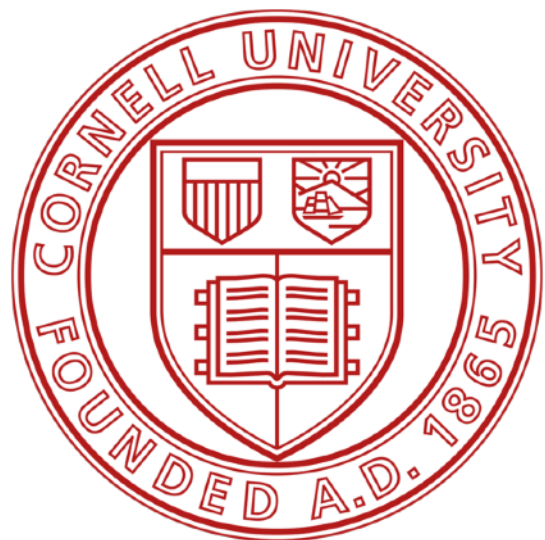
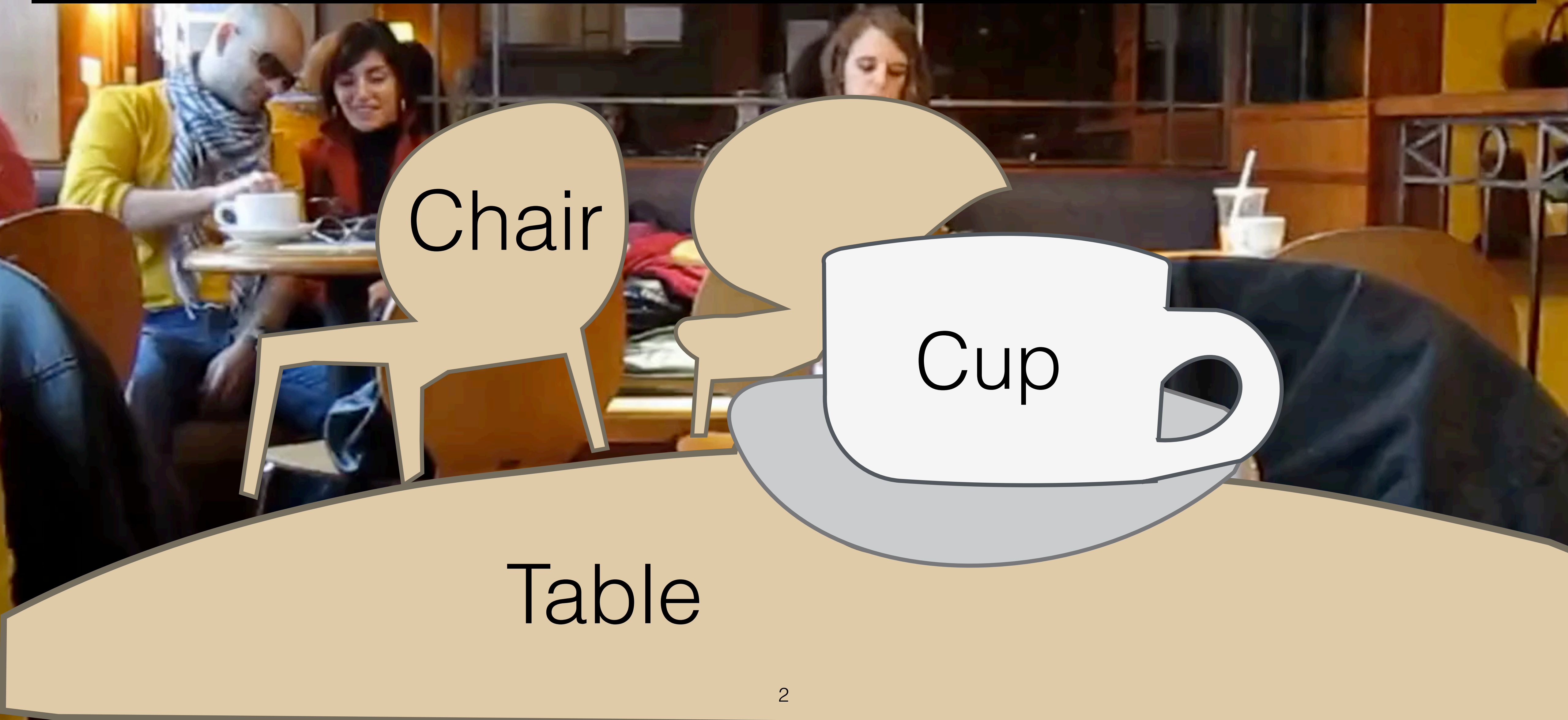


Lecture 26: Sight and sound

CS 5670: Introduction to Computer Vision



“A coffee cup resting on a table in a cafe”









(McGurk & McDonald 1976)

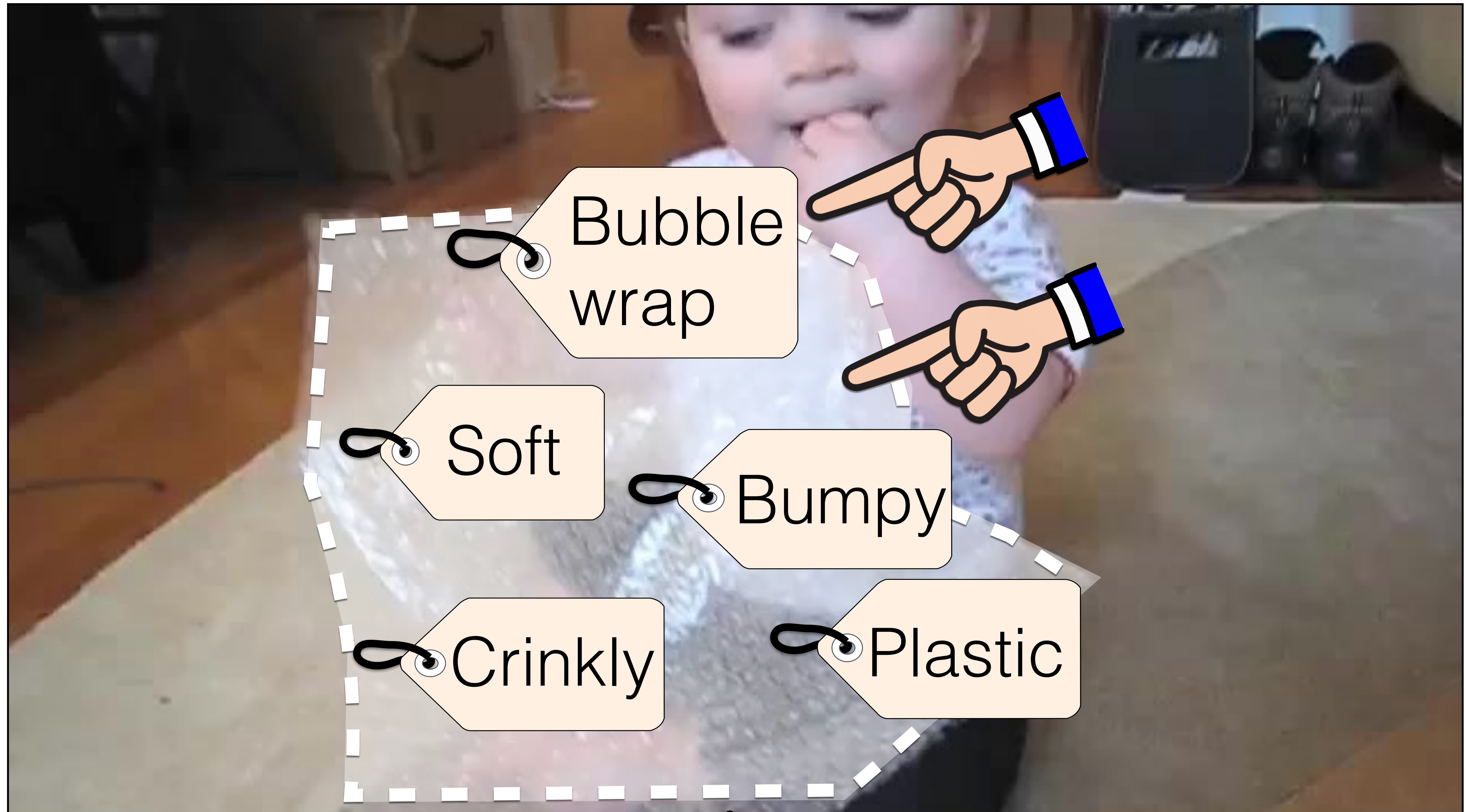




(McGurk & McDonald 1976)



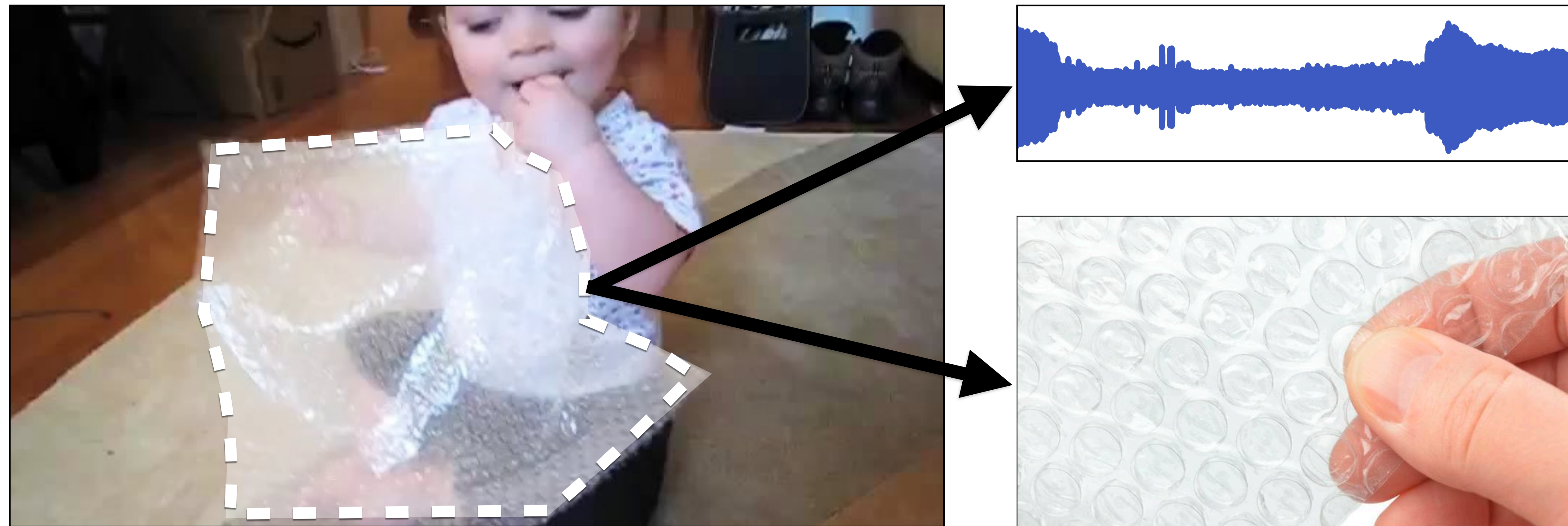
Supervised learning



Human supervision



Self-supervision

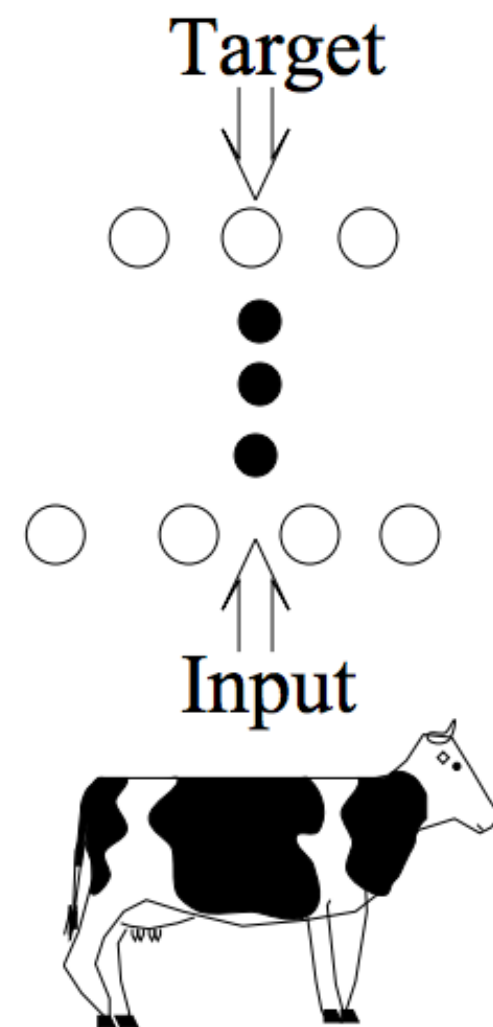


Multisensory self-supervision

Supervised

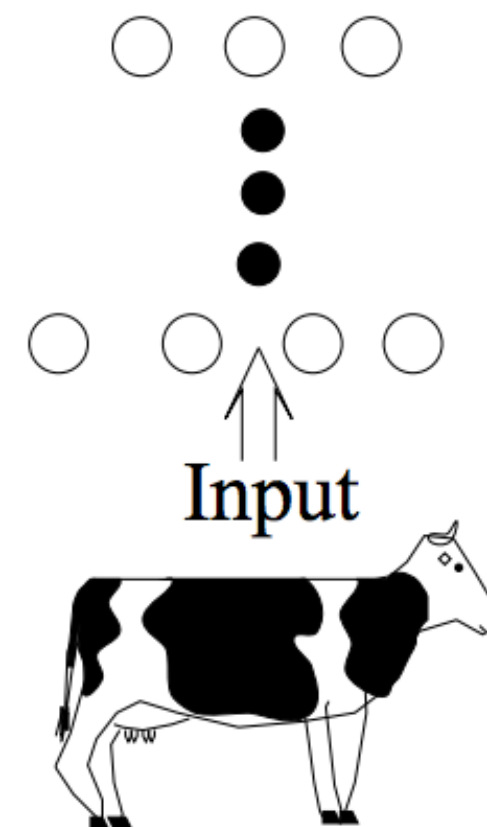
- implausible label

"COW"



