GLaMM: Pixel Grounding Large Multimodal Model

Hanoona Rasheed[1*], Muhammad Maaz[1*], Sahal Shaji[1], Abdelrahman Shaker[1], Salman Khan[1,2]
Hisham Cholakkal[1], Rao M. Anwer[1,3], Eric Xing[1,4], Ming-Hsuan Yang[5,7], Fahad S. Khan[1,6]

[1]Mohamed bin Zayed University of AI, [2]Australian National University, [3]Aalto University
[4]Carnegie Mellon University, [5]University of California - Merced, [6]Linköping University, [7]Google Research
hanoona.bangalath@mbzuai.ac.ae, muhammad.maaz@mbzuai.ac.ae
https://mbzuai-oryx.github.io/groundingLMM, https://grounding-anything.com

Abstract

Large Multimodal Models (LMMs) extend Large Language Models to the vision domain. Initial LMMs used holistic images and text prompts to generate ungrounded textual responses. Recently, region-level LMMs have been used to generate visually grounded responses. However, they are limited to only referring to a single object category at a time, require users to specify the regions, or cannot offer dense pixel-wise object grounding. In this work, we present Grounding LMM (GLaMM), the first model that can generate natural language responses seamlessly intertwined with corresponding object segmentation masks. GLaMM not only grounds objects appearing in the conversations but is flexible enough to accept both textual and optional visual prompts (region of interest) as input. This empowers users to interact with the model at various levels of granularity, both in textual and visual domains. Due to the lack of standard benchmarks for the novel setting of visually Grounded Conversation Generation (GCG), we introduce a comprehensive evaluation protocol with our curated grounded conversations. Our proposed GCG task requires densely grounded concepts in natural scenes at a large-scale. To this end, we propose a densely annotated Grounding-anything Dataset (GranD) using our proposed automated annotation pipeline that encompasses 7.5M unique concepts grounded in a total of 810M regions available with segmentation masks. Besides GCG, GLaMM also performs effectively on several downstream tasks, e.g., referring expression segmentation, image and region-level captioning and vision-language conversations.

1. Introduction

Fueled by the generative AI wave, Large Multimodal Models (LMMs) have emerged as a pivotal advancement, bridging the gap between vision and language tasks [2]. Initial efforts like [5, 6, 17, 22, 41, 48] demonstrate effective textual responses based on input images. Although these models are sophisticated, they cannot still ground their responses in the visual context. Such grounding is crucial for advanced applications like detailed visual understanding, interactive embodied agents, and localized content manipulation. Recent efforts have started to address this limitation by enabling models to process user-defined regions specified via bounding boxes [4, 24, 27, 28, 46].

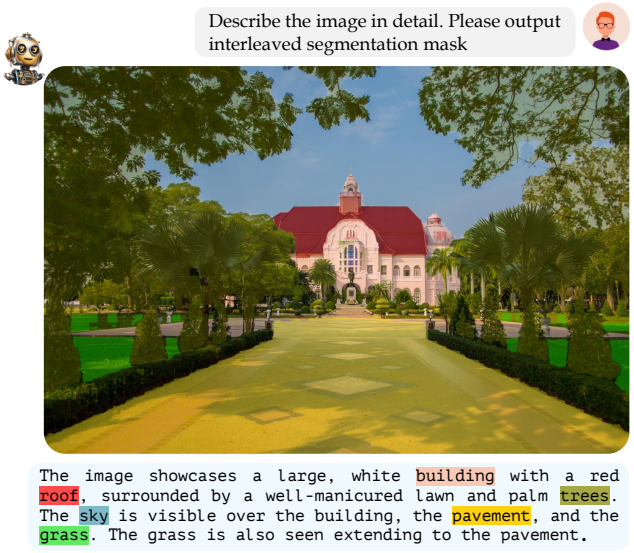

Figure 1. **Grounded Conversation Generation with GLaMM**. Our multimodal conversational model can provide natural language responses grounded at the pixel-level in the input image. Different levels of granularity are depicted in the output groundings, e.g., things (*building*, *tree*), stuff (*grass*, *sky*, *pavement*), and object parts (*roof* as a subpart of the building) alongside the object attributes (*white house*, *red roof*, *well-manicured lawn*) and object relationships (*grass extending to the pavement*, *sky over the building*). Existing LMMs, open-source (e.g., LLaVa, miniGPT4, Shikra, Kosmos-2) and closed-source (e.g., GPT4-V, Bard), do not offer pixel-level grounded conversational capability.

*Equally contributing first authors.

A few recent works have explored grounded text response generation [4, 16, 27, 47] but do not provide detailed *pixel-level* groundings. Parallel to these, efforts have been made in the referring segmentation literature to ground textual descriptions in natural images [16]. However, they are limited to grounding a single object and cannot engage in natural, coherent *conversations*, thereby restricting their practical applicability in interactive tasks that demand a deep understanding of both visual and textual content. To address these limitations of existing works, we introduce *Grounding LMM* (GLaMM), that simultaneously provides in-depth region understanding, pixel-level groundings, and conversational abilities through an end-to-end training approach (see Fig. 1 and Tab. 1).

