Towards Semantic Equivalence of Tokenization in Multimodal LLM

Shengqiong Wu[1,2] Hao Fei[1] Xiangtai Li[2] Jiayi Ji[1] Hanwang Zhang[3] Tat-Seng Chua[1] Shuicheng Yan[4,1]

▶ [1]National University of Singapore ▶ [2]ByteDanse Seed
▶ [3]Nanyang Technological University ▶ [4]Skywork AI, Singapore

https://sqwu.top/SeTok-web/

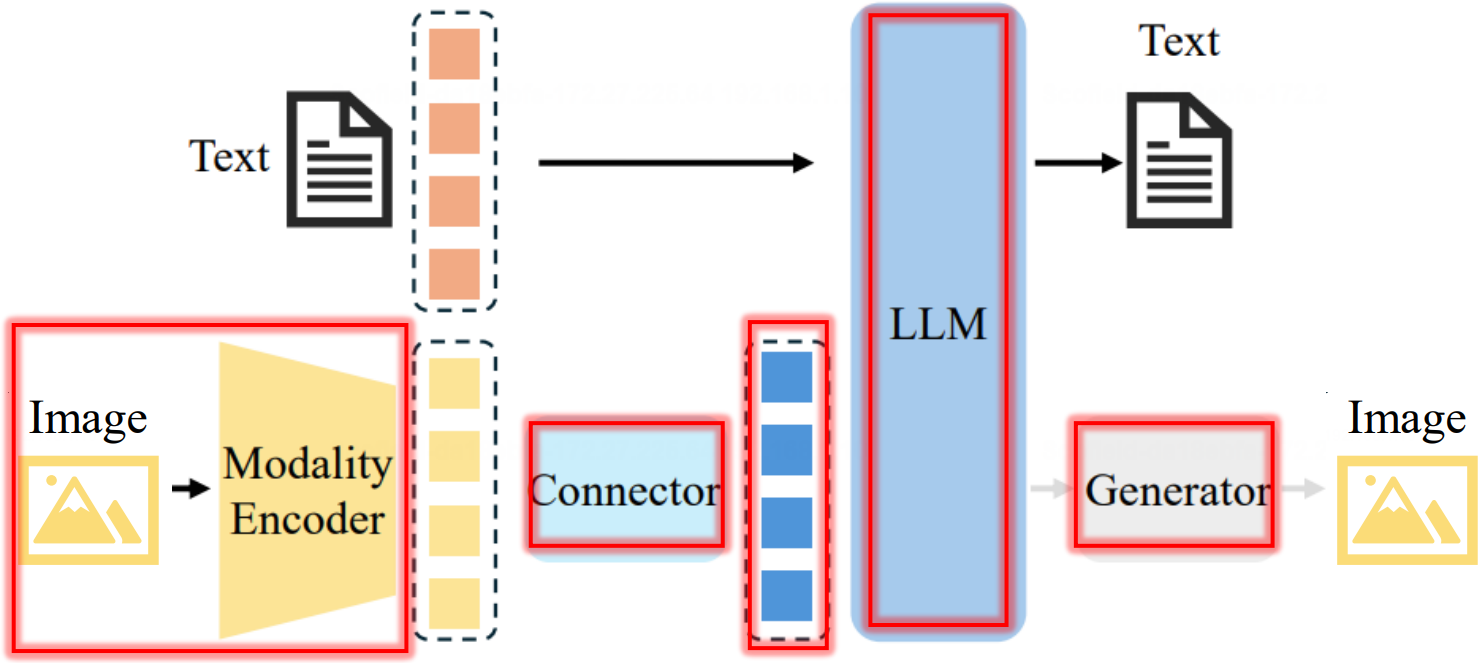
Content

Content

Introduction

➢ Existing MLLMs

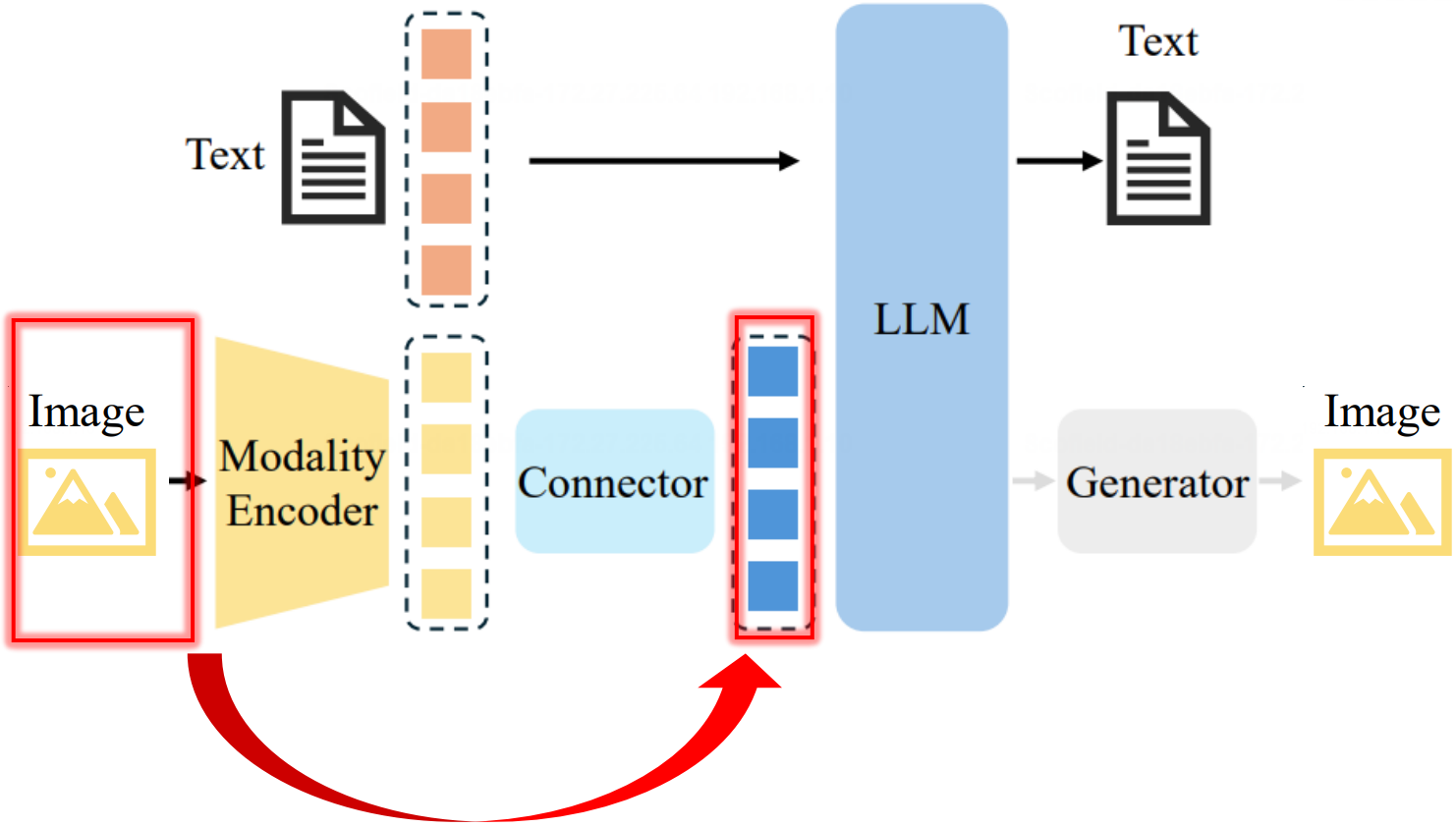
- *LLM* as the core decision-making module
- Adding external non-textual encoder and decoder for perceiving and generating non-textual modalities



