

RegionGPT: Towards Region Understanding Vision Language Model

Supplementary Material

A. Data

A.1. Instructions for region-level understanding.

The list of instructions used to briefly describe the region content are shown in Tab. 21. For detailed region description, the instructions are shown in Tab. 22. To describe the relationship between the regions, the instructions in Tab. 23 are used. Tab. 24 illustrates the instructions for region classification. For referring expression comprehension, we convert the location task to choice problem, selecting the regions which match the query description.

A.2. Instruction Tuning Data.

We list the region-level instruction tuning data in Tab. 9 and Tab. 10 for the Pre-training and Fine-tuning stage. For multiple task dataset, we integrate all the instruction-following data into a multi-turn conversation format. This approach enhances training efficiency and ensures the model’s capability in multi-round dialogues. We perform random selection across all annotations for each category, retaining a target number of annotations per category. Images with no annotations selected are discarded.

Pre-train Data	Size	Task	Random Sampling
V3Det [47]	177K	Classification	100 per class
VG [24]	108K	Caption & Relationship	No
RefCOCO [23]	25.8K	Caption & REC	No

Table 9. Region-level training data in the Pre-training Stage.

Fine-tuning Data	Size	Task	Random Sampling
V3Det [47]	98K	Classification & Caption	10 per class
COCO [28]	1.5K	Classification	20 per class
LVIS [19]	52K	Classification	20 per class
VG [24]	108K	Caption & Relationship	No
RefCOCO [23]	25.8K	Caption & REC	No

Table 10. Region-level training data in the Fine-tuning Stage.

B. More Ablation Studies

Instruction for region classification. For the region classification task, we have developed three distinct instruction modes. As shown in Tab. 13, the first mode involves a one-turn conversation for all RoIs, inputting all RoIs into a single instruction, with the LLM outputting all categories simultaneously. The second mode is a multi-turn conversation for all RoIs, where the LLM conducts multiple rounds of dialogue, classifying one RoI per round.

The third mode is a one-turn conversation for one RoI, with the LLM classifying only one RoI per dialogue. The results in Tab. 11 show that multi-turn conversation mode outperforms the other modes, because the previously predicted box provides conditions for the subsequently predicted box, and only one prediction is made at a time, reducing the difficulty.

Mode	One-turn for all RoIs	Multi-turn for all RoIs	One-turn for one RoI
mAP	70.0	73.8	71.5

Table 11. **Ablation study on the instruction mode for region classification.** The object classification results on COCO 2017 *val* set are reported. We use ViT-B/16 from [55] as our visual backbone, whose input size is 512×512. All region instances in COCO are used as training data without random sampling.

The number of sample for each concept. To assess the impact of the number of annotations per category on classification performance, we conducted experiments on the COCO dataset with varying annotation quantities. We randomly sampled 10, 20, 50, and 200 annotations per category for training. As indicated in the Tab. 12, a consistent enhancement in performance was observed with an increasing number of sampled annotations. However, the marginal gain in performance diminished with more data. Notably, increasing annotations from 20 to 200 per category resulted in only a 4 mAP increase.

Num	10	20	50	200
mAP	52.7	56.8	57.3	60.9

Table 12. **Ablation study on the annotation quantities for region classification.** The object classification results on COCO 2017 *val* set are reported. We use ViT-B/16 from [55] as our visual backbone, whose input size is 512×512. Different from the other experiments, only the COCO classification region-level data is used to train model.

Ablation study on inaccurate annotations. Besides mask, RGPT accepts user-drawn bounding box as inputs. While it doesn’t directly accept points or scribbles, these can be easily converted into usable masks via SAM. RGPT is robust to inaccurate annotations, like dilated / eroded mask due to our refinement module, as shown in Tab 15 and 16.

C. More Qualitative Results

To demonstrate the multi-turn conversation and complex reasoning capabilities of RGPT, we provide more



One-turn for all RoIs

User: What category name best describes the region represented by $\langle region1 \rangle$, $\langle region2 \rangle$ and $\langle region3 \rangle$? **Answer the question using COCO-80 category names.**

RegionGPT: Car, person, kite.

Multi-turn for all RoIs

User: What category name best describes the region represented by $\langle region1 \rangle$? **Answer the question using COCO-80 category names.**

Assistant: Car.

