Supplementary Material: Instruction-guided Multi-Granularity Segmentation and Captioning with Large Multimodal Model

We have included supplementary material to facilitate a more comprehensive understanding and in-depth analysis of the primary paper. The supplementary material is organized as follows:

A. Motivation
B. Pipeline And Datasets
C. Unified SegCap Data Format
D. Implementation Details
E. More Experiments
F. Additional Qualitative Results

A. Motivation

Limited to the space constraints of main paper, we present a detailed comparison figure here to illustrate our motivation. Although existing methods (Lai et al. 2024; Zhang et al. 2024; Ren et al. 2024; Rasheed 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. Here we take GLaMM as an example, we compare qualitative results our MGLMM and GLaMM on Panoptic and Fine-grained SegCap. As shown in Figure 1, it illustrates such a case where GLaMM overlooks the tennis racket, tennis ball, and microphone in both mask and text responses. However, our MGLMM can perceive these targets well and caption them. Further, even we provide a specific prompt like “The woman is wearing a hat and dress, holding a rocket.”, GLaMM still can hardly perceive the fine-grained objects, such as the hat, wristband, and skirt of the player in Figure 1, even provided with detailed textual cues. The missing of the above abilities would limit the universality and comprehension of the LMMs.

B. Pipeline and Datasets

In this section, we provide a more detailed example version to illustrate the construction process of MGSCData.

B.1. Data Labeling

As shown in Figure 2, we illustrate how we leverage the GPT-4o to generate the short caption and detailed caption. Through this way, we obtain the semantic label of each masked object. Figure 3 shows all semantic labels about each object in the image.

B.2. Dense Context Organization.

Figures 4 and 5 show how we use GPT-4o to organize our semantic labels to obtain a caption that is aligned with the mask. In this process, we primarily utilize GPT-4o to reorganize the input information, enabling the generation of a coherent caption that aligns with the designated mask.

C. Unified SegCap Data Format

This section will introduce how to convert the semantic, referring, reasoning, and panoptic segmentation datasets into the proposed Unified SegCap Data Format (USCDF). Specifically, the GranD Dataset and MGSCData have already been formatted to comply with the unified SegCap data format.

C.1. Semantic/Referring Segmentation

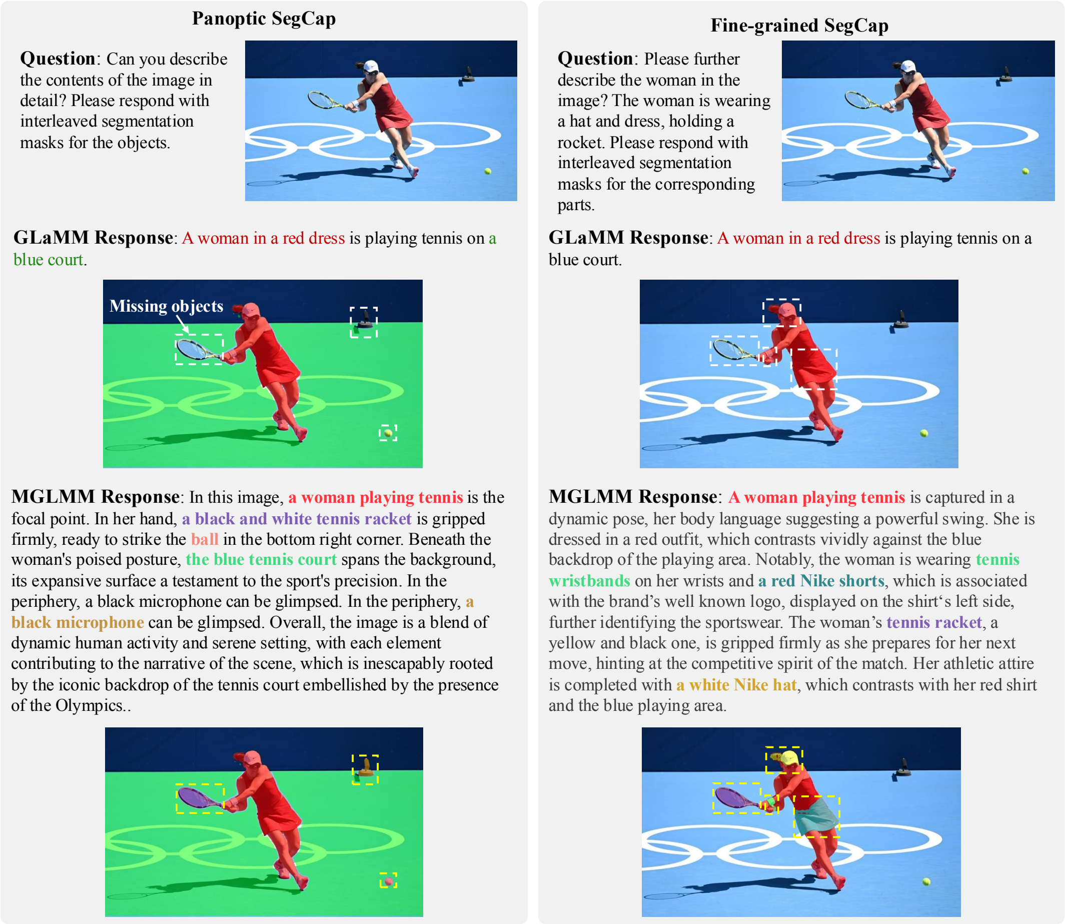
Semantic and referring segmentation datasets typically consist of images, corresponding labels, or explicit short descriptions of the target objects. During training, previous work (Lai et al. 2024; Rasheed et al. 2024) employ the following question-answer template to train their models.

User: <IMAGE> Can you segment the {object} in this image.
Assistant: Sure, it is [SEG].

Here, <IMAGE> denotes the placeholder for tokens of image patches, and {object} is the chosen semantic category or referring description.

To transform this template into our USCDF, which can explicitly align object concepts with their corresponding segmentation masks in the output format, we need only to embed the semantic category or referring description encapsulated within <p> and </p> before the [SEG] token. For example:

User: <IMAGE> Can you segment the {object} in this image.
Assistant: Sure, it is <p> {object} </p> [SEG].


Panoptic SegCap

Question: Can you describe the contents of the image in detail? Please respond with interleaved segmentation masks for the objects.

GLaMM Response: A woman in a red dress is playing tennis on a blue court.

