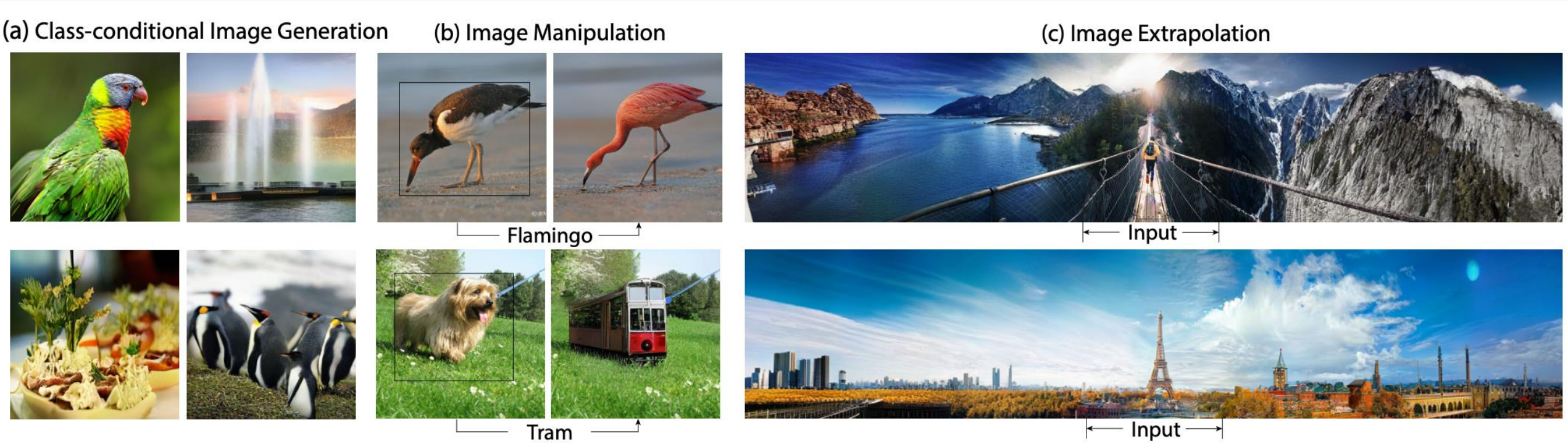
CVPR 2022

MaskGIT

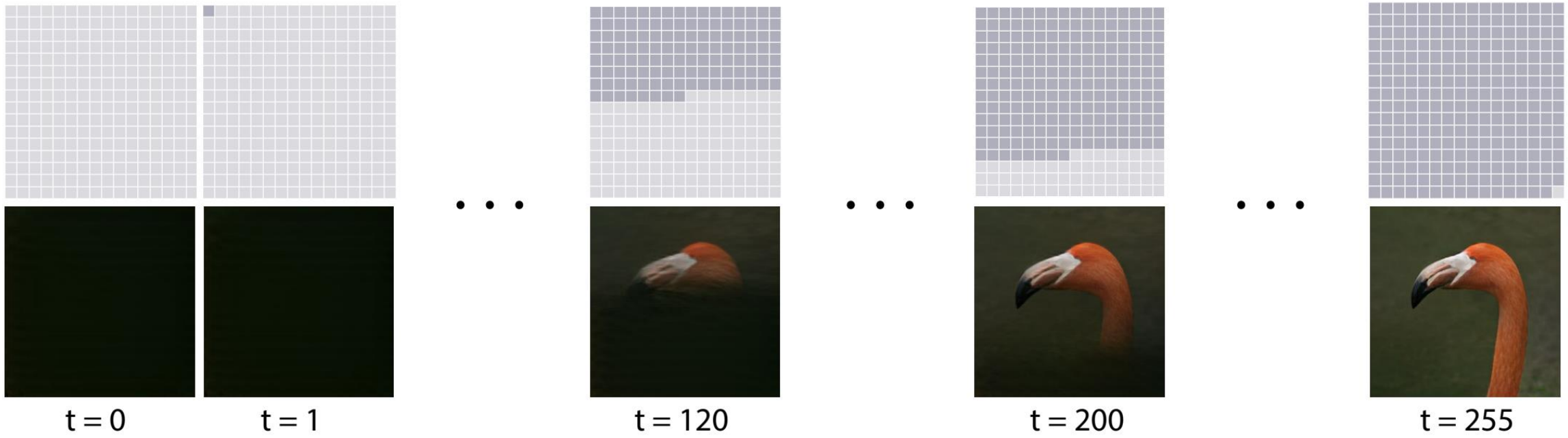
Masked Generative Image Transformer

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, William T. Freeman



