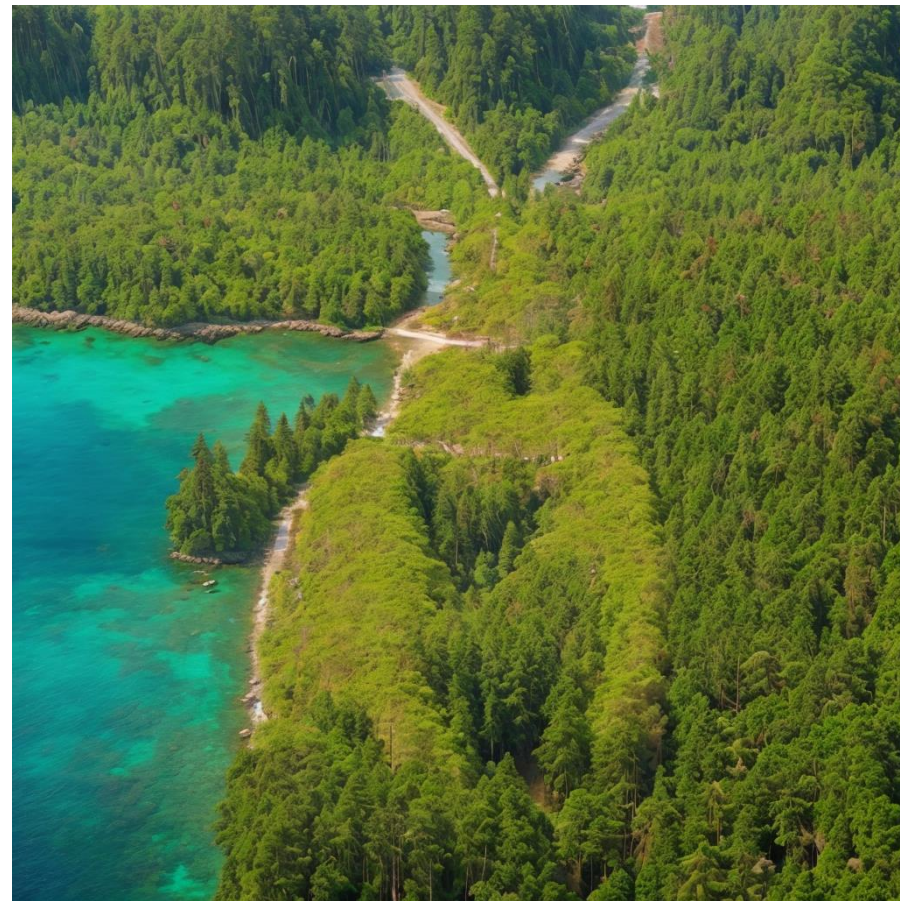
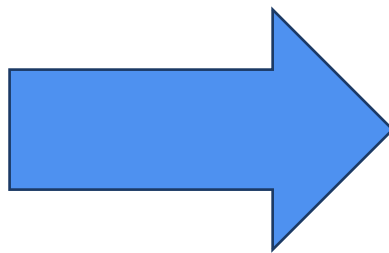


# Illusory VQA: Benchmarking and Enhancing Multimodal Models on Visual Illusions

Mohammadmostafa Rostamkhani,  
Baktash Ansari, Hoorie Sabzehvari,  
Farzan Rahmani, Sauleh Eetemadi

# Visual Illusion (Pareidolia)





# Illusory VQA

- Importance
- Task Definition
- Main Challenges

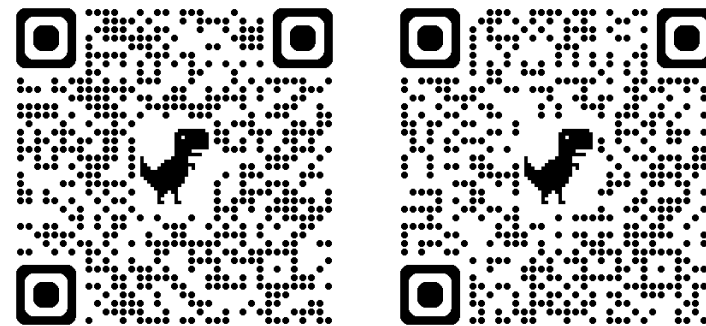


- (II): image on the left
- (RC): A bustling train station with passengers rushing to catch their trains
- (IC / Ground Truth): butterfly
- (Q): “There might be an illusion of something in the image or not. These are the classes that an illusion might belong to: {illusion\_class\_names\_str}. Just choose the correct class without any extra explanation.”
- (A): butterfly