MGLMM Response: In this image, **a woman playing tennis** is the focal point. In her hand, **a black and white tennis racket** is gripped firmly, ready to strike the **ball** in the bottom right corner. Beneath the woman's poised posture, **the blue tennis court** spans the background, its expansive surface a testament to the sport's precision. In the periphery, a black microphone can be glimpsed. In the periphery, **a black microphone** can be glimpsed. Overall, the image is a blend of dynamic human activity and serene setting, with each element contributing to the narrative of the scene, which is inescapably rooted by the iconic backdrop of the tennis court embellished by the presence of the Olympics..

Fine-grained SegCap

Question: Please further describe the woman in the image? The woman is wearing a hat and dress, holding a rocket. Please respond with interleaved segmentation masks for the corresponding parts.

GLaMM Response: A woman in a red dress is playing tennis on a blue court.

MGLMM Response: **A woman playing tennis** is captured in a dynamic pose, her body language suggesting a powerful swing. She is dressed in a red outfit, which contrasts vividly against the blue backdrop of the playing area. Notably, the woman is wearing **tennis wristbands** on her wrists and **a red Nike shorts**, which is associated with the brand's well known logo, displayed on the shirt's left side, further identifying the sportswear. The woman's **tennis racket**, a yellow and black one, is gripped firmly as she prepares for her next move, hinting at the competitive spirit of the match. Her athletic attire is completed with **a white Nike hat**, which contrasts with her red shirt and the blue playing area.

Figure 1: The visualization results of GLaMM and our MGLMM On Panoptic and Fine-grained SegCap.

The format of visual question answering in the above template contains only a single target existing in the image, which hinders LMMs from learning to segment multiple targets and reject non-existing referents. However, it is significant for LMMs to handle multiple/empty targets in real-world applications. For this reason, we convert 30% of the single-reference samples to multi-reference samples by randomly sampling positive and negative semantic categories or reference expressions. Using referring segmentation as an example, an image typically encompasses multiple target objects adorned with referring expressions, from which we randomly select a certain number as positive referring targets. As for the empty referents, we utilize the referring expressions found in the RefCLEF dataset (RefCOCO series) as negative referring candidates of the RefCOCO series (RefCLEF dataset) and sample negative references from them. We use {obj-n} and {neg-m} denote the positive and negative referring targets, respectively. The template supporting multiple/empty segmentation is shown as follows:

User: <IMAGE> Please segment the {obj-1}, {neg-1}, {neg-2}, {obj-2}, ..., {neg-m}, and {obj-n} in this image.
Assistant: Sure, <p> {obj-1} </p> [SEG], <p> {obj-2} </p> [SEG], ..., and <p> {obj-n} </p> [SEG]. But I don't see any {neg-1}, {neg-2}, ..., and {neg-m} in this image.

C.2. Reasoning Segmentation

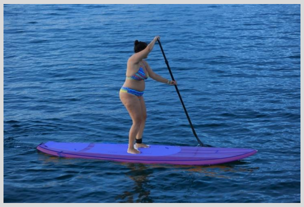
To enhance the reasoning segmentation capability of MGLMM, we use a multi-target reasoning segmentation dataset (MUSE) (Ren et al. 2024) during the pre-training phase. The following is an example from MUSE:

User: After a long day, I want to relax and hydrate at the same time. What arrangement will best suit this need considering the items in the image?
Assistant: You should sit comfortably on the large grey cushioned armchair [SEG] and reach out for the filled water bottle [SEG] that is situated within arm's length for easy access.

To transform MUSE data into USCDF format, we embed <p> and </p> at the appropriate positions by match-

Prompt:

You are an AI visual assistant that can analyze a single image. You receive an image with the masked object. Please provide a short description and detailed description of the object. The short description is limited to a few words. The detailed caption is limited to two or three sentences. In addition, the masked object locations within the image are given. The coordinate is in the form of bounding boxes, represented as (x1, y1, w, h) with floating numbers ranging from 0 to 1. Coordinate:[0.11,0.69.0.74,0.12]

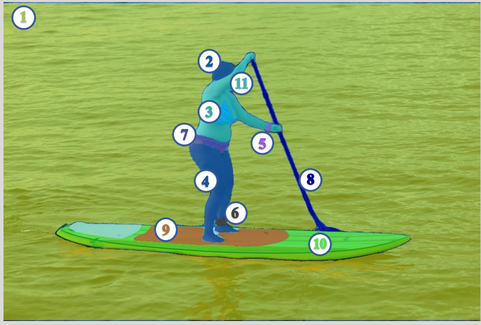


Answer:

Short Description: a surfboard in the water

Detailed Caption: A long, yellow surfboard is seen floating on the water. The surfboard is being used by a woman, who is paddling through the water with a paddle. The surfboard is large and noticeable, extending across a significant portion of t he image.

Figure 2: Illustration of Data Labeling using GPT-4o.



Object 1. **Short Caption**: the water is blue; **Detailed Caption**: The image captures a large body of water, which is the ocean, covering most of the scene. The water is blue and appears calm, with no visible waves or ripples. The woman is paddling her surfboard through this still water.

Object 2. **Short Caption**: the hair is black in color; **Detail Caption**: A woman with a ponytail is visible in the image. She is wearing a bikini and seems to be enjoying her time on the beach.

Object 3. **Short Caption**: a woman bikini top; **Detail Caption**: A woman in a bikini is visible, and she's wearing a colorful top with a floral pattern. The top has a floral design, and it's a combination of blue and green colors.

Object 4. **Short Caption**: the legs of a woman; **Detail Caption**: A woman is standing on a surfboard, paddling through the water. She is wearing a bikini bottom and appears to be in the middle of a surf session. Her stance and the surfboard beneath her indicate that she is actively participating in a water sport.

Object 5. **Short Caption**: a wrist band on the woman wrist; **Detail Caption**: The woman is wearing a necklace.

Object 6. **Short Caption**: ankle strap on surfboard; **Detail Caption**: Ankle strap attached to surfboard.

Object 7. **Short Caption**: blue and yellow bikini bottoms; **Detail Caption**: A woman in a bikini is standing on a surfboard, paddling through the water.

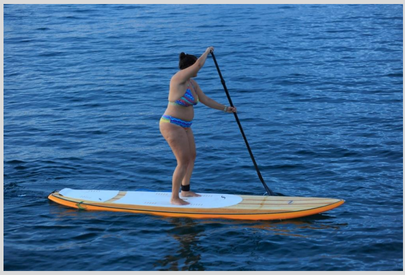
Object 8. **Short Caption**: a paddle holding by a woman; **Detail Caption**: A woman is seen standing on a surfboard, holding a paddle in her hands. The paddle is likely used for propelling the surfboard through the water.

