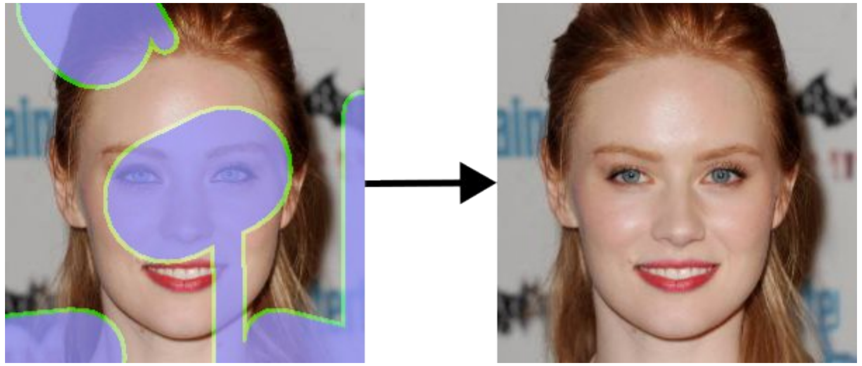
How I met Your Bias: Investigating Bias Amplification in Diffusion Models

Nathan Roos[1], Ekaterina Iakovleva[1], Ani Gjergji[2], Vito Paolo Pastore[2,3], Enzo Tartaglione[1]

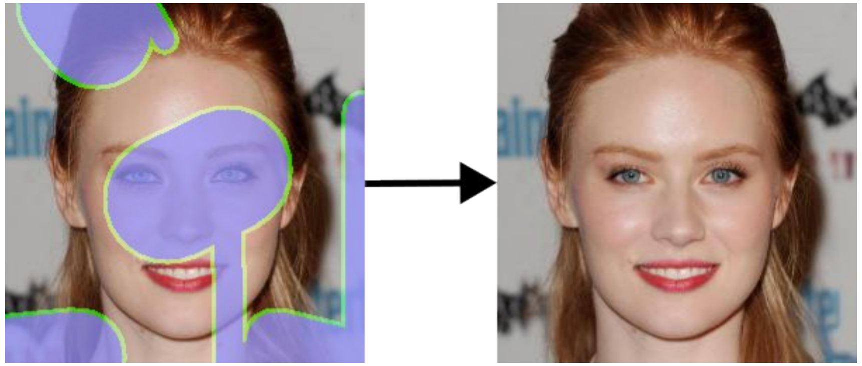
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2]MaLGa-DIBRIS, University of Genova, Italy
[3]AIGO, Istituto Italiano di Tecnologia, Italy

Introduction



RePaint [1]

Introduction

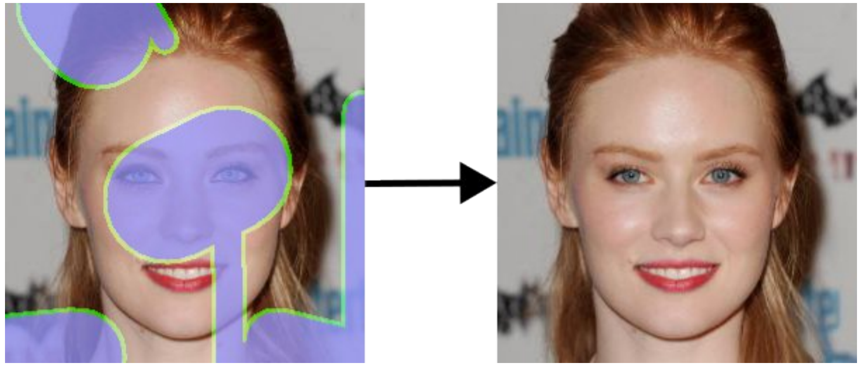


RePaint [1]

Insertion of this video at editing :
https://www.youtube.com/watch?v=lEcg6AJ6DVY

Sora 2, OpenAI (2025)

Introduction



RePaint [1]

Insertion of this video at editing : https://www.youtube.com/watch?v=lEcg6AJ6DVY

Sora 2, OpenAI (2025)

Prompt: “A portrait photo of a lawyer”

StableDiffusion v2.1-base

DDIM sampling with 50 steps

Generated images

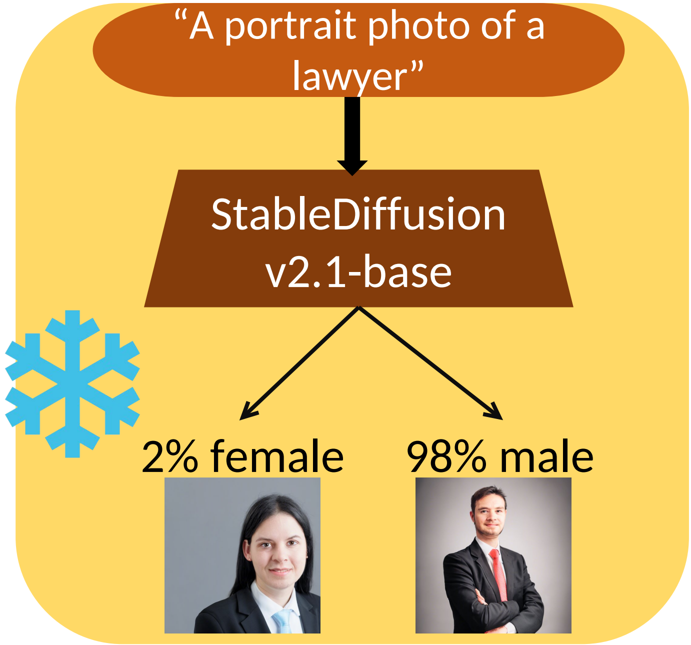
2% female lawyers

98% male lawyers

Research Motivation

Previous works:

- bias amplification is **a fixed, intrinsic feature of the model** [2, 3, 4]