Introduction

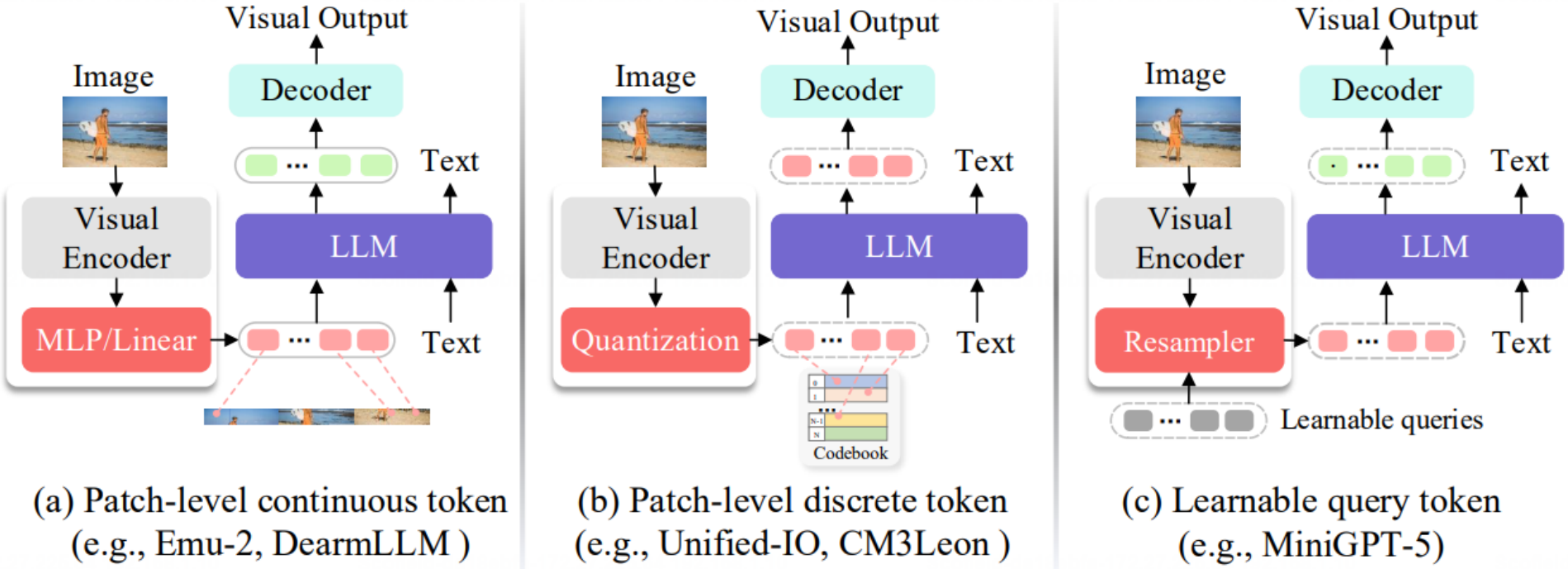
➢ Existing MLLMs

- *Key design*: **vision tokenization**, i.e., converting input visual signals into visual tokens



Introduction

- Existing MLLMs
  - *Vision Tokenization*
    - Patch-level continuous token
    - Patch-level discrete token
    - Learnable query token



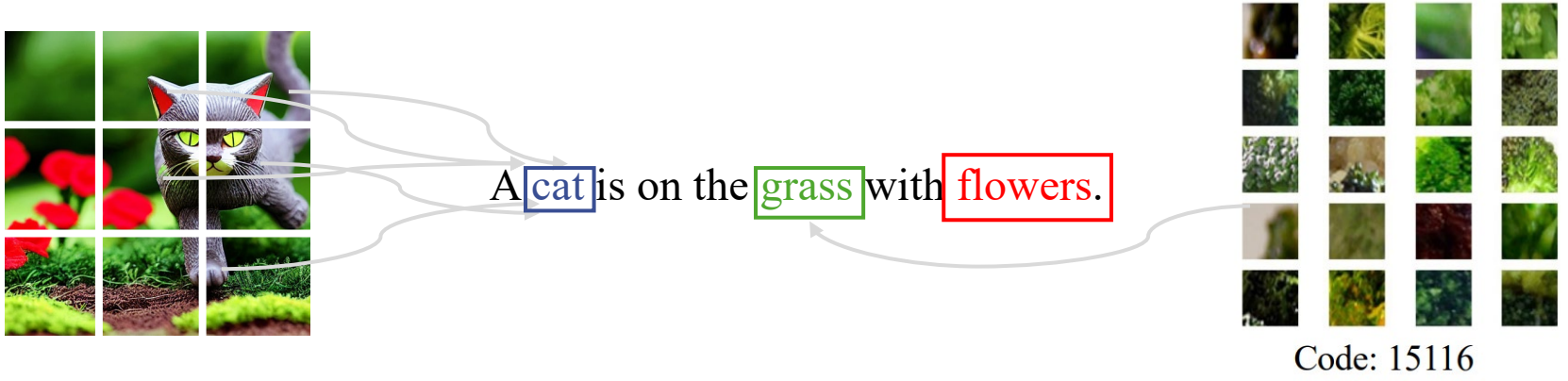
(a) Patch-level continuous token (e.g., Emu-2, DearmLLM)

