Unlocking the Generative Capabilities of Masked Generative Models for Image Synthesis via Self-Guidance

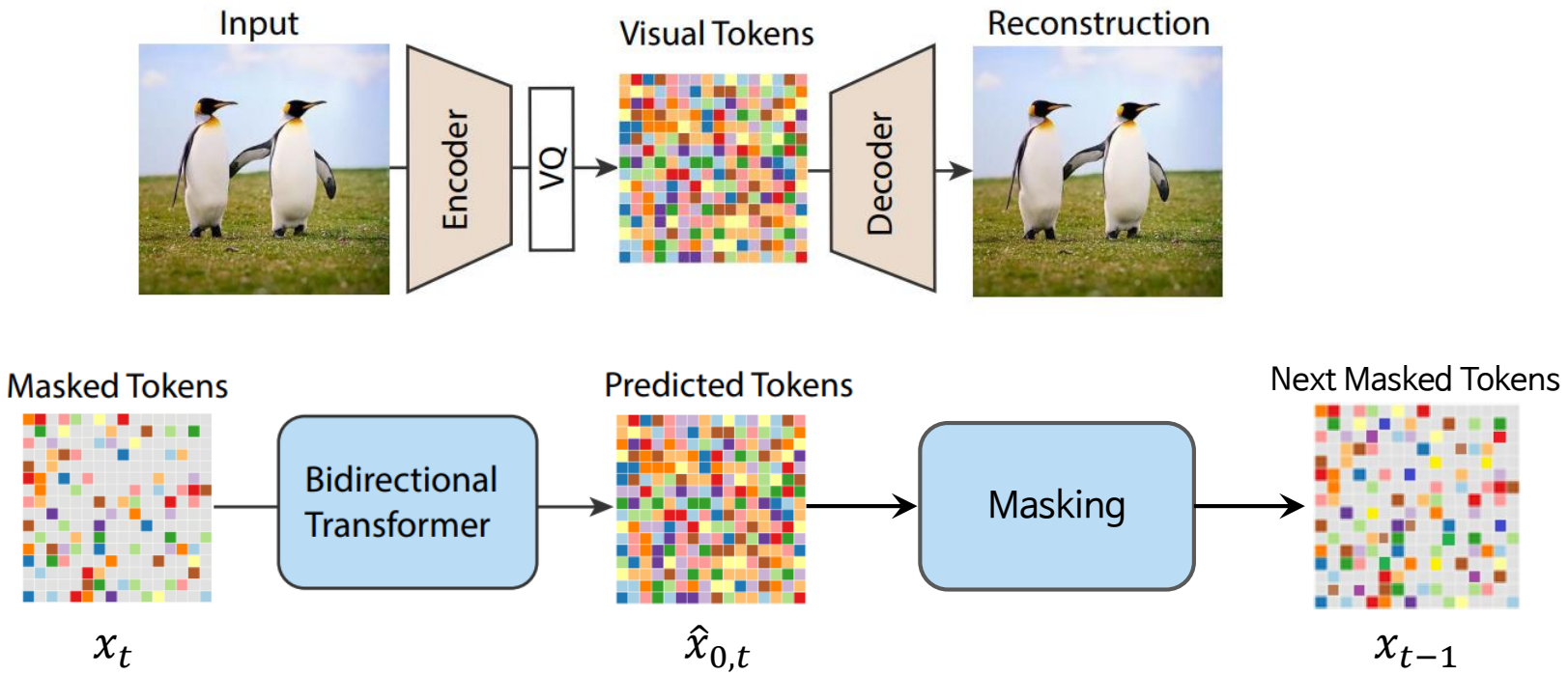
Jiwan Hur[1], Dong-Jae Lee[1], Gyojin Han[1], Jaehyun Choi[1], Yunho Jeon[2†], and Junmo Kim[1†]

