

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 $\langle |mt_0000| \rangle \sim \langle |mt_0511| \rangle$. Among them, $\langle |mt_0000| \rangle \sim \langle |mt_0255| \rangle$ correspond to the first-level codebook, and $\langle |mt_0256| \rangle \sim \langle |mt_0511| \rangle$ correspond to the second-level codebook. In addition, we introduce two special tokens, $\langle |mt_start| \rangle$ and $\langle |mt_end| \rangle$, 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]	20.0	9.1
VCTree [12]	20.6	9.7
PSGFormer [18]	18.0	14.8
Qwen2.5VL-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	<u>0.69</u>	<u>78.1</u>
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	<u>0.72</u>	77.3
256×2	0.70	77.6
512×2	0.71	<u>77.8</u>

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 (1024⁴ vs. 1024²) 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.

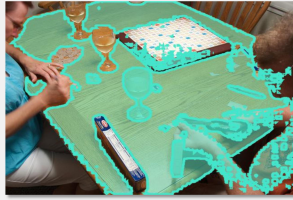
GRES visualizations. Fig. A4 illustrates GRES predictions. SAMTok allows the MLLM to resolve complex referring expressions, including attribute-dependent, context-dependent, and part-level descriptions. The generated masks accurately track the intended region even under heavy occlusion or clutter.

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.

SFT (left) vs. RL (right)

Question: The glasses that are on the tabletop
Answer: <164-438>



Question: The glasses that are on the tabletop. {deepseek-r1-style prompt}
Answer: <251-276>, <147-510>, <131-338>



Question: Right skater and two skateboards.
Answer: <238-481>, <244-381>



Question: Right skater and two skateboards. {deepseek-r1-style prompt}
Answer: <148-262>, <238-482>, <244-381>



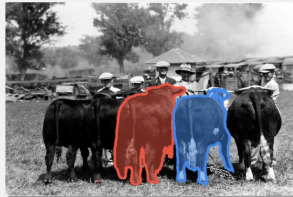
Question: Middle row second kid from right.
Answer: <125-508>



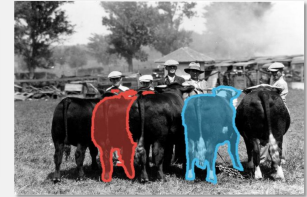
Question: Middle row second kid from right. {deepseek-r1-style prompt}
Answer: <163-404>



Question: The cow second from the right (white tail) and 2nd cow from left.
Answer: <25-480>, <68-472>



Question: The cow second from the right (white tail) and 2nd cow from left. {deepseek-r1-style prompt}
Answer: <118-472>, <41-323>



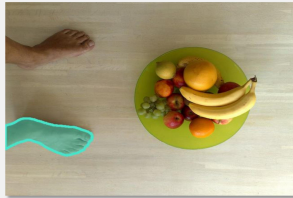
Question: Girl in red coat with hood up.
Answer: <163-303>



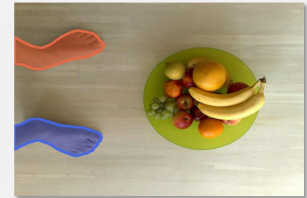
Question: Girl in red coat with hood up. {deepseek-r1-style prompt}
Answer: <163-381>



Question: The persons left foot and right foot by the fruit bowl.
Answer: <234-412>



Question: The persons left foot and right foot by the fruit bowl. {deepseek-r1-style prompt}
Answer: <234-412>, <11-471>



MLLM Prompt & Response

Mask De-tokenization Result

MLLM Prompt & Response

Mask De-tokenization Result

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.

