Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

Shengbang Tong[1] Zhuang Liu Yuexiang Zhai[2]
Yi Ma[2] Yann LeCun[1] Saining Xie[1]

[1]New York University [2]UC Berkeley

CVPR 2024 Accepted Paper

Tong, Shengbang, et al. "Eyes wide shut? exploring the visual shortcomings of multimodal llms." arXiv preprint arXiv:2401.06209 (2024).

Seil Kang
Apr. 17. 2024

Teaser

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

The dog is **facing to the right** from the camera's perspective. You can tell by the position of its tail and the orientation of its body.

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.

Q: Is the door of the truck cab open?

No, the door of the truck is **not open** in the image provided.

Q: In this image, how many eyes can you see on the animal?

The image shows **one eye** of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.

Q: How many wheels can you see in the image?

In the image provided, I can see **two wheels** on the visible side of the car.

Q: Does the keyboard have a backlight?

Based on the image, it appears that the keyboard **does have a backlight**. This is indicated by the illumination seen around the keys, which suggests that there is a light source beneath them to enhance visibility in low-light conditions.

Q: From the camera's perspective, is the piano's back panel on the left or right side?

From the camera's perspective, the piano's back panel is **on the left side**. The keys visible in the foreground suggest that we looking at the piano from front-right side, thus placing the back panel on the left from this angle.

Q: Is the edge of the heart in the picture white or dark-colored?

The image you've provided is small and of low resolution, but it appears that the hearts have a **dark-colored edge** or outline. This outline varies in visibility, likely due to the quality of the image.

Q: Is the school bus parked facing the camera or away from the camera?

The school bus is parked **facing away from the camera**. You can tell because the rear lights and the back of the exterior rearview mirror are visible. The text "SCHOOL BUS" is also reversed, which is typically seen on the front of a bus, designed so that it reads correctly in a rearview mirror.

Q: Are the butterfly's feet visible?

The butterfly's feet are **not visible** in this image. The wings and part of the body of the butterfly are in view, but the feet are either hidden beneath the body or out of the frame of the picture. Butterflies have delicate, slender feet which are often not the most noticeable feature when they are perched, especially from the angle shown in this image.

Teaser

Is vision good enough for language?

Given problem

- Visual Shortcomings
  - Visual component typically depends only on the instance-level contrastive language-image pre-training (CLIP).
- They discover "MLLMs face challenges in nine prevalent patterns."
  - Orientation and Direction, Presence of Specific Features, State and Condition, Quantity and Count, Positional and Relational Context, Structural Characteristics, Texts, Viewpoint and Perspective, Color and Appearance

Contributions

- Exploring the gap between the visual embedding space of CLIP and vision-only self-supervised learning.
  - DINOv2
- Construction of MMVP benchmark
  - Multimodal Visual Patterns
- Mixture of Features (MoF)
  - Enhancing prior work's visual grounding capabilites
  - Linealy mix CLIP and DINOv2
  - Spatially mix visual tokens from both CLIP and DINOv2

MMVP Benchmark

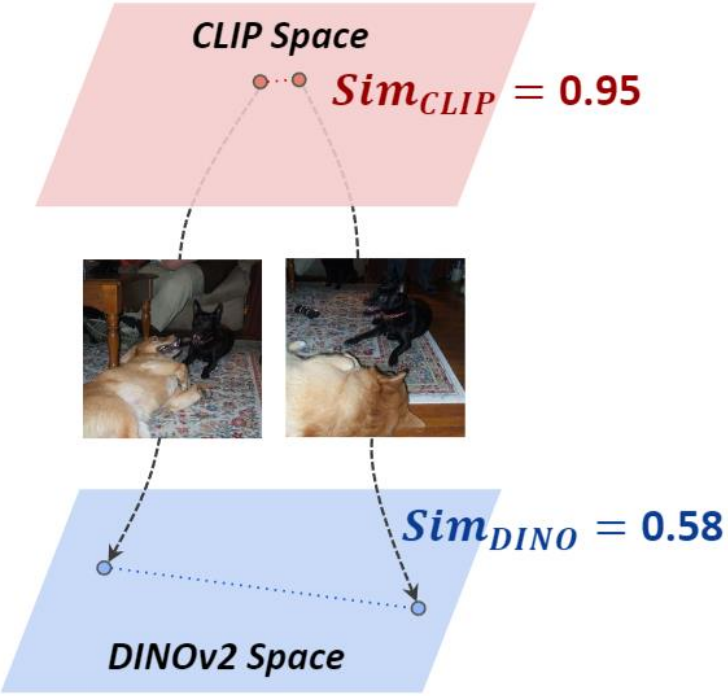
Seil Kang
Apr. 17. 2024

The MMVP Benchmark
- Finding CLIP-blind pairs

Step 1

Finding CLIP-blind pairs.

Discover image pairs that are proximate in CLIP feature space but distant in DINOv2 feature space.



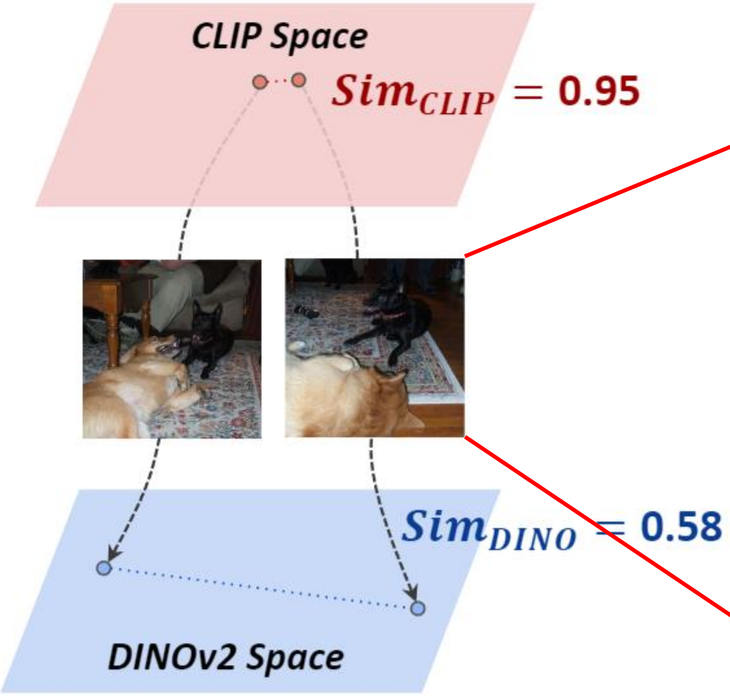
- The underlying principle is simple: if two images, despite having stark visual differences, are encoded similarly by the CLIP vision encoder, then one of them is likely encoded ambiguously.
- Self-supervised model trained without any language guidance.
  - DINOv2
- Collecting Image
  - ImageNet, LAION-Aesthetics
  - Cosine similarity ≥ 0.95 in CLIP but ≤ 0.6 in DINOv2

