Quantifying Semantic Priors vs Visual Evidence in Visual Language Models via Psychometric Curve Analysis

Hetao Zheng[1], **Yifan Zeng**[1], **Jian Zhao**[1], **Yufei Zhou**[1],**Huiyu Zhou**[2*], **Peijia Zheng**[1*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China
[2]Guangxi Key Laboratory of Digital Infrastructure, Guangxi Zhuang Autonomous Region Information Center, Nanning 530000, China
{zhenght8@mail2, zengyf53@mail2, zhaoj326@mail2, zhouyf55@mail2, zhpj@mail}.sysu.edu.cn
zhouhy3296@gmail.com

Abstract

Visual illusions are a classic tool in cognitive science for probing perceptual inference. Recent studies suggest that when visual facts conflict with semantic prior knowledge, vision-language models (VLMs) often make systematic errors in which priors override visual evidence. To quantify this effect, we treat VLMs as artificial participants and introduce IlluQuant, an illusion-diagnostic dataset that operationalizes template similarity and visual evidence strength as continuous, controllable variables. By fitting psychometric functions, we measure semantic prior dominance in VLMs. Results show that stronger semantic cues substantially increase the probability of prior-driven errors, while visual evidence must be amplified several-fold to partially offset this bias. We further propose MdCoT, an intervention that mitigates the effect. IlluQuant provides a psychophysics-style framework for quantitatively characterizing the balance between semantic priors and visual evidence in multimodal models under controlled illusion manipulations. The IlluQuant dataset is available at haitoooo/IlluQuant

Keywords: Vision language models; psychophysics; visual illusions; artificial intelligence; perception

Introduction

Illusions have long been used to study human vision because they help disentangle sensory and cognitive processes (Day (1984), Eagleman (2001), and Gregory (1997)). Inspired by this line of work, recent studies have begun to use visual illusions and "Illusion-Illusions" to probe machine vision (Gomez-Villa et al. (2019), Guan et al. (2024), Ullman (2024), and Y. Zhang et al. (2025)). These studies report an anomalously poor performance of vision-language models (VLMs) on Illusion-Illusion images: even when an image contains no illusory effect, the model may still forcefully produce the canonical illusion conclusion based on the configuration, as if it "sees an illusion that is not there." (Guan et al. (2024), Ullman (2024), and Y. Zhang et al. (2025)).

Existing literature generally attributes this error to semantic priors learned by pretrained language models overpowering visual evidence (Guan et al. (2024) and Leng et al. (2024)). During training, models may acquire high-frequency co-occurrences among "typical configurations–illusion names–standard explanations," which makes them more likely to enter an "illusion-template" mode at inference time. Consequently, even when the physical facts in the current image contradict the template, the model tends to rely on semantic prior knowledge and ignore the visual evidence at hand (Guan et al. (2024), Rohrbach et al. (2018), and Y. Zhang et al. (2025)).

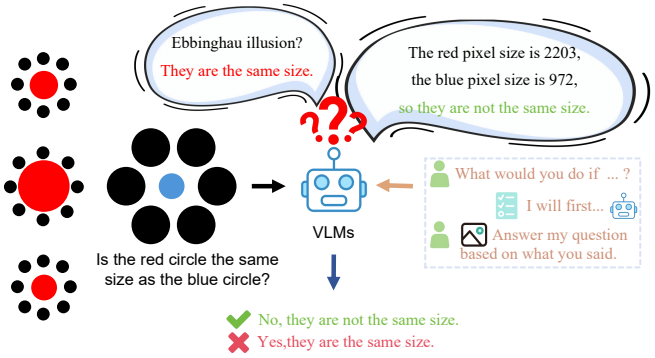

Figure 1: Overall framework of our evaluation

However, most prior work remains at the level of discrete contrasts (Illusion vs. Control/Illusion-Illusion) and qualitative description, lacking quantitative criteria that can precisely measure how strong visual evidence must be to help VLMs break free from the constraints of "illusion-template" mode. To address this gap, we define **Semantic Prior Dominance (SPD)** as the extent to which VLM responses are governed by semantic priors rather than visual evidence (Kersten and Yuille (2003), Kersten et al. (2004), and Knill and Richards (1996)). Drawing on psychophysical stimulus manipulation, we operationalize "semantic cue strength" and "visual evidence strength" as controllable variables, and quantify semantic prior dominance via behavioral curves obtained under continuous manipulations (Prins et al. (2016), Treutwein (1995), and Wichmann and Hill (2001a, 2001b)).

