

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

## Supplementary Material

Stage	Dataset	Data Number
PT	LLaVA-CC3M-Pretrain-595K [22]	595K
IT	LLaVA-158K [22]	158K
IT	VQAv2 [8]	83K
IT	A-OKVQA [34]	17K
IT	Visual Genome [15]	108K
IT	Flickr30K [30]	30K
IT	Object365* [35]	1M

Table 3. Statistics of training datasets. PT and IT denote the pre-training and instruction tuning stage, respectively.

### A.1. Discussion of Limitations

Our method relies on the LLM. Therefore, it has some shortcomings from the LLM, such as bias or unfair response, and hallucination. Like most MLLMs with RC ability, we also find that our model is not good at object detection in complex scenarios, *e.g.*, identifying multiple tiny objects in the image. It may be because the input resolution of the image is low. However, increasing the input resolution of the image is not a trivial task. Many efforts are still needed to deal with such tasks.

### A.2. Statistics of Training and Evaluation Datasets

The statistics for the training datasets and evaluation datasets are summarized in Table 3 and Table 4, respectively. We also provide an overview of the generated referring-expression-bounding-box pairs in Object365 in Table 5. To ensure the quality of data, we apply certain filters during preprocessing. Firstly, we exclude images containing more than 15 objects. Moreover, for the purpose of bounding box description bootstrapping, we only consider objects that cover an area of more than 2,000 pixels. As a result, our dataset comprises 1,063,034 images, with a total of 4,961,822 generated referring-expression-bounding-box pairs. To further enhance the reliability of our dataset, we perform a self-consistent method that filters out 2,528,619 low-quality referring-expression-bounding-box pairs.

### A.3. More experimental results

The results that leverages a more advanced LLM LLaMA-2 [38] are shown in Table 6 and Table 7. We can observe an improvement in performance on most datasets.

Similar to MMBench, SEED-Bench [16] is a recently proposed benchmark to evaluate the abilities of MLLMs on various evaluation dimensions including Scene Understanding (SU), Instance Identity (IId), Instance Attribute (IA), Instance Location (IL), Instance Counting (IC), Spatial

Dataset	Split	Metric
VQAv2 [8]	test-dev	VQA Score
IconQA [25]	multi-text-choice test	Accuracy
VSR [19]	zero-shot test	VQA Score
OK-VQA [28]	val	VQA Score
GQA [12]	test-dev	VQA Score
RefCOCO [14]	val & testA & testB	Accuracy
RefCOCO+ [14]	val & testA & testB	Accuracy
RefCOCOG [27]	val & test	Accuracy
PointQA-Local [26]	test-dev	VQA Score
PointQA-LookTwice [26]	test	VQA Score
Visual-7W [51]	which box test	Accuracy
MMBench [23]	test	Accuracy

Table 4. Summary of evaluation datasets.

Images	Referring-expressions	Avg Expression Length
1,063,034	2,433,203	3.6

Table 5. Statistics of generated referring-expression-bounding-box pairs in Object365.

Models	VQAv2	IconQA	VSR	OK-VQA	GQA
Pink	78.7†	47.8	66.3	59.5	52.6
Pink-LLaMA-2	78.8†	49.1	67.9	60.2	52.0

Table 6. Results on the conventional multi-modal reasoning tasks with LLaMA-2 [38] as the LLM. † denotes the training set of corresponding dataset is included.

Relation (SR), Instance Interaction (IIn), Visual Reasoning (VR), and Text Recognition (TR). We further conduct experiment on this benchmark. The results are shown in Table 8. Pink also exhibits the best performance among compared MLLMs. Our model has significant advantages in IIn, SR, and IL tasks, which require strong fine-grained image perception ability of MLLMs. These results further validate the effectiveness of our method to enhance the fine-grained perception ability of MLLMs.

### A.4. Qualitative Analysis

We show qualitative results on different types of vision-language tasks and RC tasks. The results of mPlug-Owl [45], Shikra [4] and Qwen-VL [1] are from their official online demos.