Autoregressive image generation

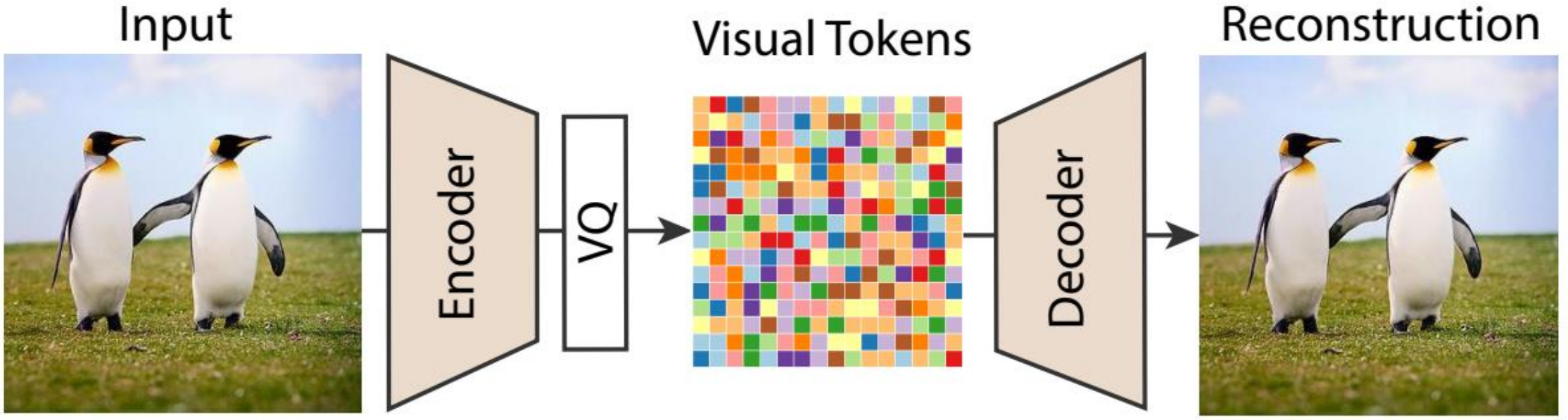
Sequential decoding with AR transformers



256 tokens at 256 × 256

One sequential sampling decision per token

Stage 1: Images as discrete visual tokens (VQ-GAN)



