SAMTok: Representing Any Mask with Two Words

Supplementary Material

A. Overview

In this file, we present additional experimental results due to space constraints in the main paper. Here are the details:

- In § B, we provide more implementation details for both the tokenizer and the multi-modal large language models.
- In § C, we provide more experimental results.
- In § D, we present ablation studies on tokenizer designs and their effectiveness.
- In § E, we provide more visualizations.

B. Implementation Details

We initialize the encoder and decoder of SAMTok with the pretrained weights of SAM 2.1 [10] (Large), while the quantizer's codebook is randomly initialized. During training, the parameters of the SAM image encoder and SAM prompt encoder are frozen, whereas the SAM decoder is trainable. Unless otherwise specified, the SAMTok codebook size is set to 256, and the number of quantization steps is 2, with the two codebooks non-shared. We train SAMTok using Xtuner [3] and the AdamW optimizer [7], with a global batch size of 1024, an initial learning rate of 4e-5, and a cosine decay schedule [6]. We employ Qwen-VL series [1] models as the base model for the main experiments. The only modification made to the base models is the addition of mask tokens in the vocabulary. By default, we introduce 512 mask tokens, formatted as <|mt_0000|> ∼ <|mt_0511|>. Among them, <|mt_0000|> ∼ <|mt_0255|> correspond to the first-level codebook, and <|mt_0256|> ∼ <|mt_0511|> correspond to the second-level codebook. In addition, we introduce two special tokens, <|mt_start|> and <|mt_end|>, to denote the start and end positions of mask token sequences. The embeddings of these newly added tokens are randomly initialized using the mean and variance statistics of the original token embeddings. During supervised fine-tuning (SFT) and reinforcement learning (RL), we freeze the MLLM image encoder and fine-tune the projection layer and LLM. SFT is conducted using the official Qwen-VL [1, 15] implementation, the AdamW optimizer [7], a global batch size of 256, and a learning rate of 2e-5 with a cosine decay schedule [6]. All training and evaluation experiments are performed on NVIDIA A100 GPUs (80 GB). For RL, we employ the Easy-R1 [20] framework with the GRPO [11] algorithm, using a learning rate of 1e-6.

Table A1. Results on panoptic scene graph generation.

Method	R@20	mR@20
IMP [17]	16.5	6.5
GPSNet [5]	17.8	7.0
MOTIF [19]	<u>20.0</u>	9.1
VCTree [12]	**20.6**	9.7
PSGFormer [18]	18.0	<u>14.8</u>
Qwen25VL-SAMTok-7B	19.8	**15.4**

C. Additional model experiments

We present more experiments for more vision language tasks.

Panoptic scene graph generation. We further evaluate our model's interleaved text–mask generation capability beyond the GCG benchmark on the more challenging PSG benchmark [18]. This task requires predicting all subject–relation–object triplets in an image, where each subject and object includes its category label and 2D mask. As shown in Tab. A1, our Qwen2.5VL-SAMTok-7B achieves performance comparable to expert models (R@20 = 19.8 vs. 20.6, mR@20 = 15.4 vs. 14.8). This result indicates that the mask-token interface provided by SAMTok enables effective task generalization by converting 2D masks into specialized word tokens. Note that, our model **does not** need any task-specific designs, compared with previous expert models.

Visual Grounding. To assess whether the mask-token interface provided by SAMTok outperforms the native text-box interface of MLLMs on grounding tasks, we also conduct experiments on the RefCOCO, RefCOCO+, and RefCOCOg benchmarks. The results are shown in Tab. A2. Specifically, we de-tokenize mask words into 2D masks and then derive bounding boxes for evaluation. Across both 3B and 7B model sizes, SAMTok yields substantial accuracy improvements while preserving the same natural-language interaction capabilities as the native text-box interface. This further verifies our motivation: the new mask representation (SAMTok) performs better than the point format for visual grounding.

SAMTok integration across MLLMs. SAMTok and MLLM are decoupled: once SAMTok is trained, it can be deployed with any MLLM. To substantiate this, we train and evaluate on two types of MLLMs: (1) models with tile-based visual encoders (e.g., PerceptionLM [2]), and (2) models with adaptive-resolution encoders (e.g., the Qwen-VL [1] series). We use the same training data across settings, and report results in Tab. A3. The mask-token inter-

Table A2. Results on grounding task (REC). We de-tokenize the mask words generated by Qwen25VL-SAMTok into 2D masks and then derive bounding boxes for evaluation.

