PAT 498/598 (Fall 2024)

Special Topics: Generative AI for Music and Audio Creation

Lecture 17: Latent-based Audio Synthesis

Instructor: Hao-Wen Dong

SCHOOL OF MUSIC, THEATRE & DANCE
PERFORMING ARTS TECHNOLOGY
UNIVERSITY OF MICHIGAN

Project Pitch

- Tell us briefly about
  - **What do you plan to do?**
    - Task, input/output, motivation, use cases, target users, etc.
  - **How do you plan to approach it?**
    - Papers to read, expected challenges, types of datasets/models to use, etc.
- **Send me a short paragraph or some bullet points by tomorrow** so that I can provide some feedback and suggest relevant papers to look into!

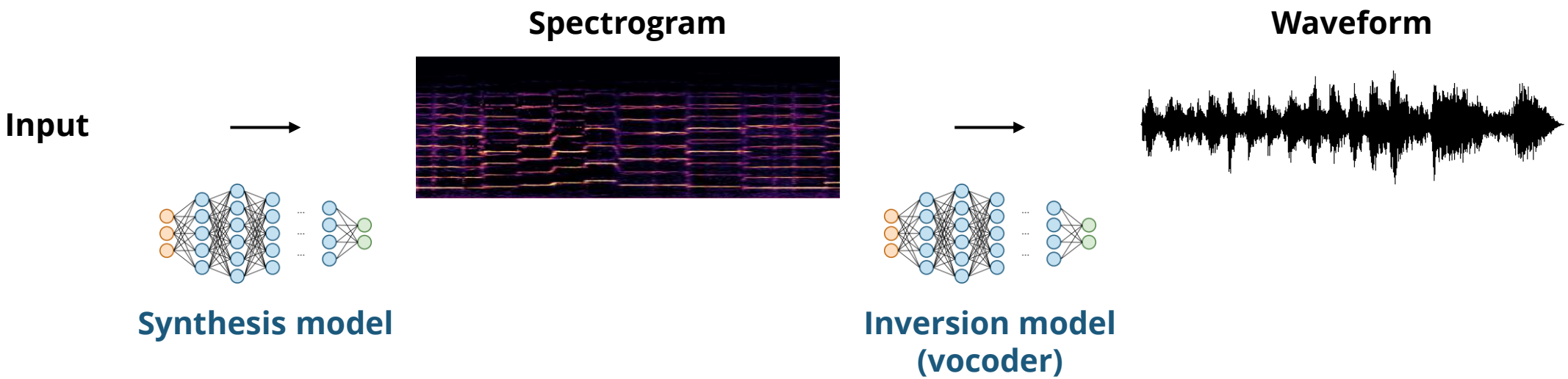
Final Project

- Milestones (all due at the specified date at **11:59 PM ET**)

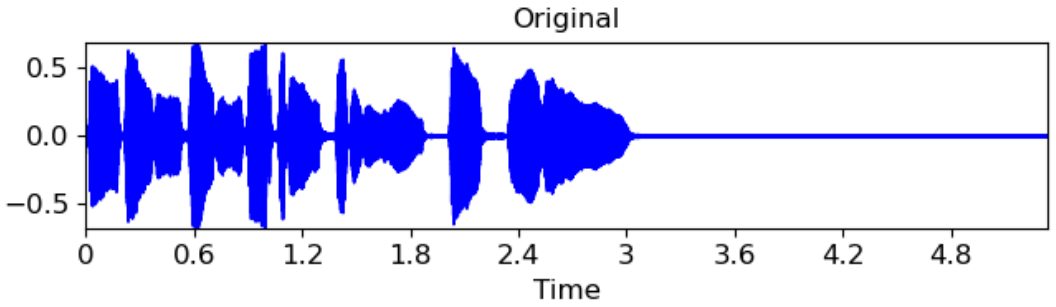
Pitch	November 6	Topic & high-level plans	
Proposal	November 18	Survey & plans (1 page)	**Next milestone!**
Presentation	December 9	Showcase & report	
Final report	December 15	Full report (3-5 pages)	

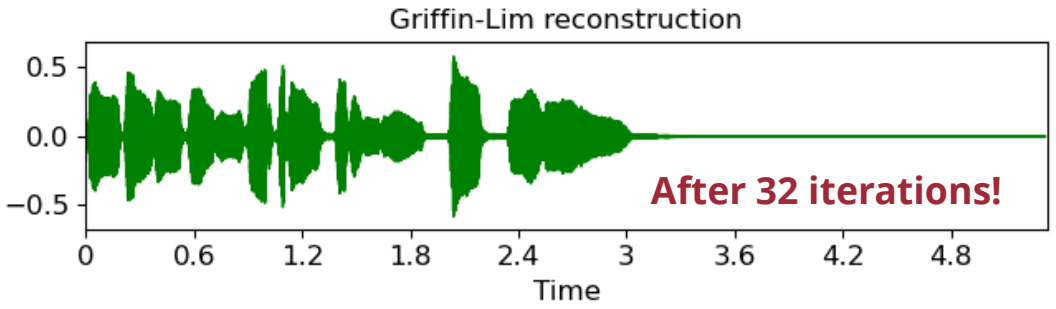
- Instructions will be released on Gradescope
- Late submissions: **NOT accepted**

(Recap) Frequency-domain Audio Synthesis



(Recap) Griffin-Lim Algorithm (Griffin & Lim, 1984)





(Source: librosa documentation)

Given a magnitude-only STFT matrix

↓

Randomly initialize the phase

↓

$$y' = \arg\min_{y} (M - \mathrm{STFT}(y))^2$$

Find the signal y that minimize the MSE between the input and STFT(y)

↓

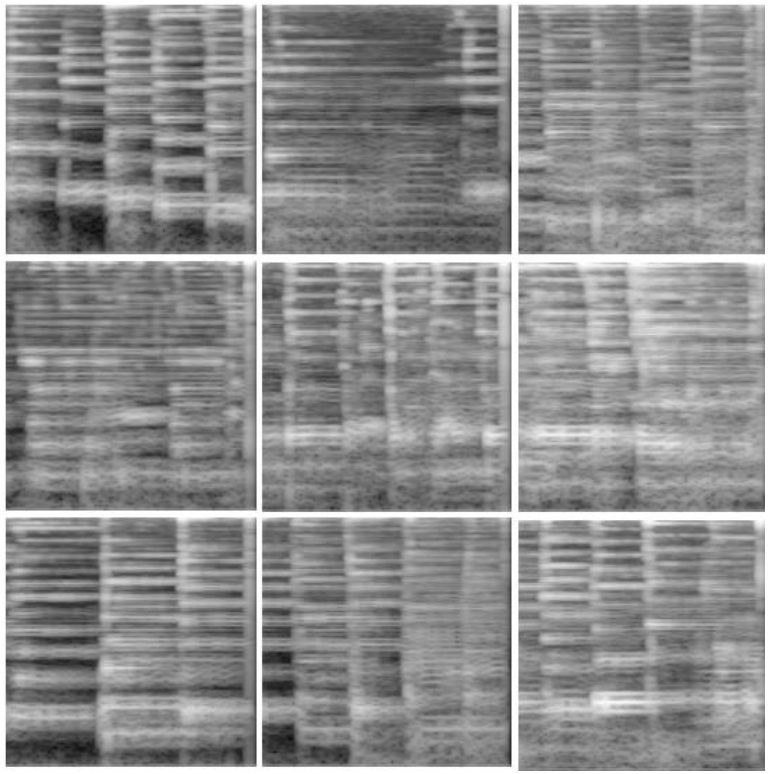
$$M' = \mathrm{STFT}(y')$$

Find the STFT of the signal y

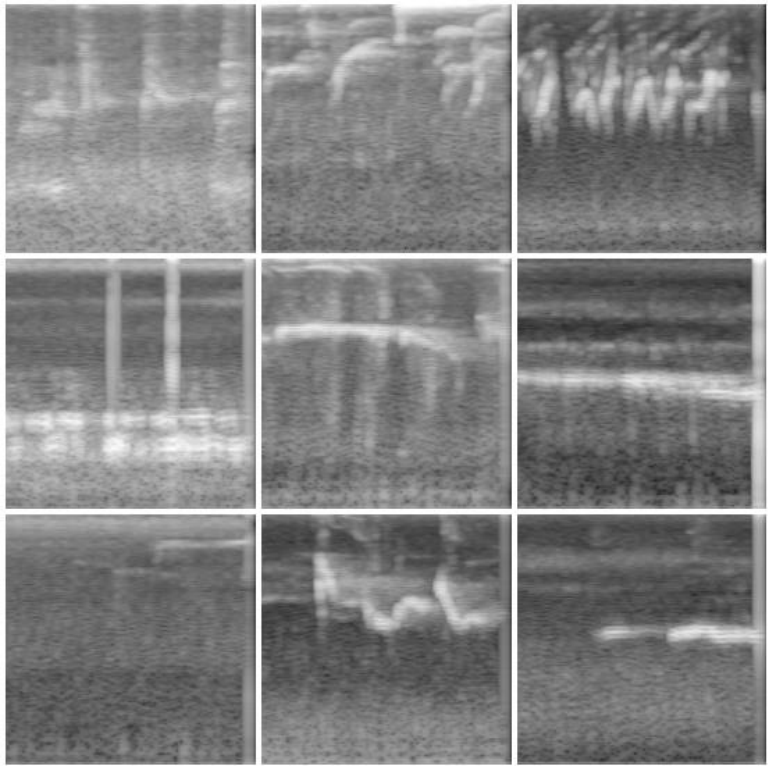
Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE TASSP*, 1984.
librosa.org/doc/main/generated/librosa.griffinlim.html

(Recap) Example: SpecGAN (Donahue et al., 2019)

Piano sounds



Bird sounds

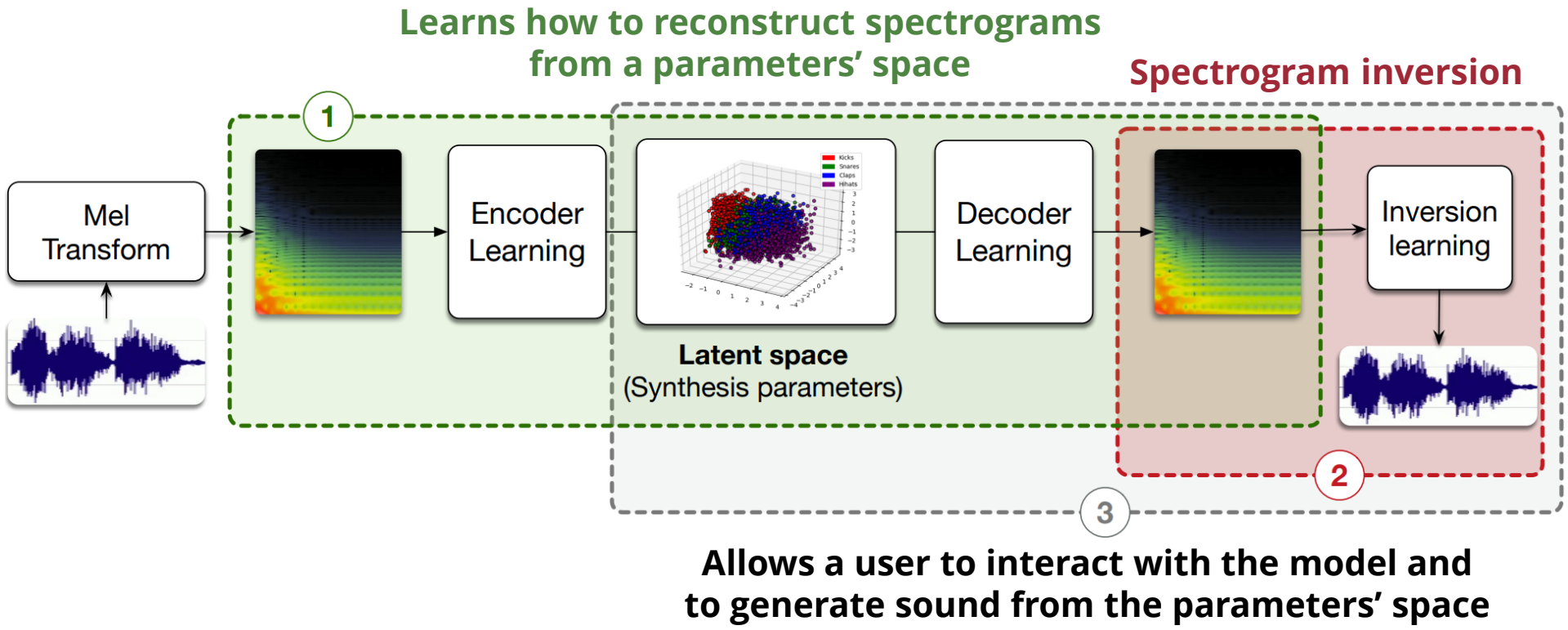


Example of generated music

(Source: Donahue et al., 2019)

Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial Audio Synthesis," *ICLR*, 2019.

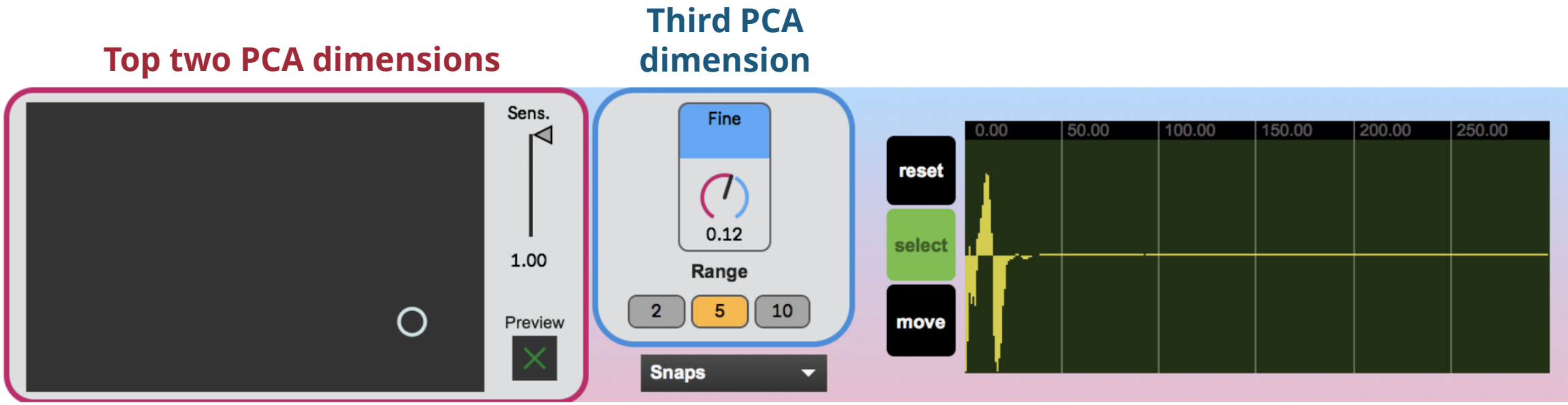
(Recap) Example: Neural Drum Machine (Aouameur et al., 2019)



(Source: Aouameur et al., 2019)

Cyran Aouameur, Philippe Esling, and Gaëtan Hadjeres, "Neural Drum Machine : An Interactive System for Real-time Synthesis of Drum Sounds," *ICCC*, 2019.

(Recap) Example: Neural Drum Machine (Aouameur et al., 2019)

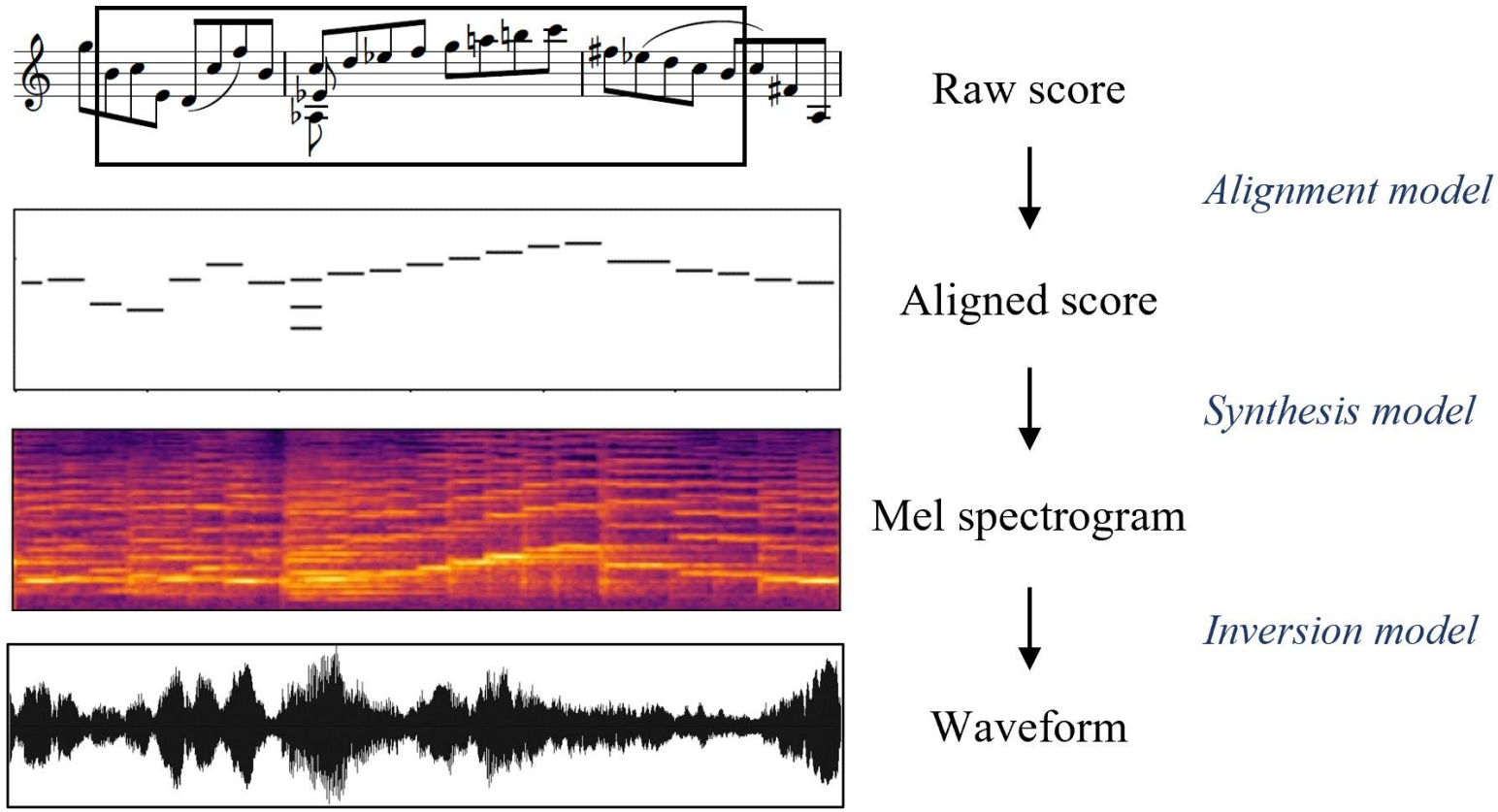


(Source: Aouameur et al., 2019)

drive.google.com/file/d/1DDo0_KnwkWirCM4t0PT8cp6uotsfuufj/view

Cyran Aouameur, Philippe Esling, and Gaëtan Hadjeres, "Neural Drum Machine : An Interactive System for Real-time Synthesis of Drum Sounds," *ICCC*, 2019.

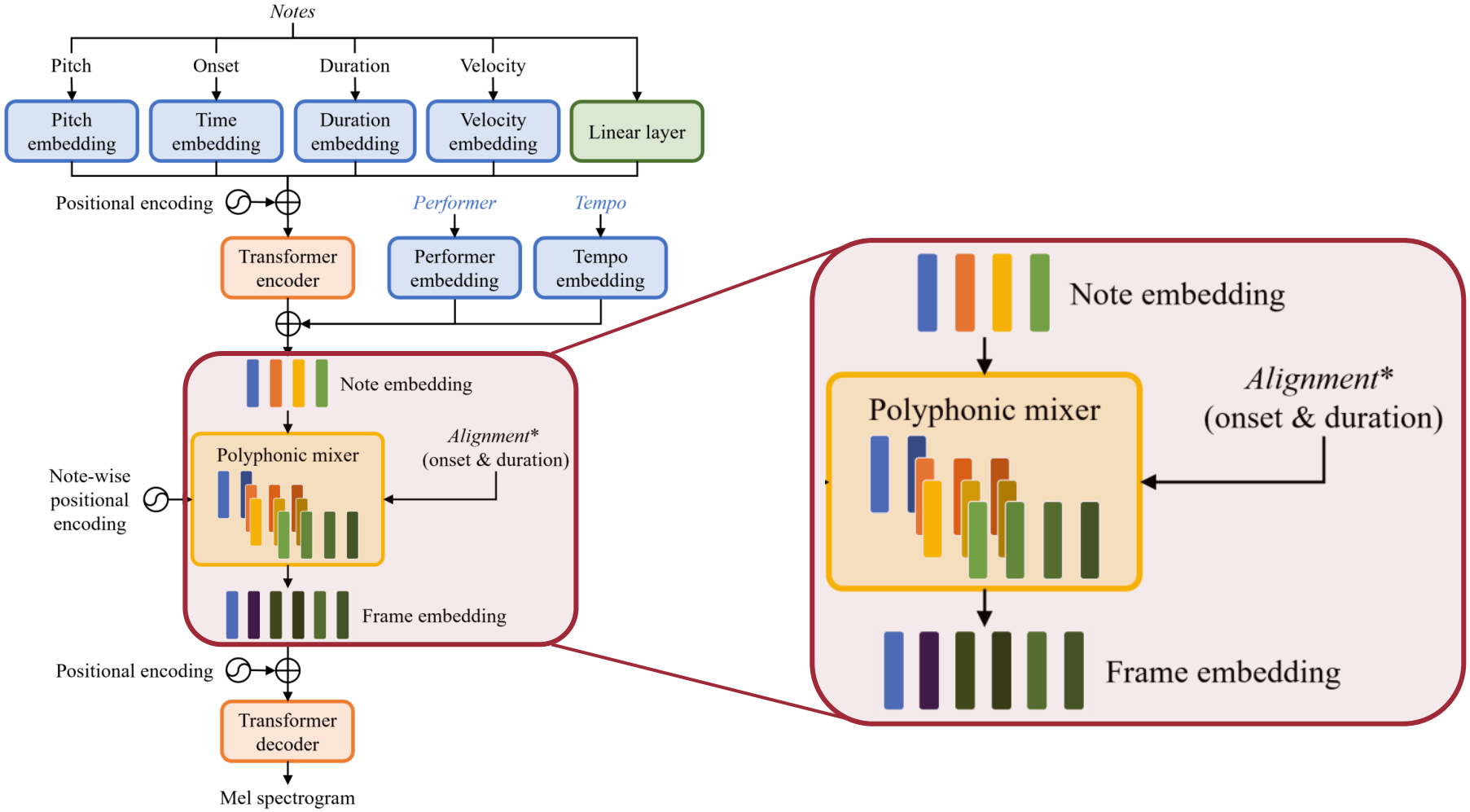
(Recap) Example: DeepPerformer (Dong et al., 2022)



(Source: Dong et al., 2022)

Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *ICASSP*, 2022.

