Groma: Localized Visual Tokenization for Grounding Multimodal Large Language Models

Chuofan Ma[1⋆]   Yi Jiang[2†]   Jiannan Wu[1]   Zehuan Yuan[2]   Xiaojuan Qi[1†]

[1]The University of Hong Kong   [2]ByteDance Inc.

Abstract. We introduce Groma, a Multimodal Large Language Model (MLLM) with grounded and fine-grained visual perception ability. Beyond holistic image understanding, Groma is adept at region-level tasks such as region captioning and visual grounding. Such capabilities are built upon a localized visual tokenization mechanism, where an image input is decomposed into regions of interest and subsequently encoded into region tokens. By integrating region tokens into user instructions and model responses, we seamlessly enable Groma to understand user-specified region inputs and ground its textual output to images. Besides, to enhance the grounded chat ability of Groma, we curate a visually grounded instruction dataset by leveraging the powerful GPT-4V and visual prompting techniques. Compared with MLLMs that rely on the language model or external module for localization, Groma consistently demonstrates superior performances in standard referring and grounding benchmarks, highlighting the advantages of embedding localization into image tokenization. Project page: https://groma-mllm.github.io/.
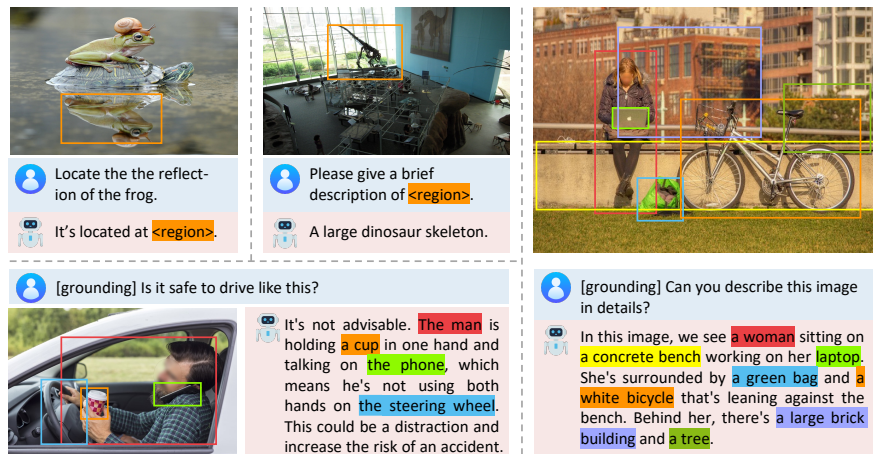
Fig. 1: Groma is a multimodal large language model with exceptional region understanding and visual grounding capabilities. It can take user-defined region inputs (boxes) as well as generate long-form responses that are grounded to visual context.

⋆ Part of the work was done during Chuofan's internship at ByteDance.

† Corresponding authors

1   Introduction

Multimodal Large Language Models (MLLMs) have spread the sparks of artificial general intelligence [5] from language to the visual domain [13,33,50,57,67]. Owing to the foundational capabilities of Large Language Models (LLMs) [12,36, 37,45,46], MLLMs excel in vision-language tasks that require advanced understanding and reasoning, such as image captioning and visual question answering. However, despite these achievements, current MLLMs typically fall short of localization capabilities, thus cannot ground understanding to the visual context. Such limitations constrains the model from fulfilling its potential in real-world applications like robotics, autonomous driving, and augmented reality.

In light of the gap, one stream of research attempts to augment the LLM to directly output quantized object coordinates for localization [3,7,8,39,50,58] (Fig. 2(a)). While this method is simple in design, the substantial computational demands of LLMs make it challenging to process high-resolution image inputs, which are essential for accurate localization. Besides, the nature of sequence outputs in LLMs is not well-suited for dense prediction tasks such as segmentation. These concerns elicit another stream of research, which incorporates an external localization module (*e.g.*, SAM [22]) to decode bounding boxes or masks [26,40,43,62] (Fig. 2(b)). This approach circumvents aforementioned issues, but introduces additional model complexity.

The above motivates us to explore a new paradigm for grounded MLLMs. Drawing inspiration from open-vocabulary object detection [66], we decompose the grounding task into two sub-problems: discovering the object (localization) and relating the object to texts (recognition). We notice that localization alone requires little semantic understanding but demands perceptual skills, which is typically out of the scope of an LLM's expertise. This inspires us to decouple localization and recognition within MLLMs. But instead of using external modules, we propose exploiting the spatial understanding capability in the visual tokenizer of MLLMs for localization (Fig. 2(c)). This perceive-then-understand design also resembles human vision process.

Building upon this concept, we introduce Groma[1] (**Gro**unded **M**ultimodal **A**ssistant), an MLLM with localized and fine-grained visual perception abilities. Specifically, Groma incorporates region tokenization alongside standard image tokenization to identify and encode potential regions of interest (ROIs) into region tokens. During this process, location information is extracted from the image and associated with region tokens, with each region token anchored to the underlying ROI. This allows Groma to ground its textual output by simply referring to region tokens, alleviating the need for the LLM to meticulously regress object coordinates. Moreover, the tokenizer of Groma can also encode user-specified region inputs (*i.e.*, bounding boxes) into region tokens, which are directly inserted into user instructions to initiate referential dialogue.

