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

Abstract

Large Multimodal Models (LMMs) extend Large Language Models to the vision domain. Initial efforts towards LMMs used holistic images and text prompts to generate ungrounded textual responses. Very recently, region-level LMMs have been used to generate visually grounded responses. However, they are limited to only referring a single object category at a time, require users to specify the regions in inputs, 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 generating visually grounded detailed conversations, we introduce a comprehensive evaluation protocol with our curated grounded conversations. Our proposed Grounded Conversation Generation (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.

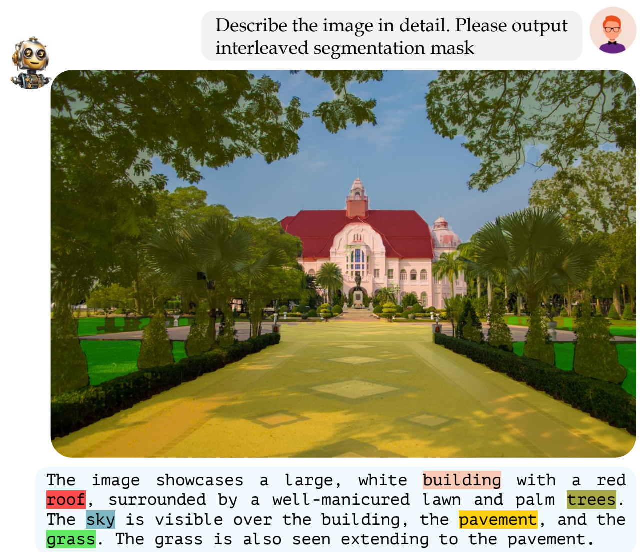


Figure 1. **Grounded Conversation Generation with GLaMM**. Our multimodal conversational model can provide natural language responses grounded at the pixel-level in the input image. Different levels of granularity are depicted in the output groundings, e.g., things (*building*, *tree*), stuff (*grass*, *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) as well as closed-source (e.g., GPT4-V, Bard), do not offer pixel-level grounded conversational capability.

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 versions like LLaVa [33], miniGPT4 [66], Otter [25], InstructBLIP [8], LLaMA-Adapter v2 [11] and mPLUG-OWL [57] demonstrate effective textual responses based on input images. Although these models are sophisticated, they still lack the ability to ground their responses in the visual context. Such grounding is crucial for advanced appli-

*Equally contributing first authors.

cations 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 [7, 34, 39, 40, 62].

A few recent works have explored grounded text response generation [7, 24, 39, 64], however, they 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 [24]. 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 (see Tab. 1), 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 (Fig. 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. To alleviate the manual labelling effort, we propose an automated pipeline to annotate the large-scale *Grounding-anything Dataset* (GranD). Leveraging the automated pipeline with dedicated verification steps, GranD comprises 7.5M unique concepts anchored in a total of 810M regions, each with a segmentation mask. Using state-of-the-art vision and language models, the dataset annotates SAM [21] images through a multi-level hierarchical scheme that enhances annotation quality. With 11M images and attributes like 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 [19, 41, 54] for GCG using GPT-4 [37] in-context learning. We refer to the high-quality dataset as GranD$_f$, denoting its suitability for fine-tuning, and the large-scale automatically generated data as GranD$_p$. GranD$_f$, together with GranD$_p$ is used to train GLaMM in pretraining-finetuning steps.

In summary, our work has three main contributions:

- *Introduction of GLaMM*: We present the Grounding Large Multimodal Model (GLaMM), the first-of-its-kind model capable of generating natural language responses that are seamlessly integrated with object segmentation masks. Unlike existing models, GLaMM accommodates both textual and optional visual prompts, facilitating enhanced multimodal user interaction.
- *Novel Task and Evaluation Metrics*: Recognizing the lack of standardized benchmarks for visually grounded conversations, we propose a new task of Grounded Conversation Generation (GCG). Alongside, we introduce a comprehensive evaluation protocol to measure the efficacy of models in this novel setting that unifies multiple isolated tasks, filling a significant gap in the literature.
- *Grounding-anything Dataset (GranD)*: To facilitate model training and evaluation, we create 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 re-purposing existing open-source datasets.

2. Related Work

Large Multimodal Models (LMMs) have effectively served as a versatile interface for a diverse array of tasks, encompassing language, vision and other modalities. Prominent models such as LLaVA [33], InstructBLIP [8] and MiniGPT-4 [66] first conduct image-text feature alignment followed by instruction tuning. BLIP-2 proposes encoding image features with a visual encoder, which are then fed into the LLM alongside text embeddings. Other representative works include Otter [25], mPLUG-Owl [57], LLaMa-Adapter [61], VideoChatGPT [35], InternGPT [34]. However, these approaches do not offer region-specific conversations or visual grounding capabilities.

Region Aware LMMs: Recent works like KOSMOS-2 [39], Shikra [7], GPT4RoI [62], VisionLLM [49] and All-Seeing Model [50] aim to allow region-specific conversation. Kosmos2 [39] and Shikra [7] give input regions as location bins and bounding-box coordinates along with the image representation for region-level understanding and depend solely on the LLM for region understanding. GPT4RoI [62] goes a step further and introduces spatial boxes with RoI-aligned features as input and trains the model on region-text pairs.

Grounding LMMs: The grounding capability of multi-

Method	Image	Input / Output		Region Enc. / Dec.	Pixel-Wise Grounding	Multi-turn Conversation	End-End Model
		Region	Multi-Region				
MM-REACT (arXiv-23) [56]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✗
LLaVA (NeurIPS-23) [33]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
miniGPT4 (arXiv-23) [66]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
mPLUG-OWL (arXiv-23) [57]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
LLaMA-Adapter v2 (arXiv-23) [11]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
Otter (arXiv-23) [25]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✗	✓
Instruct-BLIP (arXiv-23) [8]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
InternGPT (arXiv-23) [34]	✓	✓ / ✗	✗ / ✗	✗ / ✗	✗	✓	✗
Bubo-GPT (arXiv-23) [64]	✓	✗ / ✓	✗ / ✓	✗ / ✗	✗	✓	✗
Vision-LLM (arXiv-23) [49]	✓	✗ / ✓	✗ / ✓	✗ / ✗	✗	✗	✓
Det-GPT (arXiv-23) [40]	✓	✓ / ✓	✓ / ✓	✗ / ✗	✗	✓	✓
Shikra (arXiv-23) [7]	✓	✓ / ✓	✗ / ✗	✗ / ✗	✗	✗	✓
Kosmos-2 (arXiv-23) [39]	✓	✓ / ✓	✓ / ✓	✗ / ✗	✗	✗	✓
GPT4RoI (arXiv-23) [62]	✓	✓ / ✗	✓ / ✗	✓ / ✗	✗	✓	✓
ASM (arXiv-23) [50]	✓	✓ / ✗	✗ / ✗	✓ / ✗	✗	✗	✓
LISA (arXiv-23) [24]	✓	✗ / ✓	✗ / ✗	✗ / ✓	✓	✗	✓
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. A few, however, integrate both 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.

modal models enable the models to respond with visual answers like segmentation masks, and bounding boxes along with language reasoning to support a range of tasks including referring expression comprehension, segmentation and phrase grounding. Existing works like KOSMOS-2 [39], Ferret [58], All-Seeing Model [50], LISA [24], BuboGPT [64], Shikra [7] demonstrate that LLMs are capable of fine-grained image comprehension and open-world referring and grounding. However, models like Kosmos-2 [39], Shikra [7], All-Seeing [50] and Ferret [58] rely extensively on the capability of the Language Model to generate the groundings. BuboGPT [64] utilizes an off-the-shelf grounding model and matches the groundings on the generated language response. On the contrary, LISA [24] utilizes embeddings from the vision language model and the SAM [21] decoder to generate output segmentation masks. However, LISA cannot comprehend specific regions within the image or handle multiple instances.

To classify the LMM landscape, methods can be partitioned into four distinct categories, as outlined in Tab. 1 (separated via dotted lines). The first encompasses models effective in textual responses but lacking in region-specific capabilities [8, 11, 25, 32, 56, 57, 66]. In contrast, among models that handle region inputs or offer visual grounding, *three* more categories emerge. The first of these incorporates external vision modules [34, 64], the next relies exclusively on LMMs for region understanding [7, 39, 40, 49], and the last category combines specialized vision modules with LMMs, trained end-to-end for a comprehensive understanding of regions [24, 50, 62]. Our approach belongs to the last category is distinct in terms of offering 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 that are seamlessly integrated with object segmentation masks. This enables a visually grounded human-machine conversation.

