

VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks

Wenhai Wang^{*1}, Zhe Chen^{*2,1}, Xiaokang Chen^{*3,1}, Jiannan Wu^{*4,1}, Xizhou Zhu^{5,1}

Gang Zeng³, Ping Luo^{4,1}, Tong Lu², Jie Zhou⁶, Yu Qiao¹, Jifeng Dai^{†6,1}

¹OpenGVLab, Shanghai AI Laboratory ²Nanjing University ³Peking University

⁴The University of HongKong ⁵SenseTime Research ⁶Tsinghua University

Code: <https://github.com/OpenGVLab/VisionLLM>

Demo: <https://github.com/OpenGVLab/InternGPT>

Abstract

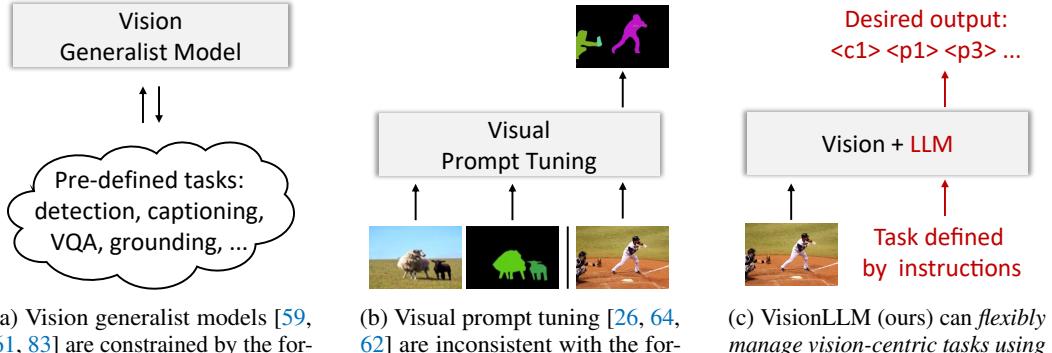
Large language models (LLMs) have notably accelerated progress towards artificial general intelligence (AGI), with their impressive zero-shot capacity for user-tailored tasks, endowing them with immense potential across a range of applications. However, in the field of computer vision, despite the availability of numerous powerful vision foundation models (VFs), they are still restricted to tasks in a pre-defined form, struggling to match the open-ended task capabilities of LLMs. In this work, we present an LLM-based framework for vision-centric tasks, termed VisionLLM. This framework provides a unified perspective for vision and language tasks by treating images as a foreign language and aligning vision-centric tasks with language tasks that can be flexibly defined and managed using language instructions. An LLM-based decoder can then make appropriate predictions based on these instructions for open-ended tasks. Extensive experiments show that the proposed VisionLLM can achieve different levels of task customization through language instructions, from fine-grained object-level to coarse-grained task-level customization, all with good results. It's noteworthy that, with a generalist LLM-based framework, our model can achieve over 60% mAP on COCO, on par with detection-specific models. We hope this model can set a new baseline for generalist vision and language models. The code and demo shall be released.

1 Introduction

The emergence of large language models (LLMs) like ChatGPT [41] has revolutionized the landscape of artificial general intelligence (AGI), showcasing their impressive zero-shot capabilities in addressing various natural language processing (NLP) tasks through user-tailored prompts or language instructions. Despite these advancements, it's essential to note that the triumph of LLMs does not effortlessly extend to pure vision and vision-language tasks, due to the inherent disparities between modalities and task formats.

The field of computer vision presents a unique set of challenges and paradigms that differ from those of NLP. The traditional paradigm of vision foundation models is pre-training followed by fine-tuning [59, 12, 51, 61, 18, 52], which is effective but comes with significant marginal costs when adapting to diverse downstream scenarios. As shown in Figure 1a, while approaches such as multi-task unification [44, 58, 1, 57, 81] have been used to achieve generalist capability, they often struggle to overcome the limitations imposed by pre-defined tasks, resulting in a gap in open-ended

^{*}Equal contribution. This work is done when Zhe Chen, Xiaokang Chen, and Jiannan Wu are interns at Shanghai AI Laboratory. [†] Corresponding to Jifeng Dai <daijifeng@tsinghua.edu.cn>.



(a) Vision generalist models [59,

61, 83] are constrained by the for-

mat of pre-defined tasks.

(b) Visual prompt tuning [26, 64,

62] are inconsistent with the for-

mat of LLMs.

(c) VisionLLM (ours) can *flexibly*

manage vision-centric tasks using

language instructions like LLMs.

Figure 1: **Comparison of our VisionLLM with popular paradigms.** Unlike current vision generalist models that depend on pre-defined task formats and visual prompt tuning models that are inconsistent with large language models (LLMs), VisionLLM leverages the power of LLMs for open-ended vision tasks by using language instructions.

task capabilities compared to LLMs. Recently, visual prompt tuning [26, 74, 79, 76, 62] has emerged as a way to flexibly outline some pure vision tasks (see Figure 1b), such as object detection, instance segmentation, and pose estimation, using visual masking. However, the format of visual prompts considerably deviates from that of language instructions, making it challenging to directly apply the reasoning abilities and world knowledge of LLMs to vision tasks. Therefore, *there is an urgent need for a unified generalist framework that can seamlessly integrate the strengths of LLMs with the specific requirements of vision-centric tasks*.

In this work, we present VisionLLM, a novel framework that aligns the definitions of vision-centric tasks with the methodologies of LLMs. Leveraging the reasoning and parsing capacities of LLMs, VisionLLM is designed to empower open-ended task capabilities for vision-centric tasks. Specifically, it comprises three core components: (1) a unified language instruction designed for vision and vision-language tasks, (2) a language-guided image tokenizer, and (3) an LLM-based open-ended task decoder that orchestrates various tasks using language instructions. With this framework, a wide range of vision-centric tasks can be seamlessly integrated, including object detection, instance segmentation, image captioning, and visual grounding. In addition, the framework also facilitates task customization at different levels of granularity, allowing for the customization of target objects, output formats, task descriptions, etc.

Compared to current popular API-based applications [68, 73, 50, 35, 30], our model takes a unified, end-to-end approach to integrate VFM and LLMs, streamlining and enhancing the overall efficiency of the overall process, and leveraging the strengths and data of both VFM and LLMs within a single, cohesive system. Furthermore, our model surpasses the limitations of generalist vision models pre-trained on pre-defined tasks. VisionLLM can effectively manage vision-centric tasks through language instructions, embodying a flexible and open-ended approach that is not constrained by pre-set tasks. This versatility makes VisionLLM a robust and powerful generalist model for vision and vision-language tasks, opening up new possibilities for the development of unified generalist models that bridge the domains of vision and language.

In summary, our main contributions are as follows:

(1) We propose VisionLLM, the first framework that leverages the power of LLMs to address vision-centric tasks in an open-ended and customizable manner. By aligning the definitions of vision-centric tasks with LLM methodologies, VisionLLM breaks new ground in enabling the unified modeling of vision and language, opening up possibilities for advancing the field.

(2) We overcome many difficulties when porting LLMs to vision-centric tasks, by designing unified language instruction that matches the format of language models and covers various vision-centric tasks including visual perception. Correspondingly, we develop a language-guided image tokenizer and an LLM-based task decoder that can handle open-ended tasks according to the given language instructions based on the LLMs' reasoning and parsing capabilities.

Human: "Identify the objects in <image> that belong to {'What is the child eating?': <c0>, 'red gamepad': <c1>} and draw a bounding box around each one. The output should be a list of tuples in

the format [(c, x1, y1, x2, y2), where c is the class label and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box. The coordinates should be within the range 1024]."

VisionLLM: "The bounding boxes are [(<c0>, 226.4, 229.8, 363.1, 347.4), (<c1>, 441.1, 183.5, 538.6, 269.9)]."



