GenAI for Sound Design

April 7th, 2025

SALMA Workshop @ ICASSP25

Oriol (Uri) Nieto (he/they)

onieto@adobe.com

Artwork by Daniel Mercadante

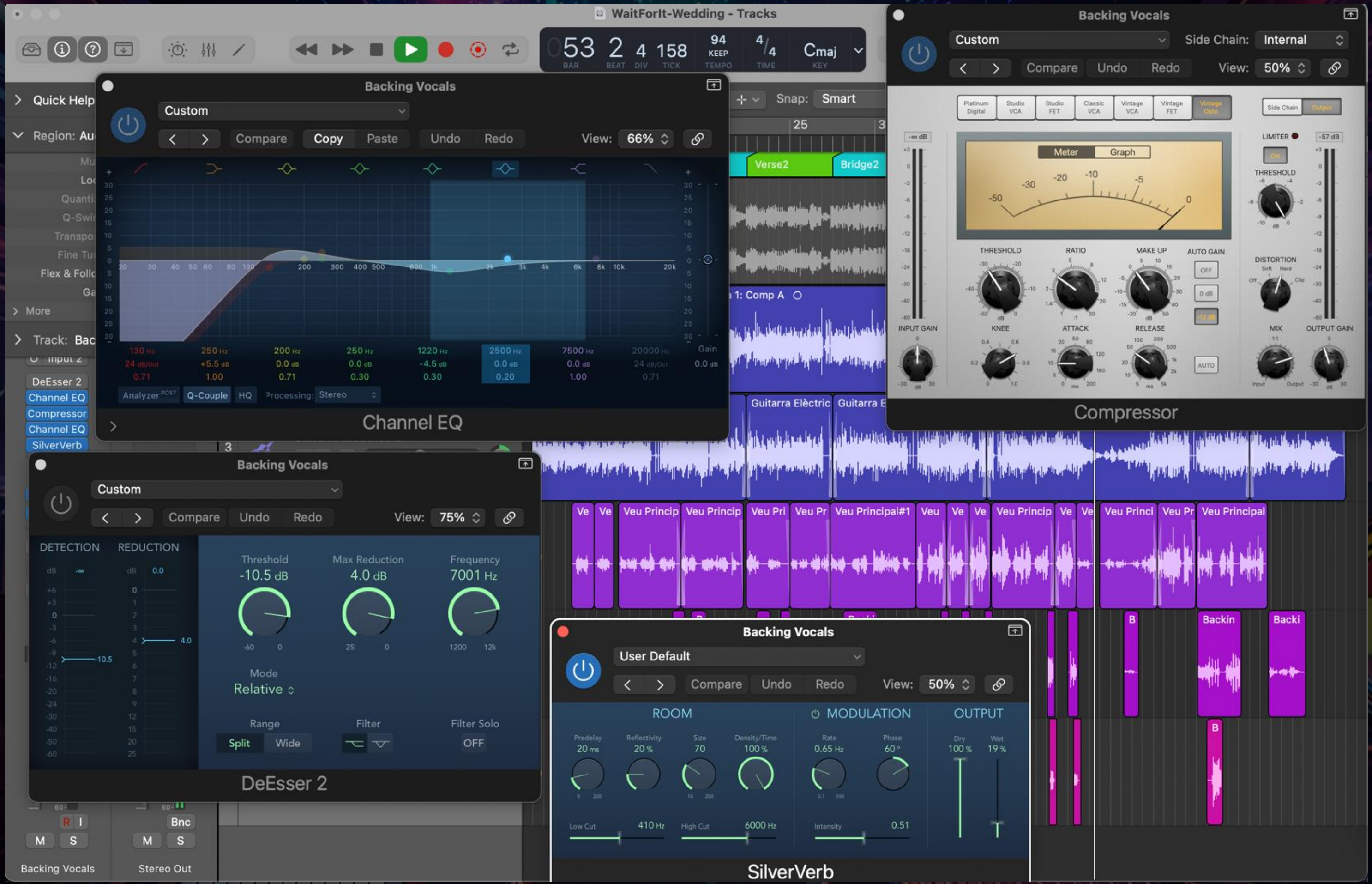
Sound Design

Art and practice of creating audio elements for various media, including films, television, video games, theater, etc.

Sound Design

Art and practice of creating audio elements for various media, including films, television, video games, theater, etc.

Sound Design



Threshold -10.5 dB
Max Reduction 4.0 dB
Frequency 7001 Hz
Mode Relative
Range Split Wide
Filter
Filter Solo OFF

Predelay 20 ms
Reflectivity 20 %
Size 70
Density/Time 100 %
Rate 0.65 Hz
Phase 60 °
Dry 100 %
Wet 19 %
Low Cut 410 Hz
High Cut 6000 Hz
Intensity 0.51

130 Hz 24 dB/Oct 0.71
250 Hz +5.5 dB 1.00
200 Hz 0.0 dB 0.71
250 Hz 0.0 dB 0.30
1220 Hz -4.5 dB 0.30
2500 Hz 0.0 dB 0.20
7500 Hz 0.0 dB 1.00
20000 Hz 24 dB/Oct 0.71
Gain 0.0 dB

Art and practice of creating audio elements for various media, including films, television, video games, theater, etc.

Sound Design AI Group (SODA)