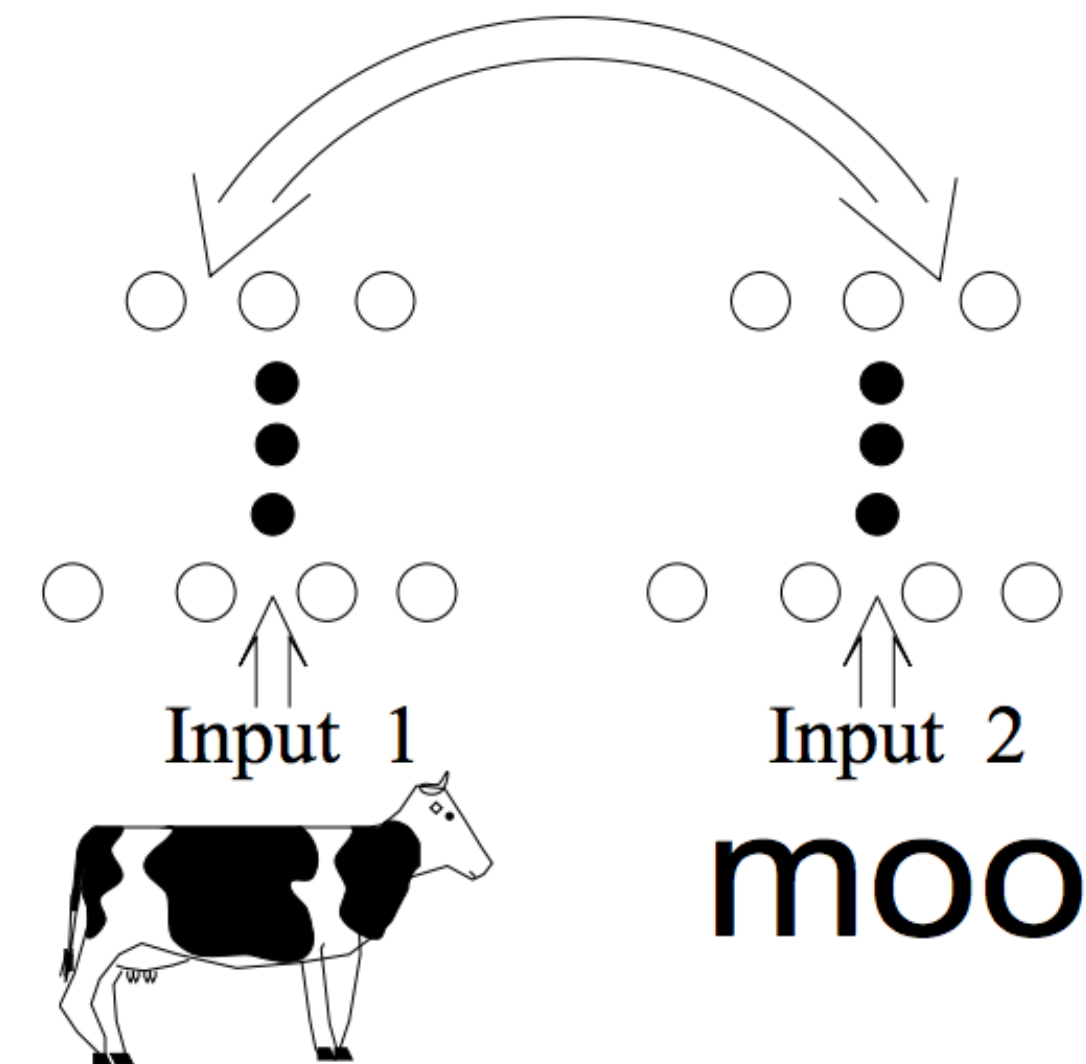
Unsupervised

- limited power



Self-Supervised

- derives label from a co-occurring input to another modality

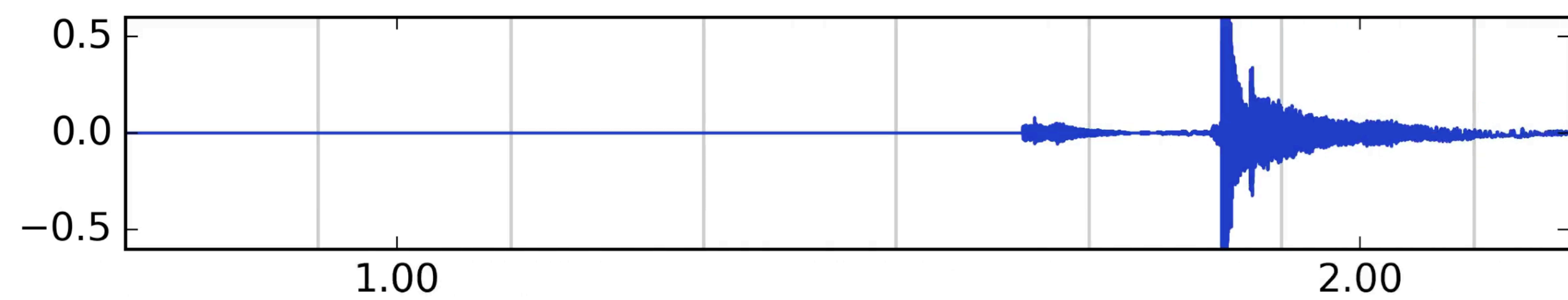


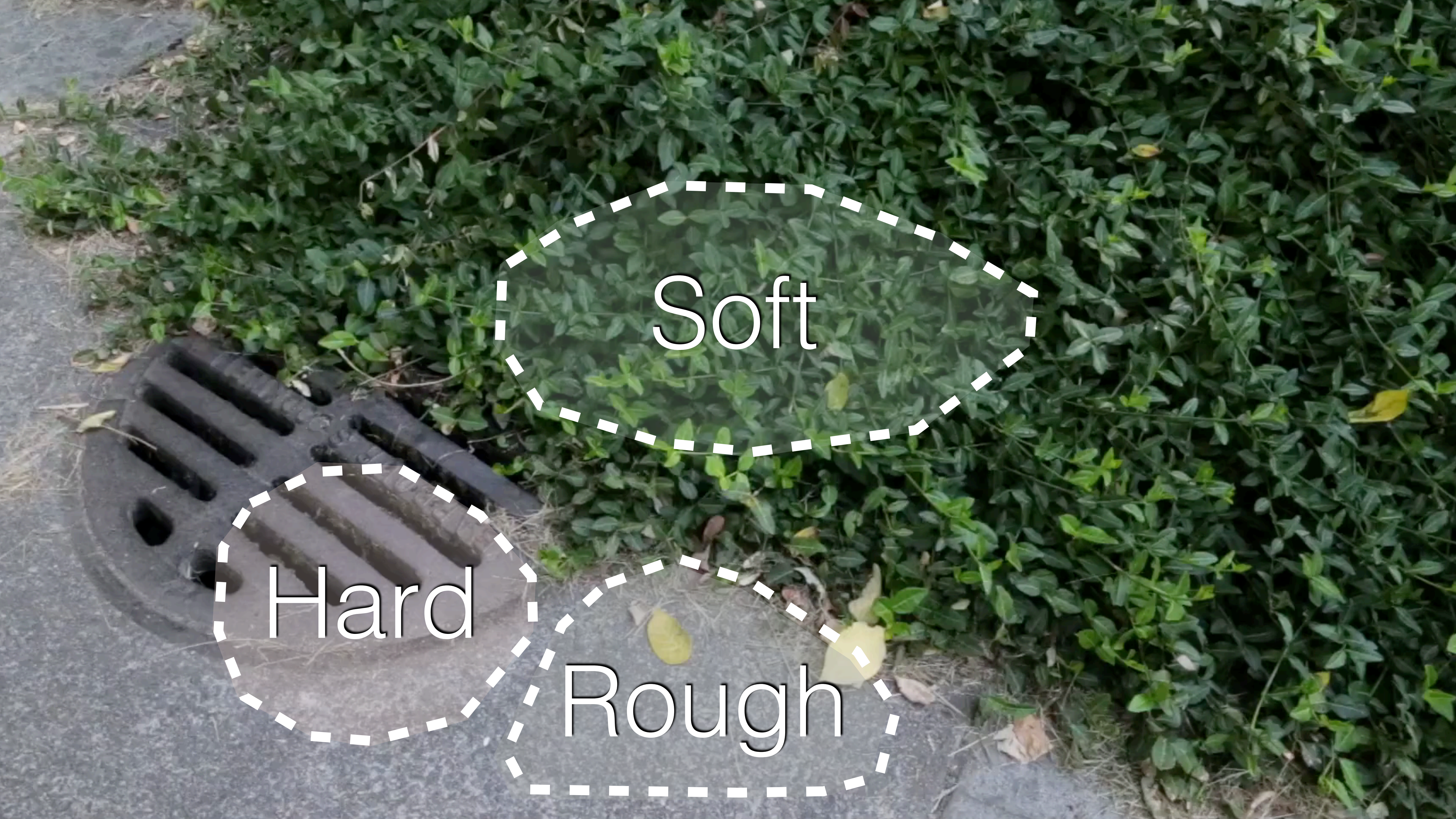
Multimodal models

- Have access to “orthogonal” sources of information.
- Can exploit cues in from multiple senses
- Can use one modality to train another, using self-supervision
- Can be more robust to sensor failures

Multimodal models

- Have access to “orthogonal” sources of information.
- Can exploit cues in from multiple senses
- **Can use one modality to train another, using self-supervision**
- Can be more robust to sensor failures





Soft

Hard

Rough

Learning about physical interactions

• Silent video

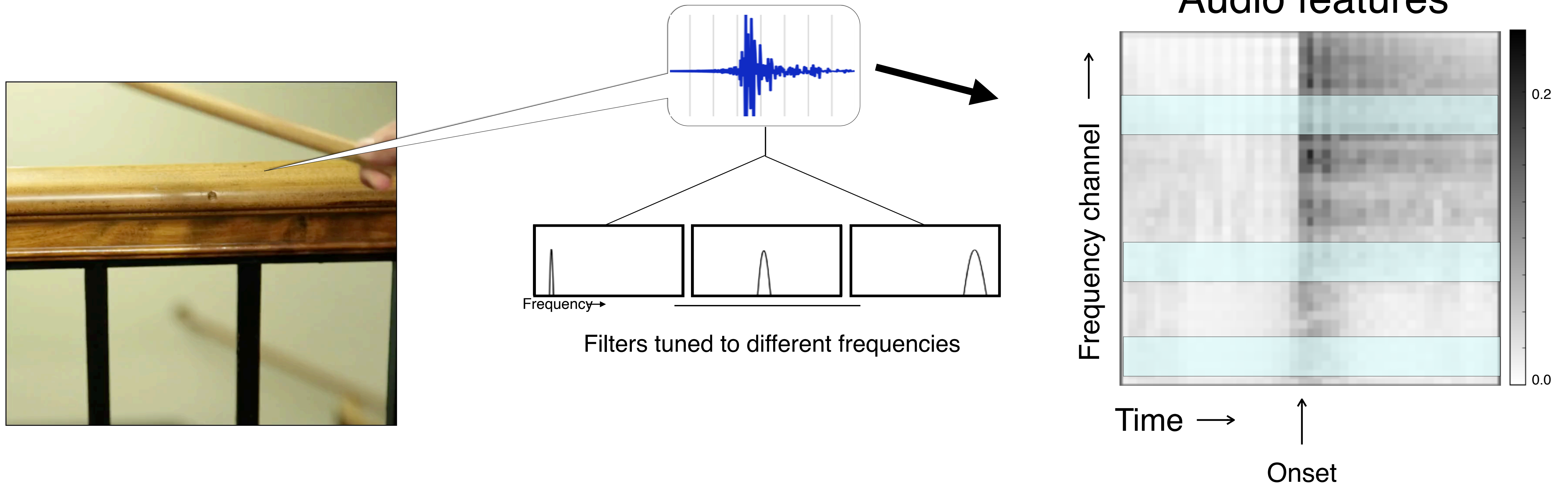
• Predicted soundtrack





The *Greatest Hits* Dataset: Volume 1

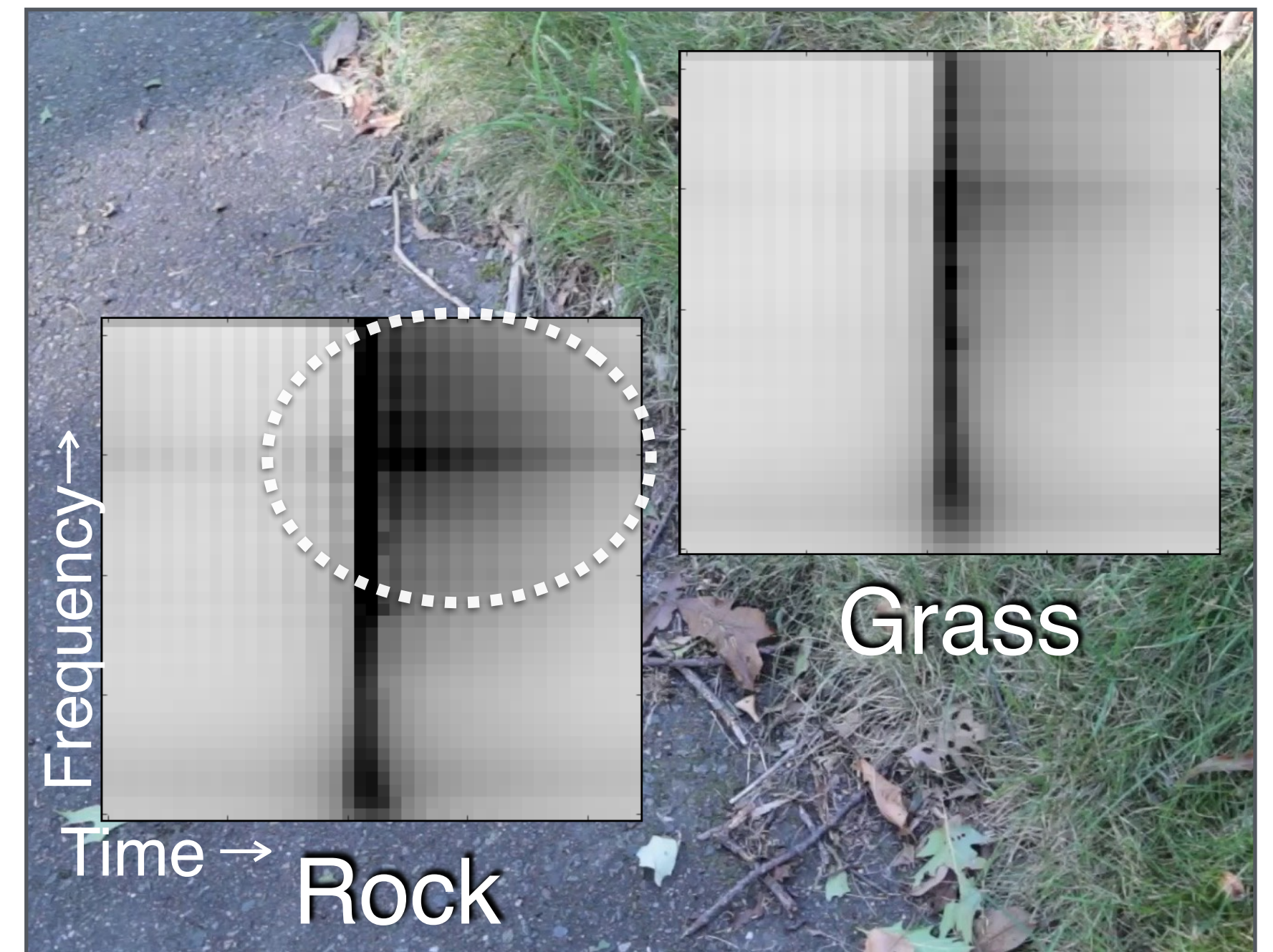
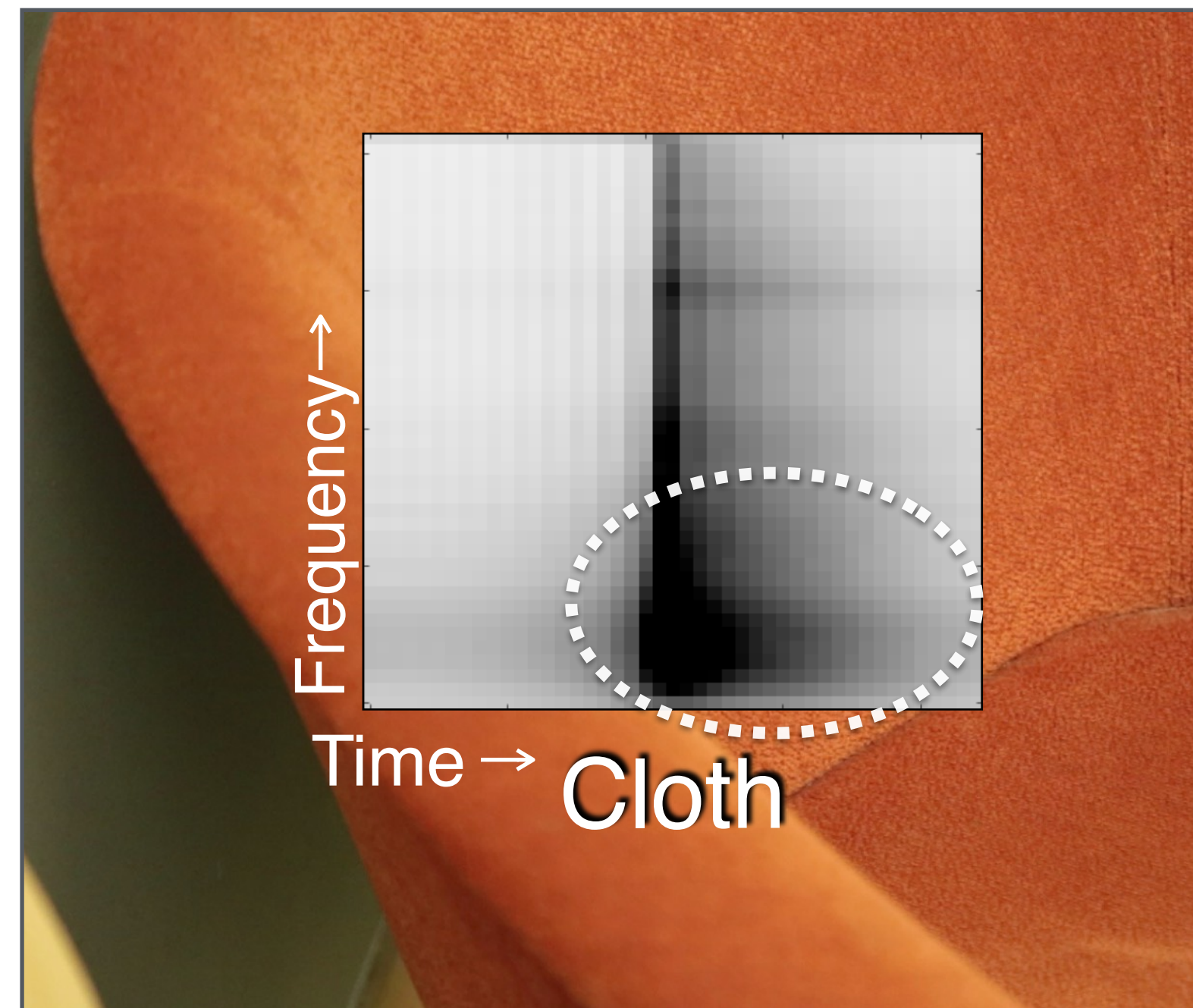
Can we predict physical properties from sound?



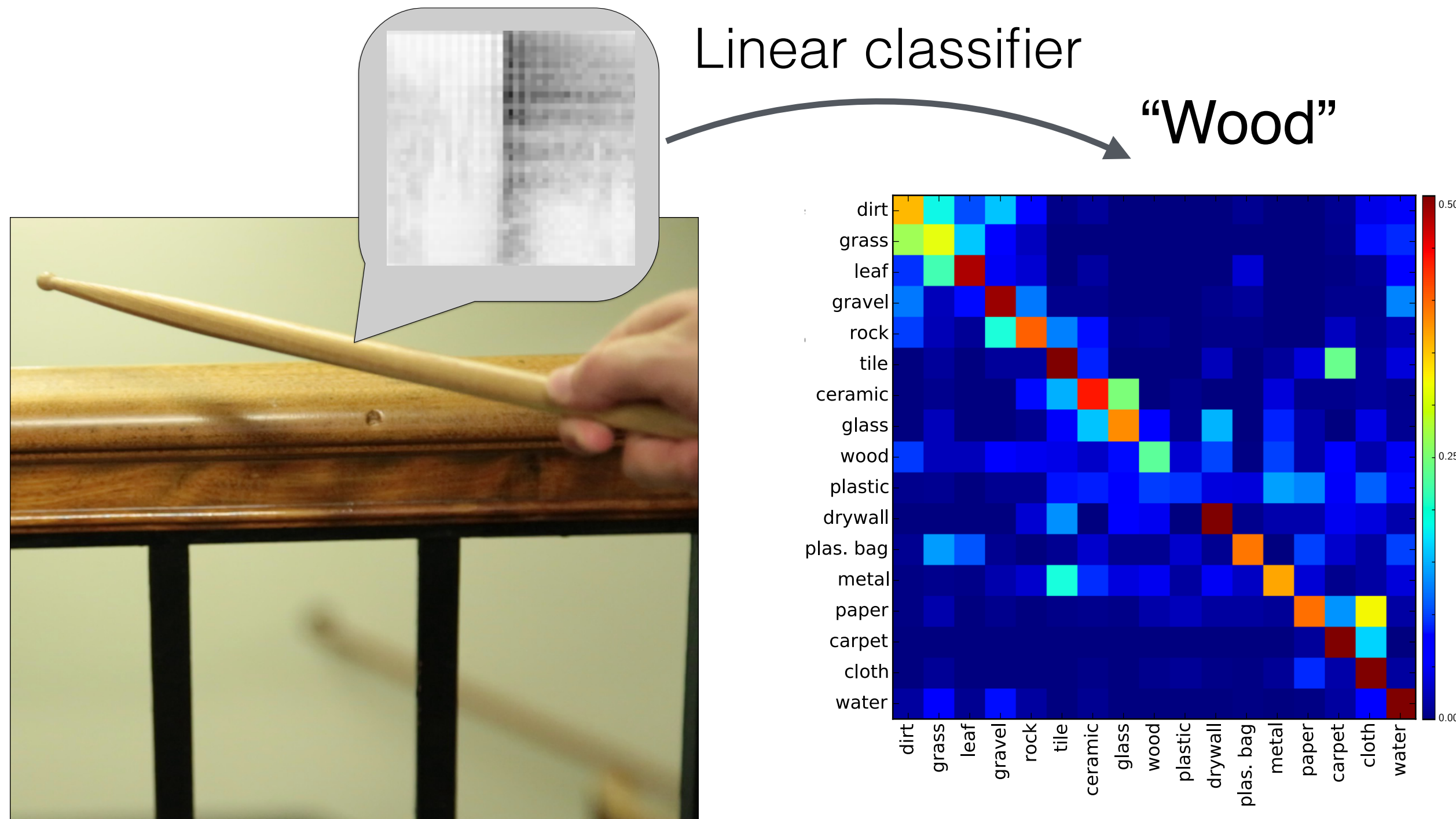
- 40 bandpass filters (+ high/low pass)

Can we predict physical properties from sound?

