Hugging Visual Prompt and Segmentation Tokens: Consistency Learning for Fine-Grained Visual Understanding in MLLMs

Supplementary Material

We provide an overview of the Appendix below. Each section includes a brief description of its contents.

S1. Additional Implementation Details

S1.1. Training Details

We instantiate our base models using the 3B and 7B versions of Qwen2.5-VL and employ the pretrained SAM-vit-h as the mask decoder. The hidden unit size inside the prompt encoder is set to 256. To reduce GPU memory consumption and accelerate training while improving detection performance, both the image input and the decoder segmentation operate at a long-side resolution of 1024. For each segmentation sample, up to five objects are randomly selected to compute the mask loss. During training, we perform full-parameter fine-tuning on both the large language model (LLM), with an input sequence length of 4096 tokens. Training is conducted on eight RTX H20 GPUs, with a global batch size of 32 for Stage 1. In Stage 2, to efficiently perform consistency learning, we set the per-GPU batch size to 2 (each sample is used to construct both the DL-RES and DLC tasks for consistency training), resulting in a global batch size of 16. The learning rate in Stage 1 is configured as 2×10^{-5} for all modules, whereas in Stage 2 only the targeted modules are updated, using a learning rate of 1×10^{-5}.

S1.2. Dataset and Evaluation Metrics

Datasets. We train our FCLM on a comprehensive dataset encompassing three categories: captioning, grounding, and GCG. Table 7 provides an overview of the data composition used in our two-stage training pipeline. Stage 1 integrates three types of supervision, including segmentation-level grounding (RES/ReasonSeg/MUSE), region-level captioning (ROC/DLC), and grounded conversation generation (GCG), covering a total of 688K samples. These datasets provide balanced signals for grounding, captioning, and multimodal reasoning. Stage 2 introduces the Describe Anything dataset (DLC/DL-RES), which contains 704K samples and accounts for 50.6% of all training data. This stage emphasizes dense region-level annotation, incorporating COCO-Stuff, Mapillary, OpenImages, and a large set of SAM-generated masks. Together, the two stages supply complementary supervision for unified training across grounding, captioning, and dense localization tasks.

Table 7. Composition of the training datasets across two stages: Grounding, Captioning, Grounded Conversation Generation (GCG), and Detailed Localized datasets.

Stage	Dataset	#Samples	Ratio
Stage 1	***Grounding***	**194,097**	**13.9%**
	RefCOCO/+/g	55,885	4.0%
	MUSE	80,000	5.7%
	ReasonSeg	40,000	2.8%
	Pixel2Cap	18,212	1.3%
	Captioning	**297,225**	**21.4%**
	VG	77,000	5.5%
	PACO_LVIS	113,212	8.1%
	Osprey_ShortForm	27,454	2.0%
	Osprey_LVISPosNeg	20,008	1.4%
	Osprey_Conversations	30,217	2.1%
	Osprey_DetailedDescription	29,334	2.1%
	Grounded Conversation Generation	**196,938**	**14.1%**
	GranDf	1,000	0.1%
	OpenPSG	28,584	2.1%
	Flickr30K	148,157	10.6%
	RefCOCOg	19,327	1.4%
Stage 2	***Detailed Localized Dataset***	**703,823**	**50.6%**
	COCO-Stuff	28,365	2.0%
	Mapillary	17,762	1.3%
	OpenImages	64,874	4.7%
	SAM (10%)	592,822	42.6%

Evaluation Metrics. For **referring object classification**, we follow a similar approach as DAM [29] and Osprey [67] by employing *semantic similarity* (SS) and *semantic IoU* (sIoU) to evaluate the classification capability of the model. SS measures the similarity between predicted and ground-truth labels in a semantic embedding space, capturing how semantically close the predictions are to the references. sIoU, on the other hand, quantifies the overlap of words between predicted and ground-truth labels, providing a complementary measure of lexical agreement. For **detailed localized captioning**, following DAM [29], the large language model (Llama 3 [17]) acts as the evaluator using predefined questions. *Positive questions* assess attributes that should appear in the target object description, awarding a point for correct inclusion, zero for omission,

and penalizing factual errors. *Negative questions* assess details that should not appear, giving a point for correct omission and penalizing inclusion. Points are granted only when the target object is correctly identified, preventing incorrect descriptions from scoring highly. For **grounded conversation generation**, we follow the same setup as the original GLaMM [43]. For **segmentation-based tasks**, we adopt the standard evaluation metrics used by LISA [27] and most existing methods, including cIoU and gIoU.

