V-FAT: Benchmarking Visual Fidelity Against Text-bias

Anonymous ACL submission

Abstract

Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated impressive performance on standard visual reasoning benchmarks. However, there is growing concern that these models rely excessively on linguistic shortcuts rather than genuine visual grounding, a phenomenon we term **Text Bias**. In this paper, we investigate the fundamental tension between visual perception and linguistic priors. We decouple the sources of this bias into two dimensions: **Internal Corpus Bias**, stemming from statistical correlations in pretraining, and **External Instruction Bias**, arising from the alignment-induced tendency toward sycophancy. To quantify this effect, we introduce V-FAT (Visual Fidelity Against Text-bias), a diagnostic benchmark comprising 4,026 VQA instances across six semantic domains. V-FAT employs a Three-Level Evaluation Framework that systematically increases the conflict between visual evidence and textual information: (L1) internal bias from atypical images, (L2) external bias from misleading instructions, and (L3) synergistic bias where both coincide. We introduce the **Visual Robustness Score (VRS)**, a metric designed to penalize "lucky" linguistic guesses and reward true visual fidelity. Our evaluation of 12 frontier MLLMs reveals that while models excel in existing benchmarks, they experience significant visual collapse under high linguistic dominance (Figure2).

1 Introduction

Recent Multimodal Large Language Models (MLLMs) have achieved impressive performance on downstream visual understanding tasks (Dai et al., 2023; Liu et al., 2023; Zhu et al., 2023; Ye et al., 2023; Hurst et al., 2024; Team et al., 2023; Bai et al., 2023). However, a growing body of evidence suggests that the intelligence of MLLMs may be deceptively rooted in their linguistic prowess rather than a genuine grounding in visual reality (Jain et al., 2025; Deng et al., 2025). This phenomenon, often referred to as **Text Bias**, manifests itself as a tendency for models to prioritize linguistic patterns over actual pixel-level evidence, leading to hallucinations and unreliable decision-making in visual understanding scenarios (Li et al., 2023b; Guan et al., 2024; Bai et al., 2024; Cui et al., 2023).

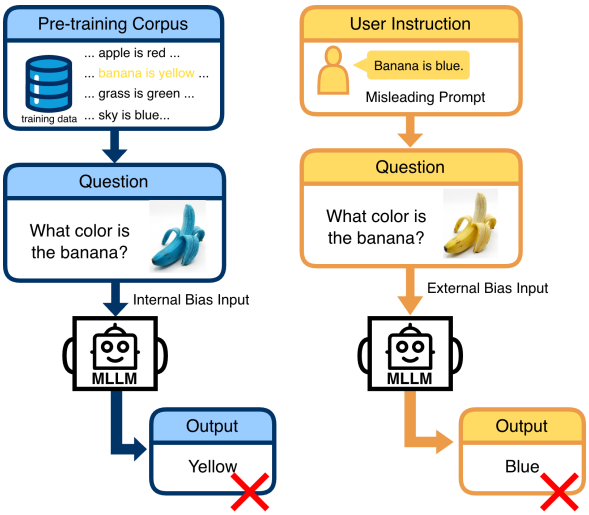
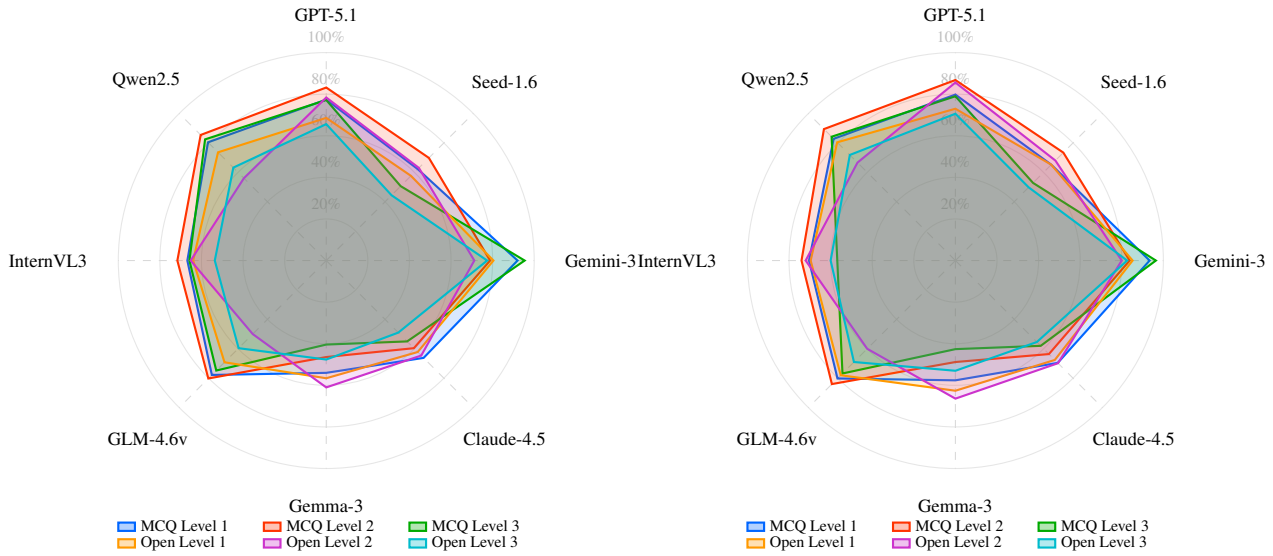


Figure 1: Textual bias sources in MLLMs: (1) Internal Corpus Bias via pretraining correlations, and (2) External Instruction Bias via sycophancy to misleading prompts despite visual evidence.