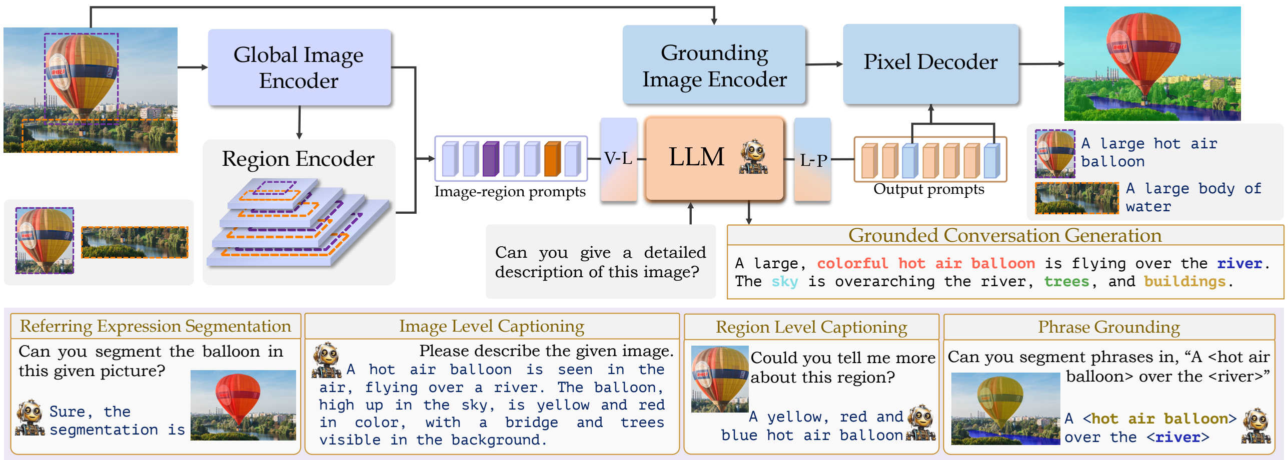


Figure 2. **GLaMM's architecture.** The figure illustrates our model architecture showcasing its ability to offer scene-level understanding, region-level interpretation, and pixel-level grounding. **Top:** The core components of GLaMM including the global image encoder, region encoder, LLM, grounding 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.

3.1. GLaMM Architecture

GLaMM consists of five core components to achieve visually grounded conversations: 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 of interest), allowing for interaction at multiple levels of granularity, and generating grounded text responses (see Fig. 2). The above five blocks together enable scene-level understanding, region-level understanding and pixel-level grounding, as explained next.

Scene-Level Understanding: To achieve a holistic understanding of the scene, we employ ViT-H/14 CLIP [42] 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 setup maps image features into the language space, enabling GLaMM to offer holistic scene understanding. This is achieved through specific prompts like, "`The <image> provides an overview of the image. Could you please give me a detailed description of the image?`" Here, the <image> token is replaced with 256 tokens generated by 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 [62], the *region 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 layers of the CLIP global image encoder, followed by RoIAlign [13] 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 will be 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 $\mathcal{P}$,

Figure 3. **Qualitative results of GLaMM on grounded conversation generation (GCG).** Given user queries, the LMM not only generates textual responses but also grounds objects, object parts, attributes and phrases using pixel-level masks showing its detailed understanding.

GLaMM facilitates fine-grained pixel-level object grounding, allowing it to visually ground its responses. We instantiate $\mathcal{V}$ with a pretrained SAM encoder [21], 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,

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

GLaMM stands out by offering comprehensive region understanding, pixel-level grounding in its responses, conversational capabilities, and an end-to-end training approach. This allows the model to be used for downstream tasks like image-level captioning, region-level captioning and referring expression segmentation. 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. By introducing the GCG task, we bridge the gap between textual and visual understanding, thereby enhancing the model's ability for fine-grained visual grounding alongside natural language captioning.

GCG Output Representation: A sample prompt employed for querying the model in this task is: "Could you please 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 is shown in Fig. 3.

Evaluation Criteria: We introduce a benchmarking suite specifically for GCG, comprising a validation set of 2.5K images and a test set of 5K images. Four key aspects are evaluated: i) the quality of the generated dense captions, ii) the accuracy of mask-to-phrase correspondence, iii) the quality of generated masks, and iv) the model's ability to produce region-specific groundings. Evaluation metrics include METEOR and CIDEr for caption quality, class-agnostic mask AP for grounding accuracy, mask IoU for segmentation quality, and mask recall for assessing region-specific grounding.

Mask Recall: To quantify region-specific grounding, we propose a 'mask recall' metric, utilizing a two-tiered validation approach. Initially, predicted masks are mapped to ground-truth masks via a one-to-one set assignment, followed by IoU computation for these pairs. Pairs surpass-

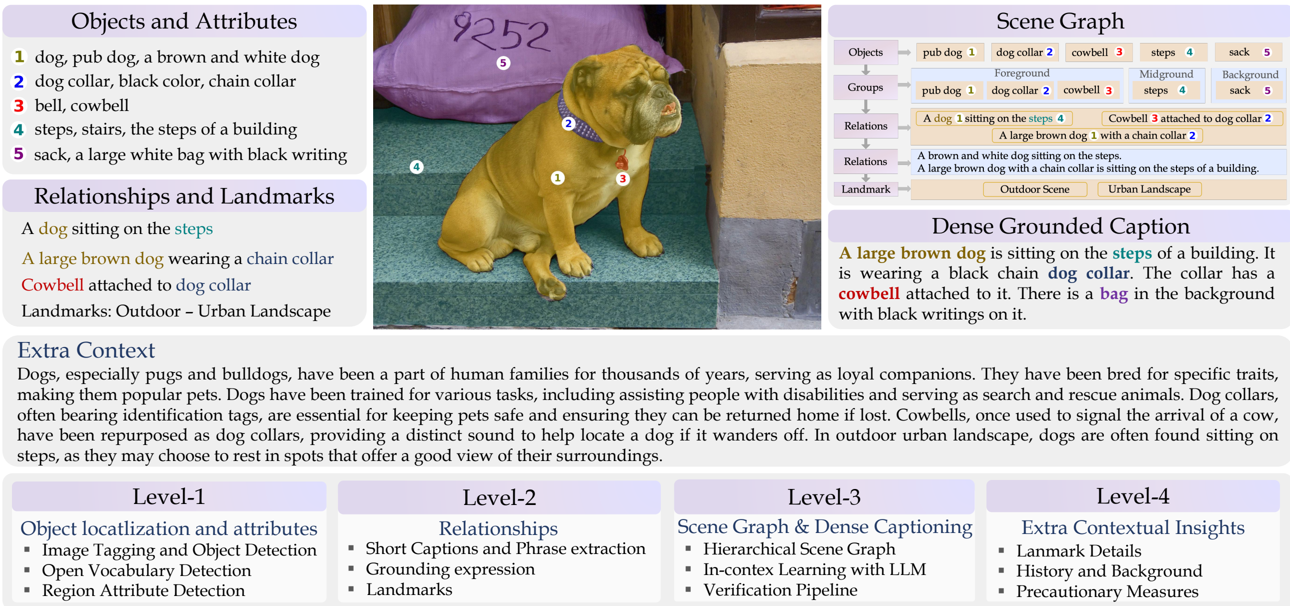


Figure 4. **Automatic Annotation Pipeline of the Grounding-anything Dataset (GranD).** Comprising four levels, this pipeline plays a pivotal role in generating GranD's 7.5M unique concepts grounded in 810M regions. level-1 details objects and attributes, level-2 includes short captions and relational 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.

ing a 0.5 IoU threshold proceed to a textual similarity assessment using BERT. A pair is considered a true positive (TP) only if both IoU and BERT similarity exceed their 0.5 thresholds; otherwise, it is classified as a false positive (FP). The mask recall is subsequently calculated using the standard formula, normalizing the number of TPs by the total ground-truth mask count.

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

4. Data Annotation Pipeline

In this section, 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. The dataset aims to overcome the challenges in image understanding and dense pixel-level grounding, thereby expanding the capabilities of visual instruction tuning in LMMs.

