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 downstream tasks (Appendix B
3. Additional qualitative results (Appendix C)
4. Dataset visualizations (Appendix D)
5. Limitations and future work (Appendix E)
6. Ethics and societal impact (Appendix F)

A. Additional Implementation Details

A.1. Evaluation Metrics

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 surpassing 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.

A.2. Model Architecture and Training

In all of our experiments, we use Vicuna LLM [60] with 7B parameters. The design of region encoder is motivated from GPT4RoI [57] and grounding image encoder and pixel decoder are inspired from LISA [21]. The V-L and L-P layers are implemented using 2 layer MLP with GELU activation as in LLaVA-v1.5 [28]. We use PyTorch to implement our GLaMM and use Deepspeed zero-2 optimization during training.

Specifically, our model is trained using two types of losses: auto-regressive cross-entropy loss for text generation and a linear combination of per-pixel binary cross-entropy loss and DICE loss for segmentation. During training, the global image encoder and grounding image encoder are kept frozen and the region encoder, projection layers (V-L and L-P) and the pixel decoder are fully finetuned, while the LLM is LORA finetuned with $\alpha = 8$. Our codes and pretrained models will be publicly released.

A.2.1 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 [57] 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.3. 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 [34]. Please refer to Appendix. D 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 [29] dataset. For LLaVA-bench, the model is finetuned on LLaVA-Instruct-80K [29] instruction set. We use eight NVIDIA A100-40GB GPUs in all of our pretraining and finetuning experiments.

A.4. Automated Dataset Annotation Pipeline

Our automated annotation pipeline incorporates diverse state-of-the-art models at various levels. For Level-1, we use Tag2Text [14] and RAM [58] for image tagging, Co-DETR [62], EVAv02 [7], OWL-ViT [33], and POMP [40] for object localization, GRiT [48] and GPT4RoI [57] for attribute generation, and MiDAS [39] for depth estimation. Level-2 leverages BLIP-2 [24] and LLaVA-v1.5 [28, 29] for scene descriptions and landmark categorization, SpaCy [11] for phrase extraction, and MDETR [15] for phrase grounding. For both Level-3 and Level-4, we use Vicuna-v1.5 [60] with 13B parameters, supplemented with in-context examples. Please refer to Appendix A.4 for further details on implementation and LLM prompts used across different pipeline levels.

We design a fully automated dataset annotation pipeline using multiple hierarchical levels in the visual domain to construct GranD dataset. The segmentation masks for

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 6. **Prompts used to construct GranD dataset.** The figure shows the prompts used to query Vicuna [60] to generate dense captions and the extra context in our automated training pipeline. We provide in-context examples to guide the LLM.

most of the regions are obtained from SAM [18] 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.4.1 LLM Prompts and In-context Learning

Landmark categorization: We use LLaVA-v1.5-13B [28] model to assign landmark categories to each image. Please refer to Tab. 7 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 7. Summary of landmark categories and their corresponding fine-grained categories. We use LLaVA-v1.5 [28] for assigning landmark categories to images.

Dense Captioning: We arrange objects, attributes and relationships hierarchically to construct a visual scene graph, that is used to query Vicuna-v1.5-13B [60] model along with in-context examples to generate dense captions. The designed prompt is shown in Fig. 6 (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. 6 (b).

B. Additional Downstream Tasks

B.1. Phrase Grounding

In order to adapt the GLaMM model for phrase grounding, we repurpose the GCG dataset to suit this particular task. Specifically, the answers in the GCG dataset are now used as questions, and the parts of the captions containing groundings are regarded as phrases. The model is subsequently trained to locate pixel-level groundings for these phrases, which are enclosed within <p> and </p> tokens. The results of this adaptation are shown in the following figure.

Please generate a segmentation mask for the specific phrase highlighted within the image caption. A <p>woman</p> in <p>a navy blue jacket</p> and <p>hat</p> with a hair ribbon in <p>her hair</p>.

A woman in a navy blue jacket and hat with a hair ribbon in her hair.

B.2. Conversational Style Question Answering

We evaluate our model on the LLaVA-Bench [28, 29] 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. 8 present a comparison of GLaMM with previous open-source models. We note that GLaMM performance is on par with the recently released LLaVA-1.5 which leverages additional data for vision-to-language alignment. Qualitative results are shown in Fig. 11 and Fig. 13.

