When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It?

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, James Zou
Stanford University
Stanford, CA 94305
{merty, fede, pkalluri, jurafsky, jamesz}@stanford.edu

ICLR 2023 oral

The Hard Positive Truth about Vision-Language Compositionality

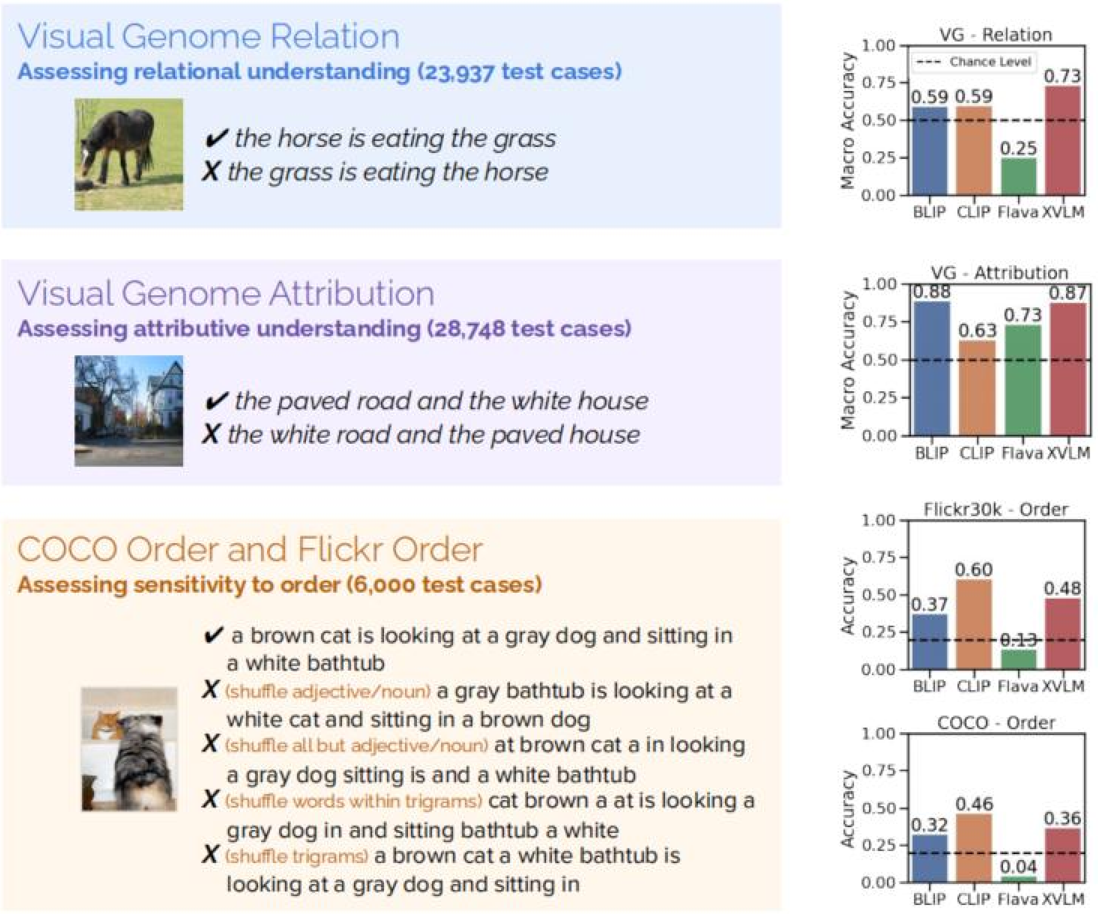
Amita Kamath[1,2] **Cheng-Yu Hsieh**[1] **Kai-Wei Chang**[2] **Ranjay Krishna**[1,3]
[1] University of Washington
[2] University of California, Los Angeles
[3] Allen Institute for AI
https://github.com/amitakamath/hard_positives

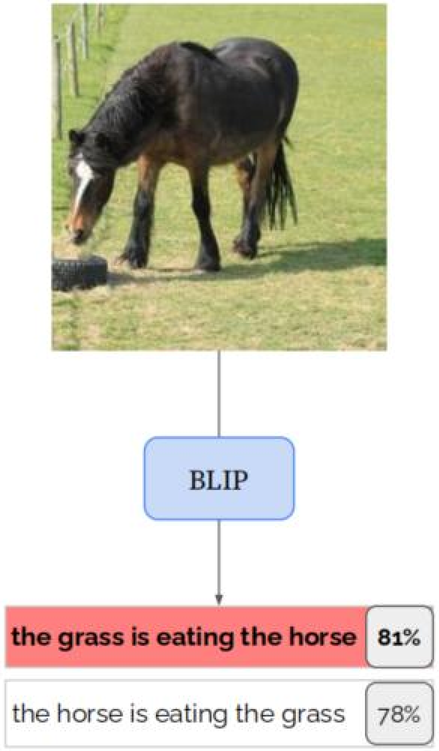
ECCV 2024

彭天天

2025/12/05

1.1 VLMs behave like Bags-of-Words?





1.大多数模型对于组合理解（Relation、Attribution）表现接近随机水平（XVLM训练数据包含VG）

2.模型对于格式正确的文本没有偏好 (只要词都对，它们几乎不在乎顺序是不是乱的)

