Supplemental Material: Activation-Norm Maximization to Accelerate Training in Flow-Matching Transformers

Yash Belhe Wesley Chang Tzu-Mao Li Ravi Ramamoorthi Michaël Gharbi
University of California, San Diego Reve

Hyperparameters and configurations. For all our experiments, we strictly follow the exact same experimental settings from the original SiT [1] codebase. Importantly, we do not tune any hyperparameters to favor our method, except for the extra ones α and λ introduced by our method. See Tab. 1 for a consolidated summary of model and training hyperparameters.

SDE-sampling-based FID comparison with baseline. With SDE sampling for the XL/2 model, our method with (cfg=1.42) achieves FID=1.95 after just 2.5M steps, which surpasses the FID=2.06 of the baseline at 7M steps.

Flow-matching loss. Not only does our method improve the FID for a set of generated images, but it also improves its actual training target, the flow-matching loss; see Tab. 2.

Effect of α and λ. We sweep α Tab. 3 and λ Tab. 4 for SiT-B/4 at 200k steps with a 2k-step warm-up. For λ, we find that for $\lambda \geq 10$ works well and for α, all values we tested $300 \geq \alpha \geq 10$ work well for the B/4 model, making our method quite robust to its hyperparameters.

Example images from sinusoid image dataset. We visualize sample images from the sinusoid dataset in Fig. 1.

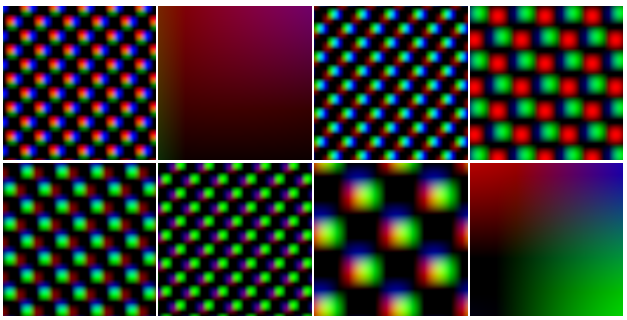

Figure 1. Example images from the sinusoid dataset used in our dataset ablation.

References

[1] Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. 2024.

Table 1. SiT configurations on ImageNet-256. B/4 is our ablation model.

model	S/2	B/4	B/2	L/2	XL/2
model configurations					
params (M)	33	130	130	458	675
depth	12	12	12	24	28
hidden dim	384	768	768	1024	1152
patch size	2×2	4×4	2×2	2×2	2×2
heads	6	12	12	16	16
training configurations					
steps	400k	400k	400k	400k	400k–2.5M
batch size			256		
optimizer			Adam		
optimizer β_1			0.9		
optimizer β_2			0.999		
weight decay			0.0		
learning rate (lr)			1×10^{-4}		
lr schedule			constant		
lr warmup			none		
ema decay			0.9999		
ODE sampling					
steps			250		
sampler			Heun		
t schedule			linear		
last step size			N/A		
SDE sampling					
steps			250		
sampler			Euler		
t schedule			linear		
last step size			0.04		
Activation-Norm Maximization					
α			70.0		
λ			10.0		

Table 2. **Flow-matching loss (XL/2 at 400k).** Averaged over the ImageNet dataset with time sampled uniformly at random.

Model	Steps	Loss↓
SiT-XL/2 (Baseline)	400k	0.7529
SiT-XL/2 (Ours)	**400k**	**0.7456**

Table 3. **Effect of** α on SiT-B/4 at 200k steps (holding λ=10, 2k steps warm-up).

Model	Steps	α	FID↓
SiT-B/4 (Baseline)	200k	-	70.61
SiT-B/4 (Ours)	**200k**	**10**	**62.03**
SiT-B/4 (Ours)	**200k**	**50**	**61.75**
SiT-B/4 (Ours)	**200k**	**70**	**62.85**
SiT-B/4 (Ours)	**200k**	**100**	**63.37**
SiT-B/4 (Ours)	**200k**	**300**	**62.39**

Table 4. **Effect of** λ on SiT-B/4 at 200k steps (holding α=70, 2k steps warm-up).

Model	Steps	λ	FID↓
SiT-B/4 (Baseline)	200k	-	70.61
SiT-B/4 (Ours)	200k	2.0	66.06
SiT-B/4 (Ours)	200k	5.0	65.49
SiT-B/4 (Ours)	**200k**	**10.0**	**62.85**
SiT-B/4 (Ours)	**200k**	**15.0**	**62.45**
SiT-B/4 (Ours)	**200k**	**20.0**	**62.98**