# Instruction-guided Multi-Granularity Segmentation and Captioning with Large Multimodal Model

**Li Zhou**[1*], **Xu Yuan**[2*], **Zenghui Sun**[1], **Zikun Zhou**[3], **Jinsong Lan**[1†],

[1]TAO Technology, Alibaba Group
[2]The Hong Kong Polytechnic University
[3]Peng Cheng Laboratory
{pengye.zl,zenghui.szh,jinsonglan.ljs}@taobao.com, xuyuan127@gmail.com, zhouzikunhit@gmail.com


Figure 1: MGLMM is a versatile and sophisticated LMM, which can handle various tasks involving textual and pixel-level mask responses. We show its visualization results in the following scenarios: multi-granularity segmentation and captioning, referring segmentation, multiple/empty segmentation, panoptic segmentation, reasoning segmentation, image-level captioning, and conversation.

## Abstract

Large Multimodal Models (LMMs) have achieved significant progress by extending large language models. Building on this progress, the latest developments in LMMs demonstrate the ability to generate dense pixel-wise segmentation through the integration of segmentation models. Despite the innovations, the textual responses and segmentation masks of existing works remain at the instance level, showing limited ability to perform fine-grained understanding and segmentation even provided with detailed textual cues. To overcome this limitation, we introduce a Multi-Granularity Large Multimodal Model (MGLMM), which is capable of seamlessly adjusting the granularity of Segmentation and Captioning (SegCap) following user instructions, from panoptic SegCap to fine-grained SegCap. We name such a new task Multi-Granularity Segmentation and Captioning (MGSC). Observing the lack of a benchmark for model training and evaluation over the MGSC task, we establish a benchmark with aligned masks and captions in multi-granularity using our customized automated annotation pipeline. This benchmark comprises 10K images and more than 30K image-question pairs. We will release our dataset along with the implementation of our automated dataset annotation pipeline for further research. Besides, we propose a novel unified SegCap data format to unify heterogeneous segmentation datasets; it effectively facilitates learning to associate object concepts with visual features during multi-task training. Extensive experiments demonstrate that our MGLMM excels at tackling

*These authors contributed equally.
†Corresponding author.

more than eight downstream tasks and achieves state-of-the-art performance in MGSC, GCG, image captioning, referring segmentation, multiple and empty segmentation, and reasoning segmentation tasks. The great performance and versatility of MGLMM underscore its potential impact on advancing multimodal research. Code and dataset will be released at https://github.com/lizhou-cs/mglmm.

## Introduction

Leveraging the commonsense reasoning and understanding abilities of Large Language Models (LLMs) (Chiang et al. 2023; Touvron et al. 2023), Large Multimodal Models (LMMs) (Zhu et al. 2023; Alayrac et al. 2022; Bai et al. 2023; Liu et al. 2024a) have notably advanced cross-modality understanding and vision-language alignment.

Recently, several studies (Lai et al. 2024; Xia et al. 2024) have explored the instruction-based LMMs capable of producing pixel-level segmentation masks as responses to user queries. More recent researches (Rasheed et al. 2024; Zhang et al. 2024a) concentrated on Grounded Conversation Generation (GCG) which aims to ground the main objects appearing in the conversations. Although these methods (Zhang et al. 2024a; Lai et al. 2024; Xia et al. 2024; Ren et al. 2024) integrate a powerful segmentation model capable of panoptic segmentation, they still have difficulty generating mask-text-aligned responses for all the instances in the image, resulting in limited panoptic segmentation performance. Figure 2 (a) shows such a case where GLaMM overlooks the tennis racket, tennis ball and microphone in both mask and text responses. Besides, these models only possess the ability to describe the image at the instance level and produce corresponding instance masks aligned with the output texts. Hence, these models can hardly perceive the fine-grained objects, such as the hat, wristband, and skirt of the player in Figure 2 (b), even provided with detailed textual cues. The missing of the above abilities would limit the universality and comprehension of the LMMs.

To overcome these limitations, we introduce the Multi-Granularity LMM (MGLMM), which is capable of seamlessly adjusting the granularity of Segmentation and Captioning (SegCap) following user instructions, from panoptic SegCap to fine-grained SegCap. To be specific, for the query requiring describing the overall contents of an image, MGLMM outputs the precise panoptic segmentation masks with captions, offering a coarse-grained understanding of the entire image. For the instruction demanding to describe a certain object in the image, MGLMM can produce a detailed response including segmentation masks of the sub-parts of the object as well as corresponding descriptions, which reveal the components of the target object. We name such a task Multi-Granularity SegCap (MGSC), which assesses the ability of progressive cognition from coarse-grained to fine-grained. Overall, MGLMM excels at tackling more than eight downstream tasks such as panoptic SegCap, fine-grained SegCap, GCG, and multiple and empty segmentation, as presented in Figure 1 and Table 1.

