

# Eyes Wide Shut?

## Exploring the Visual Shortcomings of Multimodal LLMs

ArXiv Preprint, January 2024 (7 Citations)

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, Saining Xie

Presented by Group 3:

Tyler VanderMate, Ashton Frias, Nicholas Gray, Wen-Kai Chen, Abhinav Kotta

# Outline

- Introduction
- Related Works
- Multimodal Visual Patterns (MMVP) Benchmark
- MMVP – Visual Language Model (MMVP-VLM) Benchmark
- Experiments
- Results
- Conclusion

# Introduction

- Is vision good enough for language?
- Many Multimodal LLMS are known to have qualitative errors.
  - LLaVA
  - InstructBLIP
  - GPT-4v
- What are these shortcomings?
- How can we measure them quantitatively?

# MLLM Visual Component

- LLaVA, InstructBLIP, GPT-4v all struggle with visual detail
- Use CLIP for their base vision encoder
- CLIP's vision encoder is likely the bottle-neck

# Examples of Incorrect VQA

- Here we use GPT-4V
  - Incorrect VQA responses
  - Poor visual grounding
  - Hallucination on incorrect responses
- How can we measure these inaccuracies?



**Q: Do you see any window in this image?**



No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.

# Related Works

## Other Related Quantitative Benchmarks

1. POPE
  - Hallucination
2. MM-Bench
  - Perception and reasoning
3. MM-Vet
  - Multiple visual language capability

# Key Contributions

1. Implemented a new benchmark that exposes CLIPs bottle neck
  - a. MMVP
  - b. MMVP-VLM
2. Mixture of Features
  - a. Additive
  - b. Interleaved

# Constructing MMVP Benchmarks

- Datasets:
    - ImageNet & LAION-Aesthetics
- 

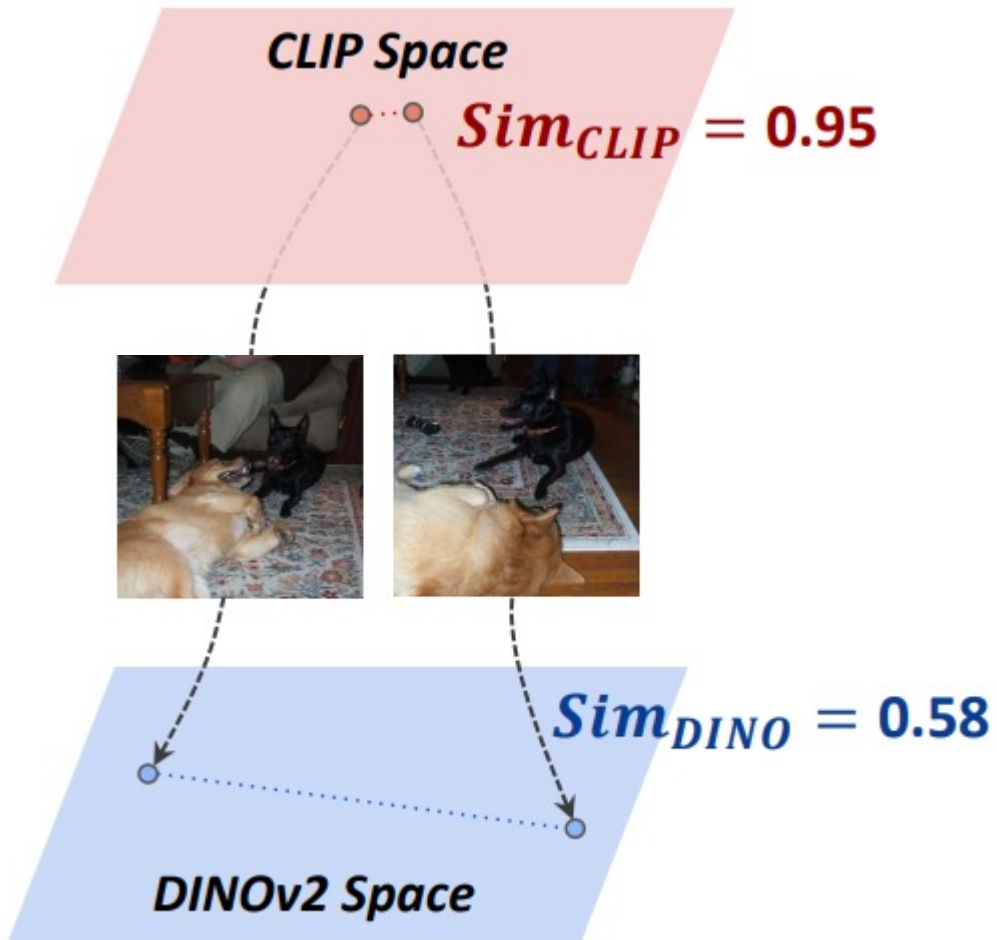
## 1. Finding CLIP-blind pairs

- CLIP Space
- DINOv2 Space

# Clip-blind pair example



# Clip-Blind Pair Search



- Pass image pairs into CLIP and DINOv2 vision encoder
- CLIP Similarity  $\geq 0.95$  ✓
- DINOv2 Similarity  $\leq 0.6$  ✓
- Clip Pair discovered