[1] KAIST, South Korea

[2] Hanbat National University, South Korea

Masked Generative Models

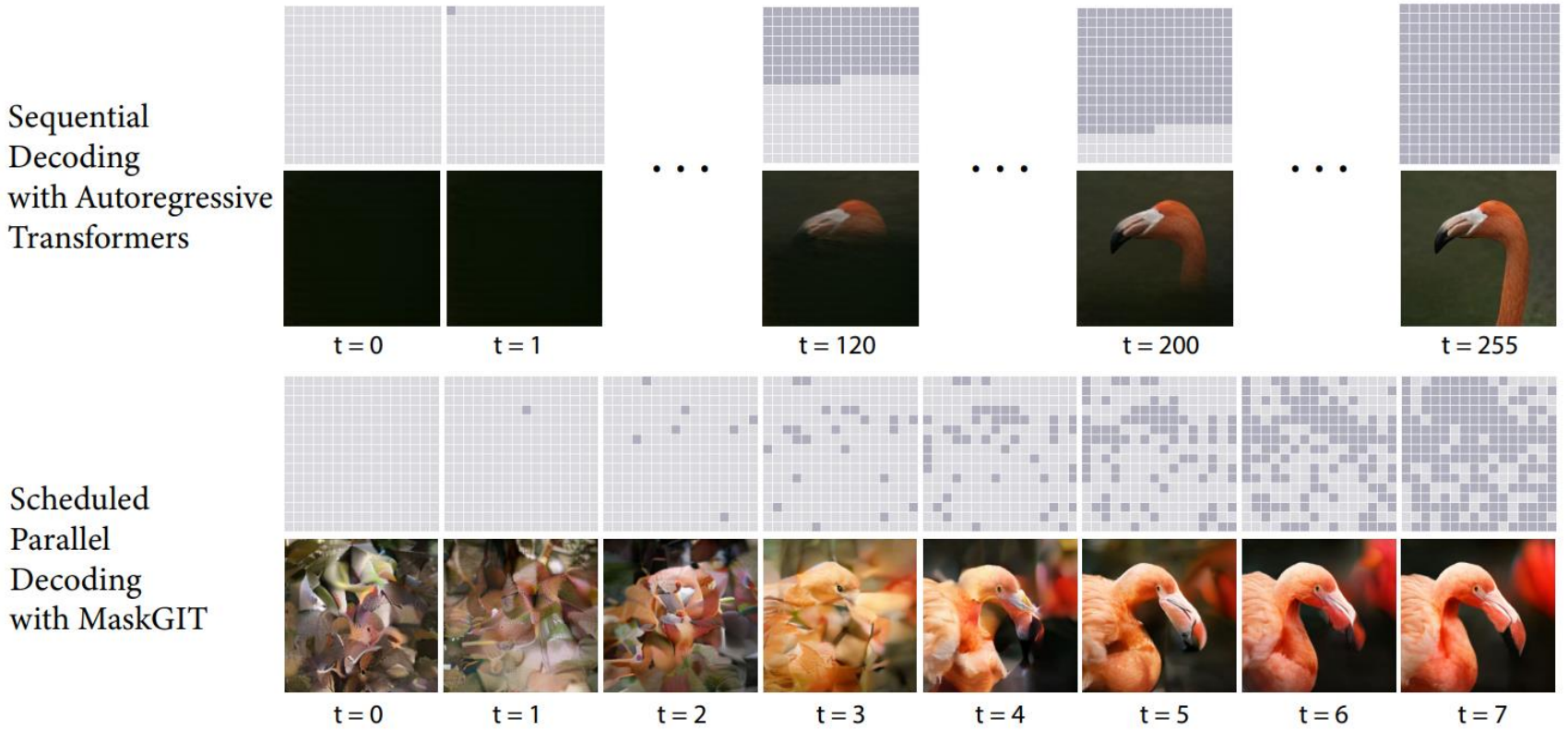
- Masked Generative Models (MGMs) are a family of discrete diffusion models, which aim to generate discrete data by predicting masked regions.
- Recently, vector-quantized (VQ) token-based MGMs have shown impressive performance.



Chang et al. "MaskGIT: Masked Generative Image Transformer" @ CVPR 2022

Masked Generative Models

- VQ token-based MGMs have shown efficient generative capabilities compared to diffusion models. (~18 steps).
- However, MGMs underperform well-improved continuous diffusion models such as LDM.