Our automated annotation pipeline operates at four distinct levels of granularity to achieve intricate image-text grounding (see Fig. 4). **i)** *Level-1* focuses on object localization and provides semantic labels, segmentation masks, attributes, and depth information. **ii)** *Level-2* includes relationships, that are short scene-level descriptions aligned with the objects in level-1. **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).

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 individual 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 different 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 serve to articulate connections between multiple objects or to define an object's role

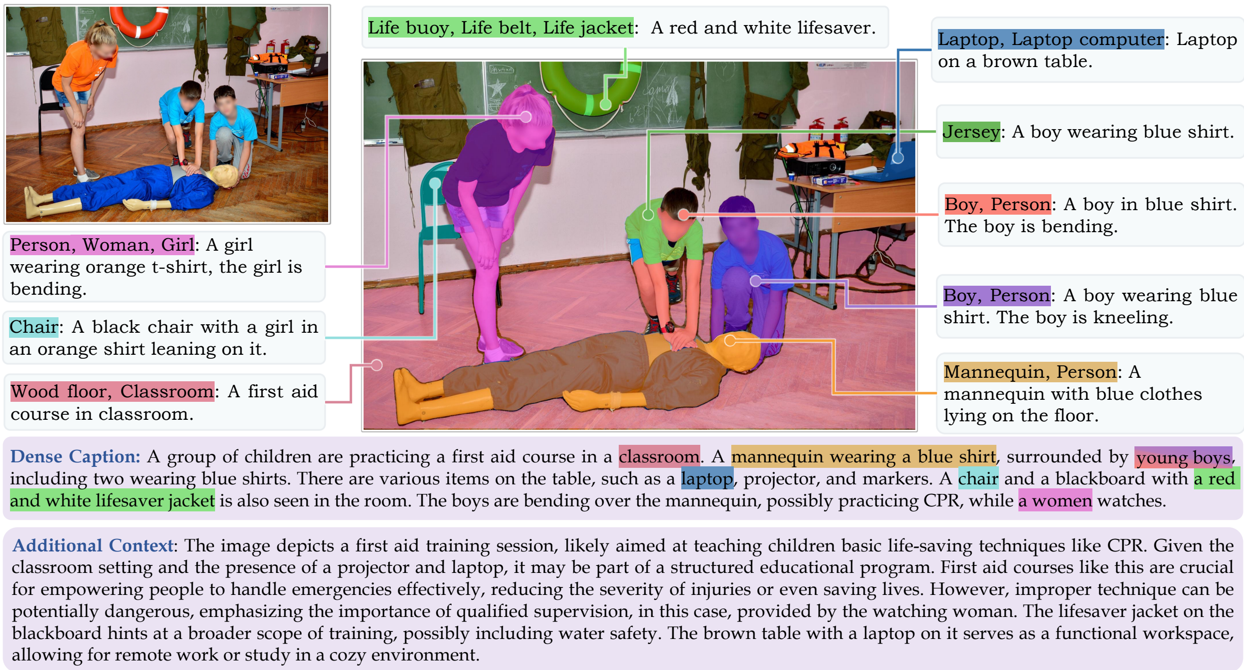


Figure 5. **Dataset sample from GranD dataset**. Our automated annotation pipeline provides multiple semantic tags and attributes for objects along with segmentation masks. The dense caption thoroughly describes the visual scene with part of the text grounded to the corresponding objects. The additional context provides a deeper understanding of the scene, going beyond what's observed.

within the scene. Further, each scene is assigned a landmark category, that includes a primary and a more specific sub-category (see Tab. 8 in Appendix. 8).

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, that are expected to be present in the image. If any object specified in the checklist is absent from the scene graph, the associated caption is flagged as inaccurate. 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 generated in level-3 to obtain a more detailed visual understanding. By querying LLM with the scene graph, we extract extended contextual insights that go 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 enriched contextualization, we optimize LLM response with carefully designed prompts and in-context learning.

Utilizing our automated annotation pipeline, we annotate a corpus of 11M SAM images [21], which are inherently diverse, high-resolution, and privacy-compliant. The resulting dataset comprises a total of 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 for dataset sample visualizations).

Implementation Details: Our automated annotation pipeline incorporates diverse state-of-the-art models at various levels. For Level-1, we use Tag2Text [17] and RAM [63] for image tagging, Co-DETR [67], EVAv02 [10], OWL-ViT [36], and POMP [44] for object localization,

Dataset	Images	Regions	Concepts	Tokens	Captions†
Image Level					
ImageNet-22K [9]	15M	-	22,000	-	-
COCO Cap. [28]	0.1M	-	-	8.4M	-
SBU [38]	0.8M	-	-	14.6M	-
CC12M [6]	12.4M	-	-	250.9M	-
YFCC15M [47]	15M	-	-	1.0B	-
COYO700M [4]	700M	-	-	15.0B	-
Laion-5B [45]	5B	-	-	135.0B	-
Region Level					
COCO [28]	0.1M	0.9M	80	-	-
LVIS [12]	0.1M	1.5M	1,203	-	-
Objects365 [46]	0.6M	10.1M	365	-	-
Open Images [23]	1.5M	14.8M	600	-	-
BigDetection [5]	3.5M	36.0M	600	-	-
V3Det [48]	0.2M	1.5M	13,029	-	-
VG [22]	0.1M	0.3M	18,136	51.2M	-
SA-1B [21]	11M	1.1B	-	-	-
AS-1B [50]	11M	1.2B	3.5M	132.2B	-
GranD (Ours)	11M	810M	7.5M	5.0B	33M

Table 2. **Comparison of GranD with available image and image-text datasets.** Our GranD dataset uniquely offers three †grounded captions per image and include segmentation masks for each annotated region. Note that AS-1B dataset is shaded in gray to indicate that it is a concurrent work and dataset is not publicly accessible as of the time of this publication.

GRiT [53] and GPT4RoI [62] for attribute generation, and MiDAS [43] for depth estimation. Level-2 leverages BLIP-2 [27] and LLaVA-v1.5 [31, 32] for scene descriptions and landmark categorization, SpaCy [14] for phrase extraction, and MDETR [18] for phrase grounding. For both Level-3 and Level-4, we use Vicuna-v1.5 [65] with 13B parameters, supplemented with in-context examples. Please refer to Appendix A.3 for further details on implementation and LLM prompts used across different pipeline levels.

4.5. Building GranD$_f$ for GCG

Motivated by the need for higher-quality data in fine-tuning stage, we introduce GranD$_f$. Explicitly designed for the GCG (Grounded Conversation Generation) task, this dataset encompasses approximately 214K image-grounded text pairs. Of these, 2.6k samples are reserved for validation and 5k for testing. GranD$_f$ comprises two primary components: one subset is manually annotated and the other subset derived by re-purposing existing open-source datasets.

Starting with the repurposed data, the key to our approach is the utilization of GPT-4's in-context learning to merge referring expressions in these datasets with scene-level information from image captions, thereby generating grounded captions for GCG tasks (see Fig. 14 for dataset sample visualizations).

We extend open-source datasets—namely Flickr-30K [41], RefCOCOg [19], and PSG [54] 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 make use of the 158K Flickr captions and their referring expressions alongside associated bounding boxes. These boxes are then accurately segmented using HQ-SAM [20].

In addition to these large-scale, repurposed datasets, we also contribute a small set of manually annotated but high-quality samples, aimed at benchmarking the GCG task. Leveraging images and annotations from our GranD dataset, we use our automatic annotations as starting points to streamline the annotation process. Annotators are tasked with matching the generated referring expressions to SAM GT masks, creating a focused set of approximately 1000 samples suitable for evaluation.

By maintaining the original referring expressions while augmenting them with rich scene-level information, and by supplementing with a targeted set of manually annotated samples, we have evolved existing open-source datasets into a comprehensive, high-quality dataset suite suited for GCG tasks, which we collectively term as GranD$_f$. Please refer to Appendix C for more details and visualizations.

5. Experiments

We perform a quantitative evaluation of GLaMM on six different benchmarks, including our newly introduced task, Grounded Conversation Generation (GCG). In the GCG task, the objective is to generate image-level captions wherein specific phrases are directly anchored to corresponding segmentation masks in the image. This task is evaluated alongside five other benchmarks to showcase GLaMM's effective performance and adaptability. The other tasks evaluated are: i) referring expression generation / region-level captioning, ii) referring-expression segmentation, iii) image-level captioning, and iv) conversational-style question answering. In subsequent sections, we elaborate on the task setup, designed prompts, and datasets employed, presenting quantitative results for each setting.