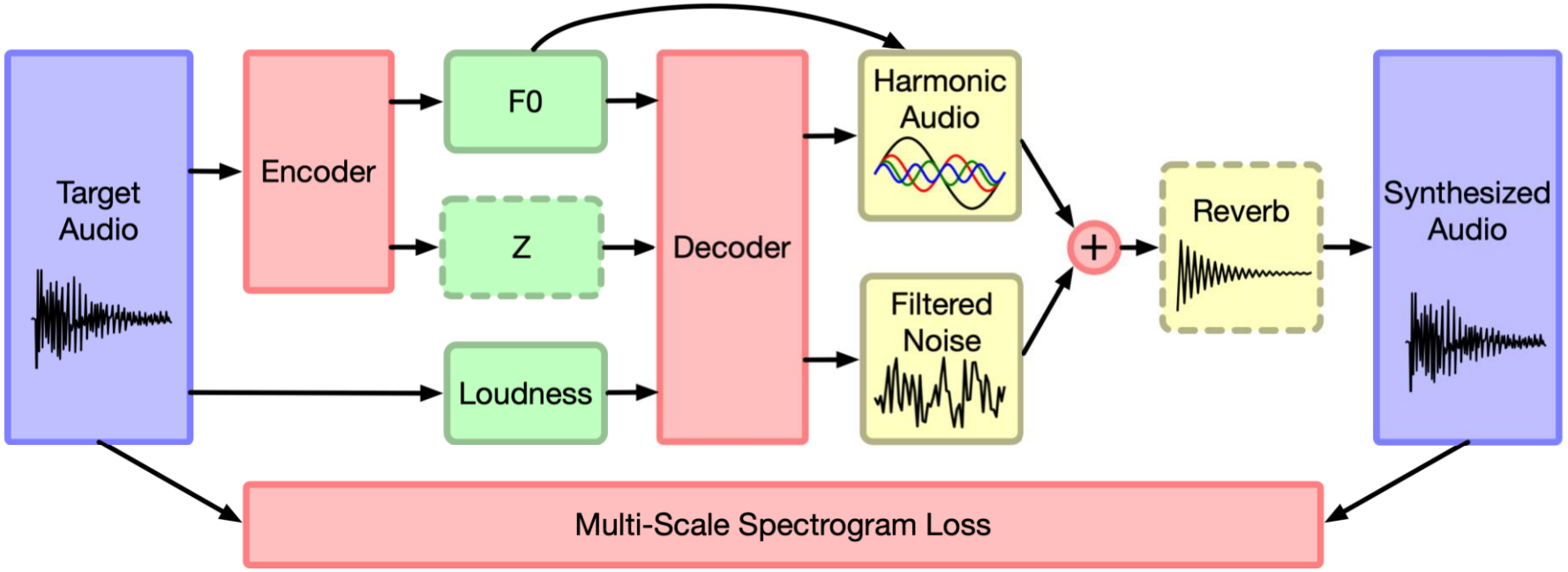
(Recap) Example: DeepPerformer (Dong et al., 2022)



(Source: Dong et al., 2022)

Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *ICASSP*, 2022.

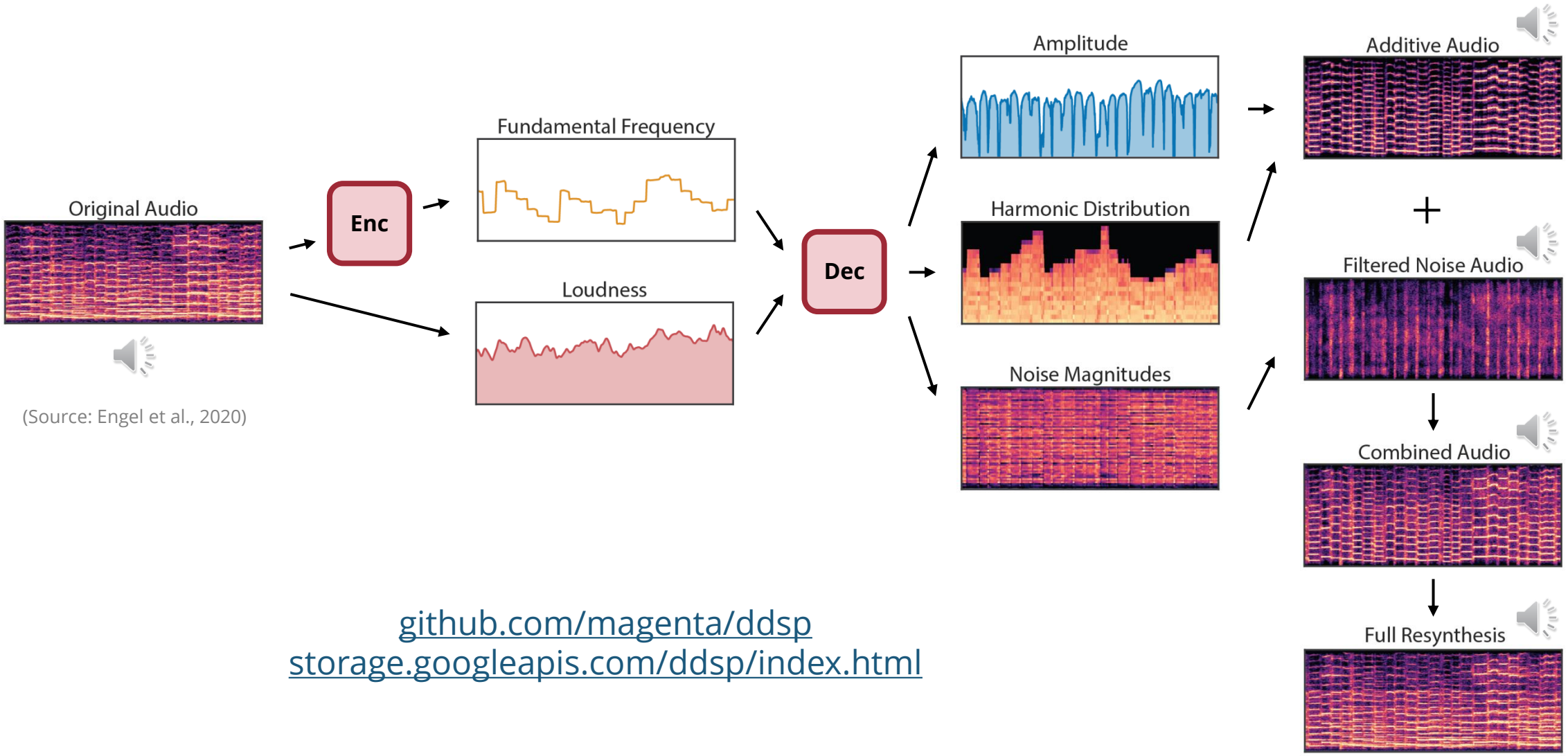
(Recap) Example: Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, "DDSP: Differentiable Digital Signal Processing," *ICLR*, 2020.

(Recap) Example: Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

github.com/magenta/ddsp
storage.googleapis.com/ddsp/index.html

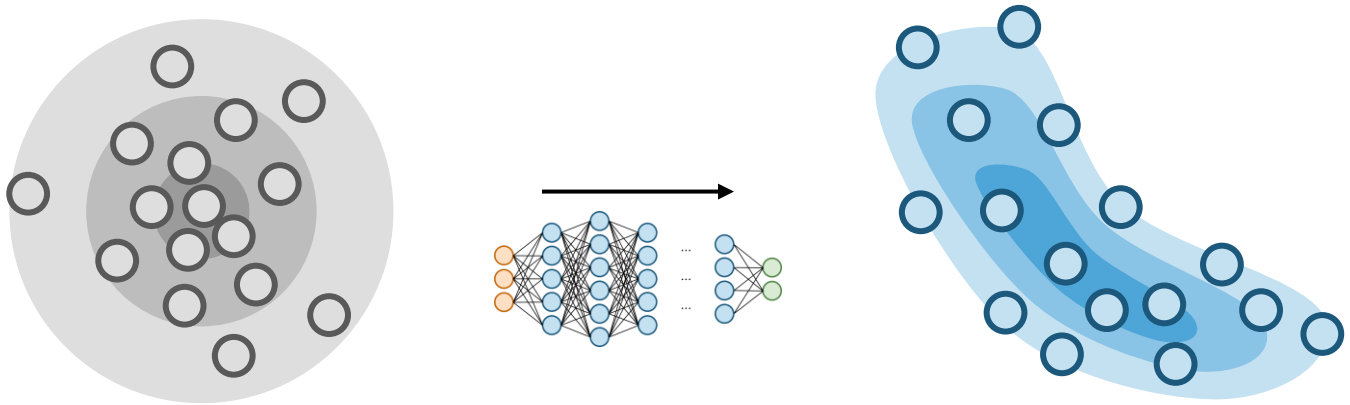
Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, "DDSP: Differentiable Digital Signal Processing," *ICLR*, 2020.

Latent-based Audio Synthesis

(Recap) Generating Data from a Random Distribution

Random distribution

Data distribution

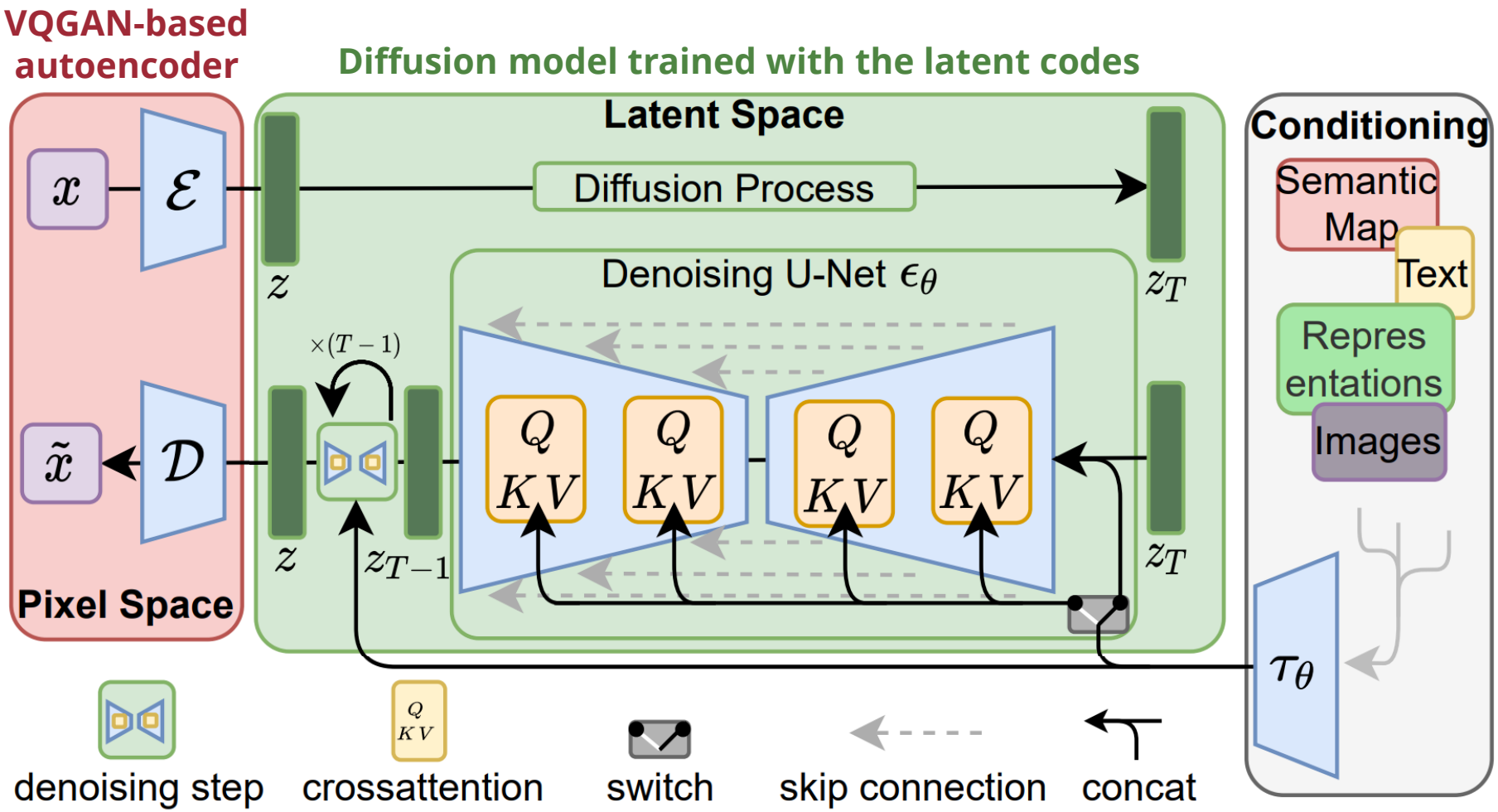


$P(z)$

$P(x)$

If we can learn this mapping, we can easily generate new samples from the data distribution

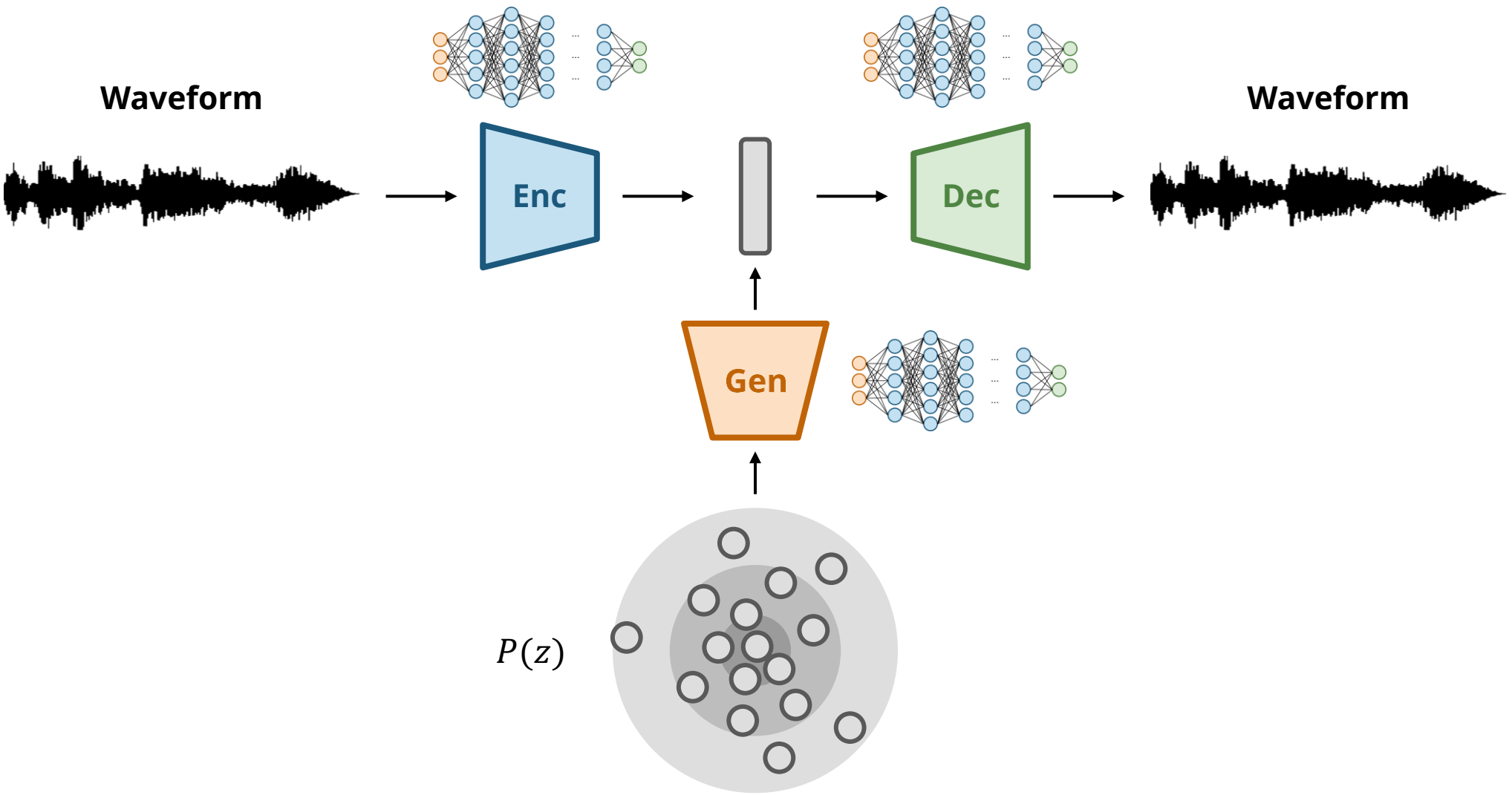
(Recap) Latent Diffusion Models (LDMs)



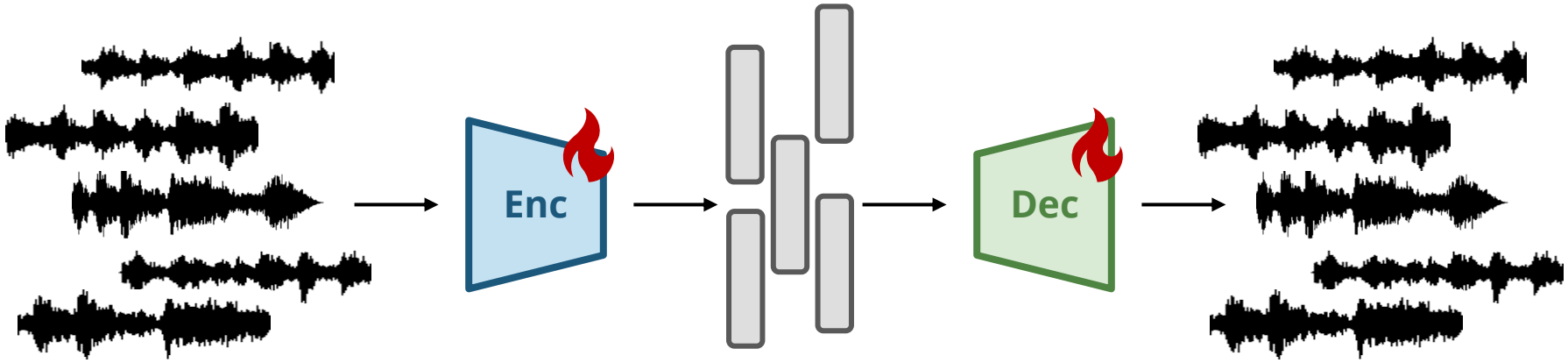
(Source: Rombach et al., 2022)

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *CVPR*, 2022.

