Beyond Semantics: Rediscovering Spatial Awareness in Vision-Language Models

Arxiv 2025

提纲

- 作者介绍
- 研究背景
- 研究方法
- 实验效果
-  总结&思考

作者介绍

Jianing Qi

关注

PhD student, CUNY Grad Center
在 gradcenter.cuny.edu 的电子邮件经过验证
AI CV

创建我的个人资料

标题	引用次数	年份
Exploring an affective and responsive virtual environment to improve remote learning J Qi, H Tang, Z Zhu Virtual Worlds 2 (1)	9	2023
VerifierQ: Enhancing LLM Test Time Compute with Q-Learning-based Verifiers J Qi, H Tang, Z Zhu arXiv preprint arXiv:2410.08048	2	2024
Beyond Semantics: Rediscovering Spatial Awareness in Vision-Language Models J Qi, J Liu, H Tang, Z Zhu arXiv preprint arXiv:2503.17349		2025

引用次数

	总计	2020 年至今
引用	11	11
h 指数	2	2
i10 指数	0	0

Zhigang Zhu

关注

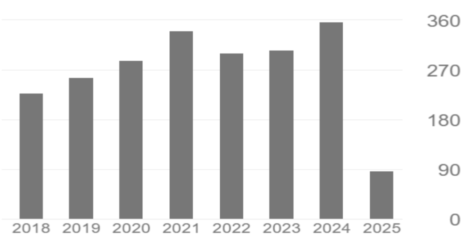
Herbert G. Kayser Professor of Computer Science, CUNY City College and Graduate Center
在 cs.ccny.cuny.edu 的电子邮件经过验证 - 首页
Computer vision multimodal sensing human-computer interaction assistive technology

创建我的个人资料

标题	引用次数	年份
Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing W Li, F Abtahi, Z Zhu Proceedings of the IEEE conference on computer vision and pattern …	204	2017
Eac-net: Deep nets with enhancing and cropping for facial action unit detection W Li, F Abtahi, Z Zhu, L Yin IEEE transactions on pattern analysis and machine intelligence 40 (11), 2583 …	161	2018
VISATRAM: A real-time vision system for automatic traffic monitoring Z Zhu, G Xu, B Yang, D Shi, X Lin Image and Vision Computing 18 (10), 781-794	155	2000

