



澳門大學  
UNIVERSIDADE DE MACAU  
UNIVERSITY OF MACAU

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

XuTaosi  
taosixu@umac.mo  
2024.8.21

[Submitted on 1 Oct 2023 (v1), last revised 13 Mar 2024 (this version, v3)]

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

Shiyu Xuan, Qingpei Guo, Ming Yang, Shiliang Zhang

Multi-modal Large Language Models (MLLMs) have shown remarkable capabilities in various multi-modal tasks. Nevertheless, their performance in fine-grained image perception is still limited. We propose a new method for constructing the instruction tuning dataset at a low cost by leveraging annotations in existing datasets. A self-consistent bootstrapping method includes a wide range of fundamental abilities essential for fine-grained image perception. Moreover, we argue that the visual encoder should be tuned to better understand the visual information. Our model exhibits a 5.2% accuracy improvement over Qwen-VL on GQA and surpasses the accuracy of Kosmos-2 by 24.7% on RefCOCO\_val. We have all code and data publicly available at [this https URL](https://github.com/SHYuan/Pink).

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Artificial Intelligence (cs.AI)

Cite as: [arXiv:2310.00582](https://arxiv.org/abs/2310.00582) [cs.CV]

(or [arXiv:2310.00582v3](https://arxiv.org/abs/2310.00582v3) [cs.CV] for this version)

<https://doi.org/10.48550/arXiv.2310.00582> 

### Submission history

From: Shiyu Xuan [[view email](#)]

[v1] Sun, 1 Oct 2023 05:53:15 UTC (4,302 KB)

[v2] Tue, 21 Nov 2023 10:32:13 UTC (6,205 KB)

[v3] Wed, 13 Mar 2024 03:42:31 UTC (5,974 KB)

