Image Understanding Makes for A Good Tokenizer for Image Generation

Luting Wang Yang Zhao[1] Zijian Zhang[1] Jiashi Feng[1] Si Liu[*] Bingyi Kang[1*]

[1]ByteDance [*]Corresponding Authors

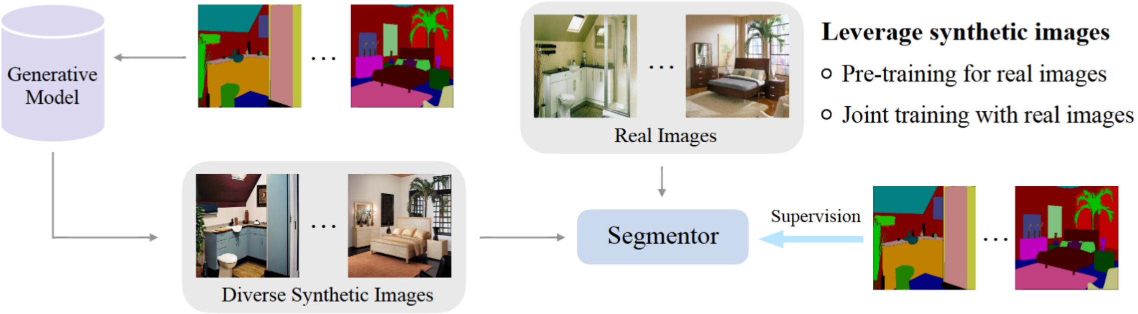
Friday, December 13

Table of Contents

Image Generation Benefits Image Understanding

Studies have shown that IG models can benefit IU tasks in various ways.

1. **Data augmentation** through synthetic data generation

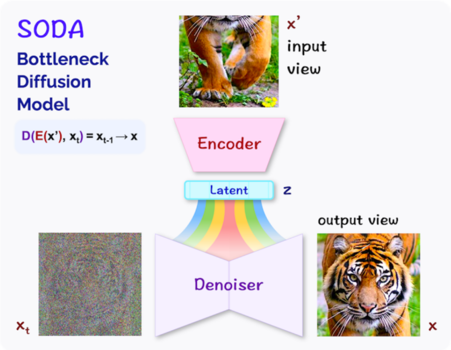


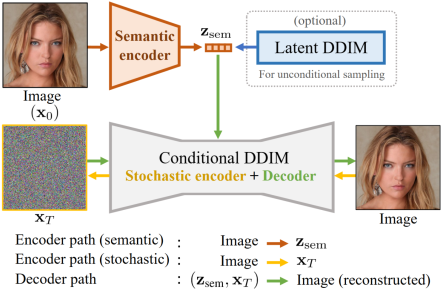
Lihe Yang, *et al.* "FreeMask: Synthetic Images with Dense Annotations Make Stronger Segmentation Models." NeurIPS, 2023.

Image Generation Benefits Image Understanding

Studies have shown that IG models can benefit IU tasks in various ways.

2. Improved **representation learning**





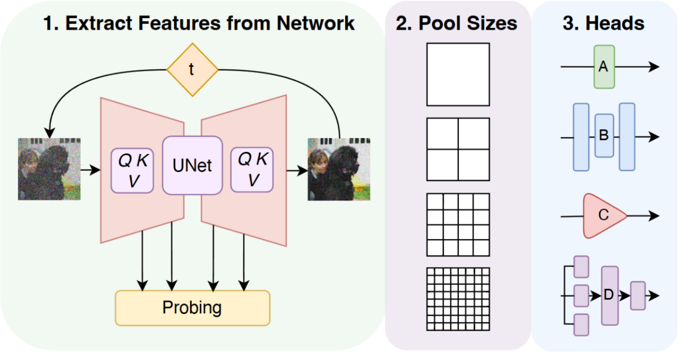
Drew A. Hudson, *et al.* "SODA: Bottleneck Diffusion Models for Representation Learning." arXiv:2311.17901, 2023.

Konpat Preechakul, *et al.* "Diffusion Autoencoders: Toward a Meaningful and Decodable Representation." CVPR, 2022.

