Generative Diffusion Models for Audio Inpainting

SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Ingegneria dell'informazione, informatica e statistica
Engineering in Computer Science

Andrea Rodriguez
1834937

Advisor
Prof. Danilo Comminiello

Generative Diffusion Models for Audio Inpainting

1. Background: Task, Diffusion models and Spectrograms
2. State-of-the-Art: AudioLDM
3. Selected technique: Tango + RePaint
4. Additional use case: Denoising in communication scenarios

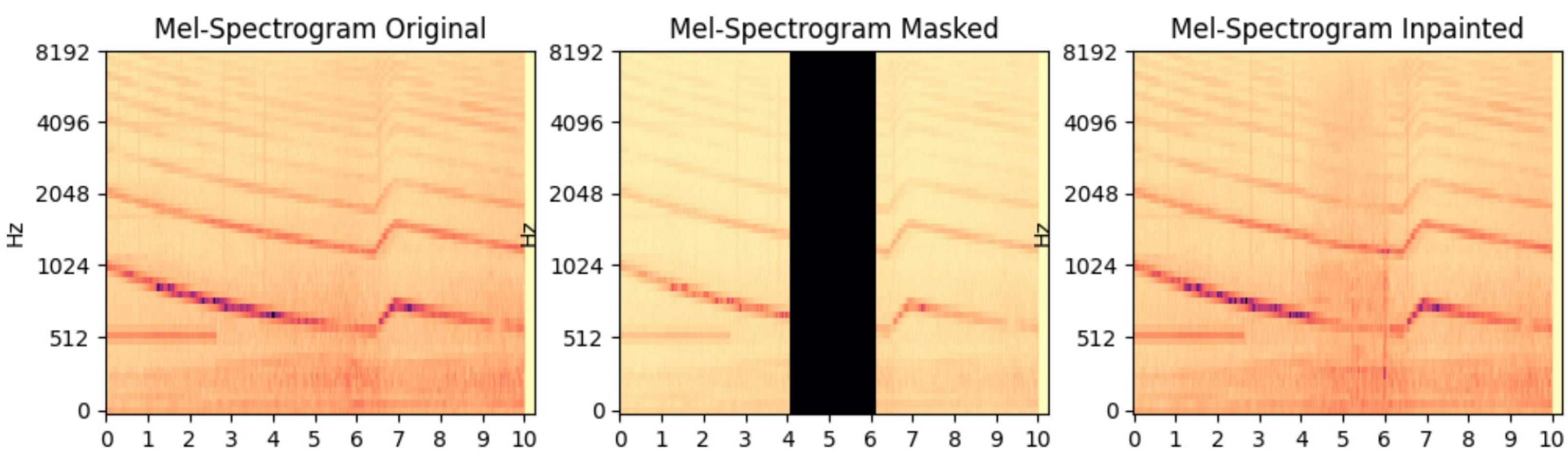
Audio generation and Audio inpainting

Audio Generation

- Produce audio from textual descriptions that is indistinguishable from human-created or real-world audio
- Generate audio samples with similar characteristics to the training data, also showcasing innovative attributes

Audio Inpainting

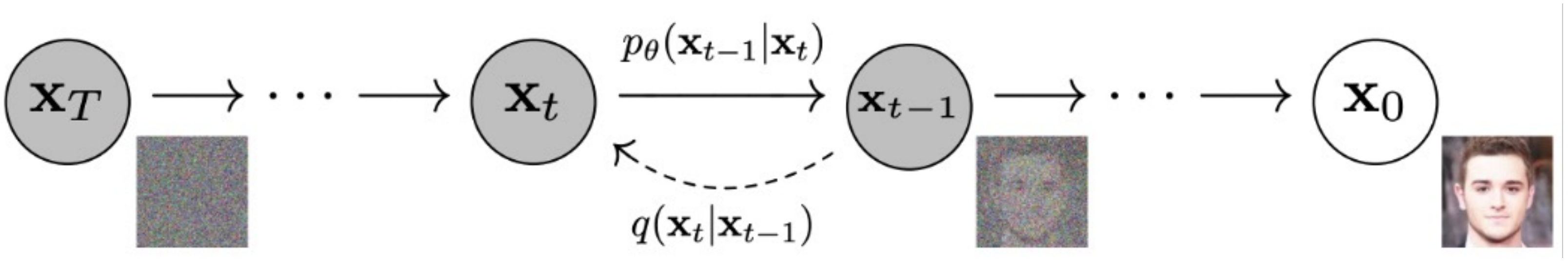
- Reconstruct missing or corrupted portions of audio signals and restore the original audio content