Observing the lack of a benchmark for training and evaluating LMMs for the MGSC task in the community, we establish a new benchmark, dubbed MGSCData, with

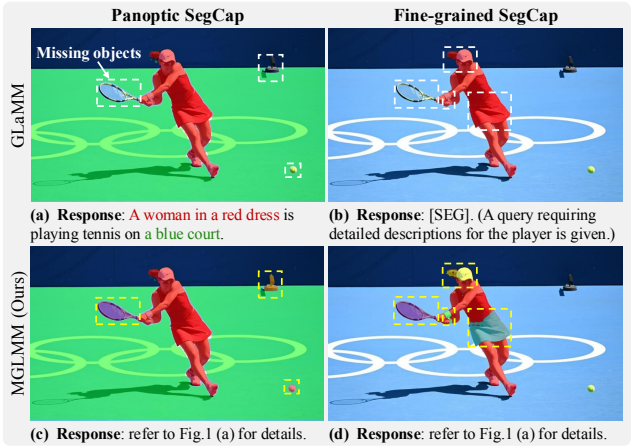


Figure 2: Qualitative comparison of GLaMM and our MGLMM. Please refer to **Appendix. A** for more details.

aligned masks and captions in multi-granularity using the customized automated annotation pipeline. It consists of 10K images and over 30K image-question pairs, encompassing both panoptic and fine-grained segmentation. To be more specific, the dataset includes more than 300K segmentation masks, each annotated with a semantic label and an accompanying detailed description. MGSCData effectively facilitates the training and assessment of the ability to associate object concepts and visual features in multi-granularity. We will release MGSCData and expect it to benefit academia.

Besides the benchmark, another key challenge in unifying segmentation tasks across granularities lies in the significant variation in both the format and semantic level of the queries and outputs. Typically, existing studies directly incorporate the heterogeneous data of different tasks into model training, overlooking the task discrepancies and complicating multimodal alignment further. To handle this issue, we propose a Unified SegCap Data Format (USCDF) to explicitly guide the model in learning the alignment relationships between object concepts and segmentation masks in different granularities during training. Specifically, USCDF unifies the output formats of different segmentation tasks, bridging the gap between them and reducing the difficulty of multi-task learning for the model. The right part of Figure 3 illustrates the instantiation of the unified data format on tasks including multi-referring reasoning, panoptic SegCap, and fine-grained SegCap. Experimental results demonstrate that USCDF benefits multi-task learning and vision-language learning. We also evaluate MGLMM across a variety of benchmarks. The experiments demonstrate that it achieves state-of-the-art results on six benchmarks.

In conclusion, our work has four main contributions:

- We propose MGLMM, the first model capable of seamlessly switching between multi-granularity segmentation and captioning, especially including panoptic and fine-grained segmentation and captioning.
- We introduce a novel benchmark MGSCData to train and evaluate the ability of multi-granularity segmentation and captioning for LMMs, which comprises over 30K high-

quality image-question pairs.

- We propose a unified data format, which facilitates learning the alignment relationships between object concepts and segmentation masks in multiple granularities.
- We achieve state-of-the-art performance across various tasks, including MGSC, GCG, image captioning, various segmentation tasks, etc.

## Related Work

Recently, there has been an increasing focus on fine-tuning pre-trained LLMs for visual instructions. These approaches, including BLIP-2 (Li et al. 2023), InstructBLIP (Dai et al. 2023), LLaVA (Liu et al. 2024b), MiniGPT-4 (Zhu et al. 2023), Qwen-VL (Bai et al. 2023), typically employ a pre-trained visual encoder to embedding visual input, utilize an LLM as the base model to comprehend user instructions and generate textual responses, and include an adapter to bridge the features of the vision encoder with those of the language model. The integration of visual and linguistic modalities within LLMs aims to enhance their capacity to understand and respond to complex, visually guided tasks. Although these methods have significantly facilitated the development of multimodal language models, their mechanisms fail to achieve pixel-level alignment and a comprehensive understanding of both images and language.

Furthermore, several works, including (Lai et al. 2024; Ren et al. 2024; Rasheed et al. 2024; Zhang et al. 2024a), explore more complex tasks driven by instructions, involving segmentation or captioning as responses to achieve effective pixel-level alignment of images and text. Although these methods perform well in various segmentation tasks, they are limited to learning only instance-level vision-language alignment, preventing them from perceiving fine-grained objects. Furthermore, all these methods integrate a mask decoder capable of panoptic segmentation into their methods but fail to generate coherent mask-text-align responses, resulting in suboptimal performance.

For the reasons mentioned above, our goal is to develop an LMM that can seamlessly perform panoptic and fine-grained segmentation and captioning based on user instructions. Further, we establish a high-quality benchmark called MGSC that fills the gap for panoptic and fine-grained segmentation and captioning and introduce our automated annotation pipeline. Last, we propose a unified data format that facilitates explicit learning of alignment relationships between object concepts and segmentation masks. MGLMM achieves state-of-the-art performances on over six tasks and ablation results also prove the effectiveness of our methods.

## Method

In this section, we introduce the model architecture of our MGLMM, as illustrated in Figure 3. We then introduce the unified SegCap data format used during training.

### Model Architecture

To achieve multi-granularity segmentation and captioning, we utilize two foundational models to construct our model: (1) an LMM for comprehending input images and user instructions and generating natural language responses, and (2) a segmentation model based on an encoder-decoder architecture for pixel-level visual understanding.

