
All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning

Gokul Swamy¹ Sanjiban Choudhury^{2,3} Wen Sun² Zhiwei Steven Wu¹ J. Andrew Bagnell^{3,1}

All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning

Topic: 本文探讨了在基础模型微调中，强化学习的作用和价值，尤其是与直接优化策略参数的离线最大似然估计相比的优势。

强化学习从人类反馈 (RLHF)： 先训练一个奖励模型 (Reward Model, RM)，然后使用该模型为下游的强化学习过程提供反馈，以优化策略。

离线微调： 通过离线最大似然估计对策略参数在数据集上进行优化。

问题提出： 从信息论的角度来看，通过奖励模型传递信息会导致信息损失，而在线策略采样并不能创造新信息。那么，为什么RLHF比直接使用MLE进行离线微调更有效呢？

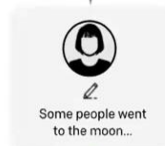
Step 1

Collect demonstration data, and train a supervised policy.

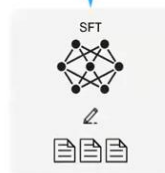
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



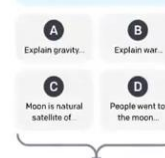
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

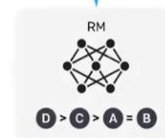
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



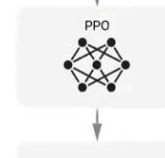
Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

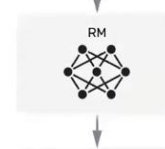


The policy generates an output.

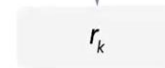


Once upon a time...

The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Q: What is the value of a two-stage, interactive FT procedure if we just want to maximize data likelihood?

All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning

统一的微调目标

从高层次看，各类微调任务（如监督微调SFT、偏好微调PFT）均可表述为以下反向KL正则化的策略优化问题：

$$\pi^* = \operatorname{argmin}_{\pi \in \Pi} \underbrace{\mathbb{D}_{\text{KL}}(\mathbb{P}_{\mathcal{D}} || \mathbb{P}_{\pi})}_{\text{Data Likelihood}} + \underbrace{\beta \mathbb{D}_{\text{KL}}(\mathbb{P}_{\pi} || \mathbb{P}_{\pi_{\text{ref}}})}_{\text{Prior Reg.}}. \quad (2)$$

其中第一项前向KL散度衡量学习策略 π 对数据集 \mathcal{D} 样本的拟合程度，第二项反向KL散度约束 π 的生成概率分布需接近参考策略的在线分布。

直观而言，若 \mathcal{D} 能完全覆盖所有轨迹对，则无需第二项；但由于有限样本限制，需引入正则项避免策略偏离过远。为简化表述，设 $\beta=1$ 并将第二项暂时替换为熵正则化）：

$$\pi^* = \operatorname{argmin}_{\pi \in \Pi} \mathbb{D}_{\text{KL}}(\mathbb{P}_{\mathcal{D}} || \mathbb{P}_{\pi}) - \mathbb{H}(\pi), \quad (3)$$

同时由于奖励模型和策略模型的起点都是同一个SFT checkpoint，并且在同一个数据集上进行训练，所以作者假设他们是同构的。

作者在数学上证明了，在线与离线PFT方法均可视为对该目标的优化，尽管实现机制存在本质差异。在线方法通过与环境的实时交互来优化策略，而离线方法直接在数据集上优化策略。

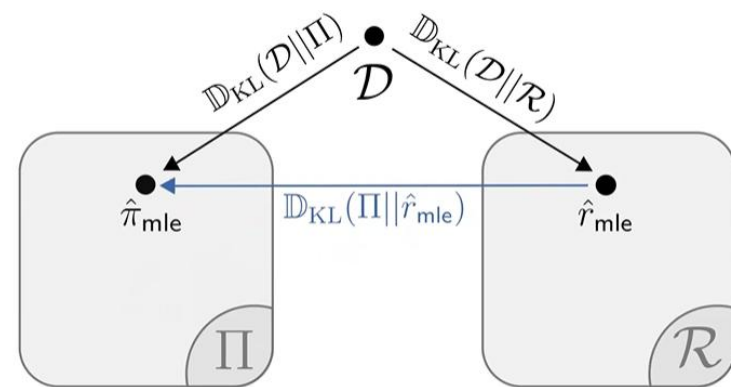
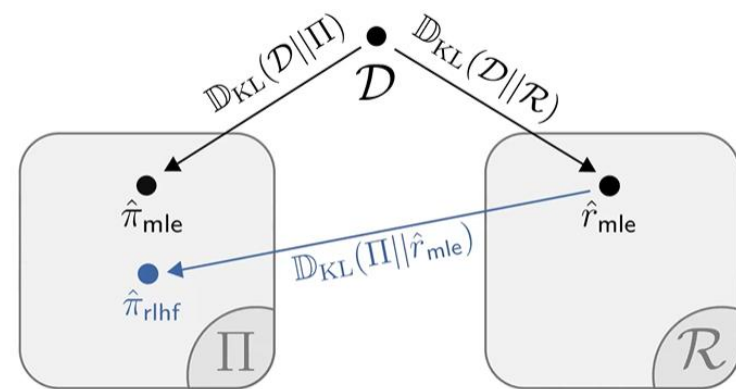
All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning

引理2.1： 软RL问题可以被看作是一个逆向 KL 散度投影问题（当我们通过软策略优化来调整策略时，实际上我们是在做一件事情：让策略的行为分布尽可能接近一个由奖励模型定义的理想分布。）

引理 2.2： 当奖励模型类和策略类同构时，RLHF 和 MLE 产生相同的解。

在理想化的假设条件下(策略与奖励模型都属于同一个可逆映射的函数类)，不同的优化方法（无论是在线还是离线）最终都会达到相同的目标，即最大化数据的似然性（likelihood）。换句话说，在这些假设下，不同的方法在理论上是**等价**的。

然而，实际经验却表明：在线RLHF总是比离线直接微调效果好。这就是文章所讨论的悖论：**理论上，它们可以学到同样的最优策略，为何实践中偏要多绕这么一步？**



All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning

作者接下来通过系列受控实验来对这些理想的假设进行质疑。

● 实验设计：

任务：tl;dr 数据集上的文本摘要。

相同的优化器：dpo；相同的数据和模型（RM就是在最后一层不同）。

● 实验结论：

1. 同样的偏好数据、同样的初始SFT模型：使用在线DPO做微调，往往比做离线DPO的性能好。
2. 增加或改变数据源（如扩增prompt、在策略上采样更多样本）并不能让离线微调性能逼近在线RL。
3. 当任务本身的生成难度显著降低时（如只生成极短文本），在线与离线方法的性能差距就会消失。

这些对比指出：在线的优势并非因为它“多拿”了信息，而是它在某些更深层的机制上，更能利用一个相对易学的验证函数去搜索策略。

All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning

H6: 微调中的生成验证差距

- **生成 (Generation)** : 生成高质量内容 (如符合人类偏好的文本) 可能非常复杂。
- **验证 (Verification)** : 判断生成内容的质量 (如判断文本是否合理) 可能相对简单

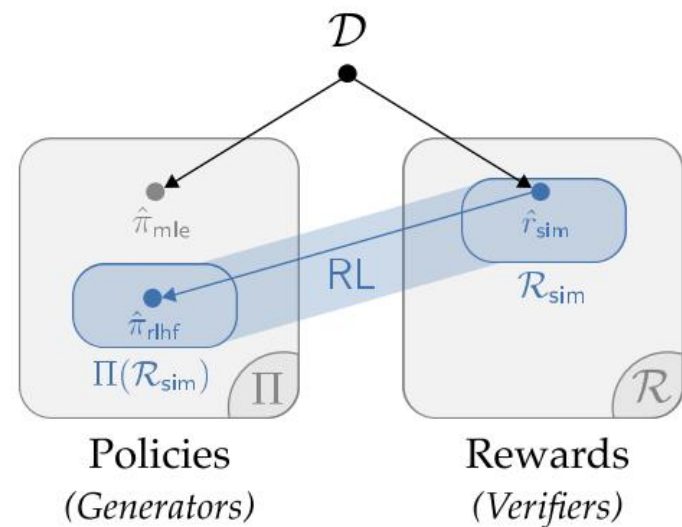
在线微调的两步过程:

1. 第一步: 找到一个相对简单的奖励模型 (是RM的子集) ;
2. 第二步: 找到 (近似软) 最优策略, 他在简单的奖励模型上表现最优;

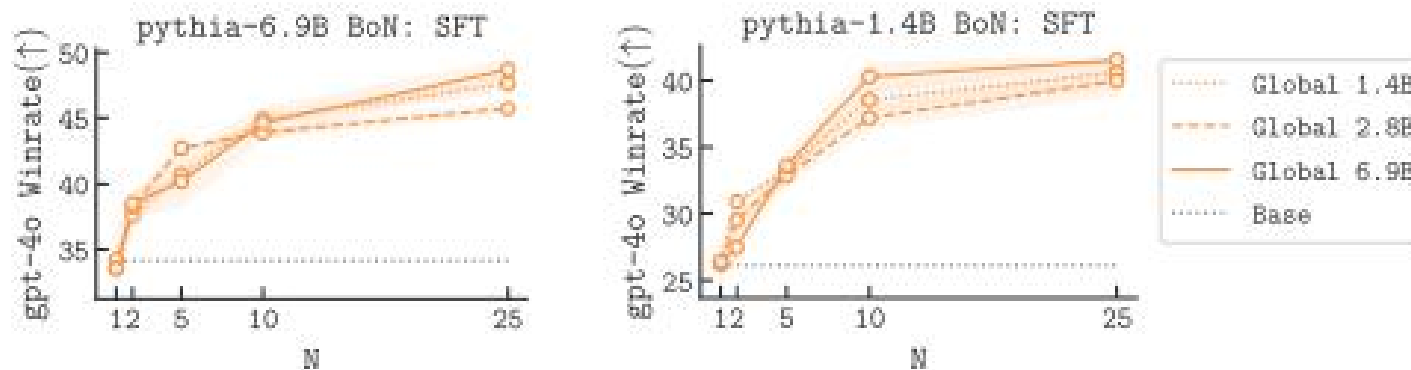
这样, 下游RL的作用不是“创造新信息”, 而是在策略空间中高效筛选出符合RM标准的最优策略。由于RM足够简单, RL可以快速收敛到满足RM要求的策略子集, 而无需遍历整个复杂策略空间。