Method	**LLM**	**LLaVA^{W}**
BLIP-2 [24]	Vicuna-13B	38.1
InstructBLIP [6]	Vicuna-7B	60.9
Qwen-VL [3]	Qwen-7B	63.4
Qwen-VL-Chat [3]	Qwen-7B	58.6
LLaVA-1.5 [27]	Vicuna-7B	63.4
GLaMM	Vicuna-7B	<u>63.3</u>

Table 8. **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.

C. Additional Qualitative Results

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

C.1. Grounded Conversation Generation (GCG)

Fig. 7 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.

C.2. Referring Segmentation

Fig. 8 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.

C.3. Region-level Captioning

Fig. 9 shows the qualitative results of GLaMM for region-level understanding. Our model can generate detailed descriptions about the user-specified regions in an image.

C.4. Image-level Captioning

Fig. 10 shows GLaMM's qualitative results on captioning tasks. Our model can generate dense captions for images.

C.5. Conditional Image Generation

Fig. 12 shows GLaMM's seamless integration for generative tasks. We use the Stable Diffusion inpainting model *stable-diffusion-xl-1.0-inpainting* [41] for this task. We first generate a segmentation mask using our GlaMM model based on the user query. This segmentation mask along with the user prompt is given as the input to the Stable Diffusion inpainting model, which generates the final output.

C.6. 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 form of text and/or region and can answer in the form of

Figure 7. **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.

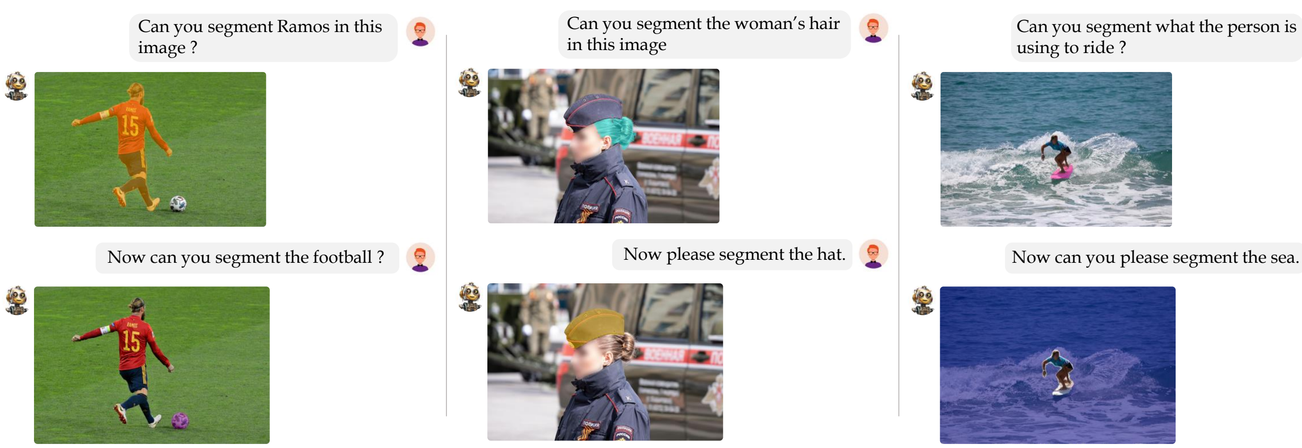

Figure 8. **Qualitative results of GLaMM's capability in referring expression segmentation.** The figure illustrates how GLaMM effectively translates text-based referring expressions into corresponding segmentation masks. Leveraging its training on the GranD dataset, the model can provide pixel-grounded reasoning and operate across various levels of granularity.