**Knowledge-intensive QA.** Fig 5 depicts an example of this task, where the model is required to identify an album cover and provide an introduction for the album. mPlug-Owl and

Models	RefCOCO			RefCOCO+			RefCOCog		Visual-7W	LookTwice
	val	testA	testB	val	testA	testB	val	test		
Pink	88.3	91.7	84.0	81.4	87.5	73.7	83.7	83.7	85.1	73.5
Pink-LLaMA-2	89.0	92.1	84.6	82.6	88.3	74.5	83.8	84.4	85.1	73.8

Table 7. Fine-tuning results on RC tasks with LLaMA-2 [38] as the LLM.

Models	Overall	SU	Iid	IA	IL	IC	SR	IIn	VR	TR
Kosmos-2 [29]	54.4	63.4	57.1	58.5	44.0	41.4	37.9	55.7	60.7	25.9
Instruct-BLIP [6]	58.8	60.2	58.9	65.6	43.6	<b>57.2</b>	40.3	52.6	47.7	43.5
mPlug-Owl [45]	37.9	49.7	45.3	32.5	36.7	27.3	32.7	44.3	54.7	28.8
Qwen-VL [1]	62.3	71.2	66.4	67.7	53.5	44.8	43.8	62.9	<b>74.9</b>	51.2
Pink	66.2	73.1	69.1	69.1	60.5	55.2	<b>51.0</b>	<b>76.3</b>	70.4	<b>58.8</b>
Pink-LLaMA-2	<b>67.0</b>	<b>75.2</b>	<b>70.1</b>	<b>70.1</b>	<b>63.3</b>	53.8	50.2	69.1	74.3	50.0

Table 8. Results on the image split of SEED-Bench [16].

Qwen-VL can identify the album *The Wall* correctly but begins to make hallucination when introducing this album. Some introductions provide by mPlug-Owl do not align with the actual facts associated with the album. The answer of Shikra focuses on the content of the image, indicating a poor instruction-following ability of Shikra. In contrast, our model not only correctly identifies the album but also delivers an accurate and relevant introduction, showcasing its strong instruction-following capability.

**Fine-grained QA.** As shown in Fig. 6, mPlug-Owl fails to output the correct number of people in the image. It may be because that its lack of RC ability hinders its fine-grained understanding of images. Despite Shikra has RC ability, it also fails to give a correct answer. This result can validate that trained with limited RC tasks Shikra struggles to achieve a fine-grained perception of images. In contrast, our model makes a correct answer to the question, demonstrating its strong capability in fine-grained image perception.

**Referential comprehension reasoning.** We show a case of RC reasoning in Fig. 7. In this case, our model accurately locates Steve Nash. Shikra seems to be unaware of who is Steve Nash. Surprisingly, Shikra’s output coordinates point to the face of Dirk Nowitzki. We also need to point out the result of Shikra is generated using the instruction template that is identical to the one used during training. Shikra shows limited instruction-following ability when using a template that is different from the one used during training. Our model also shows a ability for multi-round conversation. It can understand what “him” refers to in the second question. As Shikra makes a wrong response for the first question, in the second question, we give the correct coordinates manually. However, Shikra fails to follow the instructions and begins to hallucination. Qwen-VL fails to locate Steve Nash correctly. In the multi-round conversation, it also gives a wrong response. These results can further validate the RC

