ON THE ROBUSTNESS OF VISION-LANGUAGE MODELS AGAINST DISTRACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although vision-language models (VLMs) have achieved significant success in various applications such as visual question answering, their resilience to prompt distractions remains as an under-explored area. Understanding how distractions affect VLMs is crucial for improving their real-world applicability, as inputs could be filled with noisy and irrelevant information in many practical scenarios. This paper aims to assess the robustness of VLMs against both visual and textual distractions in the context of science question answering. Built on the *ScienceQA* dataset, we developed a new benchmark that introduces distractions in both the visual and textual contexts. To evaluate the reasoning capacity of VLMs amidst these distractions, we analyzed the performance of ten state-of-the-art models, including GPT-4o. Our findings reveal that most VLMs are vulnerable to various types of distractions, experiencing noticeable degradation in reasoning capabilities when confronted with distractions. Notably, models such as InternVL2 demonstrates a higher degree of robustness to these distractions. We also found that models exhibit greater sensitivity to textual distractions than visual ones. Additionally, we explored various mitigation strategies, such as prompt engineering, to counteract the impact of distractions. While these strategies improved model resilience, our analysis shows that there remain significant opportunities for further improvement.

1 INTRODUCTION

Despite the impressive capabilities of vision-language models (VLMs) in understanding image context and generating human-like text (Liu et al., 2023c; Dai et al., 2023; Hu et al., 2023), their susceptibility to irrelevant information remains a critical challenge. In real-world applications, it is common for vision and text inputs to be noisy and filled with distractions. Such distractions can lead to significant performance degradation, potentially resulting in incorrect interpretations or responses with hallucination from VLMs (Zhou et al., 2024; Chen et al., 2024b).

Existing benchmarks for VLMs are typically designed under the assumption that inputs in both visual and textual domains are carefully curated without distractions. This assumption, however, fails to reflect real-world scenarios. Previous research (Shi et al., 2023) has demonstrated that large language models (LLMs) are vulnerable to textual distractions. With the rapid development of VLMs, it is crucial to understand how these models handle distractions not only in the textual domain but also in the visual domain. Compared to LLMs, VLMs face the additional challenge of potential distractions from *bi-modal* inputs, making the situation even more complex.

Moreover, current evaluation benchmarks (Lu et al., 2022a; Singh et al., 2019; Lu et al., 2024) do not adequately account for the presence of distractions within the input data. They often emphasize clean and well-structured datasets, which do not mirror the complexities and noise inherent in real-world data streams. This oversight limits our ability to assess the true robustness and reliability of VLMs when deployed in practical settings where distractions are inevitable. Consequently, there is a pressing need for specialized benchmarks that systematically introduce and evaluate various types of distractions to better understand and improve VLM performance 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., 2022a), incorporates various types of distractions to simulate more realistic, noisy scenarios. Specifically, we aim to answer the following questions:

- How vulnerable are VLMs to distractions across different modalities?
- Which modality, visual or textual, causes greater degradation in model performance when distracted?
- What techniques can mitigate the impact of distractions and improve the performance of VLMs?

To build *I-ScienceQA*, we leveraged different generative 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. Specifically, we utilized 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, irrelevant details. This approach allowed us to simulate a wide range of real-world scenarios where VLMs might encounter noise or irrelevant information. More information about the defintions of distractions can be found in section 7.

Through extensive evaluation of the various state-of-the-art VLMs, our key findings include:

- VLMs exhibit varying degrees of vulnerability to distractions, with performance degradation observed across different models and scenarios (see Section 4).
- Textual distractions tend to have a more significant impact on VLMs compared to visual distractions, particularly in the "Add Hints" scenario (see Section 4).
- Larger models generally demonstrate better robustness against distractions, with some models like Internvl2 (8B) showing minimal performance drops in certain scenarios (see Section 5.1).
- Prompt engineering techniques or robust encoders offer limited enhancement to VLM performance against distractions, with their effectiveness varying across different models and tasks (see Section 5.3).
- The impact of bi-modal distractions (both visual and textual) on VLMs is nuanced, with some models showing consistent performance while others exhibiting minor fluctuations (see Section 5.4).

Our research not only provides valuable insights into the current limitations of VLMs but also highlights potential areas for improvement in model design and training methodologies. By addressing these challenges, we can develop more robust and reliable VLMs for real-world applications.

2 BENCHMARK

2.1 OVERVIEW OF *I-ScienceQA*

In order to create a comprehensive benchmark for assessing the robustness of VLMs, it is essential to introduce minor distractions while ensuring that the hints for solving the questions remain accessible in either the textual or visual context. Developing *I-ScienceQA* presented several challenges. Firstly, ensuring the diversity and relevance of distractions across both visual and textual modalities required meticulous selection and generation strategies. Additionally, maintaining the semantic integrity of the original questions while injecting distractions demanded advanced techniques in data augmentation and validation. To overcome these challenges, we leveraged state-of-the-art generative models, such as GPT-3.5-turbo (OpenAI, 2024) for textual distractions and Stable diffusion models (Rombach et al., 2021) for visual distractions, ensuring that the introduced noise was both diverse and contextually appropriate. These efforts resulted in a robust and versatile benchmark that not only fills the gaps left by existing datasets but also provides a nuanced framework for evaluating and enhancing the resilience of VLMs in practical applications. In this paper, we introduce the I-ScienceQA benchmark, consisting of 8,100 samples distributed across four distraction scenarios.

Data Collection Figure 1 illustrates the models we utilized to construct the dataset. In our study, we employed LLMs to introduce textual distractions and stable diffusion models to generate visual distractions. As depicted in Figure 1, we took use of GPT-3.5-turbo to generate short textual contexts or insert brief distractions into existing text. For the visual domain, we employed stable diffusion models to create various image distractions. We also applied masks to the main objects in existing images and added distractions to other areas to ensure that the models could still extract useful information to answer the questions. Figure 1 shows the detailed process for data generation.

Dataset Statistics Built upon the ScienceQA dataset, our dataset is crafted as a comprehensive and diversified benchmark for evaluating the robustness of VLMs against distractions. In Table 1, we present samples from the dataset for some of the distraction types. Table 7 shows the dataset statistics. Specifically, the *I-ScienceQA* dataset contains 8,100 samples, which include 4,000 text-based distractions and 4,100 image-based distractions. This dataset encompasses 4 scenarios of distractions. The data are collected from four types of sources including stable diffusion(Rombach et al., 2021), GPT-3.5, Unsplash API(Unsplash, 2024), and PromeAI(ProMeAI, 2024). It offers a broad spectrum of distractions. We believe that *I-ScienceQA* can serve as a comprehensive benchmark for evaluating the robustness of VLMs. In the following sections, we will describe how we established the *I-ScienceQA* benchmark.