---

[1] In Latin, Groma refers to an instrument used for accurate measurement, which implies our focus on accurate localization for MLLMs.

By settling localization to the image tokenization process, Groma circumvents the heavy computation of LLMs when handling high-resolution input. Specifically, Groma uses high-resolution images for tokenizer input but downsamples image tokens for LLM input, which saves computation without sacrificing localization accuracy. Besides, unlike methods adopting separate designs for modeling grounding outputs and referring inputs [43,62], Groma seamlessly unifies the two capabilities with the use of region tokens.

From the data perspective, to improve the localized understanding of Groma, we adopt an extensive collection of datasets with region-level annotations for training, which encompasses a range of region semantics from objects and relationships to detailed region descriptions. In addition, to remedy the lack of long-form grounded data, we construct a visually grounded chat dataset called Groma Instruct for instruction finetuning. Groma Instruct is the first grounded chat dataset constructed with both visual and textual prompts, leveraging the powerful GPT-4V for data generation.

Our comprehensive experiments demonstrate the superiority of the design of Groma, with results showing that it outperforms all comparable MLLMs on established referring and grounding benchmarks. We also showcase that Groma maintains strong image-level understanding and reasoning abilities on the conversational VQA benchmark. Moreover, to assess the ability to localize multiple, diverse, and variably-sized objects, we adapt the LVIS [15] detection benchmark for object grounding evaluation. On this challenging benchmark, Groma surpasses alternative methods by a significant margin (over 10% AR), highlighting its robust and precise localization capabilities.
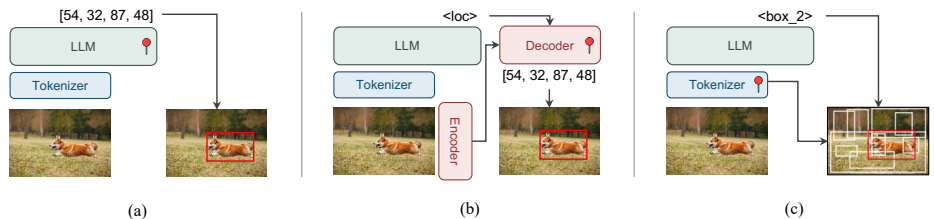

Fig. 2: Different paradigms of grounded MLLMs. We mark the modules for localization with ⚲. (a) LLM for localization (*e.g.*, Kosmos-2 [39] and Shikra [8]); (b) External modules for localization (*e.g.*, Lisa [26]); and (c) Localized visual tokenization (Ours).

2   Related Work

Image-level MLLMs. Large language models (LLMs) such as GPT series [1, 53] and LLaMA [45,46] have recently undergone rapid development and sparked a revolution in the field of natural language processing. Such progress inspires the community to extend the foundational capabilities of LLMs to the visual

domain, giving birth to multimodal large language models (MLLMs). The pioneering works [2, 13, 27, 28, 57, 62, 67] of MLLMs typically follow a tripartite architecture, comprising a visual encoder, a vision-language connector, and a large language model. Specifically, BLIP-2 [28] and Flamingo [2] first propose the Q-Former/Resampler to bridge vision and language. LLaVA [62] and MiniGPT4 [67] streamline this vision-language connector to a linear layer, and introduce visual instruction tuning to enhance the instruction-following ability of MLLMs. Following works [10, 50] further showcase the immense potential of MLLMs by scaling up the visual components to the magnitude as LLMs. While these works have exhibited impressive visual understanding capabilities, they are predominantly constrained to image-level tasks, such as image captioning and image visual question answering. This necessitates the research into region-level MLLMs, which unlock more nuanced and granular visual-language interactions.

Region-level MLLMs. In pursuit of fine-grained and grounded image understanding, recent studies further integrate region-level data into the training of MLLMs [7, 8, 39, 51, 52, 60, 65]. In particular, to model box inputs and outputs, Kosmos-2 [39] and Shikra [8] directly quantize bounding boxes into discrete location tokens or numeric representation of positions. GPT4RoI [64] and RegionGPT [14] use a simple pooling operation to extract the features within boxes or masks as the region representations. While Ferret [58] proposes a spatial-aware visual sampler to deal with free-form region inputs. Besides, to achieve more accurate localization, some works [26, 43, 63] resort to off-the-shelf models for pixel-level grounding. For instance, LISA [26] takes the segmentation token generated by the MLLM as the prompts for SAM [22] to produce the segmentation masks. GLaMM [43] and LLaVA-Ground [62] further advance the concept and enable grounded conversation generation. Our work shares the same focus with the aforementioned methods on region-level understanding and grounding. Yet, we distinguish ourselves from existing studies by proposing a novel perspective in enhancing the localization ability of MLLMs.

3   Method

In this section, we present Groma, a grounded multimodal large language model capable of understanding user-defined region inputs and generating visually grounded outputs. We first illustrate the model architecture of Groma in Sec. 3.1. Then we introduce how to format region input and output in Sec. 3.2. Finally, we detail the learning pipelines Sec. 3.3.