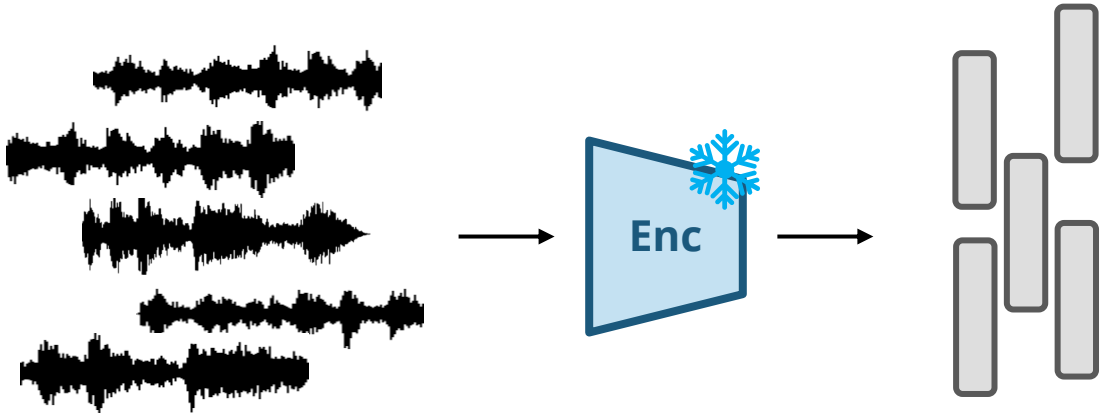
Latent-based Audio Synthesis



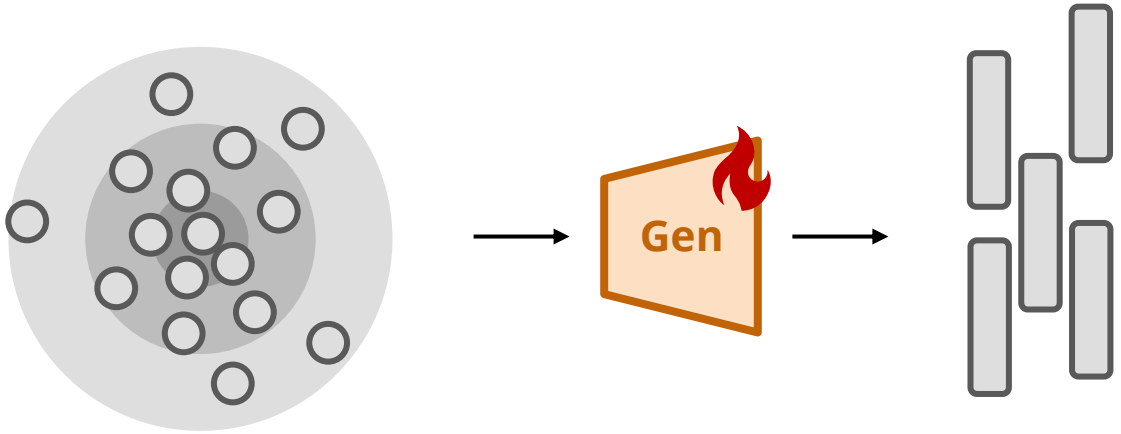
Step 1: Train an Autoencoder



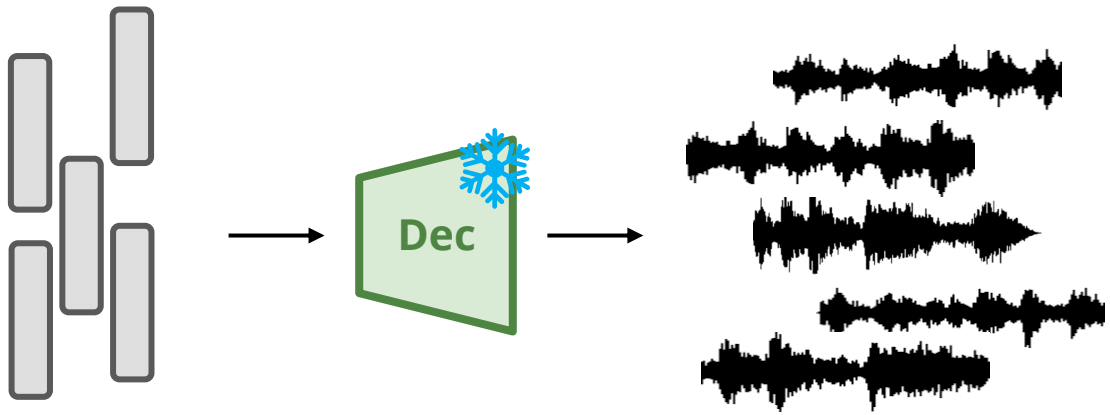
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

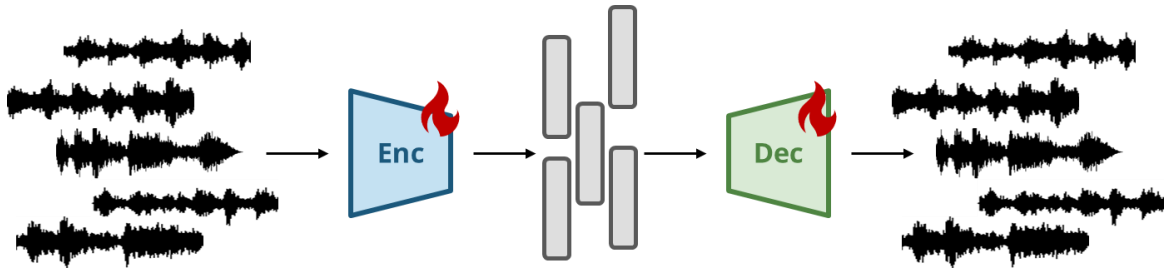


Step 4: Decode the Latent Vectors

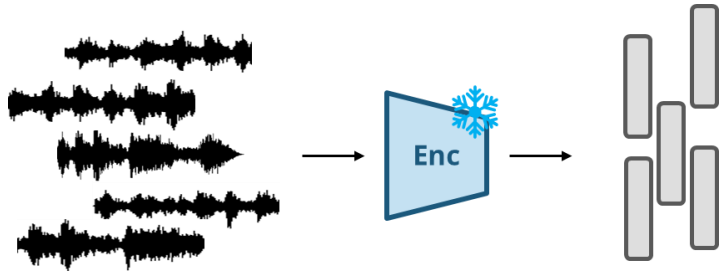


Pipeline

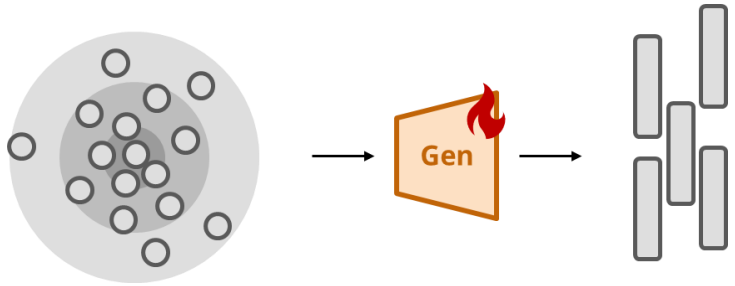
Step 1: Train an Autoencoder



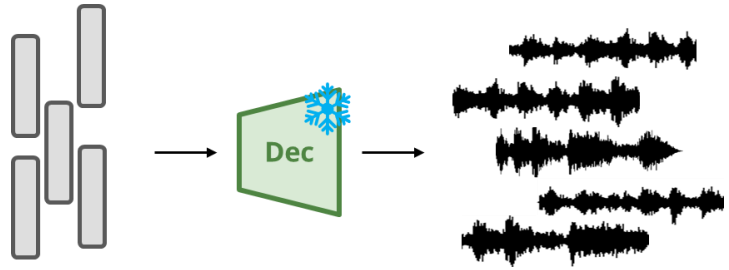
Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model

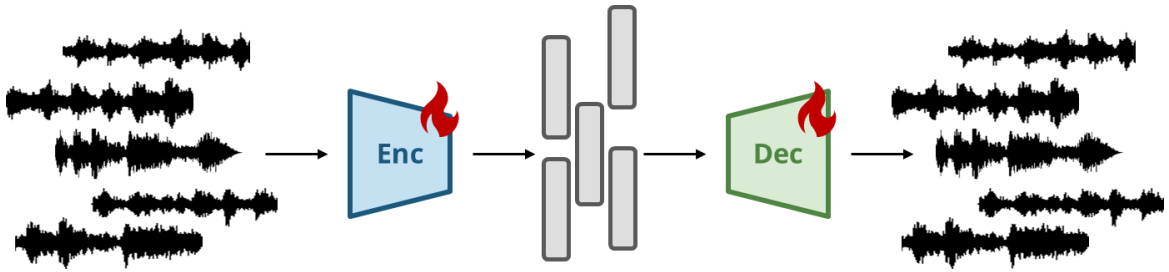


Step 4: Decode the Latent Vectors

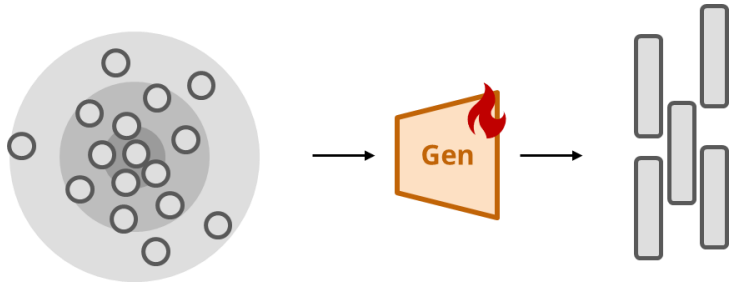


Training

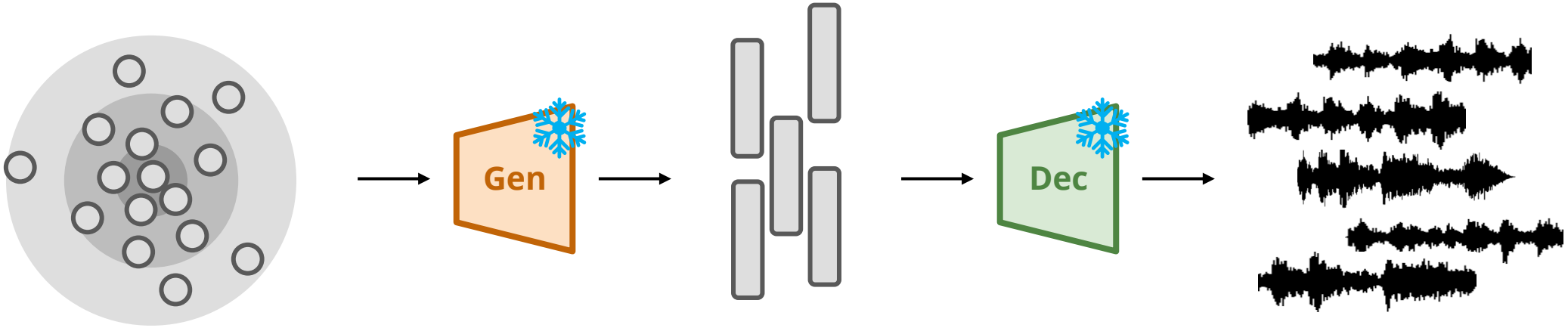
Autoencoder



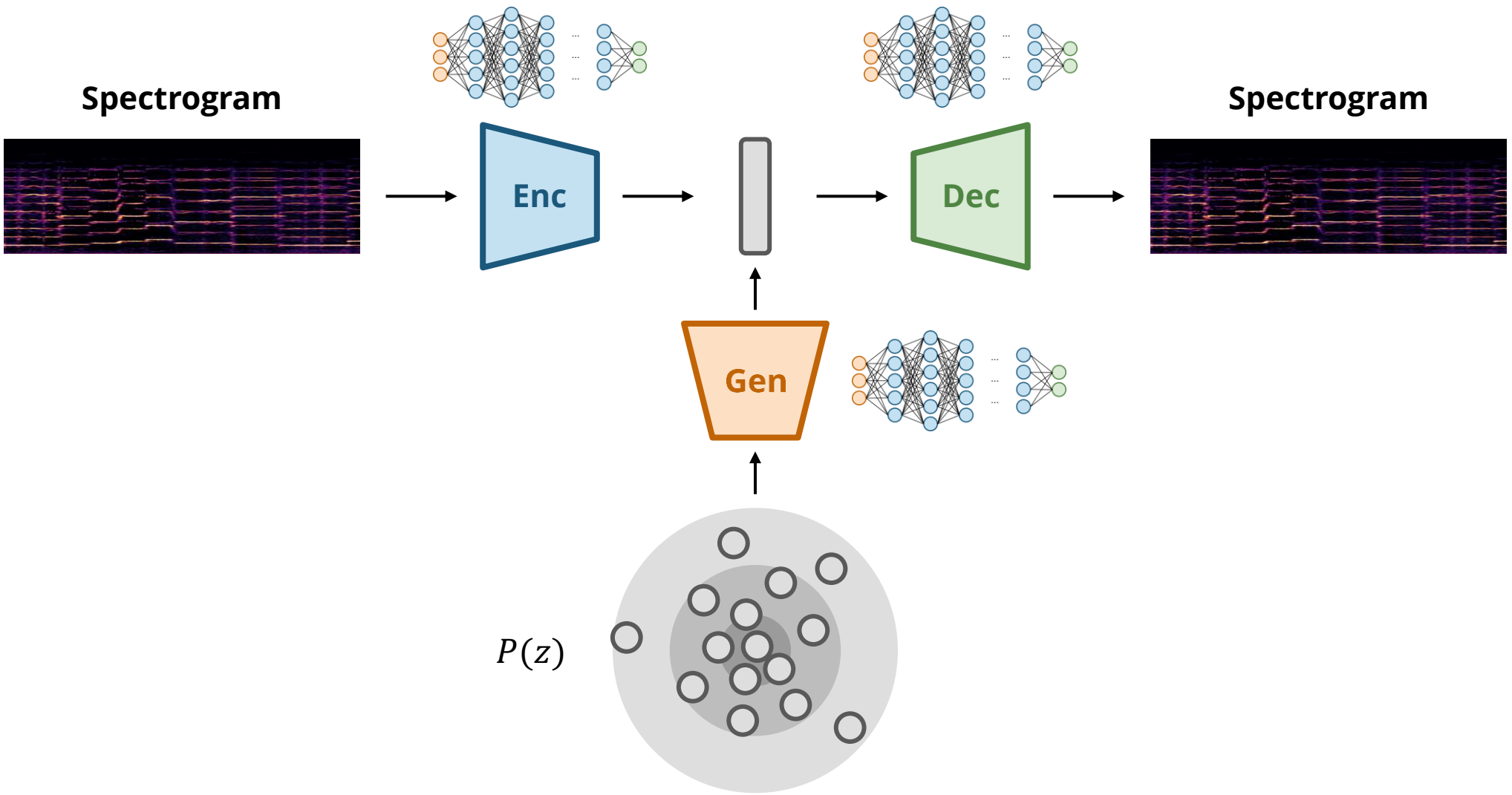
Latent Generative Model



Inference

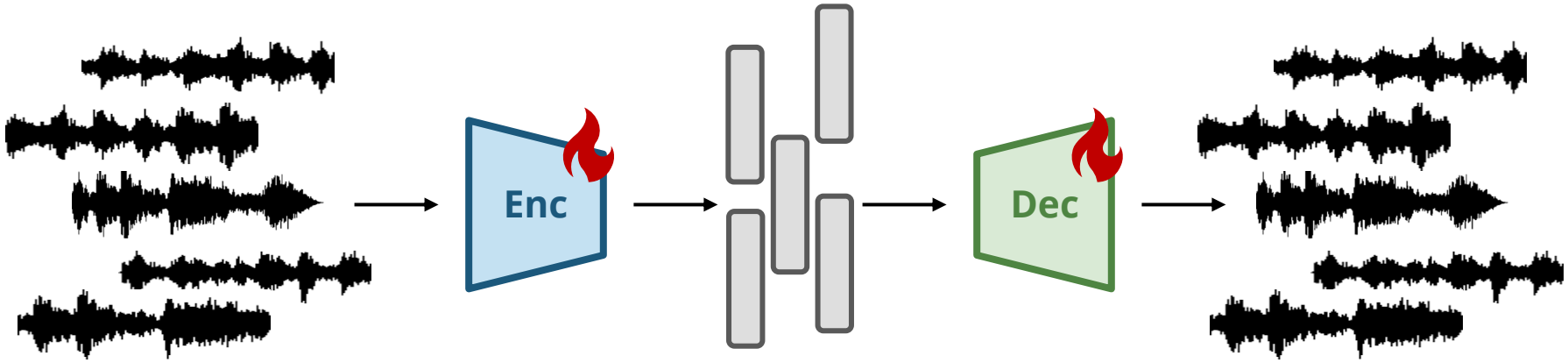


Latent-based Audio Synthesis

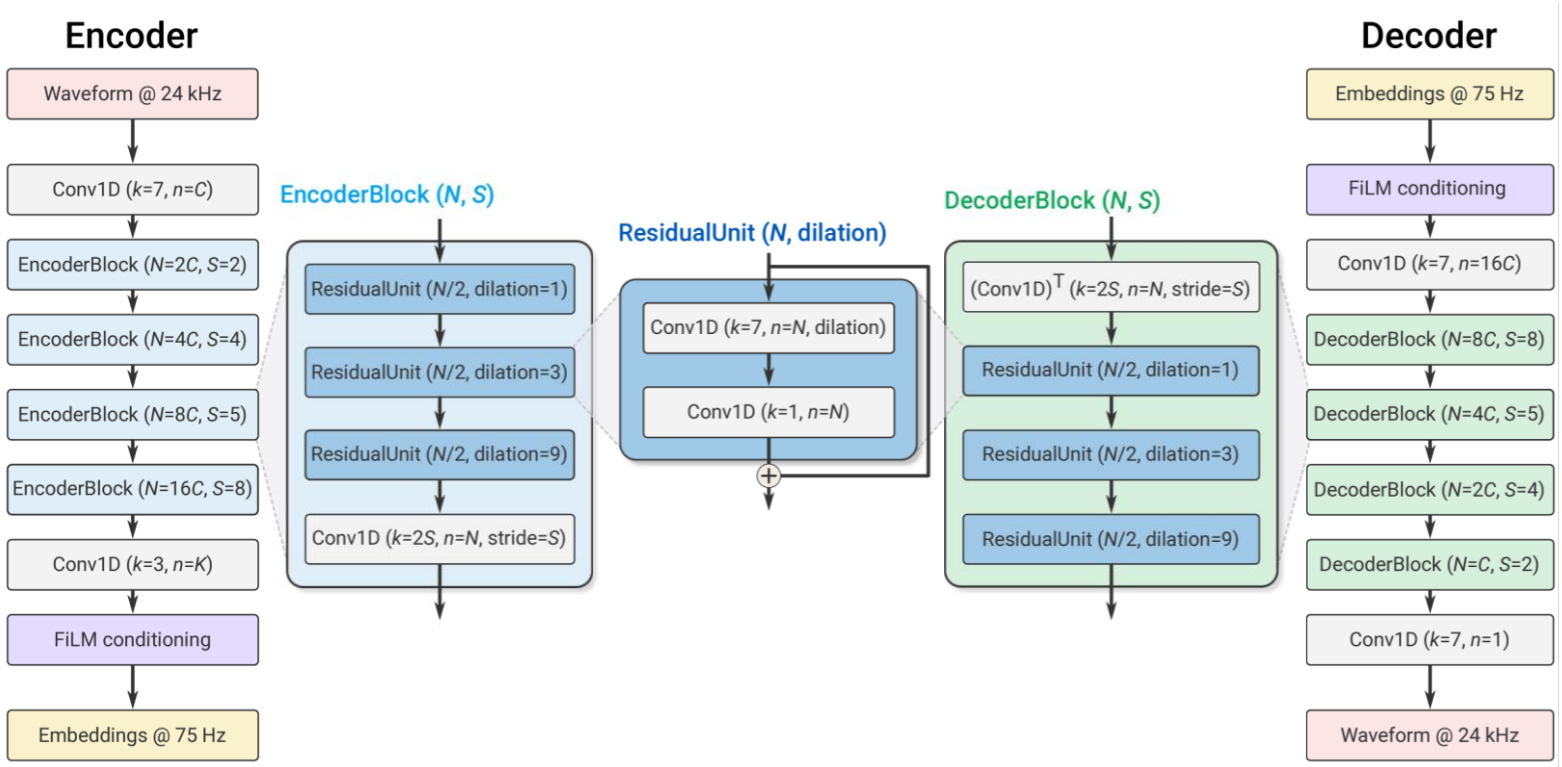


Neural Codecs

(Recap) Step 1: Train an Autoencoder



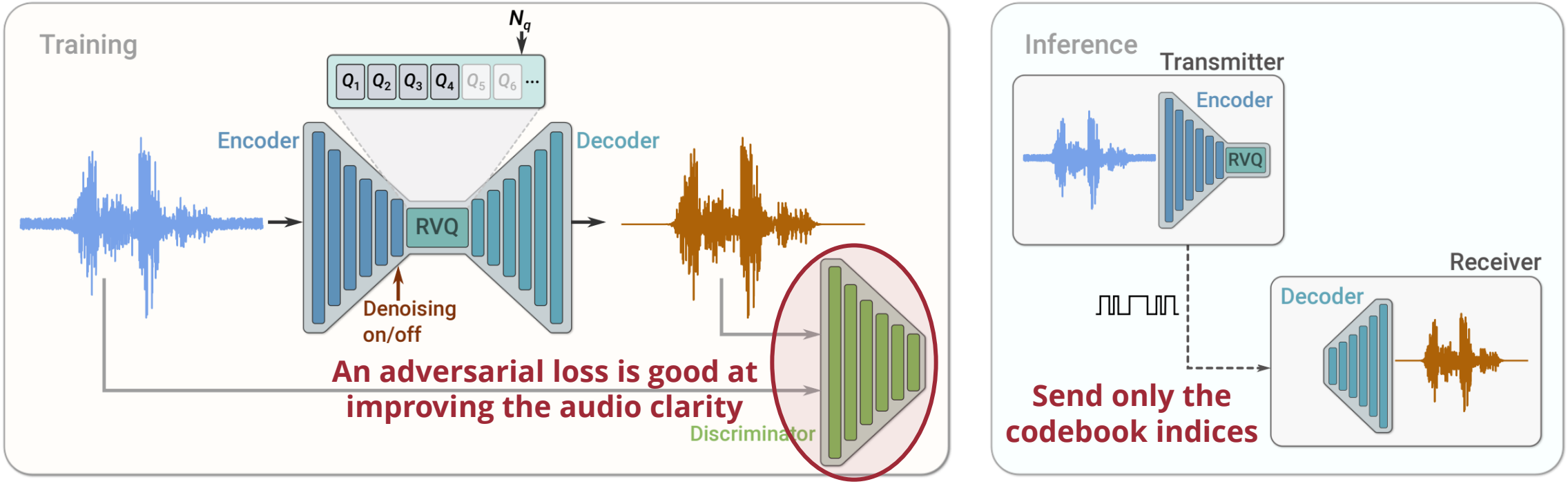
Example: SoundStream (Zeghidour et al., 2021)



(Source: Zeghidour et al., 2021)

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *TASLP*, 2021.