AdobeResearch

CAVAORG

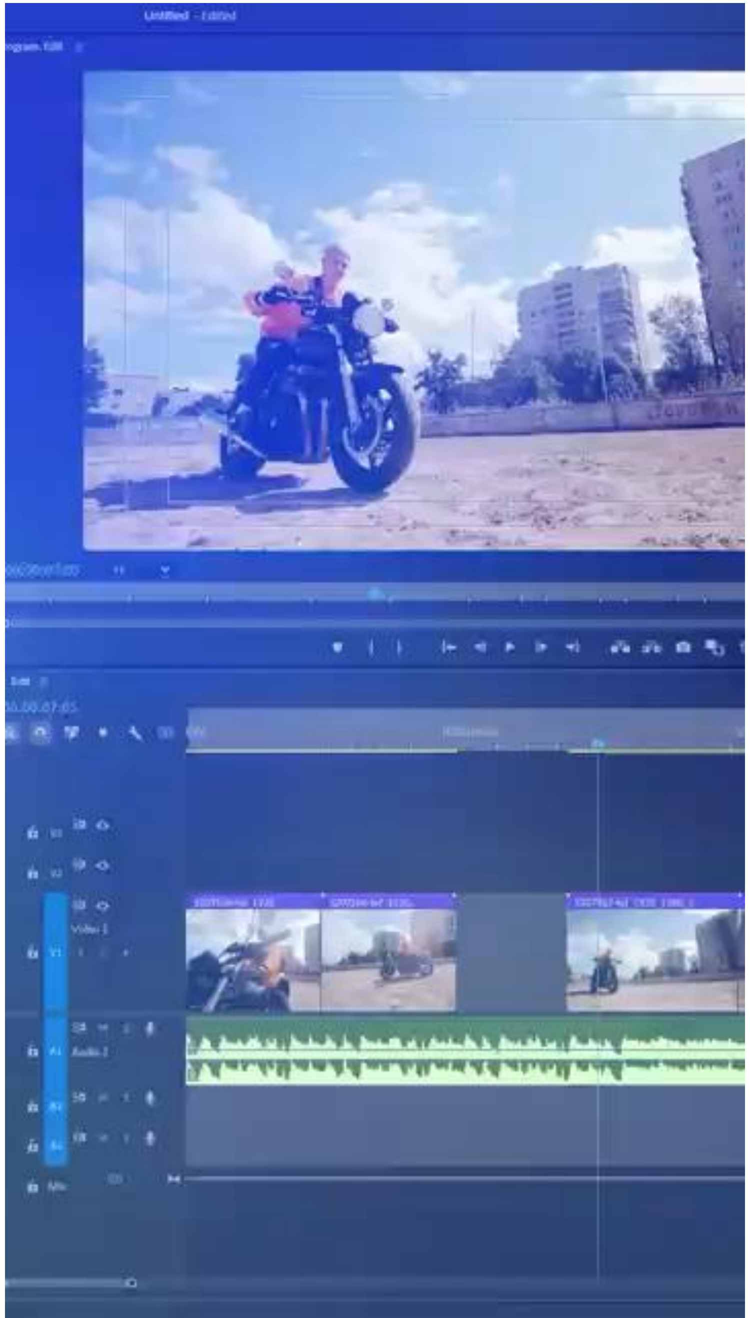
The SODA Team

Justin Salamon

Prem Seetharaman

Oriol Nieto

Generative Extend in Premiere Pro



https://www.youtube.com/shorts/PY4PKBI2Zac

Adobe

Outline

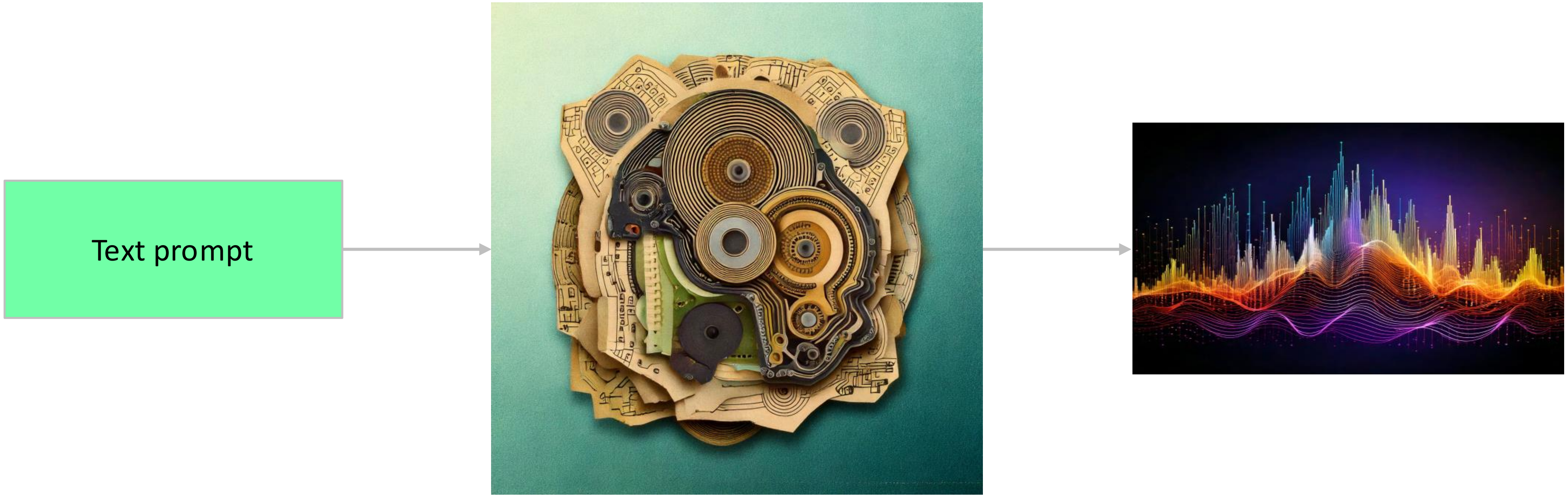
Diffusion Models for Audio Generation

SILA

Large Audio Language Models

