Illusory VQA: Benchmarking and Enhancing Multimodal Models on Visual Illusions

Mohammadmostafa Rostamkhani
mo_rostamkhani97@comp .iust.ac.ir

Baktash Ansari
baktash_ansari@comp.iust.ac.ir

Hoorieh Sabzevari
h_sabzevari@elec.iust.ac.ir

Farzan Rahmani
farzan_rahmani@comp.iust.ac.ir

Sauleh Eetemadi
Department of Computer Engineering
IUST
sauleh@iust.ac.ir

Abstract

In recent years, Visual Question Answering (VQA) has made significant strides, particularly with the advent of multimodal models that integrate vision and language understanding. However, existing VQA datasets often overlook the complexities introduced by image illusions, which pose unique challenges for both human perception and model interpretation. In this study, we introduce a novel task called Illusory VQA, along with four specialized datasets: IllusionMNIST, IllusionFashionMNIST, IllusionAnimals, and IllusionChar. These datasets are designed to evaluate the performance of state-of-the-art multimodal models in recognizing and interpreting visual illusions. We assess the zero-shot performance of various models, fine-tune selected models on our datasets, and propose a simple yet effective solution for illusion detection using Gaussian and blur low-pass filters. We show that this method increases the performance of models significantly and in the case of BLIP-2 on IllusionAnimals without any fine-tuning, it outperforms humans. Our findings highlight the disparity between human and model perception of illusions and demonstrate that fine-tuning and specific preprocessing techniques can significantly enhance model robustness. This work contributes to the development of more human-like visual understanding in multimodal models and suggests future directions for adapting filters using learnable parameters.

1. Introduction

In recent years, there has been significant progress in the field of Visual Question Answering (VQA) and multimodal models, which involves answering questions about images using both vision and language understanding. VQA requires models to interpret visual content and comprehend natural language in order to provide accurate answers. Most existing VQA datasets focus on traditional image understanding and do not consider the challenges posed by novel image illusions.

(a) Raw image without illusion (b) Image illusion

Figure 1. Illustration of illusion in image

Image illusion refers to the phenomenon where an image initially portrays one thing, yet upon closer examination and a slight adjustment of focus, it reveals an entirely different representation. An example of this can be found in Figure 1. The phenomenon we are discussing is called ”pareidolia,” a psychological effect where people perceive familiar patterns or images (such as faces, animals, or objects) in random or unrelated stimuli. In this instance, you initially see one image, but upon closer inspection, your brain detects another image hidden within the scene.