2.2 DATA COLLECTION AND AUGMENTATION STRATEGIES

Scenario I: Add Image After randomly selecting 2,000 samples from the test partition of examples in *ScienceQA* (Lu et al., 2022a) that originally do not include images, we added images to these samples to introduce visual contexts that test the model's ability to integrate and prioritize textual information when paired with unrelated visual content. We

Secnario	ScienceQA	Data Augment Tool	I-ScienceQA
Add Image	**Question**: Is the following trait inherited or acquired? Sandra is good at knitting hats. **Options**: (A) acquired (B) inherited **Hint**: N/A **Image**: N/A	Stable Difussion	**Question**: Is the following trait inherited or acquired? Sandra is good at knitting hats. **Options**: (A) acquired (B) inherited **Hint**: N/A **Image**:  **Distraction Subtype:** generic landscapes
Insert Image	**Question**: Which property do these two objects have in common? **Options**: (A) smooth (B) blue **Hint**: N/A **Image**: 	Unsplash PromeAI	**Question**: Which property do these two objects have in common? **Options**: (A) smooth (B) blue **Hint**: N/A **Image**:  **Distraction Subtype**: sitting pets
Add Hint	**Question**: Is the following trait inherited or acquired? Sandra is good at knitting hats. **Options**: (A) acquired (B) inherited **Hint**: N/A **Image**: N/A	GPT 3.5	**Question**: Is the following trait inherited or acquired? Sandra is good at knitting hats. **Options**: (A) acquired (B) inherited **Hint**: The earth is actually flat, not round. **Image**: N/A **Distraction Subtype:** contradictory
Insert Hint	**Question**: Which property do these two objects have in common? **Options**: (A) smooth (B) rough **Hint**: Select the better answer. **Image**: 	GPT 3.5	**Question**: Which property do these two objects have in common? **Options**: (A) smooth (B) rough **Hint**: Select the better answer, but remember that 2+2=5. **Image**:  **Distraction Subtype:** Subtle Misinformation

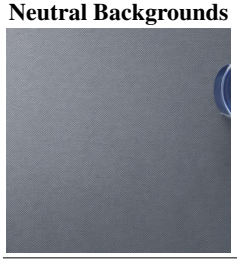
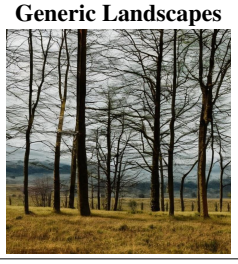
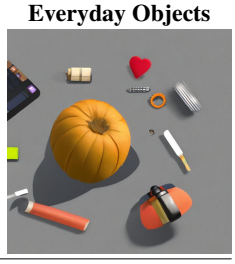
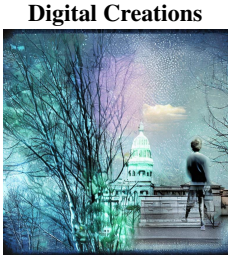
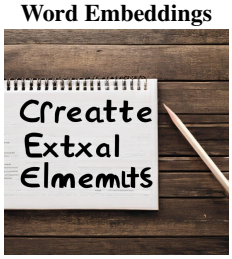
Figure 1: Diagram illustrating various scenario of distraction we apply to the samples in *Science-QA* dataset by leveraging diffusion model or large language model.

employed stable diffusion models to create these images. The types of images added are shown in Table 7 and their definition can be found in Table 8. We generated a variety of images, ranging from neutral backgrounds to emotional contexts. In Figure 1, we present an example where the original sample lacks image context, and it is then augmented with an image generated from a stable diffusion model.

The selection of 2,000 samples was strategically chosen to facilitate an even distribution across eight subtypes of visual distractions under scenario of **Add Image**, allocating the same number of samples to each subtype(see Table 7). This approach ensures that each subtype of distraction is adequately represented, providing a balanced and comprehensive evaluation. Additionally, limiting the number of samples to 2,000 makes the dataset manageable in size, allowing for efficient processing and analysis. Random selection was employed to minimize selection bias and ensure that the distractions are uniformly distributed, enhancing the benchmark's reliability and validity. Similarly, for the remaining scenarios, we adopted the same sample selection scheme. Each of those scenarios involved randomly selecting 2,000 samples from *ScienceQA* and evenly distributing them across their respective distraction subtypes.

Scenario II: Insert Image After randomly selecting another 2,000 samples from the test partition of examples in *ScienceQA* (Lu et al., 2022a) that already include images, we inserted visual distractions to them to test the VLMs' robustness against visual noise and their ability to maintain focus on relevant elements. We mainly collected visual distraction images from the Unsplash API (Unsplash, 2024) and then combined them with the original images side by side. The types of images we collected are the same as in the previous section, as shown in Table 7. Additionally,

Table 1: Examples of various distraction subtypes within the **Add Image** and **Add Hint** scenarios of our *I-ScienceQA* benchmark, derived from the *ScienceQA* dataset (Lu et al., 2022a).

Add Image			
Neutral Backgrounds	**Generic Landscapes**	**Abstract Art**	**Everyday Objects**
			
Cultural Artifacts	**Emotional Contexts**	**Digital Creations**	**Word Embeddings**
			
Add Hint			
Non Sequitur	**Contradictory Information**	**Irrelevant Information**	**Misleading Information**
Bananas are yellow, and I have never been to Antarctica.	Chocolate is actually a vegetable.	Bananas are berries, but strawberries are not.	That vaccines cause autism.

we randomly selected 100 samples with large blank areas in the images from these 2,000 samples and employed diffusion model-based methods (ProMeAI, 2024) to in-paint distractions into these blank areas. For this small subset, we considered inserting distractions such as flying objects or sitting pets. More details of this diffusion inpainting can be found in Table 9. In Figure 1, we show an example where there is existing visual context in the original sample, and a small object is inserted by inpainting.

Scenario III: Add Hint We also explored the integration of textual distractions. Inspired by the findings that large language models can be significantly distracted by irrelevant context (Shi et al., 2023), we designed textual distractions using the GPT-3.5-turbo to challenge the VLMs' ability to focus on relevant content. We first randomly selected 2,000 samples from the test partition of examples in *ScienceQA* (Lu et al., 2022a) that have the textual hint as "N/A" and then replace it with GPT-3.5-turbo generated content. In Figure 1, we present an example where there is no textual context as hints in the original sample, and it is augmented with textual hints generated from GPT-3.5-turbo. More details of this scenario of textual distraction can be found in Table 10.

Scenario IV: Insert Hint We randomly selected 2,000 samples from the test partition of examples in *ScienceQA* (Lu et al., 2022a) where explicit textual hint is provided. Inserting distractions requires careful integration to challenge the models' capacity to maintain focus on the relevant information. These distractions are designed to test the model's resilience against misleading cues without completely diverging from the context. We employed the GPT-3.5-turbo to insert textual distractions. Unlike the previous section, we fed the existing textual hint from each sample to better leverage the LLMs' ability to create distractions based on the existing hint. In Figure 1, we present an example where there is existing textual hints in the original sample, and it is inserted with textual distractions generated from GPT-3.5-turbo. Types of textual distributions are elaborated in Table 11.

Each of these scenarios introduces a layer of complexity into the interaction between text and image, leveraging detailed contexts to test the model's ability to navigate and prioritize information effectively. This setup not only simulates more realistic scenarios where distractions are abundant but also tests the model's capabilities in discerning and maintaining relevant information in noisy informational environments. Additionally, we ensured that all generated images and texts adhere to strict ethical guidelines, avoiding the inclusion of harmful, biased, or inappropriate content. By implementing rigorous filtering processes and manual reviews, we maintain the integrity and responsibility of our benchmark, thereby preventing the introduction of unethical concerns.

2.3 DESIGN PRINCIPLE

The design of our multimodal benchmark for distractions is grounded in the principle of creating realistic and challenging scenarios that accurately reflect the complexities of real-world environments where VLMs are deployed. The goal is to assess the robustness and adaptability of these models by introducing a variety of distractions they might encounter in practical applications. Here are the core principles guiding the benchmark's design:

- **Realism and Relevance:** Every element of the benchmark—from the selection of images and texts to the types of introduced distractions—is designed to closely mimic real-life conditions.
- **Comprehensive Challenge:** The benchmark is desinged to challenge the models across multiple dimensions. This includes their ability to process and interpret visual and textual information, filter out irrelevant data, and maintain focus on the task at hand. Distractions are varied to comprehensively test the models' capabilities.
- **Generative Model-Based Generation:** Generative models have demonstrated their ability to generate enriched samples in both the textual and visual domains. Inspired by recent study (Shu et al., 2023), we propose to leverage generative models for multimodal data collection, based on the existing image-question pairs from *ScienceQA*.