# Constructing MMVP Benchmarks

- Datasets used:
    - ImageNet & LAION-Aesthetics
- 

## 1. Finding CLIP-blind pairs

- CLIP Space
- DINOv2 Space

## 2. Human annotates differences between each image pair

# Spotting Difference Between Pairs



"The dogs head in the top image is resting on the carpet, while the dog's head in the bottom image is lying on the ground."

Where is the yellow animal's head lying in this image?  
(a) Floor (b) Carpet

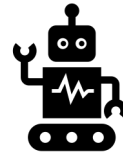
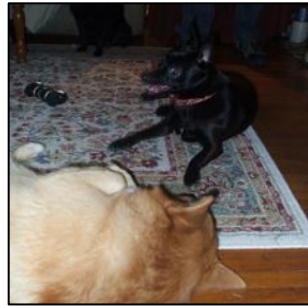
# Constructing MMVP Benchmarks

- Datasets:
    - ImageNet & LAION-Aesthetics
- 

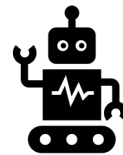
1. Finding CLIP-blind pairs
  - CLIP Space
  - DINOv2 Space
2. Human annotates differences between each image pair
3. Benchmark against various multimodal LLMs and compare results

# Benchmarking

Where is the yellow animal's head lying in this image?  
(a) Floor (b) Carpet



(a) Floor



(b) Carpet



(Correct) (Incorrect)



# MMVP Benchmark Examples

Uses close ended question-answer pairs

Is the dog facing left or right from the camera's perspective?



(a) Left (b) Right

Is the needle pointing up or down?



(a) Up (b) Down

Is the cup placed on a surface or being held by hand?



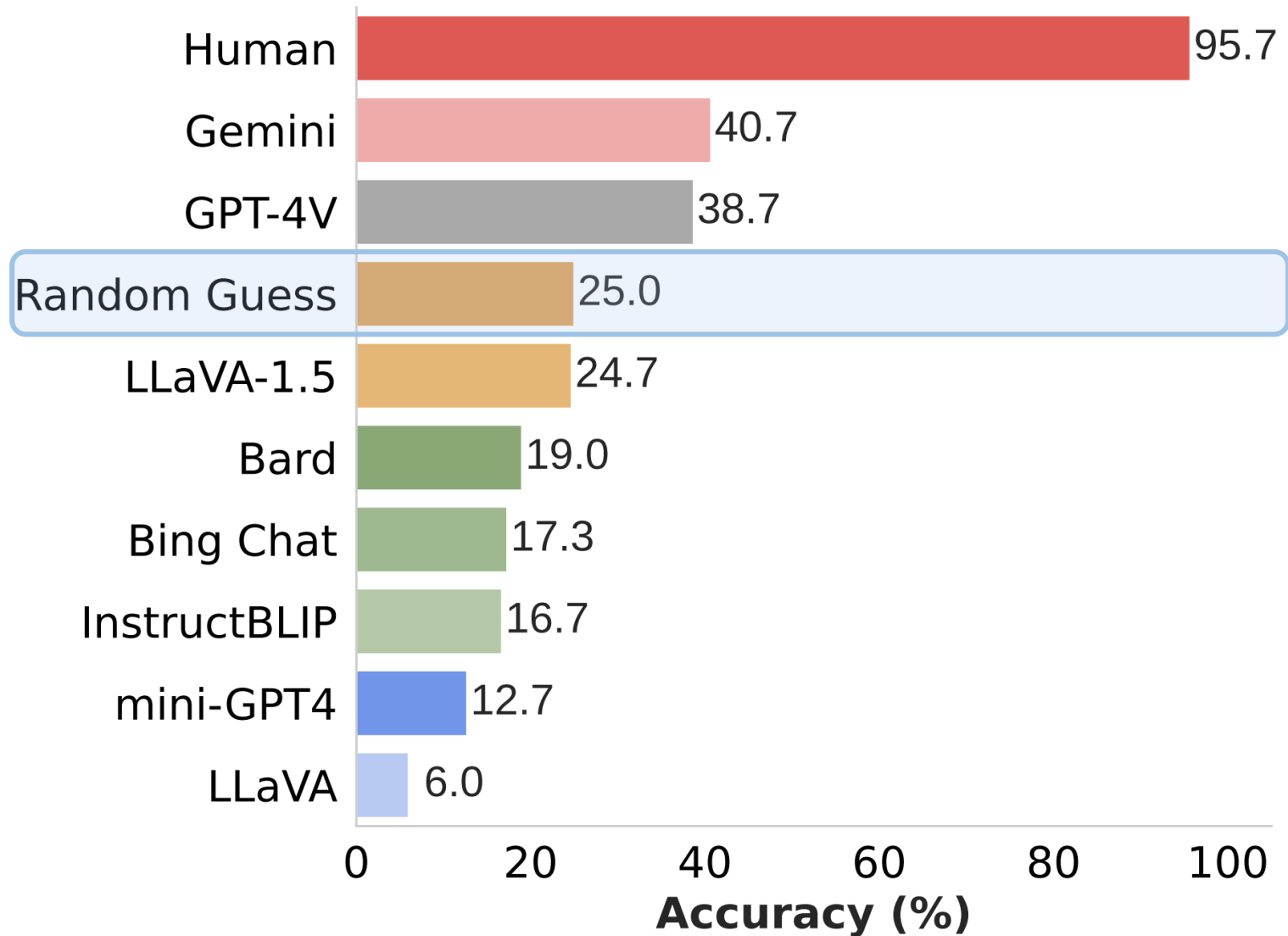
(a) Placed on a surface (b) Held by hand