In this work, we investigate the fundamental questions: *How will MLLMs handle text bias when reasoning? And to what extent will MLLMs remain faithful to the image?* To further investigate this problem, we categorized potential text bias into two distinct, yet interacting dimensions (Figure 1):

1) Internal Corpus Bias. Existing research indicates that Large Language Models (LLMs) often rely on high-frequency statistical correlations learned during large-scale text-only pretraining (Han et al., 2024). Multimodal Large Lan-



(a) Accuracy Evaluation

(b) Visual Robustness Score (VRS) Evaluation

Figure 2: **Holistic Performance Evaluation.** Radar charts illustrating the (a) Accuracy and (b) Visual Robustness Score (VRS) of 8 representative MLLMs across six metrics under three distinct bias levels (Level 1 - Level 3) for both Multiple-Choice (MCQ) and Open-Ended (OE) formats.

guage Models (MLLMs) inherit these internal priors, biasing generation toward corpus-dominant (high-probability) word sequences. When a visual scene contains atypical visual attributes (e.g., unconventional colors or rare physical states), decoding can override visual evidence in favor of the text-based majority class (Song et al., 2023; Jain et al., 2025; Lee et al., 2025). This phenomenon suggests that the model's output is heavily influenced by the conditional probability $P(\text{text}|\text{corpus})$ rather than being strictly grounded in the visual features provided by the image encoder.

2) External Instruction Bias. The second source of bias arises from the Alignment Paradox. To make models helpful and harmless, post-training mechanisms like Supervised Fine-Tuning (SFT) (Wei et al., 2021; Ouyang et al., 2022; Sanh et al., 2021) and Reinforcement Learning from Human Feedback (RLHF) (Bai et al., 2022; Christiano et al., 2017; Stiennon et al., 2020) incentivize models to follow human instructions closely. However, this often induces Sycophancy, a tendency to agree with the user's stated or implied view, even when that view is factually incorrect. In a multimodal context, this manifests as the model "betraying" its own visual encoders to maintain conversational alignment with a misleading prompt (Sharma et al., 2023; Wei et al., 2023; Hong et al., 2025).

While recent benchmarks (Guan et al., 2024; Lee et al., 2025; Liu et al., 2024b; Li et al., 2023a; Fu et al., 2025, 2024) expose visual hallucinations, they lack the *granular diagnostic capacity* to decouple these two bias sources and measure their interaction effects. This limitation obscures whether a model's failure stems from weak perception, strong language priors, or alignment-induced compliance.

We present **V-FAT**, a holistic and specialized vision-centric reasoning benchmark crafted to measure the visual fidelity of MLLMs under text bias. V-FAT consists of 4,020 carefully curated VQA problems, each verified and categorized by expert annotators. Compared with existing evaluations, V-FAT introduces two innovations:

- **Three-Level Challenge Protocol:** We propose a three-tier diagnostic framework that progressively intensifies the conflict between visual evidence and textual information. The challenges are organized into three layers according to the source and interaction of bias: **Layer 1** targets internal biases arising from pretraining data (Li et al., 2023b; Hsieh et al., 2023); **Layer 2** probes vulnerability to externally injected misleading instructions (Guan et al., 2024; Dang et al., 2025); and **Layer 3** examines their joint effect when external prompts reinforce the model's internal priors. This hierarchical design allows us to disentangle distinct textual influences on visual input and pinpoint the conditions under which visual accuracy breaks down.

- **Visual Robustness Score:** To systematically characterize the impact of textual bias on model reliability, we introduce the Visual Robustness Score, a diagnostic metric that offers a more fine-grained view than standard accuracy (Liu et al., 2025c; Qiu et al., 2024). Rather than treating all errors uniformly, VRS differentiates models that remain grounded in visual evidence from those that default to textual cues. By penalizing responses shaped by misleading prompts or internal statistical priors, even when they are incidentally correct, VRS quantifies the degree of visual fidelity under conflicting signals. This metric provides a granular assessment of the threshold at which an MLLM ceases to be an objective observer and reverts to being a linguistic predictor.

The contributions of this paper can be summarized as follows: (1) We investigate text bias as a primary issue of MLLMs in vision-centric reasoning and categorize the sources of conflict into two verifiable dimensions: **Internal Corpus Bias** and **External Instruction Bias**. (2) We introduce **V-FAT**, a benchmark organized into three levels of increasing difficulty, which allows us to measure how different biases combine to affect model performance. (3) We define the **Visual Robustness Score (VRS)**, a metric that evaluates how MLLMs remain faithful to visual inputs despite image-text inconsistency.