**Large Multimodal Model.** Considering the simplicity and consistency with previous works (Lai et al. 2024; Rasheed et al. 2024), LLaVA emerges as our preferred choice. Specifically, we employ the CLIP model as the vision encoder, denoted as $\mathcal{F}_v$, in conjunction with the Vicuna-7B model as a decoder-based LLM, denoted as $\mathcal{F}_{llm}$. As illustrated in Figure. 3, the vision encoder is responsible for extracting visual features from the input image $x_{img}$, after which a projector $\phi$ is applied to map the extracted image features into the word embedding space of $\mathcal{F}_{llm}$. Formally:

$$z_{img} = \phi(\mathcal{F}_v(x_{img})). \tag{1}$$

It is worth noting that the projector $\phi$ plays a crucial role in aligning image features with the linguistic modality. Specifically, it consists of two linear layers with a GELU nonlinearity and is initialized randomly. Meanwhile, the text input is encoded into text tokens by the tokenizer $T$ of $\mathcal{F}_{llm}$. Subsequently, we integrate image tokens $z_{img}$ and text tokens $z_{txt}$, which are then fed into the $\mathcal{F}_{llm}$ to generate final textual output $y_{txt}$, *i.e.*,

$$\hat{y}_{txt} = \mathcal{F}_{llm}(z_{img} \| z_{txt}). \tag{2}$$

Following LISA (Lai et al. 2024), we adopt the embedding-as-mask paradigm to bridge these two modules. In this paradigm, the vocabulary of the model is augmented with a specialized token '[SEG]', designed to explicitly activate the segmentation behavior of the segmentation model. When the LMM intends to generate a segmentation mask based on the user instruction, it inserts the '[SEG]' token in the output sequence $y_{txt}$ to indicate the presence of a target to segment. For example:

> **User**: <IMAGE> Please segment the dog in this image.
> **Assistant**: Sure, the segmentation result is dog [SEG].

**Segmentation Model.** This work employs SAM (Kirillov et al. 2023) as our foundation segmentation architecture because of its promising pixel-level modeling capability. As shown in Figure. 3, the pixel encoder $\mathcal{E}_{pixel}$ is instantiated using a frozen SAM encoder, while the pixel decoder $\mathcal{D}_{pixel}$ is initialized from the pre-trained SAM decoder. The former takes the high-resolution image as input to extract fine-grained visual information, while the latter generates the desired segmentation masks prompted by the embedding of the '[SEG]' token from the LLM. Specifically, we select the output embedding $\hat{z}_{seg}$ corresponding to the '[SEG]' token $\hat{y}_{txt}([SEG])$ and transform it into the feature space of decoder using a projector $\psi$. Notably, the structure and initialization of projector $\psi$ are identical to those of projector $\phi$. The entire process can be formulated as:

$$\hat{y}_{mask} = \mathcal{D}_{pixel}(\mathcal{E}_{pixel}(x_{img}), \psi(\hat{z}_{seg})). \tag{3}$$

| Method | Textual Response | | Mask Response | | | Textual & Mask Response | | |
|---|---|---|---|---|---|---|---|---|
| | Caption | Conversation | Referring Seg | Generic Seg | Multiple/Empty Seg | Reasoning Seg | GCG | MGSC |
| LISA (Lai et al. 2024) | ✓ | ✓ | ✓ | | | | | |
| PixelLM (Ren et al. 2024) | ✓ | ✓ | ✓ | | ✓ | ✓ | | |
| GSVA (Xia et al. 2024) | ✓ | ✓ | ✓ | | ✓ | | | |
| Osprey (Yuan et al. 2024) | ✓ | ✓ | | | | | | |
| LaSagnA (Wei et al. 2024) | | | ✓ | ✓ | ✓ | | | |
| PSALM (Zhang et al. 2024b) | | | ✓ | ✓ | ✓ | | | |
| OMG-LLaVa (Zhang et al. 2024a) | ✓ | ✓ | ✓ | ✓ | | | ✓ | |
| GLaMM (Rasheed et al. 2024) | ✓ | ✓ | ✓ | | | | ✓ | |
| MGLMM (Ours) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison of the capabilities of MGLMM with multiple representative methods. Here, "Generic Seg" comprises semantic segmentation, instance segmentation, and panoptic segmentation; "Reasoning Seg" requires the model to segment images based on queries involving complex reasoning and provide the corresponding textual interpretation.

