Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

Shengbang Tong[1] Zhuang Liu[2] Yuexiang Zhai[3] Yi Ma[3] Yann LeCun[1] Saining Xie[1]

[1]New York University [2]FAIR, Meta [3]UC Berkeley

Premise

Is Vision Good Enough For Language? - Are Vision Models a Bottleneck in MultiModal systems.

- Progress in MLLMs largely stems from the reasoning capabilities of LLMs.
- Visual Component is heavily reliant on CLIP

Shortcomings:

- CLIP vision encoders overlook crucial visual details in images, and systematically fail to sort important visual patterns (orientation, counting, etc)
- Scaling alone does not rectify the inherent deficiencies in CLIP models.

Root Causes:

- Gap in Visual Embedding: Misalignment between CLIP's visual embedding space and more detailed representations from vision-only self-supervised models. Highlighting CLIP's limitations in capturing detailed visual patterns.

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

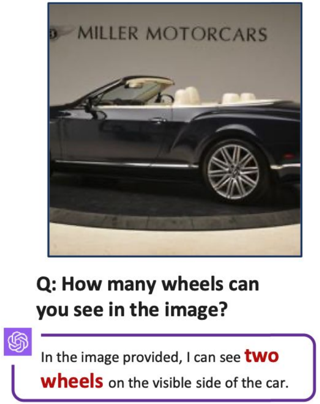
The dog is **facing to the right** from the camera's perspective. You can tell by the position of its tail and the orientation of its body.

Instances are from GPT-4V (accessed on Nov 04, 2023)
Text in Red - Incorrect Response
Text in Green - Hallucinated Response
Image Source: Fig 1, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

Overview of the Paper…

What? - Identifying the Problem

- Where does the Weakness in MLLMs stem from:
  - Deficiency in Visual Modality
  - Language Understanding
  - Alignment between the Visual and Language World



- The Paper studies the Deficiency in Visual Modality in MLLMs:
  - MLLMs fail at elementary visual tasks – object orientation, counting, identifying specific features.
  - Suggests a lack of accurate visual grounding in MLLMs
  - Most MLLMs are built on pre-trained vision and language models
    - Any Limitations in the pre-trained Vision Models can cascade into the downstream MLLM

CLIP - Recap

- Goal: To create a joint multimodal embedding space by aligning image and text representations through contrastive pre-training.

- Characteristics of the Visual Representations:

  - Semantic Alignment: Visual representations are closely aligned with textual semantics, enabling strong cross-modal understanding.

  - Global Feature Focus: CLIP's training objective prioritizes learning representations that capture overall semantic meaning rather than local details.

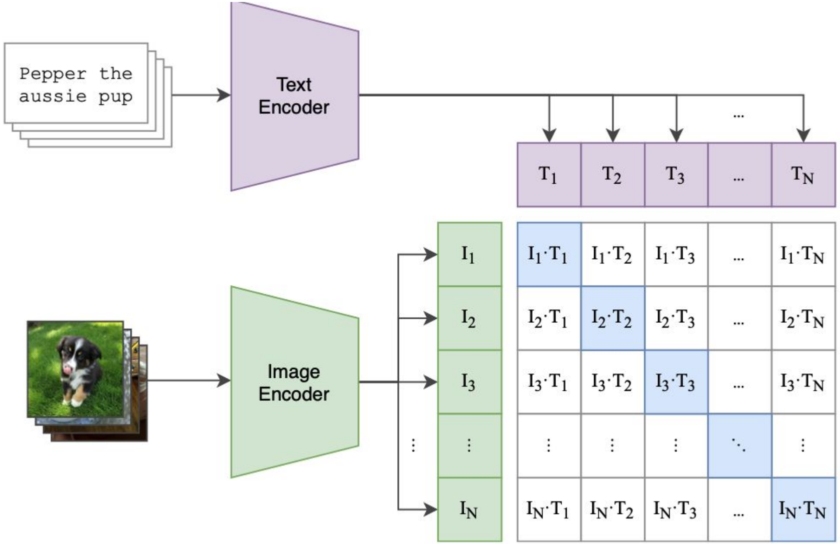


Image Source: Fig 1, [Learning Transferable Visual Models From Natural Language Supervision]

DINO - Recap

- Goal: To learn robust and detailed visual representations through self-supervised knowledge distillation, where the model predicts its own representations under different augmented views.