User: Identify the object or feature present in the region denoted by $\langle region2 \rangle$.

Assistant: Person.

User: Can you specify the type of object or landscape within the bounds of $\langle region3 \rangle$.

Assistant: Kite.

One-turn for one RoI

User: What category name best describes the region represented by $\langle region1 \rangle$? **Answer the question using COCO-80 category names.**

Assistant: Car.

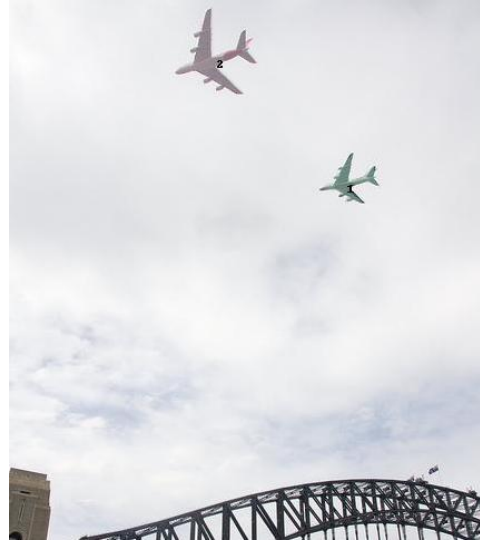
User: Identify the object or feature present in the region denoted by $\langle region2 \rangle$.

Assistant: Person.

User: Can you specify the type of object or landscape within the bounds of $\langle region3 \rangle$.

Assistant: Kite.

Table 13. In one-turn for all RoIs mode, all the regions are input to RGPT within one instruction. For multi-turn for all RoIs, the regions are split and construct sequential conversation. In one-turn for one RoI mode, the conversation is closed for after each instruction.



Ground Truth

User: What category name best describes the region represented by $\langle region1 \rangle$ and $\langle region2 \rangle$? **Answer the question using COCO-80 category names.**

RegionGPT: Airplane, airplane.

Prediction

User: What category name best describes the region represented by $\langle region1 \rangle$ and $\langle region2 \rangle$? **Answer the question using COCO-80 category names.**

RegionGPT: Aircraft, aircraft.

Table 14. Task-guided instruction prompt sometimes can not restrict the response format well. The prediction may not be within the COCO-80 classname set.

Model	w/ Feature Refinement		w/o Feature Refinement	
	w/ seg	w/ box	w/ seg	w/ box
mAP	71.0	70.4	65.6	60.0

Table 15. The object classification results on COCO 2017 val set. We use SigLip ViT-SO400M as the visual backbone and input region in the box or mask format.

Model	Box	Seg	Seg w dilation	Seg w erosion
mAP	69.3	70.0	71.0	68.2

Table 16. The object classification results on COCO 2017 val set. We input box, mask, dilated mask, and eroded mask to RGPT. The kernel is 5 for mask dilation and erosion for 336×336 input.

qualitative results. The conversations in Tab. 17 show that our model can infer relationship between regions. Tab. 18 demonstrates that our model can identify the target from multiple similar objects based on positional cues and

perform complex reasoning. Tab. 19 illustrates that our model can describe the information of a region in detail, it can even identify the brand of a motorcycle. Tab. 20 indicates that our model can do complex reasoning with multiple regions as input query.

D. Limitations and Future Work

The current task-guided instruction prompt sometimes can not restrict the response format well. As shown in Tab. 14, for region classification task, the output can be the synonym of ground truth classname. The evaluation of object classification can be reformulated as the semantic similarity between the prediction and ground truth name via a pre-trained text encoder.

E. Ethics Concerns

The large language model in our method is pre-trained with the corpus collected by previous works. Despite applying filtering, there may still be potential bias in its response.

User: What can be inferred about $\langle region1 \rangle$ and $\langle region2 \rangle$ relationship?

RegionGPT: From the image, it can be inferred that the woman and the horse share a close relationship, likely as a rider and a horse. The woman is walking towards the horse, which is standing in a stable, and they are in a stable or barn setting. This suggests that they might be preparing for a ride, training session, or simply spending time together. The woman's presence and interaction with the horse indicate that she is familiar with the animal and possibly has a bond with it.