3.1   Model Architecture

As illustrated in Fig. 3, Groma primarily consists of (1) an image encoder for scene-level image tokenization, (2) a region proposer for discovering regions of interest, (3) a region encoder for region-level image tokenization, and (4) a large language model for modeling multimodal input and output. We detail each component in the following paragraphs.
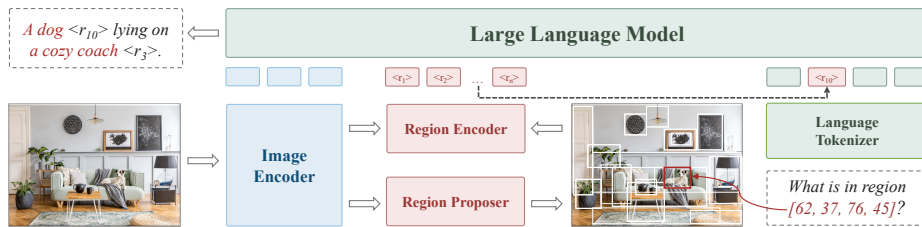
Fig. 3: Overview of Groma. Groma encodes the image input into both global image tokens and local region tokens. For region tokenization, a general-purpose region proposer is introduced to discover regions of interest, followed by a light-weight region encoder. By integrating region tokens into user instructions and model responses, Groma unlocks the referring and grounding abilities of MLLMs.

Image Encoder. Groma employs a pretrained DINOv2 [38] model as the image encoder with the input image resolution set to 448×448. Compared with the commonly adopted CLIP [42] visual encoder, DINOv2 is preferred in this work for its compatibility with high-resolution inputs and fine-grained features for localization. However, the use of higher-resolution images leads to extended sequences of visual input for the language model, *e.g.*, 1024 tokens in this case. To save computations, we further concatenate every four neighbor patch tokens into a single token following MiniGPT-v2 [7]. But slightly different from [7], we merge tokens adjacent in 2D instead of 1D, which yields better performance.

Region Proposer. To obtain localized understanding of the image, Groma innovatively incorporates a region proposer into the image tokenization process. Specifically, the region proposer is implemented as a class-agnostic detector head using the Deformable DETR (DDETR) transformer [68]. The original classification head of DDETR is replaced by a binary classifier to score region proposals based on their localization quality. Inspired by ViTDet [29], we extract feature maps from the last 4 layers of the image encoder, and rescale these feature maps to construct a hierarchical feature pyramid as the input to the region proposer. For each image, the region proposer generates 300 region proposals, which are then filtered by NMS and objectness scores before fed into the region encoder.

Region Encoder. The region encoder translates region proposals (*i.e.*, bounding boxes), coming from both user input and the region proposer, into region tokens. Akin to the previous step, we select feature maps from the last three layers of the image encoder to create a hierarchical feature pyramid. A multi-scale ROIAlign [16] module as implemented in [43, 64] is utilized to crop and fuse these hierarchical features into unified region tokens. Compared with alternative ways to represent regional inputs, such as numerical representation of positions [8] and discrete location tokens [7, 39], the region token representation offers distinct benefits as it is semantically aligned with the underlying region, which renders it more intuitive for the language model to comprehend.

LLM. We adopt pretrained Vicuna [12] as the language model of Groma. In particular, we instantiate Groma with the 7B version of Vicuna. Besides, we follow LLaVA v1.5 [32] to use an MLP layer to project the image tokens and region tokens into the feature space of the LLM.

3.2   Input and Output Formatting

Beyond textual only instructions and responses, Groma offers the flexibility to accept user-specified regions as input (referring) and generate visually grounded answers (grounding). Specifically, although different in task formulations, both referring and grounding are unified into one format with the use of region tokens.

Grounding Output. Remember in the tokenization process, each region token is inherently anchored to a concrete location in the image, corresponding to its region proposal. This connection allows the language model to ground its text output to particular regions in the image by simply referring to the associated region tokens. However, as region tokens are continuous embeddings, they cannot be directly integrated into the codebook of the language model and referenced in the text output. To bridge the gap, we further introduce a set of proxy tokens "`<r₁>`, `<r₂>`, ..., `<rₙ>`" to register region tokens. As illustrated below, Groma can refer to any region in the output by addressing the proxy token.

> **User**: Here is an image with region crops from it. Image: `<image>`. Regions: `<r₁><region>`, `<r₂><region>`, ..., `<rₙ><region>`. [grounding] Please briefly describe the image content.
>
> **Groma**: `<p>`A dog`</p>` `<roi><r₄></roi>` is jumping to catch `<p>`a frisbee`</p>` `<roi><r₇></roi>` over `<p>`a fallen man`</p>` `<roi><r₁></roi>`.

`<image>` and `<region>` stand for placeholders of image tokens and region tokens, which are replaced by corresponding visual tokens before being fed into the LLM. `<p>` and `</p>` marks the start and end of the grounded phrase. `<roi>` and `</roi>` are used to enclose the referenced regions. In addition, we use a special token [grounding] to inform the model to generate grounded responses.

Referring Input. For a region pointed out by the user, we treat it the same as region proposals from the region proposer, *i.e.*, encoding it into a region token and assigning a proxy token to it. This allows us to incorporate user-specified regions into our instructions by inserting corresponding region tokens. A simple example of referential dialogue in Groma is given below, where `<r₁₀>` comes from user-specified region input.