\mathbb{H}_6 : Online PFT is *Proper* Policy Learning.

For fine-tuning problems with a simpler underlying reward function than (soft) optimal policy and a reward model class \mathcal{R} that enables learning simpler functions with fewer samples, the first step of online fine-tuning is finding a relatively simple reward model $\hat{r}_{sim} \in \mathcal{R}_{sim} \subset \mathcal{R}$, while the second step finds the (approximately soft) optimal policy $\hat{\pi}_{rlhf}$ for \hat{r}_{sim} . Thus, end to end, online fine-tuning only has to search over policies in $\Pi(\mathcal{R}_{sim}) \subset \Pi$, rather than across all of Π like offline fine-tuning.



All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning



实验结果支持：

使用比生成策略小得多的全局RM导致与使用与策略大小相同的RM几乎相同的BoN性能。

使用比生成策略大得多的全局RM，与类似大小的RM相比，BoN性能没有明显的改善。

前者的表现整体更优。

All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning

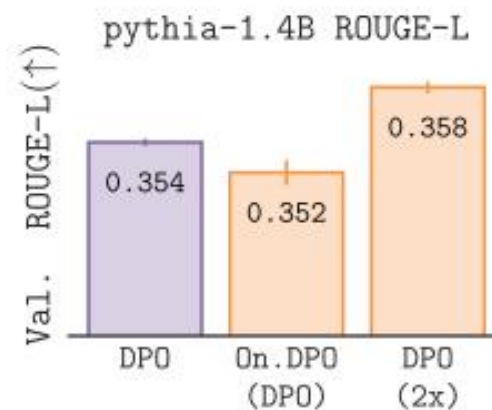
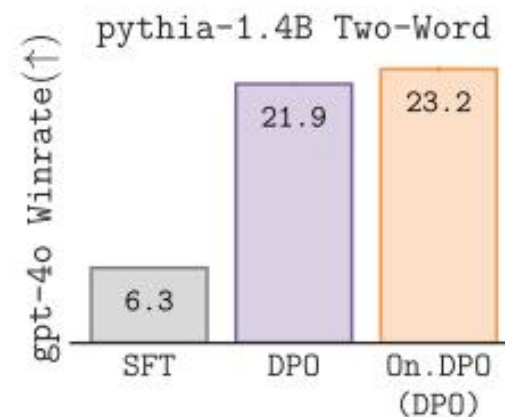
尝试证伪:

减少问题的生成-验证差距:

1. 减少问题的范围: 将摘要长度限制为两个单词来降低生成任务的复杂性。在这种情况下, 生成任务的复杂性接近于验证任务的复杂性。发现在线DPO并没有显著提升离线DPO策略的性能。这与H6的预测一致, 即当生成和验证任务的复杂性接近时, 在线PFT的优势会消失。

2. 选择复杂的奖励函数

使用ROUGE-L指标作为奖励函数。ROUGE-L指标通过计算生成摘要与参考摘要之间的重叠词数来评估摘要质量。在这种情况下, 奖励函数的复杂性与最优策略的复杂性相当。发现在线DPO并没有提升离线DPO策略的性能。这进一步支持了H6的预测, 即当奖励函数的复杂性与最优策略的复杂性相当时, 在线PFT的优势会消失。



