Flow to the Mode: Mode-Seeking Diffusion Autoencoders for State-of-the-Art Image Tokenization

arXiv 2025

Kyle Sargent, Kyle Hsu, Justin Johnson,

Li Fei-Fei, Jiajun Wu

Outline

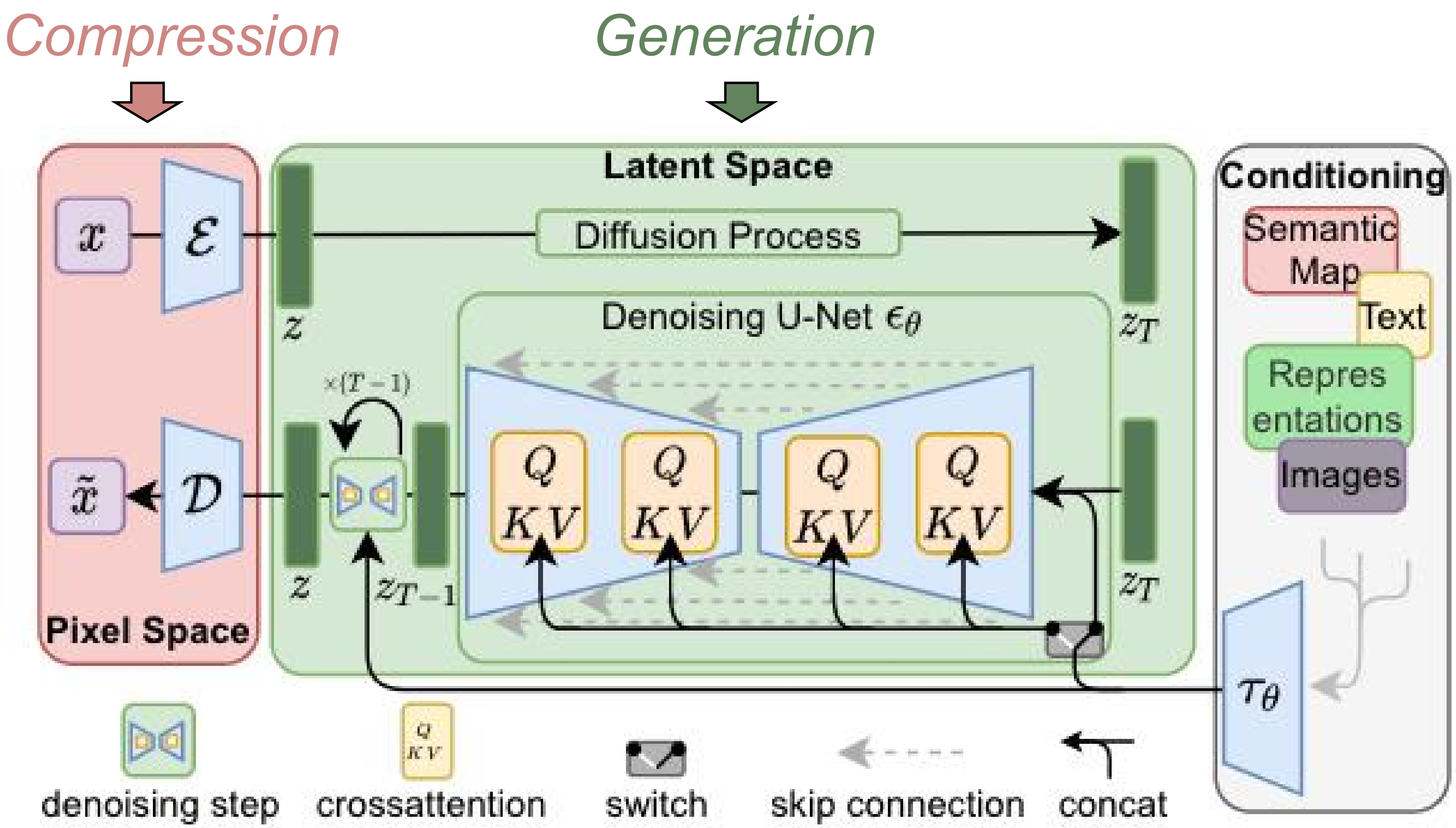
- Authorship
- Background
- Architecture
- Experiments

Two-stage Image Generation

- Tokenization / Compression
  - Image ➔ Low dimensional latent ➔ Image
- Generation
  - Learn the distribution of the latent

Latent can be continuous (*e.g.* LDM) or discrete (*e.g.* VQGAN).