Method	Size	RefCOCO			RefCOCO+			RefCOCOg	
		val	test A	test B	val	test A	test B	val	test
Qwen25VL [1]	3B	89.1	91.7	84.0	82.4	88.0	74.1	85.2	85.7
Qwen25VL-SAMTok	3B	92.7 (+3.6)	94.6 (+2.9)	89.7 (+5.7)	88.2 (+5.8)	92.2 (+4.2)	84.4 (+10.3)	89.9 (+4.7)	89.6 (+3.9)
Qwen25VL [1]	7B	90.0	92.5	85.4	84.2	89.1	76.9	87.2	87.2
Qwen25VL-SAMTok	7B	93.0 (+3.0)	95.5 (+3.0)	90.5 (+5.1)	88.6 (+4.4)	93.2 (+4.1)	84.4 (+7.5)	90.8 (+3.6)	91.2 (+4.0)

Table A3. Effectiveness of unified mask-token interface for different MLLMs. For the GCG benchmark, we report the average of each metric across the val and test splits; for the GRES benchmark, we report the average across the val, testA, and testB splits.

Method	Size	GCG					GRES			DLC-Bench		
		METEOR	CIDEr	AP50	mIoU	Recall	gIoU	cIoU	N-acc	Pos.	Neg.	Avg.
Qwen25VL-SAMTok	3B	16.8	52.2	36.4	71.3	47.0	70.1	68.9	58.2	45.2	74.8	60.0
Qwen3VL-SAMTok	4B	16.1	50.8	37.6	71.4	47.9	73.6	71.7	65.3	46.1	85.2	65.6
PLM [2]-SAMTok	1B	17.4	56.2	41.9	74.8	51.8	74.7	72.9	66.4	44.3	83.4	63.9

Table A4. Ablation study on the quantization strategy. "1024×2" means two residual codebooks with a size of 1024 each.

Quantization	Codebook Size	r-Acc	g-Acc
VQ [13]	1024	0.50	63.3
VQ [13]	65536	0.66	77.8
FSQ [8]	65536	0.69	78.1
RQ [4]	1024×2	**0.70**	**78.3**

Table A5. Ablation study on the codebook size and quantization steps when using RQ [4].

Codebook Size	r-Acc	g-Acc
1024×2	0.70	**78.3**
1024×4	**0.75**	77.3
256×4	0.72	77.3
256×2	0.70	77.6
512×2	0.71	77.8

face provided by SAMTok works effectively across diverse MLLMs.

D. Ablation Study

Set Up. We evaluate SAMTok's region-mask reconstruction capability on the EntitySeg [9] validation set, which provides 23,754 region masks with high-quality annotations. We use mask IoU as the reconstruction accuracy metric (r-Acc). To assess the mask generation capability of an MLLM integrated with SAMTok, we adopt Qwen2.5-VL-3B [1] as the base model and report the mean cIoU on the val splits of RefCOCO, RefCOCO+, and RefCOCOg in the RES setting as the generation accuracy metric (g-Acc).

Quantization. We study three quantization strategies, including VQ [13], FSQ [8], and RQ [4], as summarized in Tab. A4. When adopting standard VQ, a large codebook is essential, as region mask embeddings in images exhibit high diversity. Reducing the codebook size from 65,536 to 1,024 results in a significant drop in both reconstruction and generative accuracy. With a large codebook, FSQ improves codebook utilization, leading to better reconstruction and generation performance than standard VQ. RQ achieves comparable or even better performance while substantially reducing the codebook size. Specifically, using only two residual codebooks of size 1,024 each (i.e., 1024×2), RQ attains higher reconstruction and generation accuracy than FSQ with a much larger codebook.