Image Generation Benefits Image Understanding

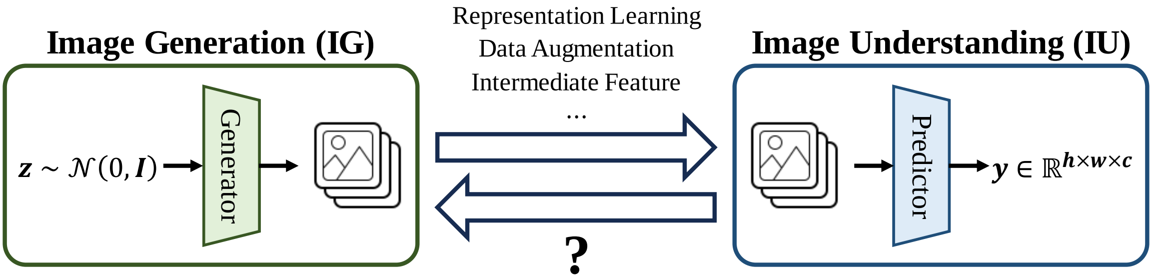
Studies have shown that IG models can benefit IU tasks in various ways.

3. Utilizing **intermediate features** from IG models for solving perception tasks



Soumik Mukhopadhyay, *et al.* "Diffusion Models Beat GANs on Image Classification." arXiv:2307.08702.

The Reciprocal Question?

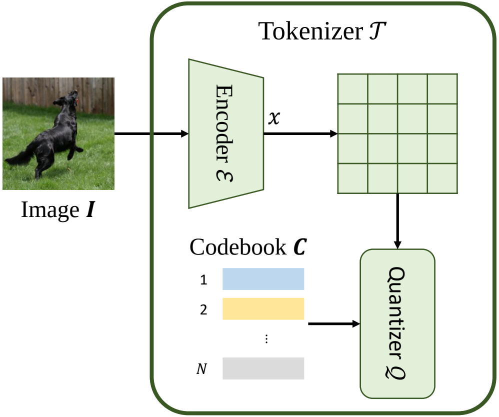


The reciprocal question remains largely uncharted:

How might IU models aid IG tasks?

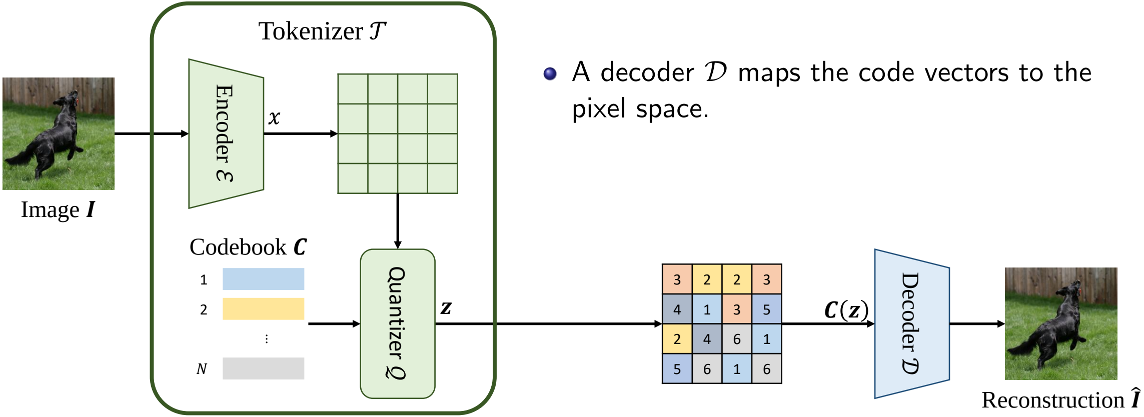
Table of Contents

Two-Stage Image Generation



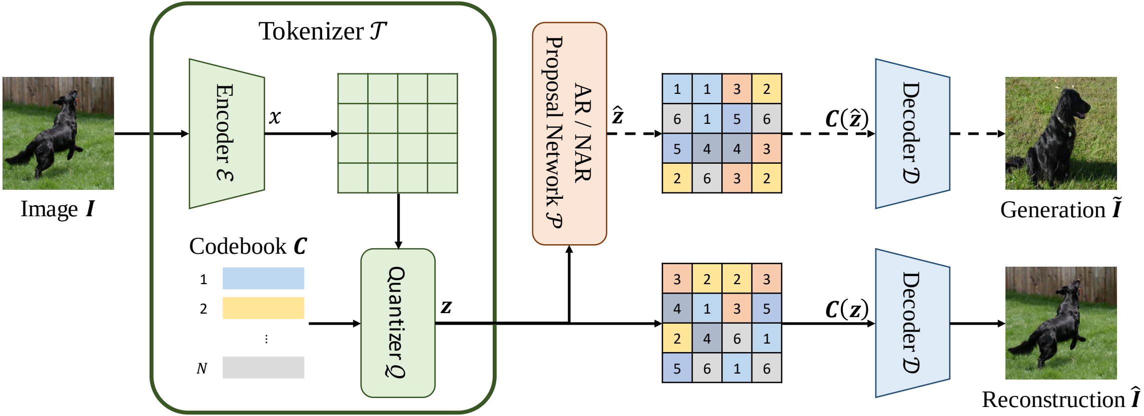
- The encoder $\mathcal{E}$, quantizer $\mathcal{Q}$, and codebook $\mathbf{C}$ collectively form an **image tokenizer** $\mathcal{T}$.
- Given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the **encoder** $\mathcal{E}$ converts this image into a feature map $x \in \mathbb{R}^{h \times w \times d}$.
- **Codebook C** is a set of N code vectors $\{c_i\}_{i=1}^{N} \in \mathbb{R}^{N \times d}$, where each code vector $c_i \in \mathbb{R}^d$ corresponds to a specific code i.
- **Quantizer** $\mathcal{Q}$ then maps x into a sequence of codes $\mathbf{z} = \{z_i\}_{i=1}^{L}$.