User: What skills might $\langle region2 \rangle$ need for the potential activity?

RegionGPT: For the potential activity of horse riding, the woman would need a variety of skills and knowledge to ensure a safe and enjoyable experience. Some of these skills and benefits include:

1. Balance and coordination: Horse riding requires maintaining balance and coordination while riding the horse, especially when navigating different terrains or performing various maneuvers.
2. Communication and understanding: The woman should be able to communicate effectively with the horse through body language, voice commands, and aids to guide the horse and establish a mutual understanding between them.
3. Confidence and assertiveness: Riding a horse can be intimidating for some people, so having confidence and assertiveness in one's abilities is essential to overcome any fears and effectively control the horse.
4. Knowledge of horse care and behavior: Understanding the horse's needs, behavior, and body language is crucial for ensuring the horse's well-being and a successful riding experience.
5. Safety awareness: Being aware of safety precautions, such as wearing appropriate riding gear, knowing how to mount and dismount the horse, and understanding the horse's temperament, can help prevent accidents and injuries during horse riding.

By developing these skills and knowledge, the woman can enjoy a safe and enjoyable horse riding experience while also building a strong bond with her horse.



Table 17. Qualitative visualization of complex reasoning capability.

User: What is $\langle region1 \rangle$ doing, and how can the surroundings affect his mood?

RegionGPT: In the image, a man in a black wetsuit is standing on the beach, with his arms outstretched. He appears to be enjoying the ocean view and the beach atmosphere. The surroundings, including the presence of other people and surfboards, can positively affect his mood by providing a sense of camaraderie and shared enjoyment of the beach environment. The ocean view and the sound of the waves can also contribute to a relaxing and refreshing atmosphere, which can enhance his overall experience and mood.



Table 18. Qualitative visualization of complex reasoning capability.

User: Describe the $\langle region1 \rangle$ in detail.

RegionGPT: The motorcycle in the image is a black Harley Davidson. It is a large and powerful motorcycle, known for its distinctive design and style. The motorcycle is parked on the side of a road, and it appears to be well-maintained and in good condition. The man standing next to the motorcycle is wearing a jacket, which suggests that the weather might be cool or windy. The motorcycle's size and style make it stand out on the road, and its presence adds a sense of adventure and freedom to the scene.



User: What challenge might $\langle region2 \rangle$ face in this environment?

RegionGPT: In the image, a man is standing next to a parked motorcycle on a road. The environment appears to be a rural or countryside setting, which might present some challenges for the man. For example, he might face difficulties in finding gas stations, restaurants, or other amenities that are typically found in urban areas. Additionally, the road might be less maintained or less frequently traveled, which could lead to potential hazards or obstacles, such as potholes, debris, or wildlife. The man might also need to be prepared for changing weather conditions, as rural areas can sometimes experience sudden changes in weather or be more prone to natural disasters. Overall, the man would need to be self-sufficient and prepared for the challenges that come with traveling in a rural environment.

Table 19. Qualitative visualization of complex reasoning capability.

User: Please tell me the roles of $\langle region1 \rangle$, $\langle region2 \rangle$ and $\langle region3 \rangle$. Give the reason.

RegionGPT: The roles of the people in the image are as follows:

1. Batter: The person holding the baseball bat, who is preparing to hit the ball.
2. Catcher: The person wearing a baseball glove, positioned behind the batter to catch the ball.
3. Umpire: The person standing behind the catcher, observing the game and making decisions on plays.



The reason for these roles is that in a baseball game, the batter's objective is to hit the ball and score runs, while the catcher's role is to catch the ball thrown by the pitcher and prevent the batter from advancing. The umpire's role is to enforce the rules of the game, make decisions on plays, and ensure fair play.

User: What might $\langle region1 \rangle$ aim to achieve during the game?

RegionGPT: During the baseball game, the batter, who is holding a baseball bat and standing in the batter's box, aims to achieve a successful hit by hitting the ball thrown by the pitcher. The batter's goal is to hit the ball with enough force and accuracy to reach the bases and ultimately score runs for their team. This can be achieved by using proper batting techniques, timing, and strategy to anticipate the type and speed of the pitch. The batter's performance can significantly impact the outcome of the game, as their ability to hit the ball effectively can contribute to their team's overall success.

