

RegionGPT: Towards Region Understanding Vision Language Model

Qiushan Guo¹, Shalini De Mello^{2†}, Hongxu Yin^{2†}, Wonmin Byeon², Ka Chun Cheung²,
Yizhou Yu¹, Ping Luo¹, Sifei Liu²

¹The University of Hong Kong ²NVIDIA

Region-level reasoning

User: How to keep `<region1>` engaged in playtime?

RegionGPT: It is a dog. To keep the dog engaged in playtime, it is essential to provide a variety of toys and activities. In the image, the dog is sitting on the grass next to a basketball and a frisbee. These toys can be used for fetching, catching, and throwing, which are all enjoyable activities for dogs. Additionally, incorporating different types of toys, such as balls, ropes, and tug toys, can help keep the dog mentally and physically stimulated. Regular playtime sessions, including walks, jogs, and interactive games, can also help keep the dog engaged and happy.



Region-level description

User: Describe in detail the object located at `<region2>` in the image.

RegionGPT: The basketball in the image is orange and appears to be a standard-sized basketball used for playing basketball games. It is placed on the grass, likely for the dog to play with or for the owner to use during a game. The basketball is likely made of rubber with a leather or synthetic cover. In the close-up view, the basketball appears to be deflated or not fully inflated.

Object classification

User: What is the class name of `<region3>`?

RegionGPT: It is a frisbee.

Referring Expression Comprehension

User: From the provided masks denoted by `<region1>`, `<region2>` and `<region3>`, which one represents the sport that Michael Jordan is best at?

RegionGPT: Basketball. Region [2].

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.

Abstract

Vision language models (VLMs) have experienced rapid advancements through the integration of large language models (LLMs) with image-text pairs, yet they struggle with detailed regional visual understanding due to limited spatial awareness of the vision encoder, and the use of coarse-grained training data that lacks detailed, region-specific captions. To address this, we introduce RegionGPT (short as RGPT), a novel framework designed for complex region-level captioning and understanding. RGPT enhances the spatial awareness of regional representation with simple yet effective modifications to existing visual encoders in VLMs. We further improve performance on tasks requiring a specific output scope by integrating task-guided instruction prompts during both training and inference phases, while maintaining the model’s versatility for general-purpose tasks. Additionally, we develop an automated region caption data generation pipeline, enriching the training set with detailed region-level captions. We demonstrate that a universal RGPT model can be effectively applied and significantly enhancing performance across a range of region-level tasks, including but not limited to complex region descriptions, reasoning, object classification, and referring

*Qiushan Guo was an intern at NVIDIA during the project. † equal contribution.

expressions comprehension. Code will be released at the [project page](#).

1. Introduction

Vision Language Models (VLMs) have marked a notable convergence between visual and linguistic domains in artificial intelligence. With the emergence of Multimodal Large Language Models (MLLMs) [1, 2, 14, 25, 28, 29, 59], there has been a notable enhancement in the field’s ability to interpret images and streamline interactions between humans and VLMs. However, despite their effectiveness in understanding entire images, these models still struggle with analyzing specific regions in detail. On the other hand, fine-grained understanding is vital for advanced vision tasks, including the analysis of object attributes and the interpretation of inter-object relations.

Addressing region-level complex understanding in VLMs demands the alignment of spatial information and semantics. To achieve this, existing works [9, 28, 35, 59] learn inputting regions of interest in textual form, e.g. $[x_1, y_1, x_2, y_2]$, which share the same model structure as that used for image-level tasks. However, this relies heavily on the language decoder to interpret the position, inadvertently overlooking the prior positional information provided by the visual encoder. Such an oversight can lead to a gap in effectively integrating visual cues with linguistic context,

which is crucial for tasks involving detailed image understanding. In a more advanced approach, GPT4RoI [56] introduces spatial boxes with RoI-aligned features, training the model specifically on region-text pairs. Despite that, the positional format is restricted to a box. And yet the potential for region-specific visual representation, which could offer more expressive fine-grained details and hence benefit downstream vision tasks, remains under-explored.

In this paper, we present RGPT, a general framework designed to facilitate complex region-level captioning and understanding. Specifically, we discover that simply refining the visual features extracted by CLIP and employing Mask Pooling to accommodate regions of interest (RoI) of any shape significantly enhances the language model performance on understanding spatial-aware semantic concepts. Furthermore, we develop task-guided instruction prompts that seamlessly integrate the vision tasks, such as closed-set classification and referring expression comprehension, into our framework. This is achieved by specifying these tasks with visual question answering and response formats. Existing available region-level captioning datasets, such as ReferCOCOg [22] and VG [23], tend to provide overly simplistic descriptions of regions, lacking detailed attributes such as color, shape, style and their spatial relation with the surroundings. To reduce the burden of manual labeling, we propose an automated pipeline for annotating detailed region-level captions, which is achieved by reformatting the existing object detection dataset and employing a two-stage GPT-assisted approach. Our annotated captions average 87.14 words per region, substantially surpassing the 8.46 words in ReferCOCOg, thereby providing richer contextual information for each region.

