EDMSound: Spectrogram Based Diffusion Models for Efficient and High-Quality Audio Synthesis

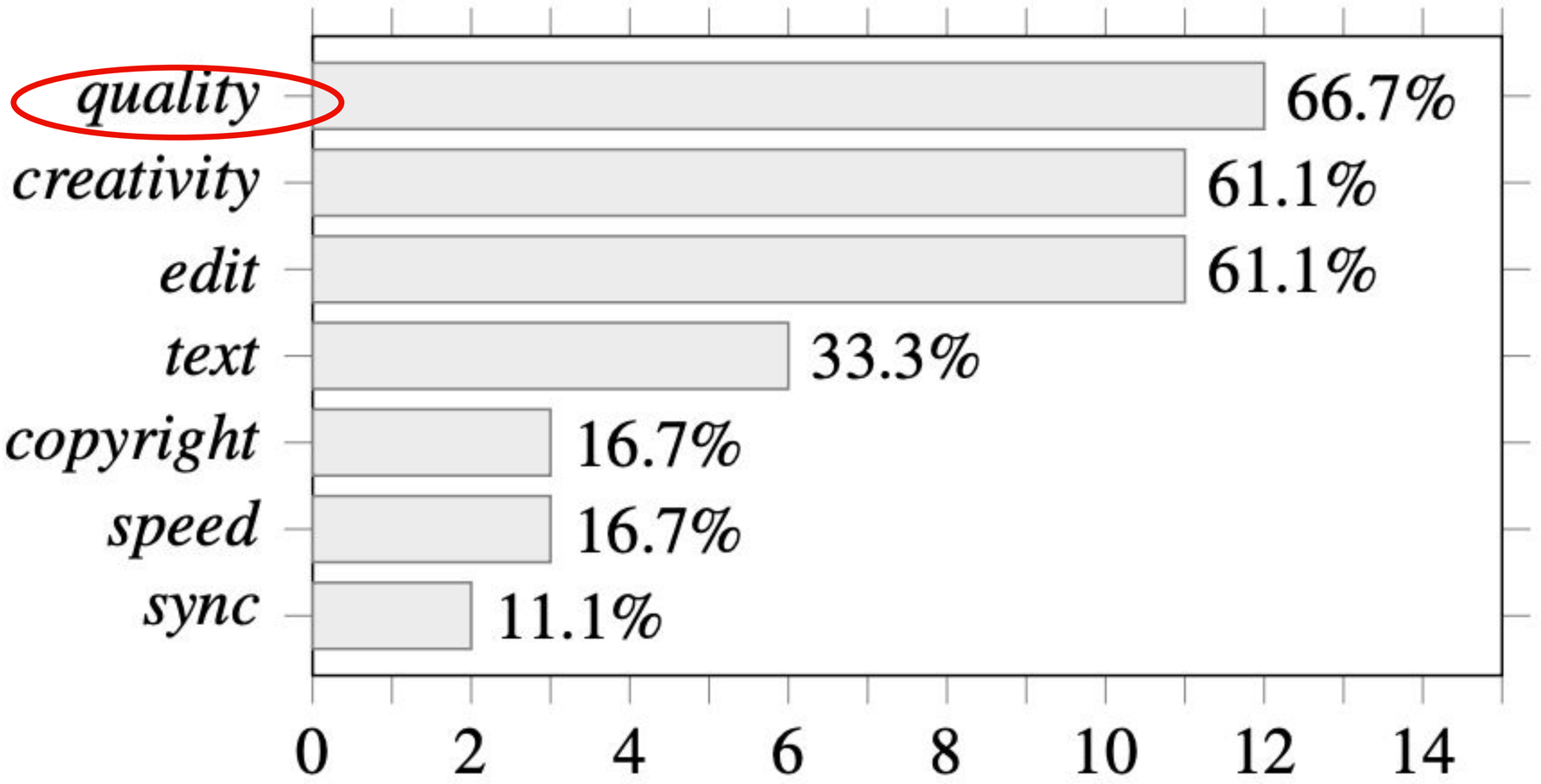
Ge Zhu[1], Yutong Wen[1], Marc-André Carbonneau[2] and Zhiyao Duan[1]

[1] ECE, University of Rochester

[2] Ubisoft La Forge

NeurIPS Workshop on ML for Audio 2023

“What is the limitation(s) of the current text-conditioned audio generation as a product?”



