Improved Training Technique for Shortcut Models

NeurIPS 2025

Anh Nguyen*

Viet Nguyen*

Duc Vu

Trung Dao

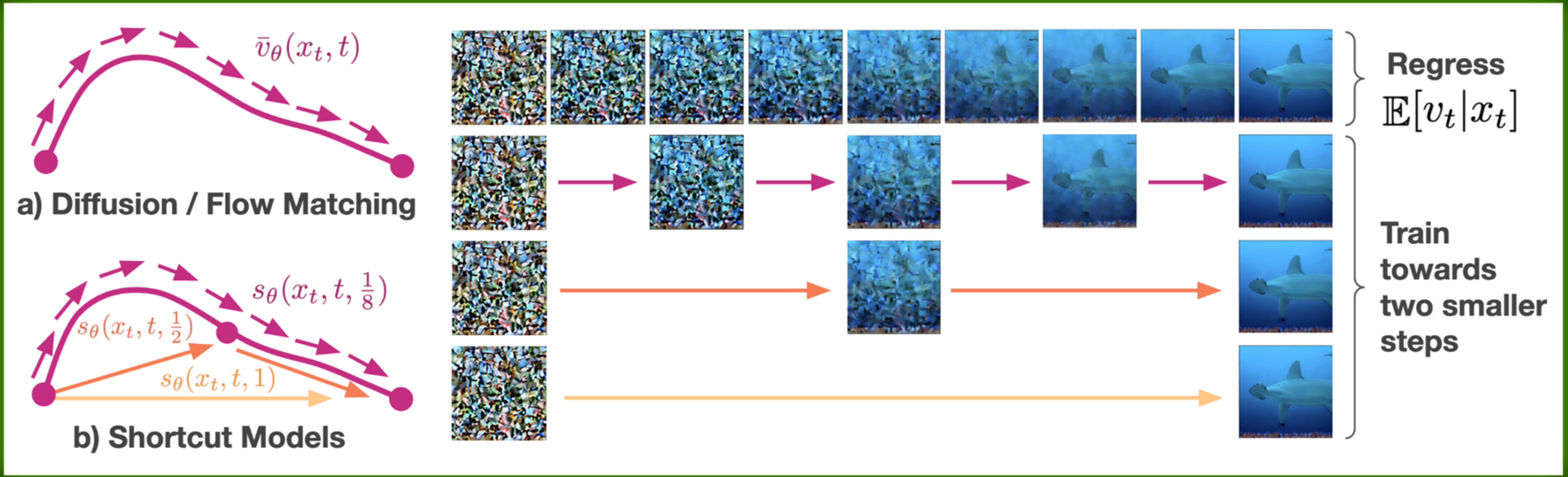
Chi Tran

Toan Tran

Anh Tran

* Equal Contribution

Shortcut Models: Promising and Unique Class of Generative Models



This is unique:

1. A single network, trained **end-to-end**, that can "jump" or "shortcut" the generation process.
2. This design inherently supports **one-step, few-step, and many-step** generation.

Shortcut Models: An Elegant Idea. Hampered by a Hidden Flaw

This paper **FIRST** tackle
the **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

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

Method	**FID** $_{N=1}\downarrow$	**FID** $_{N=4}\downarrow$
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

