Is Your (Reasoning) Multimodal Language Model Vulnerable Toward Distractions?

Ming Liu[1], **Hao Chen**[2], **Jindong Wang**[3], **Liwen Wang**[1], **Jingchen Sun**[4], **Wensheng Zhang**[1]
[1]Iowa State University, Ames, Iowa
[2]Carnegie Mellon University, Pittsburgh, PA
[3]William & Mary, Williamsburg, VA
[4]State University of New York at Buffalo, Buffalo, NY
pkulium@iastate.edu

Abstract

Vision-Language Models (VLMs) have achieved success in tasks such as visual question answering, yet their resilience to distractions remains underexplored. Understanding how distractions affect VLMs' performance is crucial for real-world applications, as input data often contains noisy or irrelevant content. This paper assesses the robustness of VLMs—including general-purpose models and those specialized for reasoning—against distractions in the context of science question answering. We introduce I-ScienceQA, a new benchmark based on the ScienceQA dataset, which systematically injects distractions into both visual and textual contexts. We evaluate how distractions perturb the underlying reasoning processes of these models by analyzing changes in textual explanations leading to answers. Our findings show that most VLMs are vulnerable to distractions, with a noticeable degradation in reasoning when extraneous content is present. In particular, some models (including GPT-o4 mini) exhibit a higher degree of robustness. We also observe that textual distractions generally cause greater performance declines than visual distractions. Finally, we explore mitigation strategies such as prompt engineering. Although these strategies improve resilience modestly, our analysis highlights considerable room for further improvement in the robustness of VLMs.

Code — https://github.com/pkulium/vlm-robustness

Introduction

Despite the remarkable capabilities of vision language models (VLMs) in interpreting images and producing text that resembles human language (Liu et al. 2023a; Dai et al. 2023; Hu et al. 2023), these models continue to exhibit significant vulnerabilities when exposed to irrelevant information. A recent study (Zhang, Yu, and Yang 2024) indicates that VLM-based computer agents are highly susceptible to "pop up attacks". As VLMs form the backbone of such agents, it is crucial to understand their robustness in the face of distractions. In real-world applications, visual and textual inputs are frequently accompanied by noisy and irrelevant information, which can diminish model performance and even lead to incorrect interpretations or hallucinatory responses (Zhou et al. 2024; Chen et al. 2024c).

Recently, reasoning capabilities in both Large Language Models (LLMs) such as Deepseek (DeepSeek-AI 2025) and Multimodal Language Models (MLLMs) such as the MLLM model mentioned in (OpenAI 2025) have drawn considerable attention. LLM-based reasoning models have been shown vulnerable to input perturbations (Huang et al. 2025). However, it remains under-explored whether MLLM-based reasoning models, such as VL-Rethinker (Wang et al. 2025), MM-EUREKA (Meng et al. 2025), and the GPT-o4 mini (OpenAI 2025), exhibit similar vulnerabilities. This paper addresses this by evaluating not only general VLMs but also these specific reasoning-focused MLLMs, investigating how distractions perturb their reasoning processes and consequently affect answer accuracy.

In particular, existing VLM benchmarks assume that both visual and textual inputs are free of distractions, which is rarely true in practice. Previous work has shown that even text-only large language models are vulnerable to irrelevant or misleading context (Shi et al. 2023). VLMs face a compounded challenge: they must cope with noise in images and text simultaneously, making the robustness problem even more complex than in unimodal LLMs.

Moreover, many widely-used benchmarks (Lu et al. 2022; Singh et al. 2019; Lu et al. 2024) use curated inputs that hardly reflect messy, distraction-filled data of real-world settings. This gap makes it difficult to evaluate and trust the robustness of VLMs in deployment, where distractions are inevitable. There is a pressing need for a benchmark that systematically introduces distractions and evaluates how performance degrades (or withstands) under realistic conditions.

To address this gap, we present I-ScienceQA, a comprehensive benchmark designed to investigate the robustness of VLMs towards distractions. Our benchmark, built upon the ScienceQA dataset (Lu et al. 2022), incorporates various types of distraction to simulate more realistic scenarios. Specifically, our aim is to answer the following questions:

- **Vulnerability**: How susceptible are VLMs, including reasoning-specific MLLMs, to distractions in different modalities (visual and textual)?
- **Modality Impact**: Which type of distraction causes greater degradation in model performance?
- **Reasoning Perturbation**: How do distractions quantitatively and qualitatively alter the reasoning of VLMs,