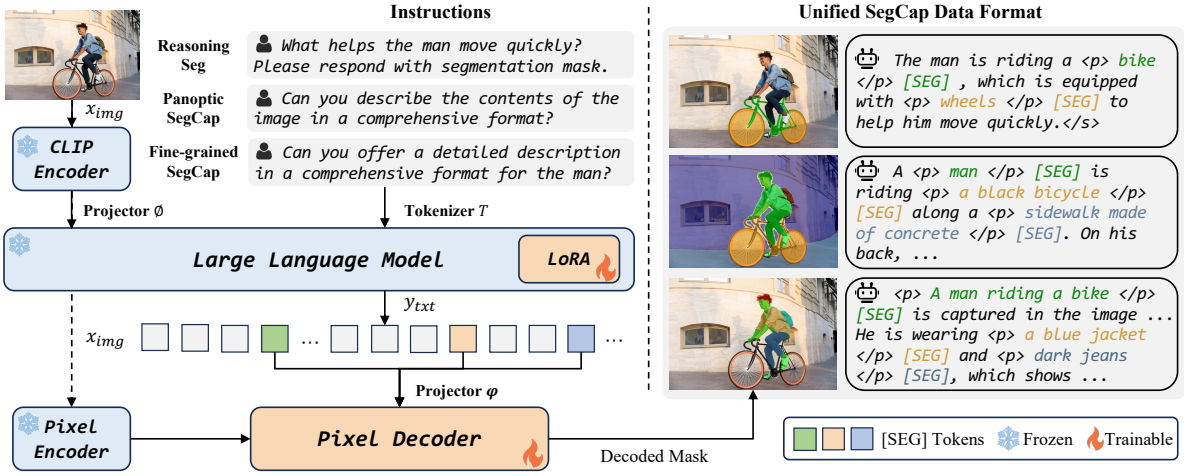


Figure 3: **Left**: The model architecture of MGLMM. **Right**: The proposed unified data format for multi-task learning.

## Design of Unified SegCap Data Format

Most existing studies primarily integrate various pixel-level segmentation capabilities into LMMs by directly extending corresponding task datasets. For example, in referring segmentation, the query may be a phrase that requires the return of segmentation masks. Conversely, in reasoning segmentation, the query can be a longer sentence or question in which the target may not be present, necessitating an answer along with segmentation masks. In different segmentation tasks, the form and semantics of queries vary. In this context, the model must adaptively align the semantic concepts of potential targets with visual features during training, which undoubtedly increases the burden on model learning. Therefore, we propose a unified SegCap data format to leverage these data, explicitly guiding the model toward improved vision-language alignment. In this manner, we unify the output formats of different segmentation tasks, bridging the gap between them and reducing the difficulty of multi-task learning for the model. Specifically, apart from the '[SEG]' token, we also introduce  and  tokens to the vocabulary of the LMM to denote the start and end of the corresponding phrases of the segmentation mask, respectively. The LLM is required to mark the corresponding description with  and  while activating the segmentation behavior using '[SEG]'. The following is an example of data format for multi-referring segmentation:

> **User**: <IMAGE> Please segment the {obj-1}, {obj-2}, ..., and {obj-n} in this image.
> **Assistant**: Sure,  {obj-1}  [SEG],  {obj-2}  [SEG], ..., and  {obj-n}  [SEG].

Here, <IMAGE> denotes the placeholder for image tokens. {obj-n} represents the semantic description of the corresponding segmentation targets.

In contrast to previous work, such a unified data pattern enables the model to explicitly learn the alignment relationships between the object concepts and the segmentation masks during training. Despite the fact that GLaMM (Rasheed et al. 2024) had adopted a similar format, it was only employed for the GCG task it presented. In contrast, we utilize this unified schema for all tasks, which reduces the modeling burden by minimizing the differences in output formats across tasks. In Figure. 3, we demonstrate our unified data format on tasks such as reasoning and multi-granularity segmentation. Notably, during the training phase, we convert the annotation format of some existing open-source datasets into the proposed unified data schema as they do not meet our requirements. For more details on this process, please refer to **Appendix. C**.

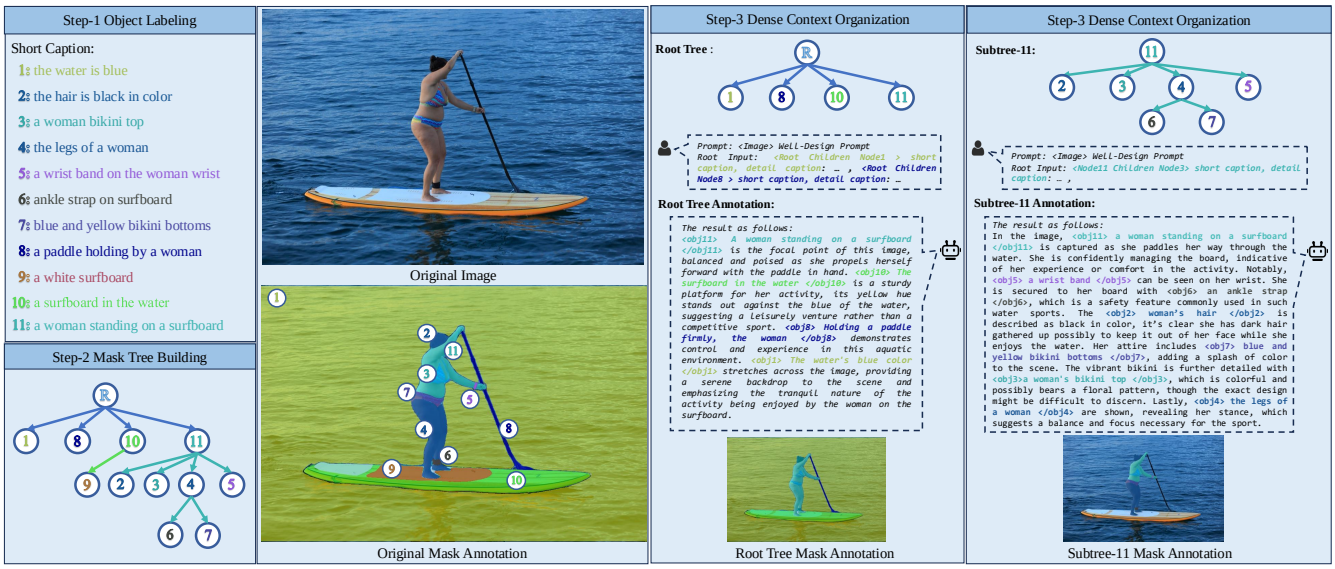


Figure 4: The overview of our proposed data auto-annotated pipeline. Due to space limitations, the detailed caption is not shown in the figure. Please refer to the **Appendix. B** for the detailed version. Best viewed with zoom-in.