	GPT-4V		(b)	(b)	✗
	Gemini Pro		(a)	(a)	✗
	LLaVA-1.5		(b)	(b)	✗
	InstructBLIP		(a)	(a)	✗

	GPT-4V		(b)	(b)	✗
	Gemini Pro		(a)	(b)	✓
	LLaVA-1.5		(a)	(a)	✗
	InstructBLIP		(a)	(a)	✗

	GPT-4V		(a)	(a)	✗
	Gemini Pro		(a)	(b)	✓
	LLaVA-1.5		(a)	(a)	✗
	InstructBLIP		(a)	(b)	✓

# MMVP Benchmark Results



# MMVP-VLM, CLIP-Blind Categorization

- Need a way to categorize the question-answer pairs
- We can use GPT-4 to find the high-level relations for us

**User**

Prompt

I am analyzing an image embedding model. Can you go through the questions and options, trying to figure out some general patterns that the embedding model struggles with? Please focus on the visual features and generalize patterns that are important to vision models

QA pairs

[MMVP Questions and Options]

# MMVP-VLM

- GPT-4 discovered 9 Visual Pattern categories in MMVP
- We use these categories to create a new balanced dataset
- VLM contains 15 Clip Blind pairs for each of the 9 categories

# MMVP-VLM Categories on EVA01 ViT-g-14

## 1. Orientation and Direction

### Orientation and Direction

a rabbit  
facing right



a rabbit  
facing left

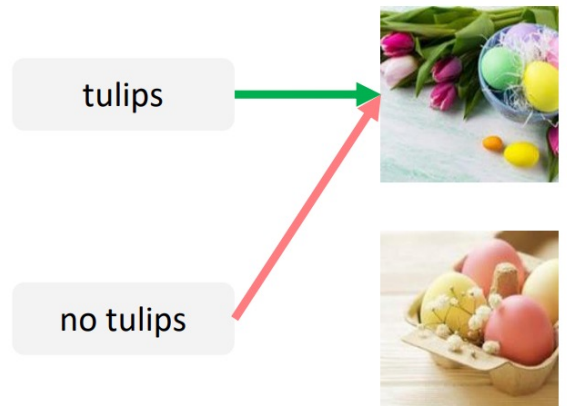


The direction  
something is facing  
or moving

# MMVP-VLM Categories on EVA01 ViT-g-14

1. Orientation and Direction
2. Presence of Specific Features

Presence of Specific Features 🔍



The existence or non-existence of certain elements or features

# MMVP-VLM Categories EVA01 ViT-g-14

1. Orientation and Direction
2. Presence of Specific Features
3. **State and Condition**

State and Condition ↻

butterfly  
with wings  
open



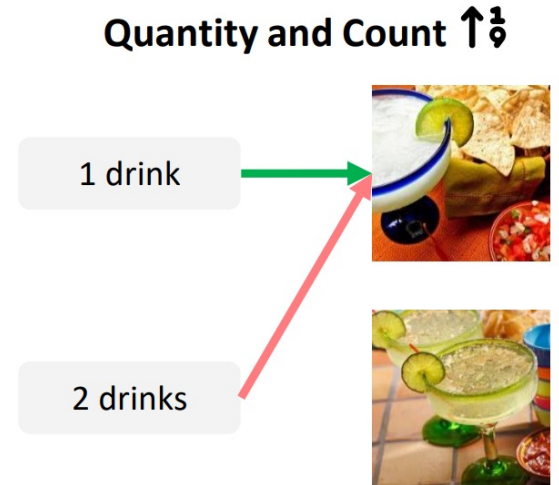
butterfly  
with wings  
closed



State or condition  
of an object

# MMVP-VLM Categories EVA01 ViT-g-14

1. Orientation and Direction
2. Presence of Specific Features
- 3. State and Condition**
4. Quantity and Count



The number of objects or features present in the image

# MMVP-VLM Categories EVA01 ViT-g-14

1. Orientation and Direction
2. Presence of Specific Features
- 3. State and Condition**
4. Quantity and Count
5. Positional and Relational Context

