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

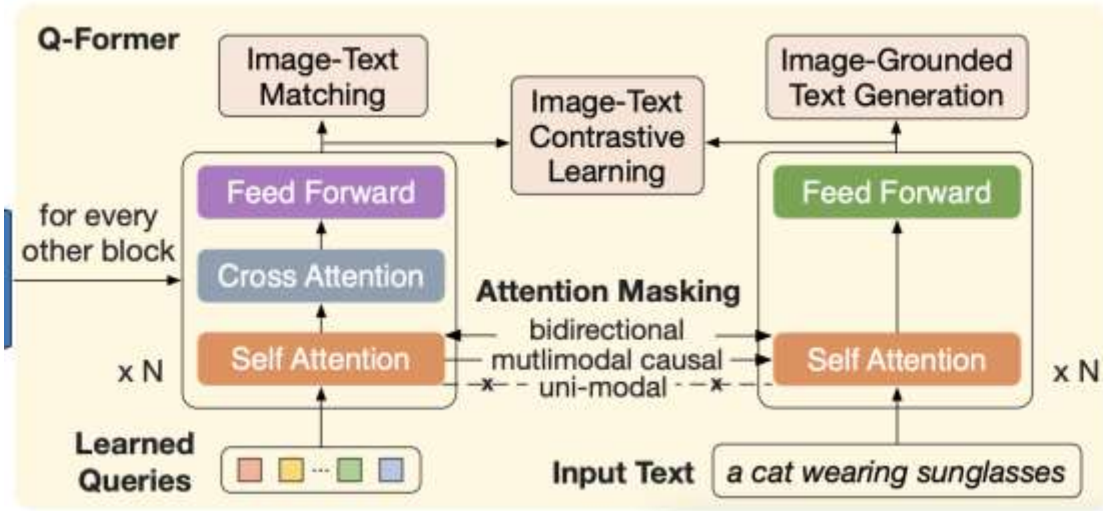
曹耘宁
2024/0319

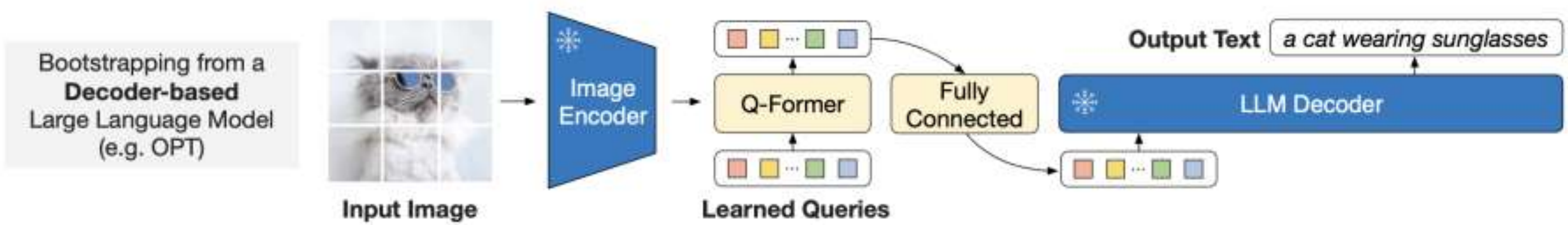
智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

- 研究背景
- 研究方法
- 实验效果
- 总结

MLLM--BLIPv2

- CLIP+LLM
- Q-former的结构
- 将query输入llm





GPT4V对简单的视觉内容描述出错

Figure 1. Instances are systematically identified where the visual question answering (VQA) capabilities of GPT-4V [41] fall short (Date accessed: Nov 04, 2023). Our research highlights scenarios in which advanced systems like GPT-4V struggle with seemingly simple questions due to inaccurate visual grounding. Text in red signifies an incorrect response, while text in green represents hallucinated

