 2024; Song et al., 2025). Critically, existing work focuses on linguistic knowledge, leaving perceptual metacognition unexplored.

Vision-Language Models. VLMs have advanced from image-text alignment (Radford et al., 2021) to sophisticated visual reasoning (Zellers et al., 2019) and chain-of-thought inference (Zhang et al., 2024). However, while no studies have directly assessed VLMs on perceptual judgment tasks analogous to human psychophysics, recent work has revealed fundamental perceptual limitations (Li et al., 2025), including overlooking their own visual representations (Fu et al., 2025), and failing at visual perspective-taking (Wang et al., 2025).

Methods

Overview of Task Paradigm

To test whether VLMs exhibit human-like perceptual metacognition, we designed a unified two-stage evaluation framework mirroring classical paradigms in cognitive psychology. In human experiments, participants make a perceptual judgment (Type-1 decision) and then rate their confidence (Type-2 evaluation). We adapted this structure to probe whether VLMs can separate decision formation from self-evaluation. All tasks followed the same two-stage paradigm (Figure 1).

Stage 1 (Perceptual Decision). The model viewed an image stimulus and made a binary forced-choice judgment (e.g., "Which one has more: O or X?"). Each prompt presented two labeled alternatives (e.g., "A. O" and "B. X"), and the model was instructed to respond with only the letter ("A" or "B") without reasoning or explanation. This ensures responses reflect pure perceptual inference rather than linguistic justification.

From the model's output logits, we derived an internal measure of evidence strength as the log-odds of the selected option:

$$\text{log-odds} = \log\left(\frac{p_{\text{chosen}}}{p_{\text{unchosen}}}\right), \tag{1}$$

where p_{chosen} and p_{unchosen} denote the softmax-normalized probabilities of the two answer tokens.

To enable direct comparison with self-report ratings, we mapped the probability difference to a 1–5 confidence scale:

$$\text{confidence} = 1 + 4 \times |p_A - p_B|, \tag{2}$$

where $|p_A - p_B|$ is the absolute difference between answer probabilities. When probabilities are equal ($p_A = p_B = 0.5$), confidence equals 1 (very uncertain); when one option dominates ($p_A = 1, p_B = 0$), confidence equals 5 (very certain). Values were rounded to integers and clipped to 1–5.

Stage 2 (Confidence Rating). After providing its perceptual answer, the model reported confidence in that choice using a standardized format (`Confidence: N`), where N ranged from 1 (very uncertain) to 5 (very certain). This structured response eliminates linguistic ambiguity (e.g., "probably," "very sure") and provides a direct analog to Likert-type scales used in human metacognition studies, representing a Type-2 metacognitive judgment.

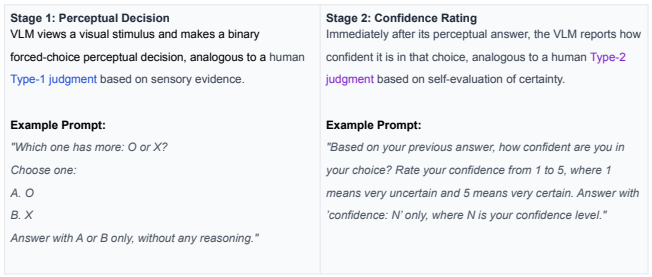

Figure 1: **Two-stage paradigm for evaluating metacognition in VLMs.** Each trial consisted of a Type-1 perceptual decision followed by a Type-2 confidence judgment. Stage 1: the model viewed a visual stimulus and made a binary choice. Stage 2: it rated confidence (1–5) using the format "`Confidence: N`". This framework enables direct comparison with human perceptual metacognition paradigms.

Together, these stages form a minimal framework for probing metacognitive behavior: Stage 1 captures evidence-based perceptual reasoning, while Stage 2 quantifies explicit self-monitoring. We compared implicit evidence (log-odds) with explicit confidence ratings to examine evidence–confidence alignment across tasks. The paradigm was applied consistently across three task families, described below.