Image Region Mask Reconstruction

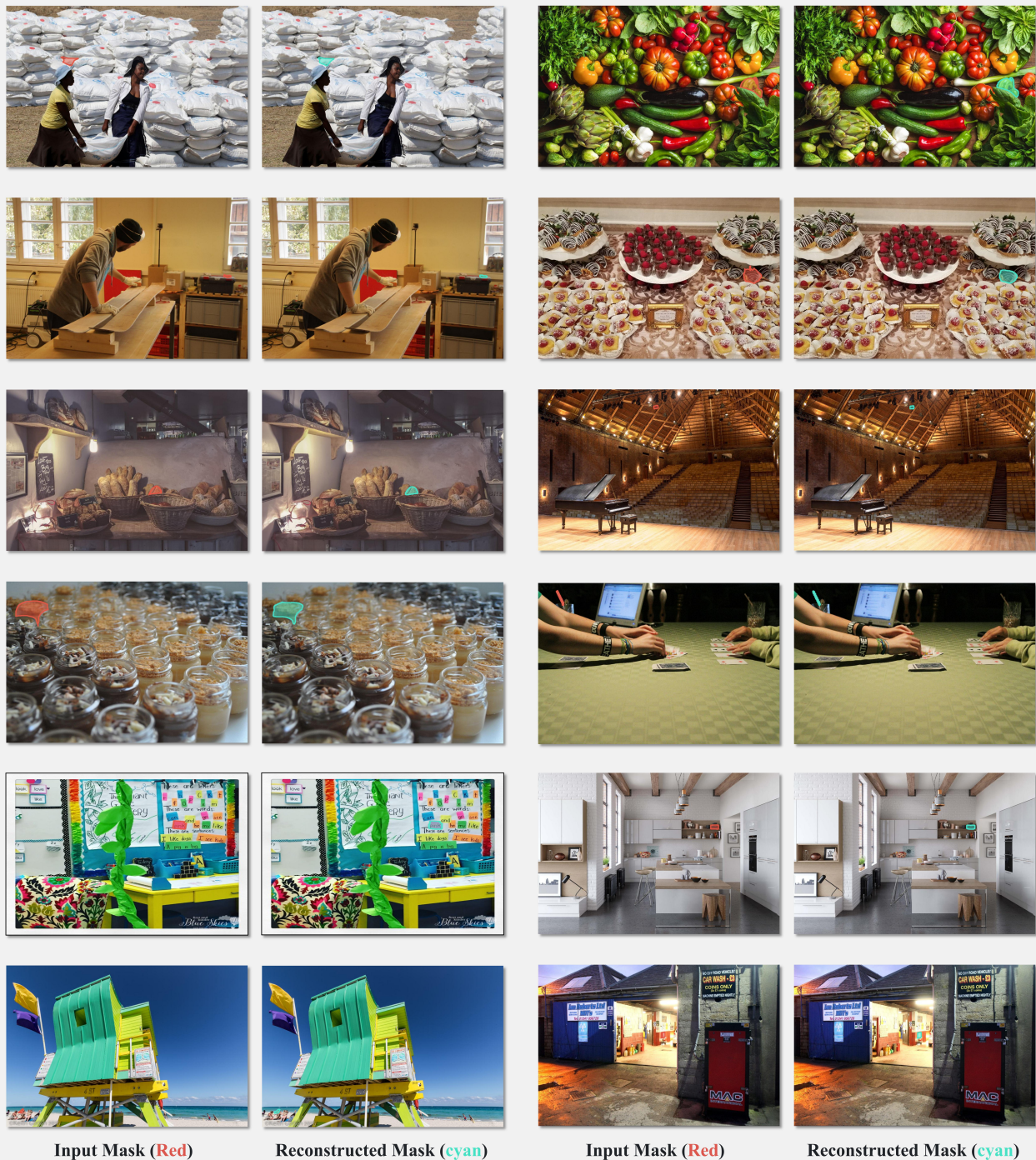


Figure A2. Region mask reconstruction examples. For each region, the ground-truth mask is tokenized into two discrete codes, and SAMTok reconstructs the mask solely from the original image and the quantized mask tokens. SAMTok preserves fine structures for small, thin, or irregular objects even under 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.