Object 9. **Short Caption**: a white surfboard; **Detail Caption**: The surfboard is white and appears to be in motion, likely due to the flowing water around it.

Object 10. **Short Caption**: a surfboard in the water; **Detail Caption**: A surfboard is prominently visible in the image, with a woman standing on it.

Object 11. **Short Caption**: a woman standing on a surfboard; **Detail Caption**: A woman is seen standing on a surfboard, paddling through the water. She is wearing a bikini and seems to be enjoying her time out in the ocean.

Figure 3: The exhibition of complete semantic labels of each object including the sub-part of the person.

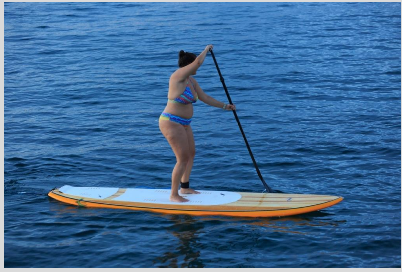


Prompt: You are an AI visual assistant, observing several objects within a single image. Each object is accompanied by a short and detailed description. Referring to the provided information, generate the descriptions that naturally flow about the image, ensuring every object entity is mentioned, and highlight the object that precedes the colon of each message with <object> in the resulting paragraph. **< Object 1 > Short Caption**: the water is blue; **Detailed Caption**: The image captures a large body of water, which is the ocean, covering most of the scene. The water is blue and appears calm, with no visible waves or ripples. The woman is paddling her surfboard through this still water.

< **Object 8** > **Short Caption**: a paddle holding by a woman; **Detail Caption**: A woman is seen standing on a surfboard, holding a paddle in her hands. The paddle is likely used for propelling the surfboard through the water.
< **Object 10> Short Caption**: a surfboard in the water; **Detail Caption**: A surfboard is prominently visible in the image, with a woman standing on it.
< **Object 11> Short Caption**: a woman standing on a surfboard; **Detail Caption**: A woman is seen standing on a surfboard, paddling through the water. She is wearing a bikini and seems to be enjoying her time out in the ocean.

Response:
<object11> A woman standing on a surfboard </object11> is the focal point of this image, balanced and poised as she propels herself forward with the paddle in hand. <object10> The surfboard in the water </object10> is a sturdy platform for her activity, its yellow hue stands out against the blue of the water, suggesting a leisurely venture rather than a competitive sport. <object8> Holding a paddle firmly, the woman </object8> demonstrates control and experience in this aquatic environment. <object1> The water's blue color </object1> stretches across the image, providing a serene backdrop to the scene and emphasizing the tranquil nature of the activity being enjoyed by the woman on the surfboard.

Figure 4: An example of generating the root caption annotation.



Prompt: You are an AI visual assistant, observing several objects within a single image. Each object is accompanied by a short and detailed description. Referring to the provided information, generate the descriptions that naturally flow about the image, ensuring every object entity is mentioned, and highlight the object that precedes the colon of each message with <object> in the resulting paragraph. **<** < **Object 2> Short Caption**: the hair is black in color; **Detail Caption**: A woman with a ponytail is visible in the image. She is wearing a bikini and seems to be enjoying her time on the beach.
< **Object 3> Short Caption**: a woman bikini top; **Detail Caption**: A woman in a bikini is visible, and she's wearing a colorful top with a floral pattern. The top has a floral design, and it's a combination of blue and green colors.
< **Object 4> Short Caption**: the legs of a woman; **Detail Caption**: A woman is standing on a surfboard, paddling through the water. She is wearing a bikini bottom and appears to be in the middle of a surf session. Her stance and the surfboard beneath her indicate that she is actively participating in a water sport.
< **Object 5> Short Caption**: a wrist band on the woman wrist; **Detail Caption**: The woman is wearing a necklace.
< **Object 6> Short Caption**: ankle strap on surfboard; **Detail Caption**: Ankle strap attached to surfboard.
<Object 7> Short Caption: blue and yellow bikini bottoms; **Detail Caption**: A woman in a bikini is standing on a surfboard, paddling through the water.
< **Object 11> Short Caption**: a woman standing on a surfboard; **Detail Caption**: A woman is seen standing on a surfboard, paddling through the water. She is wearing a bikini and seems to be enjoying her time out in the ocean.

Response:
In the image, <object11> a woman standing on a surfboard </object11> is captured as she paddles her way through the water. She is confidently managing the board, indicative of her experience or comfort in the activity. Notably, <object5> a wrist band </object5> can be seen on her wrist. She is secured to her board with <object6> an ankle strap </object6>, which is a safety feature commonly used in such water sports. The <object2> woman's hair </object2> is described as black in color, it's clear she has dark hair gathered up possibly to keep it out of her face while she enjoys the water. Her attire includes <object7> blue and yellow bikini bottoms </object7>, adding a splash of color to the scene. The vibrant bikini is further detailed with <object3>a woman's bikini top </object3>, which is colorful and possibly bears a floral pattern, though the exact design might be difficult to discern. Lastly, <object4> the legs of a woman </object4> are shown, revealing her stance, which suggests a balance and focus necessary for the sport.

Figure 5: An example of generating the subtree caption annotation.

ing the object description preceding the [SEG] tag with the category name of the corresponding mask, like the "armchair" and "water bottle" in the above example. Notably, we exclude specific samples that could not be accurately converted. Ultimately, we derive 35,104 training instances that conform to USCDF standards from the original MUSE dataset.

> **User**: Same as the previous conversation.
> **Assistant**: You should sit comfortably on the large grey cushioned <p> armchair </p> [SEG] and reach out for the filled <p> water bottle </p> [SEG] that is situated within arm's length for easy access.

C.3 Panoptic Segmentation

To adapt the model for the conventional panoptic segmentation task, we extract the panoptic segmentation annotations from MGSCData to train MGLMM. Specifically, we design a special template to organize training data. For each segmentation instance in the annotation, we use the short caption of MGSCData as its semantic label. Then, we summarize all segmentation instances into the following answer using "the image contains ...", which complies with the USCDF standard.

> **User**: <IMAGE> Can you segment all the objects in this image?
> **Assistant**: Sure, the image contains <p> {obj-1} </p> [SEG], <p> {obj-2} </p> [SEG], ..., and <p> {obj-n} </p> [SEG].

