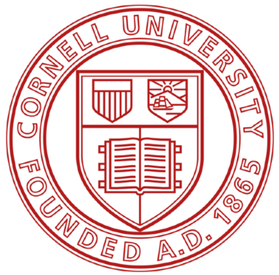
Moving beyond video in, audio out

Andrew Owens
Cornell Tech

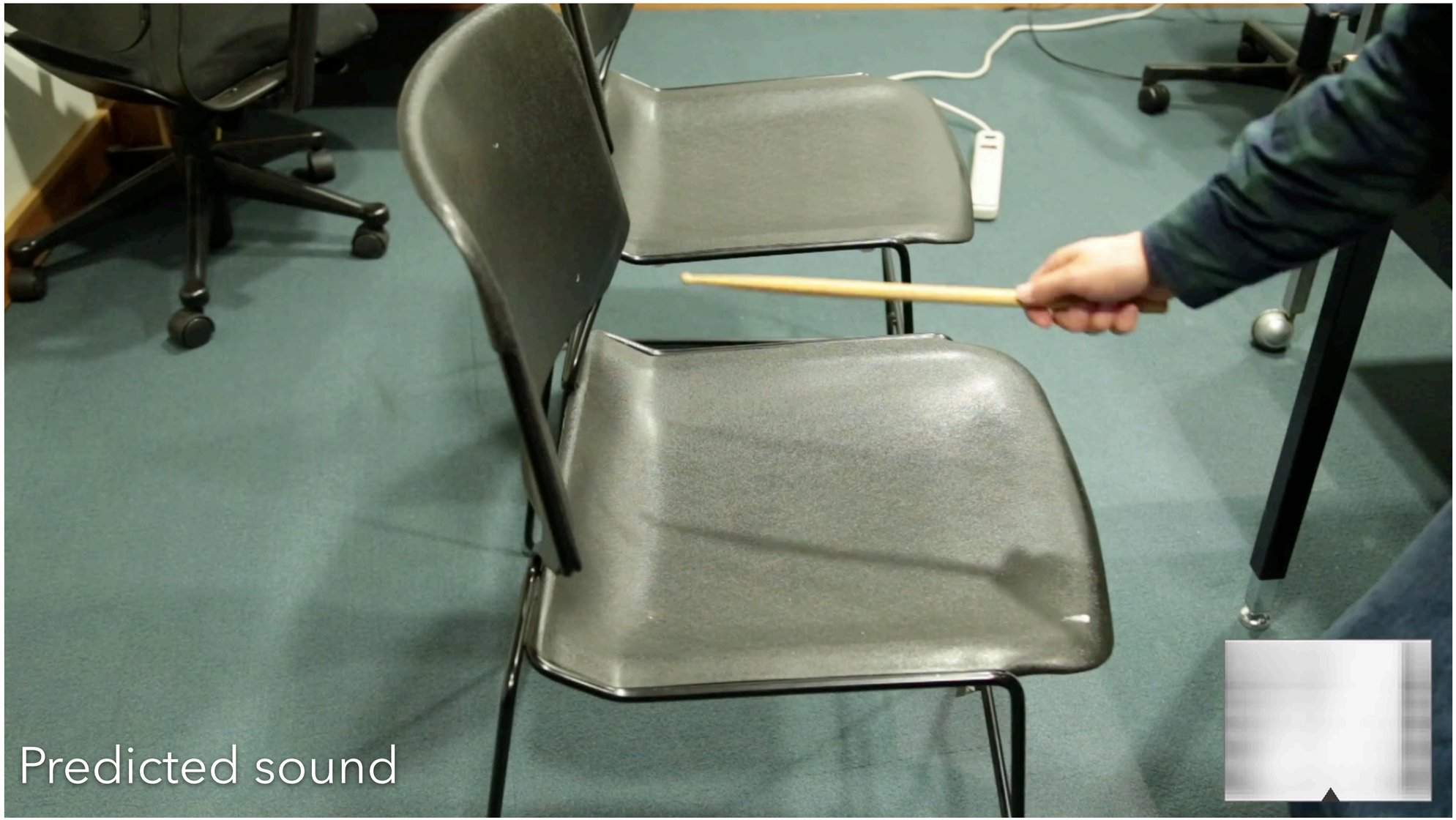


Video-to-sound (circa 10 years ago)

[Owens, Isola, McDermott, Torralba, Adelson, Freeman, "Visually Indicated Sounds", CVPR 2016]

The *Greatest Hits* dataset

Video-to-sound (circa 10 years ago)

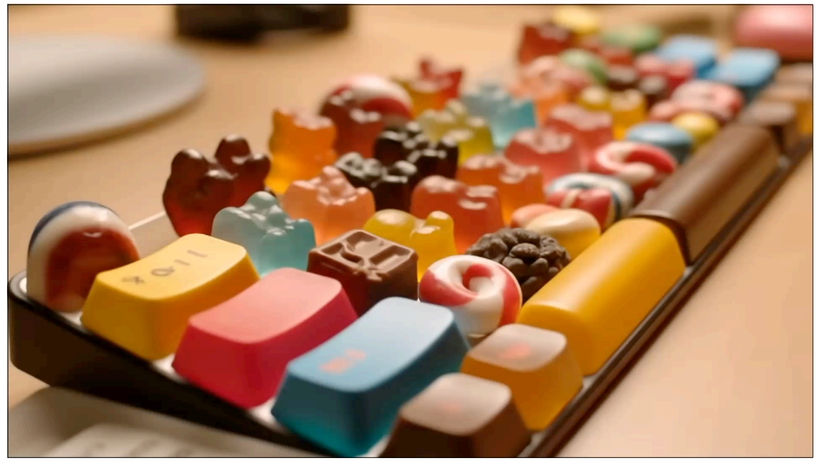


[Owens, Isola, McDermott, Torralba, Adelson, Freeman, “Visually Indicated Sounds”, CVPR 2016]

The *Greatest Hits* dataset

Moving beyond images in, audio out

Video-to-audio [Wang et al., “Frieren”, 2024]



Joint generation [Veo 3, 2025]

- Generative audio-visual models are rapidly improving.
- But also a lot of missing capabilities!
  - User control
  - Physical constraints
  - Ability to interact with a visual scene

Why do we need user control?

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

Foley Generation with Multimodal Control

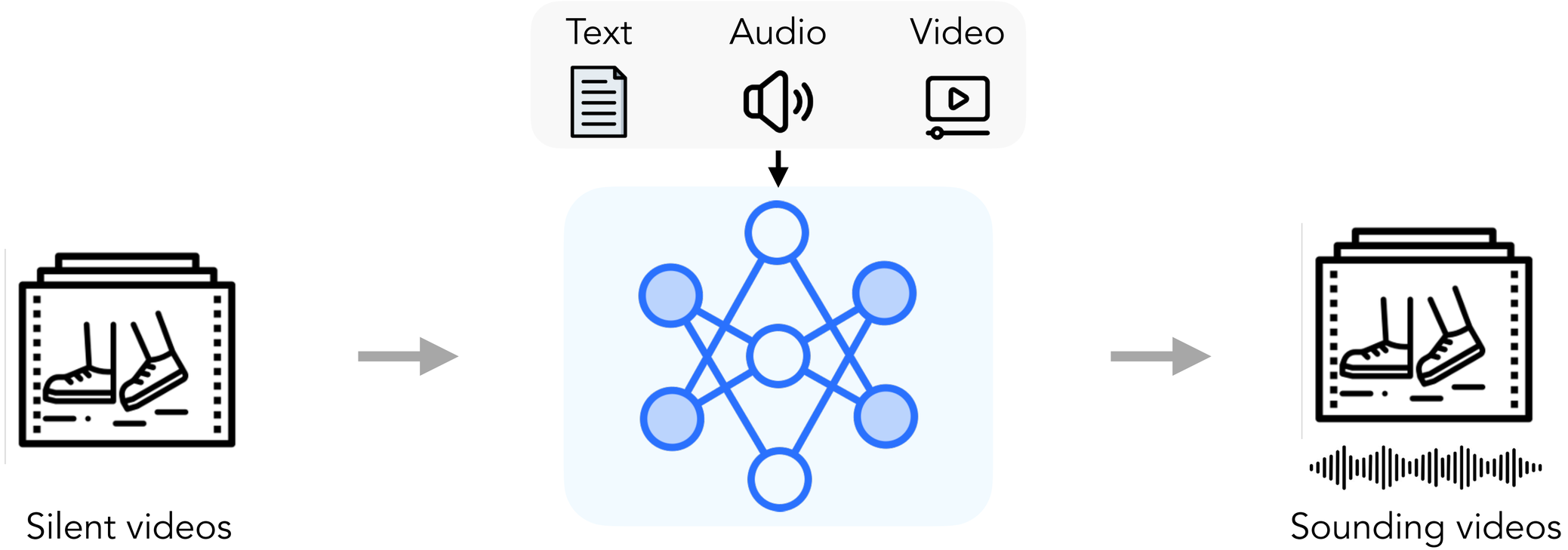
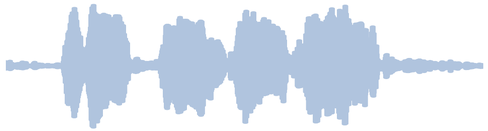


Image icons credits to Luo et al., “DIFF-FOLEY: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models.”