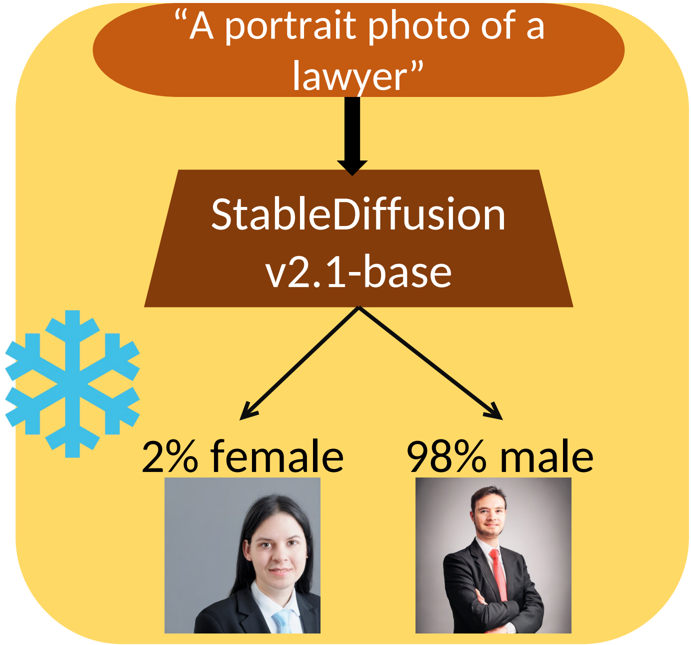


Research Motivation

Previous works:

- bias amplification is **a fixed, intrinsic feature of the model** [2, 3, 4]

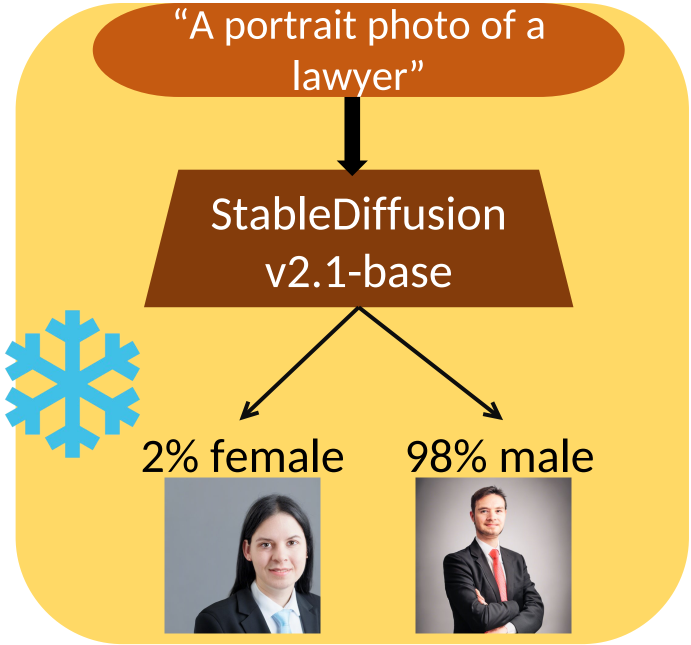
- bias mitigation (*e.g.* attribute guidance [5, 6], policy solver [7]) using **costly retraining or finetuning of the model**



Research Motivation

Previous works:

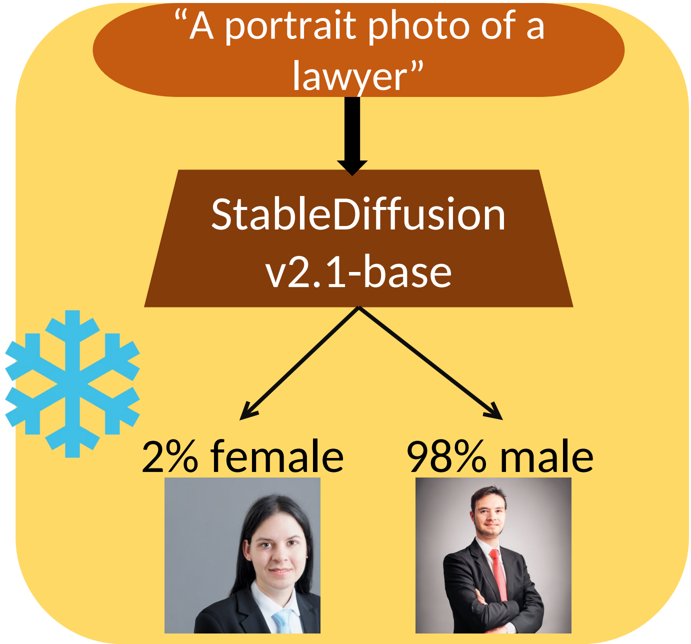
- bias amplification is **a fixed, intrinsic feature of the model** [2, 3, 4]
- bias mitigation (*e.g*. attribute guidance [5, 6], policy solver [7]) using **costly retraining or finetuning of the model**
- no systematic analysis of how the sampling process itself affects bias amplification



Research Motivation

Previous works:

- bias amplification is **a fixed, intrinsic feature of the model** [2, 3, 4]
- bias mitigation (*e.g.* attribute guidance [5, 6], policy solver [7]) using **costly retraining or finetuning of the model**
- no systematic analysis of how the sampling process itself affects bias amplification