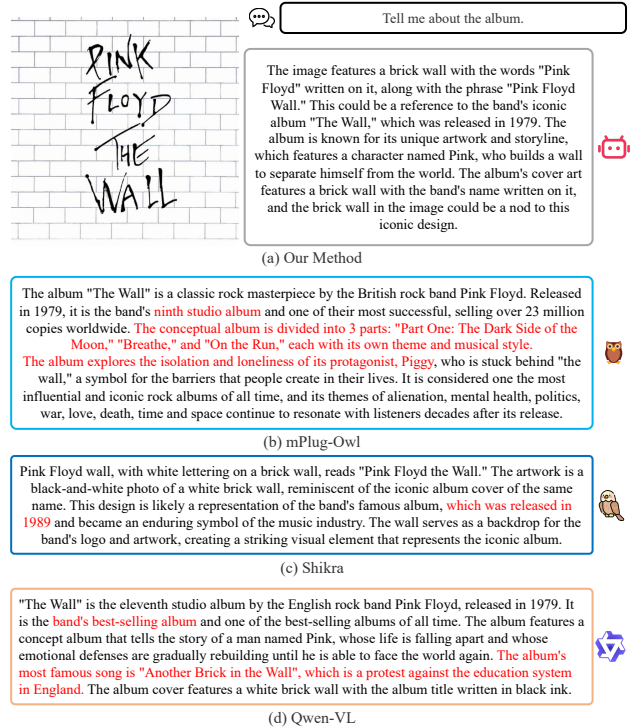


Figure 5. A comparison of knowledge-intensive QA. We mark the hallucination part in red.

ability of our model.

**More qualitative results.** More qualitative results of our model are shown in Fig. 8. Our model demonstrates the ability to generate unique descriptions with contextual information when provided with coordinates of a specific area. For instance, instead of simply outputting “helmet”, our model uses “woman’s” features for differentiation. Moreover, our

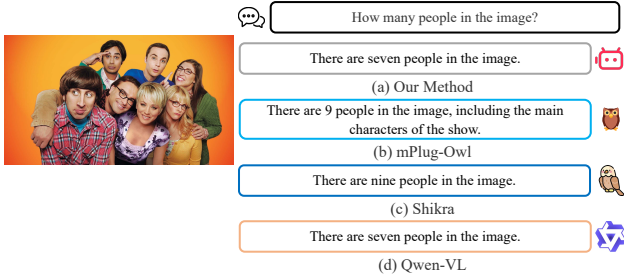


Figure 6. A comparison of fine-grained QA.



Figure 7. A comparison of referential comprehension reasoning. We mark the hallucination part in red.

model successfully locates items mentioned in descriptions that require outside knowledge. For example, it can correctly identify what can be drunk in the image. In the last two cases, our model not only provides correct answers but also locates the mentioned items in the image. This ability can achieve more applications.

We also present qualitative results of our model on multi-round conversation about RC in Fig. 9 and Fig. 10. Our method can understand the complex referential relationships in the dialogue context, *e.g.*, it, her and this instrument. As shown in Fig. 10, although the model initially gives the wrong answer about what is holding by the man, when it is told that its answer is incorrect, the model can make correc-



Figure 8. More qualitative results of our model.

tions and provide the correct answer. This result can further demonstrate the promising instruction-following ability of our model. Additionally, these qualitative results highlight that the integration of RC ability significantly expands the range of tasks our model can successfully handle, thereby broadening its application potential.

**Qualitative results of generated referring-expression-bounding-box pairs.** We illustrate some examples of generated referring-expression-bounding-box pairs in Fig. 11. For each object in the image, the bounding box description bootstrapping method can generate a description related to that object. Most of generated descriptions are correct. However, some generated descriptions exhibit incorrect or ambiguous descriptions that fail to uniquely identify an object. As shown in Fig. 11 (a) and (b), when leveraging visual grounding to locate these descriptions in the image, IOU between the predicted bounding box and the ground-truth bounding box is low. Our self-consistent filtering method can effectively filter them out. These results can further validate the effectiveness of our self-consistent bootstrap-