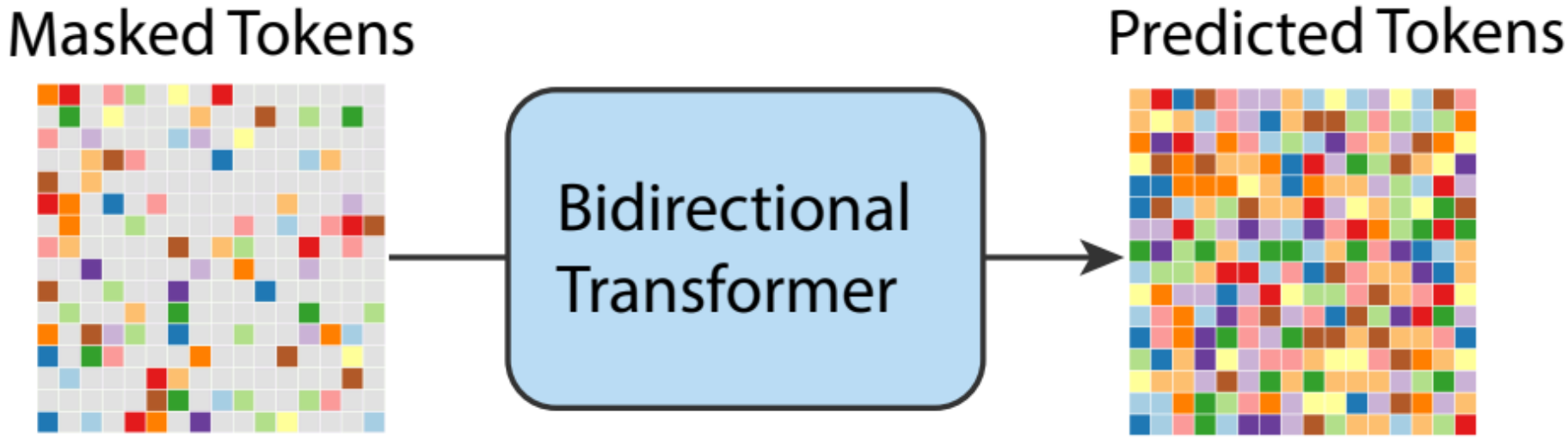
16 × spatial compression

256 × 256 pixels become 16 × 16 tokens

1,024 codebook entries

Learned codes, decoded jointly to pixels

Stage 2: Bidirectional masked prediction



Visible context can come from anywhere in the image

Confidence determines what stays

Example: six unknown positions, three remain masked

Illustrative values. Table shows deterministic ranking before the optional noise term.

Position	A	B	C	D	E	F
Sample probability	0.22	0.81	0.37	0.92	0.46	0.73
Next round	Mask	Keep	Mask	Keep	Mask	Keep

Accepted tokens become fixed context

The released sampler adds annealed randomness to confidence ranking

Training across mask ratios

$r \sim \text{Uniform}(0, 1)$

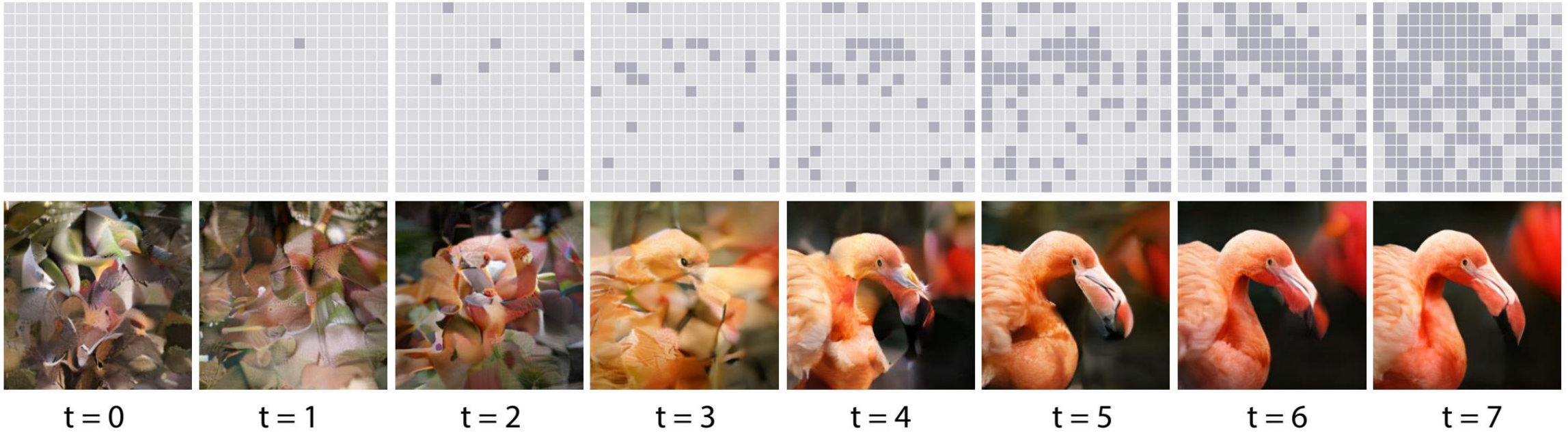
Mask approximately $\gamma(r)N$ tokens

Sampled r	Masked fraction	Available context
0.1	98.8%	Very sparse
0.5	70.7%	Partial image
0.9	15.6%	Mostly visible