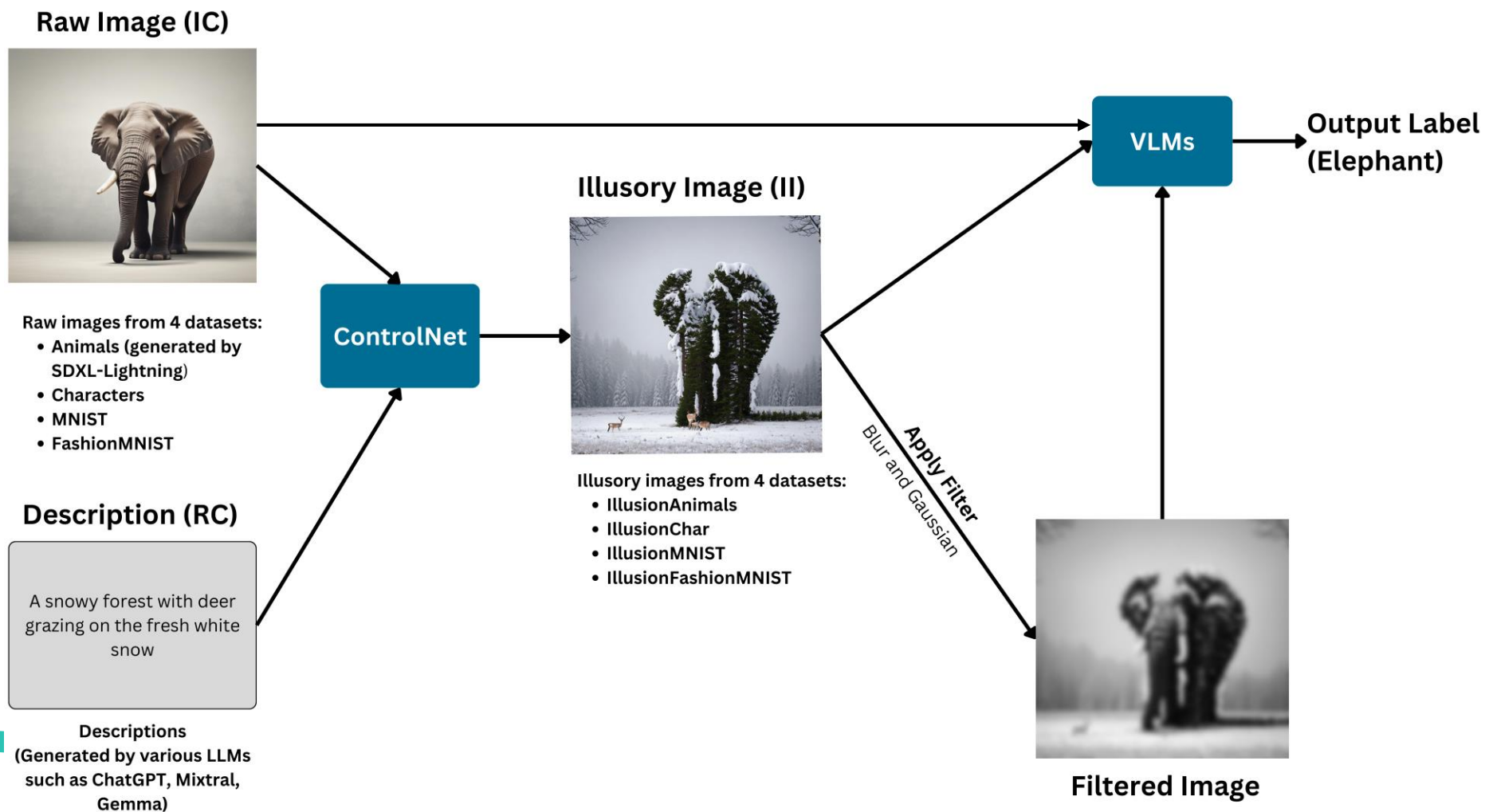
- (II): image on the left
- (RC): A bustling city street with neon lights and bustling crowd
- (IC / Ground Truth): elephant
- (Q): “There might be an illusion of something in the image or not. These are the classes that an illusion might belong to: {illusion\_class\_names\_str}. Just choose the correct class without any extra explanation.”
- (A): elephant

# Datasets



Dataset	# of training samples	# of test samples	# of classes	classes
IllusionMNIST	3960	1219	11	[0,1,2,3,4,5,6,7,8,9,No Illusion]
IllusionFashionMNIST	3300	1267	11	['T-shirt/top', 'Trouser', 'Pullover', 'Dress', 'Coat', 'Sandal', 'Shirt', 'Sneaker', 'Bag', 'Ankle boot', 'No Illusion']
IllusionAnimals	3300	1100	11	['cat', 'dog', 'pigeon', 'butterfly', 'elephant', 'horse', 'deer', 'snake', 'fish', and 'rooster', 'No Illusion']
IllusionChar	9900	3300	-	-

# Data Generation and Evaluation pipeline



# An example of IllusionMNIST



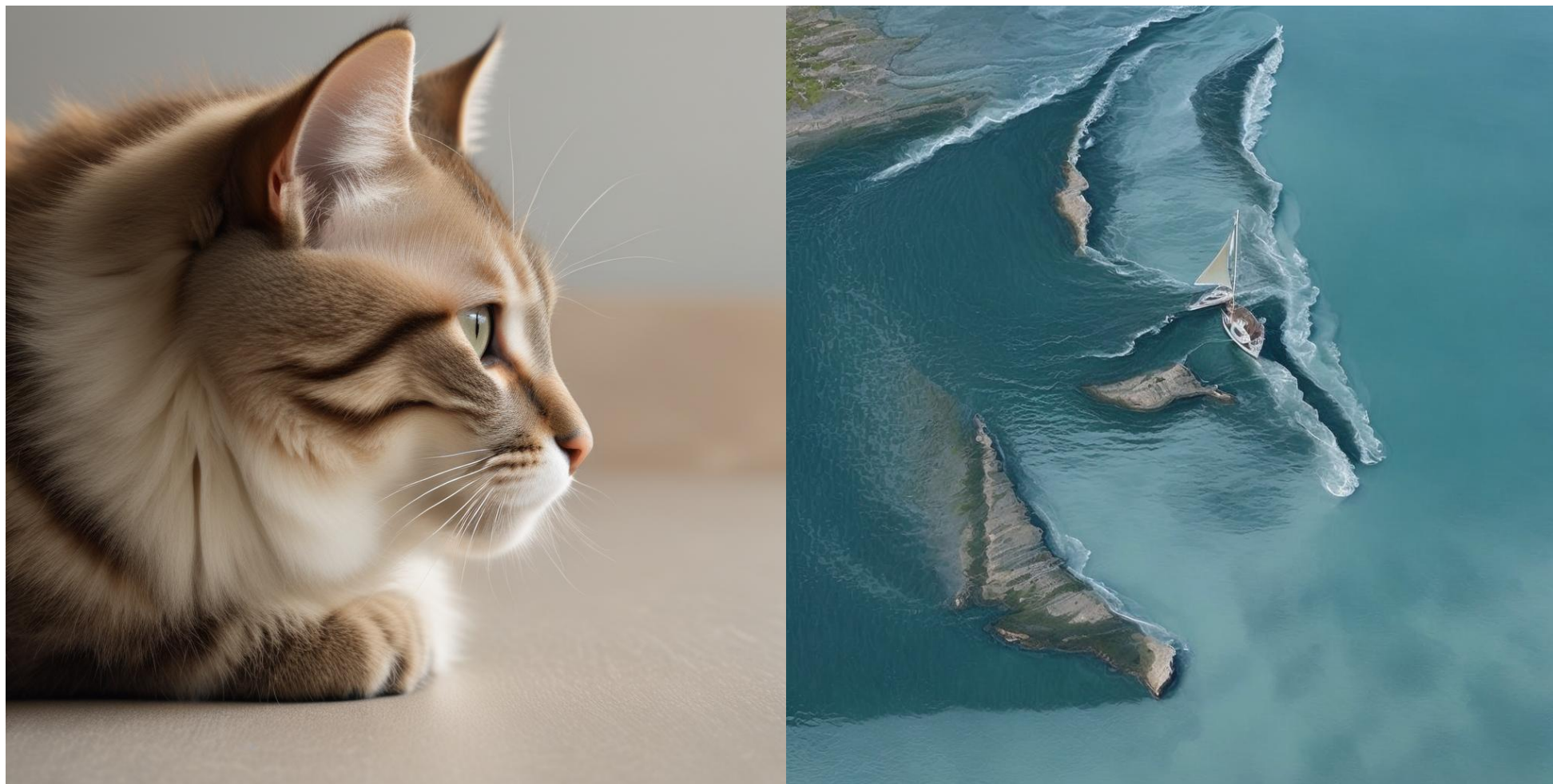
Illusory VQA: Benchmarking and Enhancing Multimodal Models on Visual Illusions