3.先前一些研究结果发现用CLIP text encoder作为图像生成的文本条件效果不好——可能原因是CLIP无法有效编码语序

1.2 Why VLMs behave like Bags-of-Words?

问题出在 预训练的目标——
不理解语序/关系，也能完成检索任务
不理解图像的空间结构，也能完成检索任务（于是模型选择走捷径）

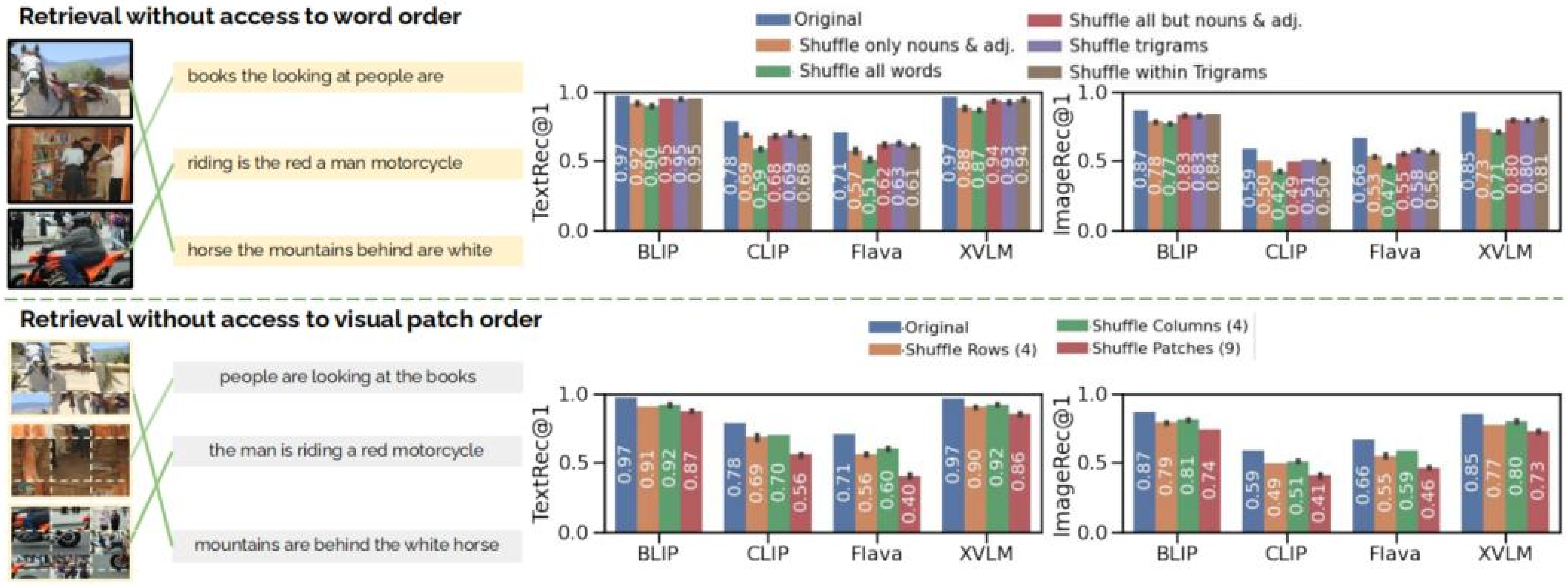


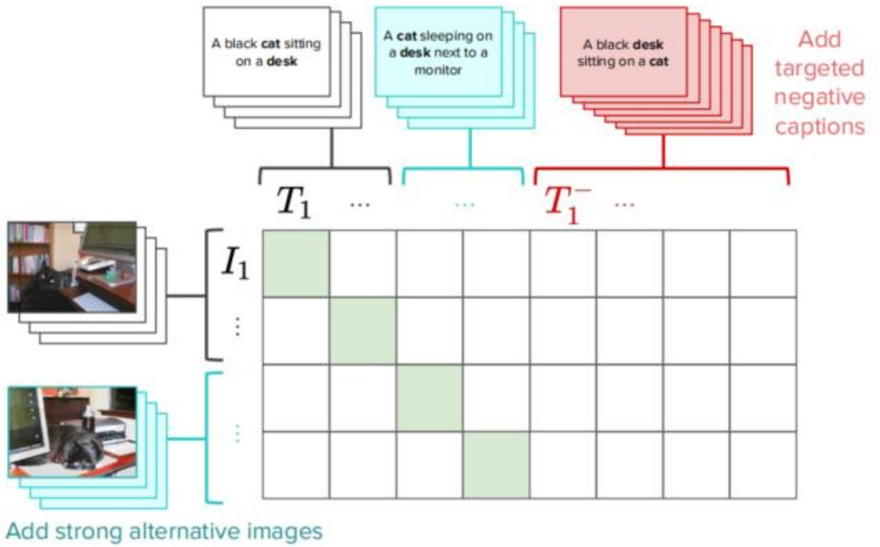
Figure 2: **Retrieval without access to order information.** We show that models can achieve substantially high performance on standard evaluations even when order information is removed. In particular, in datasets where the captions are augmented with order perturbations, models show marginal performance degradation.