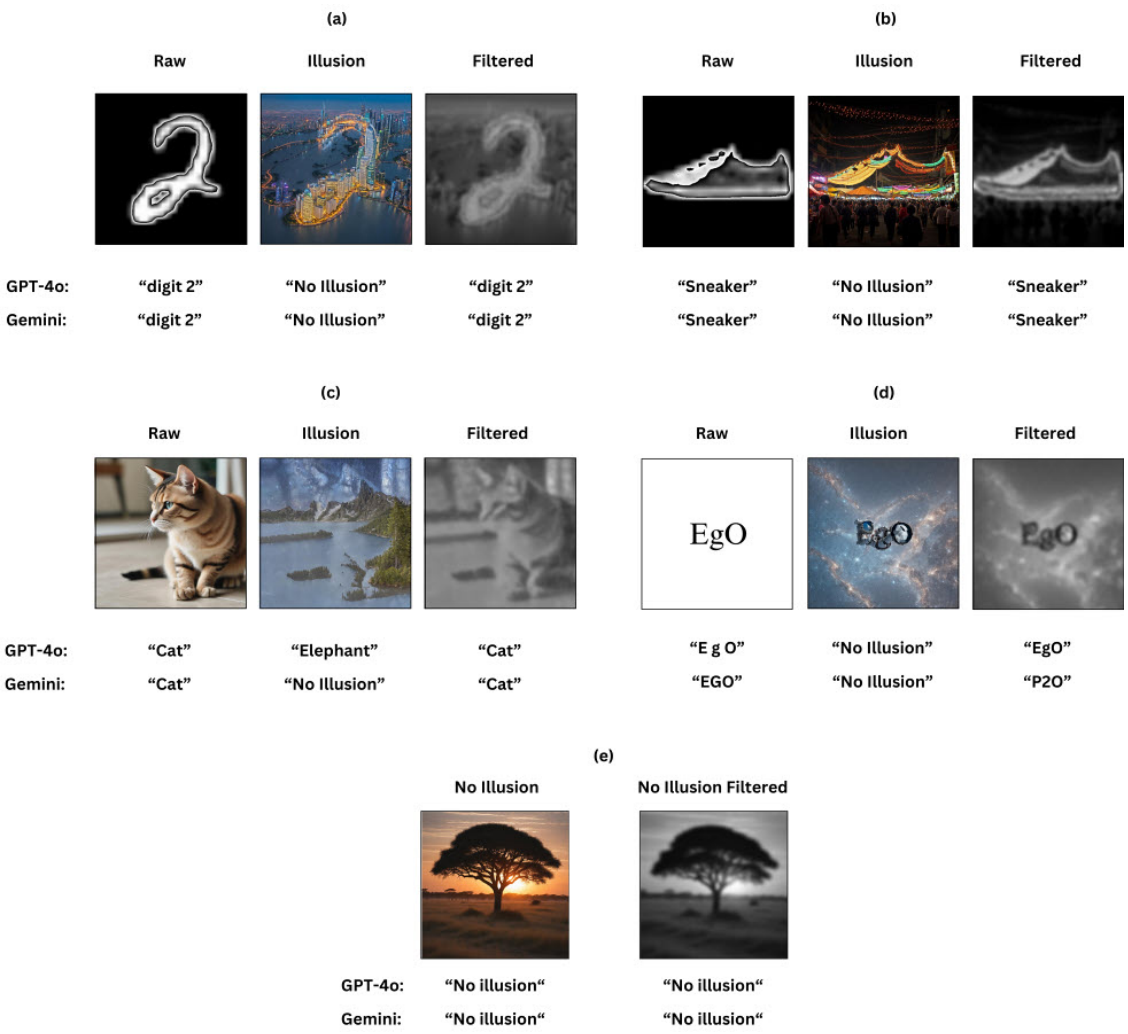

Figure 2. Examples of answers of GPT-4o and Gemini to raw images, illusory images, and applied filter one. Through the process of image minification, the human eye becomes adept at uncovering the hidden illusions within. (a): An example of IllusionMNIST, (b): An example of IllusionFashionMNIST, (c): An example of IllusionAnimals, (d): An example of IllusionChar, (e): An example of 'No illusion' class

Our brains are naturally wired to recognize patterns, which explains why we often identify familiar shapes in clouds, rock formations, or in images like the one presented here. This is also referred to as a "visual illusion" or "perceptual illusion." In this case, the image consists of an arrangement of shapes and colors that your brain interprets in multiple ways—first as a landscape, and then, upon shifting your perception, as the shape of a rabbit. This type of illusion occurs because the brain attempts to make sense of ambiguous or complex visuals by aligning them with familiar patterns. Although the image itself remains unchanged, your perception shifts, giving the impression of seeing something not intentionally designed into the image. Essentially, illusions like this illustrate how our brains actively interpret and construct reality based on visual input. In this work, we use the terms "pareidolia" and "illusion" interchangeably.

The inclusion of illusion images in the dataset is important because it opens up possibilities for applications in steganography and potentially bypassing ethical rules. These types of images for multimodal models act as brain-teasers for language models. While the image may contain certain elements, one needs to view it from a different perspective to perceive the illusion within it. It is widely recognized that the human perceptual systems are prone to visual illusions, which can be described as "consistent and persistent discrepancies between a physical state of affairs and its representation in consciousness" [4]. In recent years, there has been a growing interest in model-perceived visual illusions, inspired by the fascinating phenomena observed in human perception [7, 11, 21, 27]. These previous studies have primarily focused on vision, exploring how computer

vision models can replicate the effects of illusions by examining the internal representations and comparing them to the perceptual shifts experienced by humans. In spite of the emergence of multimodal models such as CLIP [25], BLIP [16], BLIP-2 [17], Kosmos-2 [24], LLaVA [20], MiniGPT-V2 [3, 35], Gemini [28], and GPT-4V(ision) [32], **a significant disparity remains between the way humans perceive and interpret images and how these models perceive them**. We can use these illusory datasets to make VLMs behave more like humans. We suggest that this approach can also enhance their robustness in OCR capabilities.

In this study, our main contributions are as follows.

1. Introduction of new task called Illusory VQA
2. Introduction of IllusionMNIST, IllusionFashionMNIST, IllusionAnimals, and IllusionChar Datasets: We develop these datasets specifically for benchmarking multimodal models. These datasets include separate train and test sets, providing a comprehensive evaluation framework. Some examples from these datasets can be found in Figure 2.
3. Evaluation of Zero-Shot Performance: We assess the zero-shot performance of state-of-the-art multimodal models on the aforementioned datasets. This evaluation helps us understand how well these models can generalize and interpret the illusions present in the images.
4. Fine-Tuning and Re-evaluation: We further enhance the evaluation by fine-tuning some of the multimodal models on the training set and then re-evaluating their performance on the test set. This process allows us to measure the effectiveness of fine-tuning in improving the models' ability to detect illusions.
5. Proposal of an Effective Solution for Illusion Detection: Additionally, we propose a simple yet effective solution for detecting illusions in images. This solution aims to enhance the models' capability to identify and differentiate between real and illusory elements in the datasets. Our approach involves the application of a fixed Gaussian and blur filter to illusory images. You can find the details of the filter we used in the Supplementary Material 18.