Latent Diffusion Models

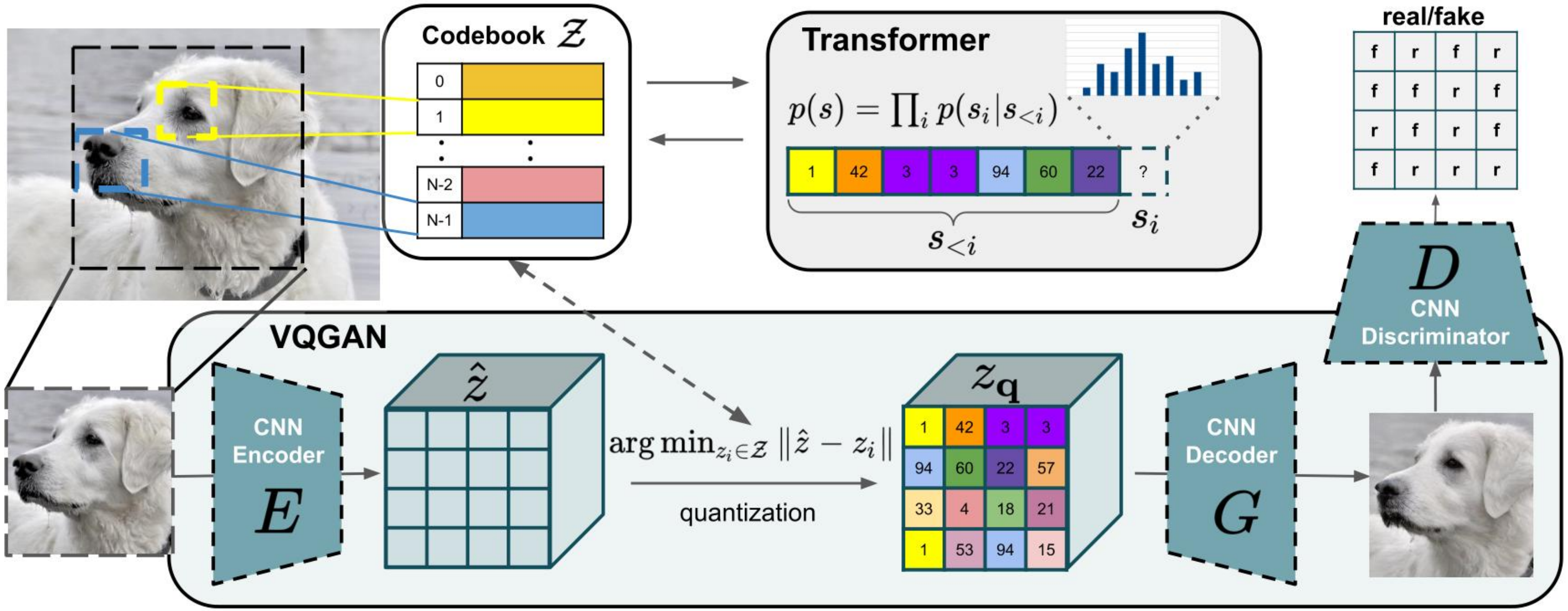


High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022

VQGAN

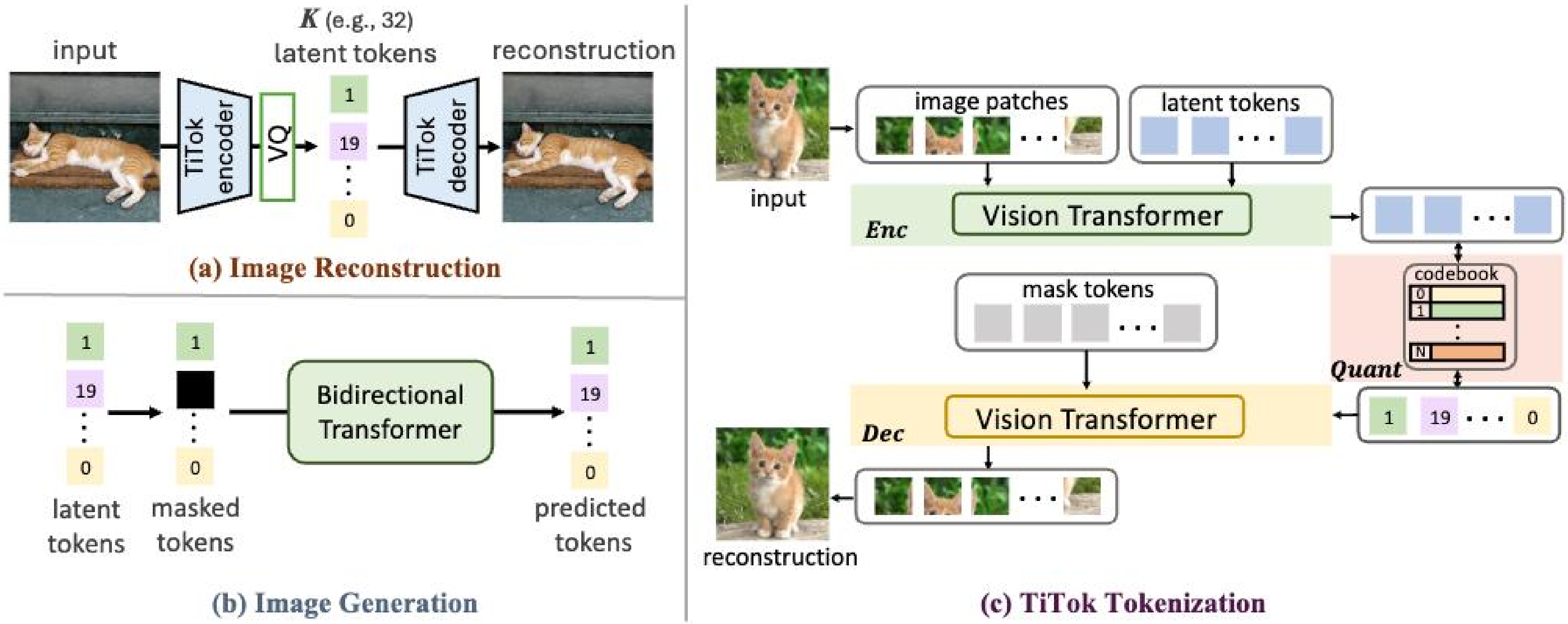
Generation

Compression



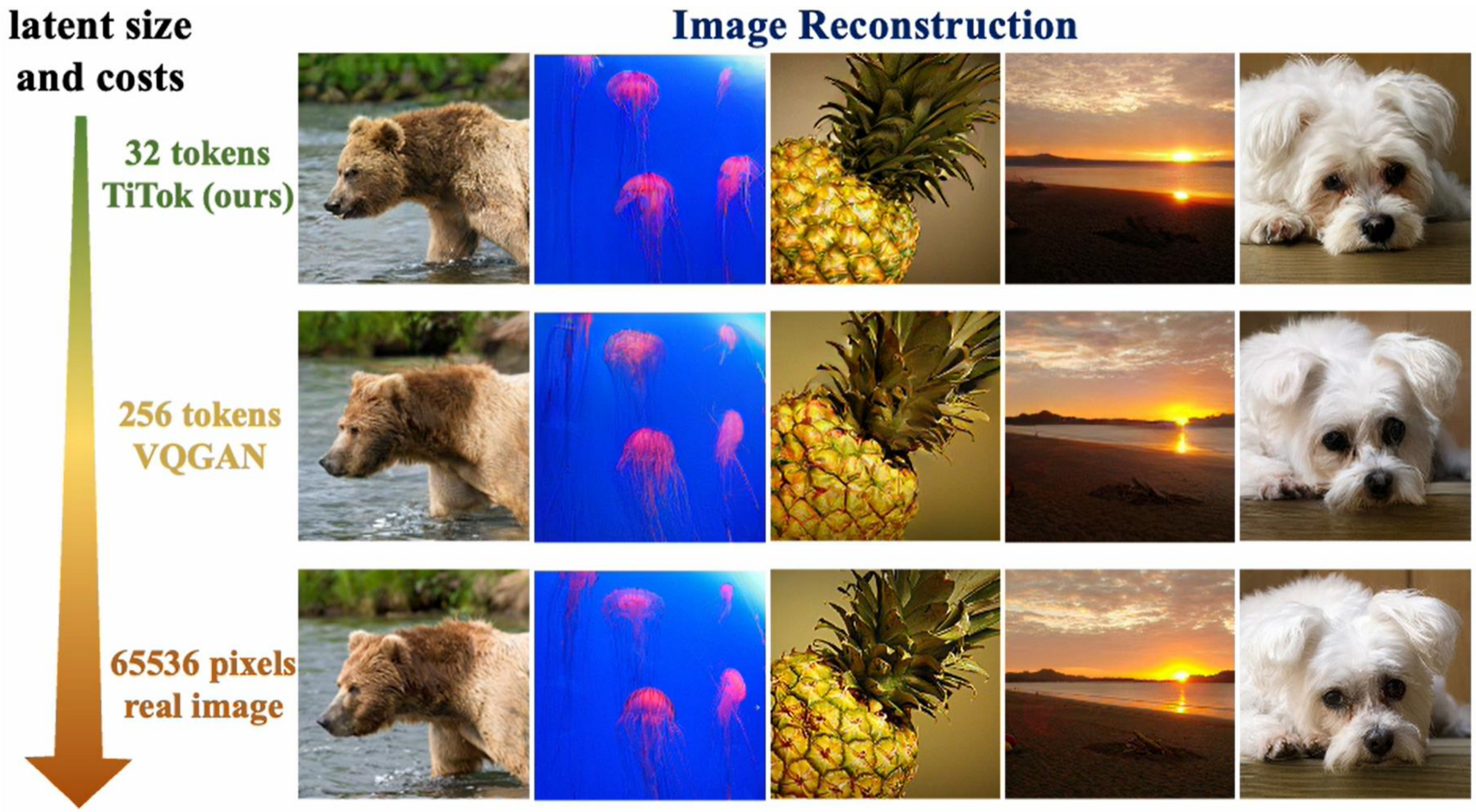
Taming Transformers for High-Resolution Image Synthesis, CVPR 2022

TiTok (1 Image = 32 Tokens)