# An example of IllusionFashionMNIST



# An example of IllusionAnimals



Illusory VQA: Benchmarking and Enhancing Multimodal Models on Visual Illusions

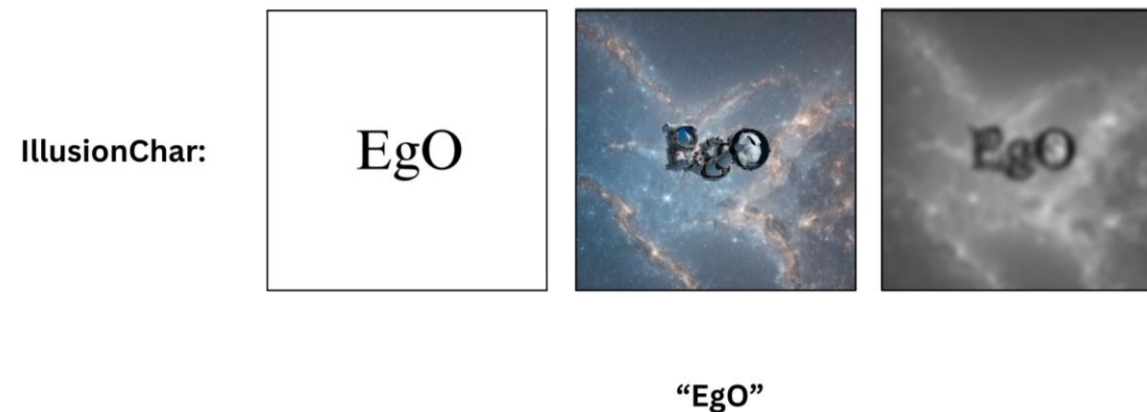
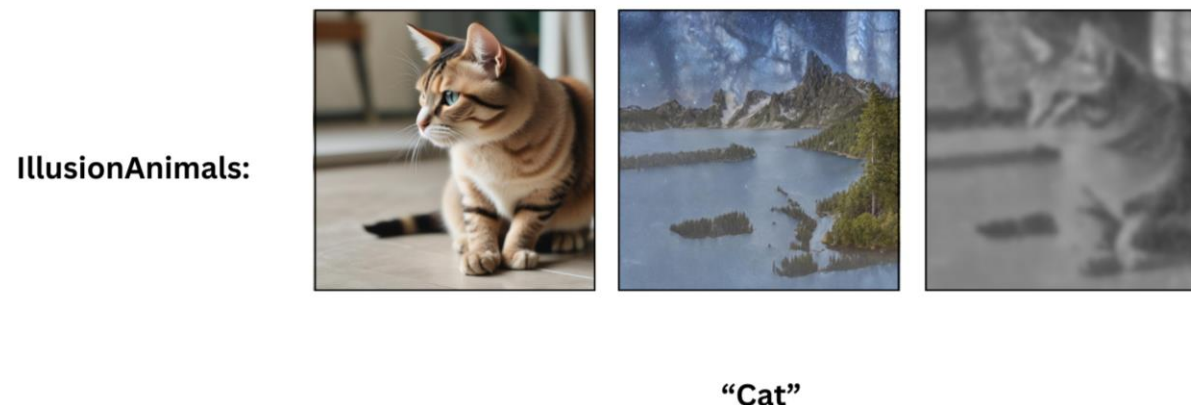
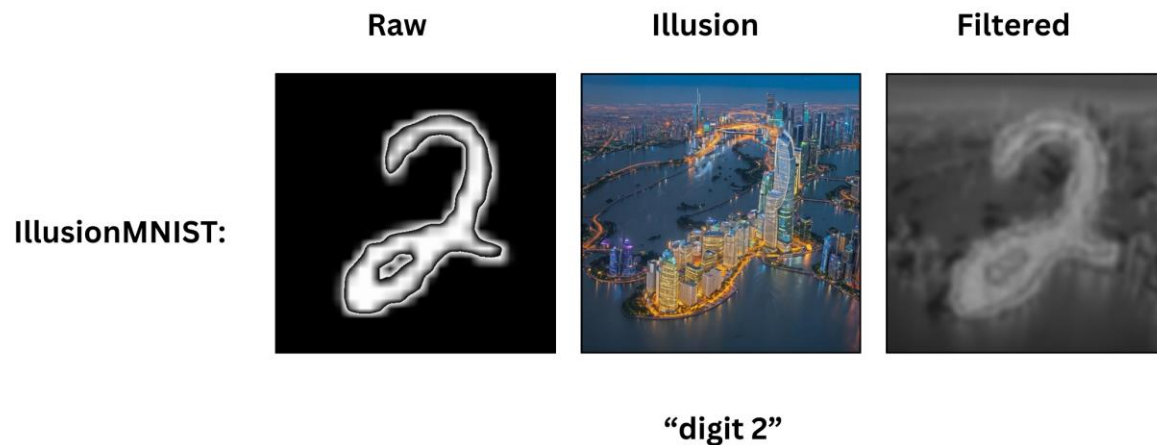