2. Related Work

[5] explores illusory contour perception in deep learning models by creating illusory contour datasets through image distortion techniques based on the abutting grating illusion. It also aims to systematically generate illusory contour samples, construct illusory contour versions of datasets like MNIST, and test these illusions in various deep learning models from TorchVision. [30] creates an optical illusion images dataset consisting of 6725 illusion images from different sources and a smaller dataset of 500 hand-picked images. [8] propose a framework for synthesizing visual illusions using deep generative models to understand the differences between vision models and human perception. It also aims to create novel visual illusions by optimizing GANs to produce illusions that have a maximum effect on a given vision model. [11] explains visual illusions, particularly lightness and color visual illusions, through the likelihood of patches in natural images. It introduces a computational model that computes the likelihood of patches based on a large dataset, providing a data-driven explanation for visual illusions. It also highlights the role of retinal image statistics in shaping visual perception and cognition, emphasizing the importance of studying visual input. [1] introduces the WHOOPS dataset and benchmark, focusing on visual commonsense reasoning by presenting unconventional and commonsense-defying images created synthetically. The dataset consists of 500 purposefully unconventional images challenging AI models' visual commonsense reasoning abilities. Tasks include image captioning, cross-modal matching, visual question answering, and a challenging explanation generation task. The dataset aims to inspire the development of AI models with stronger visual commonsense reasoning abilities. [9] introduces "HALLUSIONBENCH", a benchmark designed to evaluate image-context reasoning LVLMs and specifically designed to test the capabilities of LVLMs in handling visual illusions and language and knowledge hallucinations. This benchmark challenges advanced LVLMs like GPT-4V, Gemini Pro Vision, Claude 3, and LLaVA1.5 by emphasizing nuanced understanding and interpretation of visual data. It identifies failure modes related to visual illusion and language hallucination and provides insights into the challenges posed by visual illusion and hallucination in large vision-language models, paving the way for potential improvements in these models. Also [19] provides an in-depth analysis of the failures of state-of-the-art LVLMs like GPT-4V and LLaVA-1.5 on HALLUSIONBENCH. It identifies and categorizes the types of mistakes these models make, attributing them to either language hallucination or visual illusion. [34] investigates how VLMs align with human visual perception under the influence of visual illusions. The study explores whether models interpret visual information similarly to humans when both are subjected to illusions. They introduce the Grounding Visual Illusion in Language (GVIL) benchmark. This benchmark comprises four subtasks (Same-Difference Question Answering, Referential Question Answering, Attribute Question Answering, and Referential Localization) to assess the alignment between human and model interpretations under visual illusions. [26] evaluates the capability of VLMs in understanding and interpreting optical illusions. The study introduces a new dataset, IllusionVQA, which consists of challenging optical illusions designed to test VLMs in two specific tasks: comprehension and soft lo-

calization. IllusionVQA primarily focuses on handling unreasonable images, rather than addressing pareidolia. [2] introduces the Diffusion Illusions framework, which utilizes a frozen text-to-image diffusion model to automatically generate various optical illusions, including flip illusions, rotation overlay illusions, and hidden overlay illusions. The approach leverages score distillation and dream target losses to optimize prime images, demonstrating successful real-world fabrication and expanding the possibilities for creating and using optical illusions. [6] presents a novel method for generating multi-view optical illusions using diffusion models, creating images that reveal different interpretations when subjected to transformations like flips, rotations, or pixel rearrangements. The approach leverages the intrinsic capabilities of diffusion models to produce these illusions without the need for explicit perceptual models. [23] evaluates how various vision-language models interpret bistable images, revealing that most models exhibit strong biases toward one interpretation, largely influenced by language priors rather than visual data. The findings suggest that current models struggle to match human-like perception of ambiguous images, indicating a need for further research in handling visual ambiguity. In these types of images viewers can only see one interpretation at a time. [22] introduces CODIS, a novel benchmark designed to evaluate the context-dependent visual comprehension of Multimodal Large Language Models (MLLMs), highlighting the models' current limitations in leveraging contextual information to disambiguate images. A concurrent study [10], explores a similar idea to ours but with key differences. Their dataset relies on a limited set of raw conditioning images to generate new data, whereas we use a diverse set of images for each class, enhancing the generalizability of our dataset. Additionally, their scene descriptions are relatively limited. To introduce further challenges, we incorporate an additional "No Illusion" class, containing images that do not depict any visual illusions. Beyond fine-tuning, we apply a filtering step to the final images, which we find to be both simple and effective. Furthermore, while their approach focuses primarily on classification, our work also includes an OCR-like dataset, expanding the scope of multimodal model evaluation.

3. Task and Dataset

In this section, we propose a task definition for Illusory VQA and discuss some challenges associated with this task, as well as the process of creating our datasets.

3.1. Task Definition

In the Illusory VQA task, the goal is to analyze illusory images using a Vision-Language Model (VLM). These images, denoted as ***Illusory Images (II)***, depict both a ***Real Concept (RC)*** and ***potentially an Illusory Concept (IC)***. Given an ***Illusory Image (II)*** and a ***question (Q)*** about the image, the VLM must first detect if there is an illusory concept present and then provide an ***answer (A)*** specifically related to the illusory concept in the image.

In this study, we focus exclusively on the illusory concept, ensuring that the model can accurately detect and respond to questions about it. This means that the primary objective is to evaluate the model's ability to identify the illusory concept within the image and provide a correct answer related to that concept. You can find an example of the task definition in the Supplementary Material 8.

