

RegionGPT: Towards Region Understanding Vision Language Model

Key Ideas

Vision language models today have gotten really better in image understanding. They have gotten better to a point where humans can have good interaction with them. They have gotten really good at interpreting whole images.

Although these VLMs have become really good in interpreting whole images, they still struggle in understanding specific regions of the image. So tasks that require analyzing particular attributes on an object in a given region is difficult. You will essentially need a vlm that could connect the semantic aspects of the conversations to the spatial features really well.

One way of trying to provide this context is, you could send in the region localization information within the prompt. Like you can send in the coordinates of the object location and hope that the language decoder model can somehow learns to align the image features with the semantics you provide within the prompt. It is not very effective because it could create a gap

In the paper they introduce the following:

- A new general purpose VLM **RegionGPT** that can do really well in region level captioning and understanding, while also being really good as a general purpose VLM.
- They also additionally introduce task guided instruction prompting, which is a type of prompting for specific tasks
- Rich dataset for region level training.

RegionGPT:

The **RegionGPT** VLM takes in an image, a query and also a 2D region capturing a region in the image (could be a mask, or a box) and then provide answers that are specific to the particular region.



Figure 1. We introduce RegionGPT that enables complex region-level captioning, reasoning, classification, and expression comprehension capabilities for the multimodal large language model. Users can input regions of interest of any shape, utilizing (region) as a placeholder within the instruction at any position. Such placeholders are subsequently replaced with semantic region-level embeddings that are fed into the language decoder. Best viewed in color.

Model Architecture :

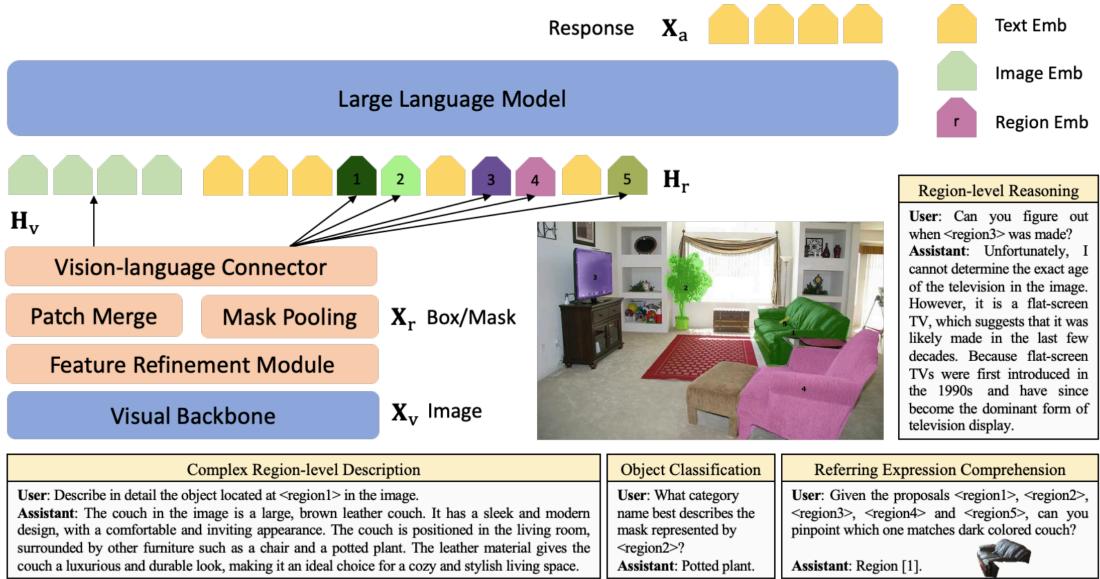


Figure 2. **Overview of the proposed RGPT architecture.** Starting from a visual backbone, we extract low-resolution semantic features from an input image X_v . Then, a feature refinement module is composed to obtain higher-resolution feature maps. With a patch merge module, the feature maps are further merged to reduce the length of input image-level sequence. The mask features are obtained by averaging the feature in the target region X_r , inputted as another branch, with Mask Pooling layer. Both the image-level feature and region-level feature share the connector for semantic consistency. The example interactions demonstrate the model's capabilities in complex region-level description, reasoning, object classification, and referring expression comprehension.

Image encoder : The visual backbone used here is a pretrained CLIP-ViT-L model to encode the image into a low resolution feature map.

Feature Refinement Module: The feature refinement module takes in the low resolution image features coming out of the encoder and upsamples it by 4.

Mask pooling : Extract the region from the upsampled feature map and perform average pooling to get region level feature.

Patch merging: Apply adaptive pooling on the whole high level image feature map to bring down to a low resolution for efficiency.