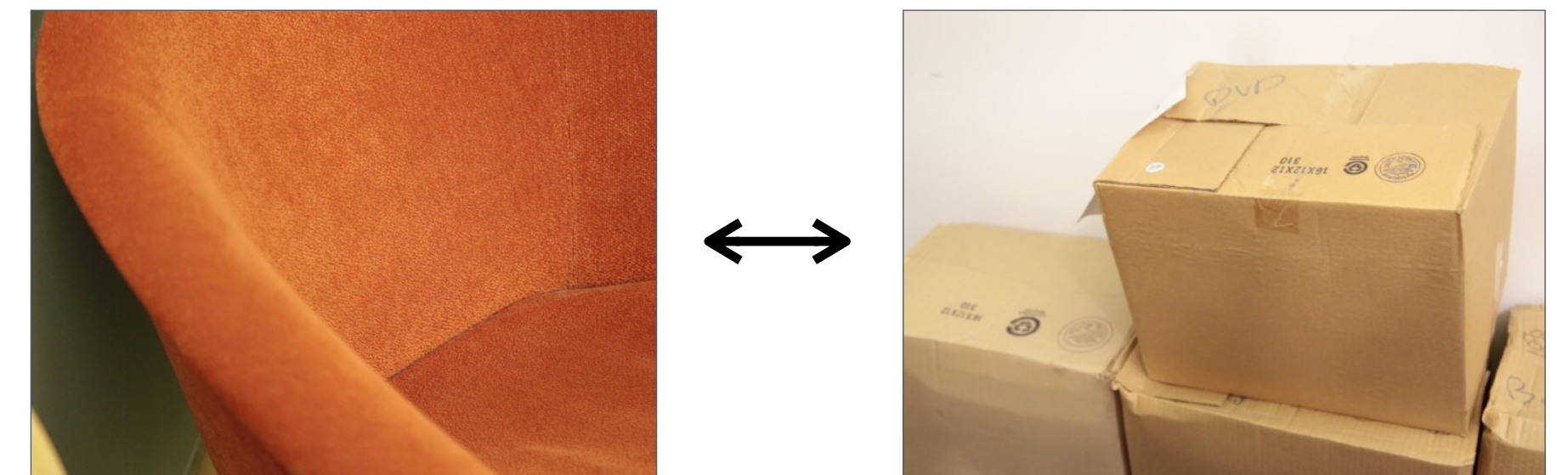
Mean sound features per category



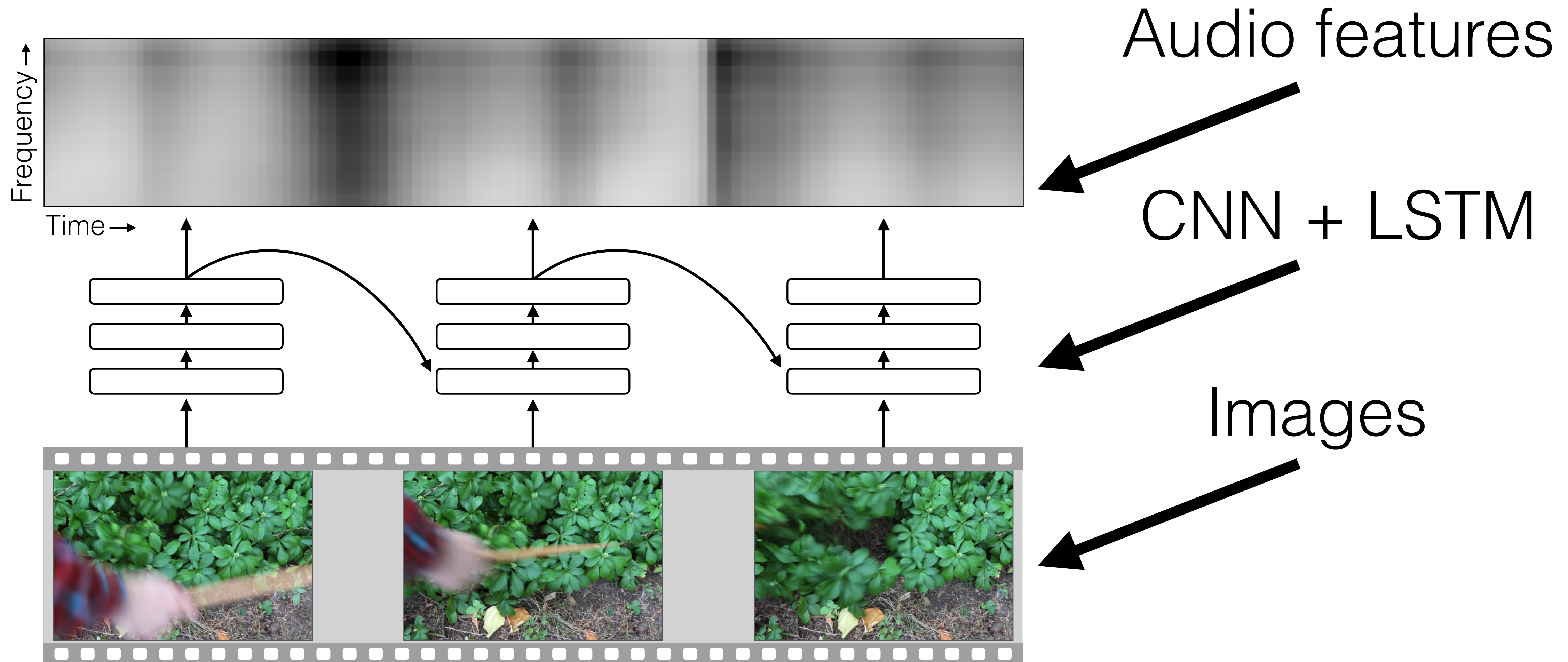
Can we predict physical properties from sound?



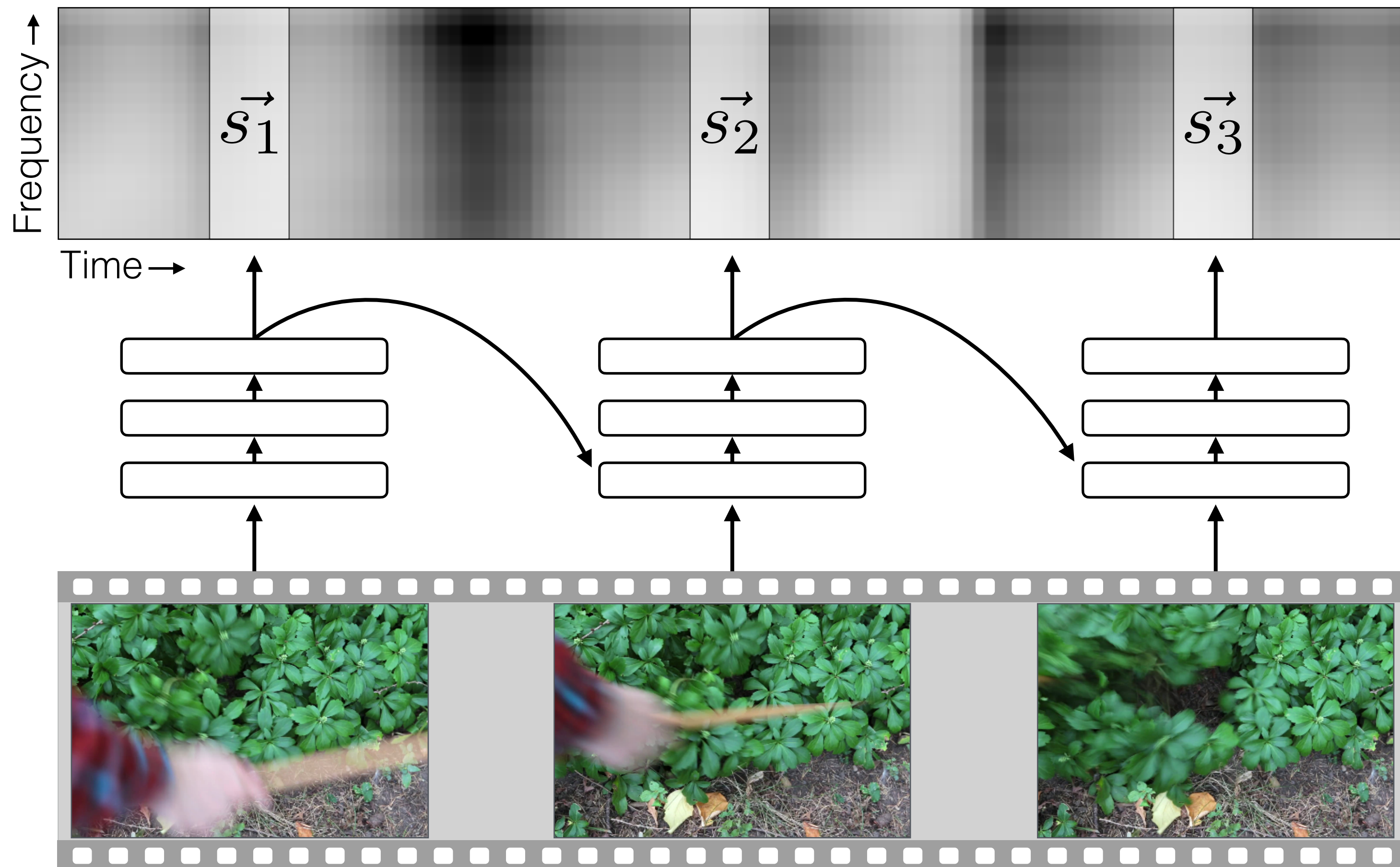
- 46% class-averaged accuracy (chance = 6%)
- Common audio confusions:
 - {cloth, paper}, {dirt, grass}, {rock, tile}



Predicting visually indicated sounds



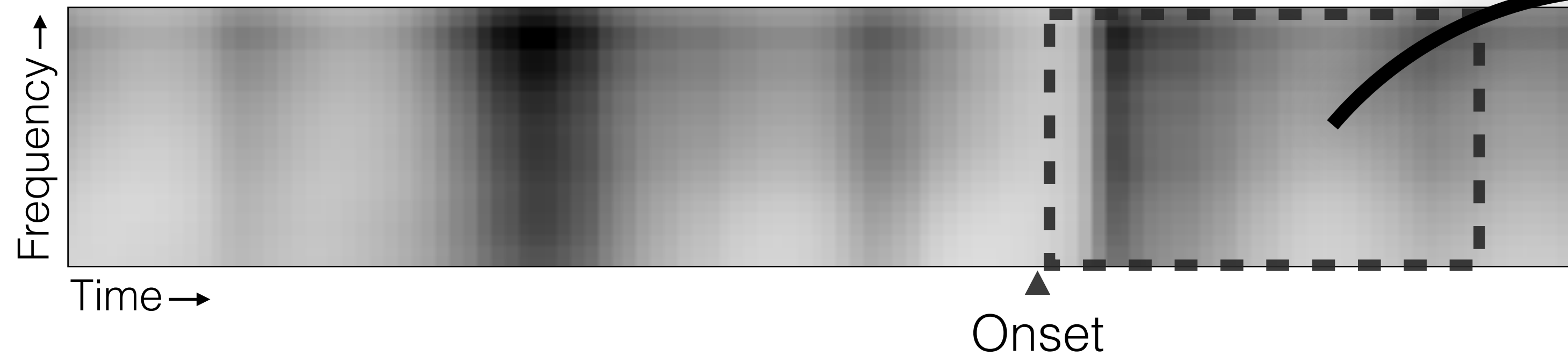
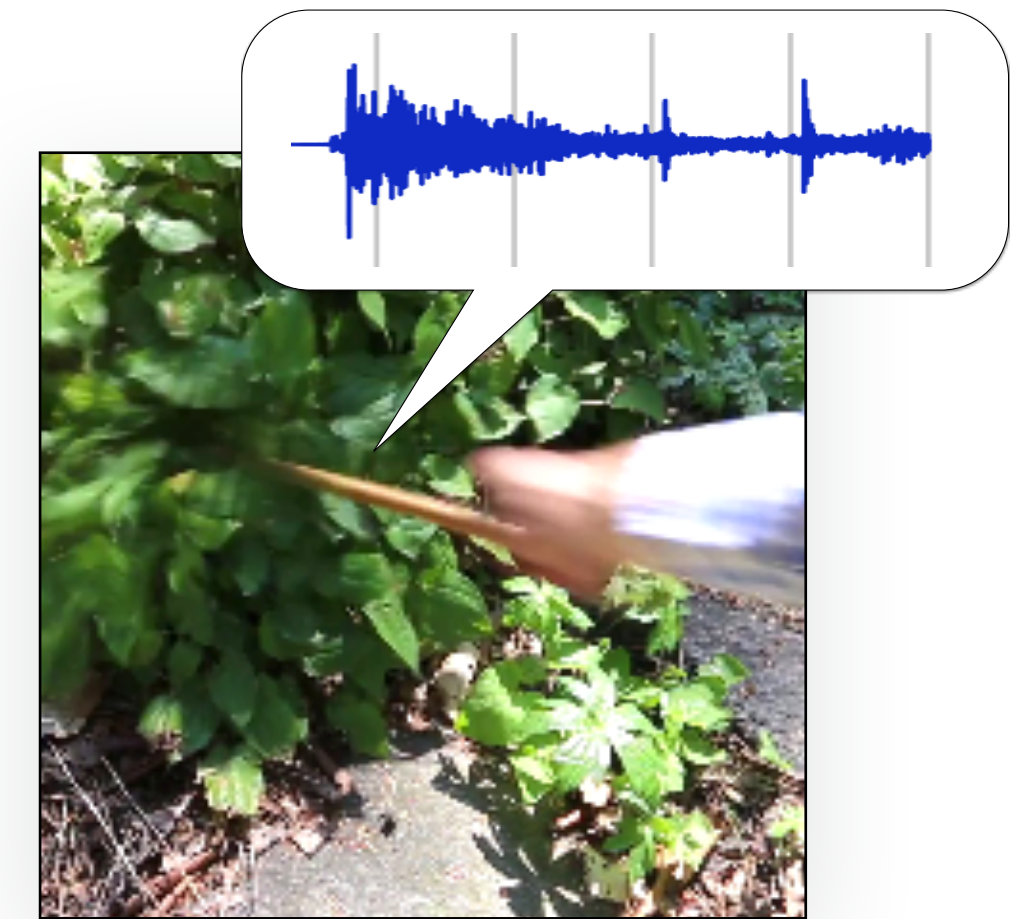
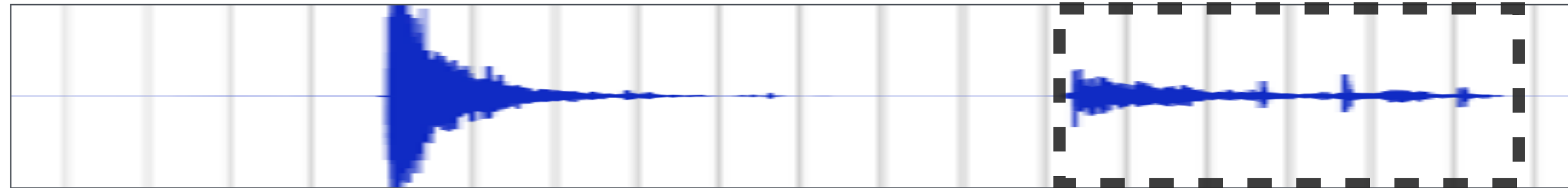
Predicting visually indicated sounds