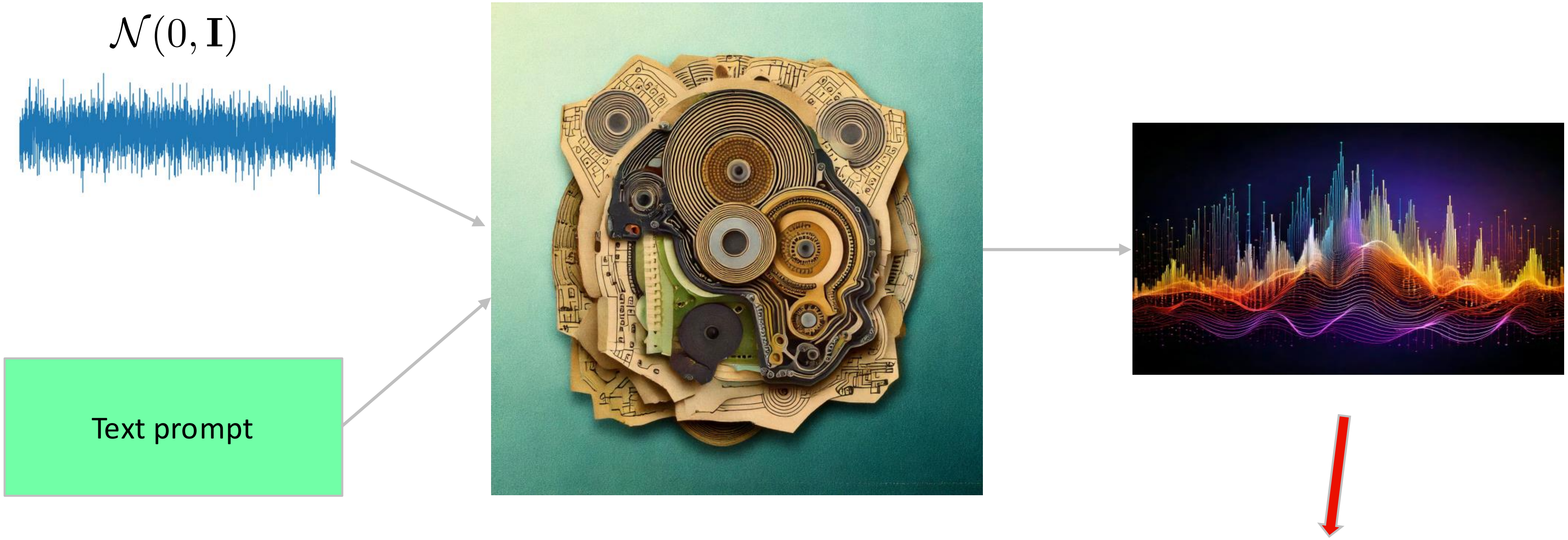
MultiFoley

Diffusion Models for Audio Generation



Text prompt

Diffusion Models for Audio Generation

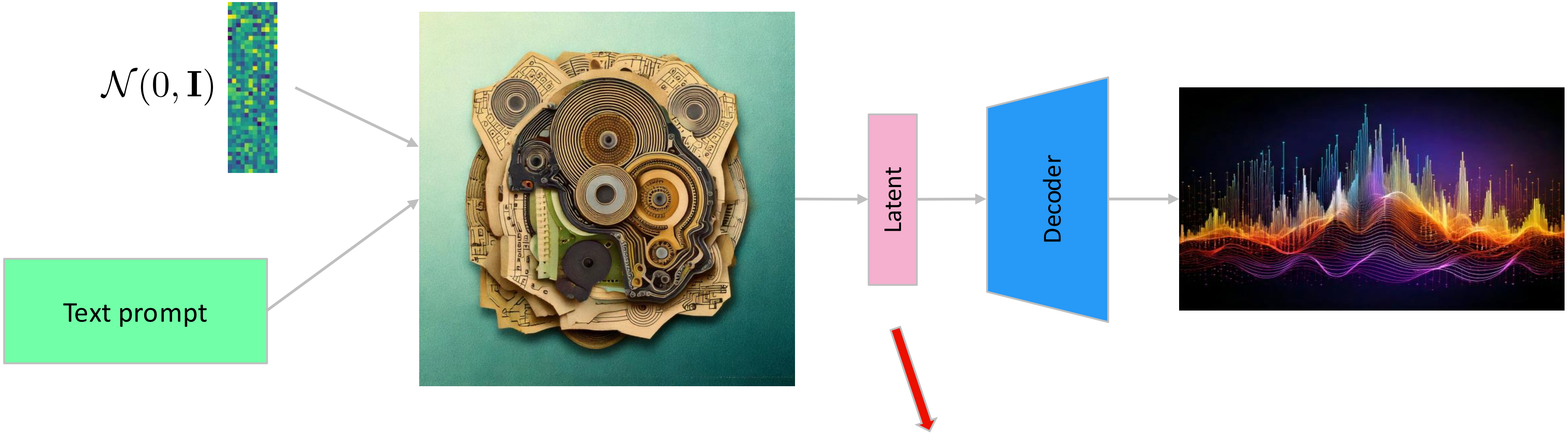


$\mathcal{N}(0, \mathbf{I})$

Text prompt

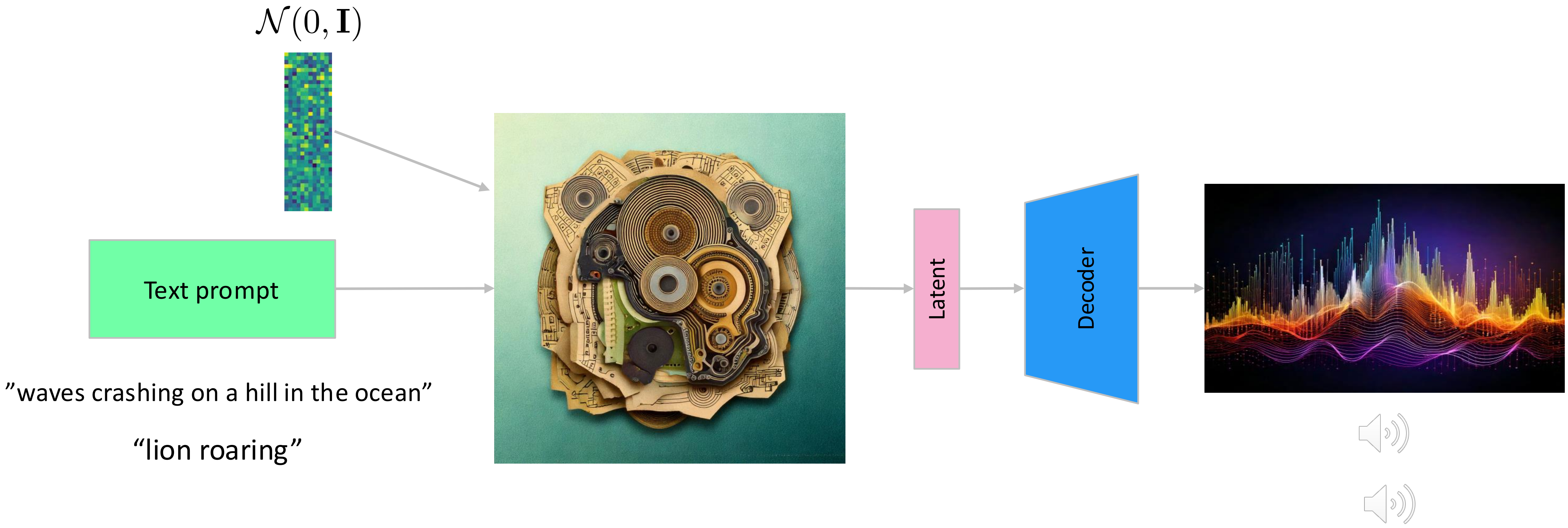
High quality audio is very high dimensional (~48kHz!)