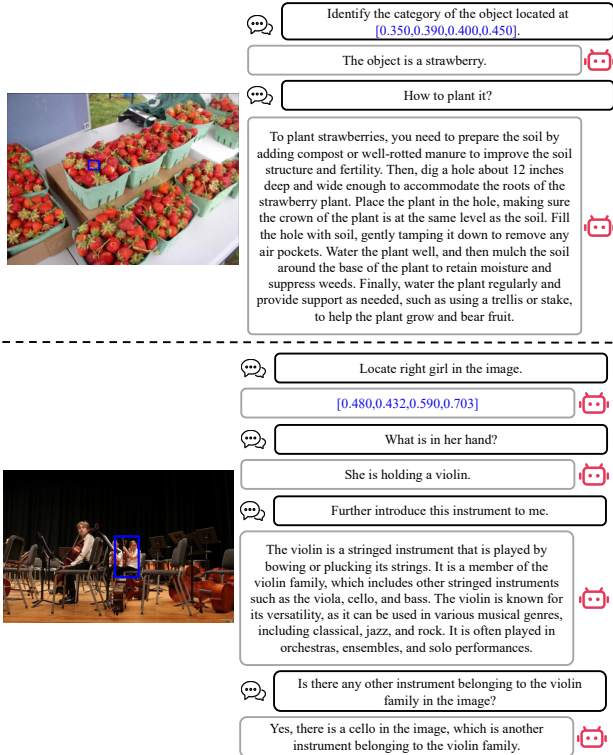


Figure 9. Qualitative results of our model on multi-round conversation.

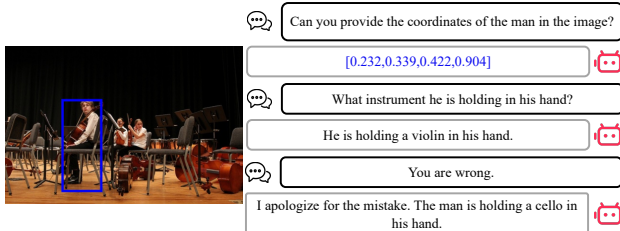


Figure 10. Qualitative results of our model on multi-round conversation with correction.

ping method to extend the object annotations to high-quality referring-expression-bounding-box pairs.

## A.5. Instruction Templates

**Instruction templates of visual relation reasoning.** We list all the instruction templates of visual relation reasoning task below. For the Task2, we use different instruction templates to ask the model output different contents (coordinates, class name, or both coordinates and class name).

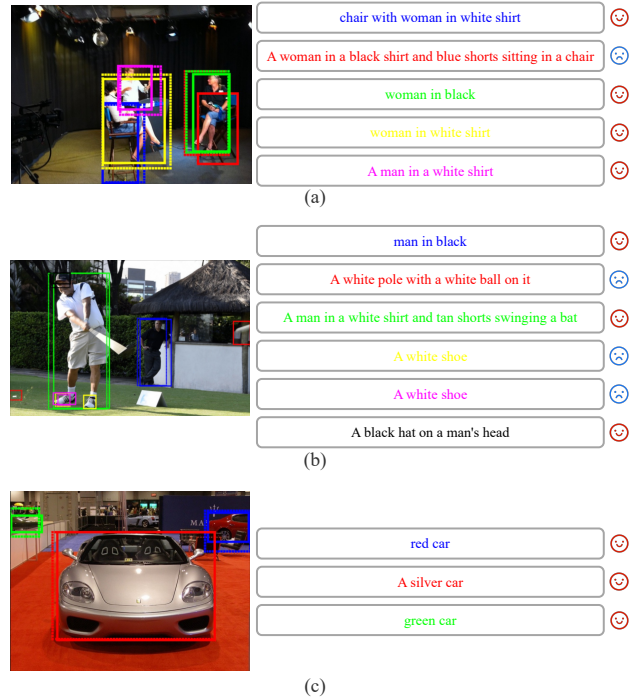


Figure 11. Qualitative results of generated referring-expression-bounding-box pairs. The solid rectangle represents the ground-truth bounding box. The dashed rectangle represents the bounding box generated by our model with visual grounding according to the generated description. ☹️ denotes the generated description is filtered out by the proposed self-consistent method.

Instruction templates of visual relation reasoning-Task 1

What is the relation between <subject> and <object>?  
 Describe the relation between <subject> and <object>?  
 Assist me in finding the relation between <subject> and <object> in the photo.  
 In the given image, could you find and tell me the relation between <subject> and <object>?  
 I need help identifying the relation between <subject> and <object>. Can you point it out in this image?  
 What is the relation between <subject> and <object> in this picture?  
 Could you describe the relation between <subject> and <object> in this image?  
 I'm having trouble identifying the relation between <subject> and <object>. Could you clarify it for me in this image?  
 Can you help me understand the relationship between <subject> and <object> in this image?  
 I'm trying to understand the relation between <subject> and <object>. Can you help me by describing it?  
 I need some assistance in identifying the relation between <subject> and <object> in this image.  
 In this image, can you find and describe the relation between <subject> and <object> for me?  
 Could you please explain the relation between <subject> and <object>.