To address the lack of benchmarks for visually grounded conversations, we introduce the novel task of *Grounded Conversation Generation* (GCG). The GCG task aims to produce natural language responses interleaved with object segmentation masks. This challenging task unifies several existing tasks in computer vision that are typically treated in isolation, i.e., referring expression segmentation, image and region-level captioning, phrase grounding, and vision-language conversations. Thereby, our unified model and proposed pretraining dataset can effectively transfer to several downstream tasks (referring expression segmentation, region-level captioning, image captioning, and conversational-style QA). We present GLaMM as the first model specifically designed for this challenging task. Unlike prior works, GLaMM can work with both textual and visual prompts and can generate visually grounded outputs, thus offering a versatile user experience.

Detailed region-level understanding requires the laborious process of collecting large-scale annotations for image regions. We propose an automated pipeline to annotate the large-scale *Grounding-anything Dataset* (GranD) to alleviate the manual labeling effort. Leveraging the automated pipeline with dedicated verification steps, GranD comprises 7.5M unique concepts anchored in 810M regions, each with a segmentation mask. Using state-of-the-art vision and language models, the dataset annotates SAM [13] images through a multi-level hierarchical scheme that enhances annotation quality. With 11M images, 84M referring expressions, and 33M grounded captions, GranD sets a new benchmark in comprehensiveness. In addition to the automatically generated dataset for the GCG, we provide the first high-quality dataset for grounded conversations obtained by revamping the existing manually annotated datasets [11, 29, 38] for GCG using GPT-4 [26] in-context learning. We refer to the high-quality dataset as GranD$_f$, denoting its suitability for fine-tuning.

Our work has three main contributions:

- We present GLaMM, the first model capable of generating natural language responses seamlessly integrated with object segmentation masks. Unlike existing models, GLaMM accommodates textual and visual prompts, facilitating enhanced multimodal user interaction.
- Recognizing the lack of standardized benchmarks for visually grounded conversations, we propose the new Grounded Conversation Generation (GCG) task. We also introduce a comprehensive evaluation protocol to measure the efficacy of models for GCG that unifies multiple isolated tasks, filling a significant gap in the literature.
- To facilitate model training and evaluation, we create Grounding-anything Dataset (GranD), a large-scale densely annotated dataset. Developed using an automatic annotation pipeline and verification criteria, it encompasses 7.5M unique concepts grounded in 810M regions. Additionally, we propose GranD$_f$, a high-quality dataset explicitly designed for the GCG task finetuning, by repurposing existing open-source datasets.

2. Related Work

LMMs provide a versatile interface for a diverse array of tasks, encompassing language and vision. Prominent models such as BLIP-2 [19], LLaVA [22], InstructBLIP [5] and MiniGPT-4 [48] first conduct image-text feature alignment followed by instruction tuning. Other representative works include Otter [17], mPLUG-Owl [41], LLaMa-Adapter [45], Video-ChatGPT [25], InternGPT [24]. However, these approaches lack region-specific understanding.

Recent works like Kosmos-2 [27], Shikra [4], GPT4RoI [46], VisionLLM [33], Ferret [42] and All-Seeing [34] aim to allow region-specific conversation. Some methods [4, 27, 34, 42] input location bins and bounding boxes with image data for region-level understanding, relying on the LLM exclusively for interpreting these regions. GPT4RoI advances this by using spatial boxes and RoI-aligned features for input and training on region-text pairs. BuboGPT [47] utilizes an off-the-shelf grounding model [23] and matches the groundings with the language response. In contrast, LISA [16] utilizes embeddings from the vision language model and the SAM [13] decoder to generate output segmentation masks. However, LISA cannot comprehend specific image regions or handle multiple instances.

To classify the LMM landscape, methods can be partitioned into four distinct categories (see Tab. 1 - separated via dotted lines). The first encompasses models effective in textual responses but lacking in region-specific capabilities [5, 6, 17, 22, 40, 41, 48]. In contrast, among models that handle region inputs or offer visual grounding, *three* more categories emerge. The first of these incorporates external vision modules [24, 47], and the next relies exclusively on LMMs for region understanding [4, 27, 28, 33]. The last category combines specialized vision modules with LMMs, trained end-to-end for a comprehensive understanding of regions [16, 34, 46]. Our approach belongs to the

Method	Image	Input / Output		Region Enc. / Dec.	Pixel-Wise Grounding	Multi-turn Conversation	End-End Model
		Region	Multi-Region				
MM-REACT (arXiv-23) [40]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✗
LLaVA (NeurIPS-23) [22]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
miniGPT4 (arXiv-23) [48]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
mPLUG-OWL (arXiv-23) [41]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
LLaMA-Adapter v2 (arXiv-23) [6]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
Otter (arXiv-23) [17]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✗	✓
Instruct-BLIP (arXiv-23) [5]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
InternGPT (arXiv-23) [24]	✓	✓ / ✗	✗ / ✗	✗ / ✗	✗	✓	✗
Bubo-GPT (arXiv-23) [47]	✓	✗ / ✓	✗ / ✓	✗ / ✗	✗	✓	✗
Vision-LLM (arXiv-23) [33]	✓	✗ / ✓	✗ / ✓	✗ / ✗	✗	✗	✓
Det-GPT (arXiv-23) [28]	✓	✓ / ✓	✓ / ✓	✗ / ✗	✗	✓	✓
Shikra (arXiv-23) [4]	✓	✓ / ✓	✗ / ✗	✗ / ✗	✗	✗	✓
Kosmos-2 (arXiv-23) [27]	✓	✓ / ✓	✓ / ✓	✗ / ✗	✗	✗	✓
GPT4RoI (arXiv-23) [46]	✓	✓ / ✗	✓ / ✗	✓ / ✗	✗	✓	✓
ASM (arXiv-23) [34]	✓	✓ / ✗	✗ / ✗	✓ / ✗	✗	✗	✓
LISA (arXiv-23) [16]	✓	✗ / ✓	✗ / ✗	✗ / ✓	✓	✗	✓
GLaMM (ours)	✓	✓ / ✓	✓ / ✓	✓ / ✓	✓	✓	✓