Codebook size and quantization steps. We further analyze the effect of codebook size and the number of residual quantization steps on reconstruction and generation performance, as shown in Tab. A5. For the same codebook size, increasing the number of quantization steps (e.g., 1024×4 vs. 1024×2) yields higher-fidelity quantization and thus better reconstruction accuracy (0.75 vs. 0.70 in r-Acc). However, the exponentially expanded search space (1024^4 vs. 1024^2) makes it more difficult for the MLLM to learn to generate mask tokens effectively. In particular, in the dense mask setting, longer words incur much higher computational costs. Therefore, we set the number of residual steps to 2 by default. Under this configuration, increasing the codebook size yields only a limited improvement in reconstruction accuracy but slightly enhances generation accuracy. We finally adopt the 256×2 configuration as our default, since it offers a compact codebook while maintaining a substantial trade-off between reconstruction and generation performance. We use this setting for further experiments, including SFT and RL processes.

E. Visualization

In all visualizations in the main paper and the supplementary material, we represent mask words using quant-code pairs: for example, "<|mt_0011|><|mt_0347|>" is denoted as "<11-347>".

SFT vs. RL. We visualize the improvements brought by RL over SFT in Fig. A1. The gains manifest in three key aspects: (1) higher recall of targets in multi-object grounding scenarios; (2) more accurate localization for expressions involving relative positions and ordering; and (3) improved mask quality.

Mask reconstruction. In Fig. A2, we demonstrate SAMTok's ability to reconstruct small objects across a variety of challenging scenarios. Since SAMTok and the MLLM are decoupled, the mask reconstruction capability is unaffected by any subsequent MLLM training. In contrast, other mask tokenizers [14, 16] require joint training with the MLLM, which ultimately leads to degraded mask reconstruction performance (with all masks reconstructed as ellipses). Thus, our method yields stronger mask reconstruction than these methods. Such mask reconstruction generalization further leads to strong generalization in MLLMs.

PSG visualizations. Fig. A3 shows PSG predictions. Each example contains subject–relation–object triplets where both subject and object include their segmentation masks. Thanks to the unified mask-token interface, the MLLM can jointly generate structured relational descriptions and pixel-aligned masks, enabling dense, panoptic-level scene understanding.

GRES visulizations. Fig. A4 illustrates GRES predictions. SAMTok allows the MLLM to resolve complex referring expressions, including attribute-dependent, context-dependent, and part-level descriptions. The generated masks accurately track the intended region even under heavy occlusion or clutter.

Region caption visualizations. Fig. A5 presents region captioning results. Given a region mask (tokenized into two special tokens), the MLLM produces detailed and contextually relevant descriptions. The compact mask tokens remove ambiguity inherent in bounding-box-based grounding and lead to more precise and consistent region-level captioning.

GCG visualizations. Fig. A6 shows examples where the model simultaneously generates textual captions and region masks for phrases mentioned in the narrative. SAMTok provides a lightweight and efficient mechanism for linking each phrase with a precise pixel region, enabling both high-quality caption generation and aligned mask prediction within one unified response.