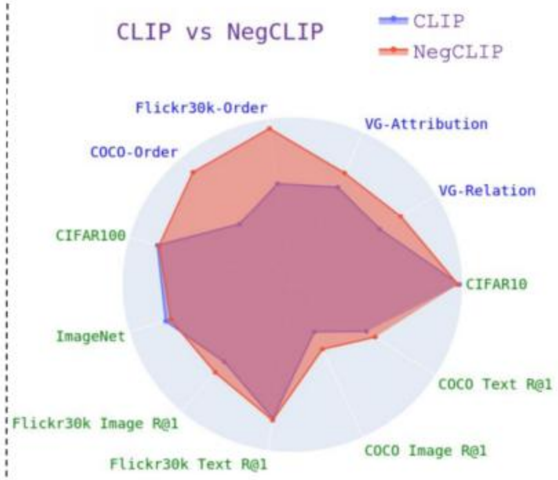
训练数据中缺乏包含相近物体并且需要顺序进行区分的样本（hard negatives）

1.3 A simple fix: NegCLIP

在COCO上加入hard negatives数据微调CLIP得到NegCLIP
直接在COCO上进行微调得到CLIP-FT

在几乎不破坏下游任务表现的前提下，提高了模型对于属性、关系、顺序的理解





	CLIP	CLIP-FT	NegCLIP
Compositional Tasks			
VG-Relation	0.59	0.63	0.81
VG-Attribution	0.62	0.65	0.71
Flickr30k-PRC	0.59	0.50	0.91
COCO-PRC	0.46	0.36	0.86
Downstream Tasks			
CIFAR10	0.95	0.95	0.94
CIFAR100	0.80	0.80	0.79
ImageNet	0.75	0.74	0.72
Flickr30k Image R@1	0.59	0.67	0.67
Flickr30k Text R@1	0.78	0.83	0.79
COCO Image R@1	0.30	0.42	0.41
COCO Text R@1	0.50	0.59	0.56

2.1 Only using hard negatives resulting oversensitive

Existing work

Image i

	Captions	CLIP	Hard Negative Finetuned	Ours
Original Caption c	brown grass	0.236	0.152	0.240
Hard Negative c_n	blue grass	0.240	0.143	0.231
Hard Positive c_p	chestnut grass	0.249	0.134	0.241

Our work

先前的工作大多只用hard negative来微调CLIP——让模型在“该降分时降降分”
而忽略了hard positive——“不该降分时不能误降分”
于是本文又构建了一个同时包含c_n和c_p的新数据集：

2.2 Results

Orig：只含c和cn，Test Acc为正确分类的比例：$s(c|i) > s(c_n|i)$

Aug：包含c、c_n、cp，Test Acc：$s(c|i) > s(c_n|i)$ and $s(c_p|i) > s(c_n|i)$

Brittleness： $s(c|i) > s(c_n|i) > s(c_p|i)$ or

$$s(c_p|i) > s(c_n|i) > s(c|i)$$

	Model	REPLACE Orig. Test Acc.	REPLACE Aug. Test Acc.	SWAP Orig. Test Acc.	SWAP Aug. Test Acc.	REPLACE Brittleness (↓)	SWAP Brittleness(↓)
(a)	CLIP ViT-B/32	61.6	46.8 (-14.9)	60.5	49.6 (-10.9)	23.2	21.7
	NegCLIP	68.6	52.1 (-16.6)	70.9	56.7 (-14.2)	21.5	26.4
	CREPE-Swap	63.5	50.4 (-13.1)	70.6	56.7 (-13.9)	**19.8**	26.0
	CREPE-Replace	73.7	53.9 (-19.8)	71.1	57.7 (-13.4)	23.9	25.4
(b)	SVLC	76.6	44.5 (-32.1)	72.4	**61.6** (-10.9)	39.9	**20.8**
	SVLC+Pos	64.3	45.0 (-19.3)	56.5	45.4 (-11.1)	29.8	22.8
	DAC-LLM	87.6	48.9 (-38.7)	72.0	61.1 (-10.9)	40.1	21.6
	DAC-SAM	86.9	**55.9** (-31.0)	69.5	56.5 (-13.0)	32.5	25.6
	Our HN	73.9	55.7 (-18.2)	74.3	60.5 (-13.8)	21.0	25.1
(c)	Our HP+HN	69.0	**58.0** (-11.0)	73.2	**61.1** (-12.1)	**16.9**	**22.9**
	Our HP+HN (Swap-only)	63.9	51.6 (-12.3)	73.0	**61.9** (-11.2)	18.6	**21.2**
(d)	Our HP+HN (Replace-only)	70.9	**59.0** (-11.9)	69.7	55.6 (-14.1)	**17.8**	26.5
	Random Chance	50.0	33.3	50.0	33.3	33.3	33.3
	Human Estimate	97	97	100	100	0	0
	0 HN	58.5	49.8 (-8.6)	64.1	51.2 (-12.9)	**15.8**	25.0
	0.25 HN	66.0	55.5 (-10.5)	71.6	59.8 (-11.8)	16.6	22.8
(b)	0.50 HN	67.3	56.9 (-10.5)	72.5	60.5 (-12.0)	16.4	22.8
	0.75 HN	68.2	**57.6** (-10.6)	72.9	**61.0** (-11.9)	16.6	**22.7**