Two-Stage Image Generation



- A decoder $\mathcal{D}$ maps the code vectors to the pixel space.

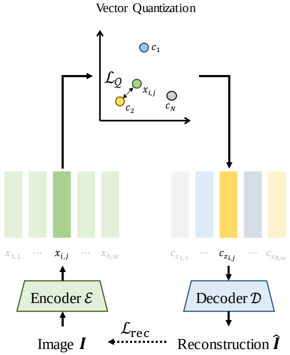
Two-Stage Image Generation



- The **proposal network** $\mathcal{P}$ models the distribution over $\mathbf{z}$, denoted as $p(\mathbf{z})$.

Image Tokenizers

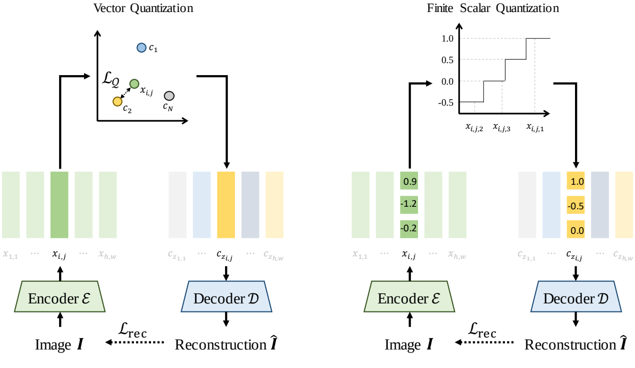
Image Tokenizers



(a) VQGAN

Patrick Esser, Robin Rombach, *et al.* “Taming Transformers for High-Resolution Image Synthesis.” CVPR, 2021

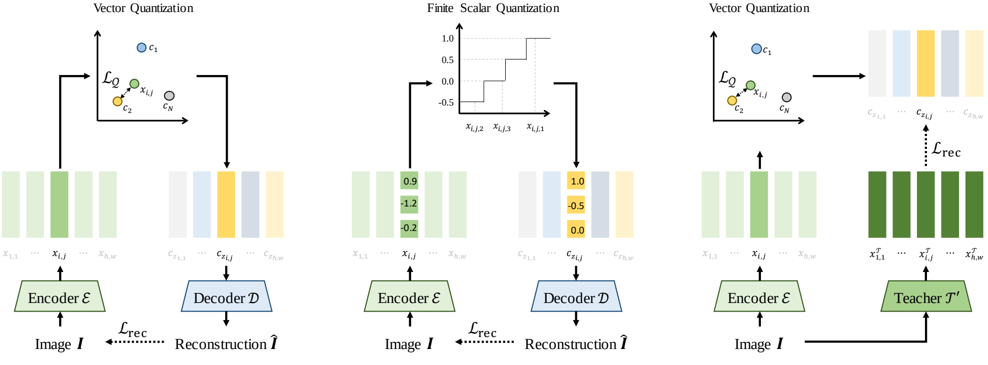
Image Tokenizers



(a) VQGAN (b) FSQ

Fabian Mentzer, David Minnen, *et al.* "Finite Scalar Quantization: VQ-VAE Made Simple." ICLR, 2024.