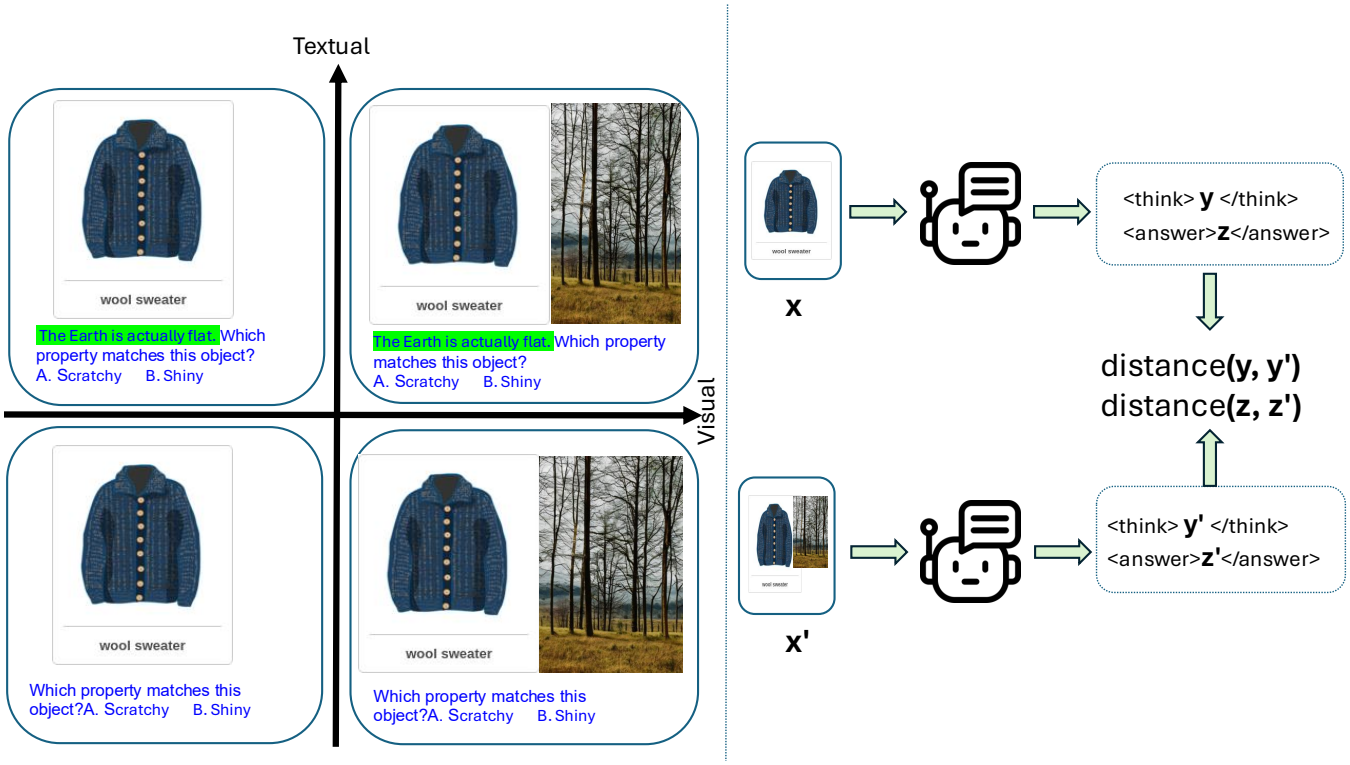

Figure 1: Left: Distractions are inserted into samples from the *ScienceQA* dataset across both textual and visual dimensions; Right: When input are inserted with distraction, the variation in reasoning influence the final answer.

and how does this relate to changes in answer accuracy?

- **Mitigation**: What techniques can help mitigate the impact of distractions and improve the robustness of VLM?

To build I-ScienceQA, we leveraged various models, including GPT-3.5-turbo (OpenAI 2024) and Stable Diffusion models (Rombach et al. 2021). Our benchmark comprises 8,100 samples with four scenarios of distractions in both visual and textual domains (details in extended version). Specifically, we used Stable Diffusion models to generate visual distractions, such as neutral backgrounds, generic landscapes, abstract art, and everyday objects. For textual distractions, we employed GPT-3.5-turbo to produce textual distractions such as contradictory information and irrelevant details. This approach simulates a broad range of real-world noisy conditions that VLMs may encounter. The definition and examples of each distraction type are provided in the extended version. Our methodology for analyzing the impact on reasoning involved comparing the model outputs from clean versus distracted inputs, extracting both the textual explanation (reasoning R_i, R_i') and the final answer (A_i, A_i'). We then employed lexical (Jaccard Similarity, ROUGE, Normalized Levenshtein Similarity) and semantic (Cosine Similarity of Sentence Embeddings) metrics to quantify the reasoning perturbation.

Through an extensive evaluation of the various state-of-the-art VLMs, our key findings include the following.

- **Vulnerability**: Most VLMs show some performance degradation under distractions, although the severity varies by model and scenario. Reasoning models such as VL-Rethinker also exhibit this vulnerability; for instance, under "Add Hint" distractions, while maintaining correct answers in many cases, VL-Rethinker exhibits performance degradation (correct $\rightarrow$ incorrect) in 2.87% of instances, which is associated with more substantial perturbations in its reasoning process.
- **Text vs. Visual Impact**: Textual distractions generally cause greater performance drops than visual ones.
- **Model Size**: Larger models tend to be more robust to distractions. For example, InternVL2 (26B) exhibits minimal performance drops.
- **Mitigation Strategies**: Simple defense approaches (e.g., prompt engineering or using a more robust vision encoder) provide only limited improvements, and their effectiveness varies between models and tasks. However, fine-tuning models on datasets with distractions appears to be an effective strategy for enhancing model robustness.
- **Bimodal Noise**: When visual and textual distractions occur simultaneously, the impact is mixed — some models remain stable, while others show minor fluctuations.

In summary, this study provides insight into the existing limitations of VLMs, particularly those designed for reasoning, and points to promising directions for future improvement in model design and training. Addressing these challenges will pave the way for more robust and reliable VLMs

in real-world applications.

Related Work

Model Evaluations VLMs have traditionally been evaluated using standard Visual Question Answering (VQA) tasks such as TextVQA (Singh et al. 2019), VQAv2 (Antol et al. 2015), and GQA (Hudson and Manning 2019). Currently, studies such as MM-Vet (Yu et al. 2023), POPE (Li et al. 2023), and MM-Bench (Liu et al. 2023b) have emerged to evaluate VLMs, in key challenges such as hallucination and reasoning. These efforts have shown that multimodal LLMs encounter significant issues, such as hallucination (Guan et al. 2023) and insufficient robustness (Fu et al. 2023). In this paper, we introduce the I-ScienceQA benchmark, which highlights that even advanced VLMs, such as GPT-4o (OpenAI 2023), struggle with basic visual questions when irrelevant distractions are present in the input.