The MMVP Benchmark
- Finding CLIP-blind pairs

Step 1

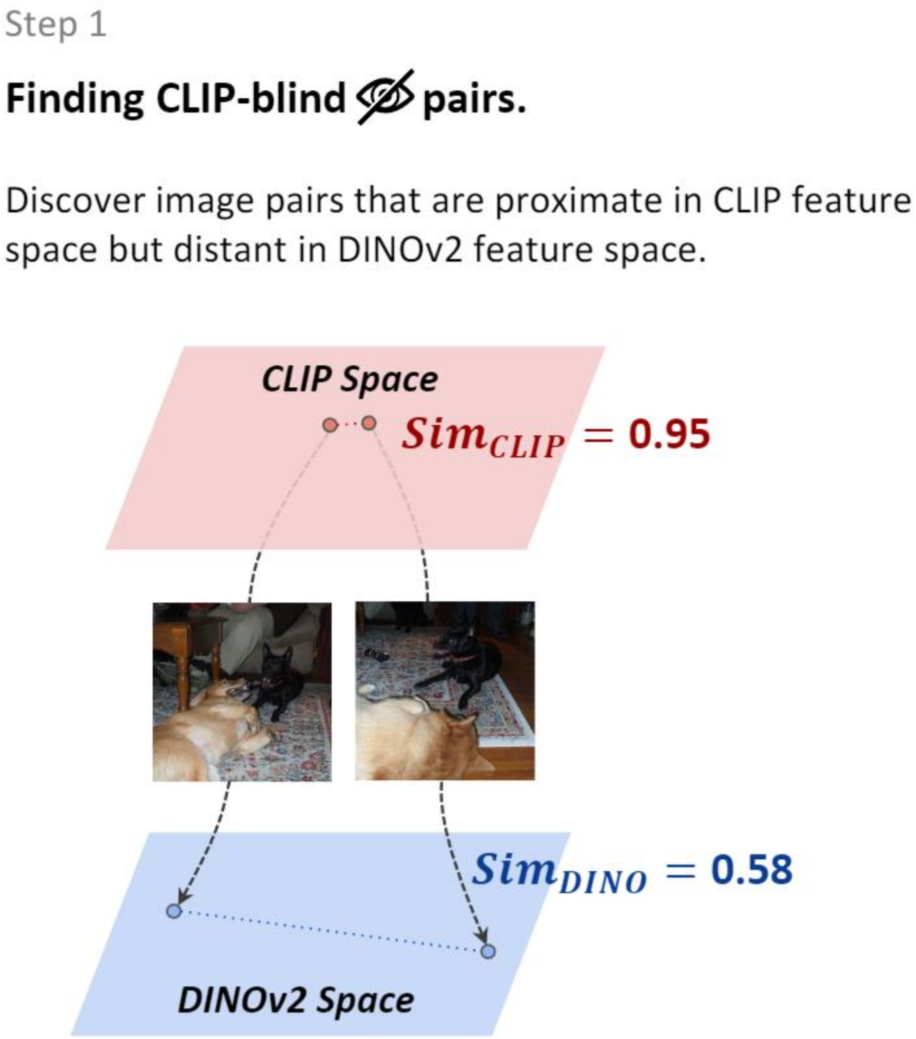
Finding CLIP-blind pairs.

Discover image pairs that are proximate in CLIP feature space but distant in DINOv2 feature space.



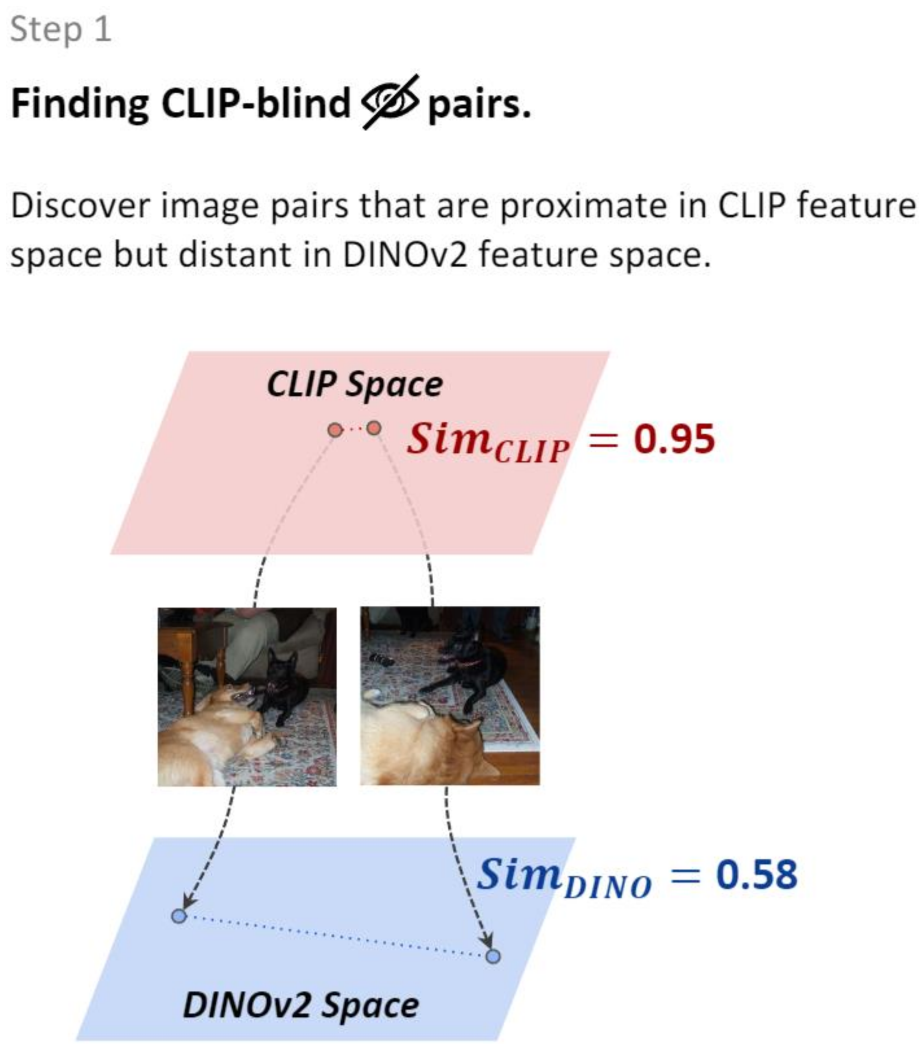
- <u>The underlying principle is simple</u>: if two images, despite having stark visual differences, are encoded similarly by the CLIP vision encoder, then one of them is likely encoded ambiguously.