Specifically, we construct an illusion-diagnostic dataset, IlluQuant, to measure how easily a model enters and exits an "illusion-template attractor" and to fit evidence–response curves, thereby estimating the evidence threshold and response slope required for judgment updating. Building on IlluQuant, we further propose ETdyn75, the evidence strength at which the fitted psychometric curve reaches 75% of its floor-to-ceiling transition, as a compact summary of how resistant a model is to updating its judgment in response to visual evidence. Through systematic experiments across multiple VLMs, we move beyond asking whether the phenomenon occurs to quantifying its magnitude, when it emerges, and

*Corresponding author.

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

when it subsides, providing an experimental framework for understanding the trade-off between priors and evidence in multimodal models. Our main contributions are as follows.

1. We treat VLMs as "artificial participants," operationalize template similarity and visual evidence strength as continuous, controllable variables, and quantify VLMs' semantic prior dominance by fitting psychophysical functions.
2. We construct an illusion-diagnostic dataset, IlluQuant, which consists of two subsets, IlluLadder and IlluScale, providing a quantitative experimental platform for studying semantic-over-visual effects.
3. We conduct a human-referenced evaluation of the semantic-over-visual effect across multiple mainstream VLMs and further propose an intervention method, MdCoT, to mitigate this phenomenon.

Related Work

Recent work has extended "hallucination" beyond a purely linguistic issue to cross-modal mismatches in large vision-language models (VLMs) (Li et al. (2023) and Pham and Schott (2024)). Guan et al. (2024) explicitly distinguishes two typical failure modes: Language Hallucination (drawing conclusions without visual grounding) and Visual Illusion (misinterpreting visual input with high confidence). They further show that answer reversals under slight visual edits can more sensitively reveal the failure pattern in which language priors override visual evidence.

Ullman (2024) further treats illusions as a diagnostic tool by testing "Illusion-Illusions" that should not elicit illusory responses. They report that when the prompt explicitly states "this is a visual illusion" (e.g., by prepending a phrase such as "in the following visual illusion" to the original prompt), the performance of frontier models under the Illusion-Illusion and control conditions collapses: models answer almost all Illusion-Illusions as if they were genuine illusions, exhibiting a strong semantic-prior-over-visual-evidence effect.

Shinozaki et al. (2025) directly leverages classic visual illusions to probe VLMs' "perceptual style," proposing three task types: Genuine, Fake and Control VQA. The authors note that some prior studies use non-abstract images, which can introduce ambiguity in model responses (Guan et al. (2024) and Shahgir et al. (2024)). A model may appear superficially competent on genuine illusions, yet on fake illusions it often follows a pattern of "getting the apparent answer right but the physical truth wrong," leading the authors to infer that some models rely more on illusion-related knowledge learned during training than on online visual perception.

Overall, prior work has established that vision-language models perform poorly when challenged by Illusion-Illusions, reflecting a reasoning mode in which semantic priors override visual evidence. However, these studies largely lack a more "psychophysics-style" characterization: how model performance varies as a function of visual evidence strength, and whether the resulting transition can be explained in terms of a quantifiable degree of semantic prior dominance.

IlluQuant Dataset

Differences in illusion stimulus strength can substantially affect the performance of vision-language models (VLMs). As illustrated in Fig. 2, scaling the physical magnitude of a key property (e.g., a 2× vs. 5× enlargement) yields markedly different visual salience.

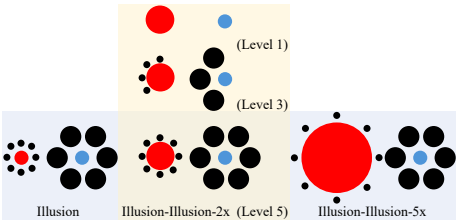

Figure 2: Different stimulus intensities

Based on these, we construct **IlluQuant**, an illusion-based stimulus dataset designed to quantify the Semantic Prior Dominance of VLMs. IlluQuant consists of two complementary subsets: **IlluLadder** and **IlluScale**.

IlluQuant includes 11 classic two-dimensional abstract illusion types, which reduces ambiguities introduced by semantic content and 3D cues in natural images. IlluScale covers all 11 illusion types. Because not every classic illusion admits a well-matched control stimulus without compromising task comparability, IlluLadder retains only 8 illusion tasks for which strict controls can be constructed; the dataset composition is summarized in Fig. 3.