5.1. Grounded Conversation Generation (GCG)

The primary aim of the GCG task is to generate image-level captions with phrases explicitly linked to their correspond-

Model	Validation Set					Test Set				
	METEOR	CIDEr	AP50	mIoU	Recall	METEOR	CIDEr	AP50	mIoU	Recall
BuboGPT [64]	**17.2**	3.5	19.6	53.5	30.3	**17.1**	3.4	17.5	53.8	27.4
Kosmos-2 [39]	16.2	27.1	17.8	55.4	28.7	15.9	26.8	17.4	56.8	29.1
LISA* [24]	13.3	35.6	26.2	**62.1**	37.6	13.1	33.0	25.1	61.6	36.1
GLaMM	13.7	**35.7**	**27.3**	62.0	**38.7**	13.3	**34.8**	**26.0**	**62.0**	**36.8**

Table 3. **Comparison of GLaMM Model Performance on GCG Task**: Metrics include METEOR, CIDEr, AP, mIoU, and Mask Recall for both validation and test sets in our proposed benchmark. LISA* indicates a modified LISA adapted for GCG.

Method	**refCOCO**			**refCOCO+**			**refCOCOg**	
	val	testA	testB	val	testA	testB	val(U)	test(U)
CRIS (CVPR-22) [52]	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT (CVPR-22) [55]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES (CVPR-23) [29]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder (CVPR-23) [68]	-	-	-	-	-	-	64.6	-
SEEM (arXiv-23) [21]	-	-	-	-	-	-	65.7	-
LISA-7B (ZS) (arXiv-23) [24]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.4
LISA-7B (FT) (arXiv-23) [24]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
GLaMM (ZS)	54.7	58.1	52.2	42.5	47.1	39.5	54.8	55.6
GLaMM (FT)	**78.3**	**81.5**	**74.4**	**68.0**	**75.7**	**61.8**	**72.5**	**72.0**

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.

ing segmentation masks. Initially, we pretrain GLaMM on our comprehensive GranD dataset and then fine-tune it using the GranD$_f$ dataset we have tailored for GCG (see Sec. 4.5).

We evaluate the model using METEOR and CIDEr for caption quality, class-agnostic mask AP for mask-to-phrase accuracy, mask IoU for mask quality, and mask recall for region-specific grounding. The comparative results between our baseline and fine-tuned models are shown in Tab.3, covering both validation and test splits of our custom GCG benchmark (refer to Sec. 3.2). Qualitative results are illustrated in Fig. 3 and Fig. 9.

5.2. Referring Expression Segmentation

In this task, the model receives an image along with a text-based referring expression, to which it outputs a corresponding segmentation mask. We instruct the model using a prompt like, "`Please segment the <referring expression> in the given image.`" The model generates a single segmentation token alongside its text response, which is subsequently decoded to obtain the mask. We present quantitative results on the validation and testing sets of refCOCO, refCOCO+, and refCOCOg in Tab. 4. The results indicate that our performance surpasses that of closely related work, such as LISA. This demonstrates the efficacy of our GranD dataset, given that the model encounters a wide-ranging vocabulary of concepts during its pre-training phase. Additionally, refer to Fig. 6 (middle) and Fig. 10 for qualitative results.

Model	**refCOCOg**		**Visual Genome**	
	METEOR	CIDEr	METEOR	CIDEr
GRIT [53]	15.2	71.6	17.1	142
Kosmos-2 [39]	14.1	62.3	-	-
GPT4RoI [62]	-	-	17.4	145.2
GLaMM (ZS)	**15.7**	**104.0**	**17.0**	**127.0**
GLaMM (FT)	**16.2**	**105.0**	**18.6**	**157.8**

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.

5.3. Region Level Captioning

In this task, the goal is to generate referring expressions, or region-specific captions. The model is input an image, a designated region, and accompanying text, and is then tasked with responding to questions about the specified region. We utilize a prompt like, "`Can you provide a detailed description of the region <bbox>?`", to instruct the model for this task. The special token <bbox> is replaced with the actual region representation, specified by

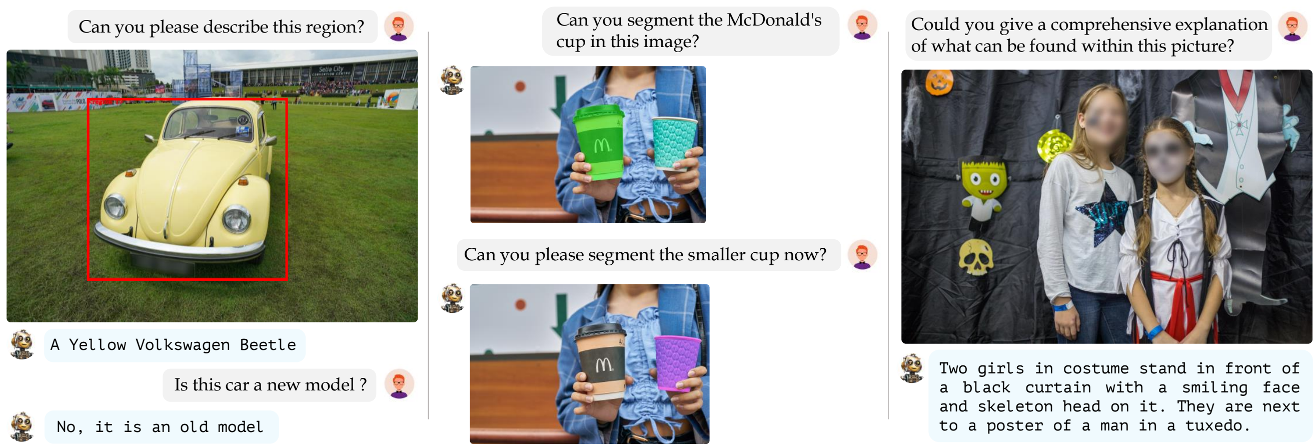


Figure 6. **Qualitative results of GLaMM's performance across downstream tasks.** The figure showcases examples from three tasks: region-level understanding (left), referring-expression segmentation (centre), and image-level captioning (right), demonstrating its capabilities in offering in-depth region understanding, pixel-level groundings, and conversational abilities through an end-to-end training approach.

Model	**NoCap**		**Flickr30k**	
	CIDEr	SPICE	CIDEr	SPICE
VinVLM [60]	95.5	13.5	-	-
LEMON [15]	106.8	14.1	-	-
SimVLM [51]	110.3	14.5	-	-
CoCa [59]	120.6	15.5	-	-
BLIP [26]	113.2	14.7	-	-
BLIP-2 [27]	121.6	15.8	-	-
InstructBLIP [8]	**123.1**	-	82.8	-
Shikra-13B [7]	-	-	73.9	-
Kosmos-1 [16]	-	-	67.1	14.5
Kosmos-2 [39]	-	-	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.

the user via a bounding box. Our GranD dataset's level-1 and level-2 data aid the model in understanding these regions, enabling it to provide relevant answers.

We conduct a quantitative assessment of region-level captioning on two OpenSource datasets: Visual Genome [22] and refCOCOg [19]. Performance is measured using the METEOR and CIDEr metrics, and the results are shown in Tab. 5. When compared to GRiT [53] and GPT4RoI [62], GLaMM, after fine-tuning on these two datasets, exhibits superior performance. Notably, the zero-shot capabilities of GLaMM also highlight the significance of having a large-scale dataset featuring region-text pairs, as facilitated by GranD. We also present qualitative examples in Fig. 6 (left) and Fig. 11.

5.4. Image Level Captioning

For image captioning, the model is presented with an image accompanied by a user query such as, "`Could you please give me a detailed description of the image?`", to which the model generates a textual description. We assess GLaMM's capabilities on two datasets in a zero-shot setting, i.e., Flickr30k [41] and NoCap [1]. Tab. 6 illustrates that GLaMM offers favourable performance when compared with recent models specialized in image captioning, as well as other LMMs. Qualitative results for image captioning are shown in Fig. 6 (right) and Fig. 12.