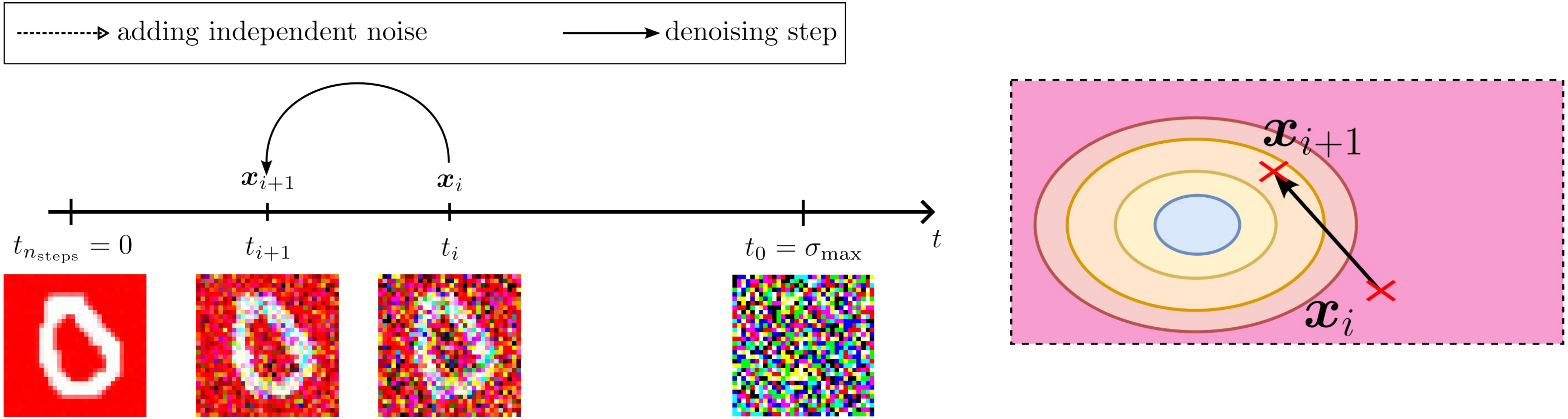


<u>Research question:</u>

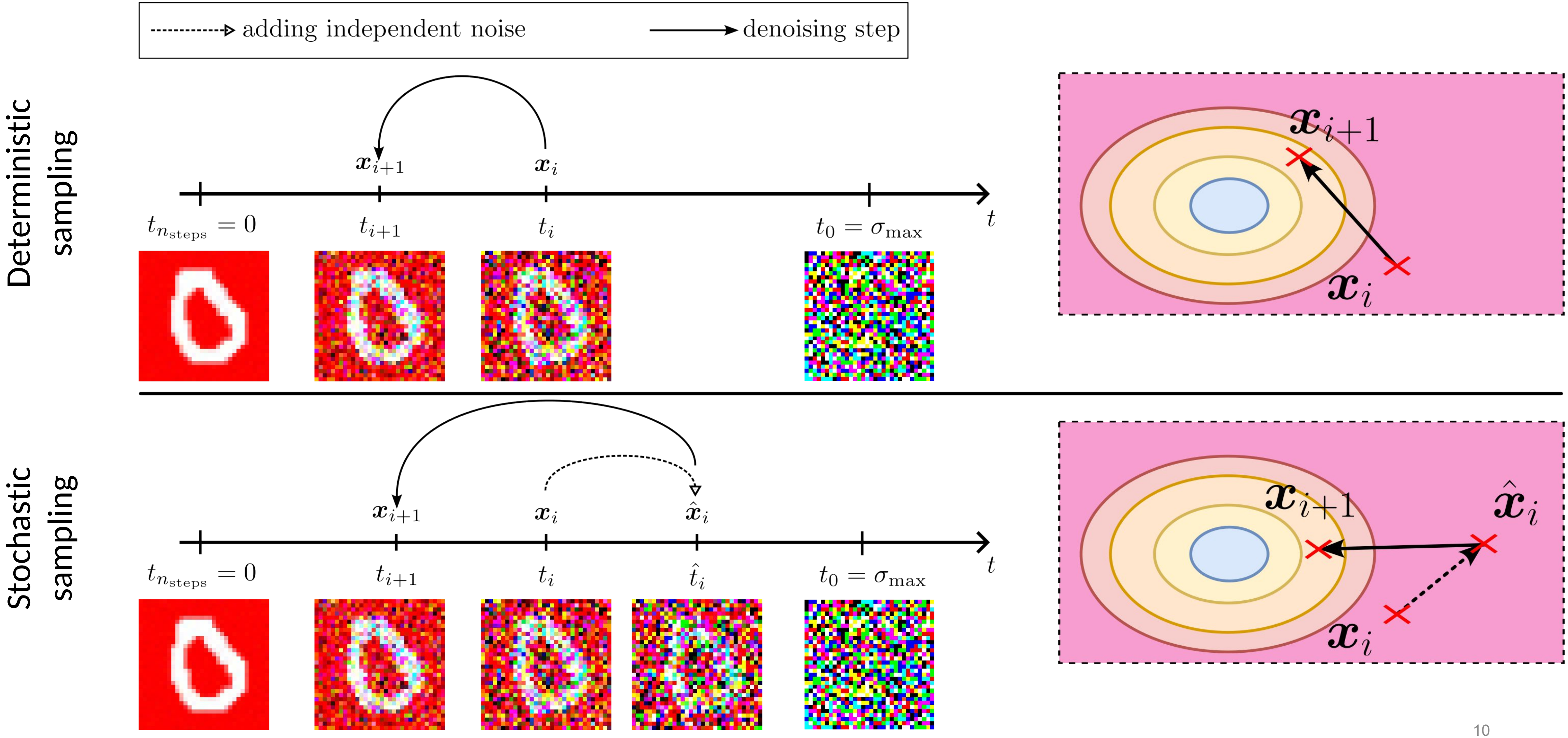
Does the choice of sampling hyperparameters influence the level of bias in images generated by Diffusion Models?

Background: Diffusion Models

Deterministic sampling



Background: Diffusion Models



Method (1/3): definition of bias

The level of bias is measured by ρ the ratio of bias-aligned samples in the generated images or in the dataset:

$$\rho = P(\text{"bias-aligned"} \mid \text{target attribute})$$

$\rho_{model} > \rho_{dataset}$: bias amplification

$\rho_{model} < \rho_{dataset}$: bias reduction

Method (1/3): definition of bias

The level of bias is measured by ρ the ratio of bias-aligned samples in the generated images or in the dataset:

$$\rho = P(\text{"bias – aligned"} \mid \text{target attribute})$$

$\rho_{model} > \rho_{dataset}$: bias amplification

$\rho_{model} < \rho_{dataset}$: bias reduction

target attribute: “lawyer”

Both images were generated using StableDiffusion v2.1-base with the prompt *“A portrait photo of a lawyer”*

bias attribute: “gender”

“male”
bias-aligned sample

“female”
bias-conflicting sample

12

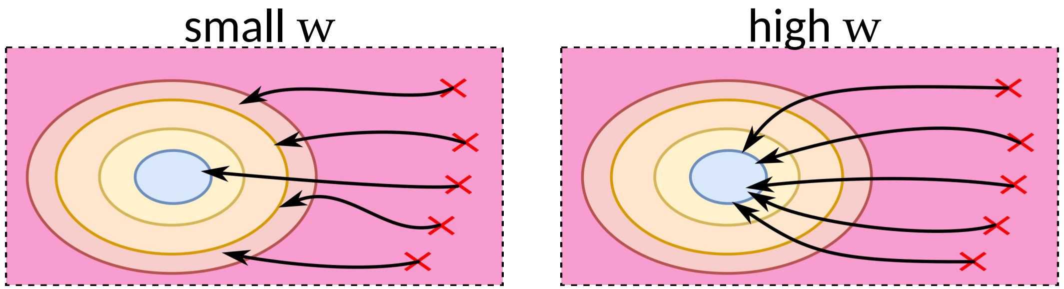
Method (2/3): selecting hyperparameters

Relevant groups of hyperparameters:

Method (2/3): selecting hyperparameters

Relevant groups of hyperparameters:

