Supplementary Material for Masked Diffusion Transformer is a Strong Image Synthesizer

Shanghua Gao[1,2] * Pan Zhou[2] † Ming-Ming Cheng[1] † Shuicheng Yan[2]
[1]Nankai University [2]Sea AI Lab
{shanghuagao,shuicheng.yan}@gmail.com zhoupan@sea.com cmm@nankai.edu.cn

A. Model Details

Network configurations. We follow the network configurations described in DiT [3] to set the total block number (*i.e.* $N_1 + N_2$), token number, and channel numbers for the masked diffusion transformer of MDT. The configurations of MDT models are given in Tab. 1. Following DiT, the MDT has models with different sizes, denoted by S/B/XL.

Network parameters and costs. The network parameters and training costs for MDT under different model scales are listed in Tab. 1. In comparison to DiT baselines, MDT introduces a negligible extra inference parameters and costs.

Size	Layers	Dim.	Head Num.	Param. (M)	FLOSs (G)
Network configurations of MDT models.					
S	12	384	6	33.1	6.07
B	12	768	12	130.8	23.02
XL	28	1152	16	675.8	118.69
Network configurations of DiT baselines.					
S	12	384	6	32.9	6.06
B	12	768	12	130.3	23.01
XL	28	1152	16	674.8	118.64

Table 1. Network configurations of MDT models. The configurations are following DiT networks [3]. The layers consist of the layer numbers of encoder and decoder, and the decoder number N_2 is set to 2 for all models. FLOSs are measured with the latent embedding size of 32×32 and p=2. The parameters and FLOSs are measured using the inference model.

B. Comparison of VAE decoders

To ensure fair comparisons with DiT [3], we use both the MSE and EMA versions of pretrained VAE decoders[1] for image sampling. Tab. 2 shows that the EMA version has slightly better performance than the MSE version. Except for the results in Table 1 of the manuscript that uses the EMA VAE decoder, we use the MSE VAE decoder by default.

Method	Decoder	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
MDT	MSE	6.65	5.07	129.47	0.72	0.63
MDT	EMA	6.46	4.92	131.70	0.72	0.63
MDT-G	MSE	2.14	4.45	259.21	0.82	0.59
MDT-G	EMA	2.02	4.46	263.77	0.82	0.60

Table 2. Comparison between the EMA and MSE version of VAE decoders. -G denotes the results with classifier-free guidance.

C. Inpainting with MDT

We verify the inpainting ability of MDT by filling the masked tokens with the side-interpolater at the first step and then conducting denoise diffusion process on masked tokens. As shown in Fig. 1, we utilize different mask ratios on the image and inpaint the masked parts with MDT. Although the MDT model is trained with the mask ratio of 30%, it can easily handle much larger masking ratios, such as 70% mask ratio. We attribute this ability to the combination of our proposed mask latent modeling and the diffusion model.

D. Improved Classifier-free Guidance

The classifier-free guidance sampling [2] enables the trade-off between sample quality and diversity. It achieves this by combining the class-conditional and unconditional estimation:

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t) + w \cdot (\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t)),$$

where $\epsilon_\theta(x_t, c)$ is the class-conditional estimation, $\epsilon_\theta(x_t)$ is the unconditional estimation, and w is the guidance scale. Generally, a larger w results in high sample quality by decreasing the diversity. MUSE [1] changes the fixed guidance scale with a linear increasing schedule during sampling,

*This work was done while S. Gao was a research intern at Sea AI Lab.
†Pan Zhou and Ming-Ming Cheng are joint corresponding authors.
[1]MSE and EMA versions of VAE models are downloaded in https://huggingface.co/stabilityai/sd-vae-ft-mse and https://huggingface.co/stabilityai/sd-vae-ft-ema.