## Positional and Relational Context

glasses on  
the right of  
the slipper



glasses on  
the left of  
the slipper



Position and  
relationship of  
objects in relation  
to each other and  
their surroundings

# MMVP-VLM Categories EVA01 ViT-g-14

1. Orientation and Direction
2. Presence of Specific Features
- 3. State and Condition**
4. Quantity and Count
5. Positional and Relational Context
6. Structural and Physical Characteristics

## Structural Characteristics

some fruits  
cut in half



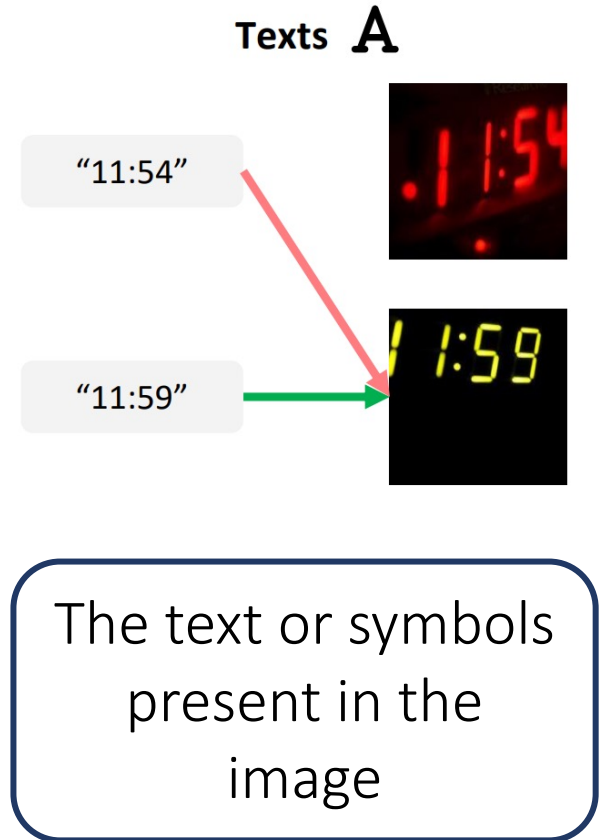
uncut fruits



Focuses on identifying and analyzing objects' physical attributes and structural features in images

# MMVP-VLM Categories on EVA01 ViT-g-14

1. Orientation and Direction
2. Presence of Specific Features
- 3. State and Condition**
4. Quantity and Count
5. Positional and Relational Context
6. Structural and Physical Characteristics
7. Texts



# MMVP-VLM Categories on EVA01 ViT-g-14

1. Orientation and Direction
2. Presence of Specific Features
3. **State and Condition**
4. Quantity and Count
5. Positional and Relational Context
6. Structural and Physical Characteristics
7. Texts
8. Viewpoint and Perspective

## Viewpoint and Perspective 📷

flowers  
seen from  
above



flowers  
seen from  
the side



The perspective  
from which the  
photo was taken

# MMVP-VLM Categories on EVA01 ViT-g-14

1. Orientation and Direction
2. Presence of Specific Features
- 3. State and Condition**
4. Quantity and Count
5. Positional and Relational Context
6. Structural and Physical Characteristics
7. Texts
8. Viewpoint and Perspective
- 9. Color and Appearance**

## Color and Appearance

light blue  
sky



dark blue  
sky











The color of certain  
objects or elements

# MMVP-VLM leads to interesting results

- The paper claims increasing model size improve 2 categories
  - Color and Appearance
  - State and Condition

# MMVP-VLM Zero-Shot Benchmarks

- Larger input image resolution show minimal improvement
- A slight advantage is observed when scaling up the network