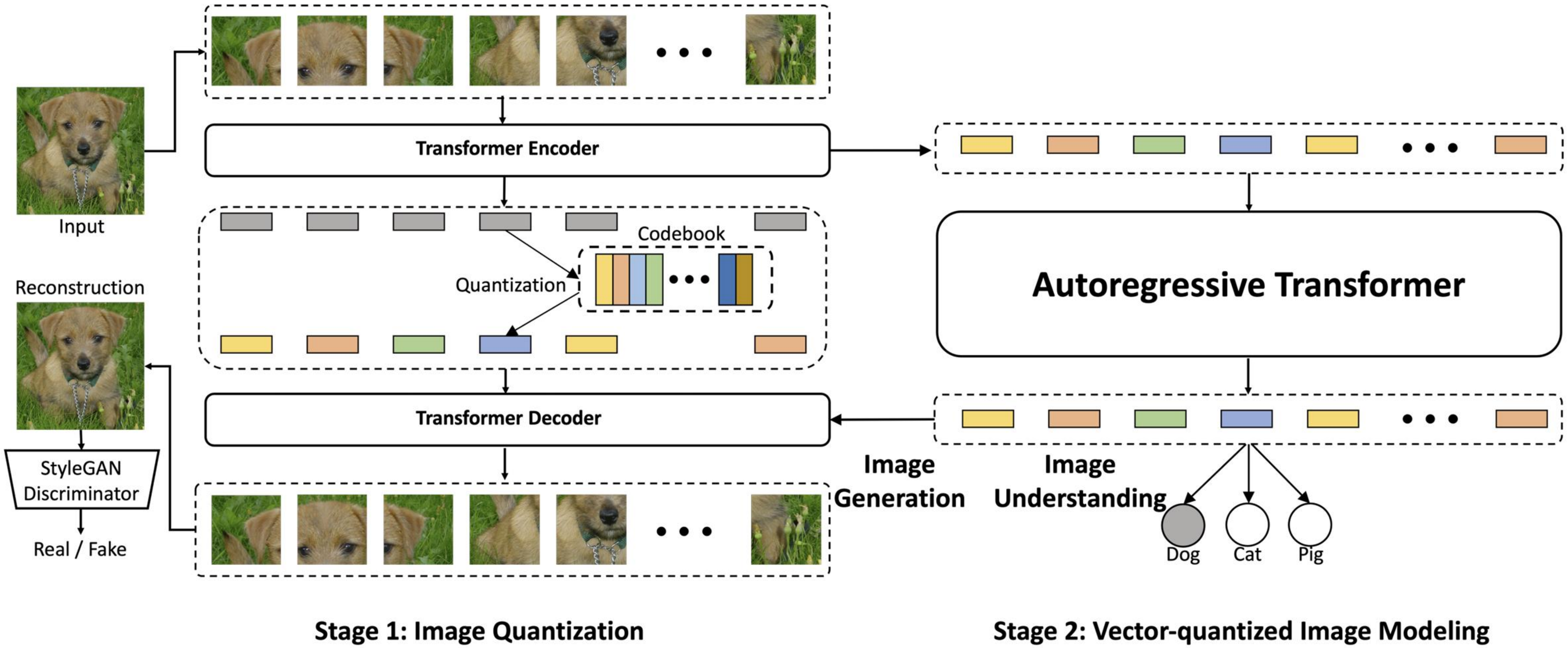
An Image is Worth 32 Tokens for Reconstruction and Generation, NeurIPS 2024

TiTok (1 Image = 32 Tokens)



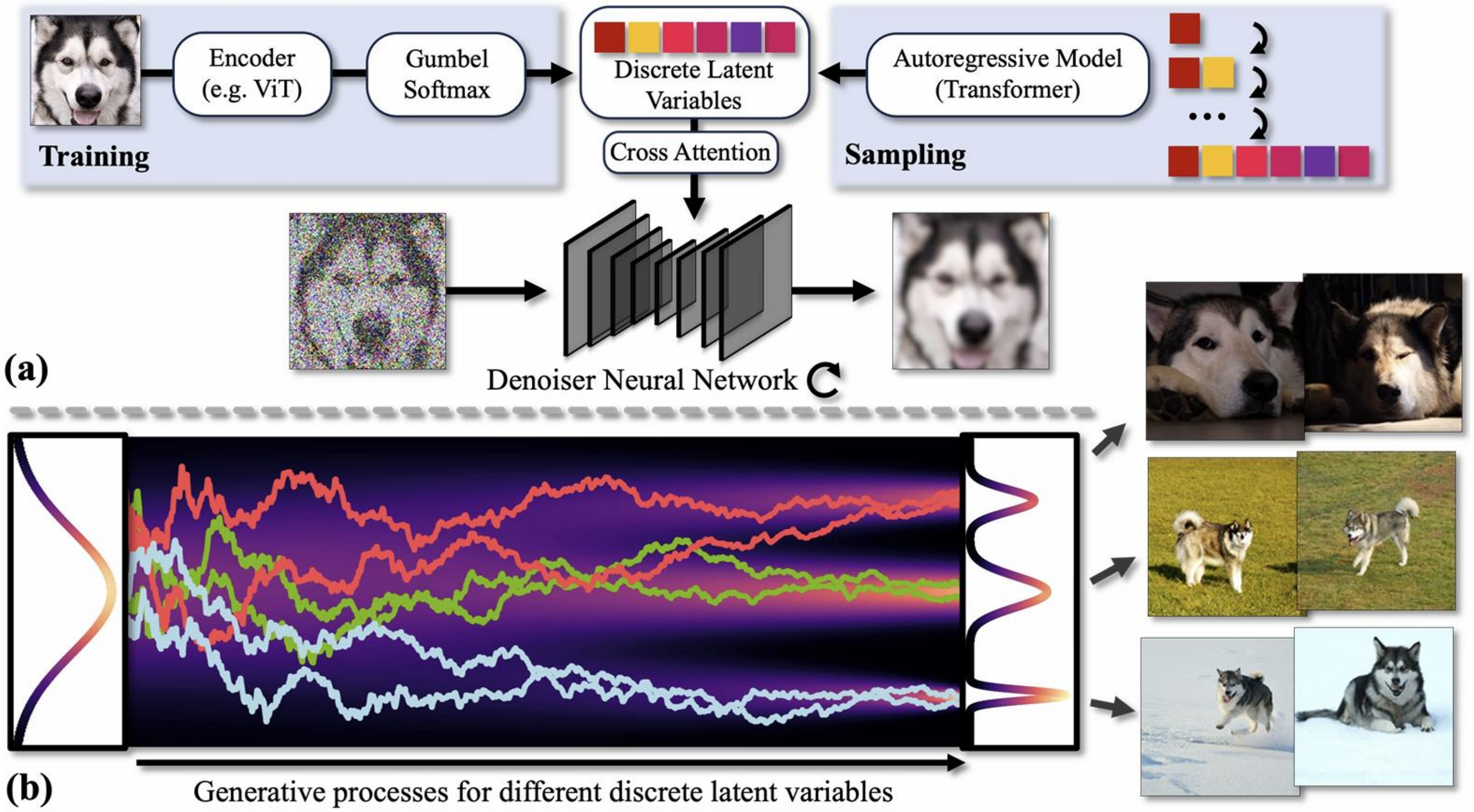
An Image is Worth 32 Tokens for Reconstruction and Generation, NeurIPS 2024

ViT-VQGAN



Vector-Quantized Image Modeling With Improved VQGAN, ICLR 2022

DisCo-Diff



DisCo-Diff: Enhancing Continuous Diffusion Models with Discrete Latents, ICLR 2024