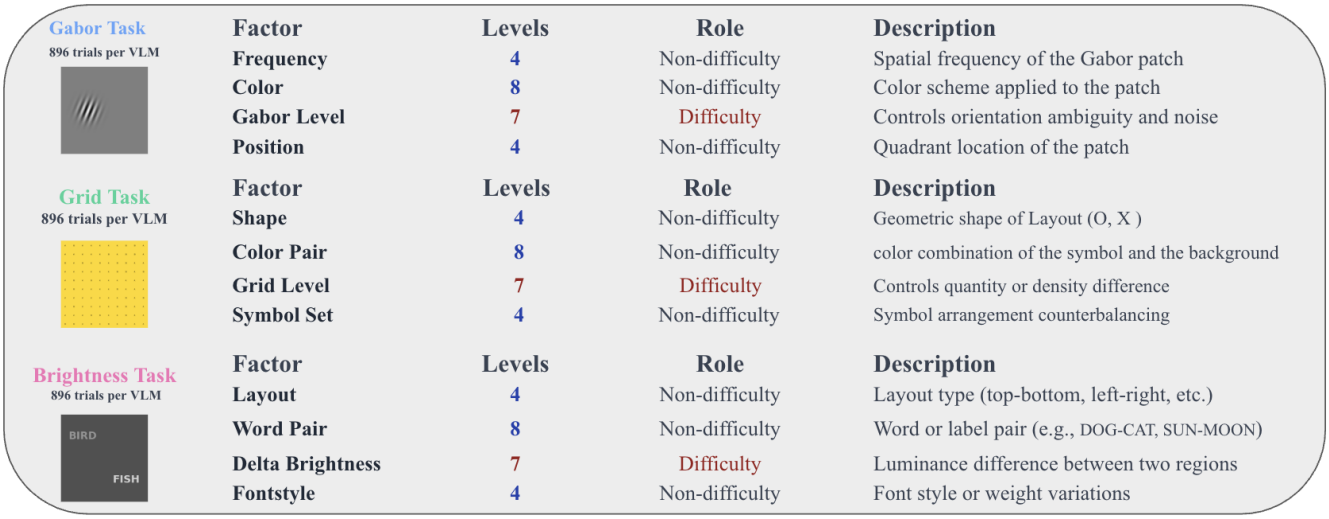

Figure 2: **Factorial structure of the three perceptual tasks.** Each task (Grid, Gabor, Brightness) followed a $7 \times 4 \times 4 \times 8$ design: one difficulty factor (7 levels) and three balanced non-difficulty factors (shape/layout, color, identity), ensuring only one perceptual dimension determined task difficulty.

Design of Perceptual Tasks

All three tasks followed the same two-stage paradigm while probing distinct visual dimensions—numerosity (Grid), orientation (Gabor), and luminance (Brightness) (Figure 2). Each task followed a factorial design ($7 \times 4 \times 4 \times 8 = 896$ trials), with perceptual difficulty isolated to a single variable while visual factors remained balanced (Figure 2). In total, each model completed 2,688 unique trials, providing a balanced dataset for metacognitive analysis.

Grid (Structured Quantity Discrimination). Each image contained two symbol categories (e.g., "O" vs. "X"). The model judged which symbol appeared more frequently. Task difficulty was determined by the proportional difference between symbols (**Grid Level**, 7 levels): closer ratios to 50/50 increased difficulty. Three non-difficulty factors were counterbalanced: (1) *Shape* (4 categories)—boundary contours (circle, diamond, square, triangle) ensuring visual diversity; (2) *Color Pair* (8 categories)—foreground/background combinations modulating contrast but not numerosity; and (3) *Symbol Set* (4 categories)—distinct symbol pairs (e.g., A/B, X/O) preventing lexical memorization. This task captures mid-level perceptual comparison analogous to human numerosity discrimination.