MLP Vision-language Connector: The vision language connector is basically a MLP layer that maps the vision features to word embedding space. Takes in the image level and region level features and maps them to the word embedding space.

Large Language Model: They use Vicuna 7B for language model. The text input is tokenized and then transformed into word embedding space. The image and region level embeddings are also sent in directly into the LLM alongside the text embeddings.

Training Prompt and Task guided Instruction Prompting:

General prompt format:

The prompt used for training is a multi-turn conversational data.

For each image \mathbf{X}_v , we generate multi-turn conversation data as follows:

($[X_v, X_1-q], X_1-a, \dots, X_t-q, X_t-a$)

Where:

- T is the number of conversation turns
- X_{t-q} is the t -th instruction (question)
- X_{t-a} is the t -th response (answer)

To allow region-level responses, they introduce `<region>` tag in the user input that is replaced by the region embeddings.

They also introduce a second type of prompting called **Task-guided instruction prompt**

Task-guided instruction prompt:

COCO Object Detection User: What category name best describes the region represented by <code><region1></code> ? Answer the question using COCO-80 category names. Assistant: TV User: How would you label the section <code><region3></code> ? Assistant: Laptop	Referring Expression Comprehension User: From the provided masks denoted by <code><region1></code> , <code><region2></code> , <code><region3></code> , <code><region4></code> and <code><region5></code> , which one fits “a desktop monitor with Bruce Lee photo on it”? Answer the question using the template Region [x]. Assistant: Region [2] . User: Match the object with its description: a black Xbox 360. Assistant: No corresponding options found.	
---	--	---

Table 1. **Task-guided instruction prompt** to indicate the response format. Two specific tasks are illustrated here. The guided prompt is highlighted in red. We empirically show that instruction prompt is able to adjust the output format and significantly improves the mAP and accuracy on COCO 2017 val set.

For certain tasks, like coco object detection, when you provide the vlm with an image and region, you would expect it to provide an class name within the 80 classes available in coco dataset. For such closed set problems, they tuned the vlm with specific prompts that would push vlm into providing outputs in a particular format.

The perform pre-training and finetuning of their VLMs using different datasets.

Pre-training Phase:

They perform joint pre-training at both global and region level tasks. vision and language backbones are frozen. Feature refinement layer and MLP are train.

Global training:

LAION-CCSBU-558K (Image captioning)

Region level training:

Visual Genome, ReferCOCOg, and V3Det (these are turned as multi-turn data before training)

Fine-tuning phase:

In this phase, the visual encoder is frozen and the rest of the network is learnable.

Here for COCO2017 and ReferCOCOg datasets, the task guided prompts are used when training for the closed set classification task.

Additionally, the model is also finetuned on a richer dataset that has more dense region level information

RecapD Dataset generation:

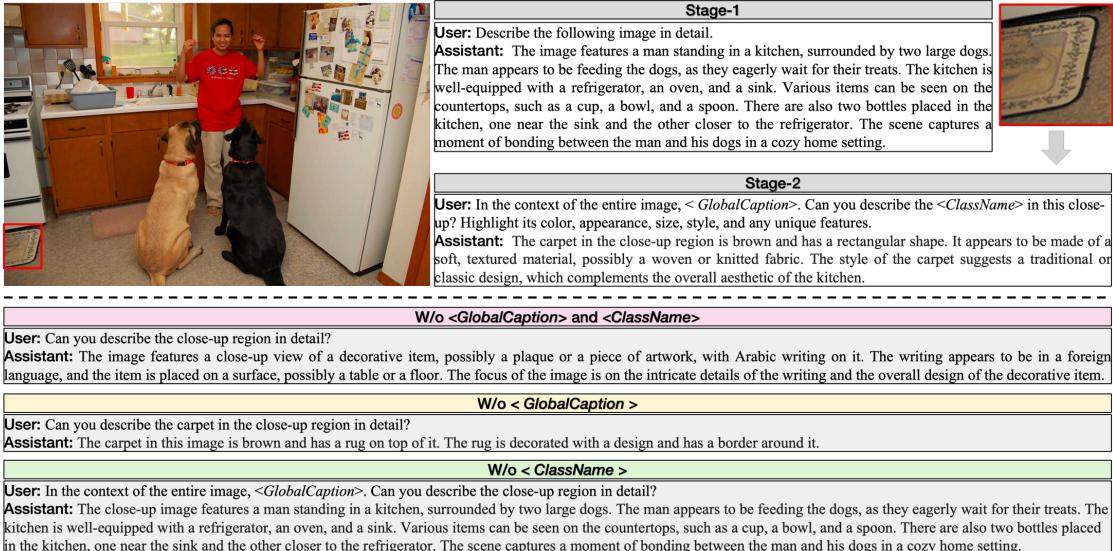


Figure 3. **Overview of the GPT-assisted region caption generation.** In the upper block, we show our two-stage paradigm in which the final output from the assistant accurately described the local region in terms of color, size and style. In contrast, without the global caption and/or the class name, the assistant either generates vague or over-simplified description, or fails to focus on the region but instead repeating the global context.

