ICLR

BiGR: Harnessing Binary Latent Codes for Image Generation and Improved Visual Representation Capabilities

Shaozhe Hao[1], Xuantong Liu[2], Xianbiao Qi[3], Shihao Zhao[1], Bojia Zi[4],
Rong Xiao[3], Kai Han[1], Kwan-Yee K. Wong[1]

[1]The University of Hong Kong [2]Hong Kong University of Science and Technology
[3]Intellifusion [4]The Chinese University of Hong Kong

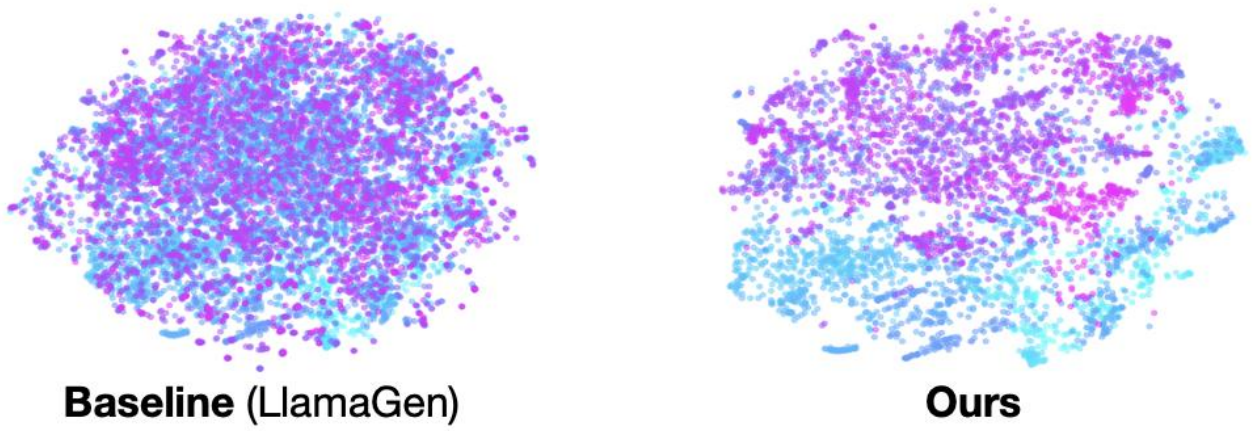
香 港 大 學
THE UNIVERSITY OF HONG KONG

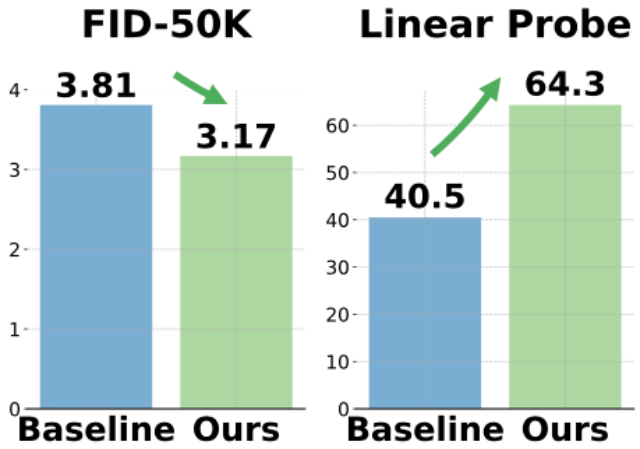
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

intellifusion
云 天 励 飞

香港中文大學
The Chinese University of Hong Kong

Can image generative models produce discriminative visual representations?