Instruction templates of visual relation reasoning-Task2#1

Assist me in locating the position of all the objects <relation> the <subject>?

I want to know the coordinates of all the objects <relation> the <subject>?

Detect all the objects have a relationship <relation> with the <subject> and output there locations.

There are some objects that are <relation> the <subject>. Could you tell me there locations?

Identify all the objects that have a relationship <relation> with the <subject>. Where are they located?

Please locate all the objects that are <relation> the <subject> and provide their coordinates.

Find all the objects that have a relation of <relation> with the <subject>. Can you give me their positions?

Point out the objects that are <relation> the <subject>. Where can I find them?

I need to locate all the objects that are <relation> the <subject>. Can you assist me with this task?

Could you help me find all the objects that have a relation of <relation> with the <subject>? Please provide their locations.

Please detect all the objects that are <relation> the <subject>. Output their positions.

Identify and provide the coordinates of all objects that are <relation> the <subject>.

Find the objects that have a relation of <relation> with the <subject>. Where are they situated?

What objects have the relation of <relation> with the <subject>? Could you locate them for me?

Can you help me locate all the objects that are <relation> the <subject> and give me their positions?

Output the positions of all objects that have a relation of <relation> with the <subject>.

Identify the objects that are <relation> the <subject>. Where are they located?

Please locate all the objects that are <relation> the <subject> and provide their positions.

Instruction templates of visual relation reasoning-Task2#2

Assist me in identifying the categories of all the objects <relation> by the <subject>?

Detect all the objects <relation> by the <subject> and output there categories, respectively.

There are some objects that are <relation> the <subject>. Could you tell me there categories?

I want to know the categories of all the objects whose relation is <relation> with the <subject>?

Identify the object categories that are <relation> the <subject>.

Find all objects that are related to <subject> using the relationship <relation>, and categorize them.

Your task is to recognize and classify all objects that are <relation> by the <subject>.

Please determine the categories of all objects that are <relation> by the <subject>.

Can you identify the categories of objects <relation> the <subject>?

Your job is to identify all objects that <relation> the <subject> and list their categories.

Detect and categorize all objects that are <relation> the <subject>.

I need you to determine the categories of all objects that <relation> the <subject>.

Identify and classify all objects that are <relation> the <subject>.

Please identify the categories of all objects that are <relation> the <subject>.

Please help me identify the object categories whose relationship is <relation> with <subject>.

Recognize and categorize all objects that are related to <subject> using the relationship <relation>.

I need you to categorize all objects that is related to <subject> with relationship as <relation>.

#### Instruction templates of visual relation reasoning-Task2#3

Your task is to locate all objects that have a relation <relation> with <subject> and classify them.

I need you to categorize and locate all objects that is related to <subject> with relationship as <relation>.

Please locate and categorize all the objects that have a relation of <relation> with <subject>.

Assist me in locating and classifying all the objects <relation> the <subject>?

Find all the objects that have a relation of <relation> with the <subject>. Can you give me their positions and categories

Your task is to locate all objects that have a relation <relation> with <subject> and classify them.

I need you to categorize and locate all objects that is related to <subject> with relationship as <relation>.

Please locate and categorize all the objects that have a relation of <relation> with <subject>.

Assist me in locating and classifying the position of all the objects <relation> the <subject>?

Find all the objects that have a relation of <relation> with the <subject>. Can you give me their positions and categories?

Your task is to locate and classify all objects that are related to <subject> using the relationship <relation>.

I need you to locate and categorize all objects having a relationship <relation> with the given <subject>.

Find all objects related to <subject> with the relationship <relation>. Categorize and locate them for me.

Your objective is to locate and classify the objects that are related to <subject> through the relationship <relation>.

I require you to detect and categorize all objects that have a relationship <relation> with <subject>.

