Audio Editing with Non-Rigid Text Prompts

Francesco Paissan[1,2], Luca Della Libera[3,2], Zhepei Wang[4], Paris Smaragdis[4], Mirco Ravanelli[3,2], Cem Subakan[5,3,2]

[1]Fondazione Bruno Kessler, [2]Mila-Quebec AI Institute, [3]Concordia University, [4]University of Illinois at Urbana-Champaign, [5]Laval University

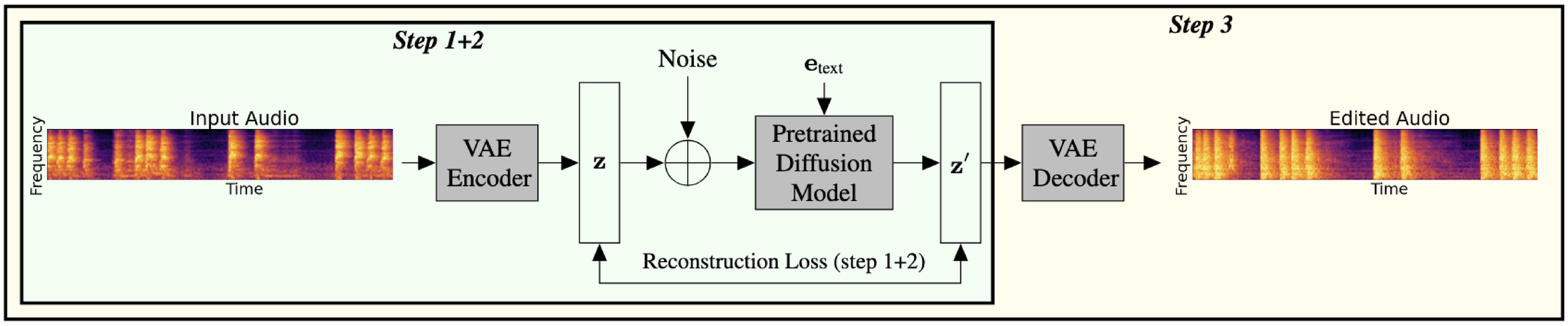
Summary

- **Task:** Audio Editing with Freeform text
- **Methodology:** Latent Variable Interpolation for Text-Conditioned Diffusion Models

What is Audio Editing?

- Fidelity to Text
- Fidelity to Input Audio
- Check out some samples samples!

The Method



Step 1: Embedding Optimization

$$\mathbf{e}_{\text{opt}} = \min_{\mathbf{e}_{\text{text}}} \mathbb{E}_{t,\epsilon}[\epsilon - f_\theta(\mathbf{z}_t, t, \mathbf{e}_{\text{text}})]$$

We learn the text embedding vector ($\mathbf{e}_{\text{opt}}$) corresponding to the original audio.

Step 2: Fine Tuning

$$\min_{\theta} \mathbb{E}_{t,\epsilon}[\epsilon - f_\theta(\mathbf{z}_t, t, \mathbf{e}_{\text{opt}})]$$

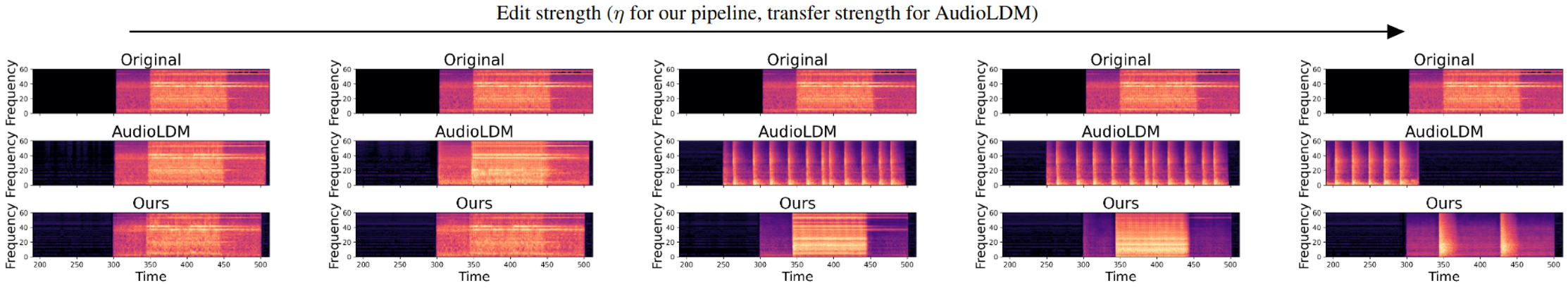
We finetune the diffusion model to reconstruct the original audio with $\mathbf{e}_{\text{opt}}$.