S1.3. Instruction Details

Table 12 summarizes the prompts used for the detailed localized tasks during training. For DLC, each prompt comprises an image, a question, and an answer, where the image is represented using the Qwen2.5-VL placeholder <|image_pad|>, and the model is required to generate a detailed localized caption. For DL-RES, the prompt consists of an image, a query, and an answer, with the query provided by the corresponding detailed localized caption of the sample. Both tasks are trained jointly on the same set of samples. To enhance model robustness, we construct multiple prompt formulations and randomly sample among them during training.

S2. Additional Ablation Studies

S2.1. Additional Ablation Studies of the Hybrid Region Extractor

Visual Prompt Embeddings. We study the roles of the mask token <mask> and position token <pos> in the visual prompt embeddings. The mask token captures the semantic details of the region, while the position token conveys shape and spatial information. Ablation results in Table 8 show that each token individually contributes to performance, and using both tokens together achieves the best results, as they provide complementary information for accurate reconstruction and fine-grained visual understanding. All subsequent SS and sIoU results are evaluated on the LVIS dataset.

Table 8. Ablation study on visual prompt embeddings for ROC LVIS.

Visual Prompt Embeddings	SS	sIoU
<mask>	85.1	72.4
<pos>	83.7	71.9
<mask> + <pos>	**87.2**	**76.4**

FPS Analysis. We analyze the impact of each component in the hybrid region extractor on both performance and inference speed (FPS) with inference GPU A30. As shown in Table 9, incorporating local aggregation and pixel enhancement slightly reduces FPS due to additional computations. However, these components consistently improve the model's performance, yielding higher SS and sIoU scores on LVIS.

Table 9. Effect of each component in the hybrid region extractor on LVIS performance and inference speed (FPS).

Local Agg.	Pixel Enh.	Semantic Guid.	LVIS		FPS
			SS	sIoU	
		✓	87.1	75.7	**1.40**
	✓	✓	88.7	77.1	1.37
✓		✓	88.1	78.0	1.38
✓	✓	✓	**89.3**	**78.9**	1.36

Deformable Attention vs. Cross Attention. We distinguish semantic guidance from pixel enhancement, as semantic guidance relies on contextual semantic correlations rather than focusing solely on local details. As illustrated in Figure 3, when analyzing a batter, in addition to local pixel-level semantics, it is crucial to adaptively capture global contextual information such as the positions of fielders and the surrounding grass. To achieve this, we employ deformable attention, which adaptively selects sampling locations to incorporate relevant contextual cues. As shown in Table 10, using deformable attention with a sampling point number of 4 yields the best performance.

Table 10. Comparison of cross attention and deformable attention for semantic guidance.

Attention Type	Sampling Points	Pos	Neg
Cross Attention	–	50.1	81.9
Deformable Attention	2	51.2	82.0
	4	**52.0**	**83.6**
	8	51.8	83.5

S2.2. Ablation on Stage 2 Configurations

Training with Additional Datasets. Our Stage 2 training incorporates the Describe Anything dataset to achieve a more fine-grained visual understanding. While our Stage 1 training already establishes general captioning and grounding capabilities, an important question is whether consistency learning remains effective for general referring expression segmentation tasks. As shown in Table 11, by finetuning on grounding data during Stage 2, we observe that consistency learning still yields performance improvements, even though the annotations in the grounding datasets are not as detailed.

S3. Additional Qualitative Results

S3.1. Visualization Comparison with UniPixel

We compare the qualitative results of our proposed FCLM with UniPixel [36] on the novel DL-RES task. As illustrated

Table 11. Ablation study on Stage 2 training strategies using additional datasets on RefCOCO.

Training Strategy	RefCOCO		
	val	testA	testB
FCLM w/o $\mathcal{L}_{latent}$	81.7	83.6	79.9
FCLM w $\mathcal{L}_{latent}$	82.4	83.9	80.5

in Figure 6, we provide several representative examples.

The first example describes a taxi with fine-grained details, which UniPixel incorrectly localizes as a bus. In the second case, for a detailed description of a watch, UniPixel erroneously includes a similarly shaped ring in addition to the correct watch. The third example shows UniPixel misidentifying a pedestrian light as a billboard. For the fourth description, which includes the detailed phrase "the left temple has a visible hinge mechanism near the frame", UniPixel, despite localizing the eyeglasses, still makes an incorrect segmentation. In the fifth case, although UniPixel identifies the shoe, its segmentation lacks precision. Across all these challenging instances, FCLM demonstrates significantly more accurate referring expression segmentation, highlighting its robust capability in comprehending detailed linguistic descriptions and precisely localizing specific targets.