Ground truth

$$\sum_{t=1}^T \|\vec{s}_t - \tilde{\vec{s}}_t\|$$

Predicting visually indicated sounds

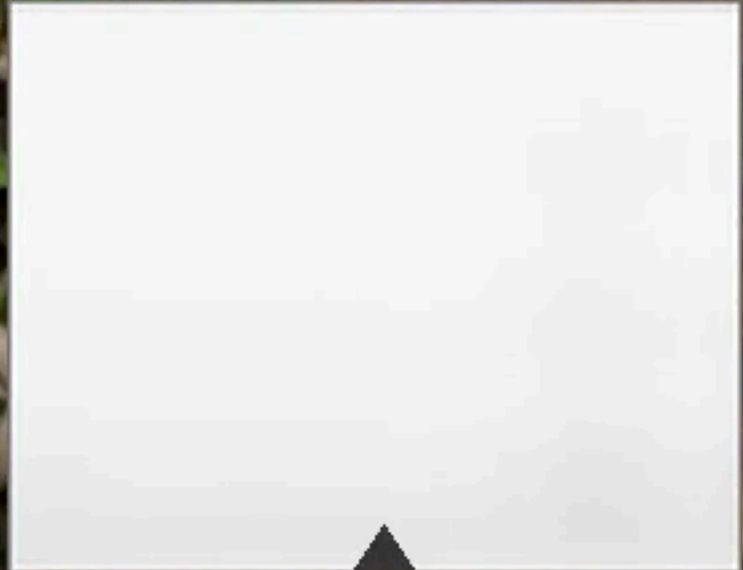


Sound database



27

Predicted sound

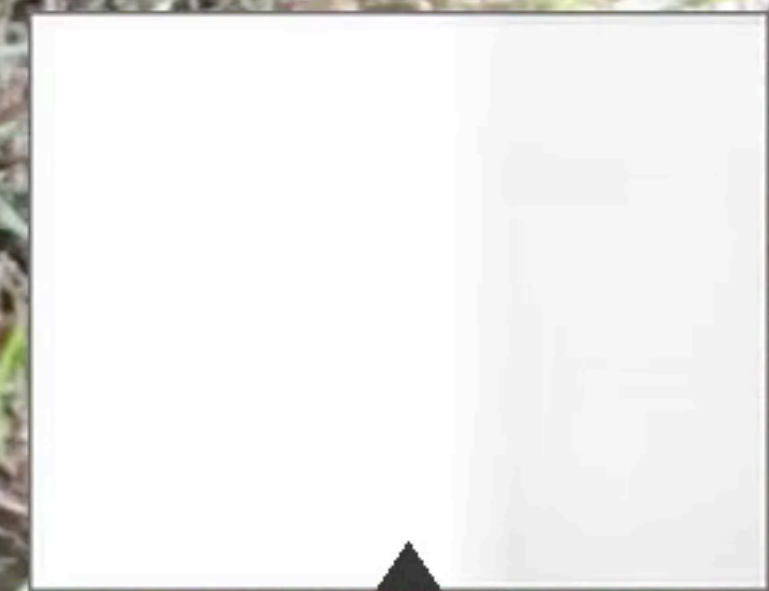




Predicted sound

28

X





Predicted sound

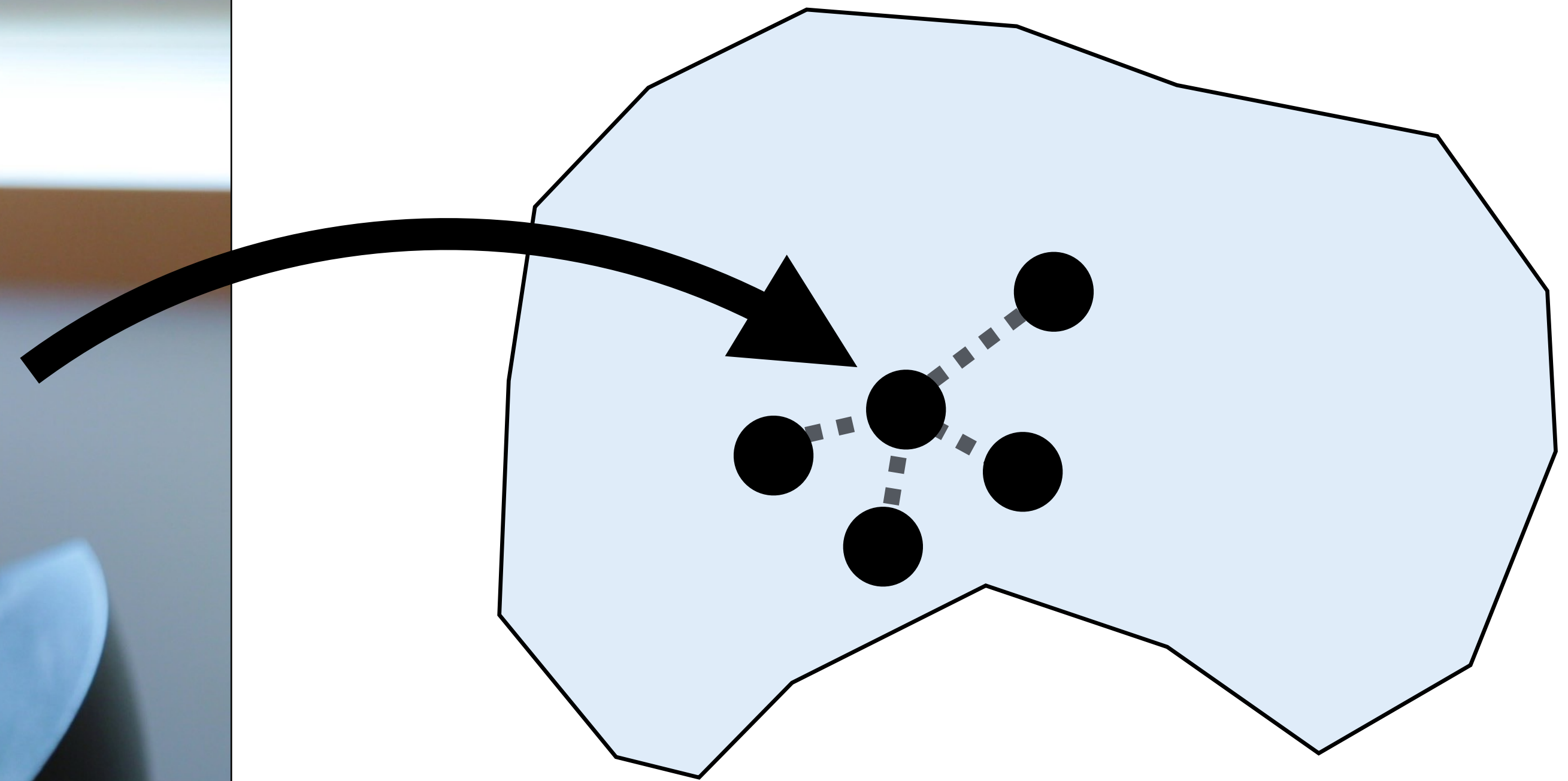


30

Failure case



Translating between modalities



Audio space

Translating between modalities



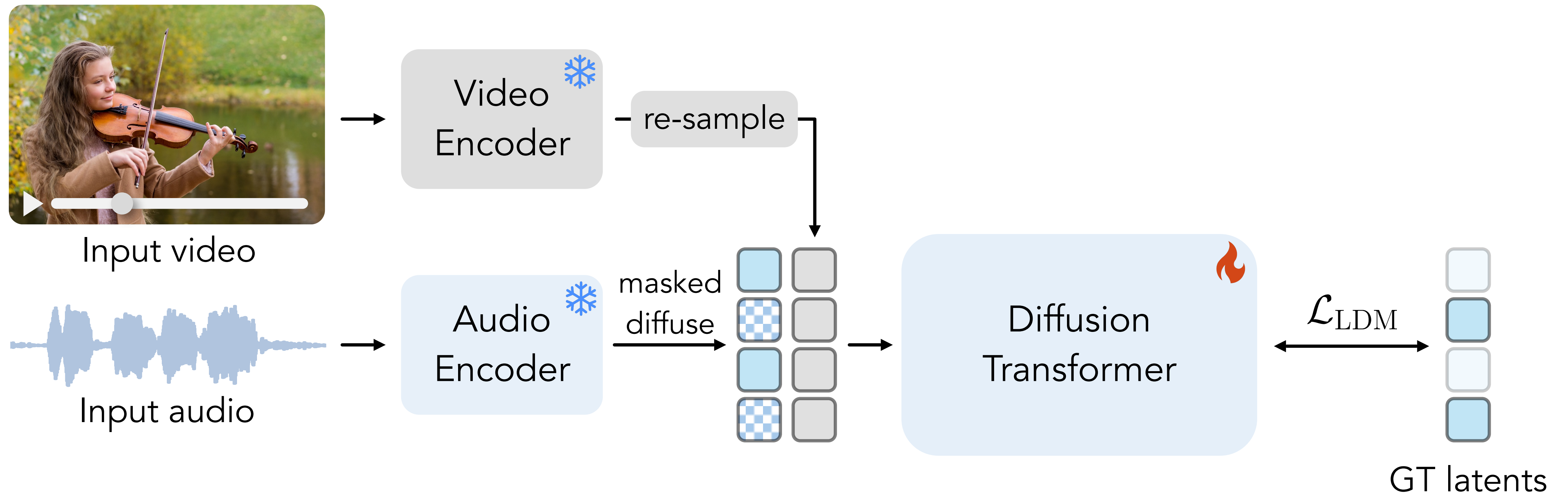
Audio nearest neighbors have similar material properties

Why do we need user control?



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

Multimodal control for video-to-audio generation



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

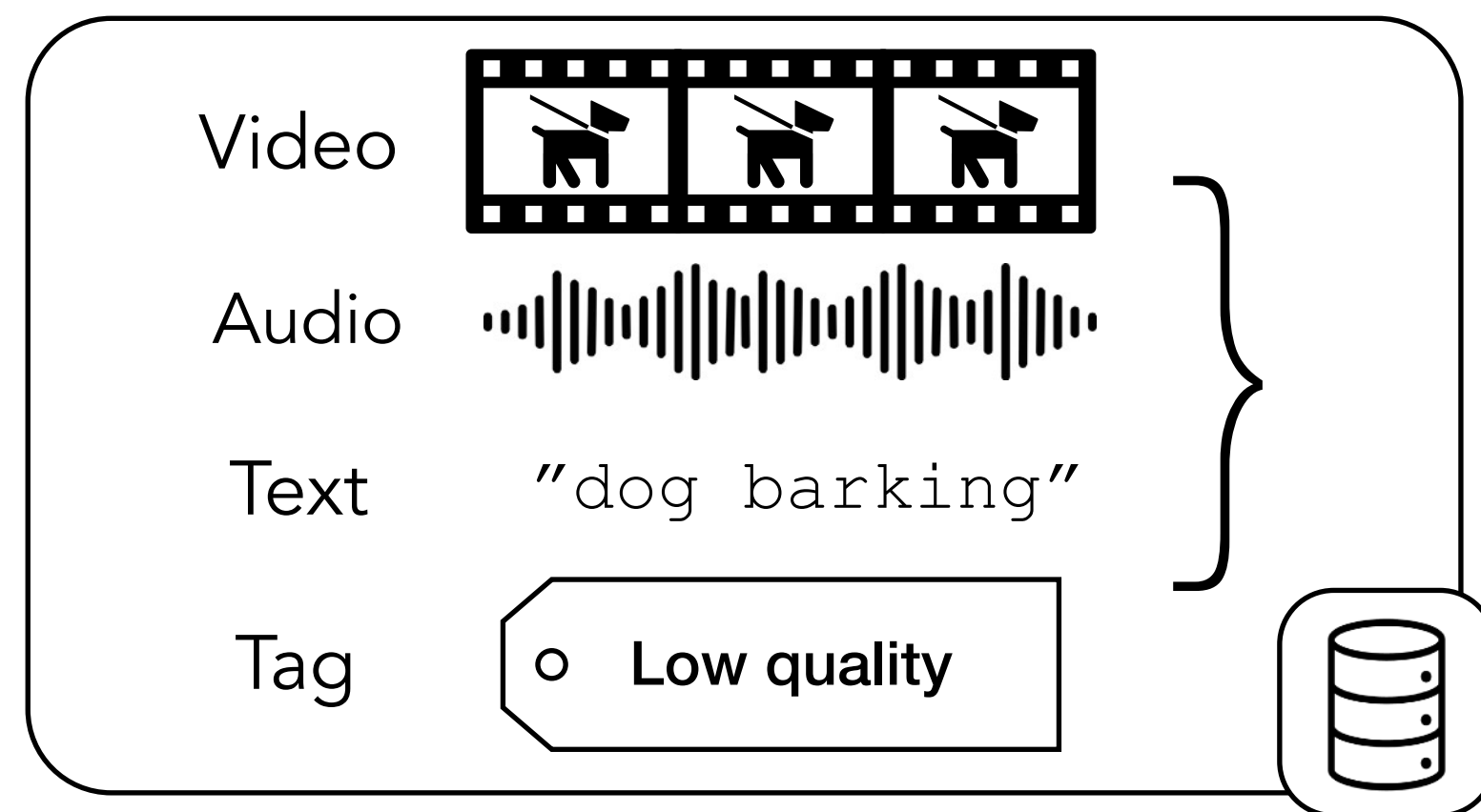


Ziyang Chen



Prem Seetharaman

Training with two data sources

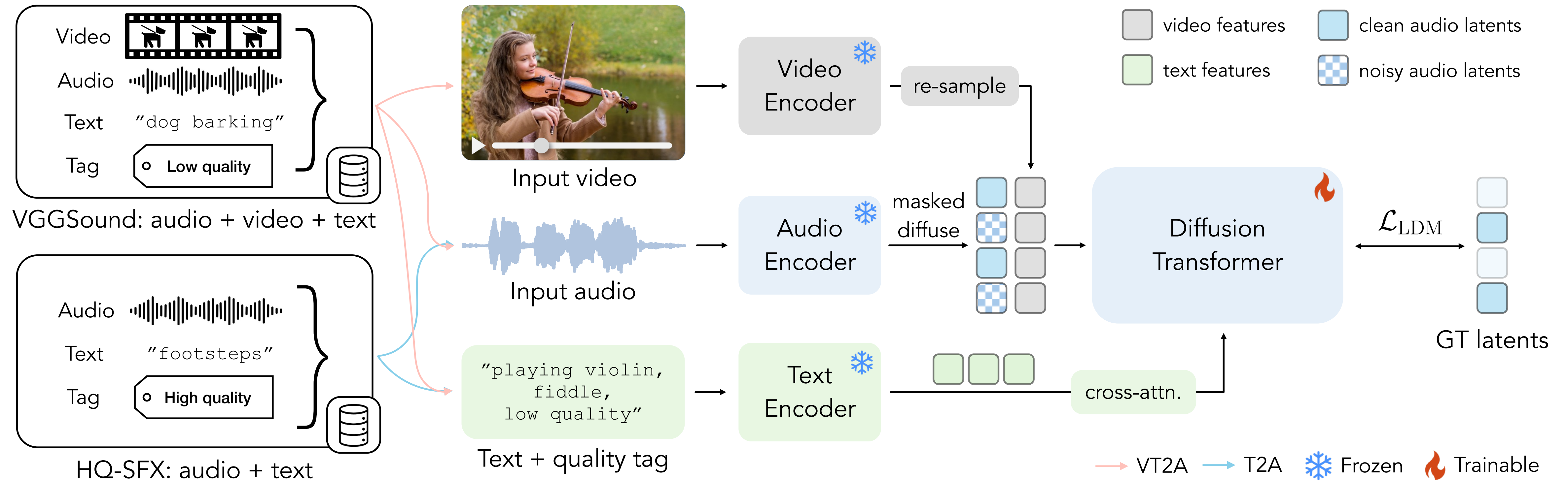


VGGSound: audio + video + text

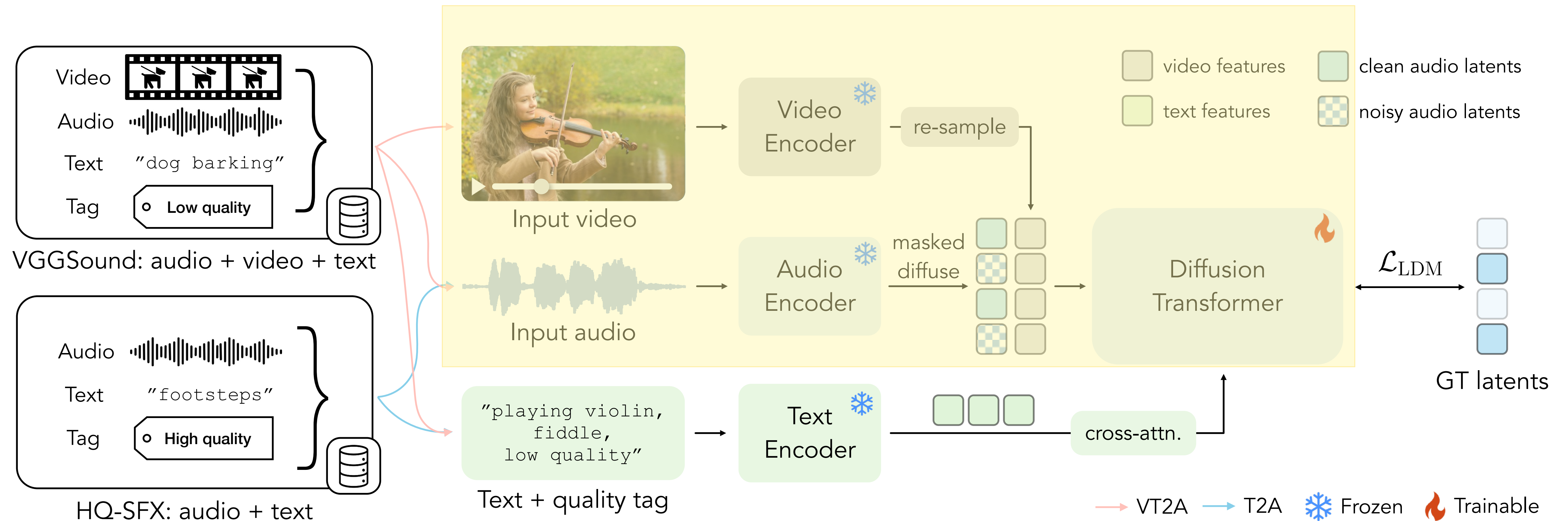


HQ-SFX: audio + text

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

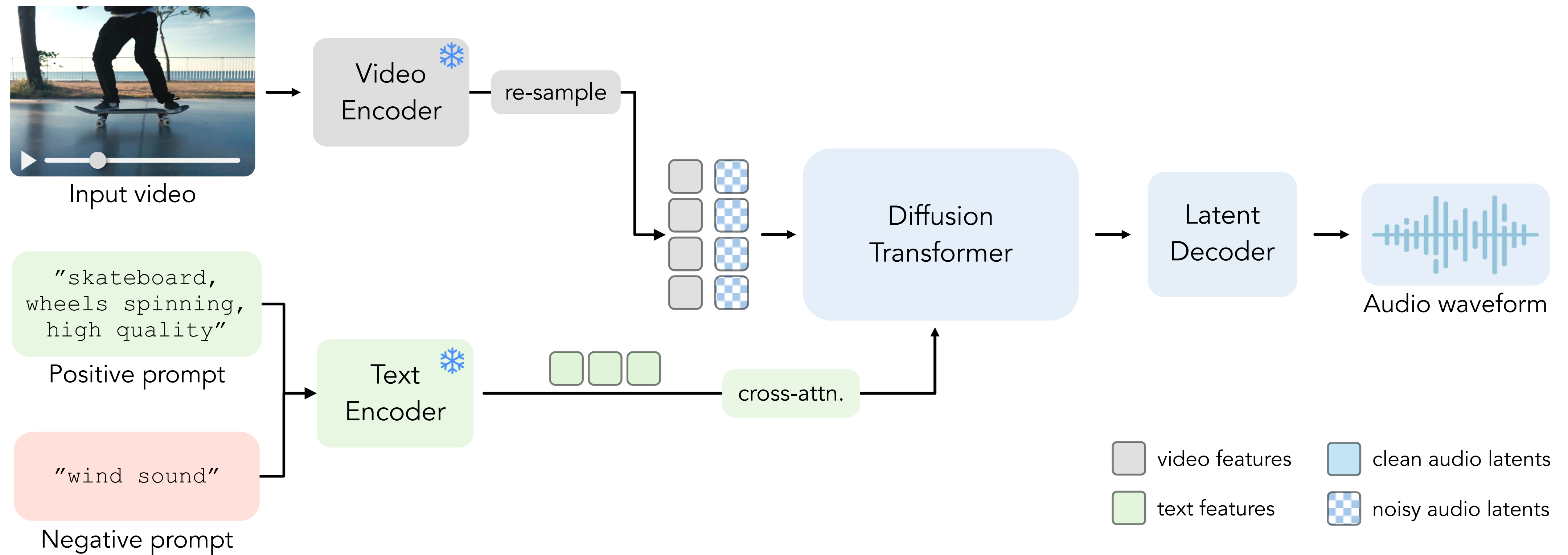
Given this reference drum audio



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”

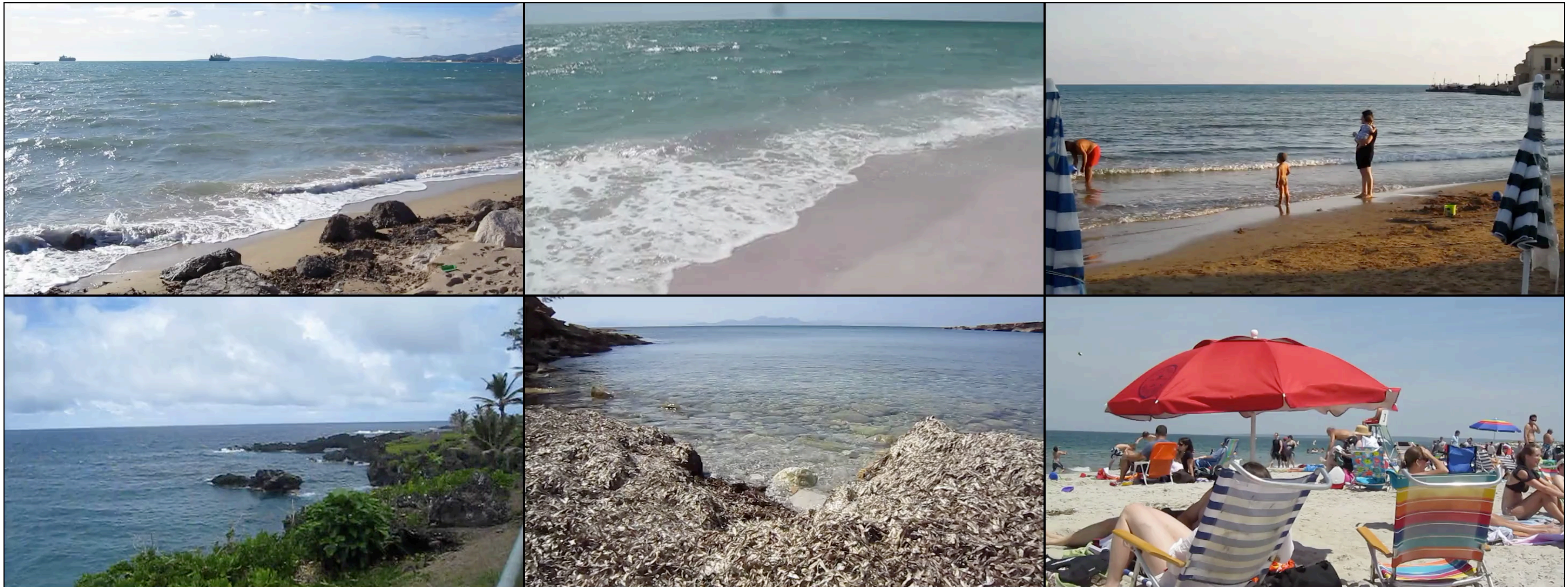
Joint audio and visual generation (Veo 3)



Joint audio and visual generation (Vevo 3)



Active vs. passive perception



A. Owens, J. Wu, J. H. McDermott, A. Torralba, W. T. Freeman.
Ambient Sound Provides Supervision for Visual Learning. ECCV 2016.

