Guiding a Diffusion Model with Itself using Sliding Windows

Nikolas Adaloglou, Tim Kaiser, Damir Iagudin, Markus Kollmann

Heinrich-Heine-Universität Düsseldorf, Germany

{adaloglou,tikai103,markus.kollmann@hhu.de}

Source Code

BMVC 2025

arXiv

Overview

Diffusion model (DM) guidance is a sampling technique that employs an auxiliary "negative" model $\boldsymbol{\epsilon}_{\text{neg}}(.)$ that generalizes more broadly than the positive model $\boldsymbol{\epsilon}_{\text{pos}}(.)$, realized as:

$$\tilde{\boldsymbol{\epsilon}}(\boldsymbol{x}, t) = \boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) + w[\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) - \boldsymbol{\epsilon}_{\text{neg}}(\boldsymbol{x}, t)].$$

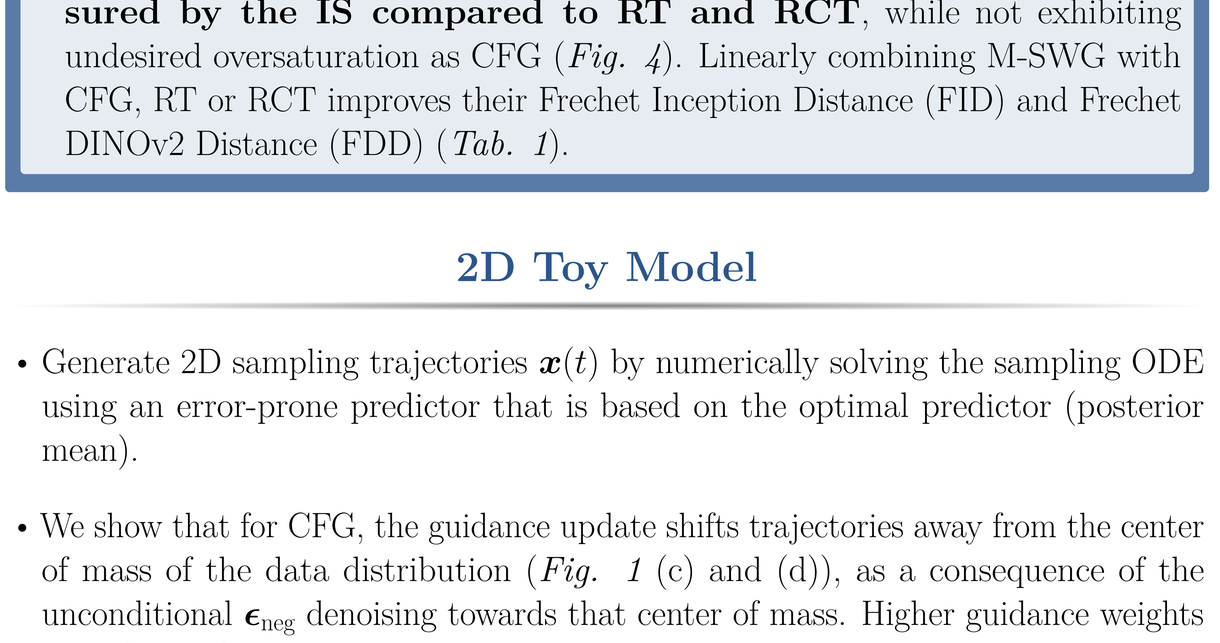
Baselines: Typical choices for $\boldsymbol{\epsilon}_{\text{neg}}(.)$ include: an unconditional model as in classifier-free guidance (*CFG*), a model of reduced capacity (*RC*), a previous checkpoint with reduced training (*RT*), or a model of reduced capacity and training (*RCT*), also known as "AutoGuidance". CFG requires joint training of the unconditional model, while RC and RCT require training a separate DM.

Main Contributions

- We introduce a 2D Toy Model to show that (i) Weak model guidance (*WMG*) methods sample closer to high likelihood regions than CFG and (ii) **guidance works best if $\boldsymbol{\epsilon}_{\text{neg}}$ makes similar errors to $\boldsymbol{\epsilon}_{\text{pos}}$ but stronger** (*Fig. 1*).
- We introduce **masked sliding window guidance (M-SWG)**, a novel guidance method designed to upweight long-range spatial dependencies (*Figs. 2, 3*). M-SWG **requires neither training nor class conditioning** and can be applied to any DM that can process multiple image resolutions.
- We show that **M-SWG achieves superior perceptual quality as measured by the IS compared to RT and RCT**, while not exhibiting undesired oversaturation as CFG (*Fig. 4*). Linearly combining M-SWG with CFG, RT or RCT improves their Frechet Inception Distance (FID) and Frechet DINOv2 Distance (FDD) (*Tab. 1*).