Human: "For each object in <image> that is a member of {'motorbike': <c0>}, output a tuple with the class label and the coordinates of a polygon with 16 points that

encloses the object. The coordinates should be within the range 1024. The output format should be [c, x1, y1, x2, y2, ..., x16, y16]."



VisionLLM: "The detected polygons are [(<c0>, 135.3, 95.7, 123.4, 53.4, 84.9, 57.6, 66.8, 60.5, 60.1, 72.3, 34.2, 71.4, ..., 124.9, 119.3)]."

(a) Object Detection with *Customized Class Set* (e.g., question, reasoning text, open-vocabulary description)

Human: "Provide a long description of the <image>."



VisionLLM: "The image features a white dog laying down on a foldable outdoor cooling pad, which is placed on top of a wooden table. The dog has its head in the middle of the cooling pad, and a fan is positioned behind it, blowing air onto the pad. The table is outdoors, in a grassy area, and there are two chairs placed around it, one on each side. The scene appears to be set in a sunny weather, with a bright and sunny sky, which contributes to the overall outdoor atmosphere."

(c) Image Description with *Controllable Text Length*

(b) Object Location with *Editable Output Format*

Human: "With the objects in the <image>, if you feel hot, what will you do?"

VisionLLM: "First, You can also try turning off the computer and other electrical appliances in the



room to reduce the heat generated by their operation. As an alternative, you can use the desk fan provided in the image to create a cooling effect in the room. You can plug in the fan and adjust its settings to ensure it blows air in a direction that reaches the room's occupants effectively."

(d) Visual Question Answer with *Complex Reasoning*

Figure 2: **Results and visualizations of our VisionLLM.** Guided by language instructions, our unified generalist framework showcases its effectiveness on diverse open-ended vision-centric tasks. The text marked with a gray background indicates the customized instructions and the desired outputs.

(3) We construct a series of tasks with different granularities to verify the effectiveness of our models, ranging from easy to hard, and from pre-defined to flexible. Through these validations, we demonstrate the remarkable generality of our models, showcasing their ability to handle diverse scenarios, including random object categories, random output formats, and random task descriptions, as shown in Figure 2. The successful outcomes of these validations underscore the tremendous potential of our model in harnessing the capabilities of LLMs to control and guide vision-centric tasks. In addition, with a generalist LLM-based framework, our model also yields promising results on various vision-centric tasks. Notably, our generalist model achieves an impressive mAP score of 60+% on the COCO dataset, surpassing many detection-specific models [82, 7, 22] and approaching the state-of-the-art record.

2 Related Work

2.1 Large Language Model

Large language models (LLMs) have gained significant attention in the field of natural language processing (NLP) and artificial general intelligence (AGI), due to their impressive capabilities in language generation, in-context learning, world knowledge, and reasoning. The GPT family, including GPT-3 [6], ChatGPT [41], GPT-4 [40], and InstructGPT [42] are most representative works of LLMs. Other LLMs like OPT [78], LLaMA [54], MOSS [15], and GLM [77] have also made substantial contributions to the field. These models achieve high performance and are open-sourced, serving as valuable resources for training large models and as foundations for further fine-tuning for specific purposes. For instance, Alpaca [53] introduces a self-instruct framework that facilitates instruction tuning of the LLaMA model, reducing the reliance on human-written

instruction data. Recently, the emergence of these LLMs has also opened up API-based applications for solving vision-centric tasks. These applications have integrated visual APIs with language models to enable decision-making or planning based on visual information, such as Visual ChatGPT [68], MM-REACT [73], HuggingGPT [50], InternGPT [35], and VideoChat [30]. However, despite the convenience of using language-based instructions to define tasks and describe visual elements, these interactive systems [68, 73, 50, 35, 30] still face limitations in capturing fine-grained visual details and understanding complex visual contexts, which hinder their ability to effectively connecting vision and language models. In summary, while LLMs have shown tremendous potential in various NLP applications, their applicability to vision-centric tasks has been limited by the challenges posed by modalities and task formats.

2.2 Vision Generalist Model

The pursuit of generalist models [83, 38, 70], which aim to handle a wide range of tasks using a shared architecture and parameters, has been a long-standing goal in the machine learning community. Inspired by the success of sequence-to-sequence (seq2seq) models in the field of NLP [44], recent advancements such as OFA [58], Flamingo [1], and GIT [57] propose modeling diverse tasks as sequence generation tasks. Unified-IO [38], Pix2Seq v2 [9], and UniTab [71] extend this idea by using discrete coordinate tokens to encode and decode spatial information for more tasks. Gato [47] also incorporates reinforcement learning tasks into the seq2seq framework, while GPV [21] develops a general-purpose vision system by combining a seq2seq module with a DETR-based visual encoder [7]. However, these methods suffer from some limitations, such as slow inference speed and performance degradation due to the non-parallel auto-regressive decoding process. Uni-Perceivers [83, 81, 28] solve these issues by unifying different tasks using the maximum likelihood target for each input based on representation similarity, regardless of their modality, making it possible to support both generation and non-generation tasks in a unified framework. Nevertheless, these generalist models are still restricted by pre-defined tasks and cannot support flexible open-ended task customization based on language instructions like LLMs.

2.3 Instruction Tuning

Language instructions are a powerful way to express various NLP tasks and examples for LLMs, as introduced by GPT-3 [6]. Following this idea, subsequent works, such as InstructGPT [42], FLAN [14, 67], and OPT-IML [25], explore the instruction-tuning method [66, 65] and demonstrate that this simple approach effectively enhances the zero-shot and few-shot capabilities of LLMs. The language instruction paradigm has also been adopted by the computer vision community to define image-to-text tasks. Flamingo [1] is a milestone work that uses vision and language inputs as prompts and achieves remarkable few-shot results in various vision-language tasks, such as image captioning [10] and VQA [3]. BLIP-2 [29] further connects the visual encoder with LLMs through a querying transformer and a linear projection layer to build strong multimodal models. MiniGPT-4 [80] and LLaVA [33] finetune the BLIP-2-style models on synthetic multimodal instruction-following data to unleash the potential of LLMs. However, these models mainly focus on image-to-text tasks and fail to address visual perception, such as object detection, instance segmentation, pose estimation, etc. To tackle image inpainting tasks, Bar *et al.* [4] introduces the first visual prompting framework that utilizes inpainting with discrete tokens on images. Painter [63] and SegGPT [64] employ masked image modeling on raw pixels for in-context learning with paired images. While these visual prompt models demonstrate good results in segmentation tasks, their applicability to numerous real-world vision tasks is challenging. Moreover, defining the visual prompts as image inpainting is inconsistent with the language instructions in LLMs, hard to leverage the reasoning, parsing ability, and world knowledge of LLMs. In this work, we aim to align vision-centric tasks with language tasks, use language instructions to unifiedly and flexibly define all tasks, and solve them with a shared LLM-based task decoder.

3 VisionLLM

3.1 Overall Architecture

This work targets to provide a unified generalist framework that can seamlessly integrate the strengths of large language models (LLMs) with the specific requirements of vision-centric tasks. As shown in

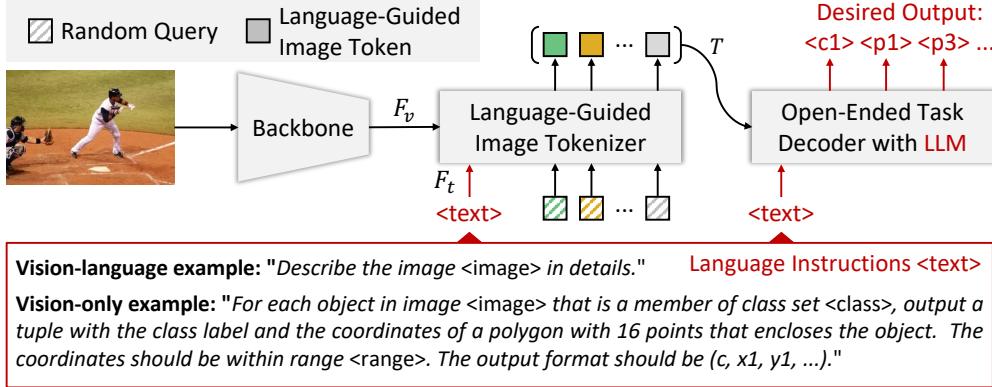


Figure 3: **Overall architecture of the proposed VisionLLM.** It consists of three parts: a unified language instruction designed to accommodate both vision and vision-language tasks, an image tokenizer that encodes visual information guided by language instructions, and an LLM-based open-ended task decoder that executes diverse tasks defined by language instructions.

