Dynamic Negative Guidance of Diffusion Models: Towards Immediate Content Removal

ICLR25

Félix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester*, **Luca Ambrogioni***

Contact: felix.koulischer@ugent.be

GHENT UNIVERSITY

Radboud Universiteit

TL;DR: A novel theoretically grounded ***D**ynamic **N**egative **G**uidance* scheme is introduced as alternative for the Negative Prompting algorithm.

Motivation

We derive a new ***D**ynamic **N**egative **G**uidance* scheme from first principles that better preserves image diversity when compared to the theoretically ungrounded Negative Prompting (NP) algorithm. Preserving image diversity is key to both improve FIDs and reduce potential ethical concerns.

Theoretical derivation

Negative Prompting originated as a simple change of sign from the Classifier-Free Guidance (CFG) equation.
This implies sampling from an ill-defined distribution:

$$p(\boldsymbol{x}, \boldsymbol{c}) \propto p(\boldsymbol{x}) / p(\boldsymbol{c}_-|\boldsymbol{x})^{\lambda} \qquad (1)$$

Instead, we remove the asymptote by reconstructing the desired conditional from the undesired conditional:

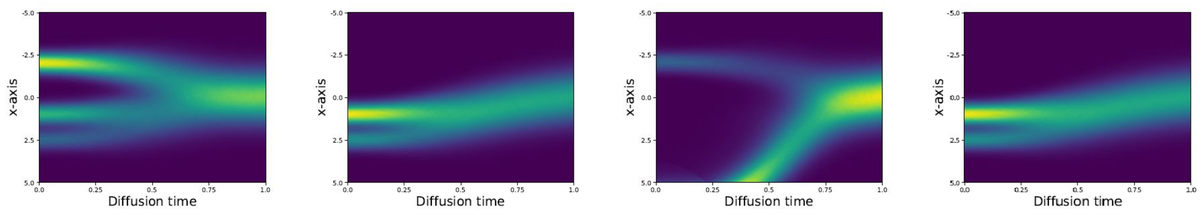
$$p_t(\boldsymbol{x}|\boldsymbol{c}_+) \propto p_t(\boldsymbol{x})\left(1 - p_t(\boldsymbol{c}_-|\boldsymbol{x})\right) \qquad (2)$$

From a score based perspective, this results in:

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}|\boldsymbol{c}_+) = \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log \left(1 - p_t(\boldsymbol{c}_-|\boldsymbol{x})\right)$$

$$\boldsymbol{\lambda(x,t)} = -\frac{p_t(\boldsymbol{c}_-|\boldsymbol{x})}{1 - p_t(\boldsymbol{c}_-|\boldsymbol{x})} \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{c}_-|\boldsymbol{x}) \qquad (3)$$

We find that a ***dynamic*** state-time modulated guidance scale is required. This guidance scale is large when the posterior probability is close to one, and small otherwise. Therefore, only regions related to the negative prompt are affected, reducing the invasiveness of the method.



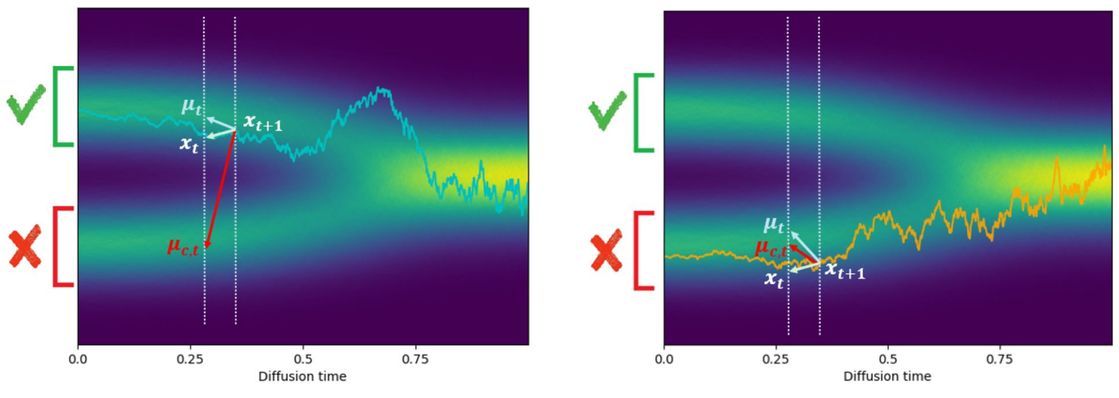
No guid. CFG (λ=1) NP (λ=1) DNG (ours)

Posterior approximation

Unfortunately, the posterior likelihood is most often unavailable at inference. We approximate it by tracking the Markov Chain along the diffusion trajectory.
We find a recursive formula that solely depends on already computed quantities:

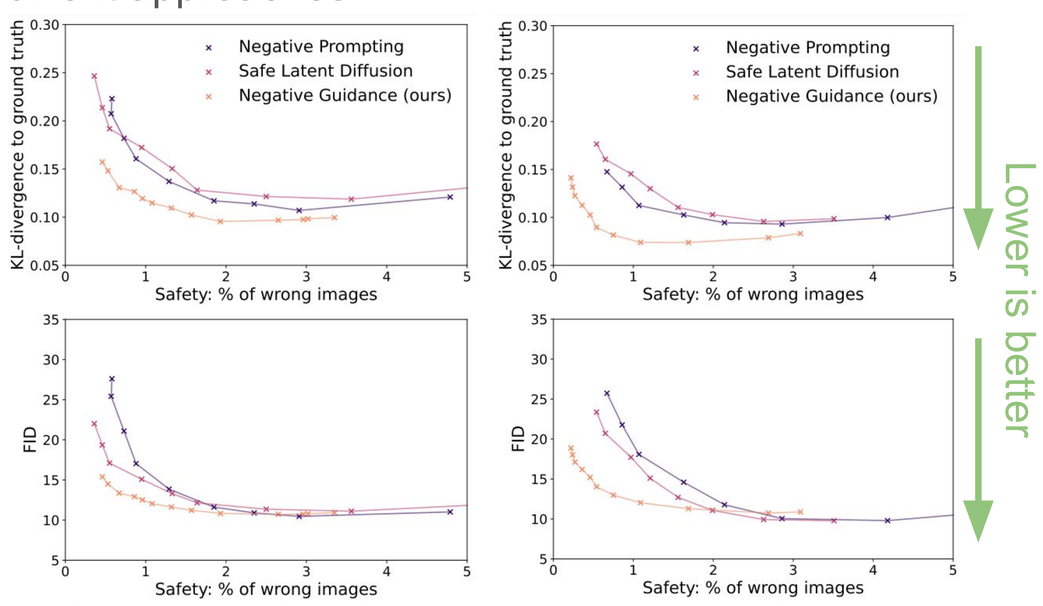
$$\log p_t(\boldsymbol{c}_-|\boldsymbol{x}_t) = \log p_{t+1}(\boldsymbol{c}_-|\boldsymbol{x}_{t+1}) - \frac{1}{2\sigma_t^2}\left(\|\boldsymbol{x}_t - \boldsymbol{\mu}_{t,\theta}(\boldsymbol{x}_{t+1}|\boldsymbol{c}_-)\|^2 - \|\boldsymbol{x}_t - \boldsymbol{\mu}_{t,\theta}(\boldsymbol{x}_{t+1})\|^2\right) \qquad (4)$$

neg. prompted unconditional



Class-removal Experiments

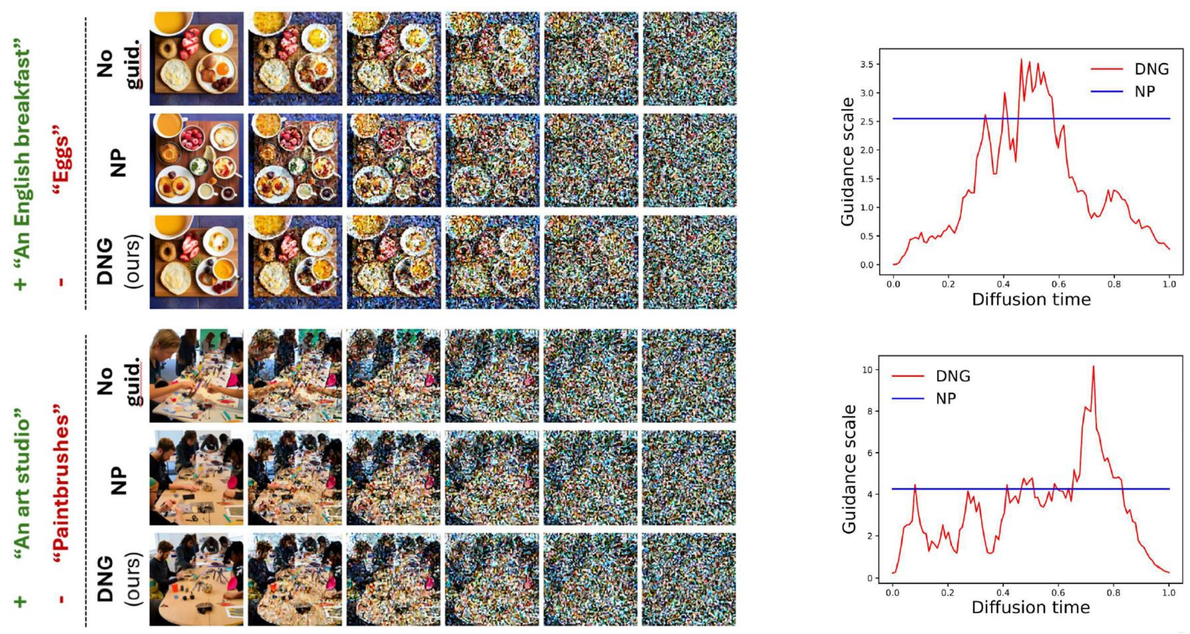
To evaluate how well the diversity of a model is affected by various negative guidance scheme, we propose a class removal setting. The goal is to avoid the generation of a specific class using solely an unconditional model and a model trained on that single class. DNG outperforms concurrent approaches



Removing: "airplane" "car"

T2I Experiments

In the context of T2I, our *dynamic guidance scale* preserves image diversity by deactivating itself for unrelated negative prompts. Here are some illustrative trajectories:



Conclusion

The theoretically optimal *dynamic* negative guidance scale can be approximated by tracking the posterior using the discrete markov chain to improve sample diversity when using negative prompting-like techniques. Our self-regulated guidance scale is further trajectory and prompt specific, providing further insights into when guidance is required

Further questions?
Scan this or ask us!

arXiv