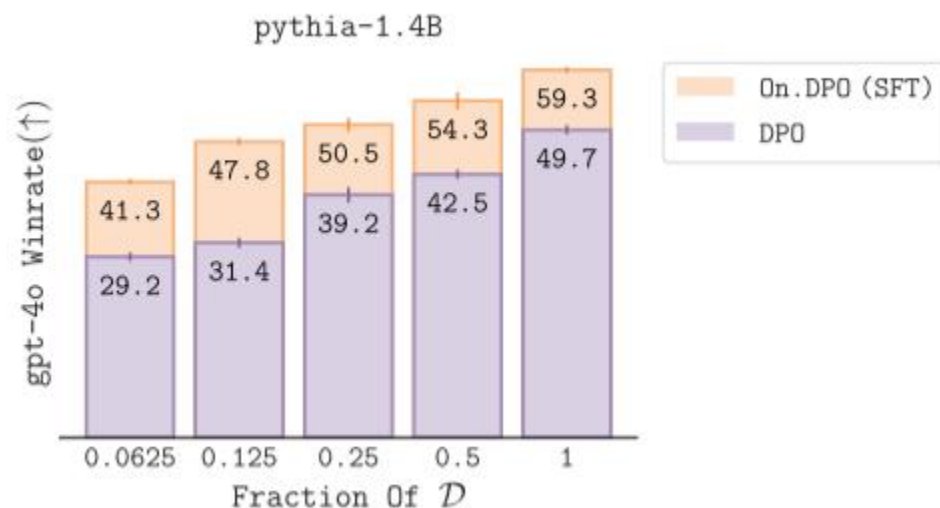
All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning

在实验中，作者观察到随着使用的人类偏好数据集的比例逐渐增加，离线PFT和在线PFT之间的胜率（winrate）差距保持相对稳定。这个结果并没有反驳假设H6。相反，它表明随着数据量的增加，虽然奖励模型（RMs）的复杂性可能在增加，但这些奖励模型仍然比对应的最优策略（soft-optimal policy）更简单。

理论分析：

无限数据极限（Infinite Data Limit）：假设H6预测，在无限数据的极限情况下，离线PFT和在线PFT之间的性能差距应该会消失。

原因是，当数据量足够大时，我们可以完全确定生成器（generator）在状态空间中的行为。在这种情况下，即使有一个完美的验证器（verifier），也不会提供任何新的信息。



All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning

作者还提出了几个基于假设H6的后续研究方向：

1. 可以应用机制可解释性（mechanistic interpretability）的技术来更精确地描述实际偏好微调（PFT）策略中的奖励模型和策略的链路复杂性。
2. 也可以将奖励模型及其诱导的策略视为一个新平台，用于测试关于深度学习背后机制的各种理论。
3. 另一个研究方向是确保当前的奖励模型架构能够准确地表示人类的偏好。评分者观点的多样性常常导致非传递性偏好，这些偏好无法用任何 Bradley-Terry 奖励模型来合理化。相反，研究不假设传递性的成对偏好模型可能会使在线 PFT 的验证器更好。
4. 假设 H6 还表明，对于越来越复杂的、需要更长期规划的问题（例如多轮强化学习人类反馈（RLHF））或代理任务，我们应该会看到在线和离线微调（FT）之间的差距进一步扩大。看看这一现象是否会在实践中得到证实，将是一件很有趣的事情。

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

Shengbang Tong¹ Zhuang Liu² Yuexiang Zhai³
Yi Ma³ Yann LeCun¹ Saining Xie¹

¹New York University ²FAIR, Meta ³UC Berkeley

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs



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 explanations for the incorrect answer. All the images referenced are sourced from ImageNet-1K and LAION-Aesthetic datasets.

近年来，多模态大语言模型（MLLMs）取得了显著进展。这些模型通过将图像信息整合到大型语言模型（LLMs）中，展现出在图像理解、视觉问答和指令遵循等任务上的强大能力。然而，尽管这些模型在某些任务上表现出色，但它们在视觉理解方面仍存在系统性的不足。

其中一些错误是比较离谱的，一个自然的问题是：这些错误的根源在哪里？它是视觉形态、语言理解或它们的对齐方面的缺陷吗？

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

CLIP盲对 (CLIP-blind pairs)

在这项工作中，作者认为，在MLLM中观察到的这些问题可能源于**视觉表示**。大多数MLLMs依赖于CLIP作为视觉编码器，所以本文从识别clip中的错误示例开始研究。

大多数MLLM 建立在预训练的视觉和语言模型上。这些模型使用Adapter来集成。一个自然的假设是，**预训练的视觉模型中的任何限制都可以级联到采用它们的下游MLLM中。**

CLIP盲对：是在视觉上存在明显差异，但在CLIP的嵌入空间中却被编码为相似的图像对。

作者通过比较CLIP和DINOv2（纯视觉模型）的嵌入来识别这些盲对。如果两个图像在CLIP嵌入空间中的相似度超过0.95，而在DINOv2嵌入空间中的相似度小于0.6，则认为这对图像是CLIP盲对。

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

构建MMVP Benchmark

基于CLIP盲对，作者构建了MMVP benchmark。该基准测试包含150对图像和300个问题，例如“狗是面向左边还是右边？”等问题。主要目标是确定当提出这些看似基本的问题并忽略关键的视觉细节时，MLLM模型是否会失败。

