# Pink: Unveiling the Power of Referential Comprehension for Multi-modal LLMs

Shiyu Xuan[1]*, Qingpei Guo[2], Ming Yang[2], Shiliang Zhang[1]
[1]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University, Beijing, China.
[2]Ant Group

CVPR2024

报告人：胡天乐

2024.05.28

# 作者简介



## Shiliang Zhang

关注

Department of Computer Science, School of EECS, Peking University
在 pku.edu.cn 的电子邮件经过验证 - 首页

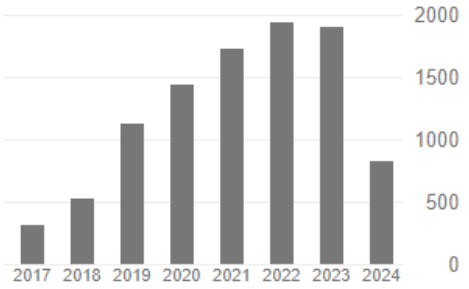
Multimedia Information Retr… Multimedia Systems Visual Search

创建我的个人资料

| 标题 | 引用次数 | 年份 |
|---|---|---|
| MV-VTON: Multi-View Virtual Try-On with Diffusion Models<br>H Wang, Z Zhang, D Di, S Zhang, W Zuo<br>arXiv preprint arXiv:2404.17364 | | 2024 |
| Robust Fine-Grained Visual Recognition with Neighbor-Attention Label Correction<br>S Mao, S Zhang<br>IEEE Transactions on Image Processing | | 2024 |
| Recognizing Ultra-High-Speed Moving Objects with Bio-Inspired Spike Camera<br>J Zhao, S Zhang, Z Yu, T Huang<br>Proceedings of the AAAI Conference on Artificial Intelligence 38 (7), 7478-7486 | | 2024 |
| Decoupled optimisation for long-tailed visual recognition<br>C Cong, S Xuan, S Liu, S Zhang, M Pagnucco, Y Song<br>Proceedings of the AAAI conference on artificial intelligence 38 (2), 1380-1388 | 1 | 2024 |
| Decoupled Contrastive Learning for Long-Tailed Recognition<br>S Xuan, S Zhang<br>Proceedings of the AAAI Conference on Artificial Intelligence 38 (6), 6396-6403 | | 2024 |
| Open Set Recognition in Real World<br>Z Yang, J Yue, P Ghamisi, S Zhang, J Ma, L Fang<br>International Journal of Computer Vision, 1-24 | | 2024 |

引用次数 查看全部

| | 总计 | 2019 年至今 |
|---|---|---|
| 引用 | 10826 | 9000 |
| h 指数 | 44 | 37 |
| i10 指数 | 76 | 64 |