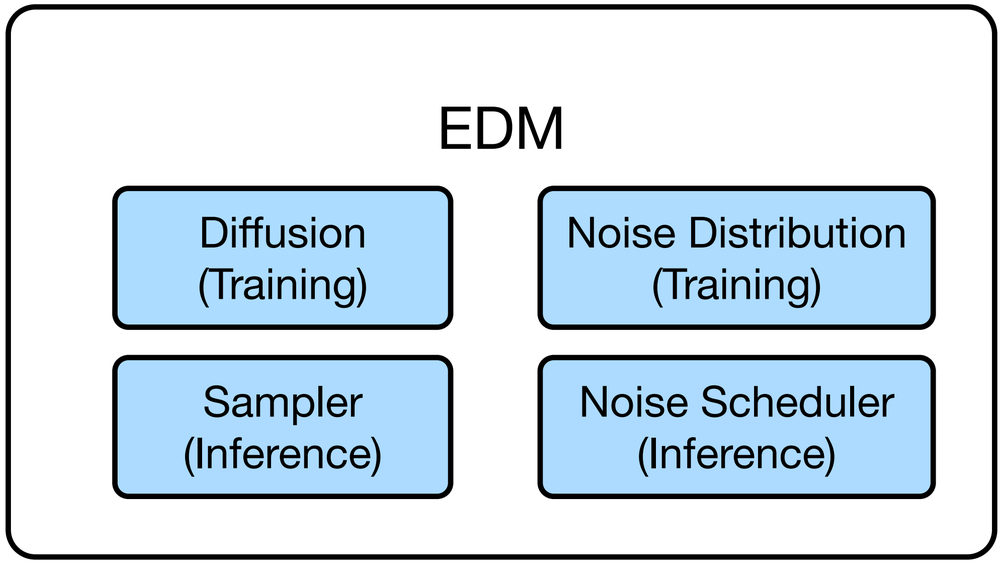
Sangshin Oh et al. A Demand-Driven Perspective on Generative Audio AI. *Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML) USA. 2023*

Overview

- EDMSound: Spectrogram based diffusion models
  - Training: Elucidated diffusion model (EDM) framework
  - Sampling: Exponential-integrator based deterministic solver
- Copy detection/memorization issue in diffusion models
  - Fine-tuned CLAP for audio

Elucidated Diffusion Model (EDM)

- Flexible diffusion training and inference with decoupled components



Karras, Tero, et al. "Elucidating the design space of diffusion-based generative models." Advances in Neural Information Processing Systems 35 (2022): 26565-26577.

Perspectives on diffusion

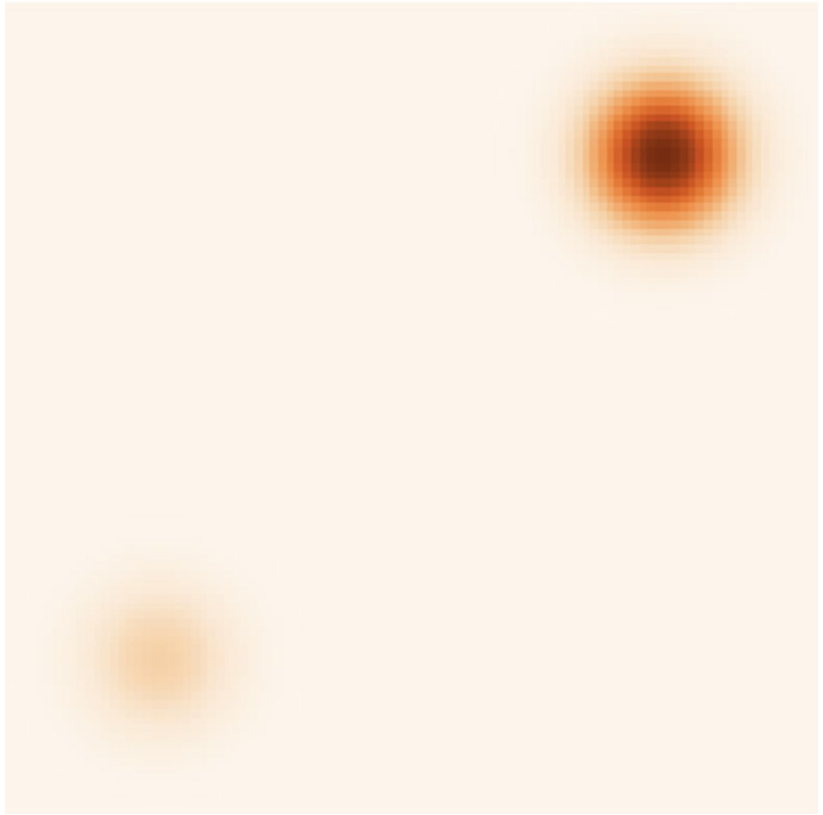
1. *Diffusion models are autoencoders*
2. *Diffusion models are deep latent variable models*
3. *Diffusion models predict the score function*
4. *Diffusion models solve reverse SDEs*
5. *Diffusion models are flow-based models*
6. *Diffusion models are recurrent neural networks*
7. *Diffusion models are autoregressive models*
8. *Diffusion models estimate expectations*