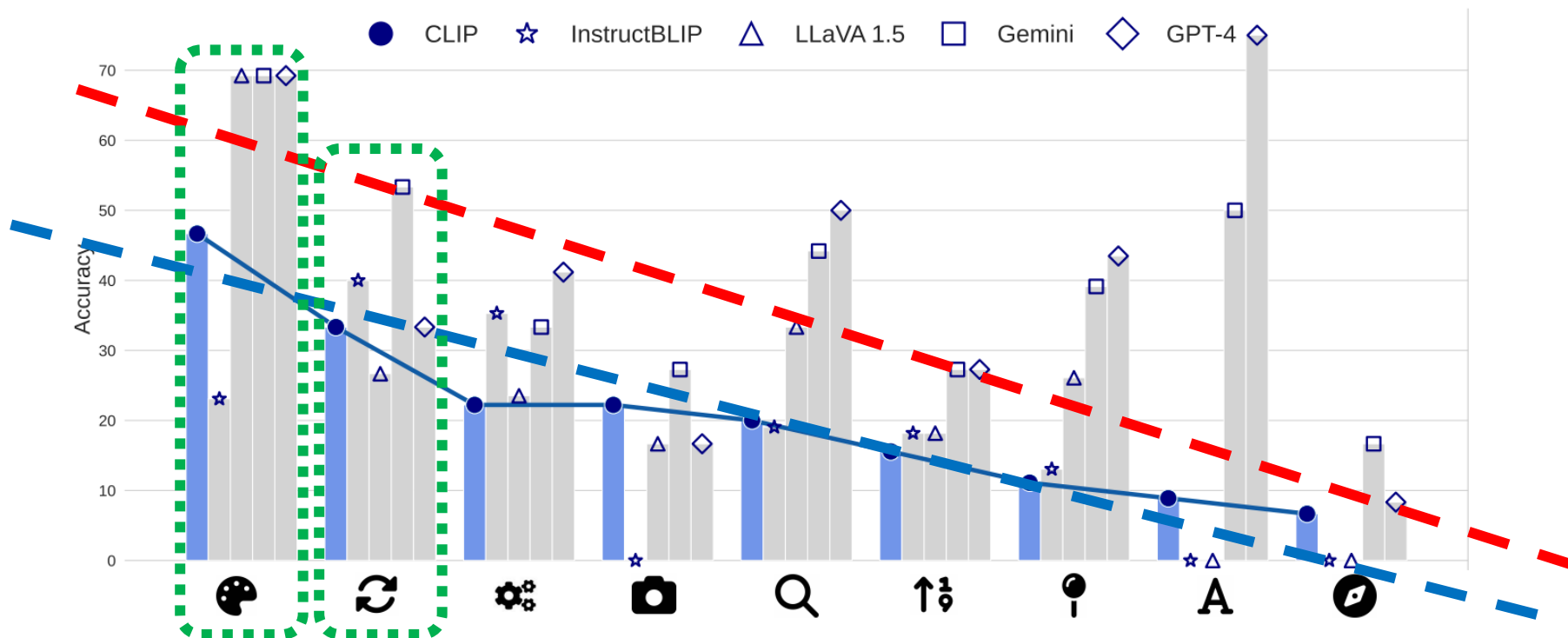
	Image Size	Params (M)	IN-1k ZeroShot								<b>A</b>		MMVP Average
OpenAI ViT-L-14 [43]	224 <sup>2</sup>	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [43]	336 <sup>2</sup>	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [66]	224 <sup>2</sup>	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [66]	384 <sup>2</sup>	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [10]	224 <sup>2</sup>	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [10]	378 <sup>2</sup>	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [62]	224 <sup>2</sup>	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [62]	224 <sup>2</sup>	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [54]	224 <sup>2</sup>	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [54]	224 <sup>2</sup>	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

# MMVP-VLM leads to interesting results

- The paper claims increasing model size improve 2 categories
  - Color and Appearance
  - State and Condition
- A correlation exists between CLIP and MLLMs on MMVP-VLM bench

# CLIP Correlations on MMVP-VLM

- If CLIP performs poorly on an MMVP-VLM class, MLLMs do too
- Order CLIP categories from best to worst performing
- State & Condition and Color & Appearance categories scale with model size
- Pearson coefficient greater than 0.7 for LLaVA 1.5 and InstructBLIP