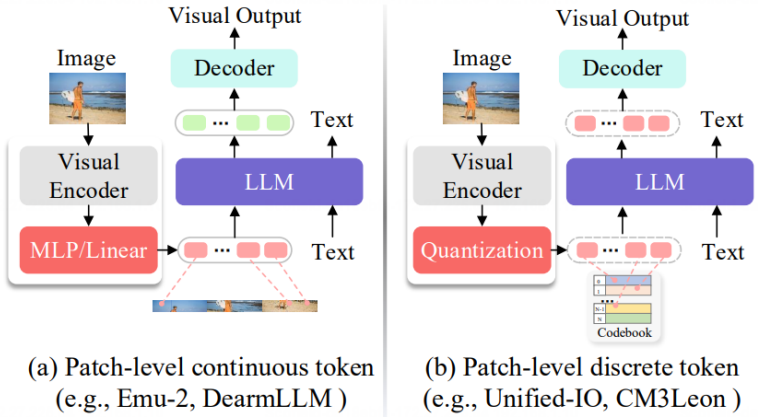
(b) Patch-level discrete token (e.g., Unified-IO, CM3Leon)

(c) Learnable query token (e.g., MiniGPT-5)

Introduction

- Existing MLLMs
  - Patch-level continuous/discrete token
    - Fixed patch squares, fragmenting objects across multiple patches and disrupting the integrity of visual semantic units
    - Codebook introduce information loss



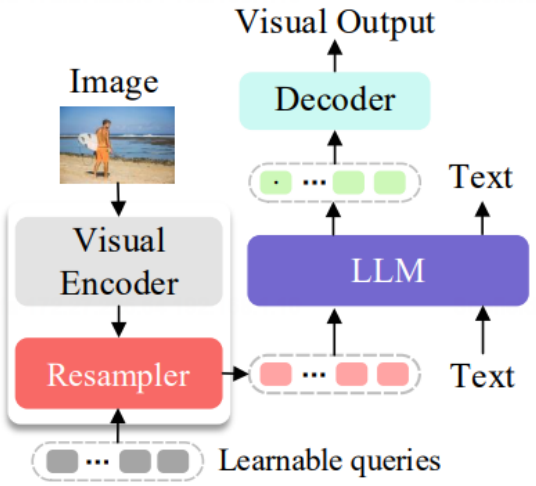


(a) Patch-level continuous token (e.g., Emu-2, DearmLLM)

(b) Patch-level discrete token (e.g., Unified-IO, CM3Leon)

Introduction

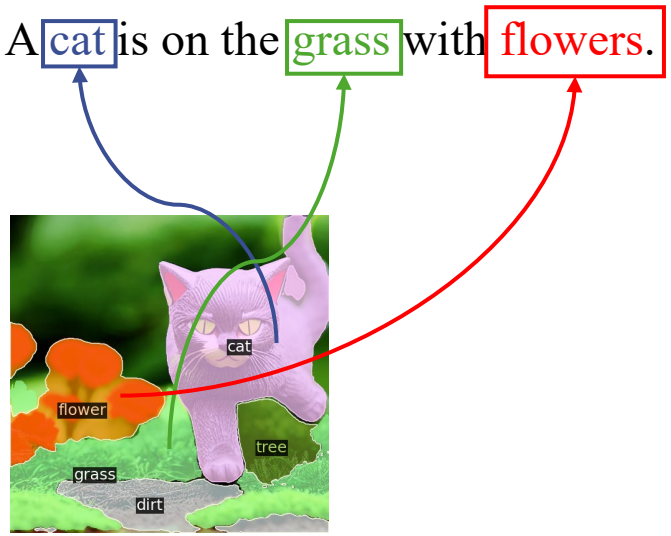
- Existing MLLMs
  - Learnable query token
    - Struggle to align with actual visual semantic units and meanwhile offer limited interpretability



(c) Learnable query token (e.g., MiniGPT-5)

Introduction

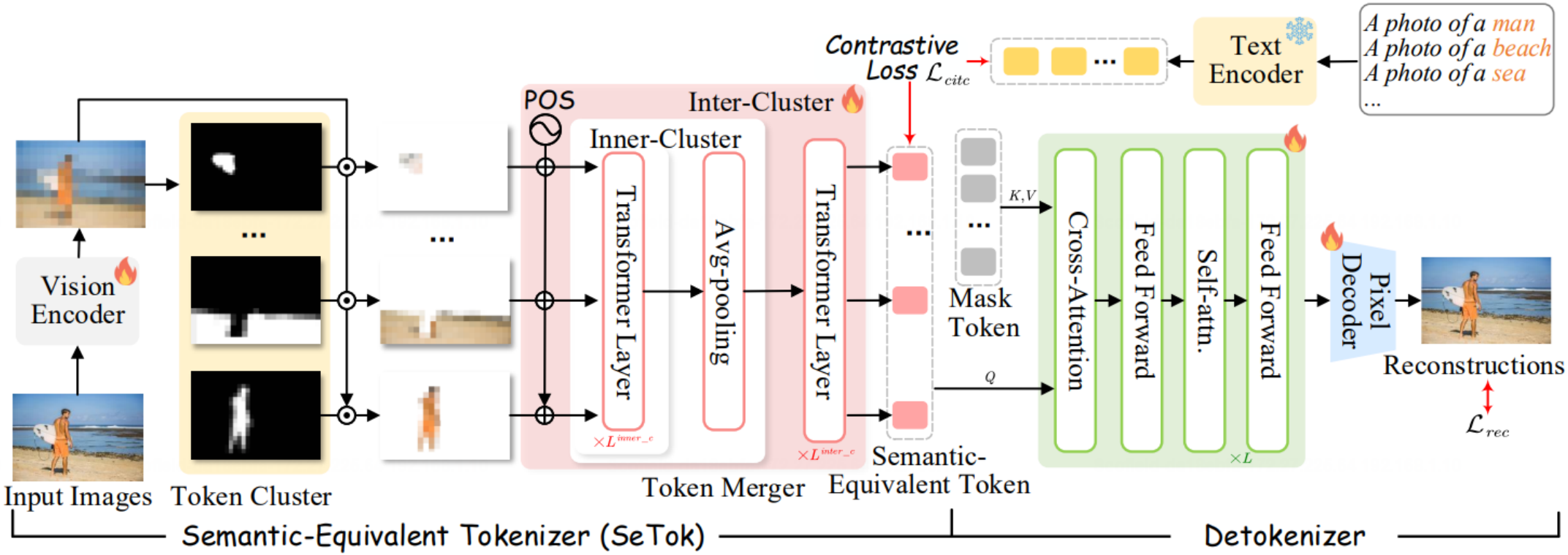
- Existing MLLMs
  - Semantic-Equivalent Token
    - Well-encapsulated semantic units, semantically align to the linguistic tokens