1. CLIP本身区分HP就有困难

2. 只基于HN进行微调，虽然在Orig上表现良好，但实际上破坏了模型对于HP的检测——模型只学会了 “检测扰动” 而不是 “理解语义”——所有被扰动的样本都是负样本

3.只基于HN进行微调——所有被扰动的样本都是正样本——会降低模型在Orig上的表现。

2.2 Results

北京大学 PEKING UNIVERSITY

在HN、HP上微调，会削弱模型在其它基础任务上的表现

Mean c Score	CLIP	Neg-CLIP	CREPE -Swap	CREPE -Repl.	SVLC	SVLC +Pos	DAC -LLM	DAC -SAM	Ours
REPL.	0.234	0.225	0.233	0.214	0.202	0.223	0.157	0.228	0.231
SWAP	0.255	0.239	0.250	0.228	0.211	0.228	0.132	0.224	0.247

Table 9: Mean image-text matching score of original caption c per benchmark of all evaluated models. All hard negative-finetuned models reduce the image-text matching score of c, nearly all more so than our model finetuned on both hard negatives and hard positives.

		ImageNet1k		COCO		Flickr30k		VTAB	
	Model	Acc@1	Acc@5	Image Recall@1	Text Recall@1	Image Recall@1	Text Recall@1	Acc@1	Acc@5
(a)	CLIP ViT-B/32	63.33	88.83	30.46	50.14	58.82	77.40	39.00	70.90
(b)	CLIP-COCO	53.18	81.98	50.34	66.76	68.48	83.40	34.67	68.55
(c)	Our HN	50.40	79.58	49.61	63.98	67.80	80.10	32.40	67.53
	Our HP+HN	49.85	79.70	49.67	65.02	67.52	80.60	33.24	67.75

1. 检索任务上，所有方法相比于直接微调CLIP，均有下降，论文所提方法（HN+HP共用）下降最少

2.分类任务上，论文所提方法均出现了性能下降

Is CLIP ideal? No. Can we fix it? Yes!

Raphi Kang Yue Song Gerogia Gkioxari Pietro Perona
California Institute of Technology
{rkang, yuesong, georgia, perona}@caltech.edu

ICCV 2025

Exploring How Generative MLLMs Perceive More Than CLIP with the Same Vision Encoder

Siting Li, Pang Wei Koh, Simon Shaolei Du
University of Washington
{sitingli,pangwei,ssdu}@cs.washington.edu

ACL 2025

彭天天

2025/12/05

3.1 Is CLIP ideal?

北京大学 PEKING UNIVERSITY

理想的CLIP模型：使用编码器 i(·)，t(·)将文本和图像映射至共享的嵌入空间中，用不同的方向表示不同的内容，嵌入之间的相似度能反映原始输入之间的关系

理想中我们希望CLIP满足：

Condition 1. (Concept Categorization) Satisfaction of this condition requires that (1.1) *C* represents basic descriptions and image content.

$$\mathbf{i}(x)\cdot\mathbf{t}(x) > \mathbf{i}(x)\cdot\mathbf{t}(y)$$
$$\mathbf{i}(x,y)\cdot\mathbf{t}(x) > \mathbf{i}(x,y)\cdot\mathbf{t}(z) \quad \forall\, x,y,z \in \mathbb{V}$$

(1.2) Images that contain the same semantic concept(s) but differ due to an attribute or scene composition, should have higher cosine similarity with each other than with an image that contains a different set of semantic concepts.

$$\mathbf{i}(x_a)\cdot\mathbf{i}(x_b) > \mathbf{i}(x_a)\cdot\mathbf{i}(y)$$
$$\mathbf{i}(x,g_1^{<loc>})\cdot\mathbf{i}(x,g_2^{<loc>}) > \mathbf{i}(x)\cdot\mathbf{i}(y)$$

Condition 2. (Attribute Binding) *C* respects attribute binding. More specifically: (2.1) concepts with different attributes are not parallel in CLIP space.

$$\mathbf{i}(x_a)\cdot\mathbf{i}(x_b) < 1 \quad \forall a,b \in \mathbb{A}$$

(2.2) Images representing a concept with a specific attribute are closer in CLIP space to its text embedding.