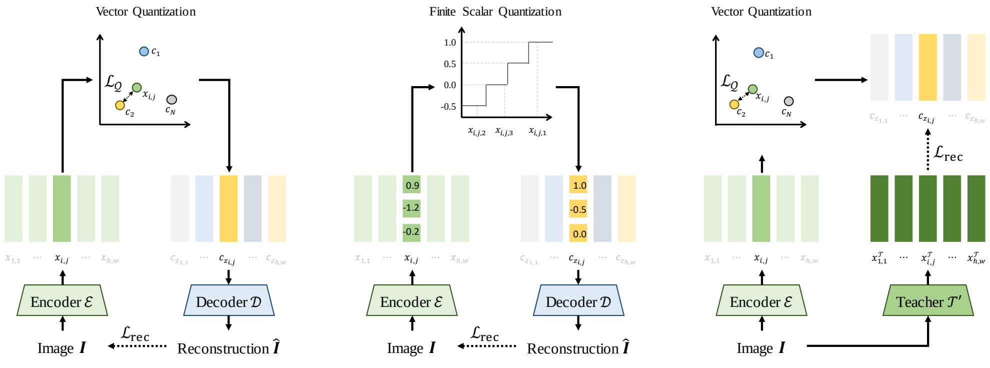
Image Tokenizers



(a) VQGAN (b) FSQ (c) VQ-KD

Zhiliang Peng, Li Dong, *et al.* "BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers." arXiv preprint arXiv:2208.06366, 2022.

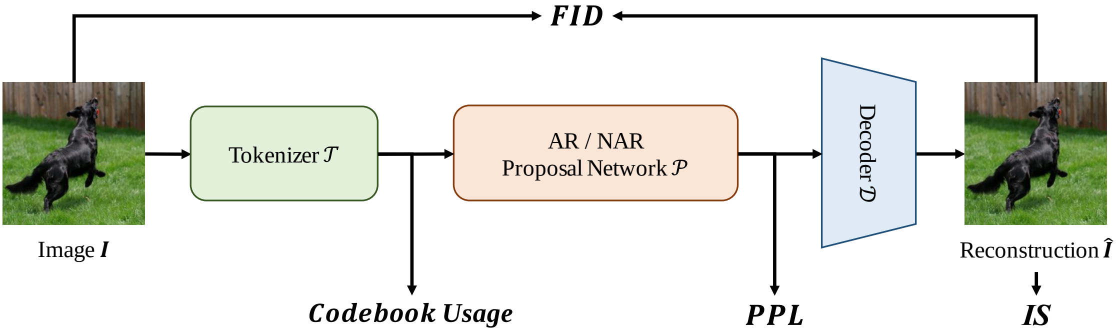
Image Tokenizers



(a) VQGAN (b) FSQ (c) VQ-KD

Pixel Reconstruction Feature Reconstruction

Benchmark



We build the above benchmark to evaluate the IG performance of tokenizers.

- For each tokenizer, we train a proposal network and a decoder to form an image generator.
- Various metrics are adopted for a comprehensive evaluation.

Table of Contents

Main Observation

1. VQ-KD significantly enhances generation quality over VQGAN.

Tokenizer	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{NAR} ↓
VQGAN	4.9	5.09	116.75	**24.11**	**20.03**
FSQ	100.0	4.96	791.56	40.17	29.78
$\text{VQ-KD}_{\text{CLIP}}$	100.0	4.96	53.73	11.78	9.51
$\text{VQ-KD}_{\text{ViT}}$	100.0	3.69	89.30	**11.40**	**8.45**
$\text{VQ-KD}_{\text{DINO}}$	100.0	3.41	74.07	13.15	10.21
$\text{VQ-KD}_{\text{MAE}}$	100.0	4.93	280.06	26.85	16.11

Main Observation

1. VQ-KD significantly enhances generation quality over VQGAN.
2. The superiority of VQ-KD is irrelevant to the quantization operation and codebook usage.

Tokenizer	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{NAR} ↓
VQGAN	4.9	5.09	116.75	24.11	20.03
FSQ	**100.0**	**4.96**	791.56	**40.17**	29.78
$\text{VQ-KD}_{\text{CLIP}}$	**100.0**	**4.96**	53.73	**11.78**	9.51
$\text{VQ-KD}_{\text{ViT}}$	100.0	3.69	89.30	11.40	8.45
$\text{VQ-KD}_{\text{DINO}}$	100.0	3.41	74.07	13.15	10.21
$\text{VQ-KD}_{\text{MAE}}$	100.0	4.93	280.06	26.85	16.11

Main Observation

1. VQ-KD significantly enhances generation quality over VQGAN.
2. The superiority of VQ-KD is irrelevant to the quantization operation and codebook usage.