Content

Method

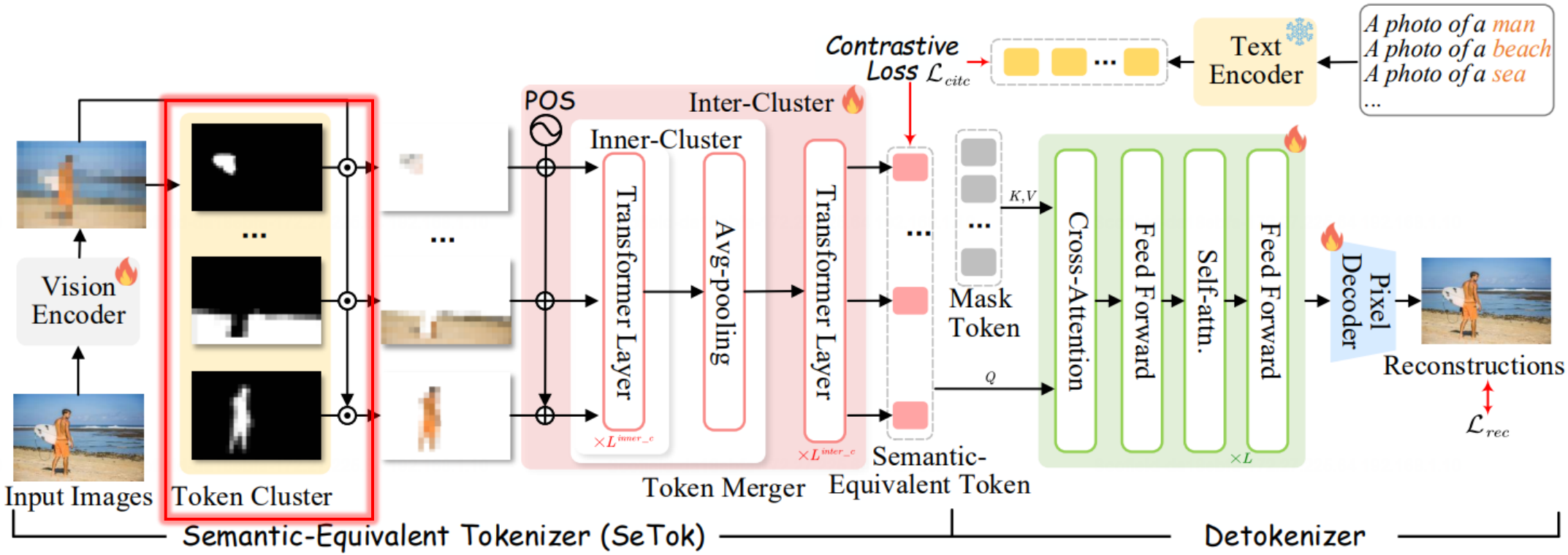
- Semantic-Equivalent Tokenizer (SeTok)
- SETOKIM integrated with SeTok



Method

- Semantic-Equivalent Tokenizer (SeTok)
  - Token Cluster

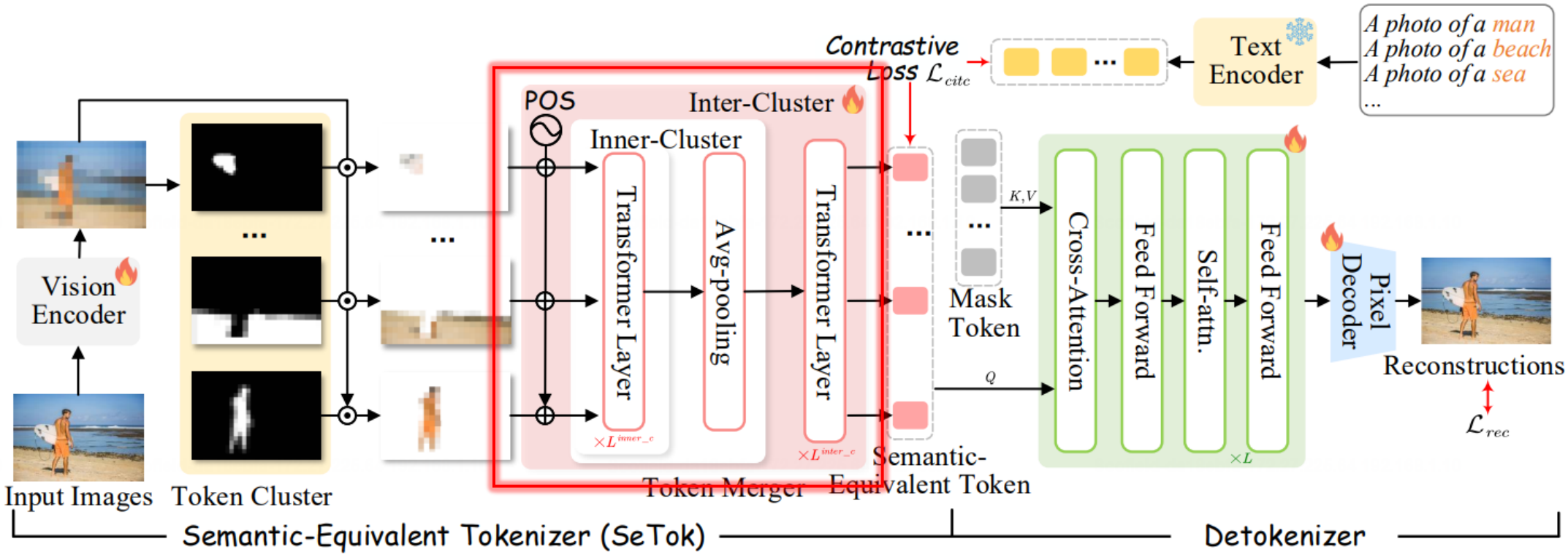
Take the visual patch embeddings X as input and then assign individual patches into a semantic complete cluster, which can be formulated as obtaining a variable number of concept masks



Method

- Semantic-Equivalent Tokenizer (SeTok)
  - Token Merger

Adopt a token merger that aggregates visual embeddings beyond merely using cluster centers as definitive vision tokens