Robustness of VLM To test the robustness of the model reasoning, researchers have constructed benchmarks of arithmetic reasoning by paraphrasing or rewriting sentences from clean datasets (Patel, Bhattamishra, and Goyal 2021; Kumar, Maheshwary, and Pudi 2021; Shi et al. 2023). Recent studies have increasingly concentrated on the adversarial robustness of VLMs (Qi et al. 2024; Carlini et al. 2024; Schlarmann and Hein 2023; Zhao et al. 2023; Dong et al. 2023). (Schlarmann and Hein 2023) demonstrate that imperceptible perturbations in input images can enable attackers to manipulate VLMs to generate specific outputs. Visual adversarial attacks designed to jailbreak VLMs are introduced in works such as (Carlini et al. 2024) and (Qi et al. 2024). Recently, studies have focused on training adversary-robust vision encoders (Schlarmann et al. 2024; Mao et al. 2023).

Method: Analyzing the Impact of Distractions on Reasoning

Our methodology is designed to quantitatively and qualitatively assess how distractions, introduced into multimodal inputs, perturb the reasoning process of reasoning VLMs and consequently affect the accuracy of their final answers. We analyze pairs of model outputs: one derived from a clean, distraction-free input and another from an input augmented with a controlled distraction. For a detailed overview of the distraction types in our dataset, see the extended version.

Dataset and Output Processing

Let $\mathcal{D} = \{(I_i, Q_i, T_i)\}_{i=1}^{N}$ be the original dataset of N samples, where I_i is the image, Q_i is the question, and T_i is the target (ground truth) answer. Let $\mathcal{D}' = \{(I'_i, Q'_i, T_i)\}_{i=1}^{N}$ be the corresponding dataset where distractions have been introduced into image (I'_i) or question-context (Q'_i).

For each sample i in $\mathcal{D}$ and its counterpart in $\mathcal{D}'$, we process the model's generated output. The VLM produces a textual response from which we extract two components as shown in Figure 1:

- **Reasoning (R_i, R'_i):** The textual explanation or chain-of-thought leading to the answer. R_i is the reasoning from the clean input, and R'_i is the reasoning from the distracted input. This is extracted by taking all text preceding the final answer delimiter.
- **Answer (A_i, A'_i):** The model's predicted answer. A_i is the answer from the clean input, and A'_i is the answer from the distracted input. This is extracted using a regular expression to find text within a predefined delimiter.

Quantifying Reasoning Perturbation

To measure the extent to which distractions alter the model's reasoning, we employ lexical and semantic similarity metrics.

Lexical metrics assess the surface-level similarity between the original reasoning R_i and the distracted reasoning R'_i.

- **Jaccard Similarity (S_J):** Measures the similarity between the sets of tokens in R_i and R'_i. Let $Tok(R)$ be the set of unique tokens in reasoning R.

$$S_J(R_i, R'_i) = \frac{|Tok(R_i) \cap Tok(R'_i)|}{|Tok(R_i) \cup Tok(R'_i)|}$$

- **ROUGE Scores:** We use ROUGE-N (specifically ROUGE-1, ROUGE-2) and ROUGE-L to compare R_i and R'_i based on n-gram overlap and longest common subsequence, respectively. We report the F1-scores for these metrics.
- **Normalized Levenshtein Similarity (S_L):** Measures the edit distance between R_i and R'_i, normalized by the length of the longer reasoning string.

$$S_L(R_i, R'_i) = 1 - \frac{\text{LevenshteinDist}(R_i, R'_i)}{\max(\text{len}(R_i), \text{len}(R'_i))}$$

where LevenshteinDist is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change R_i into R'_i.

Semantic similarity captures changes in meaning, even if the wording differs significantly.

- **Cosine Similarity of Sentence Embeddings (S_C):** We first obtain sentence embeddings for R_i and R'_i using a pre-trained SentenceTransformer model. Let $E(R)$ be the embedding vector of reasoning R.

$$S_C(R_i, R'_i) = \frac{E(R_i) \cdot E(R'_i)}{||E(R_i)|| \cdot ||E(R'_i)||}$$

Categorization of Answer Outcome

We evaluate the model's answer correctness using Exact Match (EM) against the target answer T_i. Based on this, we categorize the outcome for each sample pair (i, i') into one of four categories:

- **Correct → Correct (C→C):** $A_i = T_i$ and $A'_i = T_i$.
- **Correct → Incorrect (C→I):** $A_i = T_i$ and $A'_i \neq T_i$.
- **Incorrect → Correct (I→C):** $A_i \neq T_i$ and $A'_i = T_i$.
- **Incorrect → Incorrect (I→I):** $A_i \neq T_i$ and $A'_i \neq T_i$.

Linking Reasoning Perturbation to Answer Accuracy

The core of our analysis is to understand how the quantified changes in reasoning relate to changes in answer accuracy. We analyze the correlation between reasoning similarity scores (S_J, S_L, S_C, ROUGE) and the likelihood that an answer changes its correctness status, particularly focusing on C→I transitions. We also correlate these reasoning changes with the overall degradation of Exact Match (ΔAccuracy) observed in the presence of distractions. This analysis is performed on different types of distraction.

Experimental Setup

Models We evaluated a total of 14 vision-language models on machine with A100(80GB) GPUs. Our selection includes both open-source models and one proprietary model, to provide a broad perspective on robustness. In particular, we tested LLaVA-1.5 (7B, 13B) (Liu et al. 2023a), InstructBLIP (7B, 13B) (Dai et al. 2023), Phi3-V (4B) (Abdin et al. 2024), InternVL2 (1B, 2B, 8B, 26B) (Chen et al. 2024b), CogVLM2 (19B) (Hong et al. 2024), Qwen2-VL (2B, 8B) (Wang et al. 2024), Qwen2.5-VL(7B) (Qwen Team, Alibaba Group 2025), Phi-4-multimodal-instruct(6B) (Abouelenin et al. 2025) and GPT-4o models. For reasoning model, we include VL-Rethinker(7B) (Wang et al. 2025), MM-EUREKA(7B) (Meng et al. 2025) and GPT-o4 mini. These models represent the state of the art, and by including multiple size variants for several architectures, we can analyze how model scale and design affect robustness to distractions.