## Data Annotation Pipeline

Most existing segmentation datasets focus on instance-level objects, and although the SAM dataset provides fine-grained segmentation mask annotations, it lacks corresponding text descriptions. Therefore, to address the issue of insufficient benchmarks for evaluating models in multi-granular segmentation and captioning, we propose a novel task called Multi-Granularity SegCap. To build up this benchmark, we came up with an automated annotation pipeline that allows us to leverage the capabilities of LMMs, specifically the GPT-4 and Qwen-VL series, for data labeling. In the following section, we will introduce our automatic annotation pipeline, designed to seamlessly transform any segmentation dataset. This pipeline consists of three main steps, as illustrated in Figure 4. The first step focuses on generating short captions and detailed captions for each masked target, known as object labeling. Subsequently, the second step constructs tree relationships based on the segmentation masks. The third step organizes various levels of granular information by utilizing the raw data from different levels of the subtree. As a result, we achieve multi-granularity segmentation and captioning annotations that demonstrate high alignment between visual and textual concepts. Since the SAM (Kirillov et al. 2023) dataset provides hundreds of millions of high-quality images and fine-grain segmentation, we perform our automated pipeline on the SAM dataset.

### Object Labeling

In step 1, the key point is generating a short caption and detailed caption for each target in the images. The short caption is used as a semantic representation of the target. The detailed caption is a comprehensive and semantically rich textual representation of the target, which is primarily used to provide a reference representation to limit the divergence and randomness of LMMs. In practice, we leverage the GPT-4o to create instruction-following data to generate the semantic label of each masked object.

### Mask Tree Building

After obtaining the semantic labels of each target, we need to organize the hierarchical relationships between each target within the image. We discover that the hierarchical relationships between the targets could be effectively reflected by the Intersection of Union (IoU) relationships among the masks. Therefore, we denote the entire image as the root node and then extend the tree according to the inclusion relationship between masks. Besides, in the SAM dataset, numerous mask annotations exist within a single image, many of which share the same semantics labels. For example, in a building with many windows, each window is represented as an individual mask with the same short captions. For such nodes that share the same parent node, we merge the nodes and their masks. In this manner, we obtain a simple and hierarchical tree and significantly shorten the length of the resulting text annotations.

### Dense Context Organization

The generation of multi-granularity captions is based on the mask tree which provides semantic labels of each target and hierarchical relationships between them. First, we utilize the semantic labels of child nodes of the root node to generate an ordered text input which mainly includes the instance-level objects in the image, which aims to create a coarse-grained caption for the entire picture. Subsequently, we concatenate the well-designed prompt, the ordered text input, and the image to prompt GPT-4o and obtain an organized description in which each target is embedded in a natural and coherent sequence. We apply the same process on each subtree under the root node. In particular, we use all the descendant

| Method | Textual Response | | Mask Response | | | | Textual & Mask Response | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Flickr30k CIDEr | NoCap CIDEr | refCOCO+ cIoU | refCOCOg cIoU | gRefCOCO cIoU | reasonSeg cIoU | GCG CIDEr | GCG AP50 | MGSC CIDEr | MGSC AP50 |
| LISA (Lai et al. 2024) | – | – | 65.1 | 67.9 | – | 46.0 | 33.9 | 25.2 | – | – |
| PixelLM (Ren et al. 2024) | – | – | 66.3 | 69.3 | – | – | – | – | – | – |
| GSVA (Xia et al. 2024) | – | – | 65.9 | 72.7 | – | – | – | – | – | – |
| LaSagnA (Wei et al. 2024) | – | – | 66.4 | 70.6 | 38.1 | 47.2 | – | – | – | – |
| PSALM (Zhang et al. 2024b) | – | – | 72.9 | 73.8 | 42.0 | – | – | – | – | – |
| OMG-LLaVA (Zhang et al. 2024a) | – | – | 69.1 | 72.9 | – | – | 41.2 | 29.9 | – | – |
| GLaMM (Rasheed et al. 2024) | 95.3 | 106.8 | 72.6 | 74.2 | – | – | 47.2 | 30.8 | 8.7 | 5.4 |
| MGLMM (Ours) | **104.6** | **112.6** | **73.9** | **77.2** | **52.8** | **51.1** | **50.1** | **31.7** | **11.6** | **7.4** |

Table 2: The comprehensive comparison of MGLMM and other LMMs in terms of text description and pixel-level understanding capabilities. "–" indicates that the method does not handle this task.

nodes of the subtree to build up a description aiming to obtain a fine-grained description of the specific target. Through such a construction process, we obtain panoptic segmentation masks with aligned descriptions for each instance-level target, as well as fine-grained segmentation masks with aligned descriptions for the specific target in each image.