Table 1. **Comparison of recent Large Multimodal Models (LMMs)** emphasizing their capabilities for region-level understanding. The *Input* denotes models that can process regions defined by users via bounding boxes, with *Multi-Region* indicating models that can handle multiple such regions. The *Output* represents models capable of delivering grounded responses. While some methods employ external vision modules for region understanding, others rely solely on the LMM, which may result in imprecise localization. However, a few integrate specialized vision modules and LMMs, as indicated by the *Region Enc./Dec.*. The *End-End Model* distinction separates models that leverage LMMs for region understanding from those employing external modules. *Pixel-wise Grounding* highlights models that can respond with segmentation masks, and *Multi-turn Conversation* represents models that can hold an interactive dialogue with the user. Among these, our proposed *GLaMM* stands out by offering comprehensive region understanding, pixel-wise grounding in its responses, conversational capabilities, and an end-to-end training approach.

last category and distinctly offers pixel-level grounding together with multi-turn conversations and the flexibility to operate on both input images and specific regions. Further, we provide large-scale instance-level grounded visual understanding dataset that allows generalizability of GLaMM to multiple vision-language tasks.

3. Method

Existing Large Multimodal Models (LMMs) either generate ungrounded text or are restricted by limitations such as single-object grounding, user-specified region inputs, or the lack of dense pixel-level object grounding (see Tab. 1). Our Grounding LMM (GLaMM) aims to overcome these limitations by generating natural language responses seamlessly integrated with object segmentation masks. This enables a visually grounded human-machine conversation.

3.1. GLaMM Architecture

GLaMM consists of five core components: i) Global Image Encoder, ii) Region Encoder, iii) LLM, iv) Grounding Image Encoder, and v) Pixel Decoder. These components are cohesively designed to handle both textual and optional visual prompts (image level and region), allowing for interaction at multiple levels of granularity and generating grounded text responses (Fig. 2). These blocks together enable scene-level, region-level, and pixel-level grounding, as explained next. Refer Appendix A.2 for training details.

Scene-Level Understanding: To achieve a holistic understanding of the scene, we employ ViT-H/14 CLIP [30] as our *global image encoder* ($\mathcal{I}$), in conjunction with a vicuna-based LLM ($\mathcal{L}$) and a vision-to-language (V-L) projection layer (f). Specifically, given an image x_{img} and a text instruction x_t, the image is first encoded into a feature vector $I_x = \mathcal{I}(x_{\text{img}}) \in \mathbb{R}^{D_v}$ and projected to language space $f(I_x) \in \mathbb{R}^{D_t}$. The LLM then integrates both the projected image features and the text instruction to generate output y_t:

$$y_t = \mathcal{L}\Big(f(I_x), x_t\Big).$$

This maps image features to language space, enabling GLaMM to offer holistic scene understanding, achieved through specific prompts like, "`The <image> provides an overview of the image. Could you please give me a detailed description of the image?`" The `<image>` token is replaced with 256 tokens from the CLIP global image encoder.

Region-Level Understanding: Building on the shortcomings of existing models that can handle only image-level visual inputs, and in alignment with recent work [46], the *re-*