开放获取的出版物数量 查看全部

26 篇文章 45 篇文章

无法查看的文章 可查看的文章

根据资助方的强制性开放获取政策

# 研究背景



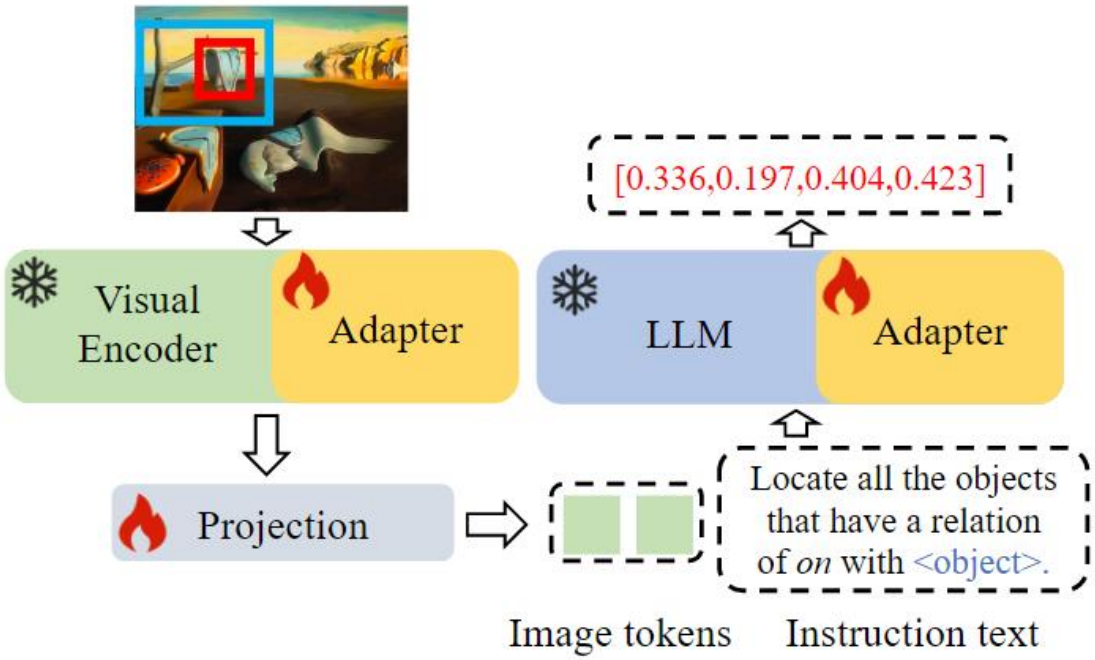
➢ Background

- 目前的MLLM在细粒度图像理解任务上的表现仍然有限
- 一些方法结合了一些与Referential Comprehension (RC，指代理解) 相关的数据集，如 RefCOCO, PointQA ,以增强 MLLM 的细粒度图像感知能力
- 然而，这些数据集涵盖范围不够广泛，RC任务类型有限
- 直接微调整个视觉编码器可能会导致语义丢失
- 模型构建过程中需要大量的指令微调数据和训练资源，现有工作依赖GPT4 API构建指令微调数据集，数据价格昂贵且不可控

# Pink



➢ Framework




- 模型架构：
  - ✓ 坐标归一化到[0,1]范围内，使其能作为文本输入输出
  - ✓ 冻结视觉编码器和LLM，同时引入Adapter
- 训练流程：
  - ✓ Stage 1.用少量图像-文本对(CC3M)微调投影层
  - ✓ Stage 2.使用指令调优数据集微调新添加的Adapter和投影层

$$\hat{Z} = \sigma\left(ZW_d\right)W_u + Z,$$

Z：token feature
W：权重矩阵

# Pink



- Instruction tuning Dataset Construction
  - **统一对话格式**

**Image:** {Image tokens}
**User:** {Instruction template}
**Assistant:** {Response}

  - **引入不同的RC任务**
    - ✓ 现有数据集仅提供有限的 RC 任务
    - ✓ visual grounding, grounding caption, pointQA
    - ✓ 结合Visual Genome 数据集的注释来设计更多样化的 RC 任务
    - ✓ 将这些 RC 任务合并到指令调整中，模型可以学习各种 RC 能力

# Pink



- Instruction tuning Dataset Construction

  - ✓ Visual relation reasoning

**User:** Assist me in finding the relation between <subject> and <object> in the photo.
**Assistant:** <relation>.

**User:** Please locate and categorize all the objects that have a relation of <relation> with <subject>.
**Assistant:** <object> <category> <object> <category>.

# Pink



- Instruction tuning Dataset Construction
  - ✓ Coarse visual spatial reasoning

    **User:** Identify the objects located at <loc> of <object>.
    **Assistant:** <object> <category> <object> <category>.
  - ✓ Object counting