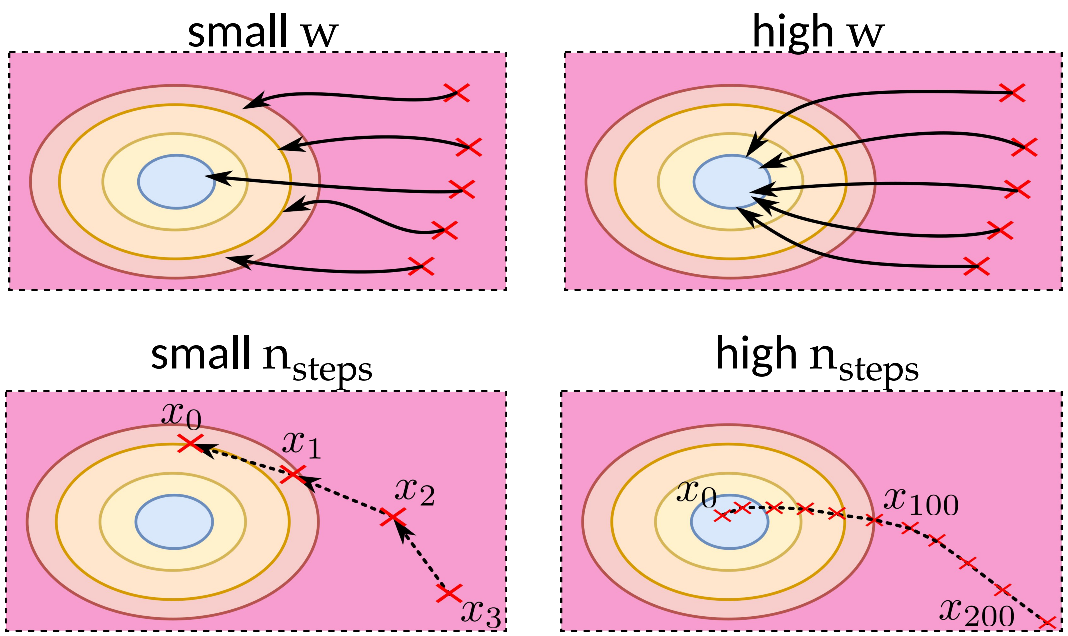
- **Conditioning strength**:
  - guidance scale w of CFG



Method (2/3): selecting hyperparameters

Relevant groups of hyperparameters:

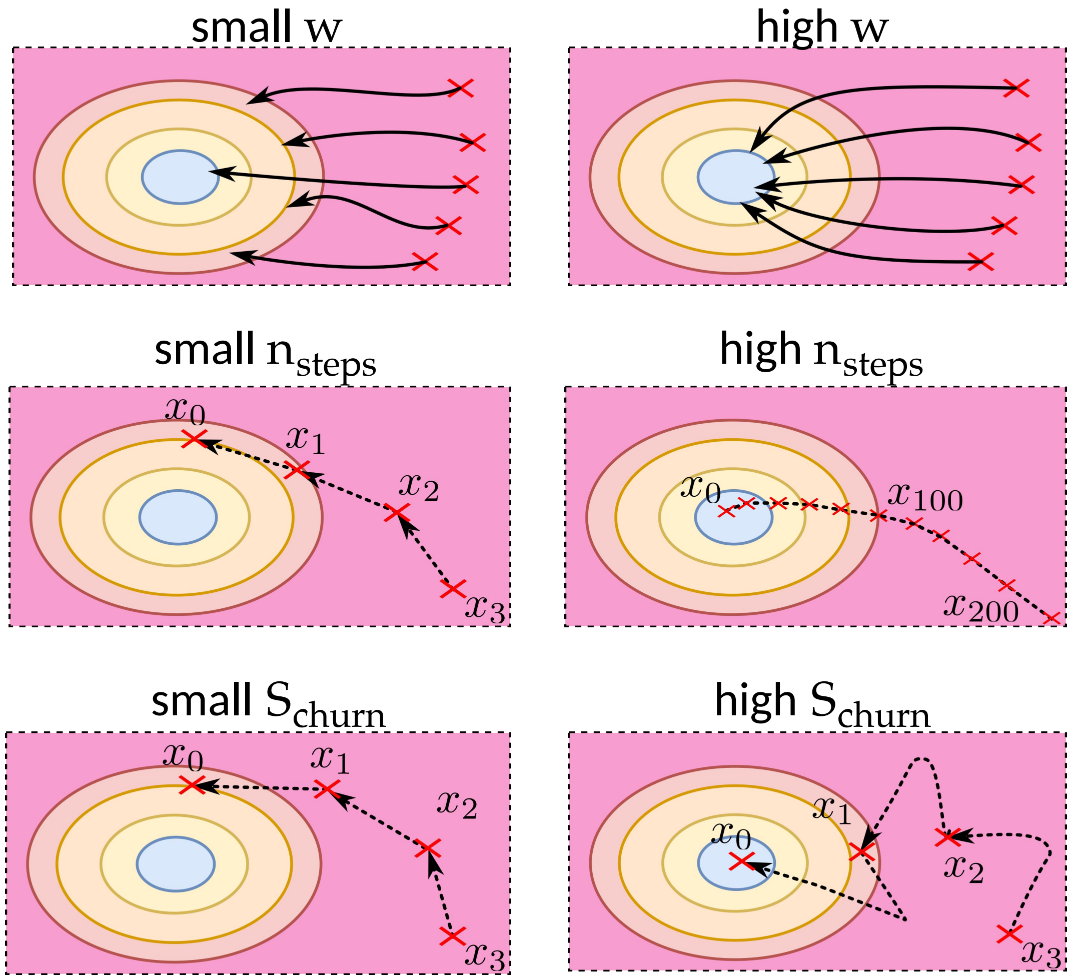
- **Conditioning strength**:
  - guidance scale w of CFG
- **Computational cost**:
  - number of sampling steps $\mathrm{n}_{\mathrm{steps}}$
  - type of integration scheme



Method (2/3): selecting hyperparameters

Relevant groups of hyperparameters:

- **Conditioning strength**:
  - guidance scale w of CFG
- **Computational cost**:
  - number of sampling steps n_{steps}
  - type of integration scheme
- **Amount of stochasticity**:
  - η or S_{churn}: variance of added noise
  - $S_{t,min}$, $S_{t,\,max}$: time window with noise