Component 1: Fixing Artifacts with Intrinsic Guidance

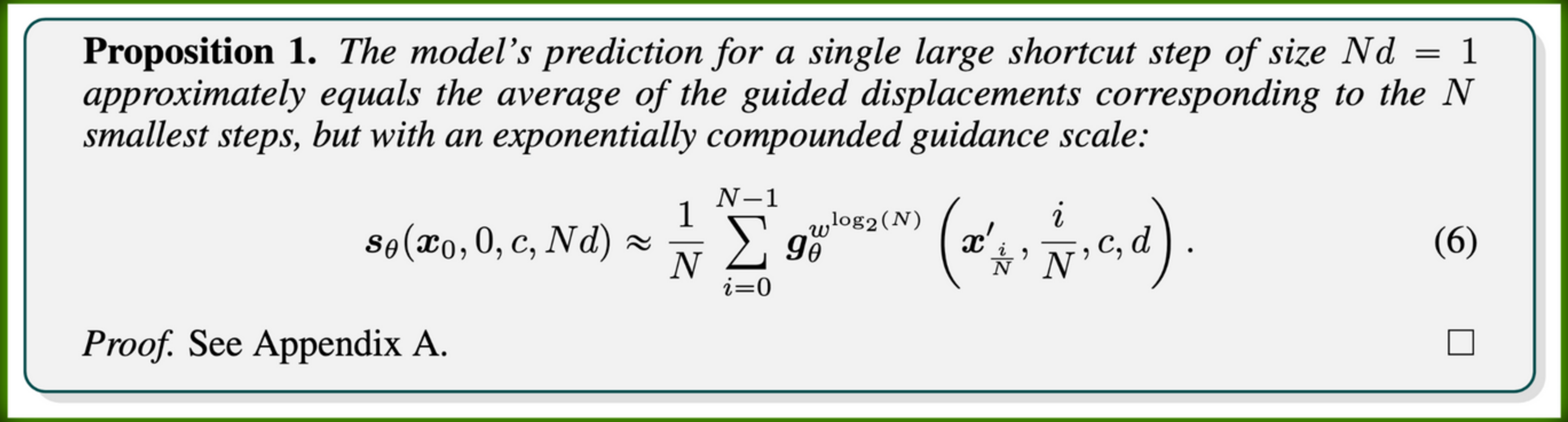
Observation:

1. **Hidden flaw** of compounding guidance, which we are the first to formalize, causing severe image artifacts.
2. **Inflexible fixed guidance** that restricts inference-time control.

Solution: We introduce Intrinsic Guidance, making the guidance scale w an explicit input to the model.

Result: Provides explicit, dynamic control resolving both the compounding guidance flaw and the inflexibility of prior models.

Original Shortcut Model



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:*

$$\boldsymbol{s}_\theta(\boldsymbol{x}_0, 0, c, Nd) \approx \frac{1}{N} \sum_{i=0}^{N-1} \boldsymbol{g}_\theta^{w^{\log_2(N)}} \left(\boldsymbol{x}'_{\frac{i}{N}}, \frac{i}{N}, c, d \right). \tag{6}$$

Proof. See Appendix A. □

*The hidden flaw of **compounding guidance***

Component 1: Fixing Artifacts with Intrinsic Guidance

Observation:

1. **Hidden flaw** of compounding guidance, which we are the first to formalize, causing severe image artifacts.
2. **Inflexible fixed guidance** that restricts inference-time control.

Solution: We introduce Intrinsic Guidance, making the guidance scale w an explicit input to the model.

Result: Provides explicit, dynamic control resolving both the compounding guidance flaw and the inflexibility of prior models.

Original Shortcut Model

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*.

$$\begin{aligned} \mathcal{L}_{\text{consistency}}(\theta) &:= \mathbb{E}_{\substack{\boldsymbol{x}_0\sim\mathcal{N},\,(\boldsymbol{x}_1,c)\sim D\\(t,w,d)\sim p(t,w,d)}}\left[\left\|\boldsymbol{s}_\theta(x_t,t,c,2d,w) - s_{\text{consistency}}\right\|^2\right], \\ \text{where} \quad s_{\text{consistency}} &:= s_{\theta^-}(x_t,t,c,d,w)/2 + s_{\theta^-}(x'_{t+d},t,c,d,w)/2 \\ \text{and} \quad x'_{t+d} &= x_t + s_\theta(x_t,t,c,d,w)d, \end{aligned} \tag{10}$$

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.

Intrinsic Guidance *Training for Shortcut Models*

Component 1: Fixing Artifacts with Intrinsic Guidance

Observation:

1. **Hidden flaw** of compounding guidance, which we are the first to formalize, causing severe image artifacts.
2. **Inflexible fixed guidance** that restricts inference-time control.