Our contributions are threefold: (1) We propose RGPT, a general framework that harnesses the capabilities of LLMs to tackle complex region-level captioning and understanding tasks. RGPT is designed for open-ended vision questions, catering to both image-level and region-level tasks. (2) We design task-guided instruction prompts to specify the output format, thereby eliminating ambiguities in the responses. By transforming vision tasks into VQA tasks, the output patterns are aligned to the language model. (3) We present a novel data reformation approach and pipeline, leveraging GPT-assistant, to create high-quality, detailed region-level captions. Our approach significantly enhances the descriptive richness of these captions, with an average word count of 87.14 words per caption.

2. Related Work

2.1. Large Language Model

Large Language Models have recently gathered considerable interest in the realm of Natural Language Processing (NLP), which is viewed as a form of artificial general in-

telligence. This surge in attention is attributable to their remarkable proficiency in several key areas: language generation, in-context learning, and the integration of extensive world knowledge and reasoning abilities. The early potential of LLM was first showcased by groundbreaking works such as, BERT [15] and GPT [36]. This initiated a trend of scaling up that led to a succession of significant advancements, including T5 [38], GPT-3 [4], Flan-T5 [13], PaLM [12], among others. As training data and model parameters expanded, this scaling-up progress culminated in the development of ChatGPT [41] by OpenAI. ChatGPT, leveraging a generative pre-trained model and refined through instruction tuning [34] based on human feedback, demonstrates unparalleled capabilities in engaging in human-like conversations. Rapid advancements in open-source LLMs, such as Llama [43], Llama-2 [44] and Vicuna [11], have also started to make them increasingly competitive with ChatGPT.

2.2. Multimodal Large Language Model

LLMs have demonstrated formidable capabilities in prior knowledge and reasoning, prompting interest in other modalities. This has led to efforts aimed at extending LLMs into the multimodal domain, where they can interact with and interpret information across various inputs beyond just text. For image modality, end-to-end instruction tuning on image-text pairs is proposed to connect the visual backbone with language decoder. Flamingo [1], BLIP-2 [25], LLaVA [29] and MiniGPT4 [59] are the pioneers to train vision-language connector or language decoder on image-level vision tasks, such as image captioning and visual question answering. Inspired by these pioneers, more recent works are emerged to construct user-friendly interaction dataset [18, 24] and lightweight trainable weights [17, 55]. Some other interesting works have made remarkable progress by extending LLM to audio [7, 20], medical VQA [31, 57] and control systems [16, 32].

2.3. Region-level Vision Language Model

Traditional region-level tasks are common practice in computer vision, such as object detection [5, 39], instance segmentation [19] and semantic segmentation [40], which aims at localizing the regions of interest and close-set classification. Open-vocabulary region-level recognition tasks [50, 51] target at understanding an object with arbitrary categories described by texts. Recently, region-aware MLLMs, like KOSMOS-2 [35], Shikra [9], MiniGPT-2 [8] and LLaVA [28], learn inputting regions information in textual form, which heavily rely on the language decoder to interpret position. We argue that incorporating a visual spatial-aware module can extract region-level features more directly and efficiently. By utilizing a visual-language connector, these features enable the complex region-level captioning and reasoning ability. VisionLLM [47], GPT4RoI