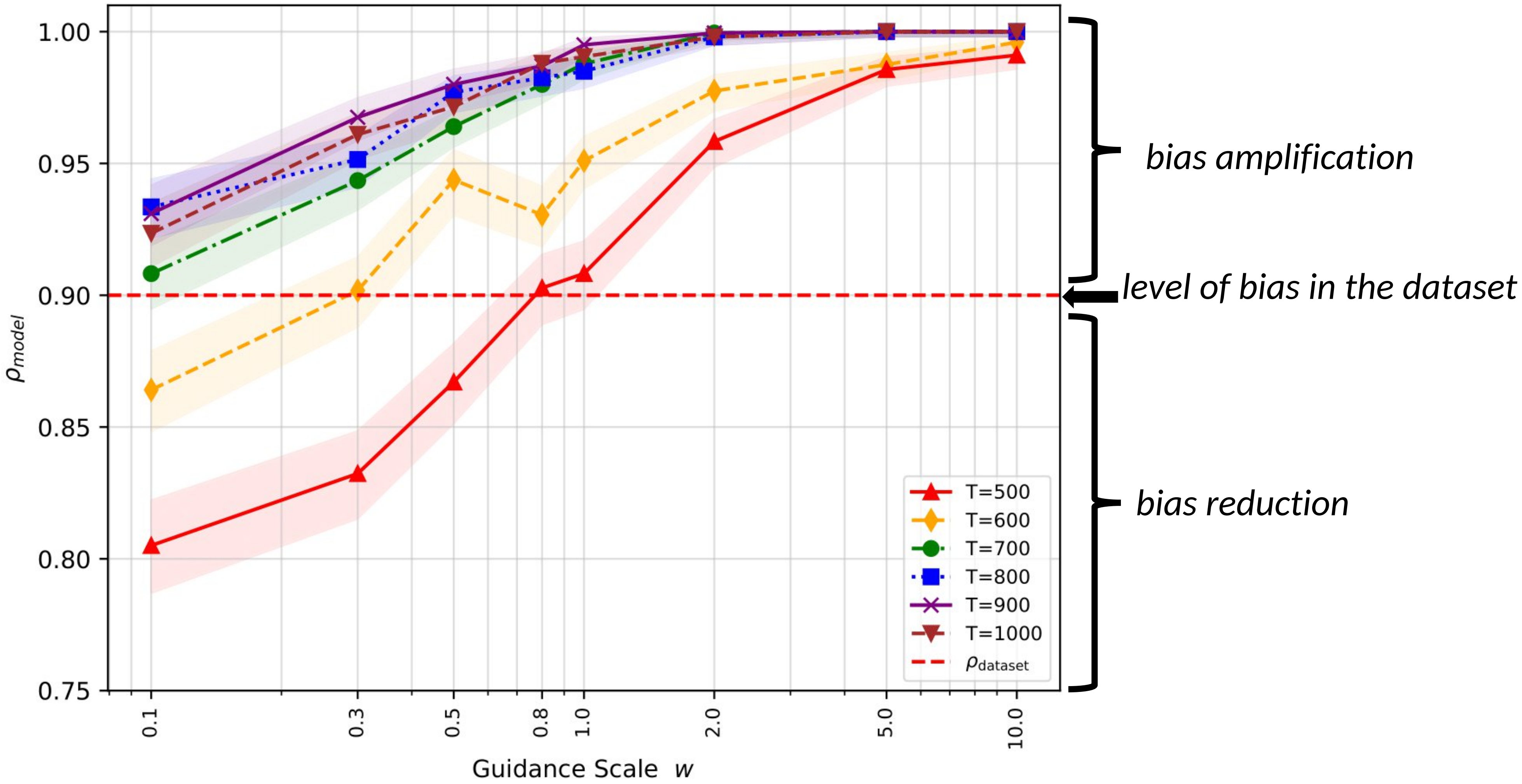


Method (3/3): experiments

Assess the impact of the hyperparameters on ρ_{model} across several models, samplers and datasets:

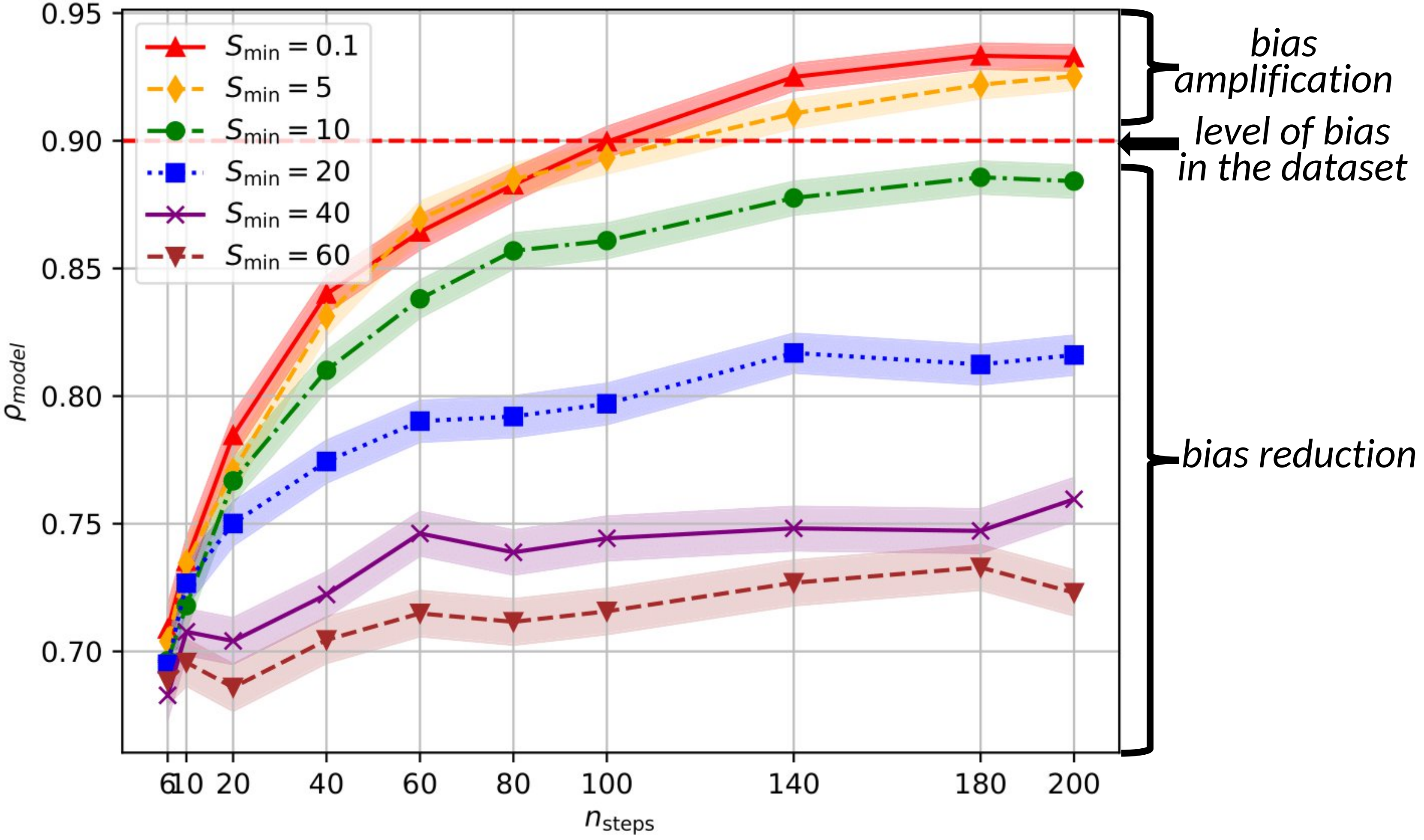
Model	Sampler	Datasets
DDPM [8], CDPM	DDPM, DDIM [9]	Biased MNIST [10], BFFHQ [11]
Continuous framework of [12]	EDM-Sampler, VP-Sampler [13], DPM-Solver [14]	Biased MNIST, Multi-color MNIST [15]
Stable Diffusion [16]	DDIM	Laion-5B [17]

