SEEING IS BELIEVING? A BENCHMARK FOR MULTIMODAL LARGE LANGUAGE MODELS ON VISUAL ILLUSIONS AND ANOMALIES

Anonymous authors
Paper under double-blind review

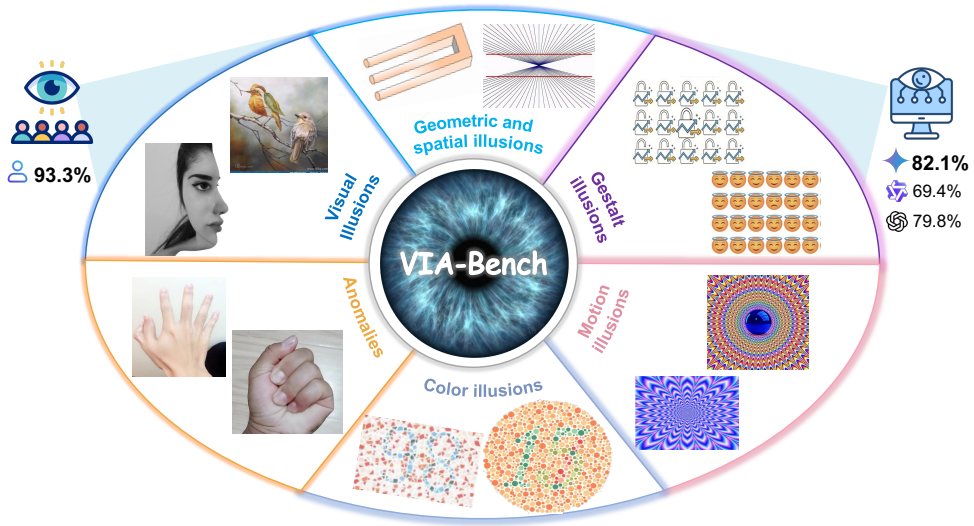


Figure 1: Overview of VIA-Bench. The benchmark includes six categories: color illusions, motion illusions, gestalt illusions, geometric and spatial illusions, general visual illusions, and visual anomalies. These perceptual-conflict scenarios probe whether MLLMs remain grounded in *objective physical reality*. On VIA-Bench, humans achieve 93.30% average accuracy, whereas the best MLLM reaches 82.10%.

ABSTRACT

Multimodal Large Language Models (MLLMs) now match or exceed human performance on many vision-language benchmarks. However, these evaluations typically rely on standard in-distribution data, leaving the robustness of MLLMs largely unexamined under prior conflict scenarios. To address this gap, we introduce VIA-Bench, a diagnostic benchmark that probes whether MLLMs remain grounded in objective physical reality under such conflicts. VIA-Bench covers six categories: color illusions, motion illusions, gestalt illusions, geometric and spatial illusions, general visual illusions, and visual anomalies. Through careful human-in-the-loop review, we construct over 1K high-quality question-answer pairs supporting both multiple-choice and open-ended evaluation. Extensive evaluation of 30 state-of-the-art MLLMs, including proprietary, open-source, and reasoning-enhanced models, uncovers significant vulnerabilities. Beyond aggregate accuracy, we further find that general-purpose Chain-of-Thought (CoT) prompting offers negligible robustness, often yielding "brittle mirages" where the model's logic collapses under illusory stimuli. Even task-specific CoT helps only marginally. To better characterize these failures, we introduce an error taxonomy that disentangles visual misperception, self-negation, and overthinking. These findings expose a structural vulnerability of current MLLMs under perceptual conflict and motivate future work on visually grounded reasoning. The benchmark data and code will be released.

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

1 INTRODUCTION

Recent years have witnessed the prominent role of Multimodal Large Language Models (MLLMs) across visual domains (Achiam et al., 2023; Gemini Team, 2025; OpenAI, 2026b; Anthropic, 2026; Team et al., 2026). Frontier proprietary systems have reached human-level performance on a wide spectrum of visual benchmarks (Yue et al., 2025). Beyond basic perception, these models are evolving into sophisticated reasoners with multi-step Chain-of-Thought (CoT) reasoning (Guo et al., 2025a; OpenAI, 2025b), and increasingly into visual agents that think with images and act in the physical world (Xu et al., 2025; Zhang et al., 2025; Team et al., 2025; Gemini Team, 2026).

To assess these increasingly powerful capabilities, numerous benchmarks have been proposed from various perspectives. Most of them, however, still emphasize standard visual contexts and stepwise reasoning (Wang et al., 2025; Zeng et al., 2025). Parallel efforts probe counterfactual logic or visual puzzles (Komanduri et al., 2025; Song et al., 2025; Gao et al., 2025). Yet, emerging evidence consistently shows that many existing multimodal benchmarks lack sufficient difficulty and discriminative power for frontier models (Roberts et al., 2025b), underscoring the gap between superficial benchmark success and genuine visual intelligence (Chen et al., 2026; Lin et al., 2026).

More precisely, these evaluations rarely stress-test models under *atypical conditions* such as cognitive illusions, perceptual conflicts, and intuition-breaking scenarios, where reliance on canonical priors can contradict visual evidence (Sun et al., 2026). For instance, models often blindly predict "five" for a six-fingered hand, prioritizing a learned prototype over what is shown (see Fig. 1). As MLLMs increasingly saturate standard benchmarks (Yue et al., 2025), this gap becomes more consequential and motivates two coupled questions: *(i) Under perceptual conflict, can MLLMs remain grounded in the* ***objective physical reality*** *of the stimulus? (ii) When they fail, does the breakdown originate from perception, reasoning, or commonsense priors?*

These scenarios (see Fig. 1) are particularly well suited for probing these questions, as they expose failure modes that standard evaluations often miss. In other words, such illusions fundamentally challenge the intuition that *"seeing is believing"*. Crucially, our goal is not to test whether MLLMs experience illusions like humans, but whether they can answer based on the objective physical reality of the image, even when its appearance or priors push elsewhere. Handling these out-of-distribution stimuli thus forces MLLMs to weigh *what is seen* (visual evidence) against *what is known* (internal priors), and ultimately deliver a physically grounded answer through deliberate reasoning.

To this end, we present `VIA-Bench`, a comprehensive and sufficiently challenging benchmark targeting visual illusions and anomalies. As depicted in Fig. 1, `VIA-Bench` spans diverse specialized categories: color illusions (CI), motion illusions (MI), gestalt illusions (GI), geometric and spatial illusions (GSI), general visual illusions (VI), and visual anomalies (VA)[1]. Despite having physical ground truth, these stimuli remain consistently challenging for SOTA MLLMs.

We construct `VIA-Bench` via careful human-in-the-loop review. For each scene, we focus on two key aspects. First, the *"What"*: the *objective physical fact* of the image (*e.g.*, identical pixel values, true counts). Second, the *"How"*: how to probe model reasoning so that the answer must be grounded in that physical fact rather than in appearance or canonical priors. To ensure data quality and diversity, we adopt a multi-stage pipeline to construct question–answer (QA) pairs (see Fig. 2).

To analyze model behavior, we conduct extensive evaluations on 30 MLLMs. We perform a systematic analysis and deliver three core insights. *First*, MLLM performance varies substantially across models and categories, and no model is uniformly robust. *Second*, general-purpose CoT prompting often degrades accuracy on this benchmark, and task-specific prompts bring only marginal gains. *Third*, failures originate from Visual Misperception ($\sim$58.8%), Self-Negation ($\sim$4.5%), and Overthinking ($\sim$36.8%), providing a diagnostic view beyond aggregate accuracy. This work highlights the following contributions:

- **The `VIA-Bench` Dataset:** We develop a challenging benchmark targeting visual illusions and anomalies, comprising six specialized categories and 1,004 meticulously curated QA pairs.
- **Extensive Benchmarking:** We evaluate 30 frontier MLLMs and find substantial performance variation across models and categories. No single model performs consistently well across the benchmark. Even leading models remain weak in some categories.

[1] In Appendix B, we provide detailed descriptions and analyses of these categories.

- **The CoT Paradox and Failure Diagnosis:** To better guide future research and development, we show that general-purpose CoT provides little robustness under perceptual conflict, and that task-specific prompting yields only marginal gains. Failures originate from visual misperception, self-negation, and overthinking.

2 Preliminaries and Problem Formulation

The primary objective of VIA-Bench is to evaluate the capacity of MLLMs to reconcile conflicting visual stimuli with common-sense priors. We focus on scenarios where standard reasoning often fails due to the deceptive nature of the visual input.

2.1 Formal Task Definition

Let $\mathcal{I}$ denote the space of images and $\mathcal{T}$ denote the space of natural language. For a given instance in VIA-Bench, we define the input as a tuple $(\mathbf{x}, q, O, i)$, where:

- $\mathbf{x} \in \mathcal{I}$ is an illusory or anomalous image;
- $q \in \mathcal{T}$ is a task-specific question designed to probe the model's perception;
- $O = \{o_1, \ldots, o_m\}$ is a set of m candidate options ($m \geq 2$), exactly one of which corresponds to the ground-truth answer y;
- $i \in \mathcal{T}$ is a formatting instruction (*e.g.*, "Please provide your answer in the following format: Answer: []") to ensure deterministic parsing.

An MLLM, denoted as a parameterized function f_θ, maps the concatenated visual and textual inputs to a predicted answer a:

$$a = f_\theta(\mathbf{x}, q, O, i).$$

In our CoT analysis (Sec. 4.3), we further augment the instruction i with system-level CoT prompting, which encourages the model to generate intermediate reasoning steps before selecting a. We also consider an *open-ended* variant where O is omitted: the model generates a free-form answer a judged against y via LLM-as-judge.

2.2 Evaluation Criteria

A central design choice of VIA-Bench is that the ground-truth answer y is anchored in the *objective physical reality* of the image $\mathbf{x}$ rather than in subjective human percepts. Concretely:

- For **CI**, correctness is defined by the objectively encoded visual content (*e.g.*, identical RGB values, or the number or word embedded in the image), regardless of the surrounding context (see Figs. 4 and 8 for examples).
- For **MI**, every image is physically *static*, despite apparent motion (see Figs. 1 and 9 for examples).
- For **GI**, the answer is the row and column of the odd one out (see Figs. 4 and 10 for examples).
- For **GSI**, the label refers to the geometric or topological fact, such as whether lines are parallel or whether a Kanizsa-type figure has a real hole (see Figs. 1 and 11 for examples).
- For **VI**, labels describe the real physical scene rather than its apparent layout, *e.g.*, nobody actually stands on the hand under forced perspective (see Figs. 4 and 12 for examples).
- For **VA**, the ground truth is the actual number of fingers or toes, even when it departs from the usual count (see Figs. 4 and 13 for examples).

Under this protocol, the ground truth is determined by the image rather than by textual or common-sense priors in q, and the question of whether a model phenomenologically "perceives" the illusion is decoupled from whether it ultimately produces a physically grounded answer.

3 The VIA-Bench Dataset

We introduce VIA-Bench, a diagnostic evaluation suite designed to probe the perceptual robustness of MLLMs. The dataset consists of 1,004 high-quality QA pairs, meticulously curated to isolate failure modes where internal model priors conflict with true visual evidence.

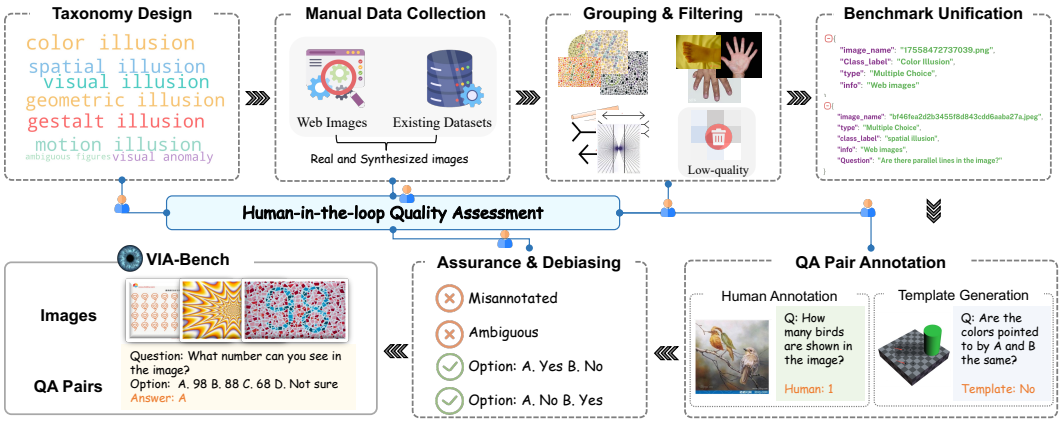


Figure 2: Benchmark construction pipeline. The workflow progresses from data collection to unification, annotation, and debiasing, forming VIA-Bench. To ensure quality, we apply human-in-the-loop assessment at all key stages.

3.1 Dataset Construction Pipeline

As illustrated in Fig. 2, we employ a systematic, multi-stage, human-in-the-loop pipeline to ensure consistently high-quality and diverse data.