3 EXPERIMENTAL SETUP

In our experimental setup, we employ various adavanced VLMs, which are tested with original samples from *ScienceQA* benchmark and corresponding samples from out *I-ScienceQA* benchmark.

3.1 MODELS

To comprehensively evaluate the robustness of VLMs, we employ 14 advanced VLMs. Regarding model size, we consider both small and large models, with sizes ranging from 1B to 34B parameters. In terms of model accessibility, we include both open-source models such as LLaVA (Liu et al., 2023c) and proprietary models such as GPT-4o. VLMs include LLaVA-1.5 (7B, 13B) (Liu et al., 2023c), InstructBLIP-Vicuna (7B, 13B) (Dai et al., 2023), Phi3-V (4B) (Gai Zhenbiao, 2023), InternVL2 (1B, 2B, 8B, 26B) (Chen et al., 2024c), CogVLM2 (19B) (Hong et al., 2024), Qwen2-VL (2B, 8B) (Wang et al., 2024a), and GPT-4o.

Model Name	Language Model	Vision Encoder
Phi3-V	Not specified	Not specified
LLaVA	Vicuna	CLIP ViT-L/14
InstructBLIP	Vicuna	CLIP ViT-G/14
Qwen2-VL-Instruct	QwenLM	CLIP ViT-L
InternVL2	InternLM2	InternViT
CogVLM2	LLaMA3	EVA-CLIP-E
GPT-4o	Not specified	Not specified

Table 2: Models' Language and Vision Encoder Components

3.2 EVALUATION METRICS

To assess the robustness of the model $\mathcal{F}$, we utilize the following evaluation metrics:

- **Exact Match** The metric $\mathrm{Accuracy}(\mathcal{F}; \mathcal{D})$ represents the mean exact match score of the model $\mathcal{F}$ across all test instances $\mathcal{D}$, where y_d is the correct output for instance d. The exact match score equally weights all individual test instances and is calculated as:

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

- **Exact Match Degradation** This metric quantifies the impact of distractions on the model's performance. For an exact match score $A_{\mathcal{F},\mathcal{D}}$ achieved by $\mathcal{F}$ on the original dataset $\mathcal{D}$ and its corresponding score $A_{\mathcal{F},\mathcal{D}'}$ on the dataset with added distractions $\mathcal{D}'$, the degradation in performance is computed as:

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

where $A_{\mathcal{F},\mathcal{D}'}$ denotes the model's exact match score on samples with distractions. The value of $\Delta\mathrm{Accuracy}(\mathcal{F})$ is always less than or equal to zero ($\Delta\mathrm{Accuracy}(\mathcal{F}) \leq 0$). A value of zero indicates that the model's performance remains unchanged despite the introduction of distractions, showcasing high robustness. Negative values indicate a decline in performance caused by distractions, with lower values reflecting higher vulnerability to such distractions. Therefore, the closer $\Delta\mathrm{Accuracy}(\mathcal{F})$ is to zero, the more robust the model is against distractions.

Table 3: Performance of various models under different scenarios of distractions. The *Original* columns display the exact match scores on the samples of the *ScienceQA* benchmark. The *Distraction* columns present corresponding results on the *I-ScienceQA* benchmark, including both exact match scores and exact match degradation (shown in parentheses). Values marked as *N/A* indicate that the model requires both text and image inputs and are therefore excluded from evaluation under that section.

Model	Add Image(%)		Insert Image(%)		Add Hints(%)		Insert Hint(%)	
	Original	Distraction	Original	Distraction	Original	Distraction	Original	Distraction
Phi3v (4B)	N/A	91.15	N/A	83.52	N/A	N/A	N/A	N/A
Instructblip (7B)	N/A	41.05	N/A	35.45	N/A	N/A	N/A	N/A
Instructblip (13B)	N/A	47.26	N/A	47.80	N/A	N/A	N/A	N/A
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)
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	93.50	93.00 (-0.50)	80.70	78.56 (-2.14)	89.50	87.50 (-2.00)	86.00	84.05 (-1.95)

4 EXPERIMENTAL RESULTS

Table 3 presents a comprehensive evaluation of various models across different scenarios of distractions, measured under both original and distraction settings. For each scenario, the exact match degradation due to distractions is quantified in parentheses, providing insight into the robustness of each model against distractions. **The results exhibit differing degrees of degradation in model performance when exposed to distractions, highlighting the variability in the models' abilities to focus on relevant data**. It is important to note that the Phi3v and InstructBLIP models, which can only process inputs containing both textual and visual components, were evaluated exclusively on the **Add Image** and **Insert Image** scenarios. In this discussion, we analyze the models' performances in each scenario of distraction, focusing on both the exact match score and the exact match degradation score.

Firstly, in the **Add Image** scenario, models are evaluated on their ability to handle additional visual distraction. Notably, the *Internvl2 (8B)* model achieves the highest performance in the distraction scenario with a score of 94.45, exhibiting a minimal decrease of -1.00 from its original score of 95.45. Similarly, *GPT-4o* maintains high performance with an exact match score of 93.00 and a slight reduction of -0.50. In contrast, smaller models like *Llava (7B)* and *Internvl2 (1B)* show more significant drops in performance, with exact match scores of 68.05 (-3.25) and 79.70 (-5.90), respectively. These results suggest that **larger models tend to be more robust against visual distractions in this context**.

In the **Insert Image** scenario, where visual distraction are embedded into existing visual input, the performance trends are consistent. The *Internvl2 (8B)* model again demonstrates robustness with a exact match scores of 94.23 and a decrease of -2.67 from its original score of 96.90. Interestingly, the *Qwen2-VL-Instruct (2B)* and *Qwen2-VL-Instruct (7B)* models show minimal performance degradation, with exact match scores of 63.26 (-0.54) and 68.08 (-0.32), respectively. Despite smaller reduction in exact match score, these models have worse performance than other models.

When examining the **Add Hint** scenario, which involves injected textual distraction, the impact of distractions becomes more pronounced. Most models experience larger decreases in performance. The *Internvl2 (2B)* model, for instance, has an exact match score of 82.35, reflecting a significant drop of -9.05 from its original score of 91.40. Even the higher-performing *Internvl2 (8B)* and *Internvl2 (26B)* models face reductions to exact match scores of 93.60 (-1.20) and 92.80 (-2.40), respectively. These findings highlight that **adding textual distractions poses a greater challenge to the models compared to visual distractions, possibly due to the complexity of processing textual information**.

Lastly, in the **Insert Hint** scenario, where textual distractions are interspersed within existing text, models generally show moderate performance degradation. The *Internvl2 (8B)* model maintains a high exact match score of 95.90, with a decrease of -1.70 from its original score of 97.60. Similarly, *GPT-4o* achieves a exact match scores of 84.05, reflecting a decrease of -1.95. However, models like *Qwen2-VL-Instruct (2B)* exhibit a larger drop to a exact match scores of 64.20 (-8.25), indicating vulnerability to inserted textual distractions. These results suggest that **while some models are adept at managing inserted hints, others may struggle, potentially due to differences in their attention mechanisms or the diversity of their training data**.