# An example of IllusionChar

3CK1L

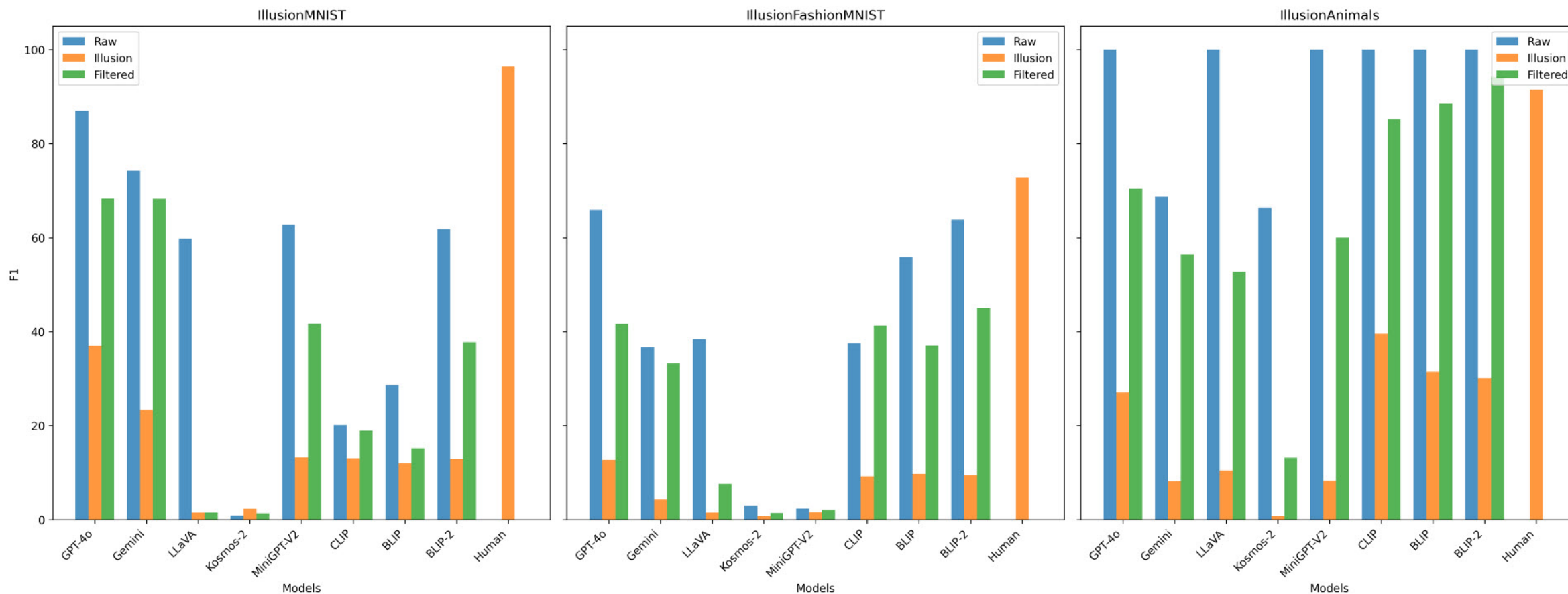


# Blur and gray-scale filter as a proposed method



# F1score performance across datasets

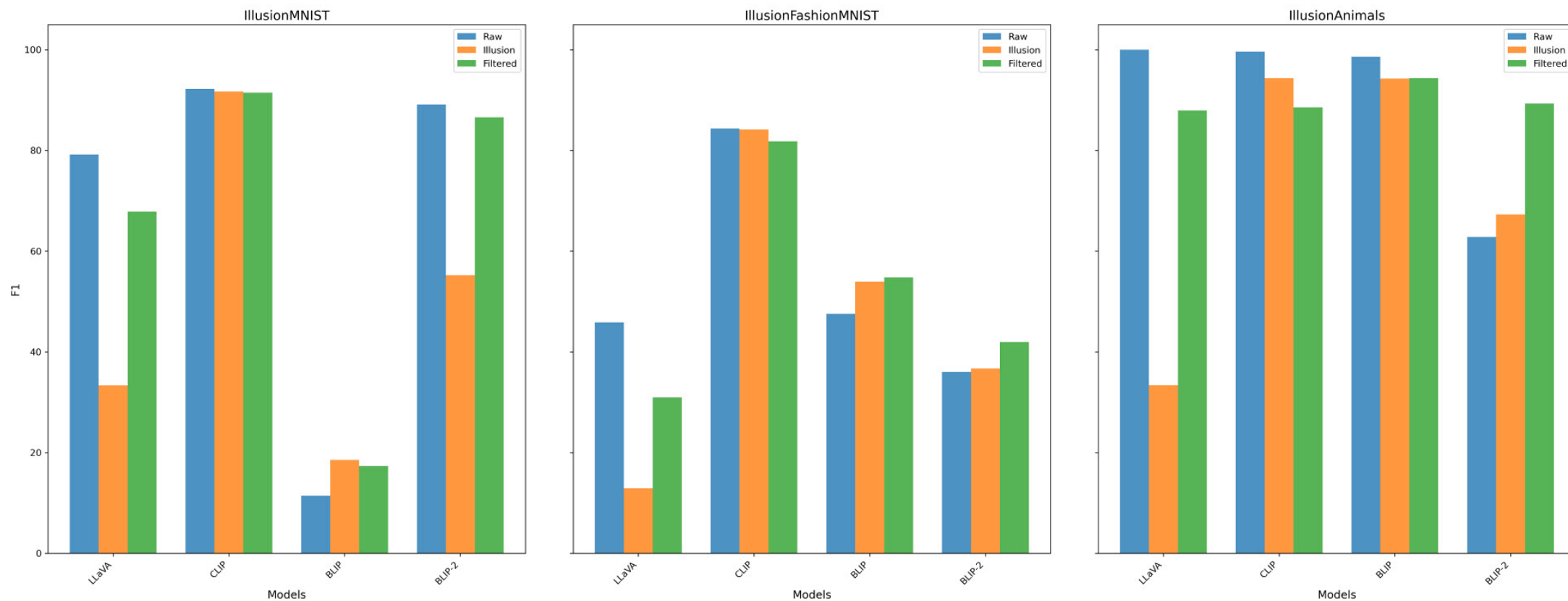
F1 Comparison Across Models and Datasets