$$\mathbf{i}(x_a)\cdot\mathbf{t}(a) > \mathbf{i}(x_b)\cdot\mathbf{t}(a)$$

(2.3) Images with the same concepts and attributes present but in different pairings are not parallel in CLIP space.

$$\mathbf{i}(x_a,y_b)\cdot\mathbf{i}(x_b,y_a) < 1$$

Condition 3. (Spatial Relationship) C respects spatial locations or relationships of objects. This requires that (3.1) images where the same object is in a different location must not have identical embeddings.

$$\mathbf{i}(x,g_1^{<loc>})\cdot\mathbf{i}(x,g_2^{<loc>}) < 1, \quad \forall g_1^{<loc>},g_2^{<loc>} \in \mathbb{G}$$

(3.2) Images with the same objects but in different spatial relationships must not have identical embeddings.

$$\mathbf{i}(x,g_3^{<rel>},y)\cdot\mathbf{i}(x,g_4^{<rel>},y) < 1, \quad \forall g_3^{<rel>},g_4^{<rel>} \in \mathbb{G}$$

(3.3) Images where an object is in the same location or relationship must be semantically closer than images where it is in a different location or relationship.

$$\mathbf{i}(x,g_1,y)\cdot\mathbf{i}(x,g_1,z) > \mathbf{i}(x,g_1,y)\cdot\mathbf{i}(x,g_2,z)$$

Condition 4. (Negation) *C* respects negation. This requires that (4.1) texts and their negated counterparts must have a similarity score lower than any other pairs.

$$\mathbf{t}(x)\cdot\mathbf{t}(\neg x) < \mathbf{t}(y)\cdot\mathbf{t}(\neg x), \quad \forall\, x,y \in \mathbb{T}$$

(4.2) An image with some concept must have a lower similarity score with the negated concept text than another text.

$$\mathbf{i}(x)\cdot\mathbf{t}(\neg x) < \mathbf{i}(x)\cdot\mathbf{t}(y)$$

3.2 Proof

理想的CLIP模型：使用编码器 i(·)，t(·)将文本和图像映射至共享的嵌入空间中，用不同的方向表示不同的内容，嵌入之间的相似度能反映原始输入之间的关系

实际上CLIP的这种用**一个**高维embedding表示图像或文本的方式无法满足我们的需求。

证明：
假设满足condition1，那么i(·)，t(·)应满足：

$$\mathbf{i}(x^1,x^2) = \operatorname*{argmax}_{\mathbf{i}(x^1,x^2)} \Big[\mathbf{i}(x^1,x^2)\cdot\mathbf{i}(x^1) + \mathbf{i}(x^1,x^2)\cdot\mathbf{i}(x^2) - \sum_{j=3}^{M} \mathbf{i}(x^1,x^2)\cdot\mathbf{i}(x^j)\Big] \quad \text{s.t.} \quad \|\mathbf{i}(x^1,x^2)\| = 1 \tag{1}$$

Here, the first two terms guide the local placement of $\mathbf{i}(x^1,x^2)$, while the last term introduces a global constraint to avoid proximity to other embeddings. The constraint ensures that all embeddings must lie on the unit hypersphere. We can expand the sum to see that:

变形得到：

$$\mathbf{i}(x^1,x^2) = \operatorname*{argmax}_{\mathbf{i}(x^1,x^2)} \Big[\mathbf{i}(x^1,x^2) \cdot \Big(\mathbf{i}(x^1)+\mathbf{i}(x^2)+\mathbf{i}(x^1)+\mathbf{i}(x^2) - \sum_{j=1}^{M} \mathbf{i}(x^j)\Big)\Big] \tag{2}$$

Since random vectors in high dimensions will be approximately symmetrically distributed, $\sum_{j=1}^{M} \mathbf{i}(x^j) \approx 0$. The optimum is then reached when $\mathbf{i}(x^1,x^2)$ is parallel to $\mathbf{i}(x^1)+\mathbf{i}(x^2)$. Thus we see $\mathbf{i}(x^1,x^2)$ is a normalized superposition of $\mathbf{i}(x^1)$ and $\mathbf{i}(x^2)$, and lies on the geodesic arc between $\mathbf{i}(x^1)$ and $\mathbf{i}(x^2)$ on the hypersphere, i.e.,

$$\boxed{\mathbf{i}(x^1,x^2) = \frac{\mathbf{i}(x^1)+\mathbf{i}(x^2)}{\|\mathbf{i}(x^1)+\mathbf{i}(x^2)\|}} \tag{3}$$

□

后续推导....得到：

$$\boxed{\begin{aligned}\mathbf{i}(x_a,y_b) &= \frac{(1-\delta)(\mathbf{i}(x)+\mathbf{i}(y)) + p\mathbf{t}(a) + q\mathbf{t}(b)}{2} \\ &= \mathbf{i}(x_b,y_a)\end{aligned}} \tag{9}$$