Diffusion models

Main concepts

- Iteratively transform an initial noise distribution into a target distribution
- Generate high-quality samples and perform denoising and inpainting



Forward process: using a variance schedule, small amounts of Gaussian noise are added to the sample in *T* steps

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

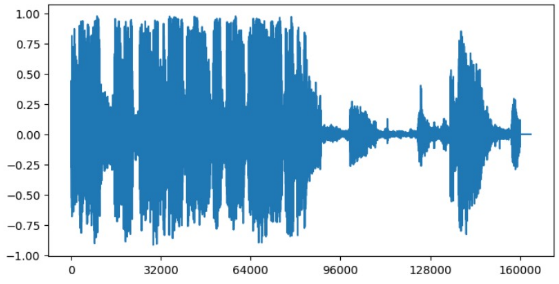
Reverse process: the noise added at each step of the forward process is predicted and removed from initial noise

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I})$$

Spectrograms

A **spectrogram** is a visual representation that displays the frequency content of an audio signal over time

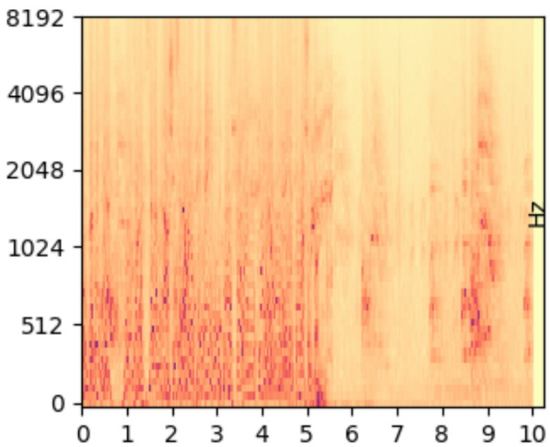
Vector 1 x 160000 (10 seconds)



(a) Audio wave

Sparsity → Computationally demanding and Risk of overfitting

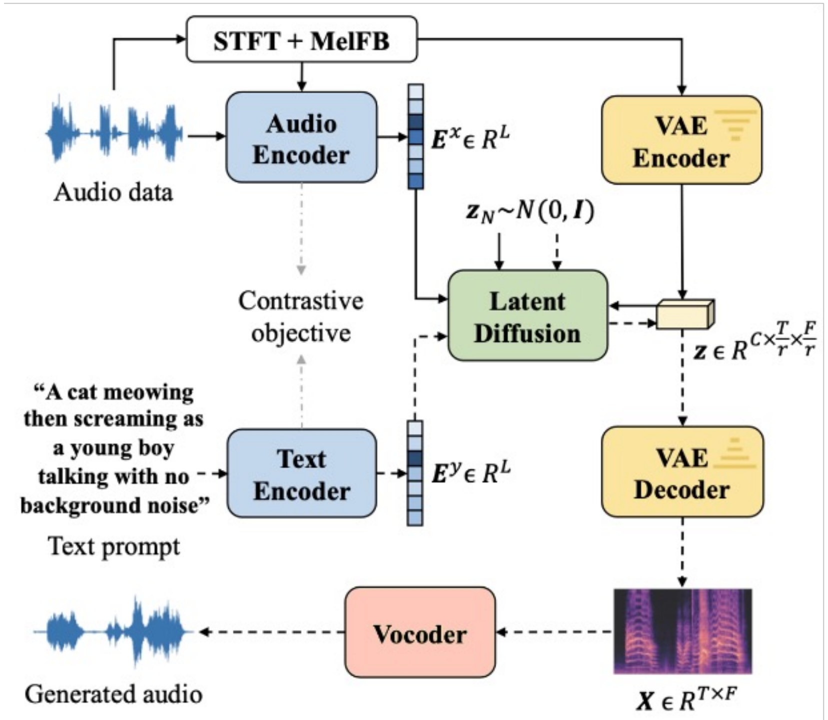
Matrix 1024 x 64 (10 seconds)