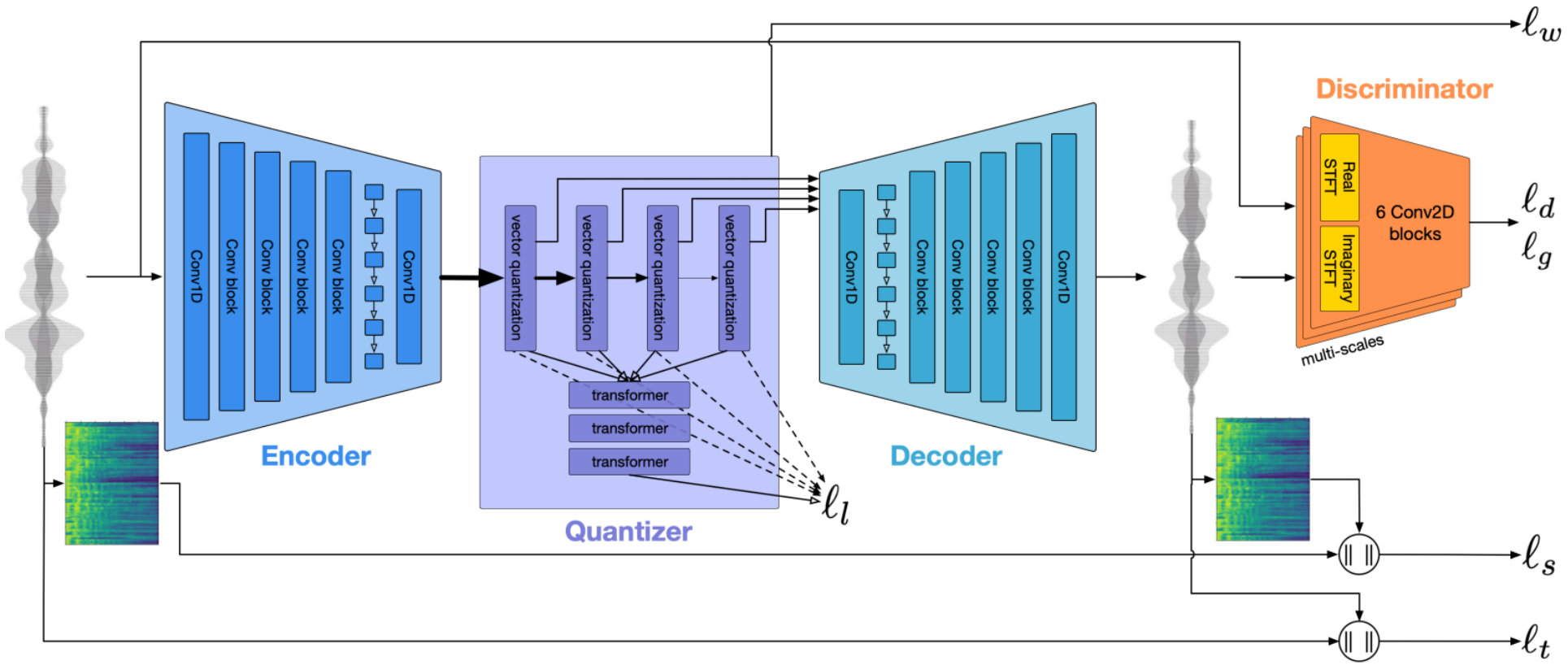
Example: SoundStream (Zeghidour et al., 2021)



(Source: Zeghidour et al., 2021)

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *TASLP*, 2021.

Example: EnCodec (Défossez et al., 2022)



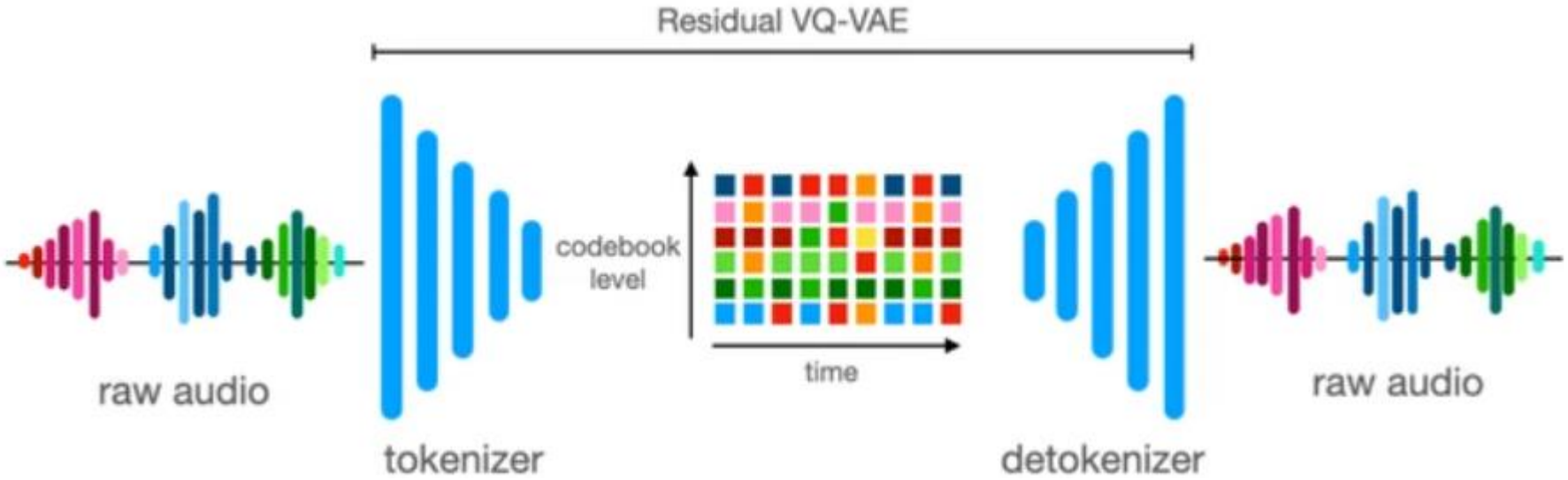
(Source: Défossez et al., 2022)

ai.honu.io/papers/encodec/samples.html

github.com/facebookresearch/encodec

Alexandre Défossez, Jade Copet, Gabriel Synnaeve and Yossi Adi, "High Fidelity Neural Audio Compression," *arXiv preprint arXiv:2210.13438*, 2022.

Example: Descript Audio Codec (Kumar et al., 2023)



(Source: Kumar et al., 2023)

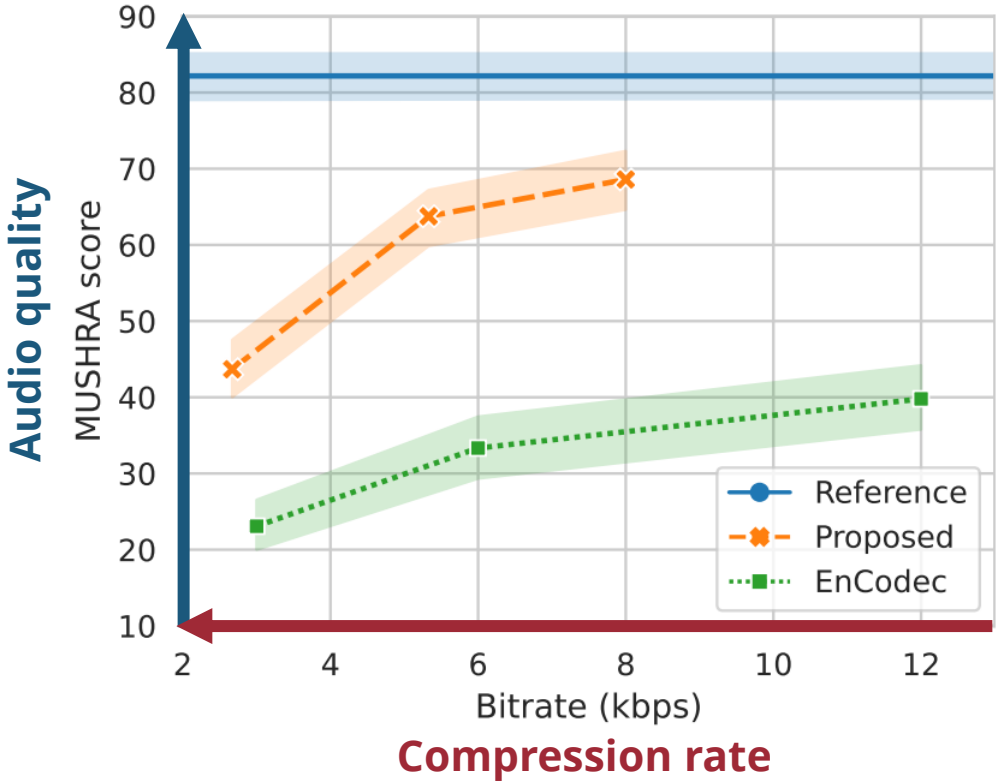
descript.notion.site/Descript-Audio-Codec-11389fce0ce2419891d6591a68f814d5

github.com/descriptinc/descript-audio-codec

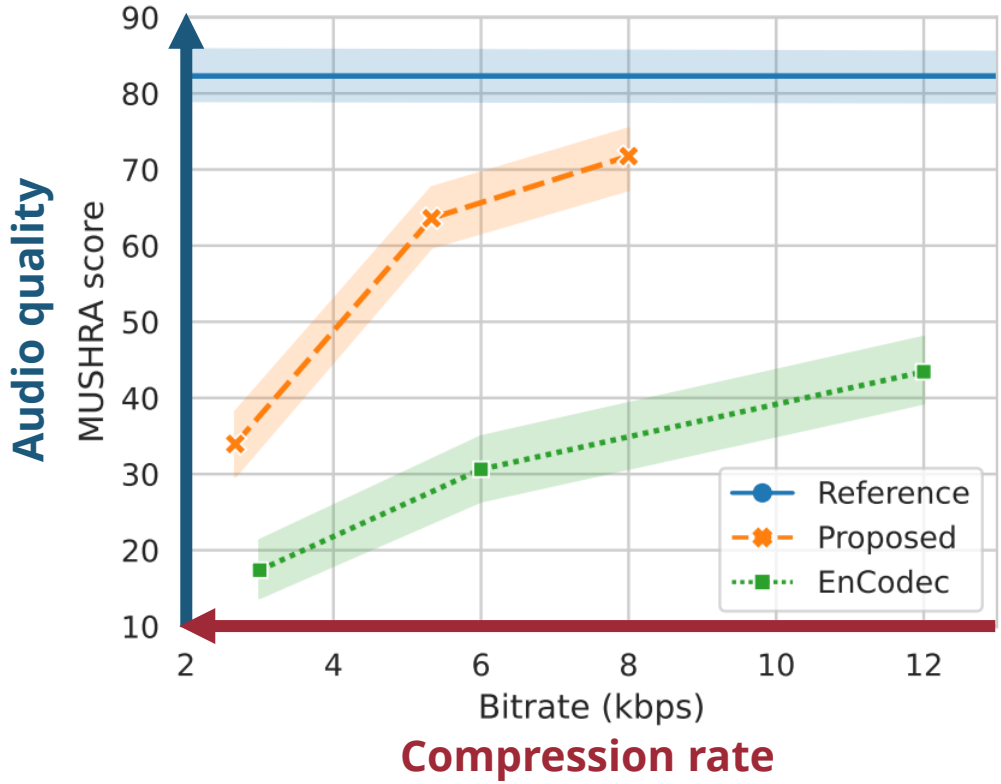
Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," *NeurIPS*, 2023.

Example: Descript Audio Codec (Kumar et al., 2023)

Listening Test Results @ 44.1 kHz



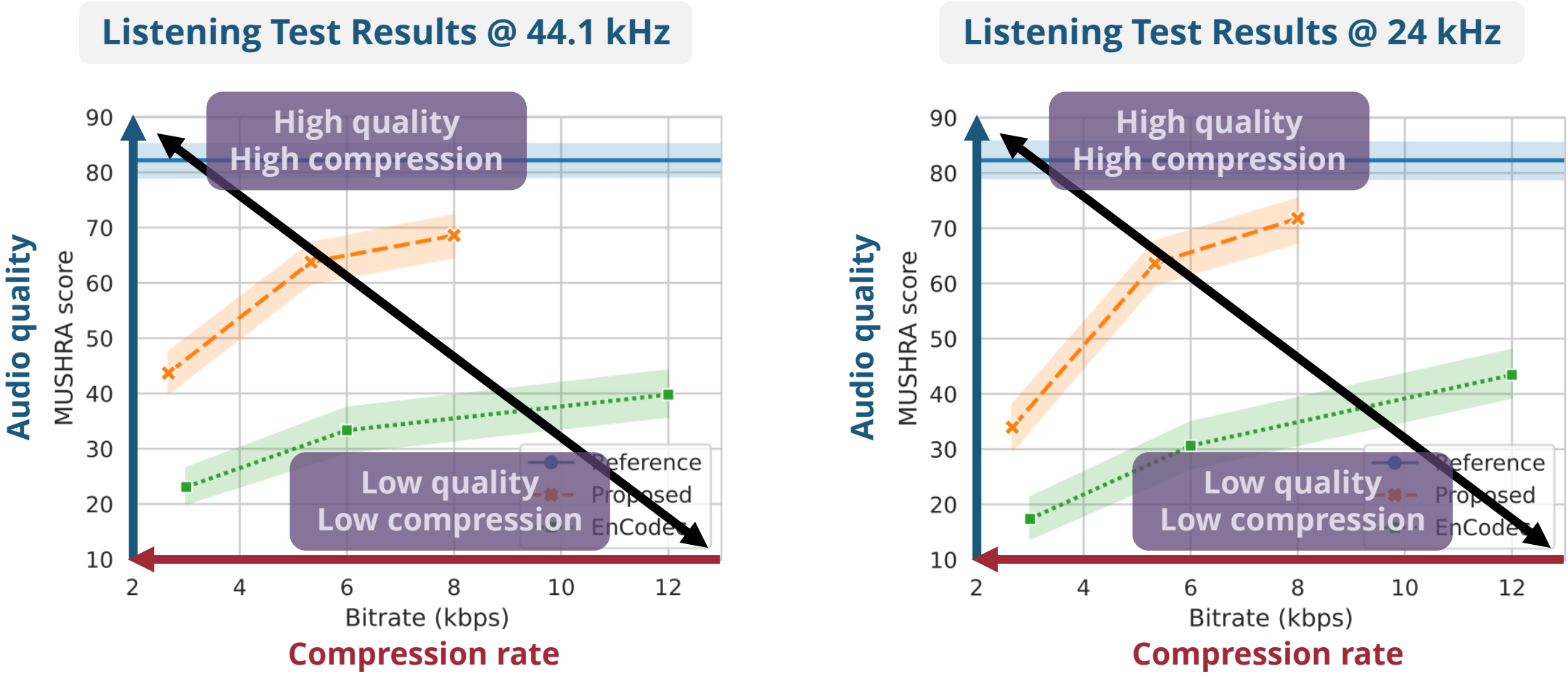
Listening Test Results @ 24 kHz



(Source: Kumar et al., 2023)

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," *NeurIPS*, 2023.

Example: Descript Audio Codec (Kumar et al., 2023)

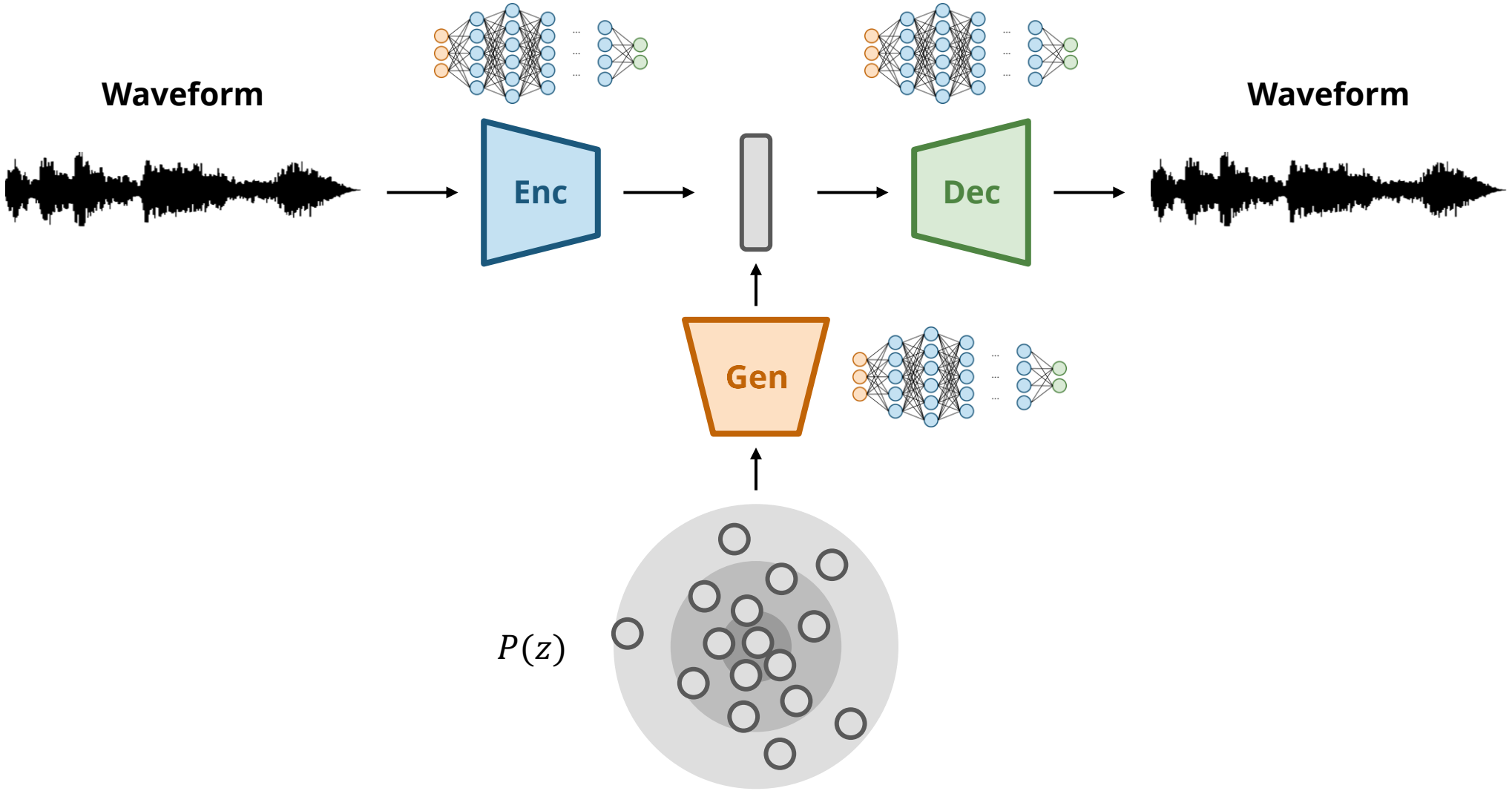


(Source: Kumar et al., 2023)

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," *NeurIPS*, 2023.

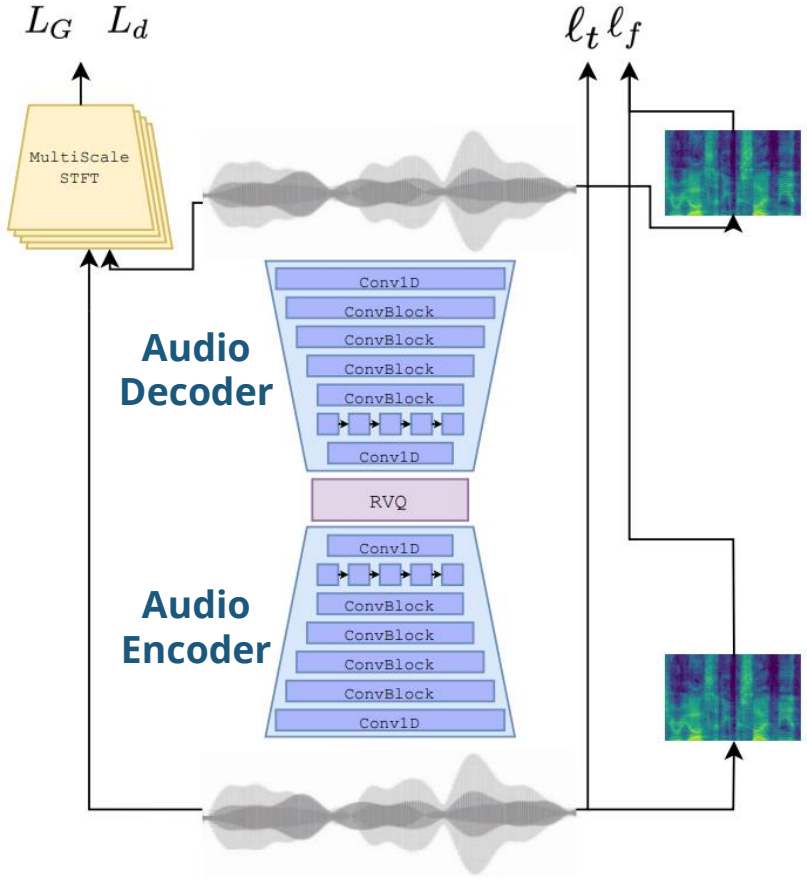
Latent-based Audio Synthesis

(Recap) Latent-based Audio Synthesis

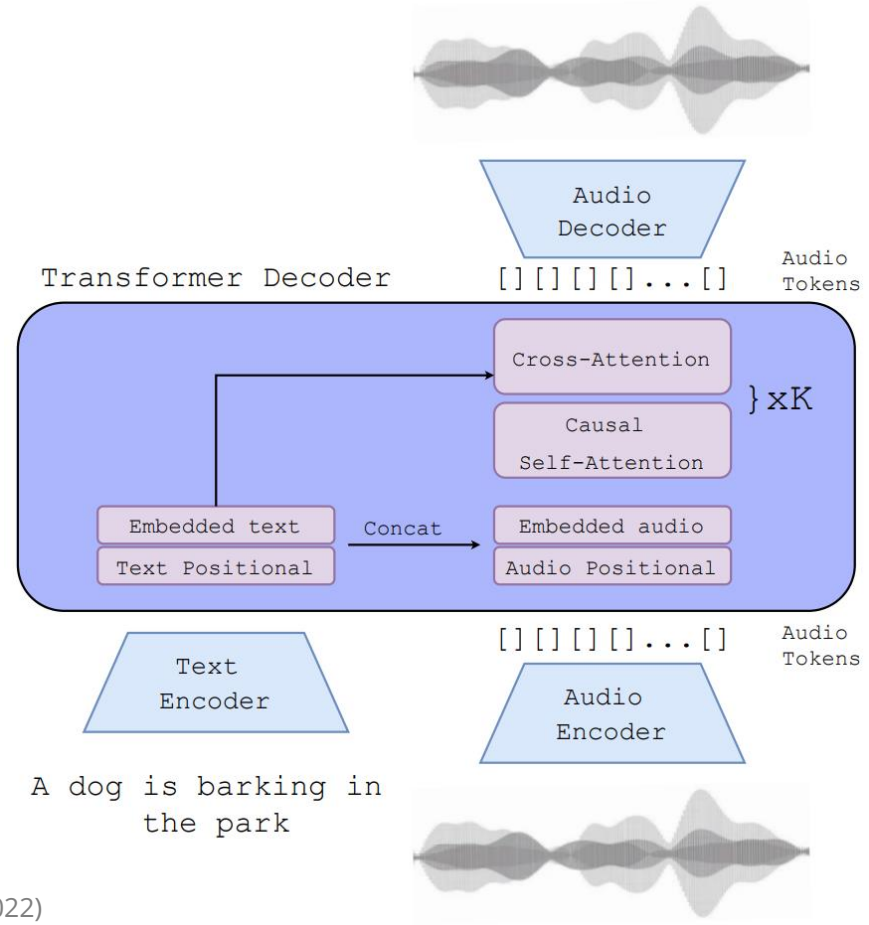


Example: AudioGen (Kreuk et al., 2023)