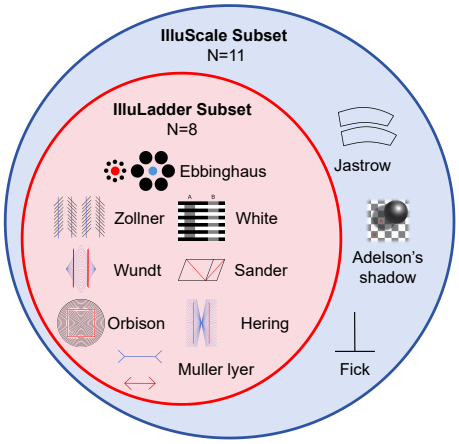

Figure 3: IlluQuant Dataset

Base configurations (Illusion → Illusion-Illusion → Control). Following the tri-partite stimulus settings used to probe VLMs with classic illusions—Illusion, Illusion-Illusion and Control (Shinozaki et al. (2025) and Ullman (2024))—we treat the transitions between these endpoints as continuous manipulation axes and discretize them into graded levels. Intuitively, Illusion images contain illusion inducers that create a discrepancy between appearance and physical truth; Illusion-Illusion images preserve the canonical configuration but make

the critical property physically consistent with the apparent percept; and Control images remove the illusion-inducing cues so that the judgment can be made without the illusion template being triggered.

IlluLadder (Template-Similarity: Illusion-Illusion → Control). We discretize the Illusion-Illusion to Control transition into five ordered similarity levels, denoted as IlluLadder. As illustrated in Fig. 2, this manipulation progressively reduces illusion-inducing cues and contextual information while keeping the target physical difference fixed. L1 is Control-like, and L5 is Illusion-Illusion-like. At each level from L5 to L1, we remove approximately one quarter of the inducers (e.g., in Fig. 2, removing 1/4 of the black spheres). This design enables a graded assessment of how strongly cue availability drives models toward illusion-template-consistent responses.

IlluScale (Visual-Evidence: Illusion → Illusion-Illusion). In parallel, we discretize the Illusion to Illusion-Illusion transition into eight evidence levels (including the original Illusion at 1.0×).As illustrated in Fig. 2, IlluScale varies visual evidence strength by manipulating only the magnitude of the underlying physical difference on the target dimension while keeping the overall inducing configuration fixed (same inducer set, layout, and styling). Concretely, starting from the original stimulus (1.0×), we increase the physical difference with multipliers 1.1×, 1.2×, 1.3×, 1.5×, 2×, 3× and 5×. We choose the 2× condition in IlluScale as the Illusion-Illusion stimulus, which also corresponds to L5 in IlluLadder, because it is the smallest multiplier at which human accuracy reaches near-ceiling overall, ensuring physical-consistency while minimizing distribution shift. The 3× and 5× conditions are further used to amplify visual evidence and test whether the model remains locked into the illusion template. We noted that the notion of "scale" is illusion-family specific: for Ebbinghaus-like stimuli it is a size multiplier, whereas for Zollner-like stimuli it corresponds to a change in tilt angle. Importantly, within each illusion family, step sizes are defined on a consistent physical parameterization so that evidence strength is comparable across levels. This axis operationalizes how much additional physical evidence is required for a model to abandon an illusion-template response and switch to evidence-consistent judgment.

For each image, we additionally provide a standardized **VQA (Visual–Question–Answer) triple**: **Question** specifies the comparison targets and the decision dimension, **Options** defines a discrete answer set to constrain output format (e.g., "same/different", we set each item as a 2AFC judgment to preserve a binary chance level of 0.5), and **Answer** gives the ground-truth label based on physical reality. For example: Question: "Are the red circle and the blue circle the same size?" Options: ["Yes, they are the same size."; "No, they are not the same size."] Answer: "No, they are not the same size."

To examine whether the manipulations in IlluLadder and IlluScale elicit the expected patterns of human perceptual behavior, we recruited 60 undergraduate participants as a human baseline sample. The participants were ordinary undergraduate students rather than individuals with specialized training in visual illusions. All participants provided informed consent prior to the experiment.

The stimulus pool consisted of 40 IlluLadder items and 88 IlluScale items. Eight items are shared between the two benchmarks (the L5 of IlluLadder is identical to the IlluScale 2× condition), yielding 120 unique 2AFC questions in total. We then replicated each question five times (120 × 5 = 600 trials) and randomly partitioned the 600 trials into 60 questionnaires, each containing 10 questions. Each participant completed exactly one questionnaire (10 trials), and by construction each unique question was answered exactly five times. Participants completed the task at their own pace, and no explicit time limit was imposed for each question.