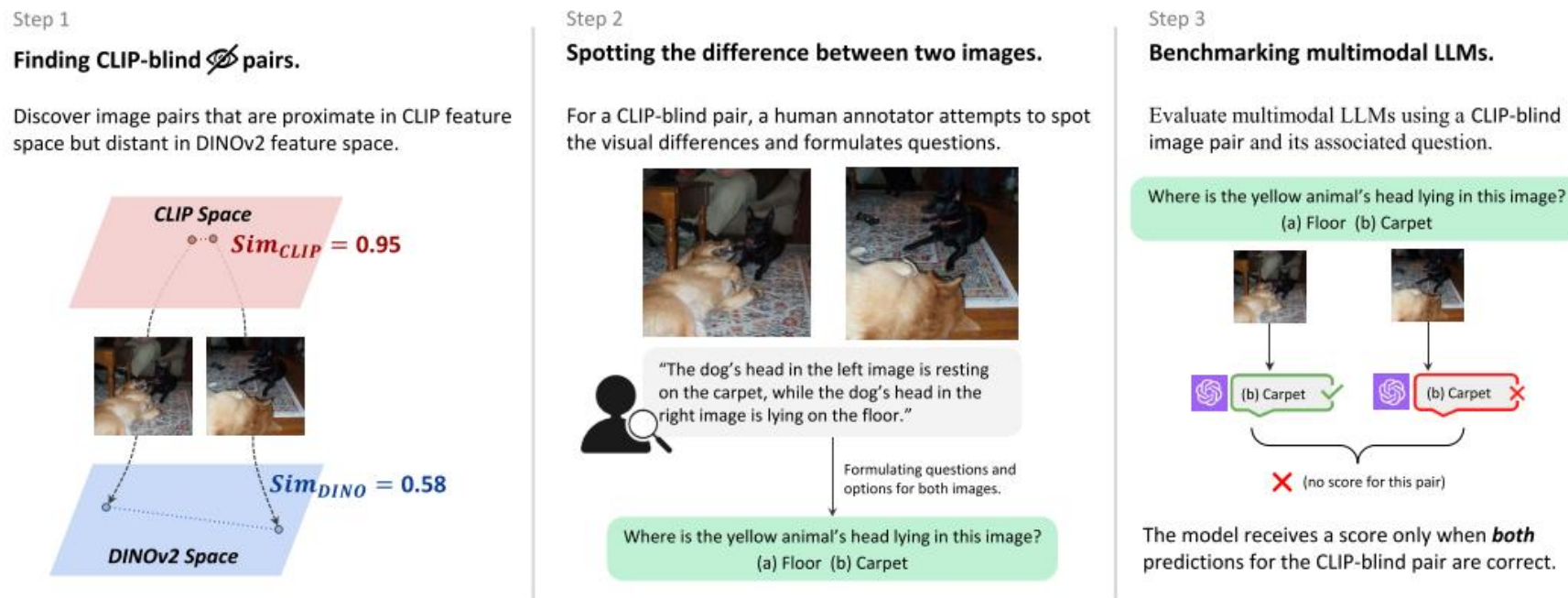


Figure 2. Constructing MMVP benchmark via CLIP-blind pairs. **Left:** We start with finding CLIP-blind pairs that have similar CLIP embedding but different DINOv2 embedding. **Center:** We manually inspect the differences between pair-wise images and formulate questions based on the differences in the 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.