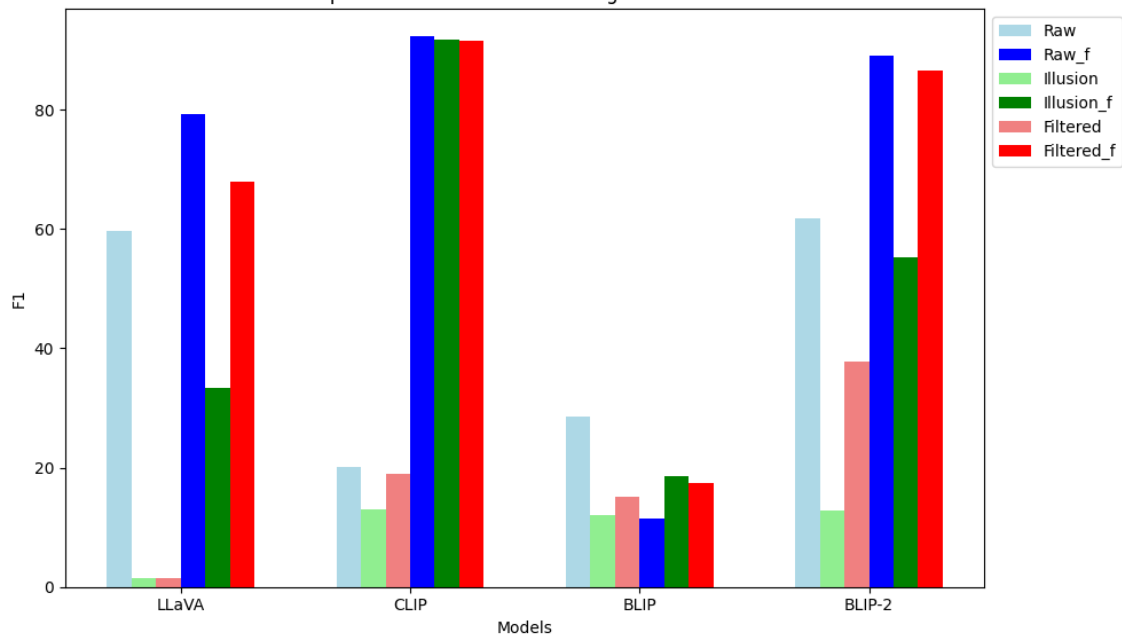
# F1score performance across datasets

F1 Comparison Across Fine-tuned Models and Datasets

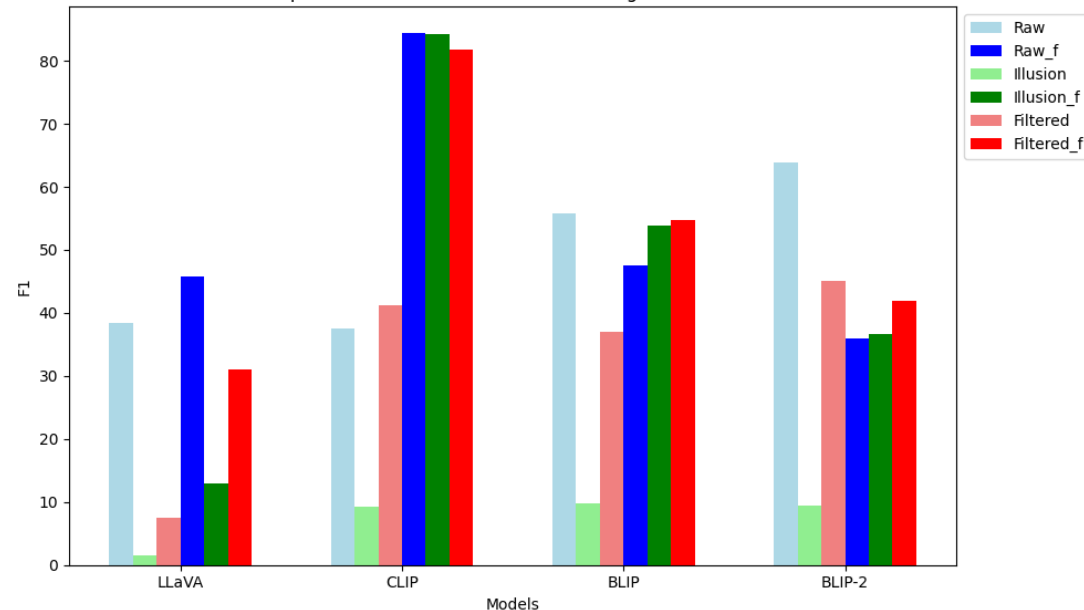


# F1score comparison between zero-shot and fine-tuning

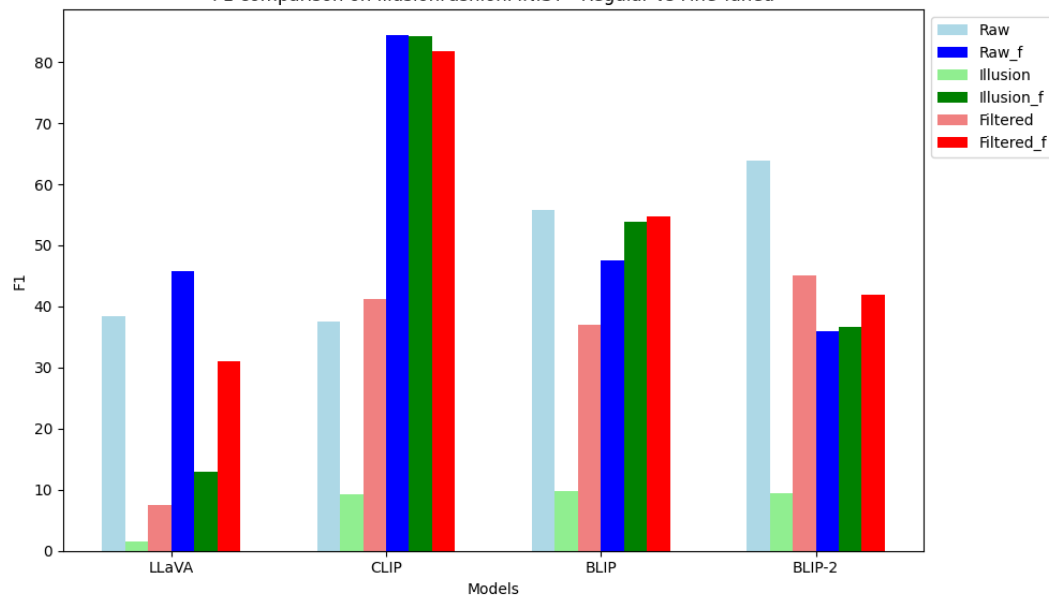