引用次数 查看全部

	总计	2020 年至今
引用	4438	1677
h 指数	34	20
i10 指数	102	39



提纲

研究背景

- MLLM在空间推理效果不好
  - Visual prompt: sft

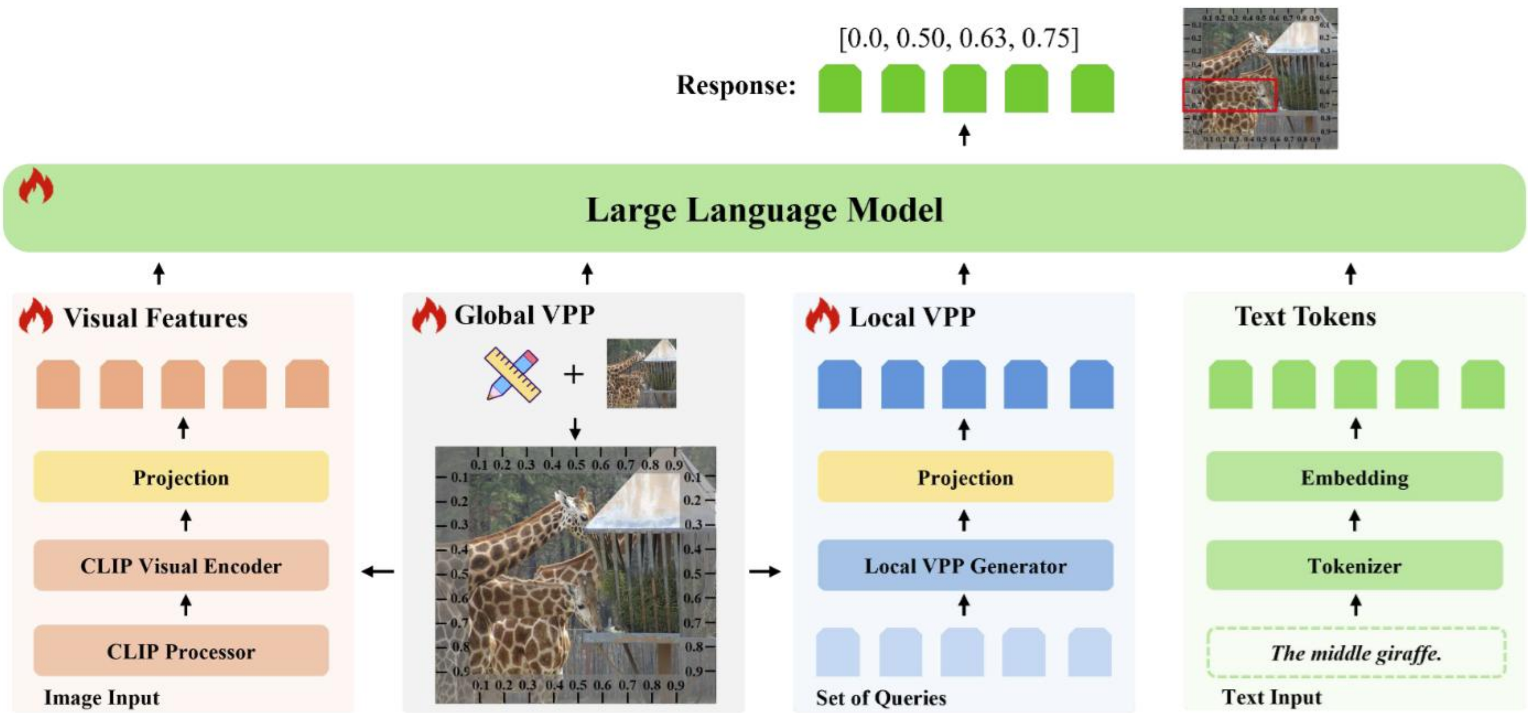


Fig. 2. An illustration of VPP-LLaVA, an MLLM-based visual grounding framework with Visual Position Prompt (VPP). We utilize the global VPP to provide a global position reference for MLLMs with foundational spatial cues. Additionally, a local VPP, serving as a local position reference, is introduced to further enhance and incorporate object spatial information. For brevity, some text instructions are omitted.

Tang W, Sun Y, Gu Q, et al. Visual Position Prompt for MLLM based Visual Grounding[J]. arXiv preprint arXiv:2503.15426, 2025.

研究背景

- MLLM在空间推理效果不好
  - 数据构造

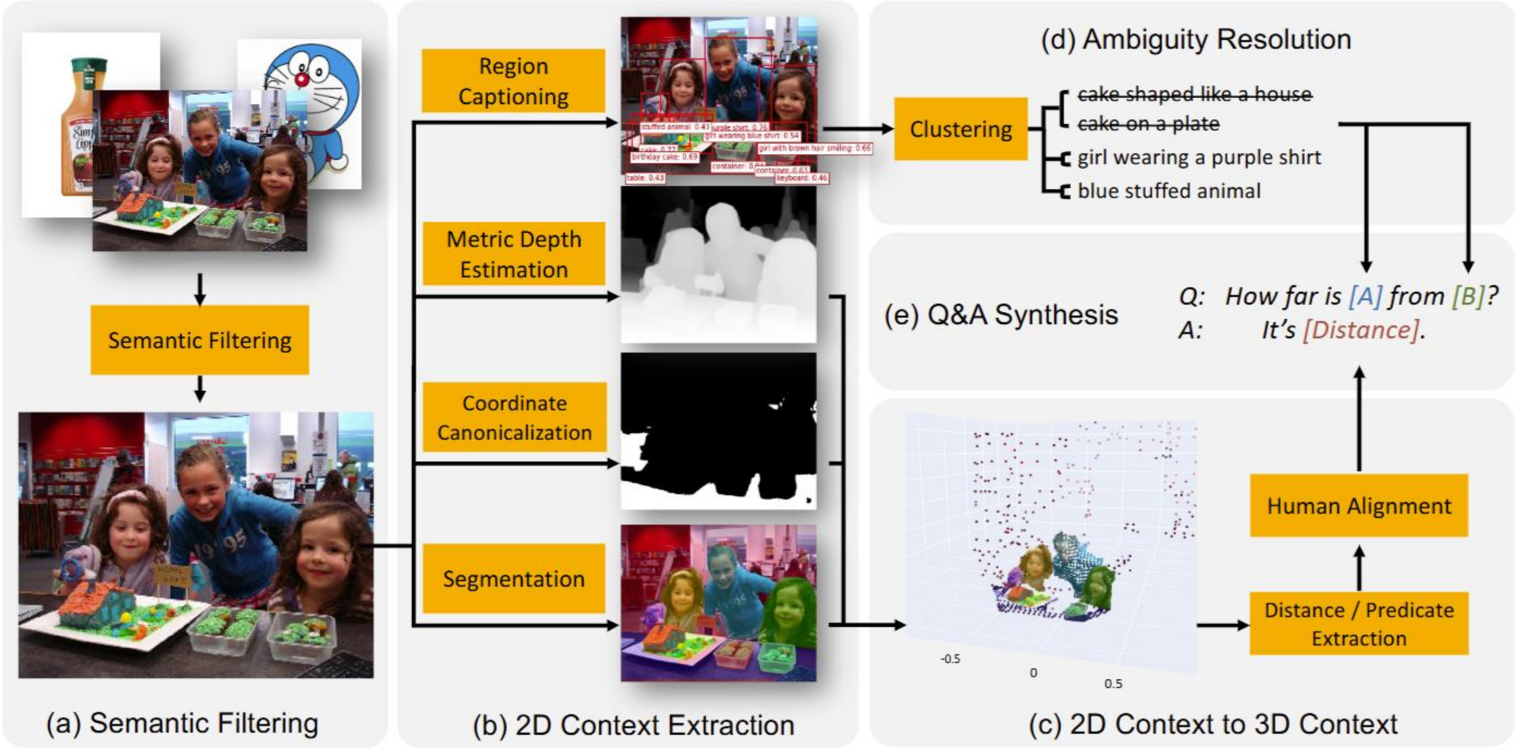
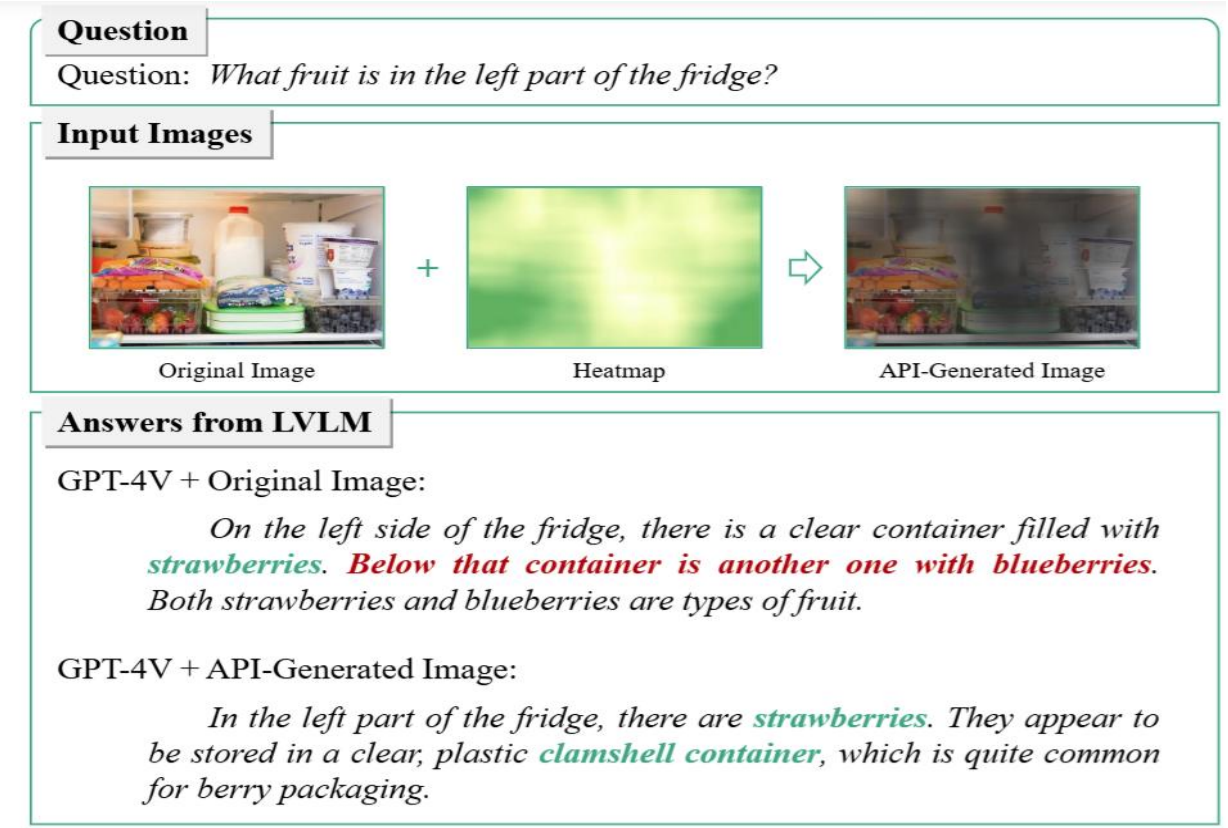