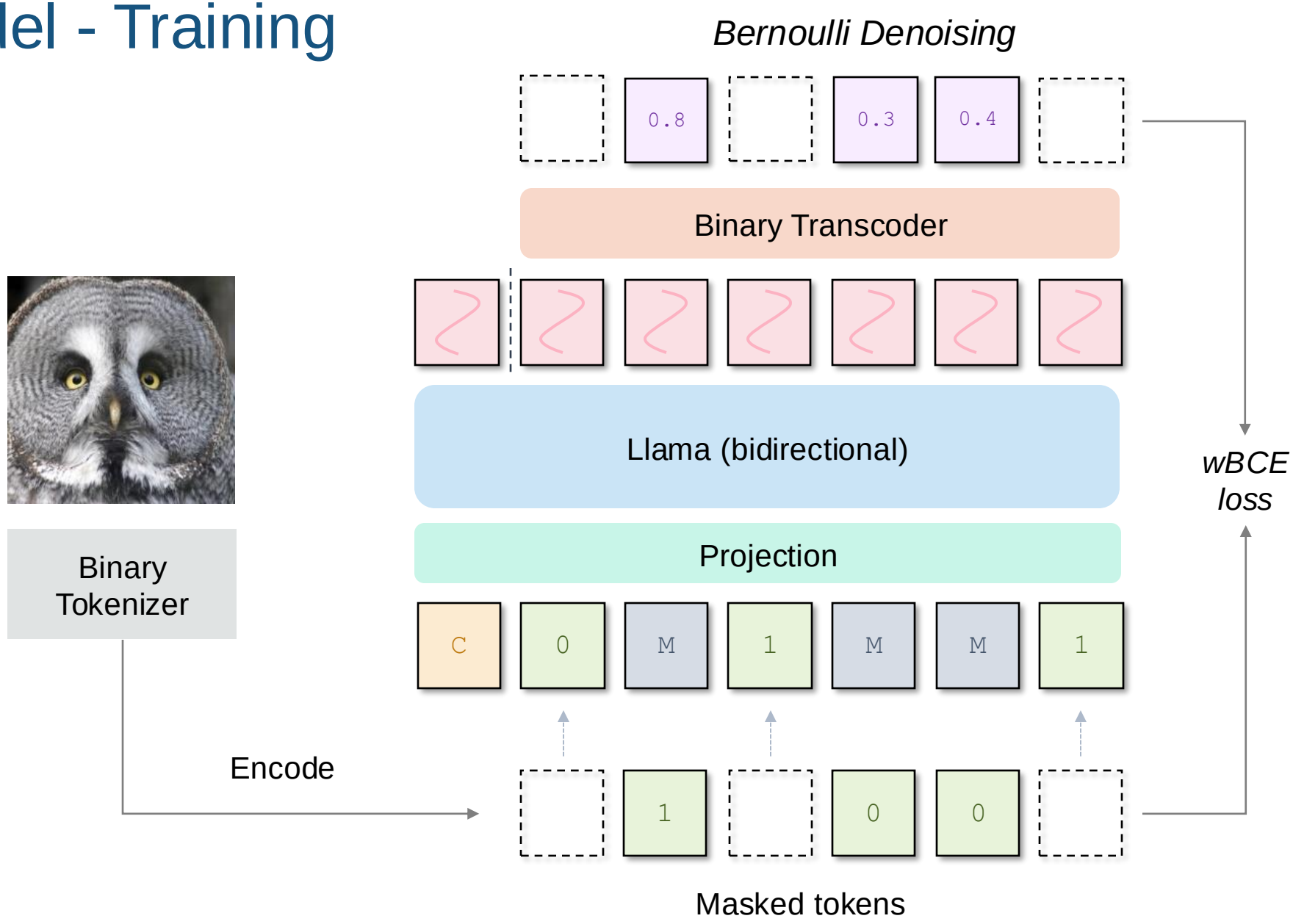


Samples from the same class are in the same color.

We introduce a novel conditional image generation model, which unifies generative and discriminative tasks.

Sun, Peize, et al. "Autoregressive model beats diffusion: Llama for scalable image generation." arXiv preprint arXiv:2406.06525 (2024).