Mean classification accuracy was computed as the primary outcome measure. Results showed that participants achieved **over 99%** accuracy on IlluLadder. In contrast, under IlluScale, mean accuracy at low evidence levels (1.0× and 1.1×) was approximately **50%**, near chance performance, indicating that the visual evidence at these scales was insufficient to support reliable judgments. As the scale factor increased, accuracy rose rapidly and approached 100% by 2×. This fast-saturating improvement with increasing evidence strength is consistent with IlluScale's intended continuous manipulation of visual evidence, providing behavioral support for the validity of the dataset design.

Experiment

Our evaluation covers both state-of-the-art proprietary multimodal models—**GPT-5.2** (OpenAI (2025)), **Gemini-3-Pro**(Google (2025)), and **Claude-Opus-4.5** (Anthropic (2025))—and a representative open-source vision-language model, **Qwen3-VL:32B**[†] (Bai et al. (2025)), to improve reproducibility. To ensure consistency and fairness, all models are tested under a standardized single-turn VQA protocol without additional chain-of-thought unless otherwise specified, and each test instance performs 5-round repeated sampling to obtain robust performance estimates.

Modality Decoupling

To demonstrate that model errors arise from prior interference rather than deficits in basic capability, we first decouple visual and language competencies via two controlled settings.

Table 1: Results of VLMs' Modality Decoupling Experiment. All metrics are reported in percentage (%).

Model	Visual-Control	Text-Only
Gemini-3-Pro	95.83	100.00
GPT-5.2	87.50	90.91
Claude-Opus-4.5	100.00	100.00
Qwen3-VL:32B	87.50	92.73

[†]https://ollama.com/library/qwen3-vl:32b

Table 2: Performance across IlluLadder levels and average conditional metrics. All metrics are reported in percentage (%). **Ac.** denotes accuracy. **Illu.MR** (Illusion Mention Rate) denotes the probability that the token "illusion" appears in the *rationale* (case-insensitive). $\overline{\text{Ac.Illu}}$ / $\overline{\text{Ac.NIllu}}$ are conditional accuracies grouped by whether the rationale mentions "illusion". $\overline{\Delta}$ measures the mean illusion-frame penalty, defined as Δ = Ac.NIllu − Ac.Illu.

Model	L1		L2		L3		L4		L5		Average Conditional		
	Ac.	Illu.MR	Ac.	Illu.MR	Ac.	Illu.MR	Ac.	Illu.MR	Ac.	Illu.MR	$\overline{\text{Ac.Illu}}$	$\overline{\text{Ac.NIllu}}$	$\overline{\Delta}$
Gemini-3-Pro	95.83	33.33	75.00	58.33	75.00	79.17	37.50	87.50	29.17	87.50	52.14	100.00	47.86
GPT-5.2	87.50	12.50	79.17	8.33	50.00	50.00	62.50	54.17	45.83	66.67	13.11	93.79	80.68
Claude-Opus-4.5	100.00	20.83	66.67	54.17	50.00	62.50	37.50	87.50	25.00	87.50	40.26	100.00	59.74
Qwen3-VL:32B	87.50	8.33	87.50	20.83	91.67	25.00	83.33	29.17	54.17	41.67	29.90	97.66	67.76

Text-Only (Semantic-Prior Baseline): We provide only the illusion name and question-options to VLMs (e.g., "Imagine an image of Hering Illusion. Are the two red lines parallel? Options[Yes, they are parallel. No, they are not parallel.]").

Visual-Control (Visual-Perception Baseline): We input control images with illusion-inducing cues removed (the control group, i.e., the IlluLadder-Level1 image set) and question-options to VLMs.

As shown in Table 1, when only textual descriptions are provided (Text-Only), all models achieve **over 90%** accuracy on illusion-related knowledge, indicating strong semantic memory. Under the Visual-Control setting with illusion cues removed, models also maintain high visual discrimination accuracy (**87.5%–100%**). Together, these results establish an "upper bound" for semantic priors and a "lower bound" for basic visual grounding, supporting the interpretation that the performance changes observed in subsequent experiments are driven by illusion-inducing perturbations rather than insufficient base-level visual capability.