问题

- CLIP视觉编码器基于实例级预训练，细节特征提取能力不足（CLIP-blind pair）

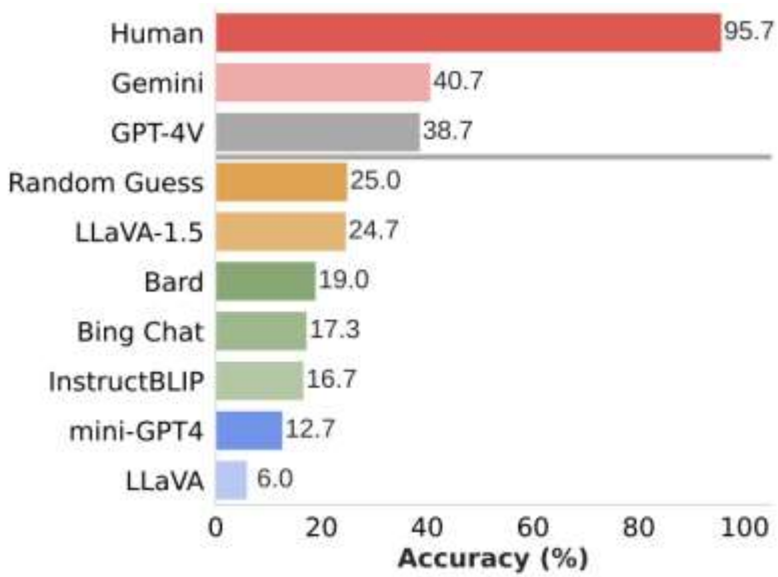


Figure 4. **Benchmark results of current *SOTA* MLLM models and humans.** We evaluate benchmark questions for current *SOTA* MLLM models and human performances through user studies.

- 研究背景
- 研究方法
- 实验效果
- 总结

构建MMVP（multimodal visual patterns）benchmark

- 1. CLIP blind pair：有明确视觉差异，但是clip特征相似
  - ImageNet，LAION-Aesthetics
  - CLIP-ViT-L-14：相似度>0.95
  - DINOv2-ViT-L-14: 相似度<0.6

Figure 2. Constructing MMVP benchmark via CLIP-blind pairs. **Left:** We start with finding CLIP-blind pairs that have similar CLIP embedding but different DINOv2 embedding. **Center:** We manually inspect the differences between pair-wise images and formulate questions based on the differences in the images. **Right:** We ask MLLMs the question alongside the CLIP-blind pair. The model receives a score only when both questions for the CLIP-blind pair are answered correctly.