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

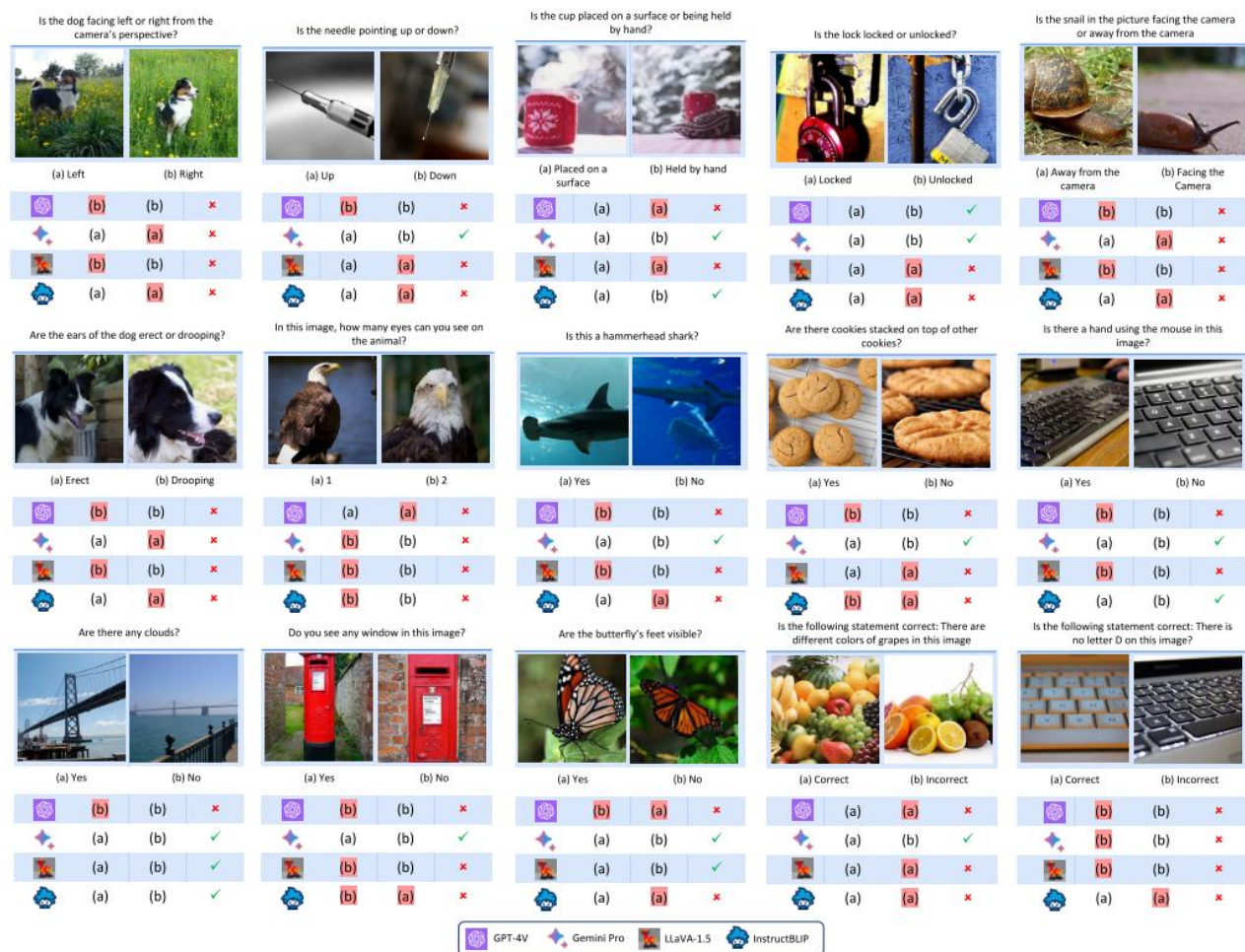


Figure 3. **Examples of Questions in the MMVP benchmark.** Incorrect answers are shaded in red. A model is considered correct only if it answers both questions in a pair correctly. Both leading closed-source models (GPT-4V, Gemini) and open-source models (LLaVA-1.5, InstructBLIP) fail these simple visual questions. (See Appendix B.2 for all the questions in MMVP benchmark.)

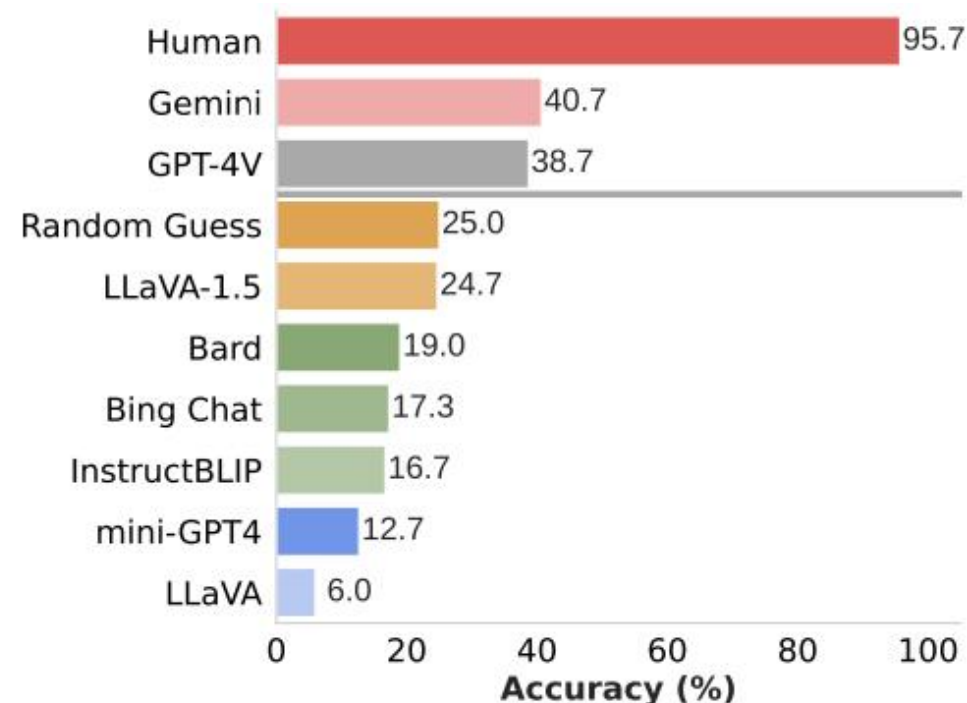


Figure 4. **Benchmark results of current SOTA MLLM models and humans.** We evaluate benchmark questions for current SOTA MLLM models and human performances through user studies.