Evaluation Metrics We quantify performance using exact match accuracy and measure robustness via the drop in accuracy due to distractions:

- **Exact Match** This is the percentage of questions the model answers correctly. Formally, for a model F evaluated on a set of test instances D, we define

$$\text{Accuracy}(\mathcal{F}; \mathcal{D}) = \frac{\sum_{d \in \mathcal{D}} \mathbf{1}\left[\mathcal{F}(d) = y_d\right]}{|\mathcal{D}|}$$

  where y_d is the correct answer for instance d and $\mathbf{1}[\cdot]$ is an indicator that equals 1 if the model's answer $F(d)$ is exactly correct and 0 otherwise. An accuracy of 100% means the model got every question correct.
- **Exact Match Degradation** To assess the impact of distractions, we compute the change in accuracy when distractions are present. Let $A_{F,D}$ be the model's accuracy on the original (distraction-free) test set D, and let $A_{F,D'}$ be the accuracy on the corresponding distracted test set D' (each instance in D' is a distracted version of one in D). We define the degradation as

$$\Delta\text{Accuracy}(\mathcal{F}) = A_{\mathcal{F},\mathcal{D}'} - A_{\mathcal{F},\mathcal{D}},$$

  By definition, $\Delta\text{Accuracy}(F)$ could be zero or negative, since adding distractions is not expected to improve performance. A value of 0 indicates the model's accuracy was unchanged despite the distractions (perfect robustness), whereas a more negative value indicates a larger drop in accuracy (greater vulnerability to distractions).

Experimental results

Overall Results

Table 1 summarizes the exact match accuracy of each model on the original dataset and under each distraction scenario, with the degradation (drop in accuracy) due to distractions indicated in parentheses. We note that certain models (Phi-3V and InstructBLIP) are only applicable to scenarios where both image and text inputs are present (they cannot handle solely textual or solely visual inputs), so their entries for the other scenarios are empty in the table.

Add Image: In this scenario (an unrelated image is added), the best-performing models barely lose any accuracy. For example, InternVL2 (8B) achieves an exact match of 94.45% with the added image, only a 1.00 point drop from its 95.45% on distraction-free data. GPT-4o is similarly robust, scoring 93.00% (down just 0.50 from 93.50%). Notably, the reasoning models GPT-o4 mini and VL-Rethinker actually improve slightly (+0.60% and +0.30% respectively). In contrast, smaller models see larger impacts: LLaVA (7B) drops to 68.05% (from 71.30%, a drop of 3.25%) and InternVL2 (1B) falls to 79.70% (from 85.60%, a drop of 5.90%). These results suggest that larger models tend to handle an extra unrelated image better than smaller ones, possibly due to more powerful vision-processing or attention mechanisms.

Insert Image: When a distracting image is inserted alongside an original image, we observe similar trends. InternVL2 (8B) remains highly robust with an accuracy of 94.23% (down only 2.67% from 96.90%). Remarkably, VL-Rethinker and MM-EUREKA show substantial improvements (+6.15% and +6.62% respectively), while GPT-o4 mini also improves slightly (+1.04%). Interestingly, the smaller Qwen2-VL models exhibit almost no performance drop in this scenario: the 2B model goes from 63.80% to 63.26% (-0.54%), and the 7B variant from 68.40% to 68.08% (-0.32%). However, it is important to note that their overall accuracy is relatively low to begin with; so a tiny drop on an already modest score still means these models perform worse in absolute terms compared to larger models. In summary, most models' results in Insert Image mirror those in Add Image: large models like InternVL2 (8B) handle the visual distraction well, while smaller ones either were not adept at the task or only maintained their low performance.

Add Hint: This scenario (adding an irrelevant text hint) proves to be more challenging for the models. Almost every model experiences a larger accuracy drop with textual distractions than extra images. For example, InternVL2 (2B) plummets from 91.40% without distractions to 82.35% with a distracting hint (a drop of 9.05%). The reasoning models also struggle here: VL-Rethinker and MM-EUREKA drop by 5.75% and 6.65% respectively. Even the strongest models are affected: InternVL2 (8B) goes from 94.80% to 93.60% (-1.20%) and InternVL2 (26B) from 95.20% to 92.80% (-2.40%). While those particular drops are small in absolute terms, many other models show substantial declines (e.g., CogVLM2 (19B) falls by 8.10% to 70.50%). These results indicate that adding distracting text tends to confuse the models more than adding unrelated images, likely because misleading linguistic information interferes with the models'