即CLIP如果满足condition1，则无法区分图像 (x_a,y_b)
和图像(x_b,y_a)——就无法满足condition2、3、4，存在矛盾

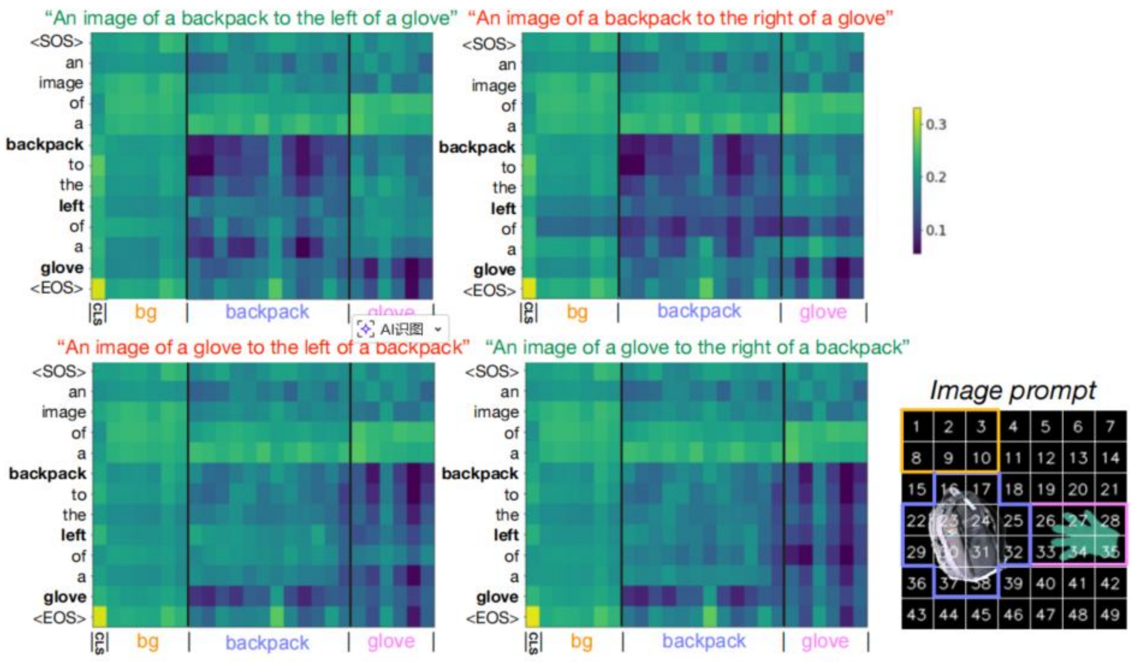
问题：

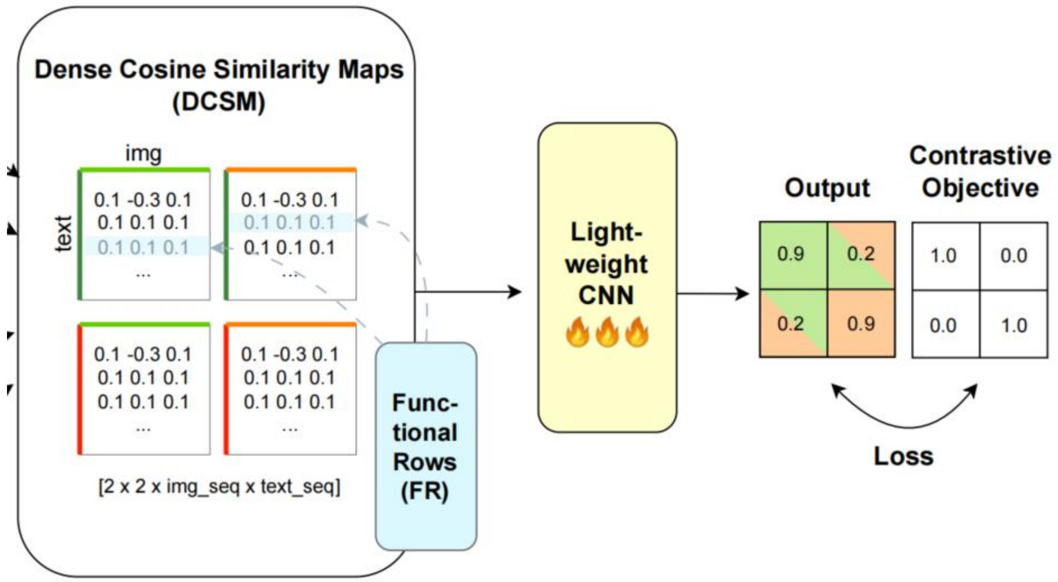
1. $\sum_{j=1}^{M} i(x^j)$ 不为0（存在锥效应）
2. 其它证明步骤引入了太多主观判断

感觉可以再探索探索

3.3 How to fix? CNN for Dense Cosine Similarity Maps

以前的CLIP：只取 [CLS] 和 [EOS] token代表图像和文本，计算余弦相似度
论文方法：所有token都算，得到一个相似度矩阵，然后训一个CNN处理这个矩阵