The MMVP Benchmark
- Finding CLIP-blind pairs



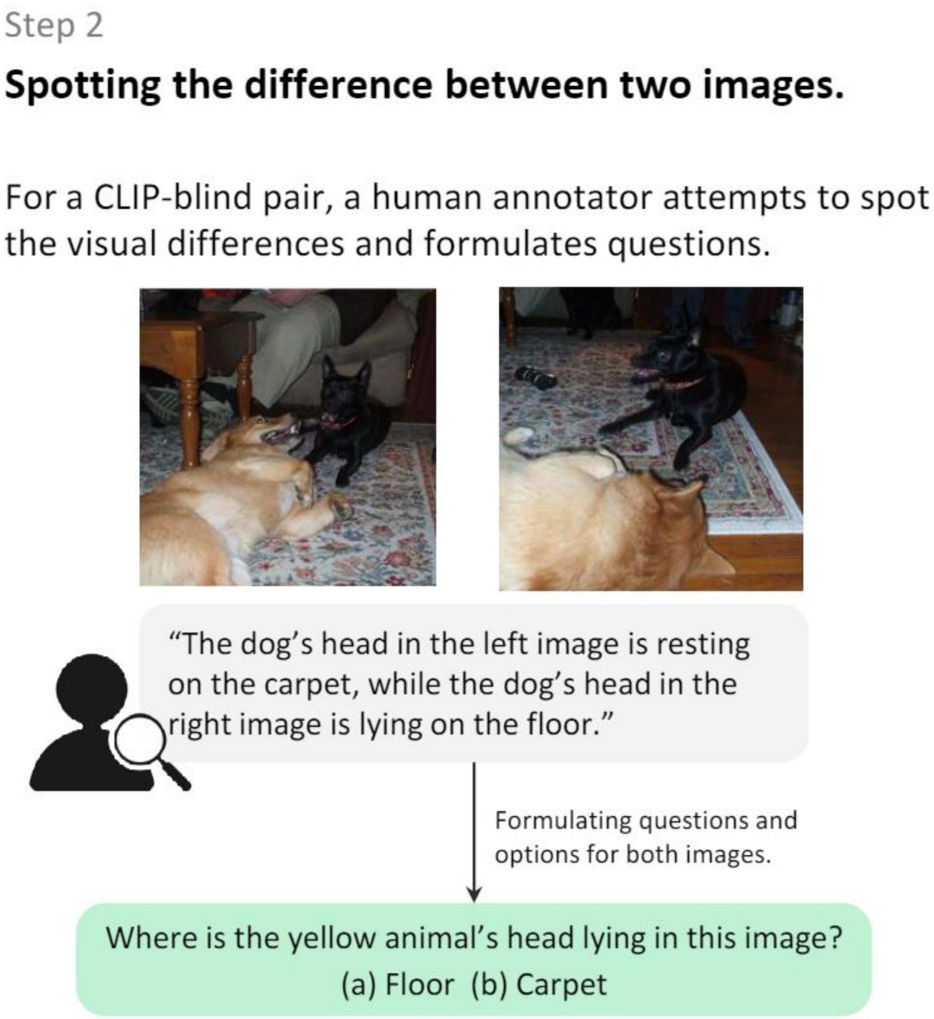
- The underlying principle is simple: if two images, despite having stark visual differences, are encoded similarly by the CLIP vision encoder, then one of them is likely encoded ambiguously.
- Self-supervised model trained without any language guidance.
  - DINOv2
- Collecting Image
  - ImageNet, LAION-Aesthetics
  - Cosine similarity $\geq$ 0.95 in CLIP but $\leq$ 0.6 in DINOv2

The MMVP Benchmark
- Finding CLIP-blind pairs



- The underlying principle is simple: if two images, despite having stark visual differences, are encoded similarly by the CLIP vision encoder, then one of them is likely encoded ambiguously.
- Self-supervised model trained without any language guidance.
  - DINOv2
- Collecting Image
  - ImageNet, LAION-Aesthetics
  - Cosine similarity ≥ 0.95 in CLIP but ≤ 0.6 in DINOv2

The MMVP Benchmark
- Finding CLIP-blind pairs



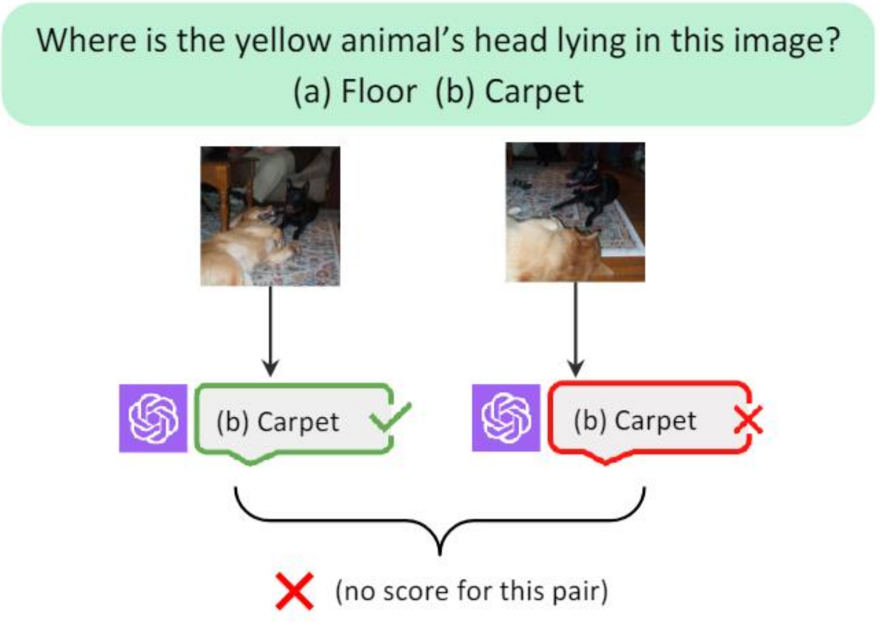
- 150 pairs with 300 questions
- The primary goal is to determine whether MLLM models would fail when posed with these seemingly basic questions and overlook critical visual details.

The MMVP Benchmark
- Benchmarking

Step 3

Benchmarking multimodal LLMs.

Evaluate multimodal LLMs using a CLIP-blind image pair and its associated question.