Multimodal control for video-to-audio generation

Input video



Input audio

Video Encoder

re-sample

Audio Encoder

masked diffuse

Diffusion Transformer

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

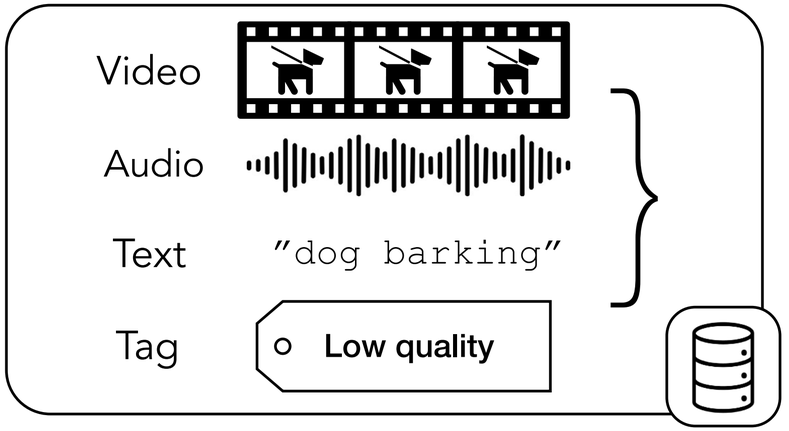
GT latents

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

Ziyang Chen

Prem Seetharaman

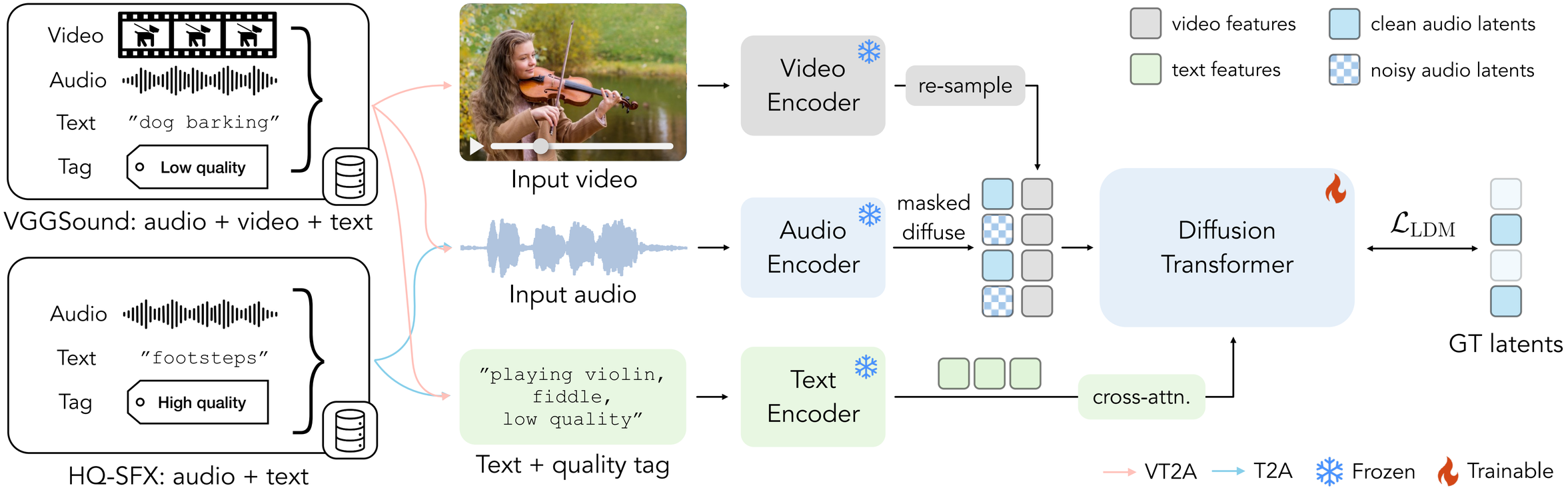
Training with two data sources



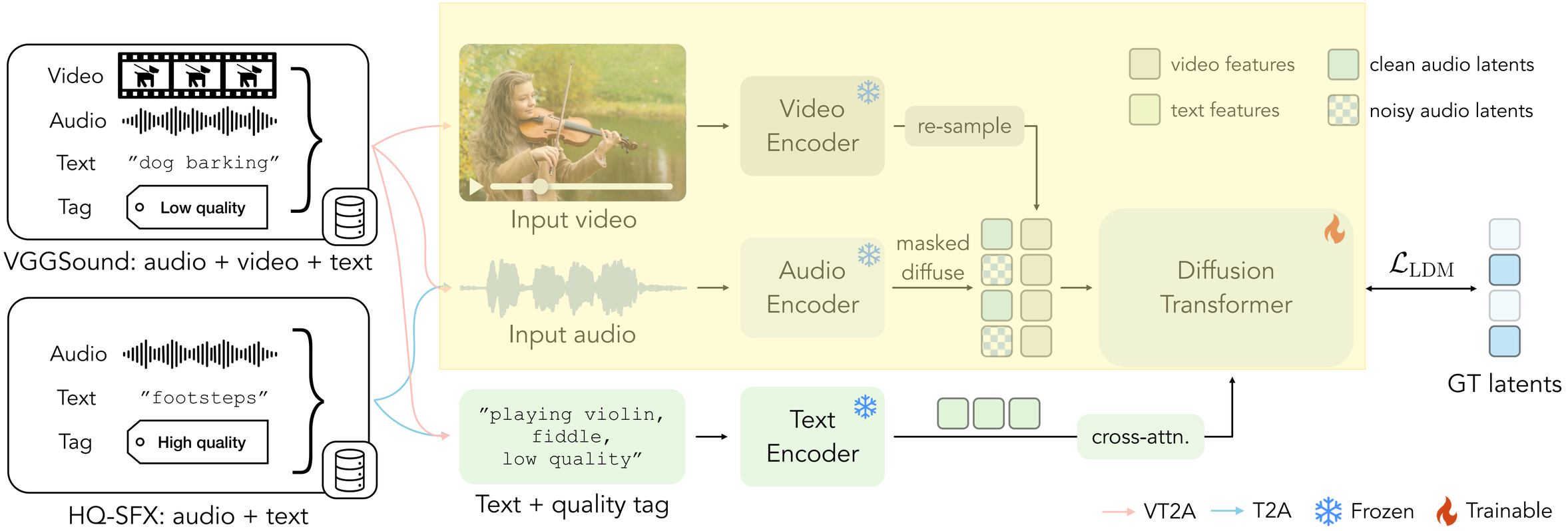
VGGSound: audio + video + text

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

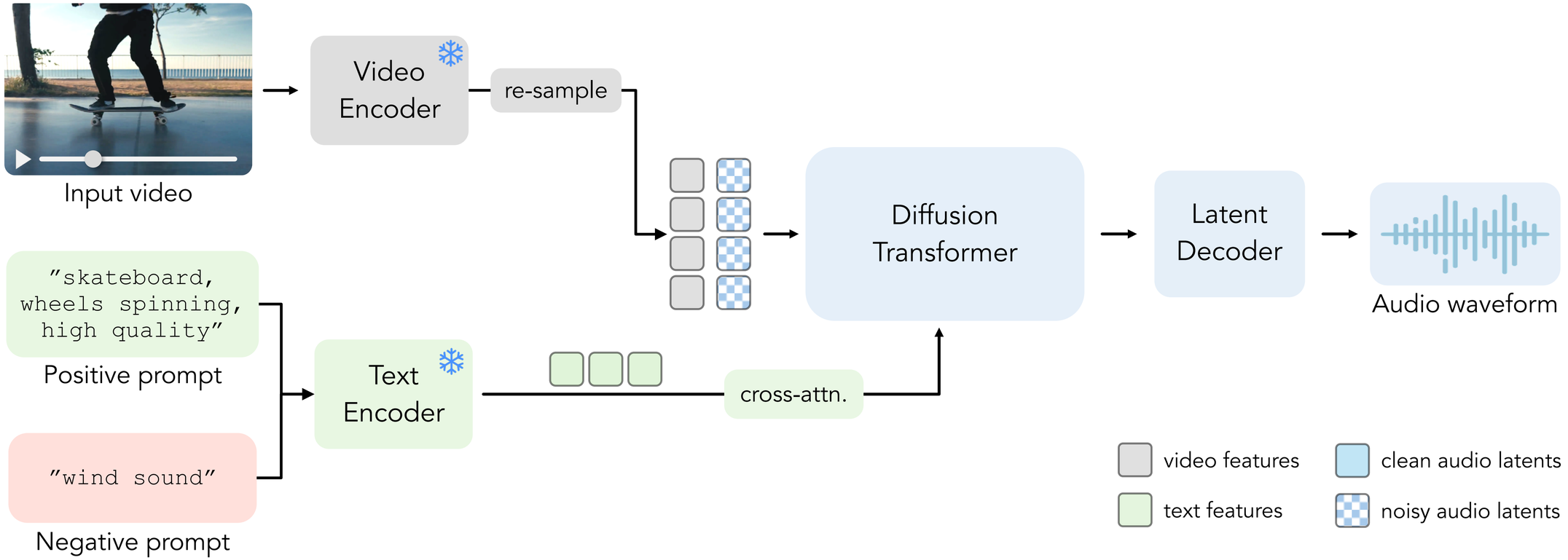
Generated sound for silent video

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`

Adding language conditioning

“bird chirping”

Adding language conditioning

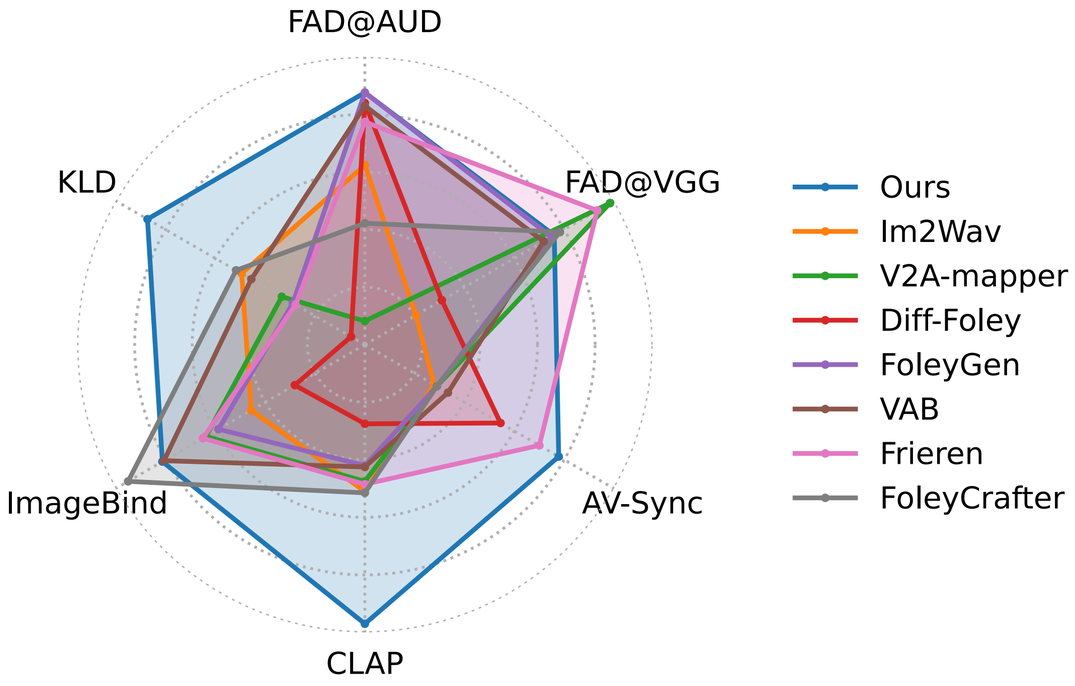
“rooster crowing”

Adding language conditioning

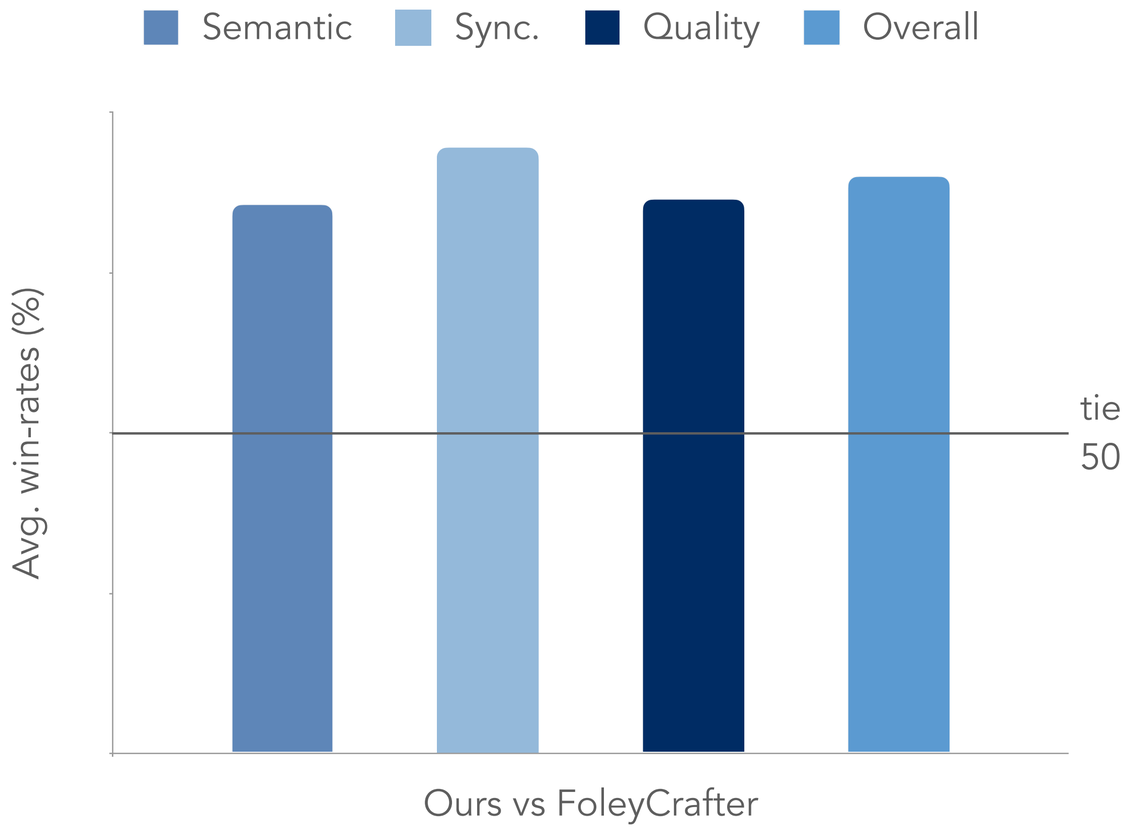
“male speaking”

Quantitative Evaluation

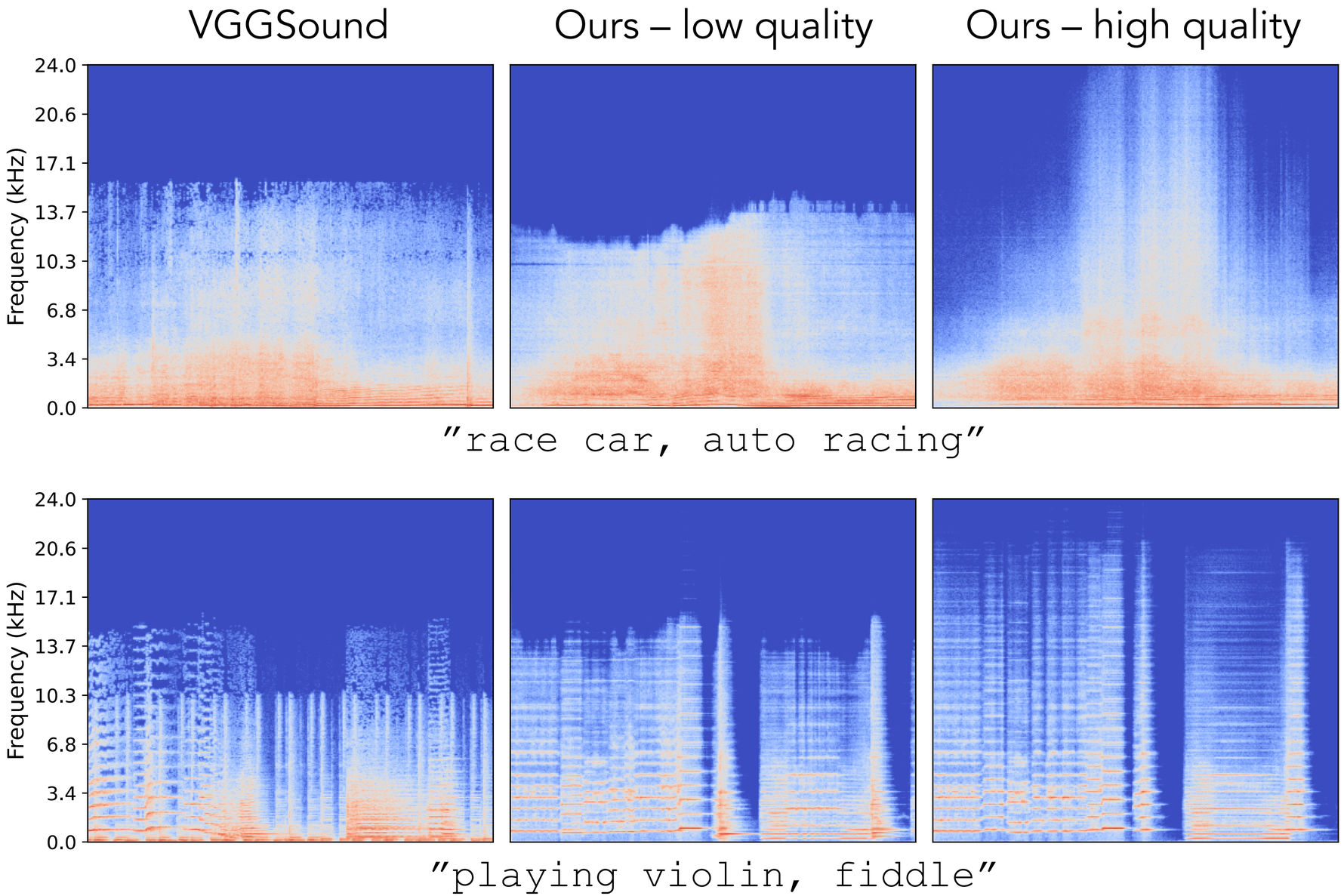
Video-to-audio



Human Study on Text Control



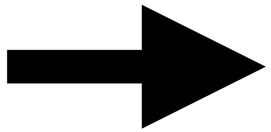
Foley Generation with Quality Control



Failure Case

Text prompt: `skateboard, water splash, high quality`

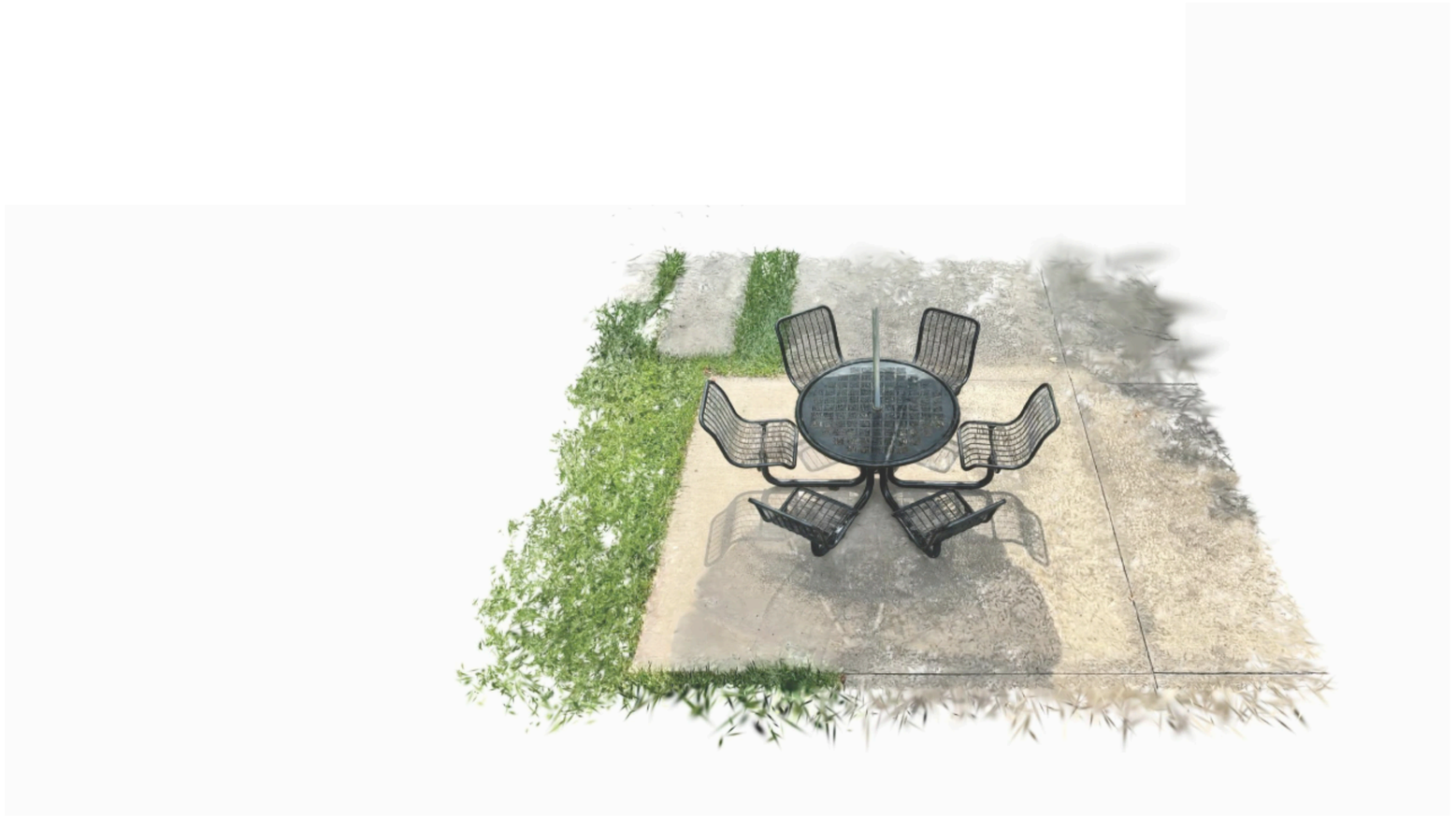
What would it sound like if I interacted with this scene?