Audio Autoencoder



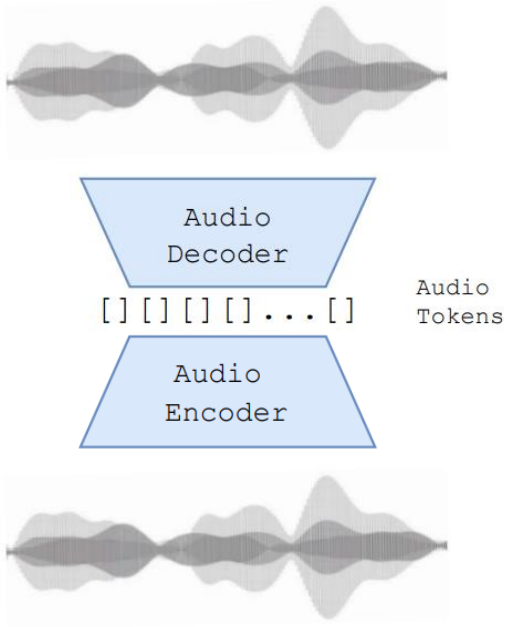
Audio Language Model



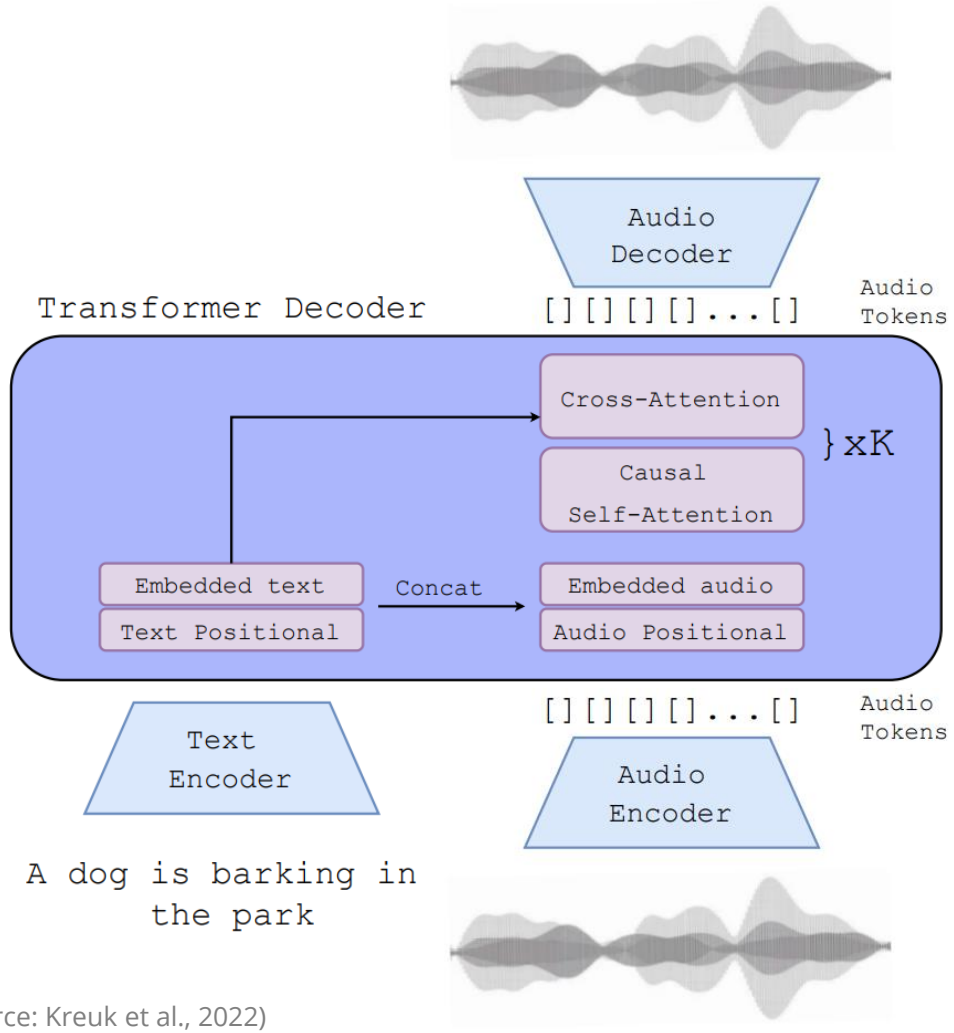
(Source: Kreuk et al., 2022)

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually Guided Audio Generation," *ICLR*, 2023.

Example: AudioGen (Kreuk et al., 2023)



4k hours
(speech, music, sound effects)



(Source: Kreuk et al., 2022)

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually Guided Audio Generation," *ICLR*, 2023.

Example: AudioGen (Kreuk et al., 2023)

(Source: Kreuk et al., 2022)

felixkreuk.github.io/audiogen

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually Guided Audio Generation," *ICLR*, 2023.

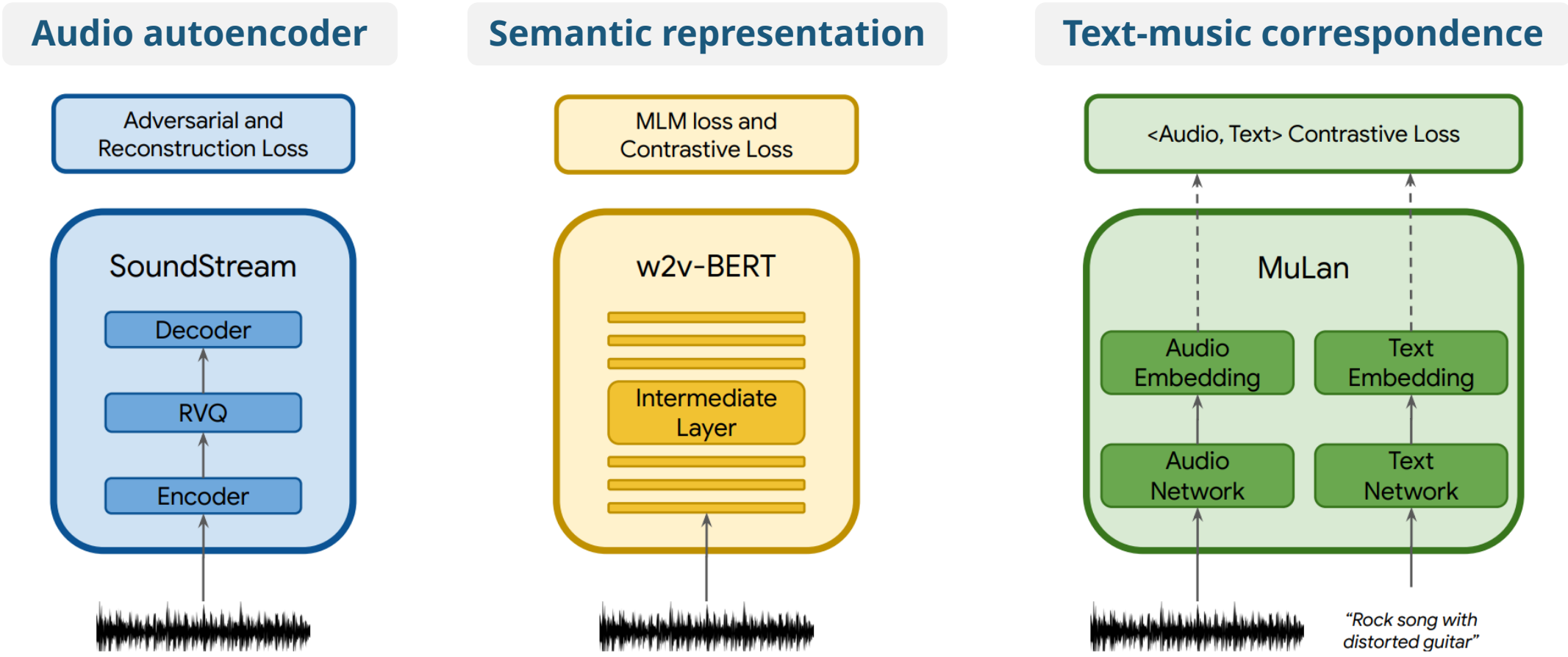
Example: MusicGen (Copet et al., 2023)

- AudioGen for Music
- Use EnCodec (Défossez et al., 2022) as the autoencoder
  - instead of SoundStream for AudioGen (Kreuk et al., 2023)
- **20k hours** of licensed music
  - Internal dataset 10k High-quality (private)
  - ShutterStock 25k Instrument-only
  - Pond5 365k Instrument-only

ai.honu.io/papers/musicgen/

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and Controllable Music Generation," *NeurIPS*, 2023.

Example: MusicLM (Agostinelli et al., 2023)



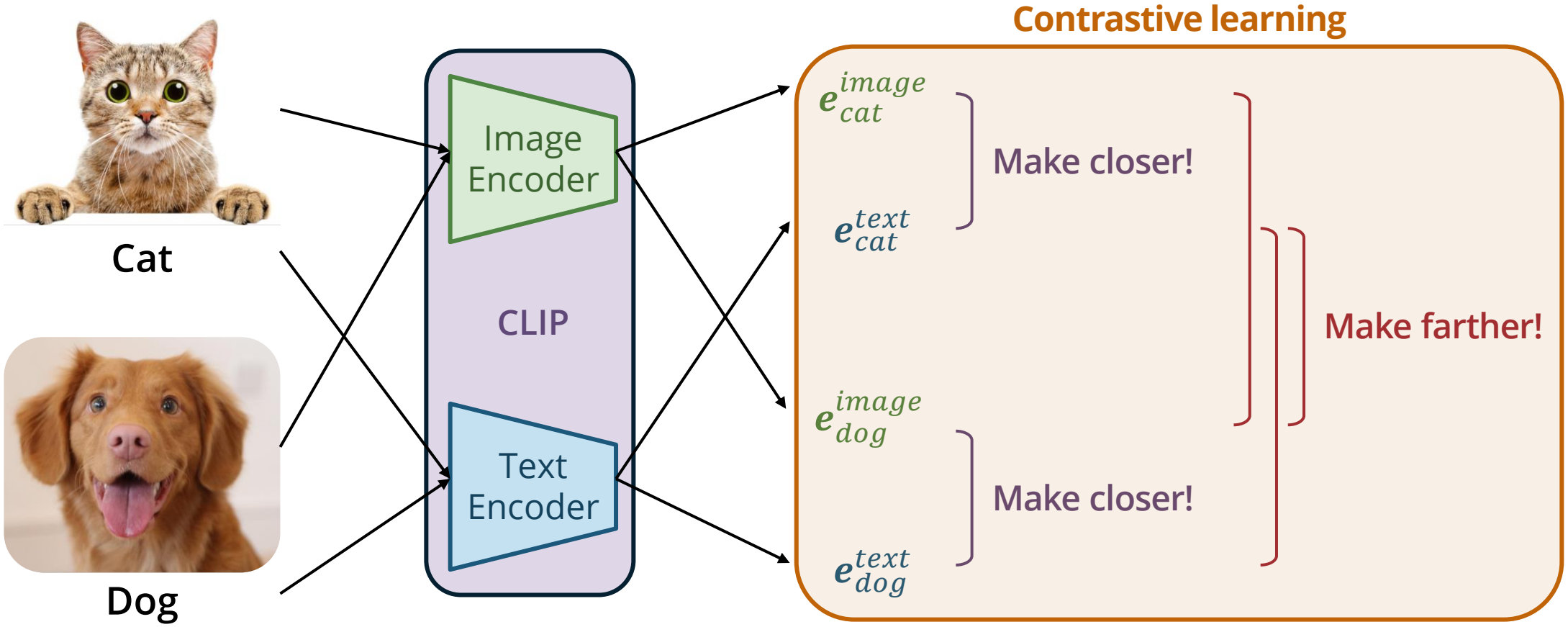
106k songs, 8.2k hours

44M 30-sec clips, 370k hours

(Source: Agostinelli et al., 2022)

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

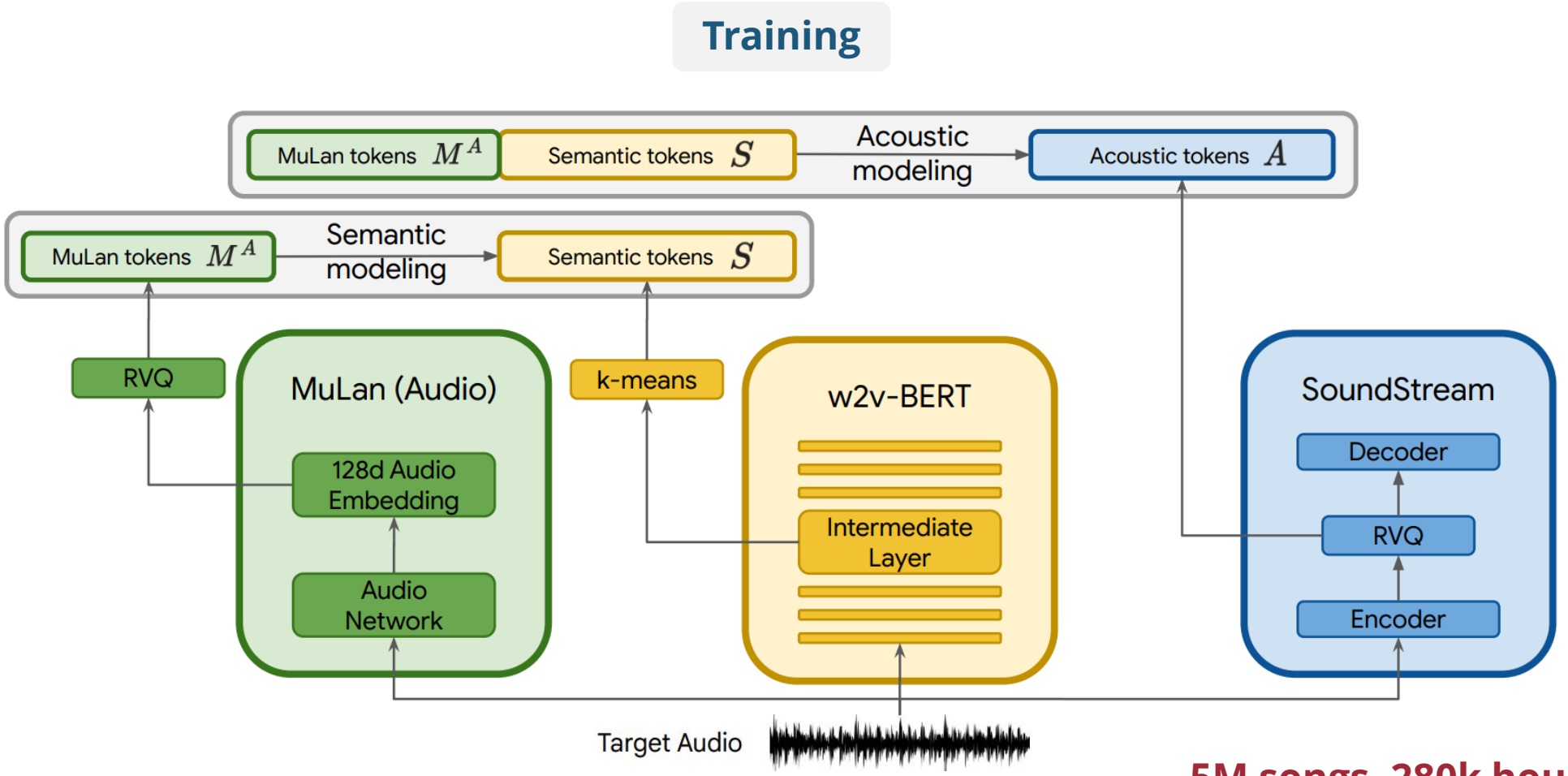
Contrastive Language-Image Pretraining (CLIP)



Learn a shared embedding space for images and texts

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *ICML*, 2021.

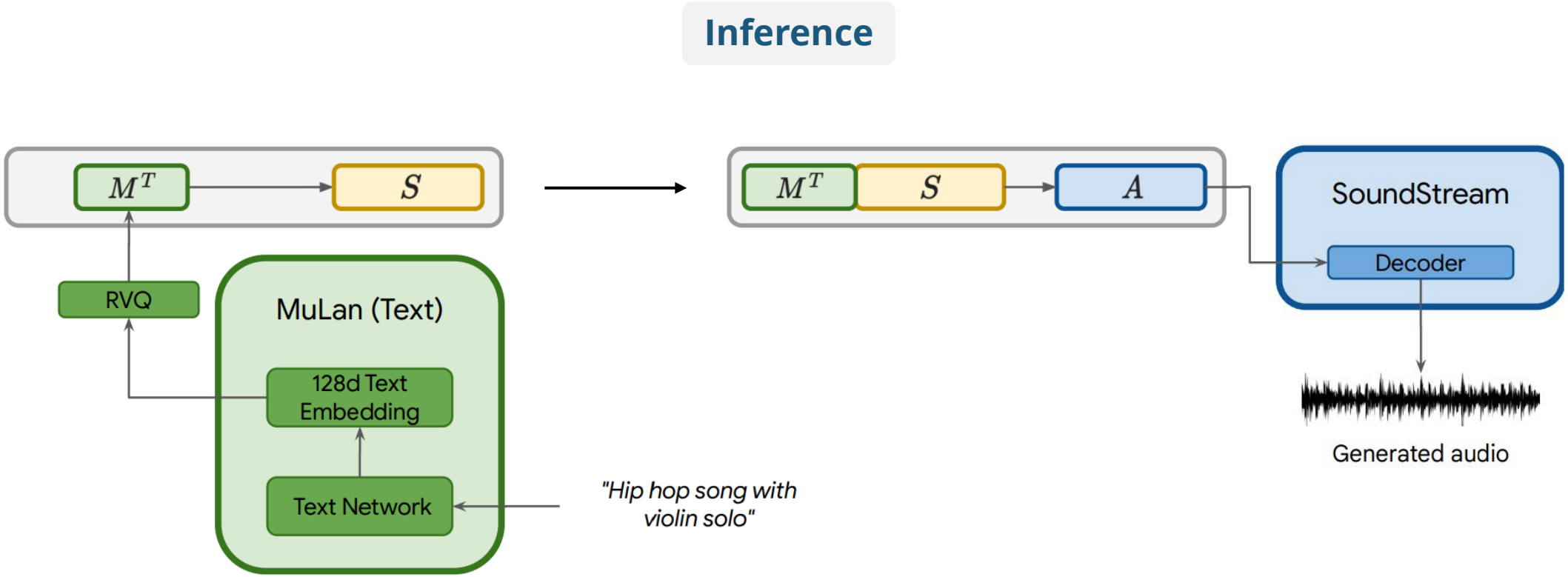
Example: MusicLM (Agostinelli et al., 2023)



(Source: Agostinelli et al., 2022)

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

Example: MusicLM (Agostinelli et al., 2023)



(Source: Agostinelli et al., 2022)

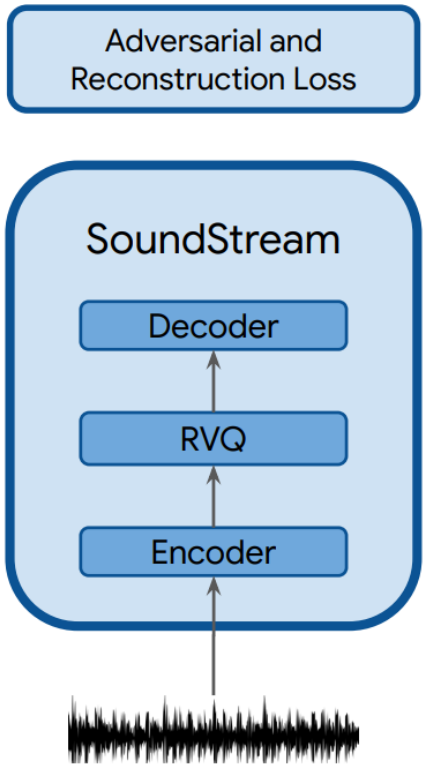
google-research.github.io/seanet/musiclm/examples/

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

Example: MusicLM (Agostinelli et al., 2023)

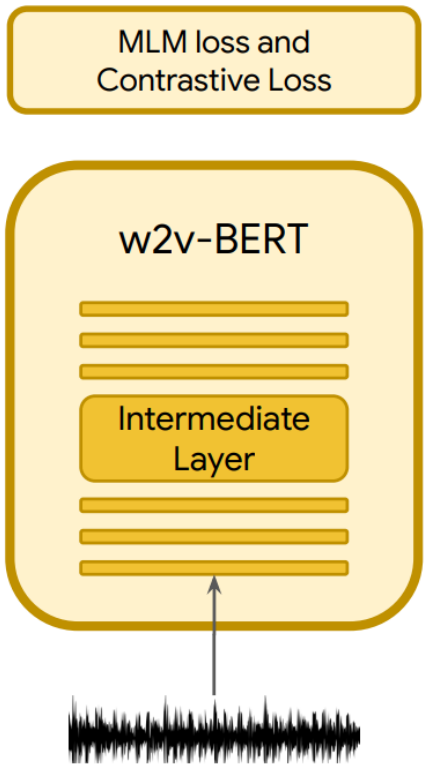
Which is the most challenging component?