In this manner, we annotate 10K SAM images, which are inherently diverse and exhibit multi-granularity. The resulting dataset comprises 30K conversations and contains over 45M tokens, totaling more than 300K segmentation masks, each accompanied by a short semantic label and a detailed caption. For more details about the pipeline and dataset, please refer to the **Appendix. B**.

# Experiments

## Experimental Settings

**Datasets.** To achieve all the capabilities of MGLMM, our training dataset is composed of six parts: (1) semantic segmentation: including ADE20K (Zhou et al. 2019), COCO-Stuff (Caesar, Uijlings, and Ferrari 2018), Maplilary Vistas (Neuhold et al. 2017), PACO-LVIS (Ramanathan et al. 2023), and PASCAL-Part (Chen et al. 2014) ; (2) referring segmentation: including RefCLEF (Jing et al. 2021) the RefCOCO series (Yu et al. 2016); (3) image-level caption: including COCO Caption (Chen et al. 2015); (4) visual question answering: including LLaVA-150k (Liu et al. 2024b) (5) grounded conversation generation including GranDf. Additionally, we also use approximately 4M captioning and referring segmentation data from Grounding-anything Dataset (GranD) [1] dataset published by GLaMM (Rasheed et al. 2024), which is annotated automatically on SAM (Kirillov et al. 2023) images. (6) multi-granularity SegCap, including MGSCData, which we proposed.

**Implementation details.** In our experiments, we use Vicuna-7B as a structure for LLM except for some ablations. We train our model on 16 Tesla A100 GPUs (80GB) for 30,000 iterations with a batch size of 16 per device. Unless otherwise specified, the model is trained with a joint training setting and without additional task-specific fine-tuning. Following the previous works, we apply the CE loss for modeling text generation, and the BCE and DICE loss to supervise high-quality mask prediction. Further implementation details, particularly regarding LORA fine-tuning, the optimizer, hyperparameter settings, and training objectives, can be found in the **Appendix. D**.

[1] Although GranD contains 11M images, only 4M are available because the authors have yet to publicize all the data.

| Model | Multi-Granularity SepCap | | | | | GCG | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | M | C | AP50 | mIoU | MR | M | C | AP50 | mIoU | MR |
| Kosmos-2 | – | – | – | – | – | 16.1 | 27.6 | 17.1 | 55.6 | 28.3 |
| LISA | – | – | – | – | – | 13.0 | 33.9 | 25.2 | 62.0 | 36.3 |
| OMG-LLaVA | – | – | – | – | – | 14.9 | 41.2 | 29.9 | 65.5 | – |
| GLaMM | 16.5 | 8.7 | 5.4 | 47.6 | 18.7 | 16.2 | 47.2 | 30.8 | **66.3** | 41.8 |
| MGLMM (Ours) | **17.8** | **11.6** | **7.4** | **51.6** | **23.2** | **16.4** | **50.1** | **31.7** | **66.3** | **45.2** |

Table 3: Performance comparison on MGSC and GCG. Following the evaluation protocol of GCG, we report the metrics including METEOR (M), CIDEr (C), AP50, mIoU, and Mask Recall (MR).

**Comparisons with State-of-the-Arts.** As shown in Table 2, we compare our MGLMM with other representative methods on various kinds of tasks and outperform all tasks. Then, we evaluate the effectiveness of our MGLMM on the following six benchmarks. Additionally, we will provide more discussion of the experimental results in the Appendix. E.

**Multi-Granularity SegCap.** The MGSC aims to evaluate the ability to seamlessly adjust the granularity of segmentation and captioning. Following the same settings, we finetune the GLaMM and our MGLMM on the training set of MGSCData and evaluate them on the same metric. As shown in Table 3, we outperform GLaMM on every metric, demonstrating the impressive capabilities of our MGLMM in multi-granularity SegCap.

**Grounded Conversation Generation (GCG).** Following GLaMM, we finetune our model on the GranDf dataset. As shown in Table 3, our MGLMM outperforms other approaches in terms of both image description and pixel understanding capabilities. It is worth noting that, despite more training data utilized by GLaMM in the pre-training phase compared to MGLMM, the latter still surpasses the former, particularly in terms of the CIDEr and Mask Recall scores.

**Referring Segmentation.** Table 4 compares our MGLMM with current state-of-the-art models on three representative datasets. We achieve significant lead performances over recent works like GLaMM, and OMG-LLaVG on the refCOCO/+/g validation and test sets in Table 4. Notably,