Flow to the mode

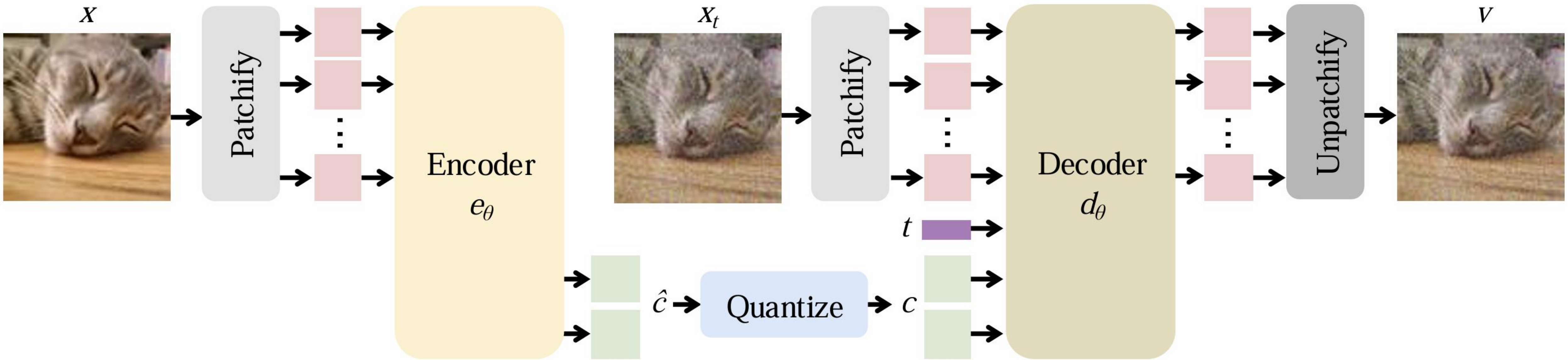
Highlights of this work

- 1D tokenizer with MMDiT
- w/o CNN, w/o any distillation
- w/o adversarial loss
- SOTA tokenizer bpp

Outline

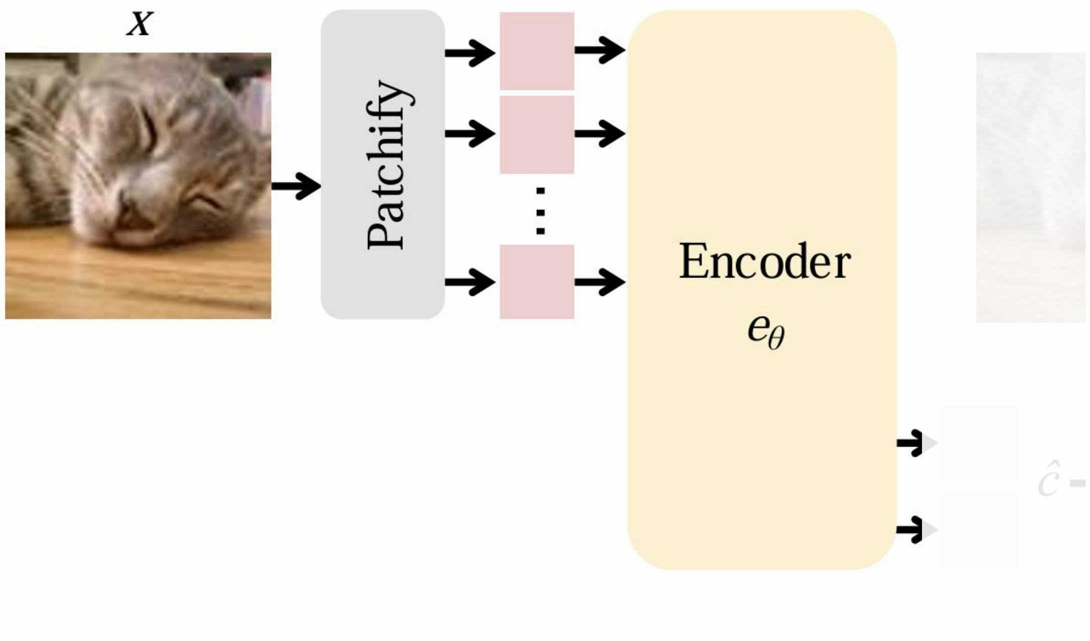
- Authorship
- Background
- Architecture
- Experiments