Solution: We introduce Intrinsic Guidance, making the guidance scale w an explicit input to the model.

Result: Provides explicit, dynamic control resolving both the compounding guidance flaw and the inflexibility of prior models.

Original Shortcut Model

N = 64

N = 4

N = 1

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*.

$$\begin{aligned} \mathcal{L}_{\text{consistency}}(\theta) &:= \mathbb{E}_{\substack{\boldsymbol{x}_0 \sim \mathcal{N},\, (\boldsymbol{x}_1, c) \sim D \\ (t,w,d) \sim p(t,w,d)}} \left[\| \boldsymbol{s}_\theta(x_t, t, c, 2d, w) - s_{\text{consistency}} \|^2 \right], \\ \text{where} \quad s_{\text{consistency}} &:= s_{\theta^-}(x_t, t, c, d, w)/2 + s_{\theta^-}(x'_{t+d}, t, c, d, w)/2 \\ \text{and} \quad x'_{t+d} &= x_t + s_\theta(x_t, t, c, d, w)d, \end{aligned} \tag{10}$$

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.

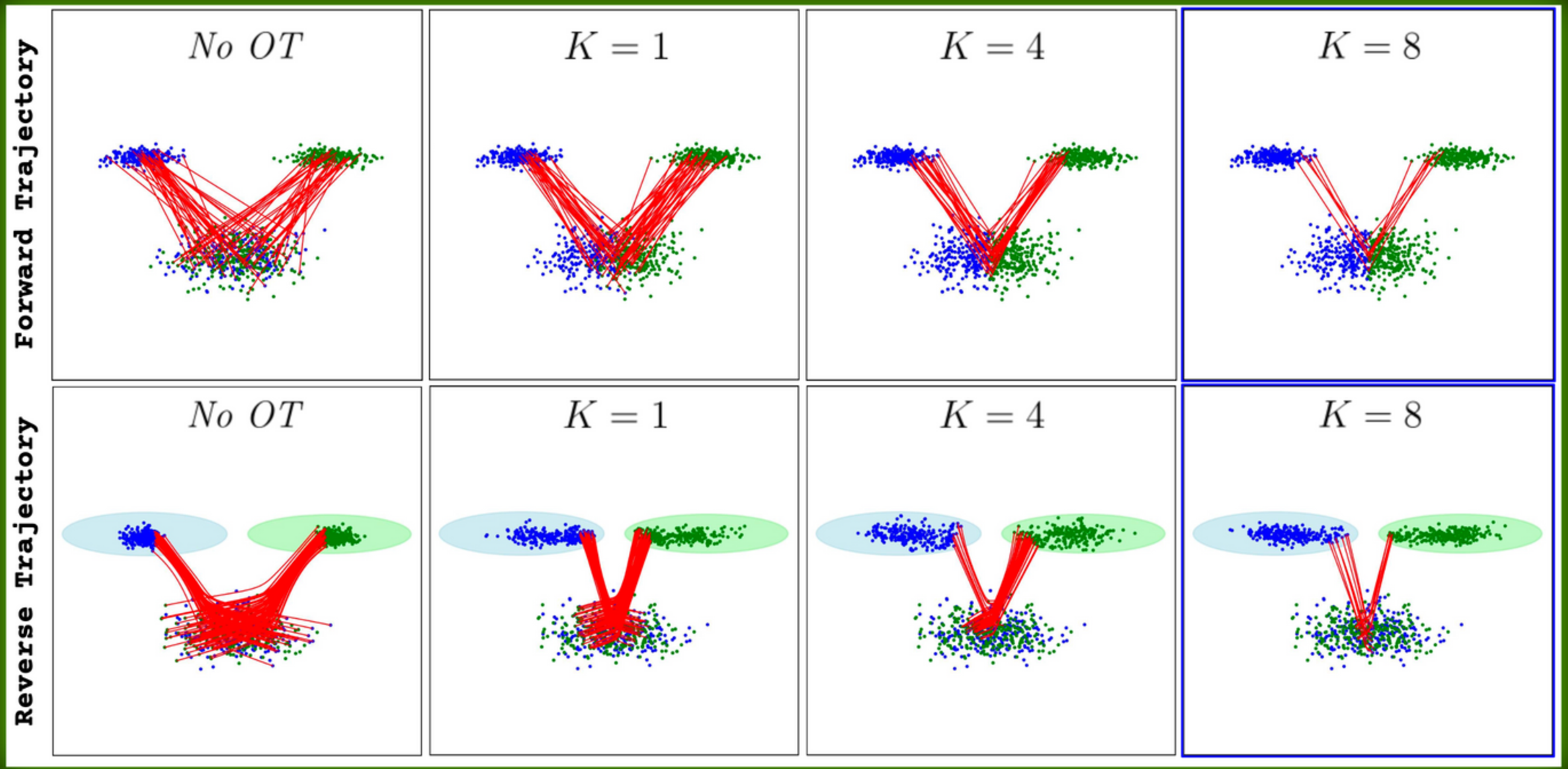
Intrinsic Guidance Training for Shortcut Models