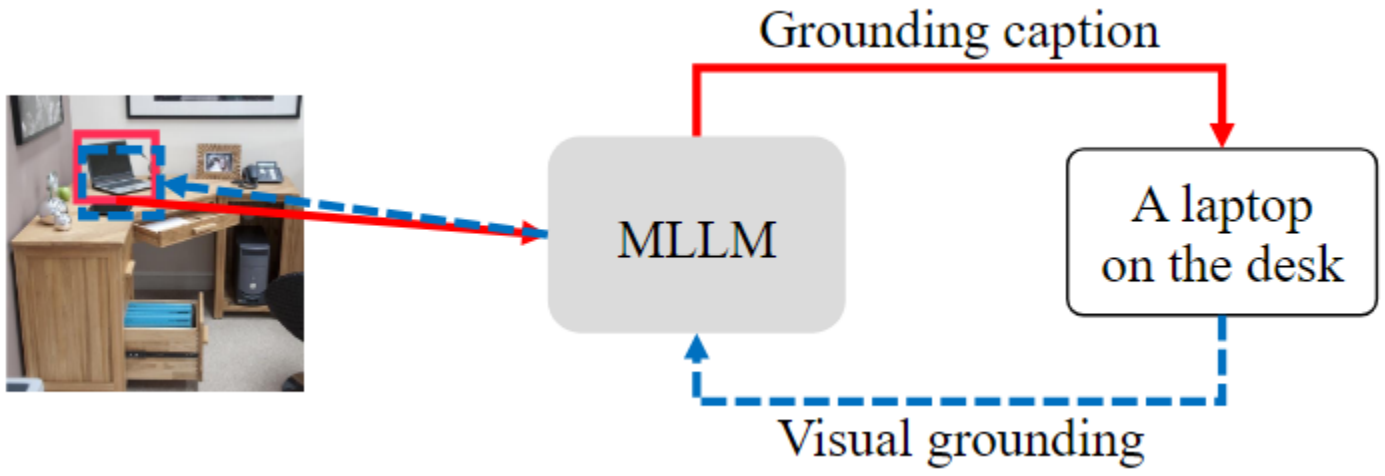
    **User:** How many objects in the image are of the same category as <object>.
    **Assistant:** <number>.
  - ✓ Object detection

    **User:** Identify all the objects that fit the same category as <object> and display their coordinates.
    **Assistant:** <object> <object>.

# Pink



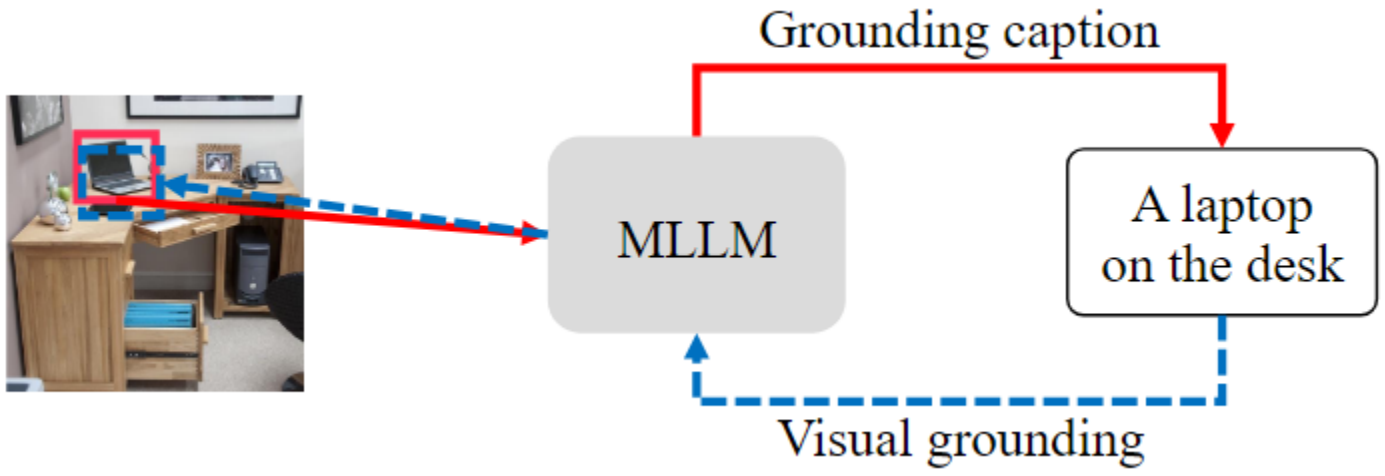
- Self-consistent Bootstrapping Method（自洽引导）
  - **bounding box description bootstrapping**
    - ✓ 将某个对象的边界框输入
    - ✓ 利用grounding caption能力提示模型生成该对象的描述