IlluLadder

As the image increasingly resembles an Illusion-Illusion, how do VLMs gradually fall into a cognitive trap? To quantify this transition as a function of template similarity, we evaluate models on the IlluLadder subset of IlluQuant and report mean accuracy under progressively increasing template similarity.

For each IlluLadder item, the VLM is required to (i) choose exactly one option from a predefined set (e.g., yes/no) and (ii) provide a brief rationale. We compute accuracy (Ac.) by matching the selected option to the ground-truth label. To quantify explicit illusion framing in rationales, we compute Illusion Mention Rate (Illu.MR) from the rationale only: it is 1 if the rationale contains the token "illusion" case-insensitive, and 0 otherwise. We emphasize that Illusion Mention Rate captures explicit framing as a lightweight surface cue and is not a definitive indicator of template/script activation.

As shown in Fig. 4 and Table 2, as template similarity increases (i.e., more illusion-inducing elements are present), VLMs are more likely to explicitly invoke the "illusion" frame in their rationales (higher Illusion Mention Rate), while accuracy decreases on average. Some models approach chance at L4, indicating that sufficiently strong illusion cues can systematically bias model decisions even under a structured option-selection protocol.

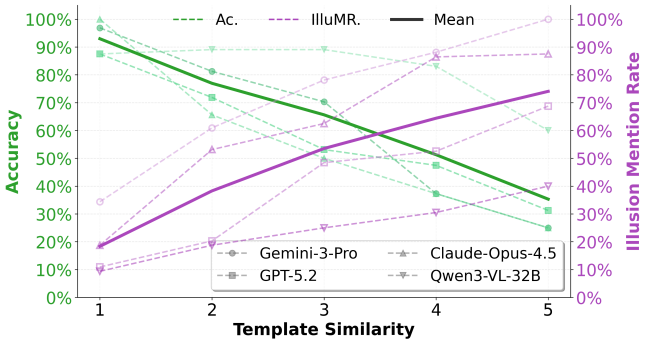

Figure 4: Accuracy (**Ac.**) and Illusion Mention Rate (**Illu.MR**) as functions of IlluLadder (L1–L5).

Table 2 further reports conditional accuracies aggregated across levels. The mean accuracy when the rationale does not mention "illusion" remains near ceiling (approximately 94%–100%), whereas the mean accuracy when the rationale does mention "illusion" is substantially lower (approximately 13%–52%).This pattern suggests that when a rationale does not explicitly invoke the "illusion" frame, models tend to answer stably from image evidence; in contrast, once a model explicitly frames an item as an "illusion" in its rationale, its decision is more likely to deviate from the visually correct option, consistent with a shift toward prior-driven or template-consistent reasoning under the Illusion-Illusion condition. We summarize this gap as the illusion-frame penalty (Δ) and report the level-averaged value $\overline{\Delta}$ in Table 2.

IlluScale

Once a model falls into a cognitive trap, how strong must visual evidence become to wake it up? To quantify VLMs' ability to *exit* a prior-driven trap, we conduct experiments on the **IlluScale** subset of IlluQuant and report mean accuracy under progressively strengthened physical differences.

As shown in Table 3 and Fig. 5, the human baseline achieves only 45.36% accuracy at 1.0× (Illusion), indicating that human judgments are readily biased by the illusion configu-

Table 3: Accuracy rates at different scale intensities. All metrics are reported in percentage (%). Note that IlluScale-2× is computed on the full IlluScale subset (including three additional illusion families not in IlluLadder), whereas IlluLadder-L5 is computed on the 8 shared families only; hence the accuracies in Table 2 and this table are not directly comparable.

Model	**1.0×**	**1.1×**	**1.2×**	**1.3×**	**1.5×**	**2×**	**3×**	**5×**
Gemini-3-Pro	100.00	0.00	3.64	5.45	5.45	34.55	58.18	56.36
GPT-5.2	100.00	1.82	12.73	3.64	9.09	32.73	56.36	56.36
Claude-Opus-4.5	82.35	25.93	20.00	25.45	27.78	37.04	36.36	45.45
Qwen3-VL:32B	78.18	14.55	21.82	29.09	38.18	52.73	63.64	67.27
Human	45.36	56.79	75.89	90.32	93.68	99.06	100.00	100.00

Table 4: Psychometric fits and derived metrics for (4PL fitted over 1.1× → 5× with $x = \log(s)$). $\mathrm{ET}_{\mathrm{dyn75}}$ is the 75% dynamic-range threshold; $\mathrm{AUC}_{\log}$ is the normalized area under the curve from log(1.1) to log(5); $p(s = 5)$is the predicted accuracy at 5×. Confidence intervals are obtained by parametric bootstrap ($B = 300$). Claude suffers from an extremely wide threshold CI due to its flat curve and parameters being close to the boundary.