Latent Diffusion Models for Audio Generation



- Audio **latent** space is much more compact (~40Hz)
- E.g., VAEs [1], RVQ [2], DAC [3]

Examples of Latent Diffusion Models for Audio Gen

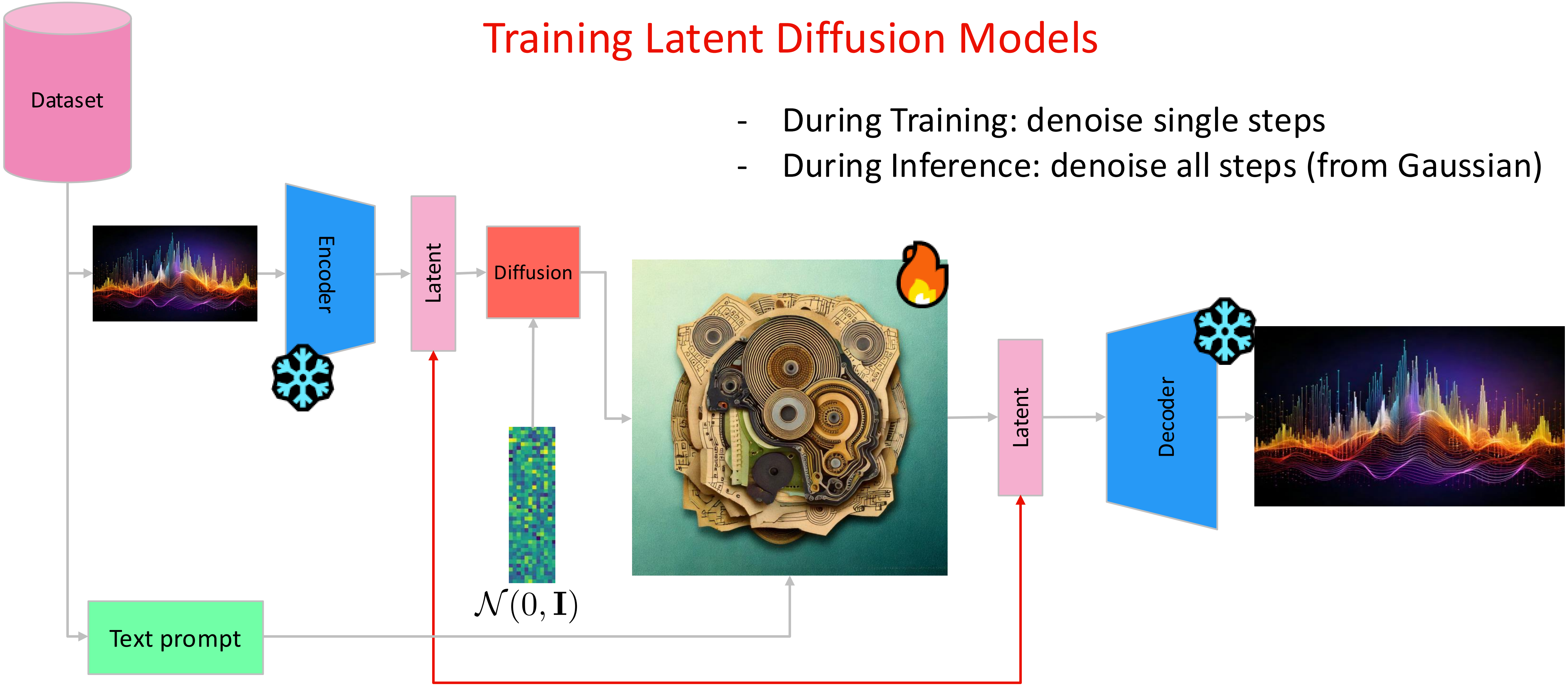


”waves crashing on a hill in the ocean”

“lion roaring”

Training Latent Diffusion Models

- During Training: denoise single steps
- During Inference: denoise all steps (from Gaussian)