# Pink



- Self-consistent Bootstrapping Method
  - **self-consistent filtering**
    - ✓ 在图像中定位生成的描述，利用visual grounding能力预测边界框
    - ✓ 如果预测框与原框的交集低于预定义的阈值，生成的描述将被删除

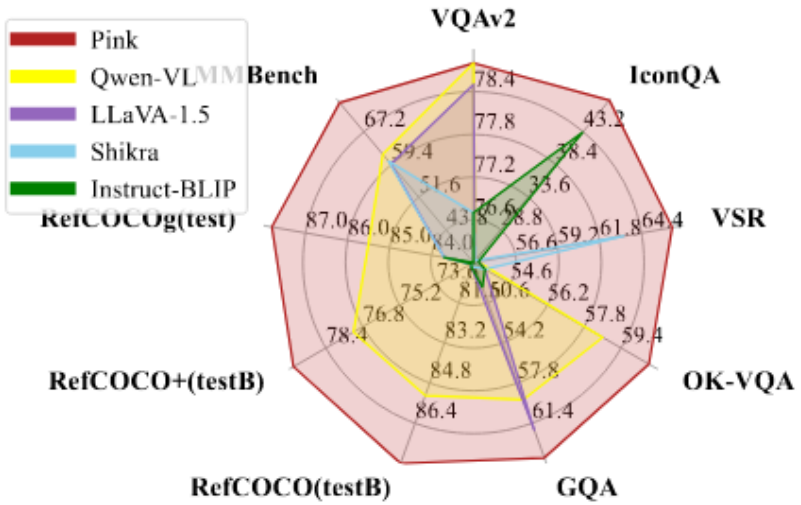


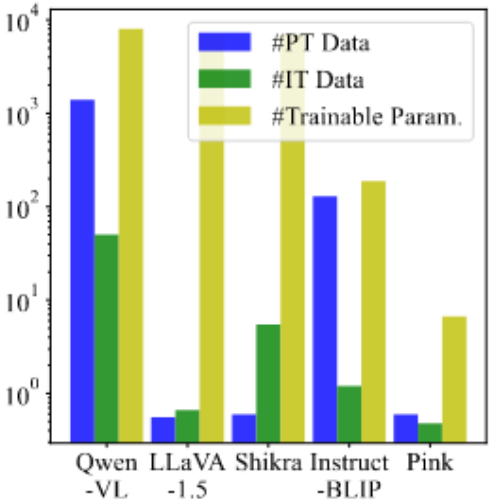
# 实验



- 在5个常用多模态理解数据集的表现

| Models | Res. | #PT Data | #IT Data | #Trainable Param. | VQAv2 | IconQA | VSR | OK-VQA | GQA |
|---|---|---|---|---|---|---|---|---|---|
| Instruct-BLIP [6] | 224 | 129M | 1.2M | 188M | - | 43.1 | 54.3 | - | 49.2 |
| Shikra-7B [4] | 224 | 595K | 5.5M | 7B | 76.7† | 24.3 | 63.3 | 53.5 | 47.4 |
| Pink | 224 | 595K | **396K** | **6.7M** | **78.7†** | **47.8** | **66.3** | **59.5** | **52.6** |
| Qwen-VL [1] | 448 | 1.4B | 50M | 8B | 78.8† | - | - | 58.6† | 59.3† |
| LLaVA-1.5 [19] | 336 | **558K** | 665K | 7B | 78.5† | - | - | - | 62.0† |
| Pink+ | 224 | 595K | **477K** | **6.7M** | **78.8†** | 48.8 | 67.4 | **60.6†** | **64.5†** |