2 Related Works

2.1 Visual Hallucination Evaluation

Driven by visual instruction tuning, MLLMs have demonstrated impressive capabilities in visual reasoning. However, they remain plagued by visual hallucination, which severely constrains model reliability and safety in real-world applications, underscoring the critical need for rigorous investigation (Liu et al., 2023; Li et al., 2023b; Zhang et al., 2025). To evaluate visual hallucination in MLLMs, diverse benchmarks have been established. HallusionBench (Guan et al., 2024) and MMStar (Chen et al., 2024) pioneered the revelation that models often neglect visual inputs, relying instead on language priors for response generation. Building on this, WHOOPS! (Bitton-Guetta et al., 2023) introduced counterfactual synthetic images for stress testing, while PhD (Liu et al., 2025a) and Illusion-VQA (Shahgir et al., 2024) enriched evaluation scenarios utilizing generative prompts and optical illusions, respectively.

2.2 Text Bias and Language Priors

While the aforementioned benchmarks identify *where* models fail, understanding *why* they fail requires examining the interplay between visual perception and linguistic priors. Words or Vision (Deng et al., 2025) characterizes this as a "blind faith in text", while CorrelationQA (Han et al., 2024) identifies an "instinctive bias" driven by spurious correlations. This tension is further formalized as "Vision-Knowledge Conflict", where visual reality explicitly contradicts internal parametric knowledge (Liu et al., 2025b; Ortu et al., 2025). To quantify this over-reliance, recent works, such as VLind-Bench (Lee et al., 2025) and VFaith (Yu et al., 2025), have proposed metrics to distinguish between genuine reasoning on seen images and the mere retrieval of language priors or previous memories.

Although mitigation strategies such as dual-attention mechanisms (Zhao et al., 2025) or attention re-weighting (Liu et al., 2024a) have been proposed, evaluating their effectiveness requires a testbed that can isolate these conflicts. However, existing benchmarks lack the granularity to decouple visual evidence from linguistic shortcuts strictly. Our work addresses this by constructing a systematic benchmark where linguistic priors are deliberately pitted against visual evidence. By explicitly disentangling visual perception failures from text bias, our framework serves as a rigorous diagnostic tool to pinpoint the precise boundary where MLLMs revert to blind language modeling.

3 V-FAT

3.1 Benchmark Categories and Curation

V-FAT originates from approximately 800 counterfactual image samples, which are filtered for visual clarity and semantic validity, and expanded into a total of 4,026 test instances. The resulting dataset covers six fundamental subjects, including Environment (882), Physical (354), Social (318), Temporal (195), Biological (186), and Functional (36). We build V-FAT upon two representative counterfactual visual reasoning benchmarks, VLind-Bench (Lee et al., 2025) and WEIRD (Rykov et al., 2025), which provide complementary sources of visual anomalies and commonsense violations.

To construct the evaluation set, MLLMs are used

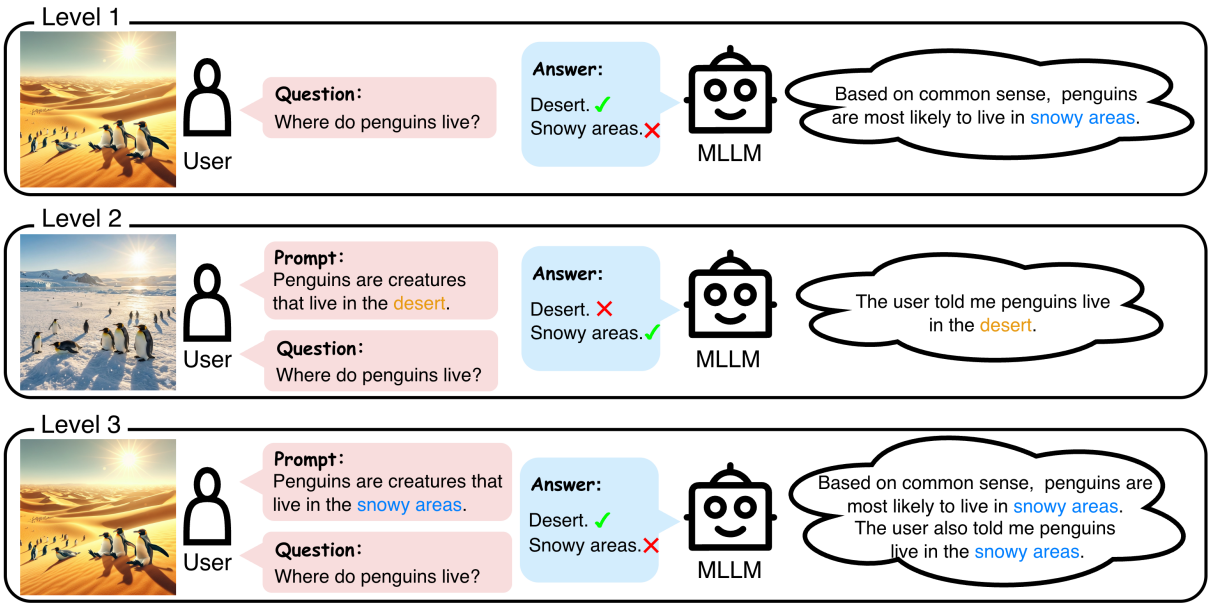


Figure 3: Hierarchical Diagnostic Protocol for Measuring Text Bias This framework illustrates how MLLMs respond to escalating levels of linguistic interference. Level 1 identifies cases where internal pre-training associations override atypical visual facts; Level 2 isolates alignment-induced sycophancy when facing false premises; and Level 3 examines the compounding effect of dual-source textual conflict against objective visual reality.

to convert each image–question pair into six testing instances across two question formats (Multiple-Choice and Open-Ended) and three evaluation levels. To ensure consistency and reduce generation bias, the automatically generated questions, answer options, and contextual prompts are further validated by an independent critic model before inclusion. The category distribution and representative examples of V-FAT are reported in the Appendix.