D. Implementation Details

D.1. Training Objective

Our model can be trained jointly on multiple tasks with a unified data schema in an end-to-end manner. Its training objective consists of two parts: (1) an autoregressive cross-entropy loss $\mathcal{L}_{txt}$ for modeling text generation; (2) a segmentation mask loss $\mathcal{L}_{mask}$ to supervise high-quality mask prediction. Like LISA (Lai et al. 2024), $\mathcal{L}_{mask}$ is a weighted combination of per-pixel cross-entropy (BCE) loss and DICE loss. Suppose the loss weights are λ_{bce} and λ_{dice}, respectively, the overall objective can be expressed as:

$$\mathcal{L} = \mathcal{L}_{txt} + \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}. \tag{1}$$

During the training process, the vision encoder $\mathcal{F}_v$ is frozen and LLM $\mathcal{F}_{lmm}$ is fine-tuned by LoRA (Hu et al. 2021), aiming to preserve the knowledge of the pre-trained LLaVA. As for the segmentation modules, we completely freeze the pixel encoder $\mathcal{E}_{pixel}$ and fully fine-tune the pixel decoder $\mathcal{D}_{pixel}$. The trainable parameters include LoRA, pixel Decoder, and project heads.

D.2. Hyperparmeter Setting

The AdamW optimizer is used with the initial learning rate and weight decay set to 0.0003 and 0, respectively. We adopt WarmupDecayLR as the learning rate scheduler, where the warmup iterations are set to 100. The weights of BCE loss λ_{bce} and DICE loss λ_{dice} are set to 2.0 and 0.5, respectively.

E. More Experiments

In this section, we provide detailed experiment results for discussion.

E.1. Multi-Granularity SegCap.

To better compare the performance differences between our method and the GLaMM model, we divide MGSCData into two tasks: Panoptic SegCap and Fine-grained SegCap, and evaluate their performance separately. Following the same evaluation protocol of GCG, we fine-tune the GLaMM and our MGLMM on the training set of MGSCData and evaluate them on the same metric. As shown in Table 1, we outperform GLaMM on two subtasks, which shows the great performance and versatility of our MGLMM.

E.2. Grounded Conversation Generation

We evaluate our MGLMM on the validation and test set of GCG and report detailed results. As shown in Table 3, we outperform in every metric except for mIoU. Specifically, we observe a significant improvement in Recall, indicating that our model can perceive more targets.

E.3. More Ablation Results

As shown in Table 2, we present more detailed ablation results to demonstrate the effectiveness of our proposed USCDF module across multiple tasks. This indicates that utilizing USCDF can significantly enhance multi-task learning.

F. Additional Qualitative Results

To better understand the capacity of MGLMM, we conduct qualitative comparisons and analyses on various tasks, including multi-granularity SegCap, grounded conversation generation, and referring segmentation. Here, we mainly compare our MGLMM with GLaMM.

F.1. Multi-Granularity SegCap

Figure 6 shows the qualitative results of the MGLMM fine-tuned on the MGSCData dataset. Our model is capable of performing panoptic and fine-grained SegCap. Our model exhibits strong capabilities in panoptic segmentation and fine-grained target perception.

F.2. Grounded Conversation Generation

Figure 8 shows that the strong performance of our model on other tasks can also be transferred to the GCG task, where our results surpass those of the GLaMM model.

F.3. Referring and Multiple/Empty Segmentation

We compared GLaMM and MGLMM on referring and multi-target segmentation tasks. The visualization result as shown in Figure 9, our model exhibits robust capabilities in fine-grained understanding and segmentation.

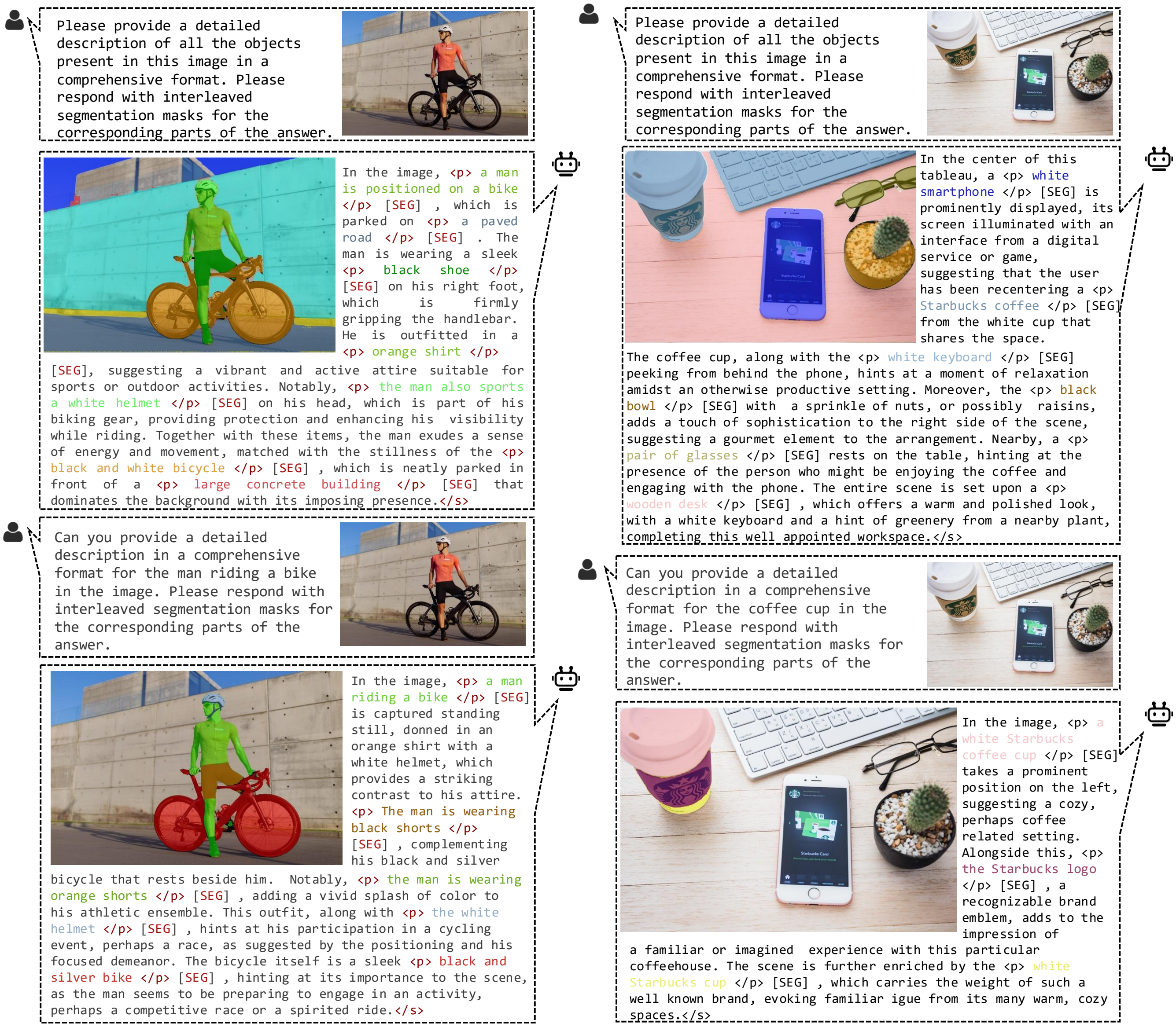
Figure 10 shows the qualitative comparisons on multiple/empty tasks for GLaMM and MGLMM. Our model provides a more accurate segmentation mask in the first example and identifies that the object does not exist with text