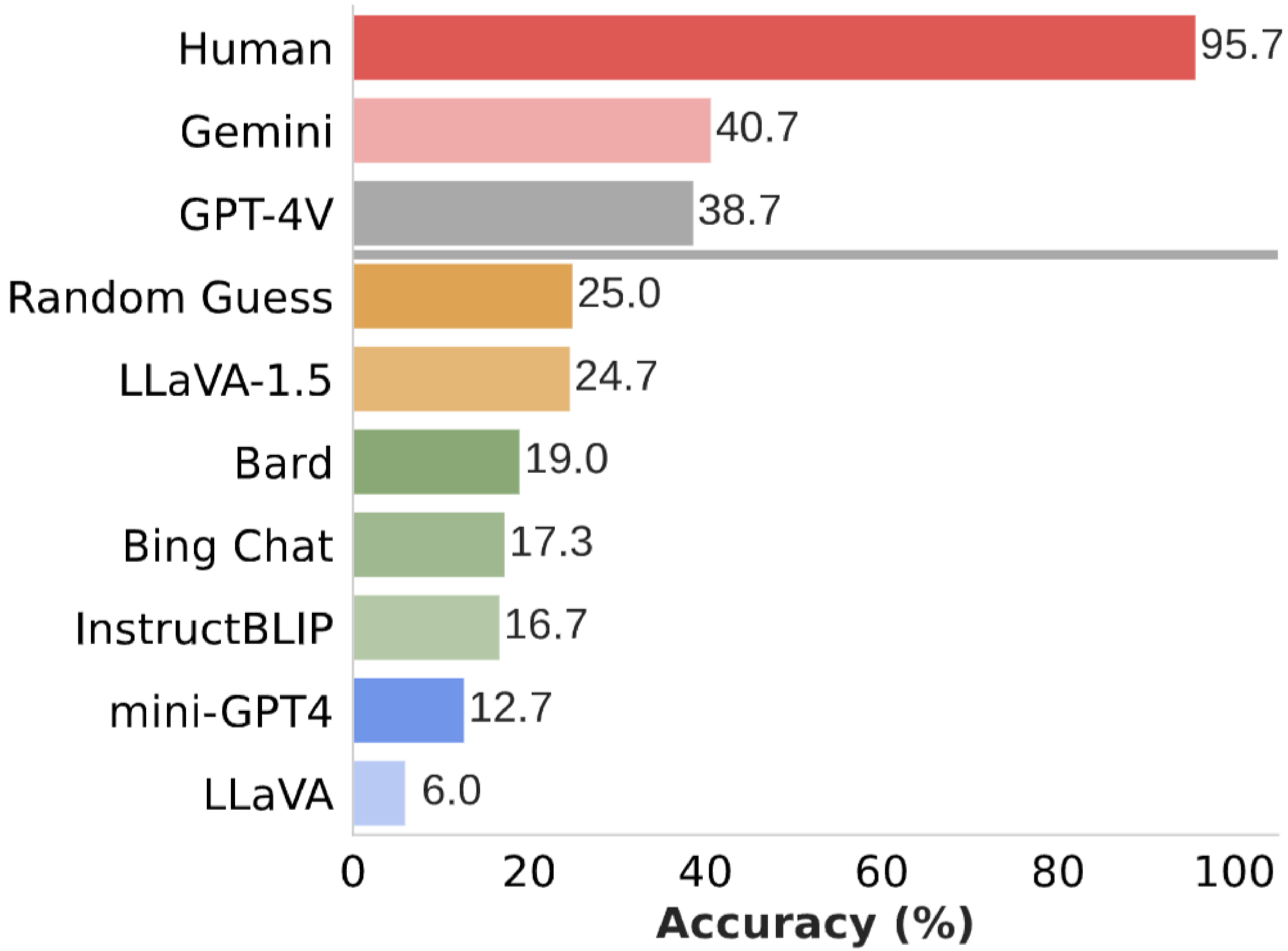


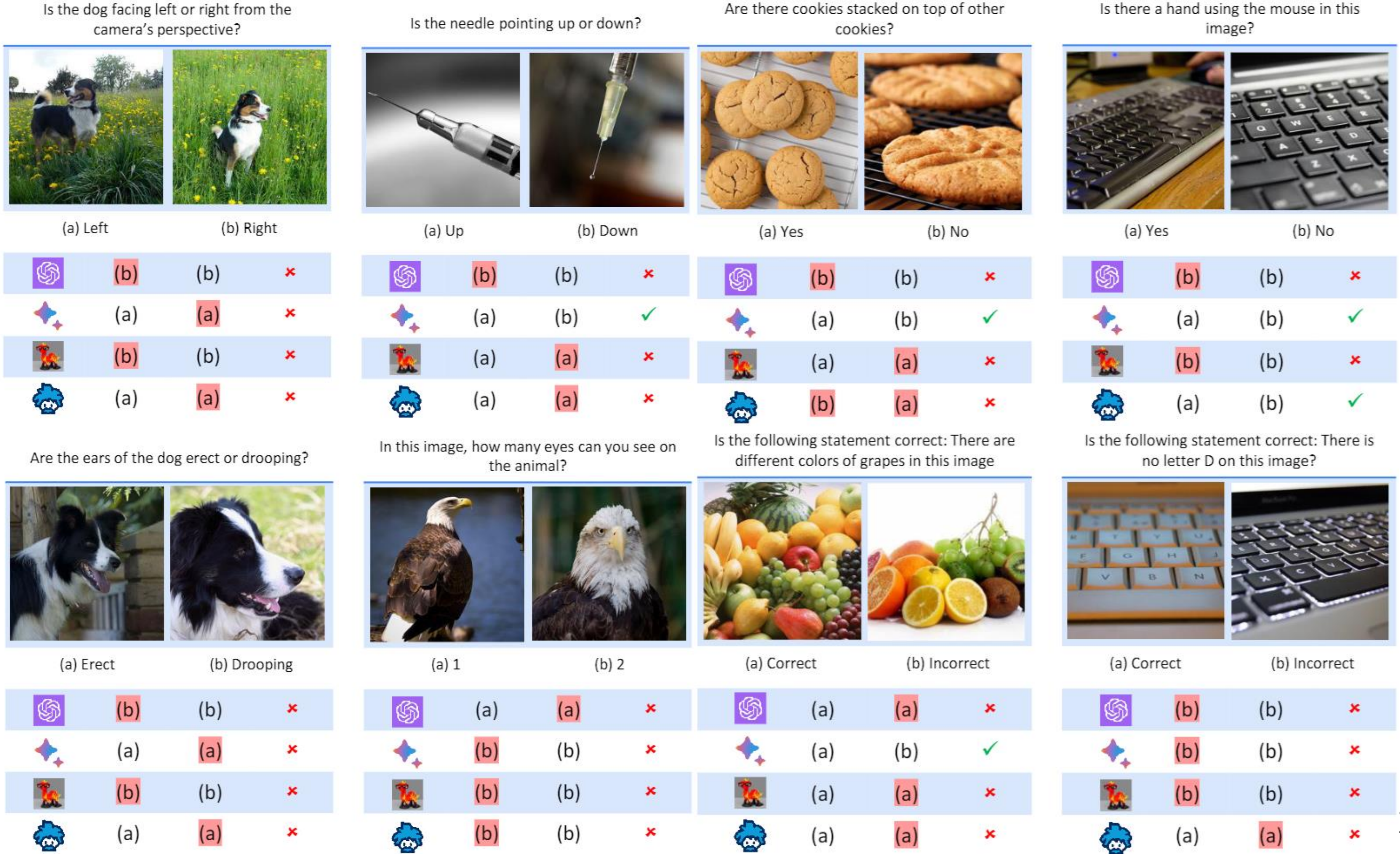
The model receives a score only when ***both*** predictions for the CLIP-blind pair are correct.

- Assess the questions on *SOTA* open-source models and closed-source models.
  - LLaVA-1.5, InstructBLIP, Mini-GPT4
  - GPT-4V(ision), Gemini, Bard
- Also, User study.
- If both the questions associated with the pair are answered accurately → a pair of images to be correctly answered.

The MMVP Benchmark

- Naïve Results of benchmarking




Is the dog facing left or right from the camera's perspective?
(a) Left
(b) Right
(b) (b)
(a) (a)
(b) (b)
(a) (a)
Is the needle pointing up or down?
(a) Up
(b) Down
(b) (b)
(a) (b)
(a) (a)
(a) (a)
Are there cookies stacked on top of other cookies?
(a) Yes
(b) No
(b) (b)
(a) (b)
(a) (a)
(b) (a)
Is there a hand using the mouse in this image?
(a) Yes
(b) No
(b) (b)
(a) (b)
(b) (b)
(a) (b)
Are the ears of the dog erect or drooping?
(a) Erect
(b) Drooping
(b) (b)
(a) (a)
(b) (b)
(a) (a)
In this image, how many eyes can you see on the animal?
(a) 1
(b) 2
(a) (a)
(b) (b)
(b) (b)
(b) (b)
Is the following statement correct: There are different colors of grapes in this image
(a) Correct
(b) Incorrect
(a) (a)
(a) (b)
(a) (a)
(a) (a)
Is the following statement correct: There is no letter D on this image?
(a) Correct
(b) Incorrect
(b) (b)
(b) (b)
(b) (b)
(a) (a)

The MMVP Benchmark
- More specifically, identify 9 systematic visual patterns.