Taxonomy and Categorization. Our taxonomy builds on classic illusion families in perception research (Eagleman, 2001) and on MLLM failures reported by prior benchmarks (Zhang et al., 2023; Shahgir et al., 2024; Guan et al., 2024) and seen in our pilot tests. In addition, recent benchmarks show that even frontier MLLMs still struggle with basic visual perception (Chen et al., 2026; Lin et al., 2026). Based on these observations, our study covers six primary categories: color illusions (CI), motion illusions (MI), Gestalt illusions (GI), geometric and spatial illusions (GSI), general visual illusions (VI), and visual anomalies (VA). Detailed descriptions of each category are given in Appendix B, and their evaluation criteria in Sec. 2.2. We treat all stimuli as *perceptual-conflict probes* under a unified physical-grounding target.

Data Curation and Unification. To better reflect real-world visual content, we manually collect 974 images (97.0%) from public web sources using Google Images, Baidu Images, Quark, and Shenma, recording the original URL for each. We supplement this collection with 30 synthetic CI images (3.0%) from the Turing Eye Test (TET) (Gao et al., 2025), whose controlled, perception-focused stimuli align with our study and directly test the perceptual abilities of MLLMs. Each retained image corresponds to one QA pair. Each candidate image underwent a rigorous cleaning process, filtered based on resolution, clarity, and non-ambiguity. To facilitate scalable evaluation, all metadata is unified into a schema before the annotation phase.

QA Pair Annotation. We formulate VIA-Bench as a multiple-choice task to allow for deterministic evaluation. Annotation focuses on three critical dimensions:

1) Inquisitive Precision: Questions are tailored to specific illusion types instead of being generic (*e.g.*, "What number is embedded in the pattern?" for CI). For spatial comparisons, we provide localized visual cues like arrows or letter labels to minimize grounding errors.

2) Distractor Design: We manually craft plausible distractors that encapsulate common machine-learning shortcuts or stereotypical priors. Crucially, we include a "Not Sure" option in every set to gauge model uncertainty and overconfidence, though it is never the ground-truth label.

3) Ground-Truth Verification: All QA pairs are manually constructed. For each image, a primary annotator first identifies a verifiable physical fact and then writes the question, answer options, and answer key. A second annotator cross-verifies all components. Disagreements are resolved through discussion, and unresolved items are discarded.

Quality Assurance and Debiasing. To mitigate shortcut learning, we apply two primary debiasing techniques: (1) randomizing the permutation of options to eliminate positional bias, and (2) for binary questions, choosing a phrasing that suppresses prior-aligned shortcut cues (*e.g.*, "Is the image moving or pulsing?" rather than "Is the image static and still?", see Appendix J).

3.2 Dataset Statistics

Fig. 3 shows the statistical summary of VIA-Bench.

Distribution: As shown in Fig. 3(a), the dataset maintains a relatively balanced distribution across categories, with VA (23.8%) and GI (11.7%) representing the upper and lower bounds.

Capability Coverage: The benchmark spans seven capability dimensions (Fig. 3(b)), testing the intersection of low-level perception and high-level reasoning.

Structural Diversity: Question lengths (Fig. 3(c)) vary significantly, reflecting the complexity of the required reasoning. Furthermore, the dataset includes ultra-high-resolution images (up to 8334×2501), providing fine-grained visual content for evaluating current and future models.

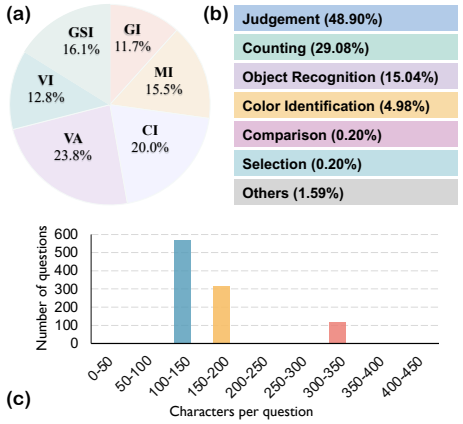


Figure 3: Statistical characterization of VIA-Bench. **(a)** Distribution across the six primary categories of illusions and anomalies. **(b)** Mapping of the dataset to specific perceptual and cognitive capabilities. **(c)** Empirical distribution of question lengths.

4 Evaluation on VIA-Bench

Benchmark Models. We conduct an extensive evaluation covering 30 MLLMs on VIA-Bench, encompassing proprietary, open-source, and reasoning-enhanced models.

- **Proprietary Systems.** We include Gemini-3.8-Flash (Gemini Team, 2026), GPT-6-astra (OpenAI, 2026b), GPT-5.6-sol (OpenAI, 2026a), Claude-opus-5 (Anthropic, 2026), Gemini-3-pro (Gemini Team, 2025), OpenAI o4-mini (OpenAI, 2025b), Gemini-2.5-pro (Comanici et al., 2025), GPT-5-mini (OpenAI, 2025a), GPT-4o-2024-11-20 (Hurst et al., 2024), ChatGPT-4o-latest (Hurst et al., 2024), and GPT-5-chat-latest (OpenAI, 2025a). These systems, primarily from OpenAI, Google, and Anthropic, represent the current frontier.
- **Open-source Models.** Our open-source suite spans the Kimi, DeepSeek, Qwen-VL, and InternVL series, including Kimi-K3 (Team et al., 2026), DeepSeek-V4.1-Flash (DeepSeek-AI, 2026), Qwen-vl-max-latest (Bai et al., 2023), Qwen2.5-VL-Instruct (3B, 7B, 72B) (Bai et al., 2025), and InternVL3.5 (without thinking) (8B, 38B) (Wang et al., 2025). These models can be deployed locally and constitute strong open-source baselines.
- **Reasoning-enhanced Models.** Recent models generate long CoT when producing answers, providing explicit reasoning traces. They are typically trained with reinforcement learning to enhance their reasoning ability. Our evaluation covers recent releases, including Qwen3.8-max-0902 (Qwen Team, 2026), OpenAI o3 (OpenAI, 2025b), Claude-opus-4.1-20250805 (Anthropic, 2025), Claude-sonnet-4-20250514 (Anthropic, 2025), Claude-3.5-sonnet-20241022 (Anthropic, 2024), InternVL3.5 (with thinking) (8B, 38B, 30B-A3B) (Wang et al., 2025), GLM-4.5V (Hong et al., 2025), and Qwen3-VL-Thinking (30B-A3B, 235B-A22B) (Yang et al., 2025). Where explicit rationales are available, we analyze whether the revealed CoT reflects deliberate reasoning.

4.1 Evaluation Metrics and Details

We report standard accuracy $a = n/N$, where n is the number of correctly answered questions and N is the total. Following (Fu et al., 2025; Zou et al., 2025; Roberts et al., 2025a), we include explicit "output format instructions" (*e.g.*, `Answer: [A/B/C/D]`) to ensure deterministic parsing. We adopt two evaluation protocols:

Match: rule-based regex matching to extract and verify the final option label.

Judge: LLM-as-a-Judge (Zheng et al., 2023; Chen et al., 2024a) using GPT-4.1-mini (Achiam et al., 2023) as arbiter.

Table 1: Evaluation on VIA-Bench. We employ two evaluation protocols–Pattern Match (*i.e.*, Match) and LLM-as-a-Judge (*i.e.*, Judge)–to assess 30 MLLMs across six categories. They are grouped into proprietary, open-source, and reasoning-enhanced models for the experiment. Each model is run 5 times, and the results are averaged to reduce randomness. Within each group, dark green indicates the best result and light green denotes the second-best result. The results show that no single model performs well across all categories (*e.g.,* even Gemini-3.8-Flash and GPT-6-astra achieve only 62.97% on VA and 59.69% on VI, respectively).

Method	Avg.	Rank	VA		CI		MI		GI		GSI		VI	
			Match	Judge	Match	Judge	Match	Judge	Match	Judge	Match	Judge	Match	Judge
Blind Evaluation														
Random Choice	29.13	–	28.45	28.45	28.36	28.36	32.69	32.69	23.08	23.08	29.63	29.63	32.56	32.56
GPT-4-Turbo (Achiam et al., 2023)	39.61	–	2.85	2.85	25.87	25.87	87.95	87.95	35.04	35.04	61.11	61.11	24.81	24.81
Proprietary Systems														
Gemini-3.8-Flash (Gemini Team, 2026)	82.10	1	62.97	62.97	75.12	74.88	100.00	100.00	90.60	90.60	95.06	95.06	68.99	68.99
GPT-6-astra (OpenAI, 2026b)	79.79	2	69.67	69.67	76.87	76.87	100.00	100.00	82.05	82.05	90.43	90.43	59.69	59.69
GPT-5.6-sol (OpenAI, 2026a)	57.23	4	45.82	45.82	26.12	26.12	66.67	66.67	66.24	66.24	84.26	84.26	54.26	54.26
Claude-opus-5 (Anthropic, 2026)	51.29	7	40.38	40.38	53.23	53.23	0.00	0.00	68.80	68.80	86.42	86.42	58.91	58.91
Gemini-3-pro (Gemini Team, 2025)	69.23	3	43.51	44.49	53.33	56.82	69.87	99.36	79.32	83.76	81.23	89.88	61.24	67.91
OpenAI o4-mini (OpenAI, 2025b)	55.25	5	24.02	24.02	27.46	28.16	94.87	94.87	26.84	26.84	97.16	97.16	60.31	61.24
Gemini-2.5-pro (Comanici et al., 2025)	55.01	6	40.42	38.74	28.46	28.46	46.54	46.54	65.81	65.81	83.95	83.95	65.74	65.74
GPT-5-mini (OpenAI, 2025a)	49.54	8	26.95	26.95	49.25	49.25	25.13	25.13	39.15	38.63	95.19	95.19	61.86	61.86
GPT-4o-2024-11-20 (Hurst et al., 2024)	43.44	9	35.31	35.31	29.25	29.55	20.13	20.13	22.39	22.39	90.86	90.86	60.93	64.19
ChatGPT-4o-latest (Hurst et al., 2024)	36.61	10	2.34	3.85	11.94	12.24	35.51	35.51	26.67	16.58	92.10	91.85	55.81	54.88
GPT-5-chat-latest (OpenAI, 2025a)	35.23	11	0.25	2.26	7.36	8.06	47.18	47.05	17.78	26.67	85.54	88.15	43.88	48.53
Open-source Models														
Kimi-K3 (Team et al., 2026)	59.59	1	43.51	43.51	46.27	46.27	62.18	62.18	69.23	69.23	82.10	82.10	54.26	54.26
DeepSeek-V4.1-Flash (DeepSeek-AI, 2026)	52.70	3	29.08	29.71	39.30	40.30	39.74	40.06	61.97	64.96	83.95	83.95	59.69	59.69
Qwen-vl-max-latest (Bai et al., 2023)	46.81	6	25.52	25.52	34.23	33.93	37.95	37.82	43.42	42.74	78.89	78.89	61.40	61.40
Qwen2.5-VL-3B-Instruct (Bai et al., 2025)	34.84	8	18.83	19.00	24.58	24.58	13.21	13.21	32.14	32.14	76.17	76.17	44.03	44.03
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	48.00	5	22.93	22.93	22.09	22.09	78.97	78.97	34.53	34.19	76.30	76.30	53.33	53.33
Qwen2.5-VL-72B-Instruct (Bai et al., 2025)	49.68	4	19.00	19.00	29.15	29.15	56.03	56.03	48.72	48.55	82.47	82.47	62.79	62.79
InternVL3.5-8B (w/o thinking) (Wang et al., 2025)	52.73	2	26.36	26.36	16.92	16.92	82.69	82.69	52.99	52.99	83.95	83.95	53.49	53.49
InternVL3.5-38B (w/o thinking) (Wang et al., 2025)	37.36	7	24.27	24.27	21.89	21.09	4.49	4.49	46.15	46.15	73.46	74.57	54.26	53.18
Reasoning-enhanced Models														
Qwen3.8-max-0902 (Qwen Team, 2026)	69.38	1	64.64	64.64	78.86	79.10	43.59	43.59	77.35	77.35	87.35	87.35	64.34	64.34
OpenAI o3 (OpenAI, 2025b)	60.26	2	23.21	22.01	51.84	51.84	90.51	90.51	36.75	36.75	92.47	92.35	67.75	67.13
Claude-opus-4.1-20250805 (Anthropic, 2025)	53.52	6	21.84	20.67	31.74	31.24	90.38	90.26	53.16	53.16	75.68	76.30	48.99	48.84
Claude-sonnet-4-20250514 (Anthropic, 2025)	46.61	8	21.09	21.09	23.98	22.59	41.40	41.28	61.20	61.20	78.64	78.64	54.26	53.95
Claude-3.5-sonnet-20241022 (Anthropic, 2024)	39.46	10	25.02	24.44	28.66	28.96	20.38	20.38	46.32	46.84	81.45	81.85	34.81	34.47
InternVL3.5-8B (w/ thinking) (Wang et al., 2025)	52.85	7	23.68	16.15	33.13	20.70	95.90	92.95	53.68	49.06	73.46	67.28	51.63	56.59
InternVL3.5-38B (w/ thinking) (Wang et al., 2025)	37.60	11	23.35	19.41	24.58	22.09	20.38	18.46	48.89	39.49	72.59	69.26	42.95	49.77
InternVL3.5-30B-A3B (w/ thinking) (Wang et al., 2025)	54.58	5	19.00	15.98	32.44	32.44	82.82	85.00	56.58	48.21	86.17	84.07	53.95	58.29
GLM-4.5V (Hong et al., 2025)	45.57	9	21.26	19.00	47.26	47.56	13.21	12.69	47.01	49.57	81.11	80.25	62.17	65.74
Qwen3-VL-30B-A3B-Thinking (Yang et al., 2025)	57.07	4	18.74	15.48	40.10	39.90	96.28	96.28	45.47	45.47	83.83	83.09	59.38	60.78
Qwen3-VL-235B-A22B-Thinking (Yang et al., 2025)	59.82	3	23.51	20.08	37.61	36.42	95.51	95.51	58.29	56.41	83.58	82.96	62.33	65.58
Human Evaluation														
Human	93.30	–	87.45	87.45	85.57	85.57	100.00	100.00	96.58	96.58	98.75	98.75	91.47	91.47

To reduce randomness, we report the average over five independent runs (per-run results in Appendix D); matching rules and judging prompts are detailed in Appendices G and H. We set the temperature to 0.8 where supported, otherwise using models' default or recommended parameters (*e.g.*, OpenAI o3, InternVL3.5); a temperature ablation is in Appendix F. Most models are accessed via API; locally deployed open-source models run on 8 H20 GPUs with tensor parallelism.