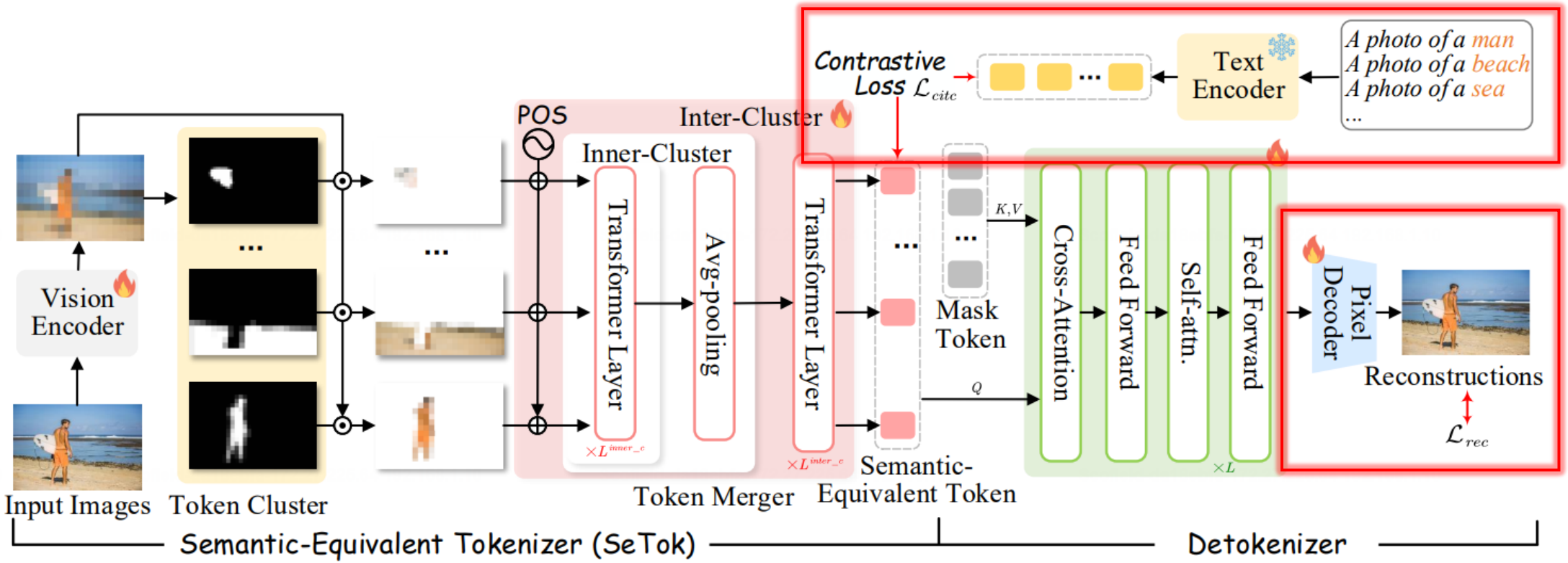


Method

- Semantic-Equivalent Tokenizer (SeTok)
  - SeTok Training

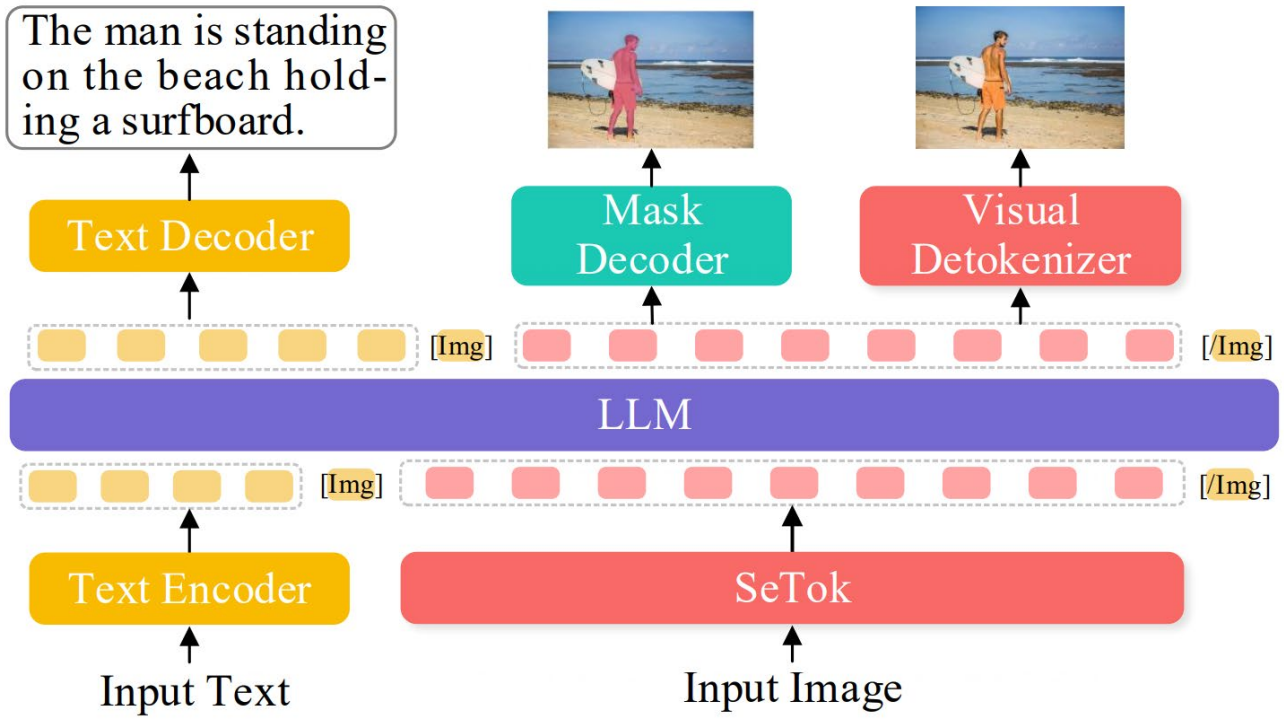
    Complete and enriched high-level semantic information - concept-level image-text contrastive loss