- Characteristics of the Visual Representations:

  - Fine-Grained Features: learns detailed and local visual features, making it effective for tasks that require precise object recognition and understanding of subtle visual differences.

  - Improved spatial understanding: By having to match both global and local views, the model learns not only high-level context but also how individual parts of objects relate to each other in space.

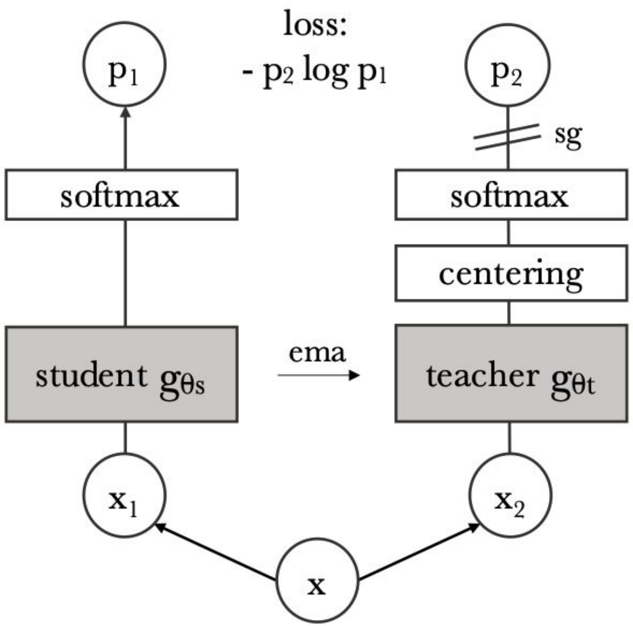


Image Source: Fig 2, [Emerging Properties in Self-Supervised Vision Transformers]

CLIP vs DINO – Visual Features

- Global vs. Local Focus:
    - CLIP: high-level semantic understanding, capturing overall scene-wide information but often overlooking fine-grained details.
    - DINO: Captures fine-grained local features, learns detailed object-level features
- Supervision Type
    - CLIP: Contrastive Learning with text-image pairs, which aligns visual features with corresponding natural language representations. – May Cause the model to prioritize text-driven semantics over vision-specific information.
    - DINO: self-supervised knowledge distillation approach, placing emphasis on the visual structure itself – learning more detailed representations that are not biased by language or external supervision.

Why? - Root Causes of Visual Shortcomings in MLLMs

- Dependence on CLIP for Visual Encoding:
  - Overemphasis on high-level semantic features
  - Deficiencies in CLIP transfer down to the MLLM.
- Gap Between CLIP and Vision-Only Self-Supervised Models:
  - DINOv2 captures more detailed and nuanced visual features
  - Differences between the visual embedding spaces of CLIP and DINOv2 – highlights CLIP's deficiencies in capturing detailed visual patterns.