Step 3: Interpolation

$$\mathbf{e}_{\text{int}} = \eta \mathbf{e}_{\text{target}} + (1-\eta)\mathbf{e}_{\text{opt}}$$

We generate the edited audio by conditioning the diffusion model with $\mathbf{e}_{\text{int}}$.

Experimental Results



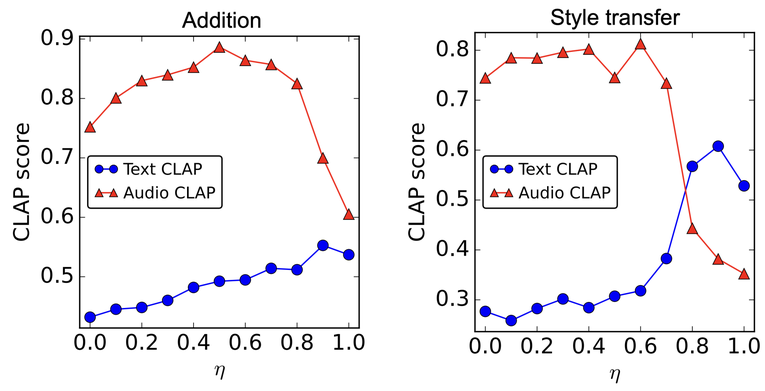
Quantitative Results

We used CLAP to measure the alignment between the target text and the edited audio and between the original and edited audio segments:

$$\text{Text-Audio Similarity} = \frac{\hat{\mathbf{h}}_a^T \mathbf{h}_t}{||\hat{\mathbf{h}}_a|| \cdot ||\mathbf{h}_t||} \qquad \text{Audio-Audio Similarity} = \frac{\hat{\mathbf{h}}_a^T \mathbf{h}_a}{||\hat{\mathbf{h}}_a|| \cdot ||\mathbf{h}_a||}$$

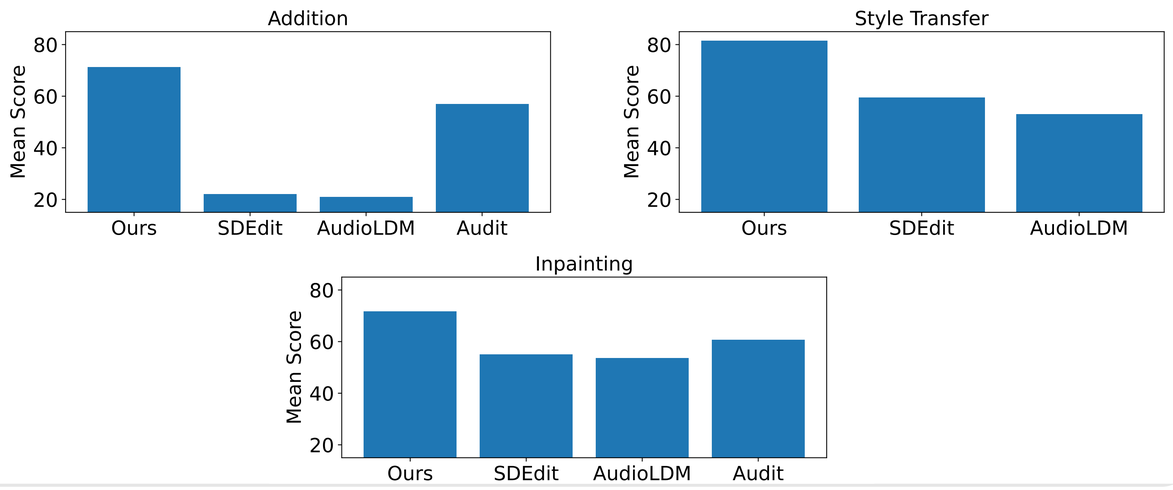
Edit type	Edit pipeline	Sum	Text	Audio
Addition	AudioLDM	1.229	0.524	0.705
	AUDIT	0.985	0.208	0.777
	SDEdit	0.925	0.237	0.688
	Ours	1.366	0.544	0.822
	Ours (LORA)	**1.407**	0.524	0.883
Inpainting	AudioLDM	1.465	0.544	0.921
	AUDIT	0.912	0.223	0.689
	SDEdit	0.979	0.280	0.699
	Ours	1.472	0.633	0.839
	Ours (LORA)	**1.499**	0.645	0.854
Style Transfer	AudioLDM	0.970	0.260	0.710
	SDEdit	0.651	0.156	0.495
	Ours	1.141	0.460	0.681
	Ours (LORA)	**1.161**	0.465	0.696

Fidelity vs Editing Strength Tradeoff



User study

- We asked 17 users their opinion on 9 samples.
- We compared with SDEdit, AudioLDM and Audit.



Conclusion

- We propose a method to perform freeform text edits of audio samples.
- We provide an optimization strategy to improve the computational complexity of our approach.
- Our edits consistently outperform current editing baselines.

Project Webpage and Paper