Intrinsic Guidance

Improved Shortcut Models

Component 2: Straightening the Curvy Generative Trajectories with sOT

Observation: Conventional random pairing of noise and data results in **high-curvature** generative paths.



***Efficacy of sOT** in improving shortcut model training, demonstrated by **varying the OT scaling factor K**.*

On a bimodal target, forward trajectories (top row) without OT exhibit **frequent intersections** (**red**), compelling the reverse generative process (bottom row) to follow **high-curvature paths** that initially average the target modes (**blue**, **green**).

Our sOT method, by increasing the effective OT scaling K, **progressively disentangles these forward couplings**, yielding substantially **straighter reverse trajectories**.

Solution: We introduce **Scaling Optimal Transport (sOT)** to straighten generative trajectories by periodically computing a large-scale transport plan.

Result: This yields straighter generative paths.

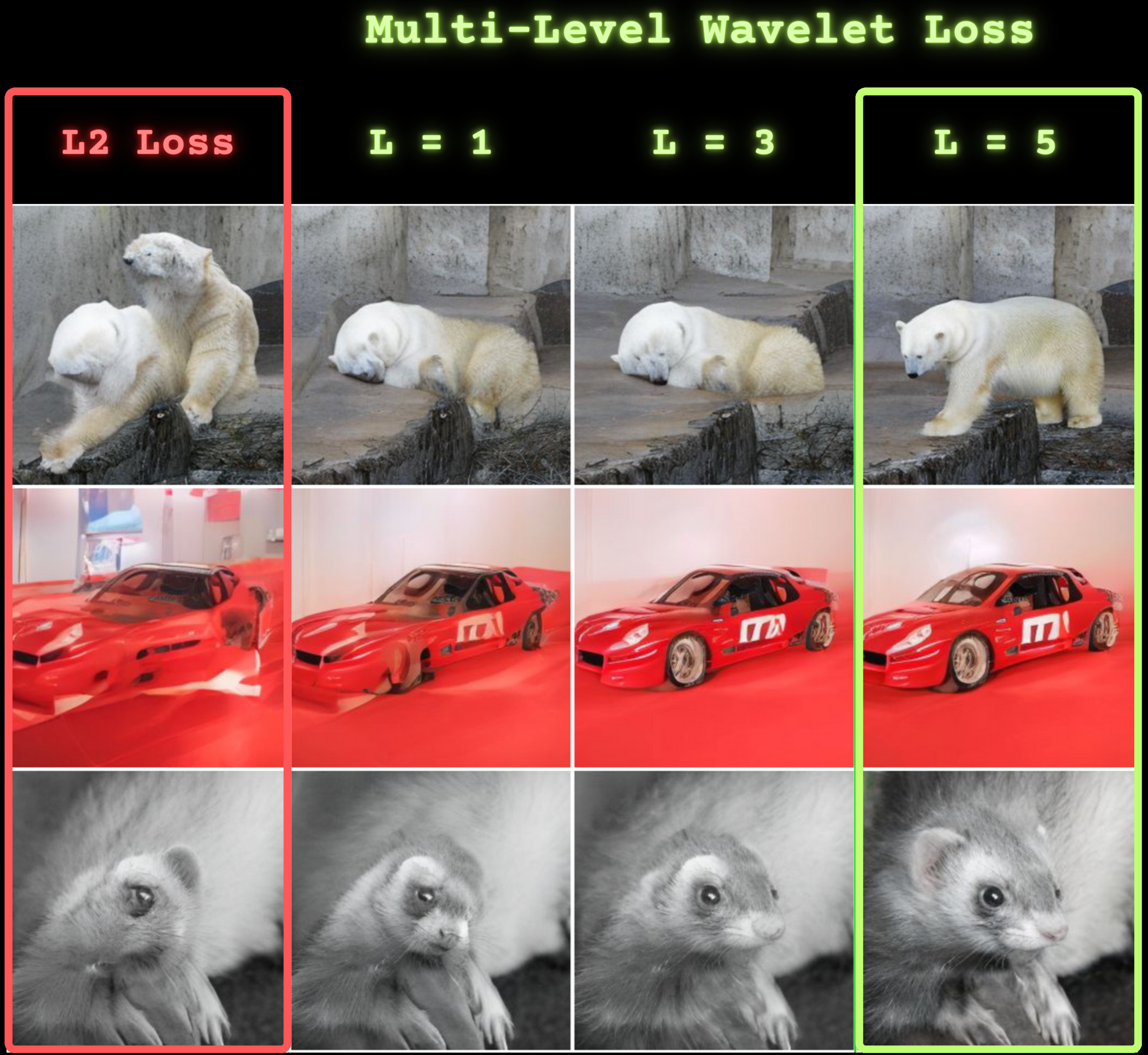
Component 3: Against Frequency Bias by Multi-Level Wavelet Function

Observation: Standard L2 loss causes a "frequency bias," prioritizing low-frequency structures and resulting in blurry textures.

*L is the number of recursive decomposition levels in the **Discrete Wavelet Transform**.*

Solution: We replace L2 with a Multi-Level Wavelet Loss, forcing the model to reconstruct details across the entire frequency spectrum.