3.2. Challenges of the Task

One of the primary challenges of the Illusory VQA task is accurately detecting the presence of an illusory concept (*IC*) in the image. Despite the presence of the real concept (*RC*) in the image, the model must still answer questions pertaining to the illusory concept. For example, in classification tasks, there may be an illusion of one class in the picture, but instances of another class may also be present. In such cases, the model must correctly detect the illusory concept and provide accurate answers to questions about it.
By addressing these challenges, the Illusory VQA task aims to enhance our understanding of how VLMs perceive and interpret illusory images, as well as their ability to reason about illusory concepts within the context of a given question.

3.3. Data Generation

To create and collect illusory images, we initially generate 1027 descriptions of scenes for transferring images to new space, using various LLMs such as ChatGPT, Gemini, Mixtral-8x7B-Instruct-v0.1 [13], and Gemma-1.1-7b-it [29] to ensure diversity of descriptions. We use these descriptions to generate the real concept of our images (*RC*). We ensured that the number of descriptions generated from each LLM was approximately equal. After collecting our raw images (*IC*), we combine them with the descriptions (*RC*) generated by the LLMs. Using a variant of ControlNet model [33], we utilized this combined data to generate illusory images (*II*). The demonstration of the whole process is shown in Figure 3. While these datasets are generated synthetically, we believe they hold significant value. The quality of the generated images has been validated by human reviewers, which enhances their reliability. Additionally, these datasets can assist multimodal models in understanding images more effectively, similar to human perception. In order to ensure the safety and appropriateness of our datasets, we took several precautionary measures. Initially, we assumed that our original datasets, namely MNIST and Fashion-MNIST, as well as the animal dataset generated by SDXL Lightning, did not contain any offensive content. Additionally, we confirmed that the Char dataset also does

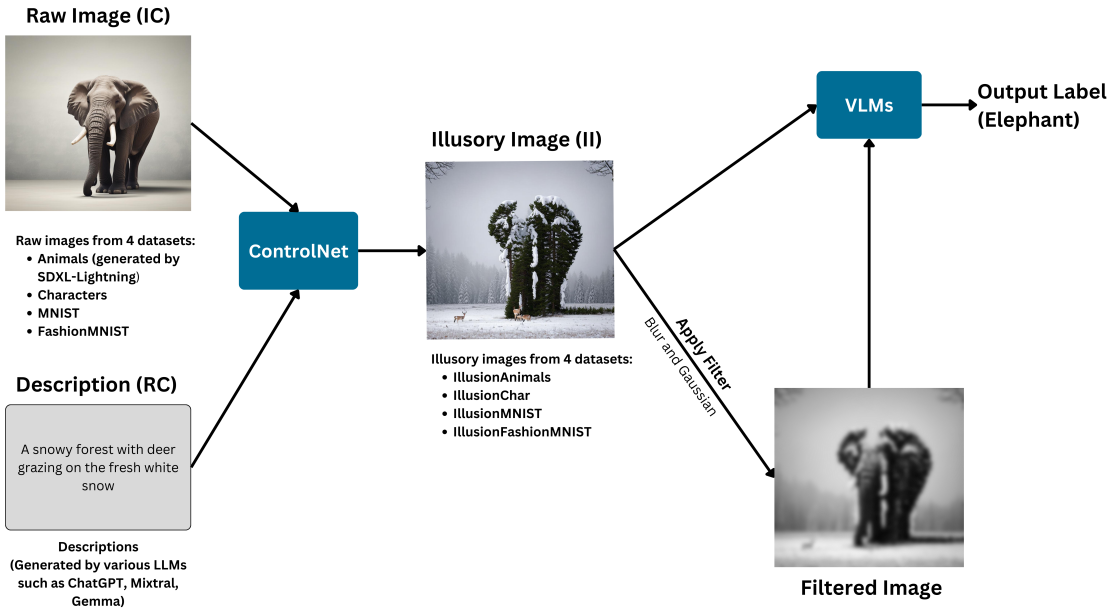

Figure 3. Demonstraion of our pipeline for generating and evaluating datasets

not include any offensive content. To further validate the absence of offensive content, we employed GPT-3.5 to examine and verify the descriptions generated for the datasets. This step was crucial in ensuring that our descriptions were free from any offensive material. Based on these measures, we initially believed that our datasets were free from offensive content. However, to provide further assurance, we conducted a sampling process. We randomly selected 10% of the data from each class in our datasets and presented them to human annotators. Their task was to identify any offensive content within the samples. This process allowed us to determine whether any offensive content was present in our datasets. Surprisingly, during this evaluation, we discovered that a portion of the images in our datasets did indeed contain offensive content. To address this, we employed NSFW (Not Safe for Work) detector models on our dataset images. By running these models, we were able to identify and filter out the images that contained offensive content. In our study, we conducted experiments using a dataset consisting of content that has the potential to be offensive. For the purpose of publishing our datasets, we have excluded the data that was flagged by the NSFW detector models. The datasets that we are making available for public use have undergone this additional filtering process to ensure that they are free from offensive content.