User
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
[MMVP Questions and Options]

We identify 9 visual patterns:
- Orientation and Direction
- Presence of Specific Features
- State and Condition
- Quantity and Count
- Positional and Relational Context
- Color and Appearance
- Structural and Physical Characteristics
- Text
- Viewpoint and Perspective

- Identifying the systematic visual patterns
- Using GPT-4

D. Visual Patterns for CLIP

Here, we provide the full description of visual patterns that pose challenges to all CLIP-based models.

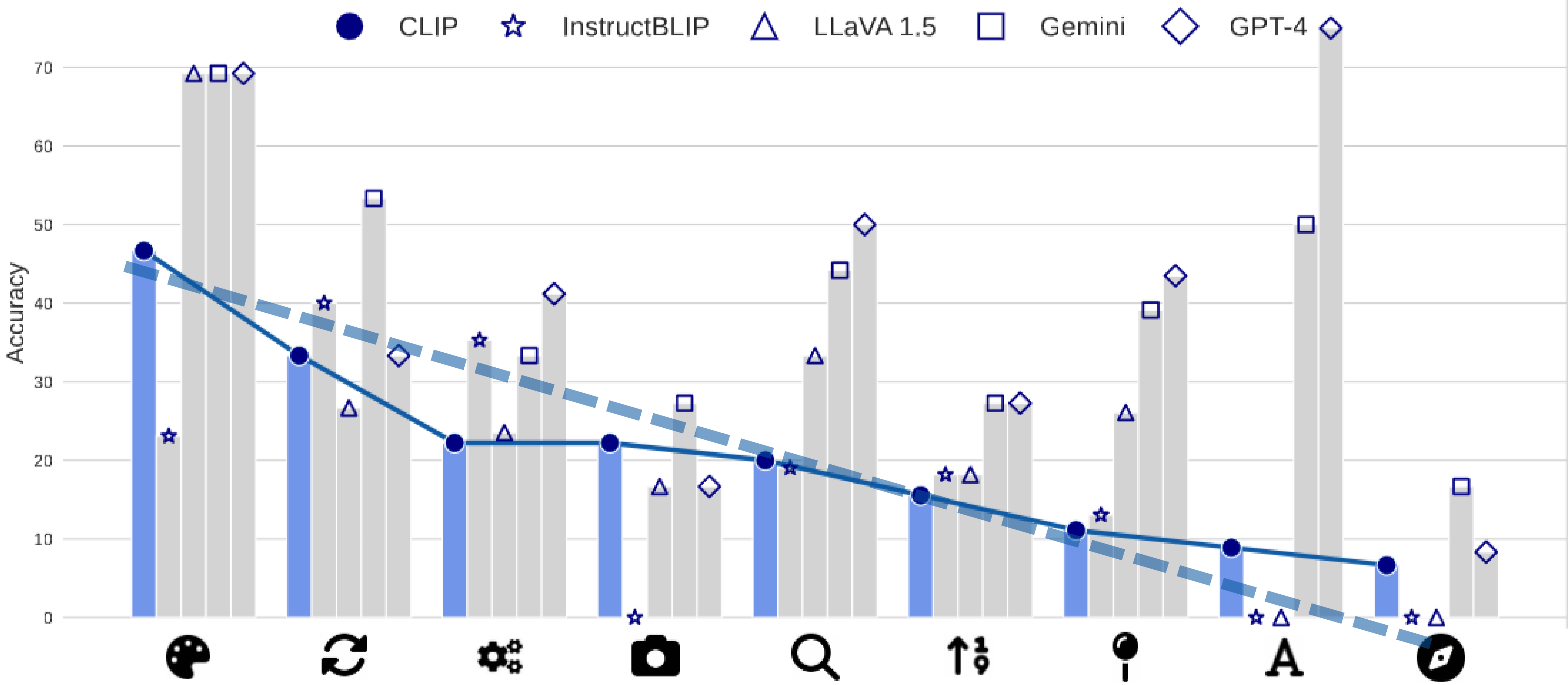
- **Orientation and Direction**: Questions about the direction something is facing or moving, such as the direction the dog or duck is facing, or the orientation of the school bus.
- **Presence of Specific Features**: Questions that focus on the existence or non-existence of certain elements or features in the image.
- **State and Condition**: Questions that pertain to the state or condition of an object, such as whether a flag is blowing in the wind or if the ground is wet.
- **Quantity and Count**: Questions about the number of objects or features present in the image.
- **Positional and Relational Context**: This aspect refers to the model's ability to understand the position and relationship of objects or elements within an image in relation to each other and their surroundings.
- **Color and Appearance**: Questions regarding the color of certain objects or elements.
- **Structural and Physical Characteristics**: This category involves the model's ability to identify and analyze the physical attributes and structural features of objects in an image.
- **Text**: Questions related to text or symbols present in the image.
- **Viewpoint and Perspective**: Questions concerning the perspective from which the photo was taken.

The MMVP Benchmark
- Results (on CLIP based models)

	Image Size	Params (M)	IN-1k ZeroShot	Orientation and Direction	Presence of Specific Features	State and Condition	Quantity and Count	Positional and Relational Context	Color and Appearance	Structural Characteristics	Texts	Viewpoint and Perspective	MMVP Average
OpenAI ViT-L-14 [43]	224^2	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]	336^2	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]	224^2	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]	384^2	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]	224^2	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]	378^2	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]	224^2	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]	224^2	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]	224^2	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]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