F1 comparison on IllusionMNIST - Regular vs Fine-Tuned



F1 comparison on IllusionFashionMNIST - Regular vs Fine-Tuned



F1 comparison on IllusionFashionMNIST - Regular vs Fine-Tuned

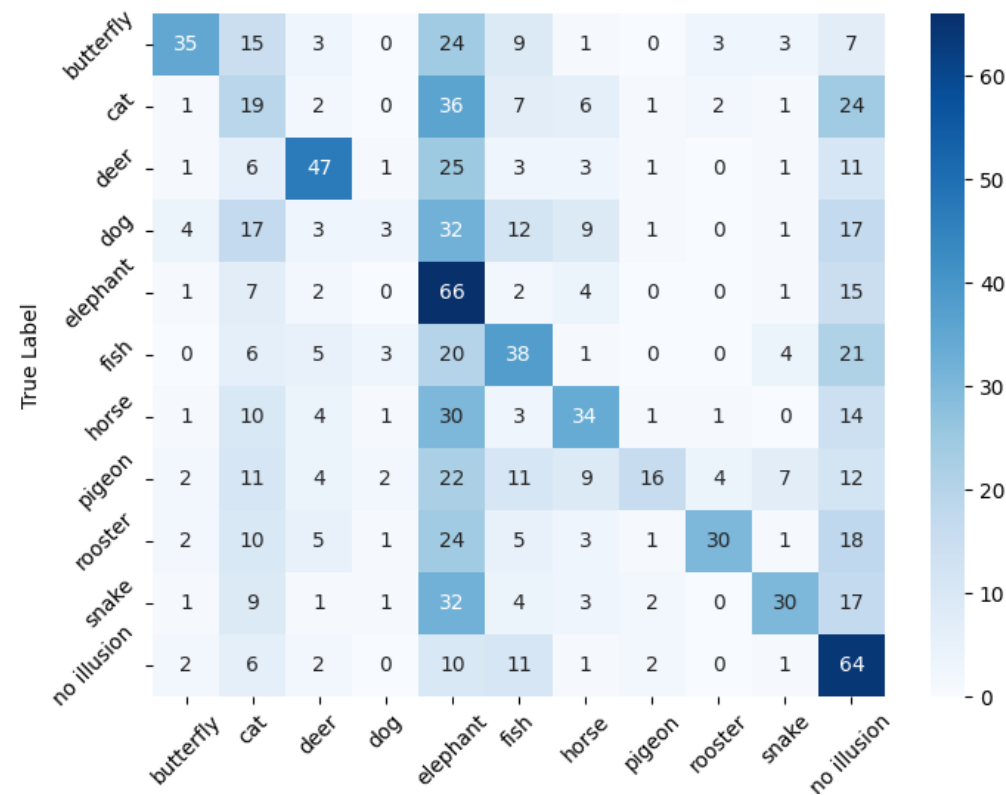
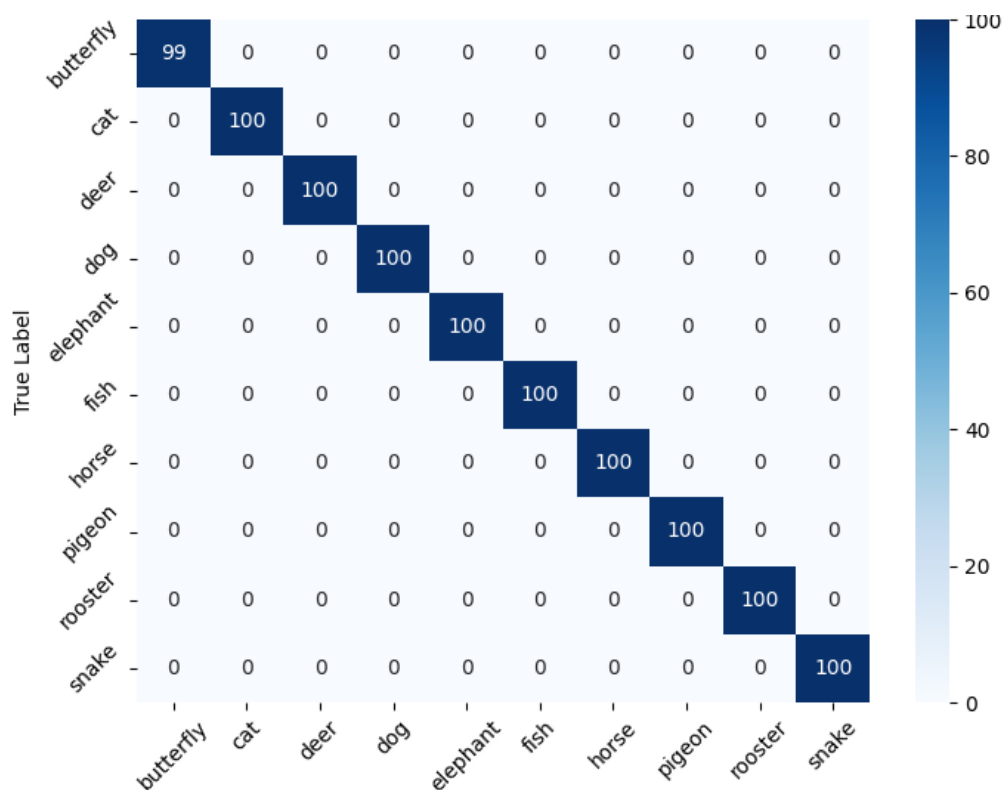