Figure 3, the overall architecture of VisionLLM consists of three key designs: (1) a unified language instruction that provides a consistent interface for vision-centric task definition and customization; (2) a language-guided image tokenizer, which encodes visual information in alignment with the given language prompt, enabling the model to comprehend and parse the visual content effectively; and (3) an LLM-based open-task decoder, which utilizes the encoded visual information and language instructions to generate satisfactory predictions or outputs. The three designs work together to achieve a flexible and open-ended framework that can handle various vision-centric tasks at different levels of task customization through language instructions.

Different from previous interactive systems [68, 73, 50, 35, 30] that rely on APIs, our VisionLLM presents a more flexible and end-to-end pipeline. Given language instructions that describe the current tasks and an input image, the model first uses a language-guided image tokenizer to encode the image tokens based on the given prompt. Then, the image tokens and language instructions are fed to an LLM-based open-ended task decoder. Finally, it evaluates the generated outputs against the task definition given by the unified language instructions, enabling the model to produce task-specific results. This seamless, end-to-end pipeline enables VisionLLM to effectively combine vision and language, achieving remarkable performance in open-ended and customizable vision-centric tasks.

3.2 Unified Language Instruction

We first introduce unified language instructions to describe vision-centric tasks. This design enables the unification of various vision-only and vision-language task descriptions and allows for flexible task customization.

Vision-Language Tasks. The instructions for vision-language tasks such as image captioning and visual question answering (VQA) are straightforward and similar to NLP tasks. Following previous methods [29, 83, 33], we describe the image captioning task like "*The image is <image>. Please generate a caption for the image:*", and the VQA task like "*The image is <image>. Please generate an answer for the image according to the question: <question>*". Here, <image> and <question> are the placeholders of the image tokens and the question, respectively.

Vision-Only Tasks. Designing effective language instructions for vision tasks is a challenging endeavor due to the differences in modality and task format between vision and language. Here, we describe vision tasks by providing a task description and specifying the desired output format via language instructions.

- (1) The task description conveys the intended task to the language model. Following self-instruct [65], we design a set of seed instructions with placeholders and employ LLMs to generate a large number of related task descriptions and randomly select one of them during training.

(2) For conventional visual perception tasks like object detection and instance segmentation, we propose a unified output format represented as a tuple (C, P) , where C denotes the class index in the category set $\langle \text{class} \rangle$, and $P = \{x_i, y_i\}_{i=1}^N$ represents N points that locate the object. To align with the format of word tokens, both the class index C and the coordinates of points x_i, y_i are transformed into discretized tokens. Specifically, the class index is an integer starting from 0, and the continuous coordinates of the points are uniformly discretized into an integer within the range $[-\langle \text{range} \rangle, \langle \text{range} \rangle]$. For object detection and visual grounding tasks, the point number N is equal to 2, representing the top-left and bottom-right points of object’s bounding box. In the case of instance segmentation, we employ multiple ($N > 8$) points along the object boundary to represent an instance mask [69]. Other perception tasks such as pose estimation (keypoint detection) can also be formulated as language instructions in this way.

An example of language instruction for the instance segmentation task is as follows: “*Segment all the objects of category set $\langle \text{class} \rangle$ within the $\langle \text{range} \rangle$ of the image and generate a list of the format $(c, x1, y1, x2, y2, ..., x8, y8)$. Here, c represents the index of the class label starting from 0, and $(x1, y1, x2, y2, ..., x8, y8)$ correspond to the offsets of boundary points of the object relative to the center point. The image is: $\langle \text{image} \rangle$.*”

3.3 Language-Guided Image Tokenizer

VisionLLM considers images as a kind of foreign language and converts them into token representations. Unlike previous works [17, 60, 34] that utilize fixed-size patch embeddings to represent images, we introduce the language-guided image tokenizer to flexibly encode visual information that aligns with task-specific language prompts or instructions.

Specifically, give an image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$ with height H and width W , we first feed it to the image backbones (e.g., ResNet [23]) and extract visual features F_v of four different scales. Additionally, we leverage a text encoder (e.g., BERT [16]) to extract the language features F_l from given prompts. The language features are then injected into each scale of visual features through cross-attention [55], yielding multi-scale language-aware visual features, enabling the alignment of features across modalities.

Afterward, we propose to adopt a transformer-based network (e.g., Deformable DETR [82]) with M random-initialized queries $Q = \{q_i\}_{i=1}^M$ to capture the high-level information of images. We build the transformer-based network on top of the multi-scale language-aware visual features to extract M image tokens $T = \{(e_i, l_i)\}_{i=1}^M$, each of which is represented by an embedding e_i and a location l_i , denoting the semantic and positional information of the token. This design not only represents the images independent of input resolution but also extracts the visual representation that is informative with respect to the language prompts.

3.4 LLM-based Open-Ended Task Decoder

We build our decoder on Alpaca [53], an LLM that is adapted from LLaMA [54], to handle various vision-related tasks with language guidance. However, Alpaca has some inherent drawbacks for vision-centric tasks, such as (1) It only has a few digit tokens (e.g., 0~9) in its vocabulary, which restricts its ability to locate objects by numbers; (2) It uses multiple tokens to represent the category name, resulting in an inefficient scheme in object classification; and (3) It is a causal model that is inefficient for visual perception tasks.

To tackle these issues, we expand the vocabulary of LLM with additional tokens specially designed for vision-centric tasks. First, we add a set of location tokens, denoted as $\{\langle p-512 \rangle, \dots, \langle p0 \rangle, \dots, \langle p512 \rangle\}$, where $\langle p i \rangle$ represents the discretized offset of $i \in [-512, 512]$ to the location l_i of the image token, and the relative value to image height or width is equal to $i/512$. These tokens successfully transform the object localization task from continuous variable prediction to more unified discrete bin classification. Second, we introduce semantics-agnostic classification tokens $\{\langle c0 \rangle, \langle c1 \rangle, \dots, \langle c511 \rangle\}$ to replace category name tokens, which overcomes the inefficiency of using multiple tokens to represent categories. The mapping between category names and the classification tokens is flexibly provided in the category set $\langle \text{class} \rangle$ of language instructions, such as $\{"\text{person}": \langle c0 \rangle, "\text{car}": \langle c1 \rangle, "\text{black cat}": \langle c2 \rangle, \dots\}$. This design allows our model to select the appropriate category name from the provided category set, facilitating efficient and accurate object classification.

Moreover, to address the inefficiency caused by the causal framework, we introduce output-format-as-query decoding. We first use LLMs to parse the structural output format from the task instructions (*e.g.*, “<cls> <x1> <y1> <x2> <y2>” for object detection, “<bos>” for image captioning), and then feed the tokens of structural output format as queries to the decoder to generate the desired output according to the queries. This simple method enables our model to not only avoid inefficient token-by-token decoding in visual perception tasks, but also keep a unified framework for vision-language tasks.

In this way, the output of object location and classification is formulated as a foreign language, thus unifying these vision-centric tasks into the format of token classification. Therefore, both vision-language and vision-only tasks can be supervised with the cross entropy loss like language tasks. In addition, for efficient training, we adopt the Low-Rank Adaptation (LoRA) approach [24], which allows us to train and fine-tune the models without excessive computational costs. It also acts as a bridge between the language and visual tokens, facilitating effective alignment between the two modalities, ensuring better task customization, and improving the convergence of the overall system.