3.3.1. IllusionMNIST

To create IllusionMNIST, we start by selecting MNIST[1] [14] as our source for raw images. From there, we sample images and resize them to a resolution of 512 pixels by 512 pixels. These images were then combined with descriptions generated by LLMs, and IllusionMNIST was formed using a variant of ControlNet. In order to further challenge the models, we introduced an additional class called 'No illusion' to the original set of classes, as we can see in Figure 2. This class enables the models to detect instances where no illusion images were present in the picture. To ensure fairness in the dataset, we ensure that an equal number of samples are included for the 'No illusion' class. Ultimately, our dataset consists of 3960 training samples and 1219 test samples. The statistics of our datasets are shown in Table 1. For more information, please refer to Supplementary Material 11.

Table 1. Size of provided datasets

Dataset	# of training samples	# of test samples
IllusionMNIST	3960	1219
IllusionFashionMNIST	3300	1267
IllusionAnimals	3300	1100
IllusionChar	9900	3300

3.3.2. IllusionFashionMNIST

For the creation of IllusionFashionMNIST, we opt to use Fashion-MNIST[2] [31] as the source for our raw images. Similar to IllusionMNIST, we sample images from the dataset and resize them to a resolution of 512 pixels by 512 pixels. These images were then combined with descrip-

[1]The MNIST dataset consists of 10 classes representing digits from 0 to 9.

[2]The Fashion-MNIST dataset contains 10 classes representing items such as 'T-shirt/top', 'Trouser', 'Pullover', 'Dress', 'Coat', 'Sandal', 'Shirt', 'Sneaker', 'Bag', and 'Ankle boot'.

tions generated by LLMs, and IllusionFashionMNIST was formed using ControlNet. Just like in the previous case, we include an additional class called 'No illusion' to provide a challenge to the models. In the end, we obtain a dataset of 3300 training samples and 1267 test samples.

3.3.3. IllusionAnimals

To generate IllusionAnimals, we create a dataset[3] consisting of raw animal images using SDXL-Lightning [18]. These images were then combined with descriptions generated by LLMs and transformed into IllusionAnimals using ControlNet. Once again, we add an extra class called 'No illusion' to push the models' capabilities. The final dataset for IllusionAnimals comprises 3300 training samples and 1100 test samples.

3.3.4. IllusionChar

To explore problems beyond classification, we develop a dataset called IllusionChar, which centers around reading characters in pictures. The creation of the IllusionChar dataset involves a two-step process. Initially, we generate images consisting of a sequence of characters, ranging from 3 to 5 characters per image. Subsequently, we apply an illusion effect to these characters using ControlNet. Just like the other datasets, we include some images with 'No illusion' to provide diversity. As a result, the IllusionChar dataset contains 9900 training samples and 3300 test samples.

Human Validation and Evaluation To assess the quality of the generated dataset, we conduct a human evaluation involving a panel of participants. For each dataset, we select three individuals to participate. In the evaluation and validation process, we randomly sample 10% of each dataset (10% of each class for classification datasets) and present the images to the participants. They were asked to identify the class or category they perceived in each picture. To ensure the validity of the generated images, we also disclose the true label after they choose their label to the participants and ask them to confirm if they perceive the indicated class or not.
please consult the Supplementary Material 10 for more details.

4. Experimental Setup

We conduct evaluations on various models for IllusionMNIST, IllusionFashionMNIST, IllusionAnimals, and IllusionChar. Specifically, we assess the performance of GPT-4o, Gemini 1.0 Pro Vision, LLaVA-1.5-7B, Kosmos-2-patch14-224, MiniGPT-V2 based on Llama2 Chat 7B (after stage-3 pretrained weights), CLIP-ViT-base-batch32, BLIP-Large, and BLIP-2 on IllusionMNIST, IllusionFashionMNIST, and IllusionAnimals. We use Hugging Face and LAVIS [15] for implementing these models. For IllusionChar, we focused on evaluating GPT-4o and Gemini 1.0 Pro Vision.
For the fine-tuning phase on IllusionMNIST, IllusionFashionMNIST, and IllusionAnimals, we utilize LLaVA, CLIP, BLIP, and BLIP-2. However, due to hardware limitations, we employed LoRA [12] for the LLaVA model. For detailed information about implementation, please refer to Supplementary Material 16.

[3]The dataset created for IllusionAnimals comprises 10 classes representing animals such as 'cat', 'dog', 'pigeon', 'butterfly', 'elephant', 'horse', 'deer', 'snake', 'fish', and 'rooster'.

5. Results

In our study, we observe the performance of various models on different datasets, both with and without the application of illusions and filters. Let's break down the findings for each dataset:

5.1. IllusionMNIST Dataset

- As illustrated in Table 2, when applying illusions to the original images, we notice a significant drop in the performance of the models. This indicates that VLMs struggle to interpret illusions in images.
- Among all the models, GPT-4o achieves the best results on the IllusionMNIST dataset, surpassing other models in both the illusion and filter-applied scenarios.
- Additionally, when we apply our simple filter to the images, we observe that almost all models, except for Kosmos-2 and LLaVA, show improved performance compared to the scenario without the filter. This demonstrates the effectiveness of our approach to this dataset.
- After fine-tuning the models on the provided training dataset, we observe a noticeable improvement in their performance, as demonstrated in Table 4.