Figure 2. **An overview of our data synthesis pipeline.** (a) We use CLIP to filter noisy internet images and only keep scene-level photos. (b) We apply pre-trained expert models on internet-scale images so that we get object-centric segmentation, depth and caption. (c) We lift the 2D image into 3D point clouds, which can be parsed by shape analysis rules to extract useful properties like 3D bounding box. (d) We avoid asking ambiguous questions by clustering object captions using CLIP similarity score (e) We synthesize millions of spatial question and answers from object captions

研究背景

- MLLM在空间推理效果不好
  - Visual prompt: training-free



Yu R, Yu W, Wang X. Attention prompting on image for large vision-language models[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 251-268.

研究背景

- MLLM在空间推理效果不好
  - Vision encoder：组合encoder

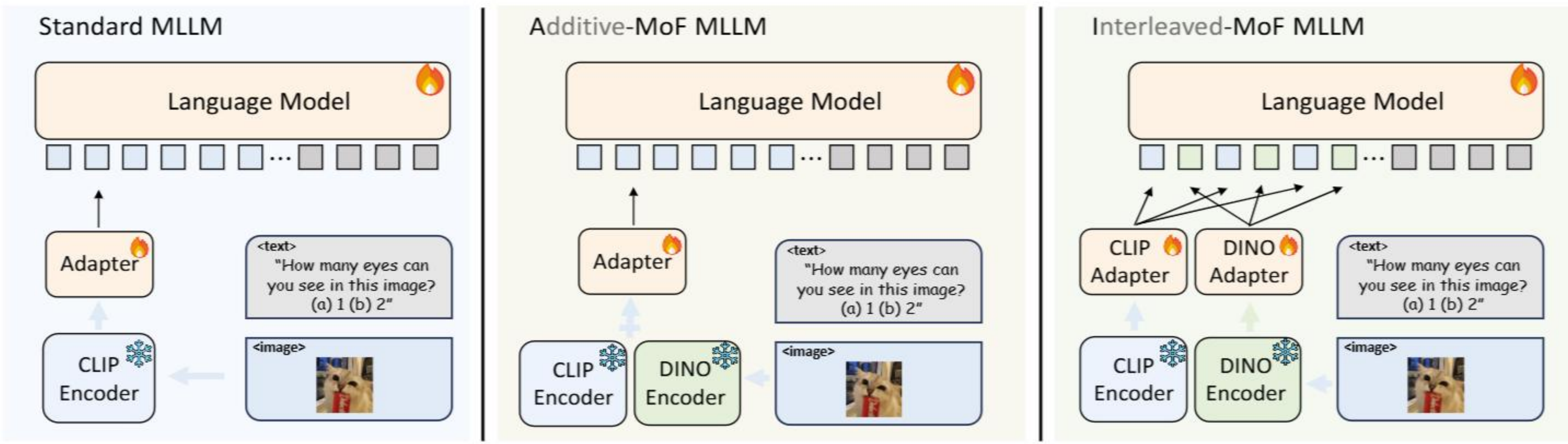


Figure 7. **Different Mixture-of-Feature (MoF) Strategies in MLLM.** *Left*: Standard MLLM that uses CLIP as *off-the-shelf* pretrained vision encoder; *Middle*: Additive-MoF (A-MoF) MLLM: Linearly mixing CLIP and DINOv2 features before the adapter; *Right*: Interleaved-MoF (I-MoF MLLM) Spatially interleaving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