构建MMVP benchmark

- 2. 人工标注question+answer
- 3. 测试MLLM准确率

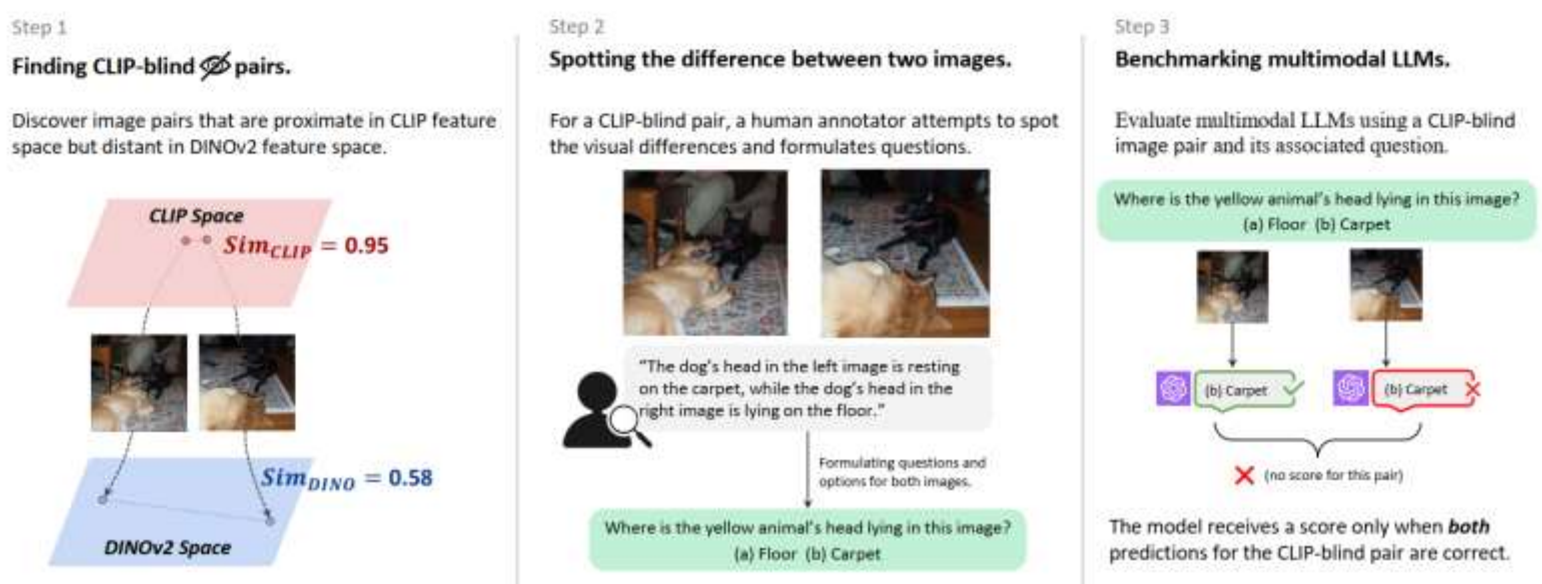


Figure 2. Constructing MMVP benchmark via CLIP-blind pairs. **Left:** We start with finding CLIP-blind pairs that have similar CLIP embedding but different DINOv2 embedding. **Center:** We manually inspect the differences between pair-wise images and formulate questions based on the differences in the images. **Right:** We ask MLLMs the question alongside the CLIP-blind pair. The model receives a score only when both questions for the CLIP-blind pair are answered correctly.