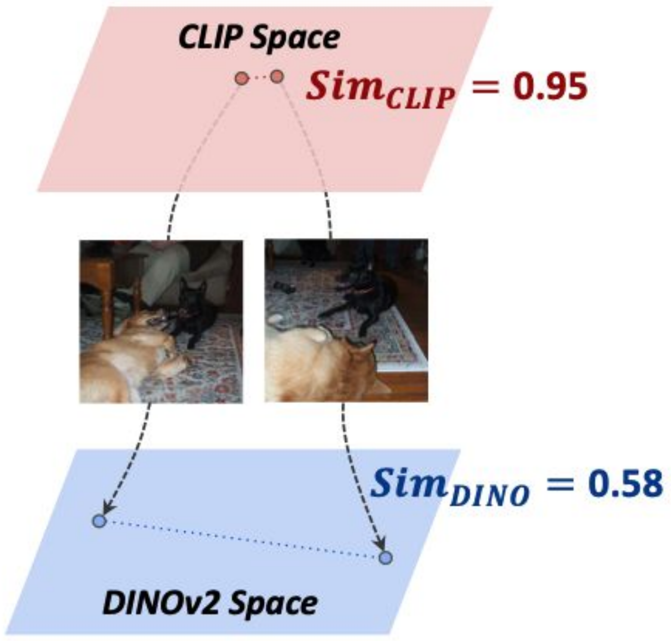
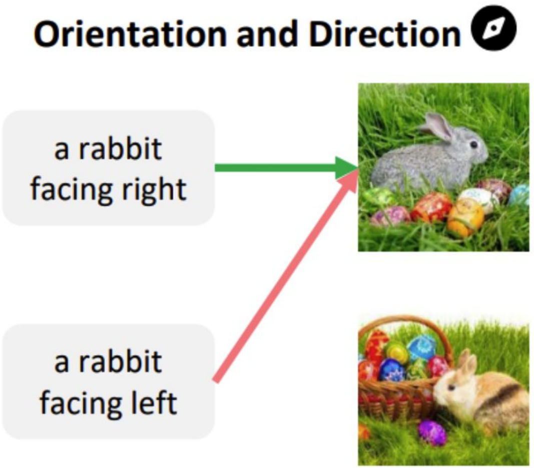


Image Source: Fig 2, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

How? - How were these shortcomings studied

- CLIP-blind Pairs
  - If two Visually Different Images are encoded close together by CLIP – means that one of the images was ambiguously encoded
  - The paper utilizes another vision encoder (DINOv2) to measure the visual similarity of the images.
  - Pairs of visually different images that are close together in CLIP space, but further apart in the DINO space are coined as CLIP-blind pairs.

	Orientation and Direction
	Presence of Specific Features
	State and Condition
	Quantity and Count
	Color and Appearance
	Positional and Relational Context
	Structural and Physical Characteristics
A	Text
	Viewpoint and Perspective



- Discovers that CLIP-blind pairs lead to errors in downstream tasks.
- Creates the MultiModal Visual Patterns (MMVP) benchmark based on these CLIP-blind pairs.
- Identifies failure instance types in MLLMs (among the CLIP-blind pairs) – patterns such as “orientation”, “counting”, “viewpoint”, etc poses challenges for CLIP models.
- Shows that scaling (training data, model) alone cannot mitigate these challenges (for 7 out of the 9 identified patterns).