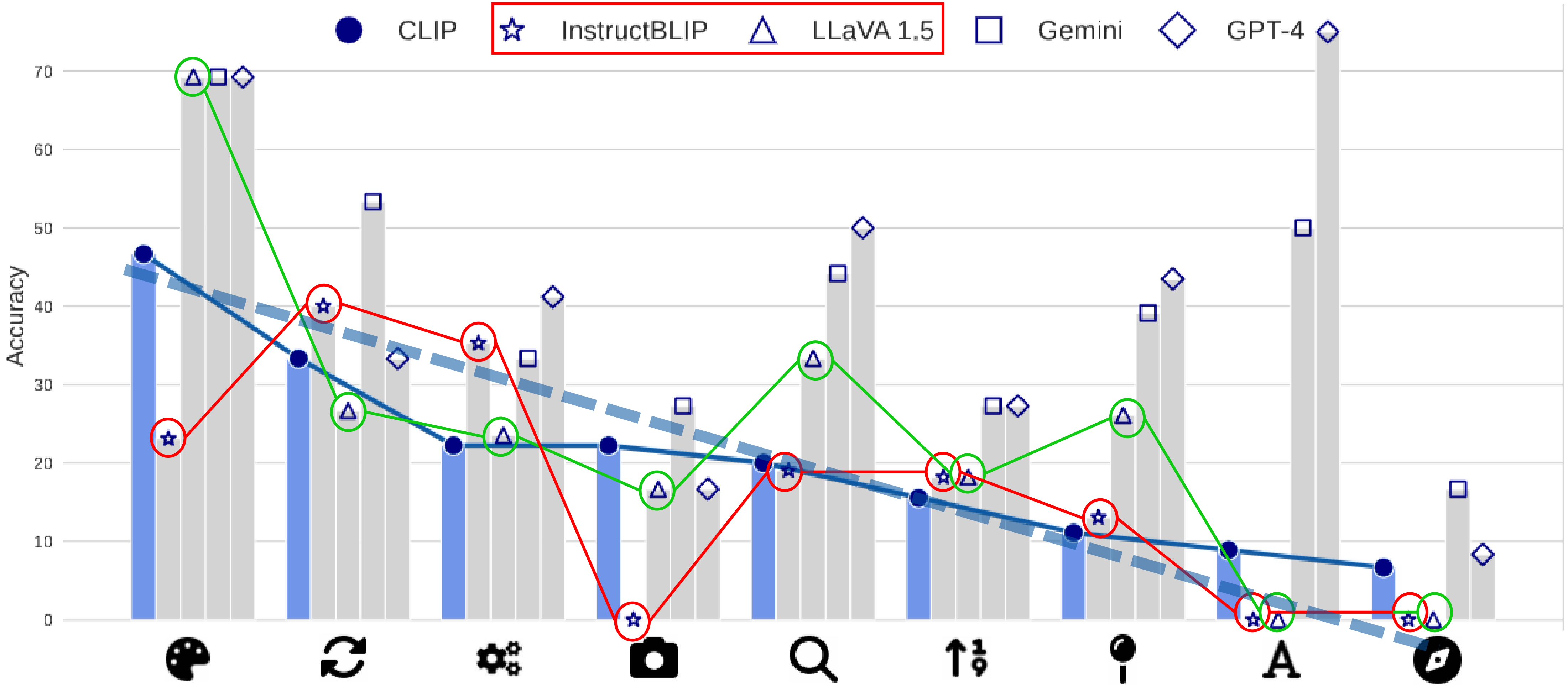
[43] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.
[66] Zhai, Xiaohua, et al. "Sigmoid loss for language image pre-training." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[10] Fang, Alex, et al. "Data filtering networks." arXiv preprint arXiv:2309.17425 (2023).
[62] Xu, Hu, et al. "Demystifying clip data." arXiv preprint arXiv:2309.16671 (2023).
[54] Sun, Quan, et al. "Eva-clip: Improved training techniques for clip at scale." arXiv preprint arXiv:2303.15389 (2023).

The MMVP Benchmark
- Results (on MLLMs and VLMs)



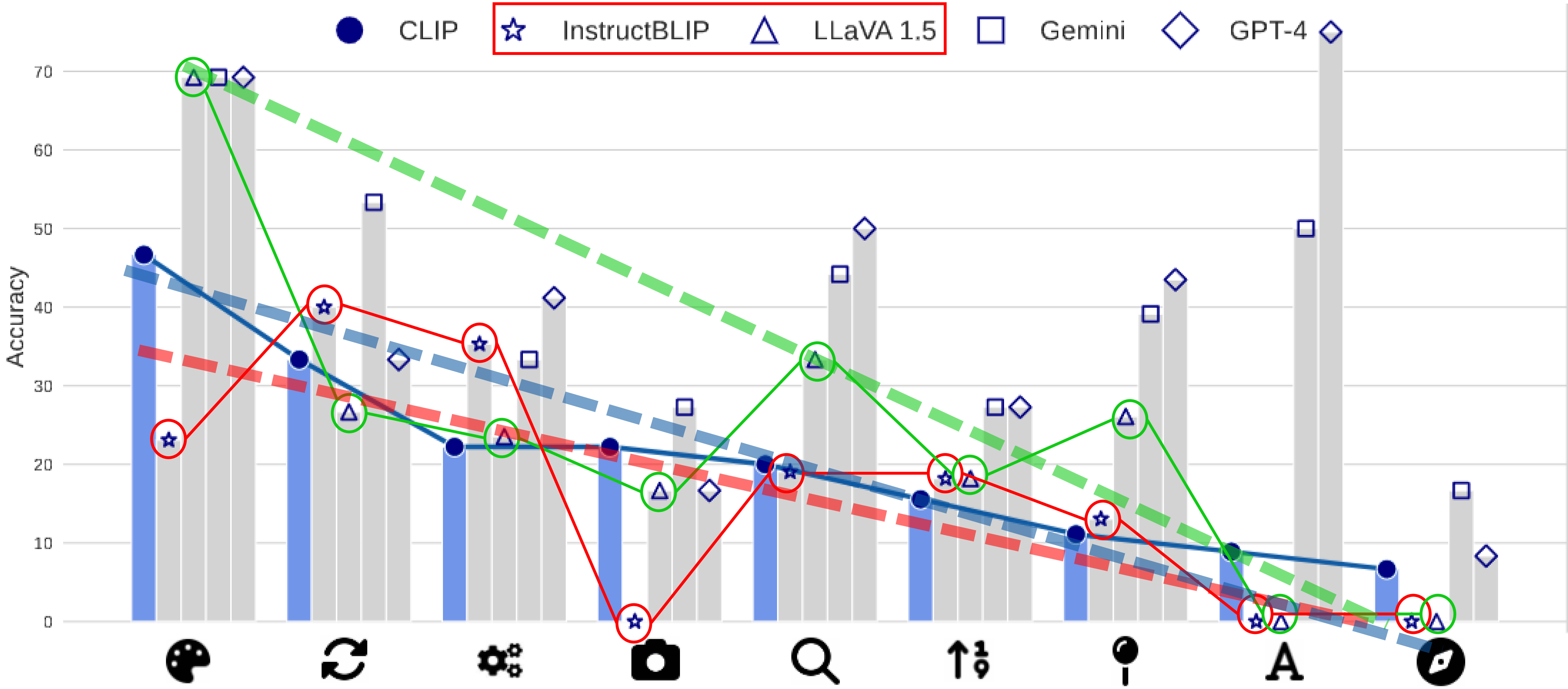
The MMVP Benchmark

- Results (on MLLMs and VLMs)



The MMVP Benchmark

- Results (on MLLMs and VLMs)



