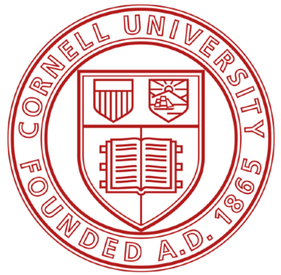
Lecture 24: Audio generation

CS 5788: Introduction to Generative Models



Reminder: Final project!

Final Project Guidelines

Posted: Tuesday, April 21, 2026	**Due:** Tuesday, May 12, 2026
Please submit your written report plus a Jupyter notebook demonstrating your code to Gradescope as a `.pdf` file.	

Deliverables. The final project will have **four** deliverables:

1. **Written report in CVPR style (template: link)**
   - This is the main way that we will evaluate your project.
   - **Page limit: 4 pages.**
2. **Code repository (as a zip file)**
3. **Jupyter notebook**
   - Supplement the written report with a Jupyter notebook demo.
   - We expect to see how you ran the code to get the results in the written report, in a similar format to homework problems.
   - *Note:* The notebook is only for demonstrating your code. All main results should be included in the written report.

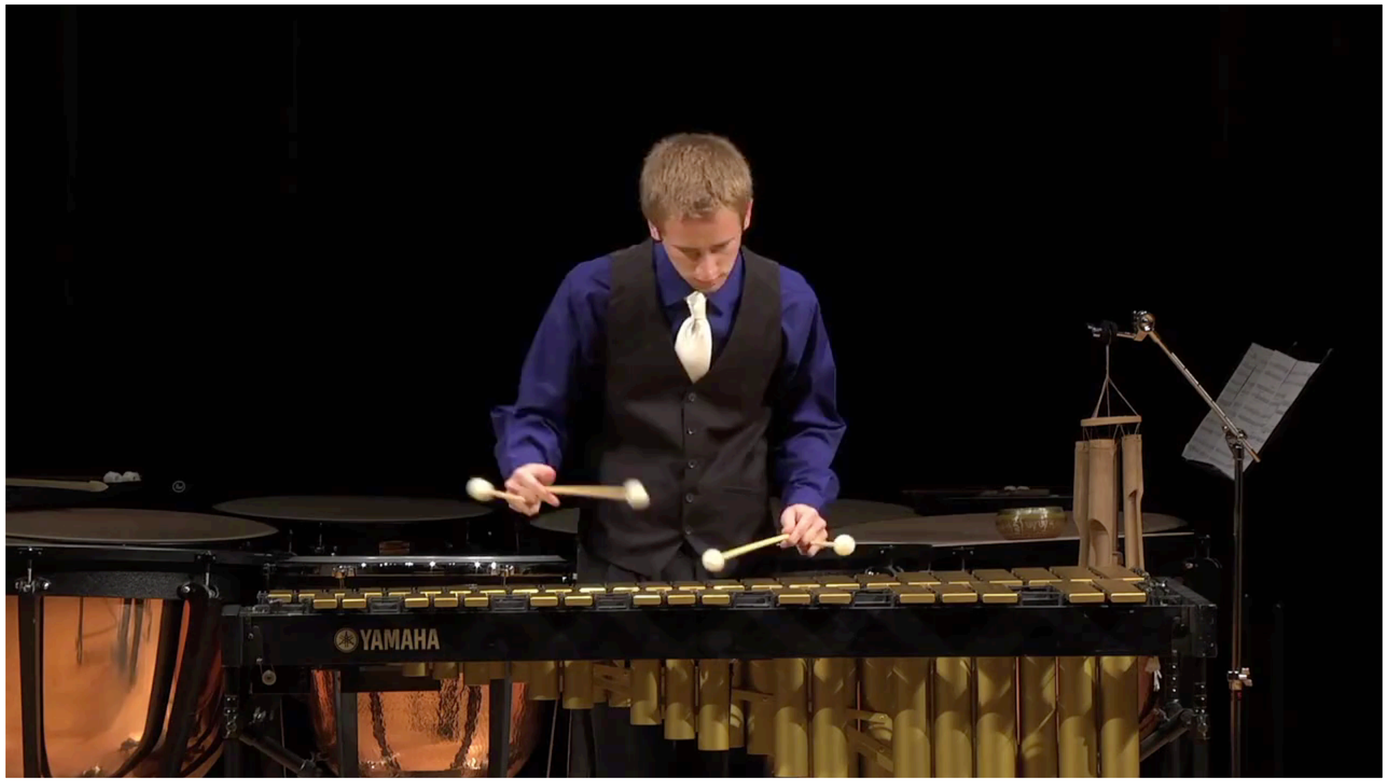
Please sign up for a presentation slot!

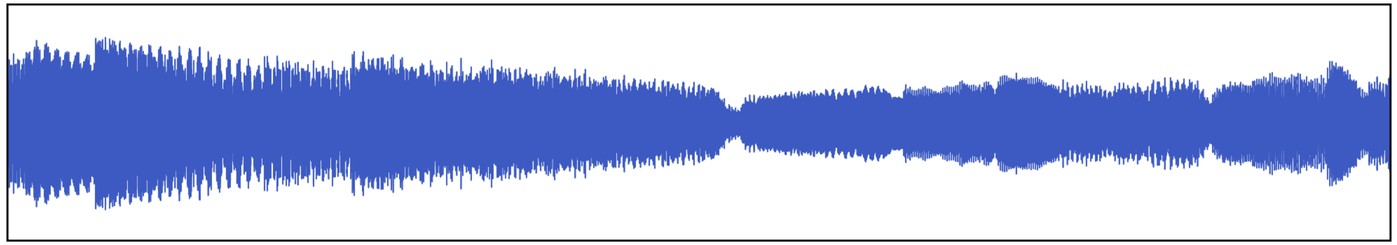
https://docs.google.com/spreadsheets/d/1h77pfR9OojyPIZSzuiTjEwA5k7SklINQupmfCIEtSyo/edit?gid=0#gid=0

Today

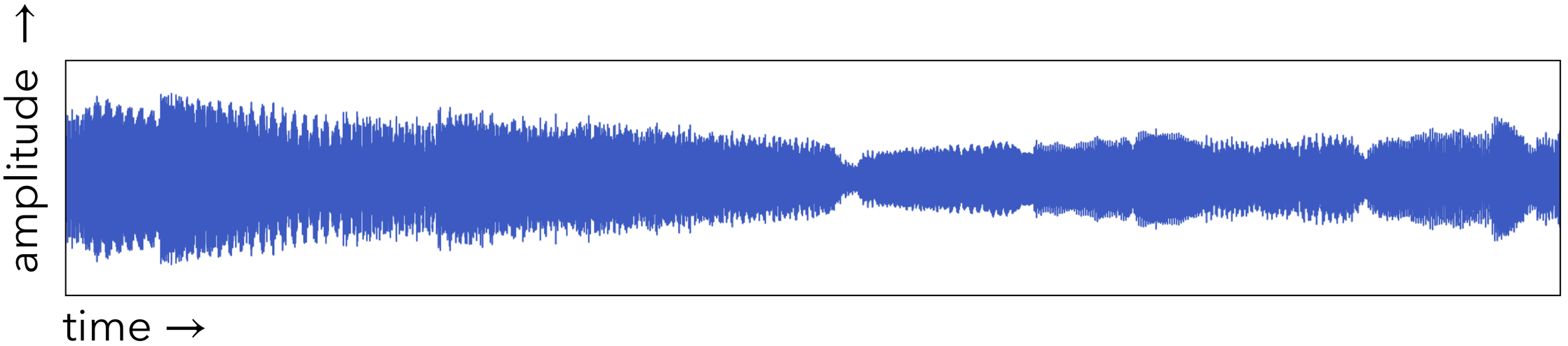
- Audio generation
- Cross-modal generation

How do we represent audio?



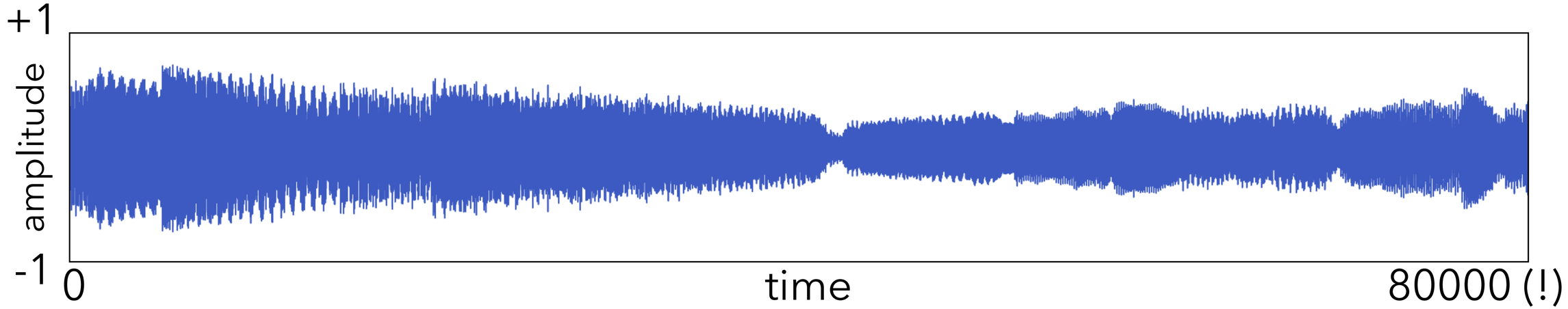


Audio waveform

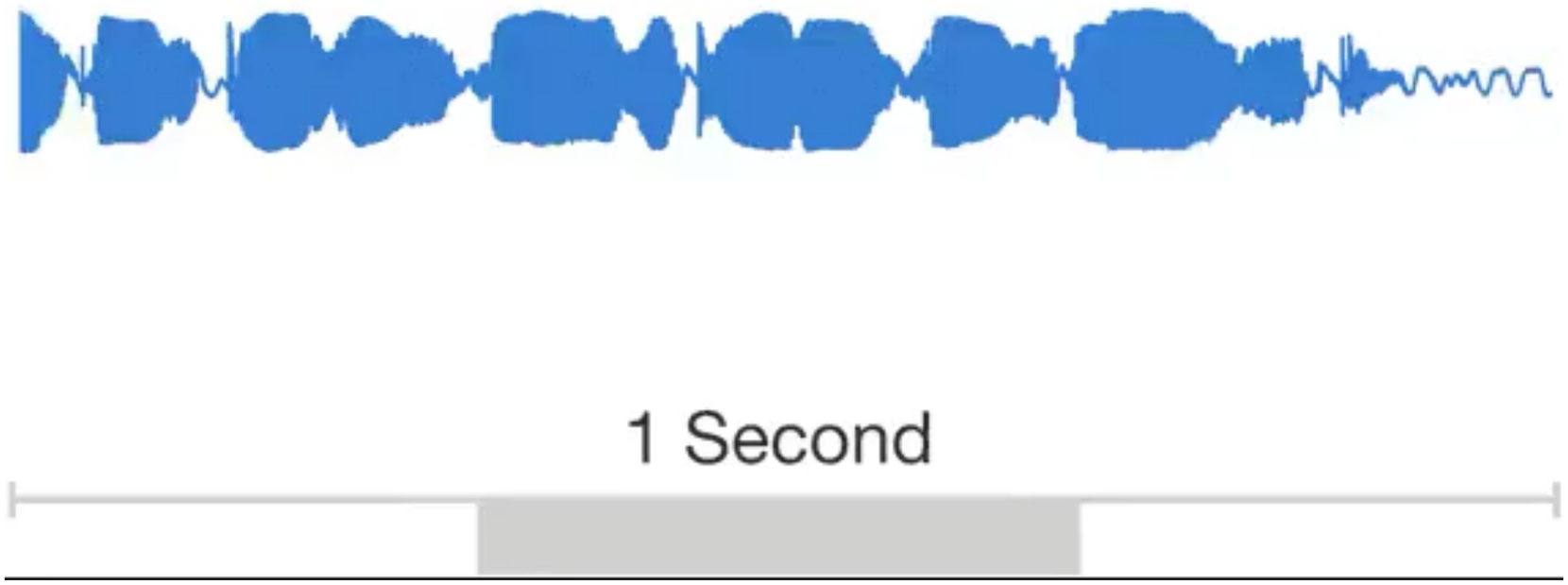


- The x axis is time, y axis is the amplitude, which conveys the sound pressure.
- What about the domain and range for this 5 second clip?

Audio waveform

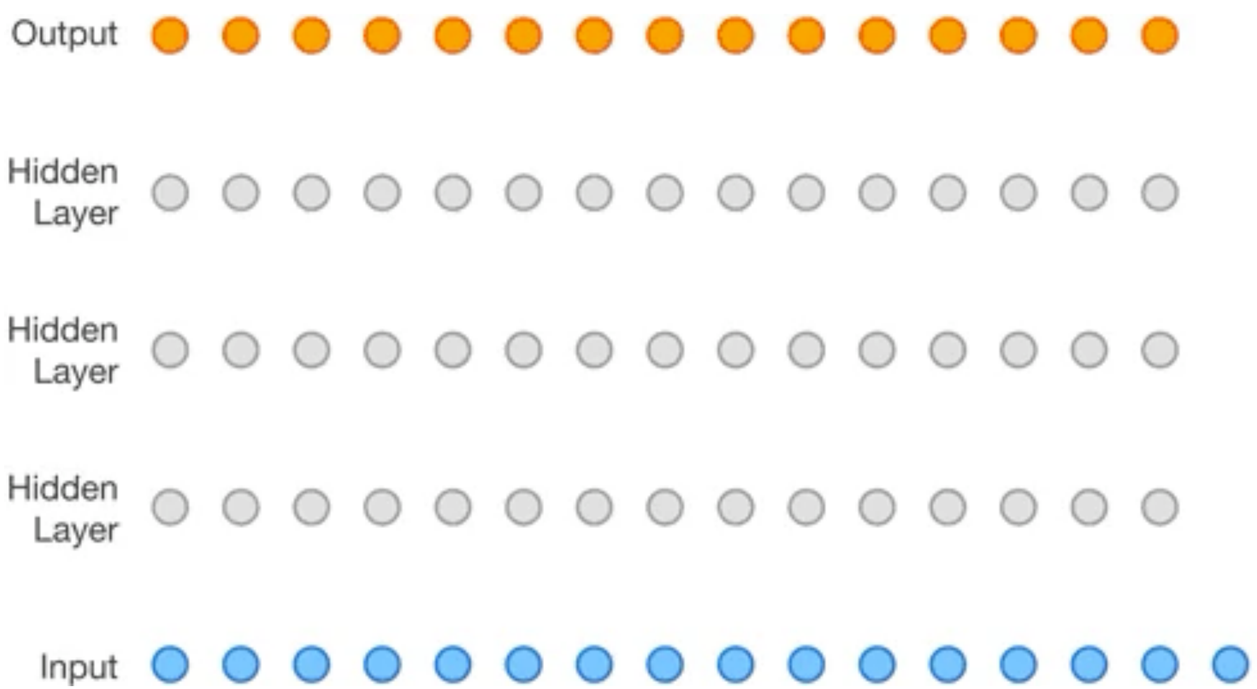


- The x axis is time, y axis is the amplitude, which conveys the sound pressure.
- What about the domain and range for this 5 second clip?
  - High dimensional! Even at 16 Khz, we have 80K samples in a 5 sec. clip.
- Can try to generate waveform directly anyway.