Model	Panoptic SegCap					Fine-grained SegCap					All				
	METEOR	CIDEr	AP50	mIoU	Recall	METEOR	CIDEr	AP50	mIoU	Recall	METEOR	CIDEr	AP50	mIoU	Recall
GLaMM (Rasheed et al. 2024)	16.5	4.7	5.4	52.5	16.3	16.6	8.9	6.5	45.5	21.6	16.5	8.3	5.4	47.6	18.7
MGLMM(Ours)	**17.6**	**16.7**	**8.0**	**58.5**	**21.5**	**17.9**	**10.0**	**7.4**	**48.7**	**25.2**	**17.8**	**11.6**	**7.4**	**51.6**	**23.2**

Table 1: Performance on multi-granularity segmentation and captioning. We compare our model with GLaMM using METEOR, CIDEr, AP50, mIoU, and mask recall metrics.

Model	+ USCDF	+ GrandD Dataset	refCOCO			refCOCO+			refCOCOg		gRefCOCO		
			val	testA	testB	val	testA	testB	val	test	val	testA	testB
MGLMM-7B			76.2	81.0	71.0	67.2	74.1	58.9	72.3	71.4	49.7	57.7	53.7
MGLMM-7B	✓		78.2	81.9	74.0	69.9	76.2	62.5	73.6	73.5	53.0	63.9	57.9
MGLMM-7B		✓	79.8	82.6	75.1	71.4	76.9	64.0	75.1	74.8	50.6	58.1	54.8
MGLMM-7B	✓	✓	80.2	83.1	76.0	73.2	78.7	66.8	76.7	77.5	52.8	61.2	56.0

Table 2: More ablation results on referring segmentation and generalized referring segmentation. We utilize cIoU as the metric.



Please provide a detailed description of all the objects present in this image in a comprehensive format. Please respond with interleaved segmentation masks for the corresponding parts of the answer.

In the image, <p> a man is positioned on a bike </p> [SEG] , which is parked on <p> a paved road </p> [SEG] . The man is wearing a sleek <p> black shoe </p> [SEG] on his right foot, which is firmly gripping the handlebar. He is outfitted in a <p> orange shirt </p> [SEG], suggesting a vibrant and active attire suitable for sports or outdoor activities. Notably, <p> the man also sports a white helmet </p> [SEG] on his head, which is part of his biking gear, providing protection and enhancing his visibility while riding. Together with these items, the man exudes a sense of energy and movement, matched with the stillness of the <p> black and white bicycle </p> [SEG] , which is neatly parked in front of a <p> large concrete building </p> [SEG] that dominates the background with its imposing presence.</s>

Can you provide a detailed description in a comprehensive format for the man riding a bike in the image. Please respond with interleaved segmentation masks for the corresponding parts of the answer.

In the image, <p> a man riding a bike </p> [SEG] is captured standing still, donned in an orange shirt with a white helmet, which provides a striking contrast to his attire. <p> The man is wearing black shorts </p> [SEG] , complementing his black and silver bicycle that rests beside him. Notably, <p> the man is wearing orange shorts </p> [SEG] , adding a vivid splash of color to his athletic ensemble. This outfit, along with <p> the white helmet </p> [SEG] , hints at his participation in a cycling event, perhaps a race, as suggested by the positioning and his focused demeanor. The bicycle itself is a sleek <p> black and silver bike </p> [SEG] , hinting at its importance to the scene, as the man seems to be preparing to engage in an activity, perhaps a competitive race or a spirited ride.</s>