Panoptic Scene Graph Generation

person<163-483> beside elephant<17-282>,
 person<163-483> standing on river<52-322>
 person<255-472> beside person<163-483>
 person<255-472> standing on river<52-322>
 elephant<17-282> in river<52-322>



cat<44-497> standing on bed<220-425>
 cup<144-311> on cabinet<65-322>
 light<12-261> beside cup<144-311>



bird<236-347> standing on tree-merged<87-356>
 bird<64-484> standing on tree-merged<87-356>
 bird<11-313> standing on tree-merged<87-356>
 bird<7-364> standing on tree-merged<87-356>
 bird<219-408> standing on tree-merged<87-356>
 bird<212-368> standing on tree-merged<87-356>
 bird<150-296> standing on tree-merged<87-356>



cat<183-384> lying on couch<44-307>
 chair<66-363> on rug<186-451>
 couch<73-452> beside couch<44-307>
 couch<73-452> on rug<186-451>
 couch<44-307> on rug<186-451>
 potted plant<31-299> on shelf<90-466>
 potted plant<31-339> on couch<44-307>
 potted plant<83-314> on window<89-484>
 vase<155-393> on shelf<90-466>
 table<25-288> on rug<186-451>



person<125-303> standing on playingfield<52-475>
 person<3-494> wearing baseball glove<222-393>
 person<3-494> walking on playingfield<52-475>
 person<57-378> holding baseball bat<149-394>



person<207-400> in bus<3-475>
 person<35-486> in bus<3-475>
 person<176-258> in bus<3-475>
 car<41-471> parked on road<186-494>
 car<111-267> parked on road<186-494>
 bus<3-475> driving on road<186-494>
 road<186-494> attached to pavement<66-356>
 sky<112-399> over tree<135-425>



person<102-486> standing on bridge<87-316>
 person<210-338> sitting on bridge<87-316>
 boat<46-494> on sea<186-360>
 bird<69-481> flying over sky<135-472>
 bird<200-446> flying over sky<135-472>
 bird<23-278> flying over sky<135-472>
 bird<33-398> flying over sky<135-472>
 bird<64-398> flying over sky<135-472>
 bird<21-406> flying over sky<135-472>
 bird<183-399> flying over sky<135-472>
 bird<77-442> flying over sky<135-472>
 sky<135-472> over bridge<87-316>



person<163-467> beside person<3-338>
 person<163-467> standing on net<156-497>
 person<73-483> carrying backpack<203-303>
 person<73-483> holding tennis racket<84-499>
 person<73-483> standing on net<156-497>
 person<230-300> beside person<125-359>
 person<192-263> beside person<209-392>
 person<192-263> standing on net<156-497>
 person<25-409> standing on net<156-497>
 person<10-401> holding tennis racket<184-258>
 person<10-401> standing on net<156-497>
 person<122-415> beside person<125-359>



person<228-505> looking at kite<203-399>
 person<228-505> standing on pavement<52-356>
 kite<203-399> beside kite<237-318>
 kite<237-318> beside kite<143-373>
 flower<44-487> beside pavement<52-356>
 pavement<52-356> attached to grass<115-269>



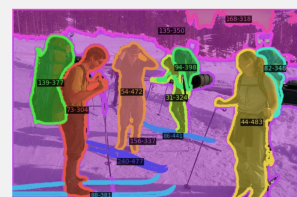
person<180-481> sitting on chair<218-412>
 boat<84-368> on sea<185-363>
 boat<248-365> on sea<185-363>
 boat<67-498> on sea<185-363>
 umbrella<71-320> over person<180-481>
 chair<218-412> on pavement-merged<186-322>
 sea<185-363> beside pavement-merged<186-322>
 sky-other-merged<135-322> over mountain-merged<197-490>



cat<25-300> on couch<52-290>
 cat<25-300> lying on couch<52-290>
 cat<25-300> beside potted plant<129-391>
 cat<25-300> beside wall<168-503>
 couch<52-290> attached to wall<168-503>
 potted plant<129-391> beside couch<52-290>
 remote<236-367> beside cat<25-300>
 remote<236-367> beside remote<77-325>
 remote<236-367> on couch<52-290>
 remote<77-325> on couch<52-290>



person<73-304> carrying backpack<139-377>
 person<73-304> standing on skis<88-381>
 person<54-472> carrying backpack<94-398>
 person<54-472> standing on skis<240-477>
 person<44-483> carrying backpack<82-348>
 person<44-483> standing on snow<156-337>
 person<31-324> standing on skis<86-441>
 skis<88-381> on snow<156-337>
 skis<240-477> on snow<156-337>
 skis<86-441> on snow<156-337>