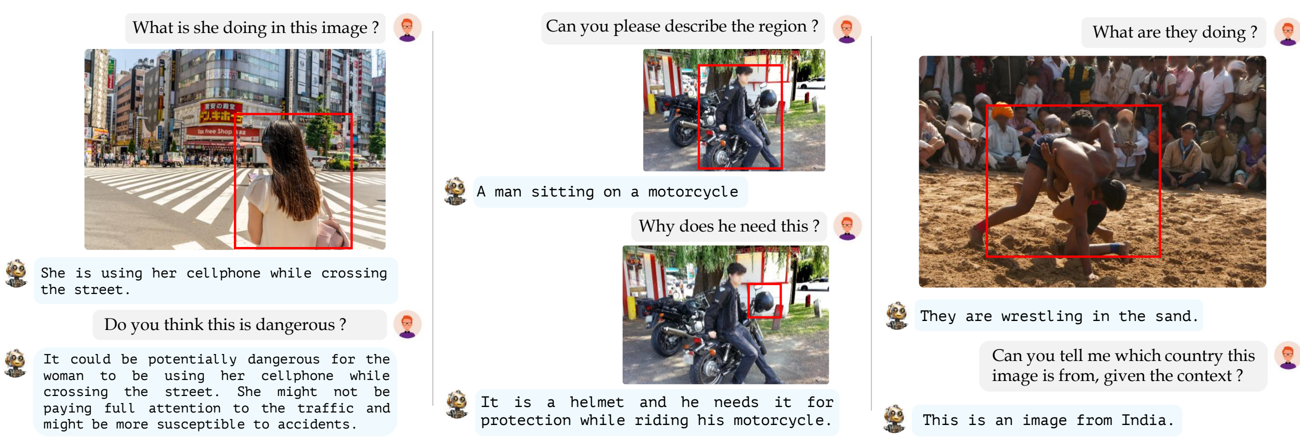

Figure 9. **Qualitative illustration of GLaMM's performance in region-level captioning.** The figure demonstrates GLaMM's ability to generate region-specific captions adeptly, translating the intricate details from designated regions into coherent textual descriptions, enriched by its training on the 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 10. **Qualitative results of GLaMM on image-level captioning tasks.** The figure shows the capabilities of GLaMM in generating detailed and context-aware captions for a diverse range of images. On the left, GLaMM demonstrates its proficiency in text recognition within images; it accurately identifies and incorporates specific textual information, such as the brand name "TESCO," into its caption. In the middle image, GLaMM's capability to discern subtleties in visual content is showcased. It can effectively distinguish between live entities and inanimate objects, such as differentiating a living creature from a statue. On the right, the figure demonstrates GLaMM's competence in reasoning about complex visual scenes. It can analyze and describe intricate details and interactions within an image, reflecting a deep understanding of both the individual elements and the overall context of the scene.