User: What might $\langle region2 \rangle$ aim to achieve during the game?

RegionGPT: The catcher, who is positioned behind the batter in the image, aims to achieve several goals during the game. First, the catcher's primary responsibility is to catch the pitches thrown by the pitcher. They must be prepared to react quickly and accurately to the ball's trajectory to prevent it from getting past them. Second, the catcher plays a crucial role in strategizing with the pitcher, suggesting pitch types and locations to keep the batter off balance and increase the chances of getting them out. Third, the catcher must be ready to throw the ball to the appropriate base to prevent the opposing team's runners from advancing or scoring. Overall, the catcher's goal is to contribute to the team's defensive efforts and help prevent the opposing team from scoring runs.

Table 20. Qualitative visualization of complex reasoning capability.

1. Provide a brief caption for the area indicated by <region>.
2. Describe in a short phrase the content within the bounds of <region>.
3. How would you succinctly caption the region highlighted by <region>?
4. Summarize the scene or object present in the section marked by <region>.
5. Can you give a concise description of what's depicted in <region>?
6. Draft a short title for the image content enclosed by <region>.
7. What brief caption would best describe the visual within <region>?
8. Offer a succinct interpretation of the area pointed out by <region>.
9. If you were to provide a short tagline for the content at <region>, what would it be?
10. Give a one-liner description of the region demarcated by <region>.
11. How would you encapsulate the essence of the segment labeled <region> in a few words?
12. Characterize the content of the image portion specified by <region> briefly.
13. Craft a mini headline for the visual element spotlighted by <region>.
14. In a few words, how would you narrate the content found within <region>?
15. Pen down a concise caption for the image section delineated by <region>.
16. A short caption of region <region>:
17. A short description of region <region>:
18. A photo containing the region <region>:
19. A region <region> that shows
20. Write a short description for the region <region>
21. Write a description for the region <region>
22. Provide a description of what is presented in the region <region>.
23. Briefly describe the content of the region <region>.
24. Can you briefly explain what you see in the region <region>?
25. Could you use a few words to describe what you perceive in the region <region>?
26. Please provide a short depiction of the region <region>.
27. Using language, provide a short account of the region <region>.
28. Use a few words to illustrate what is happening in the region <region>.
29. Provide an overview of what you see in the region <region>.
30. Can you break down the main elements present in this region <region>?
31. What are the key features or subjects captured in this region <region>?
32. Summarize the primary components of this region <region>.
33. Walk me through the different aspects of this region <region>.
34. Highlight the main points of interest in this region <region>.
35. What stands out to you the most in this region <region>?
36. If you were to give a brief overview of this region <region>, what would you mention?
37. List the primary objects or subjects you identify in this region <region>.
38. Describe the first few things that catch your attention in this region <region>.
39. How would you introduce this region <region> to someone who hasn't seen it?
40. What are the defining characteristics of this region <region>?
41. Give a concise description of the main content in this region <region>.
42. If you were to caption this region <region>, what might you say?
43. Describe the scene or setting depicted in this region <region>.

Table 21. The list of instructions for brief region description.

1. Describe in detail the object located at ⟨region⟩ in the image, including its appearance, style, and any visible details.
2. Provide a comprehensive description of the area marked by ⟨region⟩, focusing on textures, colors, and any notable features.
3. Elaborate on the artwork shown in the region indicated by ⟨region⟩, mentioning its color, appearance, size, style, and any standout features.
4. Give a detailed analysis of the scene within the boundary of ⟨region⟩, touching upon its components, ambiance, and any thematic expressions.
5. Craft a thorough narrative about the piece of the image highlighted by ⟨region⟩, from its aesthetic qualities to its possible historical context.
6. Explain in depth the characteristics and attributes of the subject found in the segment tagged with ⟨region⟩.
7. Generate a long, detailed caption for the segment of the image at ⟨region⟩, covering aspects such as its origin, material, and any symbolic meaning.
8. Paint a vivid picture with words about the region at ⟨region⟩, diving into the intricacies and nuances present in the area.
9. Zoom in on the area indicated by ⟨region⟩ and describe every discernible detail, from texture and color to form and function.
10. Offer an expanded description of the contents within the area marked by ⟨region⟩, encompassing its color, appearance, size, style, and any remarkable features.