S3.2. Multi-Task Visualization

In this section, we present additional qualitative results to further demonstrate the versatility and multi-task capabilities of our proposed FCLM model across various visual understanding tasks. We provide visual examples for captioning, grounding, and grounded conversation generation, highlighting FCLM's performance in diverse scenarios.

Captioning Tasks. Figure 7 illustrates FCLM's proficiency in various captioning tasks. Specifically, we visualize examples of *detailed localized captioning*, demonstrating FCLM's ability to generate fine-grained descriptions for specific regions, as well as *referring object classification*, where FCLM accurately identifies and categorizes referred objects. These examples collectively highlight FCLM's strong visual understanding for descriptive tasks.

Grounding Tasks. For grounding, Figure 8 presents FCLM's performance across several challenging tasks: *referring expression segmentation*, *reasoning segmentation*, and *multi-target reasoning segmentation*. These visualizations confirm FCLM's robust capability in precisely segmenting targets based on complex and multi-object referring expressions, often requiring intricate visual reasoning.

Grounded Conversation Generation. Figure 9 showcases FCLM's ability in *grounded conversation generation*. Beyond engaging in coherent dialogue, these examples specifically demonstrate FCLM's capacity to derive image-level descriptions and simultaneously provide corresponding segmentation masks, thereby unifying and highlighting its strong captioning and grounding abilities within a conversational context.

S3.3. Limitations

FCLM still presents significant room for improvement. For instance, in current visual understanding tasks, which primarily categorize into captioning and grounding, models typically either take visual prompt embeddings as input or generate segmentation tokens as output, but not simultaneously. We observe that refer to any segmentation mask group (RAS) [5] proposes a novel task that requires simultaneous input of visual prompt embeddings and output of segmentation tokens. Future work can explore how to leverage consistency learning to address and enhance such challenging tasks.

Table 12. Prompts used for detailed localized tasks.

Prompts for Detailed Localized Captioning

Question Prompt:

(1) "Can you provide me with a detailed description of the region in the picture marked by <region>?"
(2) "I'm curious about the region represented by <region> in the picture. Could you describe it in detail?"
(3) "What can you tell me about the region indicated by <region> in the image?"
(4) "I'd like to know more about the area in the photo labeled <region>. Can you give me a detailed description?"
(5) "Could you describe the region shown as <region> in the picture in great detail?"
(6) "What details can you give me about the region outlined by <region> in the photo?"
(7) "Please provide me with a comprehensive description of the region marked with <region> in the image."
(8) "Can you give me a detailed account of the region labeled as <region> in the picture?"
(9) "I'm interested in learning more about the region represented by <region> in the photo. Can you describe it in detail?"
(10) "What is the region outlined by <region> in the picture like? Could you give me a detailed description?"
(11) "Can you provide me with a detailed description of the region in the picture marked by <region>, please?"
(12) "I'm curious about the region represented by <region> in the picture. Could you describe it in detail, please?"
(13) "What can you tell me about the region indicated by <region> in the image, exactly?"
(14) "I'd like to know more about the area in the photo labeled <region>, please. Can you give me a detailed description?"
(15) "Could you describe the region shown as <region> in the picture in great detail, please?"
(16) "What details can you give me about the region outlined by <region> in the photo, please?"
(17) "Please provide me with a comprehensive description of the region marked with <region> in the image."
(18) "Can you give me a detailed account of the region labeled as <region> in the picture?"
(19) "I'm interested in learning more about the region represented by <region> in the photo. Can you describe it in detail, please?"
(20) "What is the region outlined by <region> in the picture like, please? Could you give me a detailed description?"
(21) "Please describe the region <region> in the image in detail."
(22) "Can you offer a thorough analysis of the region <region> in the image?"
(23) "Could you elaborate on the region highlighted by <region> in the picture provided?"
(24) "Please share more information about the zone emphasized with <region> in the photo."
(25) "What insights can you give about the area denoted by <region> in the image presented?"
(26) "Can you share a comprehensive rundown of the region denoted by <region> in the presented image?"
(27) "I'd like to know more about the region highlighted by <region> in the picture provided."
(28) "Work through the important details of the area <region> in the image."
(29) "Illustrate the area represented by <region> through a descriptive explanation."
(30) "Examine the region <region> closely and share its details."