> **User**: Here is an image with region crops from it. Image: `<image>`. Regions: `<r₁><region>`, `<r₂><region>`, ..., `<rₙ><region>`. What is `<r₁₀><region>`?
>
> **Groma**: A cute cat sleeping on a wooden bench.

3.3   Model Training

The training of Groma is partitioned into three stages: (i) detection pretraining for localization ability, (ii) alignment pretraining for image-level and region-level vision-language alignment, (iii) instruction finetuning for enhanced conversation capability. Tab. 1 enumerates the datasets used at different training stages. Additionally, we provide the instruction templates used to convert task-specified datasets to instruction following format in Appendix.

Table 1: Datasets used at three training stages. RefCOCO/g/+ is short for Ref-COCO, RefCOCO+, and RefCOCOg. REC means referring expression comprehension.

Training stage	Data types	Datasets
Detection pretraining	Detection	COCO, Objects365, OpenImages, V3Det, SA1B
Alignment pretraining	Image caption	ShareGPT-4V-PT
	Grounded caption	Flickr30k Entities
	Region caption	Visual Genome, RefCOCOg
	REC	COCO, RefCOCO/g/+, Grit-20m
Instruction finetuning	Grounded caption	Flickr30k Entities
	Region caption	Visual Genome, RefCOCOg
	REC	COCO, RefCOCO/g/+
	Instruction following	Groma Instruct, LLaVA Instruct, ShareGPT-4V

Detection Pretraining. This training stage only involves the image encoder and the region proposer, which collectively constitute a DDETR-like detector. The image encoder is kept frozen during training. To endow the region proposer with localization capability, an extensive collection of detection datasets, including COCO [30], Objects365 [44], OpenImages [25], and V3Det [47], is utilized for large-scale pretraining. Notably, category information is omitted from the training process, with a primary focus on box supervision.

Considering traditional detection data are typically limited to object-level annotations, we complement the training with a two million subset of SA1B [23] data filtered by GLEE [19]. Original mask annotations of SA1B are transformed into bounding boxes for consistency. The inclusion of this enriched dataset encourages the region proposer to produce region proposals encompassing not only object instances but also their constituent parts and various background stuff.

Alignment Pretraining. To align vision and language feature space of Groma, we pretrain the model on a wide range of vision-language tasks. Specifically, for image-level alignment, we leverage ShareGPT-4V-PT [9] for detailed image captioning. For region-level alignment, we engage COCO [30], RefCOCO [21], RefCOCO+ [59], RefCOCOg [35], and Grit-20m [39] for referring expression comprehension (REC), Visual Genome [24] for region captioning, and Flickr30k Entities [41] for grounded caption generation. To maintain training efficiency, we focus finetuning efforts on the MLP projection layer and the region encoder, while other modules are kept frozen throughout the training.

Instruction Finetuning. Based on alignment pretraining, we refine the training data to focus exclusively on high-quality datasets and proceed to unfreeze the language model for finetuning purposes. At this stage, LLaVA Instruct [33] and ShareGPT-4V [9] are incorporated to improve the conversational and instruction-following capabilities of Groma[2]. Besides, we curate a high-quality grounded chat dataset, named Groma Instruct (see next section for more details), to facilitate synergy of chatting and grounding abilities of Groma.

Discussions. A major difference between the training of Groma and current MLLMs is the integration of dedicated detection pretraining, which endows Groma with robust and precise localization ability. Thanks to the decoupled architecture of location and understanding within Groma, we circumvent the need to involve the LLM during detection pretraining. Such a strategic design allows Groma to benefit from pretraining on millions of bounding box annotations — a task that would be computationally prohibitive for classic MLLMs.

4   GPT4V-assisted Grounded Conversation Generation

Visual dialogue data have proven to be crucial in advancing the conversational capability of the MLLM as a visual chatbot. Previous methods mostly rely on coarse-grained image descriptions to derive free-form visual dialogues, which typically lack fine-grained region details and precise location information [33, 67]. For grounded MLLMs, such free-form dialogue data are shown to be insufficient to enable the model to generate long-form grounded responses [62] - as the format of grounded responses significantly deviates from that of normal responses, it could be challenging for the grounded MLLM to generalize its grounding capability to long-form conversations.

To bridge the gap, we have meticulously curated a dataset containing 30k visually grounded conversations for instruction finetuning, named Groma Instruct. An illustrative example from Groma Instruct is showcased in Fig. 4. Specifically, we select images with dense region annotations from Visual Genome [24] (VG), and take the following steps to construct grounded conversations with the assistance of advanced GPT-4V model:

- First, we remove highly overlapped regions (bounding boxes) from VG annotations, normally leaving 3-10 regions of interest for each image. Then we adapt the visual prompting techniques from SoM [56] to overlay a bright numeric marker at the center of each region. Using this marked image as input unleashes the grounding capabilities of GPT-4V - it can easily make references to specific image regions by addressing the corresponding numbers.

- Besides visual input, we supply GPT-4V with rich region descriptions, image descriptions, and image-based Q&A pairs, coming from COCO [30] and VG

---

[2] Since the detailed description data in LLaVA Instruct contains severe hallucinations, we replace it with ShareGPT-4V as in [9].

annotations[3]. While such textual context is optional for GPT-4V input, we empirically find it useful to reduce hallucinations in generated contents and resolve potential ambiguities in visual prompts[4].