Source: [van den Oord et al., “WaveNet”, 2016]

Autoregressive generation with WaveNet



- Quantize the samples into 256 discrete values (would be 65,536 under a naive quantization scheme).
  - Use logarithmic scale.
- Series of dilated convolutional layers, performing autoregressive generation.

Source: [van den Oord et al., “WaveNet”, 2016]

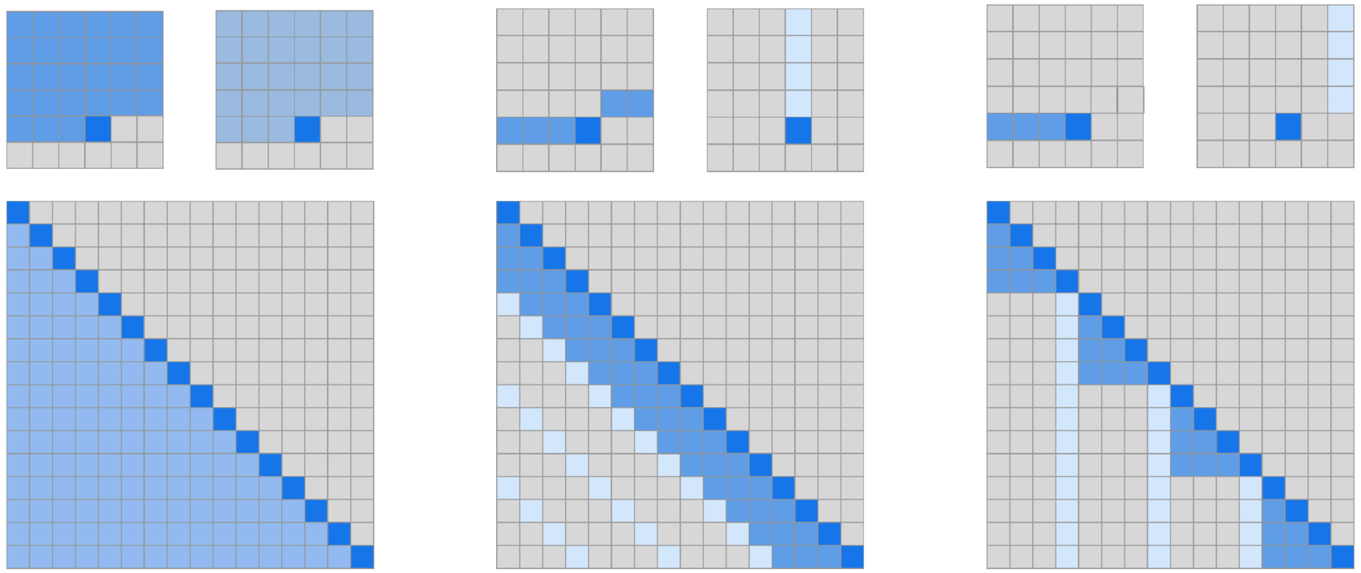
Autoregressive generation with WaveNet

Speech generation

Music generation

Source: [van den Oord et al., “WaveNet”, 2016]

More recent autoregressive waveform models



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

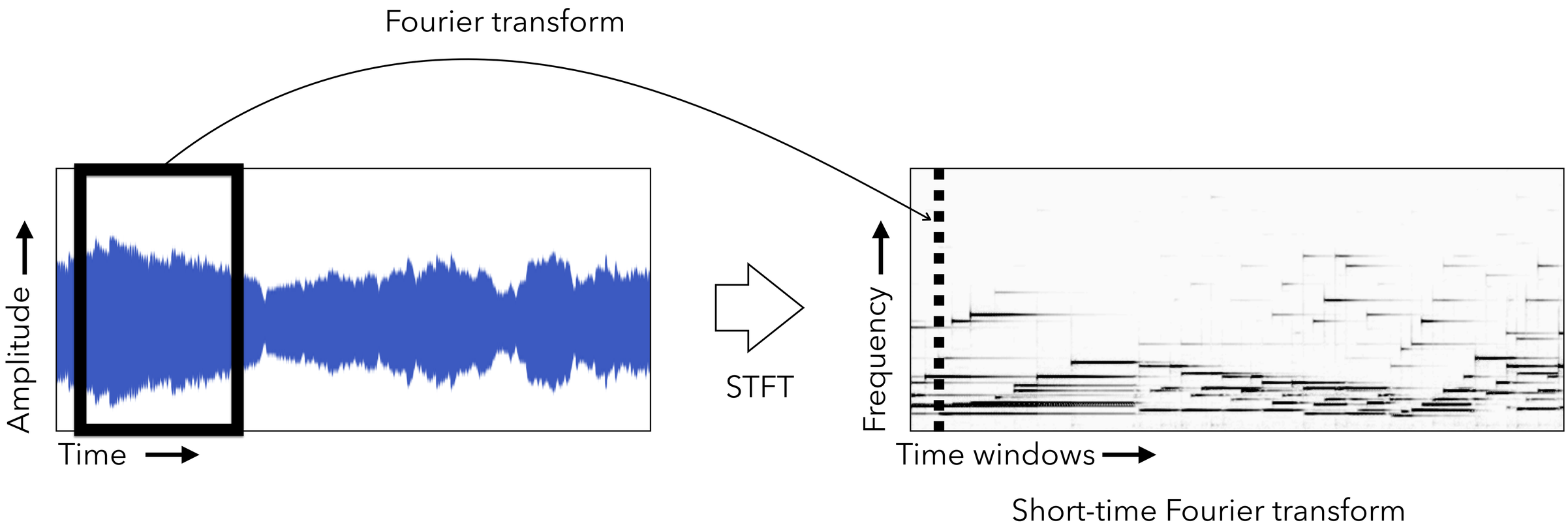
- GPT-style transformer model
- Use sparse attention matrix to deal with long sequence length.

Source: [van den Oord et al., “WaveNet”, 2016]

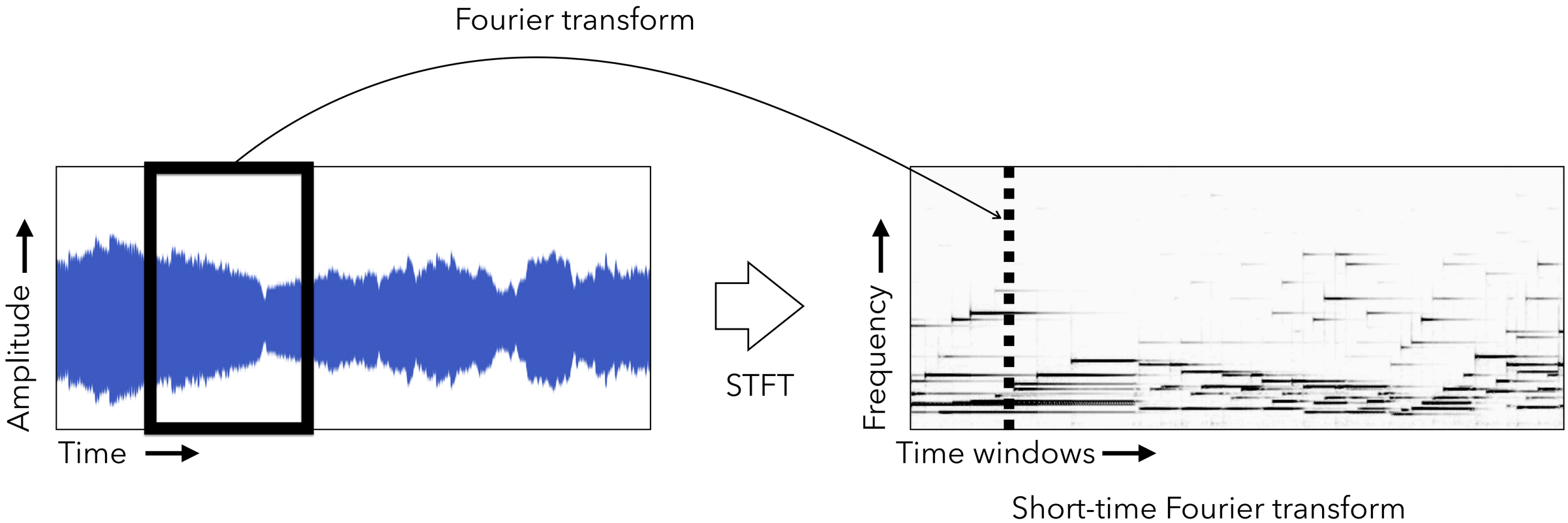
Making the space easier to work with

- In language models, we didn’t generate character by character
  - Instead, they use a tokenizers.
- Can we do something analogous for audio?
  - Compress the space into

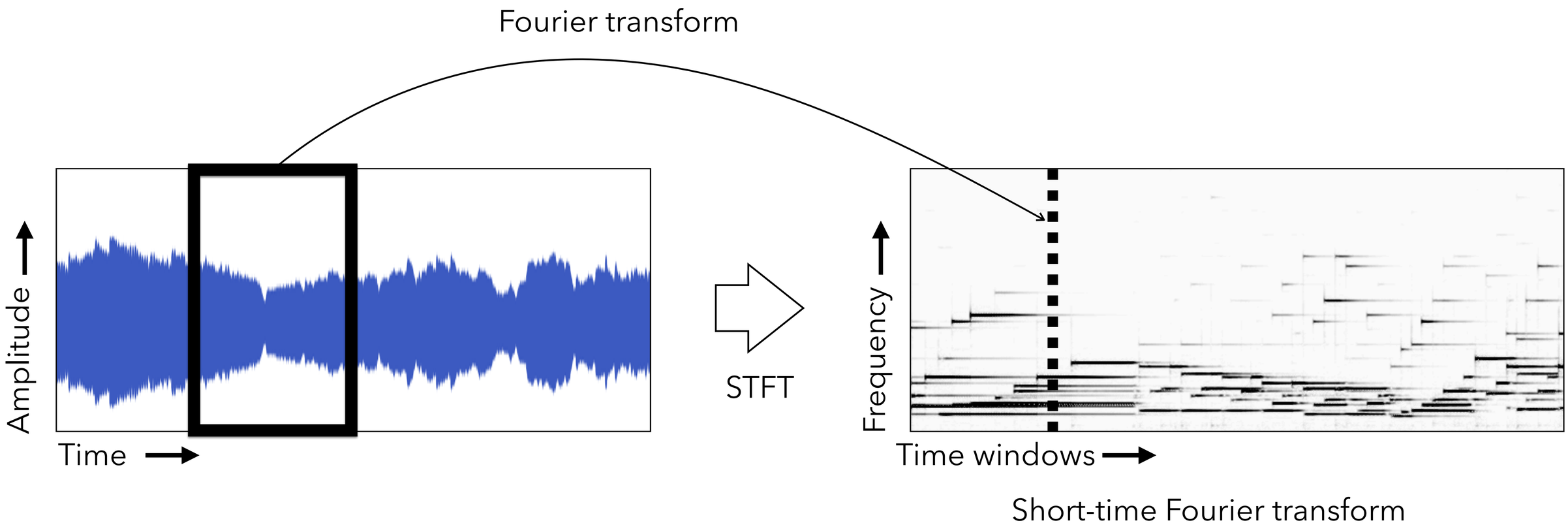
Recall: Spectrogram



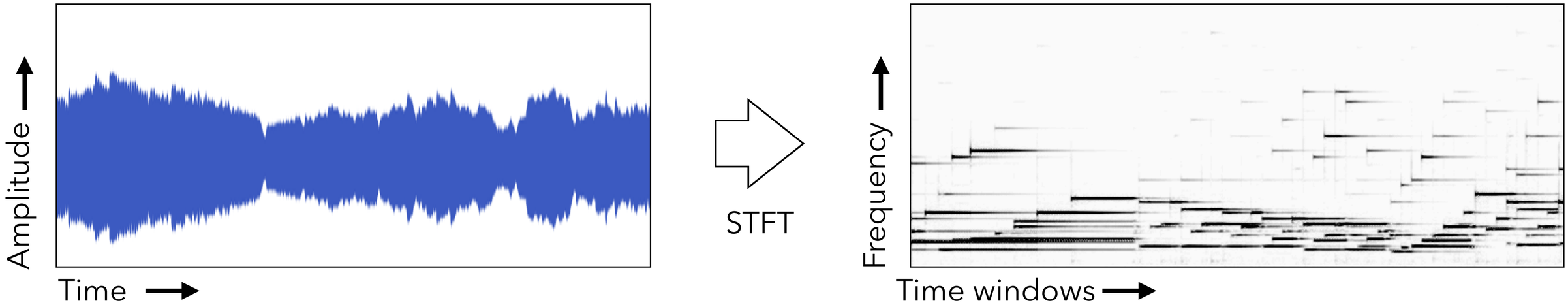
Spectrogram



Spectrogram



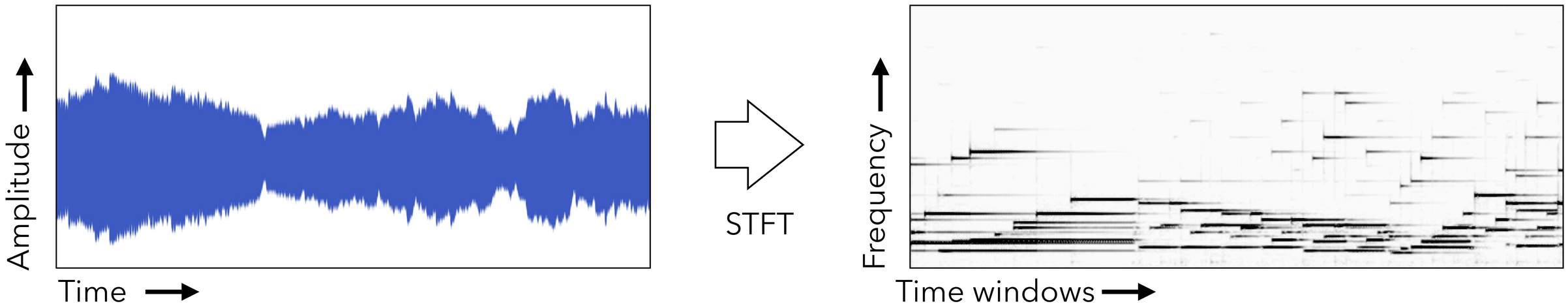
Spectrogram



Short-time Fourier transform

Spectrogram

Can treat it like an image and process with a CNN or ViT!