| Type | Model | refCOCO | | | refCOCO+ | | | refCOCOg | | ReasonSeg | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | val | testA | testB | val | testA | testB | val | test | cIoU | gIoU |
| Segmentation Specialist | LAVT (Yang et al. 2022) | 72.7 | 75.8 | 68.8 | 62.1 | 68.4 | 55.1 | 61.2 | 62.1 | – | – |
| | ReLA (Liu et al. 2023a) | 73.8 | 76.5 | 70.2 | 66.0 | 71.0 | 57.7 | 65.0 | 66.0 | – | – |
| | PolyFormer (Liu et al. 2023b) | 74.8 | 76.6 | 71.1 | 67.6 | 72.9 | 59.3 | 67.8 | 69.1 | – | – |
| LMM-based Models | LISA (Lai et al. 2024) | 74.9 | 79.1 | 72.3 | 65.1 | 70.8 | 58.1 | 67.9 | 70.6 | 46.0 | 34.1 |
| | PixelLM (Ren et al. 2024) | 73.0 | 76.5 | 68.2 | 66.3 | 71.7 | 58.3 | 69.3 | 70.5 | – | – |
| | GSVA (Xia et al. 2024) | 77.2 | 78.9 | 73.5 | 65.9 | 69.6 | 59.8 | 72.7 | 73.3 | – | – |
| | LaSagnA (Wei et al. 2024) | 76.8 | 78.7 | 73.8 | 66.4 | 70.6 | 60.1 | 70.6 | 71.9 | 47.2 | – |
| | OMG-LLaVA (Zhang et al. 2024a) | 78.0 | 80.3 | 74.1 | 69.1 | 73.1 | 63.0 | 72.9 | 72.9 | – | – |
| | GLaMM (Rasheed et al. 2024) | 79.5 | 83.2 | 76.9 | 72.6 | 78.7 | 64.6 | 74.2 | 74.9 | – | – |
| | MGLMM (Ours)† | 80.2 | 83.1 | 76.0 | 73.2 | 78.7 | 66.8 | 76.7 | **77.5** | **51.1** | **48.6** |
| | MGLMM (Ours) | **81.3** | **83.5** | **77.3** | **73.9** | **79.2** | **67.2** | **77.2** | 77.4 | – | – |

Table 4: Performance on referring and reasoning segmentation benchmarks. The table only shows the cIoU values for referring segmentation. MGLMM† indicates that the referring segmentation dataset is used only in the pre-training phase.

| Model | zero shot | Generalized Referring Segmentation | | | | | |
|---|---|---|---|---|---|---|---|
| | | val | | testA | | testB | |
| | | cIoU | gIoU | cIoU | gIoU | cIoU | gIoU |
| ReLA (Liu et al. 2023a) | ✗ | 62.4 | 63.6 | 69.3 | 70.0 | 59.9 | 61.0 |
| LISA†(Lai et al. 2024) | ✗ | 38.7 | 32.2 | 52.6 | 48.5 | 44.8 | 39.7 |
| LISA (Lai et al. 2024) | ✗ | 61.7 | 61.6 | 69.2 | 70.1 | 60.3 | 61.3 |
| GSVA†(Xia et al. 2024) | ✗ | 61.7 | 63.3 | 69.2 | 70.1 | 60.3 | 61.3 |
| GSVA (Xia et al. 2024) | ✗ | 63.3 | 66.5 | 69.9 | 71.1 | 60.5 | 62.2 |
| LaSagnA (Wei et al. 2024) | ✓ | 38.1 | 32.4 | 50.4 | 47.3 | 42.1 | 38.9 |
| PSALM (Zhang et al. 2024b) | ✓ | 42.0 | 43.3 | 52.4 | 54.5 | 50.6 | 52.5 |
| MGLMM (Ours) | ✓ | **52.8** | **50.2** | **61.2** | **58.7** | **56.0** | **54.1** |

Table 5: Performance comparison on generalized referring-expression segmentation with cIoU and gIoU metrics. LISA† and GSVA† exclusively use the gRefCOCO dataset during the pre-training phase, while MGLLM performs zero-shot learning.

| Model | Flickr30k | | NoCap | |
|---|---|---|---|---|
| | CIDEr | SPICE | CIDEr | SPICE |
| LEMON (Hu et al. 2022) | – | – | 106.8 | 14.1 |
| CoCa (Yu et al. 2022) | – | – | 120.6 | 15.5 |
| BLIP-2 (Li et al. 2023) | – | – | 121.6 | **15.8** |
| InstructBLIP (Dai et al. 2023) | 82.8 | – | **123.1** | – |
| Kosmos-1 (Huang et al. 2024) | 67.1 | 14.5 | – | – |
| Kosmos-2 (Peng et al. 2023) | 66.7 | – | – | – |
| GLaMM (Rasheed et al. 2024) | 95.3 | 18.8 | 106.8 | **15.8** |
| MGLMM (Ours) | **104.6** | **22.7** | 112.6 | 15.2 |

Table 6: Performance comparison on image-level captioning.

even without any fine-tuning on the referring segmentation dataset (MGLMM† in Table 4), our approach still surpasses GLaMM on the validation split of all benchmarks.

**Generalized Referring Segmentation and Reasoning Segmentation.** The results are shown in Table 5. Compared with PSLAM (Zhang et al. 2024b), the state-of-the-art method in the zero-shot setting, our MGLMM accomplishes average boosts of 6.0% and 6.5% in terms of cIoU and gIoU, respectively. Notably, MGLMM even outperforms LISA† in all cases, which incorporate gRefCOCO during the pre-training phase. For reasoning segmentation, we utilize the validation set of ReasonSeg dataset (Lai et al. 2024) as the benchmark. From the results reported in Table 4, we can observe that the reasoning proficiency of MGLMM surpasses that of other methods.