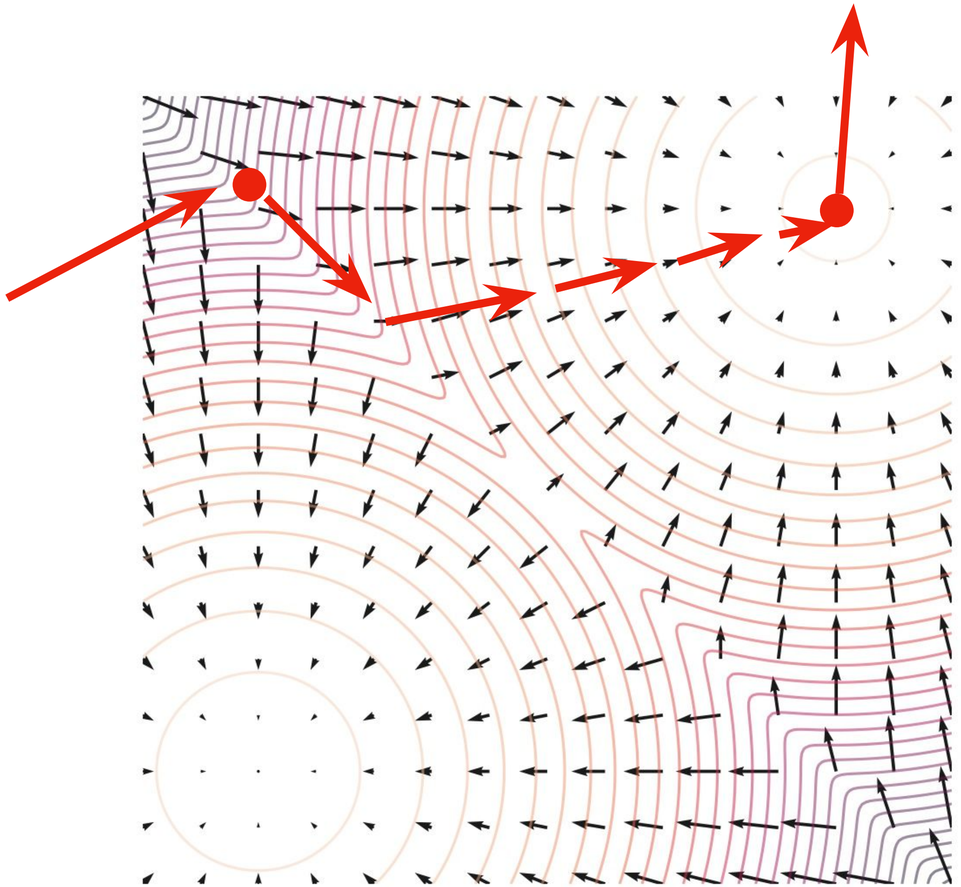
…

Dieleman, Sander “Perspectives on diffusion” https://sander.ai/2023/07/20/perspectives.html

Diffusion model

$\nabla \log p_t(x)$





- Fitting (training) error: score function estimation
- Discretization (inference) error: limited steps or large step size

Especially important in sampling with limited steps!

Elucidated Diffusion Model (EDM)

Fitting error:

- Pre-conditioning

  Predicting x_0 or ϵ objectives are problematic

  Network inputs depend on the noise level

- Loss reweighting

→ Both input and output magnitudes are fixed to unit variance

Karras, Tero, et al. "Elucidating the design space of diffusion-based generative models." Advances in Neural Information Processing Systems 35 (2022): 26565-26577.

Diffusion Exponential Integrator Sampler

Ingredient 1: Exponential Integrator over Euler method.

Original reverse sampling ODE:

$$\frac{d\hat{\boldsymbol{x}}}{dt} = \boldsymbol{F}_t\hat{\boldsymbol{x}} - \frac{1}{2}\boldsymbol{G}_t\boldsymbol{G}_t^T\boldsymbol{s}_\theta(\hat{\boldsymbol{x}}, t)$$

Semi-linear

Exact solution:

$$\hat{\boldsymbol{x}}_t = \Psi(t, s)\hat{\boldsymbol{x}}_s + \int_s^t \Psi(t, \tau)[-\frac{1}{2}\boldsymbol{G}_\tau\boldsymbol{G}_\tau^T\boldsymbol{s}_\theta(\hat{\boldsymbol{x}}_\tau, \tau)]d\tau$$

Discretization:

$$\hat{\boldsymbol{x}}_{t-\Delta t} = \hat{\boldsymbol{x}}_t - [\boldsymbol{F}_t\hat{\boldsymbol{x}}_t - \frac{1}{2}\boldsymbol{G}_t\boldsymbol{G}_t^T\boldsymbol{s}_\theta(\hat{\boldsymbol{x}}_t, t)]\Delta t$$

Discretization:

$$\hat{\boldsymbol{x}}_{t-\Delta t} = \Psi(t-\Delta t, t)\hat{\boldsymbol{x}}_t + [\int_t^{t-\Delta t} -\frac{1}{2}\Psi(t-\Delta t, \tau)\boldsymbol{G}_\tau\boldsymbol{G}_\tau^T d\tau]\boldsymbol{s}_\theta(\hat{\boldsymbol{x}}_t, t)$$

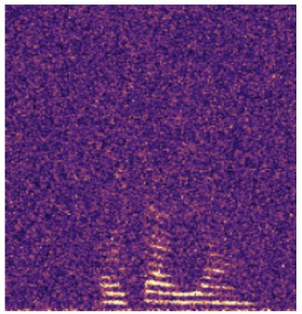
Ingredient 2: $\varepsilon_\theta(x, t)$ over $s_\theta(x, t)$

Ingredient 3: Polynomial extrapolation of ε_θ

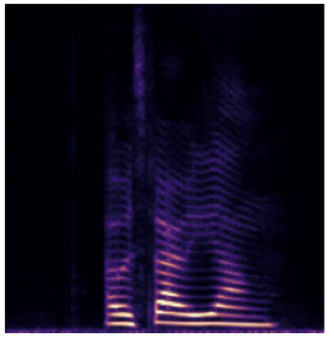
Zhang, Qinsheng, and Yongxin Chen. "Fast sampling of diffusion models with exponential integrator." 11th International Conference on Learning Representations(2022).