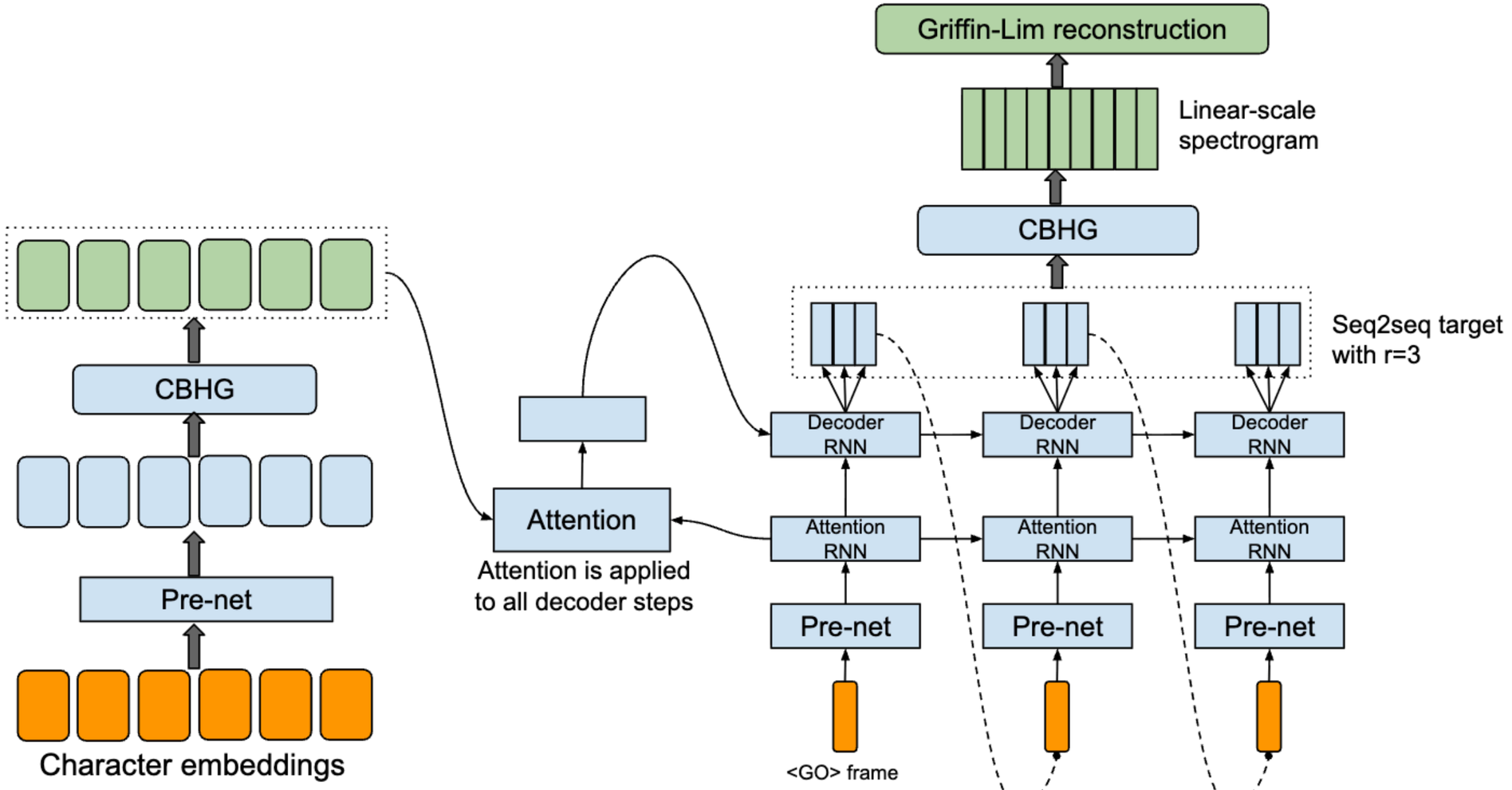
Short-time Fourier transform

Usually lower dimensional and has a 2D structure that is well-suited to image architectures.

From spectrogram to waveform

- The spectrogram discards phase and keeps only magnitude.
- Need to recover phase before you can invert it to a waveform.
- Classic approach from signal processing (Griffin-Lim algorithm):
  - Initialize phase randomly. Repeatedly encode/decode until convergence. Each iteration, force the spectrogram's magnitude to match the input.

Text-to-speech (circa 2017)



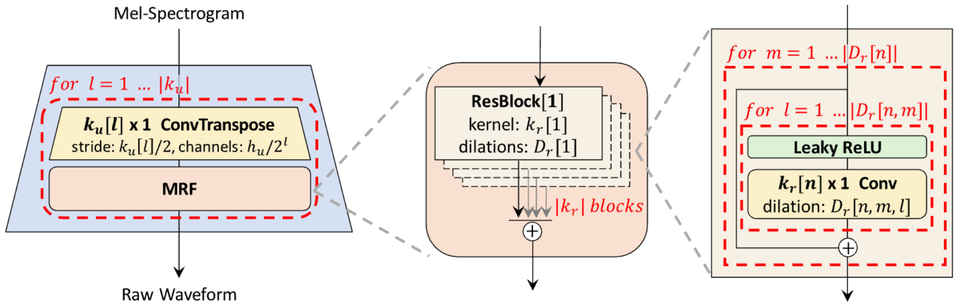
[Wang et al., “Tacotron", 2017]

From spectrogram to waveform

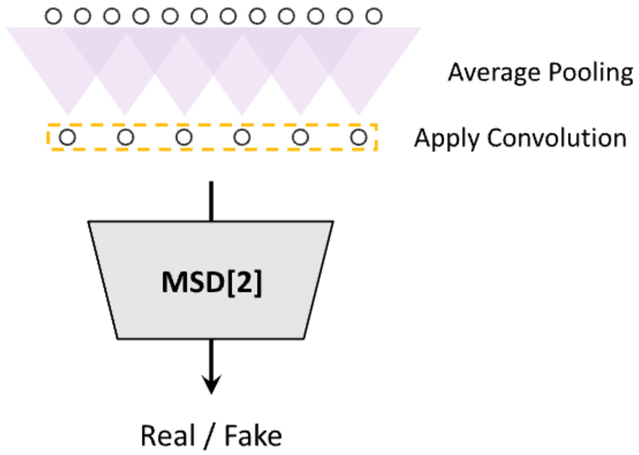
- The spectrogram discards phase and keeps only magnitude.
- Need to recover phase before you can invert it to a waveform.
- Classic approach from signal processing (Griffin-Lim algorithm):
  - Initialize phase randomly. Repeatedly encode/decode until convergence. Each iteration, force the spectrogram's magnitude to match the input.
- Or just train a network to do it!

Neural vocoder

Generator:



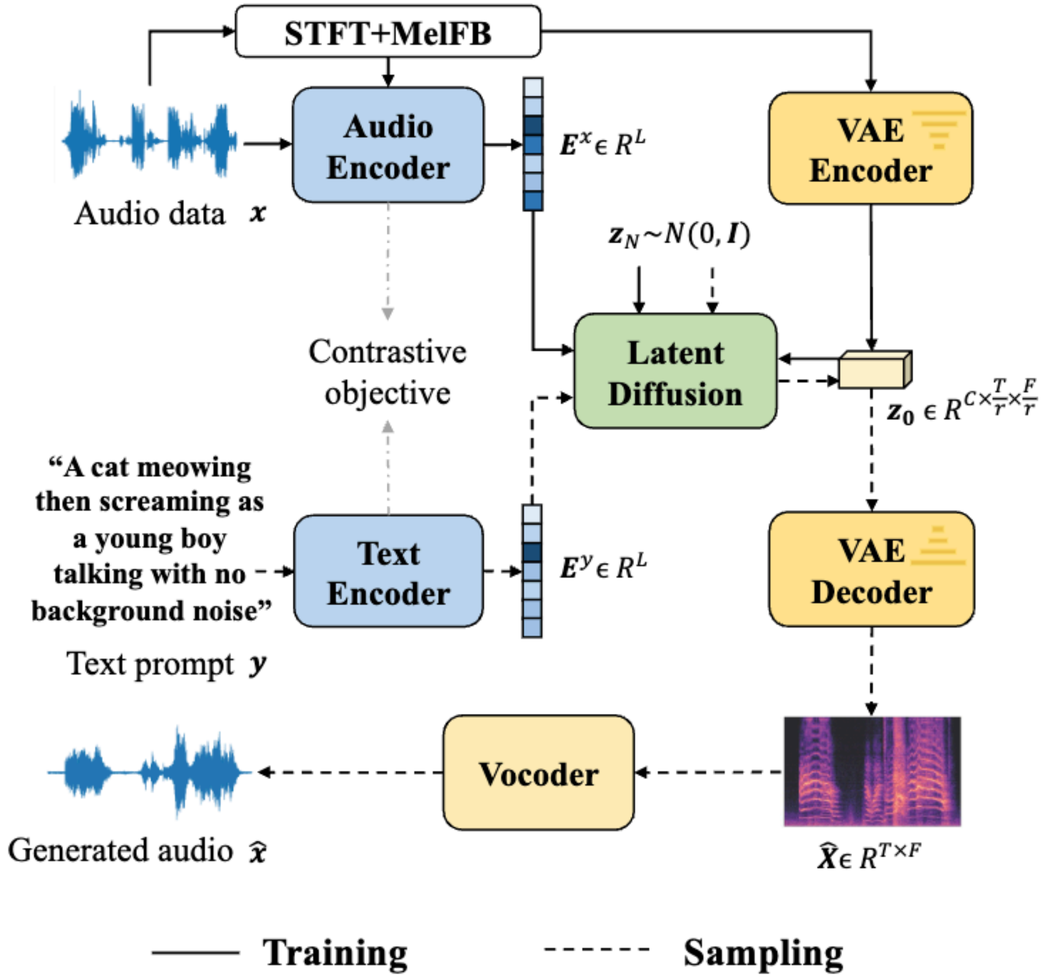
Discriminator:



- Train a separate network that maps a spectrogram to a waveform.
- For example, this could be an autoregressive model like WaveNet
- Can also be implemented more efficiently as a GAN

[Kong et al., “HiFi-GAN”, 2020]

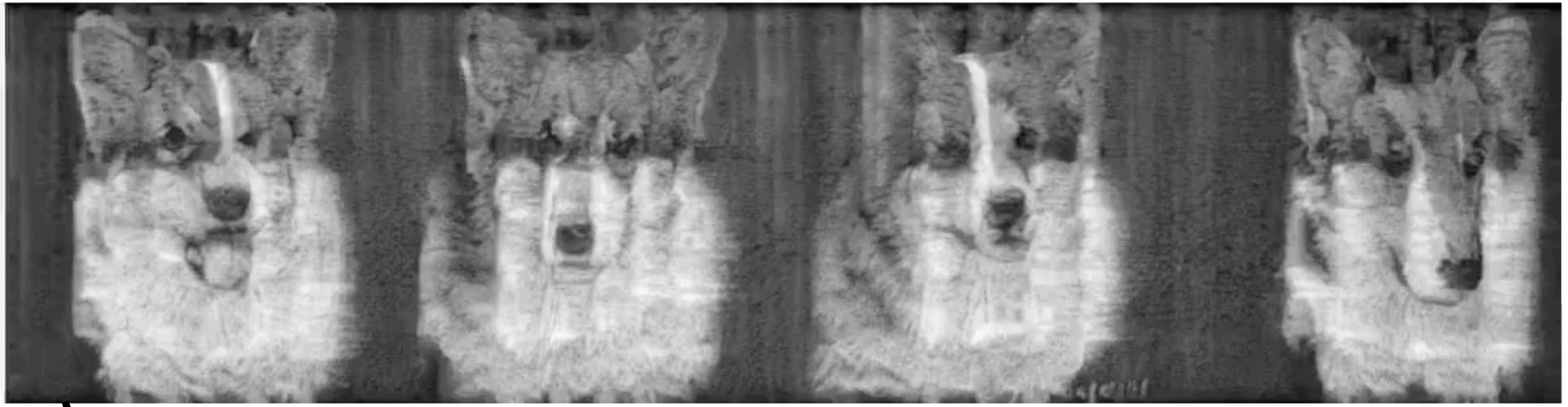
Latent diffusion on spectrograms for audio generation



- Generate audio using diffusion.
- Use latent code created from a VAE on spectrograms.
- Decode the generated latent features to get a spectrogram. Then use a neural vocoder to map the spectrogram to a waveform.
  - Are both of these intermediate representations necessary? (We'll return to this).
- What if you use an image VAE as the latents?
  - Works to some extent! [Forsgren et al., "Riffusion", 2022], [Xue et al., "Auffusion", 2024]

[Liu et al., "AudioLDM", 2023]

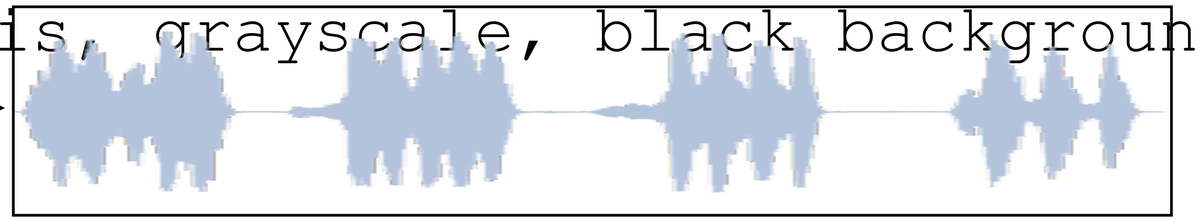
Looking at spectrograms



“a p orgis, grayscale, black background”

“dog

Spectrogram to audio

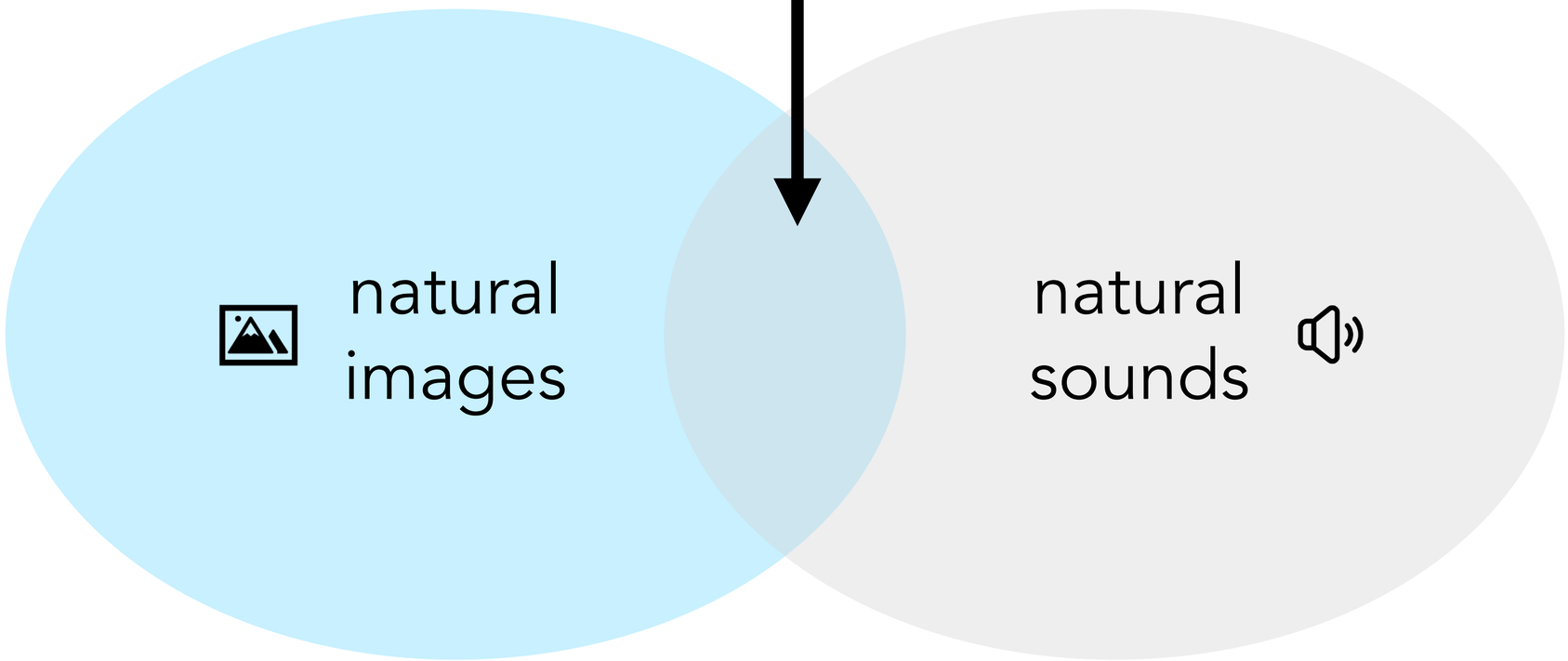


[Chen, Geng, Owens. “Images that Sound: Composing Images and Sounds on a Single Canvas.” NeurIPS, 2024]