(b) Spectrogram

Dense representation → Capture long term dependencies and Efficient training

State-of-the-Art: AudioLDM



- **CLAP** (Contrastive Language-Audio Pretraining): encode audio descriptions and audio clips into a shared audio-text embedding space
- **VAE** (Variational Autoencoder): compress the spectrogram into a compact latent space
- **Latent Diffusion**: the conditioning information is integrated into the feature extraction process
- **Vocoder**: synthesizes the audio waveform from the generated spectrogram

Inpainting:

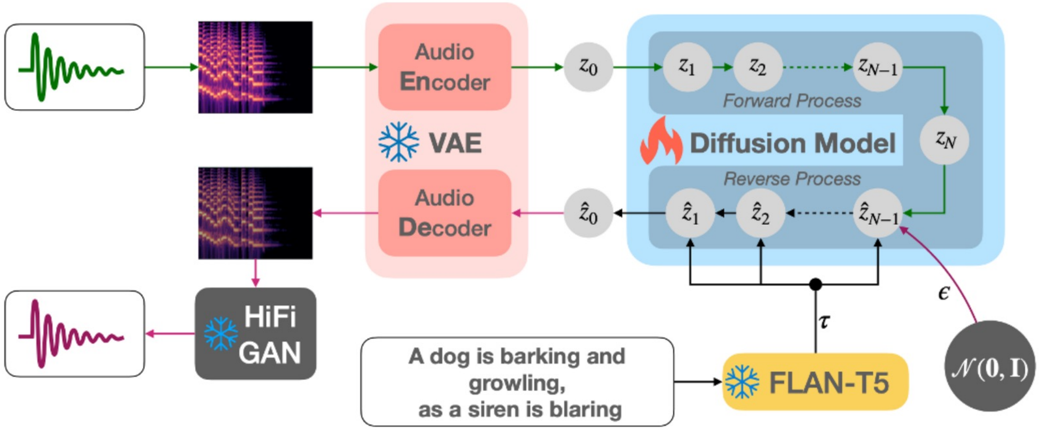
- x^{known} is sampled from the known part
- $x^{unknown}$ is sampled from the model

$$x_{t-1} = m \odot x_{t-1}^{known} + (1-m) \odot x_{t-1}^{unknown}$$

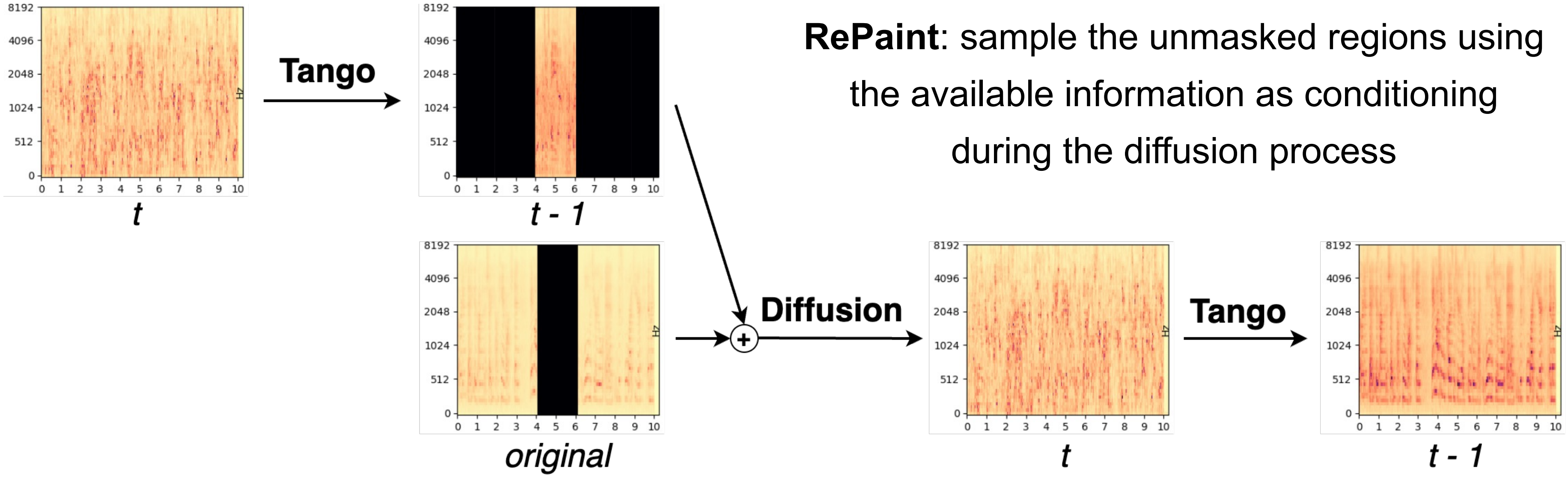
The two components are combined according to the mask *m*