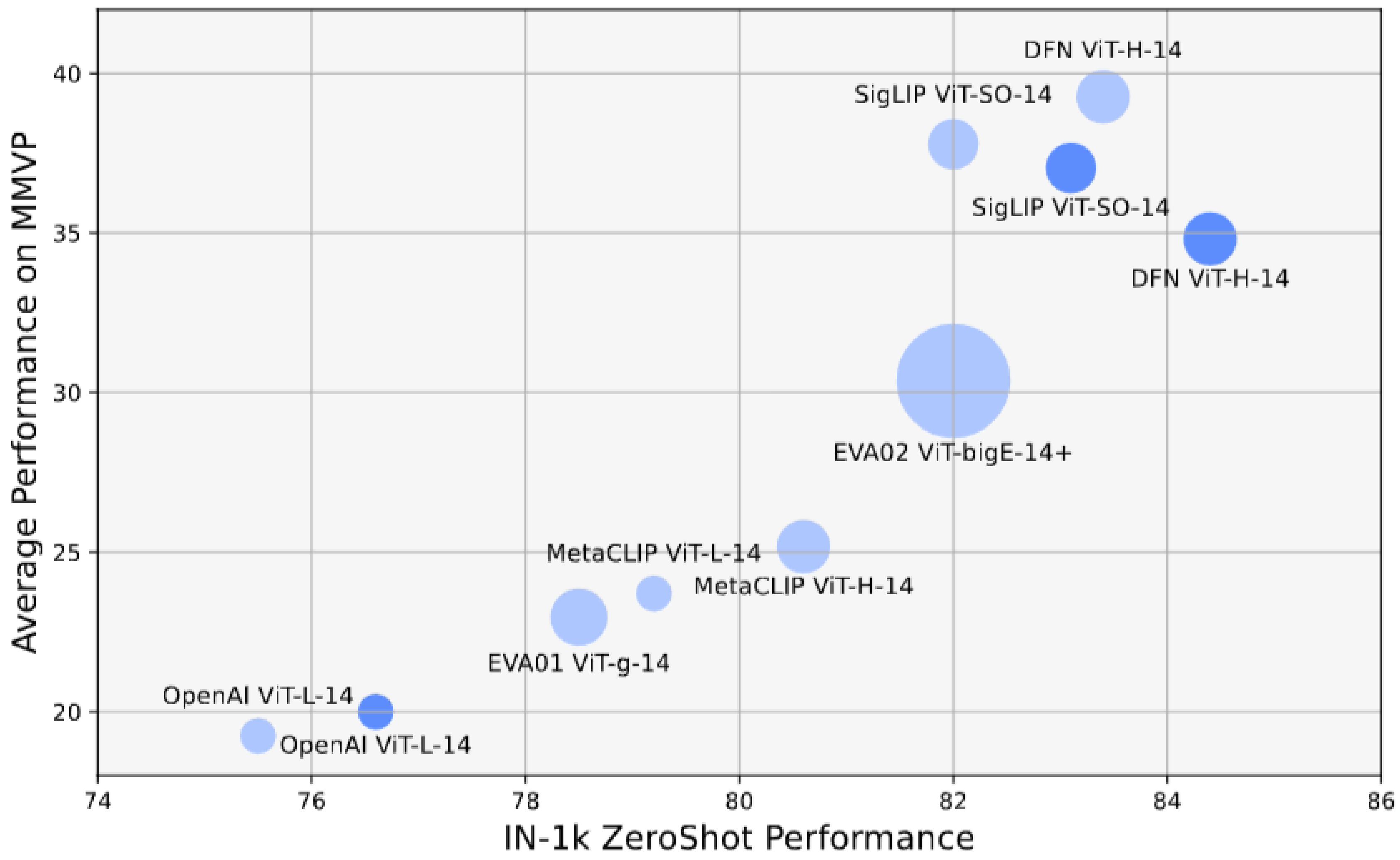
Seil Kang
Apr. 17. 2024

The MMVP Benchmark
- Results (Pearson Correlation Coefficient btw CLIP and MLLMs)

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

Table 5. Pearson Correlation between the CLIP model and MLLMs. Open-source models that explicitly use CLIP-based models are highlighted in gray.

The MMVP Benchmark
- Results (Correlation w/ ZeroShot Performances)