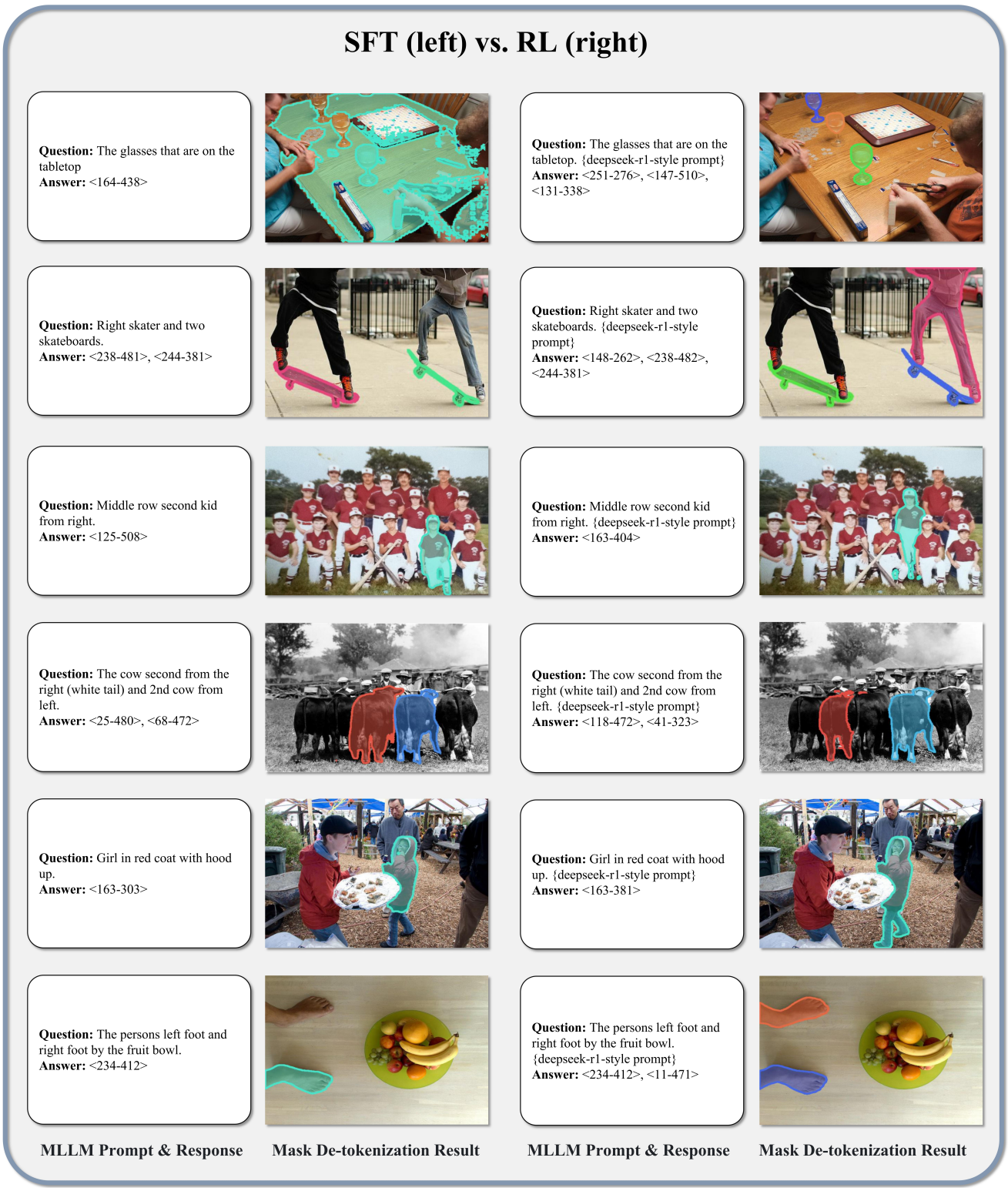

Figure A1. SFT vs. RL. Examples are sampled from the GRES benchmark. RL finds more target objects, localizes relative positions better, and produces cleaner masks than SFT across diverse scenes.