Model	**Ac.**(%) (1.0×)	**Ceiling** b(%)	$\mathrm{ET}_{\mathrm{dyn75}}$	$\mathrm{AUC}_{\log}$(%)	$p(s = 5)$(%)
Gemini-3-Pro	100.00	58.45	2.235 [1.913, 2.838]	36.2 [30.5, 40.6]	58.4 [50.0, 70.7]
GPT-5.2	100.00	71.91	2.610 [2.051, 3.503]	40.8 [35.9, 46.7]	70.8 [60.0, 81.2]
Claude-Opus-4.5	82.35	51.34	2.913 [1.842, 43.216]	34.9 [28.9, 41.9]	44.7 [35.8, 56.5]
Qwen3-VL:32B	78.18	66.42	1.815 [1.499, 2.928]	52.0 [46.5, 57.4]	66.2 [57.5, 76.9]
Human	45.36	100.00	1.192 [1.137, 1.353]	95.8 [94.6, 96.6]	100.0 [99.6, 100.0]

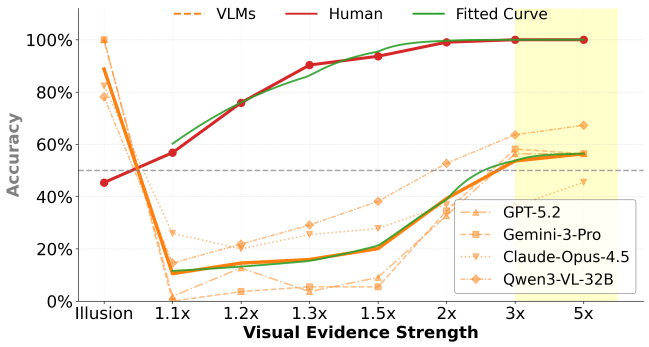

Figure 5: Accuracy as a function of visual evidence strength.

ration. However, as the true physical difference is slightly amplified, human accuracy rises rapidly: it reaches 90.32% at 1.3× and saturates at 1.0 after 3×. This monotonic, fast-saturating pattern is characteristic of psychophysical performance curves, supporting that our manipulation of "evidence strength" reliably increases perceptually accessible visual evidence rather than merely introducing noise or item-specific bias.

In contrast to humans' "low start, fast updating", VLMs often achieve high accuracy at 1.0× but exhibit a cliff-like drop in the weak-evidence regime, followed only later by a gradual recovery as evidence increases (1.1× → 5×). This asymmetric pattern of "high start, hard updating" suggests that high scores at 1.0× do not reflect more sensitive perception; instead, these results are more consistent with prior-dominated behavior: small increases in visual evidence are insufficient to reverse the response, and only larger amplifications induce a switch toward evidence-consistent judgments. Because the chance level in 2AFC is 0.5, traditional psychometric modeling often fixes the lower asymptote at $\gamma = 0.5$. However, in our experiment the model frequently exhibits $p < 0.5$ in the 1.1× → 5× range, implying that the standard form with a hard lower bound of 0.5 is mathematically incapable of capturing the observed data. Therefore, over 1.1× → 5× we fit proportion-correct using a four-parameter logistic (4PL) psychometric function under a binomial likelihood. Importantly, we do not log-transform accuracy itself; only the stimulus axis is log-transformed, with x = log(s).

$$k_s \sim \mathrm{Binomial}(n, p(x_s)), x_s = \log(s) \quad (1)$$

$$p(x) = a + (b - a) \times \sigma(\beta_0 + \beta_1 x), \sigma(z) = \frac{1}{1 + \exp(-z)} \quad (2)$$

with constraints $0 \le a \le 0.5, 0.5 \le b \le 1, \beta_1 \ge 0$ to ensure interpretability of the parameters.

In addition, we exclude the canonical condition (1.0×) from the psychometric fit. The 1.0× condition often corresponds to classic / in-distribution stimuli and may be affected by template triggering or in-distribution matching, thereby violating the assumption of a monotonic "evidence-strength axis". Accordingly, 1.0× is reported separately as CanonicalAcc.