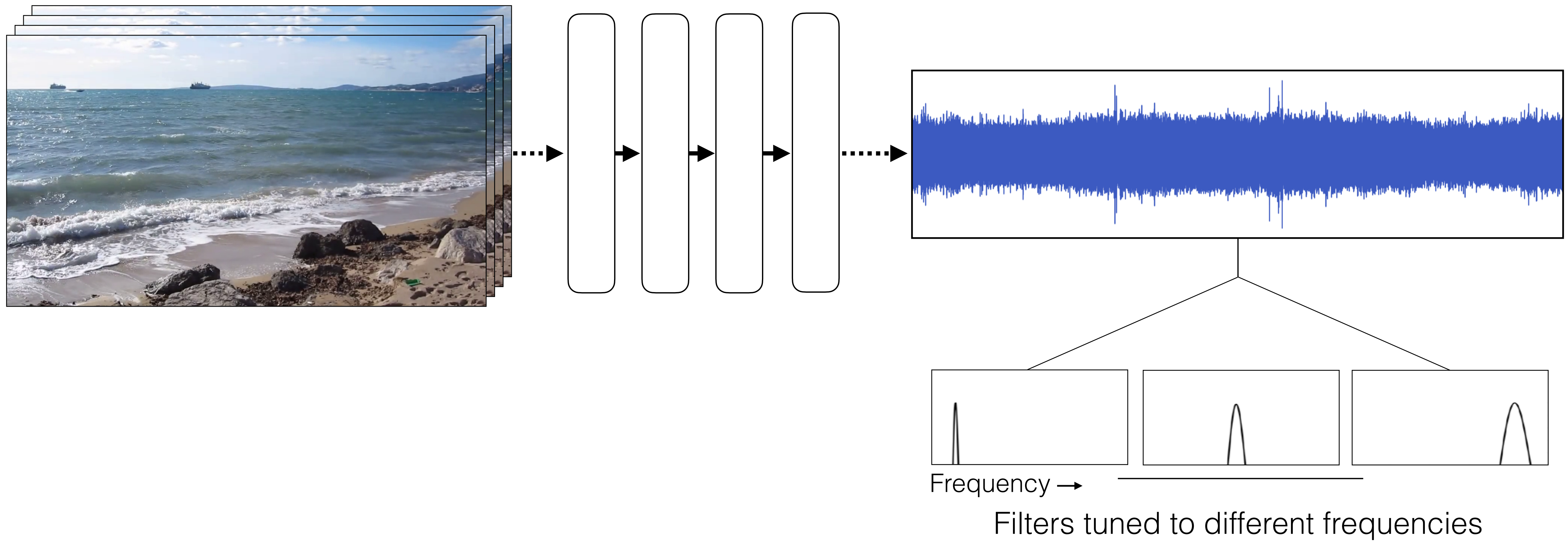
Active vs. passive perception



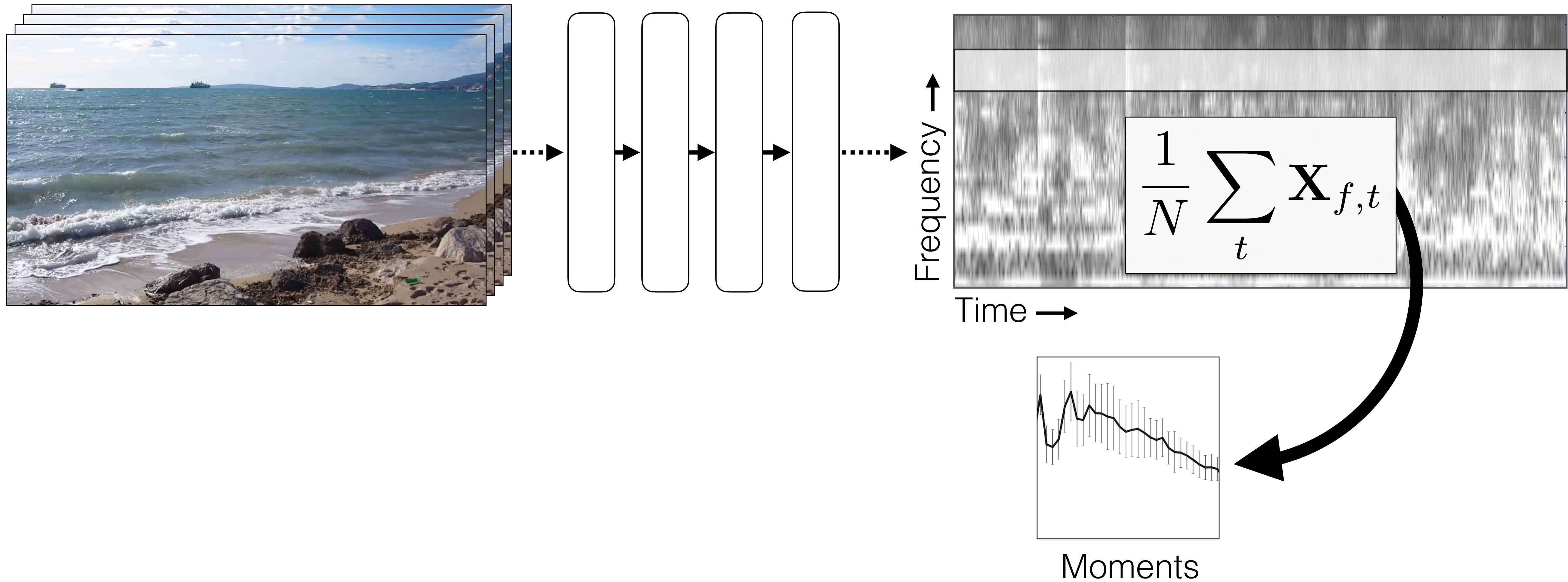
59

A. Owens, J. Wu, J. H. McDermott, A. Torralba, W. T. Freeman.
Ambient Sound Provides Supervision for Visual Learning. ECCV 2016.