Tango + RePaint

- **FLAN-T5**:
  - Instruction-tuned LLM architecture used as text encoder
  - Trained on a large-scale chain-of-thought and instruction-based dataset



RePaint: sample the unmasked regions using the available information as conditioning during the diffusion process



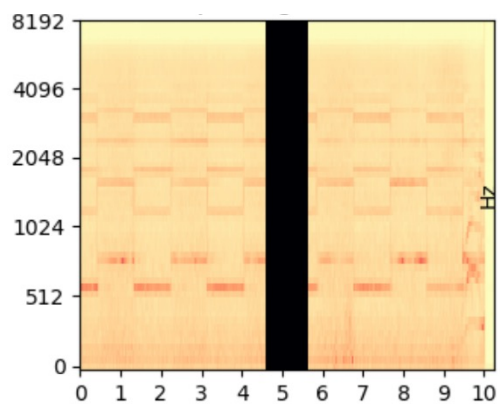
Audio samples and Inference details

Tests were performed using 24 audio clips from the **AudioCaps** dataset

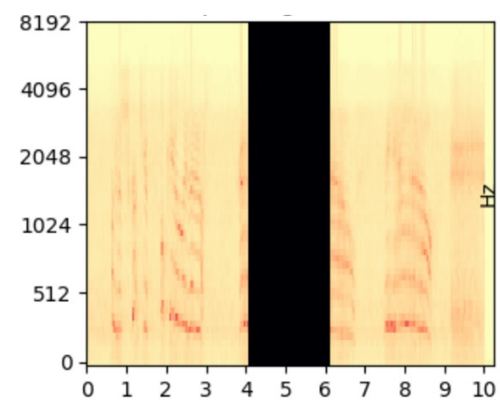
“Female and male are having conversation”

“An emergency vehicles’ siren with a brief male yell”

“Duck quacking repeatedly”