How? - How were these shortcomings studied

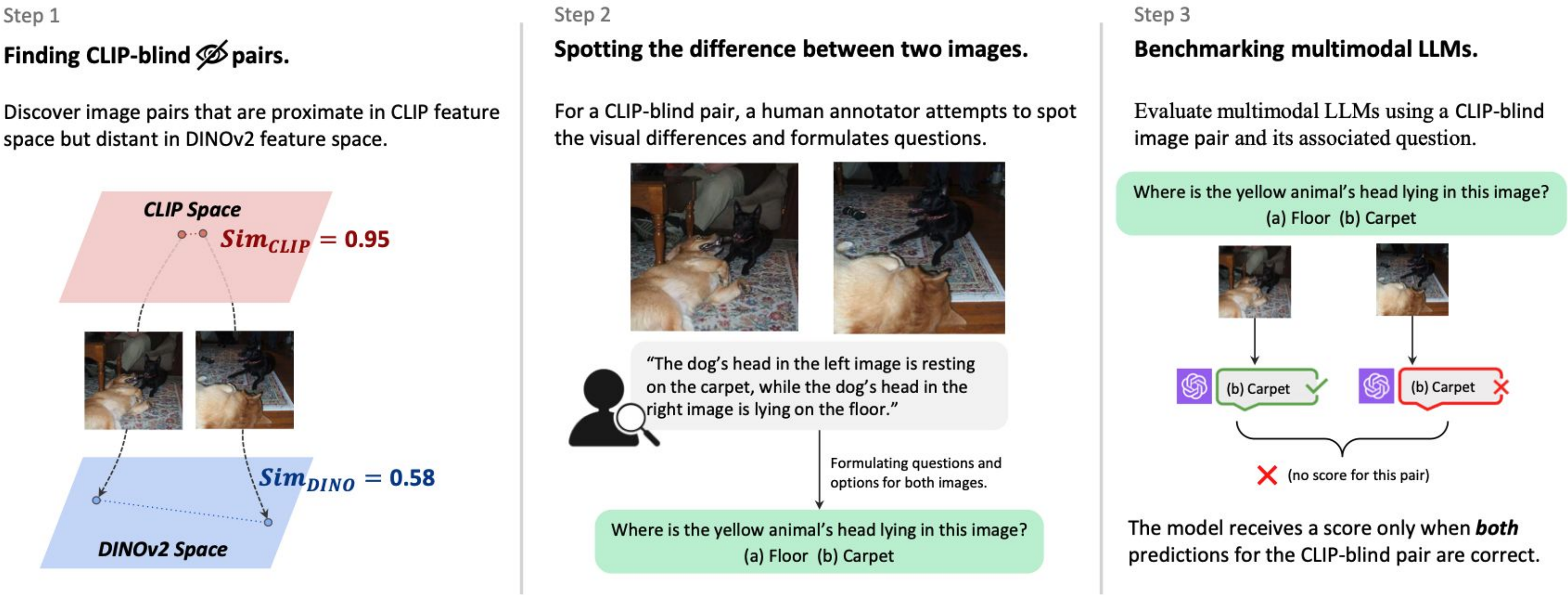


Image Source: Fig 2, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

How? - How to Improve on these Shortcomings

- Since the Visual Shortcomings of MLLMs stems from CLIP
  - Integrate vision-centric representations into MLLMs (Specifically DINOv2) - to enhance the visual grounding capabilities

- Shows that DINOv2 features are more effective in visual grounding – BUT, come at the cost of diminished instruction-following ability.
  - Solution: Interleave the tokens from CLIP and DINOv2 – which enhances visual grounding while maintaining the instruction-following capabilities.