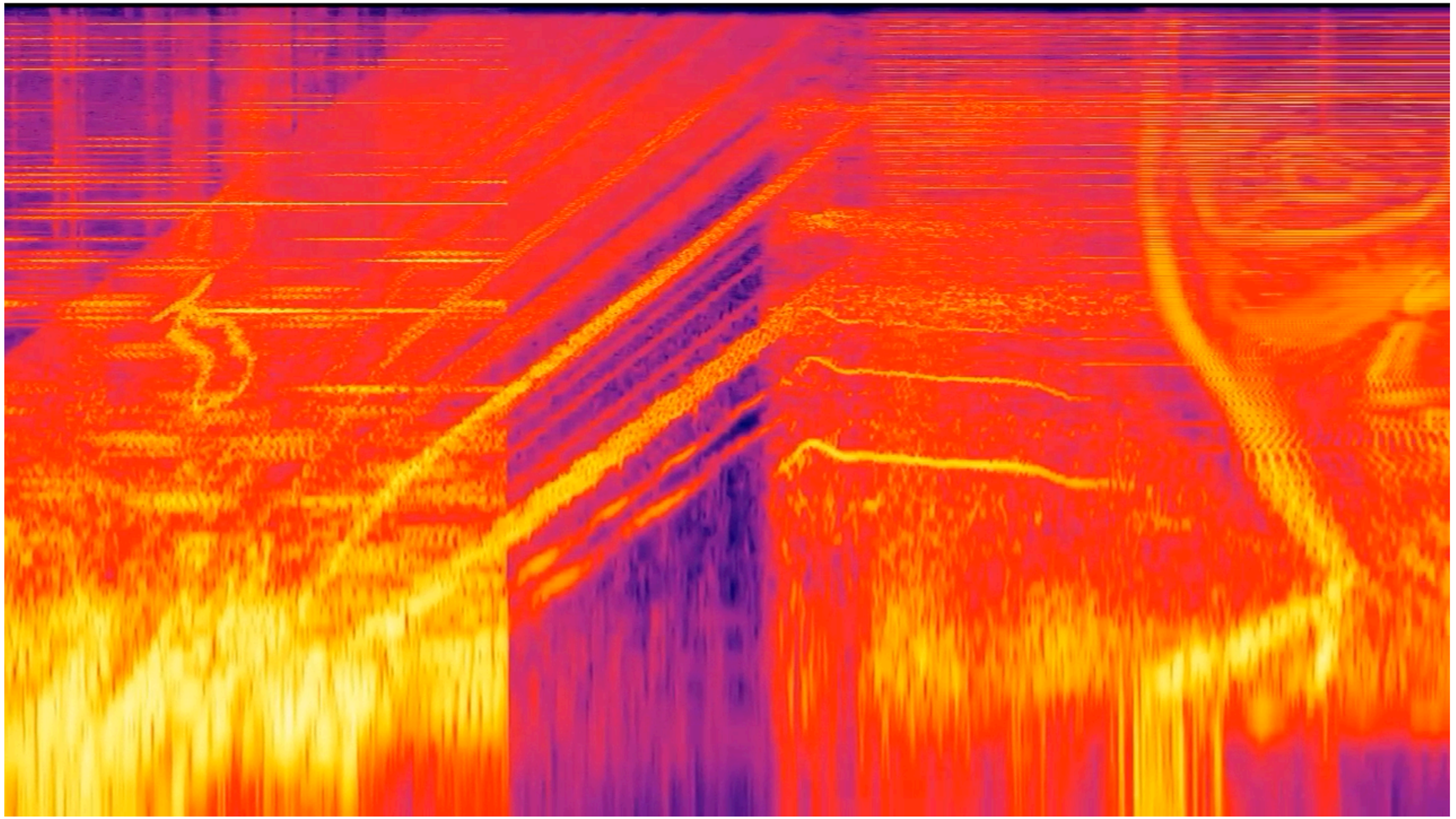
Ziyang Chen

Daniel Geng

Can we find these?

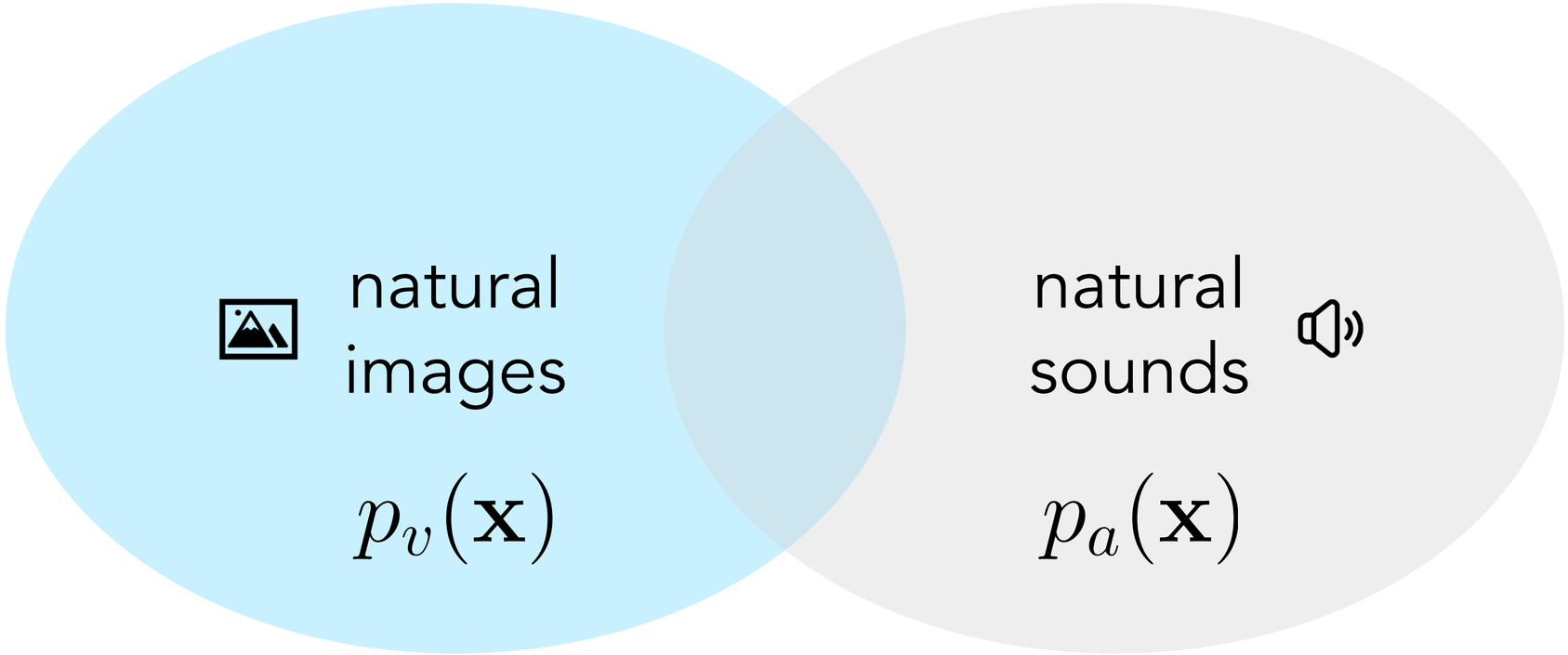


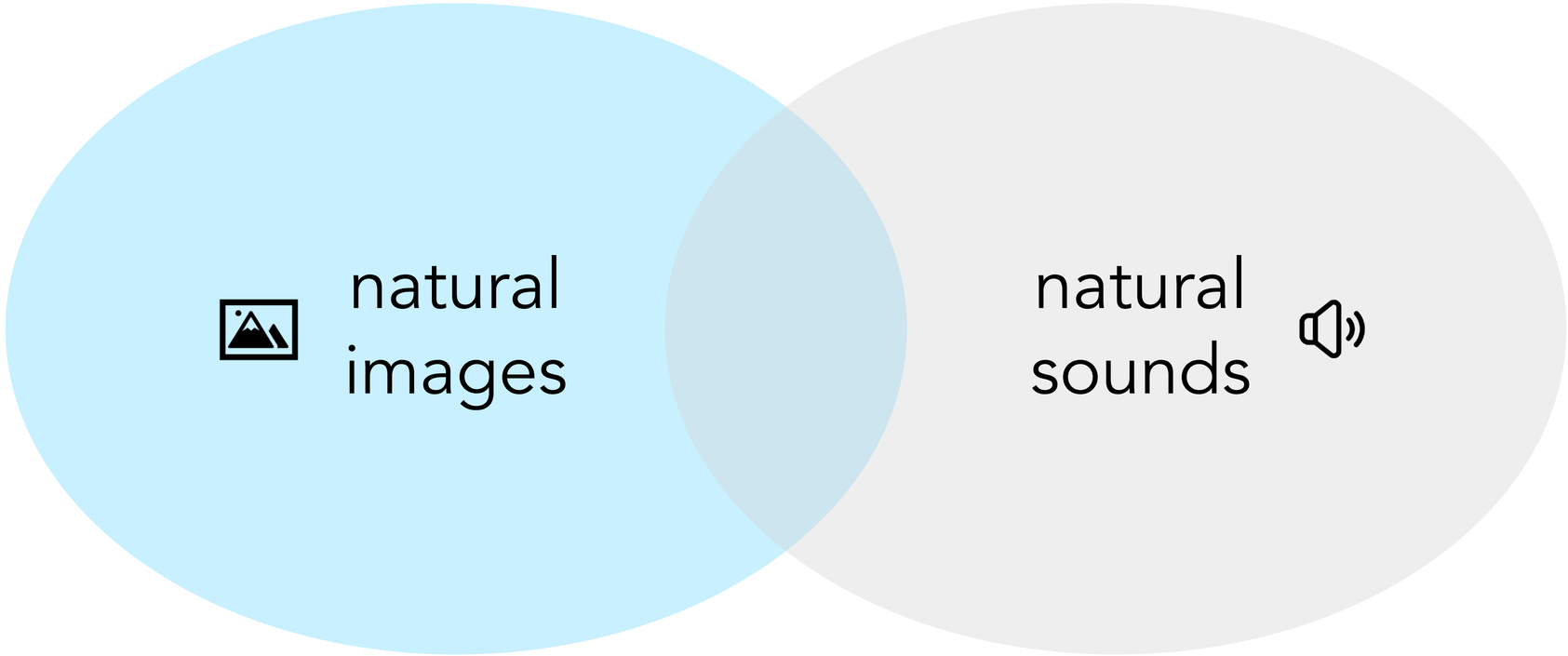
Related work: spectrogram art



Aphex Twin, *Formula, 2001*

See also: [Nine Inch Nails, “My Violent Heart”, 2007], [Venetian Snares, “Songs about my Cats”, 2006]


natural
images
$p_v(\mathbf{x})$
natural
sounds
$p_a(\mathbf{x})$

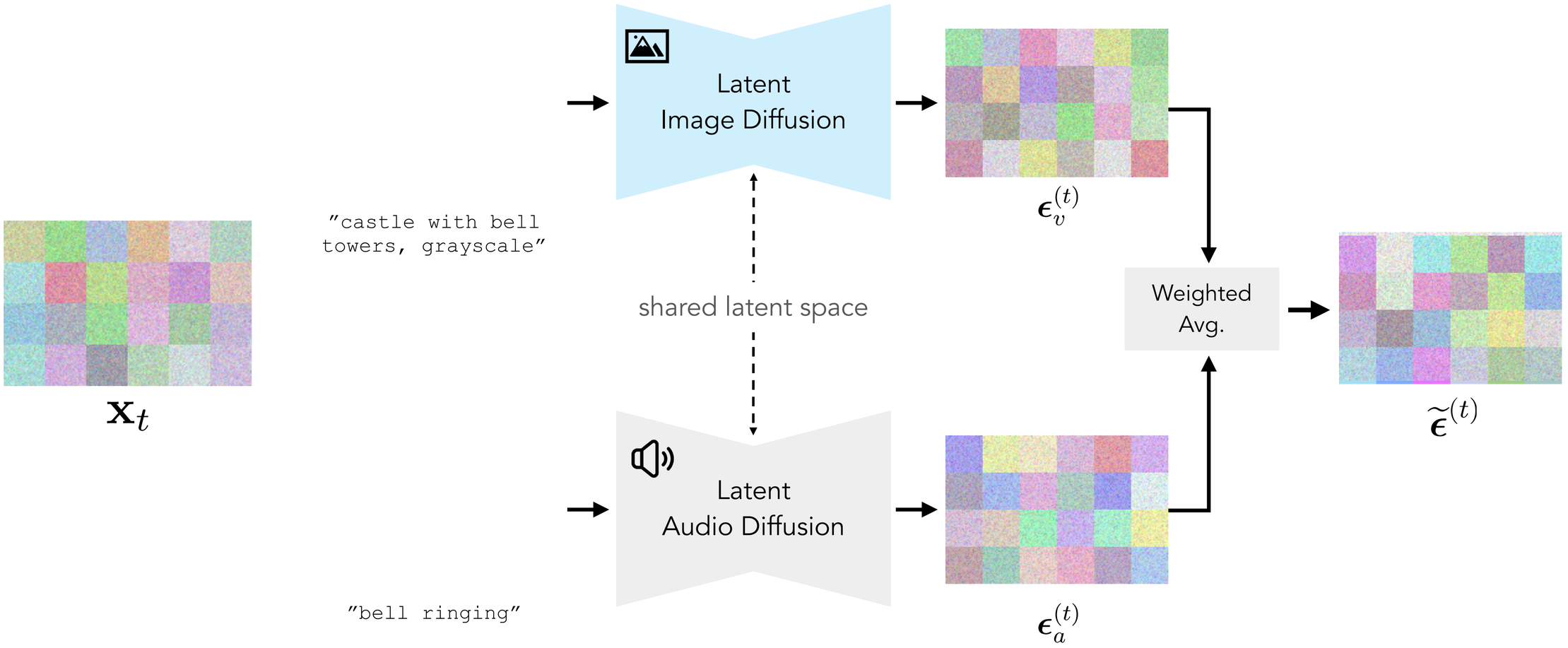


Sample from product of experts:

$$p_{av}(\mathbf{x}) \propto p_v(\mathbf{x}) p_a(\mathbf{x})$$

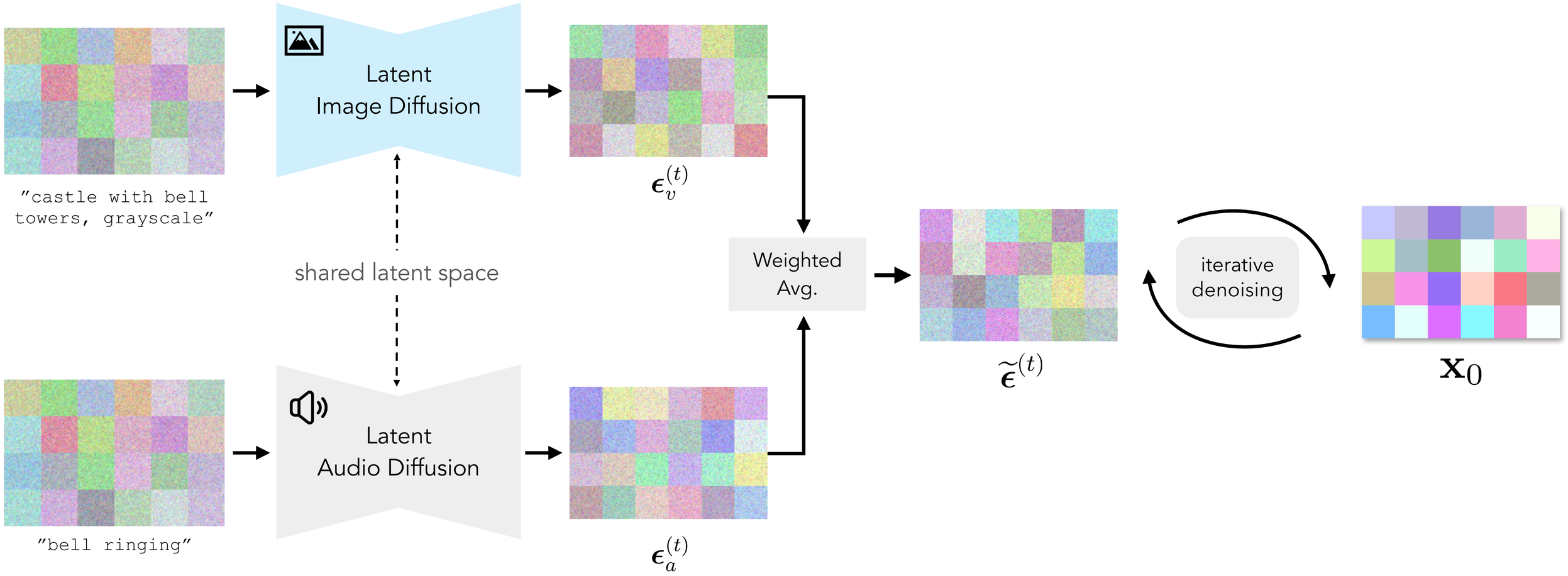
Related work on compositional generation: e.g., [Liu, Li, Du, et al. “Compositional Visual Generation.” ECCV 2022]

Composing sight with sound

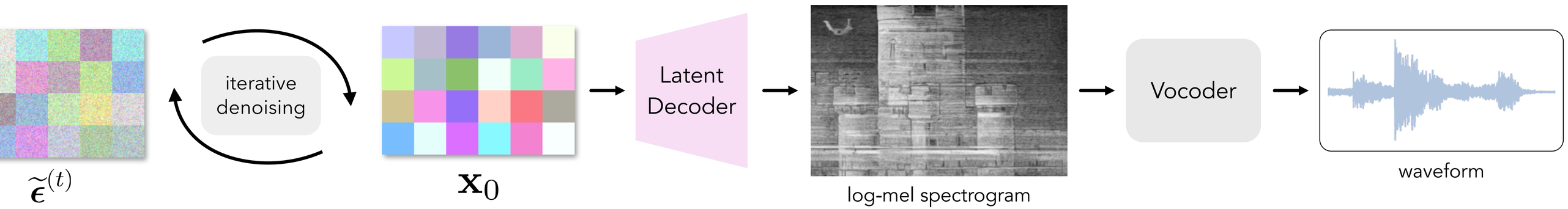


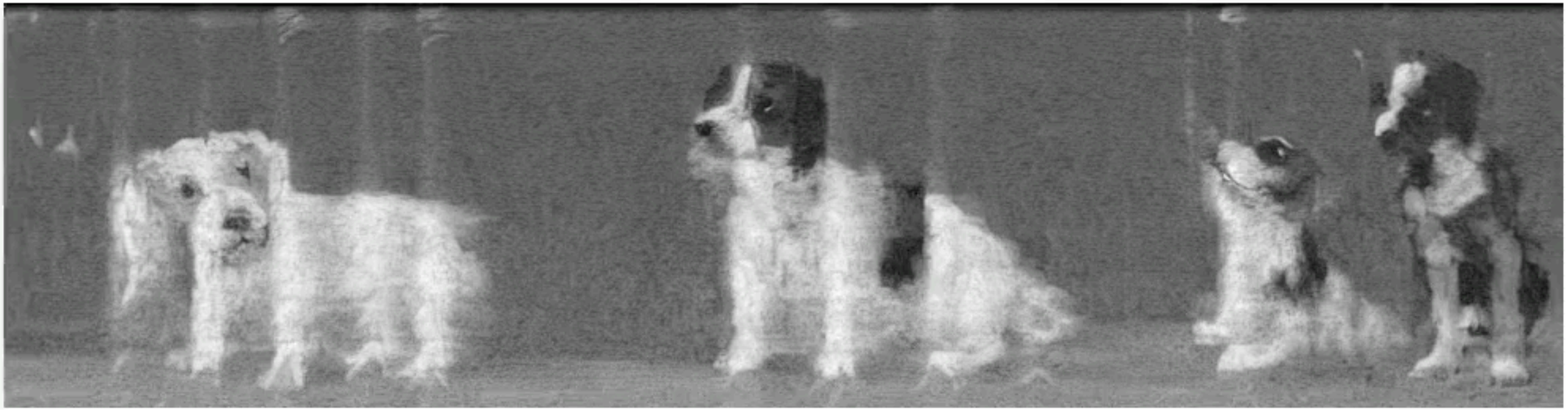
We use Auffusion [Xue et al., 2024], see also [Forsgren et al., "Riffusion", 2022]

Composing sight with sound



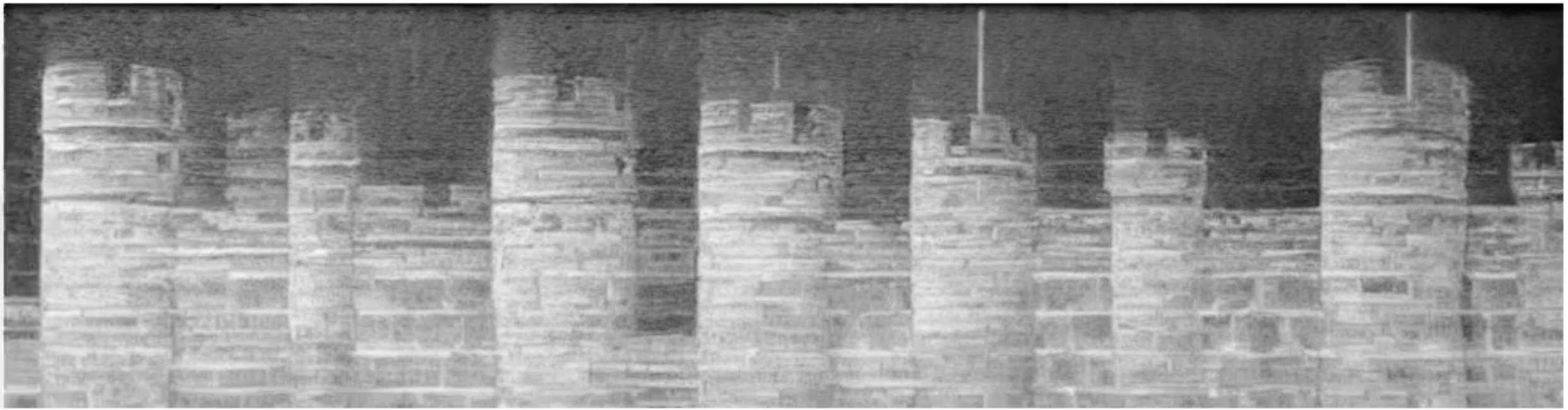
Composing sight with sound





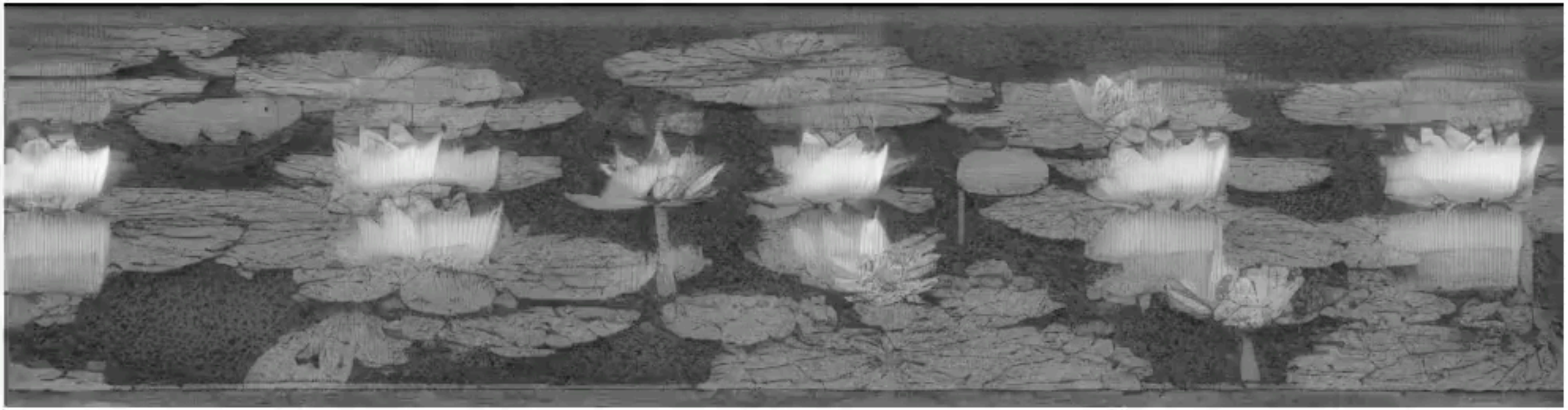
“a painting of cute dogs, grayscale”

“dog barking”



“a painting of castle towers, grayscale”

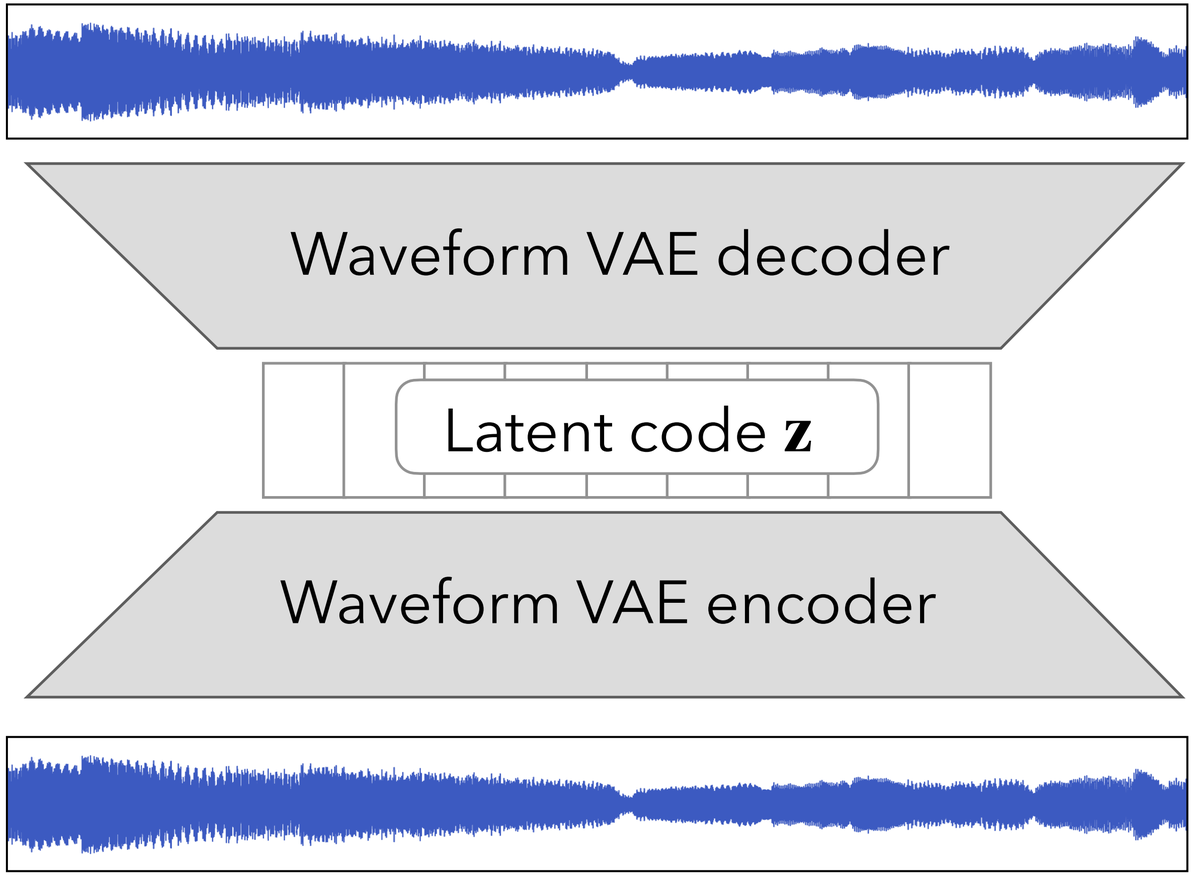
“bell ringing”



“frog croaking”

“a pond full of water lilies, grayscale, lithograph style”

Computing latent codes directly from waveforms



- Directly encode/decode the waveform, without spectrogram as intermediary.
- Use an adversarial loss and quantization along lines of VQ-VAE.
- Convert to frequency representation within *loss* function to define similarity.

