

Osprey: Pixel Understanding with Visual Instruction Tuning

- Supplementary Material -

Yuqian Yuan^{1*}, Wentong Li^{1*}, Jian Liu², Dongqi Tang², Xinjie Luo¹, Chi Qin³,
Lei Zhang⁴, Jianke Zhu^{1†}

¹Zhejiang University ²Ant Group ³Microsoft ⁴The HongKong Polytechnical University

In this document, we provides the following materials:

- Implementation details;
- More experiments;
- More qualitative results;
- Discussions on input region types.

1. Implementation Details

The AdamW [8] is used as the optimizer and the cosine annealing scheduler [9] is used to adjust learning rate. At the first training stage, we set the batch size to 128 and the learning rate to 1×10^{-3} for one epoch. At the second stage, we decrease the learning rate to 2×10^{-5} with a batch size of 4 and train for two epochs. At the final stage, the learning rate is further reduced to 1×10^{-5} with a batch size of 4 for two epochs. The maximum length of sequence in LLM is set to 2,048. All training is conducted on four NVIDIA A100 GPUs with 80GB memory. We leverage the DeepSpeed framework [10] for efficient large-scale model training. The training of the three stages costs 7, 15, and 48 hours, respectively. The input image size is set to 512×512 . All the training datasets are aggregated into a single dataloader to ensure the representational integrity. In the training process, the image and its corresponding mask-based instruction/response pairs are randomly selected from each dataset.

2. More Experiments

2.1. Additional Main Results

Region Level Captioning. We provide the quantitative comparisons on region level captioning task with box region-based approaches [11, 13, 14]. Specifically, we fine-tune Osprey-7B on the training set of RefCOCOg and employ the prompt like “Please give me a short description of <region>.” to prompt our model. The comparison results are shown in Table A1. One can see that our Osprey model exhibits competitive performance

with 16.6% in METEOR score and 108.3% in CIDEr score, thereby surpassing the recent GLaMM approach [13] by 0.4% and 3.3%, respectively. These results highlights the efficacy of Osprey with input referring pixel-level mask regions, demonstrating its superior capability in generating semantically relevant descriptions for object regions.

Method	Type	METEOR	CIDEr
GRIT [14]	Box	15.2	71.6
Kosmos-2 [11]	Box	14.1	62.3
GLaMM [13]	Box	16.2	105.0
Osprey-7B (Ours)	Mask	16.6	108.3

Table A1. Region captioning performance evaluated on the validation set of RefCOCOg.

Effectiveness of Osprey-724K. To validate the effectiveness of Osprey-724K dataset, we retrain the GPT4ROI model [16] and conduct experiments on open-vocabulary segmentation, referring object classification and detailed region description tasks. The results are presented in Table A2. It can be seen that the re-trained GPT4RoI model with Osprey-724K significantly outperforms the original one, especially on part-level region classification and Detailed Description tasks, where we observe impressive improvement of +20.89% SS and +15.33%. These results underscore the superior quality of our Osprey-724K dataset.

2.2. More Ablation Studies

Single-level vs. Multi-level Mask Features. To explore the effects of multi-scale features in Mask-Aware Visual Extractor, we carry out experiments on open-vocabulary segmentation and referring object classification tasks. A comparison between single-level and multi-level features is performed. We utilize the output of vision encoder at “res4” stage to represent single-level features. As shown in Table A3, multi-level mask features in Osprey significantly outperform single-level mask features in model training. This notable improvement demonstrates the effectiveness of Osprey with multi-level Mask-Aware Visual Extractor.

*Equal contribution.

†Corresponding author.

GPT4RoI	Cityscapes		LVIS		PACO		Detailed Description
	PQ	ADE	SS	S-IoU	SS	S-IoU	
Original	34.70	36.32	51.32	11.99	48.04	12.08	49.97
Re-trained	37.31	38.12	58.91	29.56	68.93	46.28	65.30

Table A2. Performance comparison between the original GPT4RoI [16] model and the re-trained one with Osprey-724K.