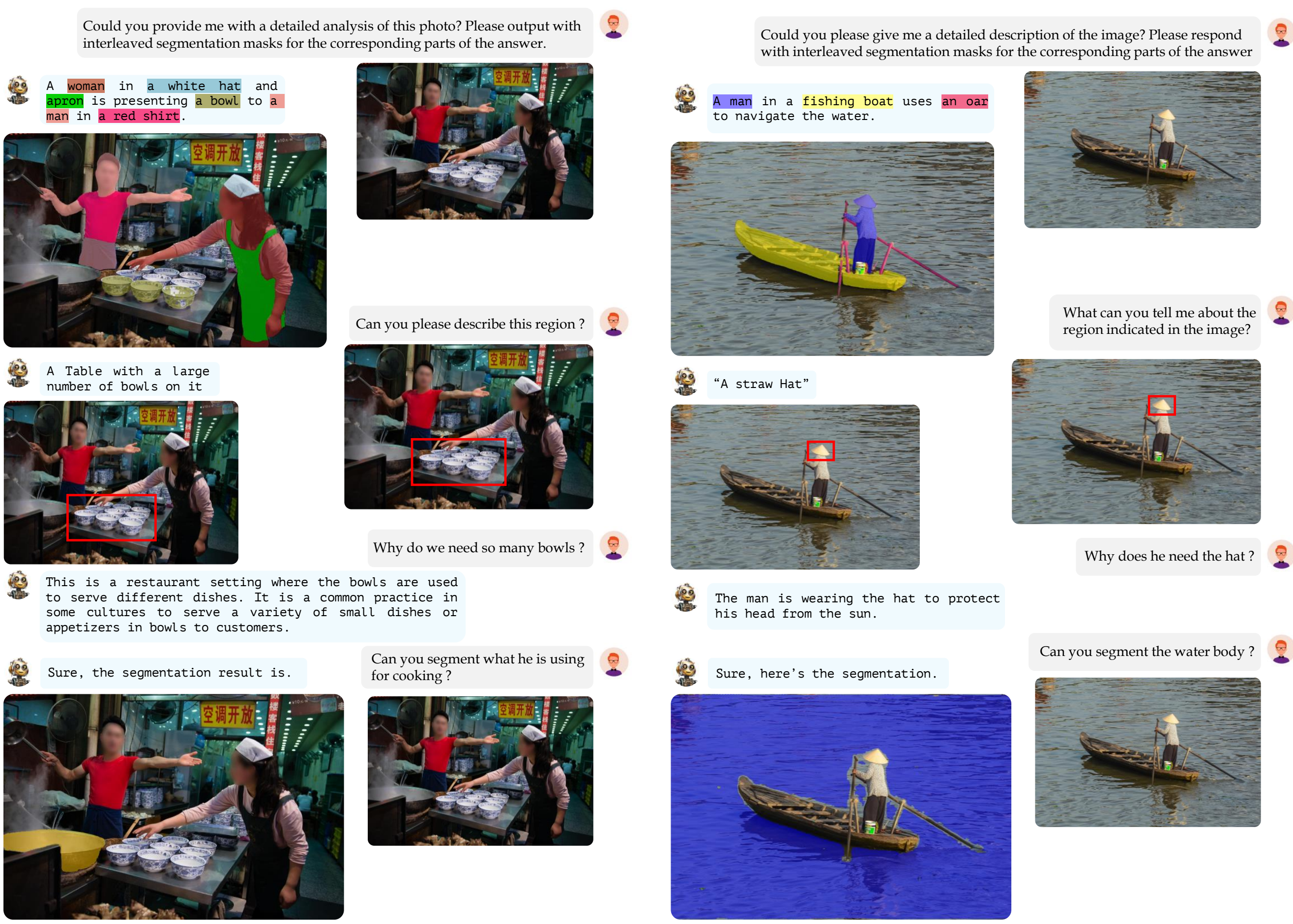

Figure 11. **Multimodal conversational interactions facilitated by GLaMM.** The figure showcases GLaMM engaging in multi-turn dialogues, providing detailed descriptions, addressing region-specific inquiries, and presenting grounded conversations. This effectively highlights its adaptability in intricate visual-language interactions and robustly retaining reasoning capabilities inherent to LLMs.

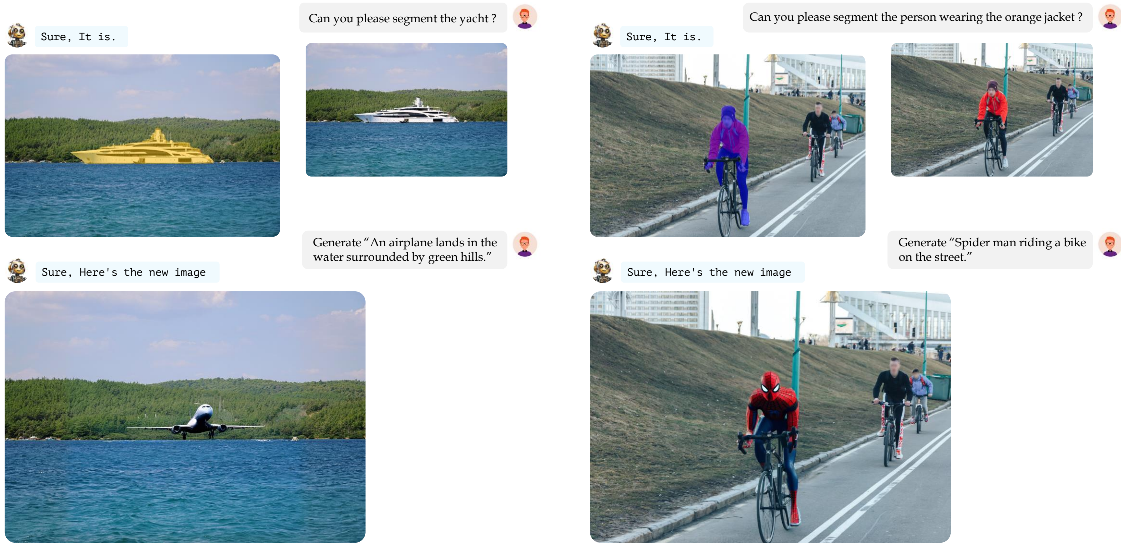

Figure 12. **Qualitative results of GLaMM on conditional image generation.** The figure shows the integration of GLaMM with an image generation model (stable diffusion). GlaMM first generates the segmentation mask (e.g. "yacht" in the left image and "person wearing orange jacket" in the right image) which is used along with a text prompt as input to the diffusion model to generate the desired images.