研究背景

- 作者假设：位置编码没有起作用
  - 实验1：token乱序测试

Dataset	Original	Permutation	Difference
VQAv2	78.2	77.35	-0.85
POPE	87.3	87.10	-0.2
GQA	61.36	58.62	-2.74
CV-Bench 2D	56.59	56.26	-0.33

LLaVA1.5-7B

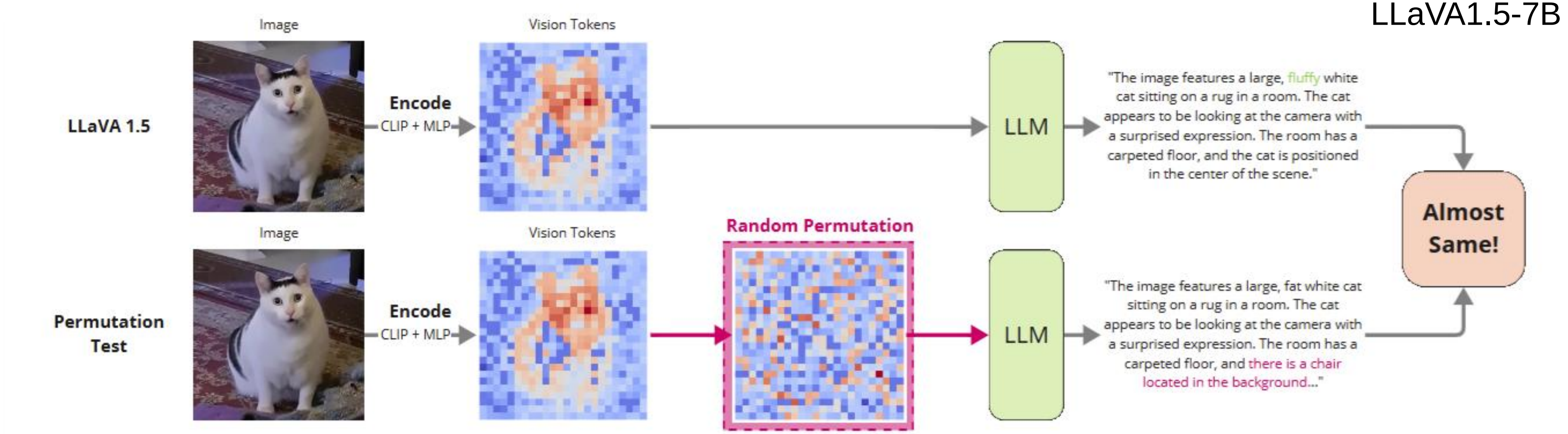
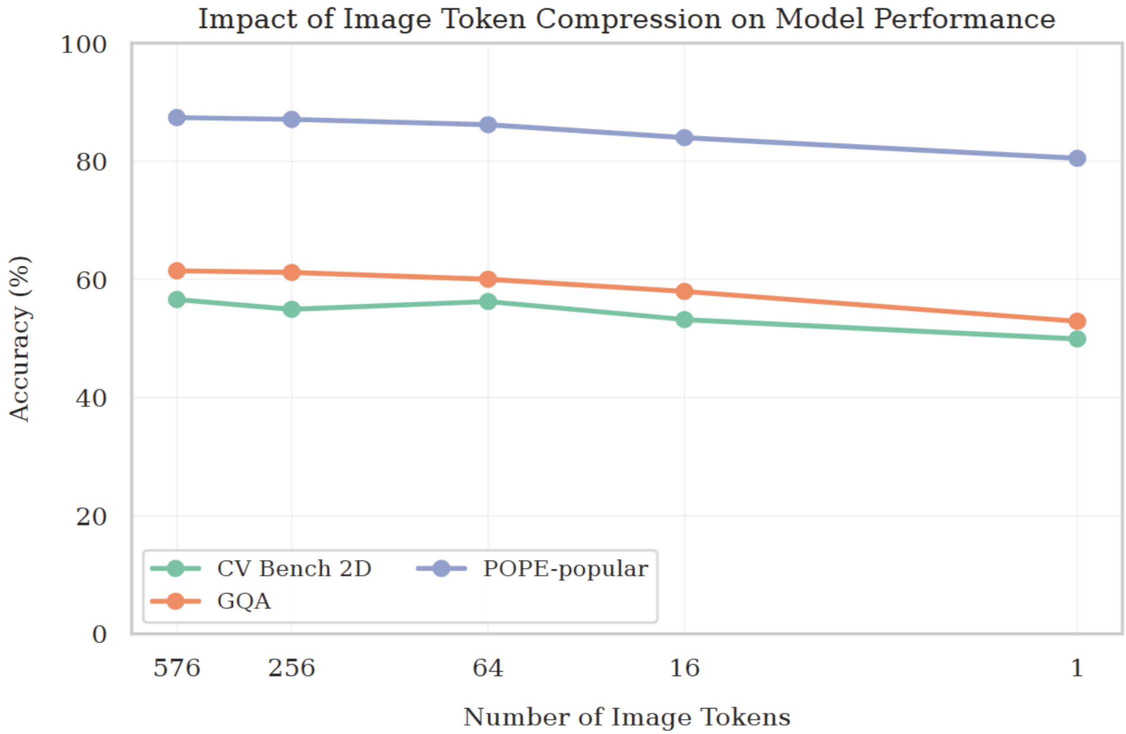


Figure 1. **Permutation Test**: Original (top) vs. randomly permuted vision tokens (bottom). Despite losing spatial ordering, the LLM accurately responds to the prompt "Describe the image," demonstrating strong robustness and a notable "bag-of-tokens" tendency. Token embeddings are visualized using cosine similarity relative to a reference token.