[Kumar et al., “RVQGAN”, 2023], [Evans et al., “Stable Audio Open”, 2024]

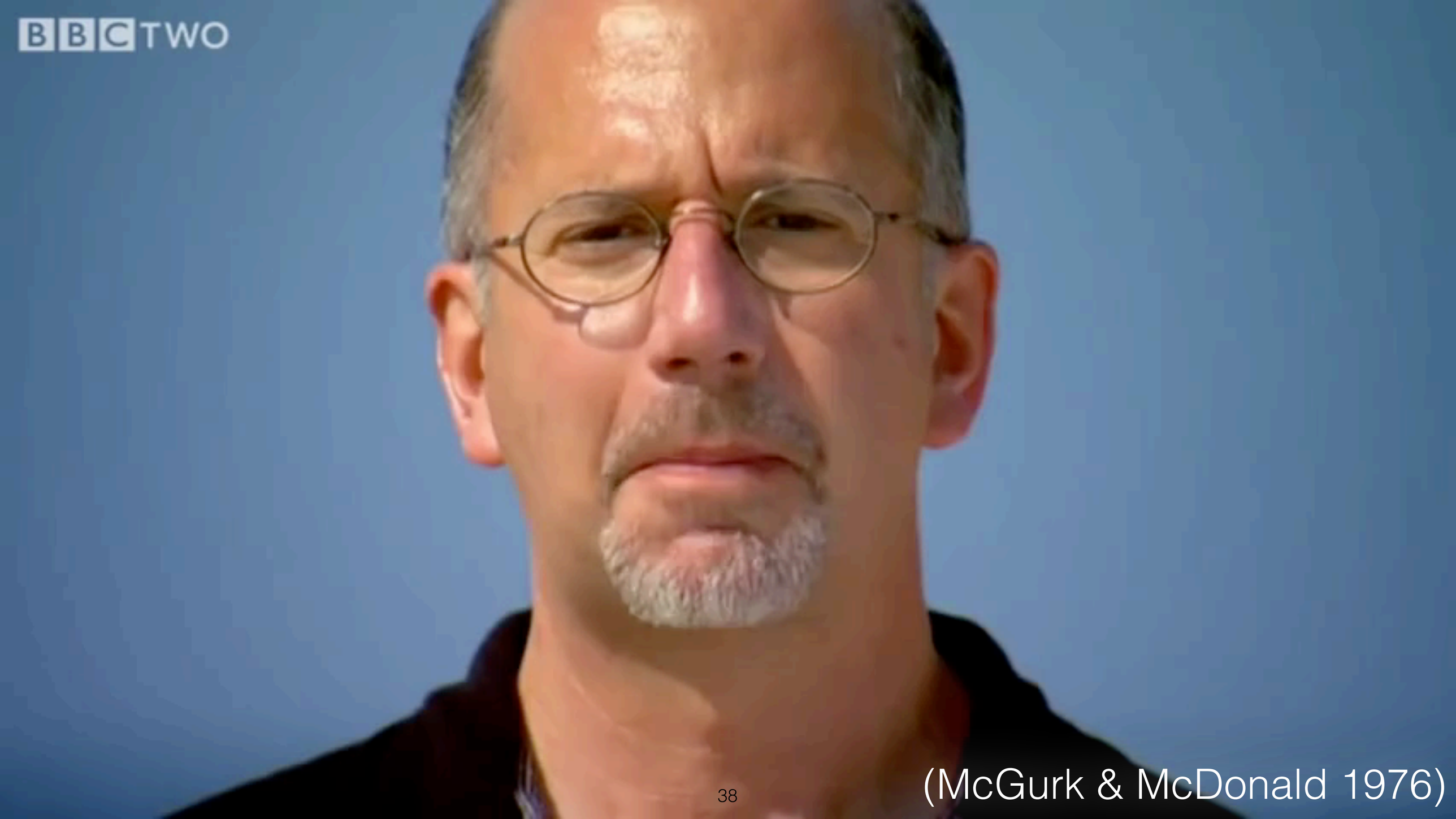
Music generation with a latent waveform diffusion

“A luxurious Indietronica instrumental perfect for a perfume advertisement featuring clean guitars, synthesizers, and a slow-tempo drum machine pattern.”

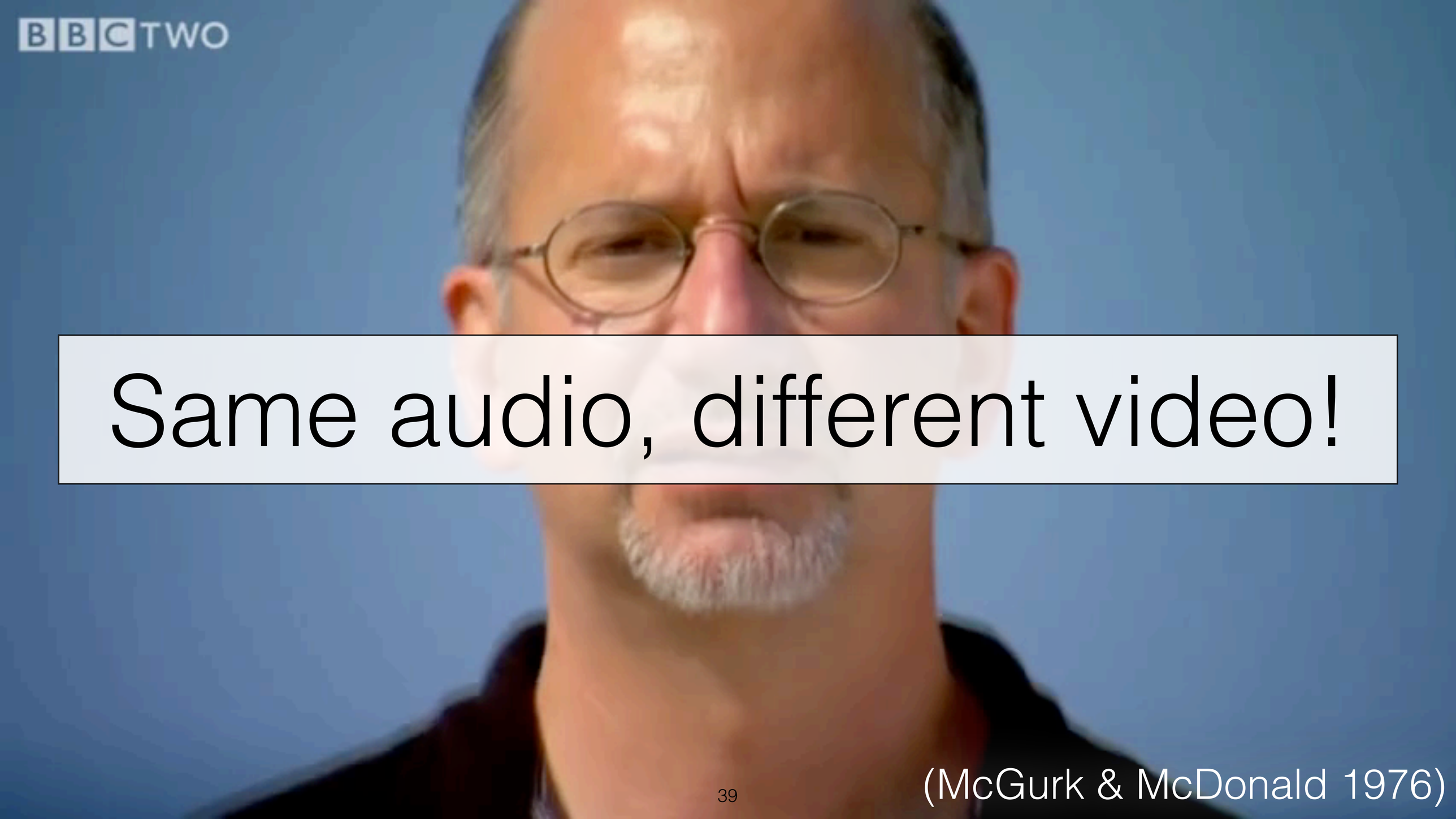
“90s garage rock instrumental with a grunge influence featuring poppy distorted guitars frantic and energetic drums and tube-distorted bass guitar”

Source: [Evans et al., “Stable Audio Open”, 2024]

Other conditioning signals?



(McGurk & McDonald 1976)



Same audio, different video!

(McGurk & McDonald 1976)

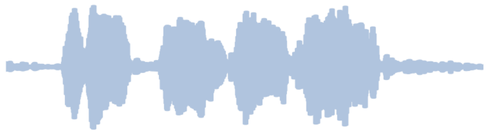
(McGurk & McDonald 1976)

Movie sound effects?

Video source: Jordan et al., https://www.youtube.com/watch?v=WFVLWo5B81w

Multimodal control for video-to-audio generation

Input video



Input audio

Video Encoder

re-sample

Audio Encoder

masked diffuse

Diffusion Transformer

$\mathcal{L}_{\text{LDM}}$

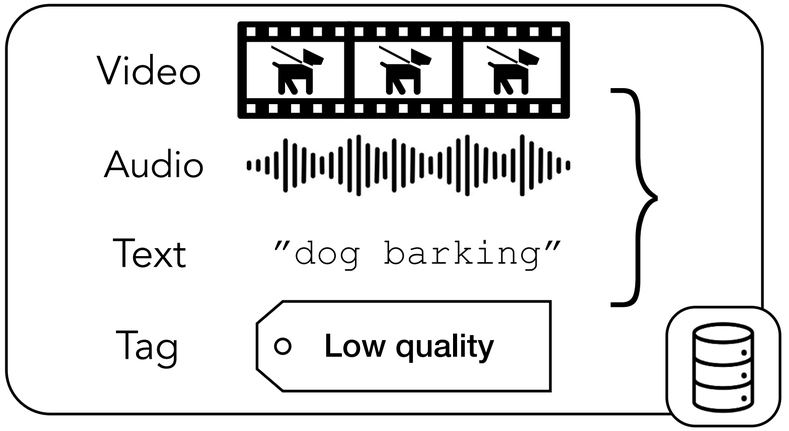
GT latents

[Chen, Seetharaman, Russell, Nieto, Bourgin, Owens, Salamon, "Video-Guided Foley with Multimodal Controls", CVPR 2025]

Ziyang Chen

Prem Seetharaman

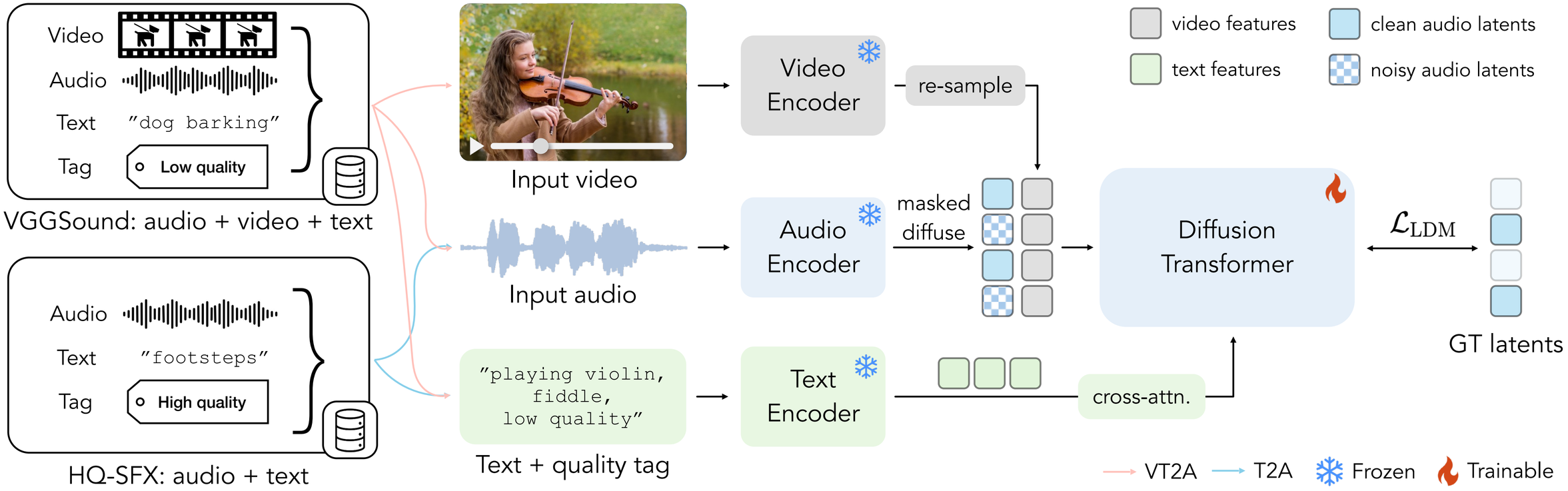
Training with two data sources



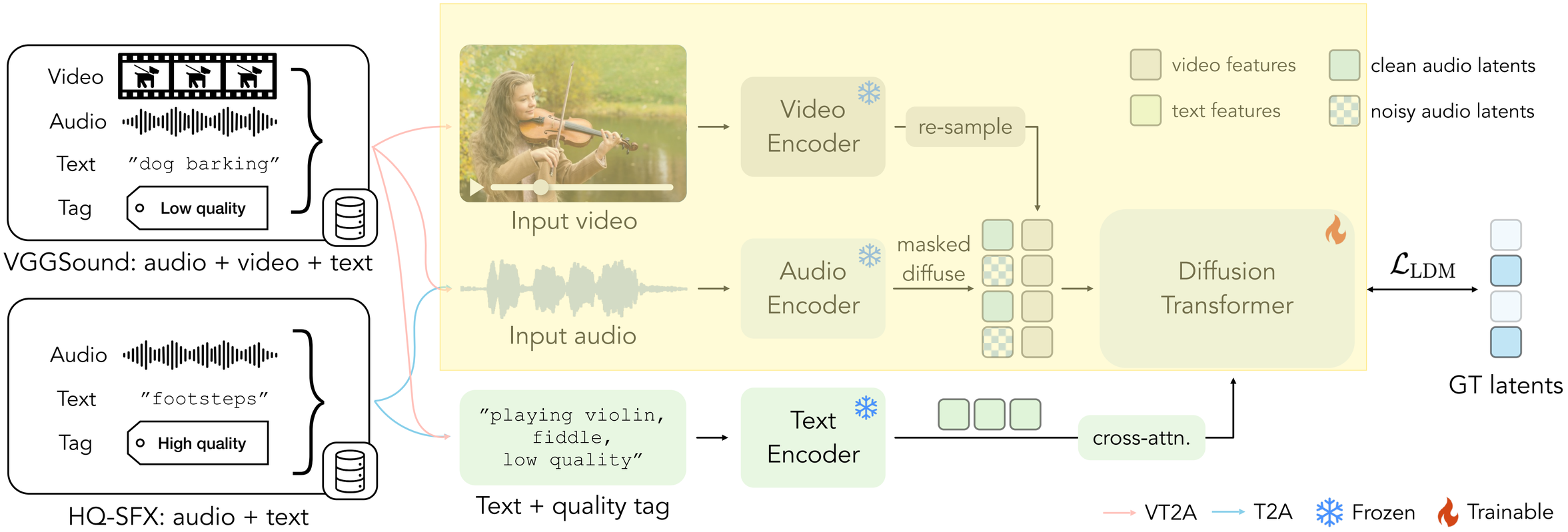
VGGSound: audio + video + text

HQ-SFX: audio + text

Multimodal Conditional Foley Generation



Multimodal Conditional Foley Generation



Application: example-based synthesis

Given this video with partial soundtrack…

Application: example-based synthesis

Conditional sound

Generated sound for silent video

Application: example-based synthesis

Conditional sound

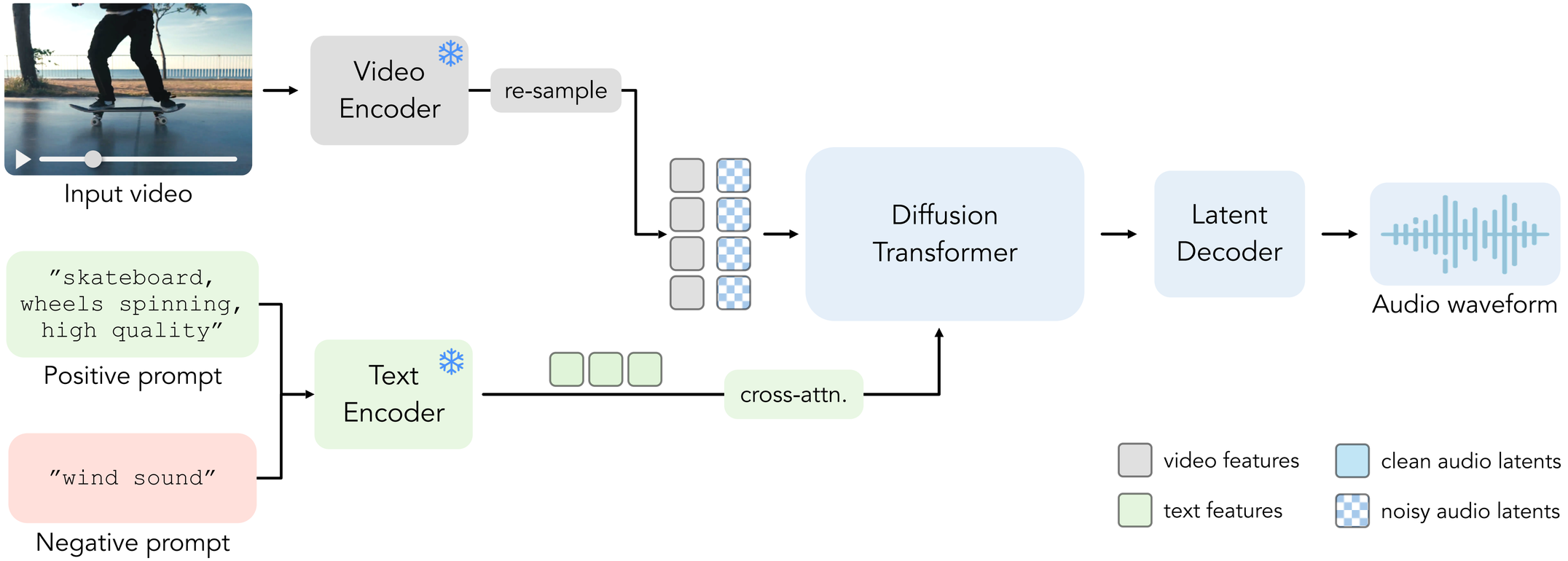
Generated sound for silent video

Application: example-based synthesis

Given this reference drum audio

We generate sound for this silent video

Foley Generation with Text Control



Adding language conditioning

“skateboard, wheels spinning, high quality”