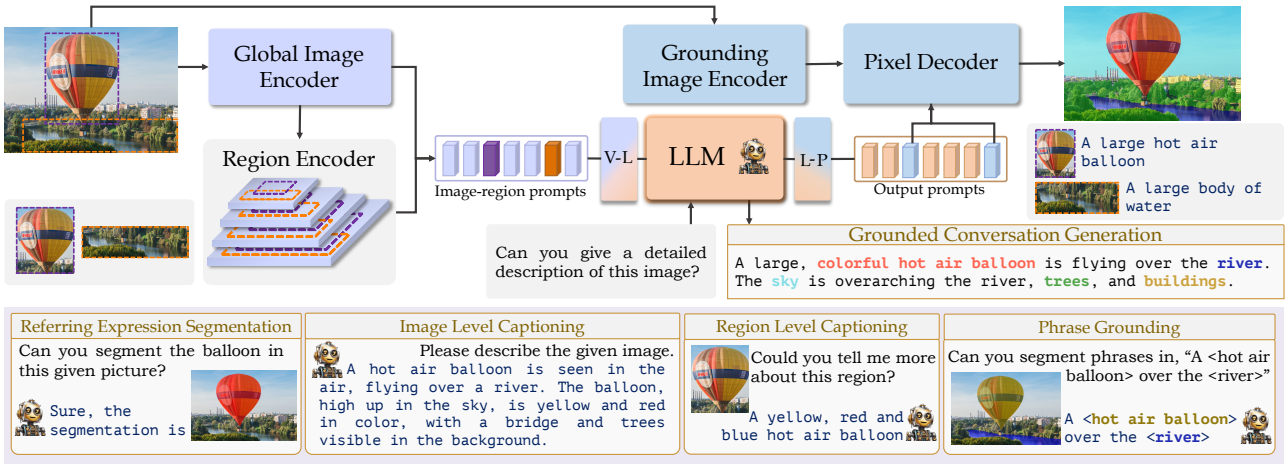

Figure 2. **GLaMM's architecture.** The figure illustrates our model architecture, showcasing its ability to offer scene-level understanding, region-level interpretation, and pixel-level grounding. **Top:** The core components of GLaMM, including the global image encoder, region encoder, LLM, grounding image encoder, and pixel decoder, are cohesively tailored for vision-language tasks across different granularities. The vision-to-language (V-L) projection layer efficiently maps image features into the language domain, and the pixel decoder utilizes the language-to-prompt (L-P) projection layer, transforming text embeddings related to segmentation into the decoder space. A major feature of GLaMM is its ability to perform our newly introduced *Grounded Conversation Generation (GCG)* task. This highlights the model's capability to anchor specific phrases to corresponding segmentation masks in the image. **Bottom**: The diverse downstream applications of GLaMM, including referring expression segmentation, region-level captioning, image-level captioning, and phrase grounding.

gion encoder ($\mathcal{R}$) extends the model's capability to interpret and interact with user-specified regions in an image. This component constructs a hierarchical feature pyramid from four selected CLIP global image encoder layers, followed by RoIAlign [8] to generate a 14x14 feature map. Combining these features yields a unified region-of-interest (RoI) representation. To facilitate region-targeted responses from GLaMM, we augment the existing vocabulary with a specialized token <bbox>. This is integrated into a prompt like, "The <image> provides an overview of the image. Can you provide a detailed description of the region <bbox>?". Here the <bbox> token is replaced with the RoI extracted features.

For the region-level understanding, alongside the global image features I_x, we also take user-specified regions r as inputs, encoded as $R_x = \mathcal{R}(I_x, r)$, followed by projection to language space through the same V-L projection layer f employed in scene-level understanding. We augment the text instruction x_t by replacing <bbox> tokens with the corresponding region features to obtain $x'_t = [x_t \leftarrow f(R_x)]$. The LLM then generates the output y_t as,

$$y_t = \mathcal{L}\Big(f(I_x), x'_t\Big).$$

Pixel-Level Grounding: Utilizing the *grounding image encoder* denoted as $\mathcal{V}$ and the *pixel decoder* represented as $\mathcal{P}$, GLaMM facilitates fine-grained pixel-level object grounding, allowing it to ground its responses visually. We instantiate $\mathcal{V}$ with a pretrained SAM encoder [13] and design $\mathcal{P}$ based on a SAM decoder-like architecture. To activate the pixel-level grounding, our model's vocabulary is augmented with a specialized token, <SEG>. Prompts, such as "Please segment the 'man in red' in the given image," trigger the model to generate responses with corresponding <SEG> tokens. A *language-to-prompt (L-P)* projection layer (g) transforms the last-layer embeddings corresponding to <SEG> tokens (l_{seg}) into the decoder's feature space. Subsequently, $\mathcal{P}$ produces binary segmentation masks M,

$$M = \mathcal{P}\Big(g(l_{seg}), \mathcal{V}(x_{\text{img}})\Big), \; s.t., M_i \in \{0, 1\}.$$

Using an end-to-end training approach, GLaMM excels in region understanding, pixel-level grounding, and conversational capabilities. However, due to the lack of standard benchmarks for the novel setting of generating visually grounded detailed conversations, we introduce a novel task, *Grounded Conversation Generation* (GCG), and a comprehensive evaluation protocol as explained next.

3.2. Grounded Conversation Generation (GCG)

The objective of the GCG task is to construct image-level captions with specific phrases directly tied to corresponding segmentation masks in the image. For example, "<A man> and <a boy> sit on <a bench> next to <an old white car>.", shown in Fig. 3 (left), features how each bracketed phrase (highlighted in the image) is anchored to a unique image segmentation mask. This creates a densely annotated caption that aligns textual descriptions with visual regions, enriching the image's contextual interpretation.

GCG Output Representation: A sample prompt for querying the model in this task is: "Could you please

Figure 3. **Qualitative results of GLaMM on grounded conversation generation (GCG).** Given user queries, the LMM generates textual responses and grounds objects, object parts, attributes, and phrases using pixel-level masks, showing its detailed understanding.

`give me a detailed description of the image? Please respond with interleaved segmentation masks for the corresponding parts of the answer.`" The model generates a detailed caption along with interleaved segmentation masks, employing the format "`<p>A man</p><SEG> and <p>a boy</p><SEG> sit on <p>a bench</p><SEG> next to <p>an old white car</p><SEG>.`" We use special tokens, namely `<p>`, `</p>` and `<SEG>`, to delineate the start and end of each phrase and its corresponding region mask, respectively.

Our GranD dataset is meticulously constructed using a stage-wise annotation pipeline, capturing annotations that range from fine-grained specifics to high-level context. This enables the automatic generation of densely annotated captions well-suited for the GCG task, thereby significantly facilitating GLaMM's training for this task. Some qualitative results of our model on the GCG task are shown in Fig. 3.