4.2 Main Results

Table 1 shows overall model performance on VIA-Bench. Our key observations are as follows:

Blind Evaluation. For reference, random choice yields an average accuracy of 29.13% on VIA-Bench. However, text-only GPT-4-Turbo (*i.e.*, with vision disabled) demonstrates surprisingly high accuracy, notably on motion illusions (87.95%) and geometric and spatial illusions (61.11%). This suggests that, even without visual evidence, the model leverages internal linguistic priors to reason about impossible image states (*e.g.*, a static image depicting motion). Table 2 shows that this shortcut is *category-specific*: the model defaults to prior-consistent answers on MI/GSI (*e.g.*, answering "No" in 84.6% of MI cases, corresponding to a static image), while abstaining ("Not Sure") on 77.0% of VA cases.

Table 2: Answer distribution of the text-only (vision-disabled) GPT-4-Turbo baseline across categories. The pattern is *category-specific*: VA is dominated by "Not Sure" (uncertainty), whereas MI/GSI are dominated by prior-consistent "No" (linguistic shortcut).

Category	Top-1 (%)	Top-2 (%)	Top-3 (%)
VA	Not Sure (77.0)	5 (11.3)	10 (8.8)
MI	No (84.6)	Not Sure (15.4)	—
GSI	No (62.3)	Yes (18.5)	Not Sure (12.3)

Human Performance *vs.* MLLMs. Humans achieve 93.30% average accuracy on VIA-Bench. In comparison, the best MLLM (Gemini-3.8-flash) reaches an average score of 82.10%, leaving a gap of 11.20%. This gap has narrowed substantially relative to earlier model generations. On MI, Gemini-3.8-flash and GPT-6-astra match human accuracy at 100.00%, whereas GPT-4o and

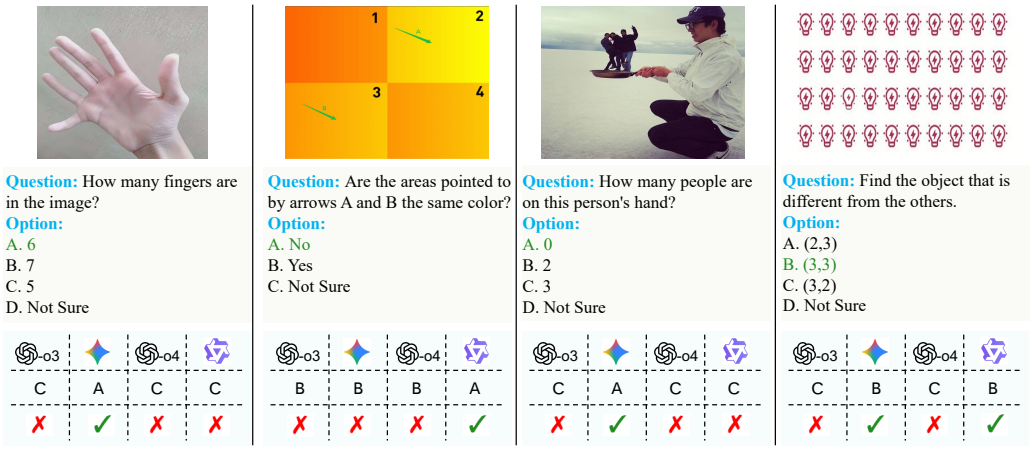


Figure 4: Visualization of model responses on VIA-Bench. These examples demonstrate that even leading models struggle to handle relatively simple tasks such as counting, color recognition, and perceiving fine-grained detail on visual illusions and anomalies. -o3: OpenAI o3; : Gemini-2.5-pro; -o4: OpenAI o4; : Qwen3-VL-30B-A3B-Thinking. See Appendix Q for more cases.

Claude-opus-5 achieve only 20.13% and 0.00%, respectively. This large variation shows that current MLLMs remain unreliable under perceptual conflict. Nevertheless, the human–MLLM gap remains large and highly uneven across categories: humans still lead by 17.78% on visual anomalies and by 22.48% on general visual illusions, precisely the categories where the conflict with canonical priors is strongest. This persistent gap underscores that VIA-Bench remains a non-trivial challenge, highlighting the limitations of current multimodal perception.

Proprietary MLLMs. Despite a significant performance gap compared to humans, proprietary models achieve human-level results on some aspects. For example, o4-mini reaches 97.16% on GSI. The remaining category leads are split between the two strongest systems: Gemini-3.8-flash is best on GI and VI, whereas GPT-6-astra, the second-best model overall at 79.79%, leads proprietary models on VA and CI. Nevertheless, no single model exhibits uniformly strong accuracy across all categories: even Gemini-3.8-flash, the best model overall, falls to 62.97% on VA and 68.99% on VI. Except for the top two models, all proprietary models score below 70% overall, indicating substantial room for improvement. These results highlight the value of VIA-Bench for evaluating MLLM robustness under OOD visual illusions and anomalies.

Open-source MLLMs. While open-source MLLMs such as Kimi-K3, Qwen-VL, and InternVL deliver competitive results, their overall performance still lags behind the strongest proprietary models. Kimi-K3 performs best in this group, reaching 59.59% average accuracy. Moreover, Qwen2.5-VL-3B-Instruct and InternVL3.5-38B achieve only 34.84% and 37.36% average accuracy, respectively, revealing persistent weaknesses on perceptual-conflict inputs.

Reasoning-enhanced MLLMs. In the reasoning-enhanced track, Qwen3.8-max-0902 and OpenAI o3 obtain the best and the second-best average accuracy, with scores of 69.38% and 60.26%. However, their performance remains highly category-dependent. Even Qwen3.8-max-0902, the best model in this track, reaches only 43.59% on MI. Apart from Qwen3.8-max-0902, the highest accuracies on VA and CI are only 25.02% (Claude-3.5-sonnet-20241022) and 51.84% (OpenAI o3), respectively. OpenAI o3 also reaches only 36.75% on GI despite ranking second overall. These results indicate that even leading reasoning models struggle to handle these challenging visual illusions and anomalies.

4.3 Does CoT Reasoning Help on VIA-Bench?

While CoT prompting typically aids complex reasoning (Wang et al., 2023; Muralidharan & Thomas, 2024), recent work shows it can backfire on counter-intuitive tasks by rationalizing deceptive priors (Liu et al., 2025; Qin et al., 2025). We evaluate two general-purpose CoT variants (zero-shot and manual; prompts in Appendix P) against direct prompting on VIA-Bench. Relative gains are shown in Fig. 5, and full numbers in Appendix Table 7. Contrary to general visual tasks, such prompting generally fails to improve—and often degrades—performance: since visual illusions trigger an

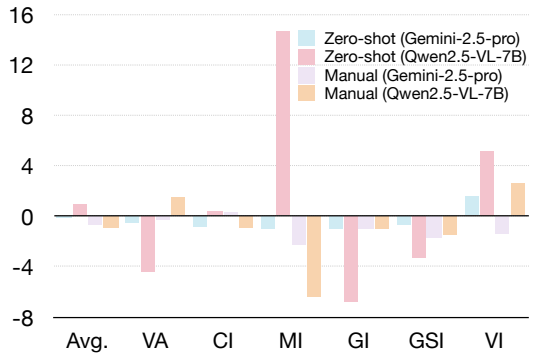


Figure 5: Relative gains of general-purpose CoT over direct prompting. Negative values indicate performance attenuation after applying CoT prompts. These results underscore that VIA-Bench presents a fundamental perceptual bottleneck that is not easily bypassed through surface-level textual prompting.

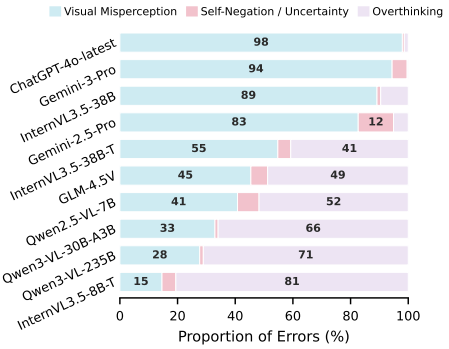


Figure 6: Error distribution on VIA-Bench. Most normal models are dominated by visual misperception, while most reasoning-enhanced models are dominated by overthinking, revealing a "perception → reasoning bottleneck" shift.

immediate misperception, subsequent textual reasoning rationalizes the error with plausible-sounding justifications rather than correcting it (Appendix Q). The +14.8% gain of Qwen2.5-VL-7B on motion illusions is a notable exception, yet closer inspection shows it stems from invoking *textual priors* (*e.g.*, static-image assumptions) rather than genuine visual analysis. In addition, we test task-specific prompts that force explicit, step-by-step enumeration. They do recover accuracy on VA and GI, by +1.3 to +3.4 points (Appendix E). The gains are small, and each category needs its own hand-written procedure. Prompt engineering thus mitigates rather than removes the perceptual bottleneck.

4.4 Open-Ended Evaluation

A common concern with MCQ benchmarks is that answer options may leak signal, inflating accuracy. To probe this, we additionally evaluate four reasoning-capable models in an *open-ended* setting, removing all options on the four categories with well-defined ground-truth values (VA, CI, GI, VI). MI and GSI are excluded because their questions are predominantly binary, where the open-ended form collapses to the MCQ form and yields no additional signal. As shown in Table 3, removing answer candidates reduces mean accuracy across all four tested models (*e.g.*, Gemini-2.5-pro: 49.69 → 35.93; OpenAI o3: 44.43 → 26.75), indicating that answer candidates provide useful cues while open-ended evaluation remains challenging.

Table 3: Open-ended evaluation on VIA-Bench for four categories with ground-truth values.

Model	Setting	Avg.	VA	CI	GI	VI
Gemini-2.5-Pro	MCQ	**49.69**	38.74	28.46	65.81	65.74
	Open-ended	35.93	26.20	7.00	56.80	53.70
OpenAI o3	MCQ	**44.43**	22.01	51.84	36.75	67.13
	Open-ended	26.75	12.30	18.90	12.80	63.00
Qwen3-VL-30B-A3B	MCQ	**40.41**	15.48	39.90	45.47	60.78
	Open-ended	25.53	14.90	12.30	29.50	45.40
GPT-5-chat-latest	MCQ	**21.38**	2.26	8.06	26.67	48.53
	Open-ended	18.85	20.90	18.20	9.40	26.90

4.5 Error Analysis

To characterize model failures on VIA-Bench, we analyze the models' incorrect responses and categorize the observed behaviors into three failure modes: **F1: Visual Misperception**, **F2: Self-Negation / Uncertainty**, and **F3: Overthinking / Error Reinforcement**. These categories are derived from recurring response trajectories (see Fig. 4 and Appendix Q for cases). Fig. 6 shows the distribution of these failure modes among each model's analyzed incorrect responses: most normal MLLMs fail predominantly via F1, whereas most reasoning-enhanced MLLMs fail predominantly via F3. Among all analyzed incorrect responses, F1, F2, and F3 account for 58.8%, 4.5%, and 36.8%, respectively. Visual misperception is thus the most common failure mode, while overthinking accounts for more than a third of the analyzed failures.