# Zero-shot performance on IllusionChar

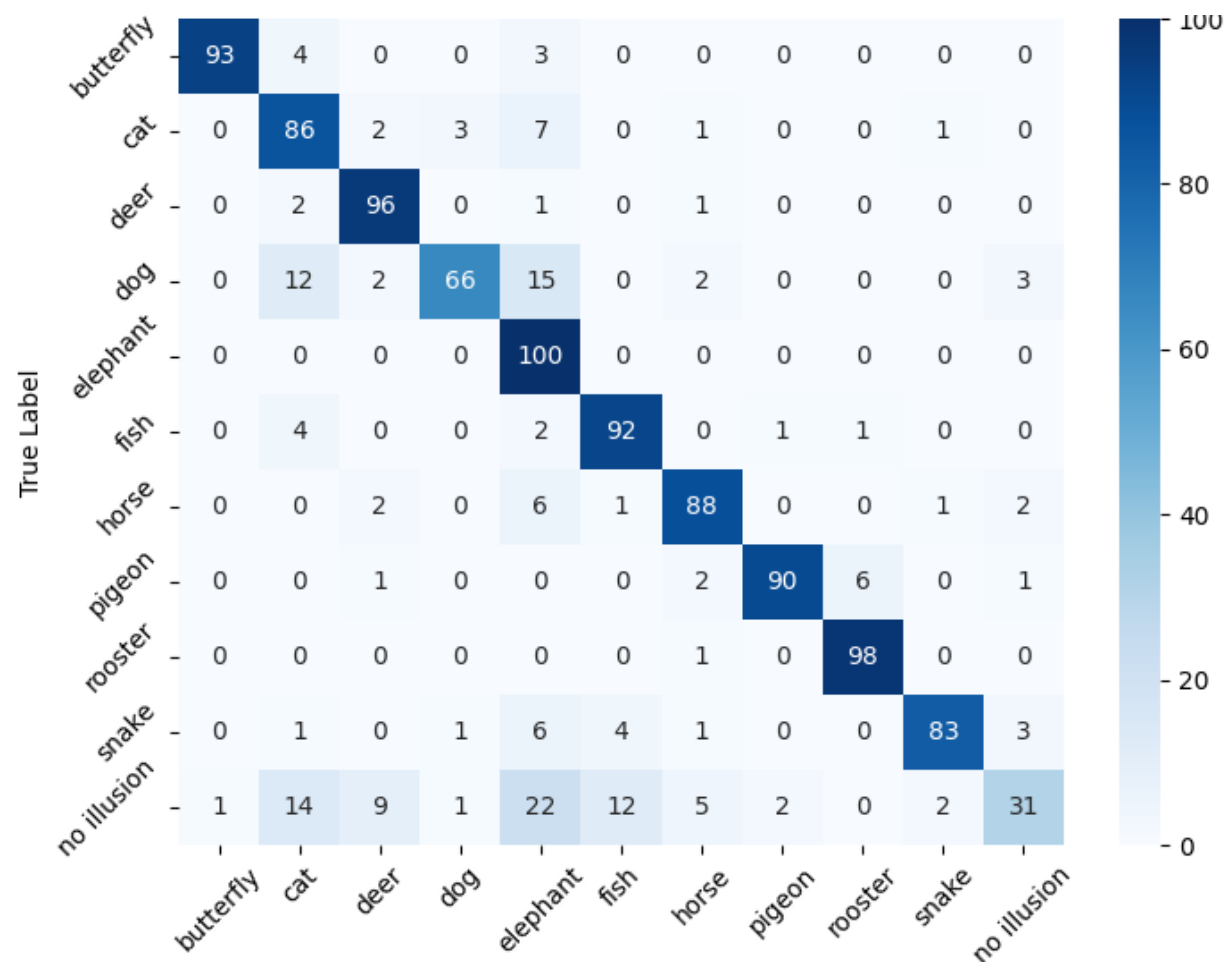
		<b>Gemini</b>	<b>GPT-4o</b>	<b>Human</b>
<b>Raw</b>	<b>WER</b>	53.73	<b>22.34</b>	-
	<b>CER</b>	56.73	<b>7.01</b>	-
<b>Illusion</b>	<b>WER</b>	90.48	<b>90.26</b>	31.94
	<b>CER</b>	175.98	<b>169.46</b>	13.32
<b>Filtered</b>	<b>WER</b>	82.73	<b>76.4</b>	-
	<b>CER</b>	<b>99.93</b>	122.18	-