Method	Cityscapes		LVIS		PACO	
	PQ	ADE	SS	S-IoU	SS	S-IoU
Single-level	46.28	38.03	62.46	34.25	68.42	46.38
Multi-level	50.64	42.50	65.24	38.19	73.06	52.72

Table A3. Comparison between single-level and multi-level mask features in Osprey model training.

CLIP Vision Encoder	224	448	672	896	1120
ViT-Surgery-L [4]	26.52	28.15	27.26	25.18	24.61
ConvNeXt-L [7]	23.35	34.36	40.57	43.04	43.33

Table A4. Panoptic segmentation comparisons (PQ) using different vision encoders with different input sizes on ADE20K-150 [17]. The ground truth mask is used for evaluation.

Comparisons on Vision Encoders. To investigate the impact of ViT-based and ConvNeXt-based CLIP vision encoders across varying input sizes, we meticulously conduct the experiments on open-vocabulary panoptic segmentation using ViT-Surgery-L [4] and ConvNeXt-L [7] models. All experimental results are obtained by directly employing CLIP as a mask classifier with ground truth masks. Table A4 reports the comparison results. The experimental results reveal that the CNN-based CLIP exhibits superior generalization performance as the input size scales up. Specifically, we observe that the ViT-Surgery-L CLIP model achieves a higher PQ at a lower resolution (*i.e.*, input size 224) while facing challenges at higher resolutions. According to this phenomenon, we adopt a straightforward solution by embracing a CNN-based CLIP as the vision encoder in Osprey.

Impacts of Short-form Prompt and Positive/Negative Data. We conduct experiments to evaluate the impacts of short-form prompt and positive/negative samples on our Osprey-724K dataset. As depicted in Table A5, Osprey trained with both short-form prompt and positive/negative samples attains 65.24% SS and 38.19% S-IoU on the object-level LVIS dataset, bringing an improvement of +8.83% and +12.54% over the model trained without short-form prompt data. On the part-level PACO dataset, the Osprey model trained with only short-form prompt achieves +22.80% SS and +29.43% S-IoU improvements over that without short-form prompt. Regarding the inclusion of positive/negative samples, Osprey model trained with them attains +1.69% SS and +1.49% S-IoU over the model trained without them on object-level LVIS dataset. On Part-level PACO dataset, +1.47% SS and +2.33% S-IoU performance improvements

Method	LVIS		PACO	
	SS	S-IoU	SS	S-IoU
w/o Short-form	56.41	25.65	50.26	23.29
w/o Pos./Neg.	63.55	36.70	71.59	50.39
Osprey-724K	65.24	38.19	73.06	52.72

Table A5. Performance comparisons with and without short-form prompt and positive/negative samples on *object-level* LVIS [2] and *part-level* PACO [12].

are obtained when positive/negative sample data are used. These experimental results underscore the effectiveness of incorporating short-form prompt and positive/negative data in our Osprey-724K for enhancing model performance.

2.3. More Qualitative Results

Fig. A2 shows some visual examples to better illustrate the effectiveness of Osprey. We see that Osprey is capable to generate accurate semantic predictions in these challenging scenarios. In Fig. A1, visual results are showcased based on the mask regions obtained from the off-the-self SAM [3] in “segment everything” mode. Furthermore, Fig. A3 provides comparisons with previous region-level and image-level methods [1, 5, 11, 16]. Our approach exhibits superior scene understanding results with fine-grained details. Please note that each box-cropped region is extracted as the input for image-level LLaVA-1.5 [5].

3. More Details of Osprey-724K

3.1. Example Illustrations

We provide several examples to illustrate the instruction-following data in our Osprey-724K dataset, including the object-level and short-form response instruction-following data in Table A6, the part-level instruction-following data in Table A7. Those data are generated through interactions with GPT-4, and the corresponding detailed prompts for GPT-4 are given from Table A9 to Table A12. Besides, Table A8 showcases positive and negative samples in robustness data.

3.2. Details on Task Prompt

Different prompt templates are used for training the Osprey model based on different instruction-following data. The question templates are randomly selected from the corresponding lists. Please refer to Table A13~Table A15 for more details.