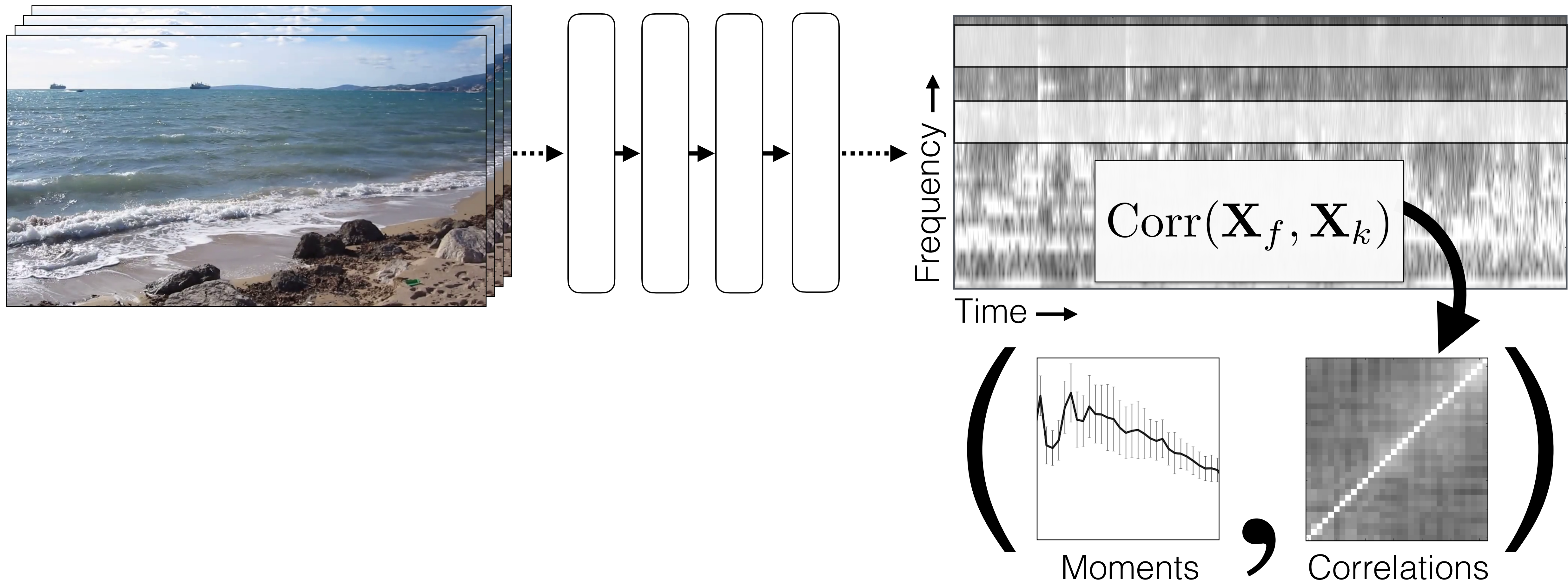
Predicting ambient sound



Predicting ambient sound



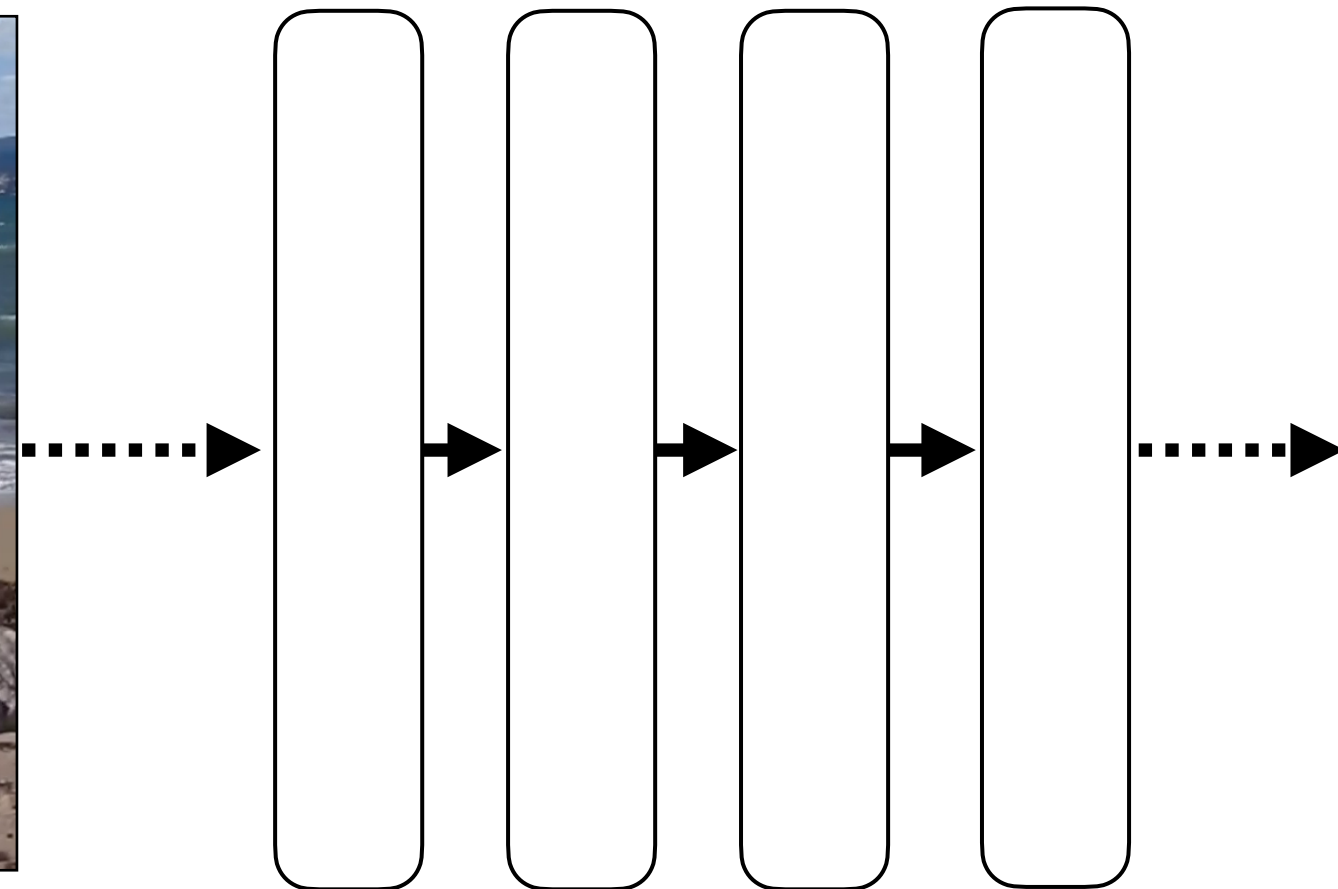
Predicting ambient sound



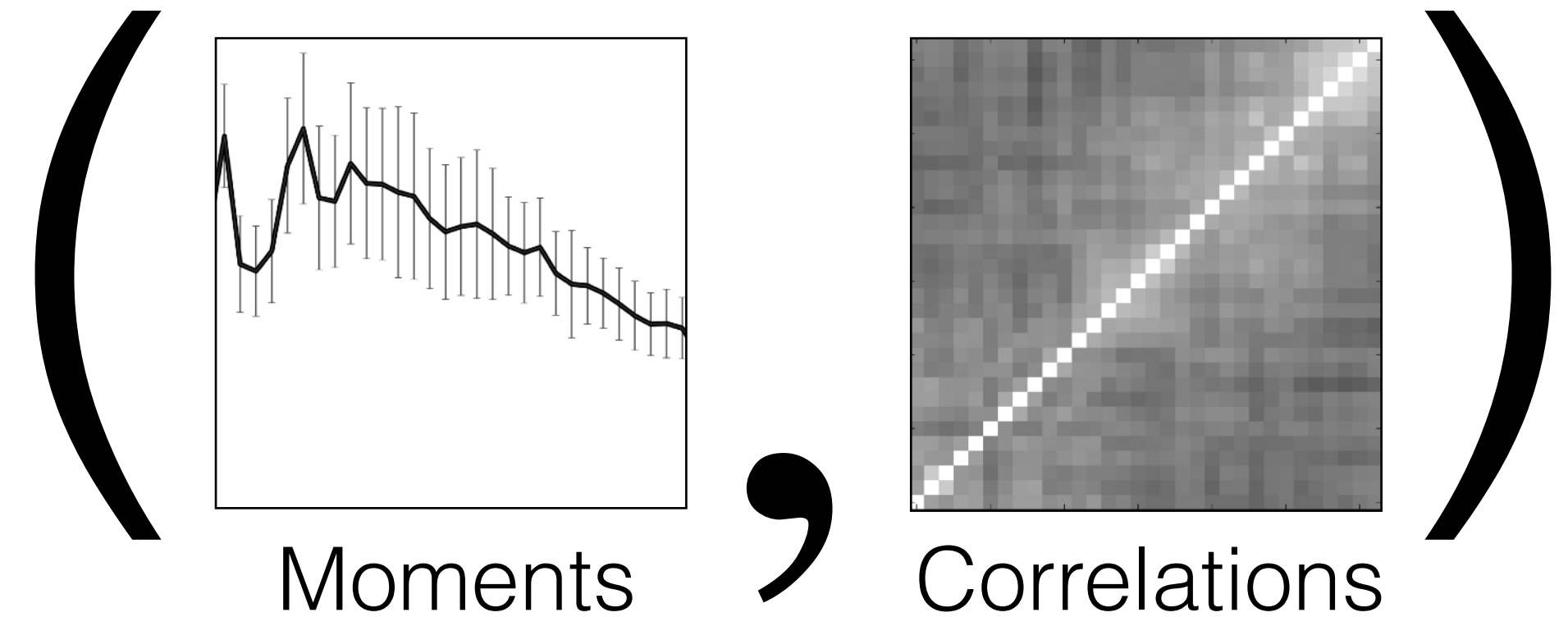
Predicting ambient sound



Video frame

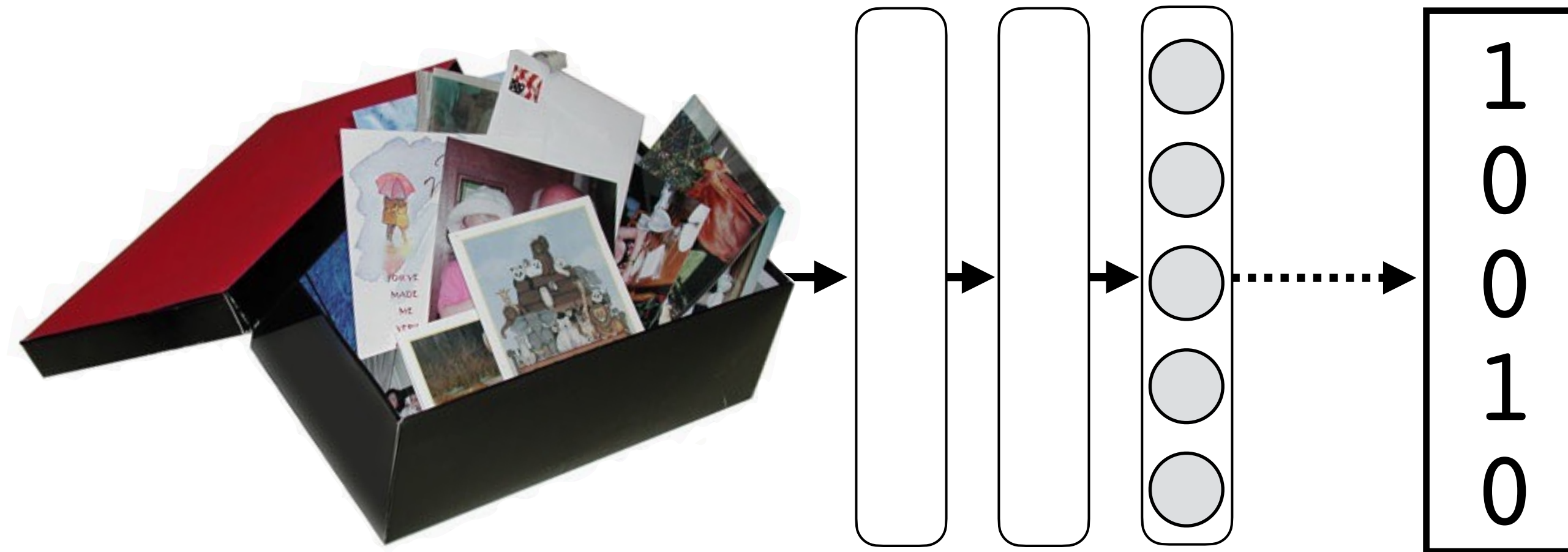


Convolutional
network

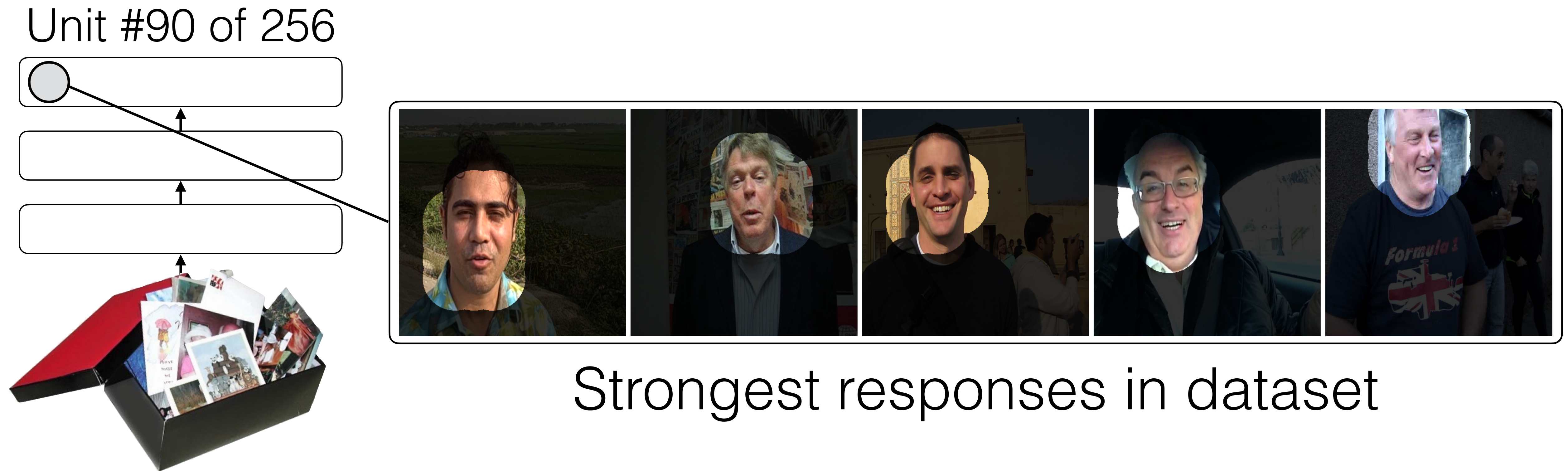


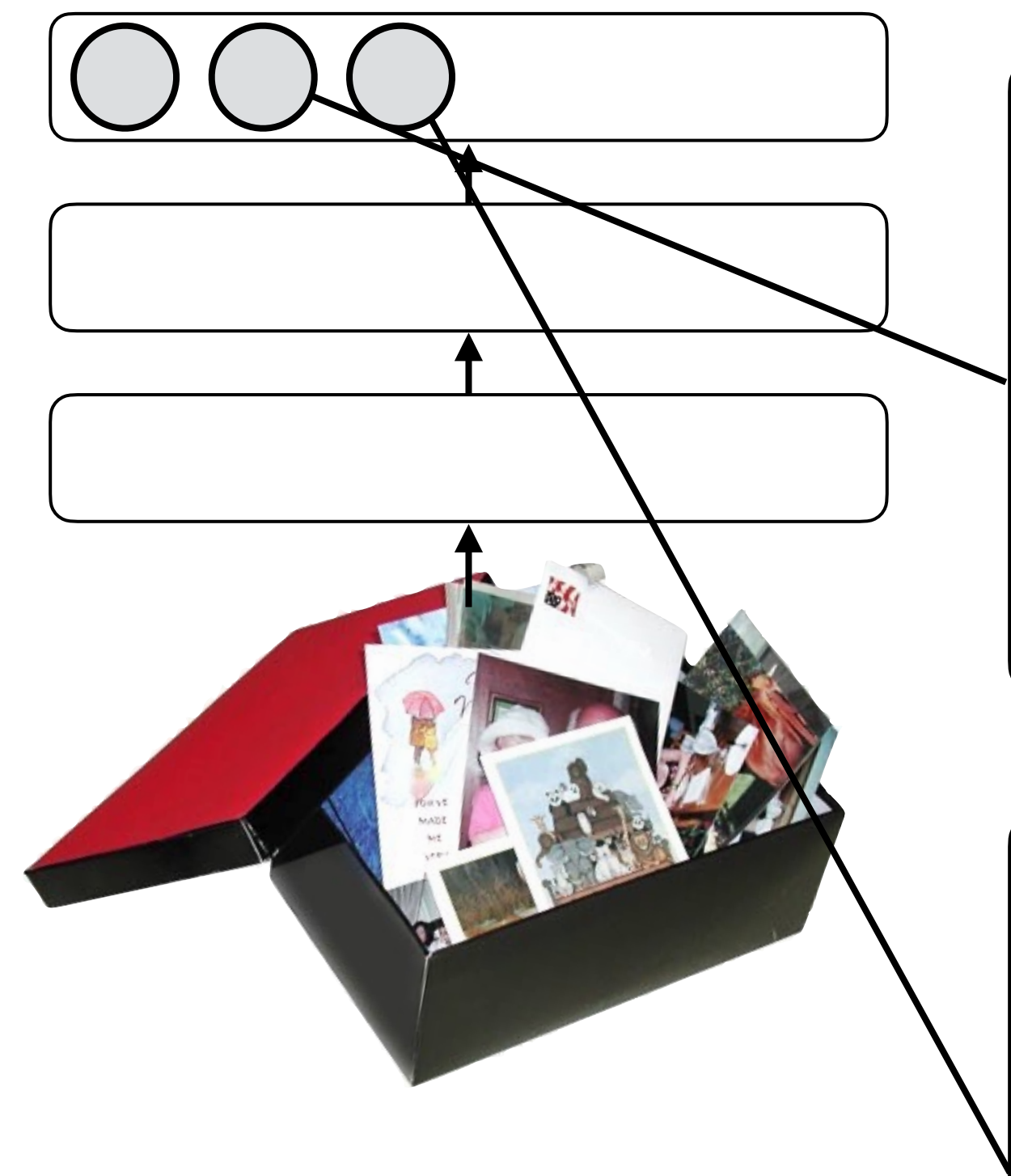
Sound texture
(McDermott & Simoncelli 2011)

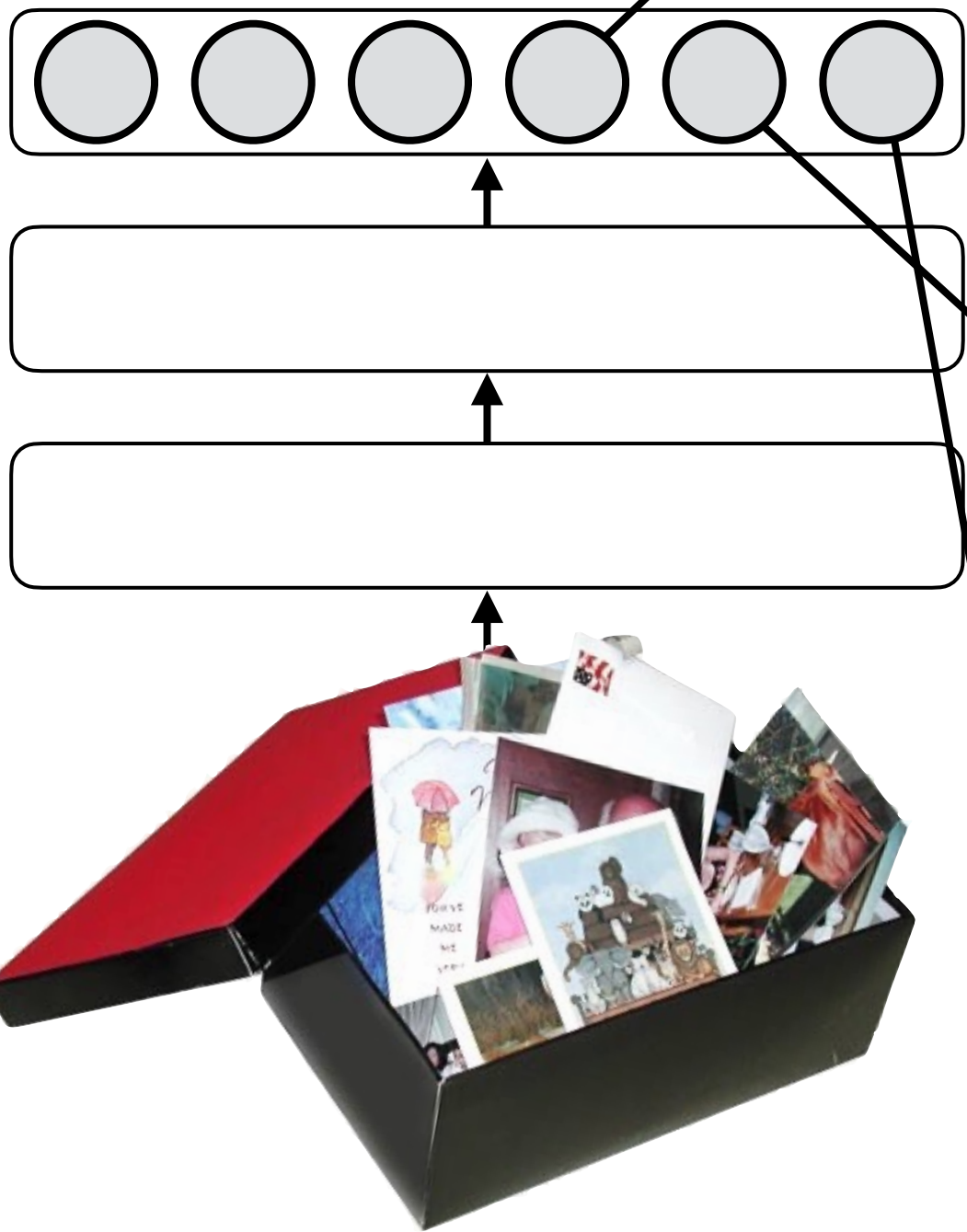
What did the model learn?



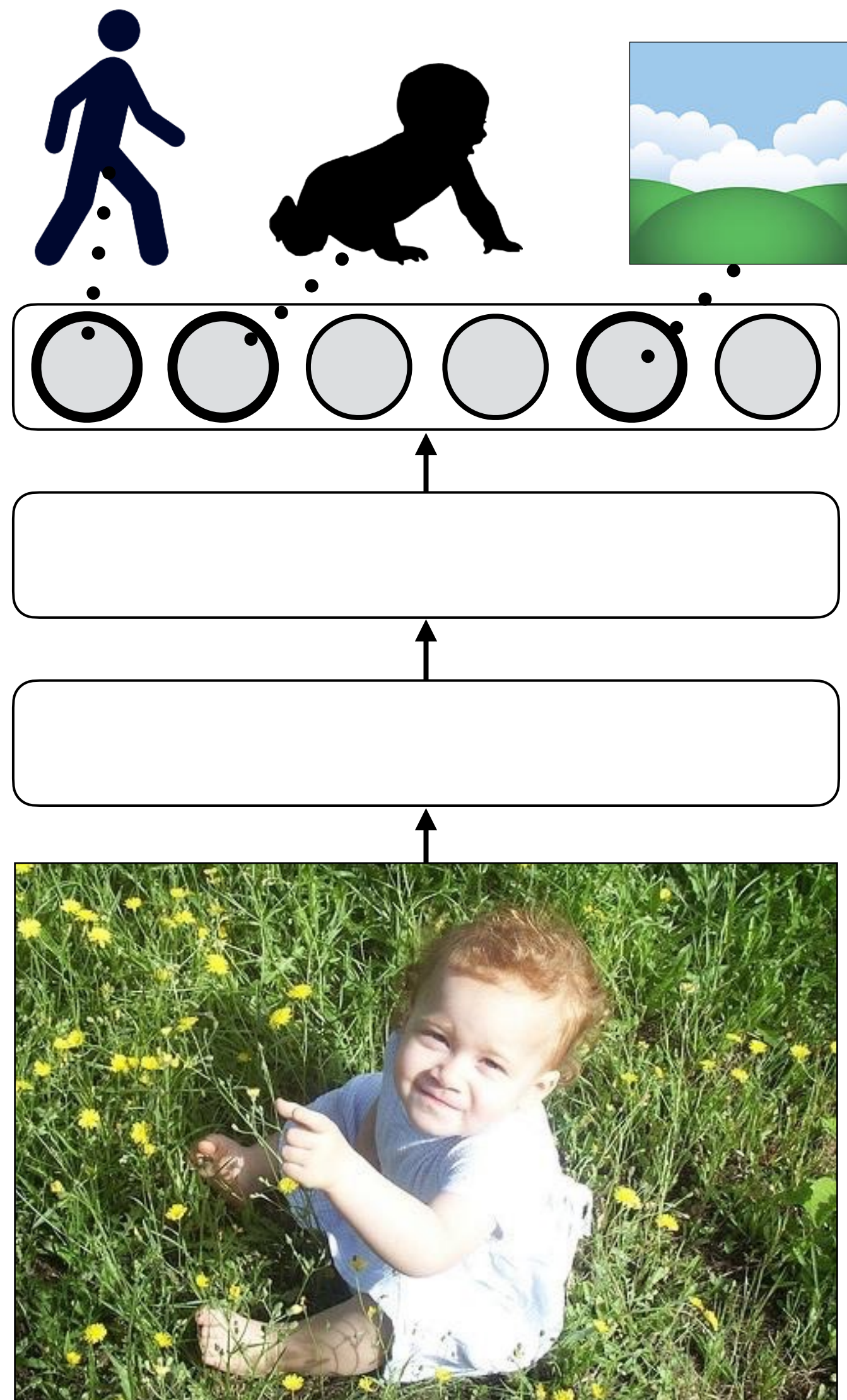
What did the model learn?



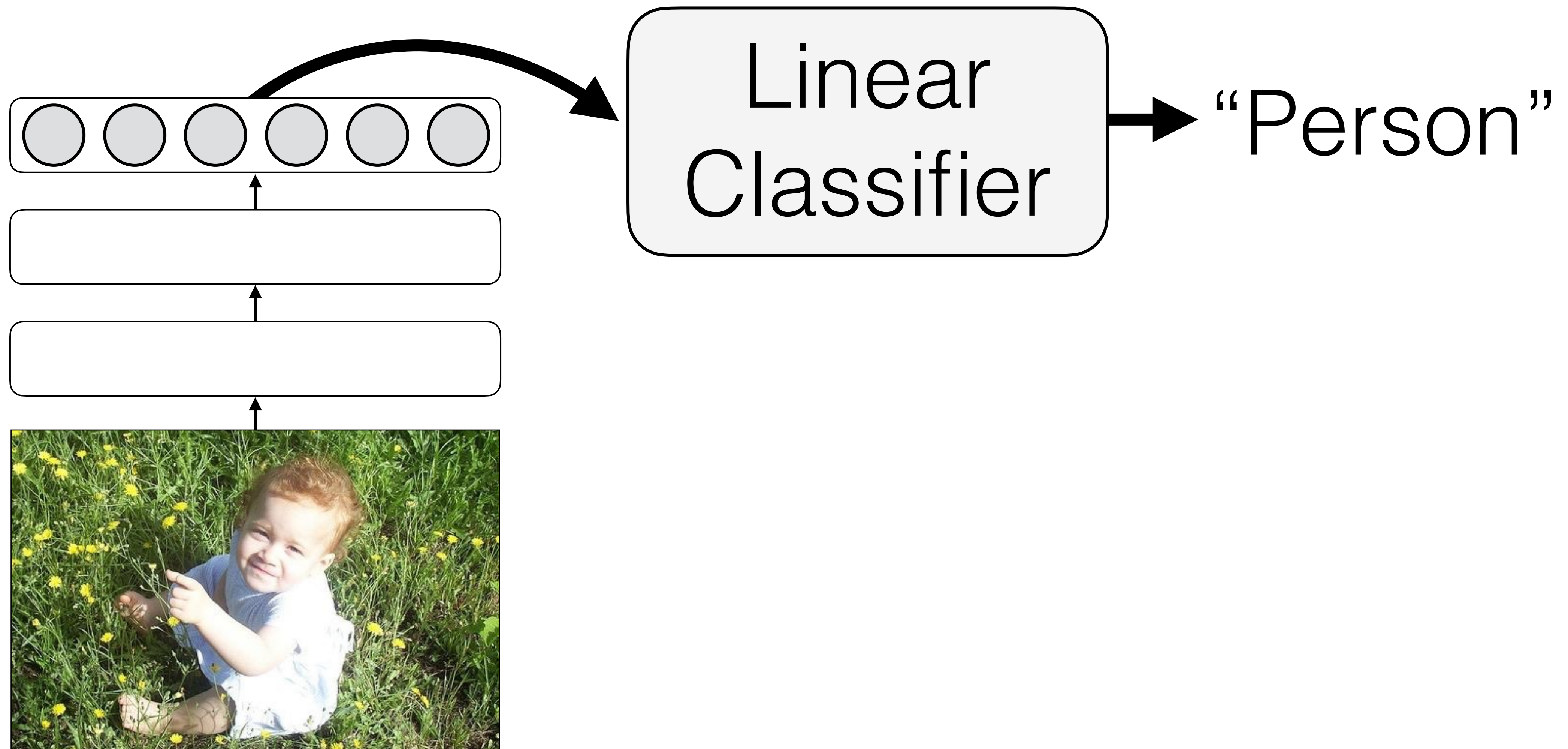




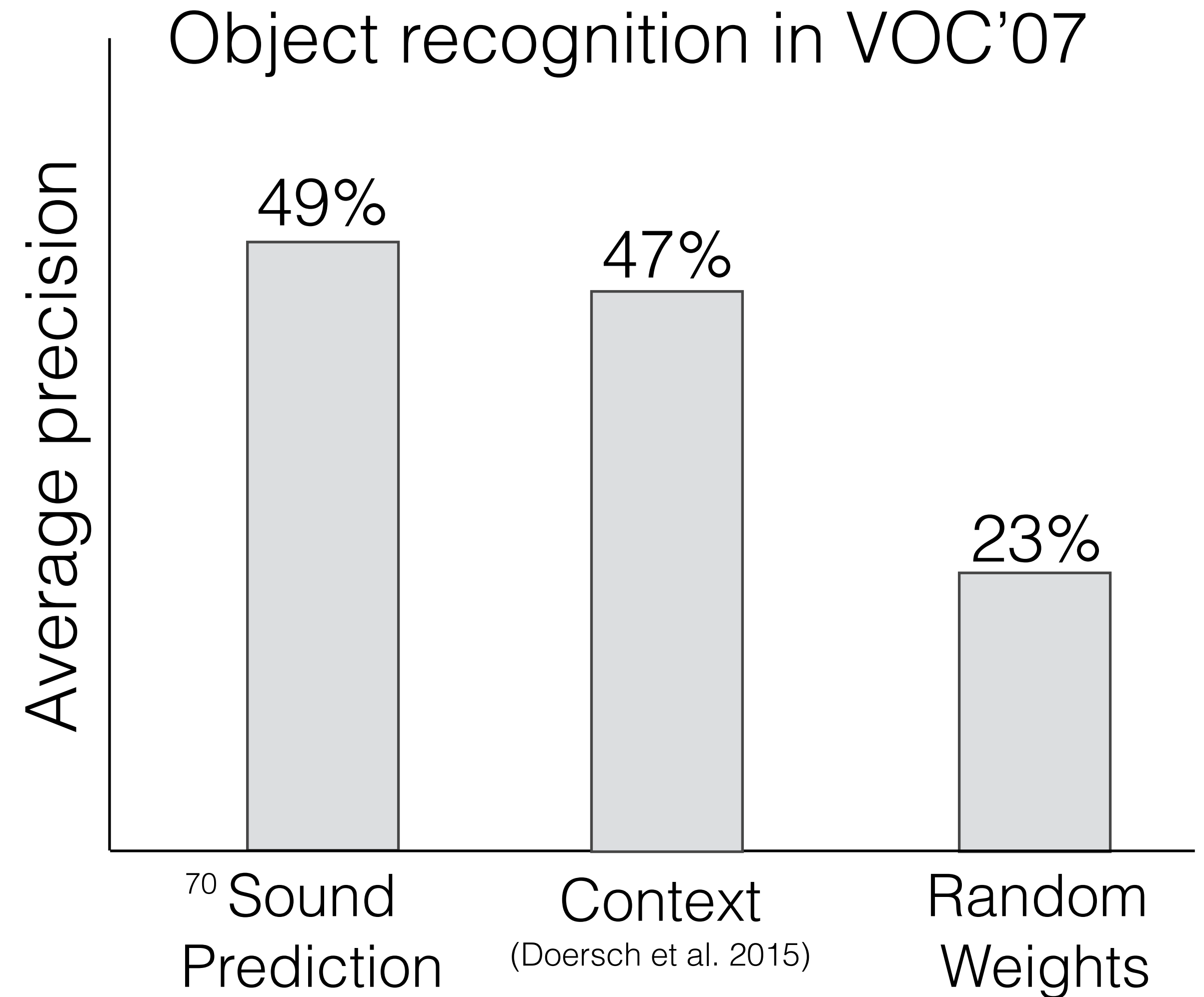
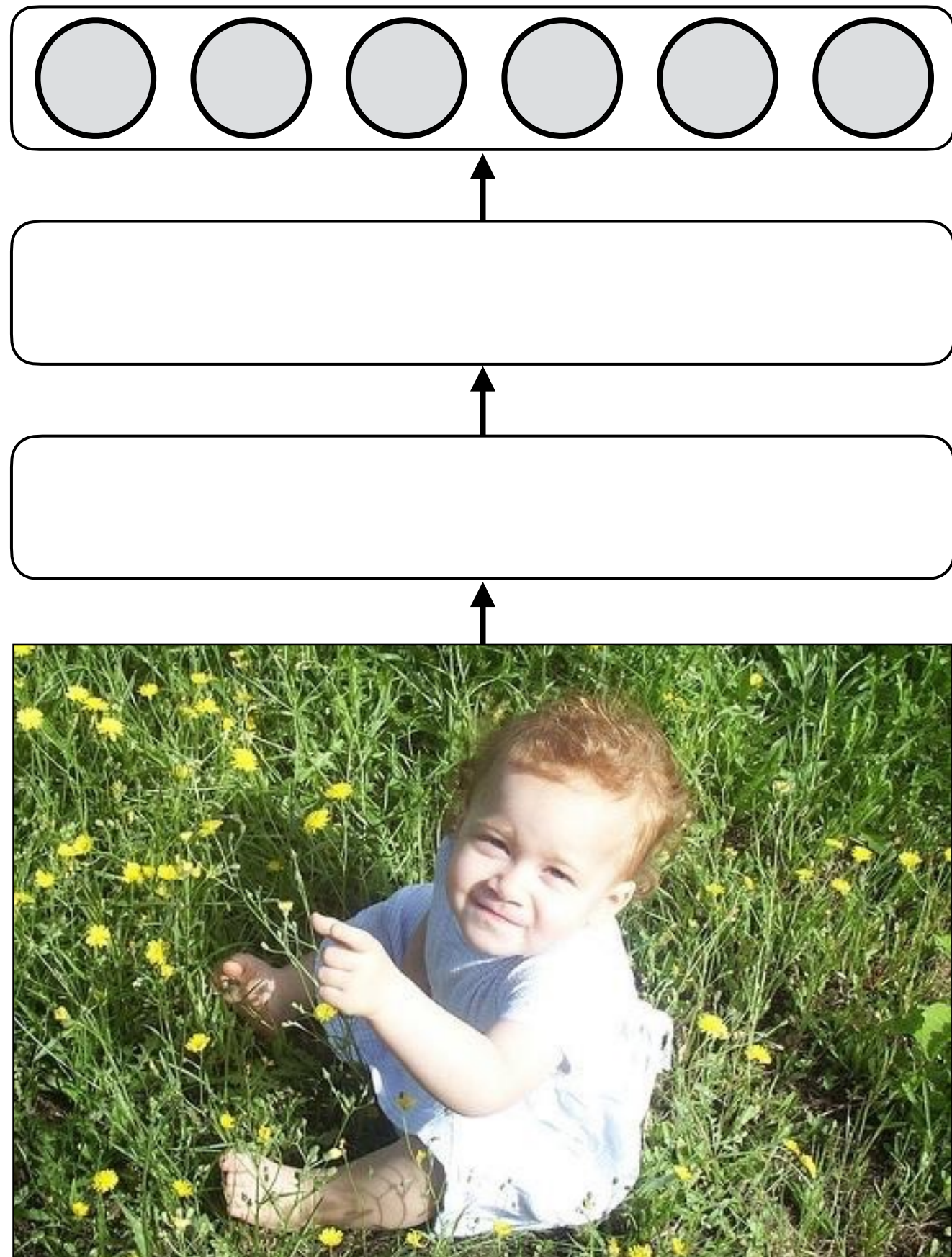
What did the model learn?