Defining the threshold at $p = 0.5$ corresponds to chance performance in 2AFC and does not carry the psychophysical meaning of a "discrimination threshold". To obtain a threshold that remains comparable across different floors and ceilings, we adopt a dynamic-range threshold, ETdyn75, defined as the evidence strength required for the curve to reach

75% of its progress from floor to ceiling:

$$p^* = a + 0.75 \times (b - a),\ p(x^*) = p^*,\ \text{ETdyn75} = \exp(x^*) \quad (3)$$

Under the 4PL parameterization, this is equivalently the point where the latent sigmoid term reaches 0.75 and uncertainty is quantified via a parametric bootstrap: at each scale factor, we generate $k'_s \sim \text{Binomial}(n, \hat{p}(x_s))$ using $\hat{p}(x_s)$, refit the model, and repeat for $B = 300$ replicates; the 95% confidence interval is constructed from the 2.5th and 97.5th percentiles.

Table 4 reports the fitted ceiling b, the dynamic threshold ETdyn75 with 95% CIs, the normalized area under the curve on the log-evidence axis (AUClog), and the predicted performance at 5×, $p(s = 5)$. Overall, the human curve exhibits a high ceiling (near 100%) and a low threshold (approximately 1.19x), while maintaining a large area across the evidence range (AUC close to 100%). In contrast, the models show substantially lower ceilings than humans (near 51%–72%) and markedly smaller AUC values; even at 5×, predicted accuracy spans roughly 45%–71%, i.e., remains near chance for several models and stays far below human-level performance. These results indicate that humans reach reliably high accuracy with mild evidence amplification (around 1.2×) and rapidly approach the ceiling, whereas multimodal models—despite often outperforming humans at the canonical condition (1.0×)—are highly fragile to slight geometric perturbations (1.1×), exhibiting systematic below-chance errors, and subsequently recover only partially with increasing evidence while remaining capped by a limited ceiling.

By adopting a 4PL psychometric function that allows floor < 0.5 and using the dynamic threshold ETdyn75, we can statistically disentangle and quantify "recovery capacity" (AUClog), and the "attainable upper limit" (ceiling). This enables a more rigorous characterization of the resistance to updating visual evidence under the influence of semantic priors or template attraction.

Intervention Method

In the preceding experiments, we observe that VLMs often exhibit a semantic-prior-over-visual-evidence pattern on Illusion-Illusion stimuli. To mitigate this effect, we propose **MdCoT** (Modality Decoupling Chain of Thought), a two-stage prompting strategy inspired by CoT (Shao et al. (2024), Wei et al. (2022), and Z. Zhang et al. (2024)) that explicitly separates textual reasoning from visual grounding. Concretely, in Stage 1 we withhold the image and ask the model to describe a step-by-step procedure for solving the task given only the question (i.e., "how to decide"). In Stage 2 we then provide the image and require the model to answer by following its own previously produced procedure. This design aims to reduce immediate template retrieval triggered by the visual configuration and to encourage an explicit evidence-checking process when the image becomes available.

Building on this benchmark, we evaluate MdCoT, a two-stage prompting strategy that consistently improves performance on Illusion-Illusion stimuli. This improvement is consistent with reduced template-consistent responding, although the present results do not isolate the underlying mechanism. As shown in Table 5, the relative gain Δ exceeds 30% for all four models. However, even after MdCoT, all models remain far below human performance at the matched 2× condition (99.06%), indicating that this particular intervention alleviates the effect but does not eliminate it.

Table 5: Effect of prompting intervention on Illusion-Illusion accuracy. All metrics are reported in percentage (%). Illu-Illu corresponds to IlluScale 2× over all 11 illusion. Human(Ref) is shown for reference only and is not included in MdCoT or Δ. $\Delta(\%) = \frac{\text{Acc}_{\text{MdCoT}} - \text{Acc}_{\text{Illu-Illu}}}{\text{Acc}_{\text{Illu-Illu}}} \times 100\%$.

Model	Illu	Illu-Illu	MdCoT	Δ
Gemini-3-Pro	100.00	34.55	47.06	↑36.21
GPT-5.2	100.00	32.73	58.18	↑77.76
Claude-Opus-4.5	82.35	37.04	52.73	↑42.36
Qwen3-VL:32B	78.18	52.73	69.09	↑31.03
Human(Ref.)	45.36	99.06	-	-

The core mechanism of MdCoT can be interpreted as a form of chain-of-thought with modality decoupling. By withholding the image during the initial reasoning phase,it may reduce immediate bottom-up priming effects where visual configurations trigger semantic templates. This intervention appears to reduce the model's tendency to automatically retrieve illusion templates. However, the residual performance gap after MdCoT suggests that this particular reasoning-based intervention is insufficient to fully eliminate the influence of semantic priors, stronger visual verification or additional grounding mechanisms may be needed.