Results: effect of the conditioning strength



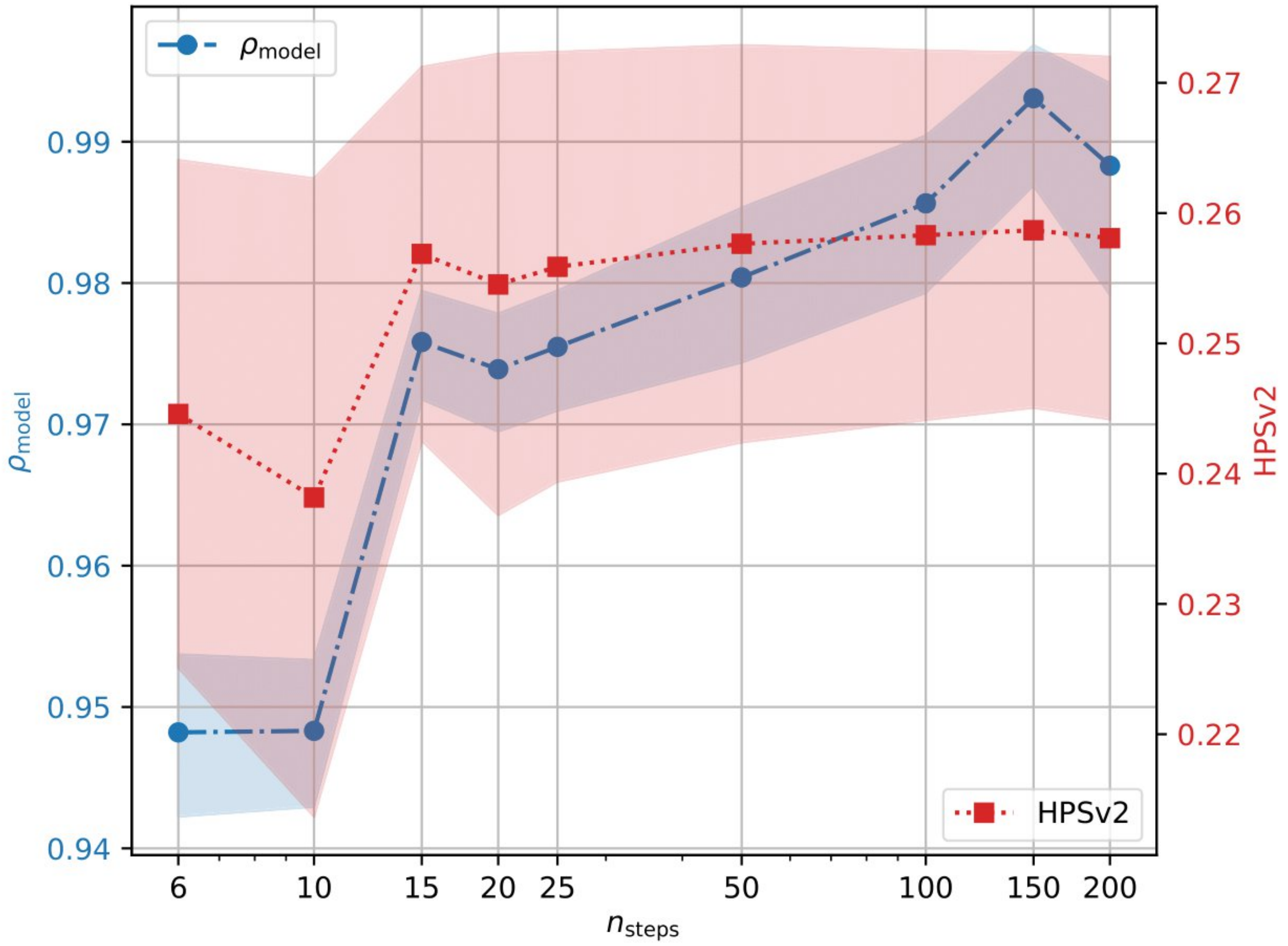
ρ_{model} vs guidance scale of CFG for various T for a fixed model trained on 2-classes Biased MNIST ($\rho_{dataset} = 0.9$), standard DDPM sampler

Results: effect of the computational cost (1/2)