Building on the blind and open-ended evaluations, we further examine model responses when the ground-truth physical fact is explicitly provided (Appendix Table 14). On VA/CI, OpenAI o3 reaches

99.6/100.0% accuracy, whereas GPT-5-chat remains wrong on 39.3/24.0% of the same items. This contrast suggests that o3 mainly struggles to extract the physical fact, whereas GPT-5-chat may override it even when explicitly provided. Separately, the MI analysis (Appendix K) shows that recognizing an illusion does not necessarily lead to a physically correct answer, reinforcing our focus on objective physical grounding.

5 Related Work

5.1 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) have emerged as a foundational paradigm for general-purpose AI. Advanced proprietary families such as GPT (Achiam et al., 2023; OpenAI, 2025a; 2026b), the o-series (Jaech et al., 2024; OpenAI, 2025b), Gemini (Comanici et al., 2025; Gemini Team, 2026), and Claude (Anthropic, 2026), alongside open-source series such as Qwen (Bai et al., 2025; Yang et al., 2025), InternVL (Wang et al., 2025), BLIP (Li et al., 2023b; Chen et al., 2025a), GLM (Hong et al., 2025), Show-o (Xie et al., 2025) have fueled rapid development. Recent efforts to enhance their understanding and reasoning capabilities through supervised fine-tuning and reinforcement learning have achieved remarkable performance across various visual benchmarks (Guo et al., 2025b; Yu et al., 2023; Chen et al., 2024b; Liu et al., 2024; Sun et al., 2024; Li et al., 2025c; Zou et al., 2025). However, studies reveal that MLLMs still struggle to handle cognitive illusions, perceptual conflicts, and intuition-breaking scenarios (Gao et al., 2025; Roberts et al., 2025b; Rostamkhani et al., 2025; Sun et al., 2026; Chen et al., 2026). This restricts their ability to generalize and be robust in real-world situations (Kim et al., 2024; Intelligence et al., 2025; Cheang et al., 2025; Gong et al., 2025; Jia et al., 2025; Liao et al., 2025; Li et al., 2025b; Team et al., 2025; Chen et al., 2025b).

5.2 Multimodal Visual Benchmarks

With the expanding capabilities of MLLMs, numerous studies have introduced innovative benchmarks to evaluate performance on multimodal tasks. Broadly, these benchmarks can be categorized into holistic evaluations (Yu et al., 2023; Liu et al., 2024; Li et al., 2023a; Dubey et al., 2024; Wang et al., 2025; Yang et al., 2025; Lin et al., 2026) and domain-specific assessments, such as MMMU/MMMU-pro (Yue et al., 2025). Despite their significance, most prior benchmarks do not explicitly test conflicts between visual evidence and learned priors. More recently, TET (Gao et al., 2025) tests four visual perception-to-reasoning tasks. OmniSpatial (Jia et al., 2025) and Space-10 (Gong et al., 2025) are grounded in basic spatial relations. VisualPuzzle (Song et al., 2025), PuzzleVQA (Chia et al., 2024), and AlgopuzzleVQA (Ghosal et al., 2024) emphasize the model's reasoning ability in visual puzzles. Hallusionbench (Guan et al., 2024) and HellaSwag-Pro (Li et al., 2025a) probe hallucination and counterfactual understanding. ZeroBench (Roberts et al., 2025b) and VibeEval (Padlewski et al., 2024) raise task difficulty to challenge current MLLMs. Recent studies also show that learned priors can override visual evidence (Vo et al., 2025; Ullman, 2024). A line of related work targets visual illusions specifically, as illustrated in Table 4. Unlike these works, VIA-Bench evaluates objectively verifiable physical facts on predominantly real-world visual illusions and anomalies, together with behavioral analyses of model failures.

Table 4: Positioning of VIA-Bench against representative illusion / perceptual-reasoning benchmarks.

Benchmark	Visual Anomalies	Diagnostic Analysis	Eval. Size
GVIL (Zhang et al., 2023)	×	Multi-task eval.	∼100
IllusionVQA (Shahgir et al., 2024)	×	Comprehension + Localization	435
Illusory VQA (Rostamkhani et al., 2025)	×	Preprocessing + Enhancement	gen.
PuzzleVQA (Chia et al., 2024)	×	Multi-stage reasoning diag.	2,000
Art of Deception (Gomez-Villa et al., 2025)	×	Diffusion-model analysis	N/A
VI-Probe (Sun et al., 2026)	×	Memory vs perception probing	N/A
VIA-Bench (Ours)	✓	**F1–F3 taxonomy + CoT analysis**	**1,004**

6 Conclusions

This paper introduces VIA-Bench, a comprehensive and challenging benchmark for evaluating frontier MLLMs on visual illusions and anomalies *grounded in objective physical reality*. VIA-Bench spans six primary categories. Through careful human-in-the-loop review, we curated 1,004 high-quality multiple-choice QA pairs. Extensive evaluations of 30 MLLMs reveal substantial performance variation across models and categories, with no single model performing consistently well across the benchmark. Furthermore, CoT prompting, whether general-purpose or task-specific, offers little robustness on VIA-Bench. We further decompose failures into visual misperception (∼58.8%), self-negation (∼4.5%), and overthinking (∼36.8%), providing a diagnostic view beyond aggregate accuracy. By exposing these gaps, we aim to push the boundaries of MLLMs, offering valuable insights for future research toward visually grounded reasoning.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing and improve readability. We have not used generative AI tools for formulating claims, designing experiments, implementing methods, reformatting the dataset, or interpreting results; these tasks were carried out entirely by the authors. Additionally, GPT-4.1-mini was used as the judge model to evaluate MLLM responses. We have reviewed all AI-assisted edits. We take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

VIA-Bench evaluates objectively verifiable physical facts in out-of-distribution visual scenarios, with the goal of identifying perceptual and reasoning weaknesses that future model optimization should address. We do not foresee major ethical or social concerns associated with this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the dataset construction process in Sec. 3, together with the matching rules, judge and system prompts, and model versions in Appendices G, H, N, and P. To estimate human performance, seven team members independently answered the questions, with at least two responses per item, and we report the average accuracy. We will release the complete benchmark data and evaluation code. We hope VIA-Bench will accelerate progress toward robust multimodal perception and reasoning.

REFERENCES

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

Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www.anthropic.com, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Anthropic. Claude opus 4.1. https://www.anthropic.com/claude/opus, 2025.

Anthropic. Claude opus 5 system card, 2026. URL https://www-cdn.anthropic.com/c5fbac3f0b1280a933ebd26d3cb8bb9f5bdeaf48/Claude%20Opus%205%20System%20Card.pdf.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

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

Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, et al. Gr-3 technical report. *arXiv preprint arXiv:2507.15493*, 2025.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024a.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Y Charles, et al. BabyVision: Visual reasoning beyond language. *arXiv preprint arXiv:2601.06521*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *arXiv preprint arXiv:2505.21996*, 2025b.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of long reasoning models. In *Forty-second International Conference on Machine Learning*, 2025c.

Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Hongcheng Gao, Zihao Huang, Lin Xu, Jingyi Tang, Xinhao Li, Yue Liu, Haoyang Li, Taihang Hu, Minhua Lin, Xinlong Yang, et al. Pixels, patterns, but no poetry: To see the world like humans. *arXiv preprint arXiv:2507.16863*, 2025.

Google Gemini Team. Gemini 3 pro model card. Technical report, Google DeepMind, November 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf. Accessed: 2026-01-24.

Google Gemini Team. Gemini 3.8 flash model card. Technical report, Google DeepMind, September 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-8-Flash-Model-Card.pdf. Accessed: 2026-09-15.

Deepanway Ghosal, Vernon Toh Yan Han, Chia Yew Ken, and Soujanya Poria. Are language models puzzle prodigies? algorithmic puzzles unveil serious challenges in multimodal reasoning. *arXiv preprint arXiv:2403.03864*, 2024.

Alexandra Gomez-Villa, Kai Wang, C Alejandro Parraga, Bartłomiej Twardowski, Jesus Malo, Javier Vazquez-Corral, and Joost van den Weijer. The art of deception: Color visual illusions and diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18642–18652, 2025.

Ziyang Gong, Wenhao Li, Oliver Ma, Songyuan Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. Space-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. *arXiv preprint arXiv:2506.07966*, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

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

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025b.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pp. arXiv–2507, 2025.

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

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. π0. 5: a vision-language-action model with open-world generalization, 2025. *URL https://arxiv. org/abs/2504.16054*, 1(2):3, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. Causalvlbench: Benchmarking visual causal reasoning in large vision-language models. *arXiv preprint arXiv:2506.11034*, 2025.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Xiaoyuan Li, Moxin Li, Rui Men, Yichang Zhang, Keqin Bao, Wenjie Wang, Fuli Feng, Dayiheng Liu, and Junyang Lin. Hellaswag-pro: A large-scale bilingual benchmark for evaluating the robustness of llms in commonsense reasoning. *arXiv preprint arXiv:2502.11393*, 2025a.

Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025b.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025c.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. PerceptionBench: Evaluating atomic visual perception in multimodal large language models. *arXiv preprint arXiv:2607.24957*, 2026.

Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *ICML*, 2025.

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

Jananee Muralidharan and Tiju Thomas. Deliberate problem-solving with a large language model as a brainstorm aid using a checklist for prompt generation. *The Journal of the Association of Physicians of India*, 72(5):89–90, 2024.

OpenAI. Gpt-5 system card. `https://cdn.openai.com/gpt-5-system-card.pdf`, 2025a.

OpenAI. Openai o3 and o4-mini system card. `https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf`, 2025b.

OpenAI. Gpt-5.6 system card, July 2026a. URL `https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf`.

OpenAI. Gpt-6 astra system card, 2026b. URL `https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf`.

Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, et al. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *arXiv preprint arXiv:2405.02287*, 2024.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Jonathan Roberts, Kai Han, and Samuel Albanie. Grab: A challenging graph analysis benchmark for large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1644–1654, 2025a.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025b.

Isabella E Rosario, Fan L Cheng, Zitang Sun, Shin'ya Nishida, and Nikolaus Kriegeskorte. Rotating snakes illusion reveals limitations of visual-motion models in explaining human vision.

Mohammadmostafa Rostamkhani, Baktash Ansari, Hoorieh Sabzevari, Farzan Rahmani, and Sauleh Eetemadi. Illusory vqa: Benchmarking and enhancing multimodal models on visual illusions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2995–3004, 2025.

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models. *arXiv preprint arXiv:2403.15952*, 2024.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. Parrot: Multilingual visual instruction tuning. *arXiv preprint arXiv:2406.02539*, 2024.

Xiaoxiao Sun, Mingyang Li, Kun Yuan, Min Woo Sun, Mark Endo, Shengguang Wu, Changlin Li, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Do vlms perceive or recall? probing visual perception vs. memory with classic visual illusions. *arXiv preprint arXiv:2601.22150*, 2026.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

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

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ICLR*, 2023.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? *arXiv preprint arXiv:2311.00047*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv preprint arXiv:2510.13759*, 2025.

APPENDIX OUTLINE

The Appendix is organized as follows.

- **Appendix A** discusses the limitations of this work.
- **Appendix B** details the taxonomy of VIA-Bench.
- **Appendix C** provides typical question examples.
- **Appendix D** presents additional experiments on OpenAI o3 and Qwen3-VL-235B-A22B-Thinking.
- **Appendix E** provides the full per-category CoT ablation table and the task-specific CoT ablation.
- **Appendix F** presents the temperature ablation study.
- **Appendix G** describes regular match patterns.
- **Appendix H** lists the input prompt for the judge model.
- **Appendix I** reports a judge-consistency audit comparing GPT-4.1-mini and OpenAI o3 as judges.
- **Appendix J** reports option-permutation and polarity-flipping debiasing ablations.
- **Appendix K** provides the four-quadrant analysis on motion illusions (illusion mention $\times$ correctness).
- **Appendix L** provides additional details on the open-ended evaluation protocol.
- **Appendix M** presents the behavioral perception–prior decomposition.
- **Appendix N** specifies detailed model versions for the 30 MLLMs.
- **Appendix O** details the safety / content-moderation audit pipeline and prompt.
- **Appendix P** shows system prompts.
- **Appendix Q** offers case studies and analysis based on models' concrete reasoning procedures.

A LIMITATIONS

VIA-Bench is a diagnostic benchmark for objective physical grounding under perceptual conflict. Performance on VIA-Bench should not be taken as evidence of human-like visual perception. By exposing critical perceptual failure modes in current systems, VIA-Bench helps improve our understanding of model robustness and reliability, and may support the development of more trustworthy multimodal AI applications. Our behavioral error analysis characterizes observable failure patterns but does not identify their internal causes, particularly for proprietary models. The current release contains 1,004 manually curated QA pairs. We prioritize annotation reliability, category balance, and controlled difficulty over scale. Larger versions could support finer-grained analysis across illusion subtypes, visual skills, model families, and deployment settings.