What did the model learn?



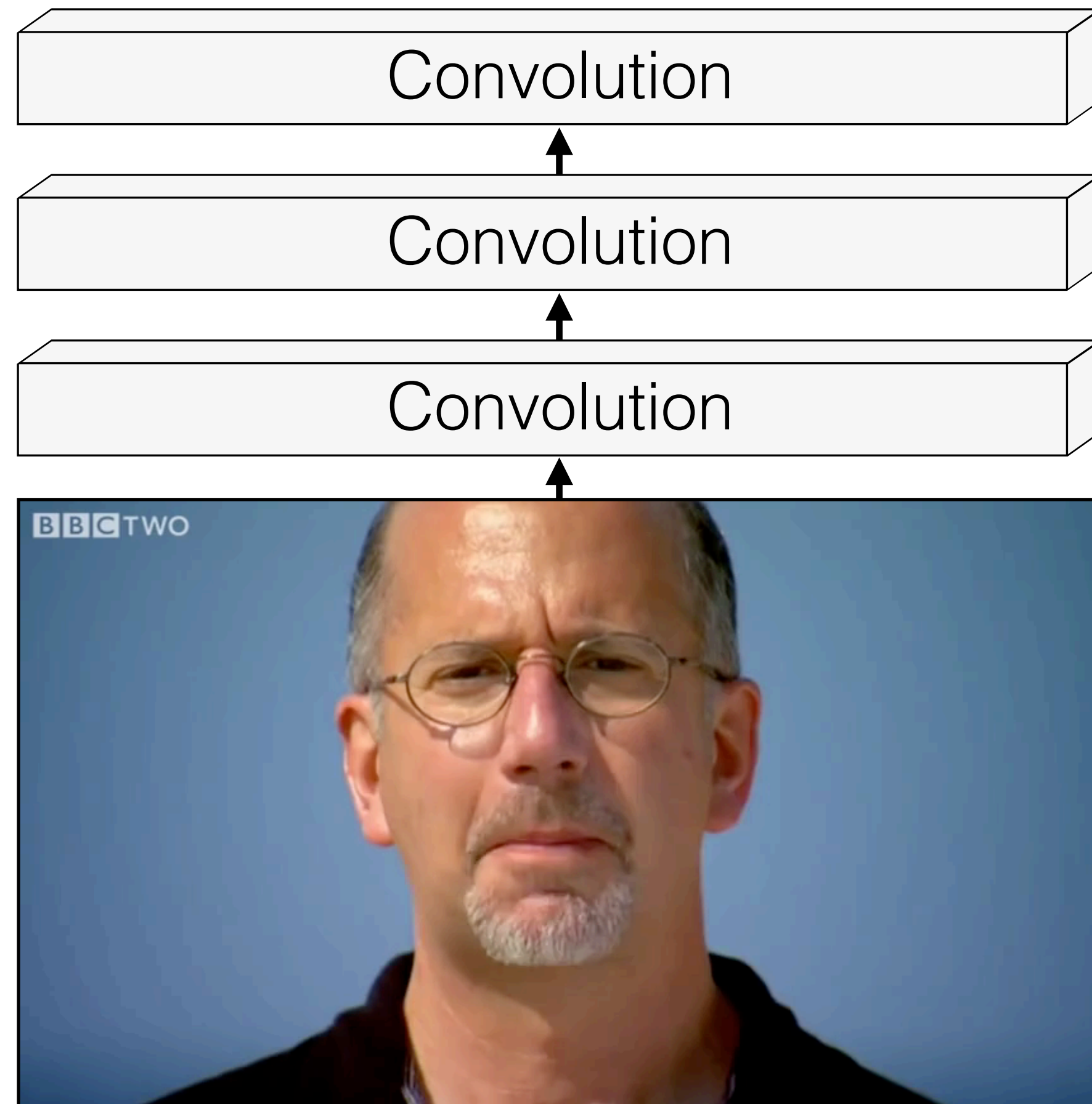
What did the model learn?



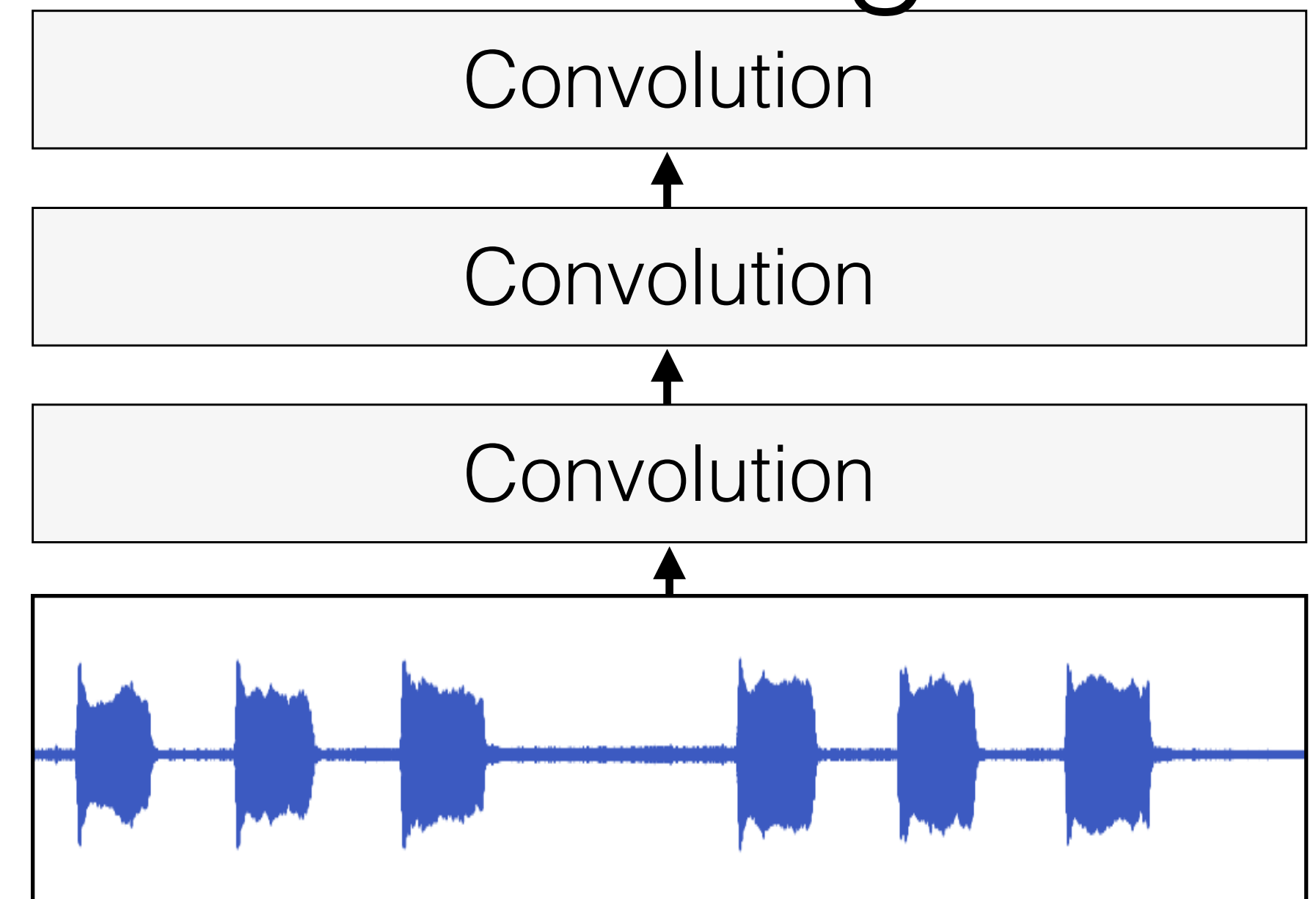
Learning multimodal representations

Single-modality representations

Vision



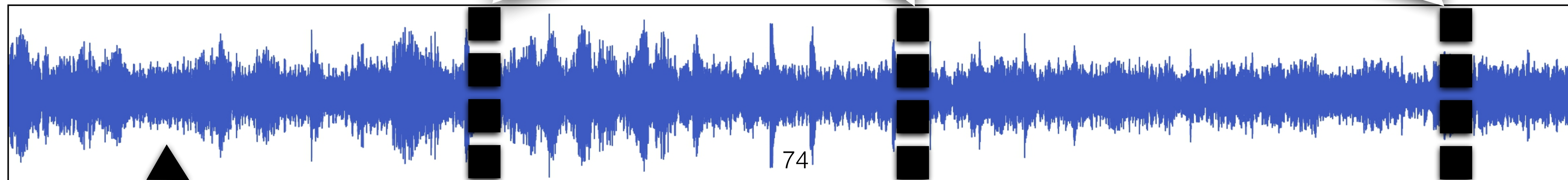
Hearing



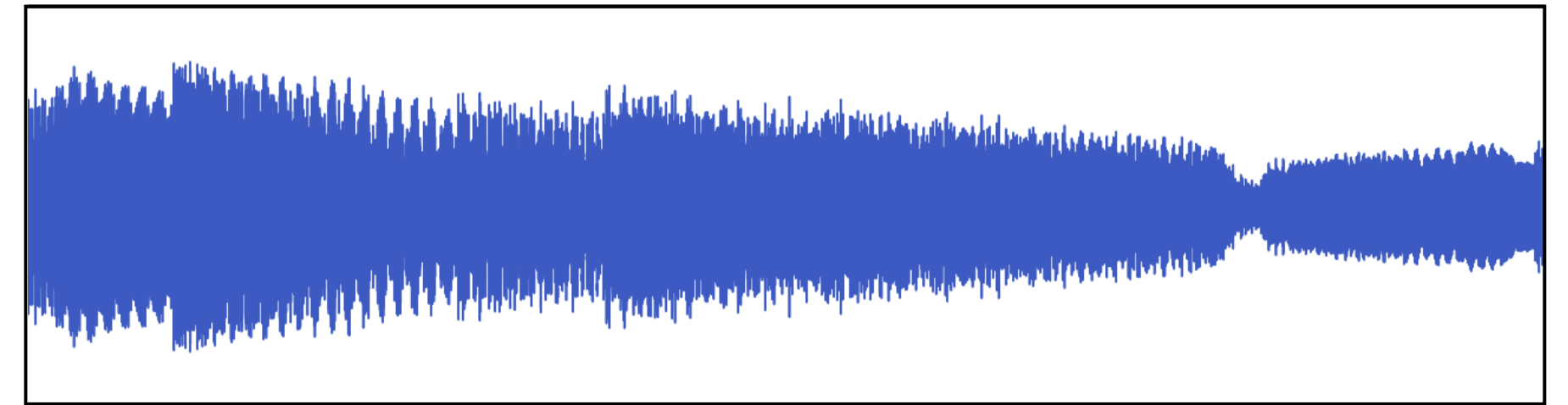
Idea: train a model to find audio-visual correspondences in video.



How do we get ground-truth correspondences?



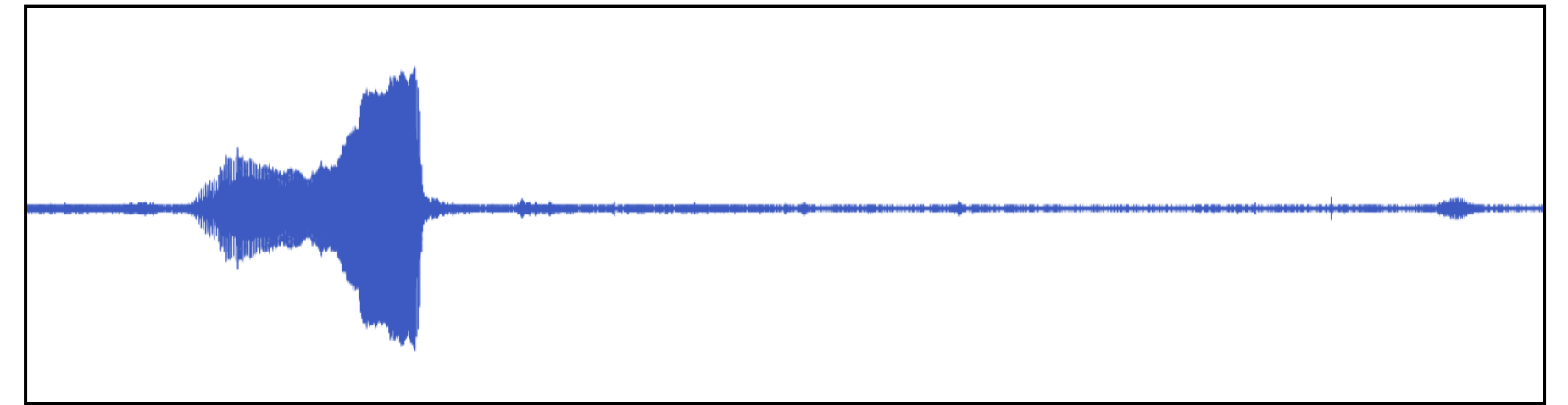
Learning audio-visual correspondences



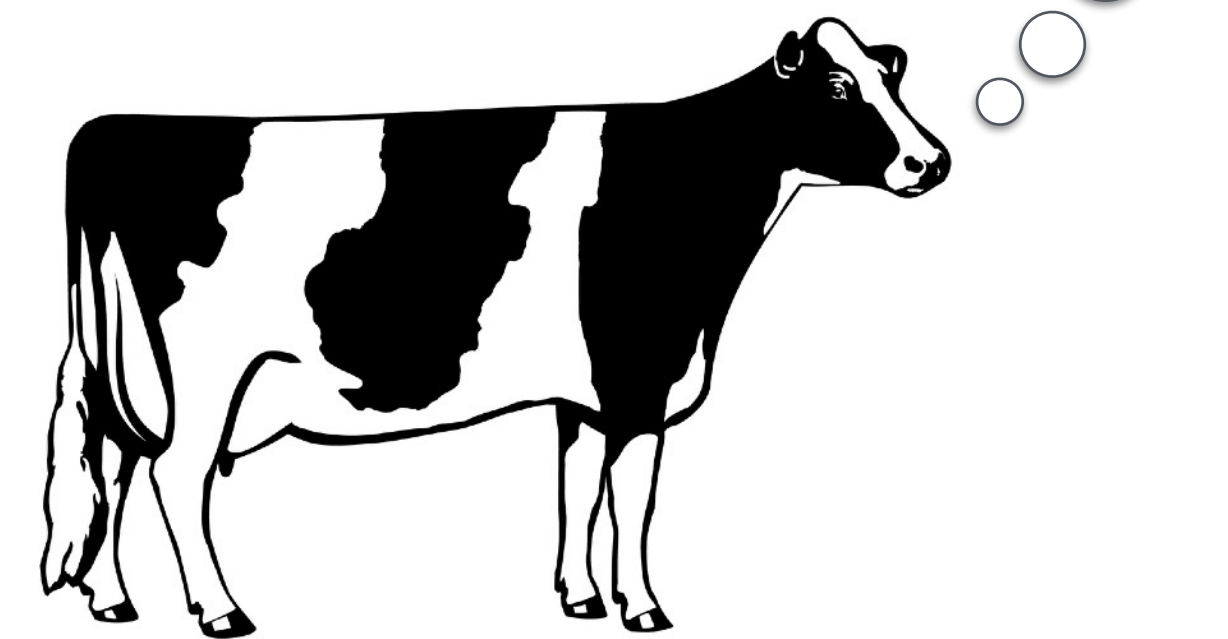
→ **real** or fake?