Chang et al. "MaskGIT: Masked Generative Image Transformer" @ CVPR 2022

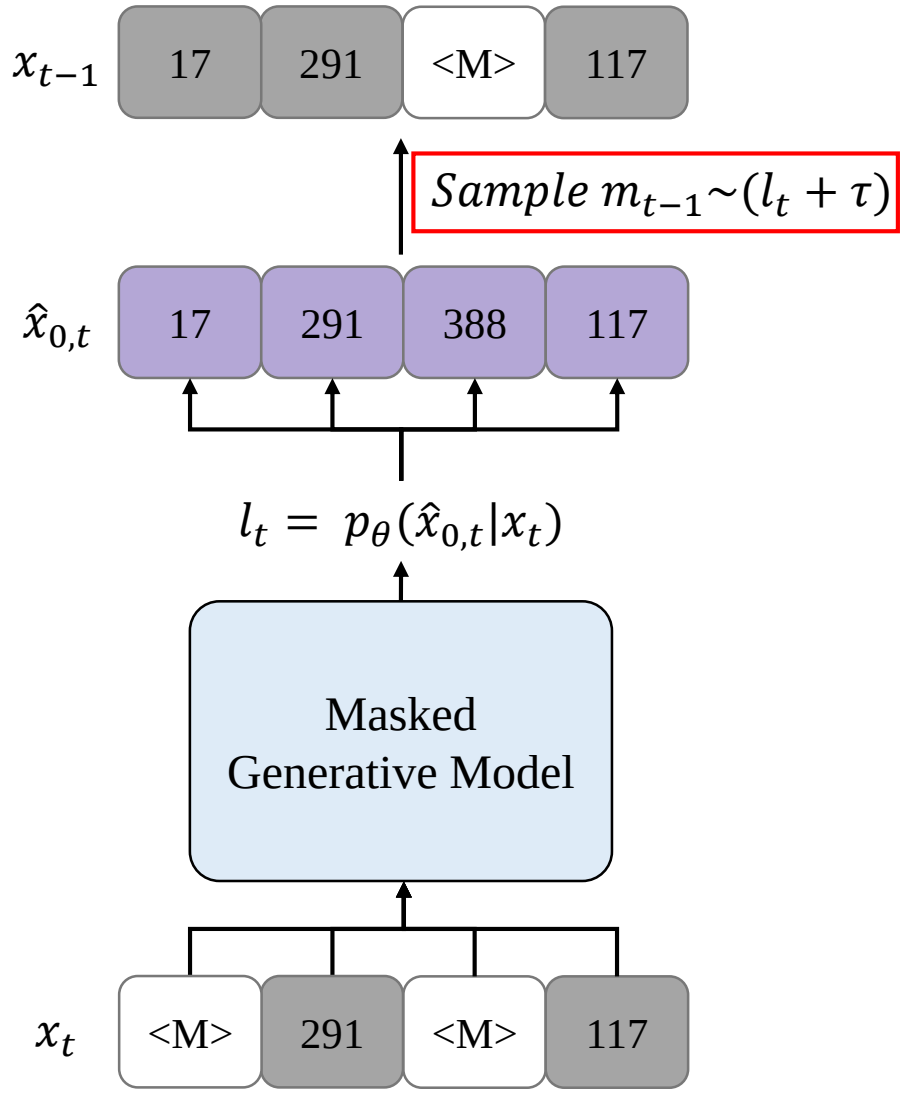
Improved Sampling Strategies of MGMs

Two approaches exist to improve the sample quality of MGMs.

1. **Low temperature sampling (τ)**
   - Restrict stochasticity of sampling procedure. (quality ↑, diversity ↓)
   - However, very low temperature harm the quality of samples due to the ***multi-modality problem***.

2. **Predictor-Corrector sampling** (Token-Critic, DPC)
   - Discern unrealistic tokens via external model.
   - Requires high training cost and more sampling steps.

- Diffusion models utilize various **guidance** techniques to improve the sample quality while sacrificing diversity.



Lezama et al. "Improved Masked Image Generation with Token-Critic" @ ECCV 2022
Lezama et al. "discrete predictor-corrector diffusion models for image synthesis" @ ICLR 2023