构建MMVP benchmark

- MMVP benchmark举例

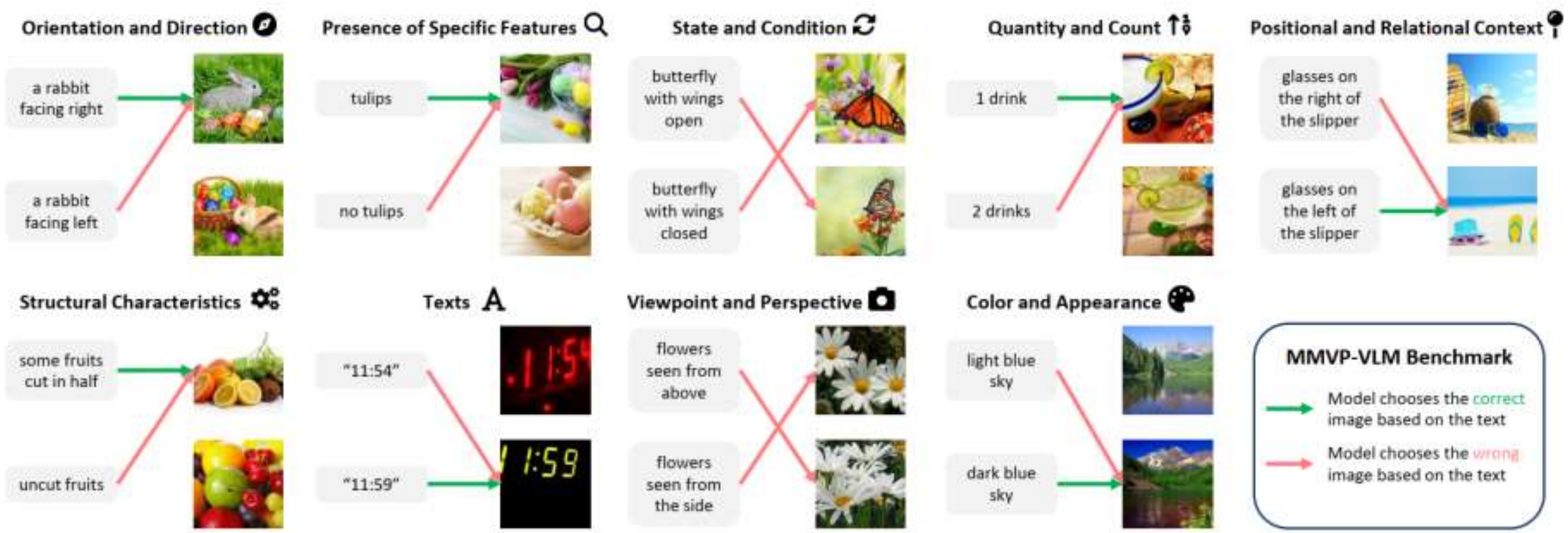


Figure 5. **Examples from MMVP-VLM.** MMVP-VLM consists of image pairs across nine visual patterns. The examples in the figure are from EVA01 ViT-g-14 model [54], one of the largest CLIP models that also fails to choose the right image given the text description.

构建MMVP benchmark

- 3. MMVP测试结果

Swapping options The first experiment swaps the two options in the MMVP benchmark. For example, we change the question from "Are the butterfly's wings closer to being open or closed? (a) Open (b) Closed" to "Are the butterfly's wings closer to being open or closed? (a) Closed (b) Open".

Empirically, we find that GPT-4V obtains a 40.3% accuracy on the option swapping in our study, as opposed to the original 38.7%. We observe that a few questions are answered differently, while the majority remain the same. This further suggests that the visual incapabilities are in the vision encoder rather than in alignment or the LLMs.

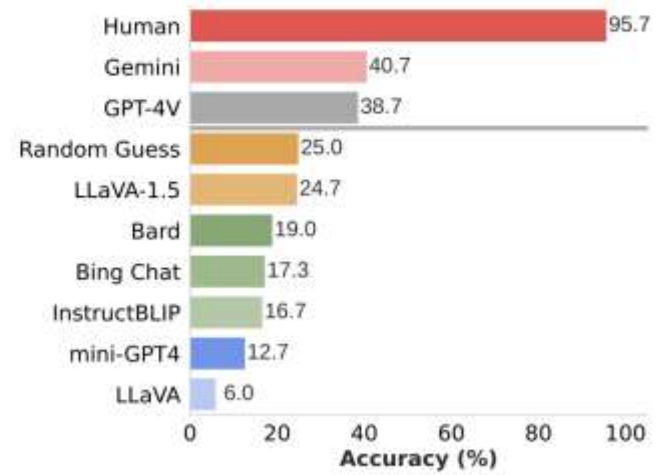
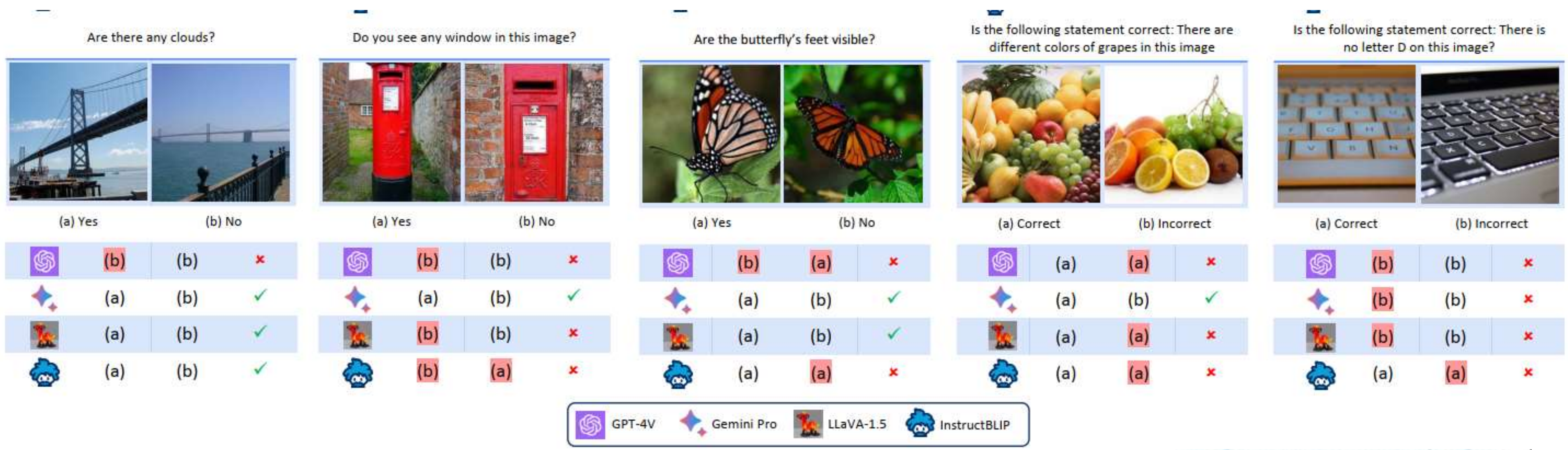