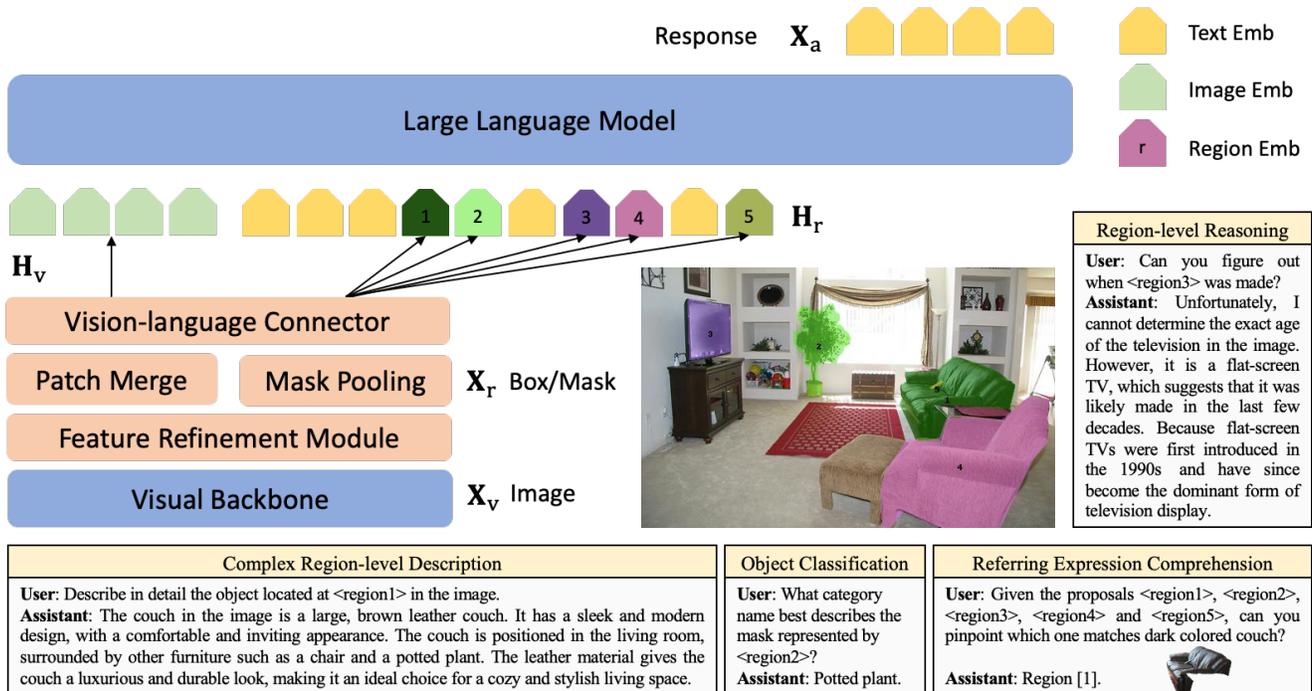


Figure 2. **Overview of the proposed RGPT architecture.** Starting from a visual backbone, we extract low-resolution semantic features from an input image X_v . Then, a feature refinement module is composed to obtain higher-resolution feature maps. With a patch merge module, the feature maps are further 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.

[56] and ASM [48] utilize spatial boxes with ROI-aligned features to align the region-level features into LLM word embedding space. However, the input positional format is restricted to a box. Besides, the region visual representation for fine-grained details remains under-explored. On the contrary, our model supports any-shape region as input and focuses on complex reasoning and captioning. Meanwhile, we introduce task-guided instruction prompts to transforming vision tasks into VQA tasks, whose output patterns are aligned with the language model.

3. Method

RGPT is a multimodal large language model with strong capabilities in understanding and referring to specific regions. It can take a inputs of any 2D region, usually in the form of a box or a mask, and provide answers based on instructions. By setting rules for how it should respond to instructions, the model is able to output in a useful and consistent format. This feature allows RGPT to classify objects at the region level in a closed vocabulary. Additionally, by giving the model region proposals, it can identify specific objects or regions given the query description. This makes RGPT a practical tool for tasks that require detailed understanding and processing of different regions within an image.

3.1. Model Architecture

An overview of our method RGPT, for region-level understanding and image-level understanding is shown in Fig. 2. It contains an image encoder to extract semantic features, a feature refinement module for the refinement of the low-resolution feature map, an MLP layer to project visual features into the word embedding space and a large language model taking both visual and text tokens.

Visual Backbone. RGPT adapts a pretrained CLIP ViT-L [37] model as the visual backbone. The visual backbone is frozen during the entire training process. Specifically, an input image X_v is encoded into a low-resolution feature map $Z_{LRes} = f(X_v)$ by the visual backbone.

Feature Refinement Module. The visual backbone yields a low-resolution feature map, which is not capable of representing small-scale regions and objects. To further refine the visual features, we introduce two deconvolution layers of stride 2 to produce feature maps up-scaled by $4\times$, i.e., $Z_{HRes} = g(Z_{LRes})$. Our method aims to understand any arbitrary-shaped region of the image, therefore, we choose Mask Pooling to extract region-level features from the high-resolution feature map. More concisely, we average the features of Z_{HRes} in region X_r to get the region-level feature $Z_r = \text{MaskPool}(Z_{HRes}, X_r)$.

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.



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.

Since the visual features are flattened as sequence input to language decoder, therefore, high-resolution feature map gets longer sequence input, which significantly lowers the training and inference efficiency. Hence, we simply use adaptive pooling layer [30] to merge image feature patches for image-level feature $\mathbf{Z}_v = \text{AdaPool}(\mathbf{Z}_{\text{HRes}}, (H, W))$, where (H, W) is the target shape of the low-resolution output feature map.