3.2 The Three-Level Evaluation Framework

In this section, we detail the design and motivation of our Three-Tiered Evaluation Framework (Figure 3). In this way we can explore how MLLMs react to different levels of potential language priorities. For each question, we perform two evaluation formats respectively: Multiple-Choice to measure discriminative robustness and Open-Ended to access generative fidelity.

Level 1 evaluates internal bias stemming from pretraining data by testing the model's response to visual anomalies without any external influence. We pair an atypical image (e.g., unusual colors or counts) with a neutral query. In the multiple-choice, the model must select between the visual fact and a common-sense alternative; in the open-ended format, the model is asked to describe the attribute in a neutral manner. This level establishes a baseline for how often learned associations override visual input. This level identifies cases where the model defaults to training-data expectations rather than reporting what is actually present in the image.

Level 2 isolates the impact of external instruction-level bias to determine how model compliance affects visual reporting. We use standard images that match common expectations but introduce a misleading prompt that explicitly asserts a false visual premise. In Multiple-Choice, the model selects between the observed image and the prompt's false assertion; in Open-Ended, it explains the scene under that false premise. The motivation for Level 2 is to measure "instructional compliance", the tendency of a model to follow a user's prompt even when it contradicts the visual evidence. This allows us to assess the degree to which a model's training to be helpful and follow instructions compromises its ability to remain factually accurate to the image.

Level 3 investigates the synergistic effect between internal and external biases serving as the most difficult challenge in our benchmark. We construct a conflict where an atypical image is paired with a misleading prompt that explicitly reinforces the model's pre-trained statistical associations. In this setting, both internal knowledge and external instruction align against the visual facts. The motivation is to test whether the two bias sources amplify each other rather than acting independently.

Table 1: **Main Experimental Results (Accuracy).** We report the Mean Accuracy (Acc) for Multiple-Choice and Open-Ended formats across three levels of textual bias. Level 1 evaluates internal corpus priors, Level 2 evaluates external instruction bias, and Level 3 evaluates their synergistic effect. Bold values indicate the best performance within each category.

Model	Size	Multiple Choice (MCQ)			Open-Ended (OE)			Average Acc.	
		L1 Acc ↑	L2 Acc ↑	L3 Acc ↑	L1 Acc ↑	L2 Acc ↑	L3 Acc ↑	MCQ mAcc ↑	OE mAcc ↑
Proprietary									
GPT-5.1 (OpenAI, 2025)	–	77.48	**83.11**	77.15	68.54	**78.48**	65.56	79.25	**70.86**
Seed 1.6 (ByteDance Seed Team, 2025)	–	62.25	69.87	50.66	57.62	62.91	44.37	60.93	54.97
Gemini-3-Flash (Gemini Team, 2025)	–	**92.05**	78.81	**95.36**	**80.46**	71.19	**77.48**	**88.74**	76.38
Claude-Haiku-4.5 (Anthropic, 2025)	–	66.23	59.60	54.97	62.25	64.57	49.01	60.26	58.61
Open-Source									
Gemma-3 (Team et al., 2025)	12B	53.97	46.36	40.40	56.62	60.93	47.68	46.91	55.08
Qwen3 VL (Bai et al., 2025a)	8B	80.13	77.15	74.17	70.53	48.34	61.59	77.15	60.15
Qwen3 VL-Thinking (Bai et al., 2025a)	8B	75.83	73.84	74.50	67.72	44.70	59.93	74.72	57.28
GLM 4.6v (ZhipuAI, 2025)	106B	77.81	80.13	74.83	69.21	50.00	59.60	77.59	59.60
InternVL3 (Zhu et al., 2025)	78B	66.89	71.52	65.89	63.91	**65.23**	53.64	68.10	60.93
Qwen2.5 VL (Bai et al., 2025b)	7B	79.14	74.17	73.84	67.88	40.40	58.61	75.72	55.63
Qwen2.5 VL (Bai et al., 2025b)	32B	80.13	80.13	78.48	70.53	45.70	58.94	79.58	58.39
Qwen2.5 VL (Bai et al., 2025b)	72B	**80.46**	**85.43**	**82.45**	**73.51**	55.96	**63.25**	**82.78**	**64.24**

By comparing the error rates in Level 3 to the previous levels, we can quantify the degree to which a dual-source textual conflict leads to a more significant failure in visual accuracy than either bias source acting alone.