Audio autoencoder

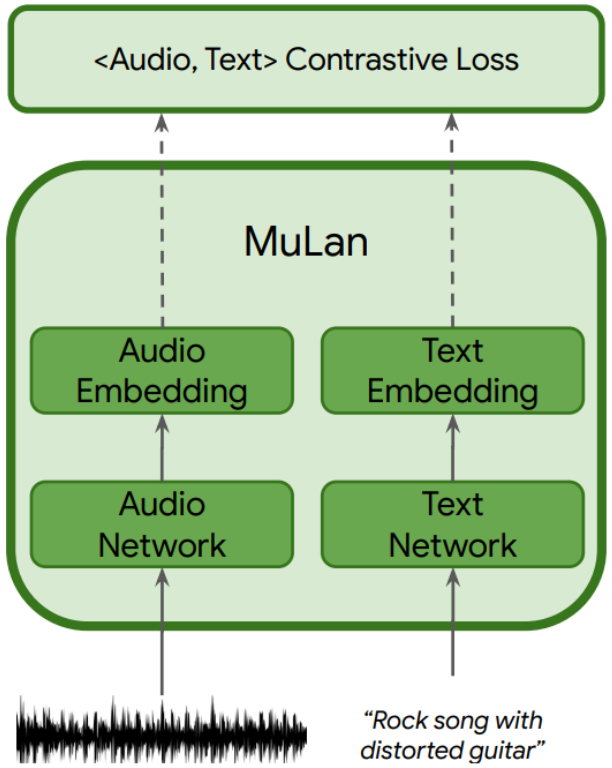


106k songs, 8.2k hours

Semantic representation



Text-music correspondence

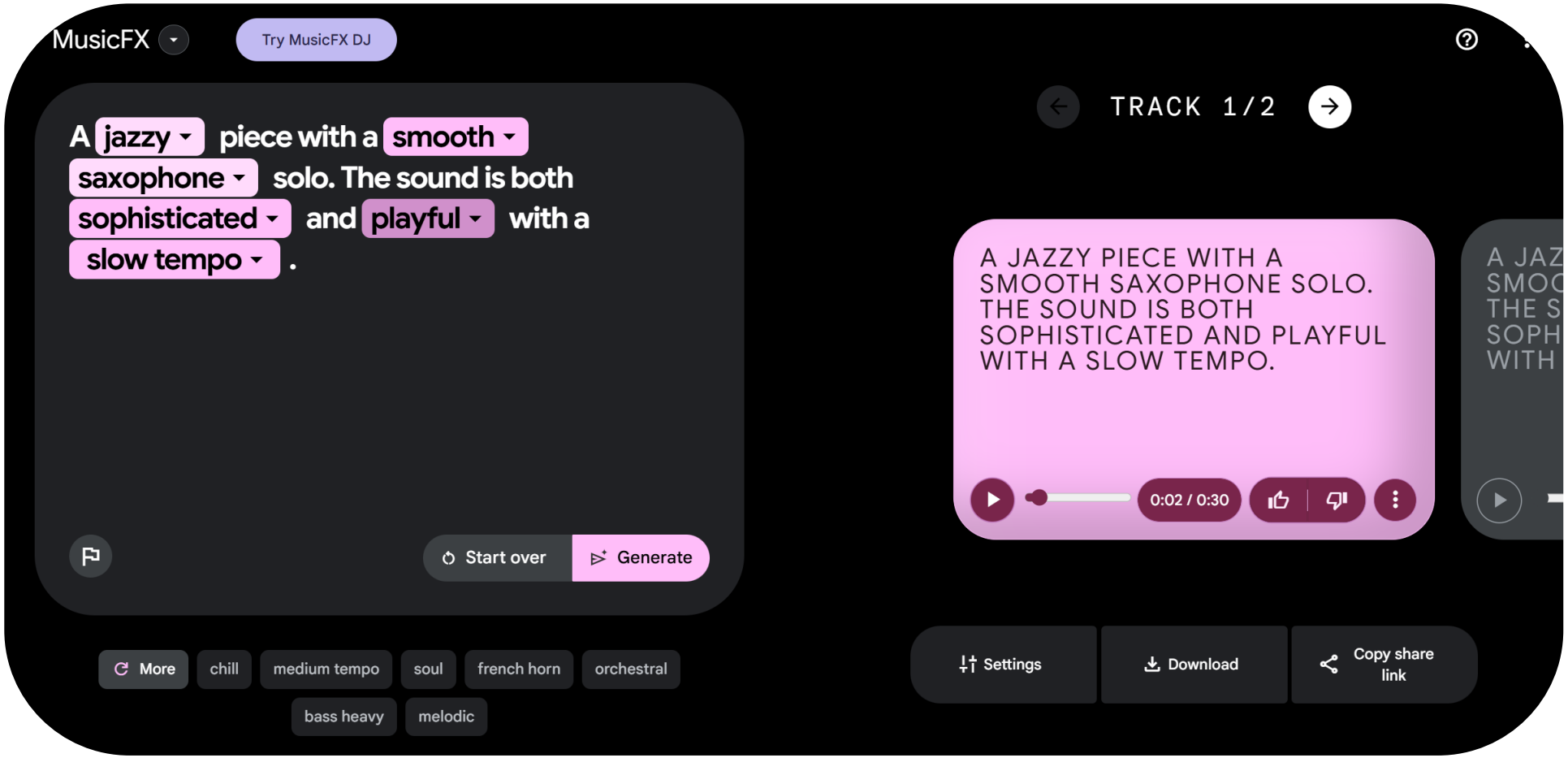


44M 30-sec clips, 370k hours

(Source: Agostinelli et al., 2022)

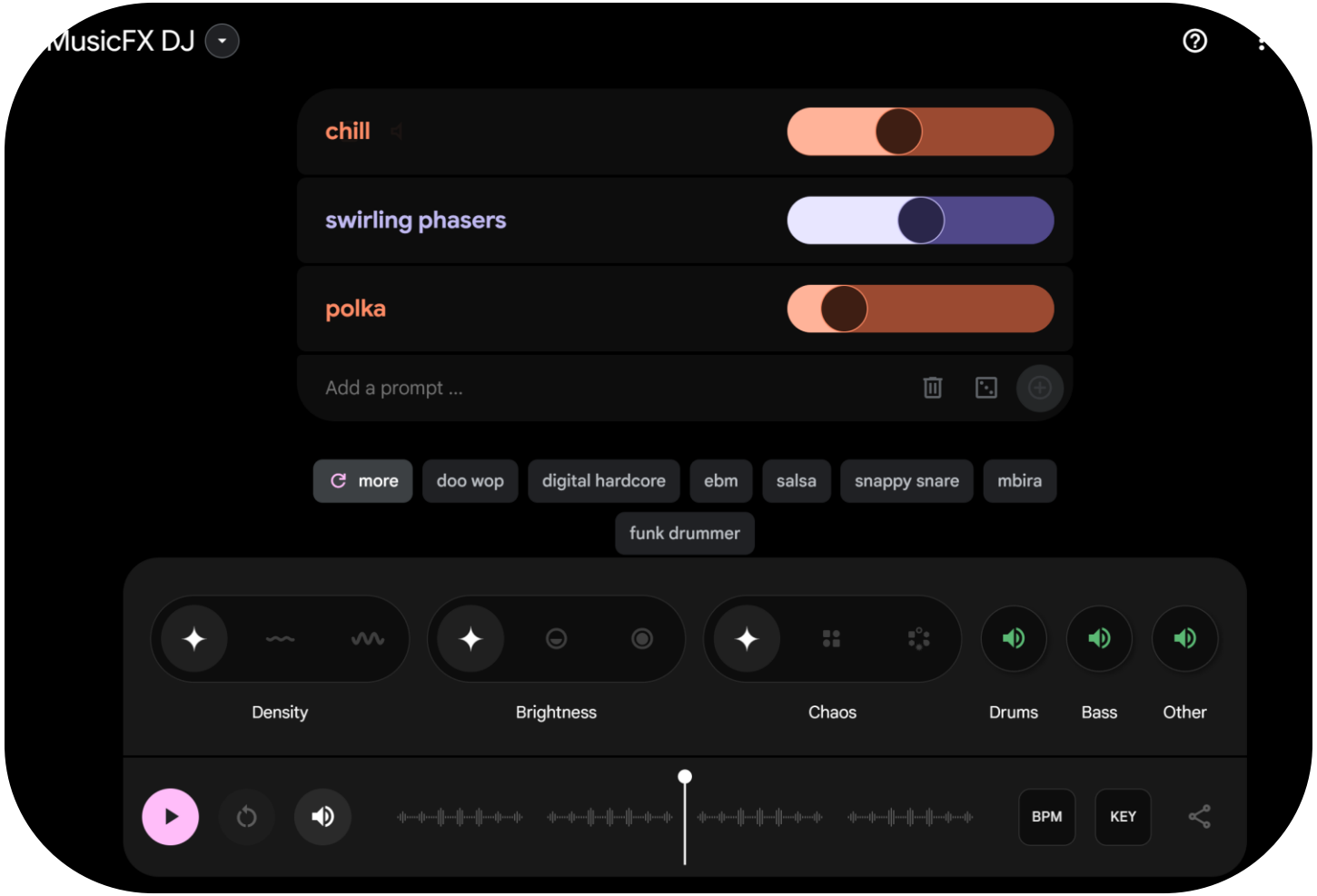
Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

Music FX (2024)



aitestkitchen.withgoogle.com/tools/music-fx

Music FX DJ (2024)

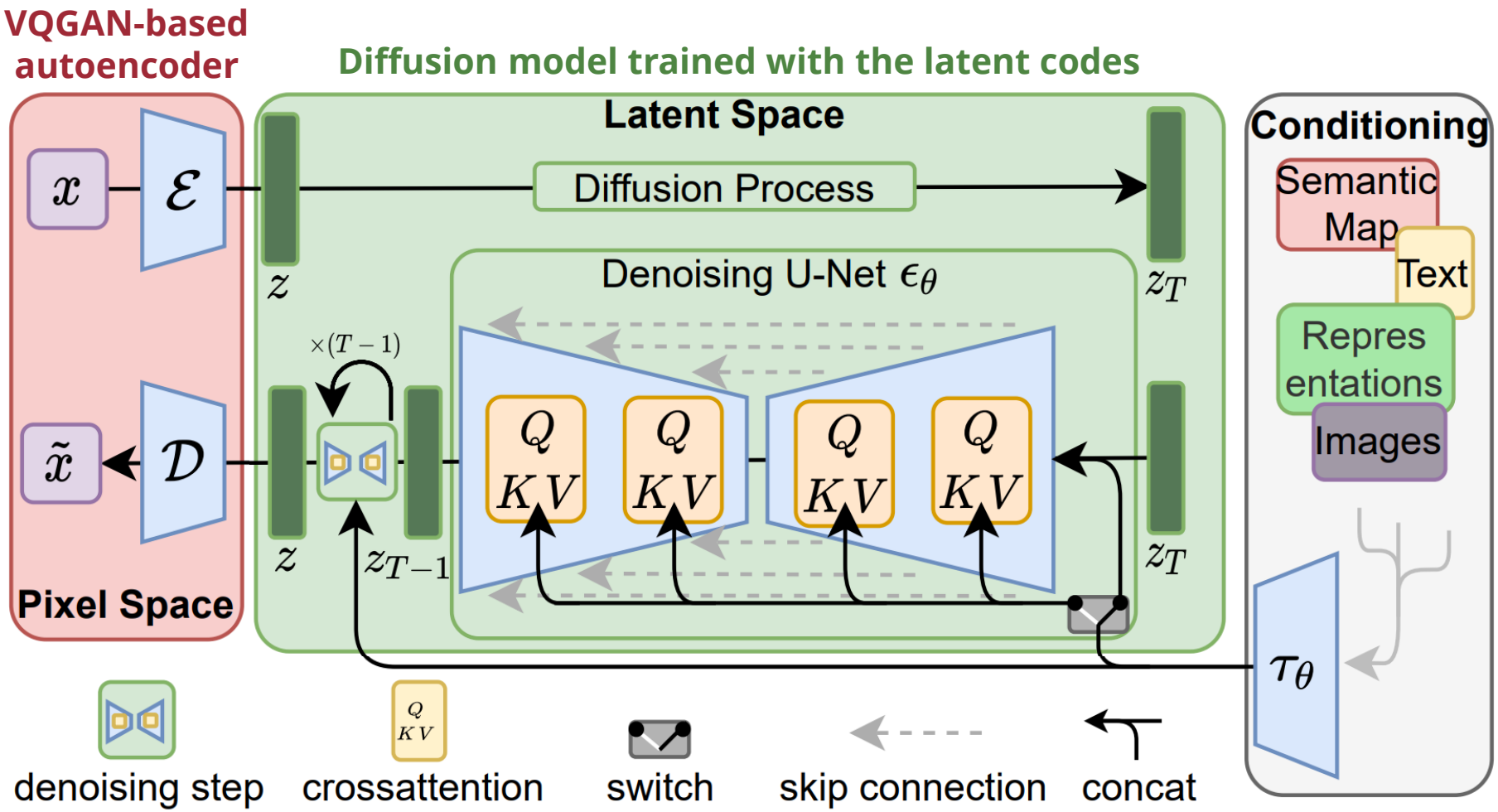


aitestkitchen.withgoogle.com/tools/music-fx-dj

Music FX DJ (2024)

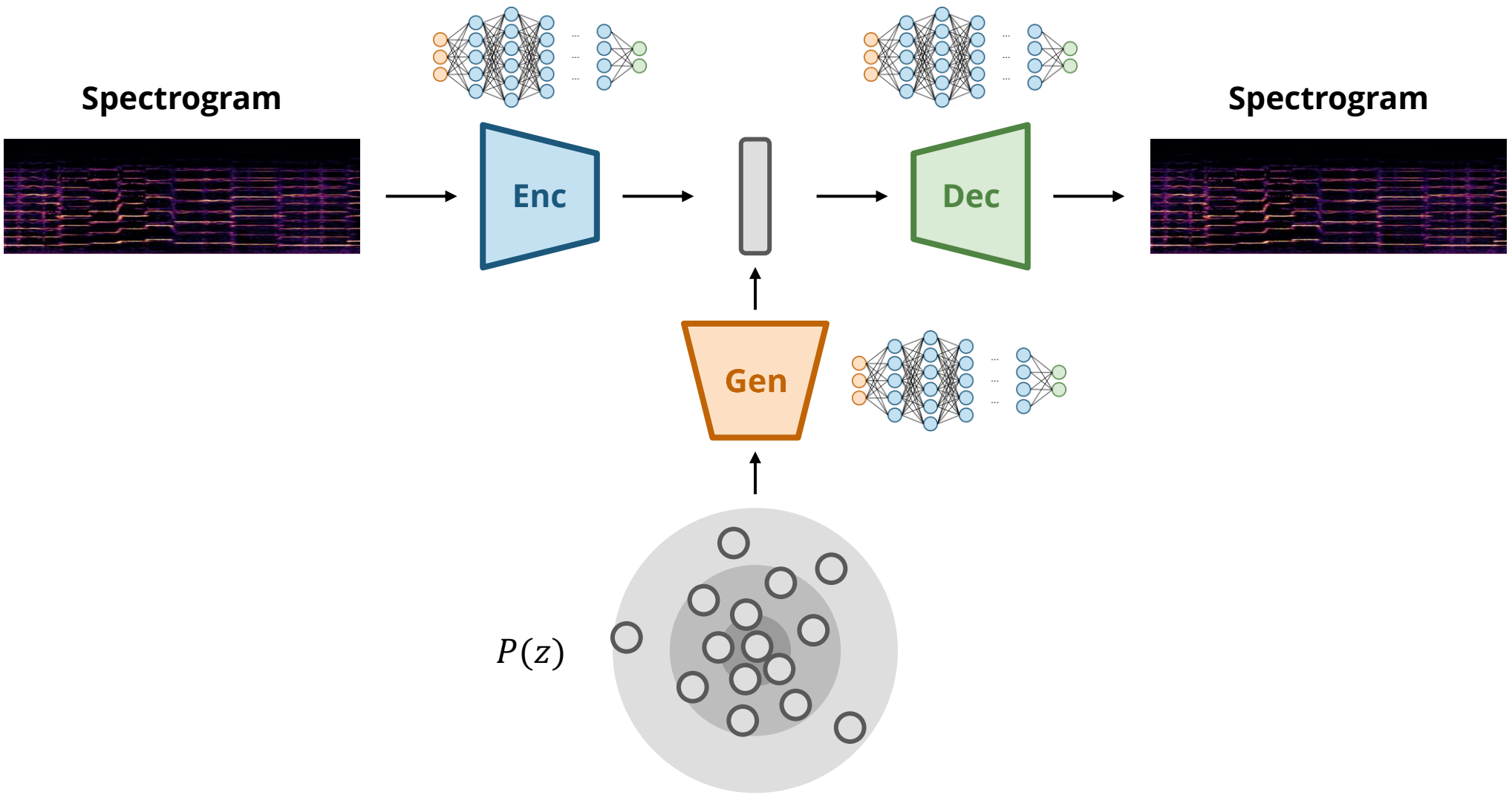
youtube.com/live/IUQW5LgBZvQ

(Recap) Latent Diffusion Models (LDMs)

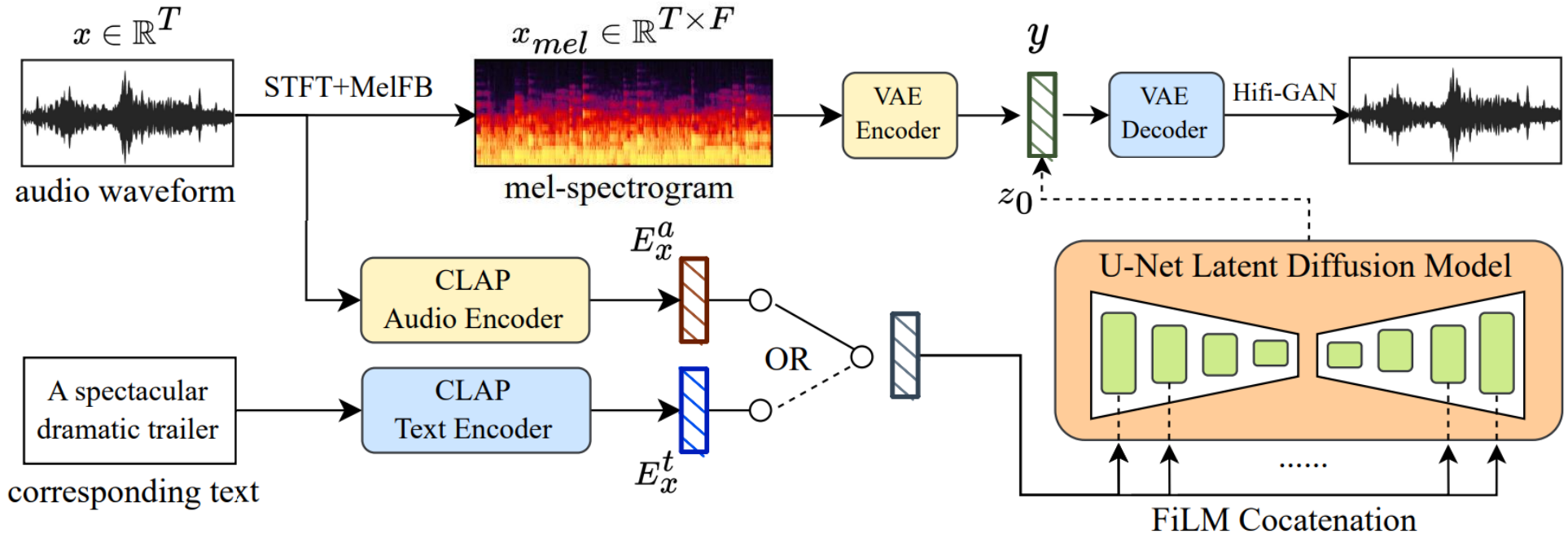


(Source: Rombach et al., 2022)

(Recap) Latent-based Audio Synthesis



Example: MusicLDM (Chen et al., 2023)

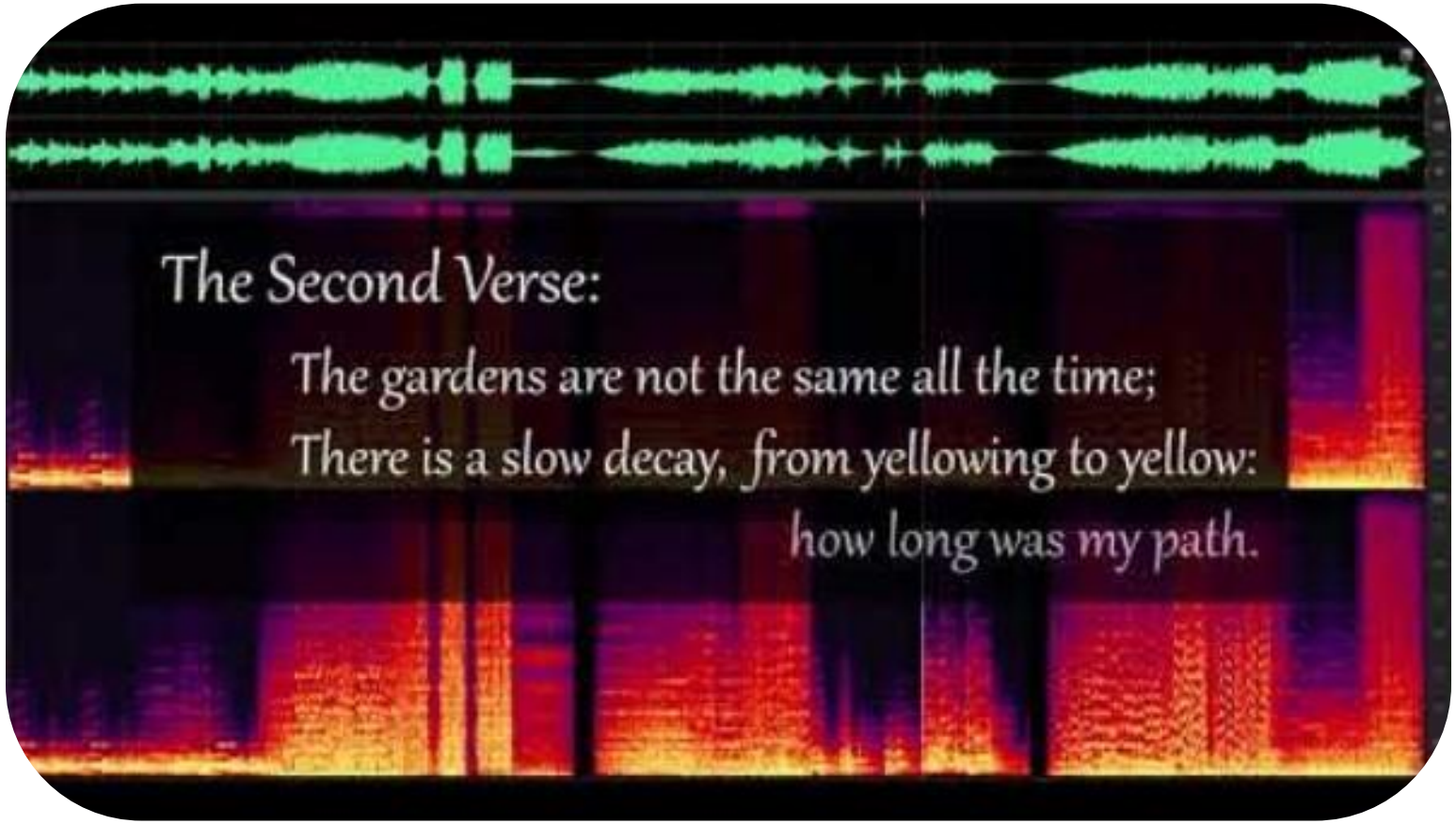


(Source: Ke et al., 2023)

musicldm.github.io

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies," *ICASSP*, 2024.