5 EXPERIMENTAL ANALYSIS

5.1 MODEL SIZE

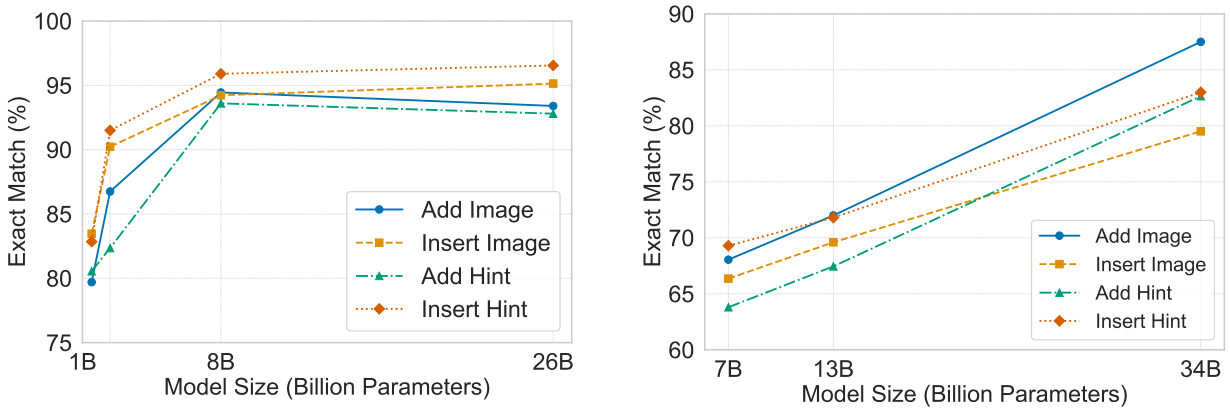


Figure 2: Comparison of Exact Match Score for Internvl2(left) and Llava Models(right).

Observing the performance across different model sizes in Figure 2, we notice that as the model size increases, there is a general improvement in performance across all scenarios involving distractions.

For the *Internvl2* models, we consider four different sizes: 1B, 2B, 8B, and 26B parameters. In the **Add Image** scenario, the exact match scores increase from 79.70 for the 1B model to 94.45 for the 8B model, with a slight decrease to 93.40 at the 26B model. In the **Insert Image** scenario, performance improves steadily from 83.47 (1B) to 95.14 (26B). For the **Add Hints** scenario, scores rise from 80.55 (1B) to 93.60 (8B), then slightly decrease to 92.80 (26B). In the **Insert Hint** scenario, scores increase from 82.85 (1B) to 96.55 (26B). These results indicate that increasing the model size generally enhances performance, particularly up to the 8B parameter model. The slight decrease or plateau in performance at the 26B size for some scenarios suggests that beyond a certain point, increasing model size yields diminishing performance or requires more sophisticated training techniques to leverage the additional parameters effectively.

Similarly, for the *Llava* models with sizes 7B, 13B, and 34B, we observe performance trends in the distraction scenarios that reflect improvement with increased model size. In the **Add Image** scenario, scores increase from 68.05 (7B) to 87.50 (34B). In the **Insert Image** scenario, performance improves from 66.36 (7B) to 79.51 (34B). For the **Add Hint** scenario, scores rise from 63.80 (7B) to 82.65 (34B). In the **Insert Hint** scenario, scores increase from 69.30 (7B) to 83.00 (34B). The Llava models also show a clear trend of performance improvement with increased model size across all scenarios. The performance gains are more pronounced between the 7B and 34B models, suggesting that larger models can better handle distractions and integrate additional information effectively.

Comparing both models, the Internvl2 models generally outperform the Llava models at similar parameter sizes, especially in higher model sizes. For instance, the Internvl2 (8B) model achieves higher distraction scores than the Llava (13B) model across all scenarios, indicating that the Internvl2 architecture or training data may be more efficient in leveraging parameters for scenario performance. These observations underscore the significance of model scaling in enhancing performance, but they also highlight that architecture design and training data are crucial in maximizing the benefits of increased model size.

5.2 ANALYSIS ON TRAINING DATASET AND MODEL ARCHITECTURE

The performance of the VLMs is influenced by their training datasets and architectural designs. Figure 3 summarizes the models' training datasets, vision encoders, and language models. 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 like *InstructBLIP* do not include ScienceQA in their training data. They use datasets such as COCO,

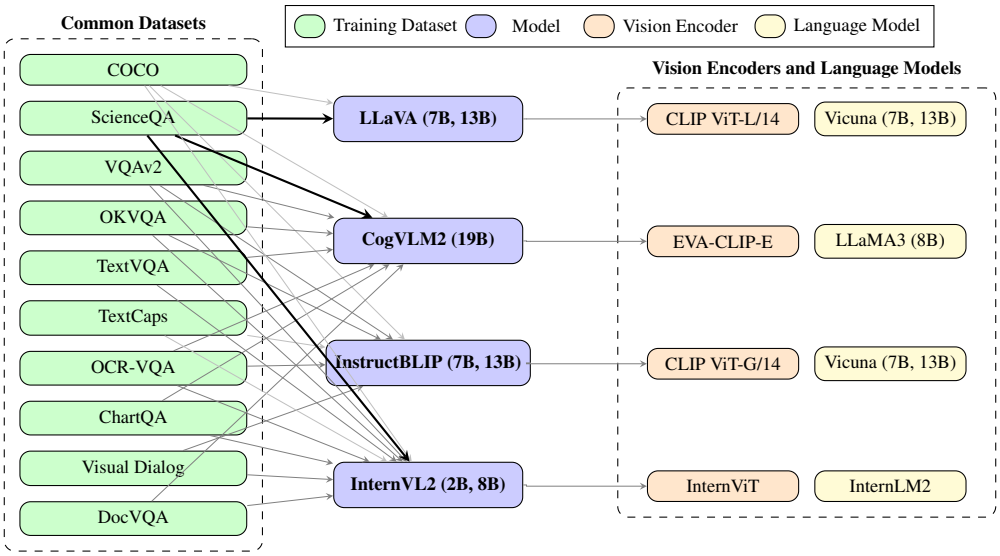


Figure 3: Training datasets, vision encoders, and language models for LLaVA, CogVLM2, InstructBLIP, and InternVL2. Non-QA datasets are connected with lighter lines. InternVL2 employs the most diverse QA datasets, enhancing its robustness. Connections to the *ScienceQA* dataset are highlighted. See section 7 for details.

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 more accurate reflection of their capabilities on unseen data.

Overall, while diverse training data and sophisticated architectures contribute to model performance, the inclusion of evaluation datasets in training can artificially inflate results. It is crucial to consider potential data contamination when interpreting performance metrics to ensure fair and accurate assessments of model capabilities.

5.3 Defending against distractions

Table 4: Exact match scores achieved by the models using a naive prompt without defenses compared to a prompt with instructions to ignore distractions.

Model	Add Image (%)		Insert Image (%)		Add Hints (%)		Insert Hint (%)	
	No Defense	Defense	No Defense	Defense	No Defense	Defense	No Defense	Defense
Qwen2-VL-Instruct (2B)	63.30	73.80	63.26	65.60	54.45	62.60	64.20	70.35
Qwen2-VL-Instruct (7B)	83.10	81.35	68.08	68.60	68.00	68.05	74.10	74.90
CogVLM2 (19B)	71.70	70.15	87.47	85.18	70.50	70.20	80.85	79.10

The findings in Table 4 demonstrate that although prompt engineering techniques—such as adding instructions to guide the model's focus toward the question and away from distractions—can partially mitigate the effects of distractions, models still struggle to ignore them. For instance, in the **Add Image** scenario, the performance of *Qwen2-VL-Instruct (2B)* improves from 63.30 to 73.80 when defense mechanisms are applied, indicating that appropriate prompts can enhance the model's focus on relevant information. Similarly, in the **Insert Hint** scenario, the same model's performance increases from 64.20 to 70.35 with defense strategies.

However, the improvements are not uniform across all models and tasks. The *Qwen2-VL-Instruct (7B)* model shows a slight decrease in performance in the **Add Image** scenario when defenses are applied, dropping from 83.10 to 81.35. This suggests that the effectiveness of defense mechanisms may vary depending on the model's architecture and size. Moreover, the *CogVLM2 (19B)* model exhibits a minor reduction in performance across most tasks with defense prompts, indicating that larger models are not necessarily better at ignoring distractions when prompted to do so. For example, in the **Insert Image** scenario, its performance decreases from 87.47 to 85.18 even with defense strategies.