Building on the insights from VIA-Bench, a key research avenue is to improve MLLM generalization in these intuition-challenging edge scenes, potentially through targeted data and appropriate learning objectives.

B TAXONOMY OF VIA-Bench

VIA-Bench aims to comprehensively evaluate MLLMs under visual illusions and visual anomalies. Generally, these scenarios are counterintuitive or countercommonsense for human observers. Such settings require MLLMs to exhibit robust visual perception and deliberate reasoning grounded in physical facts. VIA-Bench covers six major categories: Color Illusions (CI), Motion Illusions (MI), Gestalt Illusions (GI), Geometric and Spatial Illusions (GSI), General Visual Illusions (VI), and Visual Anomalies (VA). Each category targets purpose-built reasoning challenges arising from visual illusions or anomalies. We emphasize that VIA-Bench treats each category as a perceptual-conflict

probe under a unified evaluation target: the ground truth is anchored in objective physical reality (*e.g.*, pixel identity, static images, geometric facts, prior-conflicting physical counts) rather than in whether a model phenomenologically experiences the illusion as humans do. We acknowledge that several individual stimulus types (*e.g.*, Ishihara plates, rotating-snake motion patterns) involve neural mechanisms that current MLLMs do not necessarily replicate (Rosario et al.); under our protocol, such stimuli serve as conflict probes between visual appearance and physical fact, not as tests of human-like phenomenology. The following presents brief descriptions and representative examples for each category.

Color Illusions (CI): This category focuses on color constancy, contrast, illumination, and shadow, which lead to color misperception and interference. Typical examples include the Checker Shadow illusion, Ishihara-style plates (probing perceptual grouping by chromaticity), and the Bezold effect. The ground truth is defined by the objectively encoded visual content (*e.g.*, RGB identity of arrow-marked regions or the embedded digit).

Motion Illusions (MI): Although the images in this category are physically static, specific textures and contrasts elicit subjective perceptions of motion or pulsation, such as stripe-induced drift and the spiral illusion. The ground truth is that the image is *static*, regardless of subjective motion percepts.

Gestalt Illusions (GI): This category involves detecting fine-grained target objects in an image. Due to Gestalt effects, perceptual biases can arise from the overall composition. A common example is identifying the odd one out within a repetitive pattern.

Geometric and Spatial Illusions (GSI): This category probes geometric perceptual biases, such as those arising in judgments of length, parallelism, alignment, and 3D existence. Examples include impossible figures (*e.g.*, the Penrose triangle), the Zöllner illusion, geometric traps, and the Café Wall illusion. The ground truth refers to the underlying geometric or topological fact (*e.g.*, no real hole, no real 3D configuration).

General Visual Illusions (VI): This category groups scene-level naturalistic illusions that do not cleanly fit a single low-level cue. It primarily involves misperceptions and visual misdirection caused by perspective, composition, or occlusion, often combining multiple cues simultaneously. Typical examples include leaves that resemble birds, painting illusions, mirror illusions, and illusion photography.

Visual Anomalies (VA): This category is restricted to anomalies involving fingers and toes. Some images contain anomalous counts in both hands and feet, while others use clenched poses that make individual digits harder to identify. These cases conflict with canonical anatomical priors. The ground truth is the actual number of fingers or toes shown in the image (see Figs. 1 and 13 for examples).

C Question Examples

To raise the difficulty to levels appropriate to challenge the current and future generations of frontier models, we avoid trivial binary questions such as "Is this a visual illusion?" or "Is this a visual anomaly?" Instead, for each category, we design diverse questions from the perspective of a normal image to probe the model's performance on illusions and anomalies. Table 5 provides several examples. Furthermore, we construct specific multiple-choice options for each question, including misleading distractors and the ground-truth answer. This design probes both low-level perception (*e.g.*, counting, color discrimination, fine-grained detail recognition) and higher-level reasoning (*e.g.*, evidence aggregation, consistency checking, and summary judgment) in the presence of illusions and anomalies, offering a comprehensive assessment of model robustness.

Table 5: Question Examples for VIA-Bench.

Taxonomy	Question Example
Visual illusions	• *How many people are on this person's hand?* • *How many different sizes and colors of circles are in the image?* • *Can the car get through?* • *How many faces do you see in the image?* • *Is the person touching the Egyptian lion in the image?* • *Can the water flow into the cup in the boy's hand?* • *Are there people visible in the image?* • *Is the object symmetrical in the image?* • *Is the person touching the Leaning Tower of Pisa in the image?* • *Can the ball in the picture roll?*
Geometric & spatial illusions	• *Is there a real hole in the image?* • *Are there parallel lines in the image?* • *Are the objects in the image aligned straight or slanted?* • *Are the line segments near A and B equal in length?* • *Can the figure shown in the image exist in the real three-dimensional world?* • *Is this a three-dimensional figure?*
Gestalt illusions	• *Find the object that is different from the others, and mark its position.*
Motion illusions	• *Is the image moving or pulsing?*
Color illusions	• *Are the areas pointed to by arrows A and B the same color?* • *What number or word is shown in the image?*
Visual anomalies	• *How many fingers are in the image?* • *How many toes are in the image?*

D More Experiment Results

Here, we additionally report the results for each run of OpenAI o3 (OpenAI, 2025b) and Qwen3-VL-235B-A22B-Thinking (Yang et al., 2025) in Table 6. Although there are variations across runs, our evaluation process–averaging over five runs–substantially reduces model randomness. Furthermore, the models' input is formatted as follows: `[Image][Prompt]`, where `prompt` includes the question, any available options, and formatting instructions.

Table 6: Results of each run. Because model outputs vary, results can differ slightly across runs. Therefore, we report the mean accuracy over five runs to reduce randomness.

Run	VA		CI		MI		GI		GSI		VI	
	Match	Judge	Match	Judge	Match	Judge	Match	Judge	Match	Judge	Match	Judge
OpenAI o3												
1	20.50	20.50	50.75	50.75	93.59	93.59	30.77	30.77	93.21	93.21	68.22	66.67
2	27.76	21.76	51.74	51.74	91.03	91.03	35.04	35.04	92.59	92.59	66.67	65.89
3	20.08	20.08	52.74	52.74	92.31	92.31	40.17	40.17	92.59	91.98	69.77	68.99
4	23.85	23.85	53.73	53.73	87.18	87.18	40.17	40.17	92.59	92.59	67.44	67.44
5	23.85	23.85	50.25	50.25	88.46	88.46	37.61	37.61	91.36	91.36	66.67	66.67
Qwen3-VL-235B-A22B-Thinking												
1	23.43	19.67	35.32	35.82	94.23	94.23	57.26	57.26	82.72	82.72	62.02	66.67
2	23.01	19.25	36.82	34.83	94.87	94.87	56.41	55.56	82.10	81.48	62.79	65.89
3	23.85	20.50	39.30	37.81	95.51	95.51	59.83	54.70	87.04	86.42	63.57	65.89
4	23.43	20.08	38.81	37.31	94.87	94.87	58.12	57.26	80.25	80.25	62.02	64.34
5	23.85	20.92	37.81	36.32	98.08	98.08	59.83	57.26	85.80	83.95	61.24	65.12

E CoT Ablations: General-Purpose and Task-Specific

General-purpose CoT. The main paper visualizes the relative gains of CoT prompting over direct prompting in Fig. 5. For completeness, Table 7 provides the full per-category accuracy under three prompting variants (*w/o CoT*, *zero-shot CoT*, *manual CoT*) for Gemini-2.5-pro and Qwen2.5-VL-7B, averaged over five independent trials. Across both models and most categories, general-purpose CoT either yields no improvement or measurably hurts accuracy, in line with the error-type analysis in Sec. 4.5.

Table 7: Impact of CoT strategies on VIA-Bench. We evaluate the performance of Gemini-2.5-pro and Qwen2.5-VL-7B across varying reasoning configurations. Results are averaged over five independent runs to reduce randomness. We compare direct prompting with zero-shot and manual CoT prompting. The data reveals that general-purpose CoT often fails to provide robustness, highlighting the brittle nature of current reasoning mechanisms under illusory stimuli.

Method	Avg.	VA		CI		MI		GI		GSI		VI	
		Match	Judge	Match	Judge	Match	Judge	Match	Judge	Match	Judge	Match	Judge
Gemini-2.5-pro													
w/o CoT (*i.e.*, normal)	55.01	40.42	38.74	28.46	28.46	46.54	46.54	65.81	65.81	83.95	83.95	65.74	65.74
w/ zero-shot CoT	54.86	40.67	38.24	29.75	27.66	45.64	45.51	64.79	64.79	83.46	83.21	67.29	67.29
w/ manual CoT	54.32	41.09	38.49	32.34	28.76	44.23	44.23	64.79	64.79	82.22	82.22	64.34	64.34
Qwen2.5-VL-7B													
w/o CoT (*i.e.*, normal)	48.00	22.93	22.93	22.09	22.09	78.97	78.97	34.53	34.19	76.30	76.30	53.33	53.33
w/ zero-shot CoT	48.98	18.49	18.49	22.39	22.39	93.72	93.72	28.29	27.35	73.09	72.96	58.45	58.45
w/ manual CoT	47.10	24.77	24.77	21.00	21.00	72.56	72.56	33.16	33.16	74.81	74.81	56.28	56.28

Task-specific CoT. The variants above are generic. They ask the model to reason step by step without telling it *which* procedure to follow. We therefore hand-write an enumeration procedure for the two categories where such a procedure is well defined, namely VA (enumerate every visible digit) and GI (inspect the grid cell by cell), and compare it against our normal prompt. We evaluate Gemini-2.5-pro and Gemini-3.5-Flash.

Table 8 reports the result. Task-specific CoT helps consistently, on both models and both categories. This qualifies the conclusion of Sec. 4.3 to *general-purpose* prompting. The effect is nonetheless bounded. The largest gain is 3.4 points, far from closing the gap to human accuracy. Each category also needs its own hand-written procedure, because the enumeration that works for counting digits does not transfer to scanning a grid. Prompt engineering therefore mitigates rather than removes the perceptual bottleneck.

Table 8: Normal prompting *vs*. task-specific CoT on the two categories for which an explicit enumeration procedure is well defined. All values use Match accuracy; the Gemini-2.5-Pro normal baseline follows Table 1. The prompts are given below. Task-specific prompting helps consistently, but by at most 3.4 points, and each category requires its own hand-written procedure.

Model	VA			GI		
	Normal	Task-CoT	Δ	Normal	Task-CoT	Δ
Gemini-3.5-Flash	58.16	59.41	**+1.25**	88.89	92.31	**+3.42**
Gemini-2.5-Pro	40.42	43.51	**+3.09**	65.81	68.38	**+2.57**

The two procedures are given below.

```
VA_COT = """Solve this by explicit, tool-style visual enumeration -- do NOT answer
from a quick glance or from what a hand/foot 'should' look like.
Step 1: Scan the image systematically from one side to the other and locate every
finger (or toe) one at a time. For each digit, state its position (e.g., 'digit 1:
leftmost', 'digit 2: to its right', ...).
Step 2: Do not assume the normal count (5). Count only the digits that are
actually visible, including any extra or missing ones.
Step 3: After listing and numbering every visible digit, report the total count,
then select the option that matches that count."""
```

```
GI_COT = """Solve this by explicit, tool-style region inspection -- do NOT answer
from a quick global impression of the grid.
Step 1: Treat the image as a grid and mentally crop it into individual cells.
Index them by (row, column) starting at (1,1) in the top-left.
Step 2: Inspect the cells block by block, scanning row by row. For each cell,
compare its shape/orientation/color/pattern carefully against its neighbours.
Step 3: Identify the single cell whose object differs from all the others, state
its (row, column), then select the matching option."""
```

F TEMPERATURE ABLATION

To verify that our default temperature choice (T=0.8 for Gemini-2.5-pro; T=0.6 for InternVL3.5-8B) does not affect our conclusions, we sweep $T \in \{0.1, T_{\text{default}}, 1.5\}$ on the two most temperature-sensitive categories, VA and CI (Table 9). Performance changes are within $\sim$2 points across the swept range, confirming that the brittleness reported in the main paper is not an artifact of the temperature setting. All other results in the paper are averaged over five independent trials with the default T.

Table 9: Ablation on the temperature. Dark green rows mark the values used in our experiments.