$$\mathcal{L}_\theta = \mathbb{E}[||\mathbf{z} - \hat{\mathbf{z}}||^2]$$

Outline

Diffusion Models for Audio Generation

SILA

Large Audio Language Models

MultiFoley

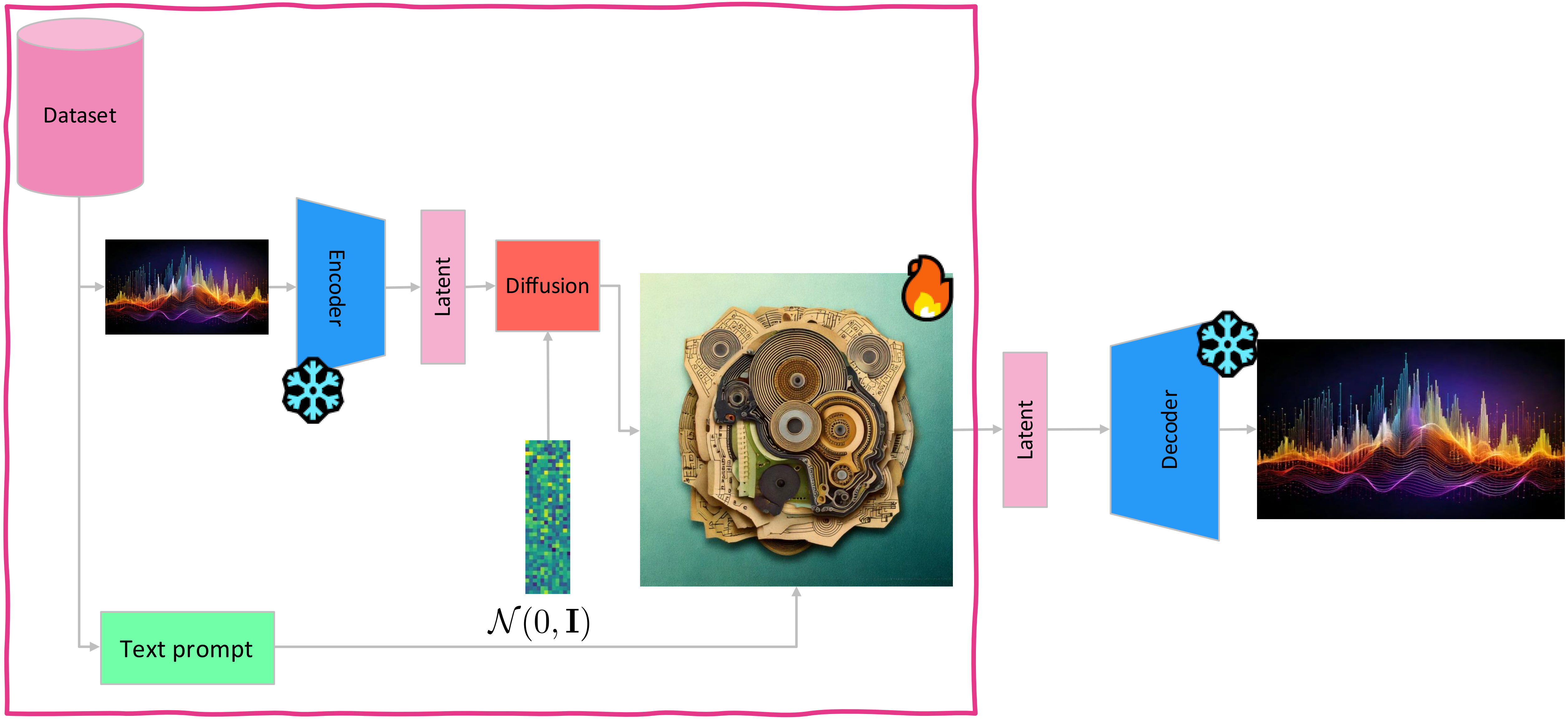
Adobe

SILA: Signal-to-Language Augmentation for Enhanced Control in Text-to-Audio Generation

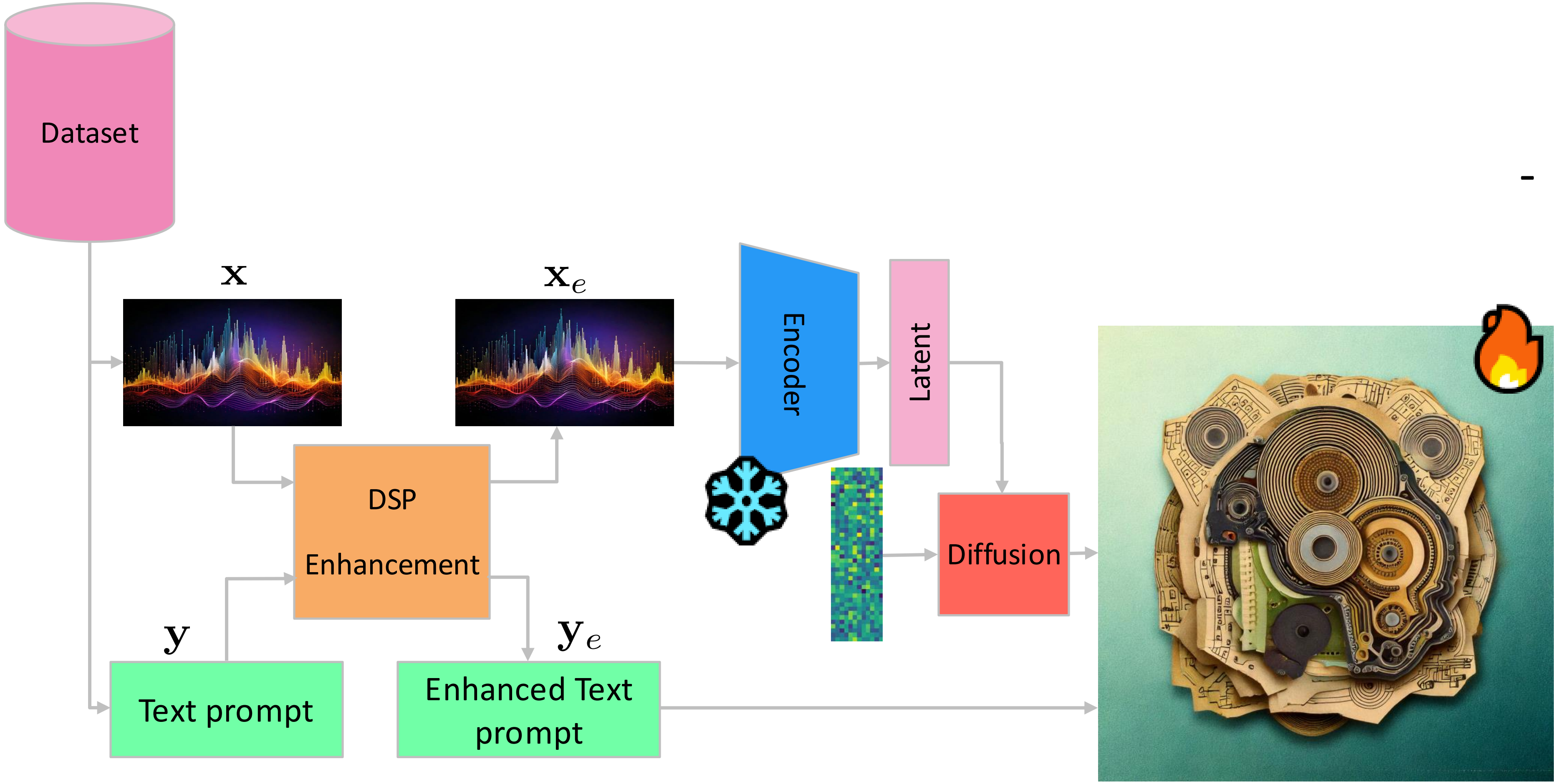
- Text-based models have limited control
- Hard to obtain desired results with a single text prompt
- Can we add control with minimal impact in architecture/performance?