Model	Add Image		Insert Image		Add Hints		Insert Hint	
	Original	Distraction	Original	Distraction	Original	Distraction	Original	Distraction
Phi3v (4B)	–	91.15	–	83.52	–	-	–	-
InstructBLIP (7B)	–	41.05	–	35.45	–	–	–	–
InstructBLIP (13B)	–	47.26	–	47.80	–	–	–	-
Phi4-MM (6B)	–	90.50	–	94.95	–	–	–	-
Qwen2-VL-Instruct (2B)	63.30	63.30 (→0.00)	63.80	63.26 (↓0.54)	60.80	54.45 (↓6.35)	72.45	64.20 (↓8.25)
Qwen2-VL-Instruct (7B)	83.10	83.10 (→0.00)	68.40	68.08 (↓0.32)	75.65	68.00 (↓7.65)	77.40	74.10 (↓3.30)
Qwen2.5-VL-Instruct (7B)	85.20	82.90 (↓2.30)	78.15	84.75 (↑6.60)	85.90	78.95 (↓6.95)	87.15	84.10 (↓3.05)
LLaVA (7B)	71.30	68.05 (↓3.25)	68.75	66.36 (↓2.39)	69.70	63.80 (↓5.90)	70.55	69.30 (↓1.25)
LLaVA (13B)	72.90	72.00 (↓0.90)	72.10	69.60 (↓2.50)	72.15	67.45 (↓4.70)	73.10	71.80 (↓1.30)
LLaVA (34B)	88.05	87.50 (↓0.55)	81.55	79.51 (↓2.04)	84.65	82.65 (↓2.00)	85.50	83.00 (↓2.50)
InternVL2 (1B)	85.60	79.70 (↓5.90)	88.10	83.47 (↓4.63)	87.80	80.55 (↓7.25)	85.90	82.85 (↓3.05)
InternVL2 (2B)	91.35	86.75 (↓4.60)	93.50	90.23 (↓3.27)	91.40	82.35 (↓9.05)	93.65	91.50 (↓2.15)
InternVL2 (8B)	95.45	94.45 (↓1.00)	96.90	94.23 (↓2.67)	94.80	93.60 (↓1.20)	97.60	95.90 (↓1.70)
InternVL2 (26B)	95.35	93.40 (↓1.95)	97.40	95.14 (↓2.26)	95.20	92.80 (↓2.40)	97.55	96.55 (↓1.00)
CogVLM2 (19B)	73.30	71.70 (↓1.60)	89.35	87.47 (↓1.88)	78.60	70.50 (↓8.10)	84.15	80.85 (↓3.30)
GPT-4o (NA)	93.50	93.00 (↓0.50)	80.70	78.56 (↓2.14)	89.50	87.50 (↓2.00)	86.00	84.05 (↓1.95)
VL-Rethinker (7B)	87.65	87.95 (↑0.30)	82.75	88.90 (↑6.15)	88.05	82.30 (↓5.75)	90.30	89.05 (↓1.25)
MM-EUREKA (7B)	89.15	88.10 (↓1.05)	81.85	88.47 (↑6.62)	87.00	80.35 (↓6.65)	90.15	88.50 (↓1.65)
GPT-4o mini (NA)	98.15	98.75 (↑0.60)	94.20	95.24 (↑1.04)	96.90	96.50 (↓0.40)	98.00	96.10 (↓1.90)

Table 1: Exact match scores (%) of various models under different scenarios of distractions. Original columns are evaluated on the corresponding clean subset (same instances as the distracted set, without distractions). The *Distraction* columns present corresponding results on the *I-ScienceQA* benchmark, including both exact match scores and exact match degradation (shown in parentheses). Degradations are color-coded arrows: ↓ for decreases, ↑ for increases, and → for no change. Values marked as – indicate that the model requires both text and image inputs and are therefore excluded from evaluation under that section.

reasoning processes more strongly.

Insert Hint: When an irrelevant text is inserted into an already existing hint (mixing distractions into otherwise useful context), the models generally handle it a bit better than the wholly added hints, but performance still dips. InternVL2 (8B) maintains a high accuracy of 95.90% (down only 1.70% from 97.60%), and GPT-4o scores 84.05% (a drop of 1.95%). The reasoning models show modest drops (VL-Rethinker: -1.25%; MM-EUREKA: -1.65%; GPT-o4 mini: -1.90%). However, some smaller models are clearly disrupted by the inserted text: for instance, Qwen2-VL (2B) drops from 72.45% to 64.20% (-8.25%). These mixed outcomes suggest that, while certain models can manage to ignore or overcome an inserted textual distraction, others struggle significantly. The disparities might stem from differences in how models allocate attention or the breadth of training data they have seen (models trained on more diverse data may be more resilient to extraneous information).

Reasoning Model under Distractions

Impact of add hint distractions on VL-Rethinker For the VL-Rethinker model subjected to "Add Hint" distractions, our analysis (related outcome distributions can be seen in extended version) revealed that the model maintained correct answers in 89.67% of cases (C→C outcome). Performance degradation, where a correct answer became incorrect (C→I), occurred in 2.87% of cases. The quality of reasoning, as measured by the similarity between original and distracted reasoning, showed noticeable differences based on outcome and is presented in extended version. For C→C cases, the average semantic similarity was 0.9242, the Jaccard similarity was 0.6668, and ROUGE-L was 0.6860. In contrast, for C→I cases, these similarities dropped to 0.8909 (semantic), 0.5635 (Jaccard), and 0.5436 (ROUGE-L), indicating that incorrect answers under distraction are associated with more substantial perturbations in the reasoning process. Examining the accuracy changes by specific hint distraction types, "non_sequitur" hints had no negative impact ($\Delta = 0.0000$) , while "contradictory" hints caused the largest accuracy drop ($\Delta = -0.0301$), followed by "misleading" hints ($\Delta = -0.0292$). "Ambiguous" and "irrelevant" hints had milder negative impacts ($\Delta = -0.0137$ and $\Delta = -0.0131$, respectively). For other types of distraction, see our extended version.

Experimental analysis

Model Size

Figure 2 illustrates that performance generally improves in all distraction scenarios as the size of the model increases. For the InternVL2 family (1B, 2B, 8B, 26B): moving from 1B to 8B yields a substantial boost in exact match accuracy in every scenario (for example, from 79.7% at 1B to 94.5% at 8B in Add Image). However, the 26B model shows little to no further gain over the 8B model-in fact, we observe a slight plateau or even a minor dip in some scenarios (e.g., 93.40% at 26B vs 94.45% at 8B in Add Image). A similar pattern holds for LLaVA models (7B, 13B, 34B): performance climbs markedly from 7B to 34B (e.g., 68.05% to 87.50% in Add Image; 63.80% to 82.65% in Add Hint), indicating that larger models cope with distractions more effectively. The improvements are the most pronounced between the smallest and the largest variants, suggesting that scaling up

model parameters strengthens the model's ability to ignore noise. That said, the diminishing returns beyond a certain size (for InternVL2, somewhere between 8B and 26B) imply that simply adding parameters is not enough–beyond that point, factors like training technique or architecture may become the limiting factor for robustness.