大多数MLLMs的表现远低于人类水平。例如，LLaVA-1.5、InstructBLIP等模型的准确率低于随机猜测水平（25%），即使是性能较好的GPT-4V和Gemini模型，其准确率也远低于人类表现。

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

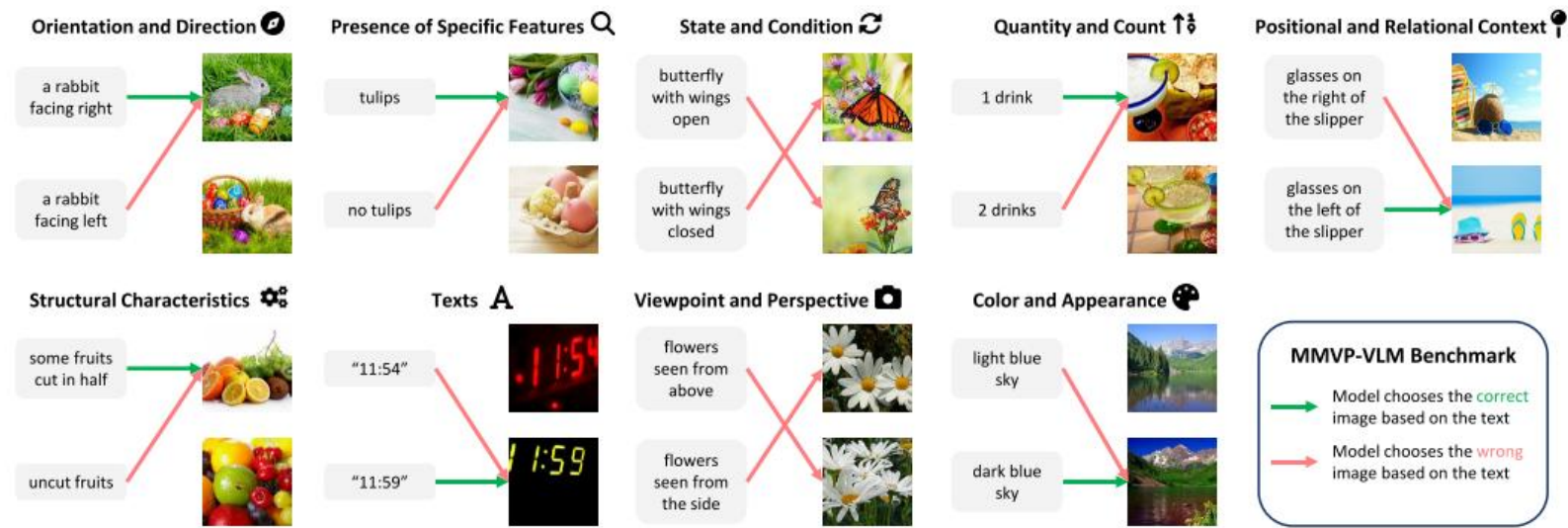


Figure 5. **Examples from MMVP-VLM.** MMVP-VLM consists of image pairs across nine visual patterns. The examples in the figure are from EVA01 ViT-g-14 model [54], one of the largest CLIP models that also fails to choose the right image given the text description.

	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

使用gpt-4来对模式进行了一个分类，分为九个类别。构建了一个新的 benchmark: mmvp-VLM。
mmvp-VLM: 9个模式，每个模式包括15对图像，旨在探究clip能否处理这些模式。

即使在大规模数据和模型尺寸扩展的情况下，CLIP模型在处理某些视觉模式时仍然存在困难。例如，在“方向”、“数量和计数”、“特定特征的存在”等模式上。这表明，仅靠数据和模型的扩展并不能解决CLIP模型在视觉理解上的根本问题。