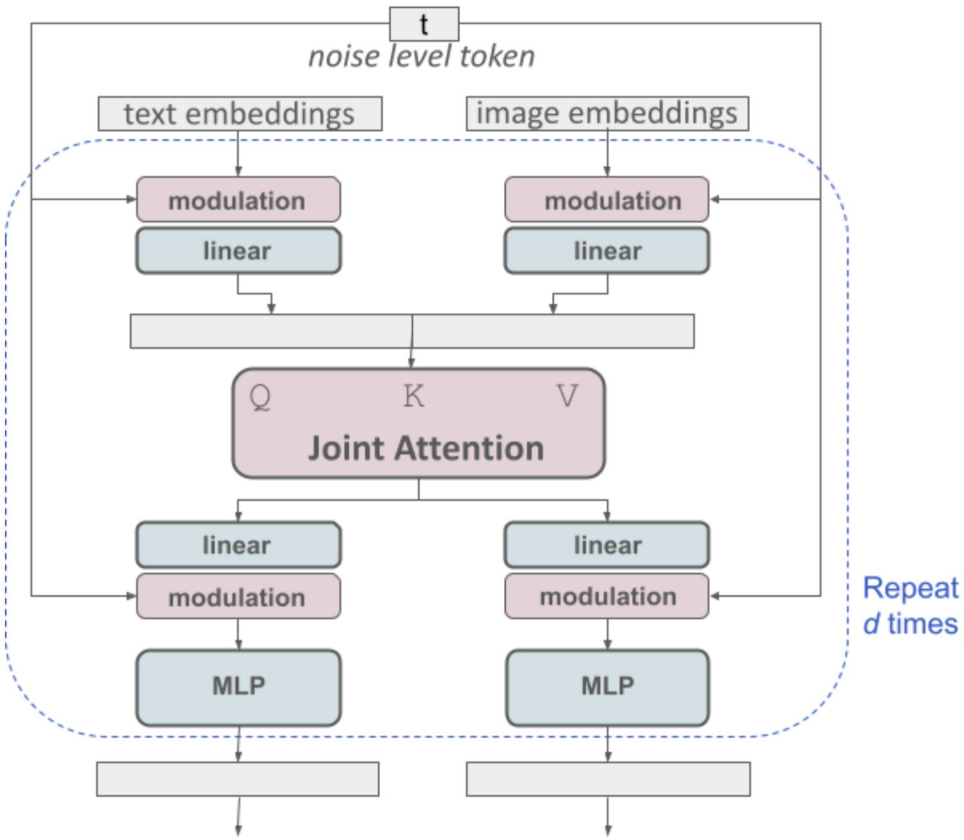
Architecture



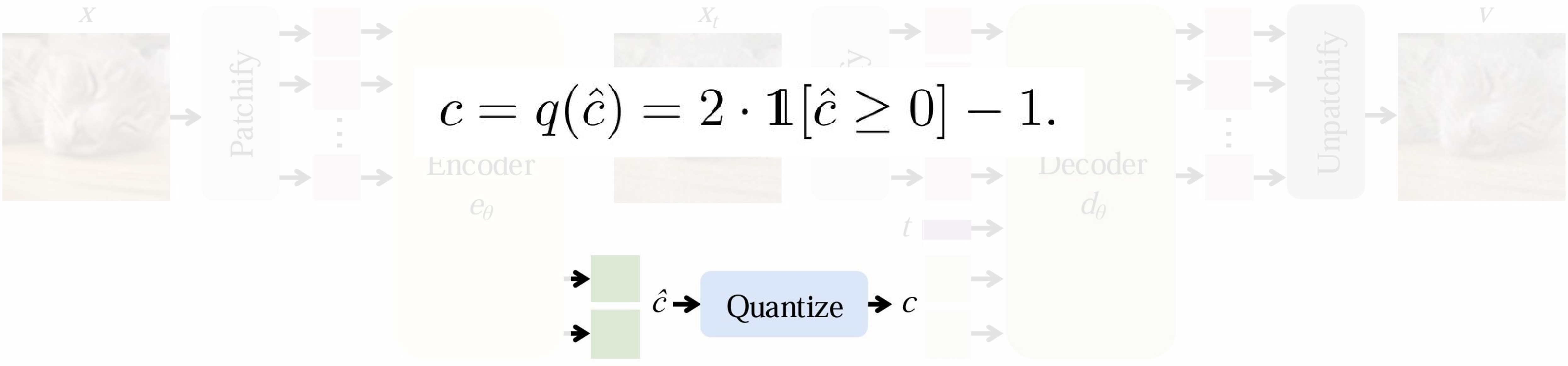
Architecture

```
_, code, aux = self.encoder(img, img_idxs, txt, txt_idxs, timesteps=None)
```





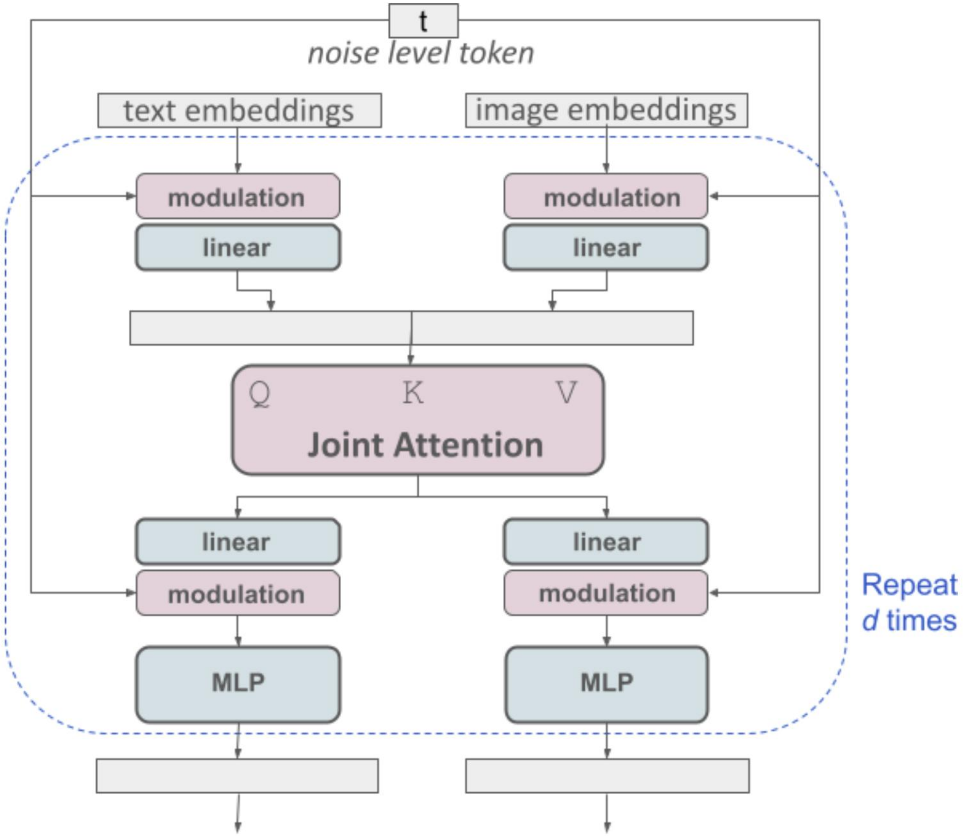
Architecture

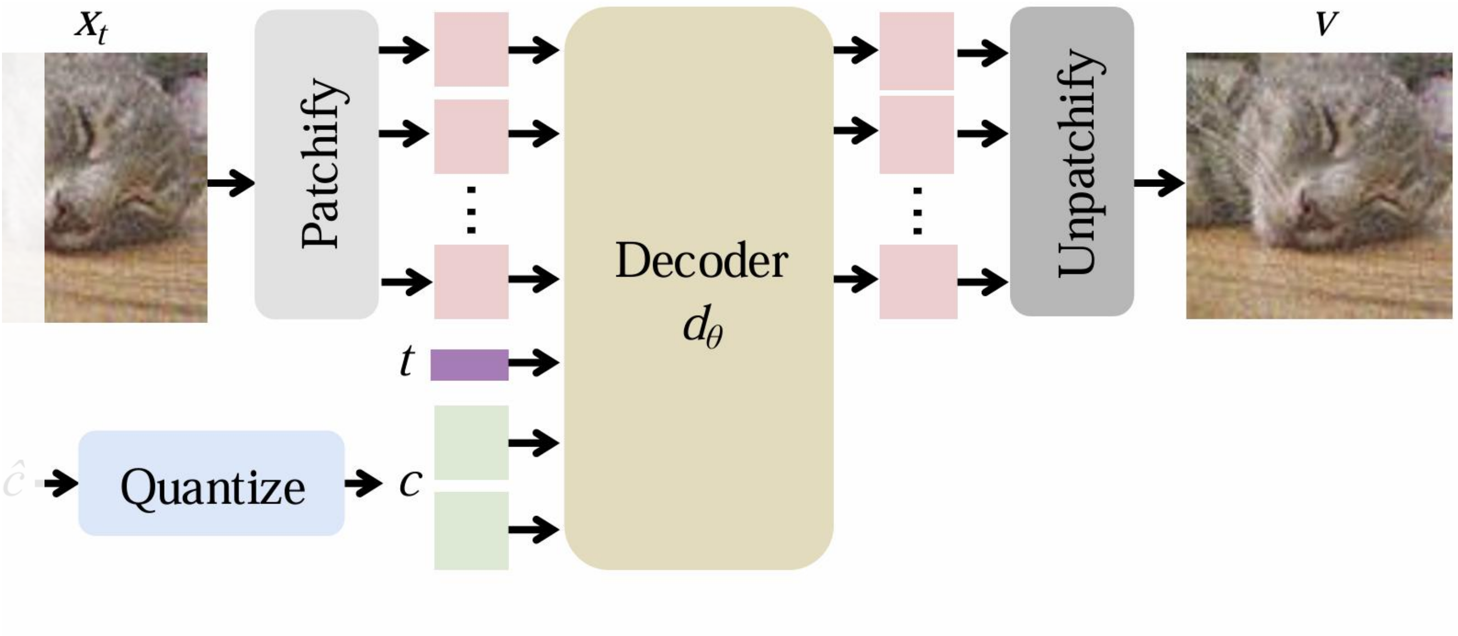


$$c = q(\hat{c}) = 2 \cdot \mathbb{1}[\hat{c} \geq 0] - 1.$$

Architecture

```
pred, _, decode_aux = self.decoder(
    img, img_idxs, code, txt_idxs, timesteps=timesteps
)
```