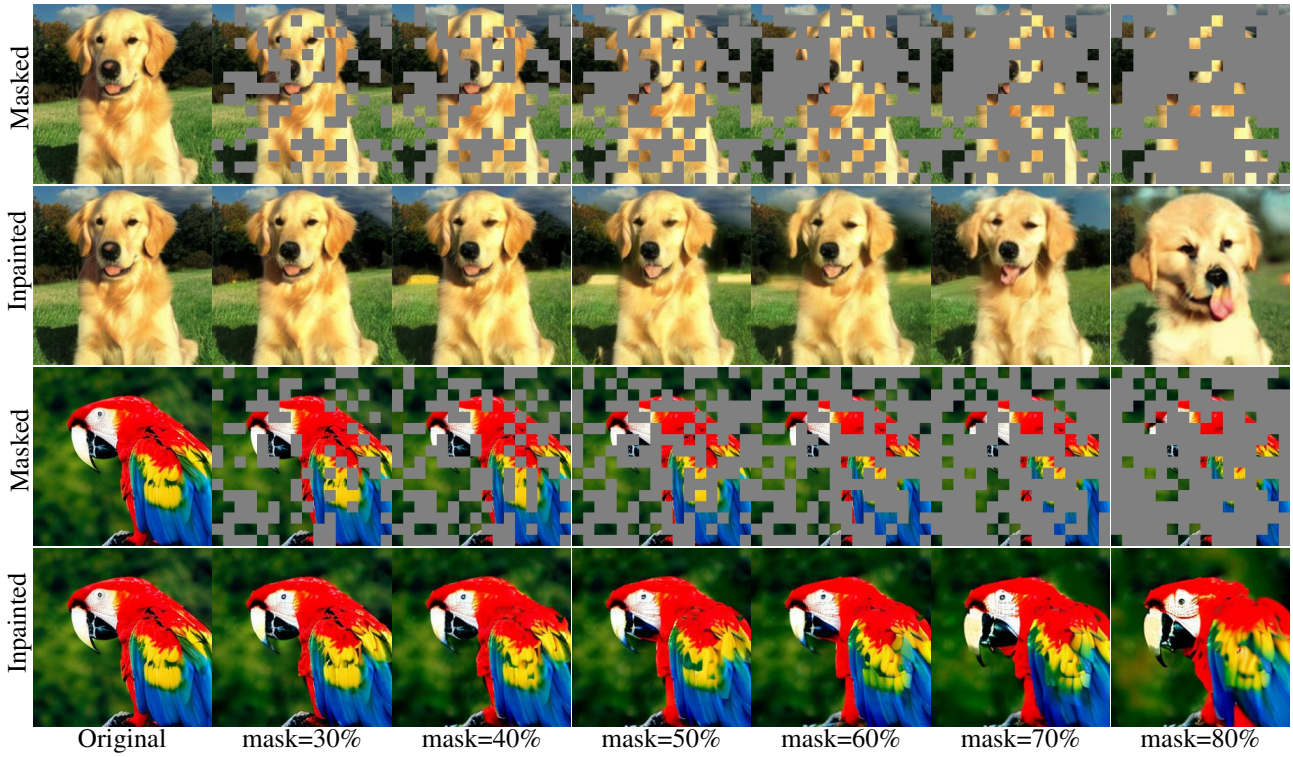


Figure 1. Inpainting under different masked ratios using MDT-XL/2.

which makes the model samples with more diversity at early steps while samples with higher fidelity at late steps. Inspired by this, we present a power-cosine schedule for the guidance scale during the sampling procedure:

$$w_t = \frac{1 - \cos \pi \left(\frac{t}{t_{\max}}\right)^s}{2} w,$$

where t is the time step during sampling, $t_{\max}$ is the maximum sampling step, w is the maximum guidance scale, and s is a factor that controls the increasing speed of the guidance scale. As revealed in Fig. 2, the power-cosine schedule enables a low guidance scale at early steps while quickly increasing the guidance scale at late steps. By increasing s, the guidance scale has a slow increase at early steps and a fast increase at late steps. The improved classifier-free guidance sampling equipped with the power-cosine guidance scale schedule enables the model samples with high diversity at early steps and high quality at late steps. In this work, s is set to 4, and the corresponding w is set to 3.8 to ensure the model generates images with high fidelity at late steps.

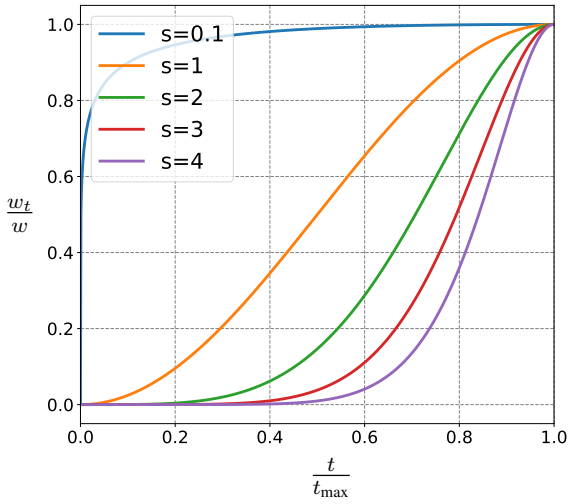


Figure 2. The power-cosine scaling schedule for guidance scale in classifier-free guidance with difference s. A larger s results in a slower increase of w at early steps and a faster increase at late steps.