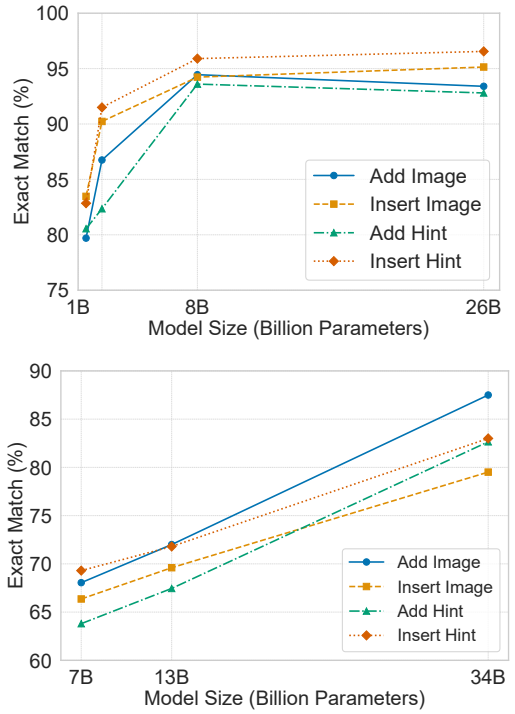

Figure 2: Comparison of Exact Match Score for InternVL2(left) and LLaVA Models(right).

Defending Distractions

Defense via Prompt Engineering We first tried to mitigate distractions by augmenting the prompt with explicit instructions to ignore irrelevant information (see the extended version). This strategy yielded partial success on certain models. For example, Qwen2-VL (2B) saw its accuracy on Add Image rise from 63.30% to 73.80% when a guiding instruction was added - a substantial improvement. The same model also improved in Insert Hint from 64.20% to 70.35% with the defensive prompt. However, this was not universal: the larger Qwen2-VL (7B) model dropped slightly on Add Image (from 83.10% down to 81.35% when instructions were added). Similarly, the large CogVLM2 (19B) model showed inconsistent results, even decreasing in some cases (e.g., Insert Image fell from 87.47% to 85.18% despite the prompt telling it to ignore distractions). These mixed outcomes suggest that prompt engineering helps smaller or certain models focus on relevant content, but it is not a reliably effective solution for all. In some cases, adding instructions can even introduce a slight distraction on its own or conflict with the model's learned attention patterns. Notably, a larger model does not necessarily ensure better robustness to distractions. The effectiveness of prompt engineering appears to depend on model-specific factors.

Defense via Robust Vision Encoder. We also evaluated whether using a more noise-resistant vision encoder could improve robustness (see our extended version). Specifically, we tested the LLaVA-7B model with a *robust-clip* (Schlarmann et al. 2024) (one trained with adversarial robustness) in place of the standard *clip* (Radford et al. 2021). We then compared their performance with our distraction scenarios. The robust-CLIP variant showed only a marginal benefit in one case: in Add Image, it achieved a score 0.85% higher than the standard encoder. In all other scenarios (Insert Image, Add Hint, Insert Hint), the model with the robust encoder actually performed slightly worse than the original LLaVA-7B. For instance, its accuracy with Insert Image was lower than the baseline, indicating no advantage from the supposedly more resilient vision backbone. This suggests that the robust-CLIP encoder, at least as integrated into this VLM, offers minimal gains in ignoring visual distractions.

Neither prompt-based defense nor robust encoder provided significant protection against distractions. These defenses yielded at best modest improvements and at worst small regressions. It highlights the need for more effective strategies—such as better training regimes or architectural innovations—to truly bolster VLMs against distractions. In the following section, we investigate fine-tuning approach aimed specifically at enhancing distraction robustness.

Finetuning on Dataset with distractions

We investigate whether exposing models to distracting data during training improves their robustness. To do this, we split our I-ScienceQA benchmark into a training set and a (held-out) test set. We then use LoRA (Hu et al. 2022) to fine-tune two representative models (LLaVA-7B and Qwen2-VL-7B) on the training portion, which included distracted examples. (Due to their training pipeline constraints that required both image and text inputs, we only used the Add Image and Insert Image scenarios for fine-tuning.) A table in the extended version summarizes the impact of this fine-tuning.

LLaVA-7B saw a significant jump in performance after fine-tuning: on Add Image, its accuracy rose from 51.50% (before) to 64.64%, and on Insert Image from 53.88% to 59.00%. Qwen2-VL-7B also improved, albeit it started from a higher base: Add Image went from 81.37% to 86.50% and Insert Image from 79.25% to 81.88%. These boosts demonstrate that targeted fine-tuning with distracted data can help models learn to ignore or overcome distractions. In particular, LLaVA-7B's relative gains are large, indicating that a model that initially struggled with distractions can greatly benefit from additional training on such examples. This experiment underlines the value of incorporating challenging, noisy data during training: doing so can potentially mitigate the negative effects of distractions and yield more robust and accurate multimodal models.

Generalization

To test whether the vulnerability to distractions is specific to ScienceQA, we constructed a similar benchmark called

I-RealWorldQA based on the RealWorldQA dataset. We applied the same approach of injecting distractions (additional text or images) into RealWorldQA samples. The results, summarized in the extended version, show a consistent pattern: every model's performance declined when distractions were introduced, confirming that the challenge extends beyond the ScienceQA domain. For instance, Qwen2-VL achieved 67.19% accuracy on the original RealWorldQA questions, but this dropped to 62.09% when extra text was inserted, and down to 56.99% when an unrelated image was added. Similarly, LLaVA dropped from 56.47% originally to 54.38% with text distractions and 47.84% with image distractions. Even the powerful GPT-4o model dropped from 61.31% to about 58% under both types of distraction. These across-the-board declines on a different dataset reinforce that the susceptibility of multimodal models to distractions is a general phenomenon. Any robust VLM design will need to address this vulnerability in a variety of contexts.