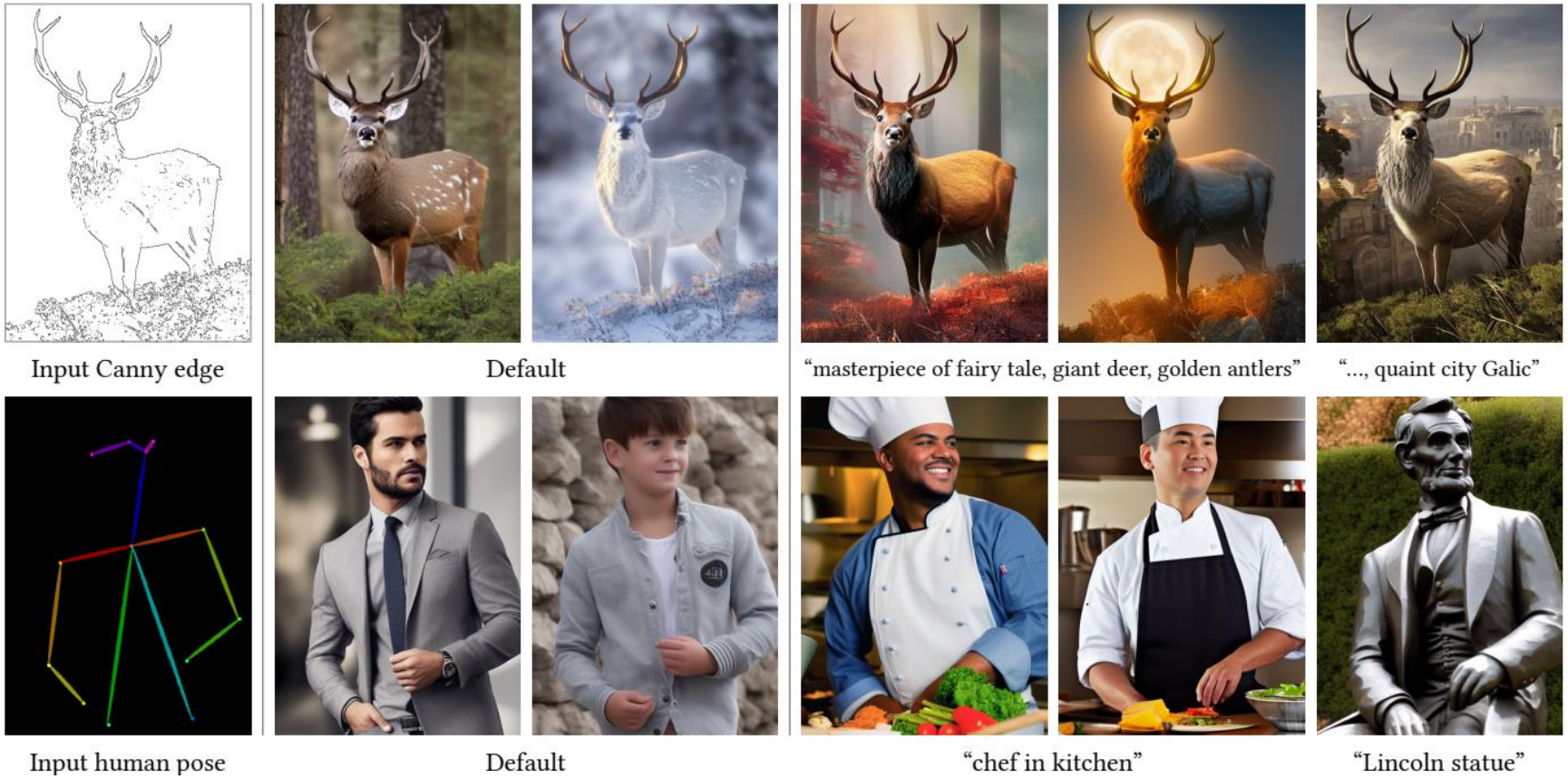
Example: MusicLDM (Chen et al., 2023)



youtu.be/DALv7ea6cv0

Controlling Music Generation Systems

ControlNet (Zhang et al., 2023)

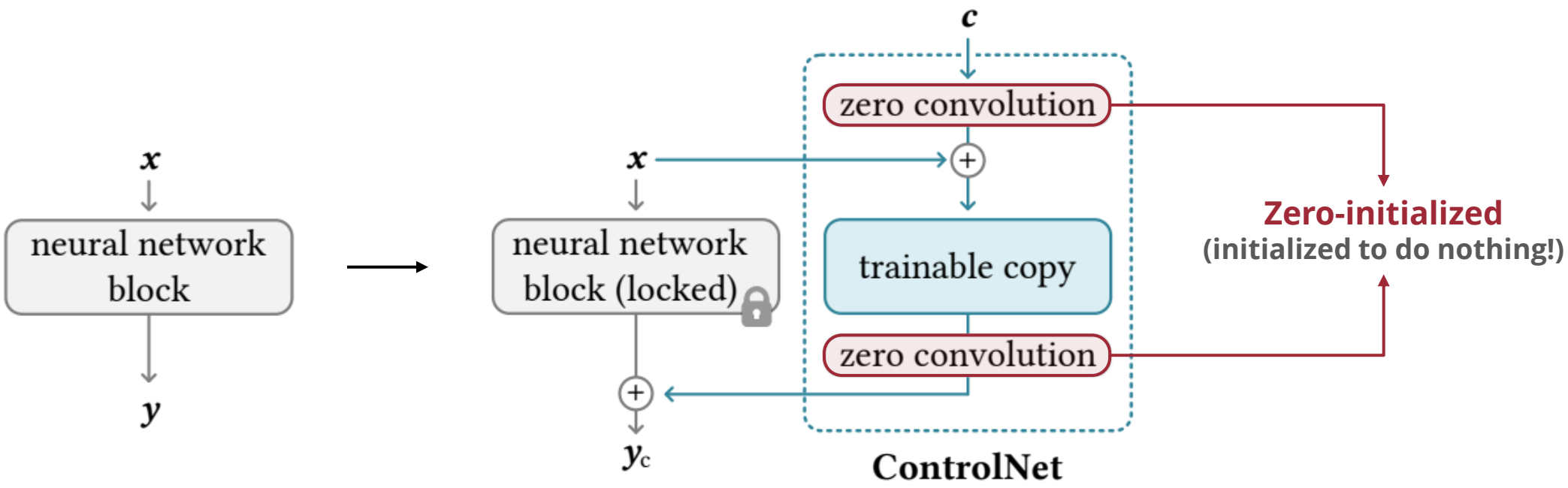


(Source: Zhang et al., 2023)

Can we add controls to a trained text-to-image diffusion model?

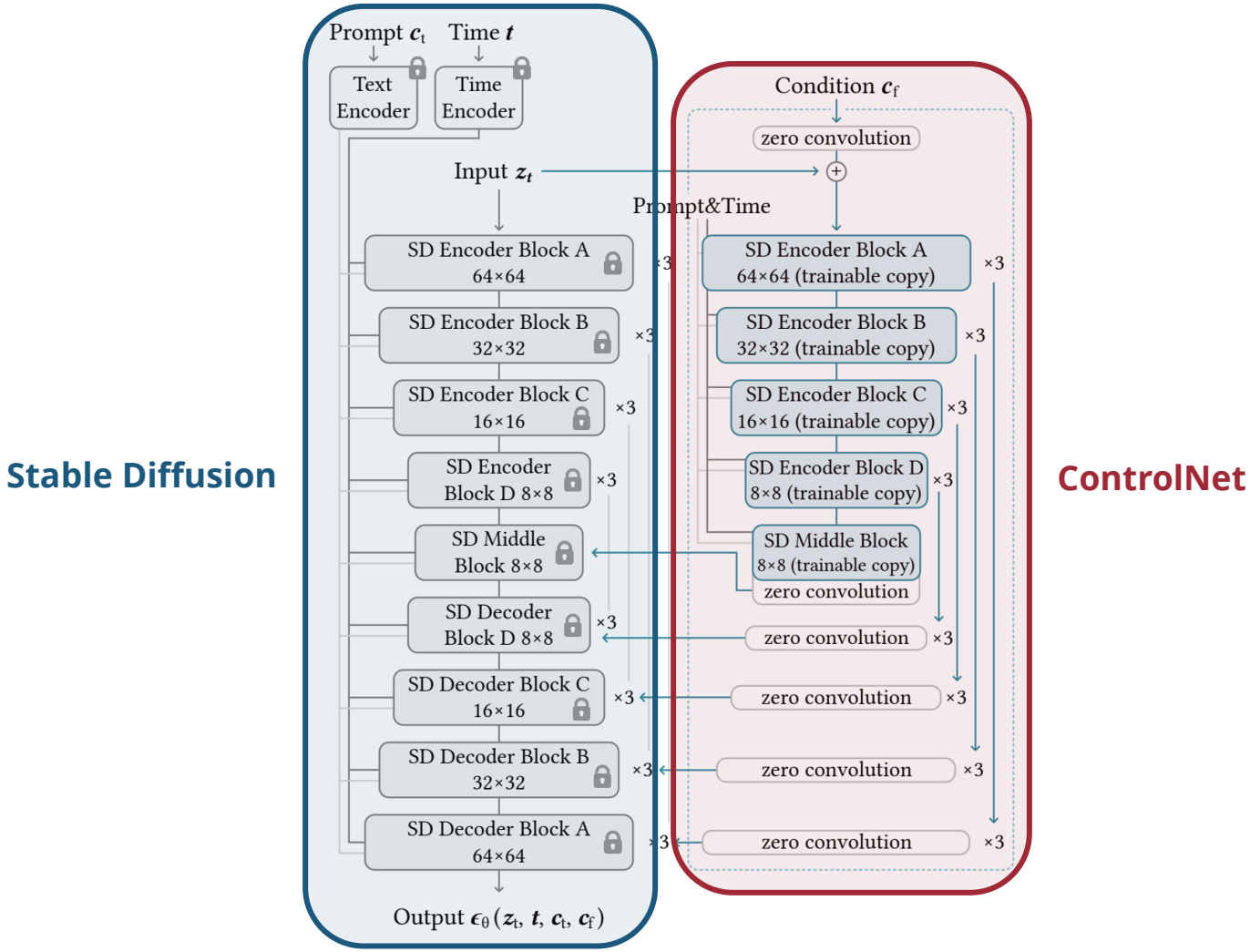
Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," *ICCV*, 2023.

ControlNet (Zhang et al., 2023)



(Source: Zhang et al., 2023)

ControlNet (Zhang et al., 2023)



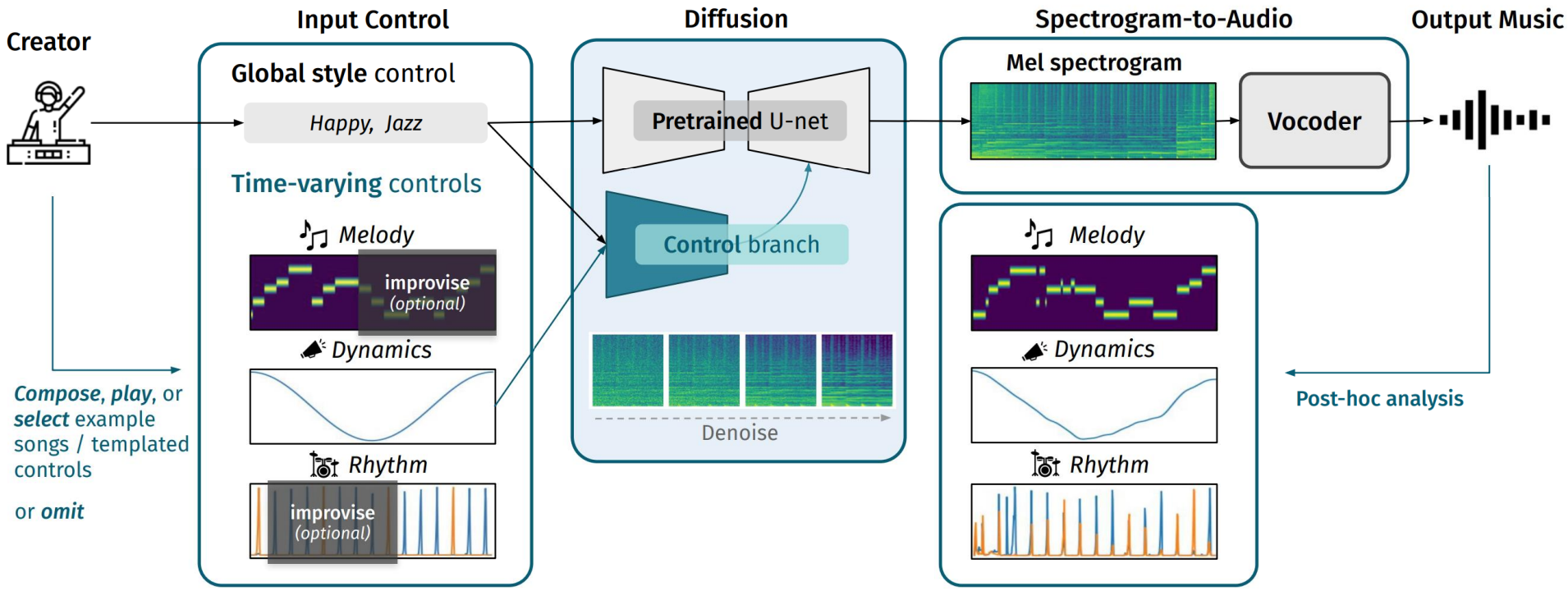
(Source: Zhang et al., 2023)

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," *ICCV*, 2023.

Synthetic Beat Brigade - How would you touch me? (2023)

youtu.be/O4cJ3acEGDw &
drive.google.com/file/d/1QTQ7P3iZI6I0anIwNQ3ewf8g3JjDjesl/view

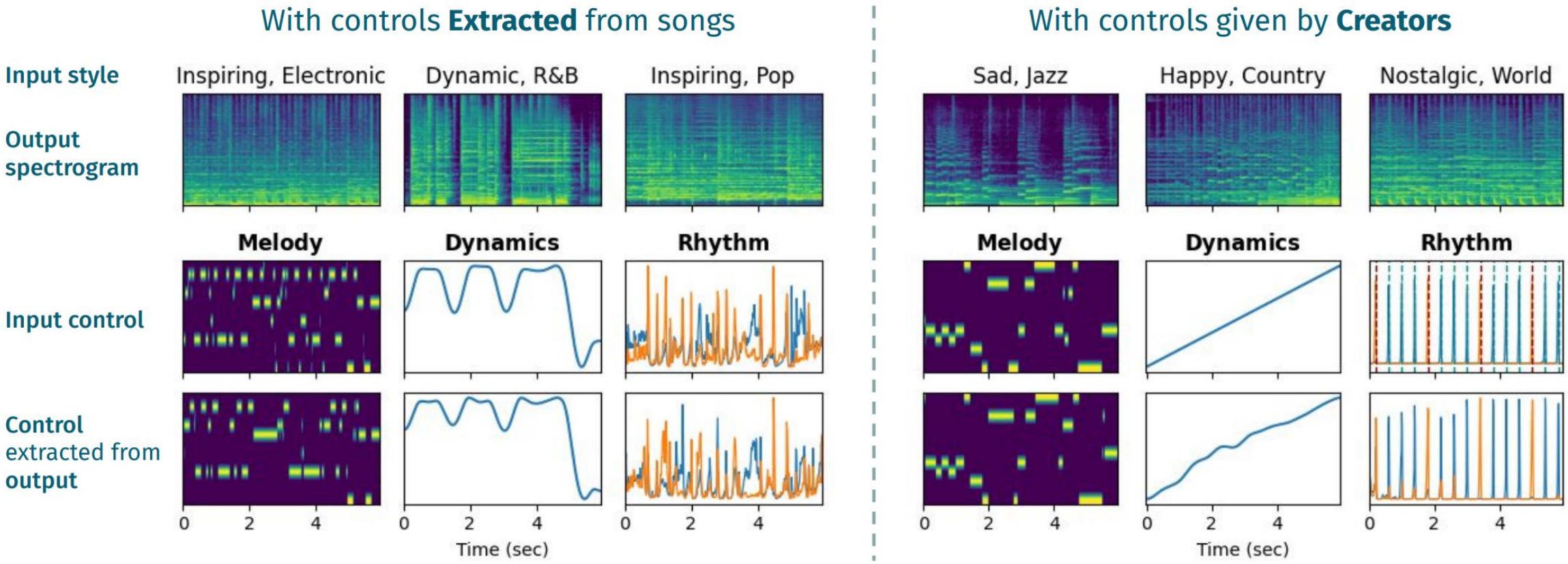
Example: Music ControlNet (Wu et al., 2024)



(Source: Wu et al., 2024)

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan, "Music ControlNet: Multiple Time-varying Controls for Music Generation," *TASLP*, 2024.

Example: Music ControlNet (Wu et al., 2024)

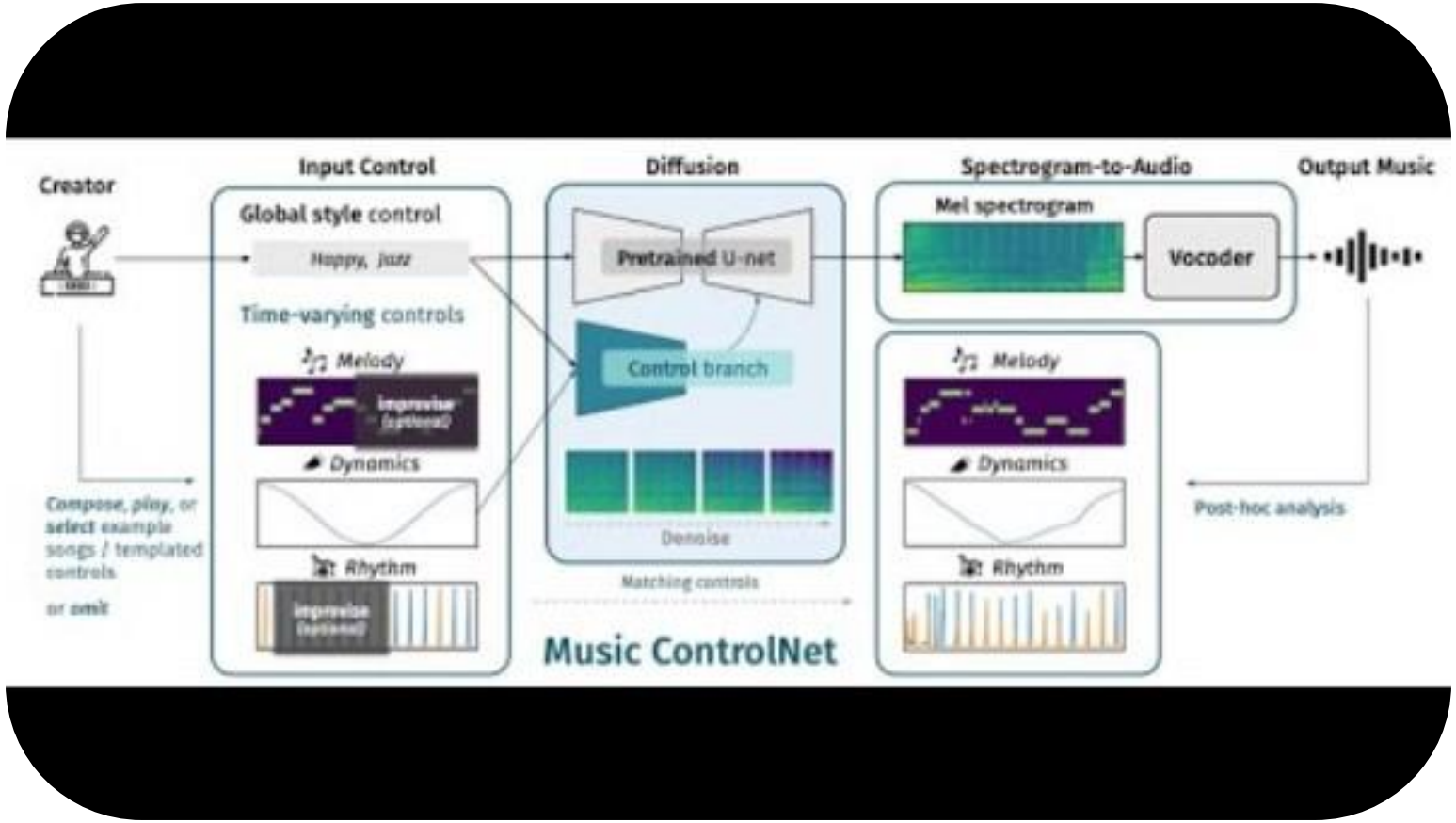


(Source: Wu et al., 2024)

musiccontrolnet.github.io/web

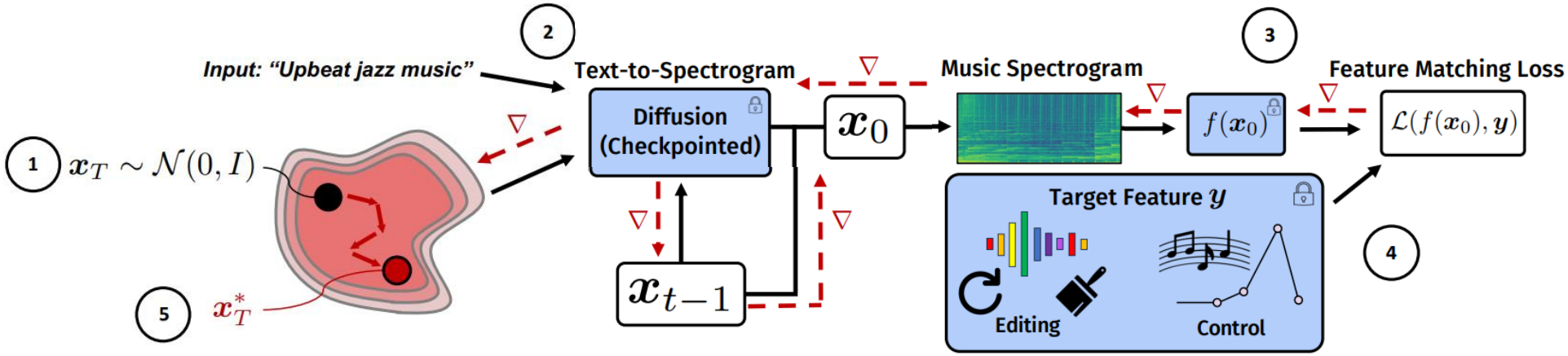
Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan, "Music ControlNet: Multiple Time-varying Controls for Music Generation," *TASLP*, 2024.