    undistorted pixel-level details - image reconstruction loss



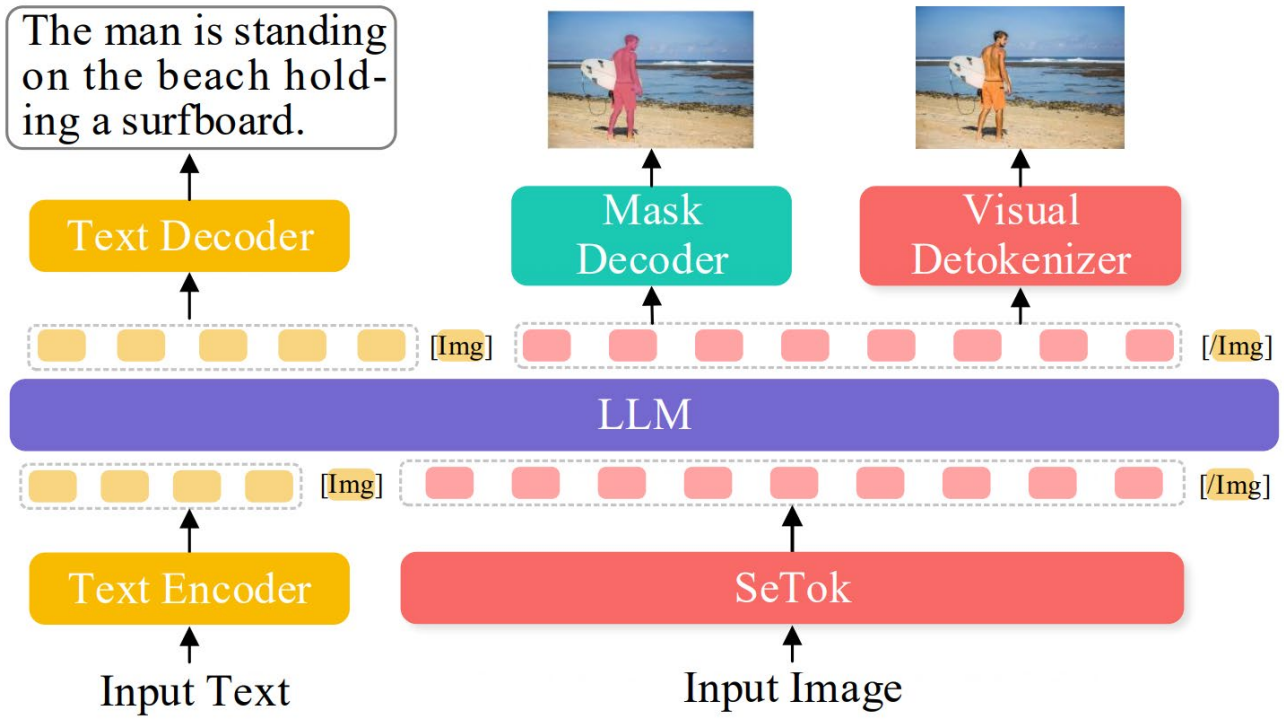
Method

- SETOKIM integrated with SeTok
  - The input image will be tokenized into a sequence of semantic-equivalent visual tokens by SeTok
  - Then concatenated with text tokens to form a unified multimodal sequence
  - The backbone LLM subsequently processes this multimodal sequence to perform multimodal understanding and generation



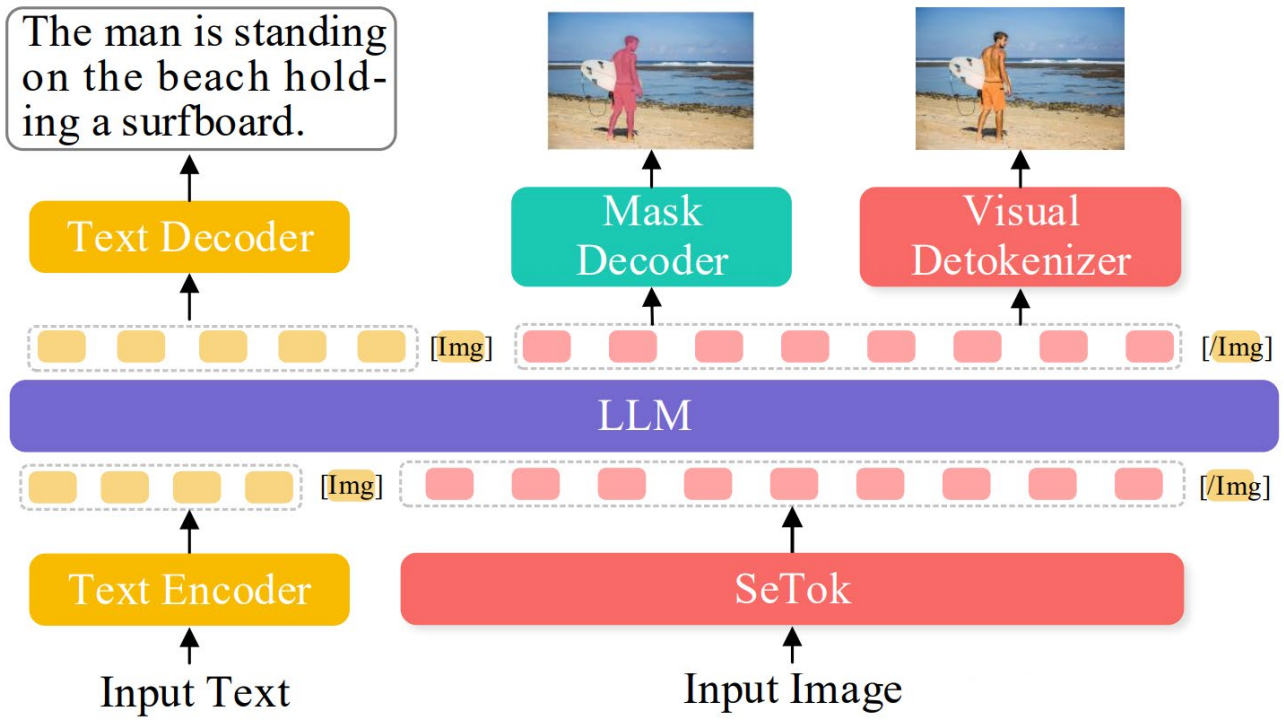
Method

- SETOKIM integrated with SeTok
  - **Stage-I: Multimodal Pretraining**
    - Leverage ImageNet-1K and 28M text-image pair dataset, to train the model for conditional image generation and image captioning



Method

- SETOKIM integrated with SeTok
  - **Stage-II: Instruction Tuning**
    - Perform multimodal instruction tuning with both public datasets covering multimodal instruction dataset



Content

Experiments

- The Quality of SeTok

Model	#Tokens	Latent size	rFID ↓	Top-1 ↑
VQ-GAN (Esser et al., 2021)	Fixed	16 × 16	7.94	-
VAE (Rombach et al., 2022)	Fixed	32 × 32	2.63	-
RQ-VAE (Lee et al., 2022)	Fixed	16 × 16	3.20	-
ViT-VQGAN (Yu et al., 2022)	Fixed	32 × 32	**1.28**	-
MQ-VAE (Huang et al., 2023)	Fixed	32 × 32	5.29	-
TiTok (Yu et al., 2024)	Fixed	32 × 1	2.21	72.6
SeTok	Dynamic	-	2.07	**75.4**