Gabor (Orientation Discrimination). Each trial presented a Gabor patch whose orientation varied across seven **Gabor Levels**. The model judged whether the pattern was closer to *vertical* or *horizontal*. Difficulty was defined by angular deviation from the 45° midpoint: stimuli near 45° were most ambiguous, whereas those near 0° or 90° were easiest. Three factors were balanced: (1) *Spatial Frequency* (4 categories)—stripe cycles per patch controlling texture granularity; (2) *Patch Color* (8 categories)—hue variations testing color-invariant sensitivity; and (3) *Position* (4 categories)—spatial location (center, left, right, top). This task probes low-level orientation sensitivity under controlled noise.

Brightness (Luminance Discrimination). Each image displayed two word-shaped stimuli in distinct regions, with one rendered slightly brighter. The model judged which word appeared visually brighter. Task difficulty was controlled by luminance difference (**Brightness Level** ΔL, 7 levels): smaller ΔL values induced higher uncertainty. Three factors were counterbalanced: (1) *Layout* (4 categories)—spatial arrangements (left–right, up–down, diagonal); (2) *Word Pair* (8 categories)—lexical pairs (e.g., DOG–CAT, STAR–MOON) ensuring visual diversity; and (3) *Font Style* (4 categories)—typographic variations (regular, bold, italic, monospace). This task measures low-level luminance sensitivity using word-shaped stimuli.

Model Inferences

We evaluated six open-source VLMs—Ovis2-34B, Qwen2.5-VL-7B/32B/72B, Gemma3-27B, and Kimi-VL-A3B—using official checkpoints. Each model completed 2,688 trials, producing Stage 1 binary choices and Stage 2 confidence ratings. Token-level logits for Stage 1 answers are translated to confidence scores (Eq. 1–2) to compare implicit evidence with explicit self-reports.

Results

First-order Task Performance

We first examined accuracy rates across the three perceptual tasks to establish baseline task difficulty. The three tasks differed substantially: Grid (0.94) was easy, while Gabor (0.36) and Brightness (0.39) were difficult (Table 1). This reveals that VLMs are not uniformly competent at simple perceptual judgments. Critically, this performance stratification enables subsequent metacognitive analyses across varying competence