MLP Vision-language Connector. To project visual features from the visual backbone into the language model’s word embedding space, a two-layer MLP is adopted as the vision-language connector. The embedding of a full image is represented as $\mathbf{H}_v = h(\mathbf{Z}_v)$ and the region embedding is $\mathbf{H}_r = h(\mathbf{Z}_r)$. Both the image-level and region-level features share the same connector for semantic consistency.

Large Language Model. RGPT incorporates Vicuna (7B) [11] as the language decoder. Textual inputs are first tokenized and transformed into word embeddings. Both image-level and region-level features, after being processed through the MLP connector, are directly input into the language decoder.

3.2. Region-level Instruction Tuning

General prompt format. For each image \mathbf{X}_v , we generate multi-turn conversation data $([\mathbf{X}_v, \mathbf{X}_q^1], \mathbf{X}_a^1, \dots, \mathbf{X}_q^T, \mathbf{X}_a^T)$, where T is the number of turns, \mathbf{X}_q^t is the t -th instruction and \mathbf{X}_a^t is the corresponding response, following [29]. The image is always used as the starting input of the first instruction to provide the contextual information. To facilitate region-level responses, we introduce the special token $\langle region \rangle$ as a placeholder in the user input prompt, which will be replaced by the corresponding region embedding \mathbf{H}_r . The training loss is the standard auto-regressive train-

ing objective. We only set the response as the learning target, ignoring the instruction parts.

Task-guided instruction prompt. The language model is trained without imposing restrictions on the range of its outputs, in pursuit of achieving flexibility and adaptability. However, certain tasks demand specific output formats. For instance, in the context of the COCO detection task, when provided with a specified bounding box, the model is required to output only the corresponding class name. This response must be selected from a predetermined set of 80 candidate categories. To tailor the model’s responses to specific tasks, we craft custom instruction prompts to guide the model to a desirable output format, as shown in Tab. 1.

The task-guided instruction ensures that the model remains both versatile and accurate in its task-specific applications. We empirically show that our carefully-designed instruction prompt significantly improves the mAP result on COCO 2017 *val* set.

Pre-training stage. To maintain and enhance the model’s capability in understanding images at both the global and regional levels, we adopt a joint pre-training strategy encompassing both image-level and region-level tasks. For global image understanding, we utilize the LAION-CC-SBU-558K dataset [28], employing image captioning as a pretext task. In parallel, to bolster the model’s proficiency in interpreting and interacting with regional aspects of images, we engage it with tasks derived from datasets like Visual Genome [23], ReferCOCOg [22], and V3Det [46]. These datasets are transformed into multi-turn conversational formats, which help the model in region-based relationship understanding, captioning, and classification.

While training, we keep the visual encoder and the language models’ weights frozen, and train the feature refine-