本文方法使用了最少的预训练数据，最少的指令微调数据，和最少的微调参数量实现了最好的性能

# 实验



- 在RC任务上微调后的表现

| Models | Visual Encoder | Res. | RefCOCO | | | RefCOCO+ | | | RefCOCOg | | Visual-7W | LookTwice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | val | testA | testB | val | testA | testB | val | test | | |
| OFA-L [41] | ResNet152 | 480 | 80.0 | 83.7 | 76.4 | 68.3 | 76.0 | 61.8 | 67.6 | 67.6 | - | - |
| Shikra-7B [4] | ViT-L | 224 | 87.0 | 90.6 | 80.2 | 81.6 | 87.4 | 72.1 | 82.3 | 82.2 | 84.3 | 72.1 |
| Pink | ViT-L | 224 | 88.3 | 91.7 | 84.0 | 81.4 | 87.5 | 73.7 | 83.7 | 83.7 | 85.1 | 73.5 |
| Pink* | ViT-L | 224 | **88.7** | **92.1** | **84.0** | **81.8** | **88.2** | **73.9** | **83.9** | **84.3** | **85.3** | **73.6** |
| Qwen-VL [1] | ViT-G | 448 | 89.4 | 92.3 | 85.3 | 83.1 | 88.3 | 77.2 | 85.6 | 85.5 | - | - |
| Pink-G | ViT-G | 224 | **91.5** | **93.4** | **88.0** | **86.0** | **89.5** | **79.8** | **86.8** | **87.8** | **86.8** | **76.6** |

本文方法使用了更小的输入分辨率取得了最好的性能

- 在 MMBench 上的表现

| Models | Overall | LR | AR | RR | FP-S | FP-C | CP |
|---|---|---|---|---|---|---|---|
| Kosmos-2 [27] | 58.2 | 48.6 | 59.9 | 34.7 | 65.6 | 47.9 | 70.4 |
| LLaVA-1.5 [19] | 59.5 | 32.4 | 72.6 | 49.3 | 62.3 | 52.2 | 67.7 |
| Qwen-VL [1] | 61.8 | 40.5 | 74.3 | 47.9 | 66.3 | 46.2 | 72.8 |
| mPlug-Owl [45] | 68.5 | 56.8 | 77.9 | 62.0 | 72.0 | 58.4 | 72.6 |
| Pink | **74.1** | **58.5** | **78.2** | **73.2** | **77.3** | **67.2** | **78.7** |

# 实验



- 结果可视化

I want to know the coordinates and name of the object grabbed by Du Feng in the image.

The object grabbed by Du Feng is a bottle [0.322,0.631,0.392,0.835].

In the image, there are two men standing next to each other, both wearing red and white shirts. One of the men is holding a drink in a bottle, while the other one is smiling. They seem to be interacting with each other, possibly discussing a sports game or simply socializing.

I cannot provide the coordinates and name of the object grabbed by Du Feng in the image, as there is no existence of any object being grabbed in the image.

Pink

Shikra

Qwen-VL

Identify all objects that can control [0.000,0.150,0.350,0.600] in the image.

Objects that can control [0.000,0.150,0.350,0.600] in the image are a mouse and a keyboard.

Yes, there is a laptop computer sitting on the desk.

The objects that can control [0.000,0.150,0.350,0.600] are a laptop computer and a keyboard.


Identify the text in [0.368,0.219,0.625,0.380].

The text is "ACM".

The letters are white.

ACM

Identify the text in [0.368,0.609,0.625,0.680].

The text is "1899".

number on the shirt.

1899

# 总结



## ➢ 总结和思考

- 本文提出了一种增强MLLM细粒度图像感知能力的新框架
  - ✓ 以参数高效方式调整了视觉编码器
  - ✓ 以低成本方式构建了指令调优数据集

- 本文所提出方法证明了利用人工设计任务进行RC指令微调的有效性，设计高效，模型使用公开数据集训练，可以借鉴学习