4. Discussion on Types of Input Region

Osprey can handle various input instructions of referring region, including point, box and scribble types, which can be considered as free-form masks. Our Mask-Aware Visual



Figure A1. Visual results of Osprey based on the class-agnostic masks from off-the-self SAM [3]. With the pixel-level mask regions and task-specific prompts, the semantic understanding results are obtained, including (a) open-vocabulary categories, (b) short descriptions, and (c) detailed descriptions. Zoom-in for better view.

Extractor is compatible with these inputs. Compared to these coarse types, fine-grained masks can more precisely represent objects, achieving pixel-level alignment for accurate semantic understanding. Besides, some efficient SAM-based models, like EfficientSAM [15] and EdgeSAM [18], have been developed to make the acquisition of masks faster with lower cost.



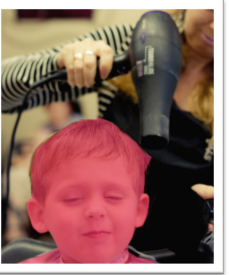





















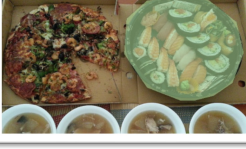







	<p>Give me a detailed description of <region>.</p> <p>A tall, slender vase is positioned in the middle of the image. It's filled with a variety of flowers, including purple and white ones. The flowers are arranged in a way that they appear to be growing from the vase, adding a sense of natural beauty to the scene.</p>			<p>Give me a detailed description of <region>.</p> <p>A young boy with brown hair is sitting in a chair, getting his hair blow dried by a woman. He is smiling and appears to be enjoying the experience.</p>	
	<p>Please give me a detailed description of <region>.</p> <p>A woman wearing a dress is standing in the grassy field, holding a kite string in her hand. She appears to be flying the kite while looking up at the sky, where a colorful kite can be seen soaring overhead.</p>			<p>Can you describe the <region> in detail?</p> <p>A woman wearing a green shirt is standing in the middle of the group. She is smiling and appears to be posing for a photo.</p>	
	<p>I'd like to know more about the region highlighted by <region> in the picture provided.</p> <p>A woman wearing a black dress is sitting on the bed, holding a cup of tea in her hand. She appears to be the mother or caregiver of the child in the image.</p>			<p>What details can you give me about the region outlined by <region> in the photo, please?</p> <p>A black double-decker bus with the words "The Ghost Bus Tours" written on the side is parked on the street. It is a tour bus, taking people on a tour of the city.</p>	
	<p>Describe the region <region>, please.</p> <p>A young boy wearing a black baseball uniform with the number 10 on the back is running towards a baseball. And he is holding a baseball glove in his hand.</p>			<p>Please describe the <region>.</p> <p>There is a pile of broccoli on a wooden tray, surrounded by other vegetables.</p>	
	<p>What is in <region>? Using a single word or phrase.</p> <p>A catcher's face mask.</p>			<p>What is in <region>? Using a single word or phrase.</p> <p>Bottle cap.</p>	
	<p>What is in <region>? Using a single word or phrase.</p> <p>Ladder step.</p>			<p>What is in <region>? Using a single word or phrase.</p> <p>Rim of a plate.</p>	
	<p>Give me a short description of <region>.</p> <p>A black plate with sushi.</p>			<p>What is in <region>?</p> <p>A chocolate donut.</p>	
	<p>What is in <region>?</p> <p>Olympic rings on the panda.</p>			<p>What is in <region>?</p> <p>Wooden bench in the snow.</p>	

Figure A2. Visual examples of Osprey on the input mask-based referring regions.



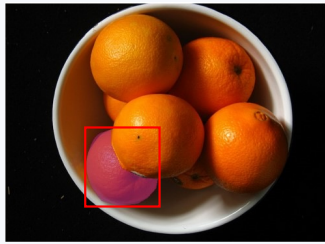
Osprey A young boy **wearing a red helmet** is riding a bicycle **with training wheels**, which is designed for children to learn how to ride a bike. He is **wearing a red shirt** and appears to be riding the bike confidently.