Kumar, S., Seetharaman, P., Salamon, J., Manocha, D., Nieto, O., SILA: Signal-to-Language Augmentation for Enhanced Control in Text-to-Audio Generation. Submitted to IEEE Signal Processing Letters, 2025

SILA: Signal-to-Language Augmentation for Enhanced Control in Text-to-Audio Generation



SILA: Signal-to-Language Augmentation for Enhanced Control in Text-to-Audio Generation



- DSP Enhancement:
  - Loudness
  - Pitch
  - Reverb
  - Noise
  - Brightness
  - Fade
  - Duration

Adobe

SILA: Signal-to-Language Augmentation for Enhanced Control in Text-to-Audio Generation

- Signal

  - Volume (LKFS): -10
  - Brightness (SC): 65
  - Reverb: Add a lot
  - …

- Signal output

- Language

- Original prompt:
  - "A thunder echoes through the sky"

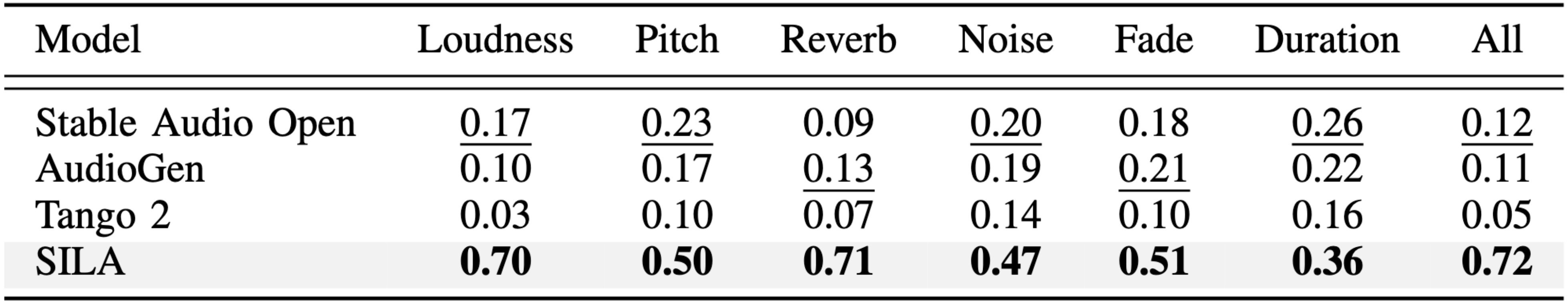
- + ", & loudness: very loud"
- + ", & brightness: bright"
- + ", & reverb: very wet"
- …

- SILA prompt:
  - "A thunder echoes through the sky, & loudness: very loud, & brightness: bright, & reverb: very wet, …"

Adobe

SILA: Signal-to-Language Augmentation for Enhanced Control in Text-to-Audio Generation

- Perceptual Evaluation Results (22 subjects)



Model	Loudness	Pitch	Reverb	Noise	Fade	Duration	All
Stable Audio Open	0.17	0.23	0.09	0.20	0.18	0.26	0.12
AudioGen	0.10	0.17	0.13	0.19	0.21	0.22	0.11
Tango 2	0.03	0.10	0.07	0.14	0.10	0.16	0.05
SILA	**0.70**	**0.50**	**0.71**	**0.47**	**0.51**	**0.36**	**0.72**

Adobe

SILA: Signal-to-Language Augmentation for Enhanced Control in Text-to-Audio Generation

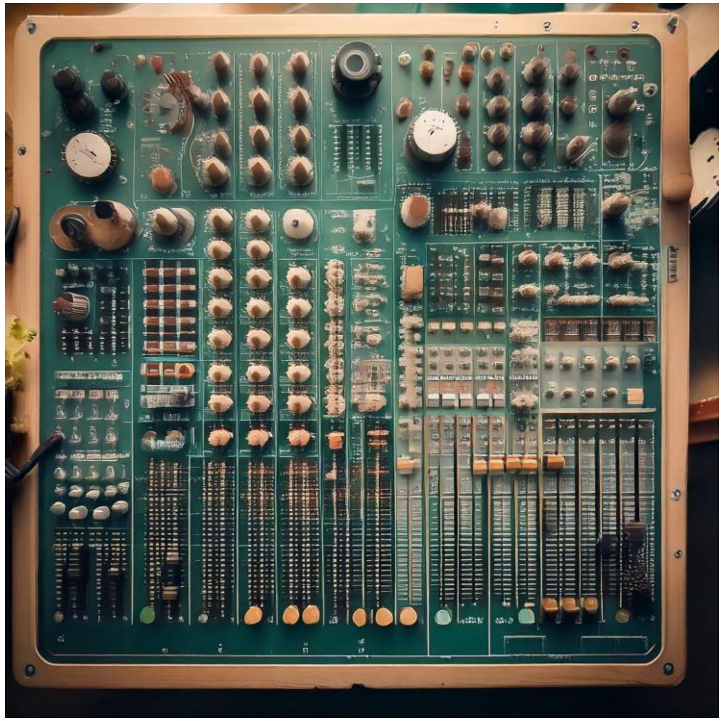
Examples

“The deep rumble of the storm echoes through the sky, & loudness: soft” 🔊

“The deep rumble of the storm echoes through the sky, & loudness: very loud” 🔊

“A dog barking nearby, & reverb: dry” 🔊

“A dog barking nearby, & reverb: wet” 🔊



SILA: Signal-to-Language Augmentation for Enhanced Control in Text-to-Audio Generation