– Inspired by prior studies on visual chat data construction [9, 33, 48, 67], we further provide GPT-4V with manually designed grounded chat as context examples. This provokes the in-context-learning ability of GPT-4V to generate grounded conversations in a uniform format. We also take a post-processing stage to filter out conversations not following the pre-defined format.

Region descriptions:
[1] dog walking on beach
[2] surfboard under woman's arm
[3] woman walking in wetsuit
[4] sand on ocean shore
[5] ocean with rolling waves
[6] hazy mountains on horizon

Q&A pairs:
What is the surfer holding?
Surf board.
===
What is the dog doing?
Following its master.
…

Image Descriptions:
A woman in a wetsuit carries a surfboard and walks with a dog.
A lady in a wet suit walking with a dog on a beach.
A woman carries a surf board as a dog walks beside her.
A woman and her dog are walking away from the ocean with a surfboard.
A person is walking with a dog and a surfboard.

Generated Grounded Conversations:
Question: Can you describe to me what you see in this image?
Answer: Sure, I see a beach scene with <p> a woman </p> [3] wearing a wetsuit. She's walking near the water's edge, and there's <p> a dog </p> [1] walking alongside her. <p> The woman </p> [3] is carrying <p> a surfboard </p> [2] under her arm. I can also see <p> the ocean </p> [5] with some waves and <p> faint mountains </p> [6] in the distance.
===
Question: What kind of dog is it, and what's it doing?
Answer: It looks like <p> a medium-sized, short-haired dog </p> [1], quite possibly a mixed breed. <p> The dog </p> [1] is just strolling beside <p> the woman </p> [3], seemingly following her as they both walk across <p> the sand </p> [4].
===
Question: Is the beach crowded, or does it seem deserted?
Answer: It actually looks quite deserted, with no one else in sight apart from <p> the woman </p> [3] and <p> her dog </p> [1]. It gives off a peaceful, solitary vibe - just them and the <p> vast ocean </p> [5] and <p> sandy shore </p> [4].
===
Question: What is the state of the ocean, and is it suitable for surfing?
Answer: <p> The ocean </p> [5] appears quite serene but with rolling waves, hence appropriate for surfing.

Fig. 4: A data example from Groma Instruct. The top part demonstrates the image input with visual prompts and contextual text input to GPT-4V. Notably, all numbered regions have corresponding bounding box annotations, which are not shown in the image. The bottom part is the grounded conversations generated by GPT-4V.

---

[3] We select VG images that also have a coco id. Thus, we can retrieve corresponding image captions from COCO Caption.

[4] There are cases where two regions highly overlap with each other and GPT-4V can hardly tell from the image which region maps to which numeric marker. For these cases, GPT-4V could rely on the numbered region descriptions to find out correspondences between regions and markers.

5   Experiments

In this section, we first quantitatively access the abilities of Groma on grounding (Sec. 5.2), referring (Sec. 5.3), and image-based conversation (Sec. 5.4) tasks. Then we provide qualitative results to exemplify the strong capabilities of Groma on a wide range of region-level tasks (Sec. 5.5). Finally, we ablate the design and training of Groma in Sec. 5.6.

5.1   Implementation Details.

We adopt DINOv2-L/14 [38] as the image encoder and Vicuna-7B v1.5 [12] as the language model. The region proposer follows an encoder-decoder architecture with 6 encoder layers and 6 decoder layers. We further employ mixed query selection and look-forward-twice scheme as in [61] to accelerate convergence. We set NMS threshold to 0.6 and filter out region proposals with objectness scores lower than 0.15. Subsequently, we select the top 100 region proposals if there are more than 100 proposals left after filtering. This results in no more than 356 visual tokens in total. For training, we sequentially proceed 12 epochs of detection pretraining, 2 epochs of alignment pretraining, and 1 epoch of instruction finetuning. More training details can be found in the Appendix.

5.2   Grounding Benchmark Results

Table 2: Results on referring expression comprehension benchmarks. We report accuracy with the IoU threshold set to 0.5. We make Qwen-VL gray because it uses a much larger visual tokenizer (1.9B ViT-bigG [17]).

Method	Model type	RefCOCO			RefCOCO+			RefCOCOg		Average
		val	testA	testB	val	testA	testB	val	test	
MDETR [20]		86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	81.81
G-DINO [34]	Specialist	90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02	86.60
UNINEXT-L [55]		91.43	93.73	88.93	83.09	87.90	76.15	86.91	87.48	86.95
VisionLLM [52]		–	86.70	–	–	–	–	–	–	–
OFA [49]		79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	72.65
Shikra [8]		87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	82.93
Ferret [58]	Generalist	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76	83.91
MiniGPT-v2 [7]		88.69	91.65	85.33	79.97	85.12	74.45	84.44	84.66	84.29
Qwen-VL [4]		89.36	92.26	85.34	83.12	88.25	77.21	85.58	85.48	85.83
Groma		**89.53**	**92.09**	**86.26**	**83.90**	**88.91**	**78.05**	**86.37**	**87.01**	**86.52**

We evaluate the localization capability of Groma on visual grounding tasks. Tab. 2 showcases our performance on three classic referring expression comprehension benchmarks: RefCOCO [21], RefCOCO+ [59], and RefCOCOg [35]. Groma notably surpasses other generalist models of similar model size across all metrics. Moreover, as a generalist model, Groma shows competitive results with

state-of-the-art specialist models [34, 55]. These findings underscore the strong capability of Groma in visual grounding.