Kosmos Two young boys on a bike.

Shikra A young boy on a bike.

GPT4RoI A boy in an orange shirt riding a bike.

LLaVA-1.5 The image features a young boy sitting on a wooden toy bike, which is designed to look like a real bicycle. The boy is wearing an orange shirt and appears to be enjoying his time on the toy bike. The bike is positioned in the middle of the scene, with the boy sitting comfortably on it.



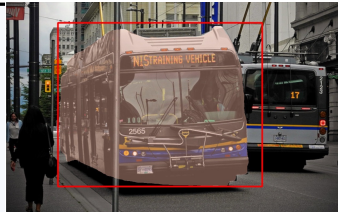
Osprey The **bottom left** orange in the bowl is visible, with its stem pointing towards the left side of the bowl. It is the **closest orange** to the viewer, and is positioned **at the 7 o'clock position**.

Kosmos Oranges in a bowl.

Shikra Fruit in the bowl.

GPT4RoI The orange on the left.

LLaVA-1.5 The image features a close-up view of an orange, which is the main focus of the scene. The orange is placed in a bowl, and it appears to be a ripe, fresh fruit. The orange is surrounded by a few other oranges, which are located in the background. The bowl containing the oranges is placed on a table, creating a visually appealing and inviting display.



Osprey A bus with the **number 2565** on it is parked on the left side of the street, next to a bus with the **number 17** on it. The bus with the number 2565 is the closest to the viewer, and it has **a yellow sign** on the front.

Kosmos A bus and a truck driving down a street.

Shikra Bus with "training vehicle" written on it.

GPT4RoI A bus heading to vancouver.

LLaVA-1.5 The image depicts a city street with a large black and white bus driving down the road. The bus is in motion, and it is passing by a bus stop where a person is waiting. The bus stop is equipped with a bus stop sign, which is visible in the scene. The street is lined with buildings, creating an urban atmosphere. There is also a traffic light present in the scene, indicating that the bus is traveling on a busy street with controlled intersections.

Figure A3. Qualitative comparisons with previous region-level and image-level approaches [1, 5, 11, 16]. The same prompt is adopted to obtain the detailed descriptions, which is selected randomly from Table A13. Our method showcases more accurate region-level semantic understanding with fine-grained details.

Context type 1: Image-level description

The image presents a lively market scene with a group of people buying fruits and bags. There are multiple individuals in the market, all browsing through the fresh produce available. A significant variety of fruits are showcased in the market, including bananas, oranges, and apples. Bananas can be seen in several groups, with some green and yellow bananas occupying different areas of the market. Meanwhile, oranges and apples are displayed in smaller sections among the fruits. In addition to fruits, handbags are also being sold at the market, attracting the attention of the customers. Overall, the market bustles with activity as people gather around the fresh fruits and bags, contemplating their purchases.

Context type 2: Boxes

person: [0.507,0.409,0.698,0.740], person: [0.243,0.496,0.558,0.746], person: [0.196,0.422,0.395,0.708], orange: [0.761,0.537,0.820,0.569], orange: [0.809,0.553,0.841,0.570], orange: [0.841,0.552,0.868,0.571], banana: [0.671,0.814,0.770,0.887], banana: [0.599,0.703,0.820,0.817], banana: [0.885,0.829,0.941,0.893], apple: [0.811,0.584,0.851,0.603], apple: [0.873, 0.568,0.900,0.586], apple: [0.778,0.580,0.819,0.601], handbag: [0.473,0.110,0.607,0.201], handbag: [0.491,0.202,0.611,0.267], handbag: [0.583,0.105,0.696,0.204].



Context type 3: Mask region captions

<region1> (person: [0.507,0.409,0.698,0.740]): gray shirt wearing glasses. // woman with gray shirt standing next to man. // woman in gray shirt facing camera on right. // the woman in the grey shirt with a watch on her wrist. // a short haired woman in jeans shopping.
<region2> (person: [0.243,0.469,0.558,0.746]): the lady with the blue shirt. // the back of an older woman with her hair in a barrette with a blue jacket on. // navy blue shirt. // woman back in blue. // a woman is wearing blue sweater.
<region3> (person: [0.196,0.422,0.395,0.708]): a woman in glasses shops in an open air fruit market. // a woman in a gray coat and scarf.