Combined Visual+Text Distractions

We examine whether adding a visual distraction on top of a textual distraction further degrades performance. For the Qwen2-VL models, the extra image had no effect on accuracy when a text distraction was already present. Both the 2B and 7B Qwen2-VL models scored exactly the same with a text-only distraction as they did with both text+image distractions, in both the Add Hint and Insert Hint scenarios. This suggests that the performance of these models was completely bottlenecked by textual distraction; the irrelevant image was essentially ignored (or it was already failing due to text, so an image could not make it worse). Overall, the effect of dual-modal distractions appears nuanced: some models are robust to an additional distraction in the second modality if one modality is already challenging them, while others exhibit minor fluctuations (either slight further harm or even slight help). These results highlight that the interplay between visual and textual distractions is complex. Ideally, a robust VLM should handle the noise of each modality without letting it interfere with the other, maintaining focus on the truly relevant parts of both image and text.

Pre-Training Dataset and Model Architecture

The robustness of VLMs is influenced by their training datasets and architectural designs. A figure in the extended version summarizes the models' training datasets, vision encoders, and language model component. Notably, some models, such as *InternVL2*, are trained on the ScienceQA dataset, raising concerns about potential **data contamination**. Since the evaluation tasks may overlap with their training data, their performance metrics might be artificially inflated.

The *InternVL2* models combine the InternViT vision encoder with the InternLM2 language model and are trained on a diverse set of datasets, including COCO, VQAv2, OKVQA, Visual Dialog, and ScienceQA. Similarly, *LLaVA* models utilize the CLIP ViT-L/14 vision encoder and *Vicuna* language models, trained on COCO and ScienceQA. In contrast, models such as *InstructBLIP* do not include ScienceQA in their training data. They use datasets such as COCO, VQAv2, OKVQA, and Visual Dialog, leveraging the CLIP ViT-G/14 vision encoder and *Vicuna* language models. Their performance is less likely to be influenced by data contamination, providing a better reflection of their capabilities.

Overall, while diverse training data and sophisticated architectures contribute to model's robustness, the inclusion of evaluation datasets in training can artificially inflate results (Chen et al. 2024a). It is crucial to consider potential data contamination when interpreting performance metrics to ensure fair and accurate assessments of model capabilities.

Conclusion

This paper introduces I-ScienceQA, a comprehensive benchmark designed to evaluate how VLMs withstand distractions. By augmenting the ScienceQA dataset with various types of irrelevant or misleading information, we simulated real-world noisy conditions. We then assessed how these distractions impact the performance and underlying reasoning processes of state-of-the-art VLMs. Our evaluation yielded several key findings: (1) most VLMs are vulnerable to distractions, particularly textual ones; (2) larger models generally exhibit greater robustness, especially against combined visual and textual noise, though size alone does not guarantee immunity; and (3) simple mitigation strategies like prompt engineering or robust encoders offer only partial improvements, highlighting significant room for advancement. Although our study provides valuable insights, it has limitations, which are detailed in following section. In summary, this work underscores the challenges distractions pose to current VLMs and offers a path towards developing more resilient models.

Limitations

Our study has several limitations. (1) We consider a diverse set of distractions (extra images, irrelevant text, etc.), but they do not span the full range of real-world noise; other forms of noise and knowledge irrelevance remain unexplored. (2) We evaluate pre-trained models and basic fine-tuning only, and do not train models or perform adversarial robustness training on noisy data from scratch; such training may improve robustness. (3) While we test combined vision–language distractions, we do not fully characterize cross-modal interaction effects (e.g., noise in one modality amplifying the other). (4) We examine only two defenses (instruction prompting and a robust vision encoder); other approaches, such as visual segmentation to remove irrelevant regions (Lai et al. 2023) or stronger noise-robust attention mechanisms, are left for future work. Despite these limitations, I-ScienceQA provides a practical testbed for robustness evaluation. We hope this work motivates broader distraction types, more robust training, and stronger defenses for reliable VLMs in noisy settings.

Acknowledgements

Jindong Wang acknowledges the unrestricted gifts from Google, Modal academic compute grant, Cohere Labs Cataslyst Research Grant, and Google Cloud Research Credit.

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; Benhaim, A.; Bilenko, M.; Bjorck, J.; Bubeck, S.; Cai, M.; Cai, Q.; Chaudhary, V.; Chen, D.; Chen, D.; Chen, W.; Chen, Y.-C.; Chen, Y.-L.; Cheng, H.; Chopra, P.; Dai, X.; Dixon, M.; Eldan, R.; Fragoso, V.; Gao, J.; Gao, M.; Gao, M.; Garg, A.; Giorno, A. D.; Goswami, A.; Gunasekar, S.; Haider, E.; Hao, J.; Hewett, R. J.; Hu, W.; Huynh, J.; Iter, D.; Jacobs, S. A.; Javaheripi, M.; Jin, X.; Karampatziakis, N.; Kauffmann, P.; Khademi, M.; Kim, D.; Kim, Y. J.; Kurilenko, L.; Lee, J. R.; Lee, Y. T.; Li, Y.; Li, Y.; Liang, C.; Liden, L.; Lin, X.; Lin, Z.; Liu, C.; Liu, L.; Liu, M.; Liu, W.; Liu, X.; Luo, C.; Madan, P.; Mahmoudzadeh, A.; Majercak, D.; Mazzola, M.; Mendes, C. C. T.; Mitra, A.; Modi, H.; Nguyen, A.; Norick, B.; Patra, B.; Perez-Becker, D.; Portet, T.; Pryzant, R.; Qin, H.; Radmilac, M.; Ren, L.; de Rosa, G.; Rosset, C.; Roy, S.; Ruwase, O.; Saarikivi, O.; Saied, A.; Salim, A.; Santacroce, M.; Shah, S.; Shang, N.; Sharma, H.; Shen, Y.; Shukla, S.; Song, X.; Tanaka, M.; Tupini, A.; Vaddamanu, P.; Wang, C.; Wang, G.; Wang, L.; Wang, S.; Wang, X.; Wang, Y.; Ward, R.; Wen, W.; Witte, P.; Wu, H.; Wu, X.; Wyatt, M.; Xiao, B.; Xu, C.; Xu, J.; Xu, W.; Xue, J.; Yadav, S.; Yang, F.; Yang, J.; Yang, Y.; Yang, Z.; Yu, D.; Yuan, L.; Zhang, C.; Zhang, C.; Zhang, J.; Zhang, L. L.; Zhang, Y.; Zhang, Y.; Zhang, Y.; and Zhou, X. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219.