Neural Network Architecture

$$\tilde{c} = \frac{|c|^{\alpha}}{\beta} e^{i\angle(c)}$$



Efficient U-Net



FiLM

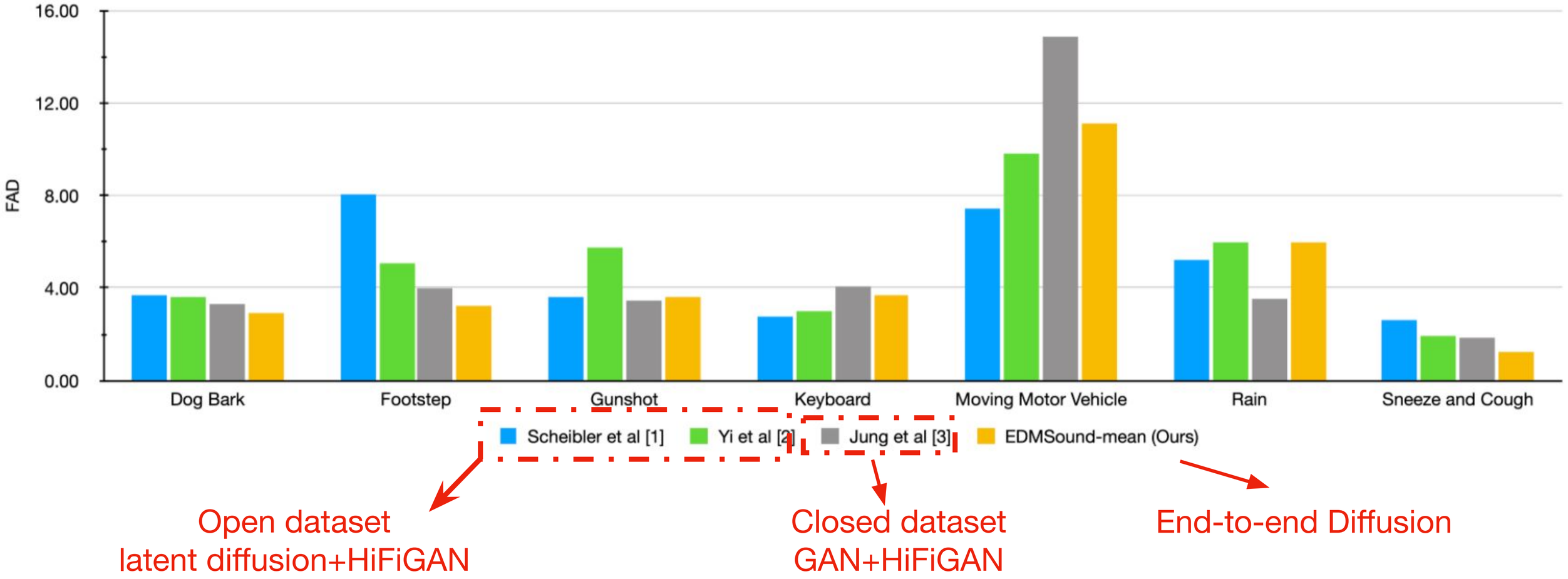
Label/text

Embedding Layer

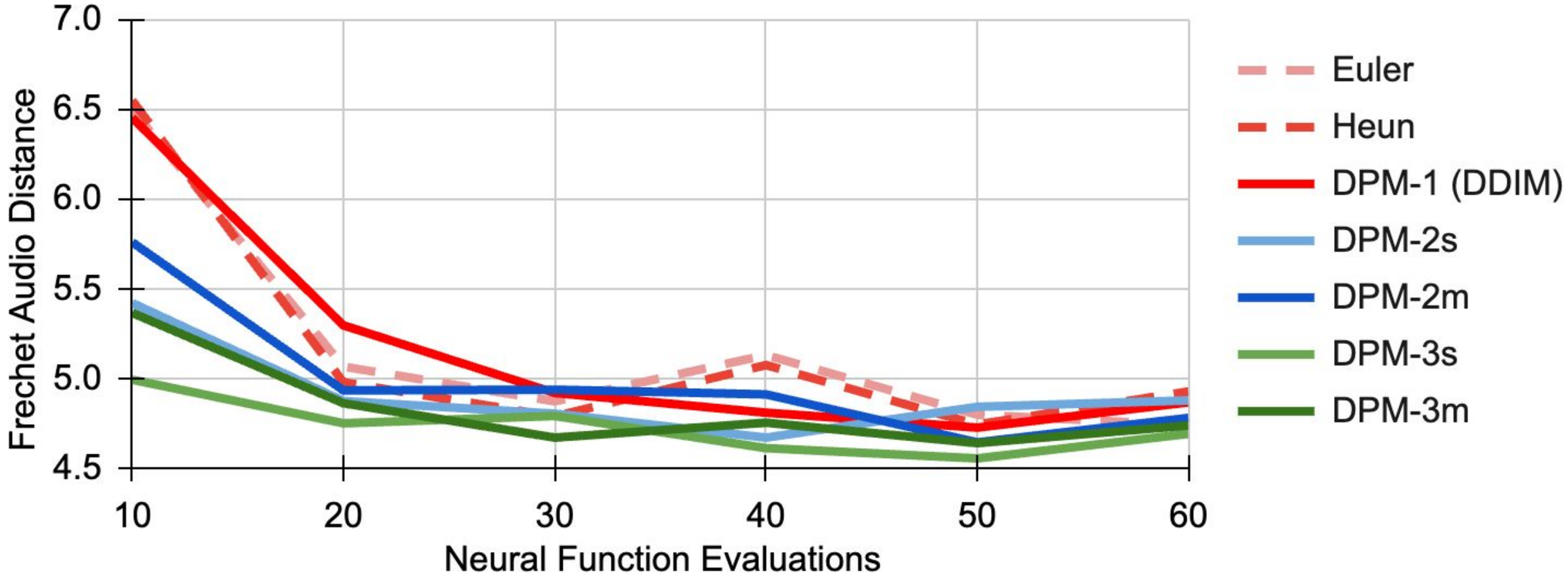
Richter, Julius, et al. "Speech enhancement and dereverberation with diffusion-based generative models." IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023).