# Introduction

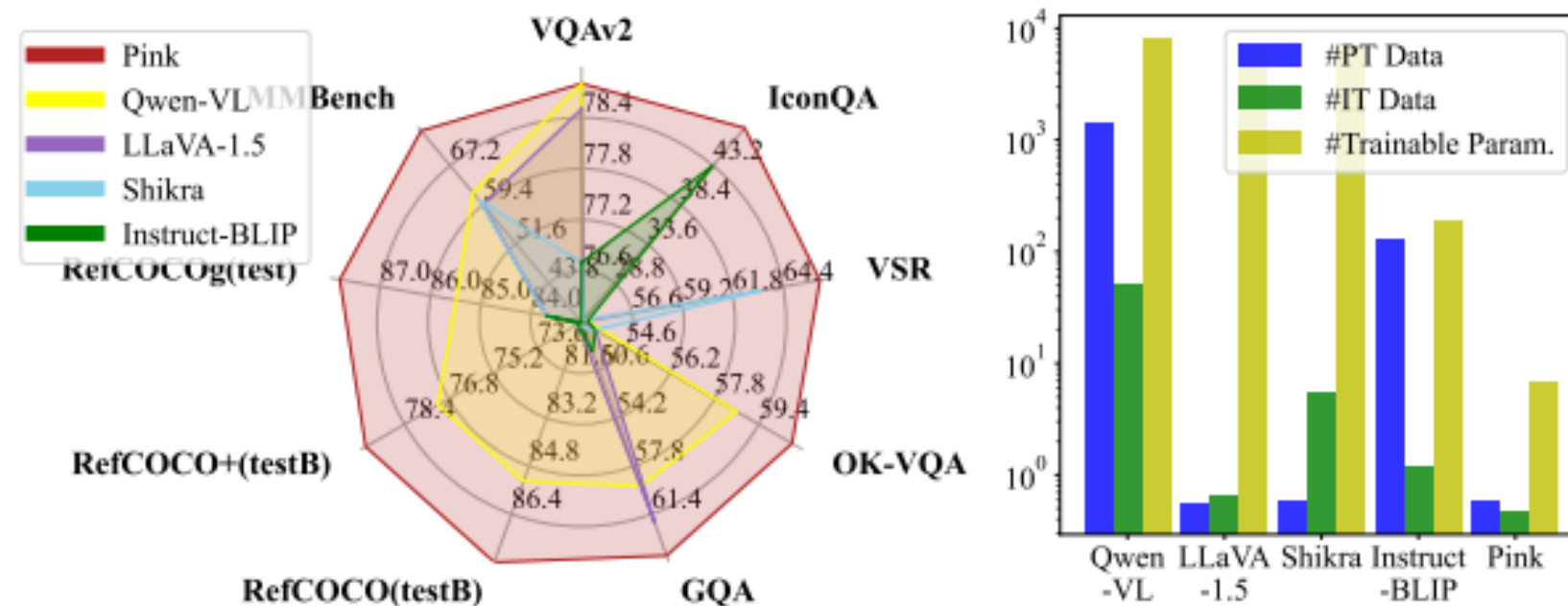


Figure 1. With fewer trainable parameters and less training data, Pink achieves the best performance on both conventional multi-modal tasks and RC tasks. “#Trainable Param.”, “#PT Data”, and “#IT Data” indicate the number of trainable parameters, the number of samples in pre-training and instruction tuning stage, respectively.

**VQAv2 (Visual Question Answering v2):** 该任务要求模型根据给定的图像回答自然语言问题，测试模型的视觉理解和语言处理能力。