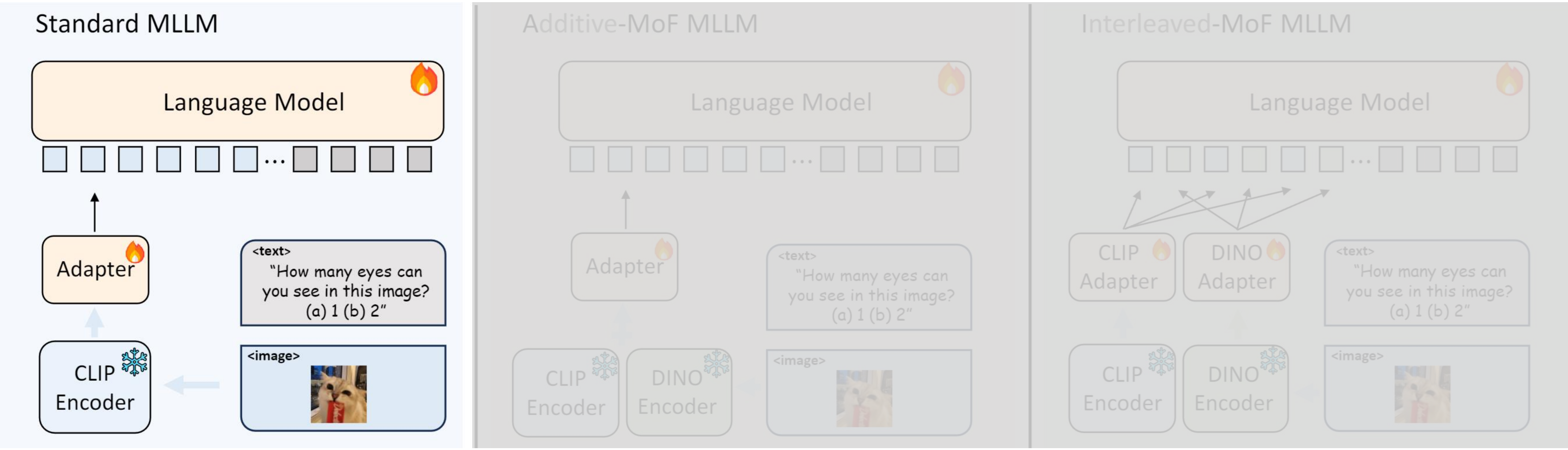


Mixture of Features

Seil Kang
Apr. 17. 2024

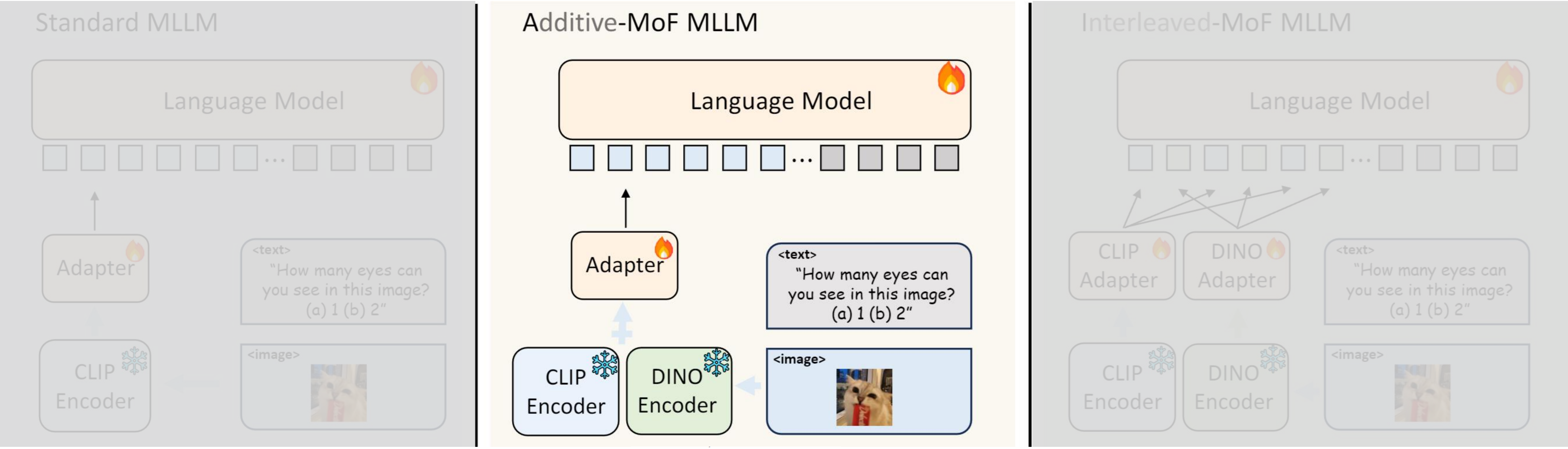
Mixture of Features

- Standard MLLM



Mixture of Features

- Additive MoF



$\{0.00, 0.25, 0.50, 0.75, 1.00\}$.
$\{0.625, 0.875\}$.

Mixture of Features

- Experiment Details

- 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 Stage 1	LLaVA Stage 2	LLaVA-1.5 Stage 1	LLaVA-1.5 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

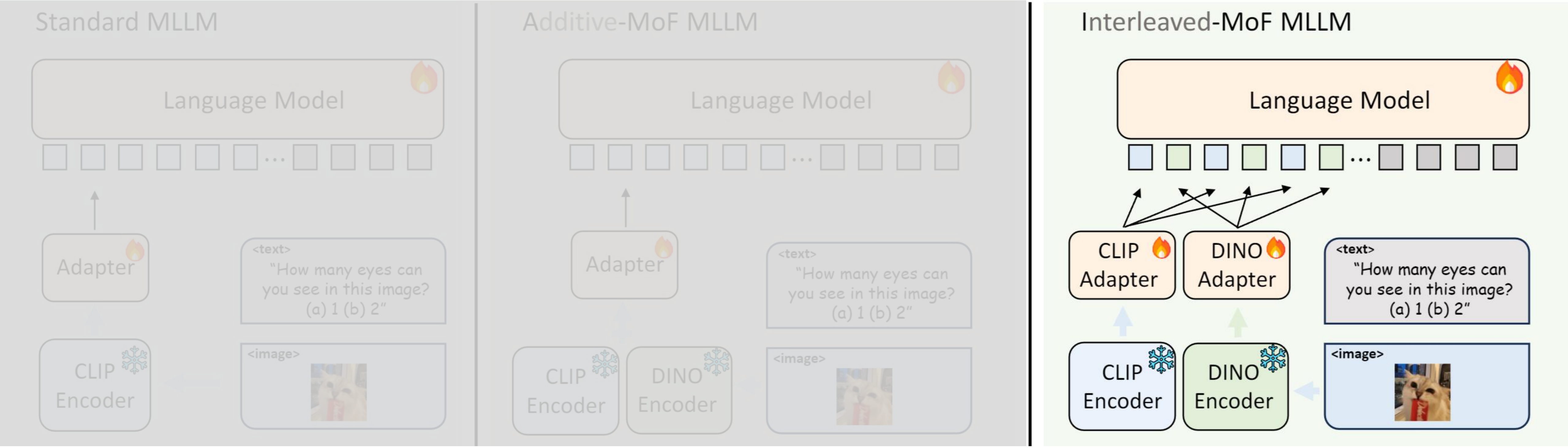
Mixture of Features

- Empirical Results of **Additive** MoF.

method	SSL ratio	MMVP	LLaVA
LLaVA	0.0	5.5	**81.8**
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)	68.5 (-13.3)

Mixture of Features

- Interleaved MoF



Mixture of Features

- Empirical Results of **Interleaved** MoF.

method	res	#tokens	MMVP	LLaVA	POPE[1]
LLaVA	224^2	256	5.5	81.8	50.0
LLaVA	336^2	576	6.0	81.4	50.1
LLaVA + I-MoF	224^2	512	16.7 (+10.7)	82.8	51.0
$\text{LLaVA}^{1.5}$	336^2	576	24.7	84.7	85.9
$\text{LLaVA}^{1.5}$ + I-MoF	224^2	512	28.0 (+3.3)		

1) *Li, Yifan, et al. "Evaluating object hallucination in large vision-language models." arXiv preprint arXiv:2305.10355 (2023).*

Mixture of Features

- Empirical Results of **Interleaved** MoF.

method	res	#tokens	MMVP	LLaVA	POPE
LLaVA	224^2	256	5.5	81.8	50.0
LLaVA	336^2	576	6.0	81.4	50.1
LLaVA + I-MoF	224^2	512	16.7 (+10.7)	82.8 +1%p	51.0 +1%p
LLaVA$^{1.5}$	336^2	576	24.7	84.7	85.9
LLaVA$^{1.5}$ + I-MoF	224^2	512	28.0 (+3.3)		

Mixture of Features

- Empirical Results of **Interleaved** MoF.

method	res	#tokens	MMVP	LLV^{B}	LLV^{W}	MMB	VQA^{T}	POPE	VQA^{V2}	MM-V
LLaVA$^{1.5}$	336^2	576	24.7	**84.7**	70.7	**67.7**	**61.3**	85.9	**80.0**	**35.4**
LLaVA$^{1.5}$ + I-MoF	224^2	512	28.0	82.7	**73.3**	61.6	55.3	86.3	77.3	33.5
LLaVA$^{1.5}$ + I-MoF	336^2	1152	**31.3**	81.8	**73.3**	65.4	58.7	**86.7**	79.3	34.6

Table 6. **Comparison with LLaVA-1.5 on 6 more benchmarks**. Interleaved-MoF LLaVA-1.5 obtains performance on par with the original method while showing improvements on benchmarks evaluating visual grounding. Benchmark names are abbreviated due to space limits. LLV^{B}: LLaVA Benchmark [31]; LLV^{W}: LLaVA-In-the-Wild [30]; MMB: MMBench [32]; VQA^{T}: TextVQA[52]; POPE: POPE [27]; VQA^{V2}: VQA-v2 [15]; MM-V: MM-Vet [64].

Conclusion

Conclusion

1. VLM and MLLM with built-in CLIP vision encoder only overlook crucial visual details in images and fail to properly classify important patterns.
2. Simply scaling data and models alone cannot resolve the inherent deficiencies in CLIP models.
3. Vision-and-language models and self-supervised learning-based vision models excel in different aspects, and these differences are difficult to capture using conventional benchmarks like ImageNet.
4. Mixture-of-Features approach can leverage the strengths and mitigate the limitations of both paradigms.
5. However, developing new evaluation metrics is necessary to facilitate the development of new visual representation learning algorithms.

Limitations

Limitations

1. MMVP requires human annotation and contations only 300 QA pairs.
2. MMVP contains only 135 Clip-blind pairs
    1. Each class only contains 15 images
    2. Possible ambiguity in the bucketing of classes from GPT-4
3. Computationally expensive for minimal performance increase
    - Two vision encoders
4. Other visual grounding performances are limited

QnA