Evaluation Criteria: We introduce a benchmarking suite for GCG, with a validation set of 2.5K images and a test set of 5K images. Four key aspects are evaluated: i) generated dense caption quality, ii) mask-to-phrase correspondence accuracy, iii) generated mask quality, and iv) region-specific grounding ability. Metrics include METEOR and CIDEr for captions, class-agnostic mask AP for grounding, mask IoU for segmentation, and mask recall for region-specific grounding (refer to Appendix A.1 for details).

Having delineated the architecture of GLaMM and the intricacies of the GCG task, it becomes imperative to address the scarcity of large-scale annotated data for region-level understanding. We next focus on devising a new, densely annotated dataset to optimize the model's performance and overcome this data limitation.

4. Data Annotation Pipeline

We introduce our automated annotation pipeline used to create the Grounding-anything Dataset (GranD). GranD is a comprehensive, multi-purpose image-text dataset offering a range of contextual information, from fine-grained to high-level details. It aims to overcome challenges in image understanding and dense pixel-level grounding, thereby expanding capabilities of visual instruction tuning in LMMs.

The pipeline contains four distinct levels (see Fig. 4). **i)** *Level-1* focuses on object localization and provides semantic labels, segmentation masks, attributes, and depth information. **ii)** *Level-2* defines relationships between detected objects. **iii)** *Level-3* organizes information from the first two levels into a hierarchical scene graph, used to generate dense captions using LLM with in-context examples. **iv)** *Level-4* offers enriched contextual information for a deeper understanding of the scene, going beyond what's observed (e.g., historical information of a landmark). Please refer to Appendix A.4 for pipeline implementation details.

4.1. Object Localization and Attributes (Level-1)

In level-1, the focus is on detailed object identification within images. First, object-bounding boxes are identified using multiple SoTA object detection models. Class-agnostic NMS is applied to each model to filter out false positives. After this step, bounding boxes from different models are compared using IoU, with a bounding box retained as an object only if detected by at least two other detection models. We also generate attributes for each filtered object using region-based vision-language models and incorporate depth information to contextualize each object's relative position within the scene.

4.2. Relationships and Landmarks (Level-2)

In level-2, multiple short textual descriptions of the overall scene are generated. Phrases extracted from these descriptions are grounded to specific objects in level-1 to form relationships. These relationships articulate connections between multiple objects or define an object's role within the scene. Further, each scene is assigned a landmark category that includes a primary and a more specific sub-category (see Tab. 7 in Appendix A.4.1).

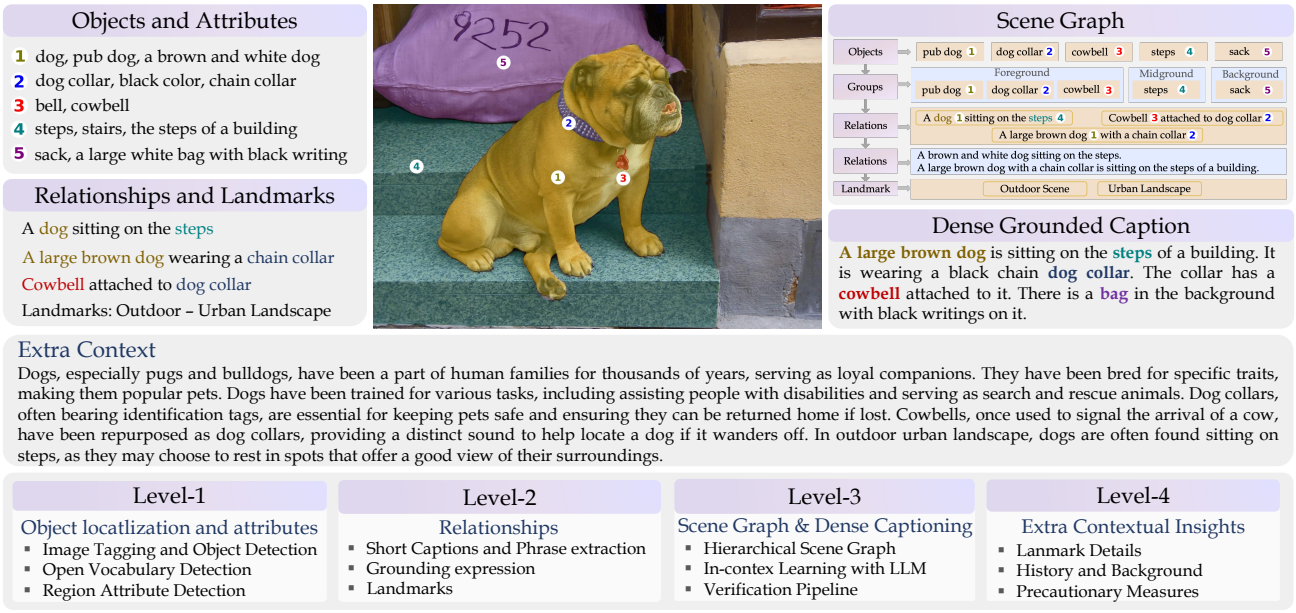

Figure 4. **Automatic Annotation Pipeline of the Grounding-anything Dataset (GranD).** Comprising four levels, this pipeline plays a pivotal role in generating GranD's 7.5M unique concepts grounded in 810M regions. level-1 details objects and attributes, level-2 includes short captions and relational markers, level-3 builds a scene graph, hierarchically organizing information from earlier levels to facilitate LLM for grounded dense captions, level-4 provides additional historical and societal context for a richer visual understanding.