4 Experiment

4.1 Experimental Settings

Datasets. VisionLLM unifies the output formats of vision and language tasks as vocabulary generation, which enables models to be jointly trained on a wide range of tasks. In the experiments, we investigate the general modeling capacities of VisionLLM on five vision-centric tasks, including object detection, instance segmentation, visual grounding, image captioning, and visual question answering. For object detection and instance segmentation, COCO2017 [32] is used for training and evaluation. For visual grounding, we combine the annotations of RefCOCO [75], RefCOCO+ [75] and RefCOCOg [39] for training, resulting in over 120k referred objects in total. And our models are evaluated on the validation set of RefCOCO. For image captioning and visual question answering, we adopt COCO Caption [10] and LLaVA-Instruct-150K [33] as the training source. We evaluate the image captioning performance on the COCO Karpathy test split following common practice [28, 58, 71]. We mainly use qualitative results (see Figure 2d) to demonstrate the VQA capability of our model, as LLaVA-Instruct-150K is not compatible with the standard VQA benchmark. These tasks differ in their granularity, ranging from coarse-grained image level to fine-grained pixel level, enabling a comprehensive evaluation of the model’s ability to adapt to different levels of customization through language instructions.

Implementation Details. We implement two variants of VisionLLM with two image backbones, *i.e.*, ResNet [23] and InternImage-H [59]. For the language-guided image tokenizer, we adopt BERT-Base [5] as the text encoder and Deformable DETR (D-DETR) [82] to capture high-level information. We set the number of queries M to 100, and the number of encoder/decoder layers to 6 for D-DETR. For the LLM, we employ Alpaca-7B [53], a LLaMA [54] model fine-tuned with instructions, and equip it with LoRA [24] for parameter-efficient fine-tuning.

The model is trained in two stages. In the first stage, we initialize the model with the pre-trained weights of D-DETR, BERT, and Alpaca-7B, and train the visual backbone and the language-guided image tokenizer, while freezing most parameters of the LLM except a few LoRA parameters. To simplify the training complexity, in this stage, we mainly focus on object detection tasks with random object categories and task descriptions. In the second stage, we freeze the visual backbone and introduce the unified supervision of multiple tasks. Unless otherwise specified, the training runs for 50 epochs on 4×8 NVIDIA A100 GPUs. AdamW [36] is used as the optimizer, with one sample per GPU. We employ the cosine annealing schedule [37] as the learning policy, with an initial learning

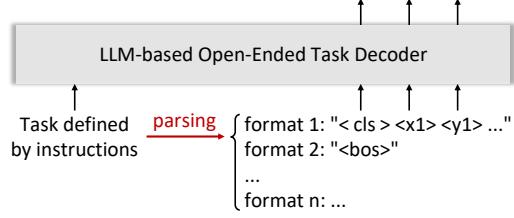


Figure 4: Illustration of the “output-format-as-query” decoding process. “<cls> <x1> <y1> ...” denote the queries of the object’s class index and boundary points, and “<bos>” denotes the beginning of string.

rate of 0.001 and a weight decay of 0.01.

4.2 Model Performance

4.2.1 Object Detection and Instance Segmentation

4.2.2 Visual Grounding

4.2.3 Image Captioning

4.2.4 Visual Question Answering

4.2.5 Qualitative Results

4.3 Ablation Studies

4.4 Comparison with State-of-the-Art

4.5 Conclusion

4.6 Limitations and Future Work

4.7 Acknowledgments

4.8 References

4.9 Appendix

4.10 Code Availability

4.11 Author Contributions

4.12 Conflict of Interest

Table 1: **Results on standard vision-centric tasks.** ‘Intern-H’ denotes InternImage-H [59]. “sep” indicates that the model is separately trained on each task.

Method	Backbone	Open-Ended	Detection			Instance Seg.			Grounding		Captioning	
			AP	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅	P@0.5	BLEU-4	CIDEr	
Specialist Models												
Faster R-CNN-FPN [48]	ResNet-50	-	40.3	61.0	44.0	-	-	-	-	-	-	-
DETR-DC5 [7]	ResNet-50	-	43.3	63.1	45.9	-	-	-	-	-	-	-
Deformable-DETR [82]	ResNet-50	-	45.7	65.0	49.1	-	-	-	-	-	-	-
Mask R-CNN [22]	ResNet-50	-	41.0	61.7	44.9	37.1	58.4	40.1	-	-	-	-
Polar Mask [69]	ResNet-50	-	-	-	-	30.5	52.0	31.1	-	-	-	-
Pix2Seq [8]	ResNet-50	-	43.2	61.0	46.1	-	-	-	-	-	-	-
UNITER [11]	ResNet-101	-	-	-	-	-	-	-	81.4	-	-	-
VILLA [19]	ResNet-101	-	-	-	-	-	-	-	82.4	-	-	-
MDETR [27]	ResNet-101	-	-	-	-	-	-	-	86.8	-	-	-
VL-T5 [13]	T5-B	-	-	-	-	-	-	-	-	-	116.5	-
Generalist Models												
UniTab [72]	ResNet-101	-	-	-	-	-	-	-	88.6	-	115.8	-
Uni-Perceiver [83]	ViT-B	-	-	-	-	-	-	-	-	32.0	-	-
Uni-Perceiver-MoE [81]	ViT-B	-	-	-	-	-	-	-	-	33.2	-	-
Uni-Perceiver-V2 [28]	ViT-B	-	58.6	-	-	50.6	-	-	-	35.4	116.9	-
Pix2Seq v2 [9]	ViT-B	-	46.5	-	-	38.2	-	-	-	34.9	-	-
VisionLLM-R50 _{sep}	ResNet-50	-	44.8	64.1	48.5	25.2	50.6	22.4	84.4	30.8	112.4	-
VisionLLM-R50	ResNet-50	✓	44.6	64.0	48.1	25.1	50.0	22.4	80.6	31.0	112.5	-
VisionLLM-H	Intern-H	✓	60.2	79.3	65.8	30.6	61.2	27.6	86.7	32.1	114.2	-

rate of 2×10^{-4} . In addition to the experiments in the main paper, more experimental settings and ablation studies are provided in the supplementary material due to space limitations.

4.2 Task-Level Customization

We first evaluate the task-level customization capability of VisionLLM. VisionLLM supports coarse-grained task customization, including visual perception tasks and visual-language tasks. Table 1 presents the evaluation results on four standard vision-centric tasks, including object detection, instance segmentation, visual grounding, and image captioning. We compare our model with task-specific methods as well as recently-proposed vision generalist models. Note that, unless specifically mentioned, *the results of our model come from a shared-parameter generalist model and switch different tasks by changing the language instructions only. Detailed instructions could be found in the supplementary material.*

Object Detection. Object detection is a fundamental computer vision task that involves identifying and localizing objects of interest within an image. Our method achieves comparable or higher results to others, 44.6 mAP, with a ResNet-50 [23] backbone. With the same backbone *i.e.* ResNet-50, our method outperforms Pix2Seq [8] by 1.4 mAP, which also discretizes the output coordinates to integers. Furthermore, benefiting from the output-format-as-query framework (see Sec. 3.4), we can decode multiple predictions in parallel during inference, making our approach more efficient. Using InternImage-H [59] as the visual backbone, we obtained 60.2% mAP, which is close to the current state-of-the-art detection-specific model [59], demonstrating the scalability of our generalist model.

Visual Grounding. Visual grounding associates textual descriptions with corresponding regions or objects within an image. Training visual grounding and object detection can potentially conflict with each other, as object detection aims to detect all the objects, while visual grounding should only localize the referred object and suppress other objects. Benefiting from our unified task instructions and the strong instruction comprehension capabilities of LLMs, our model performs both tasks effectively and achieves a result of 80.6 P@0.5 for visual grounding. With InternImage-H as the backbone, we achieve 86.7 P@0.5 on the validation set of RefCOCO.