Response type 1: mask region description

<region1>: In another part of the market, a short-haired woman in a gray shirt and jeans is shopping. She seems to be looking directly at camera. A noticeable detail is the watch worn on her wrist. This lady, who is also wearing glasses, is standing next to a man.
<region2>: An older woman is visible, her back facing us. She's donning a navy blue sweater. Her hair, neatly fastened with a barrette, adds a note of detail to her appearance.
<region3>: Another woman, clad in a gray coat and scarf, is seen shopping in the open-air fruit market. This woman, similarly wearing glasses, appears to be engrossed in the selection of fresh produce.

Response type 2: conversation

Question 1: Can you describe the woman in **<region1>** and what she is doing?
Answer 1: The woman in **<region1>** is wearing a gray shirt, jeans, and glasses. She has short hair and a watch on her wrist. She is shopping in the lively market, browsing through the fresh produce, such as fruits and bags.
Question 2: What can you tell me about the woman in **<region2>**?
Answer 2: The woman in **<region2>** is wearing a navy blue shirt or sweater. She has her back towards the camera, with her hair styled in a barrette. She appears to be an older woman, also browsing through the market, looking at the various items for sale.
Question 3: How would you compare the person in **<region1>** and the person in **<region2>**?
Answer 3: The person in **<region2>** is positioned on the left side of the image, facing away from the camera. On the other hand, the person in **<region1>** is located more towards the center-right of the image, facing the camera. Both individuals are wearing different colored shirts, with the person in **<region2>** wearing a blue shirt and the person in **<region1>** wearing a gray shirt.
Question 4: Is there anything notable about the person in **<region3>**?
Answer 4: The person in **<region3>** is described as a woman wearing a gray coat and scarf. They are wearing glasses and seem to be shopping in the open-air fruit market, providing an indication of the bustling activity taking place in the scene.
Question 5: Based on the image, how would you describe the overall atmosphere of the market?
Answer 5: The market has a bustling and lively atmosphere. There's a sense of excitement and energy as people gather around the fresh fruits and bags for sale. The variety of items available, as well as the interactions among customers, contribute to the vibrant ambience of the market scene.

Response type 3: short-form response

Question 1: What color is the shirt of the woman in **<region1>**?
Answer 1: Gray.
Question 2: What accessories does the woman in **<region1>** wear?
Answer 2: Glasses.
Question 3: What color is the woman's sweater in **<region2>**?
Answer 3: Blue.
Question 4: What is the woman in **<region2>** wearing on her head?
Answer 4: Barrette.
Question 5: How is the woman in **<region3>** dressed?
Answer 5: In a gray coat and scarf.
Question 6: What type of market is shown in the image?
Answer 6: Fruit market.

Table A6. One example to illustrate the **object-level and short-from response instruction-following data** in Osprey-724K. The top block indicates the contexts such as *image-level descriptions and object boxes* used to prompt GPT-4, and the bottom block shows *three types of responses*. Note that we only show some regions for simple illustration, and the visual image is not used to prompt GPT-4. We only show it here as a reference.

Context:
 <region1> spoon: dark grey,plain,metal,opaque
 <region2> bowl: dark green,plain,ceramic,opaque
 <region3> spoon:tip: dark grey,plain,metal,opaque
 <region4> spoon:bowl: light grey,plain,metal,opaque
 <region5> spoon:neck: dark grey,plain,metal,opaque
 <region6> spoon:handle: dark grey,plain,metal,opaque
 <region7> bowl:inner body: dark green,plain,ceramic,opaque
 <region8> bowl:rim: dark green,plain,ceramic,opaque

Response type:
 Question: What's in <region1>?
 Answer: Spoon.
 Question: What is the color of the <region2>?
 Answer: Dark green.
 Question: Tell me about the color of tip of spoon in <region3>.
 Answer: Dark grey.
 Question: What's the category of <region4>?
 Answer: Bowl of a Spoon.
 Question: What is the color of <region5>?
 Answer: Dark grey.
 Question: Is the handle of spoon plain in <region6>?
 Answer: Yes.
 Question: What is the reflectance of <region7>?
 Answer: Opaque.
 Question: What's the material of <region8>?
 Answer: Ceramic.