1 second gap

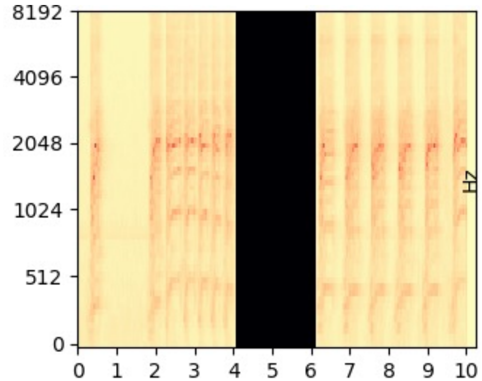


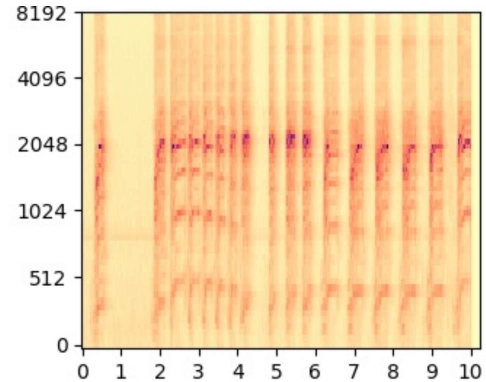
2 seconds gap

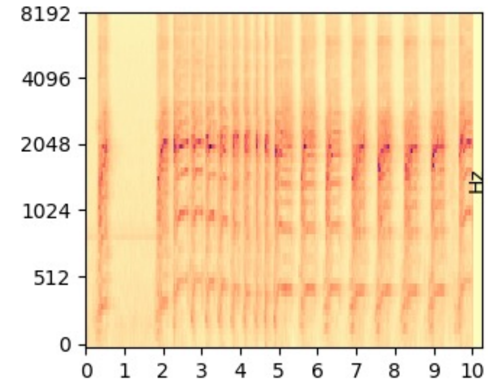
Listen to the audio clips at: https://www.andrearodriguez.it/inpainting

Results

"Duck quacking repeatedly"

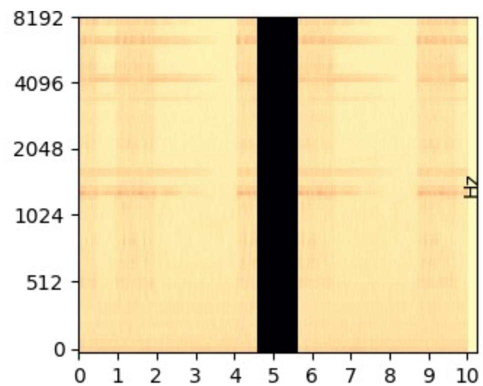


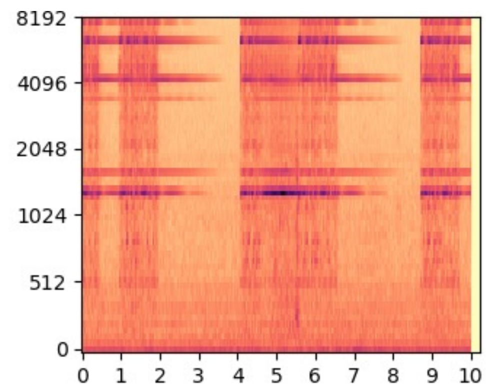


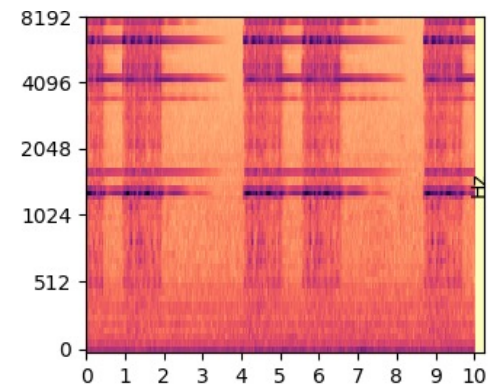


Masked **Inpainted using RePaint** **Original**

"A telephone ringing"







Average metrics (Tango + RePaint)

1 second gap

4.5s - 5.5s	AudioLDM	Tango RePaint
SDR	-3.27	**5.96**
SNR	-0.25	**6.17**
PSNR	39.46	**44.08**
SSIM	98.40	**99.18**

2 seconds gap

4.0s - 6.0s	AudioLDM	Tango RePaint
SDR	-4.97	**1.64**
SNR	-1.45	**2.71**
PSNR	35.44	**39.92**
SSIM	97.84	**98.59**

Metrics computed on **audio**:

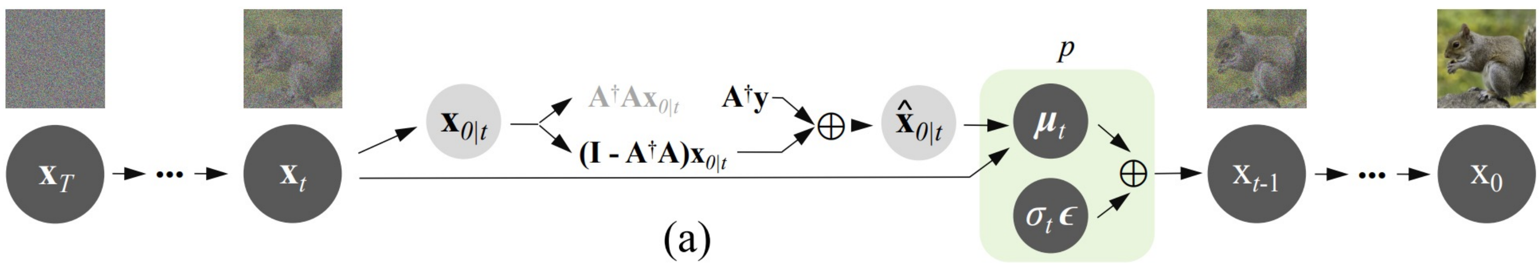
- SDR *Signal Distortion Ratio*
- SNR *Signal Noise Ratio*