Instance Segmentation. Instance segmentation involves identifying and segmenting individual objects within an image. We employ a flexible number of points (*i.e.*, 8~24) along the object boundary to represent an instance mask. Compared to mainstream models specific to instance segmentation, our model has a comparable mask AP₅₀ (61.2% with InternImage-H [59]) but relatively low mask AP₇₅. This gap could potentially arise from factors as follows: (1) We discretize the output

Table 2: **Experiments of object-level and output format customization.** We conduct these experiments based on VisionLLM-R50, and report the performance of box AP and mask AP on COCO minival for (a) and (b), respectively. “#Classes” and “#Points” indicate the number of classes and boundary points, respectively. “**” indicates that we report the mean AP of the given classes, *e.g.*, 10 classes.

(a) Object-level customization.							(b) Output format customization.						
#Classes	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	#Points	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
10*	48.9	72.6	51.2	31.7	47.5	67.3	8	18.5	45.7	11.6	9.9	19.7	28.7
20*	52.7	73.6	56.8	31.8	53.2	70.5	14	22.9	48.3	19.4	11.0	25.1	36.0
40*	49.3	70.7	53.2	33.1	53.6	63.8	16	24.2	49.9	20.9	11.5	26.3	36.8
80*	44.6	64.0	48.1	26.7	47.9	60.5	24	25.1	50.0	22.4	12.5	27.4	38.2

Table 3: **Ablation studies on language-guided image tokenizer and hyper-parameters.**

(a) Effect of text encoder in the language-guided image tokenizer.	(b) Effect of image tokenization method.	(c) Effect of the number of bins (#Bins).
w/ BERT	Freeze	COCO RefCOCO
-	-	44.7 48.1
✓	-	44.8 84.1
✓	✓	1.3 34.3
	Tokenization	AP
	Average Pooling	23.1
	Ours	44.8
	#Bins	AP
	257	34.9
	513	40.8
	1025	44.8
	2049	44.8

coordinates to integers for unifying tasks, which introduces information loss; (2) Due to the memory and computational constraint, the number of points in our model is limited, which also results in a performance drop; and (3) Point-based methods typically yield lower results compared to direct mask prediction methods, such as Mask R-CNN [22].

Image Captioning. We also evaluate our model in a representative vision-language task, *i.e.* image captioning task, and report the BLEU-4 [43] and CIDEr [56] metrics. Note that we do not adopt the beam search [2] or CIDEr optimization [49]. We can observe that VisionLLM achieves competitive performance to previous methods. With ResNet-50, we obtain a BLEU-4 score of 31.0 and a CIDEr score of 112.5. When using InternImage-H as the backbone, our model achieves a comparable BLEU-4 score of 32.1 and a CIDEr score of 114.2. These results demonstrate the effectiveness of VisionLLM in generating descriptive and contextually relevant captions for images.

4.3 Object-Level & Output Format Customization

Our VisionLLM not only allows for customizing the task description, but also for adjusting the target object and the output format using language instructions. Here, we evaluate our model’s fine-grained customization ability on COCO. In particular, to customize the target object, we modify the `<class>` in language instructions to change the model’s recognition target from 10 classes to 80 classes. Likewise, to customize the output format, we modify the number of points in language instructions to change the task output format. Table 2 shows that our method can perform well for both object-level and output format changes.

4.4 Ablation Study

In this section, we analyze the effect of key components and hyper-parameters on VisionLLM. Unless otherwise specified, we use ResNet-50 [23] backbone and perform the ablation experiments for object detection tasks with random classes and task descriptions on COCO2017 [32].

Single Task vs. Multiple Tasks. We perform an ablation study to assess the impact of multi-task learning with language instructions on VisionLLM. As shown in Table 1, the single-task trained model VisionLLM-R50_{sep} is slightly better than the jointly trained model VisionLLM-R50 except image captioning. This is due to the multitasking conflicts that also affect previous generalist models [83, 81], and it reflects a trade-off between accuracy and generalization.

Text Encoder in Language-Guided Image Tokenizer. We examine the role of text encoder (*i.e.*, BERT) in our language-guided image tokenizer in Table 3a, where we report the results for object

detection and visual grounding. The first two rows show that BERT is not essential for object detection but it is crucial for visual grounding. We also investigate the effect of freezing the text encoder during training. The last row indicates that freezing BERT hinders the alignment of vision and language modalities and thus degrades the performance for both tasks.

Image Tokenization Method. As a comparison to our query-based tokenization, we employ average pooling on the feature maps from the D-DETR encoder to obtain M patch embeddings, which serve as token representations for the image. Results in Table 3b indicate a clear advantage of our method. This is due to its ability to capture information from objects of various sizes in a more flexible way.

Number of Localization Tokens. We vary the number of localization tokens from 257 (*i.e.*, -128~128) to 2049 (*i.e.*, -1024~1024), to investigate its impact on visual perception performance. As presented in Table 3c, the model consistently exhibits improvement as the number of localization tokens increases until it reaches a saturation point. Remarkably, a substantial performance boost is observed when the number is raised from 257 to 1025 (+9.9 AP). These results indicate that a higher number of localization tokens enables the models to achieve finer localization abilities, thereby improving localization accuracy.

5 Conclusion

In this paper, we have presented VisionLLM, a novel framework that leverages the power of large language models (LLMs) to address vision-centric tasks in an open-ended and customizable manner. We have designed unified language instruction that matches the format of language models and covers various vision-centric tasks including visual perception. We have also developed a language-guided image tokenizer and an LLM-based task decoder that can handle open-ended tasks according to the given language instructions. We have verified the effectiveness of our models on a series of tasks with different granularities, demonstrating their remarkable generality and flexibility.

Broader Impact. We envision that this work will promote the fusion of visual and language tasks. In addition, since our work is built on open-source pre-trained vision foundation models and large language models, requiring low training resources, thus reducing the carbon footprint. We do not foresee obvious undesirable ethical/social impacts at this moment.

Appendix

A Example Instructions

As described in Sec. 3.2 of the main paper, we follow self-instruct [65] to design a set of seed instructions with placeholders and employ LLMs to create diverse related task descriptions for coarse-grained task-level customization. Here, we show some examples of instructions for task-level customization, including object detection, instance segmentation, visual grounding, image captioning, and visual question answering (VQA). *Following various instructions, our model can elegantly switch among different vision-centric tasks and accomplish them in a unified manner like LLMs.*

A.1 Object Detection

Example 1. “Please examine the image and identify all objects in the category set `<class>`. For each object, specify its location within the range `<range>` by determining the top-left and bottom-right corners of its bounding box. To indicate the object’s class and location, provide the output in the format (c, x_1, y_1, x_2, y_2) , where ‘ c ’ represents the class index starting from 0, and (x_1, y_1, x_2, y_2) correspond to the offsets of the bounding box corners relative to the center point. The image is: `<image>`”

Example 2. “Identify all the objects in the image that belong to the category set `<class>` and predict a bounding box around each one. The output should be a list in the format (c, x_1, y_1, x_2, y_2) , where c represents the index of the class label starting from 0, and x_1, y_1, x_2, y_2 are the offsets of the top-left and bottom-right corners of the box relative to the center point. The coordinates should be within `<range>`. The image is: `<image>`”

Example 3. “For each object in the image that is a member of the category set `<class>`, output a tuple with the index of class label starting from 0 and the offsets of corners relative to the center point that encloses the object. The offsets should be in the order of top-left and bottom-right corners of the rectangle and should be within `<range>`. The output format should be (c, x_1, y_1, x_2, y_2) . The image is: `<image>`”