- 结果：打乱视觉token顺序，输出结果几乎不受影响。

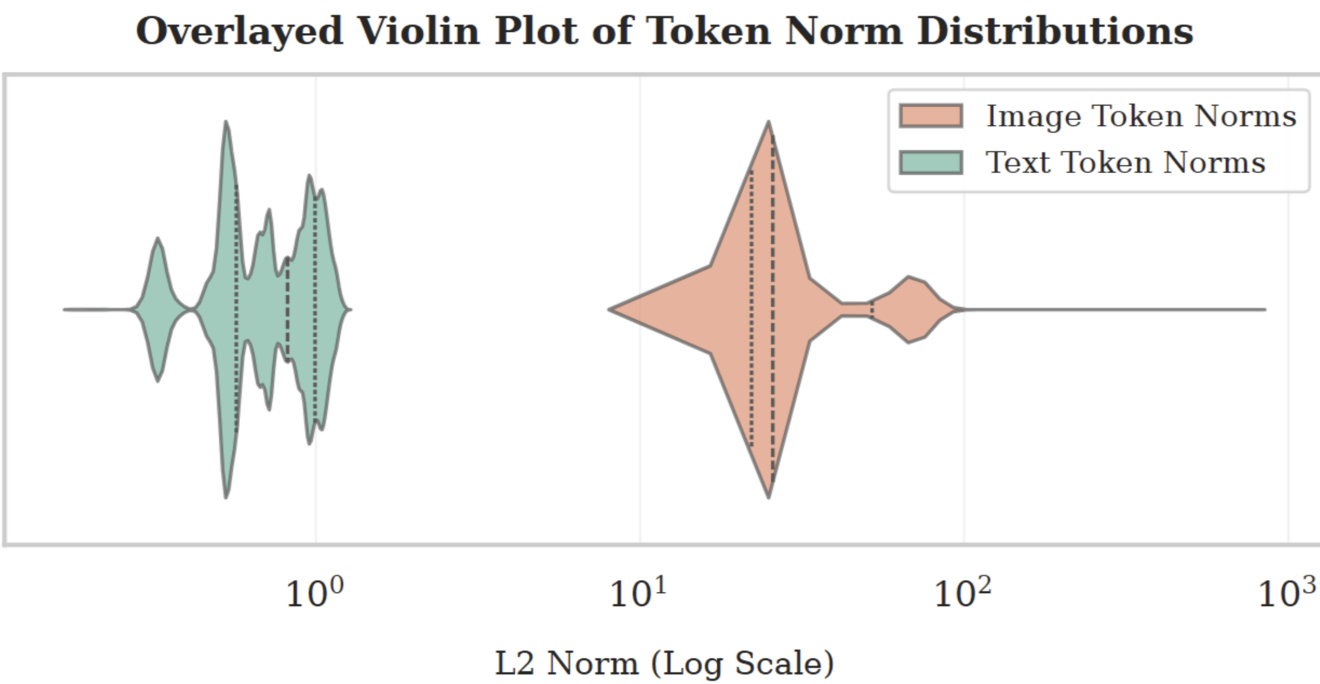
研究背景

- 作者假设：位置编码没有起作用
  - 实验2：空间token压缩测试



研究背景

- 原因分析：
  - 视觉embedding的范数通常比文本embedding的范数大1到3个数量级，这种巨大的范数差异导致位置编码在注意力机制中被掩盖。



研究背景

- 原因分析：
  - 视觉embedding的范数通常比文本embedding的范数大1到3个数量级，这种巨大的范数差异导致位置编码在注意力机制中被掩盖。

$$\mathbf{q}_i' = R(\mathbf{q}_i), \quad \mathbf{k}_i' = R(\mathbf{k}_i). \qquad \|\mathbf{q}_i'\| = \|\mathbf{q}_i\|, \quad \|\mathbf{k}_i'\| = \|\mathbf{k}_i\|. \qquad \mathrm{logit}_{\mathrm{txt,vis}} = \frac{\mathbf{q}_i' \cdot \mathbf{k}_j'}{\sqrt{d}}$$

$$\|\mathbf{q}_{\mathrm{vis}}\| \approx M\|\mathbf{q}_{\mathrm{txt}}\|, \mathbf{k}_{\mathrm{vis}}\| \approx M\|\mathbf{k}_{\mathrm{txt}}\|, \quad M \gg 1.$$

$$\mathrm{logit}_{\mathrm{txt,vis}} \approx \frac{M\|\mathbf{q}_{\mathrm{txt}}'\|\|\mathbf{k}_{\mathrm{txt}}'\|}{\sqrt{d}} \gg \frac{\|\mathbf{q}_{\mathrm{txt}}'\|\|\mathbf{k}_{\mathrm{txt}}'\|}{\sqrt{d}} \approx \mathrm{logit}_{\mathrm{txt,txt}}$$

$$\frac{\partial \alpha_{\mathrm{txt,vis}}}{\partial \phi} = \frac{\partial}{\partial \phi}\left(\frac{\exp(\mathrm{logit}_{\mathrm{txt,vis}})}{\sum_k \exp(\mathrm{logit}_{\mathrm{txt},k})}\right) \qquad \mathrm{logits}_{ij} = \frac{\mathbf{q}_i' \cdot \mathbf{k}_j'}{\sqrt{d}} = \frac{\|\mathbf{q}_i\|\|\mathbf{k}_j\| \cos\phi}{\sqrt{d}}$$

$$\frac{\partial \alpha_{\mathrm{txt,vis}}}{\partial \phi} = \alpha_{\mathrm{txt,vis}}\left(\frac{\partial \mathrm{logit}_{\mathrm{txt,vis}}}{\partial \phi} - \sum_k \alpha_{\mathrm{txt},k} \frac{\partial \mathrm{logit}_{\mathrm{txt},k}}{\partial \phi}\right)$$