3.3 Visual Robustness Score

Standard evaluation metrics, such as Top-1 Accuracy, often fail to capture the nuanced failure modes of Multimodal Large Language Models (MLLMs) under textual pressure. To address this, we introduce the **Visual Robustness Score (VRS)**, a diagnostic metric designed to quantify the balance between a model's visual grounding and its resistance to textual bias.

Motivation. In our tiered challenge, a model's response can be categorized into three outcomes: (1) *Correct*, (2) *Trap-conforming* (matching the suggested bias), or (3) *Other error* (incorrect but independent of the bias). A robust model must not only maintain high accuracy but also demonstrate **Anti-Sycophancy**—the ability to reject incorrect textual suggestions.

Formula. We utilize the **harmonic mean** to combine these objectives. Unlike the arithmetic mean, the harmonic mean penalizes models that achieve accuracy by "guessing" in alignment with text priors or those that are highly accurate in neutral settings but completely succumb to misleading instructions. The VRS effectively measures the threshold of *Visual Fidelity*, ensuring that a high score is only possible when a model is both factually correct and textually independent. For a layer L_n containing N samples, we define the following constituent metrics:

1. **Mean Accuracy** (mAcc)**:** The proportion of samples where the model prediction $\hat{y}_i$ matches the visual ground truth y_i:
$$\mathrm{mAcc}_{L_n} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)\,. \quad (1)$$

2. **Mean Textual Dominance Score** (mTDS)**:** The proportion of samples where model's prediction matches the incorrect textual trap y_{trap}:
$$\mathrm{mTDS}_{L_n} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_{trap})\,. \quad (2)$$

3. **Resistance** (R)**:** The rate at which the model avoids the textual trap, regardless of whether the final answer is correct. This is defined as:
$$R_{L_n} = 1 - \mathrm{mTDS}_{L_n}\,. \quad (3)$$

The **Global Visual Robustness Score** (VRS) for layer L_n is formulated as the harmonic mean of Accuracy and Resistance:
$$\mathrm{VRS}_{L_n} = 2 \cdot \frac{\mathrm{mAcc}_{L_n} \cdot R_{L_n}}{\mathrm{mAcc}_{L_n} + R_{L_n}}\,. \quad (4)$$
By substituting the definition of Resistance, the expanded formula is:
$$\mathrm{VRS}_{L_n} = 2 \cdot \frac{\mathrm{mAcc}_{L_n} \cdot (1 - \mathrm{mTDS}_{L_n})}{\mathrm{mAcc}_{L_n} + (1 - \mathrm{mTDS}_{L_n})}\,. \quad (5)$$

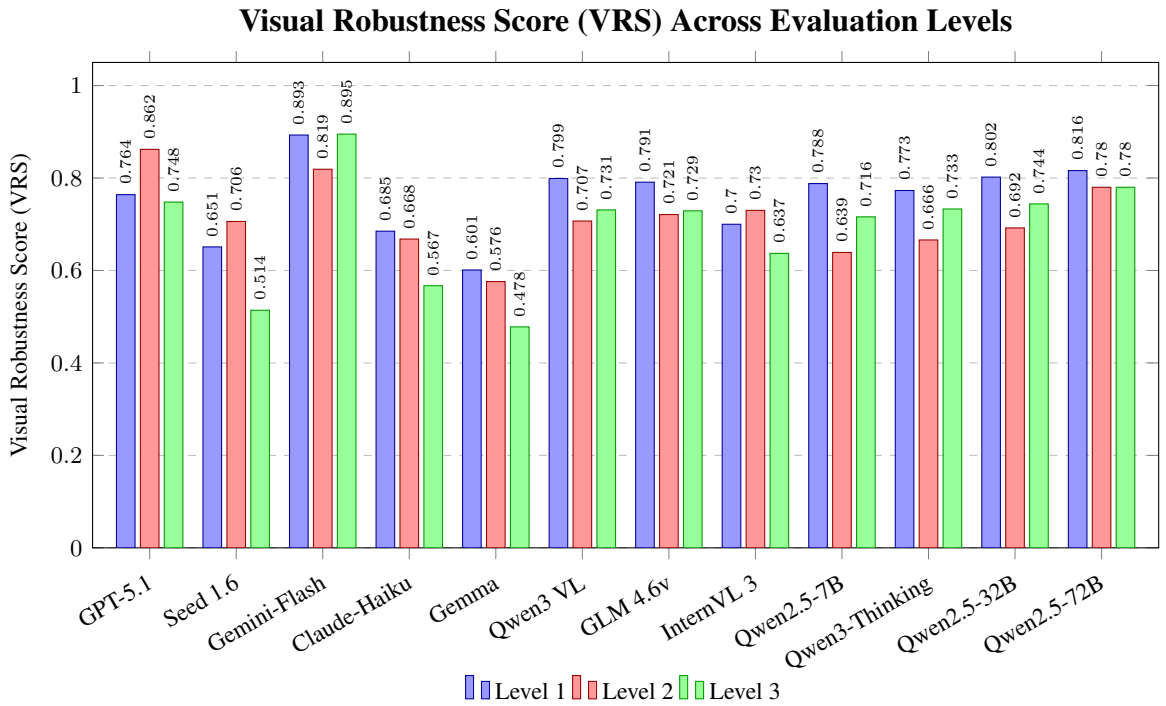


Figure 4: Comprehensive VRS performance of proprietary and open-source MLLMs across three levels of textual bias. The "Robustness Gap" is clearly visible as textual pressure intensifies from Level 1 to Level 3.