However, we notice that traditional REC benchmarks only cover a narrow range of common objects in their referring expressions, which is insufficient to thoroughly evaluate the MLLM's localization capability. Therefore, we further introduce LVIS-Ground, an object grounding benchmark converted from the LVIS [15] detection data. LVIS-Ground contains 4299 images covering 1203 categories of objects, with on average 3.7 target objects per image. Complementary to REC benchmarks, LVIS-Ground focuses on testing the model's ability to locate multiple, diverse, and variably-sized objects. For more details of LVIS-Ground, please refer to the Appendix.

Tab. 3 presents our results on LVIS-Ground. Notably, Groma demonstrates clear advantages over other grounded MLLMs, especially on the AR@0.75 metric. This evidences that the specialized design and training indeed bring more accurate localization for Groma. Moreover, it is noteworthy that current MLLMs all fall short of small object localization (AR@s metric). We conjecture this is mainly because the training data (*e.g.*, RefCOCO/g/+, Flickr30k) lack annotations for small objects. We also notice that current models typically predict only one box per image, even though more than one targets are presented. This is an expected behavior as the they heavily rely on REC data for grounding training, which only has one target object per query. These findings call for the necessity of diversifying grounding data used for training in future MLLMs.

Table 3: Results on the LVIS-Ground benchmark. We report average recall (AR) to measure performances. For each model, we use the native prompt template recommended by the paper for evaluation.

Method	AR	AR@0.5	AR@0.75	AR@s	AR@m	AR@l
Shikra [8]	4.9	14.2	2.0	0.1	3.1	18.5
MiniGPT-v2 [7]	11.4	19.8	11.2	0.3	8.0	41.1
Ferret [58]	16.8	29.6	16.3	1.6	16.7	51.1
Groma	**28.8**	**37.9**	**30.3**	**8.7**	**35.6**	**64.3**

5.3 Referring Benchmark Results

We evaluate Groma on the region captioning task to assess its fine-grained region understanding capability. To prompt the model to generate region-level descriptions, we use queries like *"Please describe <region> in details."*. Tab. 4 presents our results on two established region captioning benchmarks, RefCOCOg and Visual Genome. Without task-specific finetuning, Groma shows comparable or improved performance over GLaMM[5] [43], which has separate designs for input referring and output grounding. This exemplifies the superiority of unified refer-and-ground formulation in Groma.

---

[5] We re-evaluate the performance of GLaMM using the officially released checkpoint after fixing the bug in its original evaluation scripts.

Table 4: Results on region captioning benchmarks. We report METEOR and CIDEr scores to measure caption quality. [†]: with task-specific finetuning.

Method	RefCOCOg		Visual Genome	
	METEOR	CIDEr	METEOR	CIDEr
GRIT [54]	15.2	71.6	17.1	142
Kosmos-2 [39]	14.1	62.3	–	–
GPT4RoI [64]	–	–	17.4	145.2
GLaMM[†] [43]	16.1	101.9	**19.0**	**163.9**
ViP-LLaVA [6]	16.6	105.9	–	–
Groma	**16.8**	**107.3**	**19.0**	158.4

5.4   Conversational VQA Benchmark Results

In addition to region-level tasks, we further evaluate Groma on the conversational style VQA benchmark, LLaVA Bench (COCO) [33], which contains three types of questions, namely conversation, detailed description, and complex reasoning. As shown in Tab. 5, Groma surpasses the strong baseline method LLaVA [33] and achieves competitive performance among grounded MLLMs, especially in detailed image description. This demonstrates that Groma maintains decent image understanding and visual chatting abilities. For the underperformance in conversation and complex reasoning questions, we speculate this could be resulted from the DINOv2 features, as shown in recent studies [18, 31]. But we believe such gap can be closed by scaling up vision-language alignment pretraining.

Table 5: Results on LLaVA-Bench (COCO).

Method	Conversation	Description	Reasoning	Average
LLaVA [33]	85.4	68.3	92.1	81.9
Kosmos-2 [39]	71.7	63.4	74.9	70.0
Shikra [8]	80.6	70.7	88.1	79.9
LLaVA-G [62]	79.3	71.2	92.8	81.2
Ferret [58]	**84.4**	79.4	**96.3**	**86.7**
Groma	82.6	**84.0**	88.8	85.2