Metrics computed on **spectrograms**:

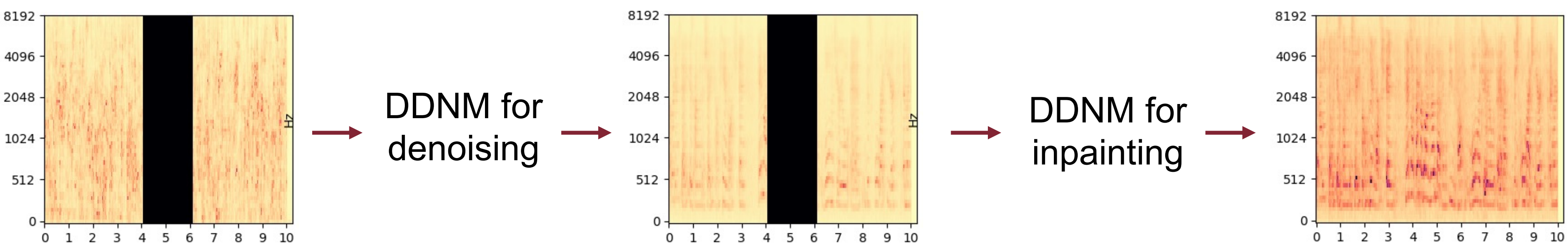
- PSNR *Peak Signal Noise Ratio*
- SSIM *Structural Similarity Index Measure (%)*

Denoising in communication scenarios: DDNM

Tests were performed adding white **Gaussian noise** to audio clips and then trying to remove it

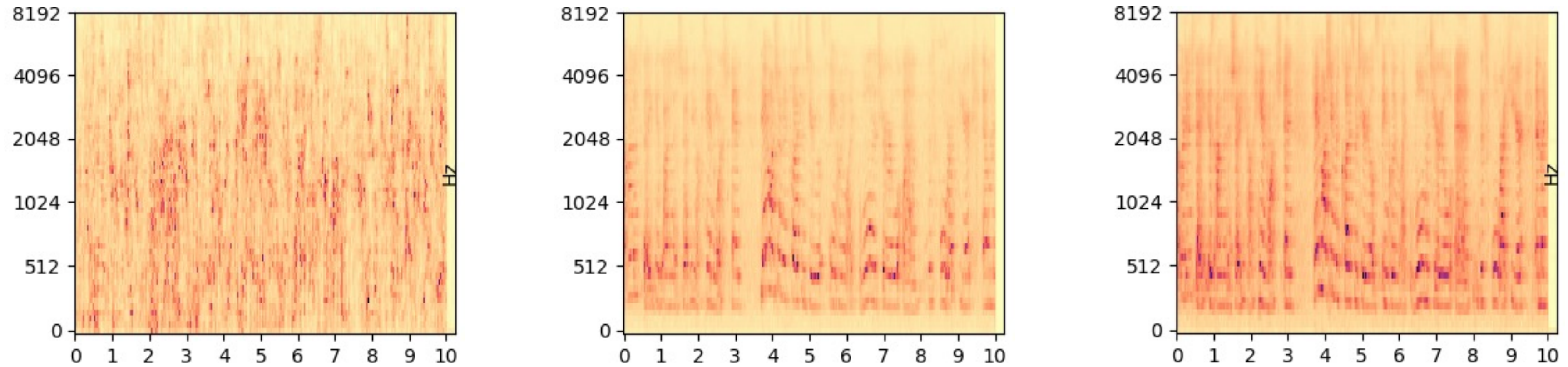


DDNM is a framework for image restoration which refines the null-space content during the reverse diffusion process to perform tasks such as denoising and inpainting