Audio

Input sceneWe need a video to do this!

Audio-visual scene reconstruction



Multimodal reconstruction


Yiming Dou

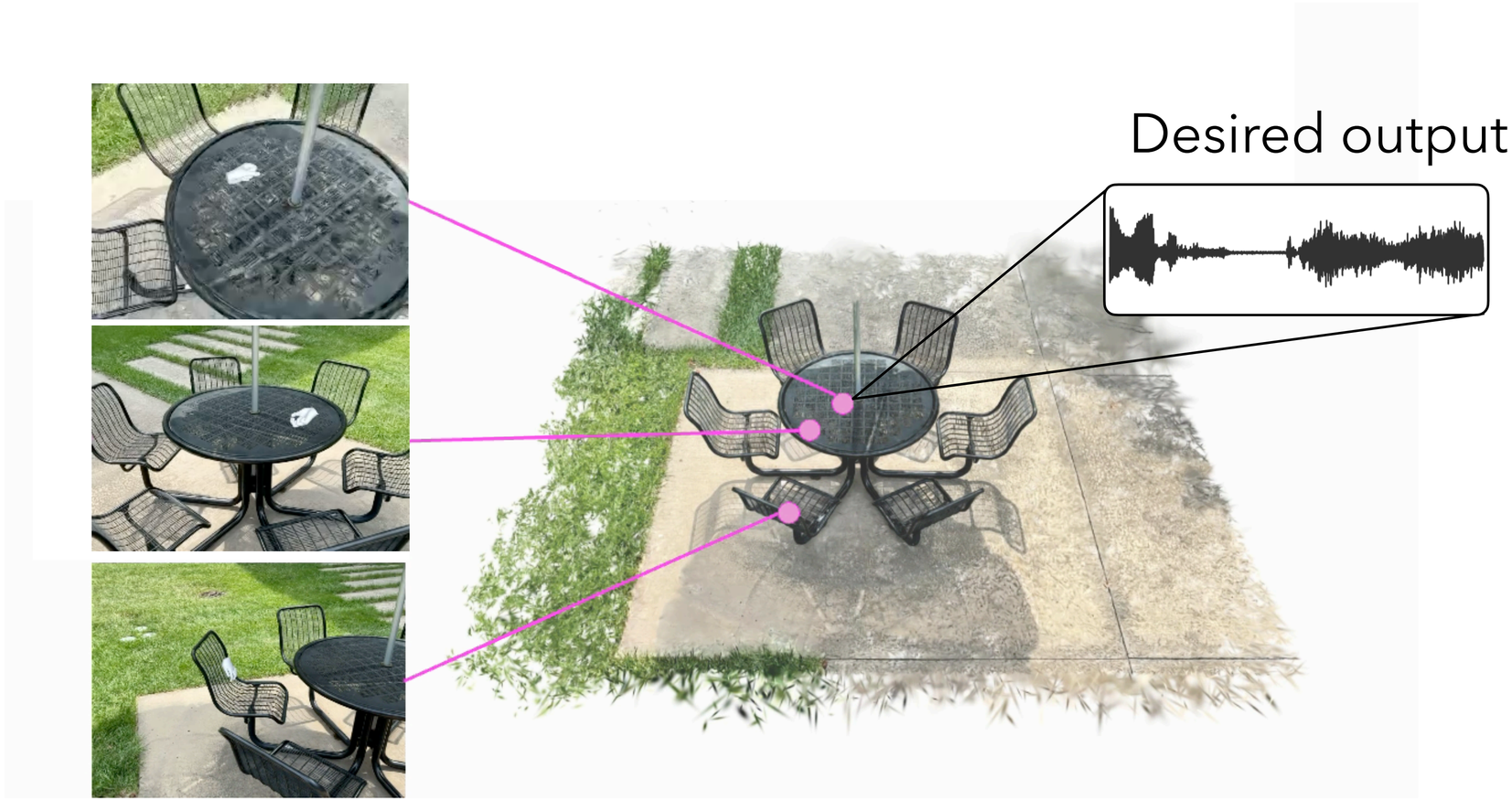

Wonseok Oh


Yuqing Luo


Antonio Loquercio

["Hearing Hands: Generating Sounds from Physical Interactions in 3D Scenes", CVPR 2025

Audio-visual scene reconstruction




Yiming Dou

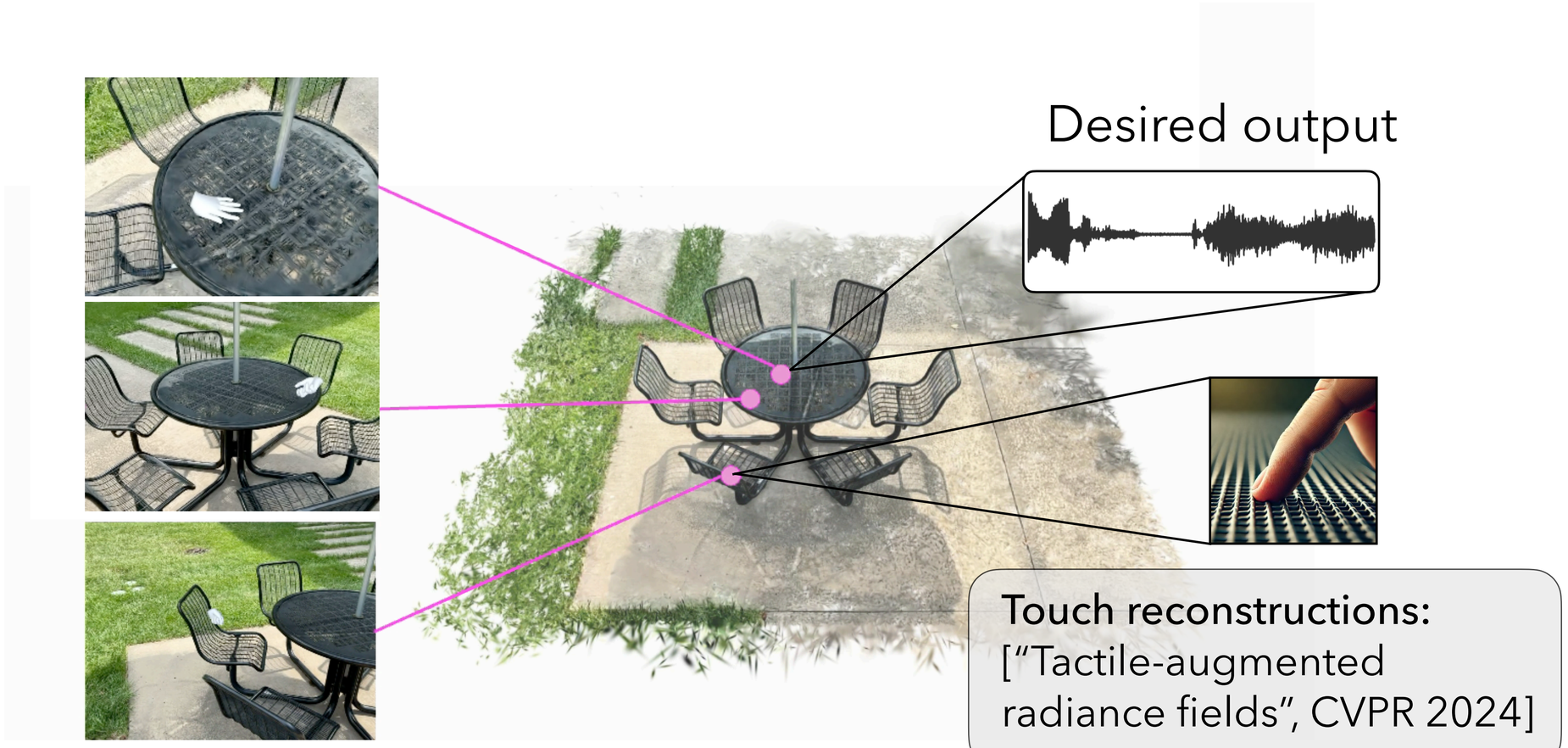

Wonseok Oh

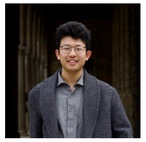

Yuqing Luo


Antonio Loquercio

["Hearing Hands: Generating Sounds from Physical Interactions in 3D Scenes", CVPR 2025

Audio-visual scene reconstruction




Yiming Dou


Antonio Loquercio

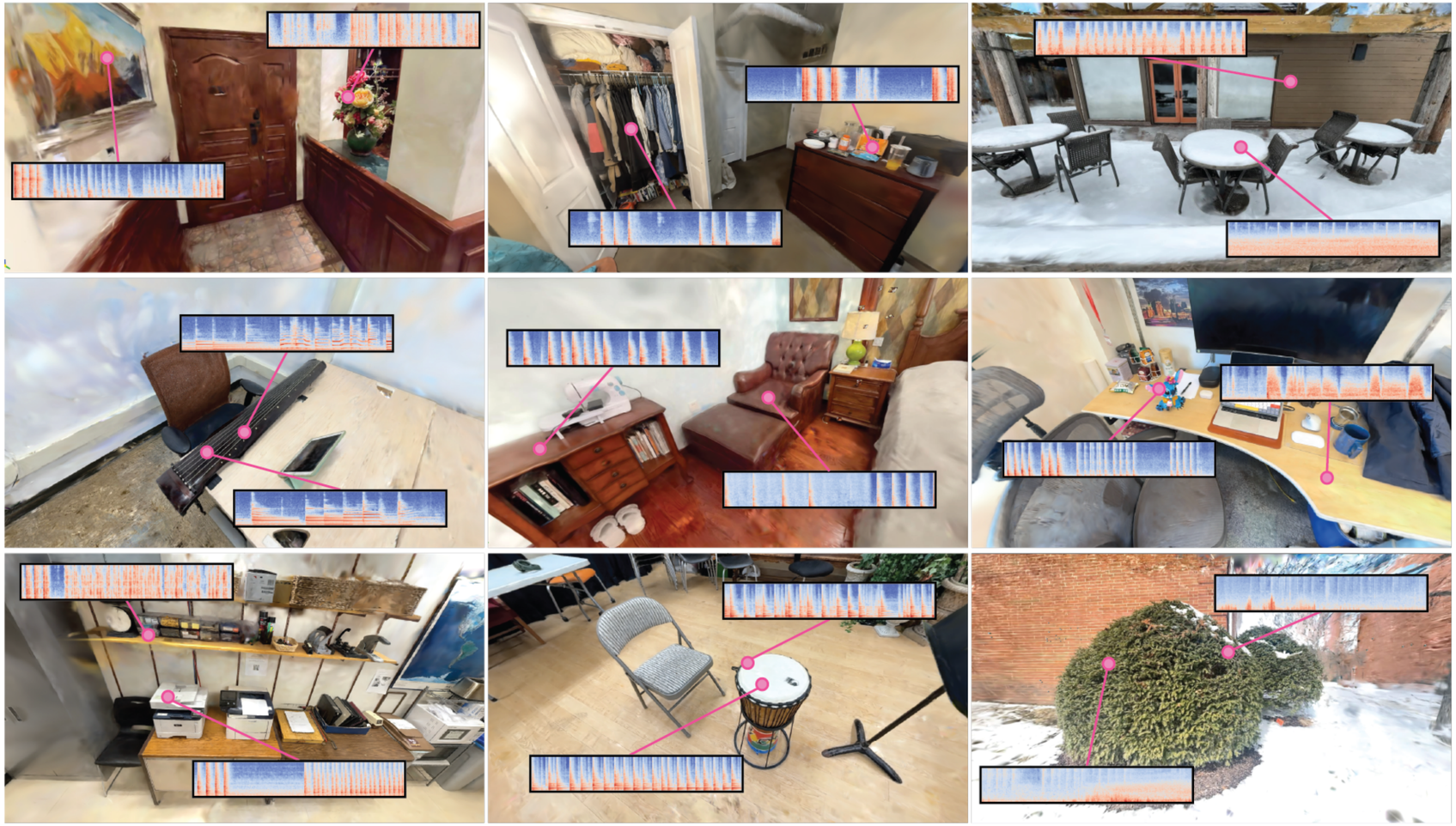
Capturing audio from interactions

Reconstructed with HaMeR [Pavlakos et al., 2023]

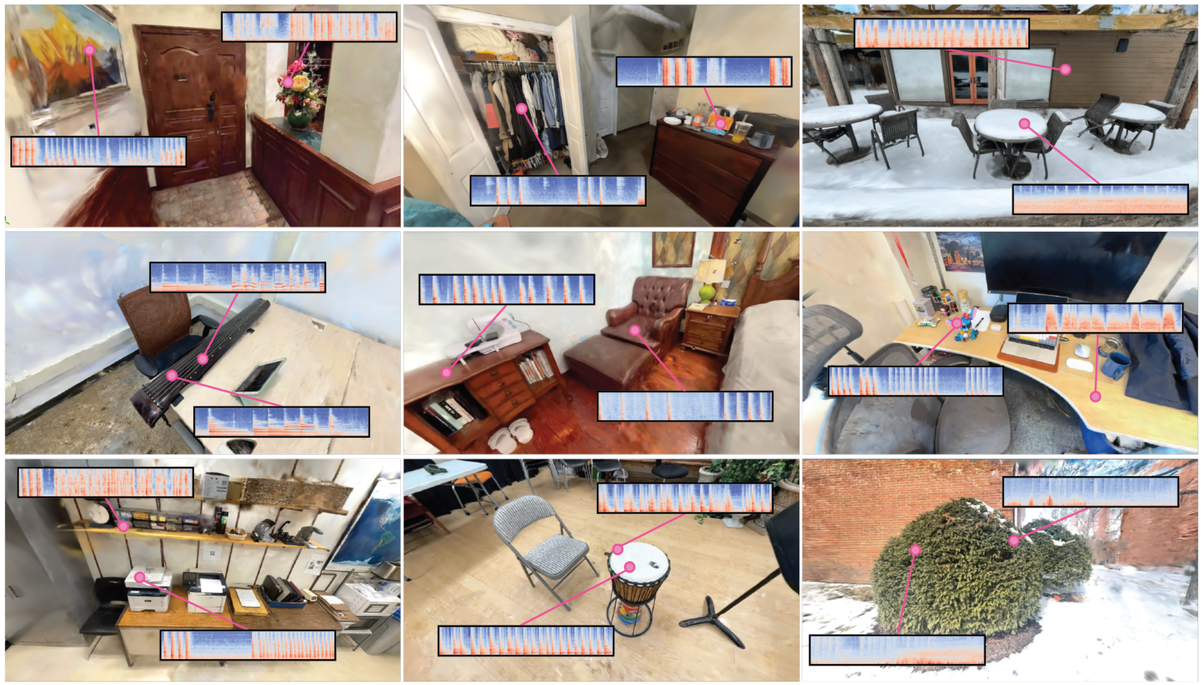
Dataset

Contents:

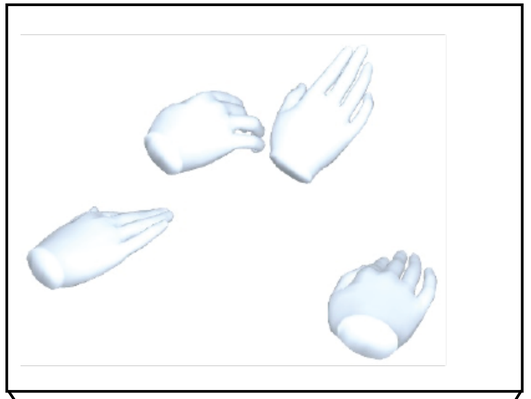
- 24 captured scenes
- Reconstructed by Gaussian Splatting
- 3D hand trajectories
- Sound recordings



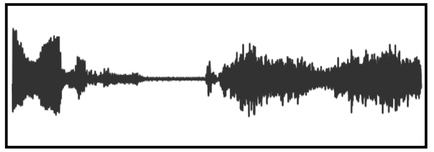
Hand-to-sound



Dataset

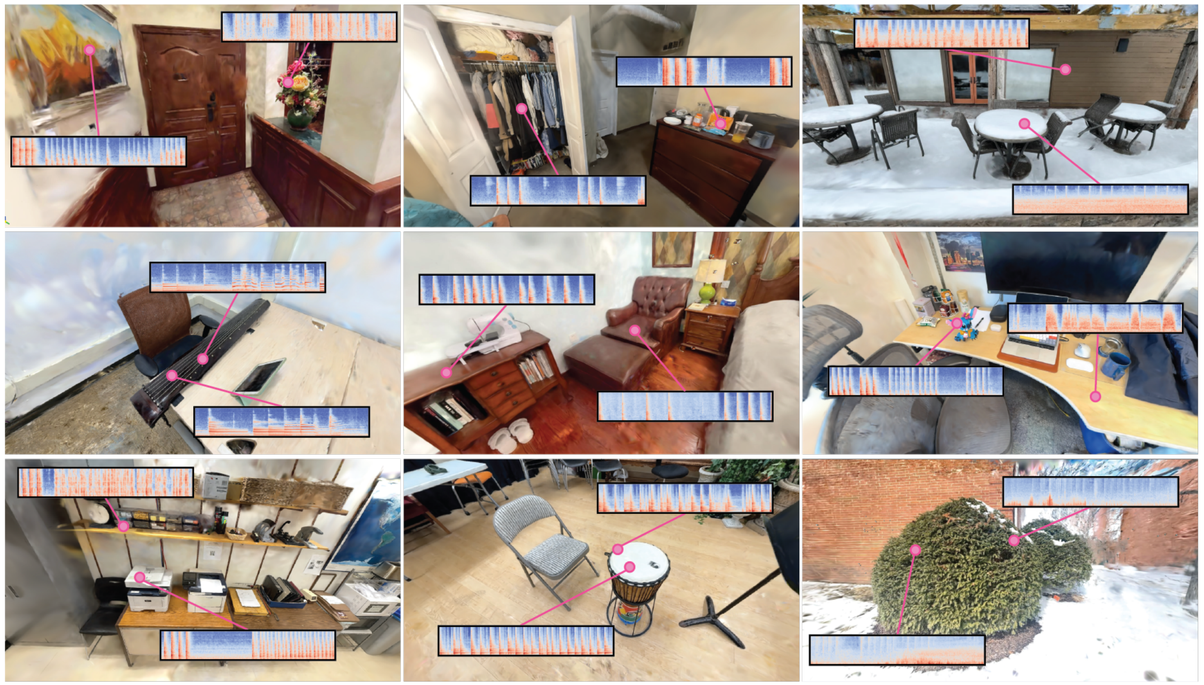


Input: action



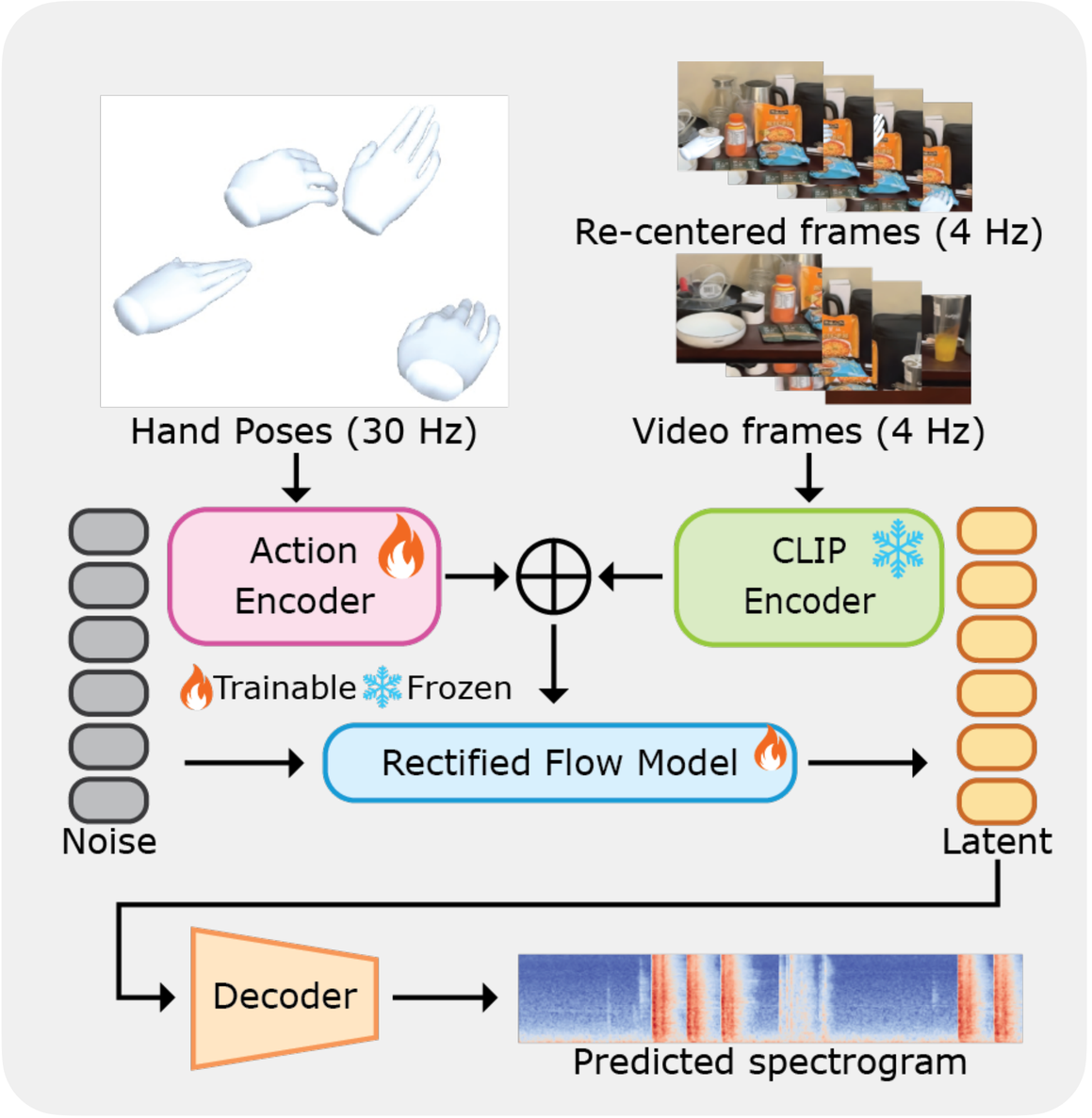
Output: sound

Hand-to-sound



Dataset

- Trained on all 24 scenes at once.
- Not aiming to generalize to new *scenes*.
- Just new *interactions* in the same scenes.




Generated sound


Generated sound


Generated sound

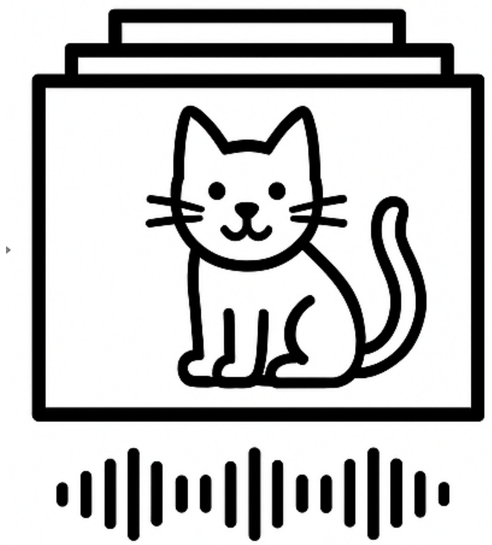

Generated sound


Generated sound


Generated sound


Step 1: Pick a location to interact with

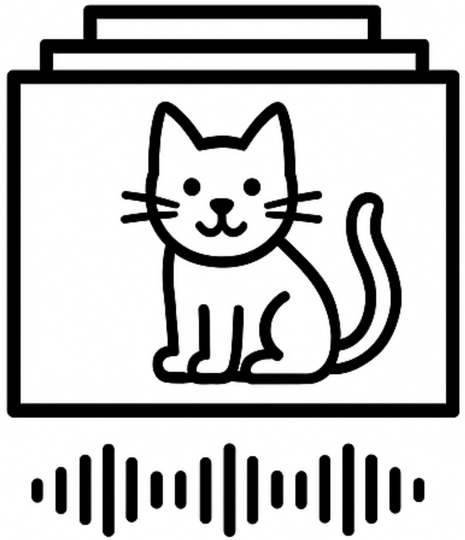
When you think of “multimodal”:



Multimodal example $X = (I, A)$:

- I is an image
- A is audio

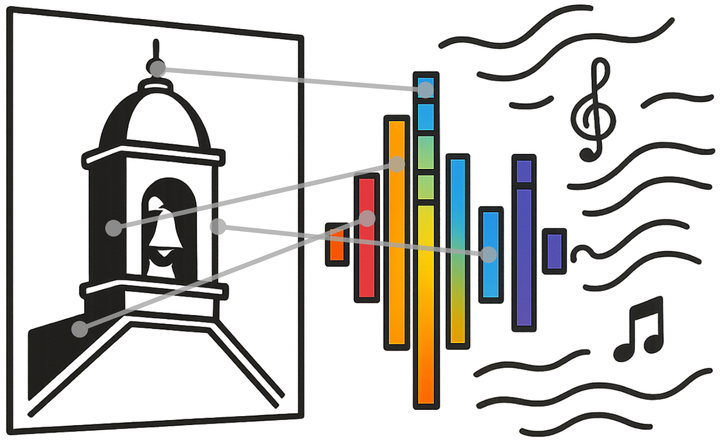
When you think of “multimodal”:



Multimodal example $X = (I, A)$:

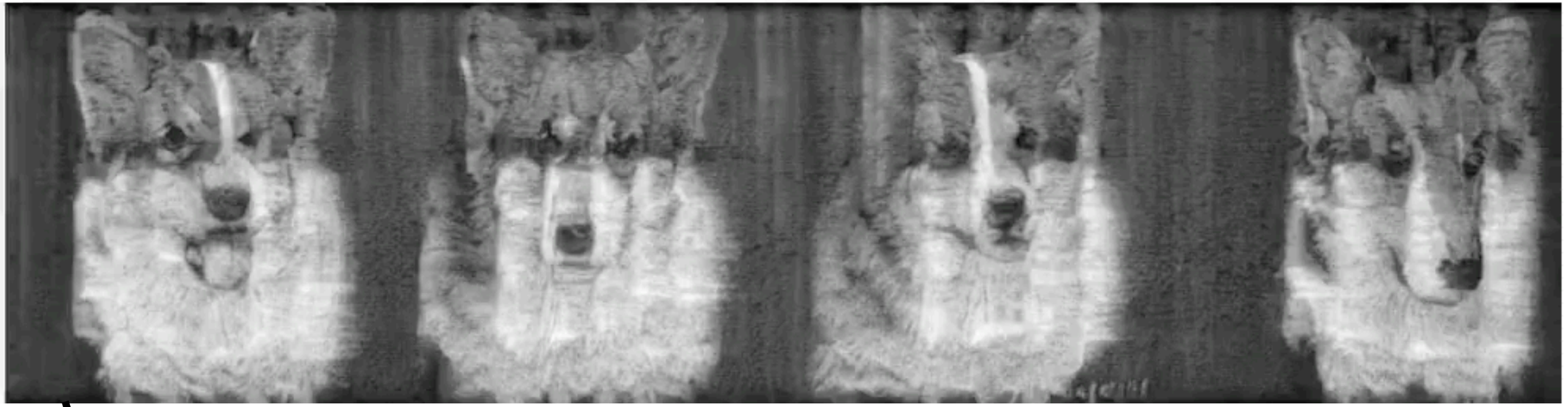
- I is an image
- A is audio

Instead:



Multimodal example X:

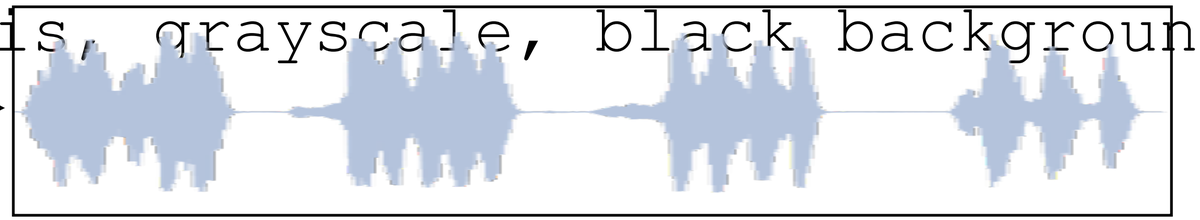
- X is an image
- X is audio



“a p orgis, grayscale, black background”