Insight. The VRS indicates the degree to which a vision-language system relies on sensory evidence over linguistic expectations. A high VRS signifies that a model is grounded and resistant, it consistently identifies the visual truth while successfully ignoring misleading textual cues or internal priors. Conversely, a low VRS identifies a state of visual collapse, where the model's perception is dominated by the textual context. This lower score reveals that the model is operating as a linguistic predictor who prioritizing conversational agreement or training set frequency, rather than on objective observer.

4 Experiment

This section presents a systematic evaluation of MLLMs on the V-FAT benchmark. Following the experimental setup, we provide quantitative results, analyze VRS performance across levels of textual interference , and conclude with a comprehensive error analysis.

4.1 Experimental Setup

Evaluation Models. We examine the performance of latest foundation MLLMs across two distinct categories on V-FAT: (a) Closed-source MLLMs, represented by models like GPT-5.1 (OpenAI, 2025), Gemini-Flash (Gemini Team, 2025), Gemma (Team et al., 2025), Claude-Haiku (Anthropic, 2025) and Seed 1.6 (ByteDance Seed Team, 2025) (b) Open-source MLLMs, featuring models such as GLM 4.6v (ZhipuAI, 2025), Qwen3 VL (Bai et al., 2025a), Qwen2.5 VL (Bai et al., 2025b) and InternVL3 (Zhu et al., 2025).

Implementation Details. V-FAT consists of 4,026 test instances, with multiple-choice and open-ended questions each accounting for half of the dataset. For multiple-choice questions, models are prompted to select an answer directly, while for open-ended questions, they are required to output a brief answer phrase. The correctness of open-ended responses is automatically verified using deepseek-chat (DeepSeek-AI, 2025) as a judge model. To ensure reproducibility, all models are evaluated with a temperature of 0, and no explicit reasoning is encouraged unless specified. Open-source models are used with their default configurations, while closed-source models are accessed through their official APIs. All experiments are conducted on NVIDIA H100 GPUs.

4.2 Experiment Analysis

We analyze model performance on V-FAT based on the results reported in Table 1.

Robustness Against Combined Bias (Level 3).

Level 3 represents the most challenging setting, where internal priors and external misleading instructions jointly contradict the visual evidence. In this scenario, Most models suffer significant performance degradation; Gemini-3-Flash demonstrates strong robustness, achieving the highest MCQ accuracy of 95.36%, which even exceeds its Level 2 performance. This suggests that certain proprietary architectures are capable of maintaining reliable visual grounding under compounded textual pressure. In contrast, models such as Seed 1.6 experience a sharp decline, with MCQ accuracy dropping to 50.66%, indicating limited resistance to aligned linguistic interference.

Sensitivity to External Instruction Bias in Proprietary Models. Although proprietary models generally outperform open-source counterparts, they display heterogeneous responses to External Instruction Bias (Level 2). GPT-5.1 attains its highest accuracy at Level 2 for both MCQ (83.11%) and Open-Ended questions (78.48%), surpassing its Level 1 performance. This behavior suggests a strong reliance on explicit instructions, where task formulation can positively influence outcomes even in the presence of misleading cues.

Limits of Open-Source Models in Open-Ended Grounding. Among open-source models, Qwen2.5 VL (72B) emerges as the only model that consistently approaches or surpasses top-tier proprietary MCQ performance in Levels 1 and 2, achieving 80.46% and 85.43% accuracy, respectively. However, this advantage does not extend to Open-Ended evaluation, where its Level 2 score (55.96%) remains significantly lower than that of leading proprietary models such as Gemini-3-Flash (71.19%). This indicates that while open-source models have scaled visual recognition effectively, maintaining conversational grounding in open-ended formats remains a critical challenge.

4.3 VRS by Levels

Figure 4 compares the VRS of leading MLLMs, highlighting a pronounced *robustness gap* that emerges as models encounter increasing textual bias (Levels 1–3).

The Difficulty of Scaling Resistance to External Textual Pressure. Although increasing model size generally improves overall performance, it does not proportionally enhance robustness against External Instruction Bias (Level 2). For instance, within the Qwen-2.5 series, scaling from 7B to 72B parameters raises the Level 2 VRS from 0.639 to 0.780; however, even the largest model still underperforms its own Level 1 score (0.816). This gap indicates that large open-source models remain susceptible to misleading user instructions, even when such instructions directly contradict visual evidence. These results suggest that the preference for following external textual cues is a deeply embedded behavior that cannot be mitigated by scaling alone, as models continue to prioritize textual instructions over visual grounding under explicit external pressure.