Model	Temperature	VA		CI	
		Match	Judge	Match	Judge
Gemini-2.5-pro	0.1	40.50	40.42	30.05	30.05
Gemini-2.5-pro	0.8	40.42	38.74	29.55	26.77
Gemini-2.5-pro	1.5	40.92	40.84	29.45	29.75
InternVL3.5-8B	0.1	23.35	14.98	34.53	20.70
InternVL3.5-8B	0.6	23.68	16.15	33.13	20.70
InternVL3.5-8B	1.5	25.27	15.06	31.74	21.00

G MATCH PATTERNS

We use the following regular expression (pattern) to match the answer:

```
patterns = [
    r'answer\s*[:\-]\s*\[([ABCD])\]',      # Answer: [A] (anywhere in text)
    r'\[([ABCD])\]',                       # [A] (anywhere in text)
    r'answer\s*[:\-]\s*([ABCD])\b',        # Answer: A (with word boundary)
    r'\b([ABCD])\b',                       # A (single letter with word boundaries)
    r'answer\s*[:\-]\s*\[([ABCD])\]\b',    # Answer: [A] (with word boundary)
    r'answer\s*is\s*\[([ABCD])\]',         # Answer is [A]
    r'answer\s*is\s*([ABCD])',             # Answer is A
    r'the\s*answer\s*is\s*\[([ABCD])\]',   # The answer is [A]
    r'the\s*answer\s*is\s*([ABCD])',       # The answer is A
    r'choice\s*[:\-]\s*\[([ABCD])\]',      # Choice: [A]
    r'option\s*[:\-]\s*\[([ABCD])\]'       # Option: [A]
]
```

H JUDGE PROMPTS

The input prompt for the judge model is in the following format:

```
judge_prompt = f"""You are a judge for multiple choice questions.
Analyze this model response and determine what answer it indicates:
Model Response: "{response}"
Based on the content of this response, what option is the model actually choosing?
Return ONLY one letter: A, B, C, or D

Examples:
- "I think it's A" -> A
- "The answer should be B" -> B
- "My choice is C" -> C
- "I'm not sure" -> D
- "Cannot determine" -> D

What option does this response indicate?"""
```

I JUDGE CONSISTENCY AUDIT

Because our evaluation uses an LLM-as-a-Judge protocol, one concern is whether the judge model itself introduces priors that bias accuracy estimates. To audit this, we re-judge the full set of Gemini-3-Pro responses on VIA-Bench using a stronger reasoning model (OpenAI o3) as an alternative judge and compare it against our primary judge (GPT-4.1-mini). As shown in Table 10, the two judges return identical aggregate accuracy (720/1004), supporting the claim that the judge primarily maps model outputs to option labels (a low-bias mapping task) rather than performing visual reasoning.

Table 10: Judge-consistency audit on Gemini-3-Pro responses. GPT-4.1-mini and OpenAI o3 return identical aggregate accuracy, supporting that the judge module introduces little additional bias.

Judge model	**Accuracy (correct / total)**
GPT-4.1-mini (primary, ours)	720 / 1004
OpenAI o3 (audit)	720 / 1004

J DEBIASING ABLATIONS

To validate the two debiasing techniques used during VIA-Bench construction (*i.e.*, option-position randomization and binary-question polarity flipping), we perform two ablations on OpenAI o3.

Table 11: Option-permutation debiasing ablation on OpenAI o3. Fixing the correct option at position A inflates accuracy relative to our randomized protocol, confirming that randomization measurably suppresses positional shortcuts. Δ is the difference.

Setting	**Avg.**	**VA**	**CI**	**MI**	**GI**	**GSI**	**VI**
Random (ours)	59.2	20.5	50.7	93.6	30.8	93.2	66.7
Position A	60.9	20.5	51.7	96.2	35.0	93.2	69.0
Δ	+1.7	0.0	+1.0	+2.6	+4.3	0.0	+2.3

Table 12: Question-polarity flipping ablation on OpenAI o3. Rephrasing the binary question to be prior-aligned changes performance noticeably.

Setting	**MI**	**GSI**	**CI**
Ours	93.6	94.3	32.0
Flipped	100.0	88.6	36.0
Δ	+6.4	−5.8	+4.0

Option permutation. Table 11 compares our randomized-position protocol against a counterfactual setting in which the correct answer is always placed at position A. Fixing the correct option at A *improves* accuracy by +1.7 points on average (and up to +4.3 on GI), confirming that without randomization models can exploit positional shortcuts. Our randomized design therefore measurably suppresses such shortcut-driven hallucinations.

Polarity flipping. Table 12 compares our binary-question phrasing to its polarity-flipped counterpart (*e.g.*, MI: "Is the image moving or pulsing?" $\rightarrow$ "Is the image static and still?"). The flipped phrasing, which aligns more naturally with commonsense priors, changes performance noticeably (*e.g.*, +6.4 on

MI, −5.8 on GSI). This indicates that our original phrasing is harder precisely because it suppresses prior-aligned shortcut cues, which is the intended effect.

K Motion Illusion Analysis

A natural question is whether *verbally recognizing* a stimulus as an illusion correlates with *physically correct* answers. We categorize Motion-Illusion responses of two reasoning-enhanced MLLMs into four mutually exclusive groups along (i) whether the response explicitly mentions the illusion, and (ii) whether the final answer matches the physical-fact ground truth (Table 13). We find that models can produce correct answers *without* explicitly recognizing the illusion (34.6% for Qwen3-VL-30B-A3B), and may also produce incorrect answers *despite* recognizing it (15.0% for InternVL3.5-30B-A3B). This shows that "illusion recognition" and "physical-fact correctness" are decoupled in current MLLMs, and motivates our choice to evaluate `VIA-Bench` under a single, objective-physical-grounding target rather than mixing two opposing goals.

Table 13: Analysis of MI responses for two reasoning-enhanced MLLMs. "Illusion mentioned" = the response verbally identifies the stimulus as a (visual) illusion / non-physical motion. "Correct" = the final answer matches the physical-fact ground truth (*i.e.*, the image is static). The two notions are decoupled: correct answers can occur without verbal recognition, and recognized illusions can still be answered incorrectly.

Response Category	**Qwen3-VL-30B-A3B (%)**	**InternVL3.5-30B-A3B (%)**
Illusion mentioned + Correct	61.7	83.2
Illusion not mentioned + Correct	34.6	1.8
Illusion mentioned + Wrong	3.7	15.0
Illusion not mentioned + Wrong	0.0	0.0
Total Correct	**96.3**	**85.0**

L Open-Ended Evaluation Details

We complement the multiple-choice protocol with an open-ended evaluation in which all answer options are removed for the four categories with well-defined ground-truth values (VA, CI, GI, VI). Models are prompted to produce a free-form answer, which is then compared against the ground-truth value via the same LLM-as-a-Judge protocol used for MCQ (with the comparison adapted to handle numeric / short-text answers). Results are summarized in Table 3 of the main paper. Across all four tested models, removing options reduces mean accuracy, indicating that answer candidates provide useful cues while open-ended evaluation remains challenging.

M Perception–Prior Probes

We combine blind, open-ended, and oracle-fact probes to distinguish failures to extract physical facts from prior override. Blind results indicate that textual priors alone are insufficient for VA and CI, while open-ended results show that the tested models often fail to extract the relevant physical facts without answer candidates. As shown in Table 14, when the ground-truth fact is provided, OpenAI o3 recovers almost fully, whereas GPT-5-chat remains wrong on 39.3% of VA and 24.0% of CI items. This contrast suggests that o3 mainly fails to extract the physical fact, whereas GPT-5-chat may override even explicitly provided facts.

Table 14: Accuracy on VA and CI under three behavioral probes. Blind evaluation removes the image, open-ended evaluation removes answer options, and oracle-fact evaluation provides the ground-truth physical fact.

Category	**Blind GPT-4-Turbo**	**Open-ended four-model range**	**Oracle OpenAI o3**	**Oracle GPT-5-chat**
VA	2.85	12.30–26.20	99.6	60.7
CI	25.87	7.00–18.90	100.0	76.0

N MODEL VERSIONS

Table 15: Model markers or links used in our evaluation.

Model	Marker/Link
GPT-6-astra	*api_openai_gpt-6-astra*
Gemini-3-pro	*api_google_gemini-3-pro-preview*
GPT-4-Turbo	*api_openai_gpt-4-turbo-2024-04-09*
GPT-4.1-mini	*api_openai_gpt-4.1-mini*
Gemini-2.5-pro	*api_google_gemini-2.5-pro*
GPT-5-chat-latest	*api_openai_gpt-5-chat-latest-response*
GPT-4o-2024-11-20	*api_azure_openai_gpt-4o-2024-11-20*
ChatGPT-4o-latest	*api_openai_chatgpt-4o-latest*
OpenAI o4-mini	*api_azure_openai_o4-mini*
GPT-5-mini	*api_azure_openai_gpt-5-mini-response*
Qwen-vl-max-latest	api_ali_qwen-vl-max-latest
OpenAI o3	*api_openai_o3*
GLM-4.5V	*api_zhipu_glm-4.5v*
Claude-3.5-sonnet-20241022	*api_anthropic_claude-3-5-sonnet-20241022*
Claude-sonnet-4-20250514	*api_anthropic_claude-sonnet-4-20250514*
Claude-opus-4.1-20250805	*api_anthropic_claude-opus-4-1-20250805*
Gemini-3.8-flash	*api_google-vertex_gemini-3.8-flash*
Gemini-3.5-Flash	*api_google_gemini-3.5-flash*
GPT-5.6-sol	*api_openai_gpt-5.6-sol*
Claude-opus-5	*api_anthropic_claude-opus-5*
Qwen3.8-max-0902	*api_ali_qwen3.8-max-0902*
Kimi-K3	*api_ali_kimi-k3*
DeepSeek-V4.1-Flash	*api_deepseek_deepseek-v4.1-flash-expires-on-0910*
Qwen2.5-VL	`https://github.com/QwenLM/Qwen2.5-VL`
InternVL3.5	`https://github.com/OpenGVLab/InternVL`
Qwen3-VL	`https://github.com/QwenLM/Qwen3-VL`

In Table 15, we provide the detailed model versions for the 30 MLLMs used in our evaluation. For models accessed via API, we report the provider's unique model identifiers (markers). For models we ran locally, we include links to the corresponding GitHub repositories. All models included in our evaluation were available as of September 10, 2026.

O SAFETY / CONTENT-MODERATION AUDIT

In addition to manual filtering during image collection and human cross-verification, we further audit every retained image with OpenAI o3 for sensitive content. The model is queried with the following prompt, and only images returning `safety=Yes` are kept in the final release. In our audit, 100% of the curated images returned `safety=Yes`, indicating that the human-in-the-loop pipeline already removed sensitive content.

```
safety_prompt = """Check whether this image contains any unsafe or
sensitive content from the following categories:
- nudity
- sexual content
- violence
- gore
- hate symbols

Return your answer in valid JSON only, with exactly this schema:
{
  "safety": "Yes" or "No",
  "reason": "",
  "categories": []
}

Rules:
- Set "safety" to "No" if any listed category is present.
- Set "safety" to "Yes" if none of the listed categories is present.
- If safety is "Yes", set "reason" to "" and "categories" to [].
- If safety is "No", give a brief reason and list the detected
  categories from the allowed set above.
- Do not include markdown fences or any extra text outside the
  JSON object.
"""
```

P SYSTEM PROMPTS

Our research indicates that prevailing general-purpose CoT methods do not effectively handle our VIA-Bench. In this section, we present all system prompts used in our experiments in Fig. 7. This is intended to showcase the carefully crafted prompts and facilitate reproducibility. Unless otherwise specified, the normal system prompt was used in all our experiments.

Q CASE STUDY

In this section, we present more case studies for human-conducted analysis of the models' concrete reasoning procedures. We include three selected frontier models: InternVL-3.5-8B (Wang et al., 2025), Qwen3-VL-30B-A3B-Thinking (Yang et al., 2025) and Gemini-2.5-pro (Comanici et al., 2025). In the analysis, we identify the categorized error types and highlight our findings in relevant parts. Results are shown in Figs. 8–13.

For the MLLMs' reasoning processes, we mark correct reasoning content in green and incorrect (or not grounded in visual evidence) reasoning content in red. The proprietary model Gemini-2.5-pro does not explicitly expose its reasoning chain. Therefore, some cases only show their final answer. Bolded words such as "wait", "no", and "but" indicate patterns of overthinking. Based on these observations, we derive three key findings.

Finding-1: One clear takeaway is that most error types stem from an initial visual misperception rather than logical fallacies. Once the model misidentifies a key visual element (*e.g.*, missing a subtle illusory cue), it creates a flawed premise, rendering the final deduction inevitably incorrect.

Finding-2: On VIA-Bench, models often exhibit high uncertainty due to the conflict between visual evidence and internalized priors. Even after successfully extracting visual information, they tend to self-negate (*e.g.*, "I'm a bit uncertain...", "maybe..."), prioritizing canonical knowledge over what is actually presented.

Finding-3: A severe overthinking phenomenon is prevalent (Chen et al., 2025c). Visual illusions trigger repetitive or self-contradictory CoT traces (*e.g.*, loops of "wait", "but", "no"), where the model struggles to resolve the visual ambiguity.

1. Normal System Prompt:

```
# ========================================================================
# Normal: Direct answer with strict format requirements
# ========================================================================
```

You are an expert at analyzing images and answering multiple-choice questions about visual content.

Your task is to carefully examine the provided **image** and **question** with multiple choice **options** (A, B, C, D), then select the most accurate answer.