**IconQA:** 该任务涉及对图标或符号的理解，模型需要识别图标并回答相关问题，评估其对视觉符号的识别能力。

**VSR (Visual Spatial Reasoning):** 该任务需要模型理解图像中物体之间的空间关系，通常涉及对位置、方向和距离的推理。

**OK-VQA (Ok Visual Question Answering):** 这个任务类似于VQA，但问题设计更侧重于常识推理，要求模型不仅依赖于图像内容，还要结合外部知识进行回答。

**GQA (Generalized Question Answering):** 该任务是一个更广泛的视觉问答任务，涵盖多种问题类型，旨在评估模型对图像和文本的综合理解能力。

**RefCOCO (Referential COmprehension):** 该任务要求模型根据给定的描述在图像中定位特定对象，测试模型的指称理解能力。

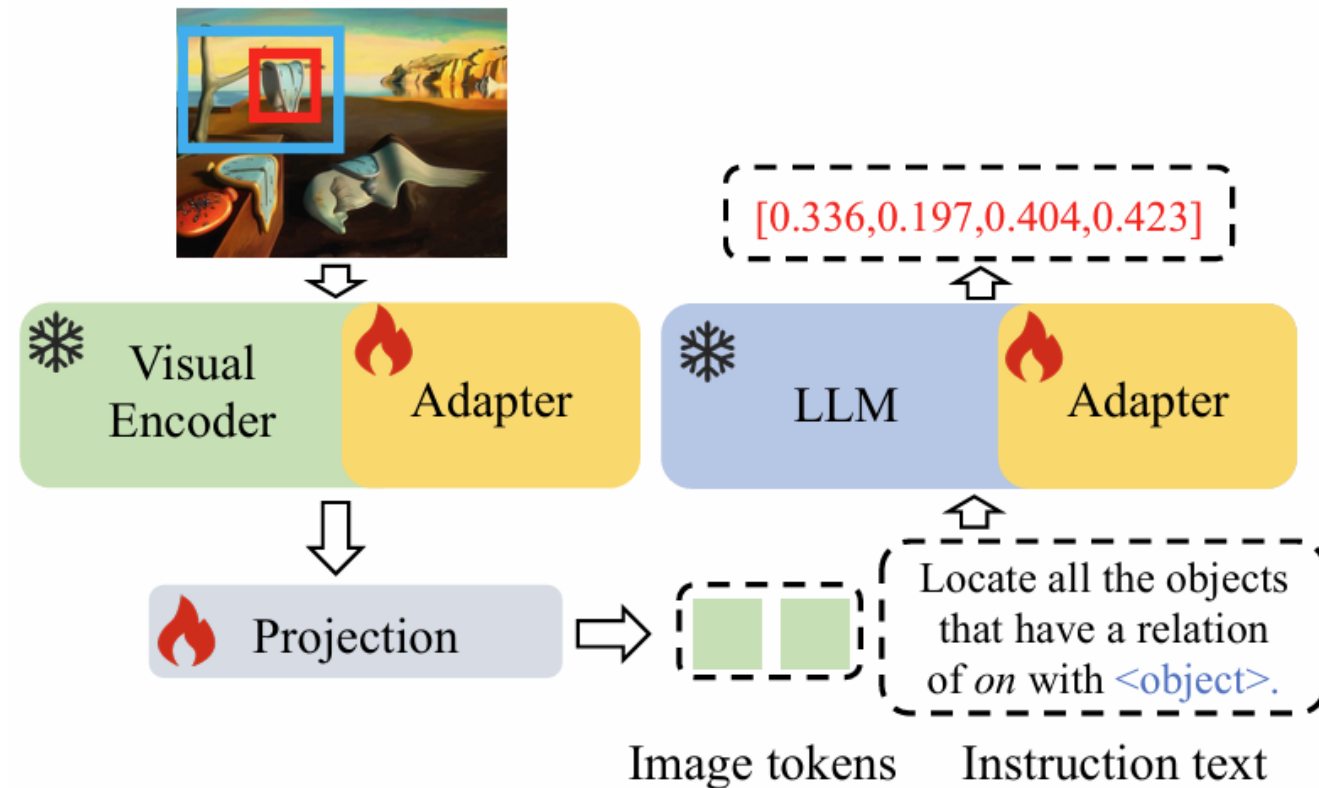
**RefCOCO+:** 与RefCOCO类似，但包含更复杂的描述和更多的场景，旨在提高模型的细粒度理解和定位能力。