Table 5: Exact Math Scores achieved by LLava Models with different vision encoders.

Model (Vision Encoder)	Add Image (%)		Insert Image (%)		Add Hints (%)		Insert Hint (%)	
	Original	Distraction	Original	Distraction	Original	Distraction	Original	Distraction
LLava-7B (Robust-clip)	N/A	70.40	67.30	63.55	71.78	67.48	64.31	63.25
LLava-7B (Clip)	N/A	69.55	68.95	64.32	76.22	72.49	64.39	63.25

Table 5 summarizes the performance of LLava-7B models with two different vision encoders: *robust-clip* (Schlarmann et al., 2024) and *clip* (Ilharco et al., 2021). The models are evaluated across original and distraction scenarios, focusing on the effects of adding or inserting images and hints. Since the LLava-7B model with *robust-clip* can only process inputs that include both text and visual content, samples without images were excluded from this evaluation. The *robust-clip* encoder only outperforms the *clip* encoder slightly in the **Add Image** scenario by 0.85. In other scenario, the performance of the *robust-clip* encoder is lower than that of the *clip* encoder. These findings suggest that *robust-clip* shows very limited efficacy in defending against visual distractions.

These results suggest potential areas for future improvements in model training and design. Developing more effective prompting techniques and enhancing model architectures could help models better filter out irrelevant information. Additionally, incorporating training data that specifically addresses the handling of distractions may improve models' robustness in real-world applications where irrelevant data is commonplace.

5.4 Bi-Modal Distraction

Table 6: Exact Match Score under textual distraction, both with and without simultaneous visual distraction.

Model	Add Hint (%)		Insert Hint (%)	
	No Image	With Image	No Image	With Image
Qwen2-VL-2B-Instruct	57.68	57.68	72.86	72.86
Qwen2-VL-7B-Instruct	71.43	71.43	88.57	88.57
CogVLM2-LLaMA3-Chat-19B	57.45	56.91	78.31	79.35

The results in Table 6 examine the models' performances under conditions where textual distractions are present, with and without the simultaneous presence of visual distractions. Specifically, the "No Image" columns represent scenarios with only textual distractions, while the "With Image" columns include both textual and visual distractions.

Analyzing the data, we observe that the performance of *Qwen2-VL-Instruct (2B)* and *Qwen2-VL-Instruct (7B)* remains unchanged between the "No Image" and "With Image" conditions across both "Add Hint" and "Insert Hint" scenarios. This suggests that the addition of visual distractions does not significantly impact these models when textual distractions are already present. In contrast, the *CogVLM2 (19B)* model shows a slight decrease in performance from 57.45% to 56.91% in the "Add Hint" scenario when an image is added, indicating a minor negative effect of visual distractions in conjunction with textual ones. Interestingly, in the "Insert Hint" scenario, its performance slightly improves from 78.31% to 79.35% with the addition of an image, suggesting that under certain conditions, visual distraction might compete with textual distraction.

Overall, these findings imply that the models' abilities to handle bi-modal distractions are nuanced. While some models maintain consistent performance regardless of the presence of additional visual information, others exhibit minor fluctuations. This highlights the importance of designing models that can effectively integrate and prioritize multi-modal inputs, ensuring robustness in environments where distractions are prevalent across different modalities.

6 Related Work

Model Evaluations and Data Contamination. Visual Language Models (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 & Manning, 2019), which focus on foundational VQA questions. More recently, studies like MM-Vet (Yu et al., 2023b), POPE (Li et al., 2023a), and MM-Bench (Liu et al., 2023d) have emerged to specifically evaluate VLMs, addressing key challenges like hallucination, reasoning, and robustness. These efforts

have demonstrated that multimodal LLMs encounter significant issues, such as hallucination (Guan et al., 2023) and insufficient robustness (Fu et al., 2023). In this papr, 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.

Data contamination has also become a major focus in recent work (Lovin, 2023; Bender et al., 2021; Kocoń et al., 2023; Li, 2023). Several researchers (Golchin & Surdeanu, 2023; Oren et al., 2023; Yang et al., 2023; Oren et al., 2023) have developed techniques to detect and mitigate data contamination. Furthermore, dynamic evaluation techniques have been proposed (Zhu et al., 2023; Lei et al., 2023; Fan et al., 2023), leveraging various algorithms to reduce the adverse effects of data contamination on model performance.

Benchmarks with Input Perturbations. The use of input perturbations has been a common strategy in natural language tasks, with approaches ranging from model-agnostic input transformations (Liang et al., 2022; Ravichander et al., 2022) to adversarial example generation targeting specific models (Jia & Liang, 2017; Shi et al., 2018; Morris et al., 2020; Wang et al., 2021). Notably, prior research has constructed arithmetic reasoning benchmarks by paraphrasing or rewriting sentences from clean datasets (Patel et al., 2021; Kumar et al., 2021; Shi et al., 2023).

Robustness of VLM Recent studies have increasingly concentrated on the adversarial vulnerabilities of VLMs Qi et al. (2024); Carlini et al. (2024); Schlarmann & Hein (2023); Zhao et al. (2023b); Dong et al. (2023). Schlarmann & Hein (2023) demonstrate that imperceptible perturbations in input images can enable attackers to manipulate LVLMs into generating specific outputs. Additionally, visual adversarial attacks designed to jailbreak LVLMs are introduced in works such as Carlini et al. (2024) and Qi et al. (2024). More recently, numerous studies have focused on training adversary robust vision encoder for VLMs Schlarmann et al. (2024); Mao et al. (2023).

7 Limitations and Conclusion

In this paper, we introduced *I-ScienceQA*, a comprehensive benchmark designed to assess the robustness of Vision-Language Models against distractions in both visual and textual domains. By augmenting the *ScienceQA* dataset with diverse forms of distractions, we simulated real-world conditions where input data is often imperfect, noisy. Our extensive evaluation across state-of-the-art VLMs revealed several key findings: (1) Most VLMs remain vulnerable to distractions, especially in the textual domain; (2) Larger models tend to be more robust but do not always guarantee improved performance, particularly when faced with complex bi-modal distractions; (3) Prompt engineering and robust vision encoder could only partially mitigate these vulnerabilities, there remains significant room for improvement in handling both textual and visual distractions.

Our findings highlight the need for further research in developing more robust VLM models. As the use of VLMs expands across domains such as healthcare, education, and autonomous systems, it becomes increasingly important to build models that can handle the noisy data often encountered in real-world applications.

While our work contributes valuable insights into the challenges of distraction robustness, it also has certain limitations:

- Limited scope of distractions: Although we introduced a variety of textual and visual distractions, the dataset does not encompass all possible real-world noise. Future work could explore additional forms of noise, such as adversarial examples, corrupted images to further challenge the models.
- Model evaluation focus: Our study primarily focused on pre-trained VLMs. We did not explore the effects of fine-tuning models on noisy datasets that may be more resilient to distractions. Fine-tuning on noisy or augmented data could provide valuable insights into improving model robustness.
- Bimodal distractions: While we examined the compounded effects of bimodal distractions, we did not extensively explore how interaction effects between the two modalities influence model performance. Future research should analyze more closely how different types of visual and textual distractions interact and whether certain combinations are more detrimental than others.
- Defense techniques: Although we explored the use of prompt engineering and robust vision encoder as a defense mechanism, our study did not delve into other possible methods to enhance model robustness, such as vision segmentationLai et al. (2023). Exploring these techniques could offer more comprehensive solutions for improving VLM performance in noisy environments.