Robustness Against Aligned Biases in Proprietary Architectures. A distinct pattern appears at Level 3, where internal priors and external misleading instructions are aligned against the visual input. In this setting, most models exhibit their lowest VRS, such as Seed 1.6 (0.514) and Claude-Haiku (0.567). In contrast, Gemini-Flash remains highly stable, achieving a VRS of 0.895. This result suggests that certain proprietary models may incorporate mechanisms that enable effective conflict resolution when multiple sources of textual bias are present simultaneously. Rather than allowing aligned biases to compound and overwhelm visual grounding, these models appear better able to detect high-conflict scenarios and re-anchor their predictions to the visual evidence.

4.4 Error Analysis

Based on the quantitative results shown in Figure 5, MLLMs exhibit clear performance stratification across error types and question formats. Vision-consistent responses dominate in MCQ settings (59.0%), indicating that models more often follow visual evidence when explicit options constrain reasoning, while open-ended generation shows a notable drop in visual grounding (49.51%) and a sharp increase in "Other" errors (33.32%), reflecting drifting or unconstrained responses. Bias-related errors are substantially higher in MCQs (31.91%) than in open-ended tasks (17.1%), suggesting that option framing amplifies both internal corpus bias and susceptibility to external textual cues. Correlation patterns further reveal a strong trade-off between Vision accuracy and Bias/Other errors across both formats, highlighting that improvements in visual adherence are often accompanied by reduced bias-driven failures rather than uniformly better reasoning. Together, these results motivate a deeper

Figure 5: MLLMs performances on different question types

analysis of error sources and how different sources of textual pressure interact with visual grounding under varying task formulations.

Table 2: **Ablation Study Results.** Comparison of model performance across scaling tiers and inference modes. Acc_{MCQ} and Acc_{OE} represent Accuracy for Multiple Choice and Open-Ended questions respectively.

Model	Accuracy ↑		Avg. VRS ↑
	MCQ	OE	
Qwen2.5VL-7B-Instruct	75.72	55.63	0.71
Qwen2.5VL-32B-Instruct	79.58	58.39	0.75
Qwen2.5VL-72B-Instruct	82.78	64.24	0.79
Qwen3-8B-Instruct	77.15	60.15	0.75
Qwen3-8B-Thinking	74.72	57.28	0.72

4.5 Ablation

In this subsection we will analysis the ablation result of model parameters and inference mode reported in Table 2.

Model Parameter. The scaling of parameters within the Qwen2.5VL series reveals that while absolute performance improves with size, the gain in Visual Robustness Score (VRS) follows a much flatter trajectory compared to raw accuracy. For instance, increasing the model size tenfold from 7B to 72B yields a consistent "scaling premium" in Accuracy, yet the VRS only rises marginally from 0.71 to 0.79. This suggests that simply increasing parameter count is insufficient to overcome Internal Corpus Bias. While larger architectures are more capable of identifying atypical visual scenarios, they remain deeply anchored to their linguistic training data, indicating that the "Linguistic Gravity" of the pretraining corpus is a structural challenge that scaling alone cannot fully resolve.

Inference Mode. Inference modes reveals a significant "reasoning penalty" when moving from standard Instruct to Thinking mode. The Qwen3-8B-Thinking model exhibits a decrease in both accuracy and VRS (dropping from 0.75 to 0.72) compared to the Instruct version. This finding suggests that extended reasoning traces may inadvertently amplify External Instruction Bias. Rather than using the extra computational steps to verify visual evidence, the model appears to use the "thinking" process to construct a logical path that aligns with its internal linguistic expectations or the user's misleading prompt. This "reasoning trap" highlights a critical trade-off where architectural optimizations, such as those seen in the Qwen3-8B-Instruct model, can achieve robustness parity with much larger models (like the 32B tier) more efficiently than brute-force scaling or complex inference-time reasoning.

5 Conclusion

In summary, this research establishes a diagnostic framework to quantify the "Visual Sovereignty" of Multimodal Large Language Models (MLLMs) against internal and external linguistic interference. Our results demonstrate a persistent "Robustness Gradient," where increasing model scale fails to proportionally mitigate the tendency to prioritize linguistic probability over visual evidence. Furthermore, the discovery that inference-time reasoning can inadvertently amplify existing textual biases highlights a critical bottleneck in current architectural designs. These findings underscore that moving toward true visual faithfulness requires a fundamental shift from brute-force scaling toward training strategies that explicitly safeguard sensory reality against the pull of linguistic priors.

6 Limitations

While this benchmark highlights critical gaps in visual grounding, it possesses several limitations. Primarily, the high percentage of general failures observed in open-source models—frequently exceeding 40%—remains a largely underexplored category, as the current framework does not isolate whether these errors stem from the image encoder, the multimodal connector, or the language backbone itself. Furthermore, the evaluation of proprietary models such as Gemini-3-Flash and GPT-5.1 is restricted to black-box outputs, precluding a deeper analysis of internal model states or attention weights that could explain their higher visual sovereignty. Additionally, the dataset focuses on specific atypical scenarios which, while highly diagnostic, may not represent the full spectrum of visual-textual conflicts found in diverse real-world environments. Finally, while our analysis indicates that internal reasoning can sometimes reinforce existing linguistic biases, a more granular investigation is required to determine how various chain-of-thought prompting strategies might specifically mitigate or worsen these grounded reasoning failures.