As shown in Fig. 1, the high-quality instruction tuning data generated by thier method enables the model to achieve promising performance with a reduced number of training samples.



# Methodology

- Model architecture.



视觉编码器提取图像特征，通过适配器转换为兼容格式，然后通过投影层生成图像标记和文本格式的边界框坐标。接着，解码器（语言模型）根据这些输入生成与指令相关的输出。在训练过程中，视觉编码器和语言模型保持冻结状态，仅更新适配器和投影层，以提高细粒度图像理解能力。

Figure 2. The illustration of our Pink model. Pink follows the architecture of LLaVA [20], which consists of three main components: a visual encoder, a projection layer, and a decoder-only LLM. The coordinates of a bounding box are converted into texts in a specific format. During instruction tuning, we freeze the visual encoder and LLM and only update the Adapters and the projection layer.

# Methodology

- **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>.

**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>.

# Methodology

- **Self-consistent Bootstrapping Method.**

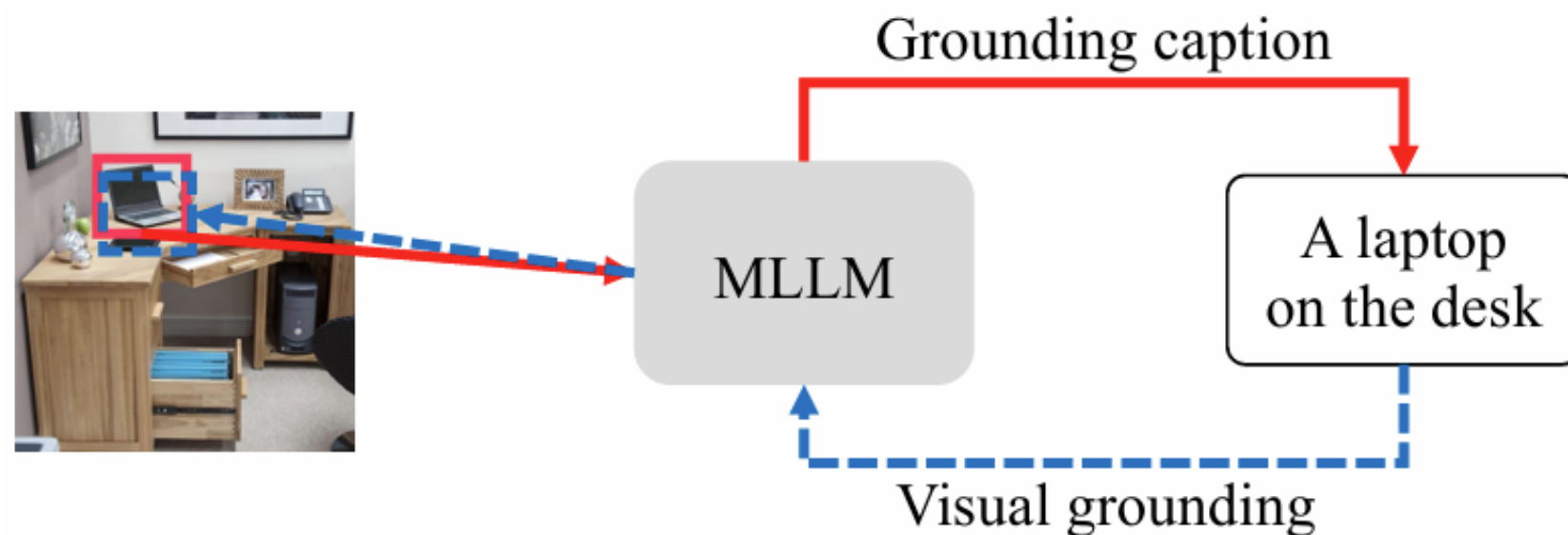



Figure 3. The illustration of self-consistent bootstrapping method. Given a bounding box, our method first generates its description by asking MLLM to perform grounding caption, then leverages the visual grounding to locate the generated description. The low-quality description will be filtered if the IOU between the predicted and ground-truth bounding box is below a threshold.

给定一个边界框，模型首先通过多模态语言模型（MLLM）生成描述（“grounding caption”）。接着，利用视觉定位（visual grounding）来确认生成的描述是否准确。如果预测的描述与真实边界框之间的重叠度（IOU）低于某个阈值，则会过滤掉低质量的描述，从而确保生成的描述更加准确和一致。整体流程旨在提高模型对图像内容的理解和描述质量。

# Methodology

- Qualitative Results on RC Reasoning.



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.

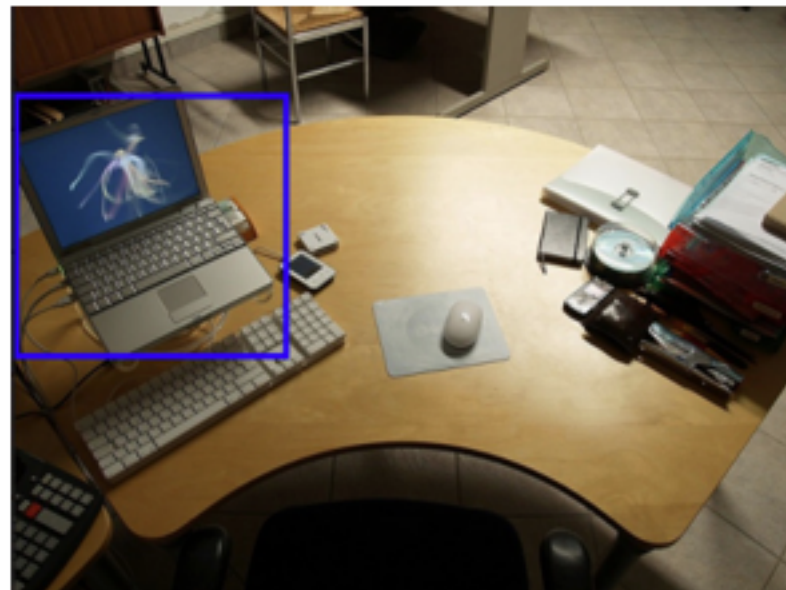
(a)

A comparison of Pink against the recent MLLMs Shikra and Qwen-VL on RC reasoning.



# Methodology

- Qualitative Results on RC Reasoning.



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.




(b)

A comparison of Pink against the recent MLLMs Shikra and Qwen-VL on RC reasoning.



# Methodology

- Qualitative Results on RC Reasoning.



The image shows a football jersey with vertical red and black stripes. In the center is a circular crest with a red cross on a white background. The letters 'ACM' are written in white at the top of the crest, and the year '1899' is written in white at the bottom. A blue bounding box is drawn around the 'ACM' text, and a yellow bounding box is drawn around the '1899' text.

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

(c)

A comparison of Pink against the recent MLLMs Shikra and Qwen-VL on RC reasoning.