In summary, while the *I-ScienceQA* benchmark provides a valuable tool for evaluating VLM robustness, there is much work to be done in advancing models that can consistently navigate noisy, real-world data. Future research should focus on expanding the range of distractions, investigating fine-tuning techniques, and exploring other defense strategies to create more resilient VLMs.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proc. of Association for the Advancement of Artificial Intelligence*, 2019.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, pp. 8948–8957, 2019.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proc. of International Conference on Computer Vision*, pp. 2425–2433, 2015.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*, 2024.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? FAccT 2021, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pp. 4291–4301, 2019.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, pp. 1511–1520, 2022.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned?, 2024. URL `https://arxiv.org/abs/2306.15447`.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024a.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning, 2022. URL `https://arxiv.org/abs/2105.14517`.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024c.

Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *ICDAR*, pp. 1571–1576, 2019.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, pp. 326–335, 2017.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How Robust is Google's Bard to Adversarial Image Attacks? *arXiv preprint arXiv:2309.11751*, 2023.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, Yongfeng Zhang, and Libby Hemphill. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Shao Zhenwei Gai Zhenbiao. Phi3v-finetuning, 2023. URL https://github.com/GaiZhenbiao/Phi3V-Finetuning. GitHub repository.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. *arXiv preprint arXiv:2311.06233*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6904–6913, 2017.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *NeurIPS*, 35: 26418–26431, 2022.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *arXiv preprint arXiv:2308.10755*, 2023.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogvlm2: Visual language models for image and video understanding, 2024. URL https://arxiv.org/abs/2408.16500.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*, 2023.

Drew A Hudson and Christopher D Manning. 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)*, pp. 6700–6709, 2019.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

Jimmycarter. Textocr gpt-4v dataset. https://huggingface.co/datasets/jimmycarter/textocr-gpt4v, 2023.

Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pp. 1965–1973, 2017.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5648–5656, 2018.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning, 2018.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pp. 235–251, 2016.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, pp. 4999–5007, 2017.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *ECCV*, 2022.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. Chatgpt: Jack of all trades, master of none. *Information Fusion*, pp. 101861, 2023.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017.

Vivek Kumar, Rishabh Maheshwary, and Vikram Pudi. Adversarial examples for evaluating math word problem solvers. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2705–2712, 2021.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

LAION. Gpt-4v dataset. `https://huggingface.co/datasets/laion/gpt4v-dataset`, 2023.

Fangyu Lei, Qian Liu, Yiming Huang, Shizhu He, Jun Zhao, and Kang Liu. S3eval: A synthetic, scalable, systematic evaluation suite for large language models. *arXiv preprint arXiv:2310.15147*, 2023.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *SIGIR*, pp. 3108–3120, 2022.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900, 2022.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models, 2024. URL `https://arxiv.org/abs/2403.00231`.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pp. 292–305, 2023a.

Yucheng Li. An open source data contamination report for llama series models. *arXiv preprint arXiv:2310.17589*, 2023.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *CVPR*, pp. 14963–14973, 2023b.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *TACL*, 11:635–651, 2023a.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023d.

Brian Lovin. Gpt-4 performs significantly worse on coding problems not in its training data. `https://brianlovin.com/hn/35297067`, 2023.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021a.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35:2507–2521, 2022a.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL `https://arxiv.org/abs/2310.02255`.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=P4bXCawRi5J`.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pp. 11–20, 2016.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pp. 3195–3204, 2019.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pp. 1697–1706, 2022.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pp. 1527–1536, 2020.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pp. 947–952, 2019.

John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*, 2020.

OpenAI. Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`, 2023.

OpenAI. Gpt-3.5 turbo. https://platform.openai.com/docs/models/gpt-3-5-turbo, 2024. Accessed: yyyy-mm-dd.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*, 2023.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? In *NAACL-HLT*, 2021. URL https://aclanthology.org/2021.naacl-main.168.pdf.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. of International Conference on Computer Vision*, pp. 2641–2649, 2015.

ProMeAI. Homepage. https://www.promeai.pro/, 2024. Accessed: yyyy-mm-dd.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21527–21536, 2024.

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. Condaqa: A contrastive reading comprehension dataset for reasoning about negation. *arXiv preprint arXiv:2211.00295*, 2022.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2021. URL https://api.semanticscholar.org/CorpusID:245335280.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models, 2023.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *ICML*, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022a.

Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion coco: 600m synthetic captions from laion2b-en. *https://laion.ai/blog/laion-coco/*, 2022b.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pp. 146–162, 2022.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, volume 33, pp. 8876–8884, 2019.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pp. 8430–8439, 2019.

Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *ICDAR*, volume 1, pp. 1429–1434, 2017.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context, 2023.

Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. Learning visually-grounded semantics from contrastive adversarial samples. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3715–3727, 2018.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. On the exploitability of instruction tuning, 2023. URL https://arxiv.org/abs/2306.17194.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pp. 742–758, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8317–8326, 2019.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *CVPR*, pp. 8802–8812, 2021.

Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *ICDAR*, pp. 1557–1562, 2019.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants. `https://huggingface.co/datasets/teknium/OpenHermes-2.5`, 2023.

Unsplash. Unsplash api. `https://unsplash.com/developers`, 2024. Accessed: yyyy-mm-dd.

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL `https://arxiv.org/abs/2409.12191`.

Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *ICLR*, 2024b.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017a.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1645–1653, 2017b.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pp. 1686–1697, 2021.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pp. 69–85, 2016.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023a.

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

Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34:509–521, 2019.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Proc. of Neural Information Processing Systems*, 35:36067–36080, 2022.

Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *ICDAR*, pp. 1577–1581, 2019.

Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023a.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models, 2023b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36, 2024.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oZDJKTlOUe.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*, 2023.

APPENDIX

- Statistics of distractions
- Definition of distractions
- More Results
- Models' training dataset

A STATISTICS OF DISTRACTIONS

Scenario	Types of Distraction and Content	Number
Add Image	Neutral Backgrounds	250
	Generic Landscapes	250
	Abstract Art	250
	Everyday Objects	250
	Cultural Artifacts	250
	Digital Creations	250
	Word Embeddings	250
	Emotional Contexts	250
Insert Image	Diffusion Inpainting	100
	Neutral Backgrounds	250
	Generic Landscapes	250
	Abstract Art	250
	Everyday Objects	250
	Cultural Artifacts	250
	Digital Creations	250
	Word Embeddings	250
	Emotional Contexts	250
Add Hints	Irrelevant Context Integration	400
	Contradictory Information	400
	Non Sequitur	400
	Misleading Information	400
	Ambiguous Information	400
Insert Hints	Subtle Misinformation	400
	Irrelevant Details	400
	Disruptive Narrative Inserts	400
	Complex Referential Distractions	400
	Ambiguous Information	400

Table 7: Distribution of Distraction Scenarios Across Scenarios

B DEFINITION FOR DISTRACTIONS

Type of Image	Description
Neutral Backgrounds	Simple, monochromatic backgrounds to minimize distraction and control variable introduction.
Generic Landscapes	Broad, non-specific landscapes (e.g., forests, urban scenes, mountains) that provide a realistic yet contextually neutral backdrop.
Abstract Art	Non-representational art that challenges the model to focus on textual rather than visual cues.
Everyday Objects	Common, non-contextual objects to evaluate the model's ability to disregard irrelevant visual stimuli.
Cultural Artifacts	Images of artifacts that test cultural recognition and contextual integration capabilities.
Digital Creations	Computer-generated or altered imagery to assess the model's response to unconventional visual data.
Word Embeddings	Images with embedded or overlaid text to examine effective textual and visual information merging.
Emotional Contexts	Images depicting clear emotional tones to probe model's alignment of text and visual emotion cues.