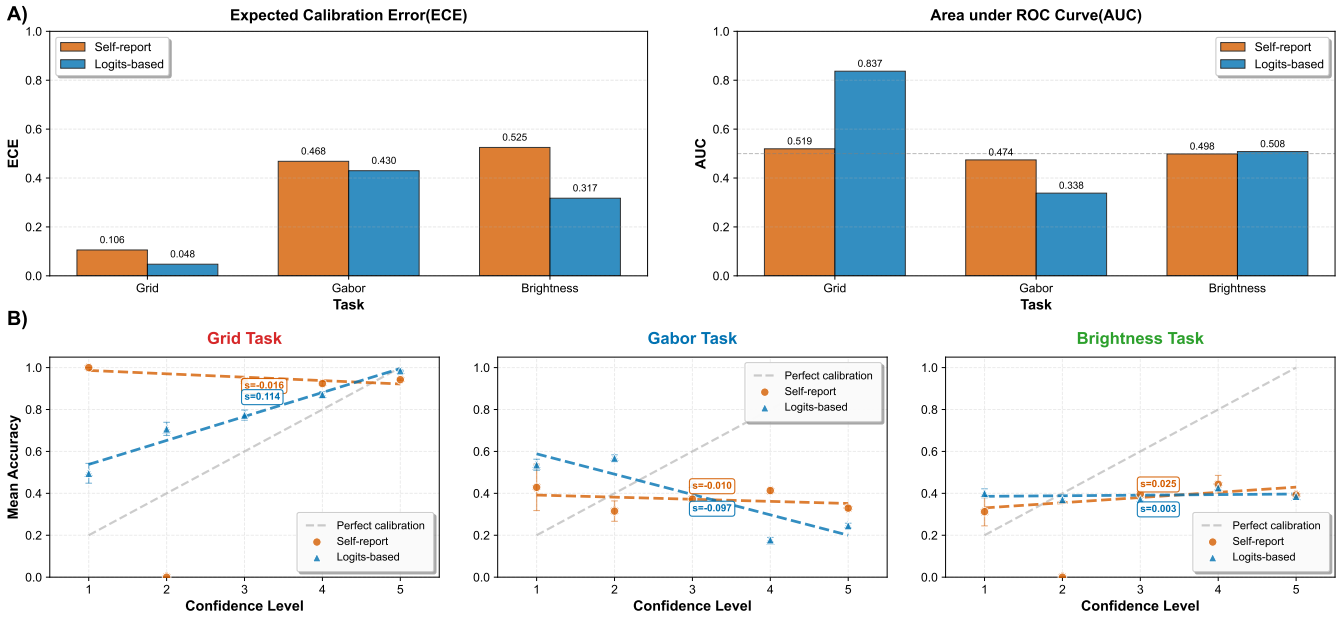

Figure 3: **Calibration analysis across tasks and confidence methods. Panel A:** Expected Calibration Error (ECE, left) and Area Under ROC Curve (AUC, right) by task and confidence measurement method. Lower ECE values indicate better calibration; AUC values above 0.5 indicate that confidence is predictive of accuracy. **Panel B:** Calibration curves for Grid, Gabor, and Brightness tasks, displaying the relationship between confidence levels (1-5) and mean accuracy. Regression lines (dashed) indicate the linear relationship, with slope values (s) displayed. The diagonal gray line represents perfect calibration.

levels, allowing us to examine whether confidence formation mechanisms differ between highly competent performance (Grid) and below-chance performance (Gabor, Brightness).

Table 1: **Accuracy across models and tasks.** Grid showed consistently high performance (>90%), while Gabor and Brightness were substantially more challenging.

Model	Grid	Gabor	Brightness
Gemma3-27b	93.75	43.42	33.59
Kimi-VL-A3b	91.85	49.89	34.71
Ovis2-34b	94.31	31.25	32.37
Qwen2.5-VL-32b	94.31	20.09	57.14
Qwen2.5-VL-72b	94.98	35.94	39.40
Qwen2.5-VL-7b	94.31	37.95	38.95
Average	**93.92**	**36.42**	**39.36**

Metacognitive Calibration

Metacognitive calibration refers to the alignment between subjective confidence and task performance, a core measure of self-monitoring reliability (S. M. Fleming, 2024). We evaluated calibration using three complementary metrics: Expected Calibration Error (ECE), which quantifies the absolute difference between confidence and accuracy (lower is better); Area Under the ROC Curve (AUC), which measures discriminative power (values >0.5 indicate confidence predicts accuracy); and calibration curve slope, which reflects the linear relationship between confidence and accuracy.

Figure 3 reveals a striking dissociation between implicit and explicit confidence measures. Logit-based confidence demonstrated robust calibration on the high-performing Grid task: AUC = 0.837, positive slope (0.114), and low ECE (0.048), indicating that higher internal confidence genuinely predicted higher accuracy. In contrast, self-reported confidence showed near-chance discriminative power across all tasks (AUC $\approx$ 0.5) with flat calibration slopes near zero (Grid: -0.016, Gabor: -0.010, Brightness: 0.025), indicating no meaningful confidence-accuracy relationship. Notably, self-report's low ECE in Grid (0.106) was purely artifactual: systematic overconfidence (mean = 4.59) numerically coincided with high accuracy (0.94), creating spurious alignment without genuine predictive ability.

Logit-based confidence demonstrated substantially superior calibration. In the high-performing Grid task, it exhibited excellent calibration (AUC = 0.837, slope = 0.114, ECE = 0.048), far surpassing self-report (AUC = 0.519, slope = -0.016). Critically, logit-based confidence maintained consistently lower ECE across all tasks (Grid: 0.048 vs 0.106; Gabor: 0.430 vs 0.468; Brightness: 0.317 vs 0.525). However, discriminative ability degraded in difficult conditions: Brightness showed minimal predictive power (AUC = 0.508), while Gabor became counter-predictive (AUC = 0.338, slope = -0.097), with higher confidence predicting lower accuracy.