Results

"An adult female is speaking in a quiet environment"



Noisy **Denoised** **Original**

SNR

PSNR 20	Clip 1	Clip 2	Clip 3	Clip 4
Noisy	-9.80	-9.11	-8.61	-10.10
Denoised	**-2.47**	**-3.80**	**-2.64**	**-2.58**

SNR

PSNR 30	Clip 1	Clip 2	Clip 3	Clip 4
Noisy	-3.53	-2.45	-3.54	-3.61
Denoised	**-1.82**	**-1.14**	**-1.32**	**-2.20**

Conclusions

- The proposed combination of **Tango + RePaint** consistently outperforms the baseline results for inpainting with clean audios as input
- The use of **DDNM** enables remarkable denoising capabilities, even in scenarios with substantial levels of noise

and Future works

- Create an **automated communication system** to perform denoising, identify problematic segments and inpaint them
- **Remove conditioning** from text and perform inpainting based only on the known portion of audio

Paper in progress!

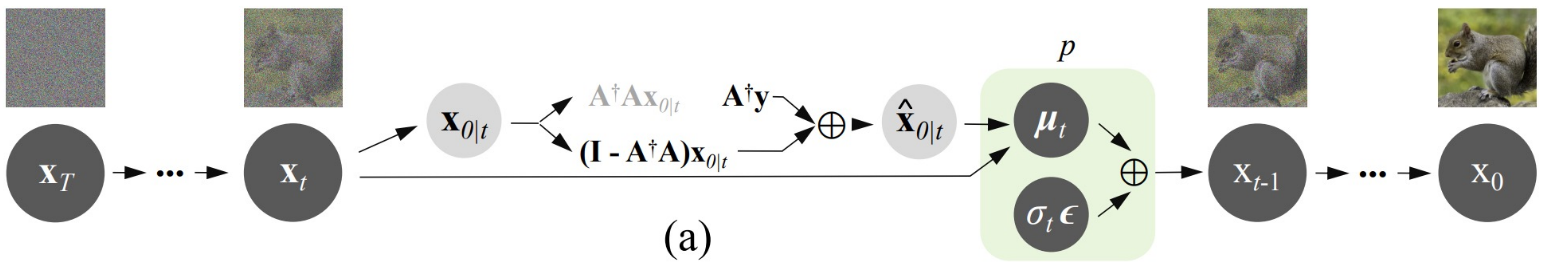
References

1. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. 2020. arXiv: 2006.11239 [cs.LG].
2. Haohe Liu et al. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. 2023. arXiv: 2301.12503 [cs.SD].
3. Deepanway Ghosal et al. Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model. 2023. arXiv: 2304.13731 [eess.AS].
4. Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. 2022. arXiv: 2212.00490 [cs.CV].
5. Andreas Lugmayr et al. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. 2022. arXiv: 2201.09865 [cs.CV].
6. Chris Dongjoo Kim et al. “AudioCaps: Generating Captions for Audios in The Wild”. In: Proceedings of the 2019 Conference of the North American Chapter of he Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 119–132. doi: 10.18653/v1/N191011. url: https://aclanthology.org/N19-1011.

Backup slides

DDNM and DDNM+

DDNM (Denoising Diffusion Null-Space Model) is a framework for image restoration which refines the null-space content during the reverse diffusion process to produce results that satisfy data consistency and realism

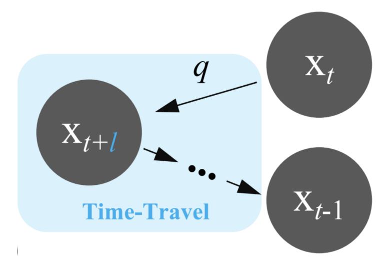


(a)

Task: reconstruct $\hat{x}$ from y where $y = Ax$

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \mathcal{Z}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\sqrt{1-\bar{\alpha}_t}\right)$
4: $\hat{\mathbf{x}}_{0|t} = \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}_{0|t}$
5: $\mathbf{x}_{t-1} \sim p(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_{0|t})$
6: **return** $\mathbf{x}_0$