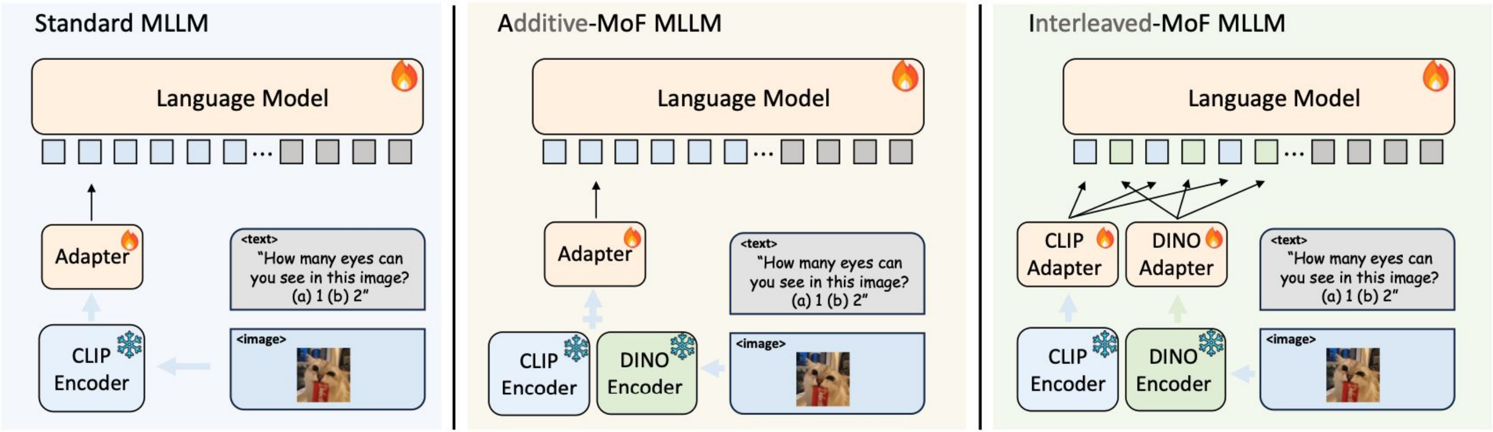


Image Source: Fig 7, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

More in depth (benchmark creation, experiments, results) …

Designing the MMVP Benchmark

- CLIP-blind pairs
  - Obtained from the ImageNet and LAION-Aesthetics Datasets
  - Pairs are obtained such that the cosine similarity exceeds 0.95 for CLIP embeddings and less than 0.6 for DINOv2 embeddings.

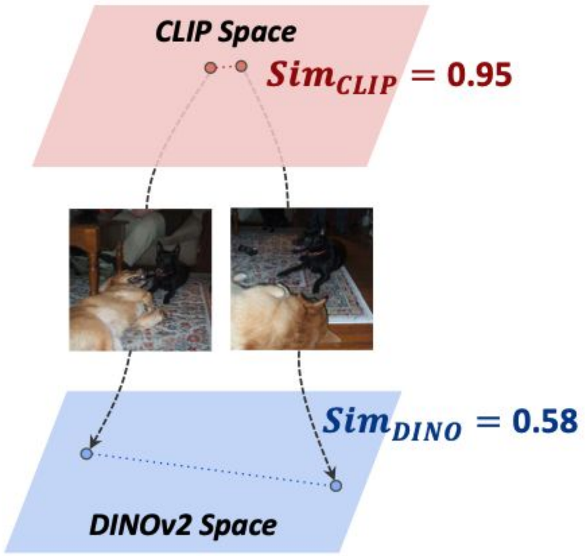


Image Source: Fig 2, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

Designing the MMVP Benchmark

- Using these CLIP-blind pairs:
  - Designs 150 pairs with 300 questions
  - Crafts questions that probe visual details that CLIP tends to overlook. (Basic questions, but MLLMs tend to fail on these, and overlook crucial visual details)

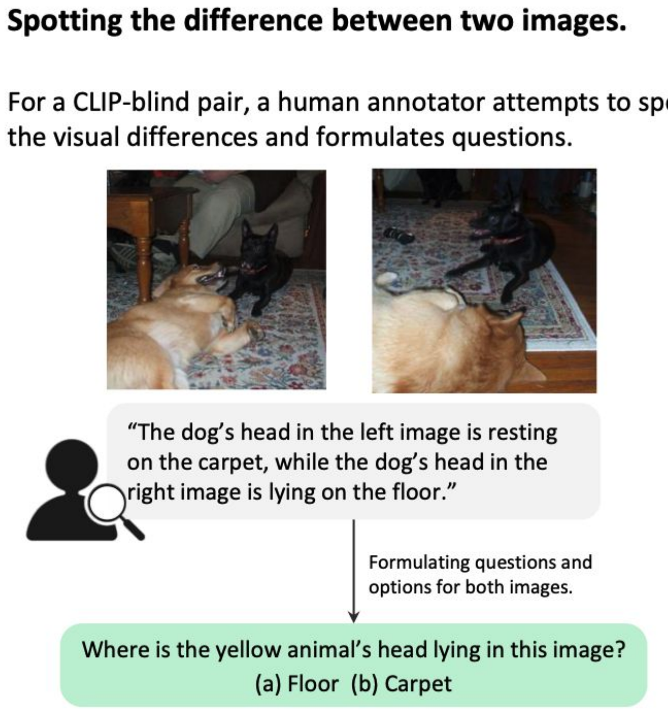
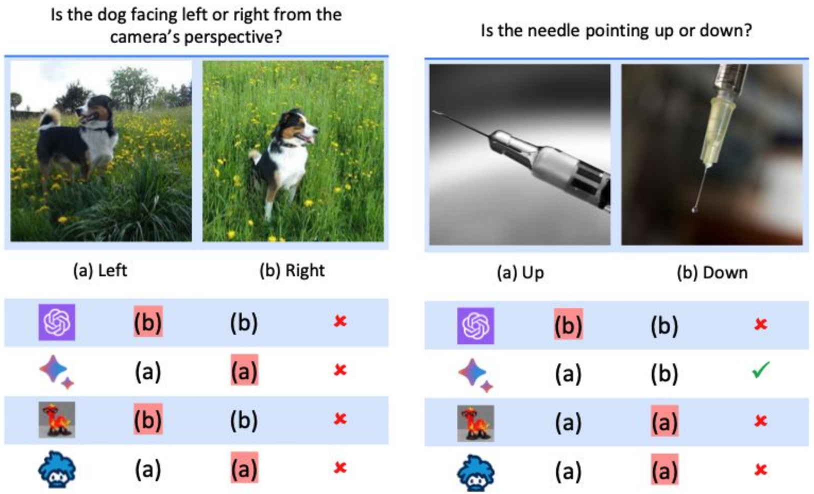