Cosine $\gamma(r) = \cos(\pi r / 2)$ exposes the model to nearly empty inputs

Eight rounds in pictures

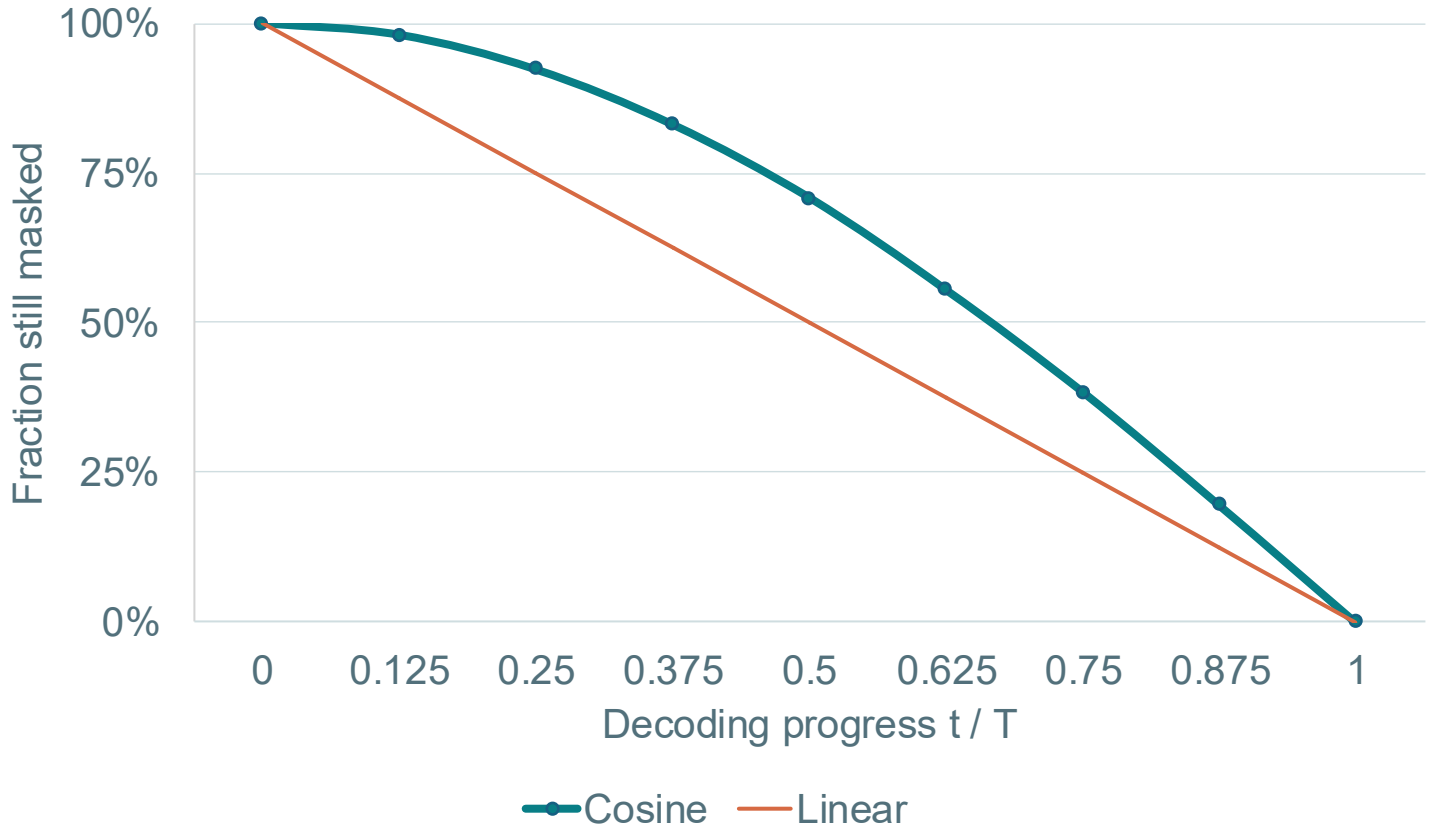
Top: input mask at each round. Bottom: decoded candidate predictions. Figure 2.