Our model - Training



Binary transcoder - Bernoulli diffusion

- Forward diffusion process:

$$q(z^t|z^{t-1}) = \mathcal{B}\left(z^t; z^{t-1}(1-\beta^t) + 0.5\beta^t\right) \quad t = 1, 2, \cdots, T.$$

z^t: z at timestep t
β^t: predefined coefficient at timestep t

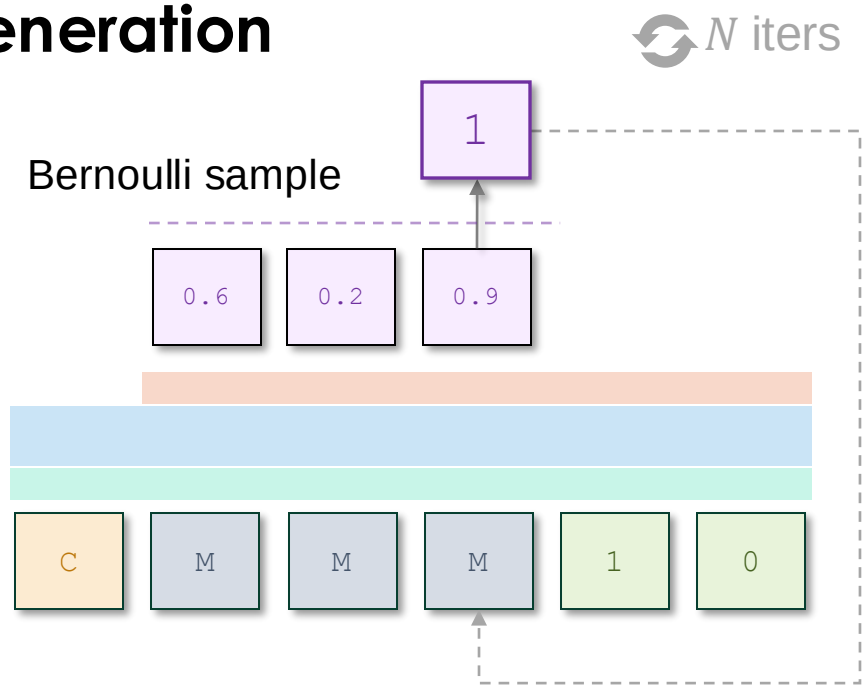
- Backward denoising process:

$$p_\phi(z^{t-1}|z^t) = \mathcal{B}\left(z^{t-1}; S(g_\phi(z^t, t, h))\right)$$

S: sigmoid function
g_θ: denoising network (MLPs)

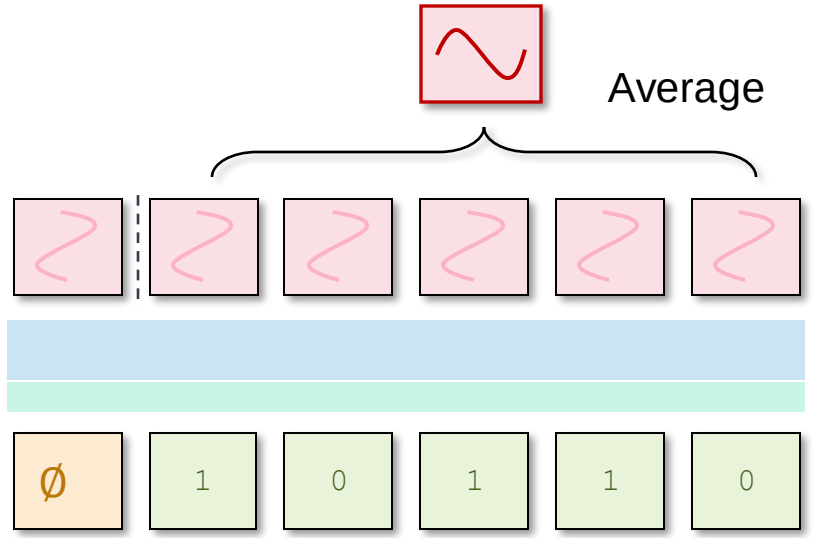
Inference

Generation



$$\mathcal{H} = -\frac{1}{K}\sum_{k=1}^{K} p_k \log_2 p_k + (1-p_k)\log_2(1-p_k),$$

Representation



Evaluation

Table 1: **Uniformity comparison.** We compare the generative and discriminative performance of our model against LlamaGen (Sun et al., 2024) and three other settings, varying by tokenizers, training objectives, and modeling types. We use KV cache to accelerate all AR models.