Foley Generation with Text Control

Text prompt: `chopping wood, high quality`

Adding language conditioning

“typewriter”

Adding language conditioning

"typing on computer keyboard"

Adding language conditioning

“playing piano”

Foley Generation with Text Control

Text prompt: `cat meowing`

Foley Generation with Text Control

Text prompt: `lion roaring`

Foley Generation with Text Control

Text prompt: `playing cello`

Foley Generation with Text Control

Text prompt: `playing erhu`

Foley Generation with Text Control

Text prompt: `chainsawing trees`

Adding language conditioning

“bird chirping”

Adding language conditioning

“rooster crowing”

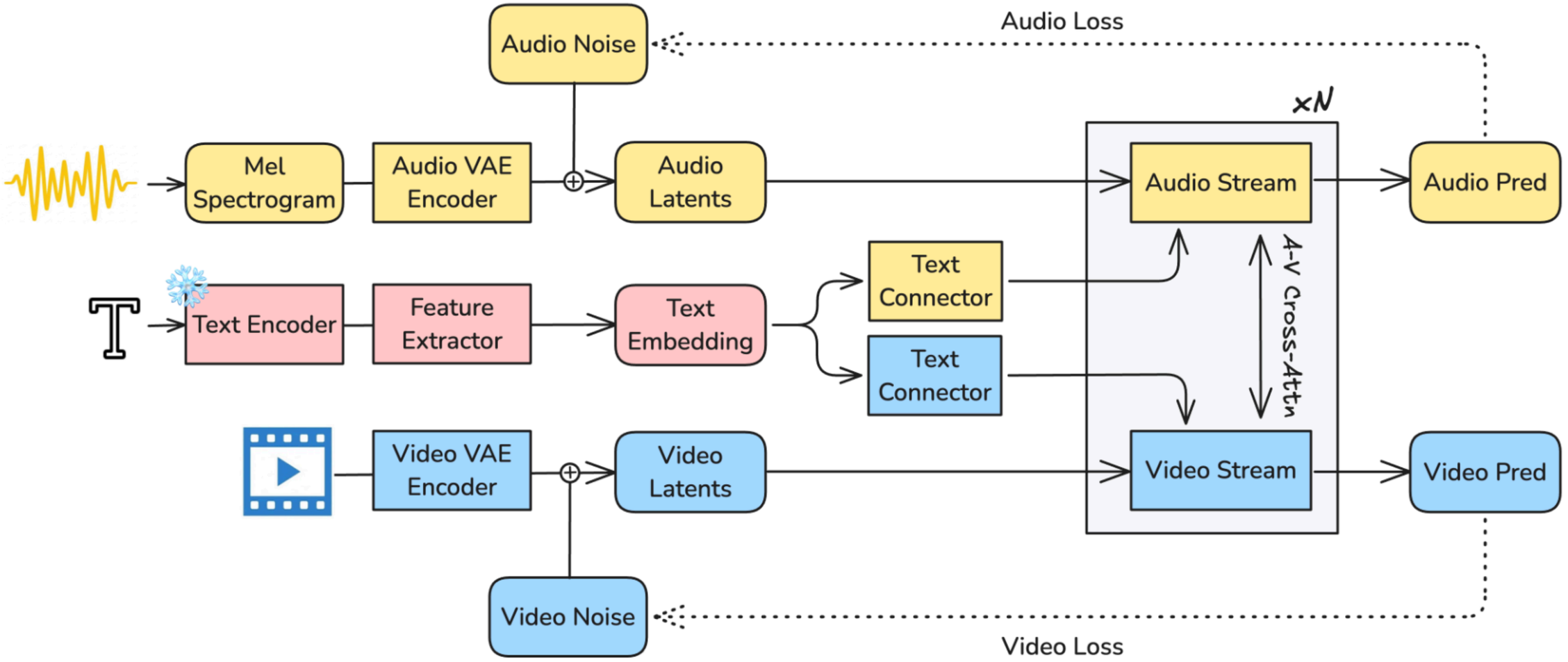
Adding language conditioning

“male speaking”

Joint audio and visual generation (Veo 3)

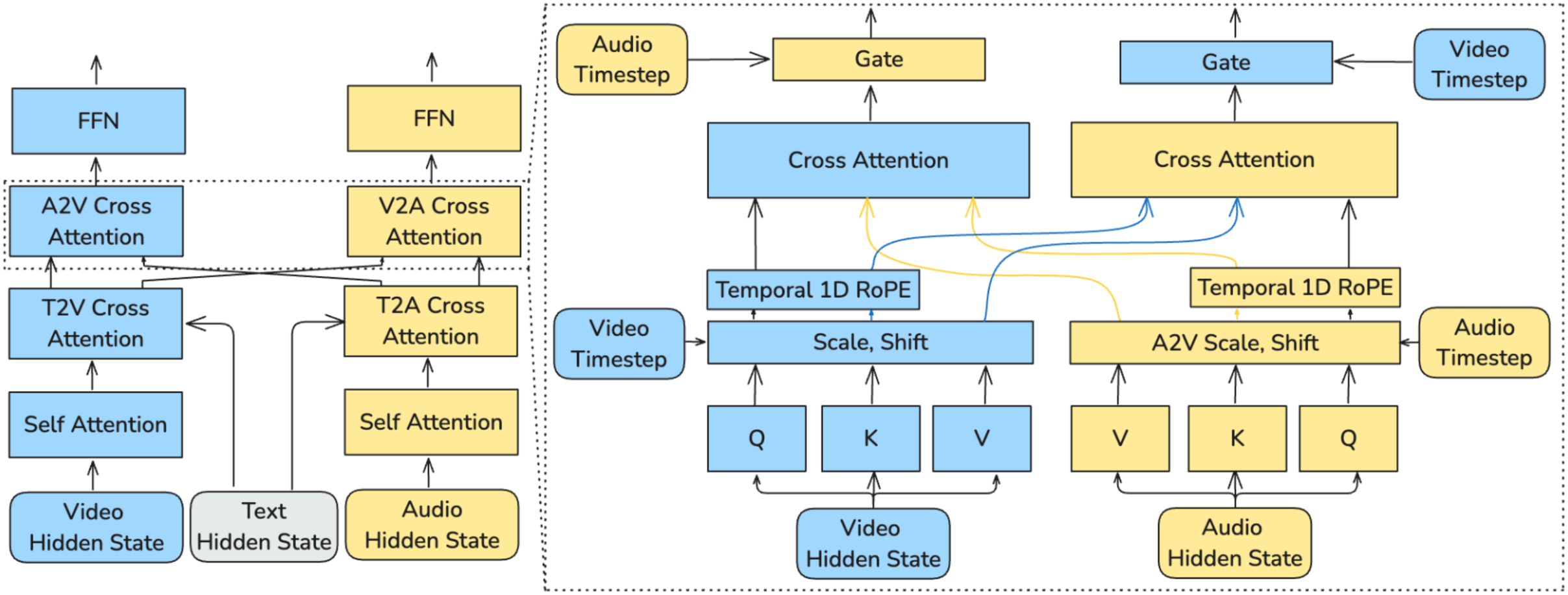
Joint audio and visual generation (Veo 3)

Joint audio-visual generation via diffusion



[HaCohen et al., "LTX-2", 2026]

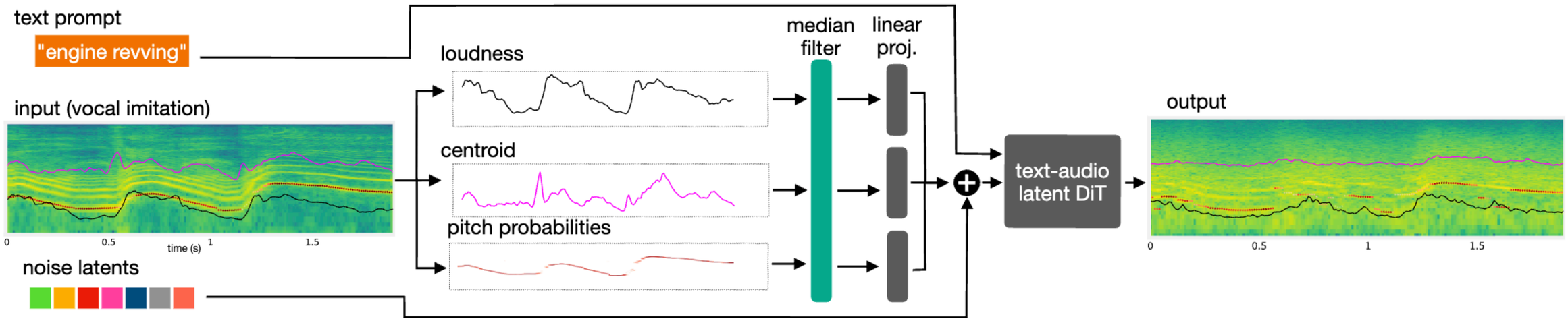
Joint audio-visual generation via diffusion



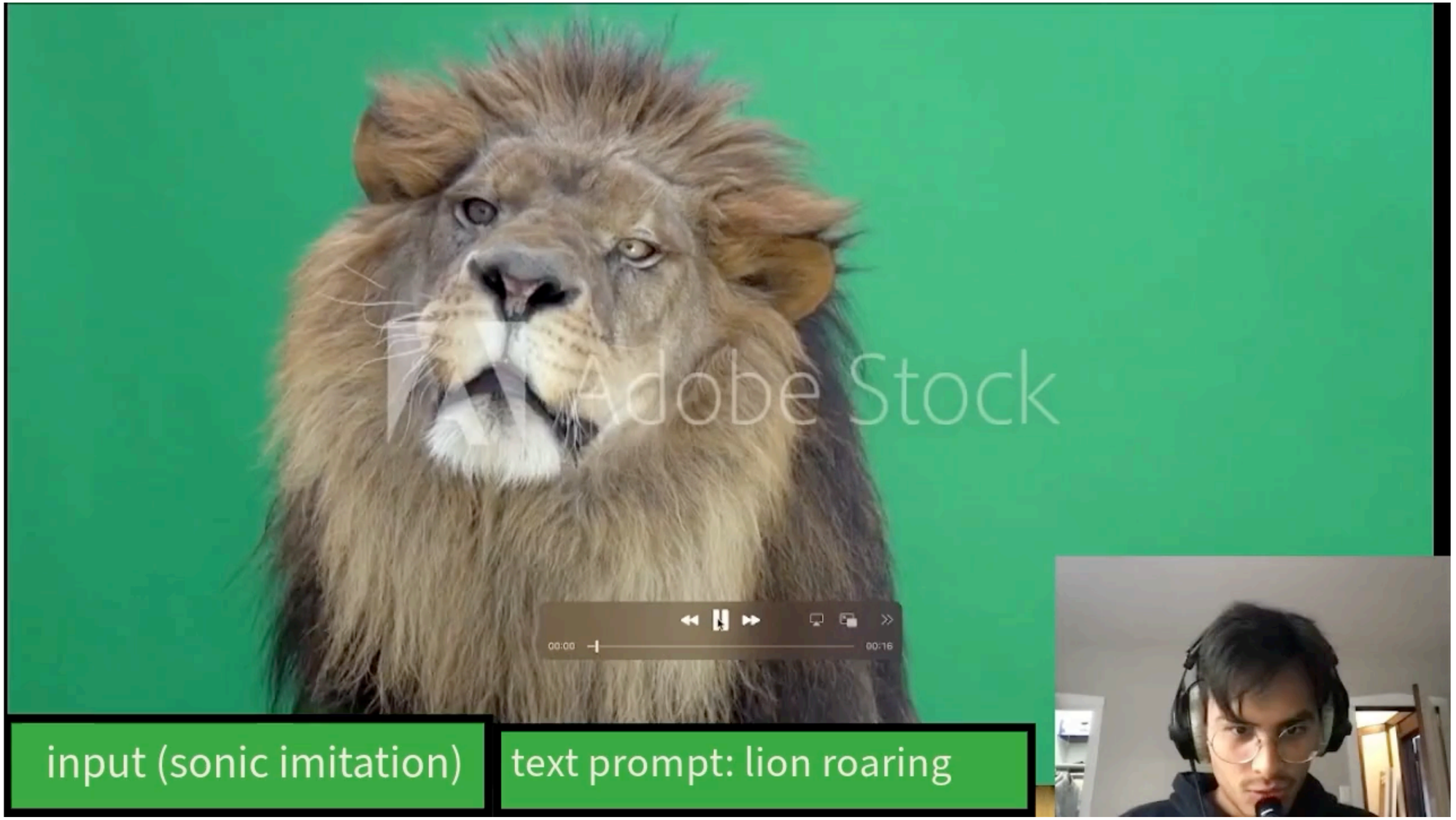
[HaCohen et al., "LTX-2", 2026]

Other cross-modal conditioning?