Generalized Sampling Guidance

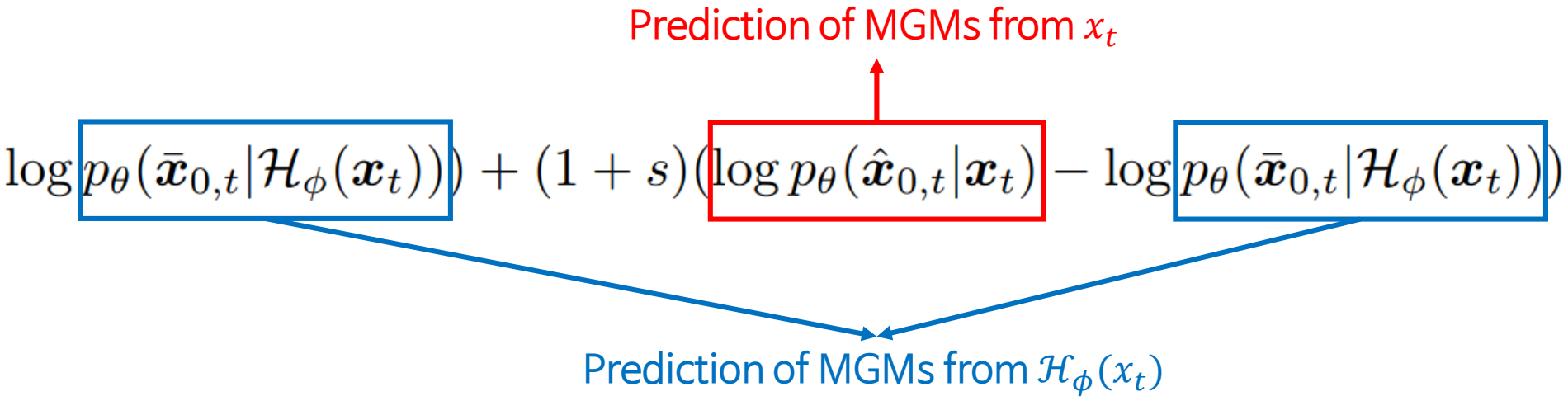
Define **Generalized guidance formulation** for discrete diffusion models from the optimization perspective:

$$\arg\max_{\phi}[\log p(\mathcal{H}_\phi(\boldsymbol{x}_t)) + (1+s)(\log p(\boldsymbol{x}_t) - \log p(\mathcal{H}_\phi(\boldsymbol{x}_t)))]$$

- $\mathcal{H}_\phi$ is information bottleneck that removes salient information h_t from x_t.
- h_t can be either internal or external information of x_t .
- It guides the sampling process toward enhancing specific information h_t.
- When $h_t = c$, it collapses to discrete classifier-free guidance (CFG) proposed in Improved VQ Diffusion.

Tang et al. "Improved vector quantized diffusion models" @ arXiv 2022

Intuition for Self-Guidance



Since we aim to improve the sample quality, we define h_t as fine scale details within x_t.

Intuition for Semantic Smoothing

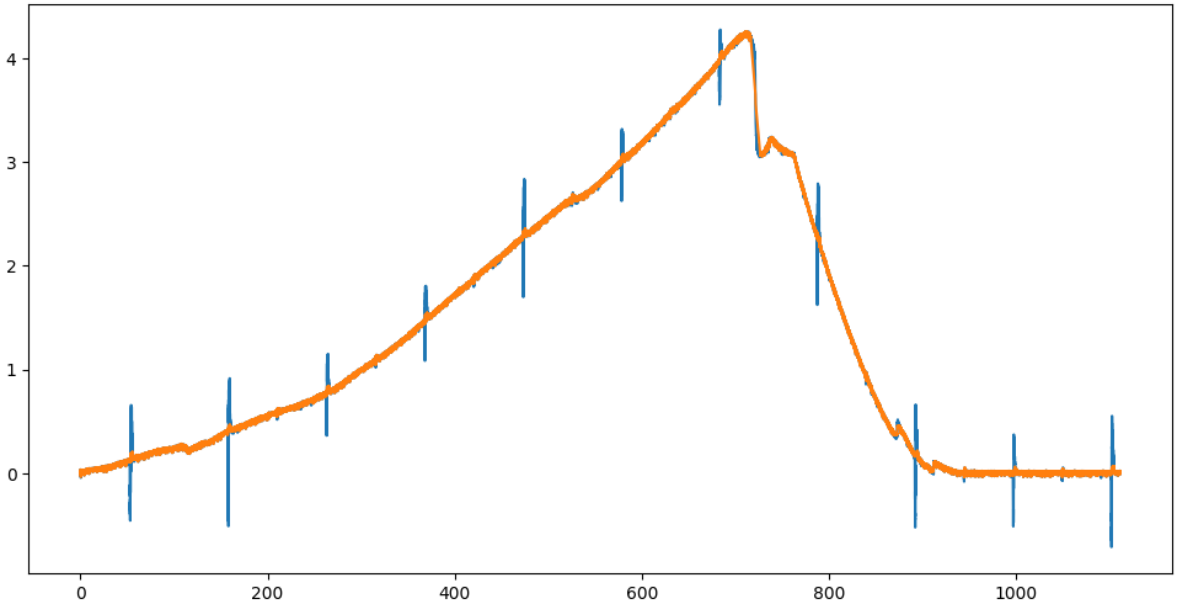
- How to generate coarse information from VQ tokens?