- Added control across several acoustic features

- Highly efficient
  - No added computation during inference

- Model agnostic

Outline

Diffusion Models for Audio Generation

SILA

Large Audio Language Models

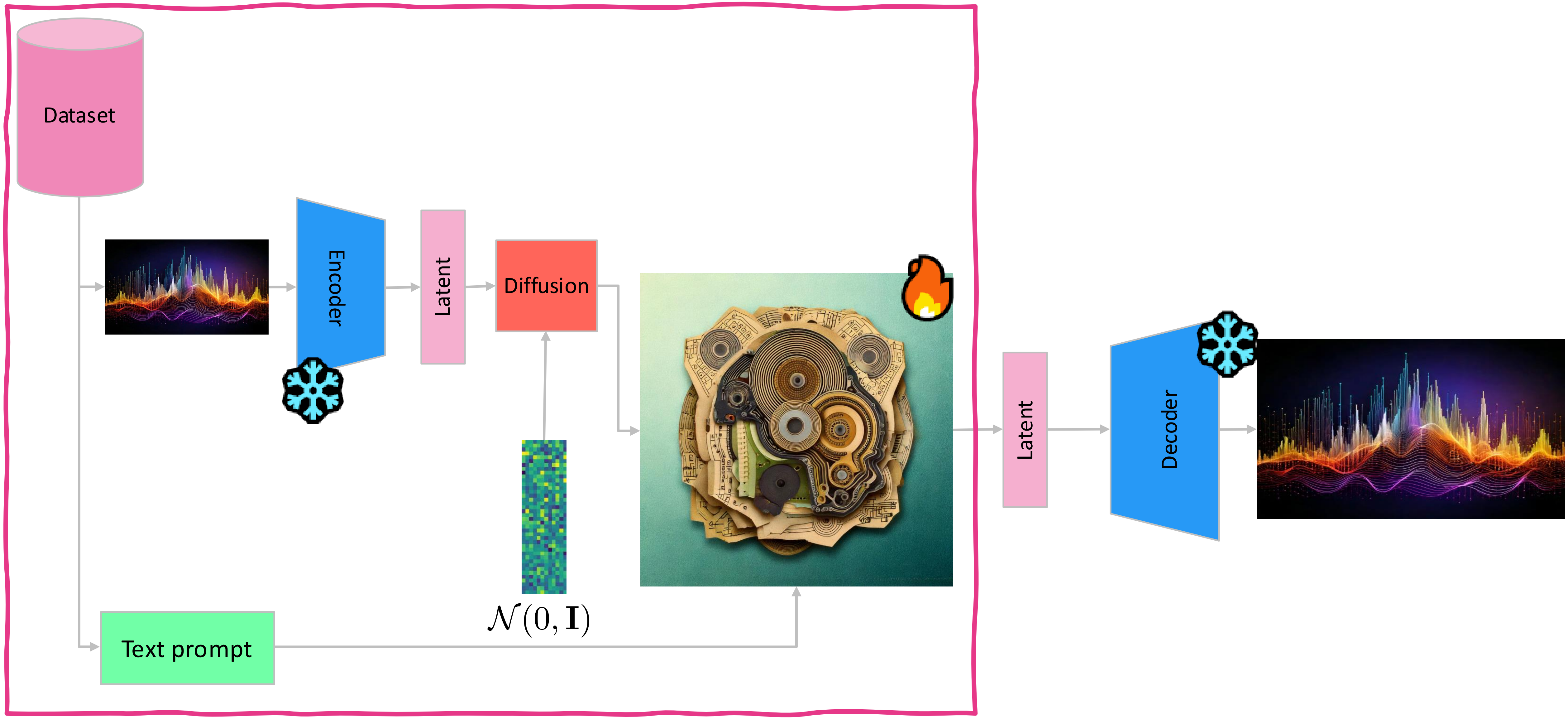
MultiFoley

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

- Perfect synchronization with video can be tedious
- Can we use videos as an additional condition for the generation?
- How about a combination of **text, audio, and video** as conditions?

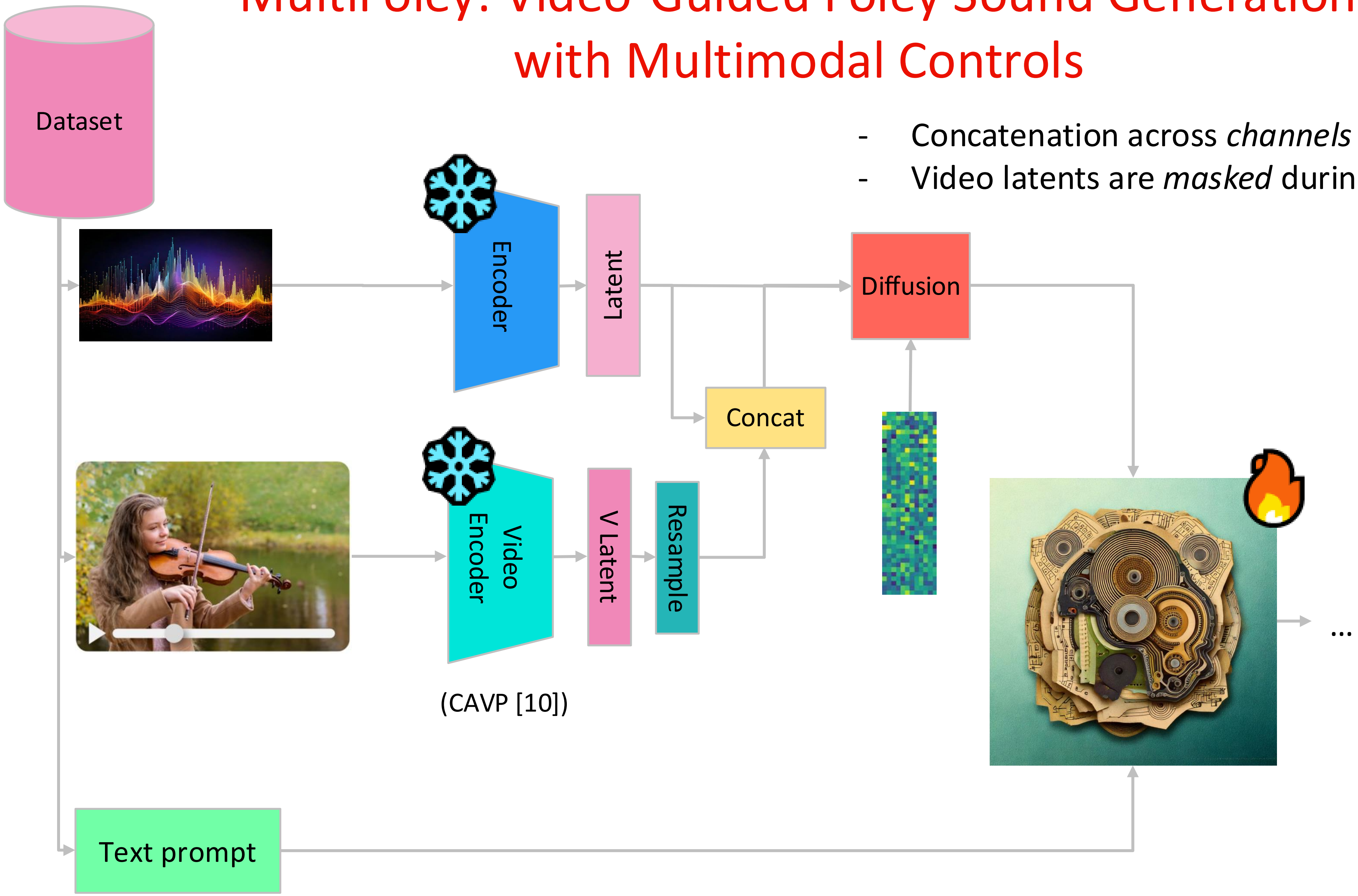
Chen, Z., Seetharaman, P., Russell, B., Nieto, O., Bourgin, D., Owens, A., Salamon, J., Video-Guided Foley Sound Generation with Multimodal Controls, Submitted to IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2025