Conditioning on time-varying sound properties

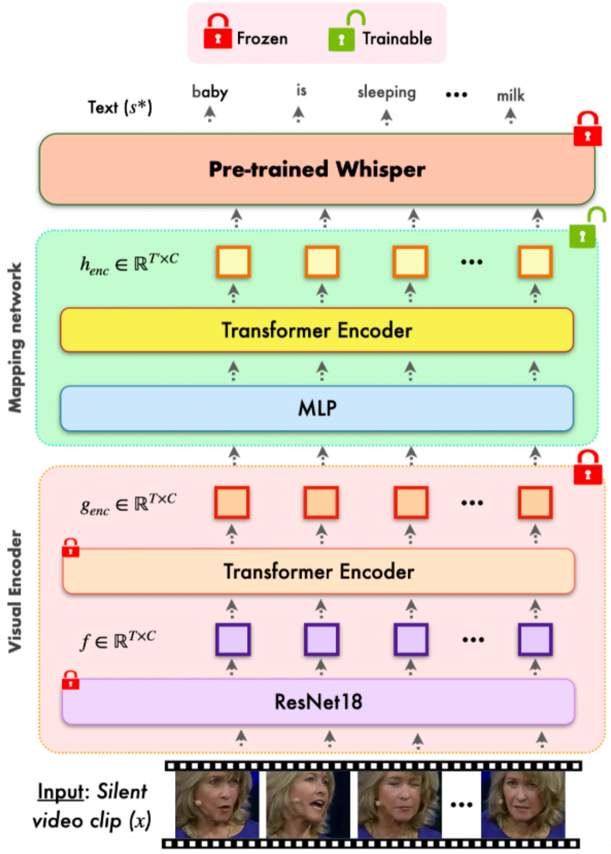


[García et al., “Sketch2Sound”, 2025]



[García et al., “Sketch2Sound”, 2025]

Lip-reading: video (and/or audio) to text



Enabling people who have lost their voice 🗣️

Lip-reading output:

Transcribing very old recordings (1960s) 📽️

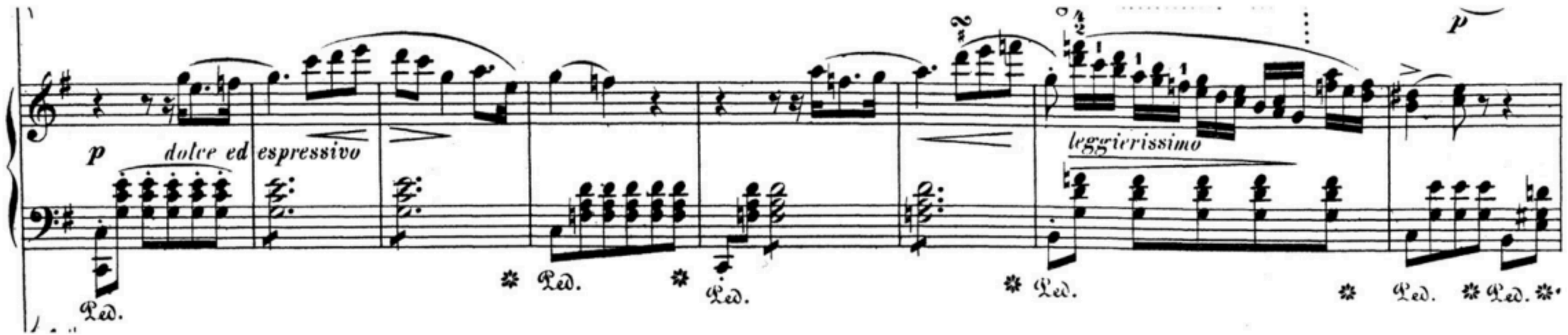
we've had this place this is the third time in my memory

(audio only played here to verify the output with the actual speech)

Source: [Prajwal et al., 2024]

Other audio represnetations

Symbolic representations of sound

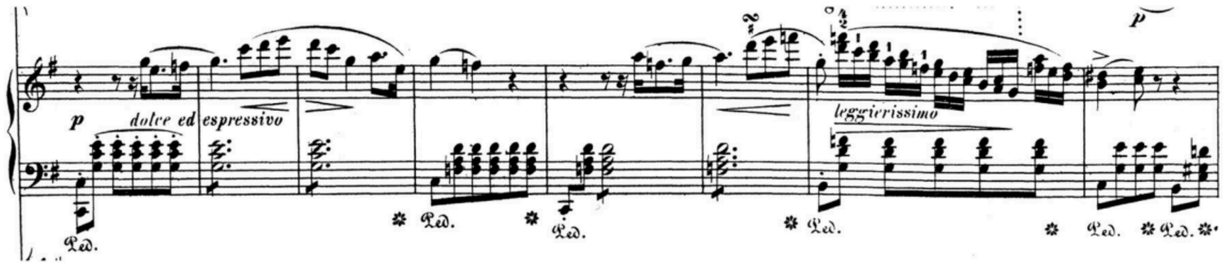


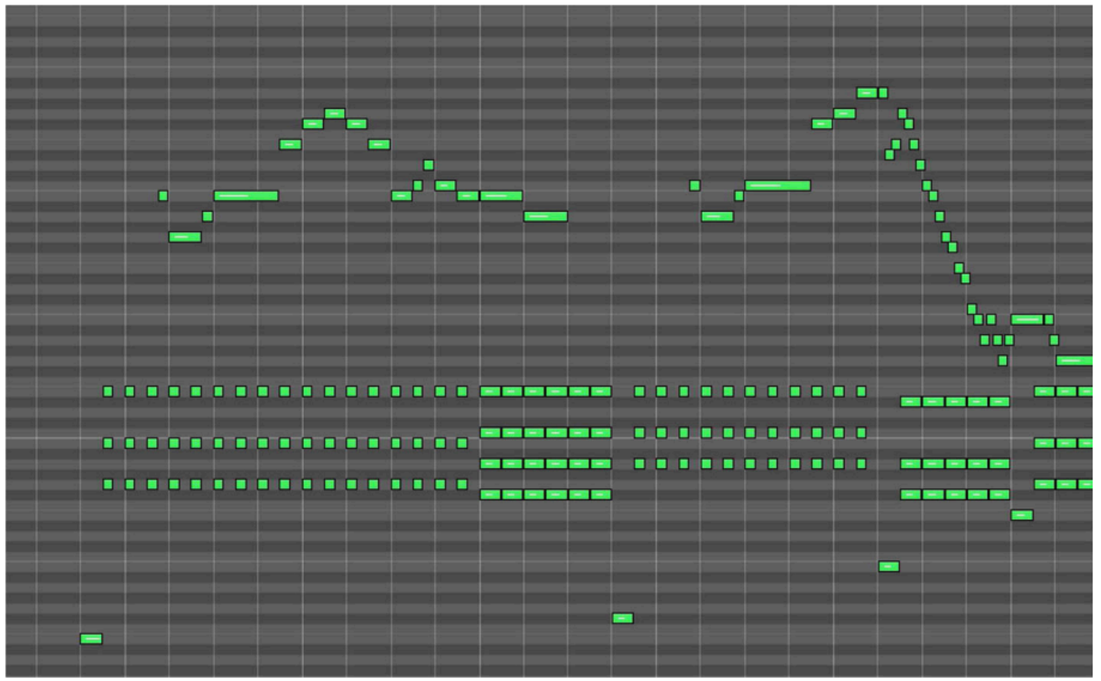
Excerpt from the score of Chopin's Piano Concerto No. 1

- A concise representation of music.
- Human editable and interpretable.
- Not one-to-one with sound. Artists vary timing, dynamics, etc.

Source: [Oore et al., 2020]

Symbolic representations of sound



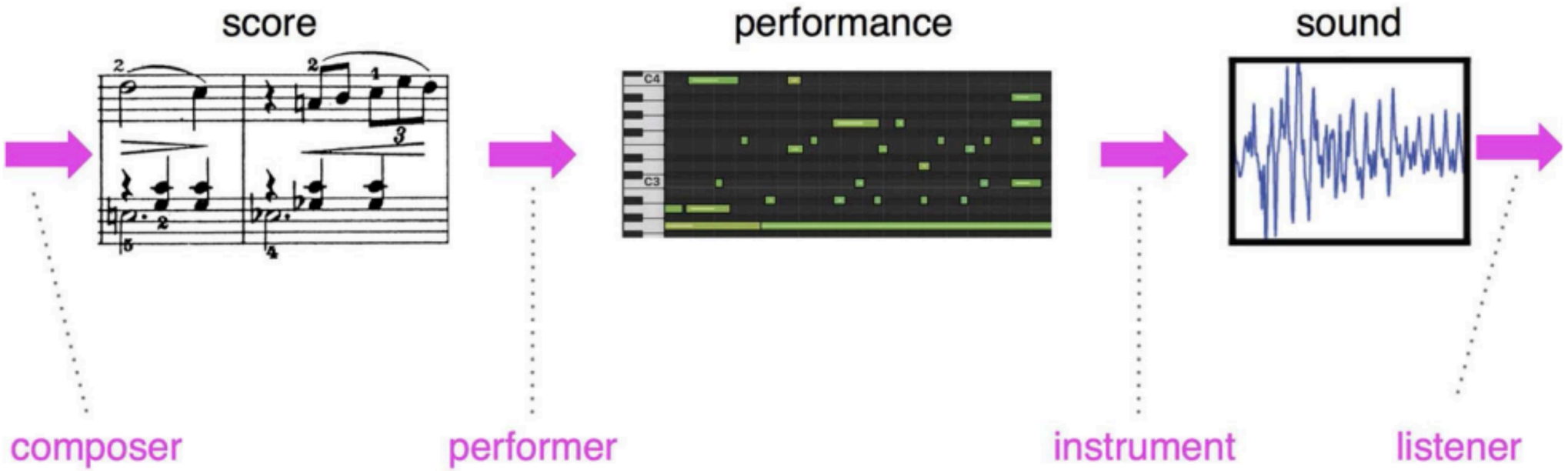


MIDI (Musical Instrument Digital Interface)

e.g., 128 possible notes, each on or off, at 125 Hz.

Source: [Oore et al., 2020]

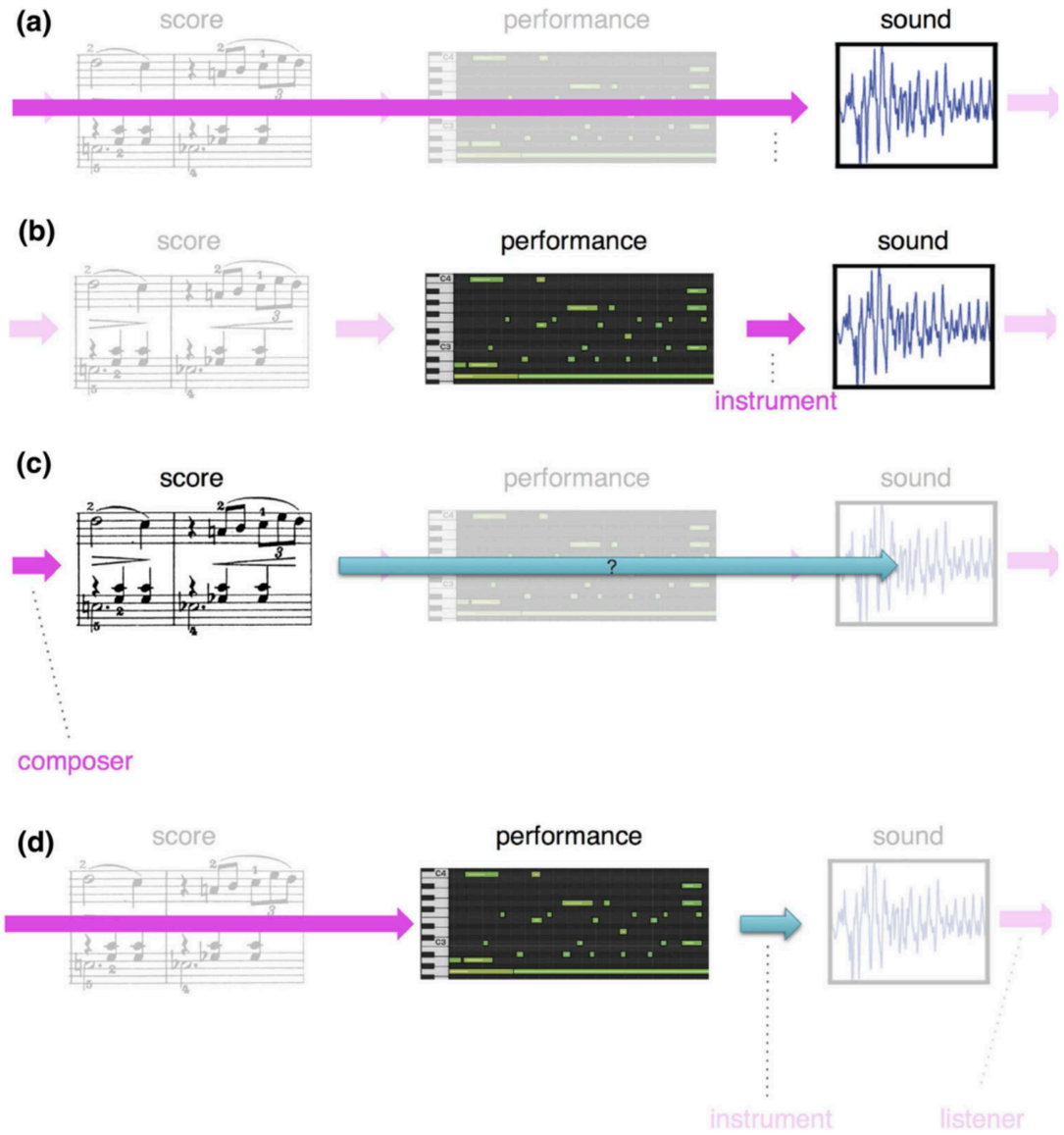
Symbolic representations of sound



Each of these steps can be learned, giving user control to the generation process.

Source: [Oore et al., 2020]

Different generation strategies

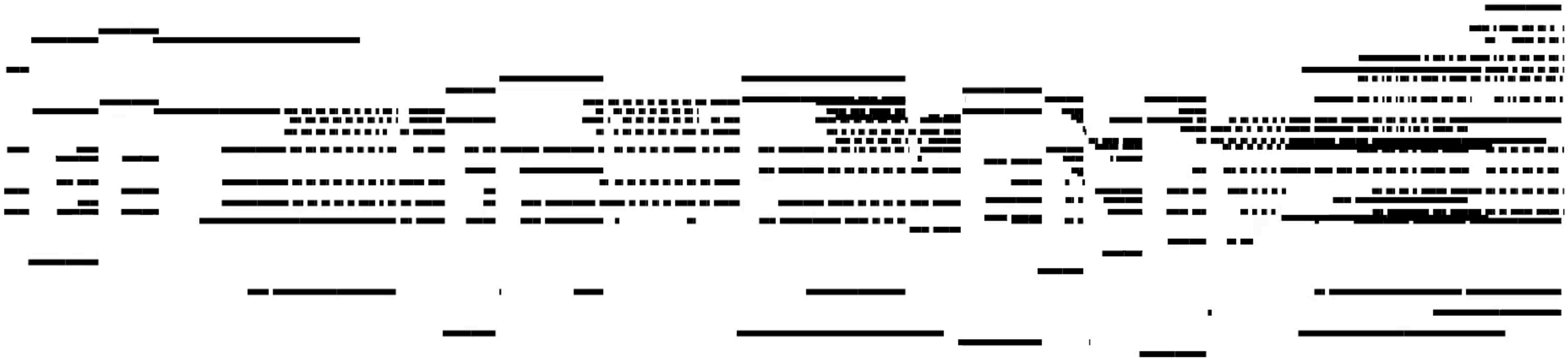


Source: [Oore et al., 2020]

Music transformer result

Source: [Huang et al., “Music Transformer”, 2018]

Music transformer



Generated sound + last layer attention visualization

Source: [Huang et al., “Music Transformer”, 2018]