Stage 1A Pre-Training

Flow matching loss

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\left\|x - z - d_\theta(x_t, q(e_\theta(x)), t)\right\|_2^2\right].$$

Perceptual loss

$$\mathcal{L}_{\text{perc}} = \mathbb{E}\left[d_{\text{perc}}(x, x_t + t d_\theta(x_t, q(e_\theta(x)), t))\right].$$

Quantization loss

$$\mathcal{L}_{\text{ent}} = \mathbb{E}\left[H(q(\hat{c})\right] - H(\mathbb{E}\left[q(\hat{c})\right]),$$

$$\mathcal{L}_{\text{commit}} = \mathbb{E}\left[\left\|\hat{c} - q(\hat{c})\right\|_2^2\right].$$

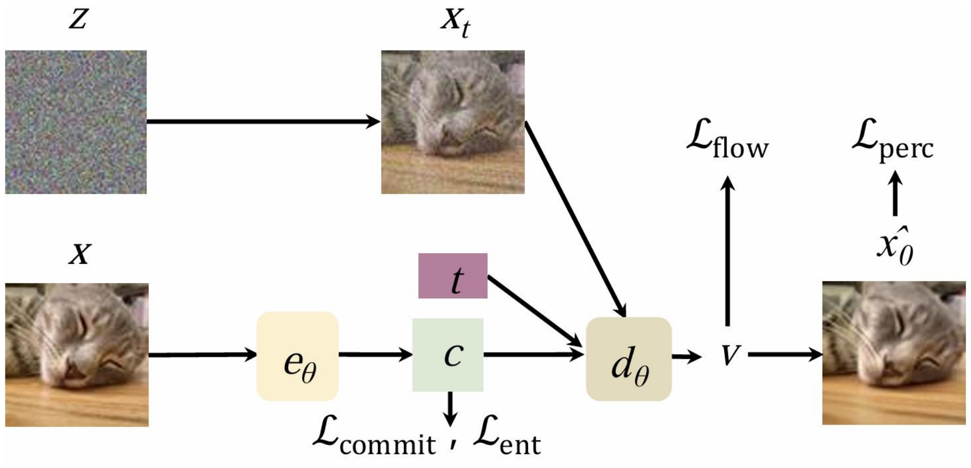


Figure 4. **Stage 1A**. The encoder and decoder are trained end-to-end with output losses $\mathcal{L}_{\text{perc}}$, $\mathcal{L}_{\text{flow}}$ and latent losses $\mathcal{L}_{\text{commit}}$, $\mathcal{L}_{\text{ent}}$.

Stage 1B Post-Training

Choose random timestep set

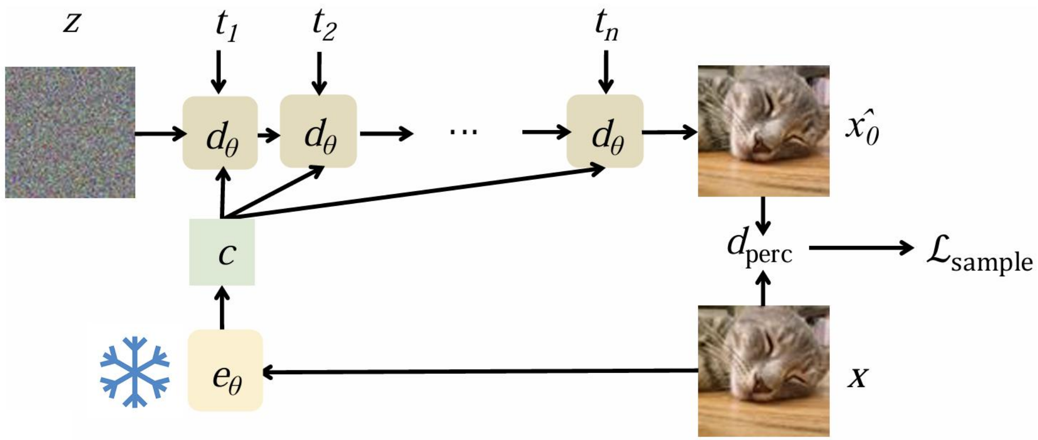
$$t_1, \ldots, t_n$$

Predict

$$d_{t_i}(x_t) = x_t + (t_{i+1} - t_i) d_\theta(x_t, c, t_i).$$

Perceptual loss (and pre-defined flow loss)

$$\mathcal{L}_{\text{sample}} = \mathbb{E}\left[d_{\text{perc}}\left(x, d_{t_n} \circ d_{t_{n-1}} \circ \cdots \circ d_{t_1}(z)\right)\right]$$