→ We aim to apply smoothing on VQ tokens x_t .

  - Continuous pixel space: Spatial Smoothing (e.g. Blur)
  - Semantic discrete space (VQ token): Semantic Smoothing

Intuition for Auxiliary Task

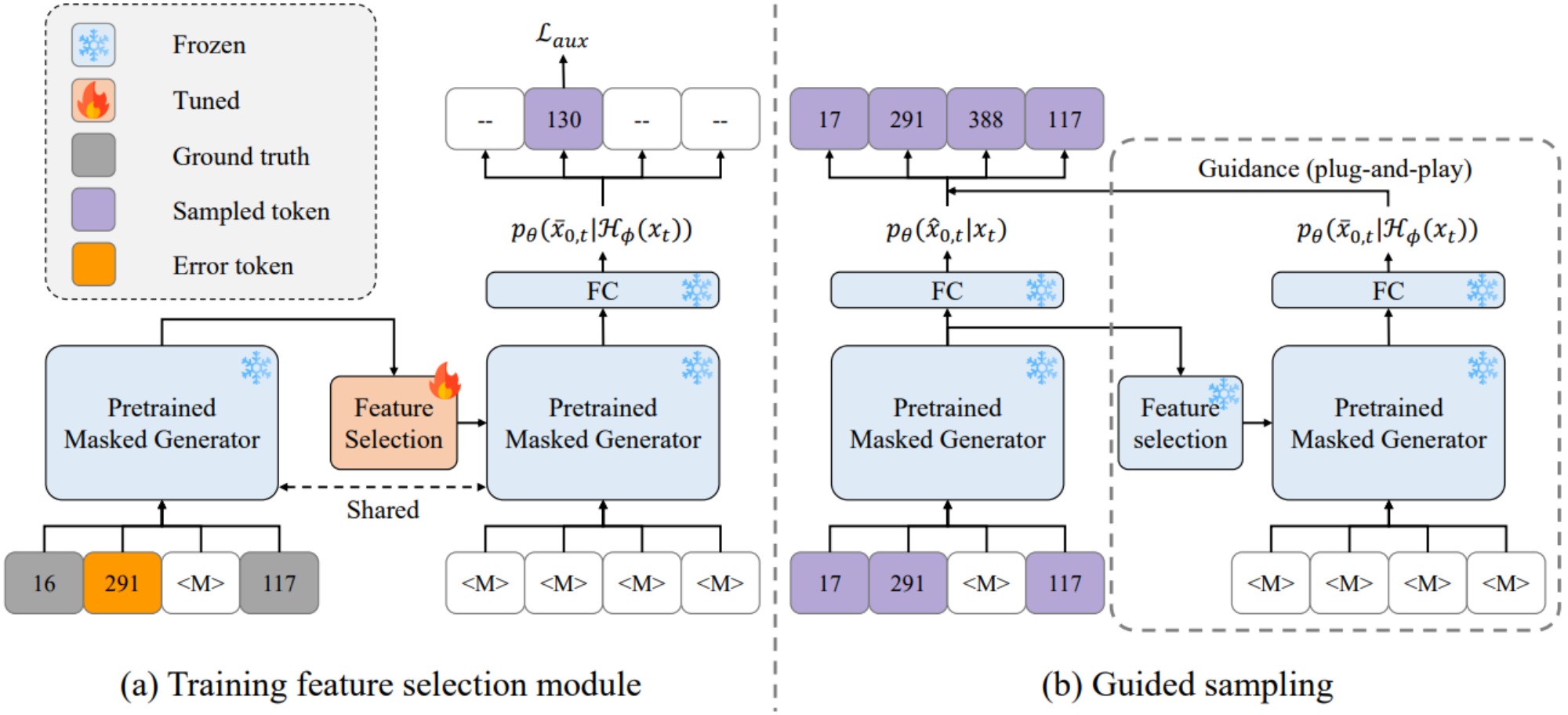
- We propose error token correction as an auxiliary task to generate semantically smoothed output of x_t.
- Error tokens often act as a semantic outlier.
- To correct the error tokens, model implicitly learns to generate semantically smoothed output to minimize the empirical risk for all data points.