Tokenizer	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{NAR} ↓
VQGAN	4.9	5.09	**116.75**	24.11	20.03
FSQ	100.0	4.96	791.56	40.17	29.78
$\text{VQ-KD}_{\text{CLIP}}$	100.0	4.96	**53.73**	11.78	9.51
$\text{VQ-KD}_{\text{ViT}}$	100.0	3.69	**89.30**	11.40	8.45
$\text{VQ-KD}_{\text{DINO}}$	100.0	3.41	**74.07**	13.15	10.21
$\text{VQ-KD}_{\text{MAE}}$	100.0	4.93	280.06	26.85	16.11

Main Observation

1. VQ-KD significantly enhances generation quality over VQGAN.
2. The superiority of VQ-KD is irrelevant to the quantization operation and codebook usage.

Tokenizer	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{NAR} ↓
VQGAN	4.9	5.09	**116.75**	24.11	20.03
FSQ	100.0	4.96	**791.56**	40.17	29.78
$\text{VQ-KD}_{\text{CLIP}}$	100.0	4.96	53.73	11.78	9.51
$\text{VQ-KD}_{\text{ViT}}$	100.0	3.69	89.30	11.40	8.45
$\text{VQ-KD}_{\text{DINO}}$	100.0	3.41	74.07	13.15	10.21
$\text{VQ-KD}_{\text{MAE}}$	100.0	4.93	280.06	26.85	16.11

Main Observation

1. VQ-KD significantly enhances generation quality over VQGAN.
2. The superiority of VQ-KD is irrelevant to the quantization operation and codebook usage.
3. Tokenizers with stronger semantic understanding tend to deliver superior IG performance.

Tokenizer	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{NAR} ↓
VQGAN	4.9	5.09	116.75	24.11	20.03
FSQ	100.0	4.96	791.56	40.17	29.78
$\text{VQ-KD}_{\text{CLIP}}$	100.0	**4.96**	53.73	**11.78**	**9.51**
$\text{VQ-KD}_{\text{ViT}}$	100.0	**3.69**	89.30	**11.40**	**8.45**
$\text{VQ-KD}_{\text{DINO}}$	100.0	**3.41**	74.07	**13.15**	**10.21**
$\text{VQ-KD}_{\text{MAE}}$	100.0	4.93	280.06	26.85	16.11

Table of Contents

Further Verification

1. The superiority of VQ-KD holds across proposal networks.

Tokenizer	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{NAR} ↓
VQGAN	4.9	5.09	116.75	24.11	**20.03**
FSQ	100.0	4.96	791.56	40.17	**29.78**
$\text{VQ-KD}_{\text{CLIP}}$	100.0	4.96	53.73	11.78	**9.51**
$\text{VQ-KD}_{\text{ViT}}$	100.0	3.69	89.30	11.40	**8.45**
$\text{VQ-KD}_{\text{DINO}}$	100.0	3.41	74.07	13.15	**10.21**
$\text{VQ-KD}_{\text{MAE}}$	100.0	4.93	280.06	26.85	**16.11**

Further Verification

1. The superiority of VQ-KD holds across proposal networks.

Tokenizer	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{NAR} ↓
VQGAN	4.9	5.09	116.75	24.11	20.03
FSQ	100.0	4.96	791.56	40.17	29.78
$\text{VQ-KD}_{\text{CLIP}}$	100.0	4.96	53.73	11.78	**9.51**
$\text{VQ-KD}_{\text{ViT}}$	100.0	3.69	89.30	11.40	**8.45**
$\text{VQ-KD}_{\text{DINO}}$	100.0	3.41	74.07	13.15	10.21
$\text{VQ-KD}_{\text{MAE}}$	100.0	4.93	280.06	26.85	16.11

Further Verification

1. The superiority of VQ-KD holds across proposal networks.
2. The superiority of VQ-KD holds across datasets.

Tokenizer $\mathcal{T}$	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{T2I} ↓
VQGAN	2.4	16.21	47.89	**38.43**	24.11
FSQ	100.0	4.62	1040.02	**44.64**	23.36
$\text{VQ-KD}_{\text{CLIP}}$	82.2	5.48	72.31	29.80	11.17
$\text{VQ-KD}_{\text{ViT}}$	100.0	3.70	117.10	23.51	15.49
$\text{VQ-KD}_{\text{DINO}}$	100.0	2.69	129.93	**17.55**	11.50
$\text{VQ-KD}_{\text{MAE}}$	100.0	3.51	317.98	44.01	15.60