Method	**LLM**	**LLaVA^{W}**
BLIP-2 [27]	Vicuna-13B	38.1
InstructBLIP [8]	Vicuna-7B	60.9
Qwen-VL [3]	Qwen-7B	63.4
Qwen-VL-Chat [3]	Qwen-7B	58.6
LLaVA-1.5 [30]	Vicuna-7B	63.4
GLaMM	Vicuna-7B	<u>63.3</u>

Table 7. **Evaluation of GLaMM on conversational style QA using LLaVA-Bench.** The table compares GLaMM's performance with previous open-source models in conversation question-answering, detailed descriptions, and complex reasoning tasks.

5.5. Conversational Style Question Answering

We evaluate our model on the LLaVA-Bench [31, 32] that uses GPT-4 for evaluation of models. This benchmark tests the model on three different types of tasks: conversation question-answering, detailed descriptions, and complex reasoning tasks. The evaluation provides insights into the model's conversational and reasoning capabilities. The results in Tab. 7 present a comparison of GLaMM with

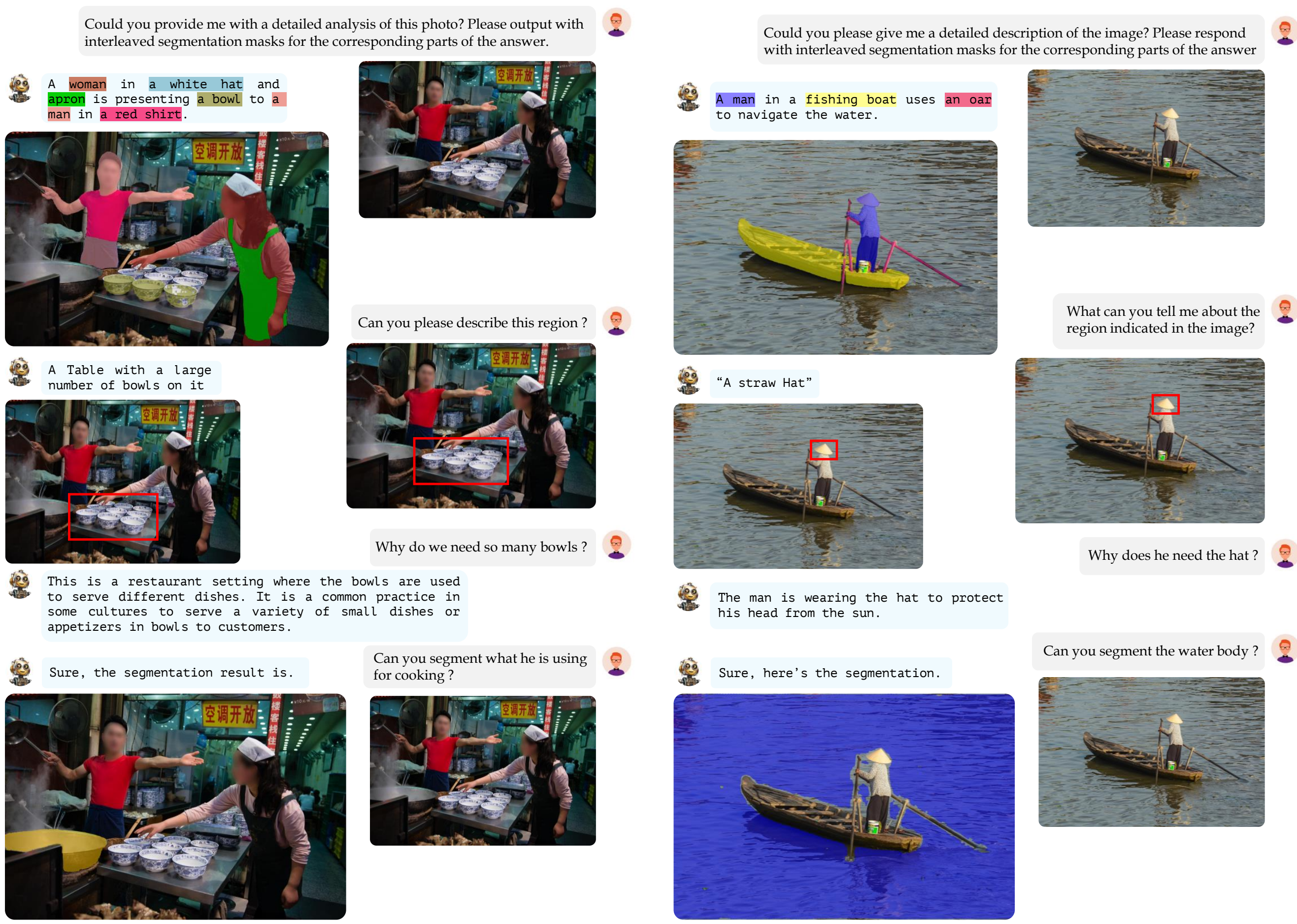


Figure 7. **Multimodal conversational interactions facilitated by GLaMM.** The figure showcases GLaMM engaging in multi-turn dialogues, providing detailed descriptions, addressing region-specific inquiries, and presenting grounded conversations. This effectively highlights its adaptability in intricate visual-language interactions and robustly retaining reasoning capabilities inherent to LLMs.

previous open-source models. Although, GLaMM uses LLaVA-1.1 as based model, its performance is on par with the recently released LLaVA-1.5 which leverages additional data and MLP for vision-to-language mapping. Qualitative results are shown in Fig. 7 and Fig. 13.

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, that is used for our model pertaining. 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 for the development of visually grounded conversations. Our model is shown to effectively perform on several downstream tasks besides GCG, including region and image captioning, referring segmentation and vision-language conversations.

Ethics and Societal Impact: Our Grounding-anything Dataset (GranD) utilizes SAM images that have de-identified personal information, with all faces and license plates obscured. To the best of our knowledge, the dataset does not portray any strong biases or discrimination. We urge for the responsible use of GranD and GLaMM, promoting research progress while safeguarding privacy.

GLaMM : Pixel Grounding Large Multimodal Model

Supplementary Material

We provide supplementary material for a deeper understanding and more analysis related to the main paper, arranged as follows:

1. Additional implementation details (Appendix A)
2. Additional qualitative results (Appendix B)
3. Dataset visualizations (Appendix C)

A. Additional Implementation Details

A.1. Model Architecture and Training

In all of our experiments, we use Vicuna LLM [65] with 7B parameters. The design of region encoder is motivated from GPT4RoI [62] and grounding image encoder and pixel decoder are inspired from LISA [24]. The V-L and L-P layers are implemented using linear layer. We use PyTorch to implement our GLaMM and use Deepspeed zero-2 optimization during training. Our codes and pretrained models will be publicly released.

A.1.1 Pretraining on GranD

During pretraining GLaMM is trained on GranD dataset for referring expression segmentation, region-level captioning, image-level captioning and grounded conversation generation (GCG) tasks simultaneously. We use a batch size of 160 and train for a total of 35K iterations during pretraining. We use LORA-8 for efficiently adapting the LLM and initialize the pretraining from GPT4RoI [62] for faster convergence. In the experiment tables in Section. 5, we refer to this model as GLaMM (ZS) which is obtained after pretraining on GranD.

A.2. Finetuning on Downstream Tasks

We finetune GLaMM on multiple downstream tasks including GCG, referring expression segmentation, region-level captioning and image-level captioning. For GCG, we finetune our model on GranD$_f$ dataset. A batch size of 160 is used and the model is trained for 5K iterations in total. It is worth noting that GranD$_f$ dataset is a combination of multiple open-source datasets that we repurposed for GCG task using GPT4 [37]. Please refer to Appendix. C for the prompts designed to query GPT4 for constructing GranD$_f$ dataset, along with the dataset visualizations.

For referring expressions segmentation, we finetune GLaMM on refCOCO, refCOCO+ and refCOCOg datasets. We represent this model as GLaMM (FT) in Tab. 4. Similarly, for region-level captioning, GLaMM (FT) is finetuned on refCOCOg and Visual Genome datasets. For image-level captioning, we fine tune GLaMM on LLaVA-Instruct-150K [32] dataset. For LLaVA-bench, the model is finetuned on LLaVA-Instruct-80K [32] instruction set. We use eight NVIDIA A100-40GB GPUs in all of our pretraining and finetuning experiments.