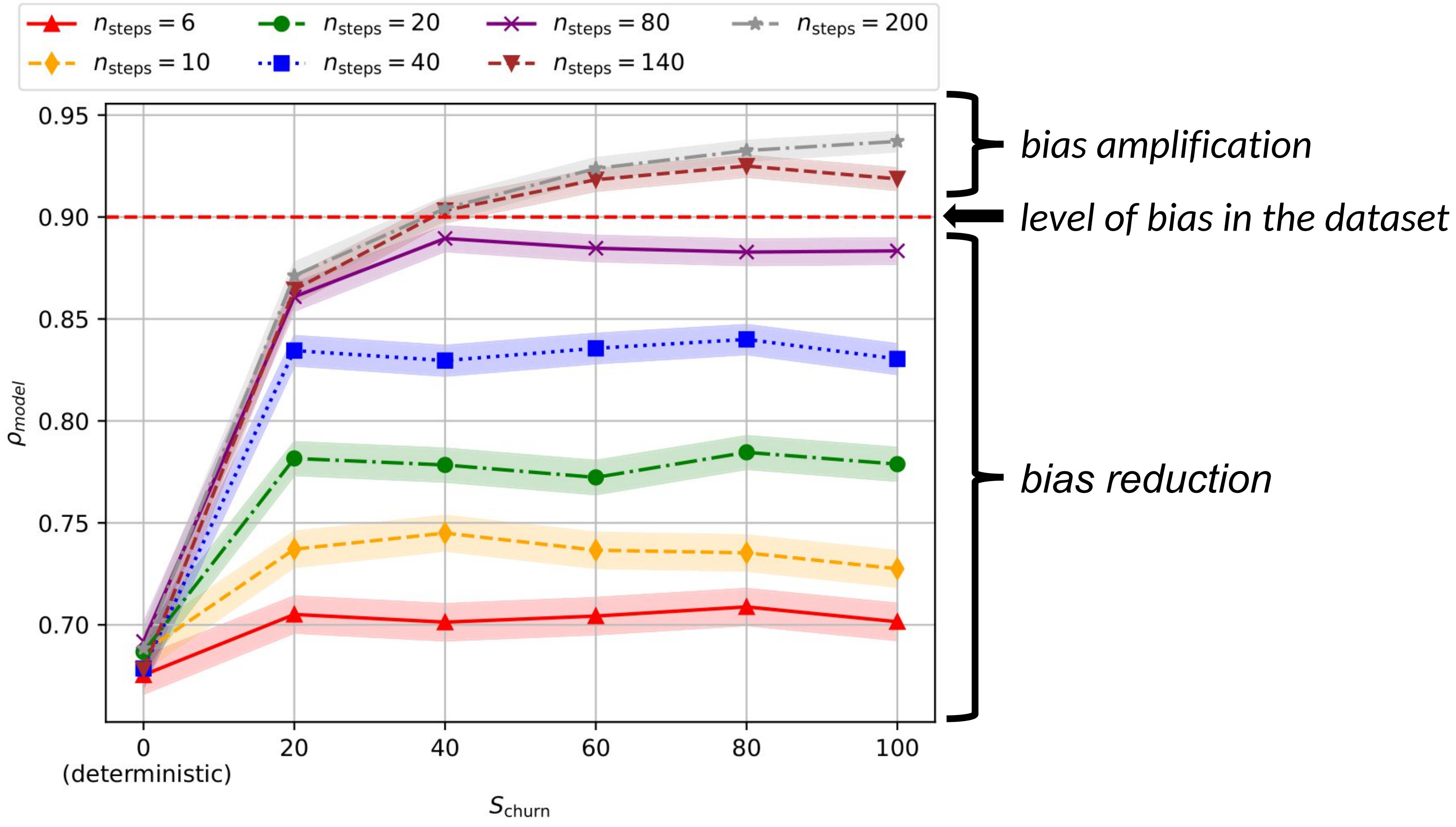
ρ_{model} vs $\mathrm{n_{steps}}$ for a fixed model trained on 10-classes Biased MNIST, EDM-Sampler

Results: effect of the computational cost (2/2)



ρ_{model} and HPSv2 vs $\mathrm{n_{steps}}$ for Stable Diffusion, DDIM sampler

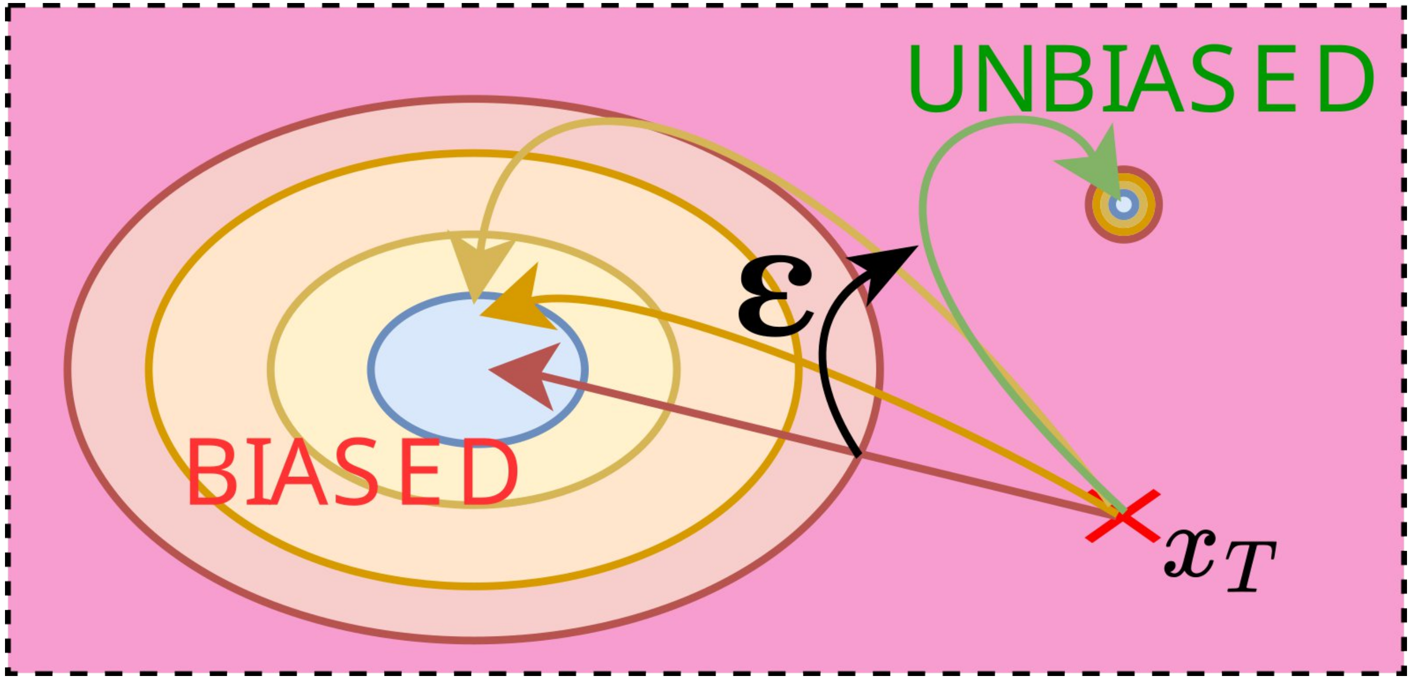
Results: effect of the amount of stochasticity