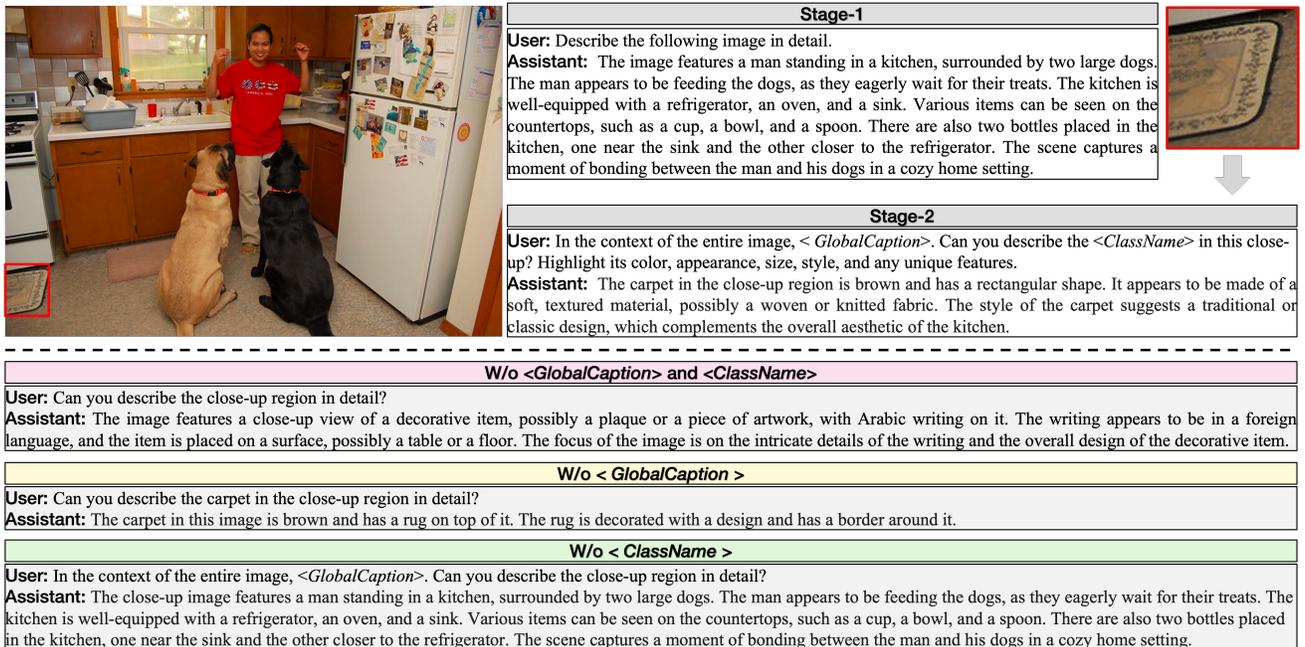


Figure 3. **Overview of the GPT-assisted region caption generation.** In the upper block, we show our two-stage paradigm in which the final output from the assistant accurately described the local region in terms of color, size and style. In contrast, without the global caption and/or the class name,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

ment module and MLP vision-language connector to align the image features with language embeddings.

Fine-tuning stage. We only keep the visual encoder weights frozen, and continue to update the feature refinement module, MLP connector and language model weights. Our objective is to develop a model capable of advanced region-level captioning and reasoning. However, the complexity of existing datasets like ReferCOCOg and Visual Genome for captioning is insufficient for our needs. To address this gap, we additionally incorporate the GPT-assisted region caption dataset (detailed in Sec. 3.3) into our training regime. Furthermore, we craft task-guided instructive prompts on COCO-2017 and ReferCOCOg *train* set to develop the model’s ability for closed-set object classification and understanding of referring expressions, as shown in Tab. 1.

Data Processing. To enhance training efficiency, we optimize the V3Det dataset by balancing the number of bounding boxes across each category. During the pre-training phase, we limit to 100 boxes per category, and in the fine-tuning phase, this is further reduced to 10 boxes per category. For the closed-set object classification task on the COCO dataset, we retained 20 boxes per category for fine-tuning. In the case of Visual Genome, we randomly sampled up to 10 boxes per image to generate dialogues. This filtering process is employed to generate dialogues that are rich in diversity and complexity. Although this filtering ap-

proach reduces the data’s volume, it is important to note that both the visual backbone and the language model have already been pre-trained on large-scale datasets. The strong prior knowledge allows the model to perform effectively even with a smaller, yet diverse set of data. Our data processing strikes a balance between training efficiency and robust model performance.

3.3. GPT-assisted Region Caption Generation

In this section, we present a GPT-assisted dense caption generation pipeline, developed to construct the Region Caption Dataset (RecapD). Distinct from traditional image-text pair datasets that typically offer a holistic description of images, RecapD provides in-depth annotations focusing on specific object regions within images. These descriptions emphasize attributes such as color, shape, material, and the spatial relationships between objects. The primary objective of RecapD is to address the challenges associated with region-level understanding and referencing in images, thereby significantly enhancing the capabilities of vision language models in detailed visual comprehension.

A two-stage approach. We explore using an existing global-level image captioning VLM, i.e., LLaVA [29] for region-specific tasks. A naive approach is to crop the region of interest (RoI) and adjust it to fit the model’s input format. However, this method often leads to inaccurate captions due to the lack of contextual information from the image’s sur-

Dataset	Images	Regions	Average words
ReferCOCO [22]	20K	142K	3.50
ReferCOCO+ [22]	20K	142K	3.53
ReferCOCOg [22]	25.8K	95K	8.46
VG [23]	82.4K	3.8M	5.09
Ours	213K	1.5M	87.14

Table 2. **Comparison of our dataset with available region-level caption datasets.** Our dataset stands out with a significantly higher average word count per region caption compared to other datasets. This richness in detail provides a robust foundation for complex region-level understanding.

rounding areas. The absence of surrounding information also makes it infeasible for conveying spatial relationships between objects.

Alternatively, we work around the limitation of the VLMs, which does not support the simultaneous input of both global images and local region patches. To circumvent this, in the first stage, we generate a global-level caption for the image using the VLM. This global description is then used as contextual information, which we include in the form of text at the beginning of the prompt. Subsequently in the second stage, by inputting the ROI, the VLM is prompted to describe the specific region represented by the image patch. We illustrate this approach with a detailed example in the following:

In the context of the entire image, $\langle GlobalCaption \rangle$,
describe the close-up region in detail.