These results reveal a fundamental dissociation in VLM

metacognition: models possess internally well-calibrated uncertainty estimates for Type-1 decisions on high-performing tasks, but fail to monitor this information to generate accurate explicit confidence reports. Self-reported confidence is fundamentally unreliable regardless of task difficulty, while logit-based confidence, though effective in simple conditions, degrades or even inverts in challenging perceptual tasks.

Metacognitive Sensitivity and Efficiency

Beyond calibration, we assessed *metacognitive sensitivity*—the ability to discriminate correct from incorrect responses via confidence, quantified through meta-d' (Maniscalco & Lau, 2012)—and *metacognitive efficiency* (M-ratio = meta-d'/d'), which measures sensitivity relative to task performance (Maniscalco et al., 2016). We focus particularly on efficiency because sensitivity measures are often "contaminated" by first-order performance, whereas M-ratio provides insight independent of task accuracy and confidence bias (Steyvers & Peters, 2025). An M-ratio of 1.0 indicates optimal efficiency; values below 1.0 indicate degraded metacognition, while values above 1.0 suggest over-discrimination relative to task performance.

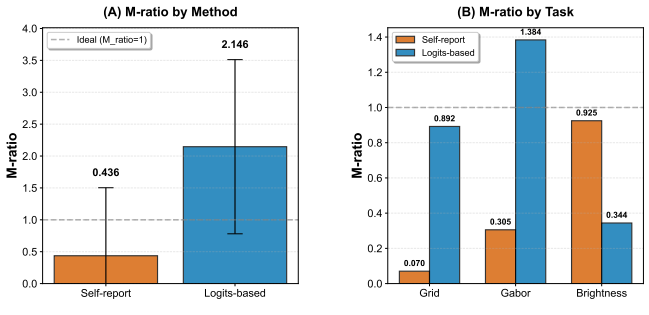

Figure 4: **Metacognitive efficiency (M-ratio) analysis. (A)** Overall M-ratio by method shows self-report (0.436) with suboptimal efficiency and logits-based (2.146) with over-discrimination. **(B)** Task-specific patterns reveal striking dissociations: logits-based dominates in Grid (0.892 vs 0.070) but fails in Brightness (0.344 vs 0.925), where self-report achieves near-optimal efficiency. Dashed line indicates ideal M-ratio = 1.0.

Figure 4A shows that overall, self-report achieved suboptimal efficiency (M-ratio = 0.436) while logits-based over-discriminated (M-ratio = 2.146). However, Panel B reveals striking task-dependent dissociations. In Grid, logits-based substantially outperformed self-report (0.892 vs 0.070), consistent with our calibration findings. In contrast, Brightness showed a complete reversal: self-report achieved near-optimal efficiency (M-ratio = 0.925) while logits-based degraded substantially (0.344)—a 2.7-fold difference favoring self-report.

This reveals a nuanced picture in which metacognitive efficiency dissociates from calibration: explicit confidence reports can effectively discriminate correct from incorrect responses (high M-ratio in Brightness) even while systematically over-estimating absolute confidence levels (poor calibration across all tasks). Conversely, logits-based confidence demonstrates strong metacognitive measures in simple tasks but fails to maintain discriminative efficiency in challenging perceptual conditions. These nuanced findings indicate that VLMs may possess distinct computational pathways for internal uncertainty estimation versus explicit confidence reporting, with fundamentally different performance profiles across perceptual tasks.