					Generative					Discriminative	
Model	Tokenizer	Objective	Type	Time↓	FID↓	IS↑	sFID↓	Pre.↑	Rec.↑	ACC1	ACC5
LlamaGen	VQGAN	Cat.	AR	0.13	3.81	248.28	8.49	0.83	0.52	40.5	64.4
S0	B-AE	Cat.	AR	0.15	3.21	239.17	**5.38**	0.83	**0.54**	23.8	44.2
S1	B-AE	Cat.	Mask	**0.10**	3.85	261.81	6.10	0.85	0.47	61.1	83.2
S2	B-AE	Bin.	AR	1.04	7.50	164.31	6.56	0.85	0.41	45.2	69.3
S3 (Ours)	B-AE	Bin.	Mask	0.69	**3.17**	**262.14**	5.59	**0.86**	0.50	**64.3**	**85.4**

Model scaling

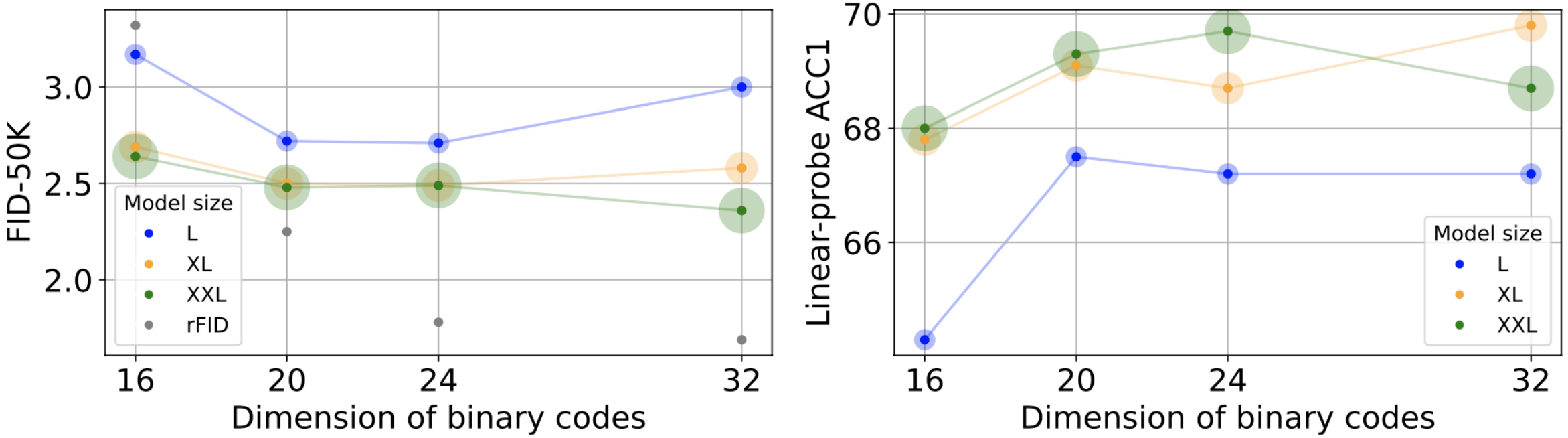
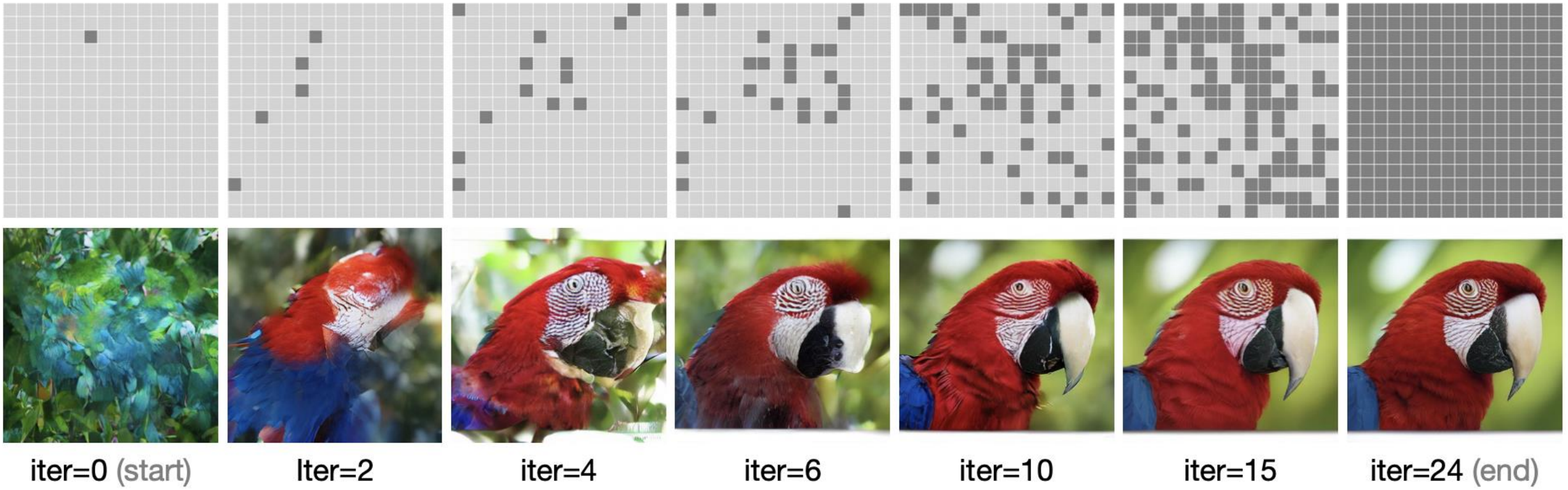
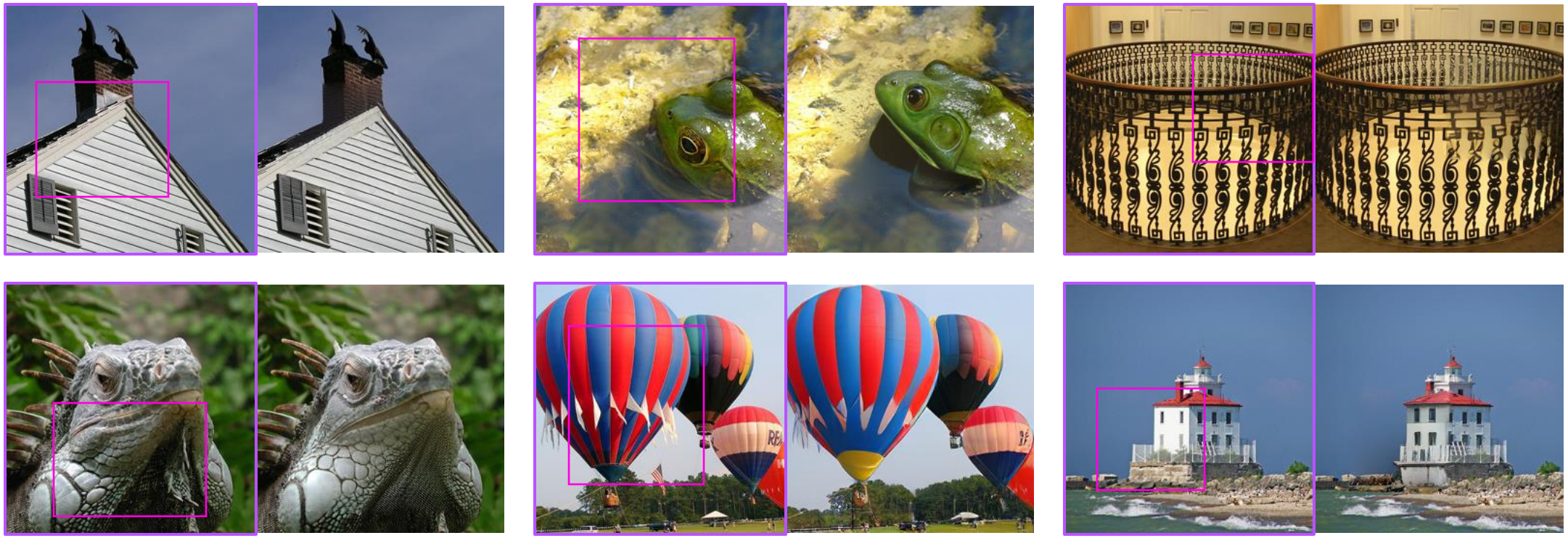


Figure 4: **Evaluation of generative and discriminative performance across different model sizes.** We report results for all tested tokenizers across four different dimensions of binary codes. We include the reconstruction FID (rFID) for each binary tokenizer for reference (grey points).