Rescontruction

Original

Experiments

- Visual Understanding

Method	Size	Vis. Tok.	Flickr30K	VQA^{v2}	OK-VQA	GQA	POPE	MME	MM-Vet
InstructBLIP (Liu et al., 2023a)	13B	Q.C	-	-	-	49.5	78.9	1212.8	-
Qwen-VL-Chat (Bai et al., 2023)	7B	Q.C	-	78.2*	-	57.5*	-	1487.5	-
Emu (Sun et al., 2024b)	7B	Q.C	77.4	57.2	43.4	-	-	-	-
DreamLLM (Dong et al., 2024)	7B	P.C	-	72.9	-	41.8	-	-	36.6
LLaVA-1.5 (Liu et al., 2024b)	7B	P.C	-	78.5*	-	62.0*	85.9	1510.7	33.1
NExT-GPT (Wu et al., 2024c)	7B	P.C	84.5	66.7	52.1	-	-	-	-
SEED-X (Ge et al., 2024)	17B	P.C	52.3	-	-	47.9	84.2	1435.7	-
LaVIT (Jin et al., 2024b)	7B	P.C	83.0	66.0	54.6	46.8	-	-	-
Unified-IO-2 (Lu et al., 2024)	6.8B	P.D	-	**79.4***	-	-	87.7	-	-
CM3Leon (Yu et al., 2023a)	7B	P.D	-	47.6	23.8	-	-	-	-
Chameleon (Team, 2024)	34B	P.D	74.7	66.0	-	-	-	-	-
SETOKIM	7B	SE.C	**86.9**	78.5*	**60.2***	**65.6***	**89.1**	**1537.8**	**45.2**

Experiments

- Visual Understanding

Where can I see this sign? What does it mean?

It's a traffic sign which is usually seen beside the street, instructing drivers to stop completely.

Do you see any person who is playing on the beach?

No, there is no one playing on the beach in this image. The beach is mostly empty except for a few items like a striped towel, a blue surfboard, a red bag, and a couple of chairs. There are also a few people in the distance, but they are not actively playing.

Describe the image in detail.

The image captures an outdoor seating area of a restaurant or cafe. The seating area is arranged around a central table, with multiple chairs placed around it. In the foreground, there is a large black umbrella providing shade for the seating area. To the left of the seating area, there is a sign that read s "A NEW EXPERIENCE COMING YOUR WAY". The sign is white with black text, and it is placed on a small stand. In the background, there is a building with a glass facade. The building appears to be a commercial or retail space, possibly related to the new experience mentioned in the sign ...

Experiments

- Visual Generation and Editing

Method	Type	Decoder	MS-COCO	MagicBrush		MA5K		EVR	
			FID ↓	$CLIP_{im}$ ↑	L_1 ↓	LPIPS↓	L_1 ↓	$CLIP_{im}$ ↑	L_1 ↓
• Diffusion-based Method									
Make-A-Scene (Gafni et al., 2022)	Autoregressive	-	11.8	-	-	-	-	-	-
Ins.P2P* (Brooks et al., 2023)	Diffusion	-	-	83.4	12.1	35.9	17.6	81.4	18.9
SD v2.1 (Rombach et al., 2022)	Diffusion	-	9.0	-	-	-	-	-	-
• MLLM-based Method									
DreamLLM (Dong et al., 2024)	LLM (Cont.)	SDv2.1	8.7	-	-	-	-	-	-
SEED-X (Ge et al., 2023)	LLM (Cont.)	SDXL	14.9	-	-	-	-	-	-
CM3Leon (Yu et al., 2023a)	LLM (Disc.)	VQGAN	10.3	-	-	-	-	-	-
LaVIT* (Jin et al., 2024b)	LLM (Disc.)	SDv1.5	**7.4**	81.1	25.3	36.9	25.1	73.8	26.8
LWM (Liu et al., 2024a)	LLM (Disc.)	VQGAN	12.6	-	-	-	-	-	-
MGIE* (Fu et al., 2024)	LLM (Cont.)	SDv1.5	-	**91.1**	8.2	29.8	**13.3**	81.7	16.3
Emu-2-gen* (Sun et al., 2024a)	LLM (Cont.)	SDXL	-	85.7	19.9	28.4	20.5	80.3	22.8
Morph-Token* (Pan et al., 2024)	LLM (Cont.)	VQGAN	-	87.9	7.6	27.9	14.6	82.6	15.3
SETOKIM	LLM (Cont.)	SeTok	8.5	89.6	**6.9**	**26.4**	15.7	**83.5**	**14.1**

Experiments

- Visual Generation and Editing

Method	Type	Decoder	MS-COCO	MagicBrush		MA5K		EVR	
			FID ↓	$CLIP_{im}$ ↑	L_1 ↓	LPIPS↓	L_1 ↓	$CLIP_{im}$ ↑	L_1 ↓
• Diffusion-based Method									
Make-A-Scene (Gafni et al., 2022)	Autoregressive	-	11.8	-	-	-	-	-	-
Ins.P2P* (Brooks et al., 2023)	Diffusion	-	-	83.4	12.1	35.9	17.6	81.4	18.9
SD v2.1 (Rombach et al., 2022)	Diffusion	-	9.0	-	-	-	-	-	-
• MLLM-based Method									
DreamLLM (Dong et al., 2024)	LLM (Cont.)	SDv2.1	8.7	-	-	-	-	-	-
SEED-X (Ge et al., 2023)	LLM (Cont.)	SDXL	14.9	-	-	-	-	-	-
CM3Leon (Yu et al., 2023a)	LLM (Disc.)	VQGAN	10.3	-	-	-	-	-	-
LaVIT* (Jin et al., 2024b)	LLM (Disc.)	SDv1.5	**7.4**	81.1	25.3	36.9	25.1	73.8	26.8
LWM (Liu et al., 2024a)	LLM (Disc.)	VQGAN	12.6	-	-	-	-	-	-
MGIE* (Fu et al., 2024)	LLM (Cont.)	SDv1.5	-	**91.1**	8.2	29.8	**13.3**	81.7	16.3
Emu-2-gen* (Sun et al., 2024a)	LLM (Cont.)	SDXL	-	85.7	19.9	28.4	20.5	80.3	22.8
Morph-Token* (Pan et al., 2024)	LLM (Cont.)	VQGAN	-	87.9	7.6	27.9	14.6	82.6	15.3
SETOKIM	LLM (Cont.)	SeTok	8.5	89.6	**6.9**	**26.4**	15.7	**83.5**	**14.1**

Experiments

- Visual Generation and Editing

I want to open my own coffee shop, which should have a very comfortable and warm environment. Can you help me design it? Show me a picture of it.