2D Toy Model

- Generate 2D sampling trajectories $\boldsymbol{x}(t)$ by numerically solving the sampling ODE using an error-prone predictor that is based on the optimal predictor (posterior mean).
- We show that for CFG, the guidance update shifts trajectories away from the center of mass of the data distribution (*Fig. 1* (c) and (d)), as a consequence of the unconditional $\boldsymbol{\epsilon}_{\text{neg}}$ denoising towards that center of mass. Higher guidance weights amplify this bias.
- In contrast, WMG drives trajectories closer to the data distribution. The crucial condition for this is that, in each step, $\boldsymbol{\epsilon}_{\text{neg}}$ exhibits a similar error as $\boldsymbol{\epsilon}_{\text{pos}}$, but stronger, causing the extrapolation to correct the prediction of $\boldsymbol{\epsilon}_{\text{pos}}$ towards the optimal prediction of $\boldsymbol{\epsilon}^*$ (*Fig. 1* (d)).



(a) $w = 0$ (b) $w = w^*$ (c) $w = 3w^*$ (d) Guidance step

Fig. 1: Inference trajectories $\boldsymbol{x}(t)$ of our 2D Toy Model. (a) No guidance, trajectories of an error-prone predictor. **(b)** Guidance with a well-performing weight w^*. **(c)** Guidance with a large weight. **(d)** Enlarged views of single trajectories. Arrows show target prediction. The guidance correction (black arrow) pushes WMG closer to the data point, unlike CFG.

Method: M-SWG

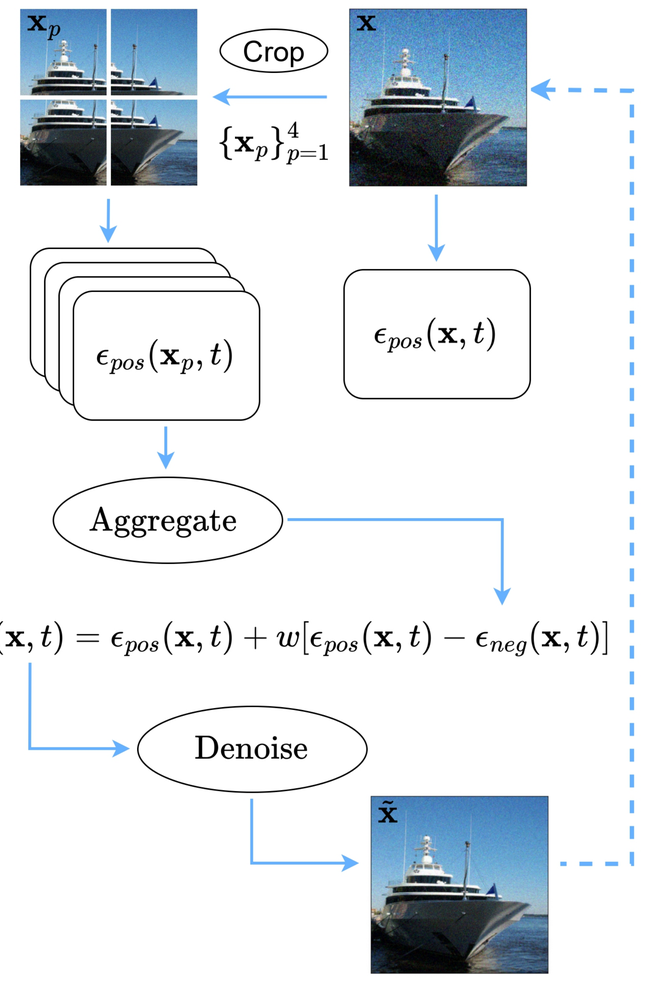


Fig. 2: SWG. Given an input size $H \times H$, we use sliding windows to generate N crops of size $k \times k$, using a fixed stride s per dimension. These crops are independently processed by $\boldsymbol{\epsilon}_{\text{pos}}$. The N predictions of $\boldsymbol{\epsilon}_{\text{pos}}$ are superimposed in the same order and at the same positions. Overlapping pixels are averaged. We use a crop size $k \approx 5/8 \times H$, and $N = 4$. For latent diffusion, SWG is performed in each denoising step in the latent space (i.e. 64×64), and the final output is passed to the decoder.