Please find and classify all objects that has a relationship <relation> with <subject>.

Assist me in locating and categorizing all objects that related to <subject> with the relationship <relation>.

Find all objects that are related to <subject> using the relationship <relation>. Categorize and locate their positions.

Your task is to identify and classify all objects related to <subject> through the relationship <relation>.

I need you to locate and categorize all objects that have a relationship <relation> with <subject>.

Assist me in locating and classifying all objects that are related to <subject> through the relationship <relation>.

Find all the objects that has a relationship <relation> with <subject>. Categorize and locate their positions for me.

#### Instruction templates of coarse visual spatial reasoning.

We list all the instruction templates of coarse visual spatial reasoning task below. Similar to visual relation reasoning, different instruction templates are used to ask the model output different contents.

#### Instruction templates of coarse visual spatial reasoning#1

Identify the objects located at <loc> of <object>. Please classify them by category and provide their locations.

I need to know what objects are present at <loc> of <object>. Can you help me locate and categorize them?

Find all the objects at <loc> of <object>. Please provide me with their categories and locations.

I want to know the categories and positions of the objects located at <loc> of <object>.

Locate and classify all the objects at <loc> of <object>. Could you tell me the categories and positions of the objects present at <loc> of <object>?

Help me locate and categorize all the objects at <loc> of <object>.

I need to know the categories and locations of the objects at <loc> of <object>.

What are the categories and positions of the objects located at <loc> of <object>?

Identify and locate all the objects at <loc> of <object>. I need their categories and positions.

I want to know the categories and positions of the objects at <loc> of <object>.

Locate and classify all the objects at <loc> of <object>. Please provide me with their categories and positions.

#### Instruction templates of coarse visual spatial reasoning#2

What are the categories of the objects located at <loc> of <object>?

Detect and classify all the objects at <loc> of <object>. I need to know their categories.

Please find and categorize all the objects present at <loc> of <object>.

Give the categories of all the objects you can find at <loc> of <object>.

I need you to find and categorize all the objects that are at <loc> of <object>.

Please provide me with the categories of all the objects present at <loc> of <object>.

What types of objects are located at <loc> of <object>? Please list their categories.

Please find all the objects at <loc> of <object> and give me their categories.

What are the categories of the objects that are present at <loc> of <object>?

I need you to classify all the objects located at <loc> of <object>.

Please give me the categories of all the objects that are located at <loc> of <object>.

### Instruction templates of coarse visual spatial reasoning#3

What are the coordinates of the objects located at <loc> of <object>?

Detect and give the coordinates of all the objects at <loc> of <object>.

Please find and locate all the objects present at <loc> of <object>.

Give the detail locations of all the objects you can find at of <loc> <object>.

Locate all the objects and give there coordinates found at <loc> of <object>.

What are the positions of all the objects at <loc> of <object>?

Can you find and list the positions of all the objects present at <loc> of <object>?

Provide the coordinates of objects located at <loc> of <object>.

List and indicate the positions of all objects at <loc> of <object>.

Enumerate and specify the positions of all objects found at <loc> of <object>.

What objects are situated at <loc> of <object> and where precisely are they located?

What are the coordinates of all objects found at <loc> of <object>?

### Instruction templates of object counting#1

Can you tell me how many <category> are present in this picture?

I need to know the number of <category> in this image.

Count how many <category> are in this picture.

Please determine the quantity of <category> shown in this image.

How many instances of <category> can you find in this picture?

I would like to know how many <category> are visible in this image.

Count the number of <category> that you see in this picture.

Please provide me with the count of <category> in this image.

How many objects of <category> are in this image?

Can you count the items of <category> in this picture?

What is the total number of <category> in this image?

How many <category> can you spot in this image?

Please determine the quantity of <category> in this image.

Count the number of <category> that appear in this picture.

How many <category> are in the picture?

Counting the number of <category> appeared in the image.

Please give me the number of <category> appeared in the image.

### Instruction templates of object counting#2

How many objects in the image are of the same category as <object>?

Count the number of objects in the image that are similar to <object> in category.