5.2. IllusionFashionMNIST Dataset

- Also, in this case, we see that applying illusion causes a drop in results.
- Prior to applying filters, GPT-4o exhibits the best results on the illusory images of the IllusionFashionMNIST dataset.
- After applying the filter, BLIP-2 achieves the best performance.
- Similarly to the previous dataset, we observe that all models show improved performance when the filter is applied to the images. This suggests that the filter has a positive impact on the models' ability to interpret the illusions.
- Likewise, we observed an improvement in performance for all models after fine-tuning.

Table 2. Zero-shot performance of different models on different datasets: The term 'Raw' refers to raw images without any illusions. 'Illusion' refers to illusory images, while 'Filtered' indicates illusory images that have been processed with our filter.

		IllusionMNIST				IllusionFashionMNIST				IllusionAnimals			
		Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
GPT-4o*	**Raw**	**89.88**	**90.26**	**85.89**	**86.99**	67.12	**71.05**	**67.89**	**65.93**	**100**	**100**	**100**	**100**
	Illusion	**37.47**	44.94	**37.40**	**36.95**	**16.67**	27.76	**16.61**	**12.72**	34.89	35.89	27.45	27.08
	Filtered	**67.97**	70.39	**67.99**	**68.32**	**48.73**	48.76	44.73	41.58	83.99	72.60	71.08	70.39
Gemini	**Raw**	89.36	76.23	74.27	74.26	**67.27**	40.56	37.43	36.74	87.20	76.92	67.08	68.68
	Illusion	21.82	**72.74**	22.15	23.36	11.05	**33.25**	8.12	4.24	25.27	19.95	7.32	8.12
	Filtered	63.82	**83.44**	64.37	68.23	40.57	54.20	31.98	33.26	85.73	59.92	55.47	56.40
LLaVA	**Raw**	55.46	83.52	56.66	59.75	38.63	60.56	39.34	38.38	**100**	**100**	**100**	**100**
	Illusion	9.02	0.82	9.09	1.50	9.00	0.82	9.01	1.51	13.91	45.61	13.91	10.42
	Filtered	9.02	0.82	9.09	1.50	13.73	26.90	13.72	7.55	51.18	90.09	51.18	52.83
Kosmos-2	**Raw**	6.22	4.54	0.49	0.82	12.07	5.82	6.97	2.99	92.40	67.79	66.00	66.35
	Illusion	8.94	1.71	8.36	2.29	8.60	0.39	3.89	0.71	8.64	2.30	1.46	0.71
	Filtered	8.86	0.74	8.18	1.36	9.23	0.78	8.33	1.42	23.00	25.11	15.81	13.17
MiniGPT-V2	**Raw**	61.59	76.60	61.63	62.76	10.42	31.02	10.27	2.38	**100**	**100**	**100**	**100**
	Illusion	15.75	48.66	16.02	13.23	9.16	0.84	9.01	1.54	12.19	**50.70**	12.18	8.23
	Filtered	36.67	64.23	36.9	41.65	9.63	1.45	8.70	2.06	58.55	83.28	58.55	59.97
CLIP	**Raw**	25.97	28.01	27.70	20.14	41.15	47.13	41.33	37.54	**100**	**100**	**100**	**100**
	Illusion	15.26	18.07	15.32	13.02	13.89	27.99	13.92	9.20	**42.64**	46.20	**42.64**	**39.56**
	Filtered	21.16	21.17	21.63	18.93	44.75	50.50	44.74	41.28	85.45	86.37	85.45	85.19
BLIP	**Raw**	29.31	31.65	29.57	28.62	60.16	64.45	60.17	55.82	**100**	**100**	**100**	**100**
	Illusion	14.44	21.05	14.45	11.98	11.84	11.09	12.02	9.72	31.90	42.26	31.90	31.45
	Filtered	16.57	22.38	17.33	15.19	40.65	53.58	41.04	37.05	89.90	90.66	89.90	88.56
BLIP-2	**Raw**	61.68	70.26	62.85	61.81	64.41	70.10	64.40	63.84	**100**	**100**	**100**	**100**
	Illusion	15.67	15.83	15.65	12.85	11.69	22.68	11.47	9.49	32.64	41.46	32.64	30.05
	Filtered	40.20	43.30	40.84	37.77	45.15	**62.46**	**45.19**	**45.08**	**93.73**	**94.54**	**93.73**	**94.23**
Human	**Illusion**	96.69	97.22	96.05	96.43	74.6	72.81	73.63	72.85	93.03	92.86	90.88	91.5

* The coverage of GPT-4o on different datasets is as follows: Coverage on IllusionMNIST for Raw, Illusion, and Filtered is 22.27%, 99.84%, and 96.8% respectively. Coverage on IllusionFashionMNIST for Raw, Illusion, and Filtered is 70.75%, 99.92%, and 99.45% respectively. Coverage on IllusionAnimals for Raw, Illusion, and Filtered is 99.90%, 99.55%, and 99.91% respectively.