MLLM Response

Mask De-tokenization Result

MLLM Response

Mask De-tokenization Result

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.

Generalized Referring Expression Segmentation

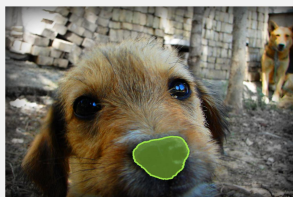
Question: The person who is speaking currently
Answer: <223-356>



Question: The area that people can walk on
Answer: <189-509>



Question: Dogs have keen sense of smell, which is why they can be used as drug-sniffing dogs. Which part in the picture gives dogs this characteristic?
Answer: <159-482>



Question: Seafood dishes often include a tangy condiment that enhances the flavor. What item in the picture can be squeezed onto the seafood as a tangy flavor enhancer?
Answer: <111-485>



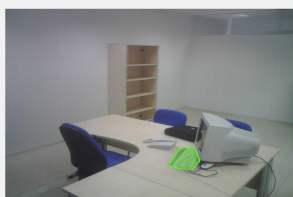
Question: Fishing is a popular activity for relaxation and leisure. What tool is the man in the picture using to catch fish?
Answer: <134-460>



Question: The people who are about to get on the vehicle
Answer: <25-409>



Question: In a modern office, what object in the picture is commonly used for inputting data and controlling the computer?
Answer: <26-392>



Question: The object that can be used by the owner to lead the dog
Answer: <139-365>



Question: The object that might reflect this person's marital status
Answer: <53-328>



Question: Someone with bare arms
Answer: <27-487>



Question: Smoking is a common activity for some people while having a hot drink. What object in the picture can be used to hold the ashes of a cigarette?
Answer: <115-361>



Question: While working at a desk, it's important to maintain a comfortable seating position. What object in the picture could be used to provide a raised seating area for this purpose?
Answer: <25-388>, <125-347>



MLLM Prompt & Response

Mask De-tokenization Result

MLLM Prompt & Response

Mask De-tokenization Result

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.

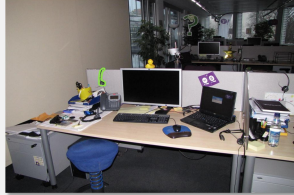
Region Caption



The table tennis racket has a light-colored wooden handle with a smooth, rounded end. The blade is rectangular with a slightly curved edge and features a blue and white design. The blade has a series of small, evenly spaced holes near the top edge.



A white rabbit figurine with long, upright ears, one of which is pink on the inside. The rabbit is wearing a yellow shirt and blue pants, and it is holding a red bucket in its front paws.



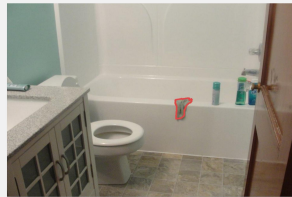
A black, over-ear headphone with a curved, padded headband and a single, flexible earcup. The earcup has a smooth, rounded design with a small, circular indentation on the side.



A roll of white adhesive tape with a smooth, glossy surface and a visible inner core.



A green, cylindrical shampoo bottle with a slightly tapered top and a rounded cap. The bottle has a smooth surface with a small, rectangular label featuring a red and white design near the top.



A light gray towel with a soft, slightly wrinkled texture, featuring a subtle, darker gray pattern along the edges.



A rectangular wooden chopping board with a smooth surface and slightly rounded edges. The board has a natural wood grain pattern and a warm, light brown color.