Figure 4. **Benchmark results of current *SOTA* MLLM models and humans.** We evaluate benchmark questions for current *SOTA* MLLM models and human performances through user studies.



Benchmark总结

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

Visual Patterns

- 更先进的CLIP只在color 和state有提升

	Image Size	Params (M)	IN-1k ZeroShot										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

Table 1. Performance of various CLIP based models on different visual patterns in MMVP-VLM benchmark. Models scaled up in resolution show minimal improvement, whereas a slight advantage is observed when scaling up the network. For each visual pattern, ImageNet-1k Zero-shot accuracy and MMVP average, we use light gray to highlight the best performance. For most of the visual patterns, all CLIP-based methods show struggle, as evident from the scores. We use symbols for visual patterns due to space limit: : Orientation and Direction, : Presence of Specific Features, : State and Condition, : Quantity and Count, : Positional and Relational Context, : Color and Appearance, : Structural and Physical Characteristics, A: Texts, : Viewpoint and Perspective.

Benchmark总结

- MLLM的性能

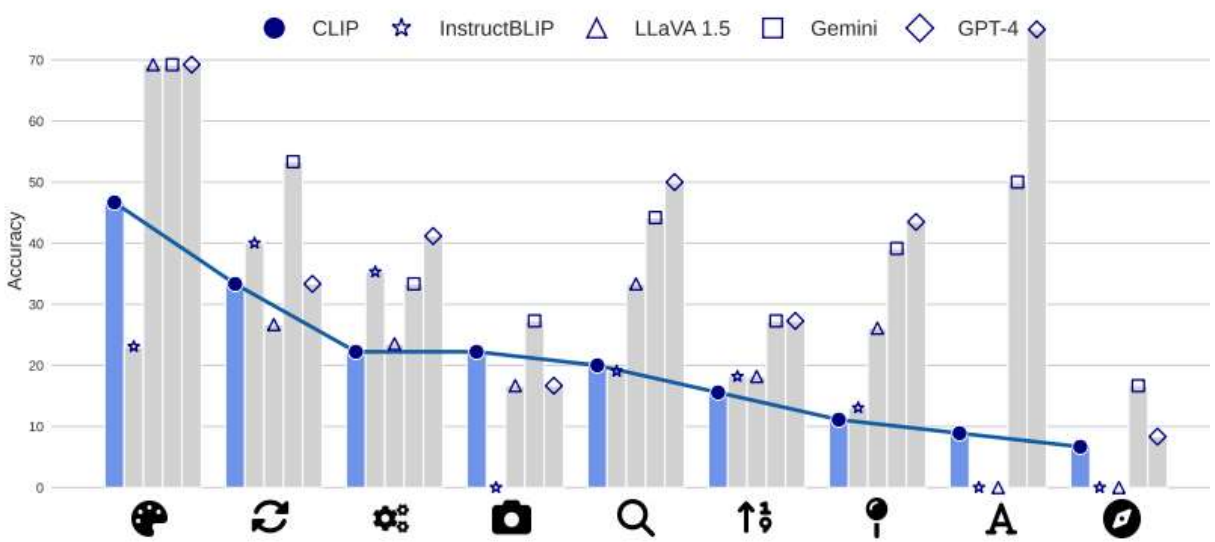


Figure 6. **CLIP and MLLM's performance on visual patterns.** If CLIP performs poorly on a visual pattern such as " orientation", MLLMs also underperform on the visual pattern.