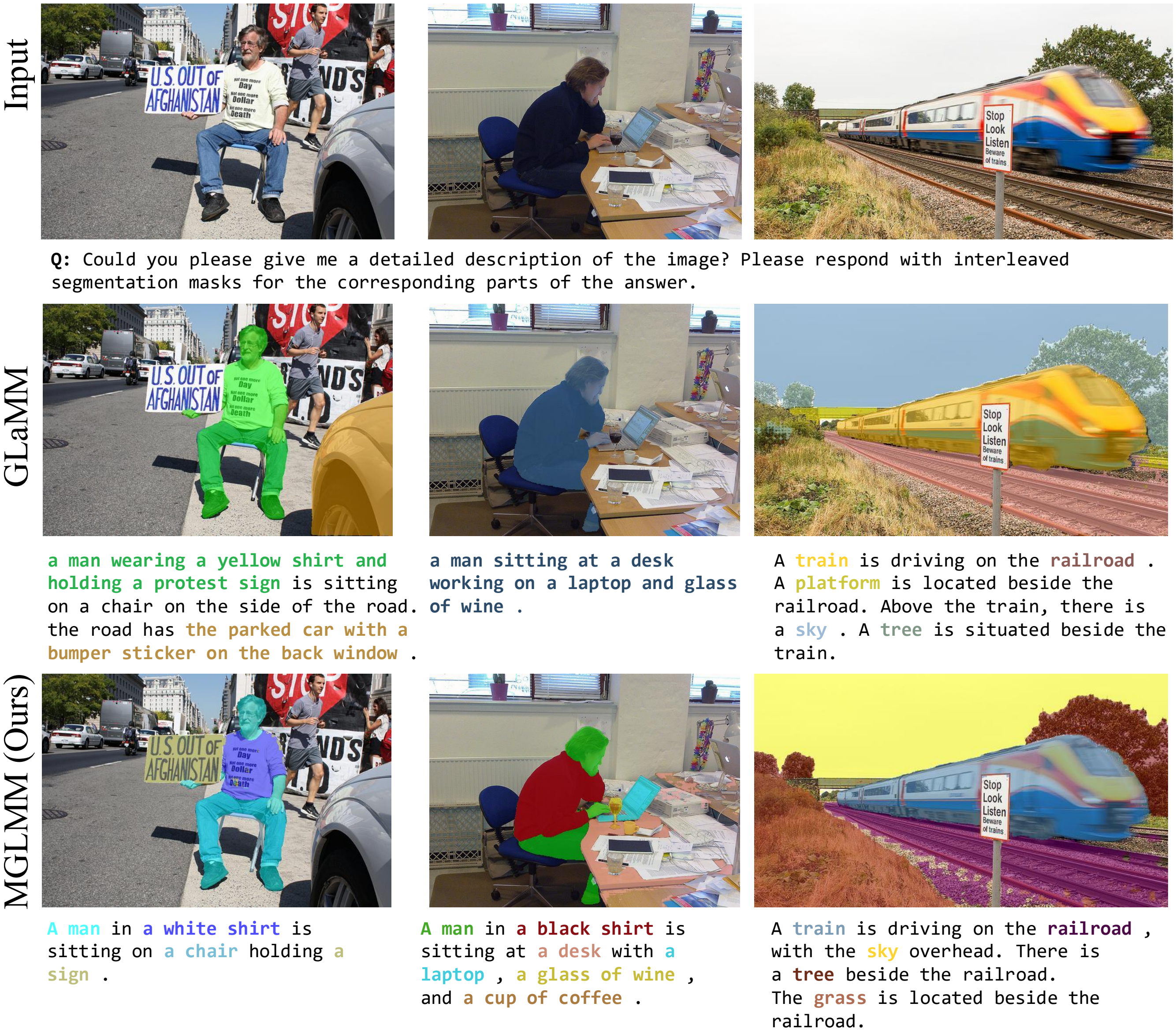
Please provide a detailed description of all the objects present in this image in a comprehensive format. Please respond with interleaved segmentation masks for the corresponding parts of the answer.

In the center of this tableau, a <p> white smartphone </p> [SEG] is prominently displayed, its screen illuminated with an interface from a digital service or game, suggesting that the user has been recentering a <p> Starbucks coffee </p> [SEG] from the white cup that shares the space. The coffee cup, along with the <p> white keyboard </p> [SEG] peeking from behind the phone, hints at a moment of relaxation amidst an otherwise productive setting. Moreover, the <p> black bowl </p> [SEG] with a sprinkle of nuts, or possibly raisins, adds a touch of sophistication to the right side of the scene, suggesting a gourmet element to the arrangement. Nearby, a <p> pair of glasses </p> [SEG] rests on the table, hinting at the presence of the person who might be enjoying the coffee and engaging with the phone. The entire scene is set upon a <p> wooden desk </p> [SEG] , which offers a warm and polished look, with a white keyboard and a hint of greenery from a nearby plant, completing this well appointed workspace.</s>

Can you provide a detailed description in a comprehensive format for the coffee cup in the image. Please respond with interleaved segmentation masks for the corresponding parts of the answer.

In the image, <p> a white Starbucks coffee cup </p> [SEG] takes a prominent position on the left, suggesting a cozy, perhaps coffee related setting. Alongside this, <p> the Starbucks logo </p> [SEG] , a recognizable brand emblem, adds to the impression of a familiar or imagined experience with this particular coffeehouse. The scene is further enriched by the <p> white Starbucks cup </p> [SEG] , which carries the weight of such a well known brand, evoking familiar igue from its many warm, cozy spaces.</s>

Figure 6: Examples of the conversation with multi-granularity masks of our model on the MGSC task.



Q: Could you please give me a detailed description of the image? Please respond with interleaved segmentation masks for the corresponding parts of the answer.

a man wearing a yellow shirt and holding a protest sign is sitting on a chair on the side of the road. the road has the parked car with a bumper sticker on the back window .

a man sitting at a desk working on a laptop and glass of wine .

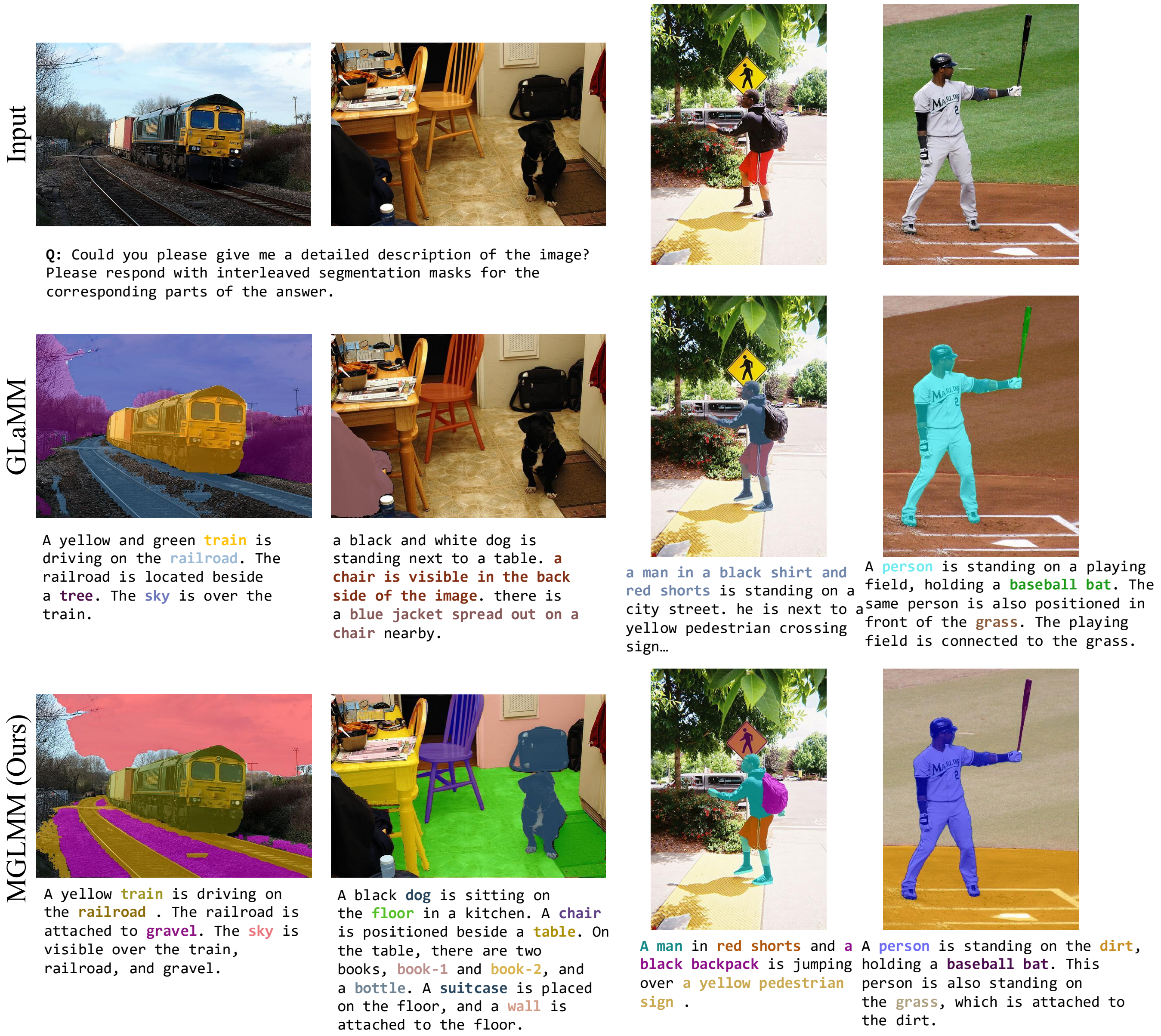
A train is driving on the railroad . A platform is located beside the railroad. Above the train, there is a sky . A tree is situated beside the train.