Example: Music ControlNet (Wu et al., 2024)



youtu.be/QVr-S-DyccU

Example: DITTO (Novack et al., 2024)

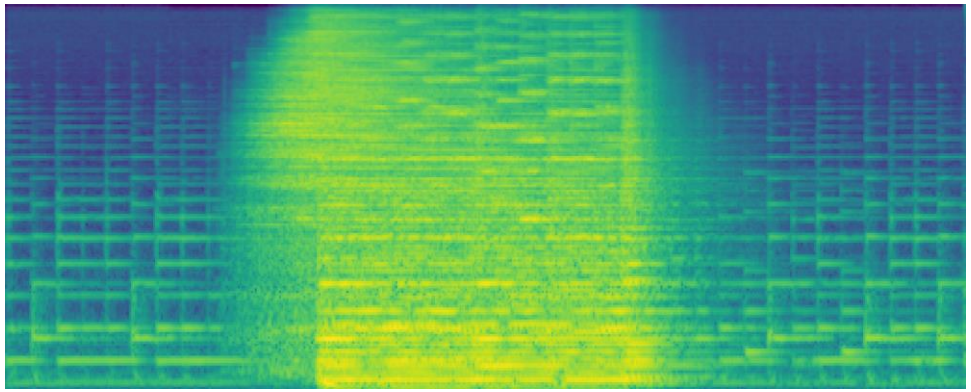


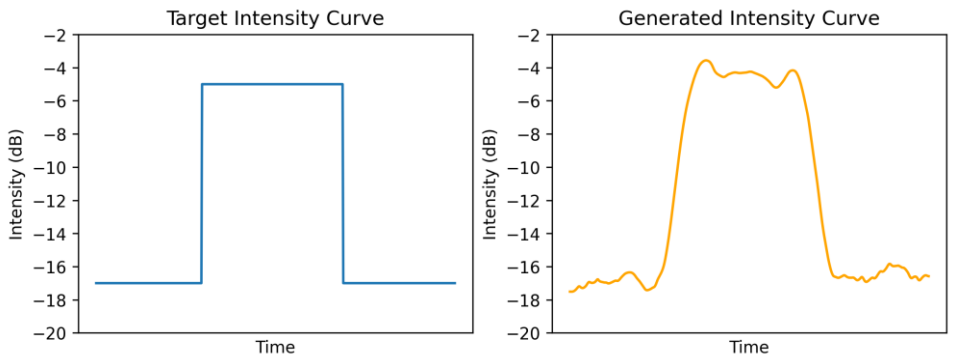
(Source: Novack et al., 2024)

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan, "DITTO: Diffusion Inference-Time T-Optimization for Music Generation," *ICML*, 2024.

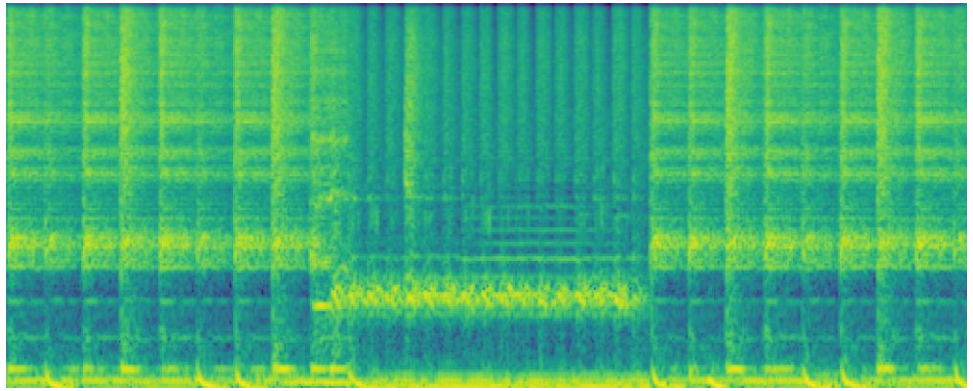
Example: DITTO (Novack et al., 2024)

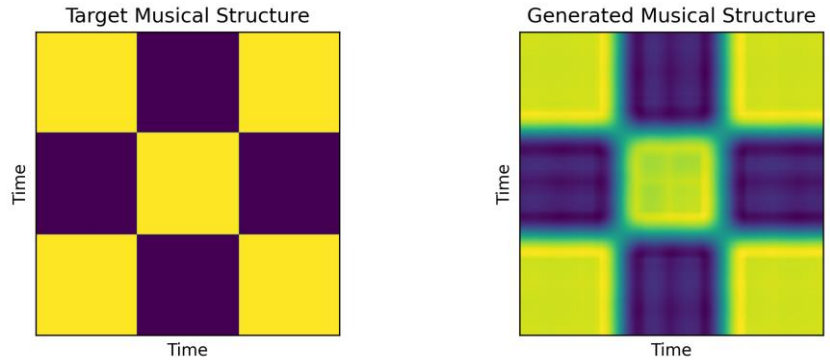
Intensity control





Structure control





(Source: Novack et al., 2024)

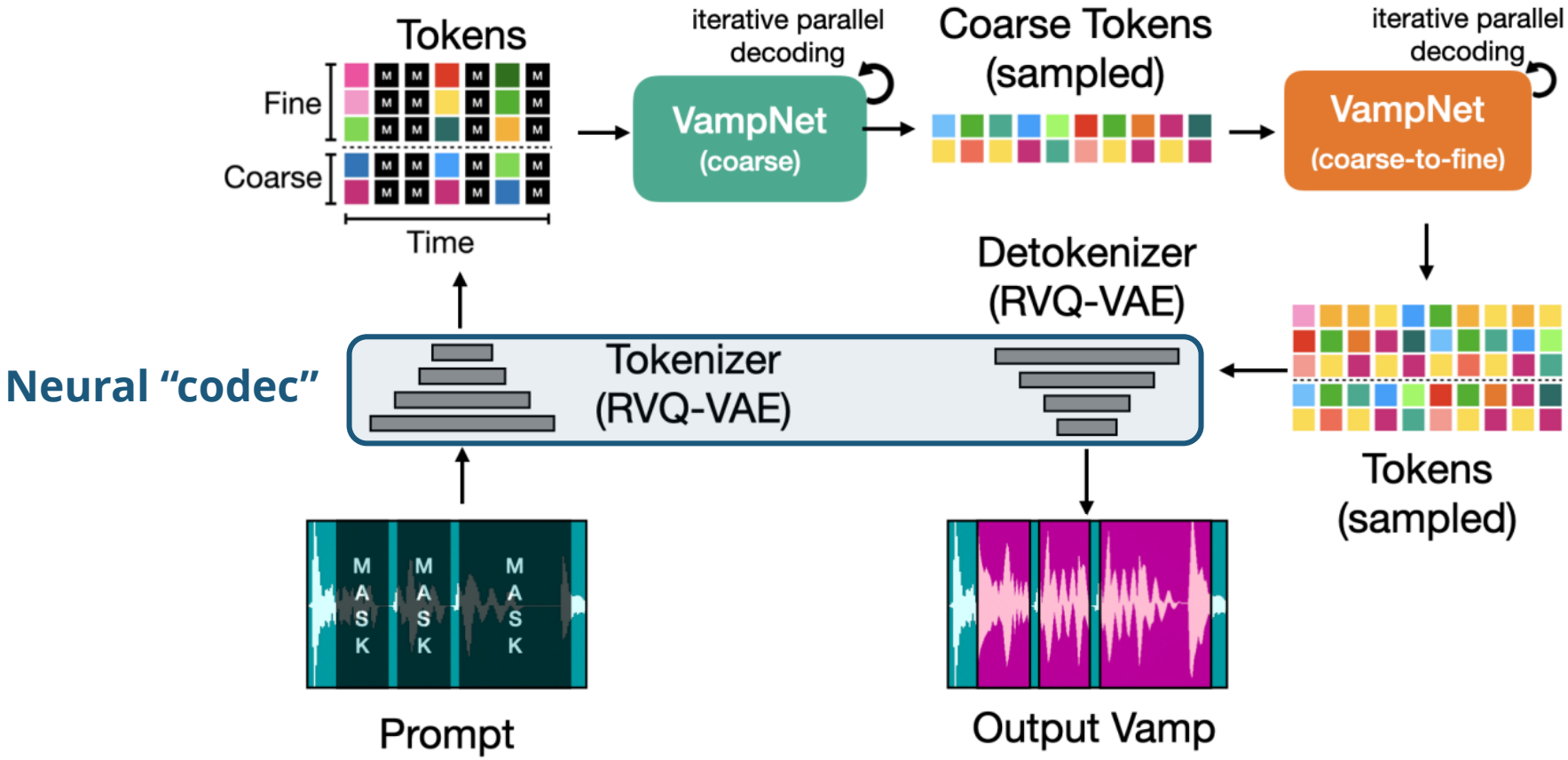
Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan, "DITTO: Diffusion Inference-Time T-Optimization for Music Generation," *ICML*, 2024.

DITTO: Diffusion for Music Generation (2024)

youtu.be/KooosSNPNo8 & ditto-music.github.io/web/

Creative Applications of Music Generation Systems

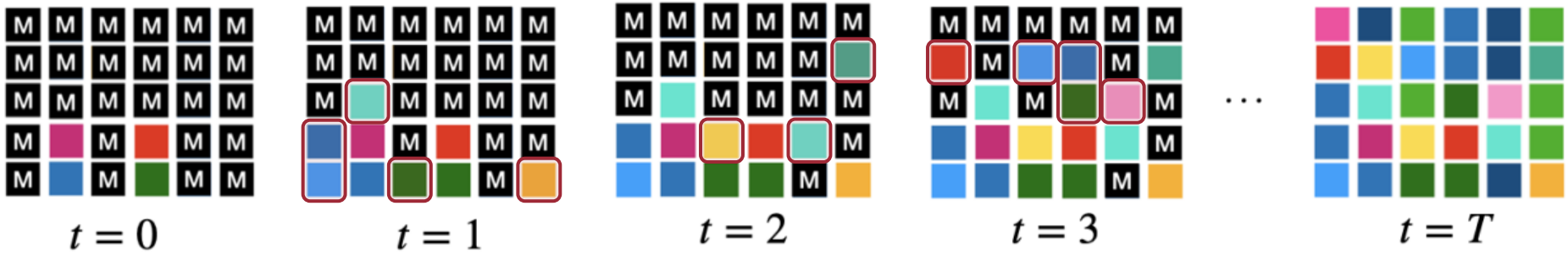
Example: VampNet (Garcia et al., 2023)



(Source: Garcia et al., 2023)

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo, "VampNet: Music Generation via Masked Acoustic Token Modeling," *ISMIR*, 2023.

Example: VampNet (Garcia et al., 2023)

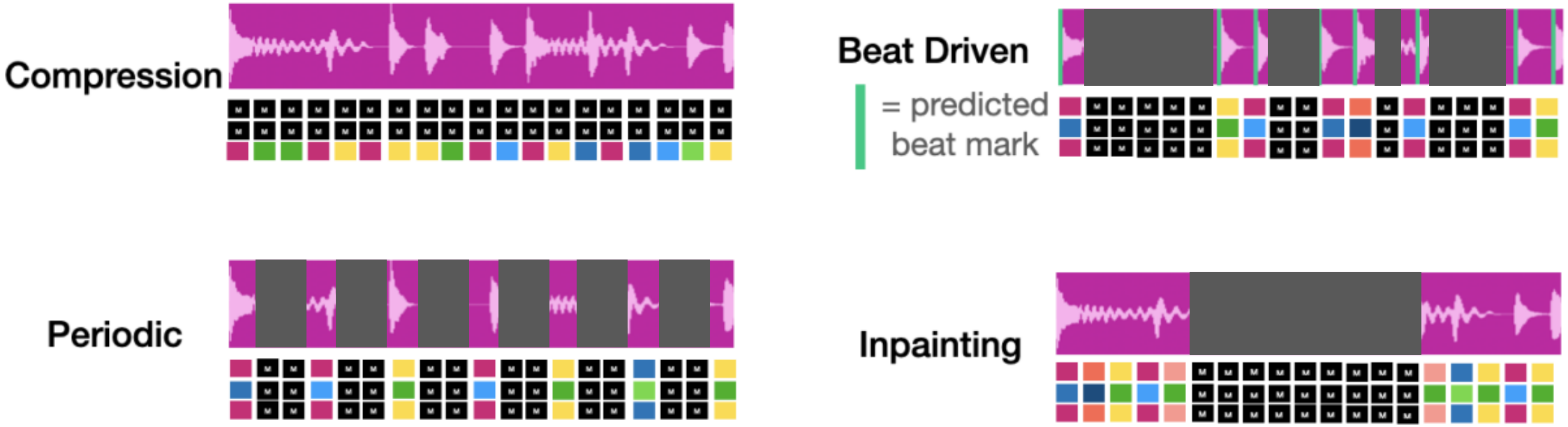


(Source: Garcia et al., 2023)

Sample a subset of the most confident predicted tokens in each iteration

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo, "VampNet: Music Generation via Masked Acoustic Token Modeling," *ISMIR*, 2023.

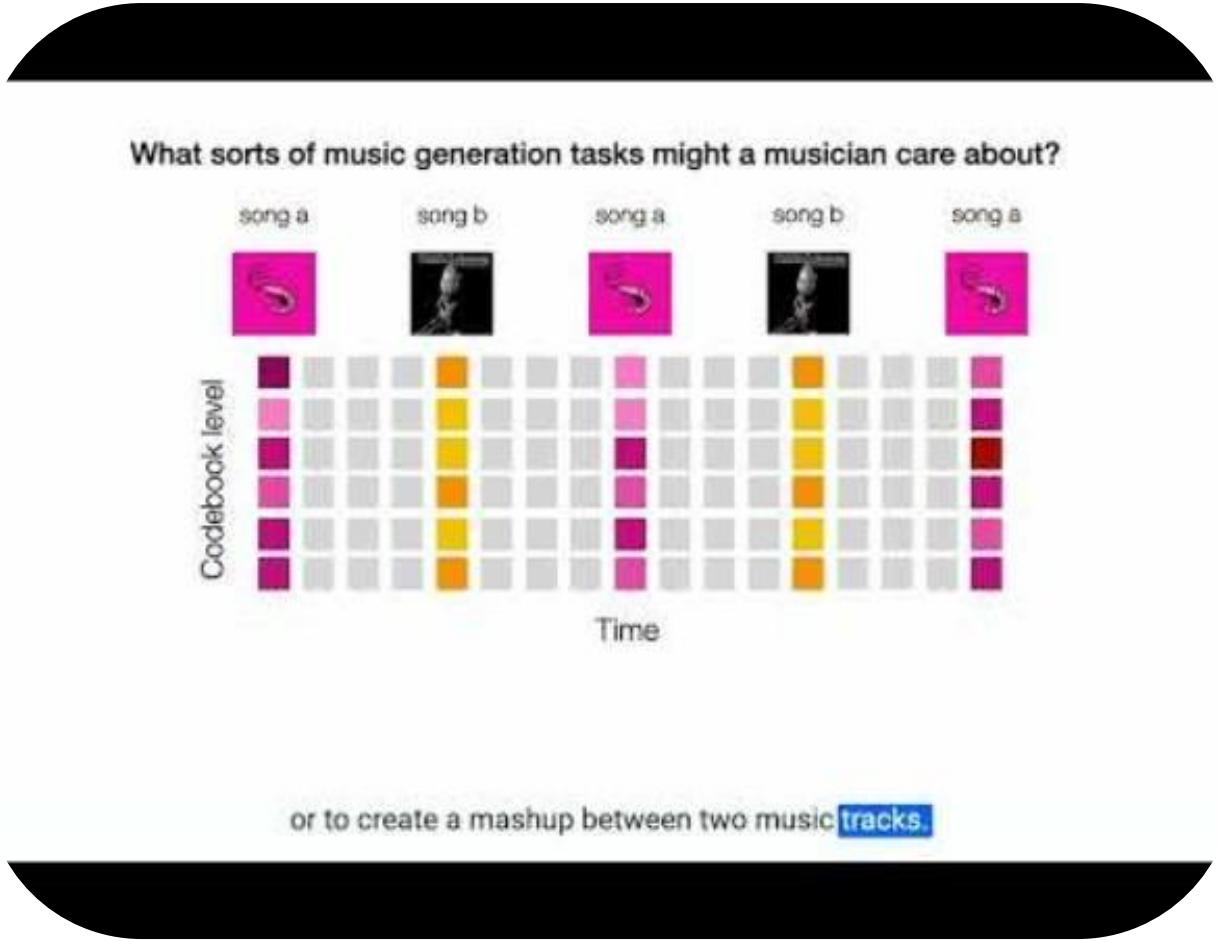
Example: VampNet (Garcia et al., 2023)



(Source: Garcia et al., 2023)

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo, "VampNet: Music Generation via Masked Acoustic Token Modeling," *ISMIR*, 2023.

Example: VampNet (Garcia et al., 2023)



youtu.be/3XfeWlV9Cp0

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo, "VampNet: Music Generation via Masked Acoustic Token Modeling," *ISMIR*, 2023.

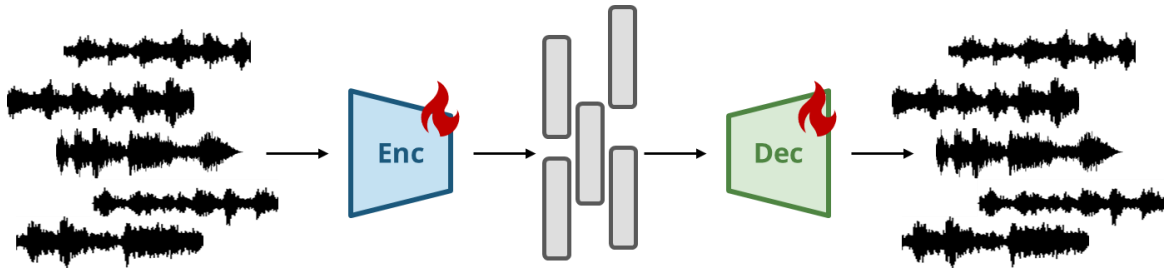
Example: unloop (Garcia et al., 2023)

youtu.be/yzBI8Vcjd2s

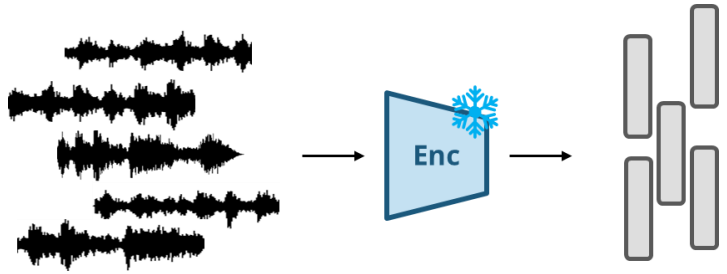
github.com/hugofloresgarcia/unloop

Pipeline

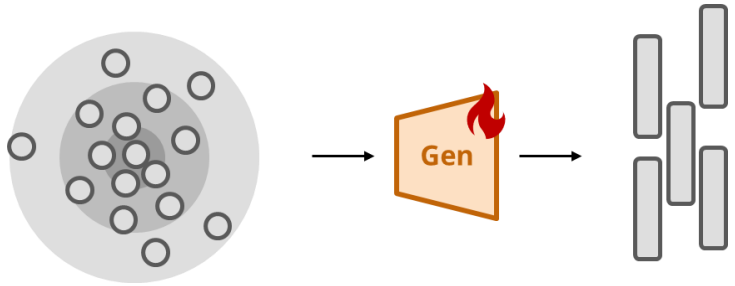
Step 1: Train an Autoencoder



Step 2: Compute the Latent Vectors



Step 3: Train a Latent Generative Model



Step 4: Decode the Latent Vectors