4.3. Scene Graph and Dense Captioning (Level-3)

In level-3, object attributes and labels from level-1 are combined with the relationships and phrases obtained from level-2 to form a hierarchical scene graph. This structured data serves as a query for LLM to generate dense image captions. To provide additional context, depth values and bounding box coordinates are used to assign each object to specific spatial layers within the scene, such as *immediate foreground*, *foreground*, *midground*, or *background*. Additionally, short scene-level captions are incorporated into the scene graph to enhance LLMs' contextual understanding.

Dense Captioning Verification: To enhance the fidelity of the LLM-generated dense captions, we implement an automatic verification pipeline using chain-of-thoughts prompting. This pipeline produces a checklist of objects derived from the generated dense caption expected to be present in the image. The associated caption is flagged as inaccurate if any object specified in the checklist is absent from the scene graph. Such captions are then regenerated, incorporating feedback from the initial assessment.

4.4. Extra Contextual Insights (Level-4)

Level-4 builds on the scene graph from level-3 to obtain a more detailed visual understanding. we query LLM to extract extended contextual insights beyond basic object identification and relationships, including details about the landmarks, historical context, guidelines for interacting with the scene, and even predictive elements about future events. To facilitate this, we prompt LLM with in-context examples.

Dataset	Images	Regions	Concepts	Tokens	Captions†
COCO [20]	0.1M	0.9M	80	-	-
LVIS [7]	0.1M	1.5M	1,203	-	-
Objects365 [31]	0.6M	10.1M	365	-	-
Open Images [15]	1.5M	14.8M	600	-	-
BigDetection [3]	3.5M	36.0M	600	-	-
V3Det [32]	0.2M	1.5M	13,029	-	-
VG [14]	0.1M	0.3M	18,136	51.2M	-
SA-1B [13]	11M	1.1B	-	-	-
AS-1B [34]	11M	1.2B	3.5M	132.2B	-
GranD (Ours)	11M	810M	7.5M	5.0B	33M

Table 2. **GranD versus existing datasets.** GranD uniquely provides three †grounded captions per image with segmentation masks for every region. AS-1B is shaded to denote its concurrent, non-public status at the time of this publication.

Utilizing our automated annotation pipeline, we annotate a corpus of 11M SAM images [13], which are inherently diverse, high-resolution, and privacy-compliant. The resulting dataset comprises 810M regions, each associated with a segmentation mask, and includes 7.5M unique concepts. Further, the dataset features 84M referring expressions, 22M grounded short captions, and 11M densely grounded captions. To our knowledge, this is the first dataset of this scale generated entirely through an automated annotation pipeline (see Tab. 2 for details and Fig. 15 in Appendix for dataset sample visualizations).

4.5. Building GranD$_f$ for GCG

Motivated by the need for higher-quality data in fine-tuning stage, we introduce GranD$_f$. It contains 214K image-grounded text pairs with 2.5K validation and 5K test sam-

Model	Validation Set					Test Set				
	M	C	AP50	mIoU	Recall	M	C	AP50	mIoU	Recall
BuboGPT [47]	**17.2**	3.6	19.1	54.0	29.4	**17.1**	3.5	17.3	54.1	27.0
Kosmos-2 [27]	16.1	27.6	17.1	55.6	28.3	15.8	27.2	17.2	56.8	29.0
LISA* [16]	13.0	33.9	25.2	62.0	36.3	12.9	32.2	24.8	61.7	35.5
GLaMM†	15.2	43.1	28.9	65.8	39.6	14.6	37.9	27.2	64.6	38.0
GLaMM	16.2	**47.2**	**30.8**	**66.3**	**41.8**	15.8	**43.5**	**29.2**	**65.6**	**40.8**

Table 3. **Performance on GCG Task**: Metrics include METEOR (M), CIDEr (C), AP50, mIoU, and Mask Recall. LISA* denotes LISA adapted for GCG. GLaMM† denotes training excluding 1K human annotated images. GLaMM shows better performance.

Method	**refCOCO**			**refCOCO+**			**refCOCOg**	
	val	testA	testB	val	testA	testB	val(U)	test(U)
CRIS [36]	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT [39]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES [21]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [49]	-	-	-	-	-	-	64.6	-
SEEM [50]	-	-	-	-	-	-	65.7	-
LISA-7B [16]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
GLaMM	**79.5**	**83.2**	**76.9**	**72.6**	**78.7**	**64.6**	**74.2**	**74.9**

Table 4. **Qualitative Assessment of GLaMM in Referring-Expression Segmentation**: Performance across refCOCO, refCOCO+, and refCOCOg in generating accurate segmentation masks based on text-based referring expressions surpasses that of closely related work, including LISA which is specifically designed for this task.

ples. GranD$_f$ comprises two primary components: one subset is manually annotated, and the other subset is derived by re-purposing existing open-source datasets.