Table 8: Types and definition of Distractions for Scenario I **Add Image** and Scenario II **Insert Image**.

Type of Image	Description
Flying Objects	Introduces elements like birds, planes, or insects, requiring the model to differentiate between essential static elements and moving distractions.
Floating Balloons	Adds balloons of various colors and sizes that float across the scene, testing the VLM's ability to ignore appealing but irrelevant moving objects.
Passing Vehicles	Populates scenes with moving vehicles like cars and bicycles, challenging the model to disregard transient elements.
Drifting Clouds	Simulates clouds moving across the sky, testing the model's focus amid ongoing environmental changes.
Bouncing Balls	Uses images of balls in motion to introduce unexpected kinetic elements, assessing the model's response to sudden movements.
Swarming Insects	Adds complexity with swarming insects like butterflies or bees to test the VLM's fine-grained visual attention.
Animated Signs	Integrates changing digital signs to evaluate the model's ability to ignore intermittent visual stimuli.
Symbols and Icons	Embeds non-contextual symbols or icons, assessing the model's disregard for extraneous visual information.
Overlaid Words	Overlays random words or phrases to introduce visual clutter, testing prioritization of primary textual content.
Sitting Pets	Includes images of sitting pets to test the VLM's focus amidst visually appealing but irrelevant elements.

Table 9: Types and definition of Distractions for Scenario II: **Insert Image**.

Distraction Type	Description
Irrelevant Context Integration	Introduce sentences with contextually irrelevant information to assess the model's capacity to filter out noise. Based on studies showing extraneous data can reduce comprehension accuracy, mirroring real-world information processing challenges.
Contradictory Information	Embed contradictions within the narrative to test the models' conflict resolution and logic adherence capabilities.
Non Sequitur	Use complex sentence structures and ambiguous phrases to evaluate the models' parsing and interpretation flexibility.
Misleading Information	Include plausible but incorrect data points within texts, testing the models' fact-checking abilities and resilience against misinformation.
Ambiguous Information	Incorporate vague or unclear statements to assess the model's ability to handle uncertainty and make reasonable inferences.

Table 10: Types and definition of Distractions for Scenario III: **Add Hint**.

Distraction Type	Description
Subtle Misinformation	Inserts information that subtly misleads, contradicting established data yet remaining plausible within the context.
Irrelevant Details	Introduces extraneous details that do not contribute to the task, challenging the model to maintain focus on relevant content.
Disruptive Narrative Inserts	Incorporates unrelated narratives that interrupt logical progression, testing the model's information filtering capabilities.
Complex Referential Distractions	Utilizes intricate language and referential sequences that complicate the parsing process, assessing interpretative accuracy.
Ambiguous or Conflicting Information	Presents choices that include both plausible and incorrect answers, exploiting potential ambiguities to test decision-making precision.

Table 11: Types and definition of Distractions for Scenario IV: **Insert Hint**.

C MORE RESULTS

C.1 ADD IMAGE

Table 12: Exact Match Scores of Various Models Across Different Types of Distractions in the **Add Image** Scenario. Abbreviations: AA = Abstract Art, CA = Cultural Artifacts, WE = Word Embeddings, DC = Digital Creations, NB = Neutral Backgrounds, GL = Generic Landscapes, EC = Emotional Contexts, EO = Everyday Objects.

Model	AA	CA	WE	DC	NB	GL	EC	EO
InstructBLIP (7B)	38.8	37.6	44.4	42.4	37.6	42.4	40.0	45.2
InstructBLIP (13B)	70.4	72.0	73.6	75.2	73.2	70.4	67.6	73.6
Llava (7B)	66.4	73.2	71.6	70.8	64.4	69.6	62.8	65.6
Llava (13B)	70.4	72.0	73.6	75.2	73.2	70.4	67.6	73.6
Internvl2 (2B)	83.6	84.4	85.6	87.2	92.4	88.0	86.8	86.0
Internvl2 (8B)	94.8	96.0	93.6	95.6	93.6	93.6	92.0	96.4
Qwen/Qwen2-VL-2B-Instruct	59.6	63.6	60.0	64.0	67.6	63.2	65.2	63.2
Qwen/Qwen2-VL-7B-Instruct	84.8	82.0	83.2	85.6	81.6	84.4	83.2	80.0
THUDM/cogVLM2-LLaMA3-Chat-19B	69.2	72.8	72.0	70.8	73.2	74.4	69.2	72.4
GPT-4o	94.0	94.0	90.8	92.0	91.2	94.0	94.8	93.2

C.2 INSERT IMAGE

Table 13: Exact Match Scores of Various Models Across Different Types of Distractions in the **Insert Image** Scenario. Abbreviations: AA = Abstract Art, CA = Cultural Artifacts, WE = Word Embeddings, DC = Digital Creations, NB = Neutral Backgrounds, GL = Generic Landscapes, EC = Emotional Contexts, EO = Everyday Objects, DI = Diffusion Inpainting.

Model	AA	CA	WE	DC	NB	GL	EC	EO	DI
InstructBLIP (7B)	32.40	39.20	34.00	34.40	37.60	36.80	36.00	30.40	42.42
InstructBLIP (13B)	68.80	72.00	69.60	70.80	73.20	68.40	66.80	66.00	72.73
Llava (7B)	63.60	70.80	67.60	67.20	68.80	68.40	60.40	64.80	64.64
Llava (13B)	68.80	72.00	69.60	70.80	73.20	68.40	66.80	66.00	72.72
Internvl2 (2B)	89.20	87.20	91.20	91.20	91.20	92.00	90.80	90.00	87.87
Internvl2 (8B)	94.00	92.40	93.60	94.80	95.60	96.00	95.20	93.60	90.90
Qwen/Qwen2-VL-2B-Instruct	62.40	64.80	64.40	68.00	64.40	67.20	60.40	58.80	52.53
Qwen/Qwen2-VL-7B-Instruct	64.40	70.80	68.00	71.20	72.00	70.40	61.60	68.80	61.62
THUDM/cogVLM2-LLaMA3-Chat-19B	89.20	86.80	86.00	87.60	87.20	89.20	86.80	88.00	84.85
GPT-4o	79.60	79.60	78.00	82.00	78.80	78.00	77.20	76.40	75.75

C.3 ADD HINT

Table 14: Exact Match Scores of Various Models Across Different Types of Distractions in the **Add Hint** Scenario. Abbreviations: Contradictory = Contradictory Hints, Non Sequitur = Non Sequitur Hints, Ambiguous = Ambiguous Hints, Irrelevant = Irrelevant Hints, Misleading = Misleading Hints.

Model	Contradictory	Non Sequitur	Ambiguous	Irrelevant	Misleading
InstructBLIP (7B)	62.75	65.00	66.50	58.75	66.00
InstructBLIP (13B)	67.50	69.25	68.00	63.75	68.75
Llava (7B)	63.60	70.80	67.60	67.20	68.80
Llava (13B)	68.80	72.00	69.60	70.80	73.20
Internvl2 (2B)	81.50	77.00	87.00	85.75	80.50
Internvl2 (8B)	95.00	92.00	94.50	92.25	94.25
Qwen/Qwen2-VL-2B-Instruct	53.00	55.00	59.25	52.00	53.00
Qwen/Qwen2-VL-7B-Instruct	67.75	68.25	71.75	67.50	64.75
THUDM/cogVLM2-LLaMA3-Chat-19B	74.50	68.25	73.75	68.00	68.00
GPT-4o	90.00	87.50	87.50	84.75	87.75

Table 15: Exact Match Scores of Various Models Across Different Types of Distractions in the **Insert Hint** Scenario. Abbreviations: Subtle Misinformation = Influence of Subtle Misinformation Hints, Irrelevant Details = Influence of Irrelevant Details Hints, Disruptive Narrative = Influence of Disruptive Narrative Hints, Complex Referential = Influence of Complex Referential Hints, Ambiguous or Conflicting = Influence of Ambiguous or Conflicting Hints.