Sure, I can show you a picture of a cozy coffee shop. Here is a picture:

show me a scene of a blue squirrel holding a flower, perched on a fence beside a cluster of purple flowers!

Here is the scene.

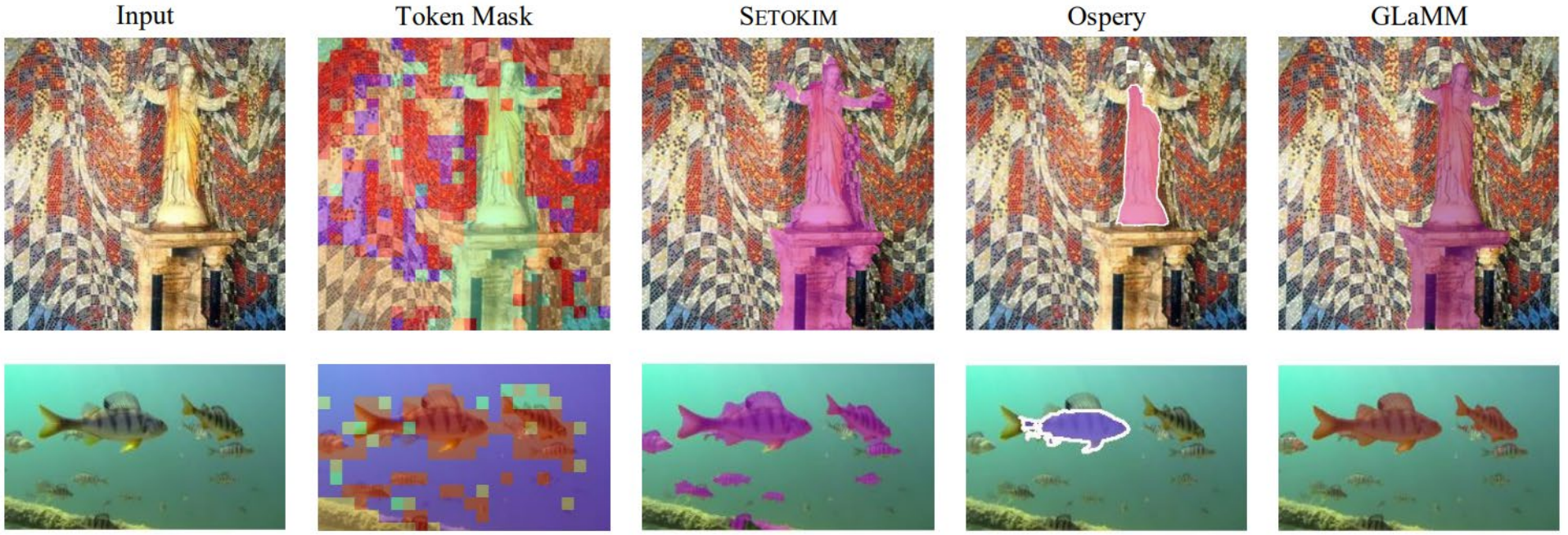
Can you change it into a cartoon style?

Sure, here are the cartoon image.

Experiments

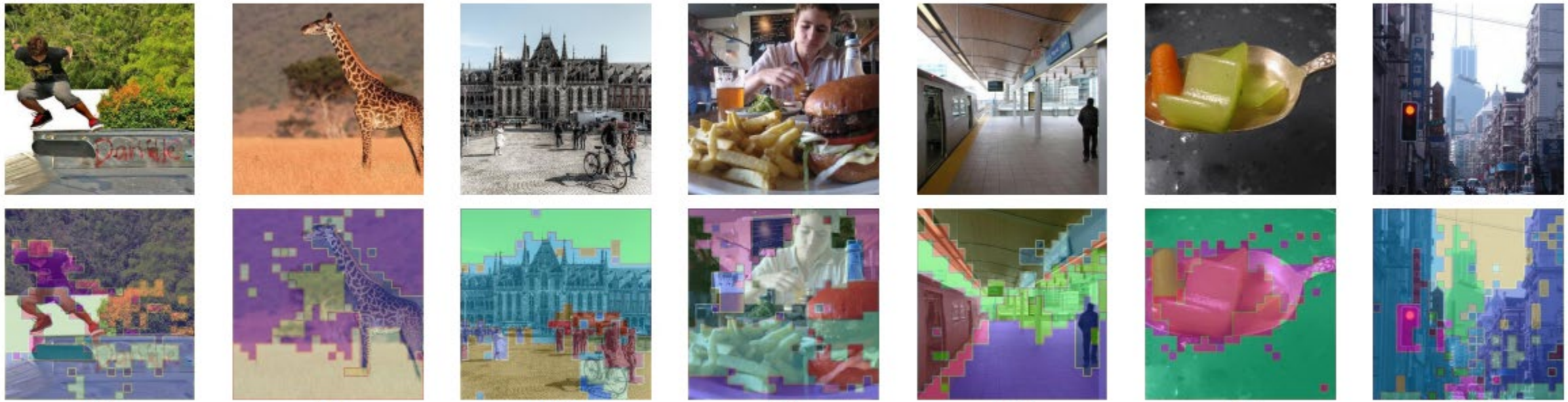
- Referring Expression Segmentation

Method	refCOCOg		refCOCO+			Reaseg	
	val(U)	test(U)	val	testA	testB	gIoU	cIoU
ReLA	65.0	66.0	66.0	71.0	57.7	-	-
SEEM	65.7	-	-	-	-	24.3	18.7
PixelLM	69.3	70.5	66.3	71.7	58.3	-	-
NExT-Chat	67.0	67.0	65.1	71.9	56.7	-	-
LISA	67.9	70.6	65.1	70.8	58.1	47.3	48.4
SETOKIM	**71.3**	**71.3**	**68.0**	**72.4**	**61.2**	**50.7**	**52.7**



Experiments

- Qualitative Analysis of Visual Tokens



(1) (2) (3) (4) (5) (6) (7)

Take-away

- Introduce SeTok, a viable **semantic-equivalent tokenizer**, that enables to tokenize automatically patch-level visual features into a variable number of semantic-complete concept visual tokens
- Integrate SeTok with a pre-trained LLM to build an MLLM, **SETOKIM**
- Extensive experiments demonstrate that **SETOKIM** performs better on a broad range of comprehension, generation, segmentation, and editing tasks