Remarkably, our observations reveal that even with this two-stage approach, the model often struggles to accurately describe the input region. This inaccuracy largely stems from its inability to correctly identify the object classes within the cropped region. Therefore, we further enhance our approach by incorporating human-annotated class names as an additional condition when prompting the VLM to describe the properties of the region:

In the context of the entire image, $\langle GlobalCaption \rangle$,
describe the $\langle ClassName \rangle$ in the close-up region in detail.

GPT-assisted prompt augmentation. To enhance the model’s adaptability to various styles and combinations of user inputs, we augmented the input prompts using ChatGPT-4 [33]. For instance, besides “describe the image in detail”, one may also ask “provide a detailed description of the given image”, or “share a thorough analysis of the image”, etc, in the first stage. To ensure a diverse range of responses, we created ten different versions of input prompts for both stages, as elaborated in the supplementary material. During data generation, one of these ten variations is randomly selected for each stage to promote diversity in the model’s responses.

Methods	PT	IT	Vision	LLM	mAP	Acc (%)
CLIP [37]	-	-	ViT-L	-	58.9	-
RegionCLIP [58]	-	-	R50x4	-	58.3	-
LLaVA [†] [29]	595K	158K	ViT-L	Vicuna-7B	-	40.04
Shikra [†] [9]	600K	5.5M	ViT-L	Vicuna-7B	-	53.91
GPT4RoI [†] [56]	266K	731K	ViT-L	LLaVA-7B	-	64.01
PVIT [†] [6]	13.7M	243K	ViT-L + R50x4	LLaVA-7B	-	64.53
ASM [48]	~22M	~22M	ViT-L	Hasky-7B	69.3	-
Ours	923K	953K	ViT-L	Vicuna-7B	70.0	80.61

Table 3. **Comparison with Region-level based methods on COCO-2017 val set.** Following RegionCLIP [58] and PVIT [6], we report the results of object classification with ground-truth box on COCO val set. [†] represents the results are imported from [6]. - means that the results are not reported in the source paper.

Region caption dataset analysis. Utilizing our automated annotation pipeline, we annotate a corpus of 213K V3Det images [46], leveraging its comprehensive object bounding boxes and class names. This dataset includes about 13,000 precisely labeled concepts, providing a rich foundation for model training. This extensive and precise labeling enhances the reliability of the generated data. To further refine our dataset, we utilize the CLIP model [37] to calculate the similarity between the image regions and the corresponding generated region captions. This process allows us to filter out noisy or irrelevant samples, ensuring that only high-quality data is used for training. As shown in Tab. 2, our dataset is distinguished by having a notably higher average number of words, 87.14 words per caption, in each region’s caption versus other datasets. This detailed richness lays a solid groundwork for an in-depth understanding at the region level.

4. Experiments

In this section, we present experimental settings and results. The experiments are primarily conducted on region classification [27], captioning [22, 23], expression comprehension [22] and object hallucination benchmark [26]. We present both quantitative and qualitative results.

Implementation details. During the entire training process, the visual backbone weights remain unchanged. We train the model with an image resolution of 336×336 during both the pre-training and fine-tuning stages. An input image is padded to achieve a square format, if it is not square. In the pre-training stage, we employ a cosine learning rate scheduler. The maximum learning rate is set at $1e-3$, with a weight decay of 0 and a warmup ratio of 0.03. The model is trained with a batch size of 256 for one epoch. In the fine-tuning stage, the maximum learning rate is reduced to $2e-5$, and the batch size is adjusted to 128. All other hyperparameters remain the same as the pre-training stage.

4.1. Quantitative Evaluation

Region Classification. We first evaluate the object classification ability of our model on COCO-2017 dataset. The

Model	RefCOCOg		Visual Genome	
	METEOR	CIDEr	METEOR	CIDEr
GRIT [49]	15.2	71.6	17.1	142.0
SLR [53]	15.9	66.2	-	-
Kosmos-2 [35]	14.1	62.3	-	-
Ours	16.9	109.9	17.0	145.6

Table 4. **Performance on the region-level captioning task on RefCOCOg and Visual Genome.** We report METEOR and CIDEr metrics, following the image-level caption task.

Method	MDETR[21]	Shikra [9]	Kosmos-2 [35]	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

Table 5. **REC on ReferCOCOg val and test set [22].** As RGPT focuses on region-level understanding rather than localization, hence, we highlight the strength of our model in interpreting complex expressions within the context of the provided regions from [60].

mAP and classification accuracy metrics are reported to quantify performance. Our focus is on region recognition, rather than object localization. Therefore, following RegionCLIP [58], we use ground-truth boxes as the input for positional information. Alongside this, we attach task-guided instruction prompts to the general instruction prompt and input only one bounding box for one-turn conversation. If the output does not fall within the predefined candidate categories of the COCO dataset, we simply discard this prediction and categorize it as a misclassification.