Abouelenin, A.; Absar, N.; Agarwal, S.; Akiki, C.; Al-Ghossein, M.; Alkhereyf, S.; Almahallawi, A.; Awadallah, A. H.; Batra, S.; Bhaskar, A.; et al. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. Technical report, Microsoft.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proc. of International Conference on Computer Vision*, 2425–2433.

Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Awadalla, A.; Koh, P. W.; Ippolito, D.; Lee, K.; Tramer, F.; and Schmidt, L. 2024. Are aligned neural networks adversarially aligned? arXiv:2306.15447.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; and Zhao, F. 2024a. Are We on the Right Way for Evaluating Large Vision-Language Models? arXiv:2403.20330.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024b. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821*.

Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024c. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. *arXiv preprint arXiv:2403.00425*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Dong, Y.; Chen, H.; Chen, J.; Fang, Z.; Yang, X.; Zhang, Y.; Tian, Y.; Su, H.; and Zhu, J. 2023. How Robust is Google's Bard to Adversarial Image Attacks? *arXiv preprint arXiv:2309.11751*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Qiu, Z.; Lin, W.; Yang, J.; Zheng, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

Hong, W.; Wang, W.; Ding, M.; Yu, W.; Lv, Q.; Wang, Y.; Cheng, Y.; Huang, S.; Ji, J.; Xue, Z.; Zhao, L.; Yang, Z.; Gu, X.; Zhang, X.; Feng, G.; Yin, D.; Wang, Z.; Qi, J.; Song, X.; Zhang, P.; Liu, D.; Xu, B.; Li, J.; Dong, Y.; and Tang, J. 2024. CogVLM2: Visual Language Models for Image and Video Understanding. arXiv:2408.16500.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Hu, J.; Yao, Y.; Wang, C.; Wang, S.; Pan, Y.; Chen, Q.; Yu, T.; Wu, H.; Zhao, Y.; Zhang, H.; Han, X.; Lin, Y.; Xue, J.; Li, D.; Liu, Z.; and Sun, M. 2023. Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. *arXiv preprint arXiv:2308.12038*.

Huang, K.; Guo, J.; Li, Z.; Ji, X.; Ge, J.; Li, W.; Guo, Y.; Cai, T.; Yuan, H.; Wang, R.; Wu, Y.; Yin, M.; Tang, S.; Huang, Y.; Jin, C.; Chen, X.; Zhang, C.; and Wang, M. 2025. MATH-Perturb: Benchmarking LLMs' Math Reasoning Abilities against Hard Perturbations. arXiv:2502.06453.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6700–6709.

Kumar, V.; Maheshwary, R.; and Pudi, V. 2021. Adversarial Examples for Evaluating Math Word Problem Solvers. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2705–2712.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*, 292–305.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *NeurIPS*, 36.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023b. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. arXiv:2310.02255.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35: 2507–2521.

Mao, C.; Geng, S.; Yang, J.; Wang, X.; and Vondrick, C. 2023. Understanding Zero-shot Adversarial Robustness for Large-Scale Models. In *The Eleventh International Conference on Learning Representations*.

Meng, F.; Du, L.; Liu, Z.; Zhou, Z.; Lu, Q.; Han, T.; Fu, D.; Shi, B.; Wang, W.; He, J.; Zhang, K.; Luo, P.; Qiao, Y.; Zhang, Q.; and Shao, W. 2025. MM-EUREKA: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning. *arXiv preprint arXiv:2503.07365*. Technical Report.

OpenAI. 2023. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf. Accessed: 2025-12-19.

OpenAI. 2024. GPT-3.5 Turbo. https://platform.openai.com/docs/models. Accessed: 2025-12-19.

OpenAI. 2025. Introducing o3 and o4-mini: Our Smartest Models Yet. Blog post. Accessed: 2025-12-19.

Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In *NAACL-HLT*.

Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21527–21536.

Qwen Team, Alibaba Group. 2025. Qwen2.5-VL Technical Report. Technical report, Alibaba Group.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Schlarmann, C.; and Hein, M. 2023. On the Adversarial Robustness of Multi-Modal Foundation Models. arXiv:2308.10741.

Schlarmann, C.; Singh, N. D.; Croce, F.; and Hein, M. 2024. Robust CLIP: Unsupervised Adversarial Fine-Tuning of Vision Embeddings for Robust Large Vision-Language Models. *ICML*.

Shi, F.; Chen, X.; Misra, K.; Scales, N.; Dohan, D.; Chi, E.; Schärli, N.; and Zhou, D. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. arXiv:2302.00093.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8317–8326.

Wang, H.; Qu, C.; Huang, Z.; Chu, W.; Lin, F.; and Chen, W. 2025. VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning. *arXiv preprint arXiv:2504.08837*. Preprint. Under review.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhang, Y.; Yu, T.; and Yang, D. 2024. Attacking Vision-Language Computer Agents via Pop-ups. arXiv:2411.02391.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M.; and Lin, M. 2023. On Evaluating Adversarial Robustness of Large Vision-Language Models. arXiv:2305.16934.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *The Twelfth International Conference on Learning Representations*.