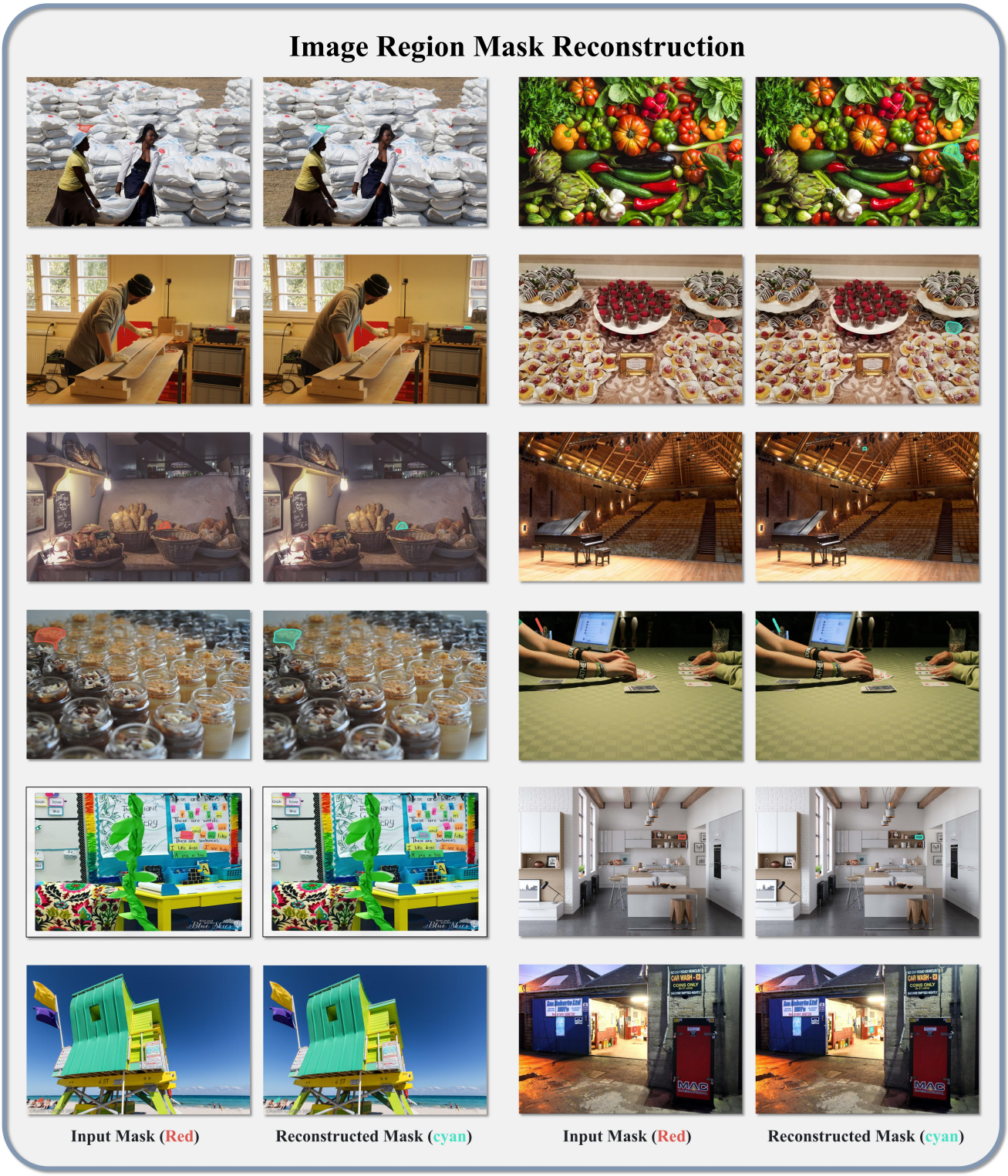

Figure A2. Region mask reconstruction examples. For each region, the ground-truth mask is tokenized into two discrete codes, and SAMTok reconstructs the mask solely from the original image and the quantized mask tokens. SAMTok preserves fine structures for small, thin, or irregular objects even under challenging lighting or clutter. Since SAMTok is fully decoupled from the MLLM, its reconstruction quality remains stable regardless of downstream model training—unlike joint-training mask tokenizers that tend to collapse to elliptical or blurred masks.

Figure A3. Panoptic scene graph generation(PSG) examples. The model predicts subject–relation–object triplets where both subject and object categories are paired with their corresponding segmentation masks, represented through mask tokens. SAMTok's interface allows the MLLM to generate consistent object masks and relational descriptions simultaneously, demonstrating strong alignment between textual predicates and pixel-grounded regions.