Mixture of Feature方法

- Additive： $\alpha F_{clip} + (1 - \alpha F_{dino})$
- interleaved :保持顺序，空间交错

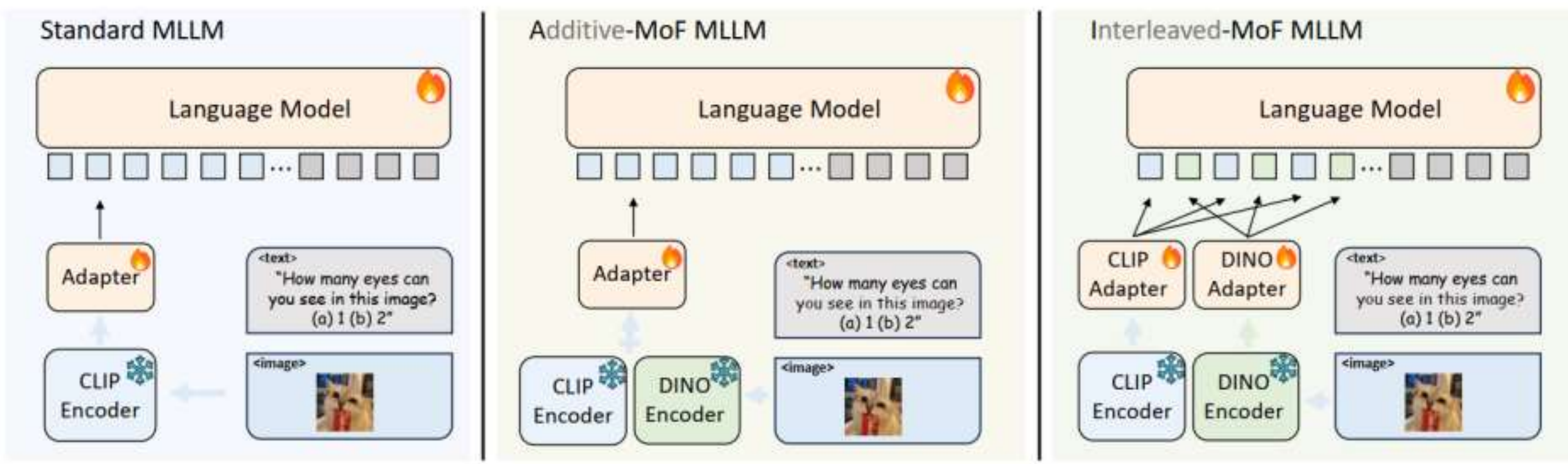


Figure 7. **Different Mixture-of-Feature (MoF) Strategies in MLLM.** *Left*: Standard MLLM that uses CLIP as *off-the-shelf* pretrained vision encoder; *Middle*: Additive-MoF (A-MoF) MLLM: Linearly mixing CLIP and DINOv2 features before the adapter; *Right*: Interleaved-MoF (I-MoF MLLM) Spatially interleaving CLIP visual tokens and DINOv2 visual tokens after the adapter.

- Additive：随着dinov2特征的比例增加，mmvp效果提升，指令跟随效果降低
- Interlead：指令跟随能力得到保持

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. **Empirical Results of Additive MoF.** We use DINOv2 as the image SSL model in our work. With more DINOv2 features added, there is an improvement in visual grounding, while a decline in instruction following ability.

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

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

- 研究背景
- 研究方法
- 实验效果
- 后续工作
- 总结

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

Table 4. Hyperparameters for MoF training on LLaVA and LLaVA-1.5.

- Stage II

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].

	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.

- Stage II

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].

	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.

- 研究背景
- 研究方法
- 实验效果
- 总结

总结

- 对MLLM在视觉问题上的错误进行详尽分析，并且聚焦到视觉编码器能力上。
- Benchmark构造与实验有说服力
- MoF方法

Thanks!