Saharia, Chitwan, et al. "Photorealistic text-to-image diffusion models with deep language understanding." Advances in Neural Information Processing Systems 35 (2022): 36479-36494.

Fréchet Audio Distance on DCASE2023 Challenge Task7



Comparison of different samplers on DCASE2023 Task 7

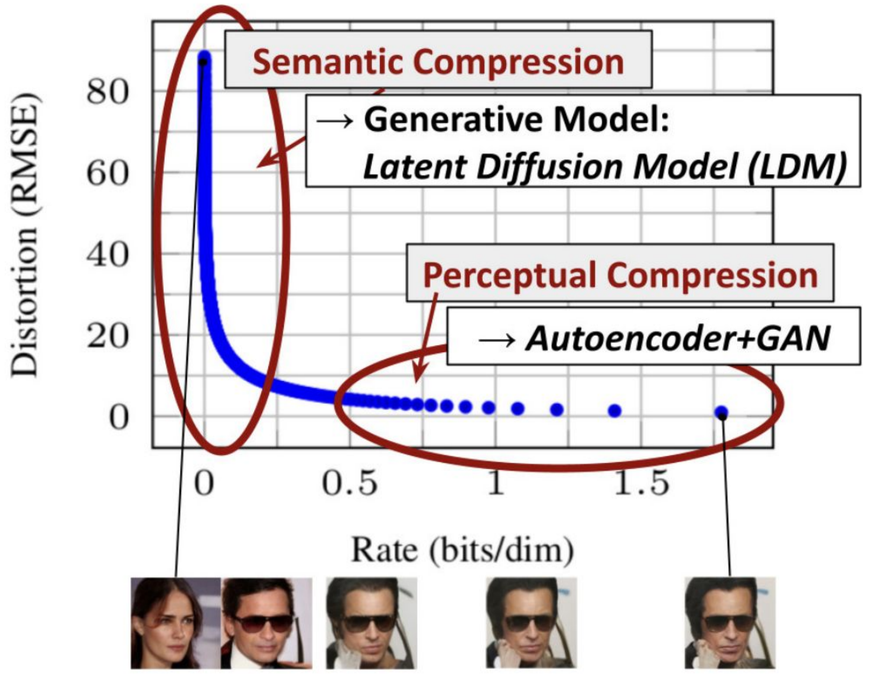


Lu, Cheng, et al. "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps." *Advances in Neural Information Processing Systems* 35 (2022): 5775-5787.

Lu, Cheng, et al. "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models." *arXiv preprint arXiv:2211.01095* (2022).