Conclusion

We treat large vision–language models as artificial subjects and examine a failure mode: even as visual evidence is gradually strengthened, models can remain pulled toward retrievable semantic templates/priors, leading to systematic biases. To study this, we introduce IlluQuant, where IlluScale modulates visual evidence strength and IlluLadder modulates template similarity, turning the classic discrete "illusion vs. control" evaluation into psychophysics-style curve analysis. Based on these curves, we derive a set of robust, curve-based quantitative measures (e.g., curve shape, dynamic range, and threshold points) to compare how different models transition between semantic prior dominance and evidence-based correction, and how recoverable they are as evidence increases. Building on this benchmark, we propose MdCoT, a two-stage prompting strategy that weakens template attraction by first externalizing the decision procedure and then enforcing evidence checking; experiments show consistent gains over the Illusion–Illusion baseline but do not fully eliminate prior lock-in, suggesting that future work should incorporate stronger visual verification and mechanisms for calibrating the balance between priors and evidence during inference and training to achieve more reliable multimodal grounding.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants W2512008, 62572500, and 62272498; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515050001; and in part by the Open Project Program of Guangxi Key Laboratory of Digital Infrastructure (Grant Number: GXDIOP2024015).

References

Anthropic. (2025). Claude opus 4.5 model card. https : / / www - cdn . anthropic . com / bf10f64990cfda0ba858290be7b8cc6317685f47.pdf

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., . . . Zhu, K. (2025). Qwen3-vl technical report. https://arxiv.org/abs/2511.21631

Day, R. (1984). The nature of, perceptual illusions. *Interdisciplinary Science Reviews*, *9*(1), 47–58.

Eagleman, D. M. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience*, *2*(12), 920–926.

Gomez-Villa, A., Martin, A., Vazquez-Corral, J., & Bertalmío, M. (2019). Convolutional neural networks can be deceived by visual illusions, 12309–12317.

Google. (2025). Gemini 3 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1358), 1121–1127.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al. (2024). Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 14375–14385.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annu. Rev. Psychol.*, *55*(1), 271–304.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current opinion in neurobiology*, *13*(2), 150–158.

Knill, D. C., & Richards, W. (1996). *Perception as bayesian inference*. Cambridge University Press.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2024). Mitigating object hallucinations in large vision-language models through visual contrastive decoding, 13872–13882.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating object hallucination in large vision-language models. https://arxiv.org/abs/2305.10355

OpenAI. (2025). Gpt 5.2 model card. https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf

Pham, N., & Schott, M. (2024). H-pope: Hierarchical polling-based probing evaluation of hallucinations in large vision-language models. https://arxiv.org/abs/2411.04077

Prins, N., et al. (2016). *Psychophysics: A practical introduction*. Academic Press.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Shahgir, H. S., Sayeed, K. S., Bhattacharjee, A., Ahmad, W. U., Dong, Y., & Shahriyar, R. (2024). Illusionvqa: A challenging optical illusion dataset for vision language models. *arXiv preprint arXiv:2403.15952*.

Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., & Li, H. (2024). Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. https://arxiv.org/abs/2403.16999

Shinozaki, T., Watahiki, A., Nishida, S., Yanaka, H., et al. (2025). Do large vision-language models distinguish between the actual and apparent features of illusions? *arXiv preprint arXiv:2506.05765*.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision research*, *35*(17), 2503–2522.

Ullman, T. (2024). The illusion-illusion: Vision language models see illusions where there are none. *arXiv preprint arXiv:2412.18613*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, *63*(8), 1293–1313.

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: Ii. bootstrap-based confidence intervals and sampling. *Perception & psychophysics*, *63*(8), 1314–1329.

Zhang, Y., Zhang, Z., Wei, X., Liu, X., Zhai, G., & Min, X. (2025). Illusionbench: A large-scale and comprehensive benchmark for visual illusion understanding in vision-language models. *arXiv preprint arXiv:2501.00848*.

Zhang, Z., Zhang, A., Li, M., hai zhao, Karypis, G., & Smola, A. (2024). Multimodal chain-of-thought reasoning in language models. https://openreview.net/forum?id=gDlsMWost9