Improved Training Technique for Shortcut Models

Qualcomm

NeurIPS 2025

Anh Nguyen* Viet Nguyen* Duc Vu Trung Dao Chi Tran Toan Tran Anh Tran

All authors are affiliated with Qualcomm AI Research. Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Improved Shortcut Models

- Shortcut models represent a **promising, non-adversarial** paradigm for generative modeling, uniquely supporting **one-step, few-step, and multi-step** sampling from a single trained network.
- However, their widespread adoption has been **stymied** by critical **performance bottlenecks**.

🚀 This paper **FIRST tackle FIVE core issues** that held shortcut models back!

1. Compounding guidance
2. Inflexible fixed guidance
3. Curvy flow trajectories
4. Frequency bias
5. Divergent self-consistency

Method	FID$_{N=1}$ ↓	FID$_{N=4}$ ↓
Shortcut Models [20]	21.38	13.46
Improved Shortcut Models (iSM)		
+ Intrinsic Guidance	9.62	3.17
+ Interval Guidance in Training	8.49	2.81
+ Multi-level Wavelet Function	8.12	2.64
+ Scaling Optimal Transport	7.97	2.23
+ Twin EMA	6.56	2.16

Our method achieves **state-of-the-art FID scores**, making shortcut models a **viable** class of generative models

Our method achieves **state-of-the-art FID scores**, making shortcut models a **viable** class of generative models capable of **one-step, few-step, and multi-step sampling**

One-to-Many Step Models			
iCT [58]	34.24	1	675M
	20.3	2	675M
SM [20]	10.60	1	675M
	7.80	4	675M
	3.80	128	675M
IMM [73]	7.77	1	675M
	3.99	2	675M
	2.51	4	675M
	1.99	8	675M
iSM (ours)	5.27	1	675M
	2.44	2	675M
	2.05	4	675M
	1.93	8	675M
	1.88	128	675M

1. **Intrinsic Guidance** conditions the network on explicit scale to enable dynamic inference control. This resolves **inflexible fixed guidance** and mathematically corrects the **compounding guidance** flaw by preventing exponential signal amplification.

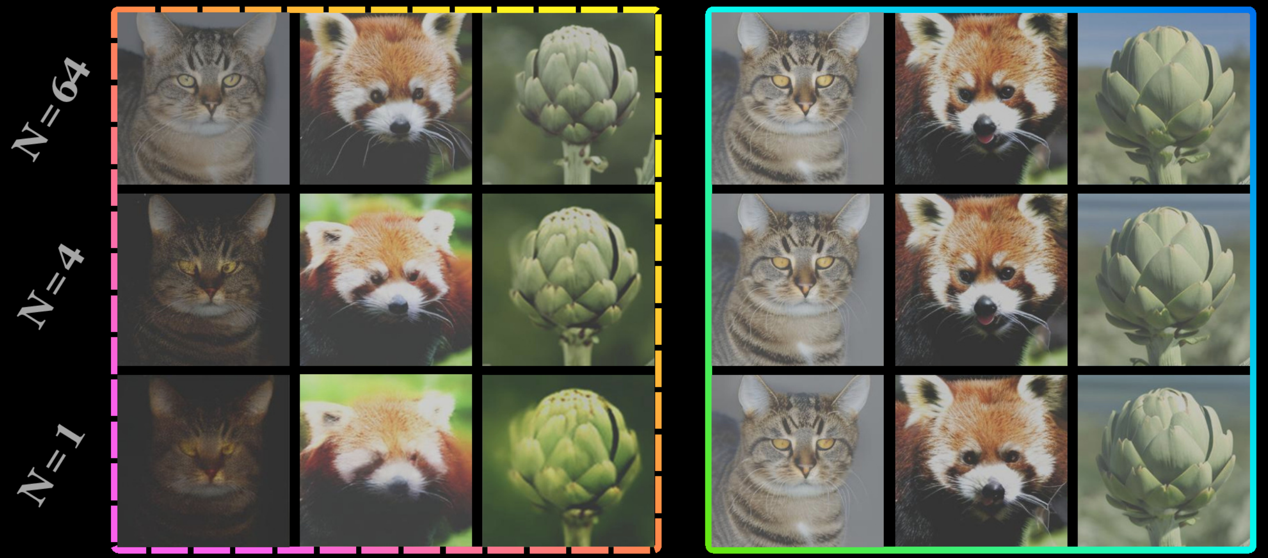
The Hidden Flaw of Compounding Guidance

Proposition 1. *The model's prediction for a single large shortcut step of size $Nd = 1$ approximately equals the average of the guided displacements corresponding to the N smallest steps, but with an exponentially compounded guidance scale:*

$$s_\theta(x_0, 0, c, Nd) \approx \frac{1}{N}\sum_{i=0}^{N-1} g_\theta^{w^{\log_2(N)}}\left(x'_{\frac{i}{N}}, \frac{i}{N}, c, d\right). \tag{6}$$