Answer Prompt:

(1) `Detailed Localized Caption.`

Prompts for Detailed Localized Referring Expression Segmentation

Question Prompt:

(1) "<images>\n <Detailed Localized Caption>. Please respond with segmentation mask."
(2) "<images>\n <Detailed Localized Caption>. Please output segmentation mask."

Answer Prompt:

(1) "It is [SEG]."
(2) "Sure, [SEG]."
(3) "Sure, it is [SEG]."
(4) "Sure, the segmentation result is [SEG]."
(5) "[SEG]."

Figure 6. Visualization comparison with UniPixel on the DL-RES task.

Detailed Localized Captioning

Can you provide me with a detailed description of the region in the picture marked by <region>?

A plush toy with large, round, black eyes outlined in white, a small black nose, and a wide, open-mouthed smile. The toy has a light blue body with a darker blue head. It has two arms and two legs, all in a light blue color.

A small, round, white ceramic bowl with a smooth, glossy finish. The bowl has a slightly flared rim and contains a mixture of white and yellow ingredients, including a red garnish.

A large, light-colored animal with a long, slender body and a slightly curved neck. The animal has a smooth, elongated head with a small, pointed beak. Its legs are short and sturdy, with visible joints and a slightly bent posture. The animal's fur appears to be short and uniform in color.

A black bowtie with a smooth, satin-like texture, featuring a classic butterfly shape. The bowtie has a central knot with two symmetrical loops and two trailing ends that are slightly tapered.

Referring Object Classification

Please identify the labels of each marked <region> in the image.

license plate

back cover of the telephone

hand of the clock

head of the dog

steering wheel

handle of the knife

Figure 7. Visualization on captioning tasks.

Referring Expression Segmentation

Please segment woman.

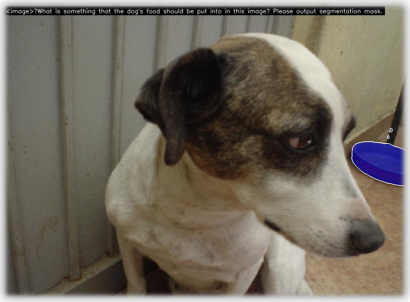
Please segment left phone case.

Please segment the woman washing dishes.

Reasoning Segmentation

Soldiers are often equipped with various tools and weapons to carry out their duties. What item in the picture can be classified as a weapon? Please output segmentation mask.

What is something that the dog' s food should be put into in this image? Please output segmentation mask.



What is something that we use to control computer games in this image? Please output segmentation mask.

Multi-target Reasoning Segmentation

How can the kid pretend that the teddy bear is a live companion while wearing something adorn? Please respond with segmentation mask.

The child can hold onto the teddy bear [SEG] and put on the headband [SEG] to make it seem like they are wearing a crown or decoration, enhancing the imaginative play experience.

If you were to take a nap, which objects in the image would provide you with warmth and comfort? Please respond with segmentation mask.

Sure, a cozy blanket located towards the bottom of the image, another soft blanket positioned above the first one are [SEG] and [SEG], separately.

Figure 8. Visualization on grounding tasks.

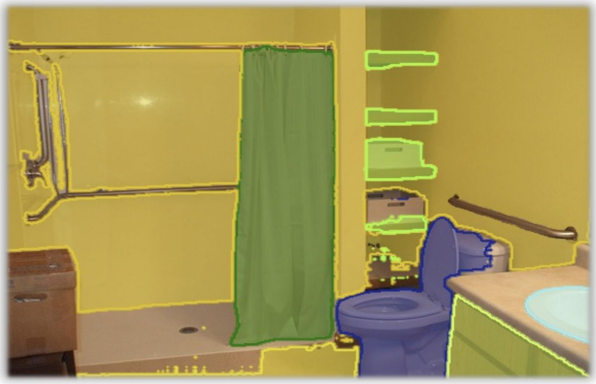
Grounded Conversation Generation

Could you please give me description of the <image>? Please respond with interleaved segmentation masks for the corresponding parts of the answer.

The elephant on the right that is facing left , is part of a group standing in the water. The elephant on the far left is facing away from the camera. there is also an elephant standing in the middle of two other elephants.

The image features a truck parked on the road , with a fence situated beside it. The truck is also positioned in front of a building. Overhead, the sky stretches above the truck and the building.

A blackberry phone next to a motorolla phone , both sitting on a table. the motorolla flip phone is also visible.



The image showcases a bathroom with a toilet placed on the floor and a sink set into a counter. A curtain is seen hanging from the wall. The bathroom also features a shelf attached to the wall.

Figure 9. Visualization on grounded conversation generation task.