A.3. Automated Dataset Annotation Pipeline

We design a fully automated dataset annotation pipeline using multiple hierarchical levels in the visual domain to construct GranD dataset. The segmentation masks for most of the regions are obtained from SAM [21] annotations by comparing our detected labeled regions with SAM-provided class-agnostic regions. For the remaining regions that do not match with any of the SAM regions, we run SAM model with a bounding box query to obtain masks.

Our automated annotation pipeline utilizes only open-source models and incorporates a feedback loop using the chain of thoughts prompting via LLM. As it does not require feedback from the human in the loop, it can be scaled to generate dense noisy labels for a larger number of images, which can then be used to pretrain a larger LMM. Given the availability of enough compute power, this could be a step towards building a larger generic large multi-modal model. We will release our GranD dataset along with the implementation of our automated dataset annotation pipeline for further research. Below we present the LLM prompts we use at different levels of our automated dataset annotation pipeline.

A.3.1 LLM Prompts and In-context Learning

Landmark categorization: We use LLaVA-v1.5-13B [31] model to assign landmark categories to each image. Please refer to Tab. 8 for primary and fine categories used.

Main category	Fine Category
Indoor scene	Living space, Work space, Public space, Industrial space
Outdoor scene	Urban landscape, Rural landscape, Natural landscape
Transportation scene	Road, Airport, Train station, Port and harbor
Sports and recreation scene	Sporting venue, Recreational area, Gym and fitness center

Table 8. Summary of landmark categories and their corresponding fine-grained categories. We use LLaVA-v1.5 [31] for assigning landmark categories to images.

Prompt: The provided prompt is a scene graph, which is a structured representation of a scene detailing its various elements and their relationships.
The scene graph consists of:
1. Layers of Depth: The scene is divided into different layers based on proximity - 'Immediate Foreground', 'Foreground', 'Midground', and 'Background'. Each layer depicts objects or entities at that depth in the scene.
2. Groups: Within each depth layer, objects and entities are clustered into groups, sometimes with specific attributes.
3. Relationships: This section illustrates the interactions or spatial relationships between various objects or groups.
4. Landmarks: It gives a broader view or categorization of the scene, defining its overarching theme or environment.

##Example - 1:
Prompt: {scene_graph_1}
Desired caption: {dense_caption_1}

##Example - 2:
Prompt: {scene_graph_2}
Desired caption: {dense_caption_2}

Please provide a simple and straightforward 2-4 sentence image caption based on the following scene graph details: {scene_graph}.
Create the caption as if you are directly observing the image. Do not mention the use of any source data like 'The relationship indicates ...' or 'No relations specified'.

(a) Illustration of LLM in-context learning for dense captioning used in the construction of our GranD dataset.

Prompt:
##Example - 1:
Prompt: {scene_graph_1}
Additional context: {caption_1}

##Example - 2:
Prompt: {scene_graph_2}
Additional context: {caption_2}

Provide context based on the typical usage, history, potential dangers, and other interesting aspects surrounding the general theme presented by the objects and elements in the following scene graph: {scene_graph}

Limit the response to one paragraph with 5-7 sentences.
DO NOT mention, refer to, or hint about "objects", "scene", or "scene graph".
ONLY focus on explaining use cases, history, potential dangers, etc.

(b) Illustration of LLM in-context learning for extra contextual insights used in the construction of our GranD dataset.

Figure 8. **Prompts used to construct GranD dataset.** The figure shows the prompts used to query Vicuna [65] to generate dense captions and the extra context in our automated training pipeline. We provide in-context examples to guide the LLM.

Dense Captioning: We arrange objects, attributes and relationships hierarchically to construct a visual scene graph, that is used to query Vicuna-v1.5-13B [65] model along with in-context examples to generate dense captions. The designed prompt is shown in Fig. 8 (a).

Extra Context: We query Vicuna-v1.5-13B model to generate additional context about the visual scene. The prompt designed for this purpose is shown in Fig. 8 (b).

B. Additional Qualitative Results

In this section, we provide more qualitative examples to better understand the capacity of GLaMM.

B.1. Grounded Conversation Generation (GCG)

Fig. 9 shows qualitative results of GLaMM finetuned on GranD$_f$ dataset. The model could produce dense captions and provide dense pixel-level groundings of the caption.

B.2. Referring Segmentation

Fig. 10 shows the effectiveness of GLaMM in understanding the natural language query and segmenting the corresponding objects. Note that GLaMM can also segment multiple objects via multi-round conversations.

B.3. Region-level Captioning

Fig. 11 shows the qualitative results of GLaMM for region-level understanding. Our model can generate detailed descriptions about the user-specified regions in an image.

B.4. Image-level Captioning

Fig. 12 shows GLaMM's qualitative results on captioning tasks. Our model can generate dense captions for images.

B.5. Conversations

Fig. 13 illustrates the unique functionality of GLaMM to engage in multi-purpose task conversations. GLaMM is a generic conversational model that can accept prompts in the

Figure 9. **Qualitative results of GLaMM's performance in grounded conversation generation.** The figure shows how GLaMM seamlessly generates detailed responses, grounding phrases using pixel-level masks showing its detailed understanding.

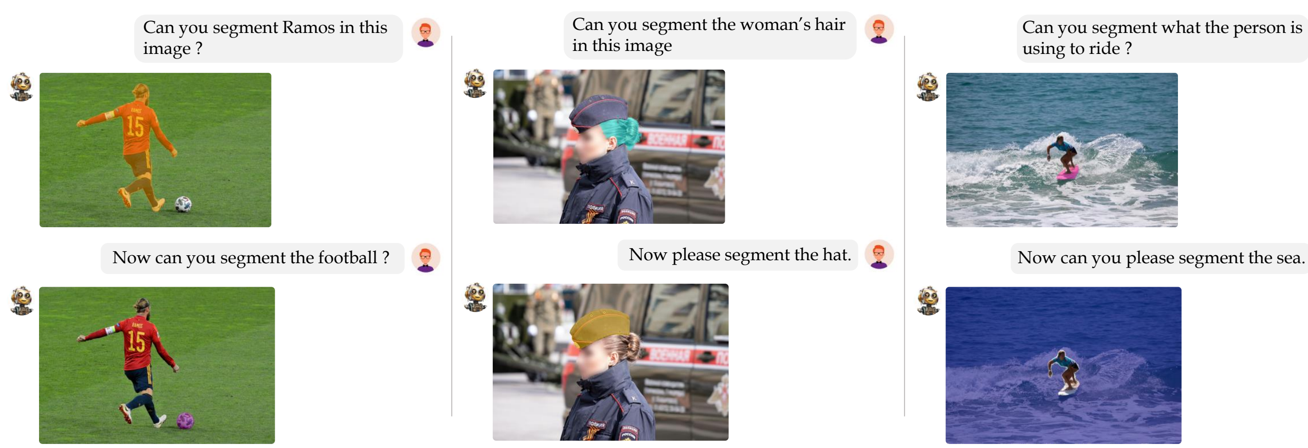


Figure 10. **Qualitative results of GLaMM's capability in referring expression segmentation.** The figure illustrates how GLaMM effectively translates text-based referring expressions into corresponding segmentation masks. Leveraging its training on the GranD dataset, the model can provide pixel-grounded reasoning and operate across various levels of granularity.