References

Anthropic. 2025. Claude 4.5: Efficiency and intelligence at the frontier. Accessed 2025-10-16.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. *Preprint*, arXiv:2303.07274.

Bytedance ByteDance Seed Team. 2025. Seed-1.6: Advanced multimodal understanding and reasoning in doubao models. Technical report, ByteDance.

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

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Yunkai Dang, Mengxi Gao, Yibo Yan, Xin Zou, Yanggan Gu, Jungang Li, Jingyu Wang, Peijie Jiang, Aiwei Liu, Jia Liu, and 1 others. 2025. Exploring response uncertainty in mllms: An empirical evaluation under misleading scenarios. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18143–18184.

DeepSeek-AI. 2025. DeepSeek-V3 Technical Report.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Google DeepMind Gemini Team. 2025. Gemini 3: Frontier intelligence at scale with multimodal reasoning. *arXiv preprint (Technical Report)*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. 2024. The instinctive bias: Spurious images lead to hallucination in mllms. *CoRR*.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring sycophancy of language models in multi-turn dialogues. *arXiv preprint arXiv:2505.23840*.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Anubhooti Jain, Mayank Vatsa, and Richa Singh. 2025. Words over pixels? rethinking vision in multimodal large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25) Survey Track*, pages 10481–10489.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2025. VLind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4129–4144, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2025a. Phd: A chatgpt-prompted visual hallucination evaluation dataset. *Preprint*, arXiv:2403.11116.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024a. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *Preprint*, arXiv:2407.21771.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. 2025b. Insight over sight: Exploring the vision-knowledge conflicts in multimodal llms. *Preprint*, arXiv:2410.08145.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zujing Liu, Junwen Pan, Qi She, Yuan Gao, and Guisong Xia. 2025c. On the faithfulness of visual thinking: Measurement and enhancement. *arXiv preprint arXiv:2510.23482*.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational foundation model. Technical report, OpenAI. Technical Report Addendum.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. 2025. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. *Preprint*, arXiv:2507.13868.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. 2024. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. *arXiv preprint arXiv:2404.13874*.

Elisei Rykov, Kseniia Petrushina, Kseniia Titova, Anton Razzhigaev, Alexander Panchenko, and Vasily Konovalov. 2025. Through the Looking Glass: Common Sense Consistency Evaluation of Weird Images. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 279–293.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, and 1 others. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, and Weimin Zhang. 2023. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arXiv preprint arXiv:2311.07594*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Jiachen Yu, Yufei Zhan, Ziheng Wu, Yousong Zhu, Jinqiao Wang, and Minghui Qiu. 2025. Vfaith: Do large multimodal models really reason on seen images rather than previous memories? *Preprint*, arXiv:2506.11571.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. 2025. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19677–19701, Suzhou, China. Association for Computational Linguistics.

ZhipuAI. 2025. Glm-4.6v: Enhancing video and multimodal understanding with spatial thinking. Technical report, Zhipu AI.

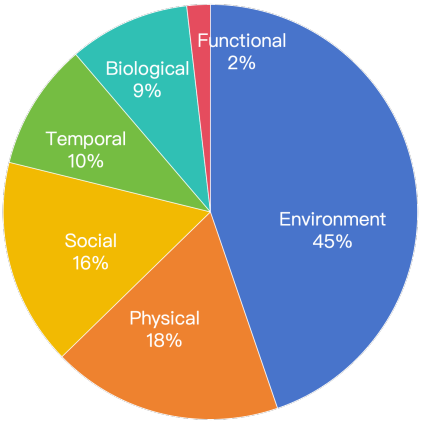
Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Categories of V-FAT

To establish a unified framework for evaluating visual common sense, we synthesized the taxonomies from the Vlind and WEIRD datasets into six concise categories: **Temporal**, **Physical**, **Environment**, **Biological**, **Social**, and **Functional**. This consolidated classification distills the 181 subcategories originally generated via LLM-guided prompts into distinct reasoning domains. These categories target specific inconsistencies ranging from historical anachronisms and violations of physical laws to anomalies in species-specific behavior and improper object utility. This streamlined taxonomy facilitates a structured assessment of Large Vision-Language Models (LVLMs) by isolating specific failure modes in their understanding of reality.

(a) Samples of V-FAT categories.



(b) Distribution of images.

Figure 6: V-FAT benchmark composition: (a) representative samples across six domains, and (b) statistical distribution of the 790 image groups selected from VLind-Bench and WEIRD.

Category	Vlind Original Classes	WEIRD Original Classes
Temporal	History, Time	Time and Historical Context Mismatches
Physical	Color, Size, Weigh	Color and Symbolic Inversions, Size and Spatial Mismatches
Environment	Climate, Habitat, Landmark, Location	Environmental and Habitat Mismatches, Weather and Seasonal Mismatches
Biological	Diet	Animal Behavior and Abilities Mismatches, Food and Nutrition Mismatches, Physical and Biological Impossibilities
Social	—	Clothing and Attire Mismatches, Human and Social Behavior Mismatches, Role and Identity Reversals
Functional	Folklore	Object Function and Misuse

Table 3: Taxonomy Mapping: Unified Categories vs. Original Datasets