# Pearson Correlation between CLIP & MLLMs

- Strong correlation between the errors made by the CLIP

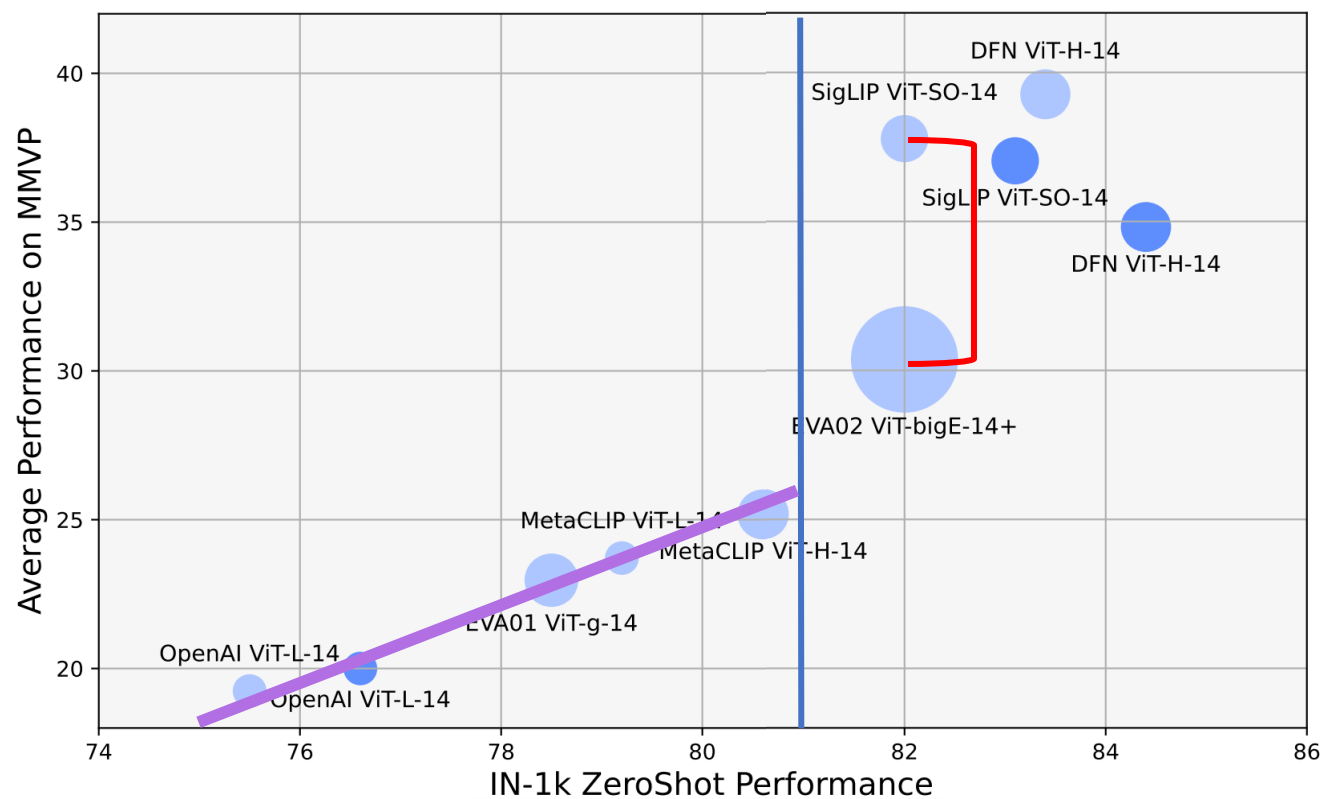
	LLaVA-1.5	InstructBLIP	Bard	Gemini	GPT-4
Correlation	0.87	0.71	0.79	0.72	0.31

# MMVP-VLM leads to interesting results

- The paper claims increasing image size and model improve 2 categories
  - Color and Appearance
  - State and Condition
- A correlation exists between CLIP and MLLMs on MMVP-VLM bench
- Shows that ImageNet accuracy can be an irrelevant measurement
  - ImageNet Zero-Shot does not necessarily measure fine-grained visual patterns
  - Shows the importance of MMVP-VLM accuracy per class
  - Emphasizes need for new metrics classifying fine-grained visual patterns