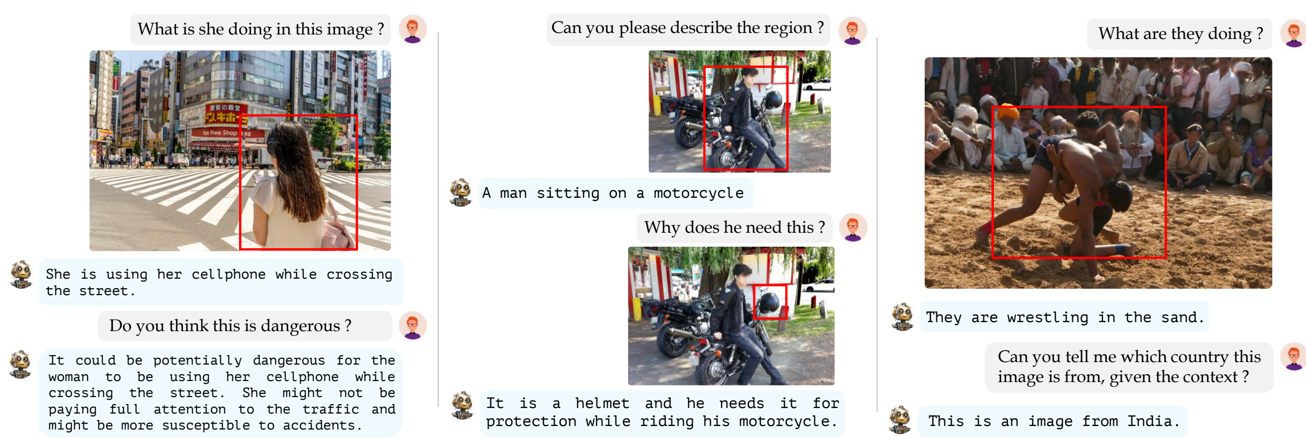


Figure 11. **Qualitative illustration of GLaMM's performance in region-level captioning.** The figure demonstrates GLaMM's ability to generate region-specific captions adeptly, translating the intricate details from designated regions into coherent textual descriptions, enriched by its training on the comprehensive GranD dataset. This capability, combined with the inherent reasoning abilities of LLMs, enables it to tackle reasoning-based visual questions about these regions.

Figure 12. **Qualitative results of GLaMM on image-level captioning tasks.** The figure shows GLaMM's ability to generate detailed captions of images. GLaMM can identify text within images (e.g. TESCO in the left image), differentiate between live objects and statues (middle image), and reason about complex visual scenes (right image).

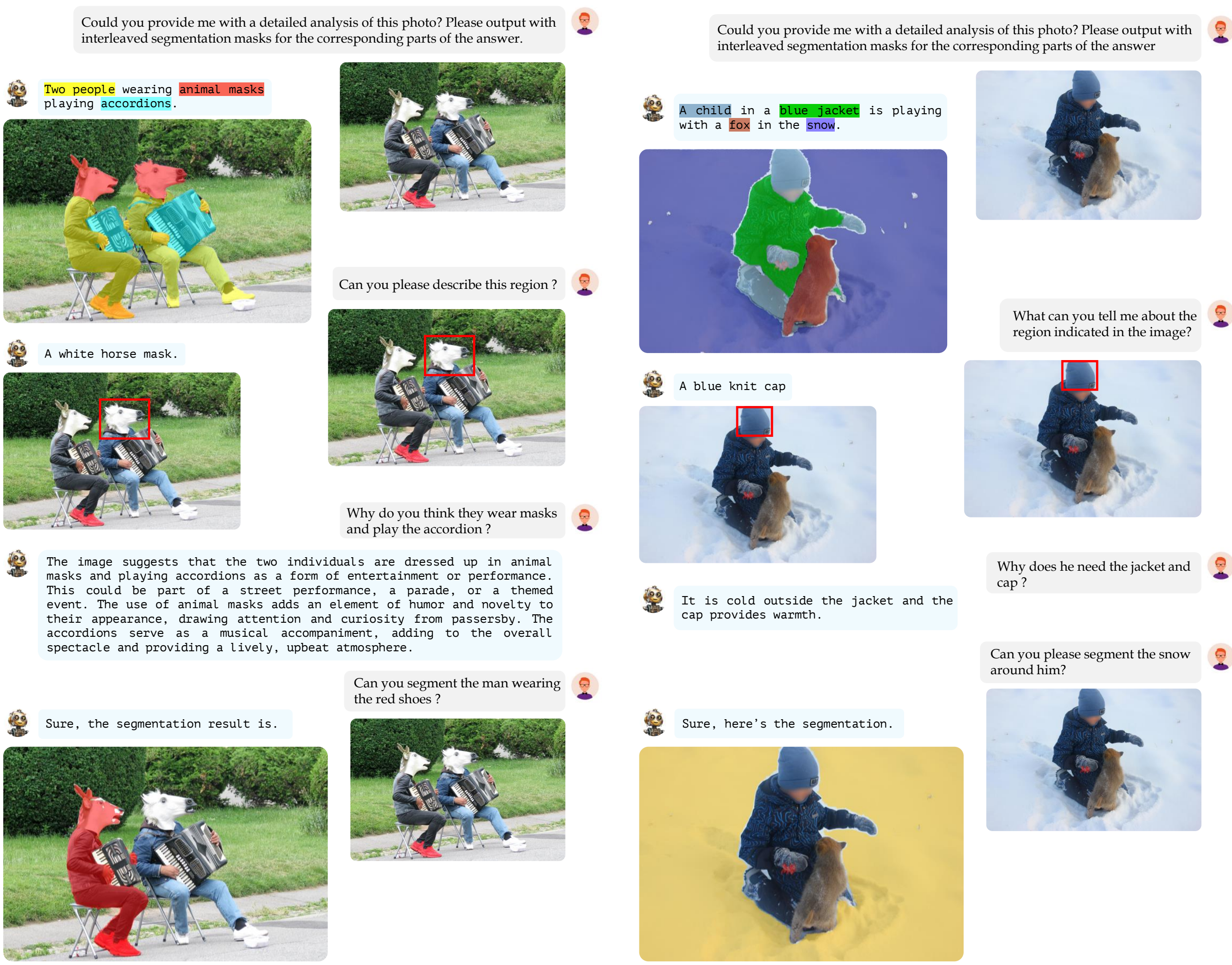


Figure 13. **Multimodal conversational with GLaMM.** The figure shows multimodal conversations generated through GLaMM. The model is flexible enough to process multimodal inputs and respond with multimodal outputs in a single conversation.

Prompt: You are given five base captions for an image briefly describing the image from different perspective. You are also given a number of relationships between objects in the image. Each relationship consists of [subject, relation/verb, object]. Note that each subject and object follows the format <entitiy name>-<object number in image>. For example, if there are five persons and two tables, they will be formatted as person-1, person-2,.., person-5 and table-1, table-2.

Provide a concise image caption that straightforwardly describes objects/things visible in the image, using provided relationships. Use the base captions for context understanding, but it is not mandatory to include them verbatim in the final description. Break down the description into shorter sentences. Importantly keep the names of the subject and object unchanged. Break down the description into shorter, clear sentences. Do not add any extra information. Keep it brief.

'person in truck', 'car parked on pavement-merged', 'truck parked on dirt-merged', 'house beside tree-merged', 'dirt-merged beside pavement-merged'

A large commercial truck is parked on a dirt area, with its lights on. A person is sitting inside this truck. Close by the truck, a car is parked on the pavement. There is a house nestled beside a tree that's near the dirt and pavement.

'person attached to dog', 'person on bed' , 'dog on bed', 'dog attached to blanket', 'bed in front of wall', 'curtain attached to wall', 'cabinet attached to curtain'

The image depicts a person on a bed, lovingly hugging a dog. Both the person and the dog are on the bed, with the dog also attached to a blanket. Further aspects of the room include a curtain attached to a wall, and a cabinet attached to this curtain.

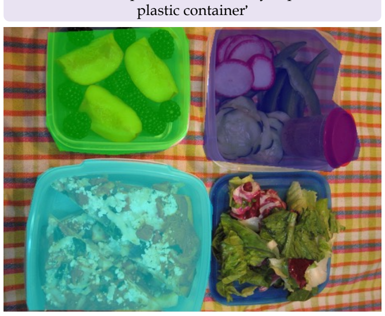
'person about to hit sports ball', 'person swinging tennis racket', 'person running on playingfield', 'banner attached to fence-merged', 'playingfield beside fence-merged'

A person is running on a playing field situated beside a fence. This individual is about to hit a sports ball, swinging a tennis racket in preparation. Attached to the fence is a banner.

(a) Samples from our GranD_f dataset: Illustrating the repurposing of the OpenPSG dataset for the GCG task.

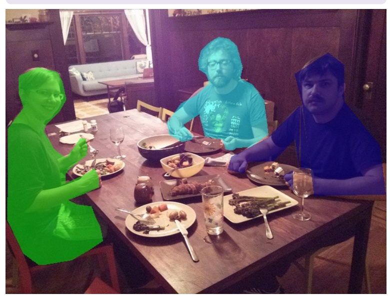
Prompt: Create a concise image caption that straightforwardly describes things visible in the image, using provided phrases verbatim without altering them in any way. Do not infer actions or relationships not explicitly mentioned in the phrases. Use the base captions for contextual understanding, but it is not mandatory to include them verbatim in the final description. Preserve the exact wording of the provided phrases in the final caption. Break down the description into shorter, clear sentences, but without sacrificing the exact wording of the provided phrases. Construct a simple, coherent, and succinct caption that accurately uses the exact phrases, organizing the information into shorter sentences without assuming or inferring additional actions or relationships.