A black van with a yellow "TAXI" sign on the rear, featuring a yellow license plate with black text. The van has a small emblem on the back and a rectangular license plate holder below the sign.



A vibrant orange with a smooth, glossy surface, featuring a gradient of colors from a deep orange at the bottom to a lighter, almost yellowish hue at the top. The orange has a small, white patch near the top left side.



A black, rectangular stapler with a smooth, matte finish. The front end features a small, rectangular metal plate with a visible hinge mechanism. The brand name "Singer" is elegantly scripted in white on the top surface.



A single, partially peeled garlic clove with a smooth, off-white surface and a slightly curved shape. The clove has a visible root end with a rough, brownish texture.



A black, rectangular handbag with a smooth, slightly shiny surface. The handbag has a structured, boxy shape with clean, straight edges and a subtle seam running along the top edge.

Image Region

MLLM Response

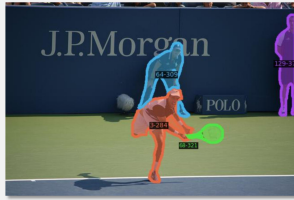
Image Region

MLLM Response

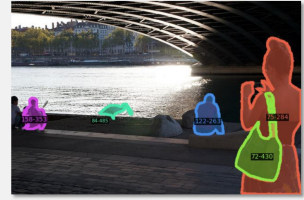
Figure A5. Region Caption examples. Each visualization shows a region mask input (tokenized as two mask tokens) and the model's generated description. SAMTok provides unambiguous spatial grounding, enabling the MLLM to generate accurate and context-aware region descriptions about attributes, roles, and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

Grounded Conversation Generation

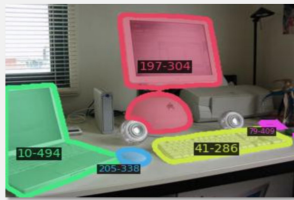
A tennis player <3-284> is in action on the court, holding a tennis racket <68-321>. He is wearing a white outfit and a cap. Behind him, a ball boy <64-309> is crouched near a J.P. Morgan sign. Another person <129-378> stands to the side, observing. The scene is set on a green and blue tennis court with a "POLO" sign visible in the background. The atmosphere suggests a professional tennis match.



A woman <75-284> stands under a bridge, holding a phone and a handbag <72-430>. She is near a riverbank where a swan <84-485> is walking on the grass. A man in a red shirt <122-263> sits on a stone bench, while another person <158-353> sits nearby. The scene is illuminated by sunlight reflecting off the water, with trees and buildings in the background. The atmosphere is calm and serene.



A desk setup features a white Apple iMac <197-304> with a large screen and a matching keyboard <41-286>. To the left, an open laptop <10-494> is visible. A white mouse <205-338> sits in front of the iMac, and another mouse <79-409> is near the laptop. The desk also holds a small speaker and a few other items, creating a functional workspace. The background includes a window with blinds and a lamp.



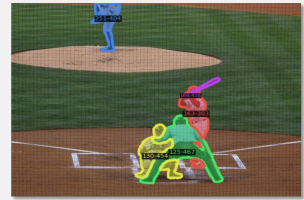
A classic red double-decker bus <198-315> is prominently featured on a city street, displaying route number 15 to Aldwych. The bus has a vintage design with a rounded front and large windows. In the background, other buses <193-317>, <39-433> can be seen, along with modern buildings and greenery. A few pedestrians <80-258>, <46-271> are visible near the sidewalk, adding to the urban scene. The overall atmosphere suggests a bustling city environment with a mix of historical and contemporary elements.



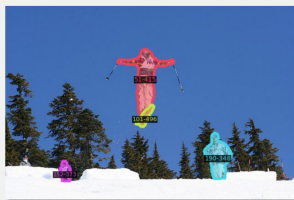
The image shows a cozy room with a bed <125-263> covered in a colorful patterned blanket. A small dining table <234-414> with a chair <191-286> is positioned near the window, which offers a view of an adjacent building. The room features a stone wall and a purple accent wall. A trash bin is visible near the bed. The space is well-lit with natural light streaming through the windows. The overall ambiance is simple and functional, with a focus on comfort and practicality.