We report the results of VLMs and feature-based vision models, as shown in Tab. 3. For our baseline, we crop the RoI from images, resize them to the input size, and then compare their features with those of the 80 classes in the COCO dataset to select the category with the highest similarity. Additionally, we consider other feature-based methods like RegionCLIP [58] and ASM [48]. RegionCLIP pre-trains CLIP model to align the CC3M [42] region-text pairs in the feature space. ASM is trained on approximately 22M images and the features are produced by the language decoder. The other VLMs use textual formats as output. On the COCO dataset, our approach achieves a mAP of 70.0 and an accuracy of 80.86%, demonstrating our method’s effectiveness in constraining output formats and its strong capability in region-level object recognition.

Region Captioning. We evaluate the region-level captioning ability of our model on the ReferCOCOg [22] and Visual Genome [23], employing the same evaluation metrics as used for image-level captioning: METEOR [3] and CIDEr [45]. As illustrated in Tab. 4, our model surpasses the region-aware VLM, Kosmos-2 [35]. The results highlight our model’s proficiency in accurately generating referring expressions for image regions.

Referring Expression Comprehension (REC). We evalu-

Arch.	Deconv	BiLinear	Deconv + BiLinear	None
AP	66.8	60.9	62.7	57.7
AP _s	51.1	52.8	53.8	42.7
AP _m	71.5	70.8	71.4	65.2
AP _l	78.0	57.9	60.3	65.4

Table 6. **Ablation study on the feature refinement module.** The object classification results on COCO 2017 *val* set are reported. We use ViT-B/16 from [54] as our visual backbone, whose input size is 512×512. Deconv represents our two deconvolution layers design for feature maps of scale 4. BiLinear indicates the use of bilinear upsampling for scale 16. Deconv + BiLinear means bilinear upsampling the Deconv output for scale 16. None refers to no module is used.

Model	AP	AP _s	AP _m	AP _l
OpenAI ViT-L-336	70.0	55.7	75.5	81.5
SigLip ViT-B-512	66.8	51.1	71.5	78.0
SigLip ViT-L-384	69.5	56.8	74.1	80.2
SigLip ViT-SO400M-384	71.0	57.9	76.5	81.6

Table 7. **Ablation study on visual backbone.** The object classification results on COCO 2017 *val* set are reported. We use SigLip models from [54] pre-trained on WebLI dataset [10] and OpenAI CLIP model [37]. The results demonstrate that our method can be further improved with more powerful visual network.

ate expression comprehension of our model on the ReferCOCOg dataset. Our method focuses on region-level understanding, rather than object localization. Therefore, we utilize bounding box proposals from [60] as candidate box sets. If the Intersection Over Union between the ground truth box and any of the candidate boxes is less than 0.5, we include the ground truth box in our set of candidates. The results in Tab. 5 only highlight the specific strength of our model in understanding complex expressions within the context of the provided regions.

Ablation Study on Feature Refinement Module. We study the effect of the feature refinement module on the object classification task. Our motivation for this module is to refine the CLIP visual features for better spatial-aware semantics. Tab. 6 shows that two-deconvolution-layer design significantly outperforms the baseline model (the last column), demonstrating the effectiveness of feature refinement. An interesting observation is that the methods of 16x upsampling (BiLinear and Deconv + BiLinear) enhance the accuracy of classification for smaller objects, though it shows a decrease in performance for larger objects. Our approach achieves a superior trade-off between these two aspects. We believe that implementing more complex and carefully designed feature optimization mechanisms could potentially lead to further improvements in performance.

Ablation Study on Visual Backbone. We study the effect of the visual backbone on the object classification task.

Datasets	Metrics	Ours	Shikra [9]	InstructBLIP [14]	MiniGPT4 [59]	LLaVA [29]	MM-GPT [18]	mPLUG-Owl [52]
Random	Accuracy (↑)	87.80	86.90	88.57	79.67	86.00	50.10	53.97
	Precision (↑)	97.75	94.40	84.09	78.24	87.50	50.05	52.07
	Recall (↑)	78.13	79.26	95.13	82.20	84.00	100.00	99.60
	F1 Score (↑)	86.85	86.19	89.27	80.17	85.71	66.71	68.39
	Yes	41.20	43.26	56.57	52.53	48.00	99.90	95.63
Popular	Accuracy (↑)	87.20	83.97	82.77	69.73	76.67	50.00	50.90
	Precision (↑)	95.44	87.55	76.27	65.86	72.22	50.00	50.46
	Recall (↑)	78.13	79.20	95.13	81.93	86.67	100.00	99.40
	F1 Score (↑)	85.92	83.16	84.66	73.02	78.79	66.67	66.94
	Yes	40.93	45.23	62.37	62.20	60.00	100.00	98.57
Adversarial	Accuracy (↑)	85.67	83.10	72.10	65.17	73.33	50.00	50.67
	Precision (↑)	91.99	85.60	65.13	61.19	69.02	50.00	50.34
	Recall (↑)	78.13	79.60	95.13	82.93	84.67	100.00	99.33
	F1 Score (↑)	84.50	82.49	77.32	76.05	66.32	66.67	66.82
	Yes	42.47	46.50	73.03	67.77	61.33	100.00	98.67