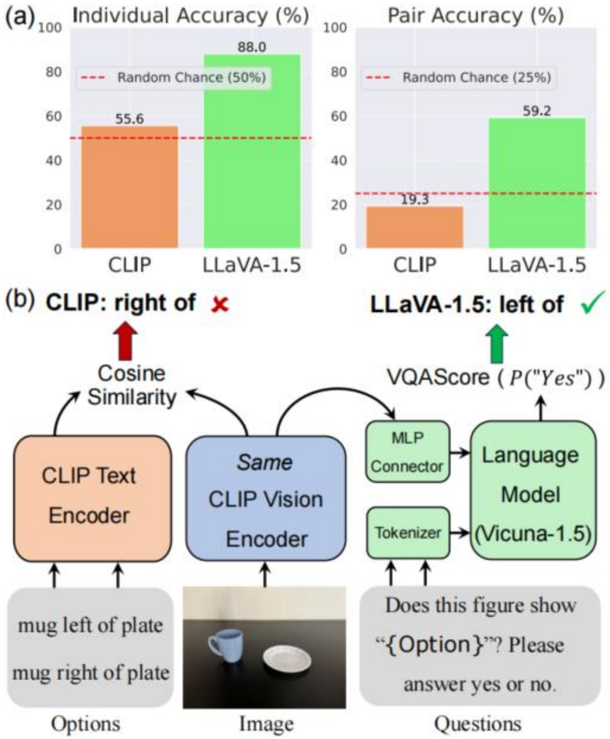


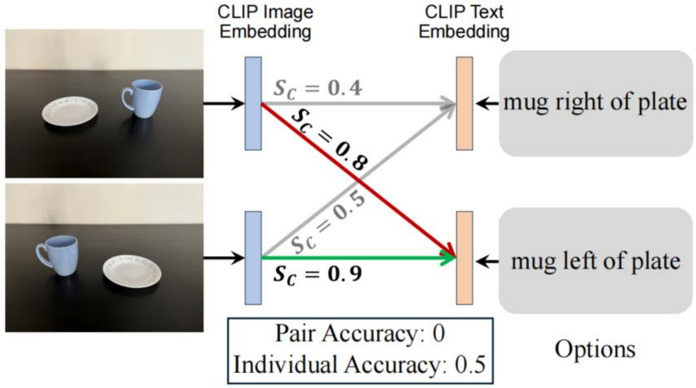


存在问题：其它 token 是否可靠？它们是否真的携带语义？

4.1 Genrative MLLMs vs CLIP

现象：即使使用相同的视觉编码器，MLLMs在各种细粒度多模态任务（空间关系、组合推理等）下总是比CLIP强很多





	What'sUp Subset A				What'sUp Subset B			
	Left/Right		On/Under		Left/Right		Front/Behind	
	Indiv.	Pairs	Indiv.	Pairs	Indiv.	Pairs	Indiv.	Pairs
CLIP-ViT-L/14-336px	49.0	1.9	61.7	23.3	54.9	10.8	51.5	7.8
LLaVA-1.5-7B	96.6	93.2	76.2	52.4	98.5	97.1	**76.0**	**52.9**
Phi-3-V-3.8B	97.6	95.1	78.6	58.3	**100**	**100**	61.8	26.5
LLaMA-3-V-8B	**98.1**	**96.1**	**81.1**	**64.1**	**100**	**100**	73.0	47.1
Random chance	50.0	25.0	50.0	25.0	50.0	25.0	50.0	25.0

Table 1: The two-way individual accuracy and pair accuracy of CLIP-ViT-L/14-336px and Generative MLLMs in percentage points on four subsets of What'sUp. Generative MLLMs outperform CLIP by a large margin.

	Winoground	NaturalBench-R	MMVP	MMVP-VLM
CLIP-ViT-L/14-336px	27.8	47.8	14.0	20.7
LLaVA-1.5-7B	39.8	52.2	36.0	**49.6**
Phi-3-V-3.8B	35.8	50.5	30.7	31.9
LLaMA-3-V-8B	**46.3**	**64.7**	**50.0**	**49.6**
Random chance	25.0	25.0	25.0	25.0

Table 2: The pair accuracy of CLIP-ViT-L/14-336px and Generative MLLMs in percentage points on several paired benchmarks. Generative MLLMs achieve substantially better performance than CLIP.

可能原因：
1.Training data
2.Token usage and position embedding and Language Models
3.Architecture design, Training objective and prompt.