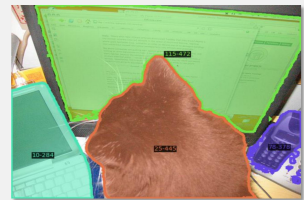
A baseball game is in progress, with the batter <163-303> preparing to swing at the incoming pitch. The catcher <130-454> is crouched behind home plate, ready to catch the ball, while the umpire <125-467> stands behind the catcher, observing the play. The pitcher <251-404> is on the mound, having just thrown the ball. The scene is viewed through a protective net, and the field's green grass and brown dirt are clearly visible. The batter holds a bat <166-436>, poised for action.



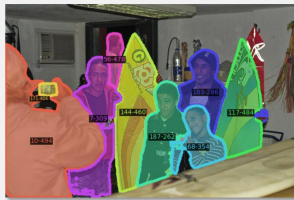
A skier <51-415> is mid-air, performing a jump with skis <101-496> crossed. The backdrop features a clear blue sky and evergreen trees. Another person <190-348> stands on the snow, observing the action. A child <152-373> is visible in the background, near the edge of the slope. The scene captures the dynamic movement and excitement of skiing in a snowy, forested area. The skier's colorful outfit contrasts with the white snow and blue sky.



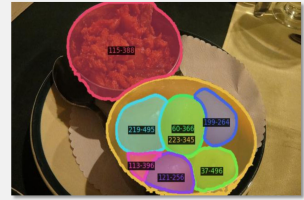
A fluffy cat <25-445> is sitting in front of a computer monitor <115-472>, partially obscuring the screen. To the left, a laptop <10-284> is open, and to the right, a cordless phone <78-378> rests on the desk. The monitor displays a webpage with text, and the scene suggests a typical home office setup. The cat appears to be the focal point, adding a touch of warmth to the workspace.



A group of people is gathered indoors, holding surfboards. One person <10-494> is taking a photo with a cell phone <171-406>. The surfboards <144-460>, <117-484> are prominently displayed, featuring vibrant designs. The individuals <187-262>, <68-354>, <183-286>, <7-309>, <56-478> appear to be posing for the picture, with some smiling and others looking at the camera. The setting includes a casual indoor environment with various objects in the background. The atmosphere suggests a fun and social gathering centered around surfing.



The image shows a plate with two bowls. One bowl <115-301> contains a chunky, red sauce, possibly a curry or stew. The other bowl <223-345> holds peeled, sliced root vegetables, likely carrots <37-496>, <60-366>, <219-495>, <199-264>, <121-256>, <113-396>. The vegetables appear to be pickled or preserved, as they have a glossy, translucent appearance. The plate is set on a table with a beige tablecloth, and part of a napkin is visible under the bowls. The presentation suggests a meal with a focus on flavorful, hearty dishes.



A motorcyclist <3-345> is riding a white motorcycle <25-281> on a road, leaning into a turn. The rider is wearing a white helmet and a matching racing suit. Spectators <27-334>, <42-443>, <89-411>, <161-467>, <192-298>, <175-373>, <110-486> stand behind a wooden fence, watching the action. The scene is set in a grassy area with a clear path for the motorcycle. The rider appears focused, navigating the curve with precision. The atmosphere suggests a competitive or exhibition event.



Two young girls <3-430>, <163-267> are interacting with two sheep <87-375>, <218-402>. The girl on the left is petting a black and white sheep, while the one on the right is smiling at the camera. In the background, a person <177-286> is walking near a fence. The setting appears to be an outdoor area with grass and trees, suggesting a farm or petting zoo environment. The scene captures a moment of joy and connection between the children and the animals.



MLLM Response

Mask De-tokenization Result

MLLM Response

Mask De-tokenization Result

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, Sibao 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 Tschanen. 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