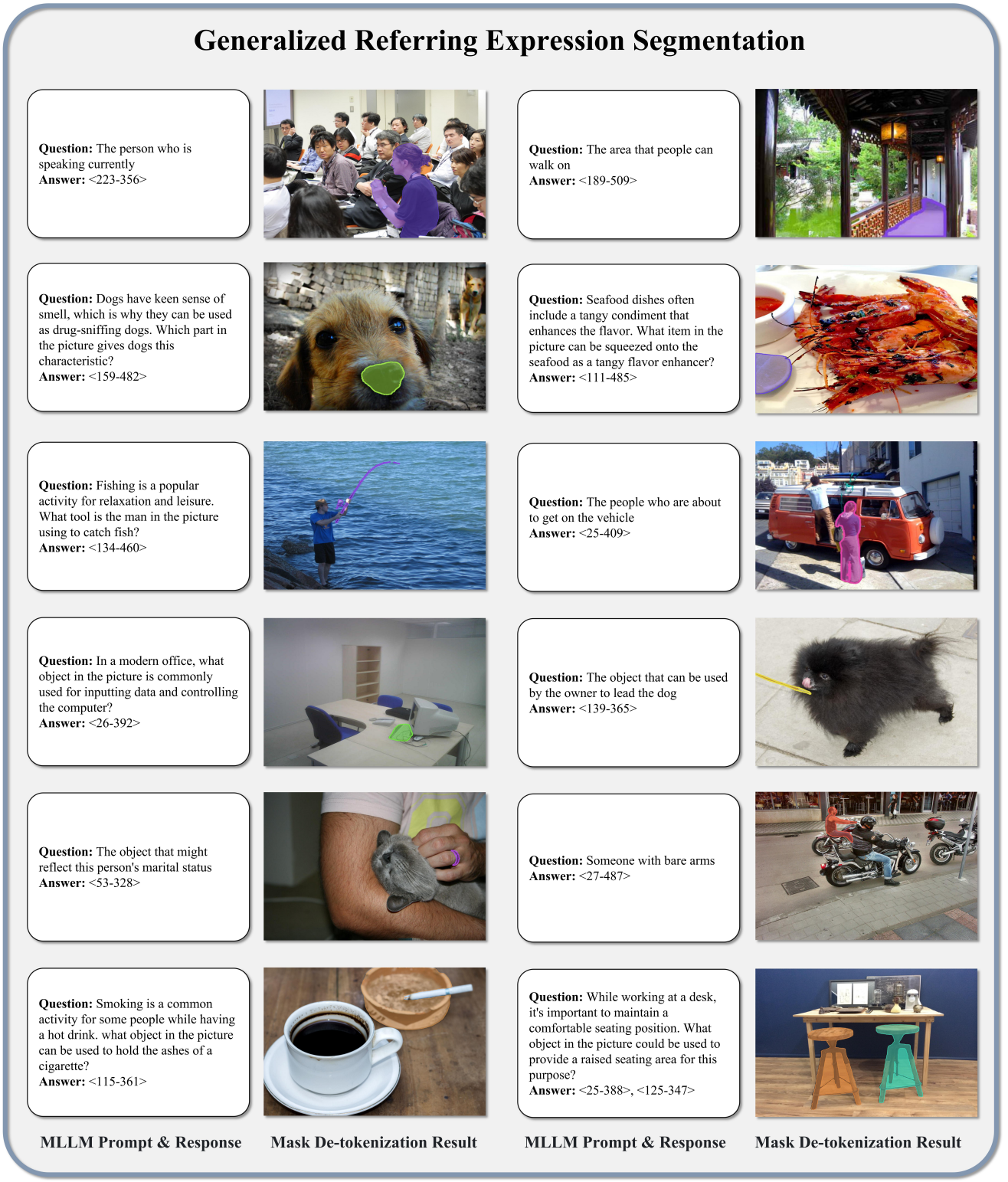

Figure A4. GRES examples. Given a natural-language referring expression, the MLLM outputs two mask tokens that decode into the final segmentation mask. SAMTok enables precise grounding for expressions involving fine attributes, part-level targets, or contextual reasoning. The examples show robustness to ambiguous descriptions, occlusion, and multi-object scenes.