1-dimensional numerical signal example

Method

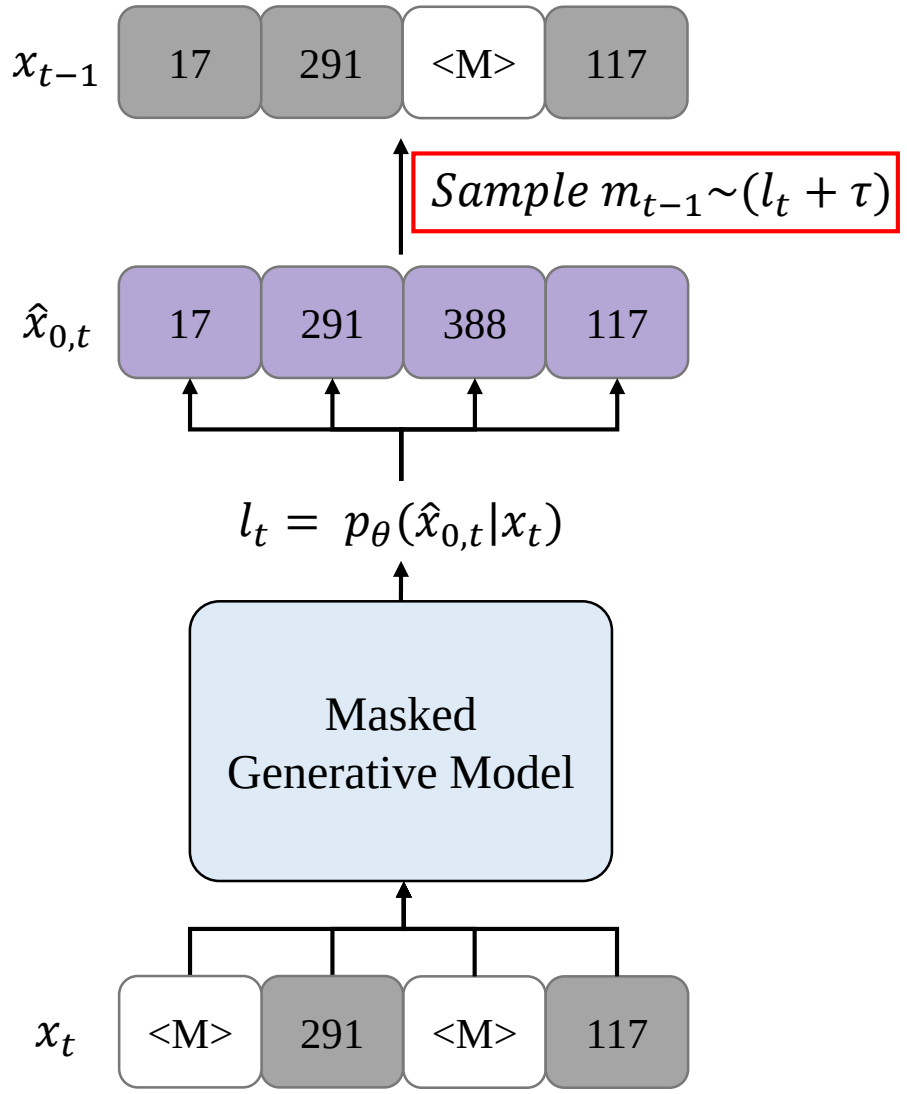
- We leverage pre-trained MGMs to utilize the generative priors.
- We adopt TOAST [1] for parameter-efficient fine-tuning (PEFT) method.



Shi, Baifeng, et al. "Toast: Transfer learning via attention steering." @ arXiv 2023.