Few fixed tokens early

More context before later commitments

Cosine scheduling delays commitment



N = 256, T = 8

Cumulative kept tokens

Round 1	≈ 5
Round 4	≈ 75
Round 7	≈ 207
Round 8	256

Counts illustrate floor-rounded mask budgets. The best iteration count is empirical.

The masked-token training objective

$$\mathcal{L}_{\text{mask}} = -\mathop{\mathbb{E}}_{\mathbf{Y}\in\mathcal{D}}\Big[\sum_{\forall i\in[1,N],m_i=1}\log p(y_i|Y_{\overline{\mathbf{M}}})\Big],$$

Y Ground-truth visual token sequence

M Randomly selected positions replaced with [MASK]

Cross-entropy applies only to masked positions

Iterative parallel decoding

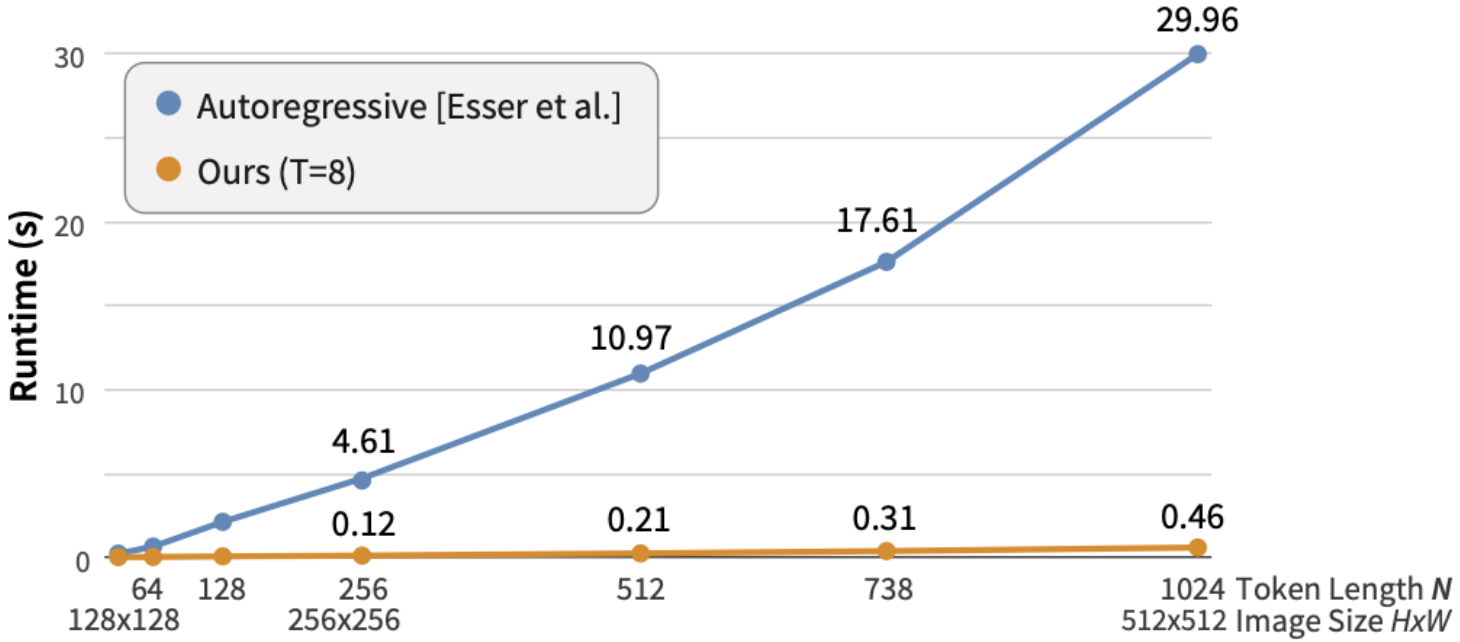
01	**Predict**	Distributions at every unknown position
02	**Sample**	A candidate token and its probability
03	**Schedule**	The number of positions to leave unknown
04	**Commit**	Keep confident candidates and remask the rest

Initialize with [MASK] everywhere. After T rounds, decode the complete token grid.