5.5   Qualitative Results

Fig. 5 presents a comparison between Groma and other grounded MLLMs on the grounded image captioning task. We choose an exemplar image that is inherently challenging with multiple and occluded instances to ground. Groma manifests exceptional grounding performance in this case with the highest recall and minimum hallucinations. In addition, we provide several visualization examples in Fig. 6 for a complementary understanding of Groma's abilities on grounded chat and referential dialogue. We show that Groma is capable of generating long-form, grounded and logically rich answers, which can be mainly attributed to the introduction of Groma Instruct data in finetuning.
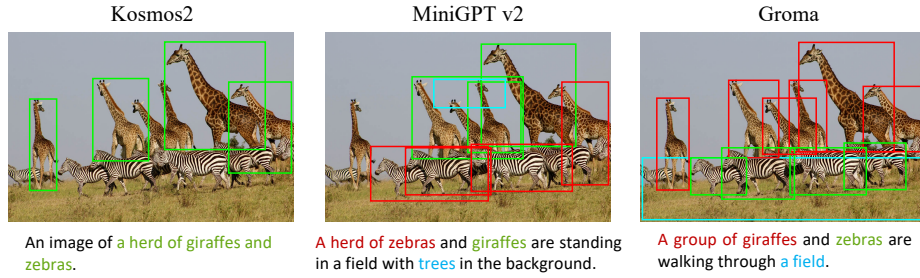
Kosmos2   MiniGPT v2   Groma

An image of a herd of giraffes and zebras.   A herd of zebras and giraffes are standing in a field with trees in the background.   A group of giraffes and zebras are walking through a field.

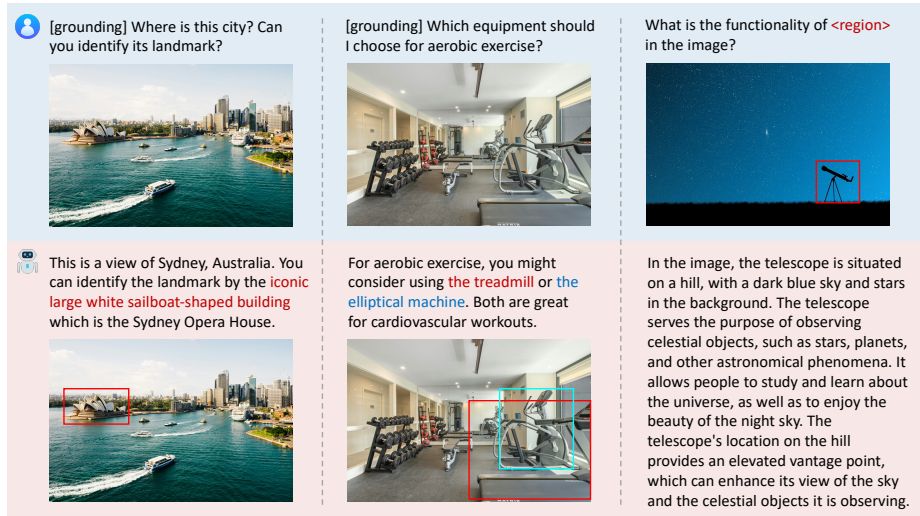**Fig. 5:** A qualitative comparison on the grounded image captioning task.

Fig. 6: Examples on visually grounded chat and referential dialogue.

5.6   Ablation

CLIP *vs*. DINOv2. To quantitatively assess the differences in localization capabilities between CLIP and DINOv2, we compare the two backbones on the COCO detection benchmark in Tab. 6. For this comparison, we equip each backbone with a DDETR [68] detection head and finetune only the detection head on COCO dataset. It can be seen that under the same resolution, DINOv2 backbone significantly outperforms CLIP backbone by 6.5 AP. Furthermore, by scaling the resolution to 448×448, DINOv2 backbone achieves a commendable performance of 43.6 AP. The results consolidate our choice of DINOv2 backbone in Groma.

Frozen LLM. In Tab. 7, we reveal that Groma retains robust localized understanding even without finetuning the LLM, *i.e.*, it demonstrates a referring ability on par with GPT4ROI [64] (148.0 *vs*. 145.2) and grounding ability comparable to Ferret [58] (84.02% *vs*. 83.91%). This finding suggests our design

effectively decouples localization and understanding within Groma, such that it requires minimum 'new knowledge' from the LLM for localized understanding.

Token Merge. To save computations, Groma by default concatenates every 4 image tokens into one as LLM inputs. Through control experiments in Tab. 8, we find that such downsampling has negligible impacts on the grounding performances (*e.g.*, less than 0.1% average accuracy drop on the REC benchmarks). The results evidence that the decoupled design is optimal in both efficiency and localization accuracy.

Table 6: Object detection performances on COCO with different backbones and image resolutions.

Backbone	Resolution	AP
CLIP	336×336	32.4
DINOv2	336×336	38.9
DINOv2	448×448	43.6

Table 7: Referring and grounding abilities with the frozen LLM. We measure referring ability with CIDEr score on Visual Genome and grounding ability with average accuracy on REC benchmarks.

Stage	LLM status	Referring	Grounding
pretraining	frozen	–	82.33
finetuning	frozen	148.0	84.02
finetuning	unfrozen	158.4	86.52

Table 8: Ablation on image token downsampling on the REC benchmarks.

Downsampling	RefCOCO			RefCOCO+			RefCOCOg		Average
	val	testA	testB	val	testA	testB	val	test	
✓	89.32	92.15	85.96	84.11	88.10	78.40	86.33	87.40	86.47
✗	89.54	92.54	86.18	83.72	88.52	78.96	86.17	86.84	86.55

6   Limitations and Conclusions

In this paper, we introduce a novel paradigm, Groma, to unleash the localized perception capabilities of MLLMs. We make the pioneering attempt to embed localization into image tokenization. Our paradigm is based on a perception-then-understand mindset that separates localization from high-level understanding and reasoning. Without introducing external modules, our approach overcomes the resolution bottleneck of using LLMs as location decoders and unifies referring and visual grounding tasks. Extensive experiments showcase the superior performance of our approach in localized perception, as evidenced by its success in referring and visual grounding tasks.