“dog

Log-mel Spectrogram to Audio

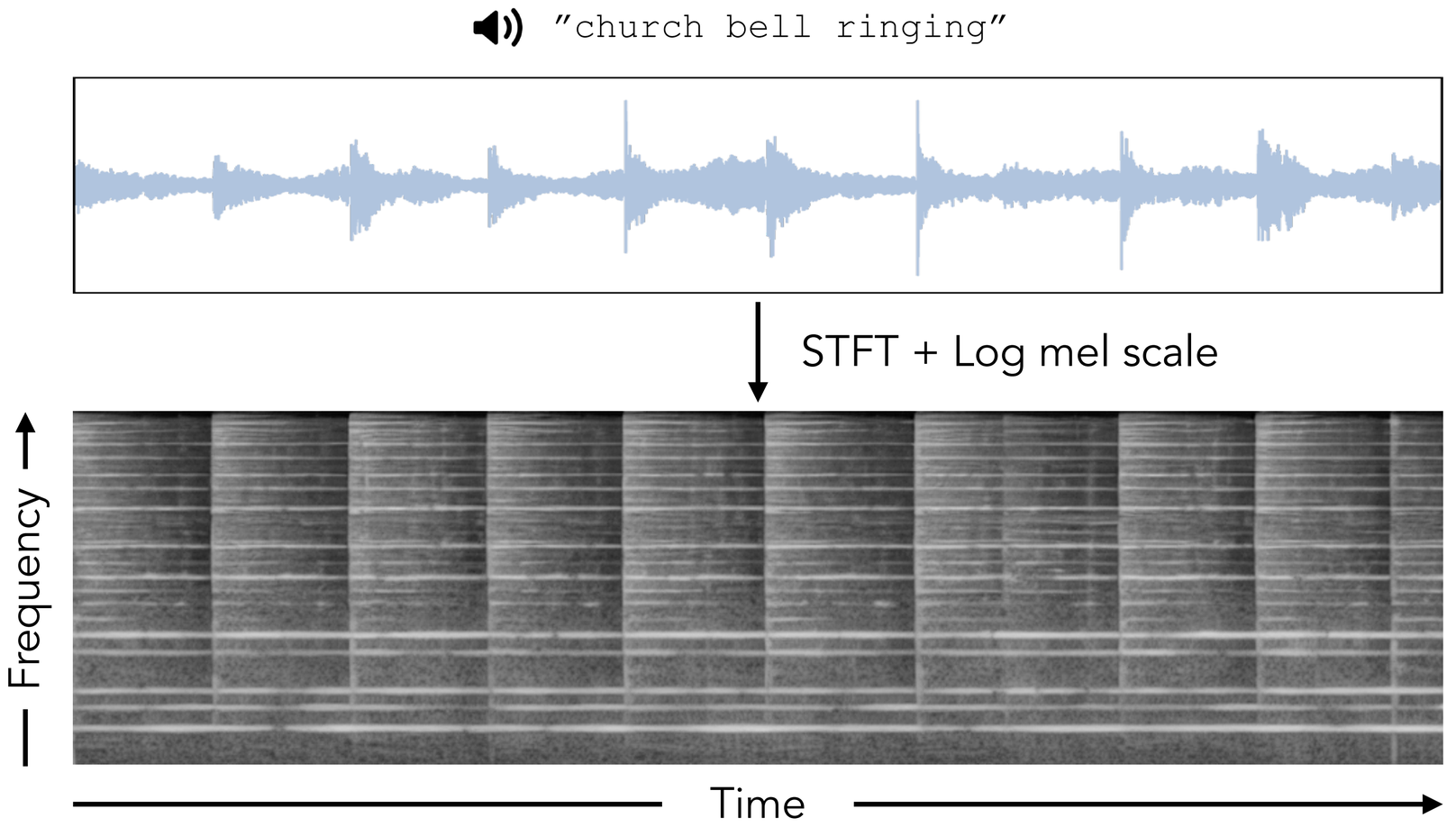


[Chen, Geng, Owens. “Images that Sound: Composing Images and Sounds on a Single Canvas.” NeurIPS, 2024]

Ziyang Chen

Daniel Geng

Audio spectrograms

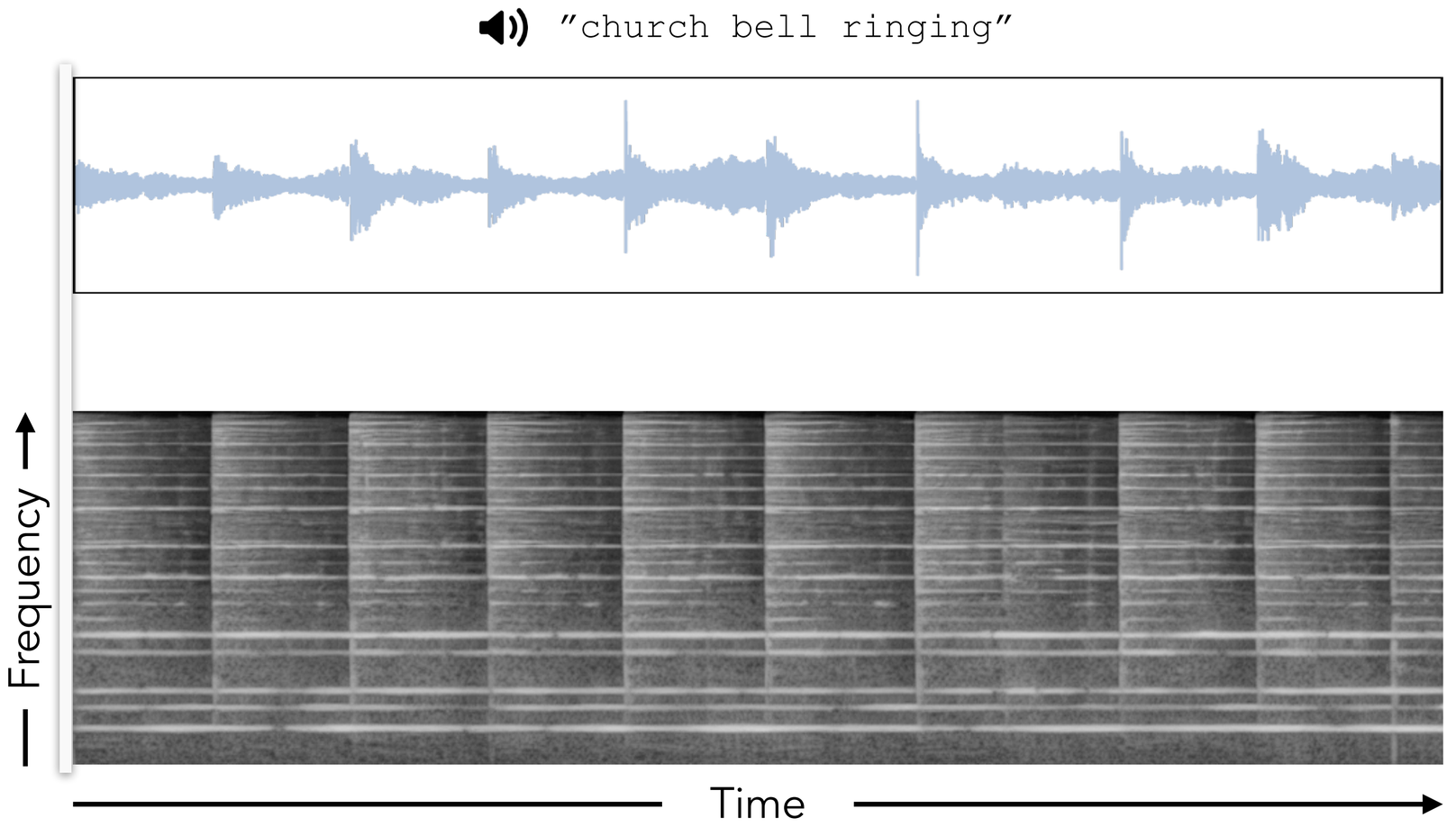


[Chen, Geng, Owens. "Images that Sound: Composing Images and Sounds on a Single Canvas." NeurIPS, 2024]


Ziyang Chen


Daniel Geng

Audio spectrograms



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


Ziyang Chen


Daniel Geng

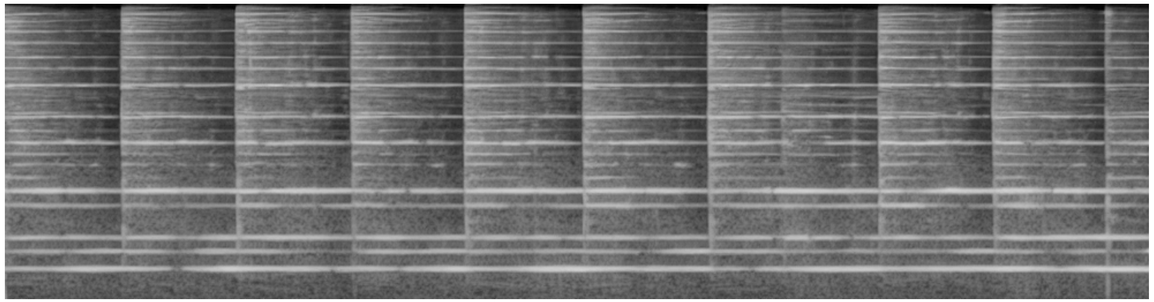
“Seeing” spectrograms

Images

”castles, grayscale”

Spectrograms

”church bell ringing”



“Seeing” spectrograms

Images

”castles, grayscale”

Spectrograms

”church bell ringing”

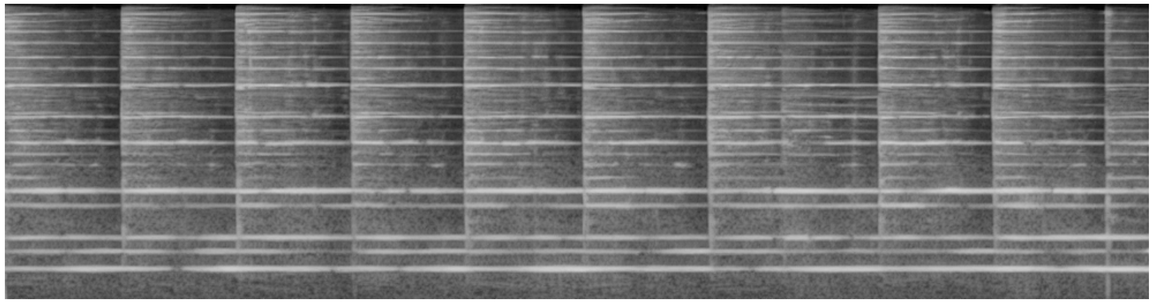
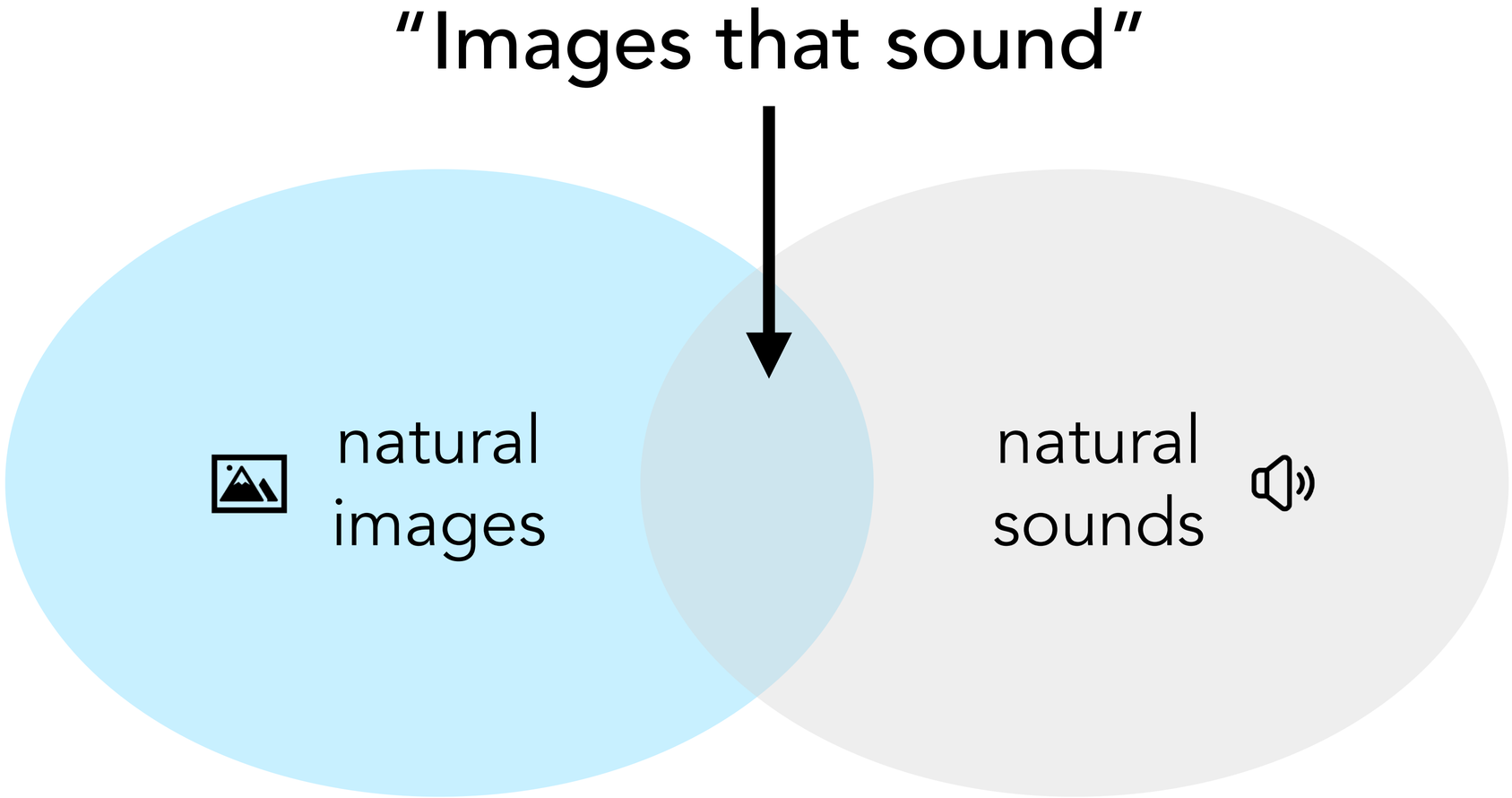
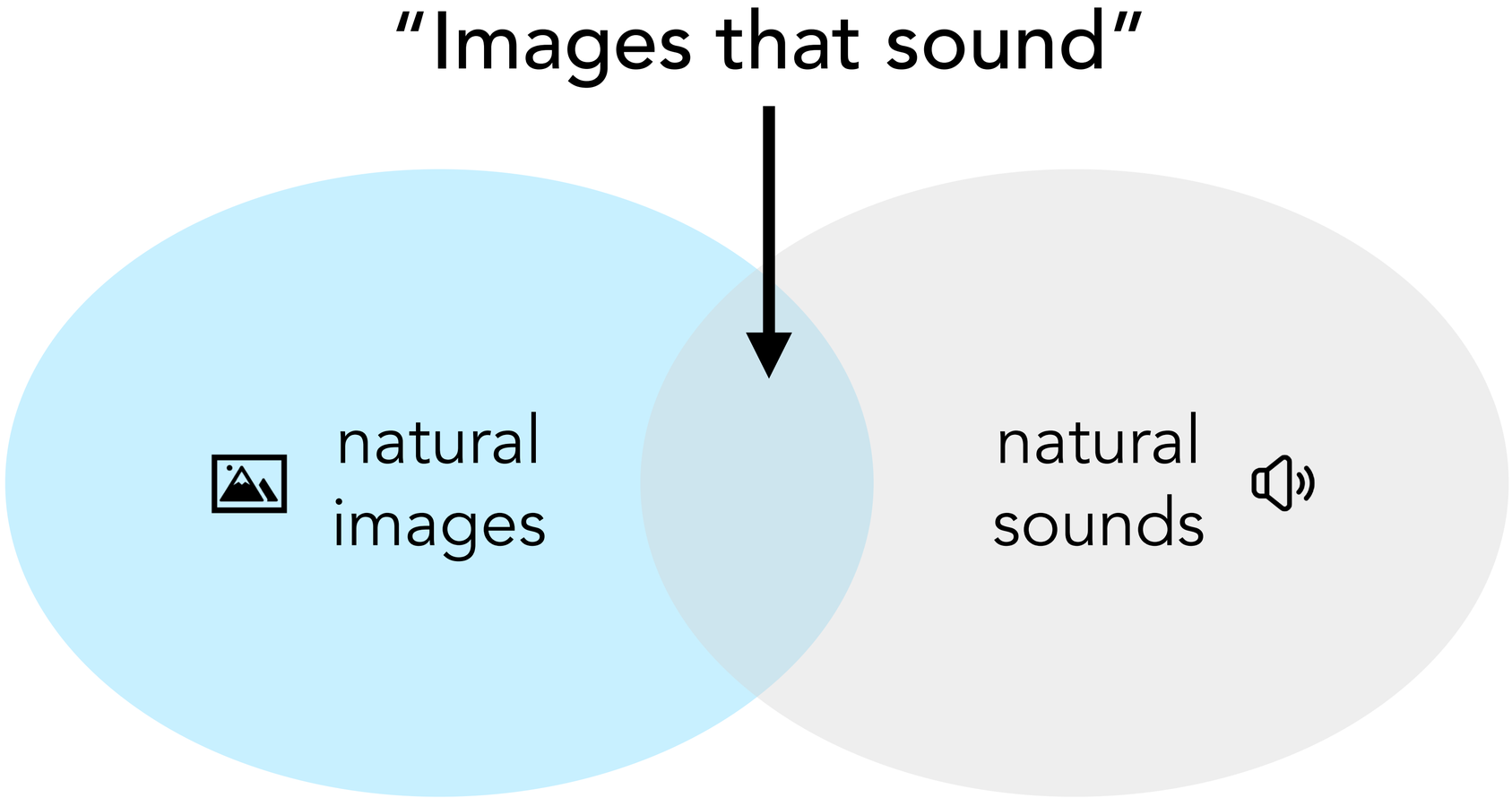