We extend open-source datasets—namely Flickr-30K [29], RefCOCOg [11], and PSG [38] by generating compatible GCG annotations. For RefCOCOg, we use the dataset's referring expressions and their connected masks. These expressions offer concise descriptions of distinct objects in the image. With the aid of GPT-4, we seamlessly blend these referring expressions with contextual information from COCO captions, crafting detailed yet accurate grounded captions while preserving the original referring expressions. This ensures zero error in matching phrases with their corresponding segmentation masks. This technique yields approximately 24K GCG samples. For PSG, we leverage the dataset's triplet structures, which describe relations between two objects in a scene. These triplets are integrated with COCO captions using GPT-4, resulting in densely annotated captions that can be mapped to segmentation masks. This gives us around 31K additional GCG samples. For Flickr-30K, we use the 158K Flickr captions and their referring expressions alongside associated bounding boxes. These boxes are then accurately segmented using HQ-SAM [12].

In addition, we contribute a minor, high-quality manual annotation set to benchmark the GCG task. Using GranD's automatic annotations as a base, annotators refine referring expressions to match SAM GT masks, yielding around 1000 focused samples for training and 1000 for evaluation (refer to Appendix D and Fig. 14 in Appendix for designed prompts and dataset visualizations).

5. Experiments

We perform quantitative evaluations of GLaMM on six benchmarks: i) Grounded Conversation Generation (GCG), ii) referring-expression segmentation, iii) region-level captioning, iv) image-level captioning, v) conversational-style question answering and vi) phrase grounding. We present the first four benchmarks next, and the remaining are discussed in Appendix B. **Grounded Conversation Generation (GCG).** We pretrain GLaMM on GranD dataset followed by fine-tuning on the GranD$_f$ dataset. The results are presented in Tab. 3 on both validation and test splits of the GranD$_f$ dataset (refer to Sec. 3.2 and Sec. 4.5 for details). GLaMM shows improved performance compared to baseline methods. Pretrained models for BuboGPT and Kosmos-2 are sourced from official releases, and LISA is adapted and trained on the GranD$_f$ dataset for the GCG task. GLaMM† denotes the variant trained on GranD$_f$ dataset excluding the 1000 human-annotated images. Qualitative results are shown in Fig. 3 and supplementary Fig. 7.
Referring Expression Segmentation. In this task, the model processes an image and a text-based referring expression to output a segmentation mask. The prompt used is, "`Please segment the <referring expression> in the image.`" The model responds with "`Sure, it is <SEG>.`", where the `<SEG>` token is decoded to obtain the mask. We achieve better results over recent works like LISA on the refCOCO, refCOCO+, and refCOCOg validation and test sets in Tab. 4. This demonstrates the efficacy of our GranD dataset, offering the model extensive concept vocabulary during pre-training (refer to Fig. 5 (middle) and supplementary Fig. 8 for qualitative results).
Region Level Captioning. In this task, models generate

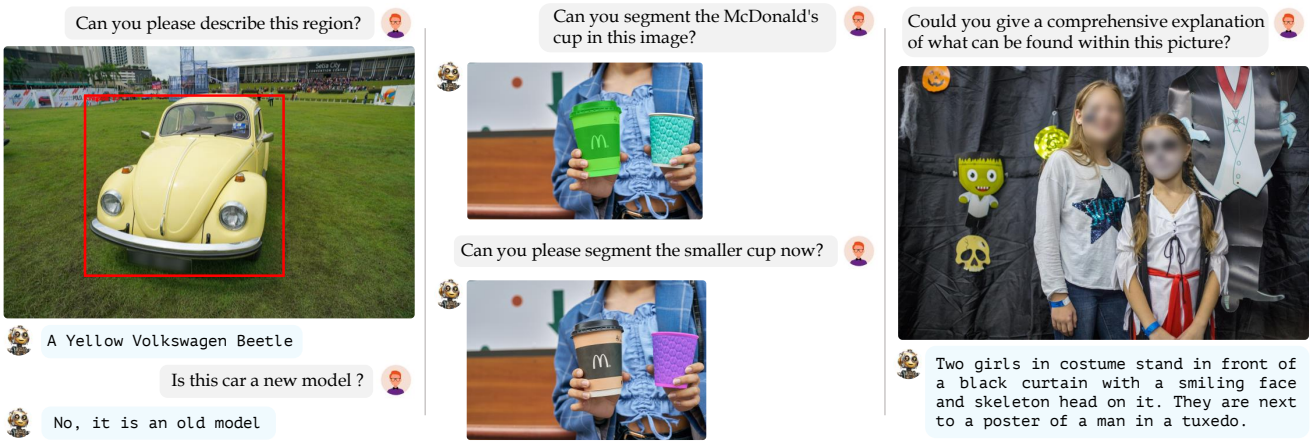

Figure 5. **Qualitative results of GLaMM's performance across downstream tasks.** The figure showcases examples from three tasks: region-level understanding (left), referring-expression segmentation (center), and image-level captioning (right), demonstrating its capabilities in offering in-depth region understanding, pixel-level groundings, and conversational abilities through an end-to-end training approach.

Model	refCOCOg		Visual Genome	
	METEOR	CIDEr	METEOR	CIDEr
GRIT [37]	15.2	71.6	17.1	142
Kosmos-2 [27]	14.1	62.3	-	-
GPT4RoI [46]	-	-	17.4	145.2
GLaMM (ZS)	**15.7**	**104.0**	**17.0**	**127.0**
GLaMM (FT)	**16.2**	**106.0**	**19.7**	**180.5**

Table 5. **Performance of GLaMM in Region-Level Captioning**: Metrics include METEOR and CIDEr scores, assessed on Visual Genome and refCOCOg Datasets, exhibiting competitive results.