# ImageNet-1k Zero-shot vs MMVP-VLM

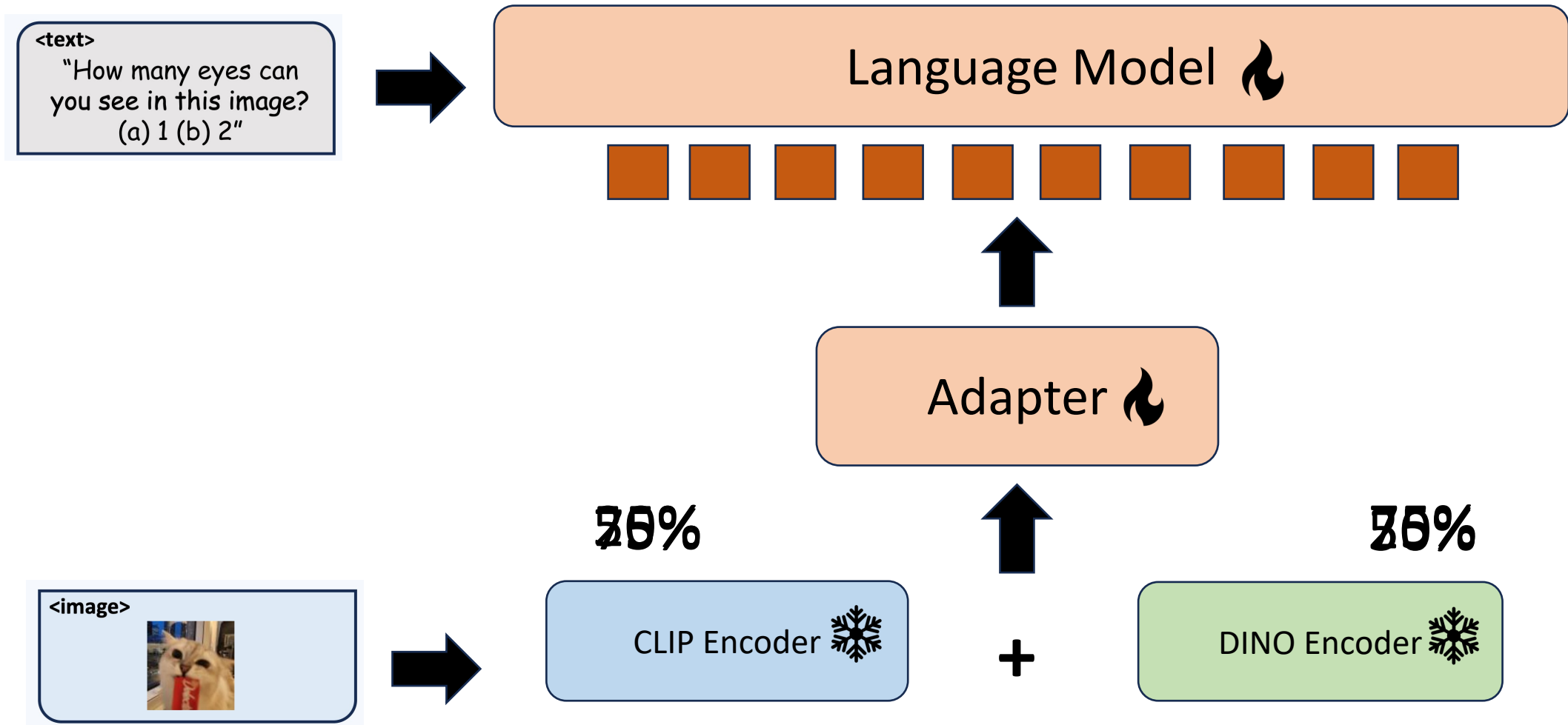
- ImageNet accuracy  $< 81\%$  indicates higher MMVP-VLM accuracy
- ImageNet accuracy  $> 81\%$  does not indicate MMVP-VLM accuracy
- ImageNet accuracy is a poor indicator of patterns within an image



# Mixture of Features

- How do we improve upon the shortcomings of CLIP blindness
- Propose two methods that utilize both CLIP and DINOv2
  - Additive MoF
  - Interleaved MoF

# Additive MoF



# Experiment

- Mixture-of-Features (MoF) for MLLM
- Setting
  - LLaVA
  - DINOv2-ViT-L-14
  - CLIP-ViT-L-14
  - 8 Nvidia A100 GPUs
  - Dataset:
    - Stage 1:
      - Both: CC595k
    - Stage 2:
      - LLaVa: LLaVA 158k
      - LLaVa-1.5: DataMix 665k