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls



MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

- Concatenation across *channels* dim
- Video latents are *masked* during diff (ie, no video gen)



Adobe

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

“Bird Chirping”

Adobe

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

“Rooster Crowing”

Adobe

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

“Male Speaking”

Adobe

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

“Typewriter”

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

“Piano”

Adobe

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

“Cello”

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

“Erhu”

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

“Chainsaw”

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

Given this reference dog bark audio

We generate sound for this silent video

MultiFoley: Video-Guided Foley Sound Generation with Multimodal Controls

- Method to generate audio from video
- Multimodal control: audio, video, and text!
- High quality output even when trained on low-quality video dataset (VGGSound)

Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations

Thursday at 11:30h: Sound Generation and Synthesis II

Adobe

Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations

Thursday at 11:30h: Sound Generation and Synthesis II

Adobe

Outline

Diffusion Models for Audio Generation

SILA

Large Audio Language Models

MultiFoley

Adobe

Large Audio-Language Models and GenAI for Sound Design?

- Lack of large-scale high-quality training data
  - “Fine-grained” synthetic data

- End-to-End LALMs with Audio Generation
  - Diffusion vs Auto-regression

- Ultimate AI Sound Design Assistant:
  - Retrieval _and_ (iterative, fine-controlled) generation

Closing Remarks

SILA
(enhanced text control)

MultiFoley
(video control)

LALMs + Sound Design

Adobe

ROYAL OAK

References

[1] Kingma, D. P., Welling, M., Auto-Encoding Variational Bayes, In Proc. of the International Conference on Learning Representations, 2014

[2] van den Oord, A., Vinyals, O., Kavukcuoglu, K., Neural Discrete Representation Learning, In Proc. of Neural Information Processing Systems (NeurIPS), 2017

[3] Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., Kumar, K., High-Fidelity Audio Compression with Improved RVQGAN, In Proc. of Neural Information Processing Systems (NeurIPS), 2023

[4] Kumar, S., Seetharaman, P., Salamon, J., Manocha, D., Nieto, O., SILA: Signal-to-Language Augmentation for Enhanced Control in Text-to-Audio Generation. Submitted to IEEE Signal Processing Letters, 2025

[5] Flores García, H., Nieto, O., Salamon, J., Pardo, B., Seetharaman, P., Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations, In Proc. of the 50th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hyderabad, India, 2025

[6] Chen, Z., Seetharaman, P., Russell, B., Nieto, O., Bourgin, D., Owens, A., Salamon, J., Video-Guided Foley Sound Generation with Multimodal Controls, Submitted to IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2025

[7] Ho, J., Jain, A., Abbeel, P., Denoising Diffusion Probabilistic Models, In Proc. of Neural Information Processing Systems (NeurIPS), 2020

[8] Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M. D., AudioLDM: Text-to-Audio Generation with Latent Diffusion Models, In Proc. of the International Conference on Machine Learning (ICML), 2023

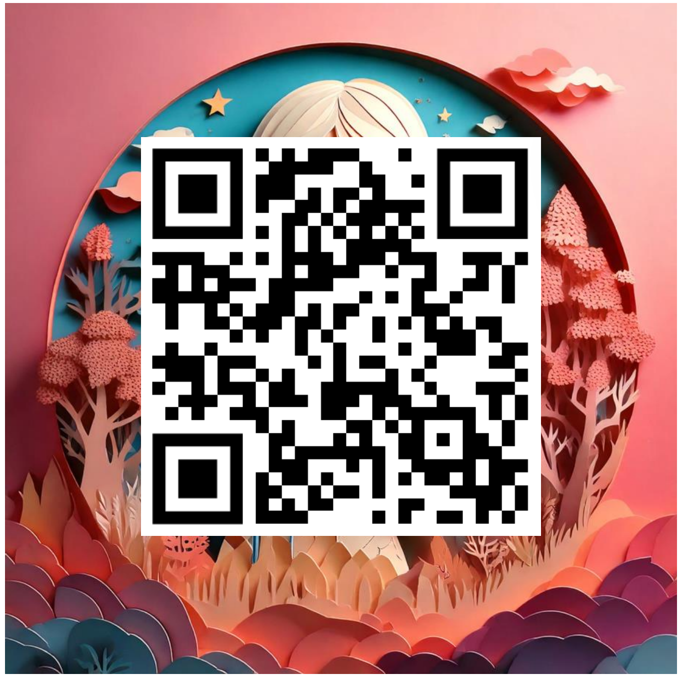
[9] Peebles,W., Xie, S., Scalable Diffusion Models with Transformers, In. Proc. of the IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2023

[10] Luo, S., Yan, C., Hu, C., Zhao, H., Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models, In Proc. of Neural Information Processing Systems (NeurIPS), 2023

Thank you!

SILA
(enhanced text control)

Sketch2Sound
(voice control)



MultiFoley
(video control)

onieto@adobe.com

@urinieto

/in/urinieto/