Image captioning
[Li et al., BLIP-2, 2023]

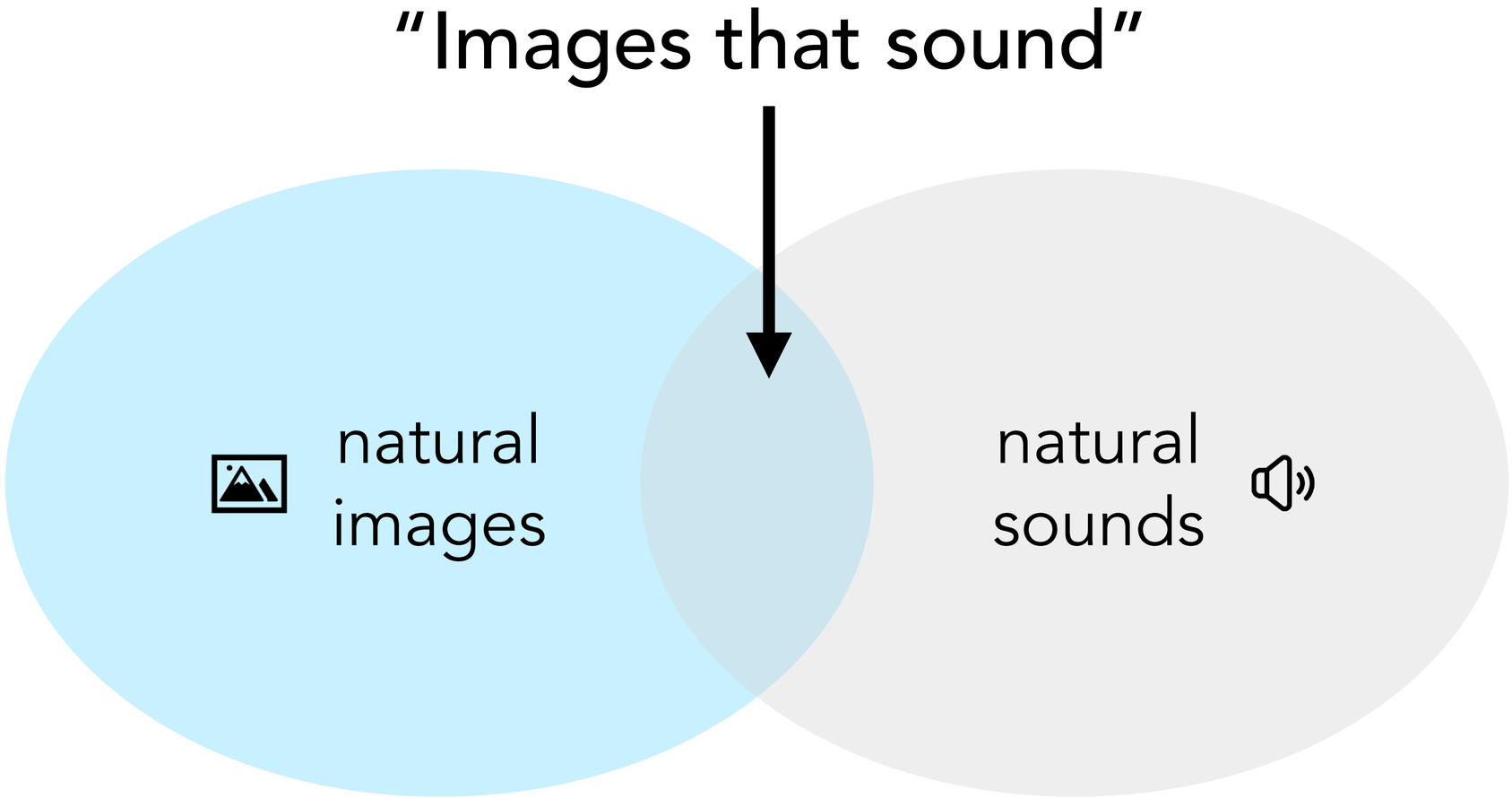
”a black and white photo of
a window with a curtain”



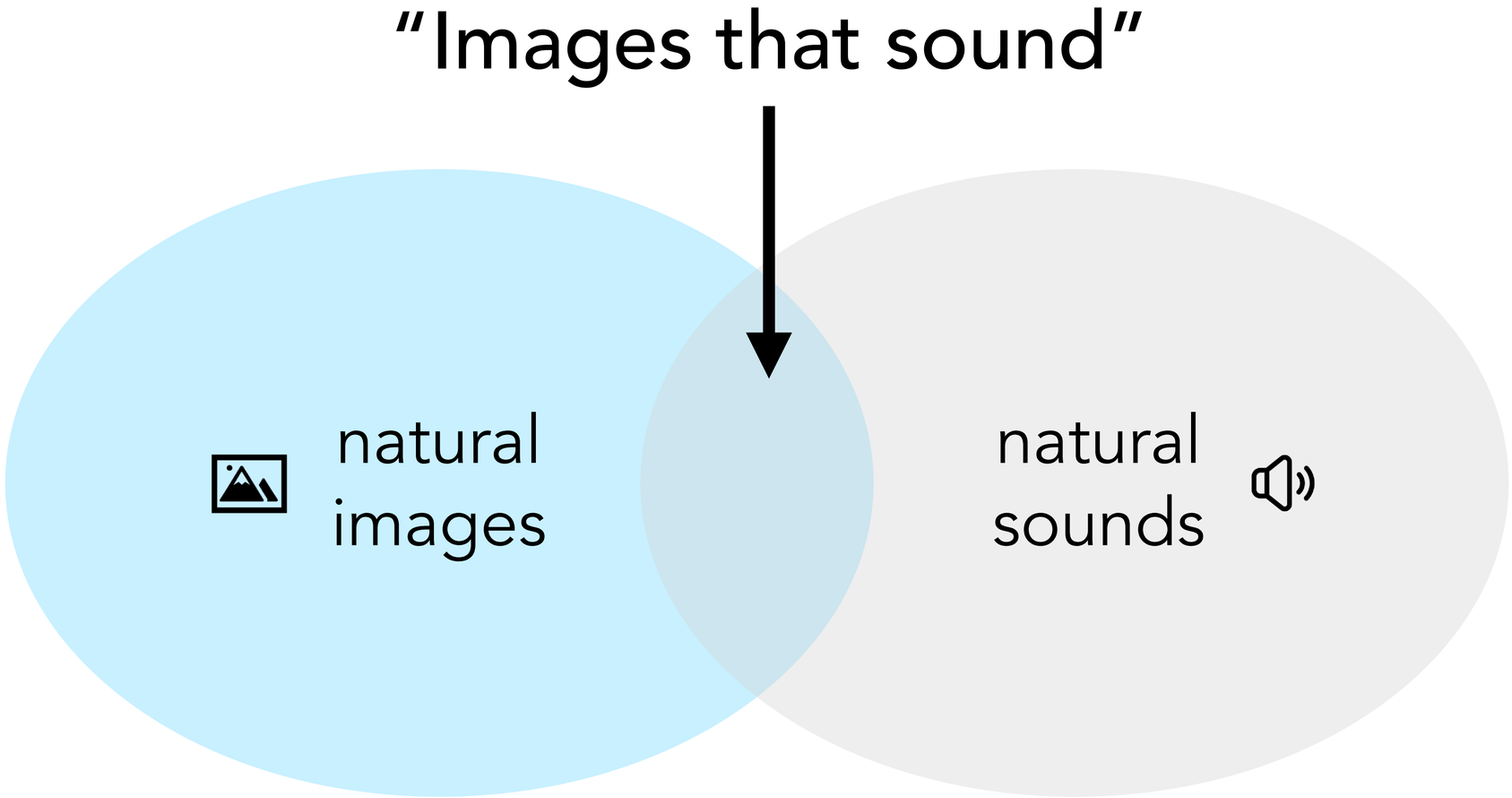
Name in homage to “Pixels that Sound” [Kidron et al., 2005] and “Objects that Sound” [Arandjelovic & Zisserman, 2018]