The RecapD dataset is generated in two stages. First an image is sent into an existing VLM, in this case LLaVA to generate a global level caption of the image. Next step is to use this caption as pretext, and then send in a ROI of an image as input and then ask the VLM to describe it. This combined approach helps with generating detailed object level data while also preserving the global context.

They also use GPT to rephrase the captions in the dataset as augmentation.

Impressions

What are highlights/impressions you have from the paper? For e.g. any ablation study/analysis that draws your attention?

Region classification : VLM ability to correctly identify a highlighted region within an image

Methods	PT	IT	Vision	LLM	mAP	Acc (%)
CLIP [38]	-	-	ViT-L	-	58.9	-
RegionCLIP [59]	-	-	R50x4	-	58.3	-
LLaVA [†] [30]	595K	158K	ViT-L	Vicuna-7B	-	40.04
Shikra [†] [9]	600K	5.5M	ViT-L	Vicuna-7B	-	53.91
GPT4RoI [†] [57]	266K	731K	ViT-L	LLaVA-7B	-	64.01
PVIT [†] [6]	13.7M	243K	ViT-L + R50x4	LLaVA-7B	-	64.53
ASM [49]	~22M	~22M	ViT-L	Hasky-7B	69.3	-
Ours	923K	953K	ViT-L	Vicuna-7B	70.0	80.61

Region Captioning:

Model	RefCOCOg		Visual Genome	
	METEOR	CIDEr	METEOR	CIDEr
GRIT [50]	15.2	71.6	17.1	142.0
SLR [54]	15.9	66.2	-	-
Kosmos-2 [36]	14.1	62.3	-	-
Ours	16.9	109.9	17.0	145.6

METEOR and CIDEr and image-level captioning evaluation metrics

Referring Expression Comprehension:

Method	MDETR[22]	Shikra [9]	Kosmos-2 [36]	MiniGPT-V2 [8]	Ours
val	81.64	82.27	60.57	84.44	86.44
test	80.89	82.19	61.65	84.66	86.96

▼ eg

COCO Object Detection

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

Assistant: TV

User: How would you label the section $\langle region3 \rangle$?

Assistant: Laptop

Referring Expression Comprehension

User: From the provided masks denoted by $\langle region1 \rangle$, $\langle region2 \rangle$, $\langle region3 \rangle$, $\langle region4 \rangle$ and $\langle region5 \rangle$, which one fits “a desktop monitor with Bruce Lee photo on it”? Answer the question using the template Region [x].

Assistant: Region [2] .

User: Match the object with its description: a black Xbox 360.

Assistant: No corresponding options found.



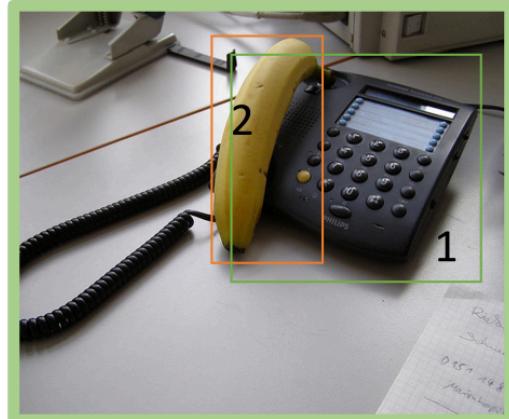
Region GPT in action:

User: What is present in *<region2>* ?

Assistant: The kite in the image is red, yellow, and blue.

User: Describe the relationship between *<region1>* and *<region2>*.

Assistant: The little girl is holding the kite and preparing to fly it.



User: Why is *<region2>* put on *<region1>* ?

Assistant: The **banana** is placed on the **phone** to create a **humorous** and **unexpected** scene. It is not a typical use for a banana, as it is not a communication device. The banana's placement on the phone might be a **playful act** or a **prank**, intending to evoke amusement or surprise from the viewer. This unusual combination of objects can be seen as a light-hearted and creative way to break away from the ordinary and make people smile or think about the scene.

User: Why is the banana is put on the phone?

GPT-4V: The image shows a banana placed on the handset of a telephone. This is likely a humorous setup, playing on the idea of communication by substituting a telephone receiver with a banana. It's a visual pun or a piece of playful staging rather than a functional setup. The juxtaposition is meant to be amusing, suggesting the banana as an alternative to the usual phone receiver.