The input consists of:
- An **image** that contains visual information
- A **question** about the image content
- Multiple choice **options** (A, B, C, D) for the answer
- **Instructions** about output format.

Based on the input, provide your answer.

2. Zero-shot CoT System Prompt:

```
# ========================================================================
# Zero-shot Chain-of-Thought: Complete reasoning process
# ========================================================================
```

You are an expert at analyzing images and answering multiple-choice questions about visual content.

Your task is to examine the provided **image**, **question**, and multiple choice **options** (A, B, C, D) carefully, then provide a complete chain-of-thought reasoning before selecting your answer.

The input consists of:
- An **image** that contains visual information
- A **question** about the image content
- Multiple choice **options** (A, B, C, D) for the answer
- **Instructions** about output format.

Think step by step and provide your answer.

3. Manual CoT System Prompt:

```
# ========================================================================
# Manual Chain-of-Thought: Systematic analysis approach
# ========================================================================
```

You are an expert at analyzing images and answering multiple-choice questions about visual content.

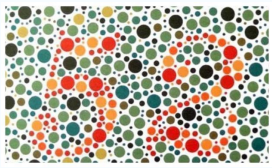
The input consists of:
- An **image** that contains visual information
- A **question** about the image content
- Multiple choice **options** (A, B, C, D) for the answer
- **Instructions** about output format.

Guidelines:
Please follow these steps to analyze the image and answer the question:
1. Carefully examine all visual elements in the **image**
2. Identify key features, patterns, and details that are relevant to the **question**
3. Evaluate each answer **option** (A, B, C, D) by comparing it with your visual observations from the **image**
4. Select the **option** that best matches the visual evidence from the **image**

After your analysis, respond with only: Answer: [X]
Where X is one of the options (A, B, C, or D).

Focus exclusively on **visual evidence from the image** and avoid assumptions not supported by what you can see in the image.

Figure 7: Illustration of three system prompts adopted in VIA-Bench evaluation.



Question: What number or word is shown in the image?
Option:
A. 25
B. 9
C. 52
D. Not Sure

Please write your answer in the following format:
Answer: []

InternVL **:** <think> So I'm looking at this image, and it's filled with a bunch of circles in different sizes and colors. The question is asking what number or word is shown, with options A through D. Let me think about how to approach this.

First, the circles are arranged in a pattern, but they don't immediately look like they form a number. They're overlapping and scattered, so maybe there's a hidden message or a number formed by their arrangement. **But** the options include numbers like 25, 9, 52, or "Not Sure."

Wait, maybe the colors correspond to something. The circles are in colors like red, green, blue, yellow, black, and some shades. **But** how would that translate to a number? Alternatively, maybe the sizes of the circles represent digits. For example, larger circles could be higher numbers, **but** that seems a bit vague.

Another thought: sometimes in these kinds of puzzles, the number of a specific color might indicate something. Let's count the red circles. **There are several red circles, but I don't see a clear pattern.** Similarly, the black circles are prominent, **but** again, not sure how that helps.

Looking at the options again, 9 is a single-digit number, while 25 and 52 are two digits. The image is a square, maybe 10x10 or something, but I don't think that's relevant. Alternatively, maybe the arrangement of the circles forms a shape that resembles a number when viewed from a distance. **But** I'm not seeing it clearly**.**

Wait, perhaps the question is a trick. The image might not actually show a number, and the correct answer would be "Not Sure." But I should check again. Let me look for any possible numbers formed by the circles. If I squint, do the circles form the shape of a number? Maybe not. **Alternatively, the density of certain colors could represent a number, but that seems too abstract.**

Considering the options, if I have to choose, maybe the answer is D. Not Sure, because I don't see a clear number formed by the circles**. But** I'm not entirely certain. Alternatively, maybe there's a pattern I'm missing. Let me think again.

Wait, sometimes these images use the concept of "circle size" to represent numbers. For example, each size could correspond to a digit. **But** without a key, that's hard to apply. Alternatively, the colors might correspond to numbers on a phone keypad or something, but that seems unlikely.

Another angle: the question might be testing if I recognize that the image is a well-known pattern that represents a number. **But** I can't recall any such pattern off the top of my head. The circles are just randomly placed with varying colors and sizes.

Given that I can't find a clear number, and the options include "Not Sure," I think the safest choice is D. However, I'm a bit uncertain because sometimes these questions have a trick. **But based on what's visible, I don't see a definitive number**. </think>

Answer: D

: So, let's look at the image. It's a color blindness test, like an Ishihara plate. The numbers are formed by colored dots. Let's count. The red and orange dots probably form a number. **Wait**, the image has two parts? **Wait, no. Wait,** the dots—let's see. **Oh, the number here is 25? Wait, no. Wait, maybe 25. Wait,** let's check. **Wait**, in Ishihara plates, the number is made by dots that are distinguishable by those with color vision. Let's see: the red/orange dots might form 25. Wait, option A is 25. Let's check the options. Option C is 52. **Wait,** maybe the number is 25. **Wait**, let's think. The image has a pattern where the dots form the number. Let me recall. Some Ishihara plates have 25. **Wait,** the correct answer for this type of plate is often 25. **Wait, no, wait. Wait**, maybe it's 25. Let's check the options. The options are A.25, B.9, C.52, D.Not Sure. So, in the image, the number formed by the colored dots—probably 25. So the answer is A. </think>

Answer: A

: * **Step 1:** Analyze the image. The image provided is a color perception test plate, similar to those used in the Ishihara test for color blindness. It consists of a pattern of multicolored dots. A figure, in this case, a number, is embedded within the dots.
* **Step 2:** Identify the figure for normal color vision. A person with normal color vision can distinguish the dots that form the number from the background dots. The number is formed by a series of red and orange dots. By tracing these dots, we can clearly identify a two-digit number. The digit on the left is a '2', and the digit on the right is a '9'. Therefore, the number shown in the image is "29".
* **Step 3:** Compare the identified number with the given options. The options are A. 25, B. 9, C. 52, and D. Not Sure. The number we identified, "29", is not among the options A, B, or C.
* **Step 4**: Evaluate the available options.
  * A. 25: This is incorrect. The image shows a '2' on the left, but the digit on the right is a '9', not a '5'.
  * C. 52: This is incorrect. Neither digit matches the position in the image.
  * B. 9: This number is a part of the number "29" shown in the image. Although it is not the complete number, the digit '9' is clearly visible.
  * D. Not Sure: This could be a valid choice since the full correct number "29" is not an option.
* **Step 5:** Select the best possible answer. Since this is a multiple-choice question and the complete answer "29" is missing, we must choose the best fit from the given options. Options A and C are definitively wrong because they contain a digit ('5') that is not present in the image**. Option B ('9') represents a digit that is clearly visible in the image.** In such a scenario where the full answer is not available, the most logical choice is the option that correctly identifies a part of the visible figure. Therefore, '9' is the most plausible answer among the choices. Some viewers with specific color vision deficiencies might also see only the '9'.

Answer: [B]

Figure 8: A demonstration of color illusions. InternVL-3.5-8B does not see a clear number and exposes a severe phenomenon of overthinking. Qwen3-VL-30B-A3B-Thinking identifies a reversed order and exhibits very low confidence in its answer, which leads to repetitive confirmation. Gemini-2.5-pro makes a recognition error, yet it preserves a readable and well-structured reasoning process.

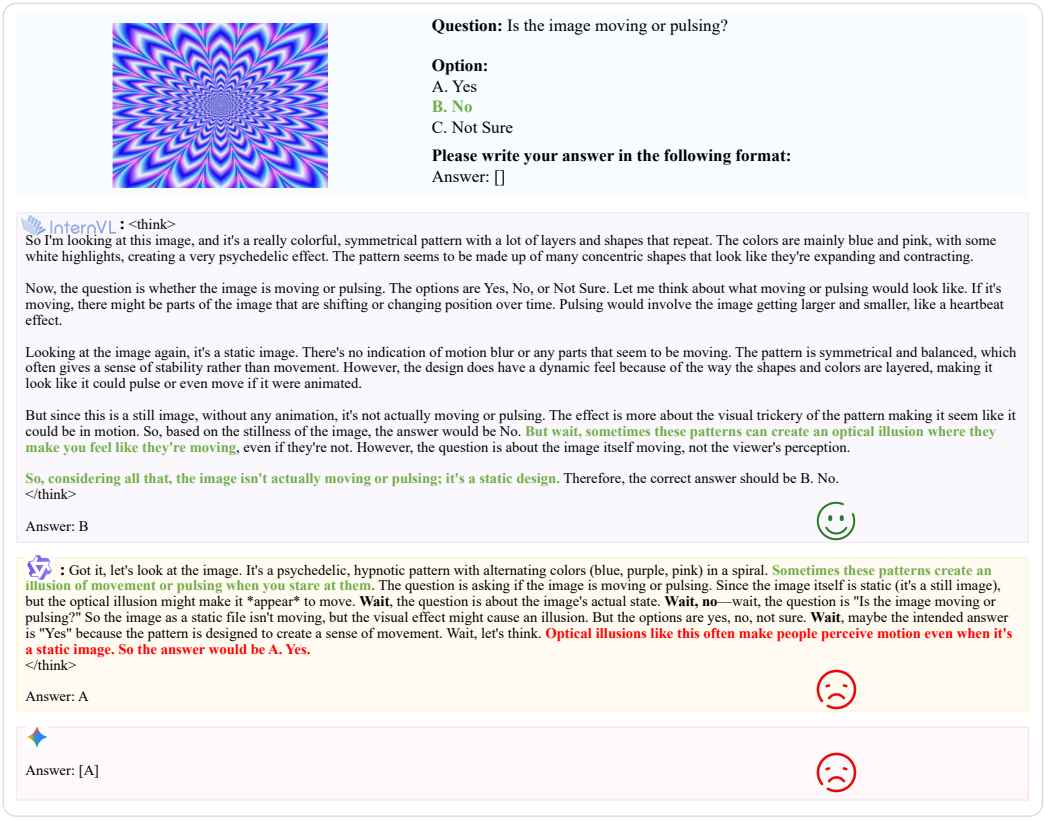


Figure 9: A demonstration of motion illusions. InternVL-3.5-8B identifies this as an optical illusion and infers from the visual content that it is a static image. Qwen3-VL-30B-A3B-Thinking recognizes the illusion, but its overthinking leads to an incorrect answer. Gemini-2.5-pro directly answers “Yes”.



Question: Find the object that is different from the others, and mark its position.
Option:
A. (3,4)
B. (3,3)
C. (3,6)
D. Not Sure

Please write your answer in the following format:
Answer: []

InternVL : <think>So I'm looking at this grid of teddy bears, and I need to find the one that's different from the others. The question says to mark its position using the row and column numbers, like (row, column). The options given are A. (3,4), B. (3,3), C. (3,6), and D. Not Sure.
First, I'll try to visualize the grid. It's a 4x10 grid, right? So there are 4 rows and 10 columns. Each teddy bear is in a specific position. I need to look for any differences among them.
Looking at the first row, all the teddy bears seem the same. Same with the second and third rows. Wait, but the fourth row might have something different. Let me check each row carefully.
In the third row, which is row 3, I'll go through each column from left to right. Let's see:
- Column 1: Normal teddy bear.
- Column 2: Normal.
- Column 3: Normal.
- Column 4: **Wait, this one looks a bit different. Maybe the bow is positioned differently?**
- Column 5: Normal.
- Column 6: Hmm, this one might be different too. Let me double-check.
- Columns 7-10: All seem normal.
Wait, maybe I'm missing something. Let me go back. The instruction says to find the one that's different. Maybe it's not just the bow but something else, like the teddy bear's posture or size.
.....
I'm getting a bit confused. Maybe I should consider that the correct answer is C. (3,6) based on the options provided, but I'm not entirely certain without a clearer visual. Alternatively, perhaps the answer is D. Not Sure, because I might be missing something. But I think I should go with the most likely option based on my observation.
</think>
Answer: C

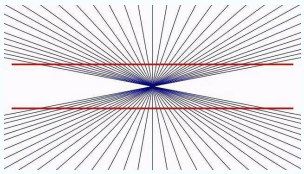
: Got it, let's look at the image. The task is to find the teddy bear that's different. Let's check each row. The teddy bears are arranged in a grid. Let's count the rows and columns.
First, let's list the rows. There are 4 rows (since the image shows 4 rows of teddy bears). Wait, the rows are from top to bottom, so row 1 is the top row, row 2 next, etc. Columns are left to right, so column 1 is leftmost.
Let's check each row:
Row 1 (top row): All teddy bears look the same.
Row 2: All look the same.
Row 3: **Let's check columns. In row 3, which is the third row from the top, let's see each column. Wait, the options are about (row, column).** Let's check the options: A is (3,4), B (3,3), C (3,6).
Wait, the image: let's count the columns. How many columns are there? Let's count the teddy bears per row. Each row has 9 teddy bears (since from left to right, 9). So columns 1-9.
.....
Now, in row 3 (third row), column 4 (fourth column), is that teddy bear different? Let's see.