Table 22. The list of instructions for detailed region description.

1. Explain the relationship between the area indicated by ⟨region⟩ and the region marked by ⟨region⟩ in terms of their visual or thematic connection.
2. Describe any functional or aesthetic connection between the elements at ⟨region⟩ and ⟨region⟩ in the image.
3. Analyze how the region ⟨region⟩ complements or contrasts with the area ⟨region⟩ in terms of design and composition.
4. Discuss the interplay between the features located at ⟨region⟩ and the attributes of the region at ⟨region⟩.
5. Detail the way in which the area labeled ⟨region⟩ interacts with or relates to the region designated by ⟨region⟩ within the image's context.
6. Assess the correlation or disparity between the segment at ⟨region⟩ and the segment at ⟨region⟩, including any observable influences or contrasts.
7. Compare the region ⟨region⟩ with the area ⟨region⟩ to determine how they either work together or differ substantially within the image.
8. Identify and elaborate on any thematic or stylistic relationships between the contents of ⟨region⟩ and ⟨region⟩.
9. Interpret the connection between the area at ⟨region⟩ and the region at ⟨region⟩, considering their positions, roles, or symbolism in the image.
10. Clarify how the part of the image within ⟨region⟩ corresponds with, or is disparate from, the part within ⟨region⟩ in terms of their visual narrative.

Table 23. The list of instructions for region relationship description.

1. Identify the object or feature present in the region denoted by ⟨region⟩.
2. What category best describes the area represented by ⟨region⟩?
3. Describe the content of the image section highlighted by ⟨region⟩.
4. Can you specify the type of object or landscape within the bounds of ⟨region⟩?
5. Which of the following categories best fits the region marked by ⟨region⟩? Provide your answer.
6. What can you discern from the area indicated by ⟨region⟩ in the image?
7. Categorize the visual element within the area designated by ⟨region⟩.
8. Give a brief description of the item or scene captured in the segment marked by ⟨region⟩.
9. Which classification would you assign to the visual content found at ⟨region⟩?
10. Determine and describe the primary subject located within ⟨region⟩.
11. How would you label the section of the image encompassed by ⟨region⟩?
12. Assess and classify the feature present within the confines of ⟨region⟩.
13. If you were to tag the section indicated by ⟨region⟩, what tag would you use?
14. What stands out to you in the region demarcated by ⟨region⟩? Please classify it.
15. Evaluate the content of the image portion pinpointed by ⟨region⟩ and provide its category.

Table 24. The list of instructions for region category description.

1. Given the mask proposals ⟨region⟩ in the image, can you pinpoint the one that matches ⟨description⟩.
2. From the provided masks denoted by ⟨region⟩ in the picture, which one best fits the description of ⟨description⟩?
3. Looking at the mask suggestions ⟨region⟩ in the image, identify the one that corresponds to ⟨description⟩.
4. In the image with mask proposals ⟨region⟩, please highlight the one that represents ⟨description⟩.
5. Considering the mask candidates ⟨region⟩ from the photo, which one would you associate with ⟨description⟩?
6. Among the mask proposals ⟨region⟩ in the visual, can you discern the one depicting ⟨description⟩?
7. From the set of masks labeled as ⟨region⟩ in the image, which one aligns with the description ⟨description⟩?
8. Based on the mask data provided as ⟨region⟩ in the photo, can you spot the one indicative of ⟨description⟩?
9. In the presented image with mask suggestions ⟨region⟩, determine which mask resonates with ⟨description⟩.
10. Given the mask assortment ⟨region⟩ in the image, please detect the one that matches the characteristics of ⟨description⟩.
11. Reviewing the mask candidates ⟨region⟩ from the picture, can you single out the one that fits ⟨description⟩?
12. From the list of mask proposals ⟨region⟩ in the image, identify the one that best encapsulates ⟨description⟩.
13. Considering the provided mask data ⟨region⟩ in the visual, which one would you say corresponds to ⟨description⟩?
14. In the snapshot with the mask proposals ⟨region⟩, please locate the mask that can be described as ⟨description⟩.
15. Based on the available mask candidates ⟨region⟩ in the image, can you pick the one that portrays ⟨description⟩?

Table 25. The list of instructions for referring expression comprehension.