Table A7. One example to illustrate the **part-level instruction-following data** in Osprey-724K. The top block shows *the attributes of each mask region* used to prompt GPT-4, and the bottom block indicates *conversations of the responses*. Note that the visual image and region masks are not used to prompt GPT-4. We only show them here as the reference.

Example1:
groundtruth:
 <region1>: sock
 <region2>: sock
 Question: Is the category of <region1> calf?
 Answer: No, it isn't.
 Question: Soccer ball is the category of <region2>, right?
 Answer: Not at all.

Example2:
groundtruth:
 <region1>: spoon
 <region2>: soup
 <region3>: soup
 Question: Does this area <region1> belong to category fork?
 Answer: This is not so.
 Question: The category of <region2> is soup, right?
 Answer: Yes, it is.
 Question: Does the category salad accurately describe this region <region3>?
 Answer: No, there is no salad in the region.



Table A8. Two examples to show **robustness data** in Osprey-724K based on the categories of LVIS. The top block shows negative instruction samples and the bottom block indicates the positive ones.


```
messages = [ {"role": "system", "content": f"""You are an AI visual assistant that can analyze a single image. You receive a detailed description/several descriptions of this image. In addition, most object locations within the image are given, along with detailed coordinates. These coordinates are in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.
```

Your role is to give a detailed description of each special region in the image. Instead of directly mentioning the bounding box coordinates, utilize this data to explain each region using natural language. Include details like **object category, object type, object color, attributes of the object, object locations, object state and other attributes**.

When using the information from the image and object region captions and coordinates, directly explain the region, and do not mention that the information source is the caption or the bounding box. Always answer as if you are directly looking at each region. Provide a direct answer without mention "this region". The answer template is: '<region1>: ...' """ }

```
]
for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample['context']})
    messages.append({"role": "assistant", "content": sample['response']})
messages.append({"role": "user", "content": '\n'.join(query)})
```

Table A9. The prompt used to generate the **detailed region description** in Osprey-724K. For each query, we show the prompt construction process for ChatGPT/GPT-4 to collect `query['response']` from `query['context']`, using few-shot in-context-learning, where examples are from `fewshot_samples`, each example including input `sample['context']` and output `sample['response']` as in [6]. `messages` is our final prompt. The prompt templates below also adopt the similar manner. Please see Table A10, Table A11 and Table A12 for the specific details.

```
messages = [ {"role": "system", "content": f"""You are an AI visual assistant, and you are seeing several object regions in a single image. What you see are provided with a detailed description for the whole image and each object region in this image, describing you are looking at. Answer all questions as you are seeing the image. The location of each object region is given in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.
```

Design a conversation between you and a person asking about each object region of this image. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers. All the regions given should be mentioned in the questions, when referring to each region, use <region1>, <region2>, etc. Include questions asking about the visual content of each object region in the image, including the **object category, object type, object color, object actions, object locations, relative positions between objects and other attributes, etc**. Only include questions that have definite answers:

(1) one can see the content in the object region of this image that the question asks about and can answer confidently;

(2) one can determine confidently from the object region of this image that it is not in the image.

Do not ask any question that cannot be answered confidently. Also include complex questions that are relevant to the content of each object region in the image, for example, asking about background knowledge of the objects, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details.

Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary. """ }

```
]
for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample['context']})
    messages.append({"role": "assistant", "content": sample['response']})
messages.append({"role": "user", "content": '\n'.join(query)})
```

Table A10. The prompt used to generate the **conversations response** data in Osprey-724K.

```
messages = [ {"role": "system", "content": f"""You are an AI visual assistant, and you are seeing several object regions in a single image. What you see are provided with a detailed description for the whole image and each object region in this image, describing you are looking at. Answer all questions as you are seeing the image. The location of each object region is given in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.
```

Design a conversation between you and a person asking about each object region of this image. The answers must be in one word or one phrase. Ask diverse questions and give corresponding answers. All the regions given should be mentioned in the questions, when referring to each region, use <region1>, <region2>, etc. Include questions asking about the visual content of each object region in the image, including the **object category, object type, object color, object actions, object locations, relative positions between objects and other attributes, etc.** Only include questions that have definite answers:

(1) one can see the content in the object region of this image that the question asks about and can answer confidently;

(2) one can determine confidently from the object region of this image that it is not in the image.

Do not ask any question that cannot be answered confidently. Do not ask any question that is not mentioned. Do not ask any question that cannot be answered with one word or phrase.

Most importantly, **the answer must be in one word or short phrase.** """ }

```
]
```

```
for sample in fewshot_samples:
```

```
    messages.append({"role": "user", "content": sample['context']})
```

```
    messages.append({"role": "assistant", "content": sample['response']})
```

```
messages.append({"role": "user", "content": '\n'.join(query)})
```

Table A11. The prompt used to generate the **short-form response** data in Osprey-724K.

```
messages = [ {"role": "system", "content": f"""You are an AI visual assistant that can analyze a single image. There are some regions in this image, each region is an object or a part of the object. You receive a short description with some words, separated by commas, for the common attributes of each region, which may contain category name, color, pattern & markings, material and reflectance etc. If a region is a part of an object, the category name is described as "object:part", like "person:body".
```

According to each description, design a conversation between you and a person asking about each region of this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include diverse questions asking about the attributes of each region including **category, part category, color, pattern & markings, material and reflectance.** Each region must involve 1-2 questions, when referring to each region, use <region1>, <region2>, etc. Answer the question using as few words as possible (single or two words). Only include questions that have definite answers: one can see the content in the region of this image that the question asks about and can answer confidently.

Do not ask any question that cannot be answered confidently. """ }

```
]
```

```
for sample in fewshot_samples:
```

```
    messages.append({"role": "user", "content": sample['context']})
```

```
    messages.append({"role": "assistant", "content": sample['response']})
```

```
messages.append({"role": "user", "content": '\n'.join(query)})
```

```
messages.append({"role": "user", "content": '\n'.join(query)})
```

Table A12. The prompt used to generate the **part-level attributes** instruction data in Osprey-724K.

- "Can you provide me with a detailed description of the region in the picture marked by <region>?"
- "I'm curious about the region represented by <region> in the picture. Could you describe it in detail?"
- "What can you tell me about the region indicated by <region> in the image?"
- "I'd like to know more about the area in the photo labeled <region>. Can you give me a detailed description?"
- "Could you describe the region shown as <region> in the picture in great detail?"
- "What details can you give me about the region outlined by <region> in the photo?"
- "Please provide me with a comprehensive description of the region marked with <region> in the image."
- "Can you give me a detailed account of the region labeled as <region> in the picture?"
- "I'm interested in learning more about the region represented by <region> in the photo. Can you describe it in detail?"
- "What is the region outlined by region in the picture like? Could you give me a detailed description?"
- "Can you provide me with a detailed description of the region in the picture marked by <region>, please?"
- "I'm curious about the region represented by <region> in the picture. Could you describe it in detail, please?"
- "What can you tell me about the region indicated by <region> in the image, exactly?"
- "I'd like to know more about the area in the photo labeled <region>, please. Can you give me a detailed description?"
- "Could you describe the region shown as <region> in the picture in great detail please?"
- "What details can you give me about the region outlined by <region> in the photo, please?"
- "Please provide me with a comprehensive description of the region marked with <region> in the image, please."
- "Can you give me a detailed account of the region labeled as <region> in the picture, please?"
- "I'm interested in learning more about the region represented by <region> in the photo. Can you describe it in detail, please?"
- "What is the region outlined by <region> in the picture like, please? Could you give me a detailed description?"
- "Please describe the region <region> in the image in detail."
- "Can you offer a thorough analysis of the region <region> in the image?"
- "Could you elaborate on the region highlighted by <region> in the picture provided?"
- "Please share more information about the zone emphasized with <region> in the photo."
- "What insights can you give about the area denoted by <region> in the image presented?"
- "Can you share a comprehensive rundown of the region denoted by <region> in the presented image?"
- "I'd like to know more about the region highlighted by <region> in the picture provided."
- "Work through the important details of the area <region> in the image."
- "Illustrate the area represented by <region> through a descriptive explanation."
- "Examine the region <region> closely and share its details."