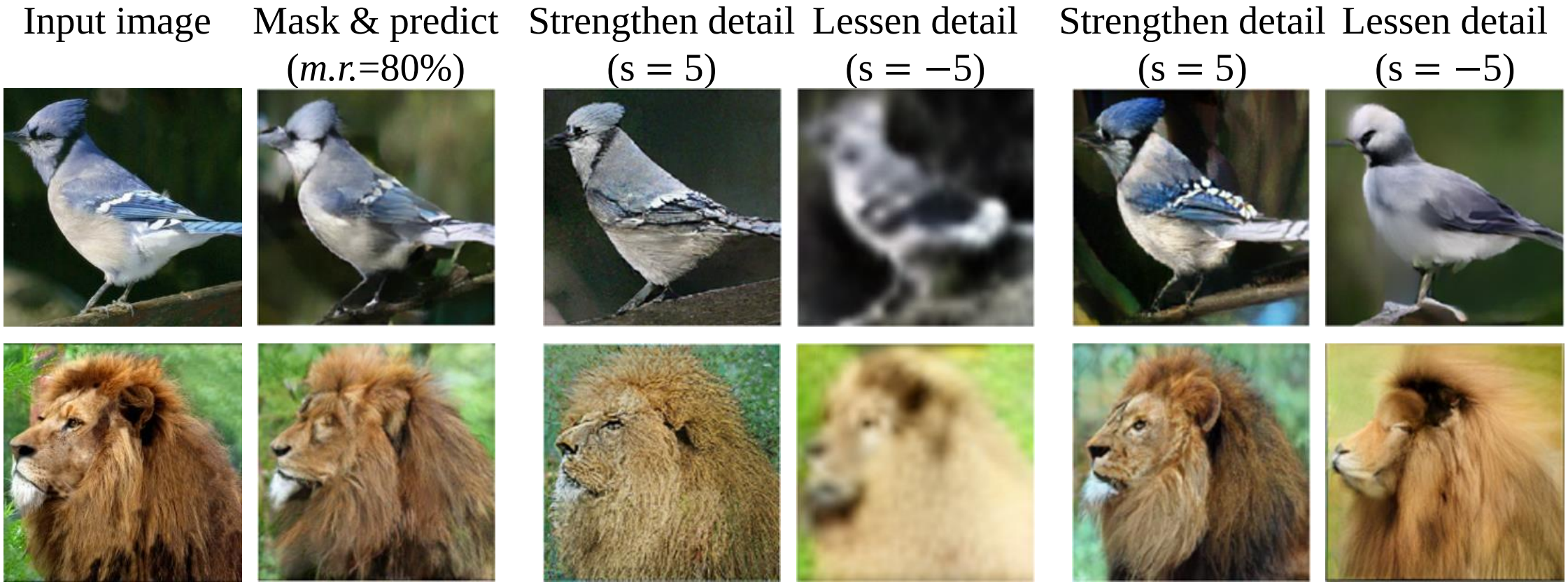
Guided Sampling with High Temperature

- High sampling temperature (τ) generally increase diversity while sacrificing the quality.
- Sampling with high temperature and self-guidance to enhance both the quality and diversity of generated samples.



Effect of Self-Guidance

- Spatial smoothing (Gaussian Blur) at pixel level vs Semantic smoothing at latent level