ImageNet quality

Model	FID ↓	IS ↑	Prec ↑	Rec ↑	# params	# steps	CAS ×100 ↑	
							Top-1 (76.6)	Top-5 (93.1)
ImageNet 256×256								
DCTransformer [32] ▫	36.51	n/a	0.36	**0.67**	738M	>1024		
BigGAN-deep [4]	6.95	**198.2**	**0.87**	0.28	160M	1	43.99	67.89
Improved DDPM [33]▫	12.26	n/a	0.70	0.62	280M	250		
ADM [12]▫	10.94	101.0	0.69	0.63	554M	250		
VQVAE-2 [37]▫	31.11	∼45	0.36	0.57	13.5B†	5120	54.83	77.59
VQGAN [15]▫	15.78	78.3	n/a	n/a	1.4B	256		
VQGAN*	18.65	80.4	0.78	0.26	227M	256	53.10	76.18
MaskGIT (Ours)	**6.18**	182.1	0.80	0.51	227M	*8*	**63.14**	**84.45**
ImageNet 512×512								
BigGAN-deep [4]	8.43	**232.5**	**0.88**	0.29	160M	1	44.02	68.22
ADM [12]▫	23.24	58.06	0.73	**0.60**	559M	250		
VQGAN*	26.52	66.8	0.73	0.31	227M	1024	51.29	74.24
MaskGIT (Ours)	**7.32**	156.0	0.78	0.50	227M	*12*	**63.43**	**84.79**

Decoding rounds and wall-clock time



Up to 64×

Reported speedup over the tested autoregressive decoder

Full-sequence attention still gets more expensive as N grows

Figure 4. Single-GPU transformer timings at T = 8. The 512-quality result uses T = 12.

Schedule ablation

γ	T	FID ↓	IS ↑	NLL
Exponential	8	7.89	156.3	4.83
Cubic	9	7.26	165.2	4.63
Square	10	6.35	179.9	4.38
Cosine	10	**6.06**	**181.5**	4.22
Linear	16	7.51	113.2	3.75
Square Root	32	12.33	99.0	3.34
Logarithmic	60	29.17	47.9	3.08

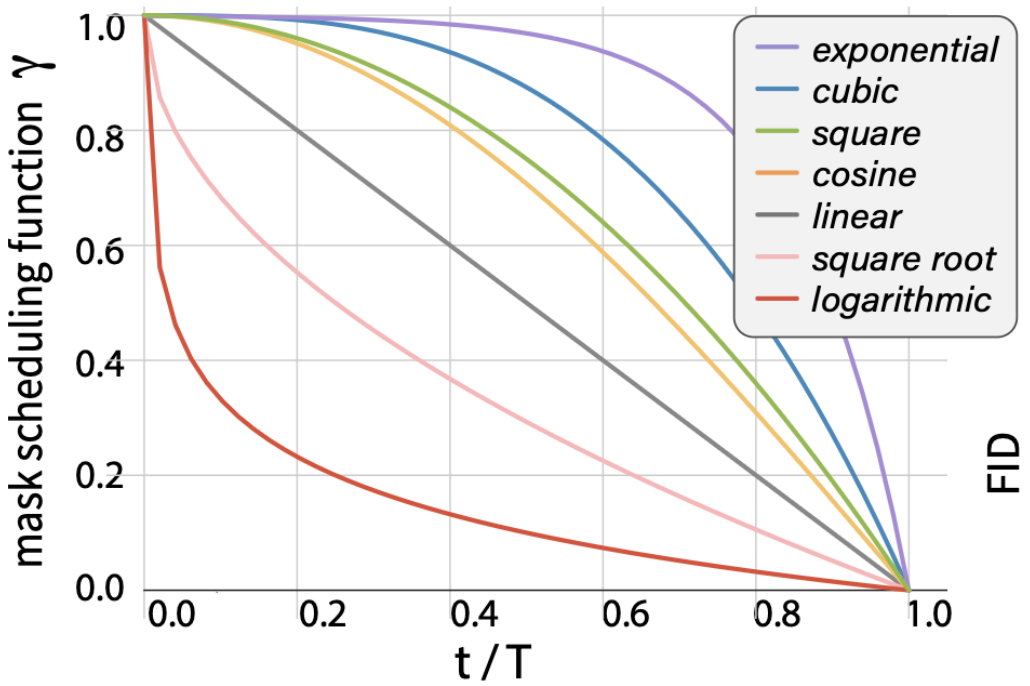


Table 3. Cosine's 6.06 uses 10 rounds. The main Table 1 result is 6.18 with 8 rounds.