Image Source: Fig 2, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

Benchmark Results

- Each question is queried independently, eliminating any biases from chat histories.
- Human performance: a user study where users are presented with 300 questions in a randomized sequence.
- For any given pair of images, they consider a pair of images to be correctly answered if both the questions associated with the pair are answered accurately.
- Human performance indicates that the task was straightforward
- Also conducted an ablation, where they swap options, reformulate questions to confirm that the poor performance is not due to hallucination by the LM, but from the visual incapability of the model.



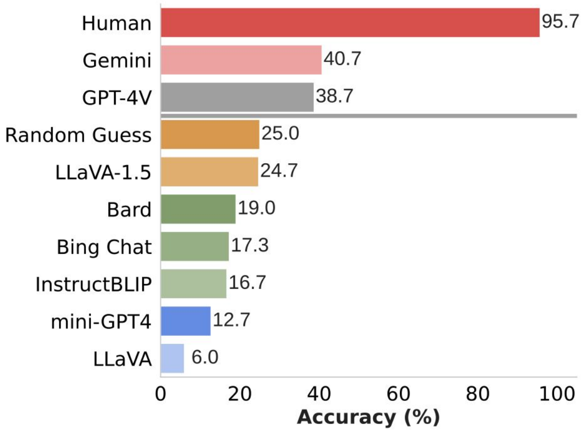


Image Sources: Fig 4, 3, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

Identifying Visual Patterns in the CLIP-Blind Pairs

- Uses the formulated Questions → Passes it to GPT-4 → To categorize and identify the patterns in the dataset

- These identified patterns are a proxy to the tasks where CLIP vision encoders fails.

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

Image Source: Sec 3.1, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]()

MMVP-VLM Benchmark

- MMVP-VLM is designed to systematically study if CLIP models can handle these visual patterns well.
- They Distill a subset of questions from the MMVP benchmark – Each Visual Pattern is represented by 15 text-image pairs.
- Increasing n/w and data size – only aids “color and appearance” and “state and condition”
- ImageNet-1k zero-shot accuracy is not a definitive indicator of a model’s performance regarding visual patterns.

	Image Size	Params (M)	IN-1k ZeroShot								A		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

Image Source: Table 1, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

How CLIP’s Errors Affect MLLMs

- When the CLIP vision encoder underperforms on a certain visual pattern, the MLLM tends to exhibit similar shortcomings.