A.2 Instance Segmentation

Example 1. “Segment the objects from the image with class labels from `<class>` and output their coordinates within range `<range>`. The coordinates should be given as the boundary points relative to the center point, and the output format should be $(c, x_1, y_1, x_2, y_2, \dots, x_{20}, y_{20})$, where c is the index of the class label that starts from 0. The image is: `<image>`”

Example 2. “Segment all the objects from the category set `<class>` in the provided image and output a tuple $(c, x_1, y_1, x_2, y_2, \dots, x_{14}, y_{14})$ for each, where c is the index of the class label in the category set that starts from 0, and $(x_1, y_1, x_2, y_2, \dots, x_{14}, y_{14})$ correspond to the offsets of boundary points on the instance mask relative to the center point which should be within `<range>`. The image is: `<image>`”

Example 3. “In the provided image, please segment all the objects in category set `<class>` within the range `<range>` by providing their coordinates in the $(c, x_1, y_1, x_2, y_2, \dots, x_{24}, y_{24})$ format, where ‘ c ’ denotes the index of the class label starting from 0, and $(x_1, y_1, x_2, y_2, \dots, x_{24}, y_{24})$ stand for the offsets of boundary points relative to the center point. The image is: `<image>`”

A.3 Visual Grounding

Example 1. “Please find the object in the category set $\{\langle\text{expression}\rangle:\langle\text{cls0}\rangle\}$ within the range `<range>`. Please provide the output in the format (c, x_1, y_1, x_2, y_2) , where c is the class index starting from 0, and (x_1, y_1, x_2, y_2) are the offsets of the top-left and bottom-right corners of the bounding box relative to the center point. The image is: `<image>`”

Example 2. “Given the input image, category set $\{\langle\text{expression}\rangle:\langle\text{cls0}\rangle\}$, and the range `<range>`, please locate the object in the image and output the corresponding coordinates in the tuple (c, x_1, y_1, x_2, y_2) , where c is the index of the class label starting from 0, and (x_1, y_1, x_2, y_2) are the

offsets of the top-left and bottom-right corners of the rectangle relative to the center point. The image is: <image>

Example 3. “*For each object in the image that belongs to the {<expression>:<cls0>} category set, please provide the class label (starting from 0) and the offsets from the center of a bounding box that encloses the object. The corner offsets should be in the order of top-left and bottom-right, and within the range <range>. The output should be in the format (c, x1, y1, x2, y2). The image is: <image>*”

A.4 Image Captioning

Example 1. “*The image is <image>. Write a caption: ”*

Example 2. “*The image is <image>. Please describe this image: ”*

Example 3. “*With the objects in the <image>, please generate a caption for the image: ”*

A.5 Visual Question Answering

Example 1. “*The image is <image>. Please generate an answer according to the question: <question>.”*

Example 2. “*The image is <image>. Please answer the question <question> according to the image.”*

Example 3. “*With the objects in the <image>, <question>.”*

B Loss Function

VisionLLM consists of two model components: language-guided image tokenizer and LLM-based open-task decoder. So the total loss \mathcal{L} of our model can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{tok}} + \mathcal{L}_{\text{dec}}, \quad (1)$$

where \mathcal{L}_{tok} and \mathcal{L}_{dec} denote the loss of language-guided image tokenizer and LLM-based open-task decoder, respectively. We introduce the two loss functions as follows:

Language-Guided Image Tokenizer. Different from the Q-Former [29], we use a supervision method similar to that of Deformable DETR [82], but with a different loss \mathcal{L}_{tok} : category-agnostic classification (focal loss [31]) and center point regression (L_1 loss). As explained in Sec. 3.3, our image tokenizer extracts M image tokens $T = \{(e_i, l_i)\}_{i=1}^M$, each of which is represented by an embedding e_i and a location l_i (*i.e.*, absolute coordinates of the center point).

LLM-Based Open-Ended Task Decoder. We handle two cases in decoding processing differently. (1) For regular word prediction, we train with standard next-token supervision [54, 45, 6, 46]; (2) For unordered set prediction (*e.g.*, bounding boxes), we first output a sequence of tokens according to the output format (see the output-format-as-query paradigm in Sec. 3.4), then use bipartite matching to align the LLM-predicted outputs with the ground truths. Despite the differences, we use cross-entropy to compute the loss \mathcal{L}_{dec} in a unified way for both cases.

C Training Schedule

As shown in Figure 5, to speed up the convergence of VisionLLM, we split the training schedule of VisionLLM into two stages:

Stage 1. In this stage, we initialize the language-guided image tokenizer by loading the pre-trained weights of Deformable DETR [82] and BERT [16]. Additionally, Alpaca [53] is employed as the LLM-based open-ended task decoder. To align visual tokens with text tokens, we make the language-guided image tokenizer trainable while freezing most parameters of the pre-trained Alpaca, with only a few LoRA [24] parameters left tunable. We only focus on object detection in this stage to simplify the training difficulty, with random task descriptions and object categories.

Stage 2. The second stage builds upon the model weights obtained from the first stage. For efficiency, we freeze the visual backbone (*e.g.*, ResNet [23]) in the language-guided image tokenizer. Notably,

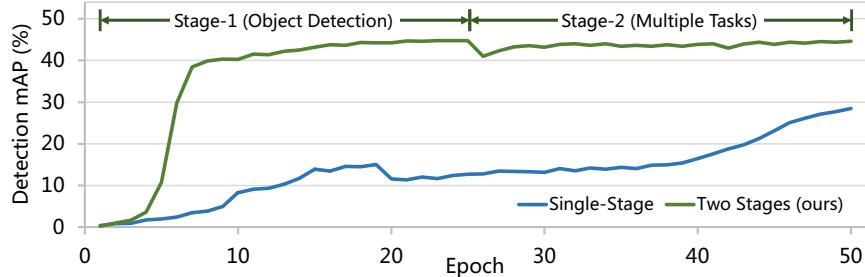


Figure 5: **Comparison of two training schedules for VisionLLM.** We found that a two-stage training from easy to hard converges faster than a single-stage training.

Table 4: **More ablation studies for VisionLLM.**

(a) Effect of randomness.		(b) Effect of LoRA [24].			(c) Effect of the number of image tokens.	
Randomness	AP					
None	45.2	✗	✗	45.2	#Tokens	AP
+ Random Task Description	45.1	✗	✓	1.2	50	44.5
++ Random Object Category	44.8	✓	✓	44.8	100	44.8
+++ Random Output Format (Multi-task Joint Training)	44.6				200	45.1
					300	45.2
(d) Effect of Seq2Seq.		(e) Large vocabulary object detection.				
Seq2Seq	AP					
✓	-				Dataset	#Classes
✗	44.8				COCO	80
					LVIS	1203
						AP
						44.8
						18.9

this stage introduces the unified supervision of multiple tasks, including object detection, instance segmentation, visual grounding, image captioning, and VQA, facilitating the model to leverage the power of LLMs to understand and manipulate visual information holistically.

D More Ablation Studies

In this section, we provide more ablation studies and analysis of VisionLLM. Unless otherwise specified, we use ResNet-50 [23] backbone and perform the ablation experiments for object detection tasks with random task descriptions and object categories on COCO 2017 [32].

Randomness. In Table 4a, we examine the effect of introducing randomness during training for VisionLLM, including randomness in task descriptions, object categories, and output formats (*i.e.*, multi-task joint training). Initially, without any randomness, the model achieves a box AP of 45.2. However, as randomness is gradually applied, interesting phenomena emerge: while there is a slight decrease (45.2 → 44.6) in the AP of standard detection with the introduction of randomness, the overall benefits of enhanced task customization and open-ended capabilities outweigh this minor trade-off. Overall, introducing randomness during training in VisionLLM positively impacts its capacity for open-ended tasks and customization.