Proof. See Appendix A. □

Original Shortcut Model | **Improved Shortcut Models**



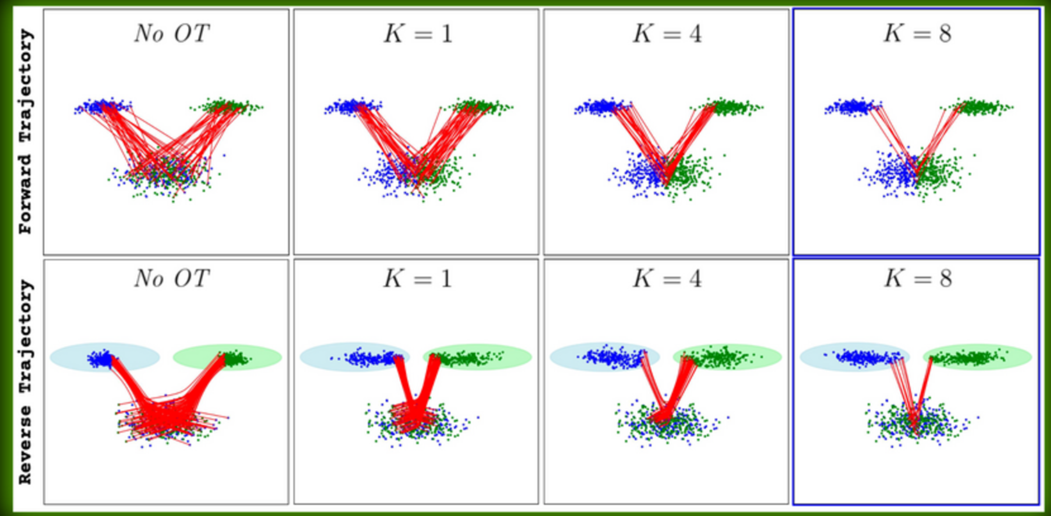
Guided Self-Consistency Objective. This objective generalizes the self-consistency principle from [20] to operate with arbitrary step sizes ($d > 0$) and any guidance scale ($w \geqslant 0$). The objective maintains the foundational properties of shortcut models, where a *single, large guided shortcut step* yields an output consistent with the composition of *two smaller, consecutive guided steps*.

$$\mathcal{L}_{\text{consistency}}(\theta) := \mathbb{E}_{\substack{x_0\sim\mathcal{N},(x_1,c)\sim D\\(t,w,d)\sim p(t,w,d)}}\left[\|s_\theta(x_t,t,c,2d,w) - s_{\text{consistency}}\|^2\right],$$
$$\text{where}\quad s_{\text{consistency}} := s_{\theta^-}(x_t,t,c,d,w)/2 + s_{\theta^-}(x'_{t+d},t,c,d,w)/2 \tag{10}$$
$$\text{and}\quad x'_{t+d} = x_t + s_\theta(x_t,t,c,d,w)d,$$

where θ^- is the EMA target network. The stop-gradient operator $\mathrm{sg}(\cdot)$ is applied to the entire consistency target to stabilize training, following standard practice for self-consistency objectives.

2. **Twin EMA** maintains a fast-decay network for fresh targets and a slow-decay network for inference. This resolves **divergent self-consistency** by eliminating the temporal lag that causes conflicts between training stability and target currency.

3. **Scaling Optimal Transport (sOT)** aggregates mini-batches to compute a global transport plan. This disentangles noise-data couplings to straighten **curvy flow trajectories**, minimizing the training variance caused by intersecting paths.



4. **Multi-Level Wavelet Function** utilizes DWT to enforce a frequency-aware error signal. This mitigates the **frequency bias** inherent in pixel-wise losses by explicitly supervising the reconstruction of neglected high-frequency details.

Multi-Level Wavelet Function