Diversity requires a separate check

ImageNet 256	Precision ↑	Recall ↑	CAS top-1 ↑
BigGAN-deep	0.87	0.28	43.99%
MaskGIT	**0.80**	**0.51**	**63.14%**

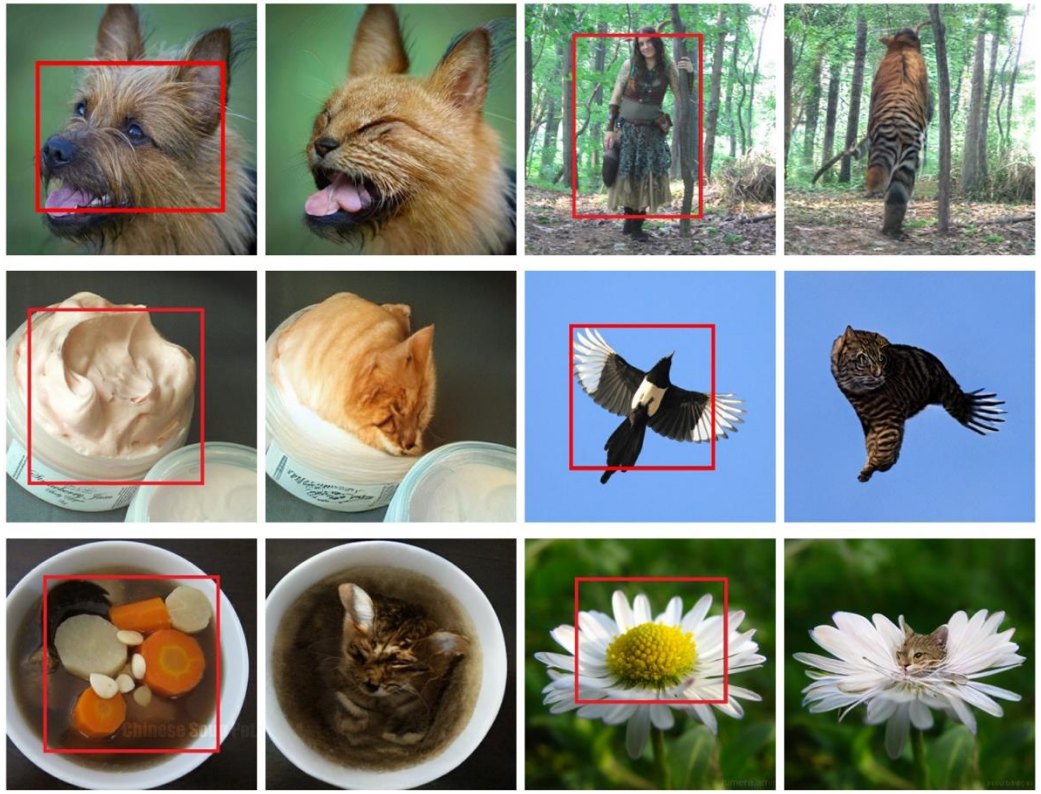
BigGAN-deep

MaskGIT

Higher recall, lower precision

CAS trains on synthetic images and tests on real images

Editing starts with a different mask



Target class: tiger cat

Class editing

Keep known tokens. Mask the target region and change the class condition.

Inpainting and outpainting

Fill an interior hole or extend beyond the visible image.

Figure 6. Scene completion uses the Places2 model and blends output near mask boundaries.

Limits of the method

Complex details

Faces, text and symmetry can distort or oversmooth

Context and commitment

Long outpainting can drift. Early accepted errors can persist.

Figure 21 and Appendix F. Tokenization also limits the detail the model can reconstruct.

Questions

- Why does using more decoding iterations reduce image quality after a certain point?
- Why does MaskGIT's FID score degrade when increasing the number of inference steps beyond 12 (Figure 8), whereas continuous diffusion models consistently improve or plateau with additional sampling steps?