# GPT-4o Confusion Matrix



# GPT-4o Confusion Matrix



# Successful examples of proposed filter for GPT-4o

(a)

Raw

Illusion

Filtered



GPT-4o:

“digit 2”

“No Illusion”

“digit 2”

Gemini:

“digit 2”

“No Illusion”

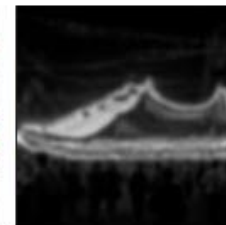
“digit 2”

(b)

Raw

Illusion

Filtered



“Sneaker”

“No Illusion”

“Sneaker”

“Sneaker”

“No Illusion”

“Sneaker”

(c)

Raw

Illusion

Filtered



GPT-4o:

“Cat”

“Elephant”

“Cat”

Gemini:

“Cat”

“No Illusion”

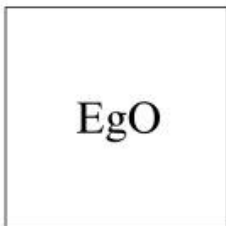
“Cat”

(d)

Raw

Illusion

Filtered



“E g O”

“No Illusion”

“EgO”

“EGO”

“No Illusion”

“P2O”

(e)

No Illusion

No Illusion Filtered



GPT-4o:

“No illusion”

“No illusion”

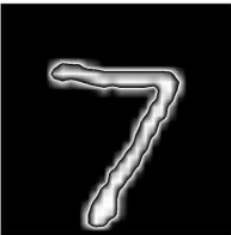


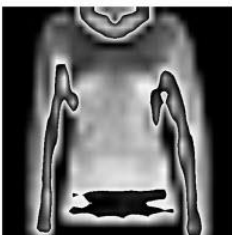

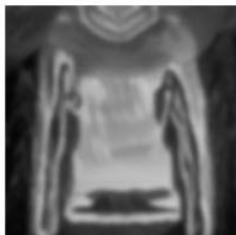






Gemini:

“No illusion”

“No illusion”

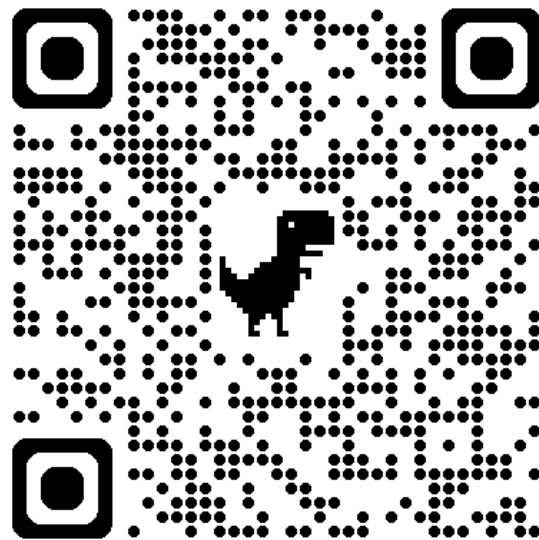


# Failure examples of proposed filter for GPT-4o

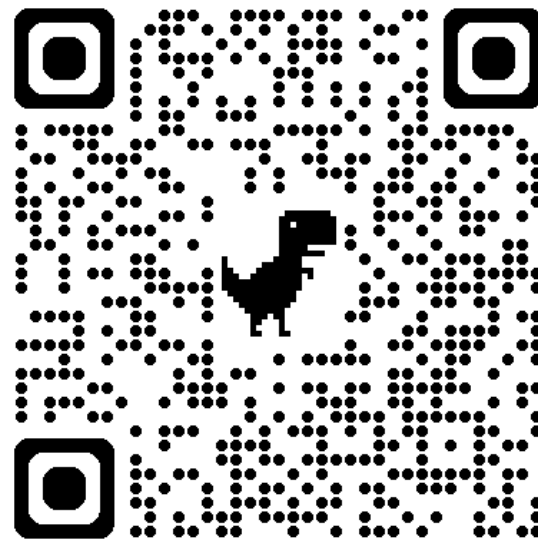
	IllusionMNIST			IllusionFashionMNIST		
	Raw	Illusion	Filtered	Raw	Illusion	Filtered
						
GPT-4o:	"digit 7"	"digit 7"	"digit 1"	"T-shirt/top"	"Pullover"	"Dress"
True Label:	"digit 7"	"digit 7"	"digit 7"	"Pullover"	"Pullover"	"Pullover"
	IllusionAnimals			IllusionChar		
	Raw	Illusion	Filtered	Raw	Illusion	Filtered
						
GPT-4o:	"dog"	"dog"	"elephant"	"WQb"	"WQb"	"No illusion"
True Label:	"dog"	"dog"	"dog"	"WQb"	"WQb"	"WQb"

Your  
Curiosity  
Fuels  
Innovation --  
Thank you

Codes:



Datasets:



Mohammadmostafa Rostamkhani

[mohammadmostafarostamkhani@gmail.com](mailto:mohammadmostafarostamkhani@gmail.com)

Codes: <https://github.com/IllusoryVQA/IllusoryVQA>

Datasets: <https://huggingface.co/VQA-Illusion>