# Experiments

Settings	IconQA	VSR	OK-VQA	RefCOCO_val	Local	LookTwice
Baseline	44.6	65.6	58.5	55.0	0.0	0.2
w/o VG	43.1	62.8	58.3	-	-	-
+ R	44.4	65.7	58.5	52.1	17.1	12.8
+ R,S	46.2	65.8	58.5	52.7	50.9	60.0
+ R,S,C	47.4	65.7	58.9	53.1	53.4	60.7
+ R,S,C,D	<b>47.8</b>	66.3	59.5	54.1	54.6	63.1
+ R,S,C,D + Object365 <sup>†</sup>	44.6	65.9	58.7	73.8	52.1	69.2
+ R,S,C,D + Object365	47.7	<b>67.1</b>	<b>59.5</b>	<b>77.0</b>	<b>57.2</b>	<b>70.3</b>
Freezing	42.9	61.5	58.3	37.2	44.9	57.5
Full-tuning	36.9	48.6	33.1	0.05	26.1	54.1
LoRA	44.3	65.4	58.9	<b>54.7</b>	<b>56.7</b>	62.2
Our	<b>47.8</b>	<b>66.3</b>	<b>59.5</b>	54.1	54.6	<b>63.1</b>

Table 1. Ablation study on instruction tuning dataset construction and training settings of visual encoder under a zero-shot setting. “Baseline” denotes leveraging Visual Genome by only performing visual grounding and grounding caption tasks. “VG” denotes Visual Genome. “R”, “S”, “C”, and “D” denote the visual relation reasoning, coarse visual spatial reasoning, object counting and object detection tasks, respectively. <sup>†</sup> denotes generated referring-expression-bounding-box pairs in Object365 are not filtered with the self-consistent method. “Freezing” and “Full-tuning” denotes freezing the visual encoder and training the entire visual encoder, respectively. “LoRA” denotes using LoRA instead of the Adapter to perform parameter-efficient tuning.

# Experiments

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	<b>396K</b>	<b>6.7M</b>	<b>78.7†</b>	<b>47.8</b>	<b>66.3</b>	<b>59.5</b>	<b>52.6</b>
Qwen-VL [1]	448	1.4B	50M	8B	78.8†	-	-	58.6†	59.3†
LLaVA-1.5 [19]	336	<b>558K</b>	665K	7B	78.5†	-	-	-	62.0†
Pink+	224	595K	<b>477K</b>	<b>6.7M</b>	<b>78.8†</b>	48.8	67.4	<b>60.6†</b>	<b>64.5†</b>

(a) Results on the conventional multi-modal reasoning tasks. † denotes the training set of corresponding dataset is included. + denotes adding the training set of OK-VQA and GQA during instruction tuning. “Res.”, “#Trainable Param.”, “#PT Data”, and “#IT Data” indicate input image resolution, the number of trainable parameters, the number of samples in pre-training and instruction tuning stage, respectively.



# Experiments

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	<b>88.7</b>	<b>92.1</b>	<b>84.0</b>	<b>81.8</b>	<b>88.2</b>	<b>73.9</b>	<b>83.9</b>	<b>84.3</b>	<b>85.3</b>	<b>73.6</b>
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	<b>91.5</b>	<b>93.4</b>	<b>88.0</b>	<b>86.0</b>	<b>89.5</b>	<b>79.8</b>	<b>86.8</b>	<b>87.8</b>	<b>86.8</b>	<b>76.6</b>

(c) Fine-tuning results on the RC tasks. Pink-G indicates the ViT-G is used as the visual encoder for a fair comparison.

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	<b>74.1</b>	<b>58.5</b>	<b>78.2</b>	<b>73.2</b>	<b>77.3</b>	<b>67.2</b>	<b>78.7</b>

(d) CircularEval results on MMBench test set [21].

Table 2. Comparison with other methods. \* denotes Object365 with generated referring-expression-bounding-box pairs is used.



# Thank You!

Avenida da Universidade, Taipa, Macau, China

Tel : (853) 8822 8833      Fax : (853) 8822 8822

Email : [info@um.edu.mo](mailto:info@um.edu.mo)      Website : [www.um.edu.mo](http://www.um.edu.mo)