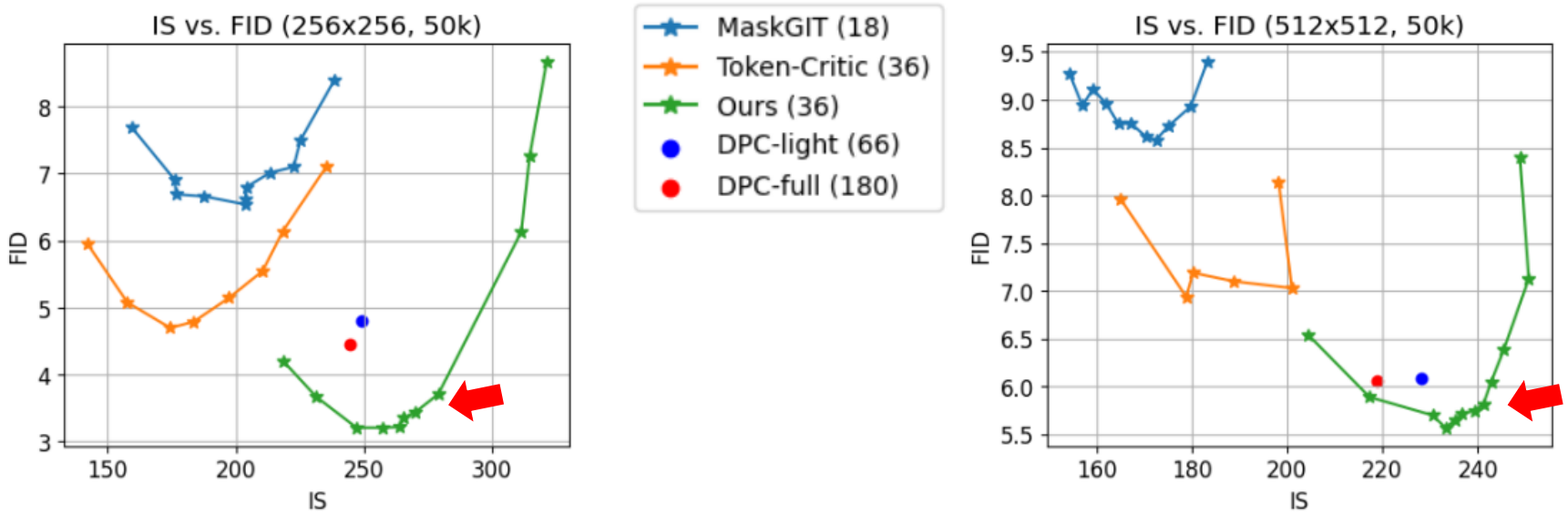


Qualitative Results

- Comparison of sampled images using 18-step MaskGIT without (top) and with the proposed self-guidance (bottom).

Quantitative Results

- Comparison with various MGMs. All methods shares same baseline generator (MaskGIT) and same sampling timestep (T=18).
- The proposed method shows outperforming quality-diversity trade-off compared to various MGMs.



Quantitative Results

- Comparison with various generative models with similar model size.

Model	Type	NFE	ImageNet 256×256				ImageNet 512×512			
			FID↓	IS↑	Prec↑	Rec↑	FID↓	IS↑	Prec↑	Rec↑
BigGAN-deep [3]	GANs	1	6.95	224.5	**0.89**	0.38	8.43	177.9	0.85	0.25
GigaGAN [22]	GANs	1	3.45	225.5	0.84	0.61	–	–	–	–
ADM [9]	Diff.	250	10.94	101.0	0.69	**0.63**	23.24	58.0	0.73	**0.60**
ADM (+ SAG) [19]	Diff.	500	9.41	104.7	0.70	0.62	–	–	–	–
CDM [18]	Diff.	250	4.88	158.7	–	–	–	–	–	–
LDM-4 [34]	Diff.	250	10.56	103.4	0.71	0.62	–	–	–	–
LDM-4 (+ CFG) [34]	Diff.	500	3.60	247.7	–	–	–	–	–	–
DiT-L/2‡ (+ CFG) [31]	Diff.	500	5.02	167.2	0.75	0.57	–	–	–	–
VQVAE-2† [33]	AR	5120	31.11	∼45	0.36	0.57	–	–	–	–
VQGAN† [10]	AR	∼1024	18.65	80.4	0.78	0.26	7.32	66.8	0.73	0.31
VQ-Diffusion [14]	Discrete.	100	11.89	–	–	–	–	–	–	–
ImprovedVQ. (+ CFG) [40]	Discrete.	200	4.83	–	–	–	–	–	–	–
MaskGIT* [4]	Mask.	18	6.56	203.6	0.79	0.48	8.48	167.1	0.78	0.46
Token-Critic [25]	Mask.	36	4.69	174.5	0.76	0.53	6.80	182.1	0.73	0.50
DPC-full [26]	Mask.	180	4.45	244.8	0.78	0.52	6.06	218.9	0.80	0.47
Ours (T=12)	Mask.	24	3.35	259.7	0.81	0.52	**5.38**	226.0	**0.88**	0.36
Ours (T=18)	Mask.	36	**3.22**	**263.9**	0.82	0.51	5.57	**233.2**	**0.88**	0.35

Paper (Arxiv)

Github

Thank you!

KAIST

SIIL Statistical Inference and Information Theory Laboratory