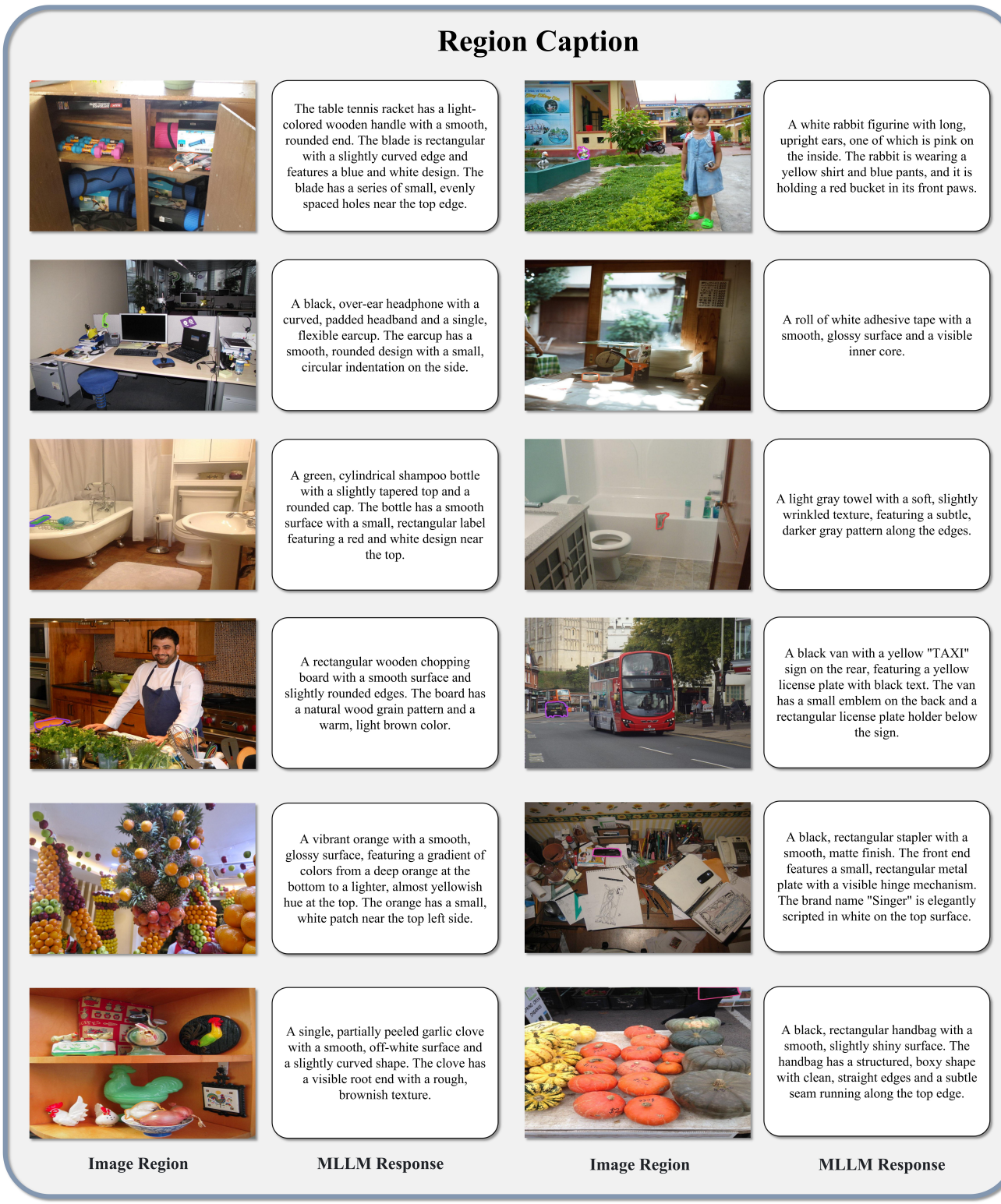

Figure A5. Region Caption examples. Each visualization shows a region mask input (tokenized as two mask tokens) and the model's generated description. SAMTok provides unambiguous spatial grounding, enabling the MLLM to generate accurate and context-aware region descriptions about attributes, roles, and interactions.

Figure A6. GCG examples. The model simultaneously describes the scene and produces region masks for phrases mentioned in the caption. For each highlighted phrase, SAMTok decodes the predicted mask tokens into segmentation masks. SAMTok's compact representation (two tokens per mask) enables efficient, aligned text–mask generation with consistent grounding across multiple phrases within long captions.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2

[2] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025. 1, 2

[3] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023. 1

[4] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022. 2

[5] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, 2020. 1

[6] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[8] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023. 2

[9] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*, 2022. 2

[10] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[11] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1

[12] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019. 1

[13] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2

[14] Lingfeng Wang, Hualing Lin, Senda Chen, Tao Wang, Changxu Cheng, Yangyang Zhong, Dong Zheng, and Wuyue Zhao. Alto: Adaptive-length tokenizer for autoregressive mask generation. *arXiv preprint arXiv:2505.16495*, 2025. 3

[15] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[16] Tao Wang, Changxu Cheng, Lingfeng Wang, Senda Chen, and Wuyue Zhao. Himtok: Learning hierarchical mask tokens for image segmentation with large multimodal model. In *ICCV*, 2025. 3

[17] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 1

[18] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, 2022. 1

[19] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 1

[20] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025. 1