提纲

研究方法

核心思想：更多保留空间线索

- 方法上：RMS归一化将视觉embedding的范数调整到与文本嵌入相近的范围
  - 文本嵌入范数的分布（均值约为0.83，最大值约为1.22）
- 架构上：利用视觉编码器中间层特征，保留更多局部信息输入模型

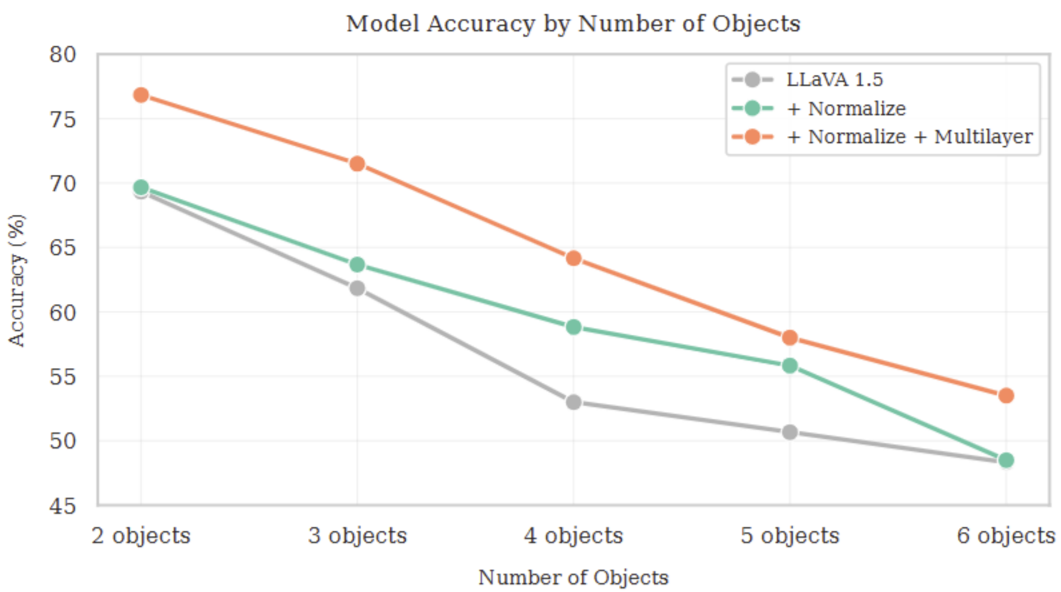
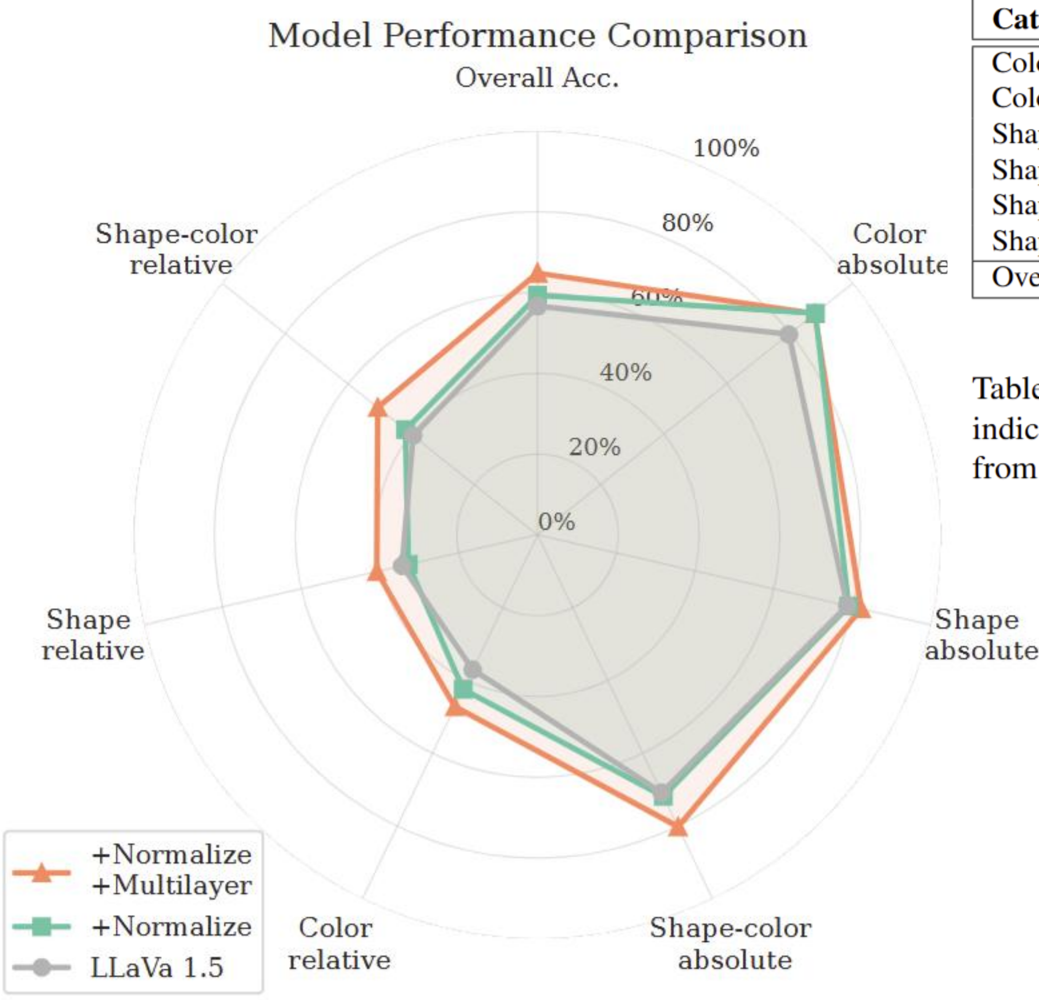