Wait, maybe

: * Analyze the image: The image displays a grid of teddy bears. There are 4 rows and 10 columns. All the teddy bears look very similar. They are brown and have a red bow tie.
* Identify the task: The goal is to find the one teddy bear that is different from all the others and state its position in the format (row, column).
* Systematic search: I will examine each teddy bear, looking for any subtle differences. The most likely place for a difference is a small detail like the eyes, nose, ears, or the bow tie.
* Scanning the bears:
  * Looking at the bears, they all have the same body shape, color, eyes, and nose.
  * Let's focus on the bow ties. Most of the bears have a red bow tie with a small, darker red knot in the center.
  * I will scan row by row, checking the bow tie of each bear.
  * Row 1: All bears have a normal bow tie with a knot.
  * Row 2: All bears have a normal bow tie with a knot.
  * Row 3:
    * The first bear has a normal bow tie.
    * The second bear has a normal bow tie.
    * The third bear has a normal bow tie.
    * **The fourth bear (at position (3, 4)) has a bow tie that is missing the central knot. It's just two red triangles. This is the difference.**
    * The fifth bear has a normal bow tie.
    * The rest of the bears in row 3 have normal bow ties.
  * Row 4: All bears have a normal bow tie with a knot.
* **Conclusion: The teddy bear in the third row and fourth column is the different one because its bow tie lacks the central knot that all other bears have. This position is denoted as (3, 4). The correct answer is (3, 4).**

Answer: [A]

Figure 10: A demonstration of gestalt illusions. The ellipsis (".....") denotes the omission of intermediate reasoning steps. In gestalt illusions, models must accurately perceive fine-grained details and infer their precise locations. InternVL-3.5-8B identifies the discrepancy in the fourth column but remains unconfident, eventually selecting the incorrect answer after analyzing the options. Although Qwen3-VL-30B-A3B-Thinking recognizes the difference, its iterative reasoning exceeds the maximum output length, failing to produce a direct answer. In contrast, Gemini-2.5-pro demonstrates effective pattern recognition, resulting in the correct outcome.



Question: Are there parallel lines in the image?

Option:
A. No
B. Yes
C. Not Sure

Please write your answer in the following format:
Answer: []

InternVL **:** So, looking at the image, there are a bunch of lines radiating out from a central point. They all seem to spread out in different directions, **but** I need to check if any are parallel. The red lines are horizontal, right? **The top and bottom red lines might be parallel if they're straight and in the same direction**. **Wait**, but they might be part of the same set. The blue lines are the ones radiating outward. Since they all come from the center, they probably aren't parallel because they diverge. **But** the red lines, if they're both horizontal and the same length, could be parallel. **Wait,** but maybe the red lines are actually part of the same line, just extended. Hmm, no, the question is about parallel lines in the image. If there are two horizontal red lines that are parallel, then the answer would be yes. **But** I need to be careful. Let me think again. The image shows lines going out from the center, but there are two horizontal red lines, one at the top and one at the bottom. If they are both horizontal and in the same direction, they might be parallel. **But wait**, if they're part of the same set of lines radiating from the center, maybe they are not parallel. **Wait, no**, parallel lines never meet, right? So if the red lines are horizontal and the same length, but placed at different vertical positions, they would be parallel. **But** in the image, the red lines might be part of the same line, just mirrored. **Wait, no**, the image shows two horizontal red lines, one at the top and one at the bottom, crossing the central point. **Wait, no**, actually, looking again, the red lines might be part of the same set. **Wait**, maybe the red lines are actually the same line but extended in opposite directions. **No**, that doesn't make sense. Alternatively, the red lines could be two separate horizontal lines. If they are both horizontal and not intersecting, then they are parallel. **But** in the image, since they cross the central point, maybe they are not parallel. **Wait, no,** if they are both horizontal and cross the center, they would intersect at the center. So if two lines intersect, they can't be parallel unless they are the same line. So if the two red lines are distinct and both horizontal, they would intersect at the center, meaning they are not parallel. **But wait**, maybe the red lines are actually part of the same line, but that seems unlikely. Alternatively, maybe the red lines are part of the same set of lines, **but** arranged symmetrically. Hmm, this is confusing. Let me think again. The main lines are the blue ones radiating from the center. The red lines are two horizontal lines, one at the top and one at the bottom. If they are both horizontal and the same length, but placed symmetrically around the center, then they would be parallel. Because parallel lines are lines in the same plane that do not intersect. If they are both horizontal and the same direction, they don't intersect unless they are the same line. **But** in the image, they might be two separate horizontal lines, one above and one below the center. **Wait, but** if they are both horizontal and cross the center, they would intersect at the center. So that would mean they are not parallel. **Wait**, **but** if they are on opposite sides of the center, maybe they are parallel. **Wait**, no, if two lines are horizontal and cross the center, they intersect there, so they can't be parallel. Therefore, the red lines might not be parallel. **But then**, are there any other parallel lines? The blue lines are all radiating from the center, so they are not parallel. **Wait,** but maybe some of the blue lines are parallel? **No**, because they all come from the same point, so they diverge. So the only possible parallel lines would be the red ones. **But** if the red lines intersect at the center, they are not parallel. So does that mean there are no parallel lines? **But** the options are A. No, B. Yes, C. Not Sure. **Wait, but** maybe I'm misunderstanding the image. Let me visualize it again. The image has a central point with lines radiating outward in all directions. There are two horizontal red lines, one at the top and one at the bottom, crossing the center. If those red lines are part of the same set, maybe they are actually the same line but mirrored. **No**, that doesn't make sense. Alternatively, maybe the red lines are part of the same line but extended in opposite directions. **Wait, but** if they are two separate horizontal lines, one at the top and one at the bottom, then they would be parallel only if they don't intersect. **But** since they both pass through the center, they intersect there, so they can't be parallel. **Therefore, there are no parallel lines. But wait,** maybe the red lines are not actually passing through the center. **Wait**

: So, let's look at the image. The question is asking if there are parallel lines. Let's check the lines. The red lines at the top and bottom—those are horizontal and parallel to each other. Also, the black and blue lines are radiating, but **the red ones are straight and parallel**. So yes, there are parallel lines. So the answer should be B. </think>

Answer: B

Answer: [B]

Figure 11: A demonstration of geometric and spatial illusions. In this case, Qwen3-VL-30B-A3B-Thinking and Gemini-2.5-pro concisely arrive at the correct choice. InternVL-3.5-8B initially suggests the red lines might be parallel. However, it subsequently gets stuck repeating this viewpoint, which leads to an incorrect conclusion. Ultimately, it fails to output a final choice because its judgment exceeds the maximum output length.



Question: Can the car get through?

Option:
A. No
B. Yes
C. Not Sure

Please write your answer in the following format:
Answer: []

InternVL **:** <think>
So, looking at this image, there's a car parked on what seems like a street or driveway. In the foreground, there's a large, three-dimensional artwork or installation made of wood planks arranged in a circular pattern with some chains and blocks on top. **The artwork gives the illusion that it's a deep pit or hole due to its perspective.**

The question is whether the car can get through this setup. The options are A. No, B. Yes, and C. Not Sure. Let's break this down.

First, the car is positioned behind the artwork. The artwork is quite large and appears to be a visual trick, making it look like a deep hole. However, since it's an artwork, it's likely not a real pit. The car is parked on the other side of the artwork, and the perspective might make it seem like the car is approaching the edge of the hole. But in reality, the artwork is just a 3D piece on the ground.

If the car tried to drive forward into the artwork, it would probably hit the base of the installation. The artwork seems to be fixed to the ground, so the car can't actually drive through it. The chains and blocks on top might also pose a physical barrier. Additionally, the car's position suggests it's on a paved area, and the artwork is part of that pavement. So, the car can't "get through" the hole because there's no actual hole; **it's just an illusion**.

Therefore, the answer should be A. **No, the car can't get through because the obstacle is a physical installation, not a real passage.**
</think>

Answer: A

: So, let's look at the image. There's a 3D street art that looks like a big pit or a deep hole, but it's just a painting on the ground. The car is positioned as if it's about to go into this "pit," **but since it's an illusion**, the ground is actually flat. **Wait, no**—wait, the question is "Can the car get through?" But the 3D art is a trick. The car is in front of the pit, but the pit is fake. **Wait, no**, maybe the question is about whether the car can drive over the pit. But the pit is an illusion, so the car can drive normally. **Wait, no,** the image shows the car is stopped at the edge. But the art is a 3D illusion, so the ground is flat. So the car can get through because the pit isn't real. Wait, no, maybe the question is tricky. **Wait,** the options: A. No, B. Yes, C. Not Sure.

Wait, **in reality, the street art is a 3D painting**, so the ground is flat. So the car can drive over it. **But** the image shows the car is near the edge of the pit. **But** since it's an illusion, the car can get through. **Wait, but** maybe the question is asking if the car can pass through the pit as if it were real. **But** the pit is fake, so the car can get through. **Wait, no**—wait, the car is parked near the edge. **But** the illusion is that it's a deep pit, but it's not. So the car can drive over. So the answer would be B. Yes? **Wait, no**. **Wait**, maybe the question is about whether the car can go through the pit (the illusion). **But since the pit isn't real, the car can go through**. **Wait, but** maybe the question is a trick. Let me think again.

Wait, the 3D art is designed to look like a deep chasm, but it's a flat surface. So the car can drive across without any problem. **Therefore, the car can get through**. So the answer is B. Yes.

</think>

Answer: B

Answer: [B]

Figure 12: A demonstration of visual illusions. Both InternVL-3.5-8B and Qwen3-VL-30B-A3B-Thinking argue that the image content is an illusion. However, InternVL-3.5-8B follows an incorrect reasoning path. In contrast, Qwen3-VL-30B-A3B-Thinking, despite excessive reasoning, correctly concludes that the car can get through.

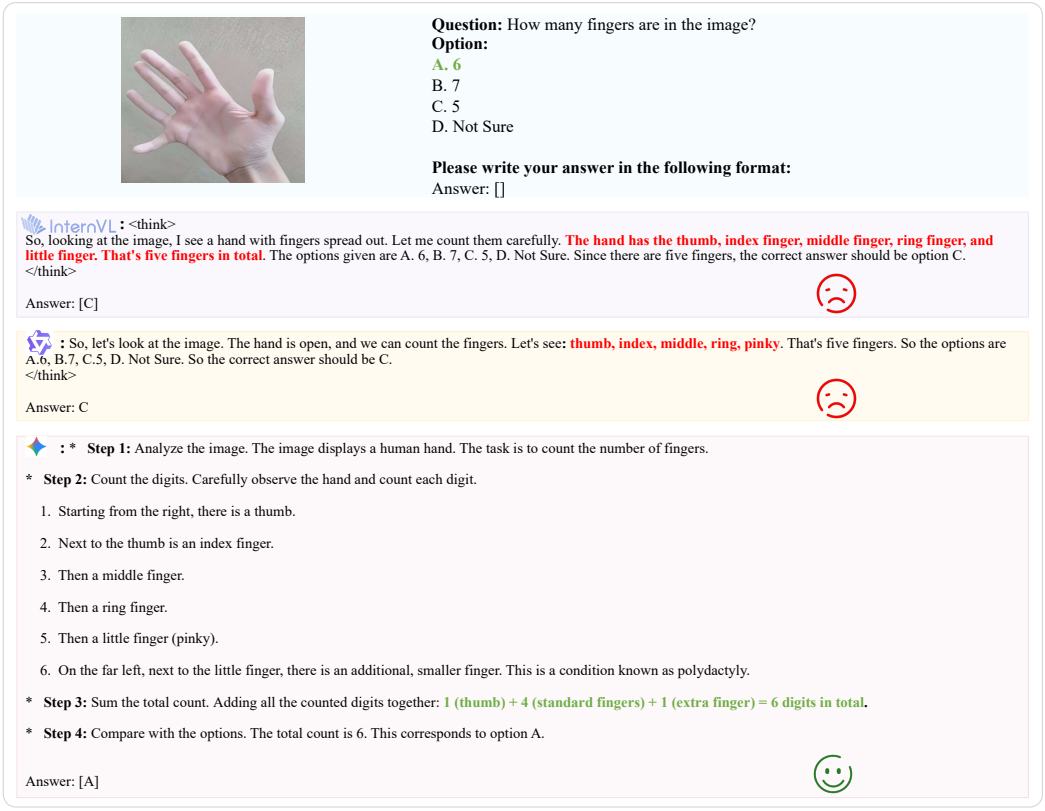


Figure 13: A demonstration of visual anomalies. Visual anomalies that violate commonsense priors, such as biological irregularities (*e.g.*, polydactyly), remain challenging for current MLLMs. As illustrated, InternVL-3.5-8B and Qwen3-VL-30B-A3B-Thinking rely on a "five-finger" prior and miscount, indicating insufficient grounding in the image. By contrast, Gemini-2.5-pro first identifies each finger, then summarizes and counts them, yielding the correct prediction.