Result: Sharper, more detailed images, especially in the challenging few-step setting.



Component 4: Resolving Divergent Self-Consistency with Twin EMA

Observation: The self-consistency objective was conflicting, as the model tried to match targets generated by an outdated, slow-moving version of itself (the EMA network).

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*.

$$\begin{aligned} \mathcal{L}_{\text{consistency}}(\theta) &:= \mathbb{E}_{\substack{\boldsymbol{x}_0\sim\mathcal{N},\,(\boldsymbol{x}_1,c)\sim D\\(t,w,d)\sim p(t,w,d)}}\left[\left\|\boldsymbol{s}_\theta(x_t,t,c,2d,w) - s_{\text{consistency}}\right\|^2\right], \\ \text{where} \quad s_{\text{consistency}} &:= s_{\theta^-}(x_t,t,c,d,w)/2 + s_{\theta^-}(x'_{t+d},t,c,d,w)/2 \\ \text{and} \quad x'_{t+d} &= x_t + s_\theta(x_t,t,c,d,w)d, \end{aligned} \tag{10}$$

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.

Solution: We use a Twin EMA strategy -- a fast-decay EMA for up-to-date training targets and a slow-decay EMA for stable inference.

Result: Resolves the training conflict, leading to faster convergence and better final performance.

Conclusion: Unique and Competitive Class of Generative Models

This paper **FIRST** tackle
the **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

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**.