4.2 Is training data?

用LLaVA-1.5的训练数据微调CLIP（冻结视觉编码器），结果几乎毫无变化

	What'sUp Subset A		What'sUp Subset B	
	Indiv.	Pairs	Indiv.	Pairs
CLIP	49.0	1.9	54.9	10.8
+ finetuning (ft)	50.5	1.9	53.9	5.9
+ ft + hard neg.	50.5	1.0	50.5	1.0
SigLIP	50.0	1.9	51.5	5.9
+ finetuning (ft)	49.0	1.0	51.0	3.9
+ ft + hard neg.	50.0	0.0	50.0	0.0
EVA-CLIP	49.0	1.0	50.1	4.9
+ finetuning (ft)	50.0	4.9	48.5	2.0
+ ft + hard neg.	50.0	1.9	48.0	2.0
Random chance	50.0	25.0	50.0	25.0

4.3 Is token usage or position embedding or language models?

北京大学 PEKING UNIVERSITY

1. 对于LLaVA，只使用视觉编码器的[CLS] token（与CLIP一样），用LoRA微调，发现掉点特别多：

	What'sUp Subset A				What'sUp Subset B			
	Left/Right		On/Under		Left/Right		Front/Behind	
	Indiv.	Pairs	Indiv.	Pairs	Indiv.	Pairs	Indiv.	Pairs
LLaVA-1.5-7B-LoRA	**84.5**	**68.9**	**76.2**	**52.4**	**89.2**	**78.4**	**86.3**	**72.5**
[CLS]-LLaVA-1.5-7B-LoRA	44.2	8.7	54.4	8.7	49.0	4.9	53.9	12.7
Random chance	50.0	25.0	50.0	25.0	50.0	25.0	50.0	25.0

Table 5: The results of [CLS]-LLaVA-1.5-7B-LoRA and reproduced LLaVA-1.5-7B-LoRA on all subsets of What'sUp, where [CLS]-LLaVA-1.5-7B-LoRA struggles with spatial reasoning.

2. 对于CLIP，使用所有视觉patch tokens，加权融合（PACL）得到一个embedding（有点用）
3. 在CLIP上加RoPE（有点用）
4. 通过SPARC方法利用多个text tokens （没用，可能原因是太难训/text encoder太弱）
5. 换更强的text encoder （没用，说明更强的text encoder也提取不到更多的信息）

$$s(\mathbf{x}, \mathbf{y}) = e_v(f_v(\mathbf{x})) \cdot e_t(f_t(\mathbf{y}))$$

$$\mathbf{v}(\mathbf{x}) = e_v(f_v(\mathbf{x}))^\top \cdot \mathrm{sigmoid}(10 \cdot s(\mathbf{x}, \mathbf{y}))$$

	What'sUp Subset A				What'sUp Subset B			
	Left/Right		On/Under		Left/Right		Front/Behind	
	Indiv.	Pairs	Indiv.	Pairs	Indiv.	Pairs	Indiv.	Pairs
CLIP-ViT-L/14-336px	49.0	1.9	**61.7**	**23.3**	**54.9**	10.8	51.5	7.8
+ Patch Tokens (PT)	47.6	9.7	52.9	10.7	52.9	9.8	51.5	6.9
+ PT + RoPE	**54.9**	**22.3**	46.1	13.6	52.0	**20.6**	45.6	12.7
+ PT + RoPE + Multiple Text Tokens	48.1	0.0	50.0	2.9	50.0	6.9	48.0	7.8
+ PT + RoPE + Stronger Text Encoder	50.5	10.7	48.5	6.8	50.0	15.7	**53.9**	**21.6**
LLM2CLIP (Huang et al., 2024)	49.5	1.0	58.7	17.4	49.0	1.0	55.4	14.7
Random chance	50.0	25.0	50.0	25.0	50.0	25.0	50.0	25.0

4.4 Is architecture design or training objective or prompt?

尝试从LLaVA中提取embedding，用于余弦相似度计算配对（VLM2Vec）

1. 图像+问题输入：” Represent the given image with the following question: {Question}”
2. 纯问题文本输入：” Find the text that can answer the given query: {Question}”

提取最后一个token的最后一层向量作为output embedding

用LoRA+对比学习微调LLaVA，结果很好——说明文本生成+自回归并非解决细粒度视觉推理的唯一方案

	What'sUp Subset A				What'sUp Subset B			
	Left/Right		On/Under		Left/Right		Front/Behind	
	Indiv.	Pairs	Indiv.	Pairs	Indiv.	Pairs	Indiv.	Pairs
CLIP-ViT-L/14-336px	49.0	1.9	61.7	23.3	54.9	10.8	51.5	7.8
LLaVA-1.5-7B-VLM2Vec-LoRA	**97.1**	**95.1**	**68.0**	**35.9**	**100**	**100**	**60.8**	**22.5**
w/o Question in Prompt	49.5	0.0	50.5	1.9	46.6	2.0	50.5	1.0
Random chance	50.0	25.0	50.0	25.0	50.0	25.0	50.0	25.0

	MMVP	MMVP-VLM
CLIP-ViT-L/14-336px	14.0	20.7
LLaVA-1.5-7B-VLM2Vec-LoRA	**30.0**	**37.8**
w/o Question in Prompt	9.3	11.9
Random chance	25.0	25.0

但是，如果修改为”Represent the given image”,prompt中不带question，则性能跌回CLIP水平
——文本能够引导特定细粒度视觉信息提取