Limitations on generation

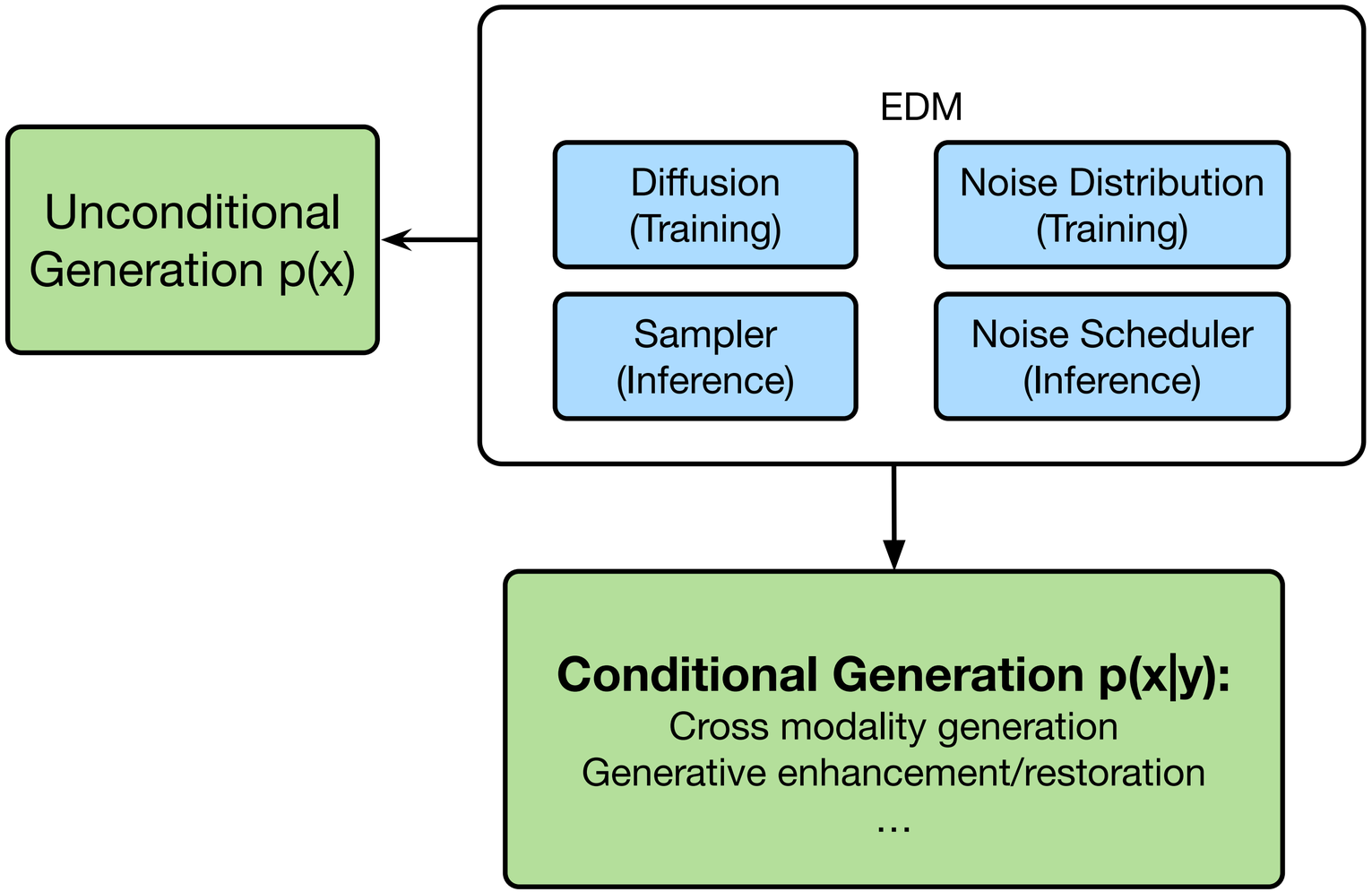
- Efficient samplers may not be as good as training-based method*
- Long context with rich semantics



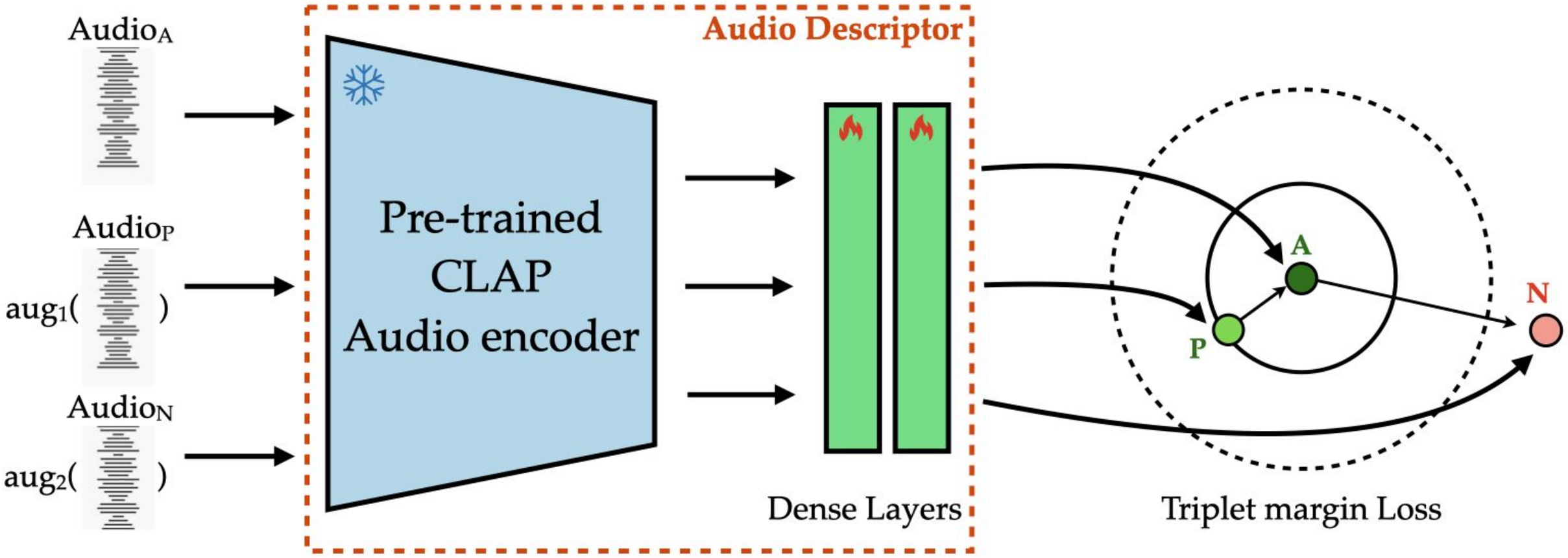
*Zheng, Kaiwen, et al. "DPM-Solver-v3: Improved Diffusion ODE Solver with Empirical Model Statistics." 37 Conference on Neural Information Processing Systems (2023).

Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Reproducibility

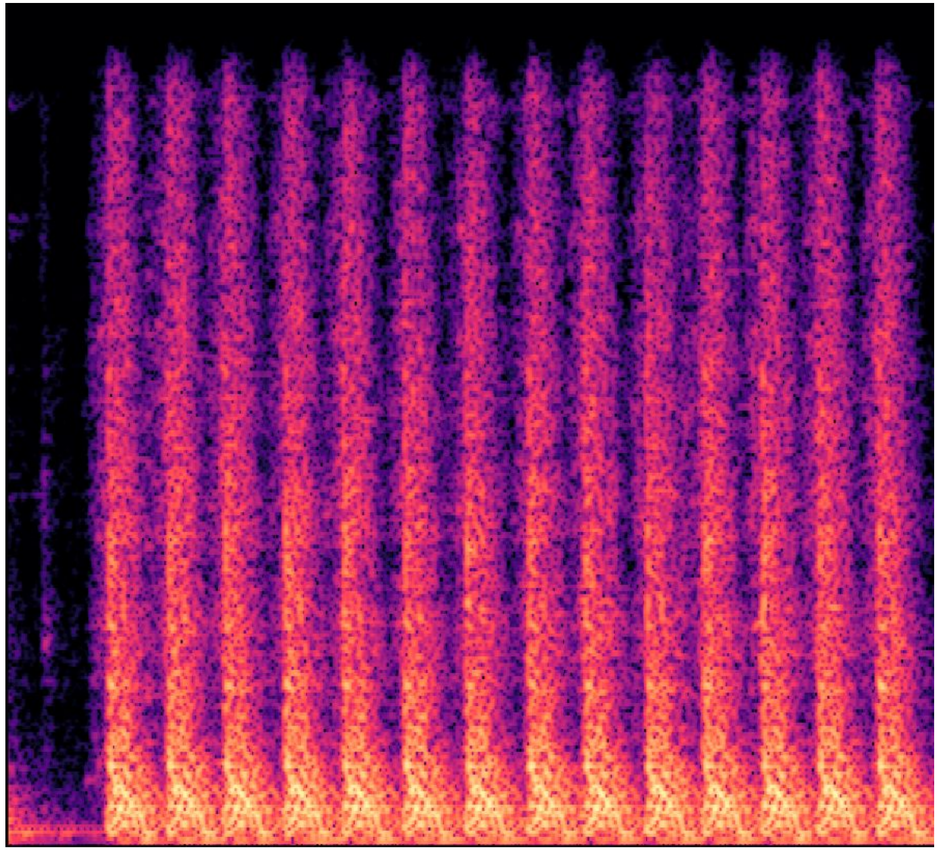


Copy detection (Memorization) in audio diffusion models

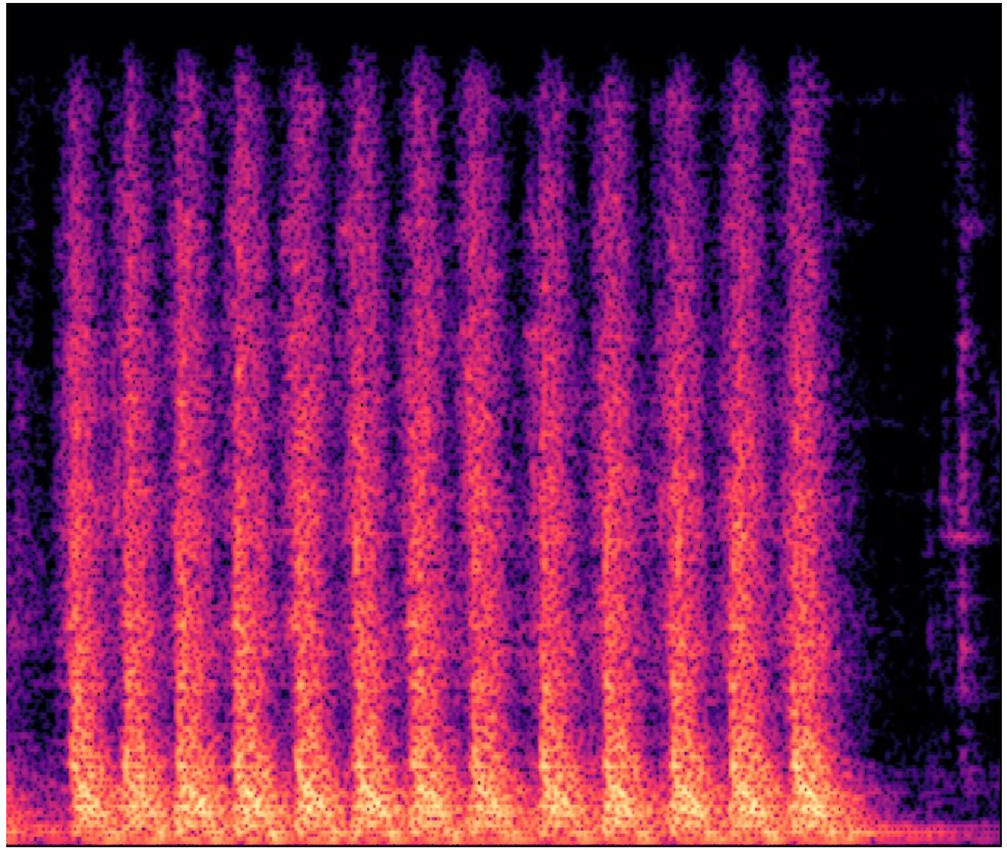


Most similar samples generated by EDMSound

https://agentcooper2002.github.io/EDMSound/



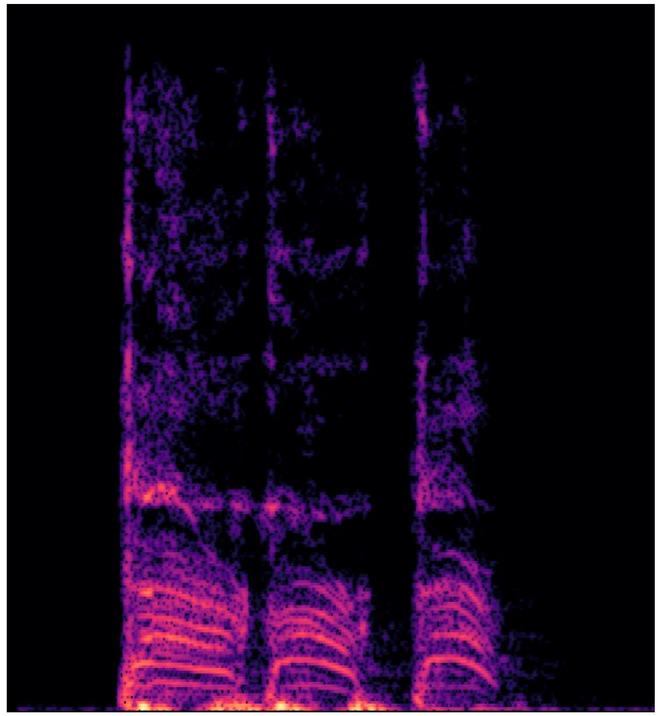
Sample from training set



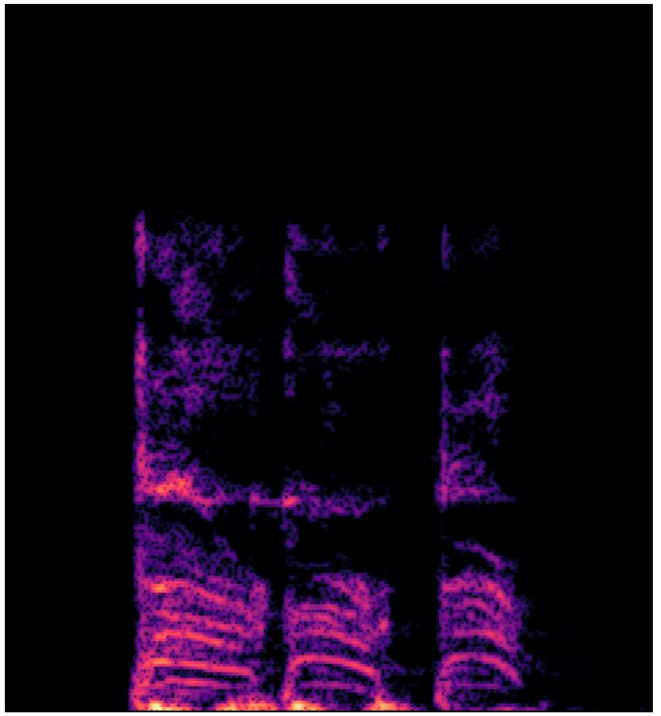
Generated Sample

Most similar samples generated by fine-tuned TANGO

https://agentcooper2002.github.io/EDMSound/



Sample from training set



Generated Sample

Ghosal, Deepanway, et al. "Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model." arXiv preprint arXiv:2304.13731 (2023).

R. Scheibler, et al. "Class-conditioned latent diffusion model for dcase 2023 foley sound synthesis challenge." Technical report, Tech. Rep., June, 2023.

Overview

- EDMSound: Spectrogram based diffusion models
  - Training: Elucidated diffusion model (EDM) framework
  - Sampling: Exponential-integrator based deterministic solver
- Copy detection/memorization issue in diffusion models
  - Fine-tuned CLAP for audio