A man in a white shirt is sitting on a chair holding a sign .

A man in a black shirt is sitting at a desk with a laptop , a glass of wine , and a cup of coffee .

A train is driving on the railroad , with the sky overhead. There is a tree beside the railroad. The grass is located beside the railroad.

Figure 7: Qualitative results on grounded conversation generation.



Q: Could you please give me a detailed description of the image? Please respond with interleaved segmentation masks for the corresponding parts of the answer.

A yellow and green train is driving on the railroad. The railroad is located beside a tree. The sky is over the train.

a black and white dog is standing next to a table. a chair is visible in the back side of the image. there is a blue jacket spread out on a chair nearby.

a man in a black shirt and red shorts is standing on a city street. he is next to a yellow pedestrian crossing sign…

A person is standing on a playing field, holding a baseball bat. The same person is also positioned in front of the grass. The playing field is connected to the grass.

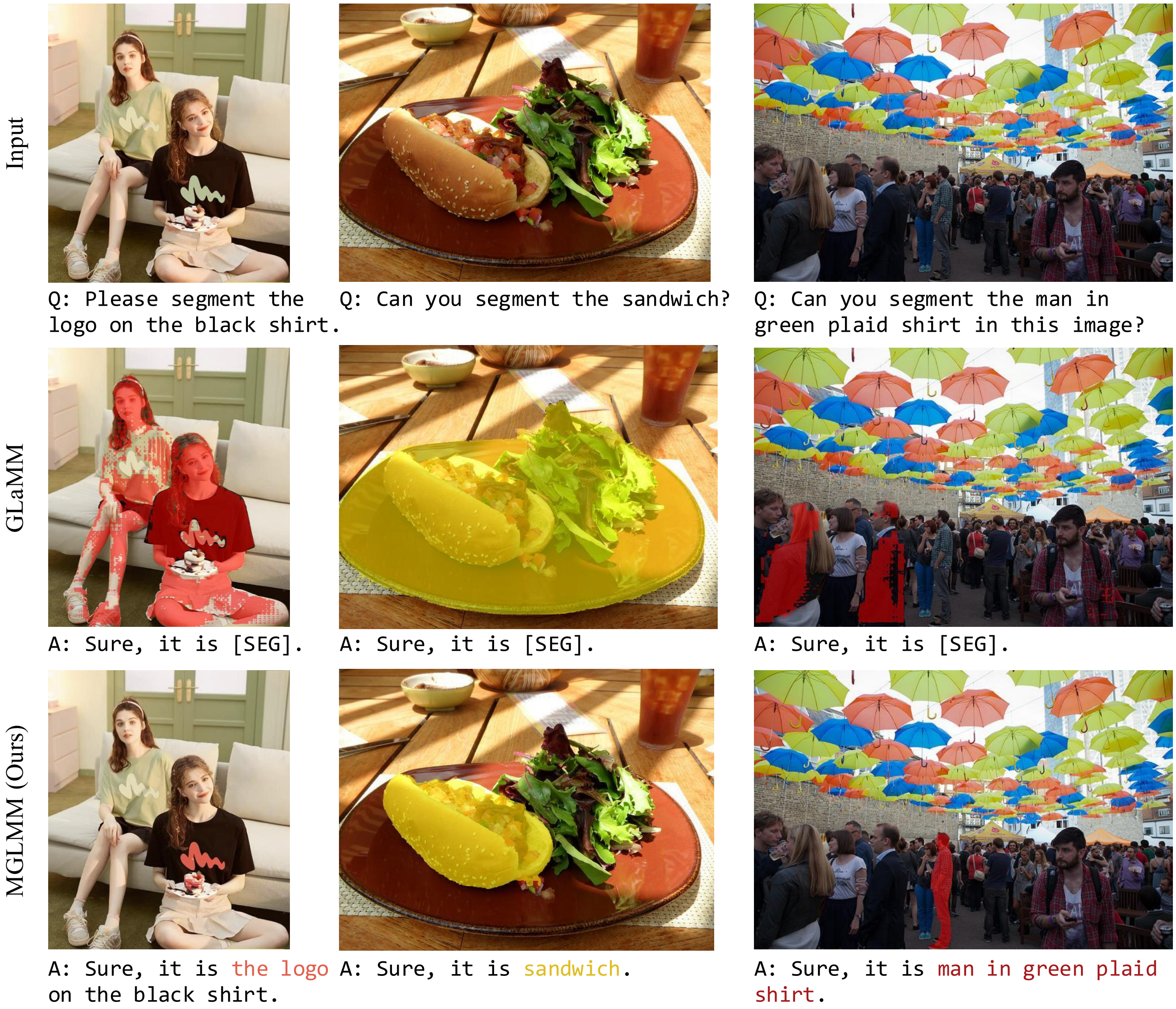
A yellow train is driving on the railroad . The railroad is attached to gravel. The sky is visible over the train, railroad, and gravel.