Low-Rank Adaptation (LoRA). As shown in Table 4b, when randomness is not applied, the model achieves 45.2 box AP without using LoRA [24]. However, when randomness is employed, it is observed that the model fails to converge without using LoRA. Conversely, when LoRA and randomness are used together, the model is able to converge. This indicates that LoRA plays a crucial role as a bridge between the language and visual tokens, enabling effective alignment between the two modalities and improving the convergence of the overall system.

Number of Image Tokens. We vary the number of image tokens from 50 to 300 to investigate their impact on the performance. Results are presented in Table 4c. As the number of image tokens increases, the performance continues to improve. This makes sense because a larger number of

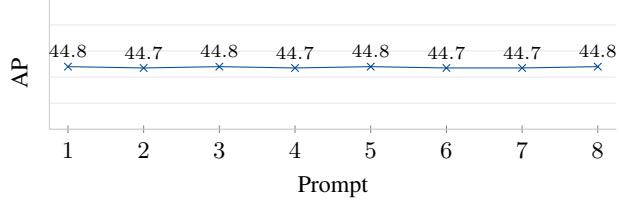


Figure 6: **Evaluation results using eight different prompts.** The first six prompts use different task descriptions of object detection, while the last two prompts employ random category orders. These results show that the performance of different prompts is similar, only a 0.1 AP gap is observed.

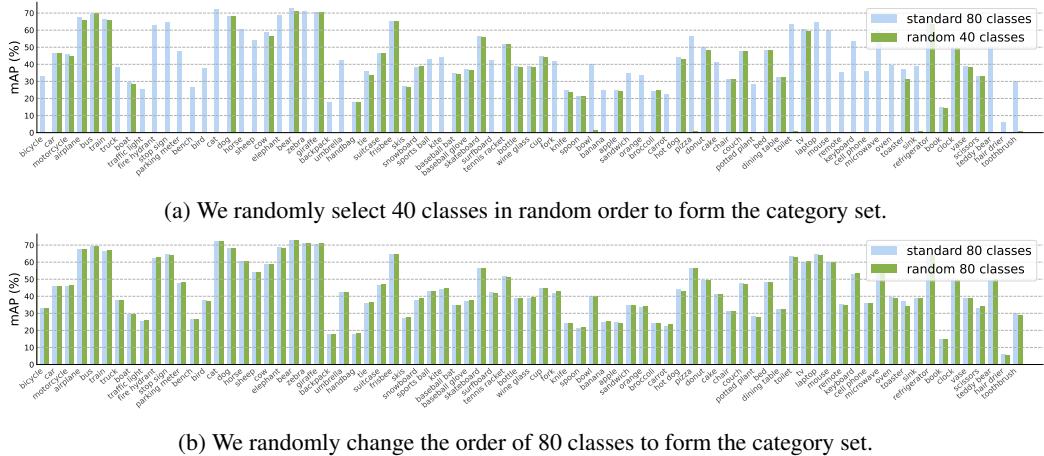


Figure 7: **Per-category AP on COCO dataset.** We randomly select some categories to form the category set $\langle \text{class} \rangle$ in language instructions.

image tokens provides a more detailed description of the image content. Considering computational complexity, we adopted 100 image tokens in our experiments.

Robustness to Prompt Changes. Since VisionLLM is trained with random prompts, including random task descriptions and random categories, one may ask whether there is a large performance variance across different prompts. To validate the stability of VisionLLM, we conduct experiments using eight different prompts. The first six prompts employ different task descriptions, while the last two prompts involve random category orders. In the case of random category orders, we map the categories back to the COCO standard category order for evaluation. As shown in Figure 6, most evaluation results are distributed closely to 44.8 AP. The performance differences among prompts are marginal, demonstrating that VisionLLM is robust to different prompts.

Instruction Following Capability. As shown in Figure 7, when the prompt only contains 40 classes, the performance for these categories remains normal, while the performance for the remaining categories is close to zero. This indicates that VisionLLM can dynamically detect objects based on the given class set $\langle \text{class} \rangle$ in instructions while disregarding the other classes that are not mentioned. This result highlights the flexibility of VisionLLM in adhering to instructions.

Output-Format-As-Query vs. Seq2Seq. In VisionLLM, we introduce the output-format-as-query framework for LLM decoder. Alternatively, we also experiment with the sequence generation method like Pix2Seq [8] for object detection with random task descriptions and object categories. However, we find that the loss is hard to converge in this paradigm, which indicates that the seq2seq decoding may need a more detailed design or a longer training schedule for the open-ended visual tasks, while the proposed output-format-as-query framework is more effective for open-ended tasks.

Large-Vocabulary Object Recognition. To validate the capacity of VisionLLM in the large-vocabulary scenario, we further conduct the experiments on the challenging dataset LVIS [20] with 1203 categories. Due to the limited number of language tokens, we randomly select 80 classes for training in each iteration. During inference, we divide the 1203 categories into 16 groups and

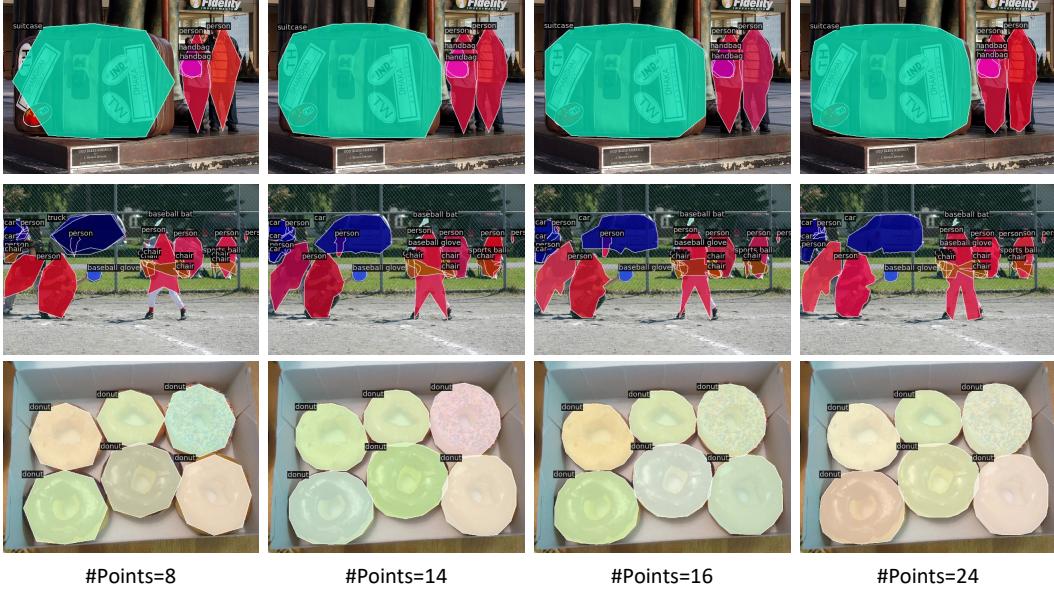


Figure 8: **Customization of instance masks using the different number of points.** Notably, we only modify the output format mentioned in the prompt, *i.e.* the number of segmentation points. For more details, please see the example prompts provided in Sec. A.2.

predict the results in a sliding-window manner. As shown in Table 4e, without tricks like federal loss, VisionLLM-R50 can achieve 18.9 mAP on LVIS.

E Qualitative Analysis

Customization of Segmentation Points. In this experiment, we focus on *modifying the output format mentioned in the prompt*, specifically the number of points for instance segmentation (see Sec. A.2). The results are visualized in Figure 8. Remarkably, by increasing the number of points for segmentation, we observe that the model successfully predicts more refined object masks. This validates the capability of our method to precisely customize the output format, showcasing fine-grained control over the segmentation process.

Customization of Category Set. We change the content of the category set `<class>` in language instructions and visualize the predictions in Figure 9. It is observed that the model can correctly predict the object category according to the provided category set. This demonstrates that VisionLLM has a strong capacity to understand and reason over the semantic information of language instructions, which allows it for flexible category customization in open-vocabulary scenarios.