Stage 2 Generation

Use generative module in MaskGiT / TiTok

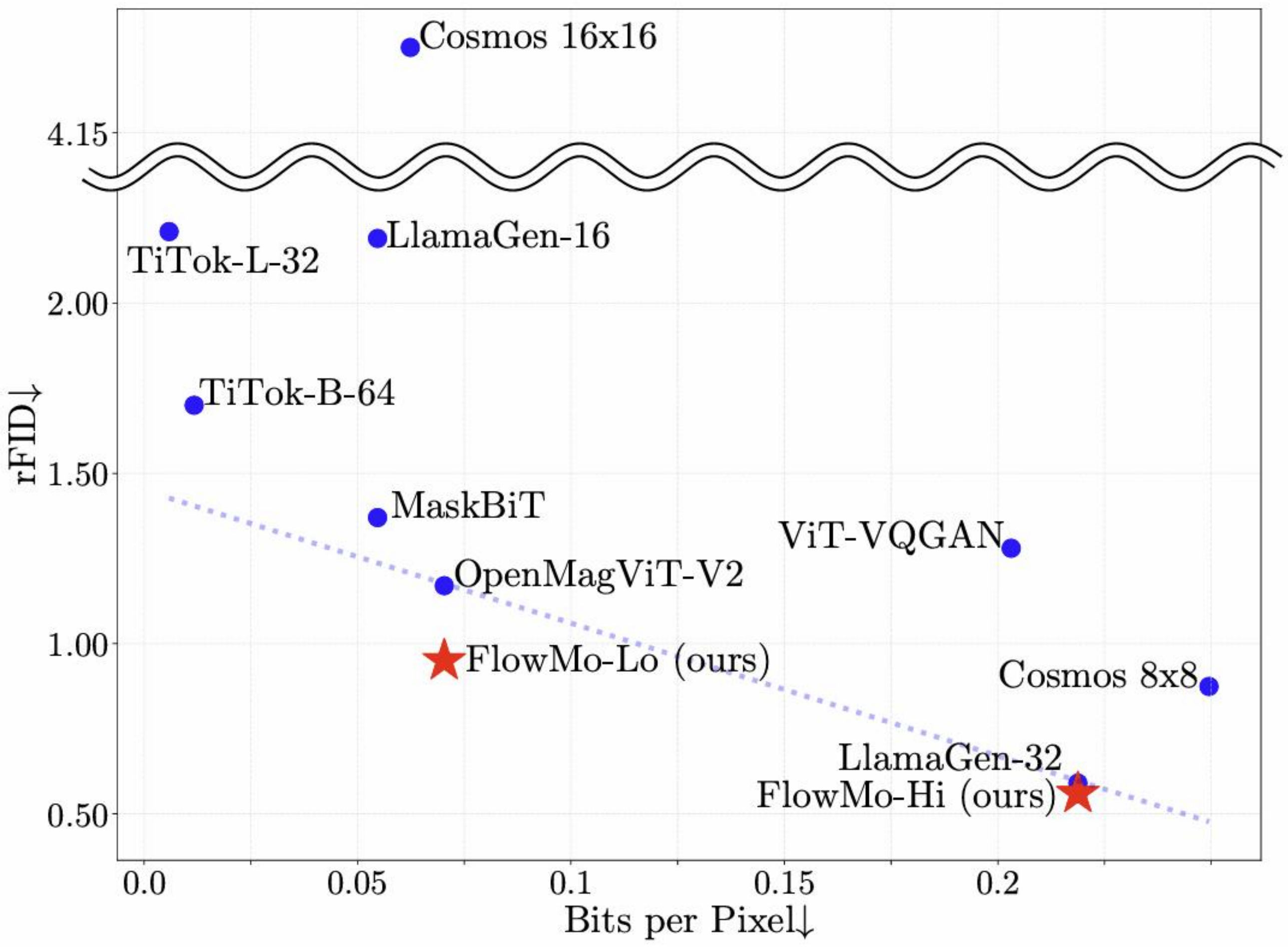
Outline

- Authorship
- Background
- Architecture
- Experiments

Reconstruction

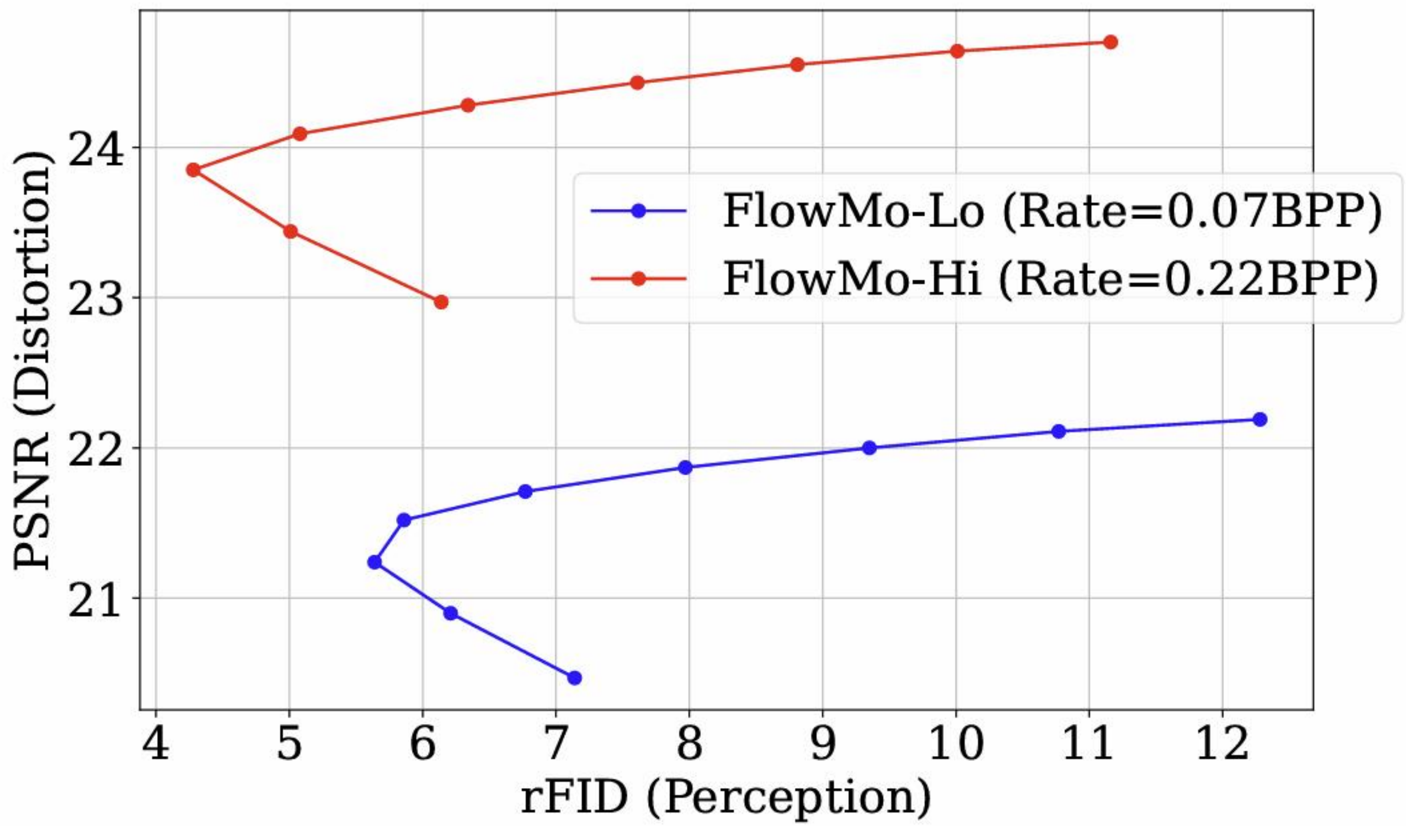
BPP	Model	Tokens per image	Vocab size	rFID↓	PSNR↑	SSIM↑	LPIPS↓
0.006	TiTok-L-32 [63]	32	2^{12}	2.21	15.60	0.359	0.322
0.012	TiTok-B-64 [63]	64	2^{12}	1.70	16.80	0.407	0.252
0.023	TiTok-S-128 [63]	128	2^{12}	1.71	17.52	0.437	0.210
0.055	LlamaGen-16 [53]	256	2^{14}	2.19	20.67	0.589	0.132
	MaskBiT† [57]	256	2^{14}	1.37	21.5	0.56	-
0.062	Cosmos DI-16x16 [63]	256	$\approx 2^{16}$	4.40	19.98	0.536	0.153
0.070	OpenMagViT-V2 [38]	256	2^{18}	1.17	21.63	0.640	**0.111**
	FlowMo-Lo (ours)	256	2^{18}	**0.95**	**22.07**	**0.649**	0.113
0.203	ViT-VQGAN† [61]	1024	2^{13}	1.28	-	-	-
0.219	LlamaGen-32 [53]	1024	2^{14}	0.59	24.44	0.768	**0.064**
	FlowMo-Hi (ours)	1024	2^{14}	**0.56**	**24.93**	**0.785**	0.073
0.249	Cosmos DI-8x8 [63]	1024	$\approx 2^{16}$	0.87	24.82	0.763	0.072