A black dog is sitting on the floor in a kitchen. A chair is positioned beside a table. On the table, there are two books, book-1 and book-2, and a bottle. A suitcase is placed on the floor, and a wall is attached to the floor.

A man in red shorts and a black backpack is jumping over a yellow pedestrian sign .

A person is standing on the dirt, holding a baseball bat. This person is also standing on the grass, which is attached to the dirt.

Figure 8: Qualitative results on grounded conversation generation.



Q: Please segment the logo on the black shirt.

Q: Can you segment the sandwich?

Q: Can you segment the man in green plaid shirt in this image?

A: Sure, it is [SEG].

A: Sure, it is [SEG].

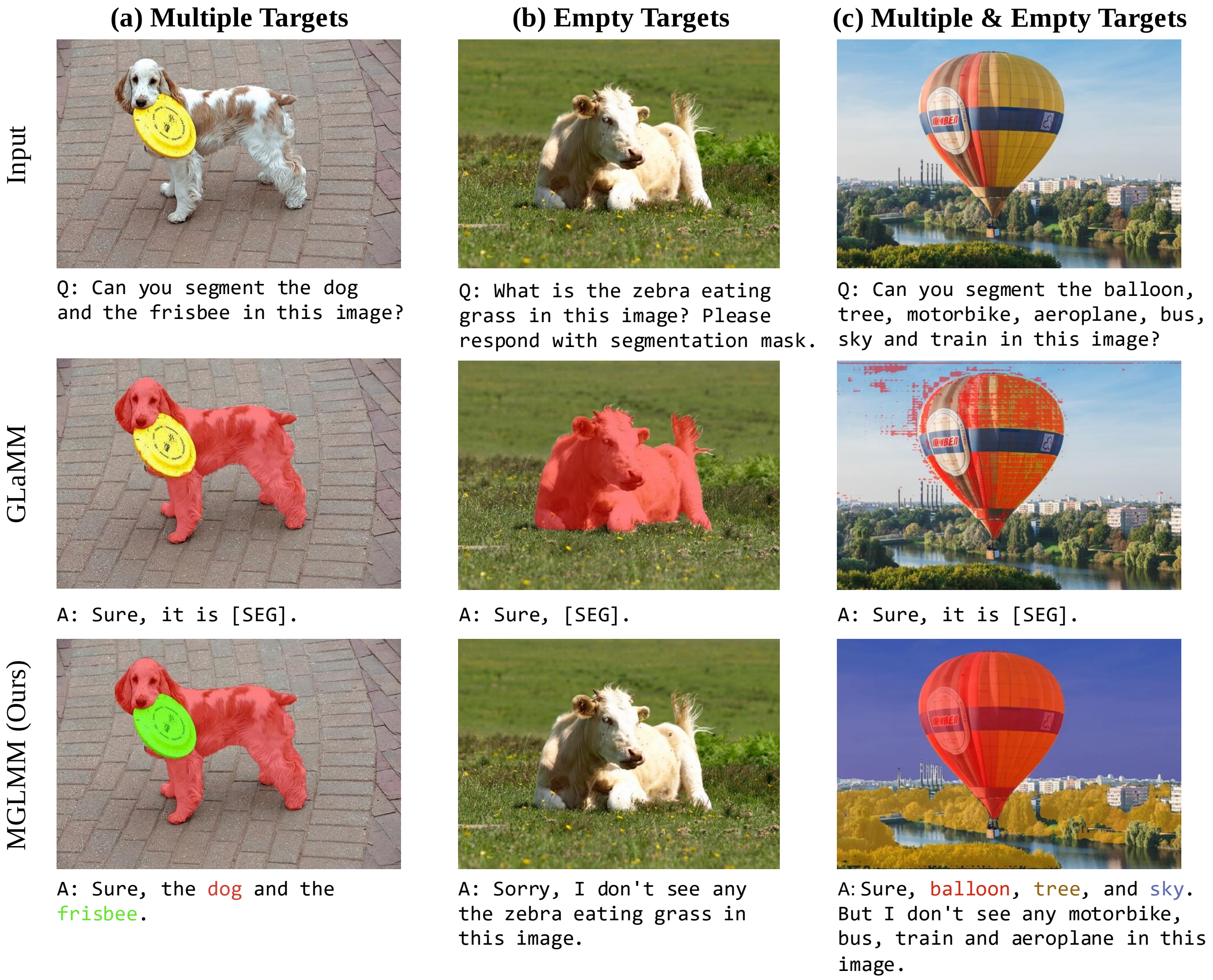
A: Sure, it is [SEG].

A: Sure, it is the logo on the black shirt.

A: Sure, it is sandwich.

A: Sure, it is man in green plaid shirt.

Figure 9: Qualitative results on the referring segmentation.



Q: Can you segment the dog and the frisbee in this image?

Q: What is the zebra eating grass in this image? Please respond with segmentation mask.

Q: Can you segment the balloon, tree, motorbike, aeroplane, bus, sky and train in this image?

A: Sure, it is [SEG].

A: Sure, [SEG].

A: Sure, it is [SEG].

A: Sure, the dog and the frisbee.

A: Sorry, I don't see any the zebra eating grass in this image.

A: Sure, balloon, tree, and sky. But I don't see any motorbike, bus, train and aeroplane in this image.

Figure 10: Qualitative comparisons on the multiple/empty target segmentation.

Model	Validation Set					Test Set				
	M	C	AP50	mIoU	MR	M	C	AP50	mIoU	MR
Kosmos-2	16.1	27.6	17.1	55.6	28.3	15.8	27.2	17.2	56.8	29.0
LISA	13.0	33.9	25.2	62.0	36.3	12.9	32.2	24.8	61.7	35.5
OMG-LLaVA	14.9	41.2	29.9	65.5	-	14.5	38.5	28.6	64.7	-
GLaMM	16.2	47.2	30.8	**66.3**	41.8	15.8	43.5	29.2	**65.6**	40.8
MGLMM	**16.4**	**50.1**	**31.7**	**66.3**	**45.2**	**16.3**	**47.5**	**30.6**	**65.6**	**45.1**

Table 3: More quantitative results on grounded conversation generation datasets. We compare our MGLMM with other models using the METEOR (M), CIDEr (C), AP50, mIoU, and Mask Recall (MR) metrics.

as responses in the second example. In the third example, which includes both multiple targets and an empty target, our model successfully segments all targets and indicates that the empty target does not exist in the image by natural language.

References

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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.

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.

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.

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