Process Modeling

To understand the computational mechanisms underlying VLM metacognition, we applied *process modeling*—a quantitative framework that tests alternative theories of how internal evidence is transformed into confidence judgments (Shekhar & Rahnev, 2021). We compared five computational models, each representing a distinct hypothesis about the evidence-confidence relationship:

- **Signal Detection Theory (SDT)**: Confidence and choice derive from the same sensory evidence, with no additional metacognitive processing (Green, Swets, et al., 1966).
- **Lognormal Meta-Noise (LogN)**: Confidence is selectively corrupted by signal-dependent metacognitive noise following a lognormal distribution (Shekhar & Rahnev, 2021).
- **Positive Evidence (PE)**: Confidence is based only on decision-congruent evidence, ignoring contradictory information (Maniscalco et al., 2016).
- **Weighted Evidence and Visibility (WEV)**: Confidence signal is a weighted sum of evidence strength and stimulus visibility (Rausch et al., 2018).
- **Bayesian Confidence Hypothesis (BCH)**: Confidence is computed as the posterior probability of being correct given the evidence (Hangya et al., 2016).

We selected these models based on two considerations. First, they have relatively few free parameters, making them appropriate for VLMs' limited perceptual abilities and suitable for an exploratory study. Second, given documented over-confidence in language models (Kumaran et al., 2025; Wen et al., 2024), we included PE to test whether VLMs selectively weight decision-congruent evidence. Model fitting followed the procedures in Shekhar and Rahnev (2021).

We fitted all five models to each VLM-task-method combination using bootstrap analysis (300 resamples), identifying the best-fitting model via Akaike Information Criterion (AIC). We then applied random-effects Bayesian Model Selection (RFX-BMS) to compute model frequencies (p_{model}) across all six VLMs, representing the posterior probability that each model best explains the data when accounting for within- and between-VLM variability.

Table 2 reveals a striking pattern: **SDT** dominated both self-report (13/18 combinations, 72%) and logits-based (12/18, 67%) confidence measures. **BCH**, which differs from SDT primarily in its Bayesian formulation, accounted for most remaining fits (self-report: 5/18, 28%; logits: 3/18, 17%).

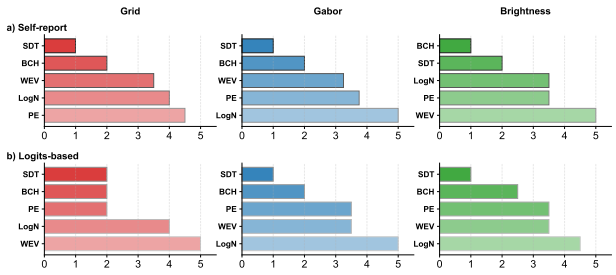

Figure 5: **Process model rankings by method and task.** Horizontal bars show model ranks (shorter = better) based on RFX-BMS across six VLMs.

VLM	Self-report			Logits-based		
	Grid	Gab.	Bright.	Grid	Gab.	Bright.
qwen2.5-vl-7b	SDT	SDT	SDT	SDT	BCH	SDT
qwen2.5-vl-72b	SDT	SDT	SDT	PE	SDT	SDT
ovis2_34b	SDT	BCH	BCH	BCH	BCH	SDT
kimi-vl-a3b	BCH	SDT	SDT	SDT	SDT	SDT
gemma3_27b	SDT	SDT	SDT	PE	SDT	SDT
qwen2.5-vl-32b	SDT	BCH	BCH	LogN	SDT	SDT
Dominant:						
PE	0	0	0	2	0	0
BCH	1	2	2	1	2	0
SDT	5	4	4	2	4	6
LogN	0	0	0	1	0	0
WEV	0	0	0	0	0	0

Table 2: **Best-fitting process models per VLM-task combination**. Blue: models in self-report; green: models in logits; red: dominant counts. SDT dominates both self-report (13/18, 72%) and logits-based (12/18, 67%) confidence, with BCH as secondary. Logits-based shows additional diversity (PE, LogN) in Grid task.