Table 3. Zero-shot OCR capability of Gemini and GPT-4o on IllusionChar dataset

		Gemini	**GPT-4o**	**Human**
Raw	**WER**	53.73	**22.34**	-
	CER	56.73	**7.01**	-
Illusion	**WER**	90.48	**90.26**	31.94
	CER	175.98	**169.46**	13.32
Filtered	**WER**	82.73	**76.4**	-
	CER	**99.93**	122.18	-

5.3. IllusionAnimals Dataset

- Similarly, in this case, we observe that applying illusion causes a drop in results.
- Before applying filters, CLIP displays the highest performance on the illusory images of the IllusionAnimals dataset.
- After applying the filter, BLIP-2 yields the best results, even better than humans.
- Once again, the application of the filter results in an increase in the performance of all models on the illusory images. This further emphasizes the effectiveness of our solution in improving model performance on illusory images.
- Similarly, we observed an enhancement in performance for all models after the process of fine-tuning.

5.4. IllusionChar Dataset

- Similarly, in this case, the application of the illusion technique leads to a drop in results, as demonstrated in Table 3.
- Before applying filters, GPT-4o demonstrates better performance than Gemini.
- After applying the filter, performance improves for both models, highlighting the effectiveness of our solution. GPT-4o exhibits enhanced performance in terms of WER, while Gemini showcases superior performance in CER.

Overall, our study highlights that the interpretation of illusions in images poses a challenge for VLMs. However, the application of filters can significantly improve model performance. You can see better visualization of results in Supplementary Material 12. One interesting observation is that, although we fine-tuned our models on illusory data, the performance of our model on raw images increased in some cases.

6. Conclusion and Future Work

In conclusion, despite the significant advancements in multimodal models, there remain notable disparities in how these models perceive and interpret images compared to human perception. One specific area where such disparities exist is in illusory images. To address this, we introduce IllusionMNIST, IllusionFashionMNIST, IllusionAni-

Table 4. Performance of different fine-tuned models on different datasets

		IllusionMNIST				IllusionFashionMNIST				IllusionAnimals			
		Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
LLaVA	**Raw**	78.99	83.75	78.95	79.16	45.66	59.45	46.2	45.85	**100**	**100**	**100**	**100**
	Illusion	40.11	38.44	33.36	33.33	12.79	13.81	12.88	12.90	32.73	36.44	32.73	33.39
	Filtered	70.55	76.71	70.29	67.84	35.04	33.96	35.18	30.98	87.91	89.79	87.91	87.92
CLIP	**Raw**	**92.25**	**93.24**	**92.22**	**92.24**	**84.20**	**85.85**	**84.99**	**84.35**	99.60	99.61	99.60	99.60
	Illusion	**91.8**	**92.47**	**91.72**	**91.69**	**83.90**	**85.17**	**84.64**	**84.15**	**94.36**	**94.87**	**94.36**	**94.37**
	Filtered	**91.55**	**92.32**	**91.47**	**91.47**	**82.00**	**83.48**	**82.92**	**81.82**	88.73	90.80	88.73	88.55
BLIP	**Raw**	12.89	12.90	12.76	11.40	51.74	64.70	51.91	47.54	98.60	98.64	98.60	98.61
	Illusion	20.43	23.45	20.30	18.55	57.70	68.00	57.82	53.94	94.27	94.35	94.27	94.28
	Filtered	19.77	17.78	19.52	17.33	58.64	67.75	58.85	54.79	**94.36**	**94.63**	**94.36**	**94.35**
BLIP-2	**Raw**	89.36	90.06	89.60	89.11	43.92	38.42	44.25	36.00	69.10	67.63	69.10	62.80
	Illusion	55.54	57.64	55.56	55.22	41.99	34.68	42.55	36.72	67.55	68.35	67.55	67.29
	Filtered	86.79	87.70	86.91	86.57	46.88	43.96	47.43	41.96	89.45	90.30	89.45	89.33

mals, and IllusionChar datasets, which serve as valuable benchmarks for evaluating the performance of VLMs on illusory images. Additionally, we propose a straightforward yet effective approach to enhance the performance of these models on such images. This involves applying a Gaussian and blur low-pass filter and converting the images to grayscale. Implementing this method demonstrates notable performance improvements. Furthermore, we suggest the possibility of fine-tuning VLMs specifically for illusory images.

