Safety-Guided Flow (SGF): A Unified Framework for Negative Guidance in Safe Generation

Mingyu Kim[1)], Young-Heon Kim[2)] and Mijung Park[2)]
Kookmin University[1)], University of British Columbia[2)]
mgyukim@kookmin.ac.kr, yhkim@math.ubc.ca, mijungp@cs.ubc.ca

ICLR

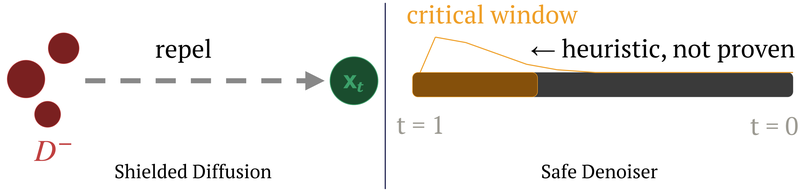
Introduction

Diffusion and flow models, widely used in various applications, can still generate unsafe results like NSFW content or inadvertently expose memorized data. Existing safety approaches are generally categorized into two types: control-barrier methods that impose geometric constraints at each step (Kirchhof et al. 2025), and negative-guidance methods that steer trajectories away from unsafe data (Kim et al. 2025).

However, prior approaches do not provide a unified probabilistic view of negative guidance, and they do not explain when such guidance should actually be active. SGF introduces an MMD-based potential that unifies Safe Denoiser and Shielded Diffusion, and explains why negative guidance should be strong only within an early critical window.

Motivation

1. Shielded Diffusion (Kirchhof et al. 2025): Radial-based negative guidance activates, but the radius, strength are all heuristic choices.
2. Safe Denoiser (Kim et al. 2025): Data-driven negative guidance is restricted to early steps heuristically, with no formal justification for why the early window works.



- SGF unifies both Shielded Diffusion and Safe Denoiser under a single MMD-based energy potential.
- We then apply control-barrier analysis to prove that a critical early window exists. It turns the heuristic schedule into a theoretical guarantee.

A Unified Framework of Negative Guidance

Safety-Guided Flow (SGF) is defined as

$$\dot{x}_t = f_\theta(x_t, t) + \lambda(t)\,\nabla_x E(x_t)$$

- $E(x_t) = \widehat{\mathrm{MMD}}^2_{k_\sigma}\big(\{x_t\}, D^-\big) = k(x_t, x_t) + \frac{1}{N^2}\sum_{i,j} k(y_i, y_j) - \frac{2}{N}\sum_{i=1}^{N} k(x_t, y_i)$
- $\nabla_{x_t} E(x_t) = \frac{2}{\sigma^2} Z(x_t)\left[x_t - \sum_{i=1}^{N} w_i(x_t)\, y_i\right], \quad Z(x_t) = \frac{1}{N}\sum_{i=1}^{N} k(x_t, y_i)$

- **Proposition 1.** The MMD gradient recovers Safe Denoiser's weighted-kernel repulsive field up to a positive scalar.
- **Proposition 2.** Under radius–bandwidth matching, it also recovers Shielded Diffusion's radial repulsive force.

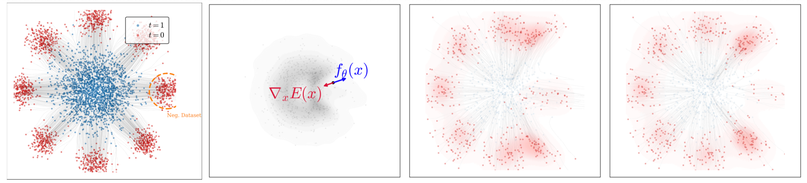
Theorem 2. (Critical window for early guidance)

$$e^{\int_0^{s_c} L(\tau)\,d\tau}\, h(x_0) + \mu \int_0^{s_c} \bar{w}_L(u)\,\beta(u)$$

- $\bar{w}_L(u)$ is decreasing in u, so shifting a fixed budget $\int\beta$ to earlier times strictly increases the left-hand side.
- It leads to *"Earlier guidance is provably better."*

Graphical Interpretation

- Applying guidance at all timesteps (c) avoids unsafe data but distorts nearby safe modes.
- Early-stop guidance (d) achieves the same safety with less distortion, confirming the critical window.



Quantitative Result

Combined with SLD or SAFREE, SGF achieves the lowest attack success rates across all benchmarks, consistently outperforming Safe Denoiser while maintaining comparable FID and CLIP scores.

Method	Fine Tuning	Negative Prompt	Negative Guidance	Ring-A-Bell		UnlearnDiff		MMA-Diffusion		COCO	
				ASR ↓	TR ↓	ASR ↓	TR ↓	ASR ↓	TR ↓	FID ↓	CLIP ↑
SD-v1.4	-	-	-	0.797	0.809	0.809	0.845	0.962	0.956	25.04	**31.38**
ESD	✓	✗	✗	0.456	0.506	0.422	0.426	0.628	0.640	27.38	30.59
RECE	✓	✗	✗	0.177	0.212	0.284	0.292	0.651	0.664	33.94	30.29
SLD	✗	✓	✗	0.481	0.573	0.629	0.586	0.881	0.882	36.47	29.28
SLD + SafeDenoiser	✗	✓	✓	0.354	0.429	0.526	0.485	0.481	0.549	36.59	29.10
SLD + Ours	✗	✓	✓	0.228	0.294	0.353	0.431	**0.297**	**0.357**	36.83	28.13
SAFREE	✗	✓	✗	0.278	0.311	0.353	0.363	0.601	0.618	25.29	30.98
SAFREE + SafeDenoiser	✗	✓	✓	0.127	0.169	0.207	0.241	0.469	0.501	**22.55**	30.66
SAFREE + Ours	✗	✓	✓	**0.051**	**0.133**	**0.164**	**0.232**	0.423	0.461	23.73	30.36

Qualitative Result