' a tupperware box filled with fruit', 'a plastic container with sliced radishes, green peppers, cucumbers and a sauce', 'slices of pizza are also visibly kept within a plastic container'



A tupperware box filled with fruit features prominently on the table, along with a plastic container with sliced radishes, green peppers, cucumbers and a sauce. Several slices of pizza are also visibly kept within a plastic container.

'a woman with glasses and a purple shirt is having a meal', 'a man with a mustache wearing a blue shirt with food on his plate', 'the man with the glasses on'

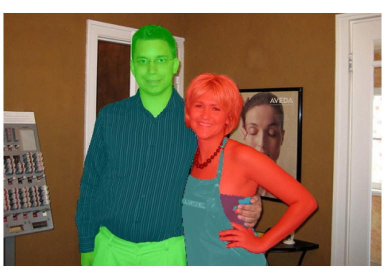


A woman with glasses and a purple shirt is having a meal at a table with a man with a mustache wearing a blue shirt with food on his plate and another man, the man with the glasses on.

'a laptop computer sitting to the left of a red book', ' laptop on the right side of the table with a scrabble game open'

A laptop computer sitting to the left of a red book. There is also a laptop on the right side of the table with a scrabble game open.

(b) Samples from our GranD_f dataset: Illustrating the repurposing of the RefCOCO-g dataset for the GCG task.



A man in a striped shirt poses with a blond girl in a black apron.

A small group of three enjoy the view of the water as a small boy wanders off and two companions enjoy a walk along the shoreline.

Toddler sits on carpet in living room touching guitar .

(c) Samples from our GranD_f dataset: Illustrating the repurposing of the Flickr-30k dataset for the GCG task.

Figure 14. **Dataset samples from GranD_f.** The figure shows the GPT4 [37] prompts used and the created dataset samples from Grand_f dataset. This repurposed human-annotated dataset provides rich semantics to GLaMM for GCG task.

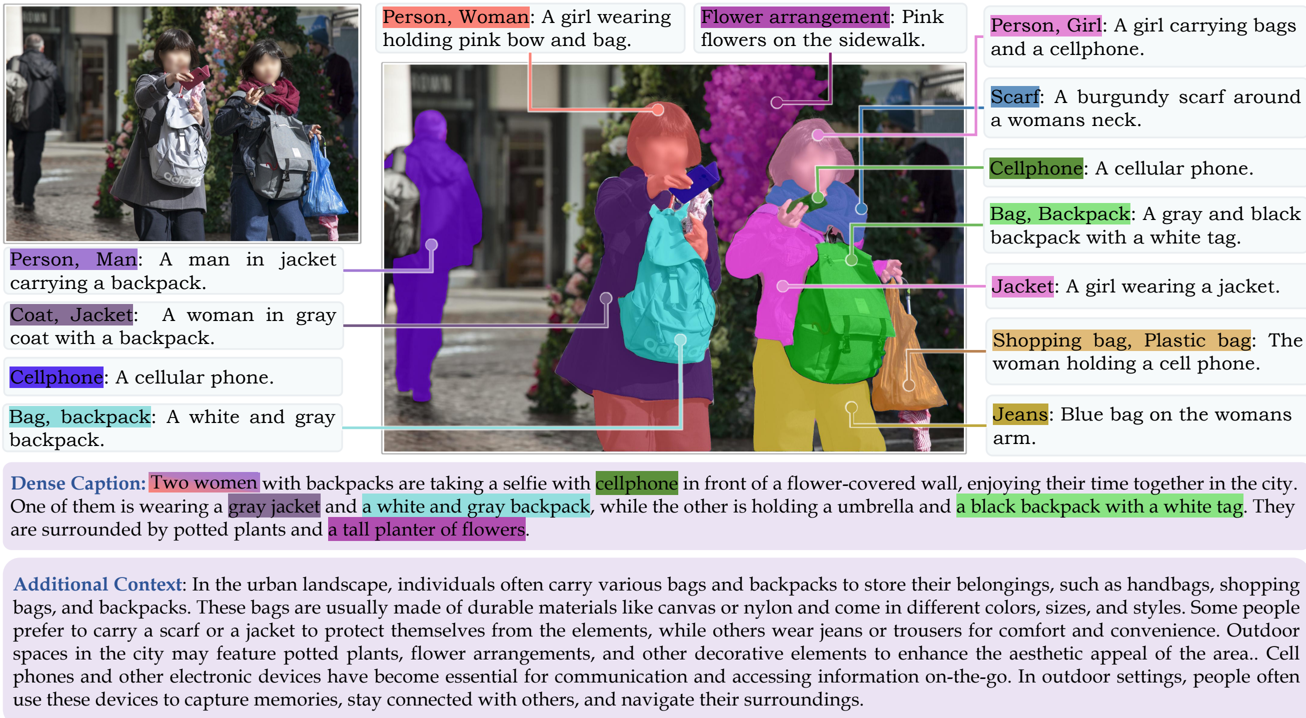


Figure 15. **Dataset samples from GranD.** The figure shows a sample from GranD dataset, generated using the automated annotation pipeline. It provides multiple semantic labels and attributes for detected objects, along with the grounded dense caption and additional context.

form of text and/or region and can answer in the form of text and/or segmentation masks. Note that our model is not explicitly trained to handle such scenarios, and this behavior emerges mainly due to our pretraining on GranD dataset, where an image is presented to LMM in different contexts.

C. Dataset Visualization

In this section, we provide additional dataset samples of our GranD and GranD$_f$ datasets to better understand the functionalities they offer. Please see Fig. 15 and Fig. 14.

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*, pages 8948–8957, 2019. 10

[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 preprint arXiv:2307.13721*, 2023. 1

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 10

[4] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 8

[5] 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. 8

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021. 8

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2, 3, 10

[8] 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 preprint arXiv:2305.06500*, 2023. 1, 2, 3, 10

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 8

[10] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 7

[11] 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 preprint arXiv:2304.15010*, 2023. 1, 3

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 8

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 4

[14] Matthew Honnibal and Ines Montani. spaCy: Industrial-strength Natural Language Processing in Python. 2020. 8

[15] 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*, pages 17980–17989, 2022. 10

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

[17] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023. 7

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 8

[19] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 8, 10

[20] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv preprint arXiv:2306.01567*, 2023. 8

[21] 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, 2023. 2, 3, 5, 7, 8, 9, 1

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 8, 10

[23] 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. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 8

[24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2, 3, 9, 1

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

[26] 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. 10

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 8, 10

[28] 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. 8

[29] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 9

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 10

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 8, 10, 1

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3, 8, 10, 1

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

[34] 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, 2023. 2, 3

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

[36] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, 2022. 7

[37] OpenAI. Gpt-4 technical report, 2023. 2, 1, 5

[38] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011. 8

[39] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2, 3, 9, 10

[40] 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, 2023. 2, 3

[41] 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*, pages 2641–2649, 2015. 2, 8, 10

[42] 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. pages 8748–8763. PMLR, 2021. 4

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 8

[44] Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alex Smola, and Xu Sun. Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition. *arXiv preprint arXiv:2304.04704*, 2023. 7

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 8

[46] 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. 8

[47] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 8

[48] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. *arXiv preprint arXiv:2304.03752*, 2023. 8

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

[50] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 2, 3, 8

[51] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 10

[52] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. 9

[53] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 8, 9, 10

[54] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, 2022. 2, 8

[55] 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*, pages 18155–18165, 2022. 9

[56] 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. 2023. 3

[57] 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, 2023. 1, 2, 3

[58] 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 preprint arXiv:2310.07704*, 2023. 3

[59] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 10

[60] 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 preprint arXiv:2101.00529*, 1(6): 8, 2021. 10

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

[62] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 2, 3, 4, 8, 9, 10, 1

[63] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 7

[64] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 2, 3, 9

[65] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 8, 1, 2

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

[67] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *ICCV*, 2023. 7

[68] 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*, pages 15116–15127, 2023. 9