Figure 5. Accuracy comparison across varying numbers of objects. Our interpretability-informed adjustments yield consistent improvements, especially as spatial complexity increases.

提纲

- 作者介绍
- 研究背景
- 研究方法
- 实验效果
-  总结&思考

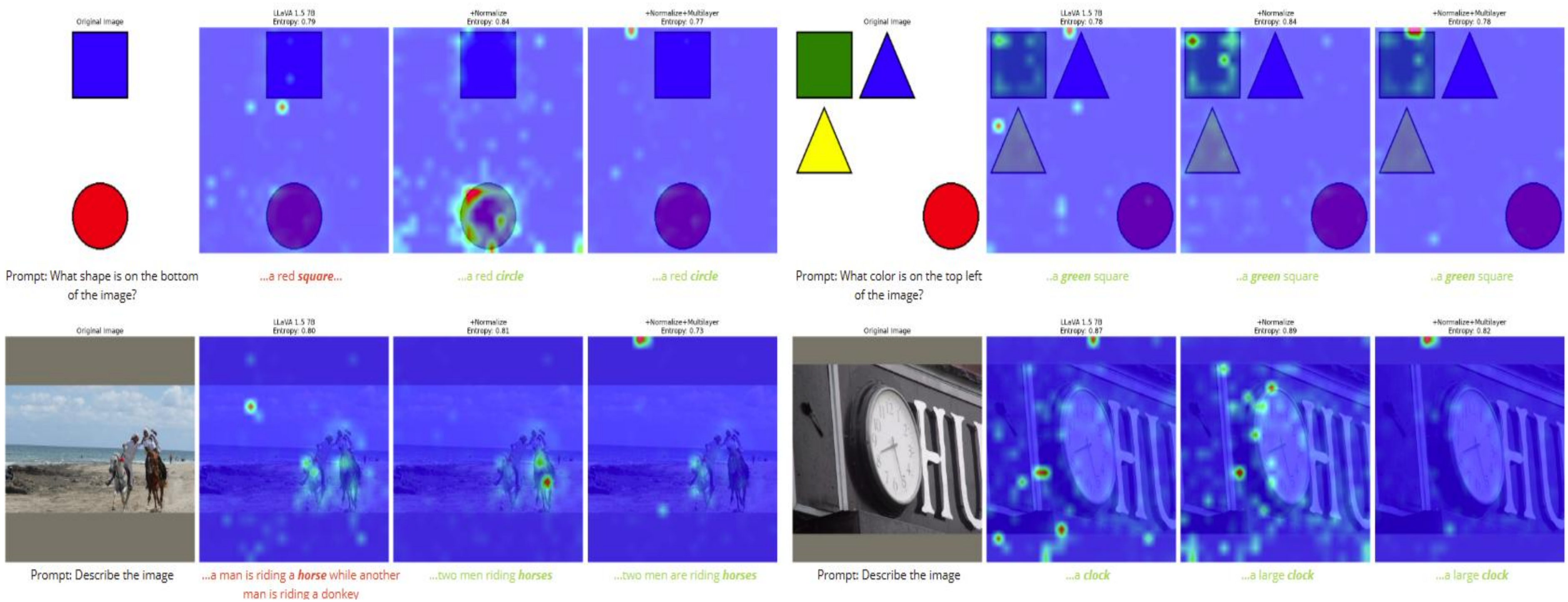
实验效果



Category	LLaVA 1.5	+ Normalize	+ Normalize + Multilayer
Color_abs. ↑	79.60	**88.00** (+8.40)	87.80 (+8.20)
Color_rel. ↑	37.00	42.40 (+5.40)	**47.20** (+10.20)
Shape_abs. ↑	78.60	79.00 (+0.40)	**82.20** (+3.60)
Shape_rel. ↑	34.40	32.80 (-1.60)	**40.80** (+6.40)
Shape_color_abs. ↑	70.80	71.80 (+1.00)	**80.20** (+9.40)
Shape_color_rel. ↑	39.40	41.80 (+2.40)	**50.60** (+11.20)
Overall Acc. ↑	56.63	59.30 (+2.67)	**64.80** (+8.17)

Table 2. Spatial reasoning accuracy (%) across 2DS categories. ↑ indicates higher is better. Values in parentheses show difference from LLaVA 1.5 baseline.