Table A13. The list of instruction templates for detailed mask-region description used in Osprey.

- "Please give me a short description of region <region>."
- "Can you give me a short description of <region>?"
- "Can you provide me with a short description of the region in the picture marked by <region>?"
- "I'm curious about the region represented by <region> in the picture. Could you describe it in few words?"
- "What can you tell me about the region indicated by <region> in the image in few words?"
- "I'd like to know more about the area in the photo labeled <region>. Can you give me a concise description?"
- "Could you describe the region shown as <region> in the picture concisely?"
- "What can you give me about the region outlined by <region> in the photo?"
- "Please provide me with a brief description of the region marked with <region> in the image."
- "Can you give me a brief introduction of the region labeled as <region> in the picture?"
- "I'm interested in knowing the region represented by <region> in the photo. Can you describe it in several words?"
- "What is the region outlined by <region> in the picture like? Could you give me a streamlined description?"
- "Can you provide me with a brief description of the region in the picture marked by <region> please?"
- "I'm curious about the region represented by <region> in the picture. Could you describe it in few words please?"
- "What can you tell me about the region indicated by <region> in the image?"
- "I'd like to know more about the area in the photo labeled <region> please. Can you give me a simple description?"
- "Could you describe the region shown as <region> in the picture in several words?"
- "What attributes can you give me about the region outlined by <region> in the photo please?"
- "Please provide me with a simple description of the region marked with <region> in the image please."
- "I'm interested in learning more about the region represented by <region> in the photo. Can you describe it in few words please?"
- "What is the region outlined by <region> in the picture like please? Could you give me a simple and clear description?"
- "Please describe the region <region> in the image concisely."
- "Can you offer a simple analysis of the region <region> in the image?"
- "Could tell me something about the region highlighted by <region> in the picture briefly?"
- "Please share some information about the zone emphasized with <region> in the photo."
- "What insights can you give about the area denoted by <region> in the image presented?"
- "Can you share a simple rundown of the region denoted by <region> in the presented image?"
- "I'd like to know some attributes about the region highlighted by <region> in the picture provided."
- "Work through the important attributes of the area <region> in the image."
- "Illustrate the area represented by <region> with some important attributes."

Table A14. The list of instruction templates for brief mask-region description used in Osprey.

- "<category> is the category of <region>, right?"
- "Is the category of <region> <category>?"
- "Does this area <region> belong to category <category>?"
- "Is <category> the appropriate classification for this area <region>?"
- "Does category <category> accurately describe this region <region>?"
- "The category of <region> is <category>, right?"
- "Is this area <region> classified under category <category>?"
- "Is it correct to say this area <region> falls into category <category>?"
- "Is the classification of this region <region> aligned with category <category>?"

Table A15. The list of instruction templates for the mask-region positive/negative categories used in Osprey.

References

- [1] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2, 5
- [2] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 2
- [3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2, 3
- [4] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 2
- [5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2, 5
- [6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 8
- [7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 2
- [8] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1
- [9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1
- [10] Microsoft. Deepspeed. <https://www.deepspeed.ai/>, 2023. 1
- [11] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1, 2, 5
- [12] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *CVPR*, pages 7141–7151, 2023. 2
- [13] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 1
- [14] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 1
- [15] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. EfficientSAM: Leveraged masked image pretraining for efficient segment anything. *arXiv preprint arXiv:2312.00863*, 2023. 3
- [16] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 1, 2, 5
- [17] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 2
- [18] Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint arXiv:2312.06660*, 2023. 3