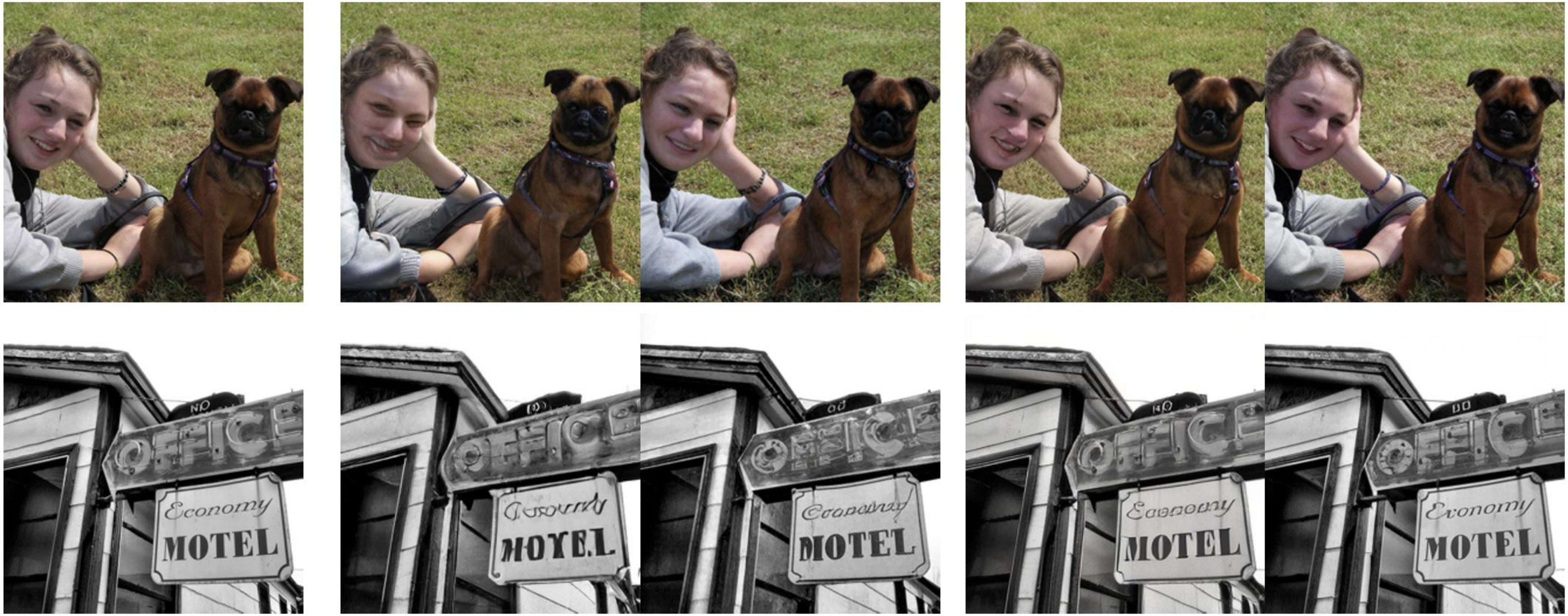
rFID - bpp



Cosmos 16x16
4.15
TiTok-L-32
LlamaGen-16
2.00
TiTok-B-64
rFID↓
1.50
MaskBiT
ViT-VQGAN
OpenMagViT-V2
1.00
FlowMo-Lo (ours)
Cosmos 8x8
LlamaGen-32
FlowMo-Hi (ours)
0.50
0.0
0.05
0.1
0.15
0.2
Bits per Pixel↓

PSNR - rFID trade-off





Original image | OpenMagViT-V2 (rFID=1.17) | FlowMo-Lo (Ours) **(rFID=0.95)** | LlamaGen-32 (rFID=0.59) | FlowMo-Hi (Ours) **(rFID=0.56)**

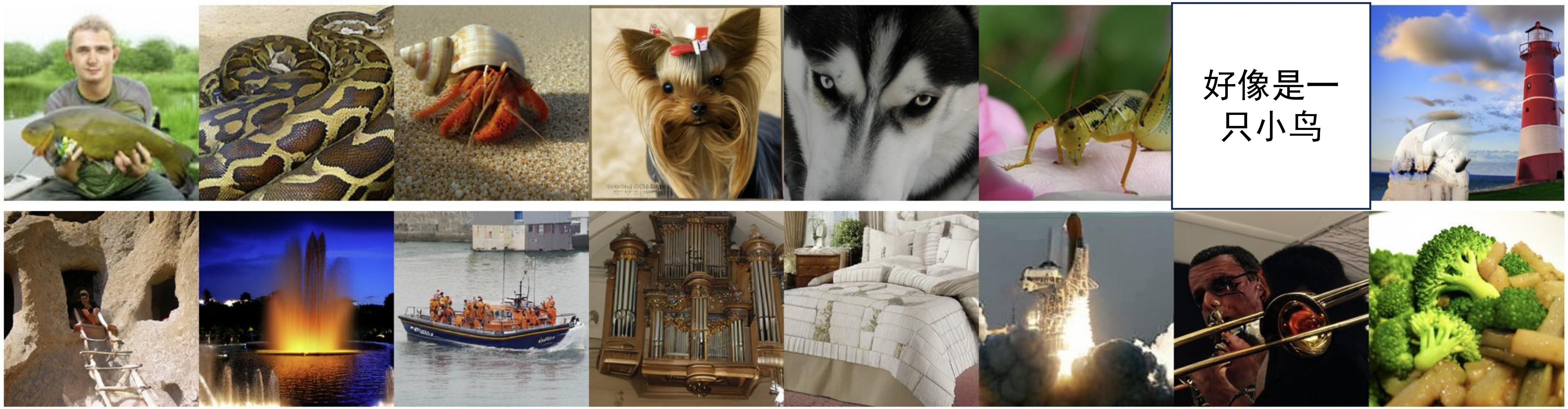
Original image | Reconstructed (OpenMagViT-V2 [38]) | Reconstructed (FlowMo-A)

Generation

Tokenizer	FID ↓	IS ↑	sFID ↓	Prec. ↑	Rec. ↑
OpenMagViT-V2	**3.73**	241	10.66	0.80	**0.51**
FlowMo-Lo (ours)	4.30	**274**	**10.31**	**0.86**	0.47

Table 2. **Generation results**. We compare two MaskGiT transformers trained atop two tokenizers at the same BPP.

Fully generated images - using OpenMagViT-V2 tokenizer (CFG=10.0)



Fully generated images - using FlowMo-Lo tokenizer (CFG=10.0)

Figure 6. **Generated images**. Example generated images from MaskGiT trained with different tokenizers. FlowMo can be used to train high-quality second-stage generative models. The corresponding class indices are identical for ease of comparison.

鸟图预警

Perceptual compression

Original image　　Reconstructed　　Variance heatmap

Figure 8. **Multimodal reconstruction.** After post-training, FlowMo reconstruction remains multimodal, but biased towards preserving the perceptually relevant details of the image, which manifests here by the variance concentrating in the background.

Ablation Study

PEKING UNIVERSITY 北京大学

Model	rFID↓	PSNR↑	LPIPS↓
FlowMo (fewer params.)	**2.87**	20.71	0.15
with doubled patch size	6.39	19.94	0.17
with MSE-trained encoder	3.82	21.40	0.15
without perceptual loss	13.86	**22.11**	0.21
with FSQ quantization	3.14	21.31	**0.14**
with logit-normal schedule	4.08	16.45	0.21
without shifted sampler	3.42	20.25	0.16
without guidance	3.28	20.67	0.16

Table 4. **Stage 1A Ablation**. Deviating from FlowMo design choices compromises either PSNR or rFID. We prioritize rFID in our model due to its correlation with perceptual quality.

Model	rFID↓	PSNR↑	LPIPS↓
FlowMo-Lo	1.10	21.38	0.134
FlowMo-Lo (post-trained)	**0.95**	**22.07**	**0.113**
FlowMo-Hi	0.73	24.02	0.086
FlowMo-Hi (post-trained)	**0.56**	**24.93**	**0.073**

Table 5. **Stage 1B ablation**. Without the post-training stage, performance is considerably worse.

Conclusion

- A 1-D tokenizer with MMDiT encoder / decoder.
- Without adversarial loss or distillation
- SOTA reconstruction performance

- What is Mode Seeking?
- Ineffective decoding
- More exploration about generation is needed

Thanks for your listening!