Image Description & VQA. Benefiting from the power of LLMs, VisionLLM exhibits a strong ability in generating long descriptions for images and answering visual questions with complex reasoning. We show the examples in Figure 10.

Instruction: “Identify the objects in the image that belong to {‘person’: <c0>, …, ‘frisbee’: <c29>, …} and draw a bounding box around each one. The output should be a list of tuples in the format (c, x1, y1, x2, y2), where ‘c’ represents the index of the class label starting from 0, and x1, y1, x2, y2 are the offsets of the top-left and bottom-right corners of the box relative to the center point. The coordinates should be within <range>. The image is: <image>”



Instruction : “Identify the objects in the image that belong to {‘frisbee’: <c0>} and draw a bounding box around each one. The output should be a list of tuples in the format (c, x1, y1, x2, y2), where ‘c’ represents the index of the class label starting from 0, and x1, y1, x2, y2 are the offsets of the top-left and bottom-right corners of the box relative to the center point. The coordinates should be within <range>. The image is: <image>”



Instruction: “Identify the objects in the image that belong to {‘person’: <c0>, ‘bicycle’: <c1>, …, ‘backpack’: <c24>, …, ‘toothbrush’: <c79>} and draw a bounding box around each one. The output should be a list of tuples in the format (c, x1, y1, x2, y2), where ‘c’ represents the index of the class label starting from 0, and x1, y1, x2, y2 are the offsets of the top-left and bottom-right corners of the box relative to the center point. The coordinates should be within <range>. The image is: <image>”



Instruction: “Identify the objects in the image that belong to {‘bicycle’: <c0>, ‘surfboard’: <c1>, ‘sports ball’: <c2>, ‘backpack’: <c3>, ‘the man wearing blue T-shirt’: <c4>} and draw a bounding box around each one. The output should be a list of tuples in the format (c, x1, y1, x2, y2), where ‘c’ represents the index of the class label starting from 0, and x1, y1, x2, y2 are the offsets of the top-left and bottom-right corners of the box relative to the center point. The coordinates should be within <range>. The image is: <image>”



Figure 9: **Customization of category set.** These results demonstrate that VisionLLM can correctly detect the specified objects according to the category set <class> provided in language instructions.

Instruction: "The image is: <image>. Analyze the image in a comprehensive and detailed manner."

VisionLLM: "The image features a busy city street with a large red double-decker bus driving down the road. There are several people walking on the sidewalk, some of whom are carrying backpacks. The bus is positioned towards the left side of the image. There are several people walking along the sidewalk, with some closer to the bus and others further away. The overall atmosphere suggests a busy urban environment with people going about their daily routines."



Instruction: "The image is: <image>. Write a detailed description of the given image."

VisionLLM: "The image features a large, open-air market filled with a variety of fruits and vegetables. There are numerous mangoes, watermelons and apples displayed in the market, with some fruits placed closer to the foreground and others further back. The market is bustling with activity as people can be seen browsing and purchasing the fresh produce. The market is likely a popular destination for those seeking a fresh and healthy diet."



Instruction: "The image is: <image>. Write a short description of the given image."

VisionLLM: "The image features a lively open-air market with a variety of fruits and vegetables. People are busy browsing and buying fresh produce."

(a) Image description with controllable text length

Instruction: "The image is: <image>. Please generate an answer according to the question: what reasons could there be for the cat squatting on the shoes?"

VisionLLM: "There could be several reasons. One possibility is that the cat is seeking warmth or comfort, as shoes can provide a cozy and warm surface for the cat to rest on. Another reason could be that the cat is seeking attention or affection from its owner, who is standing nearby. Cats are known to be curious creatures, and they may be seeking attention or attention from their owners."



Instruction: "The image is: <image>. What do you think of the working environment here?"

VisionLLM: "The working environment in this office space appears to be open, spacious, and well-equipped. There are several desks set up for employees, providing ample workspace for each individual. The office features a modern and contemporary design, which suggests a creative and collaborative work environment. The presence of several chairs and the organization of the office, including the laptops and monitors, indicate that the space is well-maintained and conducive to productivity. Overall, the office appears to be a pleasant and efficient work environment for its employees."



(b) Visual question answering with reasoning

Figure 10: Visualization of the image description and VQA capabilities of VisionLLM.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 4
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 9
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 4
- [4] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 2022. 4
- [5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, 2021. 7
- [6] 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*, 2020. 3, 4, 12
- [7] 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*, 2020. 3, 4, 8
- [8] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 8, 14
- [9] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. *arXiv preprint arXiv:2206.07669*, 2022. 4, 8
- [10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 4, 7
- [11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*. Springer, 2020. 8
- [12] Zhe Chen, Yuchen Duan, Wenhui Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations*, 2023. 1
- [13] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, 2021. 8
- [14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 4
- [15] MOSS contributors. Moss. <https://github.com/OpenLMLab/MOSS>, 2023. 3
- [16] 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. 6, 12
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6
- [18] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 1
- [19] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 2020. 8

- [20] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 14
- [21] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems. *arXiv preprint arXiv:2104.00743*, 2021. 4
- [22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 3, 8, 9
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6, 7, 8, 9, 12, 13
- [24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 7, 12, 13
- [25] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022. 4
- [26] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, 2022. 2
- [27] 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*, 2021. 8
- [28] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhui Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. *arXiv preprint arXiv:2211.09808*, 2022. 4, 7, 8
- [29] 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. 4, 5, 12
- [30] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2, 4, 5
- [31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 12
- [32] 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 *European Conference on Computer Vision*. Springer, 2014. 7, 9, 13
- [33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 4, 5, 7
- [34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 6
- [35] Zhaoyang Liu, Yinan He, Wenhui Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 2, 4, 5
- [36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*. 7
- [37] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7
- [38] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 4
- [39] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 7

- [40] OpenAI. Gpt-4 technical report. *arXiv*, 2023. 3
- [41] TB OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022. 1, 3
- [42] 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*, 2022. 3, 4
- [43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002. 9
- [44] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1, 4
- [45] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018. 12
- [46] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 12
- [47] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 4
- [48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 8
- [49] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 9
- [50] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 2, 4, 5
- [51] Weijie Su, Xizhou Zhu, Chenxin Tao, Lewei Lu, Bin Li, Gao Huang, Yu Qiao, Xiaogang Wang, Jie Zhou, and Jifeng Dai. Towards all-in-one pre-training via maximizing multi-modal mutual information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1
- [52] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022. 1
- [53] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/2023/03/13/alpaca.html*, 2023. 3, 6, 7, 12
- [54] 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. 3, 6, 7, 12
- [55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 6
- [56] 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*, 2015. 9
- [57] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1, 4
- [58] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 1, 4, 7
- [59] Wenhui Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 7, 8

- [60] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 6
- [61] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhajit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1, 2
- [62] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. *arXiv preprint arXiv:2212.02499*, 2022. 2
- [63] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. *arXiv preprint arXiv:2212.02499*, 2022. 4
- [64] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 2, 4
- [65] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 4, 5, 11
- [66] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022. 4
- [67] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 4
- [68] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2, 4, 5
- [69] Enze Xie, Peize Sun, Xiaoge Song, Wenhui Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 6, 8
- [70] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. *arXiv preprint arXiv:2303.06674*, 2023. 4
- [71] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. 4, 7
- [72] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, 2022. 8
- [73] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2, 4, 5
- [74] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 2
- [75] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*. Springer, 2016. 7
- [76] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 2
- [77] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Gilm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 3
- [78] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3

- [79] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022. [2](#)
- [80] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [4](#)
- [81] Jinguo Zhu, Xizhou Zhu, Wenhui Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *arXiv preprint arXiv:2206.04674*, 2022. [1](#), [4](#), [8](#), [9](#)
- [82] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. [3](#), [6](#), [7](#), [8](#), [12](#)
- [83] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. [2](#), [4](#), [5](#), [8](#), [9](#)