Notably, alternative models (PE, LogN) emerged only in logits-based Grid data (3/6 models), suggesting additional stochastic mechanisms for internal uncertainty in high-performing tasks. However, the overall convergence on SDT/BCH indicates that both implicit and explicit confidence derive from the same sensory evidence without the differentiated metacognitive monitoring mechanisms that characterize human perceptual metacognition (Shekhar & Rahnev, 2024).

Overall, both implicit and explicit confidence measures converged to classic Signal Detection Theory (SDT), suggesting that VLMs' perceptual decisions and confidence judgments derive from the same sensory evidence without the additional metacognitive layers that distinguish human confidence formation from perceptual decisions.

Discussion

We conducted the first investigation of perceptual metacognition in six open-source VLMs. VLMs exhibited robust implicit (logit-based) calibration on high-performing tasks but unreliable explicit (self-reported) confidence with poor calibration and task-dependent sensitivity. Process modeling revealed both measures converged on Signal Detection Theory, suggesting VLMs lack the differentiated metacognitive monitoring mechanisms that distinguish decision from confidence formation in humans (Shekhar & Rahnev, 2024). While logit-based confidence showed superior calibration in simple tasks, it degraded or became counter-predictive in challenging conditions. Self-reported confidence achieved near-optimal efficiency in specific contexts but showed systematically poor calibration across tasks, indicating fundamental limitations in explicit metacognitive reporting.

The dissociation between implicit and explicit confidence raises methodological questions about what constitutes metacognition in artificial systems (Steyvers & Peters, 2025). Logit-based confidence tracks first-order performance on a causal precedent basis rather than through post-hoc monitoring (Keeling & Street, 2025)—a critical distinction from human implicit metacognition, which operates unconsciously yet differentiates correct from incorrect judgments independent of decision strength (S. M. Fleming, 2024). Thus, whether high logit-based calibration constitutes genuine metacognition remains debatable. Moreover, counter-predictive confidence under below-chance performance (as in Gabor) violates basic credence principles. Together with SDT convergence—indicating confidence derives directly from decision evidence—these findings suggest VLMs possess limited perceptual metacognition. Self-reported confidence proves particularly unreliable, likely reflecting absent introspective access to internal states.

By adapting classical psychophysical paradigms and model-based computational analyses to VLMs, we contribute to frameworks for assessing metacognition across paradigms of intelligence (S. M. Fleming, 2023; Kammerer & Frankish, 2023; Long, 2023). Theoretically, these findings inform debates about machine self-modeling (Bongard et al., 2006) and computational requirements for genuine introspection. Practically, unreliable perceptual metacognition poses risks in high-stakes applications requiring uncertainty quantification. Speculatively, higher-order theories of consciousness predict specific functional profiles for conscious systems (S. Fleming et al., 2025); our findings suggest current VLMs lack the metacognitive architecture these theories associate with conscious awareness despite sophisticated perceptual capabilities.

Conclusion

We investigated perceptual metacognition in six VLMs using dual confidence measures. VLMs exhibited robust implicit calibration in high-performing tasks but unreliable explicit confidence with poor calibration. Process modeling revealed Signal Detection Theory convergence, suggesting VLMs lack differentiated metacognitive monitoring mechanisms. Current VLMs thus possess limited perceptual metacognition with notable implicit-explicit dissociations. Future work should consider exploring training interventions and mechanistic interpretability methods to assess perceptual metacognition in multi-modal foundation models as a complement of behavioral paradigms we explore here.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: A visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716–23736.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., & Evans, O. (2025). Tell me about yourself: Llms are aware of their learned behaviors. *arXiv preprint arXiv:2501.11120*.

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., & Evans, O. (2024). Looking inward: Language models can learn about themselves by introspection. *arXiv preprint arXiv:2410.13787*.

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, *314*(5802), 1118–1121.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, *2*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chen, H., Vondrick, C., & Mao, C. (2024). Selfie: Self-interpretation of large language model embeddings. *arXiv preprint arXiv:2403.10949*.

Cheng, Q., Sun, T., Liu, X., Zhang, W., Yin, Z., Li, S., Li, L., He, Z., Chen, K., & Qiu, X. (2024). Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*.