What is the total count of objects that share the same category as <object> in the image?

How many objects in the image have the same category as <object>?

Count all the objects in the image that fall under the same category as <object>.

What is the number of objects that share the same category as <object> in the image?

Count the objects that belong to the same category as <object> in the image.

How many objects of the same category as the object represented by <object> appear in the image?

Count all the instances whose category is the same as <object> present in the image.

**Instruction templates of object counting.** We list all the instruction templates of object counting task below. <category> will be replaced by the category name.

**Instruction templates of object detection.** We list all the instruction templates of object detection task below. <category> will be replaced by the category name.

#### Instruction templates of object detection#1

Locate and mark the positions of all <category> in the image.

Find all the instances of <category> in the image and indicate their respective locations.

Spot and record the coordinates of every <category> present in the image.

Identify the <category> in the image and provide their precise locations.

Can you determine the positions of all the <category> in the image and list them?

Pinpoint the <category> in the image and give me their exact coordinates.

Locate all the <category> in the image and provide their locations in detail.

Detect and report the locations of all the <category> present in the image.

Find and list the locations of every <category> in the image.

Please identify the <category> in the image and give me their locations.

Provide me with the precise locations of all the <category> in the image.

Detect and record the positions of the <category> in the image.

Spot all the instances of <category> in the image and give me their coordinates.

Detect all the <category> in the image, and output there location.

There are some <category> in the image, could you help me to locate them and give me their coordinates.

What are the coordinates of the <category> in the image.

Give the detail locations of all the <category> you can find in the image.

#### Instruction templates of object detection#2

Locate all the items in the picture that share the same category as <object> and provide their coordinates.

Spot every object that belongs to the same category as <object> and indicate their positions.

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

Find all the objects that have a similar classification as <object> and output their locations.

Locate and report the coordinates of all the objects that share the category with <object>.

Detect all the objects in the image that have the same classification as <object> and provide their positions.

Spot all the objects that belong to the same category as <object> and show their coordinates.

Identify every instance that falls under the same category as <object> and report their locations.

Find and output the coordinates of all the objects that have the same category as <object>.

Locate all the objects in the picture that have a similar classification as <object> and display their positions.

Detect and report the positions of all the objects that share the category with <object>.

Spot every instance that has a similar classification as <object> and indicate its coordinates.

Identify all the objects that have the same classification as <object> and output their positions.

Find all the objects that belong to the same category as <object> and report their locations.

Locate and output the coordinates of all the items that have a similar category as <object>.

Detect all the instances in the image which have the same category with <object>, and output there location.

Detect and report the locations of all the instances present in the image, these instances should have similar category with <object>.

Given an <object>, please help me to find all the instances with the same category. The output should be the coordinates of detected instances.

**Instruction templates of multi-choices VQA.** Our instruction tuning dataset also includes the multi-choices VQA, *e.g.*, A-OKVQA. Therefore, we also construct some instruction templates for this task and list them below. Placeholder <options> will be replaced by the options.



### Instruction templates of multi-choices VQA

Please take a look at the image and select the correct answer for <question> from the options given below \n<options>.

Examine the image and select the best matched answer to the question: <question> from the options given below \n<options>.

There are some options \n<options>. I have a question for you: <question> Can you select the best matched answers from the given options based on the image?

Regarding the image, you need to identify the correct answer to the question <question> from the given options \n<options>.

Analyzing the image, can you identify the best matched answer to <question> from the given options \n<options>.

Looking at the image, can you quickly answer my question: <question>. Some potential answers are given in the following options \n<options>.

Referring to the image, please select the answer for this question: <question> from the options \n<options>.

Could you please check the image and select the answer for my question: <question> from the options \n<options>.

Here is an image and a question: <question> for you. Please select an option that can answer the question from the given options \n<options>.

For this image, I want to know which option can answer my question: <question> correctly. The options are \n<options>.

Take a look at the image, can you select the best matched answer to the following question: <question> from following options \n<options>.

Considering these options \n<options>. I need a correct selection from these options that can answer this question: <question> in regards to the image.