Model	NoCap		Flickr30k	
	CIDEr	SPICE	CIDEr	SPICE
VinVLM [44]	95.5	13.5	-	-
LEMON [9]	106.8	14.1	-	-
SimVLM [35]	110.3	14.5	-	-
CoCa [43]	120.6	15.5	-	-
BLIP [18]	113.2	14.7	-	-
BLIP-2 [19]	121.6	15.8	-	-
InstructBLIP [5]	**123.1**	-	82.8	-
Shikra-13B [4]	-	-	73.9	-
Kosmos-1 [10]	-	-	67.1	14.5
Kosmos-2 [27]	-	-	66.7	-
GLaMM	106.8	**15.8**	**95.3**	**18.8**

Table 6. **Performance of GLaMM in Zero-Shot Image Captioning**: Assessed on Flickr30k and NoCap datasets, showing favorable results compared to recent models in the field.

region-specific captions given an image, a user-specified region via a bounding box and related text. We utilize a prompt like, "`Can you provide a detailed description of the region <bbox>?`", to instruct the model for this task, where the special token `<bbox>` is replaced with the actual region representations. We evaluate GLaMM on Visual Genome and refCOCOg, using METEOR and CIDEr metrics with results presented in Tab. 5. GLaMM shows improved results over GRiT and GPT4RoI after fine-tuning and demonstrates robust zero-shot performance, highlighting the significance of GranD's region-text pairs (refer to Fig.5 (left) and supplementary Fig. 9 for qualitative results).

Image Level Captioning. For this task, GLaMM responds to queries like, "`Could you please give me a detailed description of the image?`" with a textual description. We evaluate GLaMM's zero-shot performance on Flickr30k [29] and NoCap [1] datasets, with Tab. 6 showing its favorable performance against recent image captioning models and other LMMs (refer to Fig. 5 (right) and supplementary Fig. 10 for qualitative results).

6. Conclusion

We introduce GLaMM, the first model capable of generating natural language responses intertwined with object segmentation masks, allowing for enhanced multimodal user interactions. Recognizing the lack of standardized benchmarks for visually grounded conversations, we introduce the novel task of Grounded Conversation Generation and establish a comprehensive evaluation protocol. To facilitate research and model development, we create the Grounding-anything Dataset (GranD), a large-scale, densely annotated dataset with 7.5 million unique concepts grounded in 810 million regions. Our automated annotation pipeline ensures the reliability and scalability of this dataset used for our model. In addition to these contributions, we curated a dataset specifically tailored for the GCG task (GranD$_f$) by leveraging existing open-source datasets, establishing a high-quality fine-tuning dataset to develop visually grounded conversations. Our model performs well on downstream tasks besides GCG, including region and image captioning, referring segmentation, and vision-language conversations.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 8

[2] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook. *arXiv:2307.13721*, 2023. 1

[3] Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Mu Li, and Xiangyang Xue. Bigdetection: A large-scale benchmark for improved object detector pre-training. In *CVPR*, 2022. 6

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023. 1, 2, 3, 8

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*, 2023. 1, 2, 3, 8

[6] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*, 2023. 1, 2, 3

[7] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 6

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 4

[9] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *CVPR*, 2022. 8

[10] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv:2302.14045*, 2023. 8

[11] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2, 7

[12] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv:2306.01567*, 2023. 7

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 4, 6

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 6

[15] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 6

[16] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv:2308.00692*, 2023. 2, 3, 7

[17] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv:2305.03726*, 2023. 1, 2, 3

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 8

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 8

[20] 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 *ECCV*, 2014. 6

[21] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, 2023. 7

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 3

[23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*, 2023. 2

[24] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, Kunchang Li, Zhe Chen, Xue Yang, Xizhou Zhu, Yali Wang, Limin Wang, Ping Luo, Jifeng Dai, and Yu Qiao. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv:2305.05662*, 2023. 1, 2, 3

[25] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*, 2023. 2

[26] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 2

[27] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*, 2023. 1, 2, 3, 7, 8

[28] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, and Tong Zhang. Detgpt: Detect what you need via reasoning. *arXiv:2305.14167*, 2023. 1, 2, 3

[29] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2, 7, 8

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[31] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 6

[32] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. *arXiv:2304.03752*, 2023. 6

[33] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv:2305.11175*, 2023. 2, 3

[34] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv:2308.01907*, 2023. 2, 3, 6

[35] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv:2108.10904*, 2021. 8

[36] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 7

[37] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv:2212.00280*, 2022. 8

[38] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, 2022. 2, 7

[39] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 7

[40] 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:2303.11381*, 2023. 2, 3

[41] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2305.03726*, 2023. 1, 2, 3

[42] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv:2310.07704*, 2023. 2

[43] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv:2205.01917*, 2022. 8

[44] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *arXiv:2101.00529*, 2021. 8

[45] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv:2303.16199*, 2023. 2

[46] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv:2307.03601*, 2023. 1, 2, 3, 8

[47] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv:2307.08581*, 2023. 2, 3, 7

[48] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 1, 2, 3

[49] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, 2023. 7

[50] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *NeurIPS*, 2023. 7