Fleming, S. M. (2023). Metacognitive psychophysics in humans, animals, and ai: A research agenda for mapping introspective systems. *Journal of Consciousness Studies*, *30*(9-10), 113–128.

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

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

Fleming, S., Brown, R., & Cleermans, A. (2025). Computational higher-order theories of consciousness. In L. Melloni & U. Olcese (Eds.), *The scientific study of consciousness – experimental and theoretical approaches*. Springer Nature.

Fu, S., Bonnen, T., Guillory, D., & Darrell, T. (2025). Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*.

Ghandeharioun, A., Caciularu, A., Pearce, A., Dixon, L., & Geva, M. (2024). Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation*, *28*(9), 1840–1858.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. (2024). Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ji-An, L., Mattar, M. G., Xiong, H.-D., Benna, M. K., & Wilson, R. C. (2025). Language models are capable of metacognitive monitoring and control of their internal activations. *ArXiv*, arXiv–2505.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kammerer, F., & Frankish, K. (2023). What forms could introspective systems take? a research programme. *Journal of Consciousness Studies*, *30*(9-10), 13–48.

Keeling, G., & Street, W. (2025). On the attribution of confidence to large language models. *Inquiry*, 1–27.

Kumaran, D., Fleming, S. M., Markeeva, L., Heyward, J., Banino, A., Mathur, M., Pascanu, R., Osindero, S., De Martino, B., Velickovic, P., et al. (2025). How overconfidence in initial choices and underconfidence under criticism modulate change of mind in large language models. *arXiv preprint arXiv:2507.03120*.

Li, Y., Gao, Q., Zhao, T., Wang, B., Sun, H., Lyu, H., Hawkins, R. D., Vasconcelos, N., Golan, T., Luo, D., et al. (2025). Core knowledge deficits in multi-modal language models. *Forty-second International Conference on Machine Learning*.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in neural information processing systems*, *36*.

Long, R. (2023). Introspective capabilities in large language models. *Journal of Consciousness Studies*, *30*(9-10), 143–153.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, *21*(1), 422–430.

Maniscalco, B., Peters, M. A., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics*, *78*(3), 923–937.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.

Plunkett, D., Morris, A., Reddy, K., & Morales, J. (2025). Self-interpretability: Llms can describe complex internal

processes that drive their decisions, and improve with training. *arXiv preprint arXiv:2505.17120*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv: 2103.00020*.

Rahnev, D. (2021). Visual metacognition: Measures, models, and neural correlates. *American Psychologist*, *76*(9), 1445.

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics*, *80*(1), 134–154.

Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological review*, *128*(1), 45.

Shekhar, M., & Rahnev, D. (2024). How do humans give confidence? a comprehensive comparison of process models of perceptual metacognition. *Journal of Experimental Psychology: General*, *153*(3), 656.

Song, S., Hu, J., & Mahowald, K. (2025). Language models fail to introspect about their knowledge of language. *arXiv preprint arXiv:2503.07513*.

Steyvers, M., & Peters, M. A. (2025). Metacognition and uncertainty communication in humans and large language models. *arXiv preprint arXiv:2504.14045*.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, B., Li, Y., Zhou, Q., Leong, H. Y., Zhao, T., Ye, L., Deng, H., Luo, D., & Vasconcelos, N. (2025). Do vision language models infer human intention without visual perspective-taking? towards a scalable" one-image-probe-all" dataset. *ICML 2025 Workshop on Assessing World Models*.

Wen, B., Xu, C., Wolfe, R., Wang, L. L., Howe, B., et al. (2024). Mitigating overconfidence in large language models: A behavioral lens on confidence estimation and calibration. *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2023). Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. https://arxiv.org/abs/1811.10830

Zhang, R., Zhang, B., Li, Y., Zhang, H., Sun, Z., Gan, Z., Yang, Y., Pang, R., & Yang, Y. (2024). Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.