Looking ahead, there is ample room for future exploration and development. One potential avenue is the utilization of adaptive and learnable filters tailored specifically for these types of images. Furthermore, expanding the collection of more comprehensive and diverse datasets for illusory images would be beneficial, while our current study has primarily focused on image classification and OCR capability. It would be worthwhile to investigate the impact of in-context learning on these types of images. By delving into these areas, we can further bridge the gap between multimodal models and human perception, ultimately advancing the field of image interpretation and analysis. Novel architectures specifically designed to address the challenges posed by illusory images can be explored. The application of adversarial training techniques to improve the models' robustness against illusory images can be investigated. Human feedback and input can be incorporated into the training process to refine the models' understanding of illusory images, bridging the gap between model and human perception. Quantitative evaluation metrics specifically tailored for assessing performance on illusory images can be developed. Based on our observations, we believe that this topic could enhance the robustness of Vision-Language Models (VLMs) and help mitigate issues such as spurious correlations and shortcut learning. If proven effective, this approach could be utilized as a data augmentation technique in the training of VLMs. However, this would require additional work.

7. Limitations

Our work focuses on the classification and OCR aspects of illusory images, while our goal was to develop a comprehensive dataset for illusory images to support general VQA task. However, we encountered several limitations:

- **Limited Scope of Illusory Images** Our work primarily focuses on illusions in images related to classification and OCR. While we aimed to create a comprehensive dataset for illusory images for general VQA tasks, there are certain limitations that need to be acknowledged.
- **Inability to Assess Color Changes** One limitation is that we were unable to assess the color of the images using the Visual Language Model (VLM) due to the nature of illusions. Applying illusions to images can alter the appearance of objects, making it challenging to accurately determine their original colors.
- **Single Large Objects** Another limitation is that our work mainly focuses on images containing just one large-sized object. This choice was made to simplify the dataset construction process and reduce the burden of evaluation. However, it is important to note that illusions can occur with multi-object scenes and objects of varying sizes. Exploring these scenarios could provide a more comprehensive understanding of illusions in images.
- **Reduced Task Difficulty** Constructing datasets in the classification and OCR formats helps streamline the evaluation process. However, this approach may inadvertently reduce the difficulty of the task. It allows models to potentially solve the classification task by randomly choosing one of the classes, rather than genuinely understanding the underlying illusions present in the images.

When considering the limitations of our study, it is important to acknowledge the hardware constraints that have influenced our approach. Specifically, we have chosen not to fine-tune our models on IllusionChar due to these limitations. By refraining from this step, we aim to explore the potential outcomes that may arise. By acknowledging these limitations, we can better understand the context and scope of our work and identify areas for future research and improvement.

References

[1] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2616–2627, 2023. 3

[2] Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael S. Ryoo. Diffusion illusions: Hiding images in plain sight, 2023. 4

[3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023. 3

[4] R. H. Day. The nature of, perceptual illusions. *Interdisciplinary Science Reviews*, 9(1):47–58, 1984. 2

[5] Jinyu Fan and Yi Zeng. Challenging deep learning models with image distortion based on the abutting grating illusion. *Patterns*, 4(3):100695, 2023. 3

[6] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models, 2024. 4

[7] Alexander Gomez-Villa, Adrian Martín, Javier Vazquez-Corral, and Marcelo Bertalmío. Convolutional neural networks can be deceived by visual illusions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12301–12309, 2019. 2

[8] Alex Gomez-Villa, Adrián Martín, Javier Vázquez-Corral, Marcelo Bertalmío, and Jesús Malo. On the synthesis of visual illusions using deep generative models. *Journal of Vision*, 22, 2022. 3

[9] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination visual illusion in large vision-language models, 2023. 3

[10] Arshia Hemmat, Adam Davies, Tom A. Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in plain sight: Evaluating abstract shape recognition in vision-language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 4

[11] Elad Hirsch and Ayellet Tal. Color visual illusions: A statistics-based computational model. In *Advances in Neural Information Processing Systems*, pages 9447–9458. Curran Associates, Inc., 2020. 2, 3

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 6

[13] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, and et al. Mixtral of experts, 2024. 4

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[15] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022. 6

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 3

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 3

[18] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation, 2024. 6

[19] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models, 2023. 3

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[21] Ben Lonnqvist, Alban Bornet, Adrien Doerig, and Michael H. Herzog. A comparative biology approach to dnn modeling of vision: A focus on differences, not similarities. *Journal of Vision*, 21, 2021. 2

[22] Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, Peng Li, Ning Ma, Maosong Sun, and Yang Liu. CODIS: Benchmarking context-dependent visual comprehension for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10639–10659, Bangkok, Thailand, 2024. Association for Computational Linguistics. 4

[23] Artemis Panagopoulou, Coby Melkin, and Chris Callison-Burch. Evaluating vision-language models on bistable images. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 8–29, Bangkok, Thailand, 2024. Association for Computational Linguistics. 4

[24] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3

[26] Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models, 2024. 3

[27] Eric Sun and Ron Dekel. Imagenet-trained deep neural network exhibits illusion-like response to the scintillating grid, 2019. 2

[28] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, and et al. Gemini: A family of highly capable multimodal models, 2024. 3

[29] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, and et al. Gemma: Open models based on gemini research and technology, 2024. 4

[30] Robert Max Williams and Roman V. Yampolskiy. Optical illusions images dataset, 2018. 3

[31] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. 5

[32] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 3

[33] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 4

[34] Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5718–5728, Singapore, 2023. Association for Computational Linguistics. 3

[35] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3