Further Verification

1. The superiority of VQ-KD holds across proposal networks.
2. The superiority of VQ-KD holds across datasets.

Tokenizer $\mathcal{T}$	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{T2I} ↓
VQGAN	**2.4**	16.21	47.89	38.43	24.11
FSQ	100.0	4.62	1040.02	44.64	23.36
$\text{VQ-KD}_{\text{CLIP}}$	82.2	5.48	72.31	29.80	11.17
$\text{VQ-KD}_{\text{ViT}}$	100.0	**3.70**	117.10	**23.51**	15.49
$\text{VQ-KD}_{\text{DINO}}$	100.0	**2.69**	129.93	**17.55**	11.50
$\text{VQ-KD}_{\text{MAE}}$	100.0	3.51	317.98	44.01	15.60

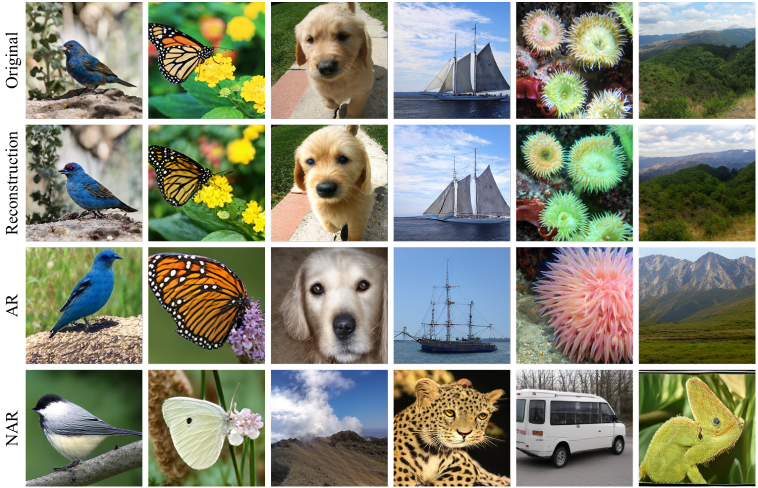
Further Verification

1. The superiority of VQ-KD holds across proposal networks.
2. The superiority of VQ-KD holds across datasets.
3. The superiority of VQ-KD holds across tasks.

Tokenizer $\mathcal{T}$	Codebook Usage (%)	rFID ↓	PPL ↓	FID_{AR} ↓	FID_{T2I} ↓
VQGAN	2.4	16.21	47.89	38.43	**24.11**
FSQ	100.0	4.62	1040.02	44.64	**23.36**
$\text{VQ-KD}_{\text{CLIP}}$	82.2	5.48	72.31	29.80	**11.17**
$\text{VQ-KD}_{\text{ViT}}$	100.0	3.70	117.10	23.51	**15.49**
$\text{VQ-KD}_{\text{DINO}}$	100.0	2.69	129.93	17.55	**11.50**
$\text{VQ-KD}_{\text{MAE}}$	100.0	3.51	317.98	44.01	**15.60**

Table of Contents

Results

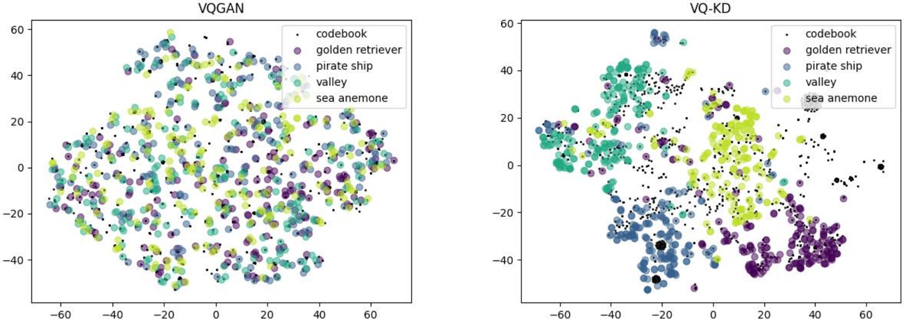


VQ-KD visualization of

- the original images,
- The reconstructed images,
- The AR generation,
- The NAR generation.

Codebook

- Compared to VQGAN, the organized feature space of VQ-KD improves the **clarity of code semantics** and helps to **better understand** image content and code interaction.



Codebook visualization of VQGAN and VQ-KD$_{\text{ViT}}$.

Contact Us

arxiv.org/abs/2411.04406

@ lutingwang.ai@qq.com

https://github.com/magic-research/vector_quantization

ByteDance