ρ_{model} vs S_{churn} with various n_{steps} for a fixed model trained on 10-classes Biased MNIST, EDM-Sampler

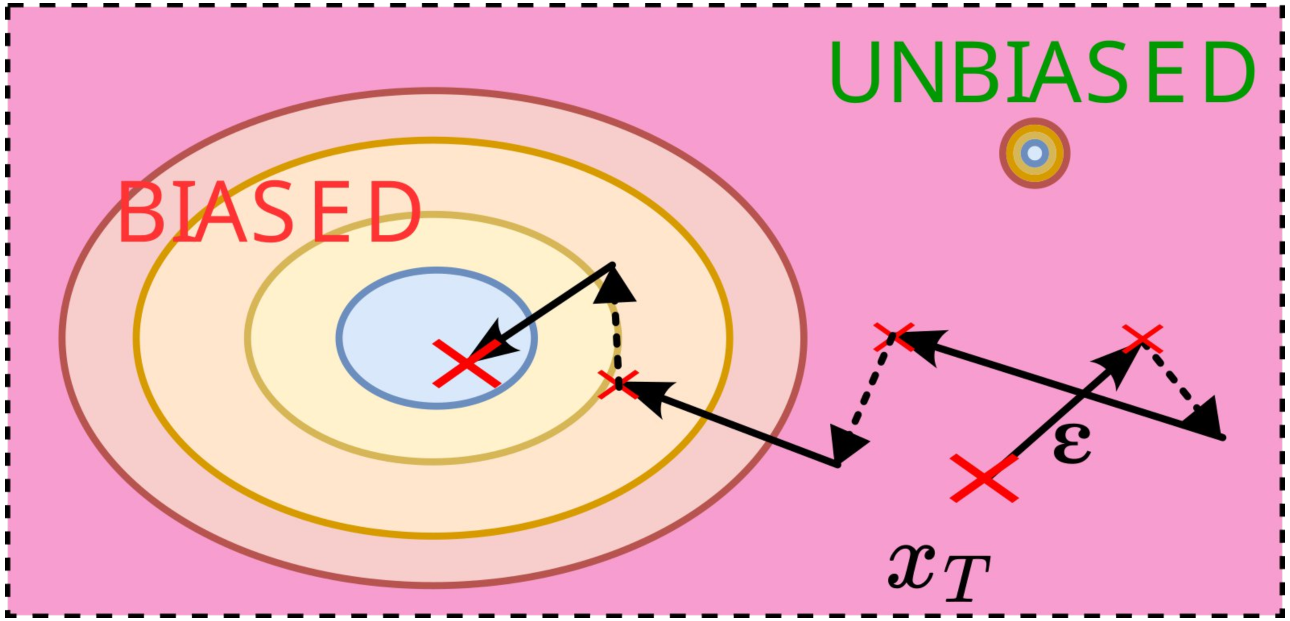
Discussion (1/2)

- Low number of sampling steps →numerical errors → sampling trajectory deviates from biased data distribution



Discussion (2/2)

- Surprisingly, the noise of stochastic sampling does not help in debiasing, but the opposite → correction of early numerical errors?



Conclusions

Does the choice of sampling hyperparameters influence the level of bias in images generated by Diffusion Models?

→YES

Conclusions

Does the choice of sampling hyperparameters influence the level of bias in images generated by Diffusion Models?

$\longrightarrow$YES

Consequences:

- Careful constructions of vanilla baselines for comparison with debiasing methods

Conclusions

Does the choice of sampling hyperparameters influence the level of bias in images generated by Diffusion Models ?

$\longrightarrow$YES

Consequences:

- Careful constructions of vanilla baselines for comparison with debiasing methods
- Possible to design a debiasing strategy by tuning the sampling hyperparameters

Thank you for your attention !

ArXiv Paper

Source code of the experiments

Bibliography

[1] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "RePaint: Inpainting Using Denoising Diffusion Probabilistic Models," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11461-11471.

[2] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan, "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale," in *Proc. 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1493–1504

[3] N. Garcia, Y. Hirota, Y. Wu and Y. Nakashima, "Uncurated Image-Text Datasets: Shedding Light on Demographic Bias," in *Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6957-6966

[4] A. Mandal, S. Leavy, and S. Little, "Generated Bias: Auditing Internal Bias Dynamics of Text-to-Image Generative Models," in *Proc. ECCV 2024 Workshops*, 2024, pp. 96–111

[5] S. Hong, D. Ahn, and S. Kim, "Debiasing scores and prompts of 2D diffusion for view-consistent text-to-3D generation," in *Proc. 37th International Conference on Neural Information Processing Systems*, 2023

[6] R. Parihar, A. Bhat, A. Basu, S. Mallick, J. N. Kundu, and R. V. Babu, "Balancing Act: Distribution-Guided Debiasing in Diffusion Models," in *Proc. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 6668-6678.

[7] R. He, C. Xue, H. Tan, W. Zhang, Y. Yu, S. Bai, and X. Qi, "Debiasing Text-to-Image Diffusion Models," in *Proc. 1st ACM Multimedia Workshop on Multi-Modal Misinformation Governance in the Era of Foundation Models*, 2024, pp. 29–36.

Bibliography

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020, pp. 6840–6851.

[9] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *9th International Conference on Learning Representations (ICLR)*, 2021.

[10] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh, "Learning de-biased representations with biased representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 528–539.

[11] E. Kim, J. Lee and J. Choo, "BiaSwap: Removing Dataset Bias with Bias-Tailored Swapping Augmentation," in *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14972-14981.

[12] T. Karras, M. Aittala, S. Laine, and T. Aila, "Elucidating the design space of diffusion-based generative models," in *Proc. of the 36th International Conference on Neural Information Processing Systems*, 2022.

[13] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *9th International Conference on Learning Representations (ICLR)*, 2021.

[14] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *Proc. of the 36th International Conference on Neural Information Processing Systems*, 2022.

[15] Z. Li, A. Hoogs, and C. Xu, "Discover and mitigate unknown biases with debiasing alternate networks," in *European Conference on Computer Vision (ECCV)*, 2022.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10674-10685.

[17] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. of the 36th International Conference on Neural Information Processing Systems*, 2022.