Visualization of the entropy-ordered sampling process

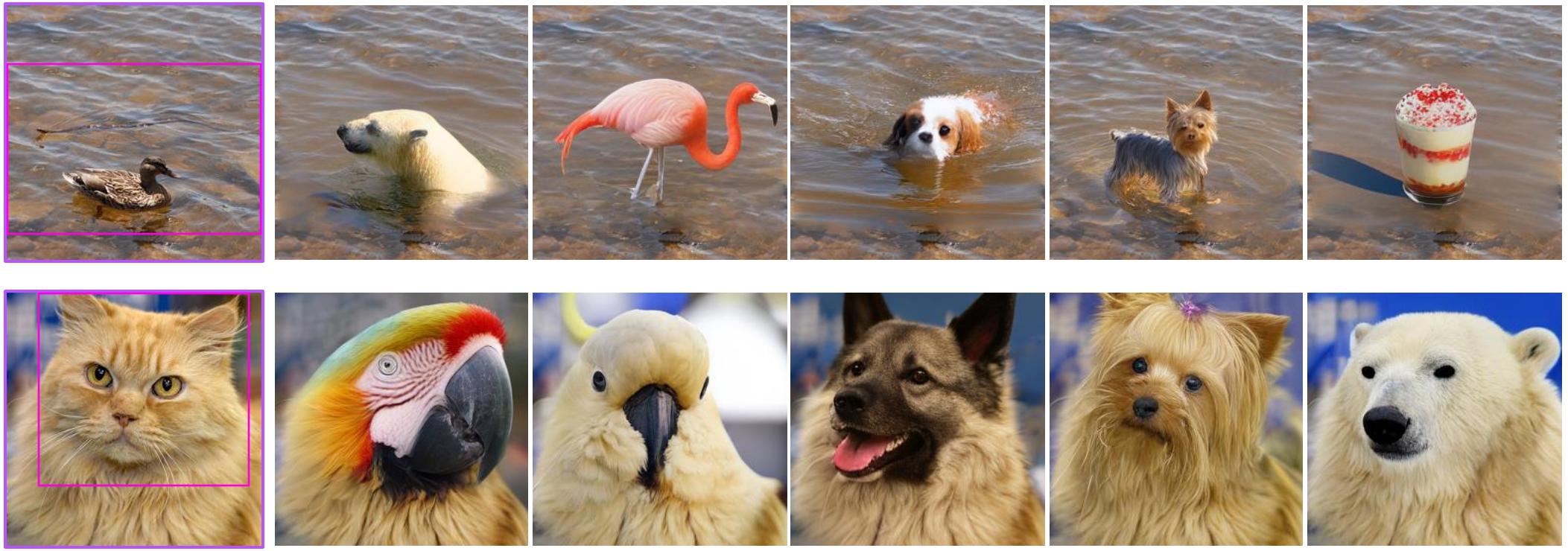


Zero-shot generalization – Inpainting



Zero-shot generalization – Outpainting

Zero-shot generalization – Editing



Zero-shot generalization – Interpolation

Text-to-image generation

A man looks up at the starry sky, lonely.

An astronaut riding a horse.

Crocodile in a sweater.

A dog is reading a thick book.

the black hole in the space.

a table top with a vase of flowers on it.

an illustration of a teapot.

the Eiffel Tower in winter

A snowy mountain.

A jellyfish rocket.

Thank you for your listening!

Welcome to our Poster:

Thu 24 Apr 3 p.m. CST — 5:30 p.m. CST

Code & Model:
https://github.com/haoosz/BiGR