natural image statistics

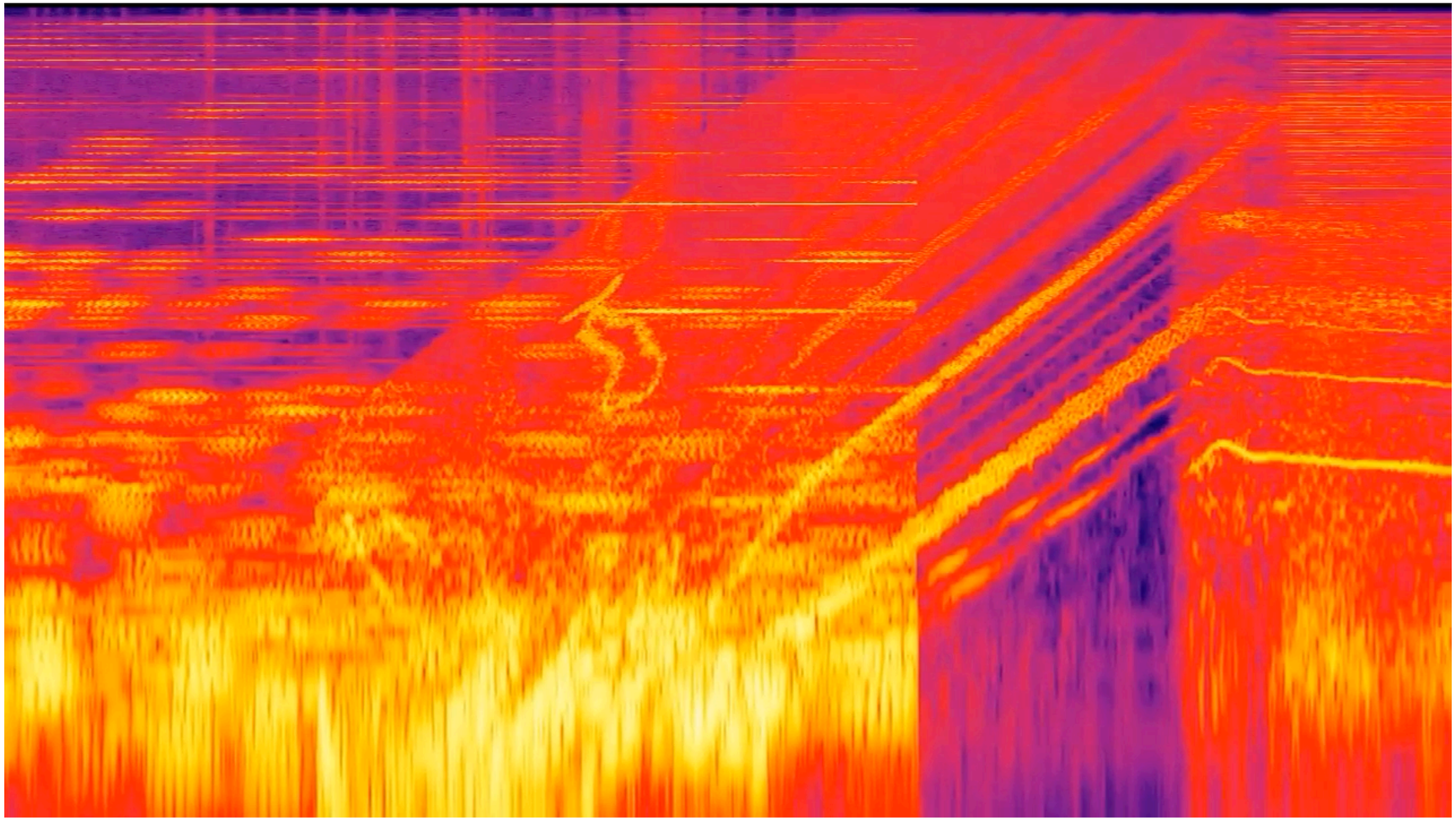


natural ~~image~~ statistics



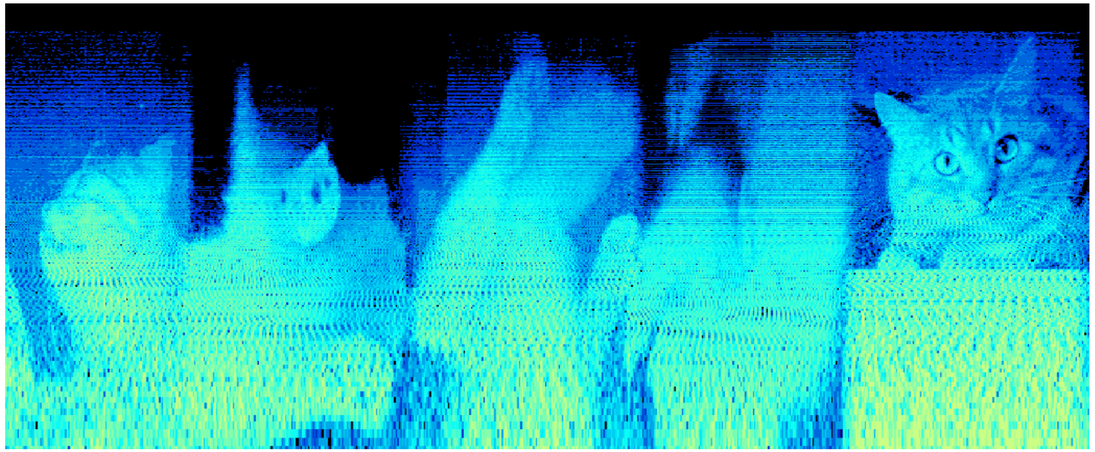
natural “image ∩ audio” statistics

Related work: spectrogram art

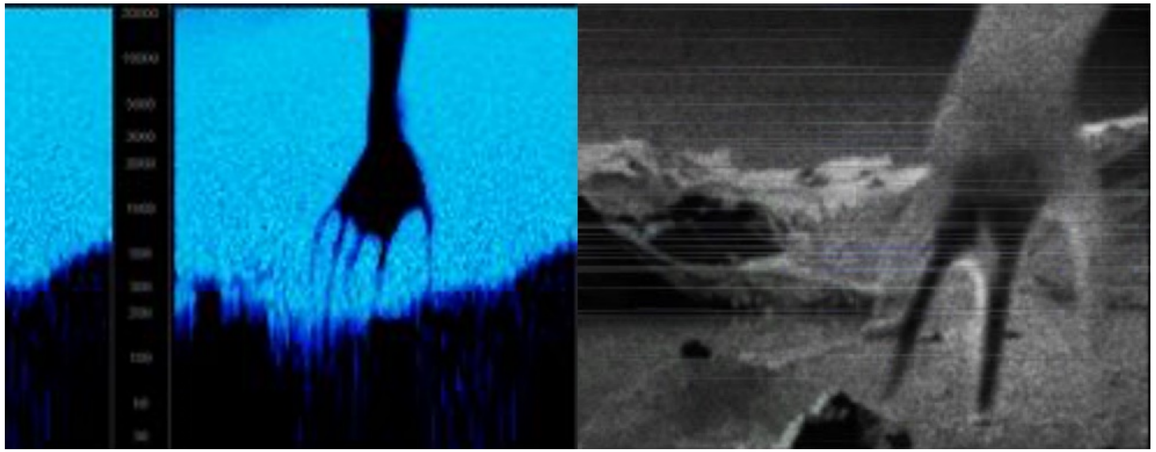


Aphex Twin, *Formula, 2001*

Related work: spectrogram art

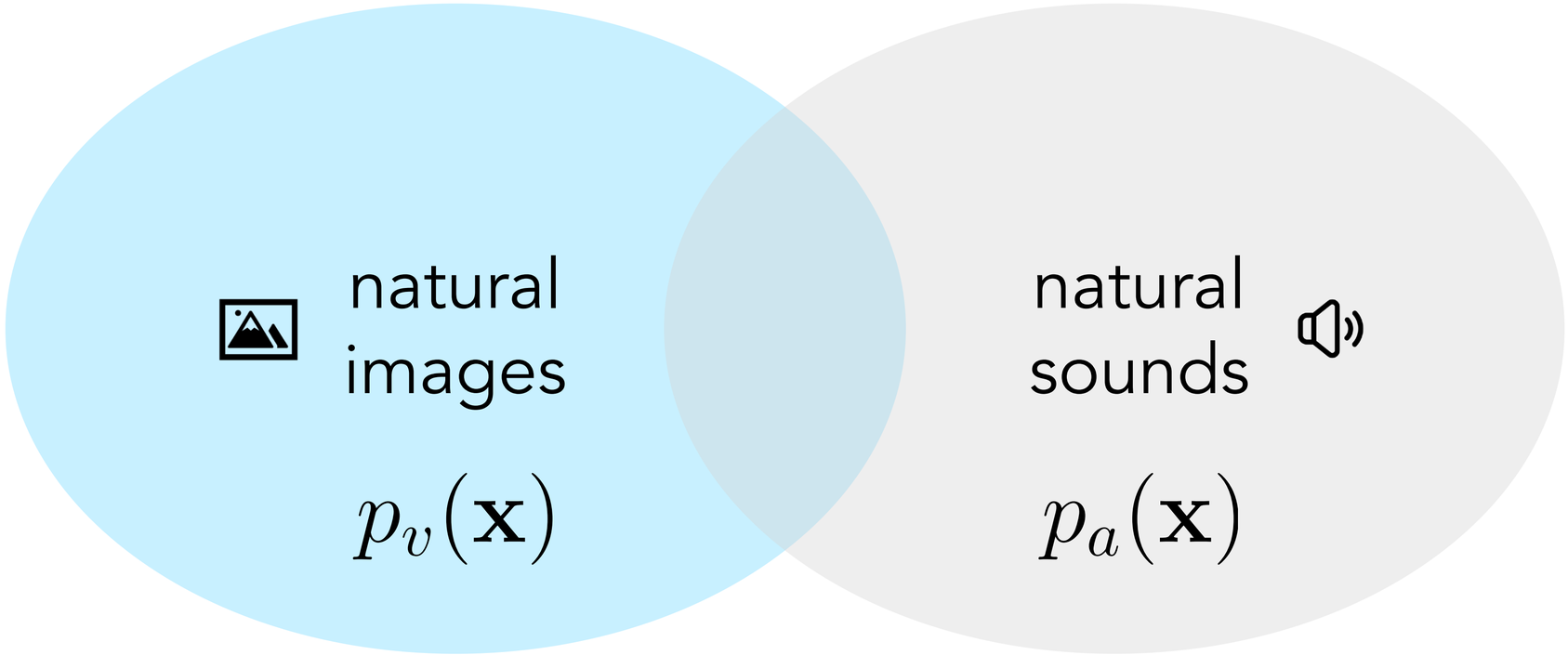


Venetian Snares, “Songs about my Cats”, 2006

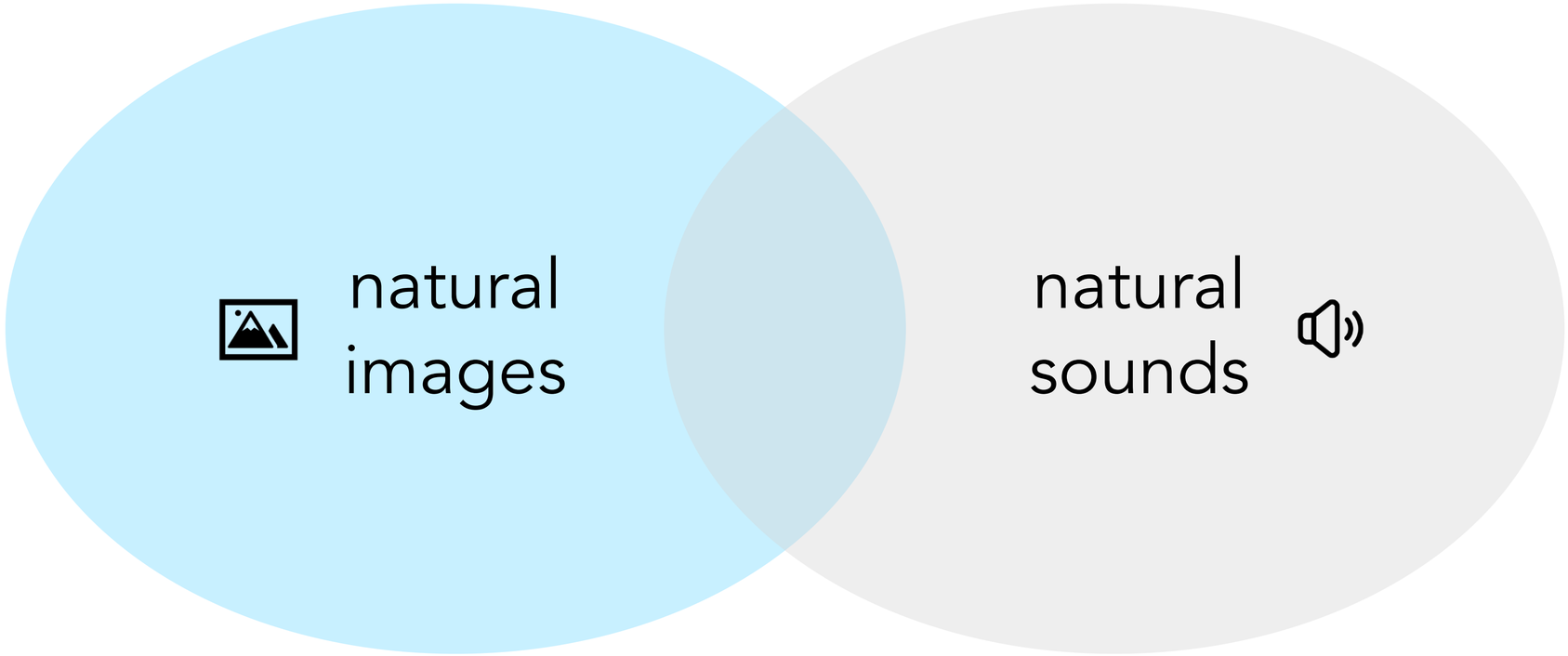


Nine Inch Nails, “My Violent Heart”, 2007

Composing sight with sound



Composing sight with sound

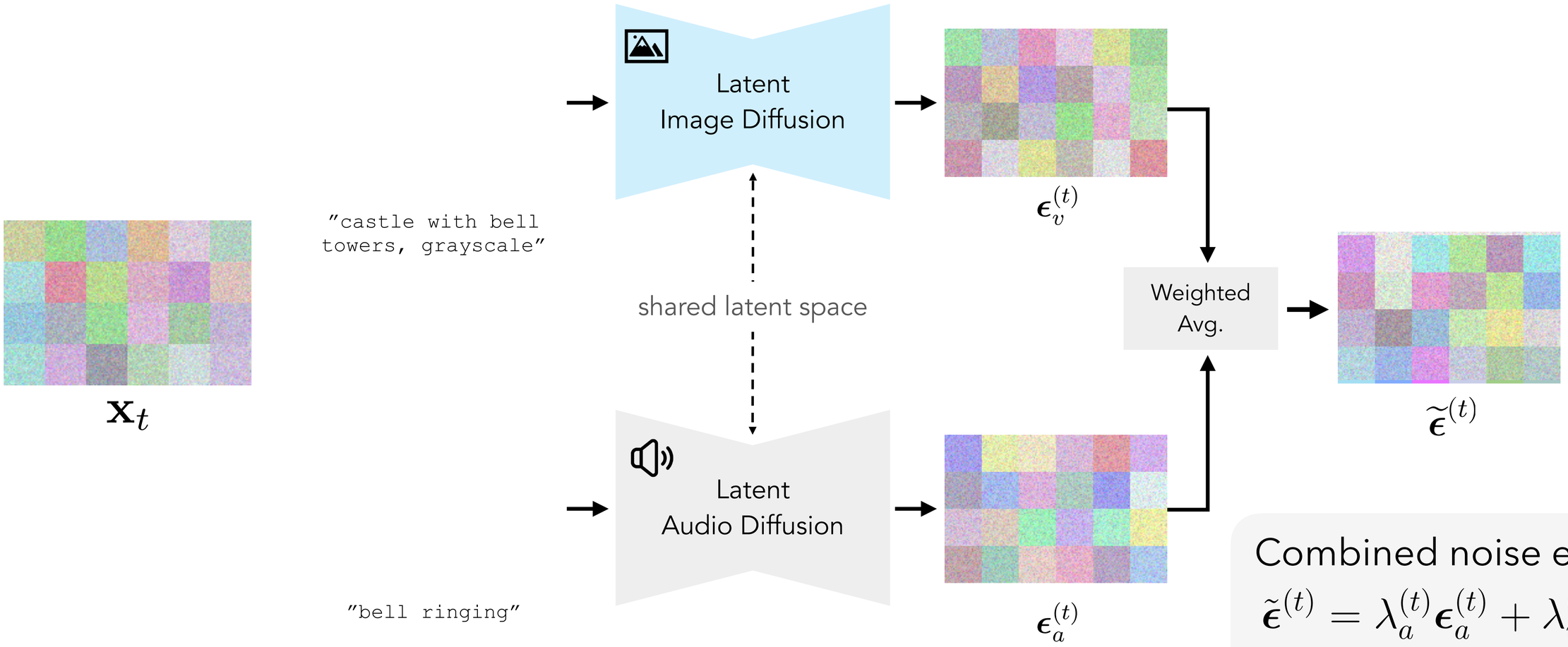


Sample from product of experts:

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

Related work on compositional generation: e.g.,
[Liu, Li, Du, et al. “Compositional Visual Generation.” ECCV 2022], [Geng et al., “Visual Anagrams”, 2024]

Composing sight with sound



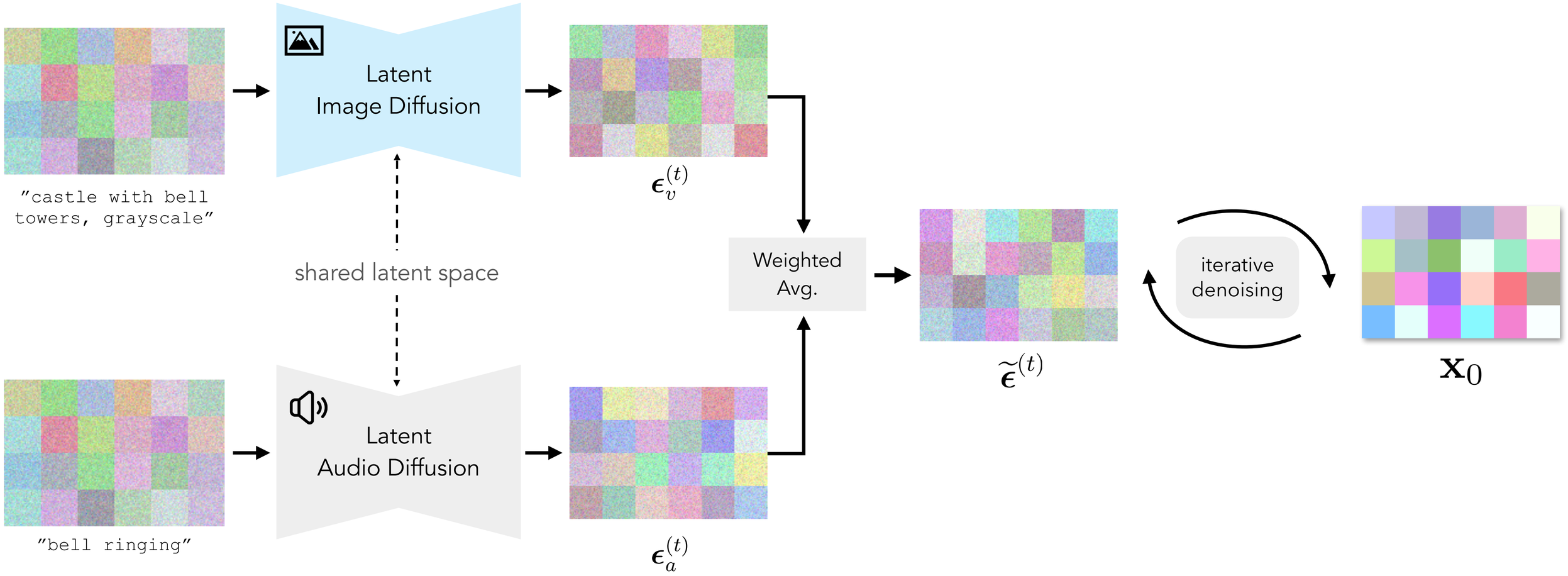
Combined noise estimate:

$$\tilde{\boldsymbol{\epsilon}}^{(t)} = \lambda_a^{(t)} \boldsymbol{\epsilon}_a^{(t)} + \lambda_v^{(t)} \boldsymbol{\epsilon}_v^{(t)}$$