However, the current implementation does not support free-form region inputs and pixel-level grounding. A promising direction to address such limitations is to re-implement the region encoder with a visual sampler as in [58,69] and replace the box region proposer by a mask region proposer like Mask2Former [11]. We leave this for future studies.

Acknowledgements

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 3

2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 4

3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023) 2

4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023) 10

5. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 (2023) 2

6. Cai, M., Liu, H., Mustikovela, S.K., Meyer, G.P., Chai, Y., Park, D., Lee, Y.J.: Making large multimodal models understand arbitrary visual prompts. arXiv preprint arXiv:2312.00784 (2023) 12

7. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023) 2, 4, 5, 10, 11

8. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023) 2, 3, 4, 5, 10, 11, 12

9. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023) 7, 8, 9

10. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023) 4

11. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022) 14

12. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) (2023) 2, 6, 10

13. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=vvoWPYqZJA 2, 4

14. Guo, Q., Mello, S.D., Yin, H., Byeon, W., Cheung, K.C., Yu, Y., Luo, P., Liu, S.: Regiongpt: Towards region understanding vision language model (2024) 4

15. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019) 3, 11

16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 5

17. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (2021), https://doi.org/10.5281/zenodo.5143773 10

18. Jiang, D., Liu, Y., Liu, S., Zhang, X., Li, J., Xiong, H., Tian, Q.: From clip to dino: Visual encoders shout in multi-modal large language models (2023) 12

19. Junfeng, W., Yi, J., Qihao, L., Zehuan, Y., Xiang, B., Song, B.: General object foundation model for images and videos at scale. arXiv preprint 2312.09158 (2023) 7

20. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021) 10

21. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014) 7, 10

22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 2, 4

23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 7

24. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017) 7, 8

25. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020) 7

26. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023) 2, 3, 4

27. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023) 4

28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 4

29. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision. pp. 280–296. Springer (2022) 5

30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 7, 8

31. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023) 12

32. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 6

33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 2, 8, 9, 12

34. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 10, 11

35. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016) 7, 10

36. OpenAI: Chatgpt. https://openai.com/blog/chatgpt/ (2023) 2

37. OpenAI: Gpt-4 technical report (2023) 2

38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 5, 10

39. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023) 2, 3, 4, 5, 7, 12

40. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Zhang, L.K.T.: Detgpt: Detect what you need via reasoning. arXiv preprint arXiv:2305.14167 (2023) 2

41. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015) 7

42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 5

43. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. arXiv preprint arXiv:2311.03356 (2023) 2, 3, 4, 5, 11, 12

44. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019) 7

45. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 2, 3

46. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 2, 3

47. Wang, J., Zhang, P., Chu, T., Cao, Y., Zhou, Y., Wu, T., Wang, B., He, C., Lin, D.: V3det: Vast vocabulary visual detection dataset. arXiv preprint arXiv:2304.03752 (2023) 7

48. Wang, J., Meng, L., Weng, Z., He, B., Wu, Z., Jiang, Y.G.: To see is to believe: Prompting gpt-4v for better visual instruction tuning. arXiv preprint arXiv:2311.07574 (2023) 9

49. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning. pp. 23318–23340. PMLR (2022) 10

50. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023) 2, 4

51. Wang, W., Shi, M., Li, Q., Wang, W., Huang, Z., Xing, L., Chen, Z., Li, H., Zhu, X., Cao, Z., et al.: The all-seeing project: Towards panoptic visual recognition and understanding of the open world. arXiv preprint arXiv:2308.01907 (2023) 4

52. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems 36 (2024) 4, 10

53. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35, 24824–24837 (2022) 3

54. Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., Wang, L.: Grit: A generative region-to-text transformer for object understanding. arXiv preprint arXiv:2212.00280 (2022) 12

55. Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., Lu, H.: Universal instance perception as object discovery and retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15325–15336 (2023) 10, 11

56. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023) 8

57. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023) 2, 4

58. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023) 2, 4, 10, 11, 12, 13, 14

59. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016) 7, 10

60. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. arXiv preprint arXiv:2312.10032 (2023) 4

61. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022) 10

62. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Zhang, L., Li, C., et al.: Llava-grounding: Grounded visual chat with large multimodal models. arXiv preprint arXiv:2312.02949 (2023) 2, 3, 4, 8, 12

63. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Zhang, L., Li, C., et al.: Llava-grounding: Grounded visual chat with large multimodal models. arXiv preprint arXiv:2312.02949 (2023) 4

64. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023) 4, 5, 12, 13

65. Zhao, Y., Lin, Z., Zhou, D., Huang, Z., Feng, J., Kang, B.: Bubogpt: Enabling visual grounding in multi-modal llms. arXiv preprint arXiv:2307.08581 (2023) 4

66. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: European Conference on Computer Vision. pp. 350–368. Springer (2022) 2

67. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 2, 4, 8, 9

68. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020) 5, 13

69. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024) 14