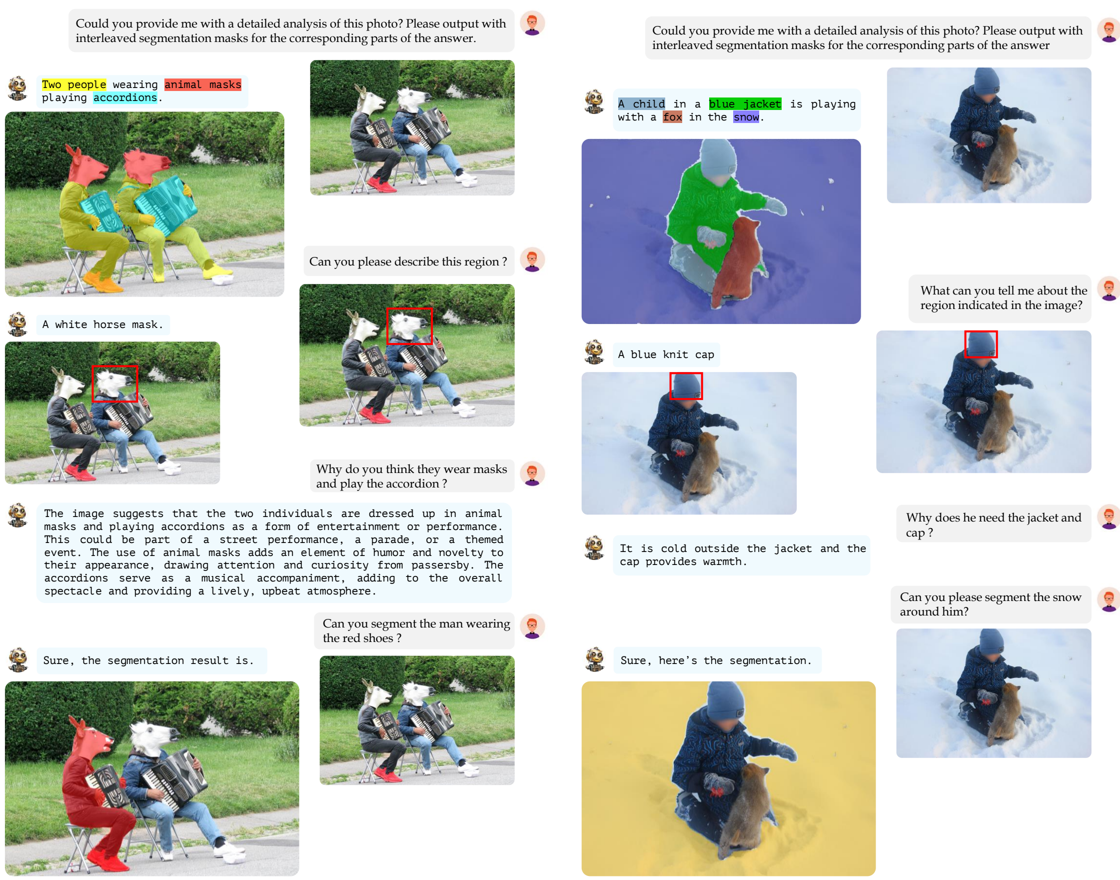

Figure 13. **Multimodal conversational with GLaMM.** The figure shows multimodal conversations generated through GLaMM. The model is flexible enough to process multimodal inputs and respond with multimodal outputs in a single conversation.

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'

'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'

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

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.

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

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 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. 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 [34] prompts used and the created dataset samples from Grand$_f$ dataset. This repurposed human-annotated dataset provides rich semantics to GLaMM for GCG task.