**Image-level Captioning.** To investigate this capability, we finetune MGLMM on the Flickr-30K (Plummer et al. 2015) and evaluate Flickr-30K and NoCap (Agrawal et al. 2019), where the latter can be considered as a **zero-shot** scene. As reported in Table 6, MGLMM is superior to the counterpart model GLaMM on several metrics.

| Model | + USCDF | + GranD Dataset | refCOCO+ | | | GCG | |
|---|---|---|---|---|---|---|---|
| | | | val | testA | testB | C | mIoU |
| MGLMM-7B | | | 67.2 | 74.1 | 58.9 | 46.5 | 65.3 |
| MGLMM-7B | ✓ | | 69.9 | 76.2 | 62.5 | 46.3 | 65.6 |
| MGLMM-7B | | ✓ | 71.4 | 76.9 | 64.0 | 48.0 | 66.2 |
| MGLMM-7B | ✓ | ✓ | 73.2 | 78.7 | 66.8 | 50.1 | 66.3 |
| MGLMM-13B | ✓ | ✓ | **73.4** | **79.8** | **68.0** | **50.5** | **66.4** |

Table 7: Ablation study results. For refCOCO+, we utilize cIoU as the metric. 'C' denotes the CIDEr score. We implement MGLMM-13B using Llama2-13B as the structure for LLM.

## Ablation Studies

To perform a thorough ablation study, we assess different variants of MGLMM using two representative benchmarks, *i.e.*, referring segmentation and GCG, which can demonstrate the models' ability to understand pixel-level details and provide image descriptions. For more details, please refer to **Appendix. E**.

**Effectiveness of USCDF.** Compared to the 1st variant in Table 7, MGLMM using USCDF obtains an improvement of more than 2% on challenging regCOCO+ benchmark. The performance difference between the 3rd and 4th variants is significant, as GranD is four times larger than the other pre-training data, which further amplifies the gains of USCDF.

**Influence of GranD dataset.** To investigate the impact of the extra GranD dataset on MGLMM, we experiment without 4M GranD samples. Comparing the 2nd and 4th variants in Table 7, we can find that the GranD dataset contributes a gain. Despite not utilizing GranD, our MGLMM remains superior to models such as OMG-LLAVA in most cases, ranking second only to GLaMM, which employed over ten times training data during the pre-training phase.

## Conclusion

We propose MGLMM, the first model capable of seamlessly adjusting the granularity of segmentation and cap-

tioning following user instructions. Realizing the lack of multi-granularity of segmentation and captioning dataset and benchmark, we introduce a novel benchmark MGSC-Data to train and evaluate the ability of multi-granularity segmentation and captioning for LMMs, which comprises over 30K high-quality image-question pairs. To facilitate aligning object concepts with visual features during various segmentation tasks, we propose a unified data format. Our model excels at tackling more than eight downstream tasks and outperforms various benchmarks.

## References


Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8948–8957. 7

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736. 2

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966. 2, 3

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1209–1218. 6

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*. 6

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1971–1978. 6

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. 2

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500. 3, 7

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17980–17989. 7

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O. K.; Patra, B.; et al. 2024. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36. 7

Jing, Y.; Kong, T.; Wang, W.; Wang, L.; Li, L.; and Tan, T. 2021. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9858–9867. 6

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026. 3, 5, 6

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589. 2, 3, 4, 6, 7

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597. 3, 7

Liu, C.; Ding, H.; Jiang, X.; and Liu, C. 2023a. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23592–23601. 7

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744. 2

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36. 3, 6

Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Satzoda, R. K.; Mahadevan, V.; and Manmatha, R. 2023b. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18653–18663. 7

Neuhold, G.; Ollmann, T.; Rota Bulo, S.; and Kontschieder, P. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, 4990–4999. 6

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*. 7

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

Ramanathan, V.; Kalia, A.; Petrovic, V.; Wen, Y.; Zheng, B.; Guo, B.; Wang, R.; Marquez, A.; Kovvuri, R.; Kadian, A.; et al. 2023. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7141–7151. 6

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13009–13018. 2, 3, 4, 6, 7

Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26374–26383. 2, 3, 4, 6, 7

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

Wei, C.; Tan, H.; Zhong, Y.; Yang, Y.; and Ma, L. 2024. LaSagnA: Language-based Segmentation Assistant for Complex Queries. *arXiv preprint arXiv:2404.08506*. 4, 6, 7

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3858–3869. 2, 4, 6, 7

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18155–18165. 7

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*. 7

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 69–85. Springer. 6

Yuan, Y.; Li, W.; Liu, J.; Tang, D.; Luo, X.; Qin, C.; Zhang, L.; and Zhu, J. 2024. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28202–28211. 4

Zhang, T.; Li, X.; Fei, H.; Yuan, H.; Wu, S.; Ji, S.; Loy, C. C.; and Yan, S. 2024a. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*. 2, 3, 4, 6, 7

Zhang, Z.; Ma, Y.; Zhang, E.; and Bai, X. 2024b. PSALM: Pixelwise SegmentAtion with Large Multi-Modal Model. *arXiv preprint arXiv:2403.14598*. 4, 6, 7

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321. 6

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592. 2, 3