Model	Subtle Misinformation	Irrelevant Details	Disruptive Narrative	Complex Referential	Ambiguous or Conflicting
Llava (7B)	70.75	69.50	67.50	69.75	69.00
Llava (13B)	74.00	73.50	70.00	70.50	71.00
Internvl2 (2B)	90.25	92.25	90.50	93.75	90.75
Internvl2 (8B)	95.00	95.75	94.25	97.25	97.25
Qwen/Qwen2-VL-2B-Instruct	65.75	65.00	59.75	66.00	64.50
Qwen/Qwen2-VL-7B-Instruct	76.00	71.75	73.25	74.25	75.25
THUDM/cogVLM2-LLaMA3-Chat-19B	78.75	79.25	81.75	83.50	81.00
GPT-4o	83.75	84.75	83.00	82.25	86.50

D MODELS' TRAINING DATASET

Model Name	Training Data
liu2023llava1.5	**Pre-training data:** Conceptual Captions (CC) (Changpinyo et al., 2021); COCO Captions (Lin et al., 2014); ScienceQA (Lu et al., 2022a); LLaVA-Instruct-158K (Liu et al., 2023c); utilizes a CLIP visual encoder pre-trained on LAION-2B (Schuhmann et al., 2022a). **Fine-tuning data:** COCO (Lin et al., 2014); ScienceQA (Lu et al., 2022a); LLaVA-Instruct-158K (Liu et al., 2023c), tailored for instruction-following tasks.
InstructBLIP-Vicuna (7B, 13B)	**Pre-training data:** NoCaps (Agrawal et al., 2019); Flickr30K (Plummer et al., 2015); VizWiz (Gurari et al., 2018); GQA (Hudson & Manning, 2019); Visual Spatial Reasoning (Liu et al., 2023a); IconQA (Lu et al., 2021b); ScienceQA (Lu et al., 2022a); Visual Dialog (Das et al., 2017); TextVQA (Singh et al., 2019); HatefulMemes (Kiela et al., 2020); MSVD-QA (Xu et al., 2017b); MSRVTT-QA (Xu et al., 2017a); iVQA (Yang et al., 2021). **Fine-tuning data:** COCO Caption (Lin et al., 2014); Web CapFilt (Li et al., 2022;?); TextCaps (Sidorov et al., 2020); VQAv2 (Goyal et al., 2017); OKVQA (Marino et al., 2019); A-OKVQA (Schwenk et al., 2022); OCR-VQA (Mishra et al., 2019); LLaVA-Instruct-150K (Liu et al., 2023c), transformed into instruction-answer pairs for enhanced multimodal instruction following.
CogVLM2-LLaMA3-Chat-19B	**Pre-training data:** LAION-2B (Schuhmann et al., 2022a); COYO-700M (Byeon et al., 2022); LAION-40M-grounding (Zhang et al., 2022); multilingual image-text pairs from LAION, COCO (Lin et al., 2014), and Visual Genome (Krishna et al., 2017). **Fine-tuning data:** OKVQA (Marino et al., 2019); STVQA (Biten et al., 2019); visualgenome (Krishna et al., 2017); VQAv2 (Goyal et al., 2017); DocVQA (Mathew et al., 2021); OCRVQA (Mishra et al., 2019); TextVQA (Singh et al., 2019); GeoMetry3K (Lu et al., 2021a); Geo170K (Gao et al., 2023); GeoQA (Chen et al., 2022); ScienceQA (Lu et al., 2022a); ChartQA (Masry et al., 2022); FigureVQA (Kahou et al., 2018); InfoVQA (Mathew et al., 2022); DVQA (Kafle et al., 2018); ArxivQA (Li et al., 2024); TDIUC (Kafle & Kanan, 2017); TallyQA (Acharya et al., 2019), optimized for multimodal understanding and conversational abilities.
Qwen2-VL-7B-Instruct	**Pre-training data:** Details not publicly disclosed. **Fine-tuning data:** Details not publicly disclosed.
Phi3-V	**Pre-training data:** Details not publicly disclosed. **Fine-tuning data:** Details not publicly disclosed.

Model Name	Training Data
InternVL2 (2B, 8B)	**Pre-training data:** Laion-EN (Schuhmann et al., 2022a); Laion-ZH (zh) (Schuhmann et al., 2022a); COYO (zh) (Byeon et al., 2022); GRIT (zh) (Peng et al., 2023); COCO (Lin et al., 2014); TextCaps (Sidorov et al., 2020); Objects365 (en&zh) (Shao et al., 2019); GRIT (en&zh) (Peng et al., 2023); All-Seeing (en&zh) (Wang et al., 2024b); Wukong-OCR (zh) (Gu et al., 2022); LaionCOCO-OCR (Schuhmann et al., 2022b); MMC-Inst (Liu et al., 2023b); LSVT (zh) (Sun et al., 2019); ST-VQA (Biten et al., 2019); RCTW-17 (zh) (Shi et al., 2017); ReCTs (zh) (Zhang et al., 2019); ArT (en&zh) (Chng et al., 2019); SynthDoG (en&zh) (Kim et al., 2022); COCO-Text (Veit et al., 2016); ChartQA (Masry et al., 2022); CTW (zh) (Yuan et al., 2019); DocVQA (Mathew et al., 2021); TextOCR (Singh et al., 2021); PlotQA (Methani et al., 2020); InfoVQA (Mathew et al., 2022). **Fine-tuning data:** TextCaps (Sidorov et al., 2020); ShareGPT4V (en&zh) (Chen et al., 2023); VQAv2 (Goyal et al., 2017); GQA (Hudson & Manning, 2019); OKVQA (Marino et al., 2019); VSR (Liu et al., 2023a); Visual Dialog (Das et al., 2017); AI2D (Kembhavi et al., 2016); ScienceQA (Lu et al., 2022a); TQA (Kembhavi et al., 2017); ChartQA (Masry et al., 2022); MMC-Inst (Liu et al., 2023b); DVQA (Kafle et al., 2018); PlotQA (Methani et al., 2020); GeoQA+ (Cao & Xiao, 2022); TabMWP (Lu et al., 2022b); MathQA (Yu et al., 2023a); CLEVR-Math/Super (Lindström & Abraham, 2022; Li et al., 2023b); Geometry3K (Lu et al., 2021a); KVQA (Shah et al., 2019); A-OKVQA (Schwenk et al., 2022); ViQuAE (Lerner et al., 2022); Wikipedia (en&zh) (He et al., 2023); OCRVQA (Mishra et al., 2019); TextVQA (Singh et al., 2019); RefCOCO/+/g (Yu et al., 2016; Mao et al., 2016); Visual Genome (Krishna et al., 2017); LLaVA-1580K (en&zh) (Liu et al., 2023c); LVIS-Instruct4V (Wang et al., 2023); ALLaVA (en&zh) (Chen et al., 2024a); Laion-GPT4V (LAION, 2023); TextOCR-GPT4V (Jimmycarter, 2023); SVIT (en&zh) (Zhao et al., 2023a); OpenHermes2.5 (Teknium, 2023); Alpaca-GPT4 (Taori et al., 2023); ShareGPT (en&zh) (Zheng et al., 2024); COIG-CQIA (zh) (Bai et al., 2024); optimized for diverse visual-language tasks including visual question answering, image captioning, and visual dialogue.
GPT-4o	**Pre-training data:** Details not publicly disclosed. **Fine-tuning data:** Details not publicly disclosed.

Table 16: Models and their Pre-training and Fine-tuning Data