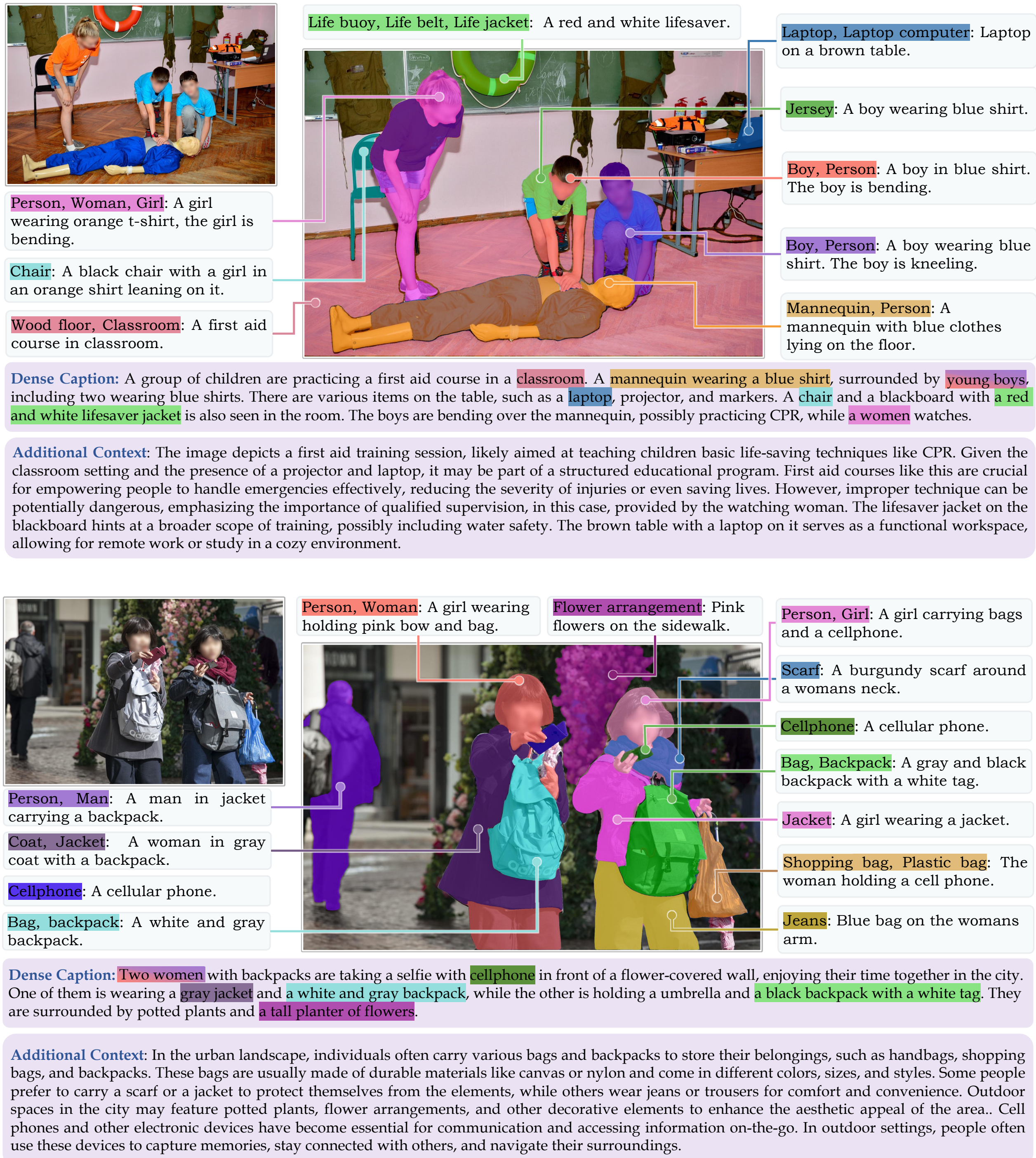

Figure 15. **Dataset samples from GranD.** The figure shows a few samples from the GranD dataset, generated using the automated annotation pipeline. It provides multiple semantic labels and attributes for detected objects, along with the grounded dense caption and additional context.

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.

D. 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.

E. Limitations and Future Work

The large-scale automated pipeline provides dense labelings that are important for our pretraining but still contains some noise. A high-quality, clean dataset could help further improve the pretrained representations, although this comes at a significantly higher annotation cost. A potential research direction is to develop a cost-effective annotation pipeline aimed at reducing noise in dense labeling. Additionally, expanding the GLaMM framework to include modalities such as video and 3D is also a future research direction.

F. 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.

G. Acknowledgement

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis partially funded by the Swedish Research Council through grant agreement no. 2022-06725, the LUMI supercomputer hosted by CSC (Finland) and the LUMI consortium, and by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.