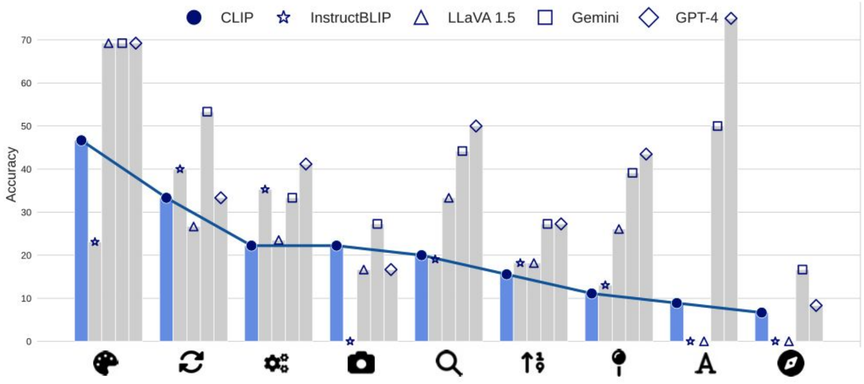


Image Sources: Fig 4, 3, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

Mixture-of-Features (MoF) for MLLM

- How to build a more component visual encoder?
  - Mix CLIP Features with vision-only SSL model features
  - Without Mixing, integrate (interleave) the features from both CLIP and SSL

- Setting: LLaVA, CLIP-ViT-L-14, DINOV2-ViT-L-14 (finetune setting same as the original LLaVA paper)

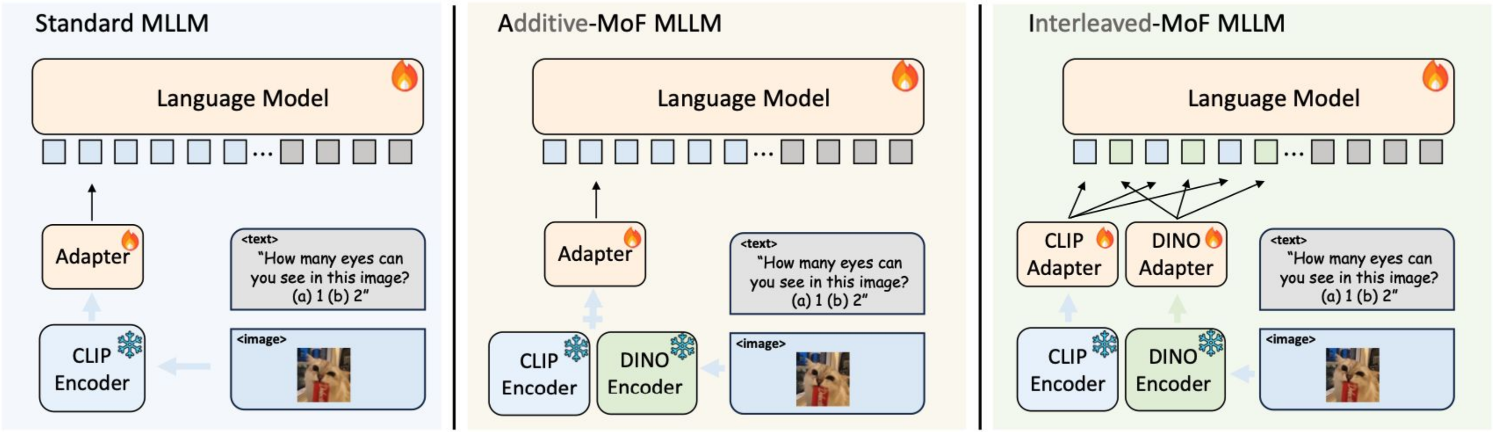


Image Source: Fig 7, [Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs]

Results

- Additive MoF
    - As the proportion of DINOv2 features increases, decline in instruction-following capability.
- Interleaved MoF
    - interleaving of MoF between vision-only SSL models and VLM models leads to improved performance in visual grounding tasks.
    - They evaluate Interleaved MoF on additional benchmarks (MM-Bench, GQA) and find that it archives similar performances.

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)

Table 2: Decrease in instruction following capability (see LlaVA benchmark scores) when DINO ratio increases

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	51.0
LLaVA$^{1.5}$	336^2	576	24.7	84.7	85.9
LLaVA$^{1.5}$ + I-MoF	224^2	512	28.0 (+3.3)	82.7	86.3

Table 3: Interleaved MoF improves visual grounding while maintaining same level of instruction following ability.