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

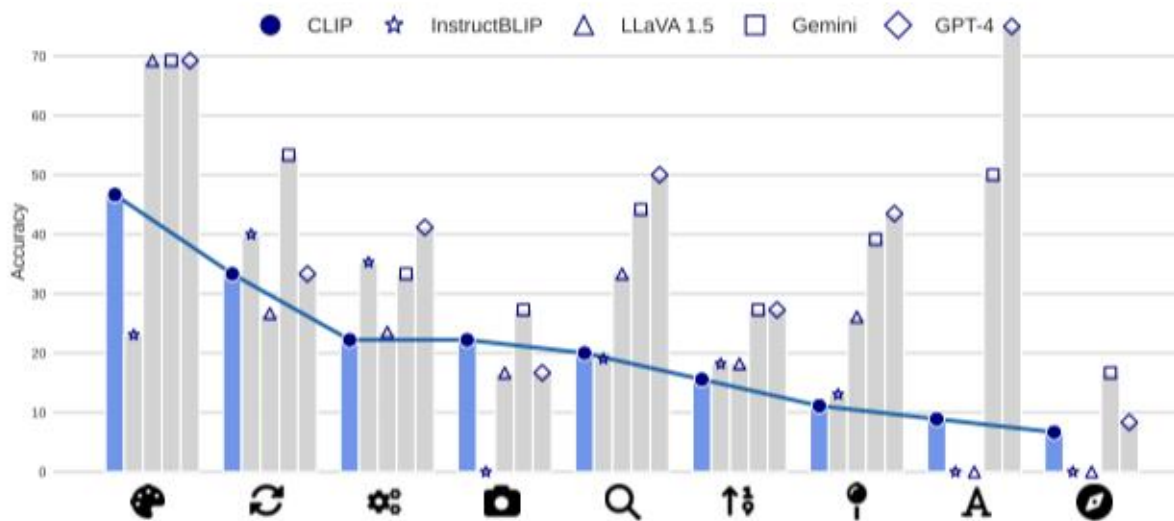



Figure 6. **CLIP and MLLM’s performance on visual patterns.** If CLIP performs poorly on a visual pattern such as “ orientation”, MLLMs also underperform on the visual pattern.

当CLIP视觉编码器在某个视觉模式上表现不佳时，MLLM往往会表现出类似的缺点。明确使用CLIP视觉编码器的开源模型，如LLaVA 1.5和InstructBLIP，在性能上显示出很强的相关性。

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

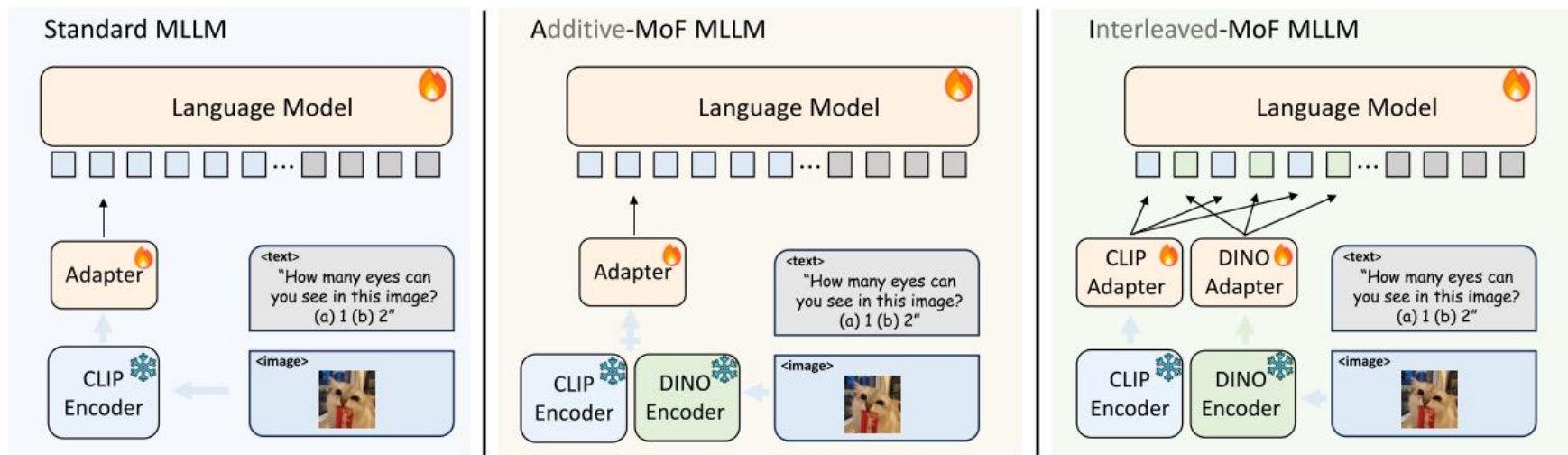


Figure 7. **Different Mixture-of-Feature (MoF) Strategies in MLLM.** *Left:* Standard MLLM that uses CLIP as *off-the-shelf* pretrained vision encoder; *Middle:* Additive-MoF (A-MoF) MLLM: Linearly mixing CLIP and DINOv2 features before the adapter; *Right:* Interleaved-MoF (I-MoF MLLM) Spatially interleaving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

method	SSL ratio	MMVP	LLaVA
LLaVA	0.0	5.5	81.8
	0.25	7.9 (+2.4)	79.4 (-2.4)
	0.5	12.0 (+6.5)	78.6 (-3.2)
LLaVA	0.625	15.0 (+9.5)	76.4 (-5.4)
+ A-MoF	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
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. **Empirical Results of Interleaved MoF.** Interleaved MoF improves visual grounding while maintaining same level of instruction following ability.

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

- 当前的视觉模型（如 CLIP）在多模态系统中存在局限性，可能成为瓶颈。
- 仅仅通过扩展数据量和模型规模无法解决这些局限性。
- 不同类型的视觉模型在不同的方面表现出色，需要结合它们的优势，或者修改模型架构。
- 需要开发新的评估指标来促进新的视觉表示学习算法的发展。