Table 8. **Results on the object hallucination benchmark using the POPE evaluation pipeline [26].** Except for our model and LLaVA [29], the other results are obtained from [9].

The results in Tab. 7 demonstrate that the performance on region-level understanding can be further improved by replacing current visual backbone with a more powerful one.

Object Hallucination. We evaluate object hallucinations, generating objects that are inconsistent with the target images in the descriptions, of our method using the POPE evaluation pipeline [26], with the results detailed in Tab. 8. Our approach significantly outperforms recent popular image-level VLMs. Given that our baseline model is the LLaVA model, we attribute this performance gain to our region-level instruction fine-tuning strategy, which significantly refines the model’s ability to interpret images with greater precision.

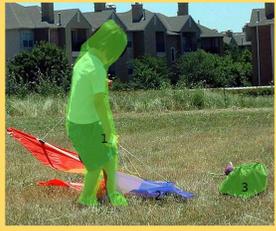
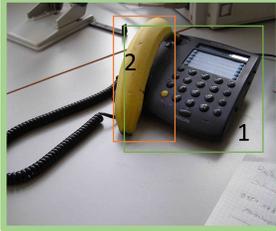
4.2. Qualitative Evaluation

As demonstrated in Fig. 4, RGPT is capable of analyzing of relationships between multiple regions within an image. Additionally, in the green example, our model exhibits complex region-level reasoning abilities akin to those seen in GPT-4V demonstrations, effectively interpreting and explaining visual content.

5. Conclusion

In this paper, we present RGPT, a general vision-language model that tackles complex region-level captioning and reasoning following user instruction. Our model employs region-level instruction tuning to align the visual feature with the language word embedding space. Besides, we carefully design task-guided instruction prompts to seamlessly blend vision tasks within GPT framework, by converting the vision tasks to VQA tasks and prompting the response format. Finally, we propose a two-stage GPT-assisted annotation pipeline to reformat the object detection dataset and create detailed region-level captions. The

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.

Figure 4. Qualitative evaluation of the multi-turn conversation of RGPT. Our model preserves the multi-turn conversation and image-level captioning ability.

results demonstrate that RGPT achieves impressive performance on the region-level understanding tasks.

Acknowledgement. This paper is partially supported by the National Key R&D Program of China No.2022ZD0161 000 and the General Research Fund of Hong Kong No.172 00622.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 2
- [2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 1
- [3] Satyanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 7
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2
- [6] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023. 6
- [7] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023. 2
- [8] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2, 7
- [9] 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. 1, 2, 6, 7, 8
- [10] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 7
- [11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2, 4
- [12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. 2
- [13] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2
- [14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 8
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2
- [16] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 2
- [17] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2
- [18] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 2, 8
- [19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2
- [20] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui,

- Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 2
- [21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 7
- [22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 4, 6, 7
- [23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2, 4, 6, 7
- [24] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2
- [26] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6, 8
- [27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6
- [28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 2, 4
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 4, 5, 6, 8
- [30] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 4
- [31] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmaia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023. 2
- [32] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhui Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023. 2
- [33] OpenAI. Gpt-4 technical report, 2023. 6
- [34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 2
- [35] 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, 7
- [36] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6, 7
- [38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2
- [39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2
- [40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2
- [41] John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, et al. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2022. 2
- [42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 7
- [43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

- [45] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. [7](#)
- [46] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. *arXiv preprint arXiv:2304.03752*, 2023. [4](#), [6](#)
- [47] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. [2](#)
- [48] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. [3](#), [6](#), [7](#)
- [49] 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. [7](#)
- [50] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. [2](#)
- [51] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. [2](#)
- [52] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. [8](#)
- [53] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7282–7290, 2017. [7](#)
- [54] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. [7](#)
- [55] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. [2](#)
- [56] 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. [2](#), [3](#), [6](#)
- [57] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. [2](#)
- [58] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. [6](#), [7](#)
- [59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [1](#), [2](#), [8](#)
- [60] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023. [7](#)