实验效果



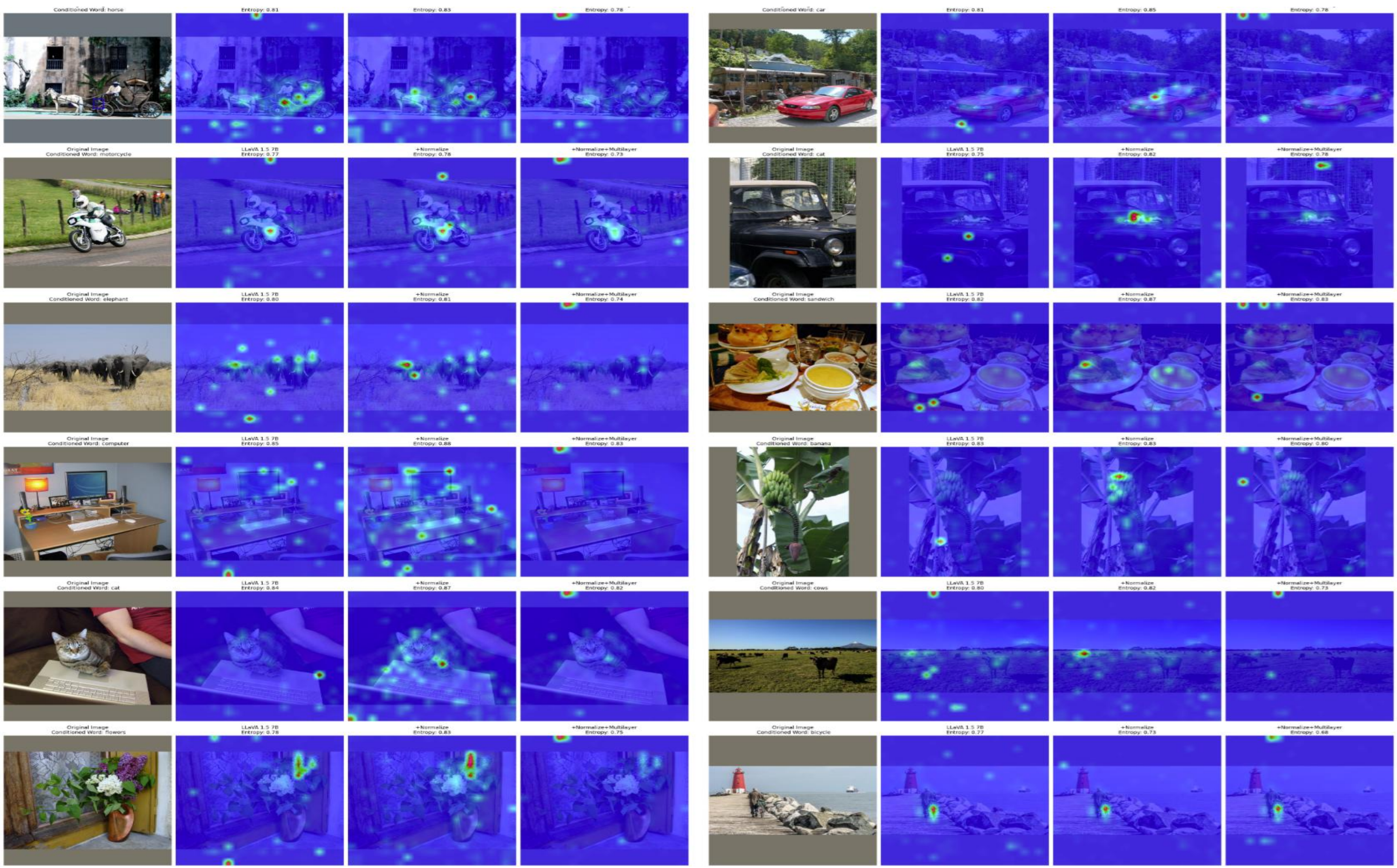
Original image
LLaVA 1.5 7B
Entropy: 0.79
+Normalize
Entropy: 0.84
+Normalize+Multilayer
Entropy: 0.77
Original image
LLaVA 1.5 7B
Entropy: 0.78
+Normalize
Entropy: 0.84
+Normalize+Multilayer
Entropy: 0.78
Prompt: What shape is on the bottom of the image?
...a red square...
...a red circle
...a red circle
Prompt: What color is on the top left of the image?
..a green square
..a green square
..a green square
Original image
LLaVA 1.5 7B
Entropy: 0.80
+Normalize
Entropy: 0.81
+Normalize+Multilayer
Entropy: 0.73
Original image
LLaVA 1.5 7B
Entropy: 0.87
+Normalize
Entropy: 0.89
+Normalize+Multilayer
Entropy: 0.82
Prompt: Describe the image
...a man is riding a horse while another man is riding a donkey
...two men riding horses
...two men are riding horses
Prompt: Describe the image
...a clock
...a large clock
...a large clock

实验效果

Category	**LLaVA 1.5**	**+ Normalize**	**+ Normalize + Multilayer**
VQAv2 ↑	78.20	78.76 (+0.56)	**79.17** (+0.97)
POPE ↑	87.30	87.30 (0.00)	**87.70** (+0.40)
GQA ↑	61.46	62.04 (+0.58)	**62.52** (+1.05)
CV-Bench2D ↑	56.59	**59.91** (+3.32)	58.69 (+2.10)

Table 3. Performance (accuracy %) on standard vision-language benchmarks. ↑ indicates higher is better. Values in parentheses show difference from LLaVA 1.5 baseline.

实验效果



提纲

作者介绍

研究背景

研究方法

实验效果

总结&思考

总结

- 假设-验证 低资源需求改进方法
- 有效性验证不充足
  - 架构只有LLaVA
  - Benchmark有限

Thank you！