DDNM+: Through the time-travel trick we generate a better "past", which in turn leads to a better "future"



RePaint and RePaint⁺

RePaint: sample the unmasked regions using the available information as conditioning during the diffusion process

$$x_{t-1}^{known} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$$

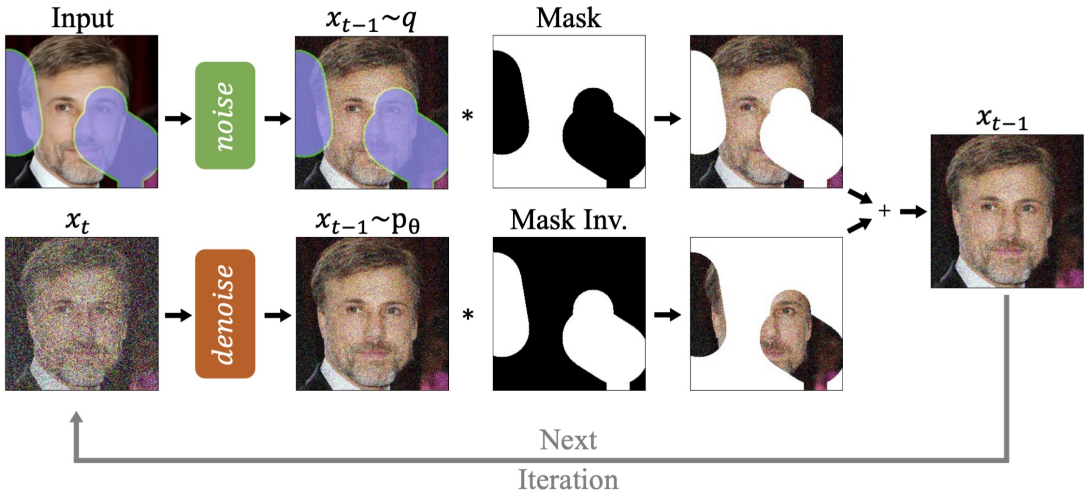
$$x_{t-1}^{unknown} \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

$$x_{t-1} = m \odot x_{t-1}^{known} + (1-m) \odot x_{t-1}^{unknown}$$

x_{t-1} is diffused back to x_t

$$x_t \sim \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

the denoising step is performed again



The model has **more time** to effectively incorporate the provided information with the generated part

RePaint⁺: The latent representation is diffused back to multiple previous steps and then all of them are sequentially performed again

Inference times

	1 Clip	Batch of 8 Clips
AudioLDM	1	4
Tango	10	40
Tango + DDNM	10	40
Tango + DDNM$^+$	120	480
Tango + RePaint	100	400
Tango + RePaint$^+$	100	400

Inference times in minutes using
one GPU NVIDIA Tesla T4

Listen to the audio clips at: https://www.andrearodriguez.it/inpainting

Metrics

- SDR = Signal Distortion Ratio
- SNR = Signal Noise Ratio
- PSNR = Peak Signal Noise Ratio
- SSIM = Structural Similarity Index Measure

1 second gap

4.5-5.5	**AudioLDM**	**Tango**	**Tango DDNM**	**Tango DDNM$^+$**	**Tango RePaint**	**Tango RePaint$^+$**
SDR	-3.27	5.48	5.03	5.47	**5.96**	4.97
SNR	-0.25	5.73	5.39	5.71	**6.17**	5.28
PSNR	39.46	43.35	42.44	43.22	**44.08**	42.38
SSIM	98.40	99.25	99.21	**99.28**	99.18	99.14

2 seconds gap

4-6	**AudioLDM**	**Tango**	**Tango DDNM**	**Tango DDNM$^+$**	**Tango RePaint**	**Tango RePaint$^+$**
SDR	-4.97	1.48	1.53	**1.99**	1.64	1.48
SNR	-1.45	2.82	**2.90**	2.21	2.71	2.02
PSNR	35.44	39.74	39.85	38.61	**39.92**	38.56
SSIM	97.84	98.34	98.46	98.45	**98.59**	98.56

* SSIM values are multiplied by 100