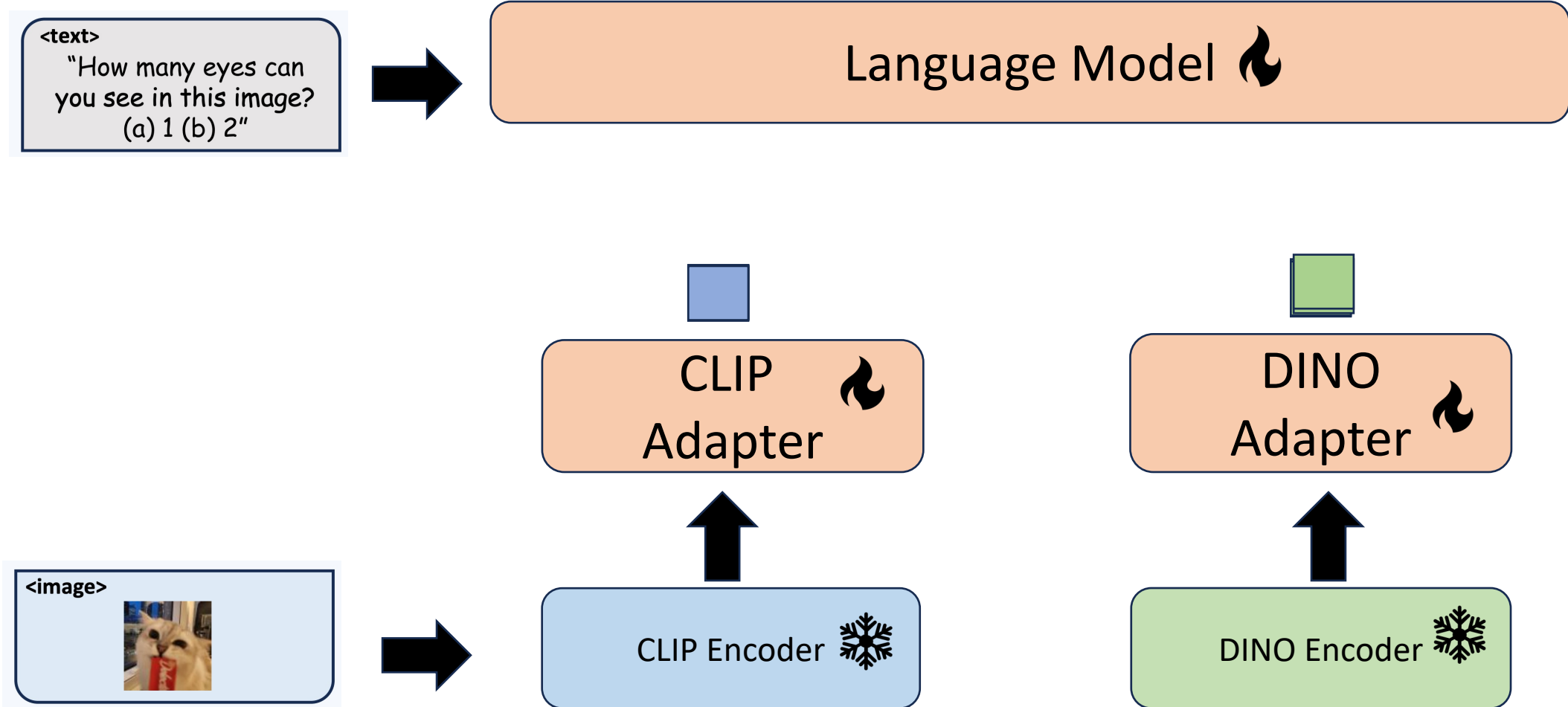
Hyperparameter	LLaVA		LLaVA-1.5	
	Stage 1	Stage 2	Stage 1	Stage 2
batch size	128	128	256	128
lr	1e-3	2e-5	2e-3	2e-5
lr schedule decay	cosine	cosine	cosine	cosine
lr warmup ratio	0.03	0.03	0.03	0.03
weight decay	0	0	0	0
epoch	1	3	1	1
optimizer		AdamW [33]		
DeepSpeed stage	2	3	2	3

# Additive MoF

- We measure against MMVP and LLaVA Bench using Additive MoF with various alpha parameters
- As DINOv2 encoding increases
  - LLaVA Bench decreases
  - Increase in MMVP until surpassing 75%

Method	SSL Ratio	MMVP	LLaVA
LLaVA	0.0	5.5	<b>81.8</b>
LLaVA + A-MoF	0.25	7.9 (+2.4)	79.4 (-2.4)
	0.5	12.0 (+6.5)	78.6 (-3.2)
	0.625	15.0 (+9.5)	76.4 (-5.4)
	0.75	18.7 (+13.2)	75.8 (-6.0)
	0.875	16.5 (+11.0)	69.3 (-12.5)
	1.0	13.4 (+7.9)	69.3 (-13.3)

# Interleaved MoF



# Interleaved MoF

- I-MoF has small changes in other metrics
- But has significant increase in MMVP accuracy

method	res	#tokens	MMVP	LLaVA	POPE
LLaVA					
LLaVA					
LLaVA + I-MoF					
LLaVA <sup>1.5</sup>					
LLaVA <sup>1.5</sup> + I-MoF					
LLaVA <sup>1.5</sup> + I-MoF					

# Limitations

- MMVP requires human annotation and contains only 300 QA pairs
- MMVP contains only 135 Clip-Blind pairs
  - Each class only contains 15 images
  - Possible ambiguity in the bucketing of classes from GPT-4
- Computationally expensive for minimal performance increase
- Paper doesn't state which Adapter they train using LLaVA
  - No ablations on using frozen LLaVA weights vs fine-tuning entire model
- As LLaVA parameter number increases
  - Interleaved MoF visual grounding improvement declines

# Conclusion

- CLIP based models have difficulty discriminating text-image visual patterns
- CLIP blindness has a detrimental cascade effect on many MLLMs
- Increasing model scale and training data can't fix most of these shortcomings
- Interleaved MoF CLIP improves text-image visual patterns
- Further research is needed to create meaningful metrics quantitatively measuring fine-grained text-image visual patterns