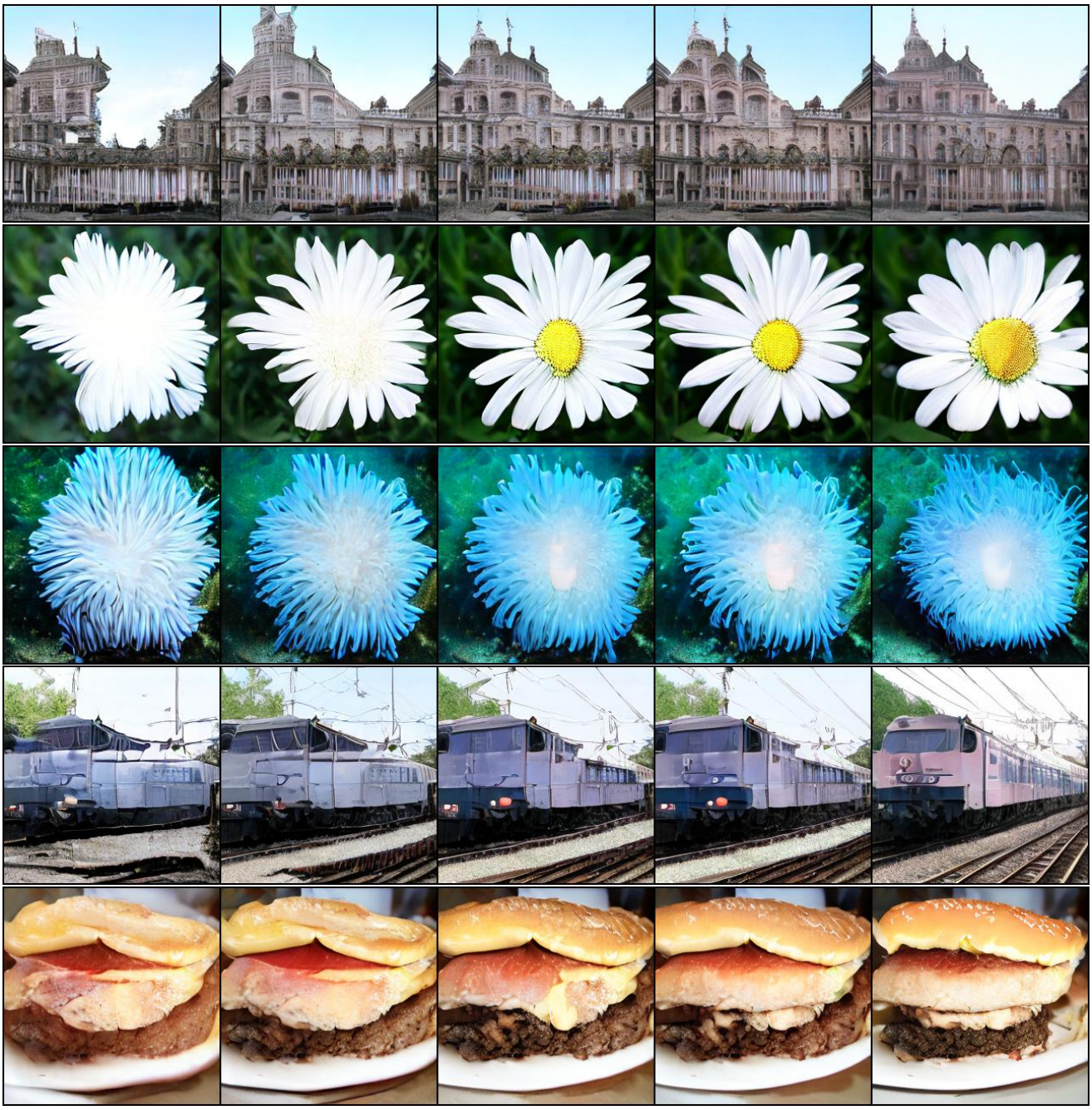
E. Visualization

We provide more visualized examples of MDT-XL/2 generated images in Fig. 3. In Fig. 4, we show more visualized examples of MDT-S/2 along with training steps.

References

[1] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy,

Figure 3. Visualization of images generated by the MDT-XL/2.



50k 100k 200k 300k 3000k

Figure 4. Visualized example of MDT-S/2 along with training steps.

William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 1

[2] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS Workshop*, 2021. 1

[3] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 1