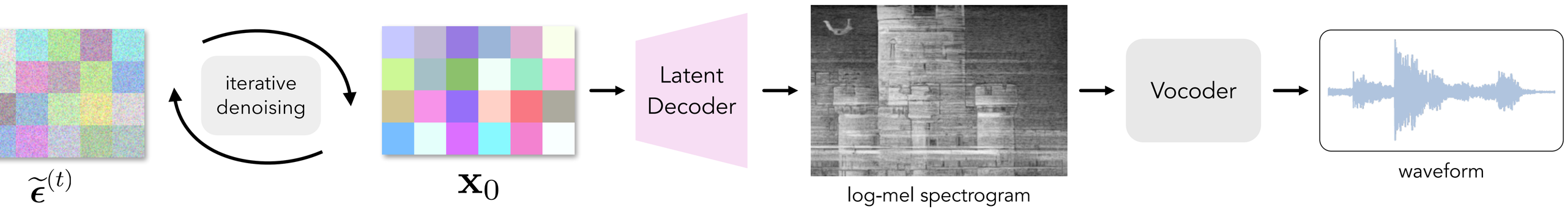
with time-varying weights: $\lambda^{(t)}$

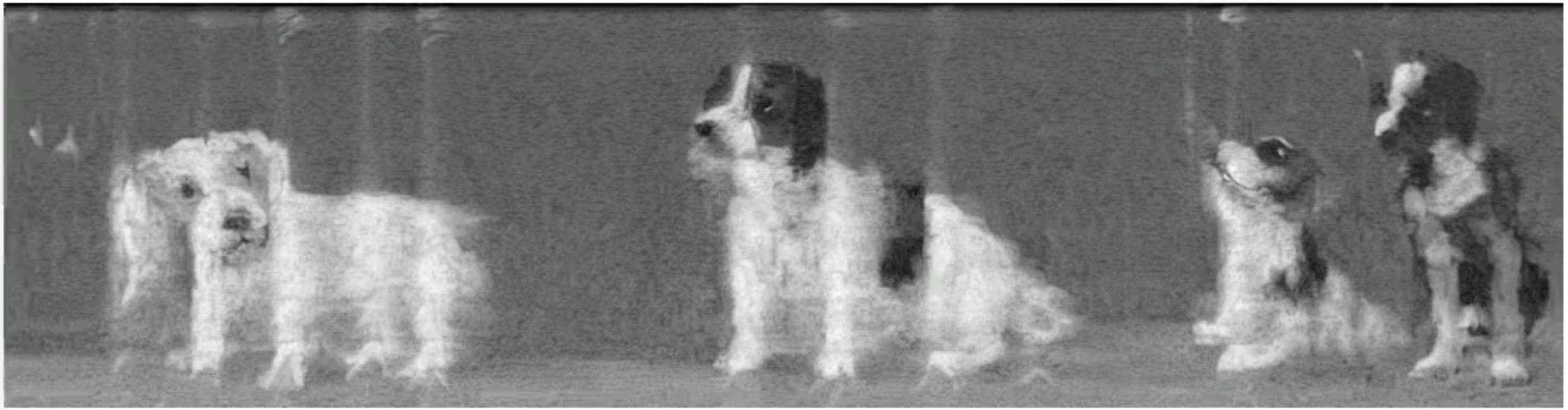
[Xue et al., "Auffusion", 2024], [Forsgren et al., "Riffusion", 2022]

Composing sight with sound



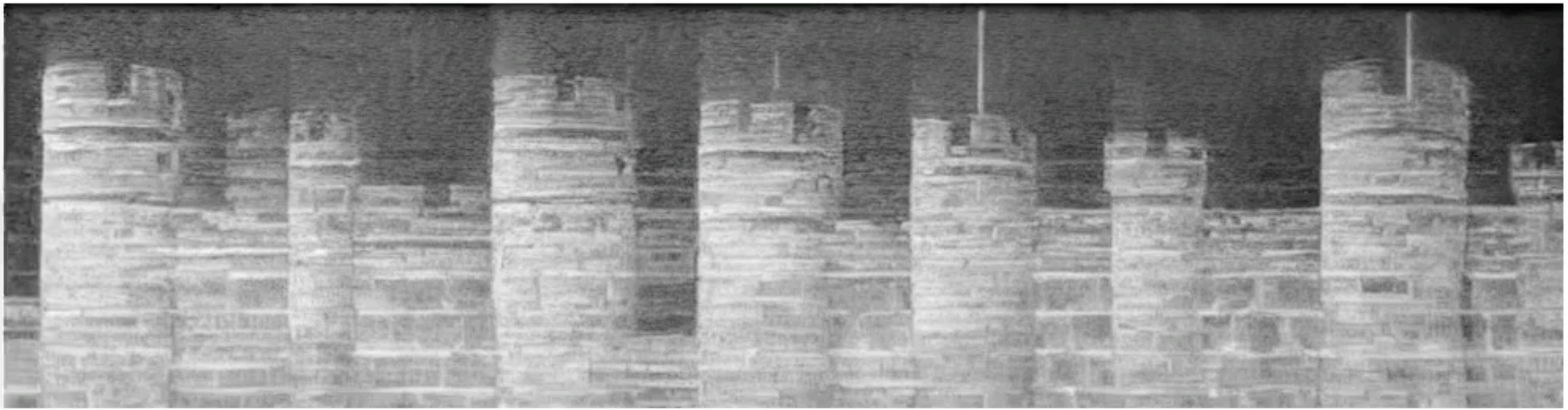
Composing sight with sound





“a painting of cute dogs, grayscale”

“dog barking”

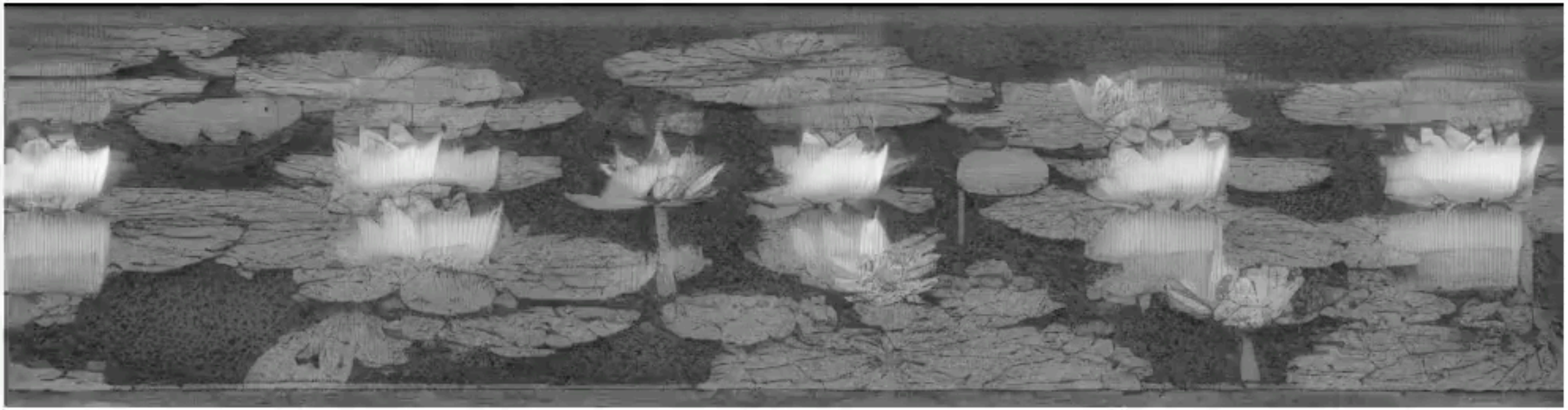


“a painting of castle towers, grayscale”

“bell ringing”

“a painting of auto racing game, grayscale”

“a race car passing by and disappearing”

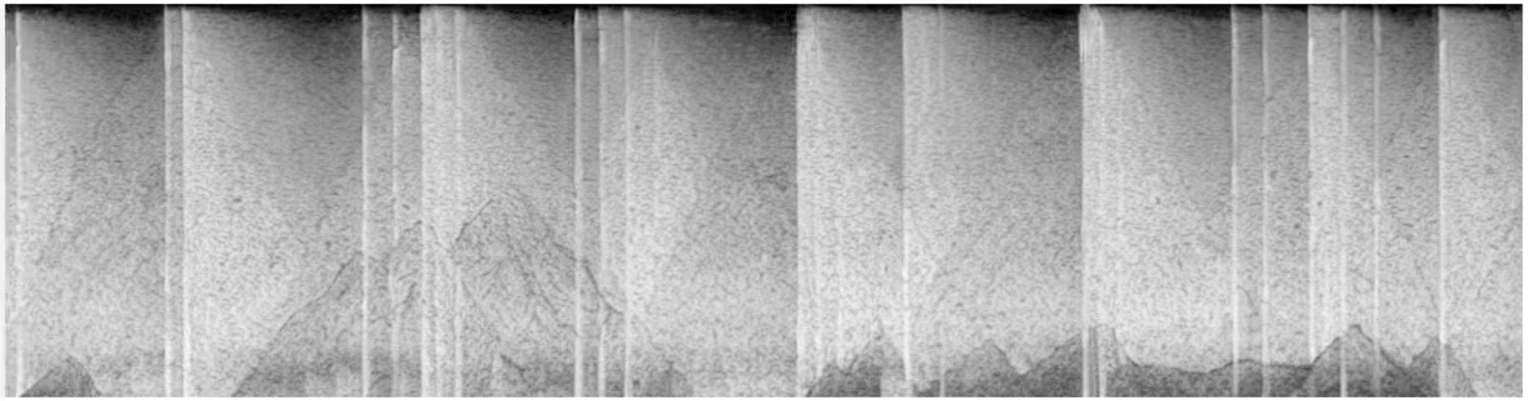


“frog croaking”

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

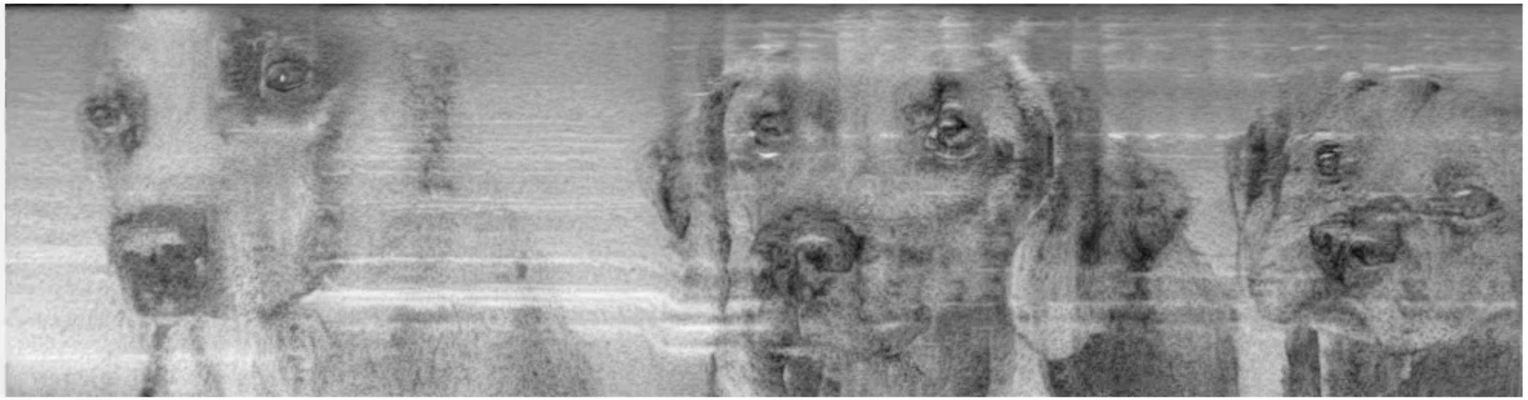
Failure Cases

Image prompt: a painting of mountains, grayscale



Audio prompt: fireworks banging

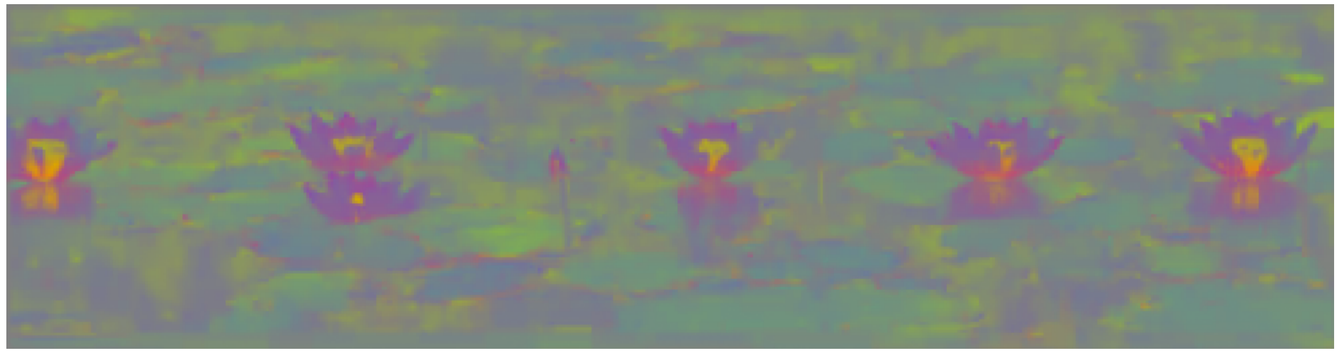
Image prompt: a painting of dogs, grayscale



Audio prompt: train whistling

Color images?

Factorized diffusion



“a colorful photo of a water-lily pond”

[Factorized Diffusion: Perceptual Illusions by Noise Decomposition, ECCV 2024]

Daniel Geng*

Aaron Park*

Image prompt: a colorful photo of tigers

Audio prompt: tiger growling

Beyond images in, audio out

Multimodal conditioning

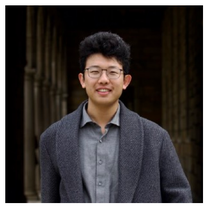
"rooster crowing"

3D scene interactions

Images that sound

Thank you!


Ziyang Chen


Yiming Dou


Daniel Geng


Wonseok Oh


Yuqing Luo

Plus Antonio Loquercio, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, and Justin Salamon