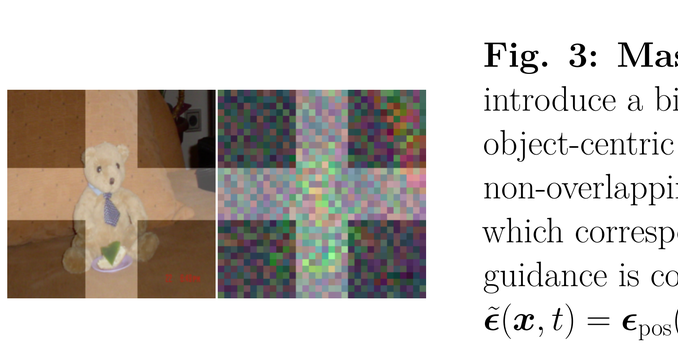


Fig. 3: Masked SWG (M-SWG). Thus, we introduce a binary mask $\mathbf{M} \in \{0, 1\}^{H \times H}$, for object-centric datasets, to exclude the non-overlapping regions when applying SWG, which correspond to the corners. The updated guidance is computed as:
$\tilde{\boldsymbol{\epsilon}}(\boldsymbol{x}, t) = \boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) + w\mathbf{M} \odot [\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) - \boldsymbol{\epsilon}_{\text{neg}}(\boldsymbol{x}, t)]$.

Experimental Results on ImageNet-1K

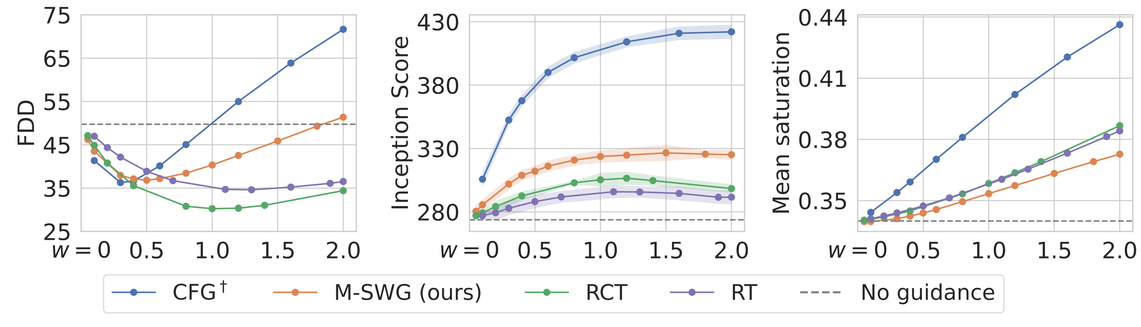


Fig. 4: Visualizing the achieved tradeoff by varying the guidance scale w (x-axis) for FDD (↓, left). Inception score (IS ↑, center), and saturation (↓, right) using EDM2-XXL. Saturation is quantified as the mean channel intensity in the HSV color space.

Model	Guidance	FID	FDD
EDM2 XXL-512	✗	2.29	49.7
	$\text{RT}_{.1/1.3}$	**2.26**	34.6
	$\text{SWG}_{.05/.2}$	2.61	37.7
	$\text{M-SWG}_{.05/.5}$	2.46	36.8
	$\text{RT}_{.1/1.05} + \text{M-SWG}_{0/.1}$	2.26	**32.9**
EDM2 XXL-512	RCT_1	-	30.3
	$\text{RCT}_{.9} + \text{SWG}_{.05}$	-	**29.8**
	$\text{RCT}_{.9} + \text{M-SWG}_{.05}$	-	**29.8**
	$\text{CFG}^{\dagger}_{.4}$	-	36.7
	$\text{CFG}^{\dagger}_{.2} + \text{M-SWG}_{.05}$	-	**36.2**
DiT XL-256	✗	9.83	213
	$\text{SWG}_{.25/.5}$	4.00	80.6
	$\text{M-SWG}_{.5/1.5}$	**3.30**	**70.3**
DiT XL-256	$\text{CFG}_{2.05}$	-	54.3
	$\text{CFG}_{.7} + \text{SWG}_{.1}$	-	50.6
	$\text{CFG}_{.6} + \text{M-SWG}_{.5}$	-	**49.1**

Tab. 1: The combinations of M-SWG and existing competitive guidance methods outperform the methods in the considered benchmarks.

- The subscripts indicate the best guidance scale based on a hyperparameter search.
- FID is only reported for training-free methods.
- M-SWG combinations similarly beat the baselines in *unconditional image synthesis* (see main paper).