Related work: L³-net (Arandjelović & Zisserman 2017), AVTS (Korbar et al. 2018)
Noise-contrastive estimation (Gutmann & Hyvarinen 2010)

Learning audio-visual correspondences



“moo”



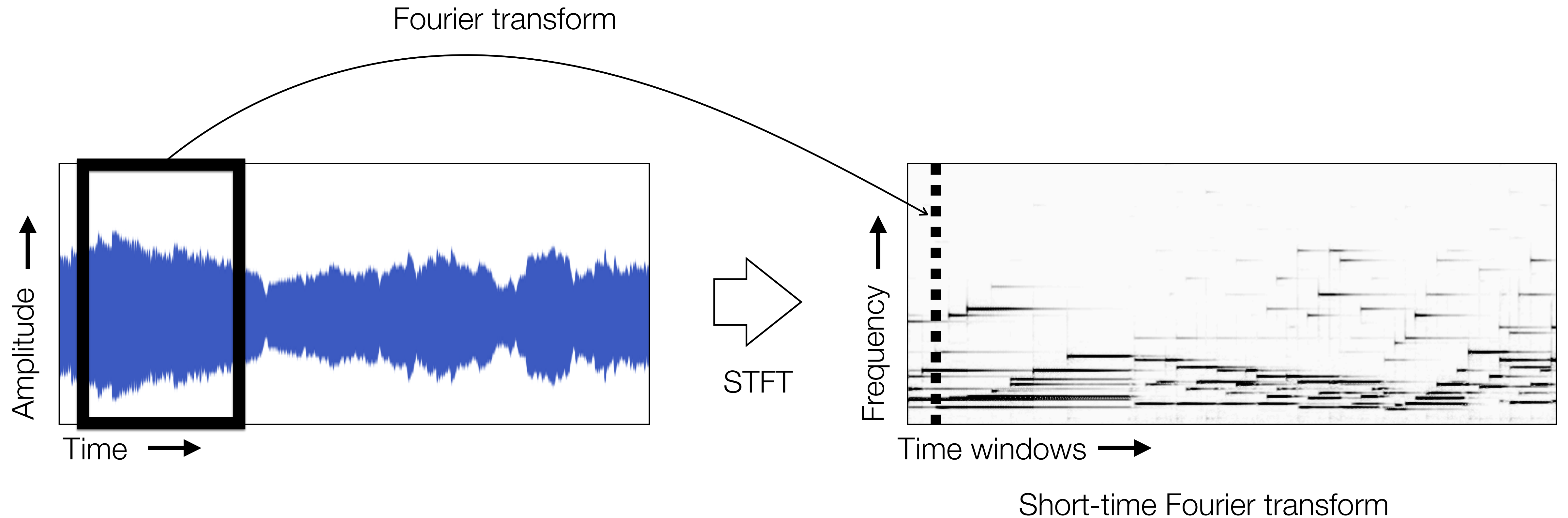
→ real or **fake**?

Related work: L³-net (Arandjelović & Zisserman 2017), AVTS (Korbar et al. 2018)
Noise-contrastive estimation (Gutmann & Hyvarinen, 2010)

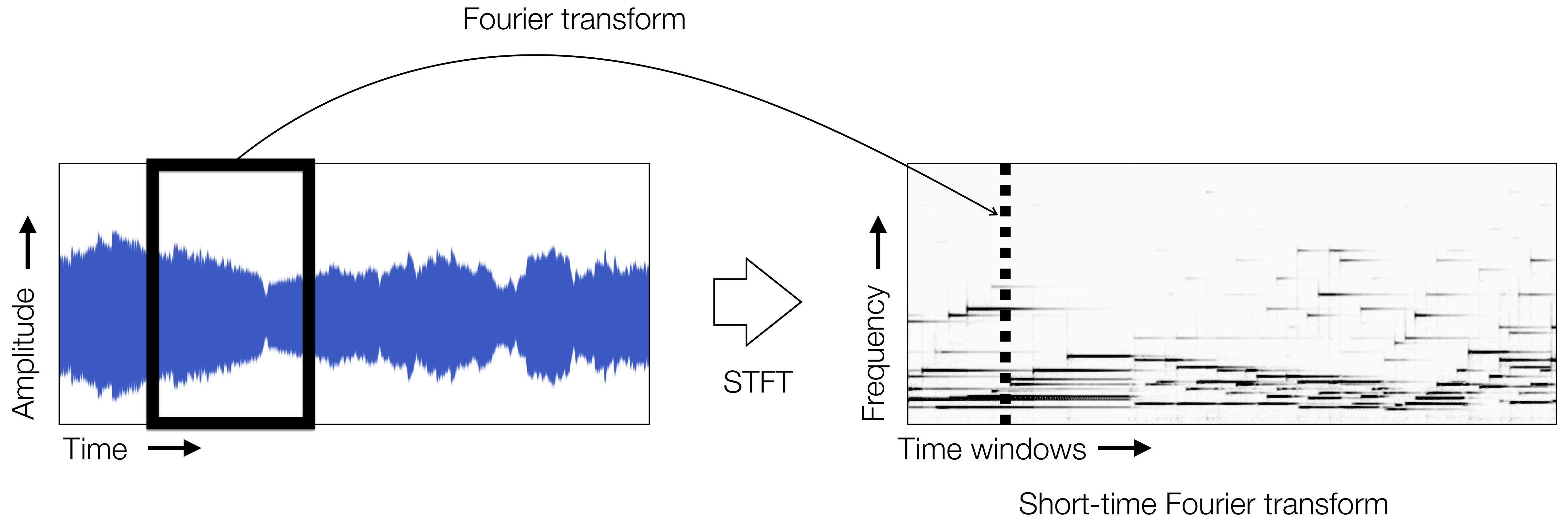
Idea #1: use random audio as negatives



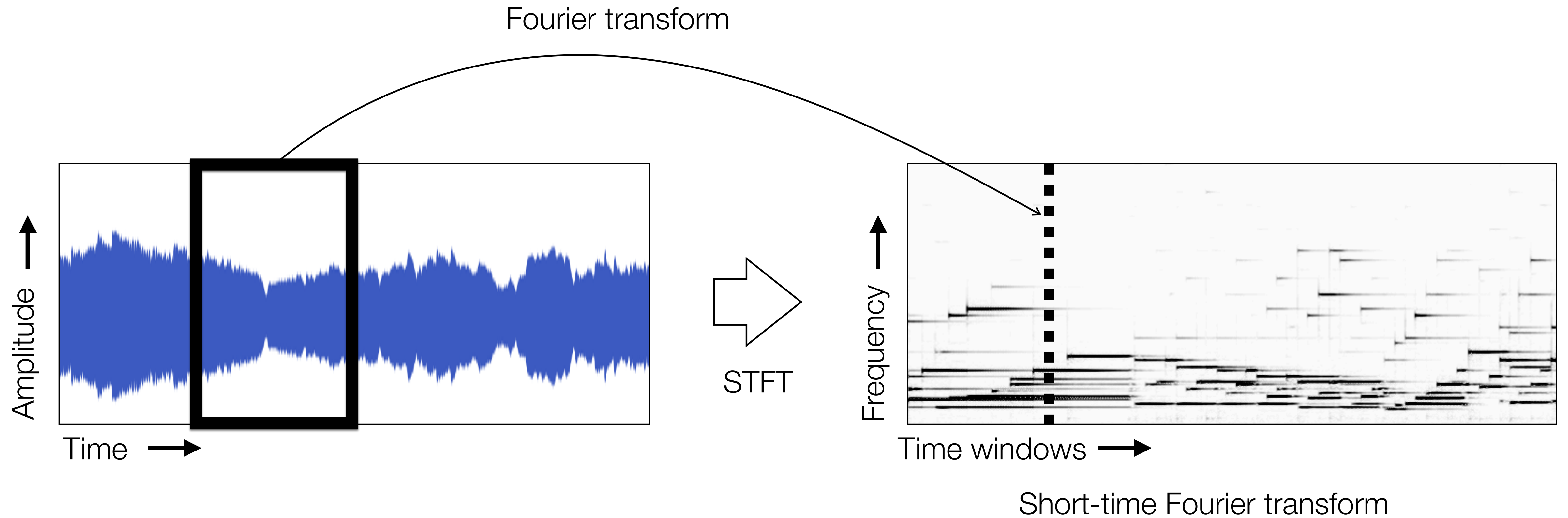
How do we represent the sound?



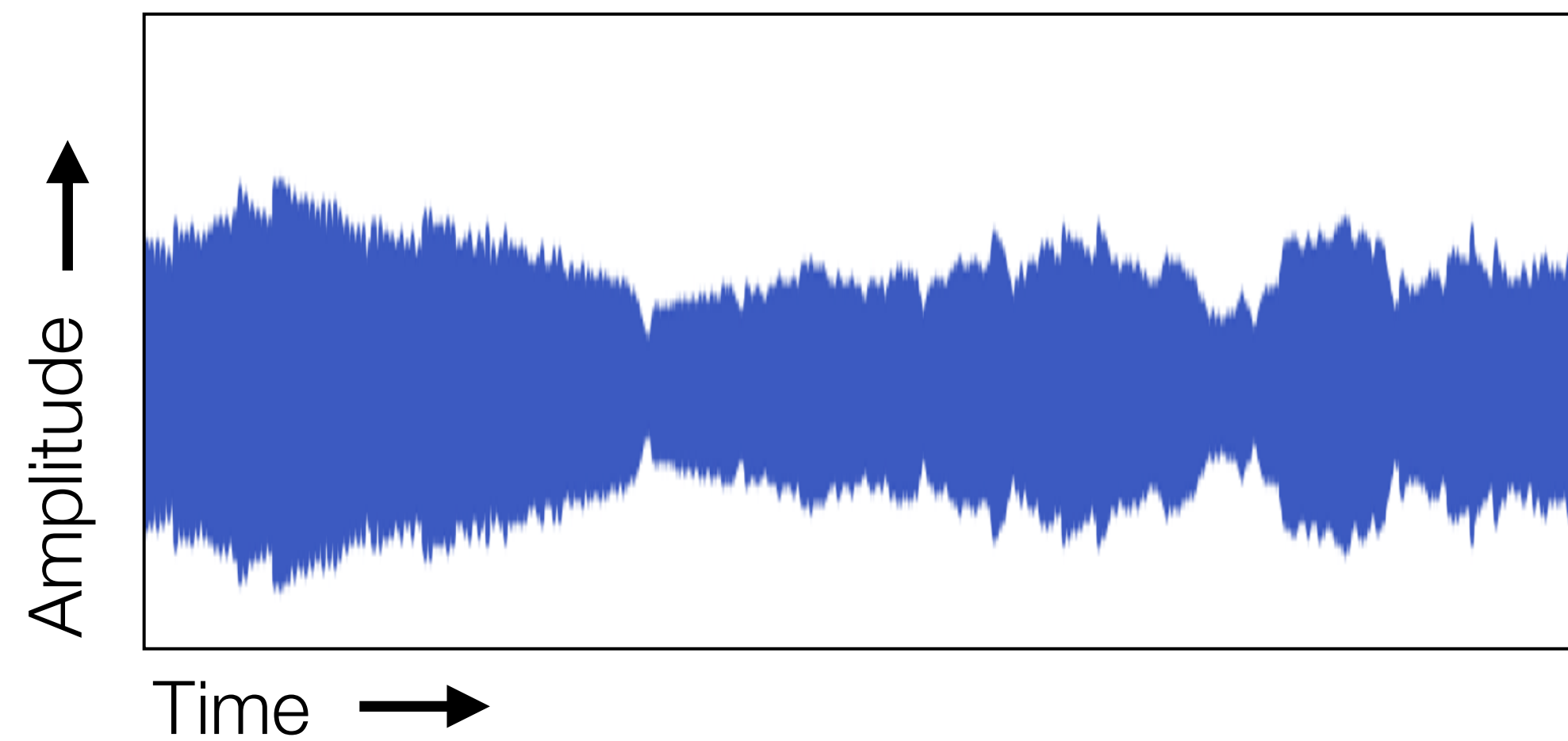
Spectrogram



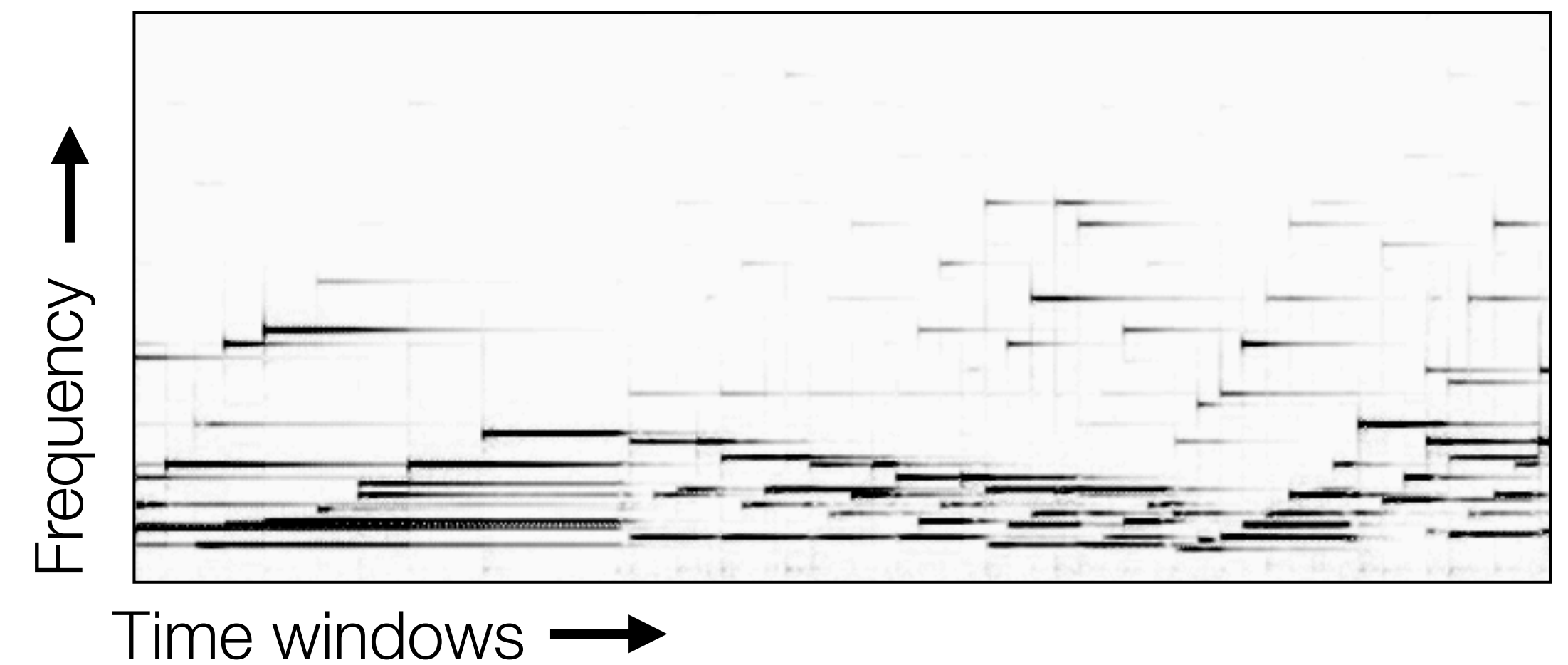
Spectrogram



Spectrogram



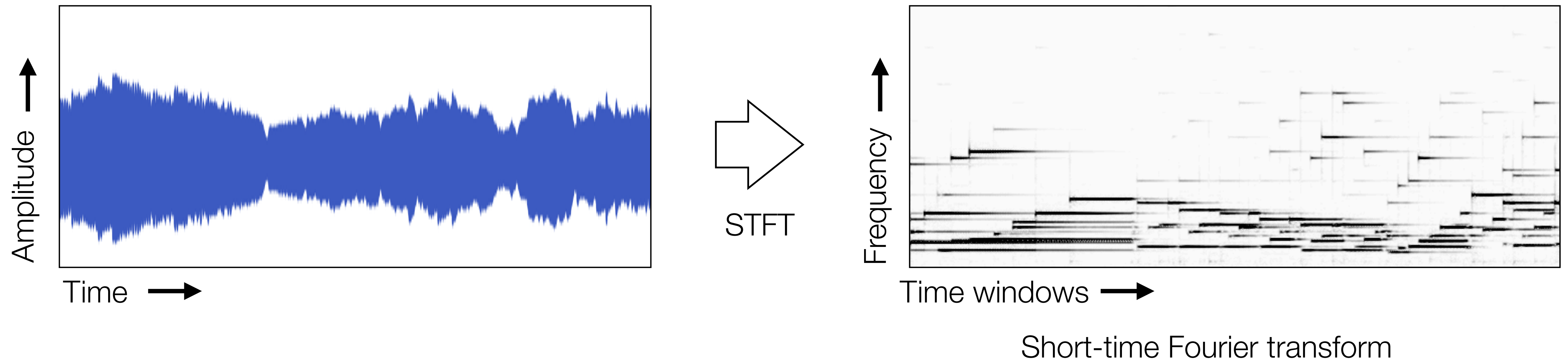
STFT



Short-time Fourier transform

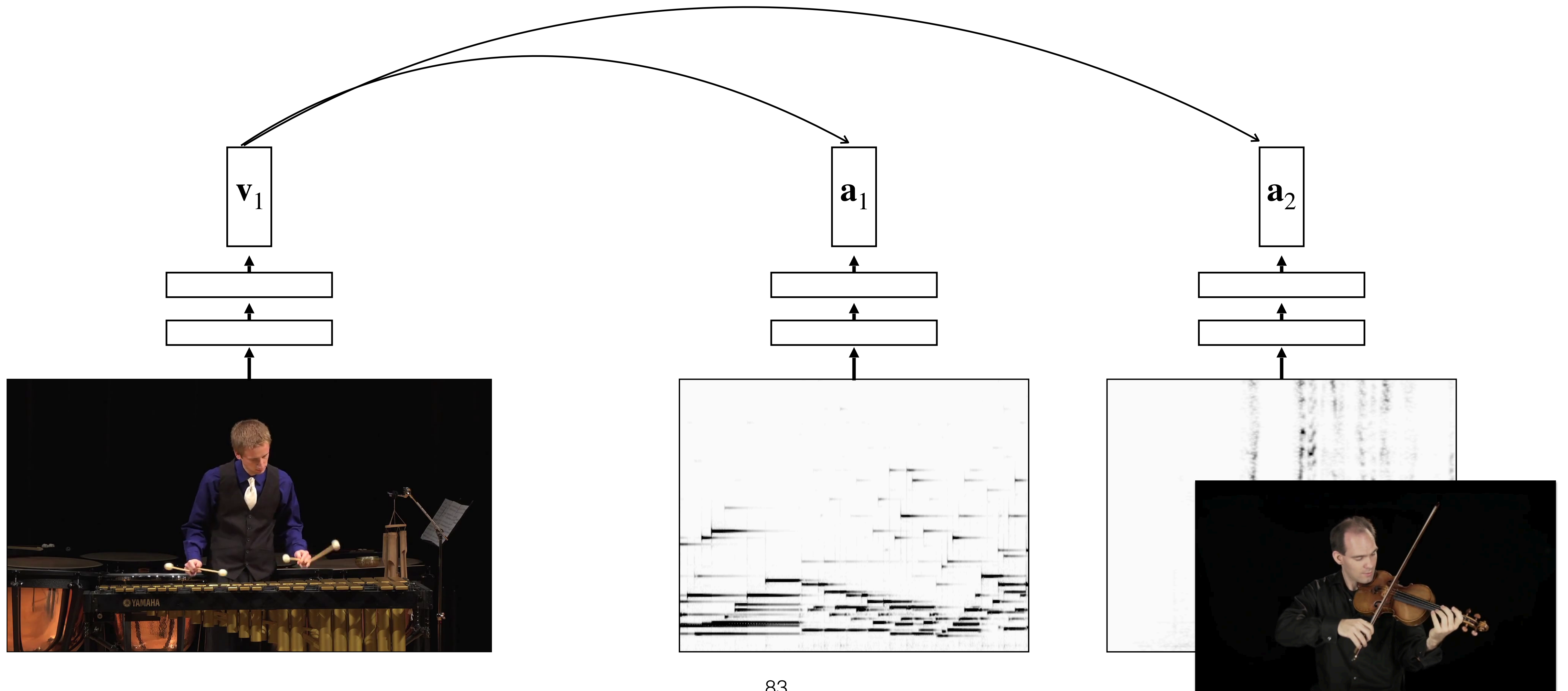
Spectrogram

Treat it like an image and
process with a CNN!



Usually lower dimensional: e.g. 16,000 samples becomes 256 x 256 spectrogram

Contrastive audio-visual learning

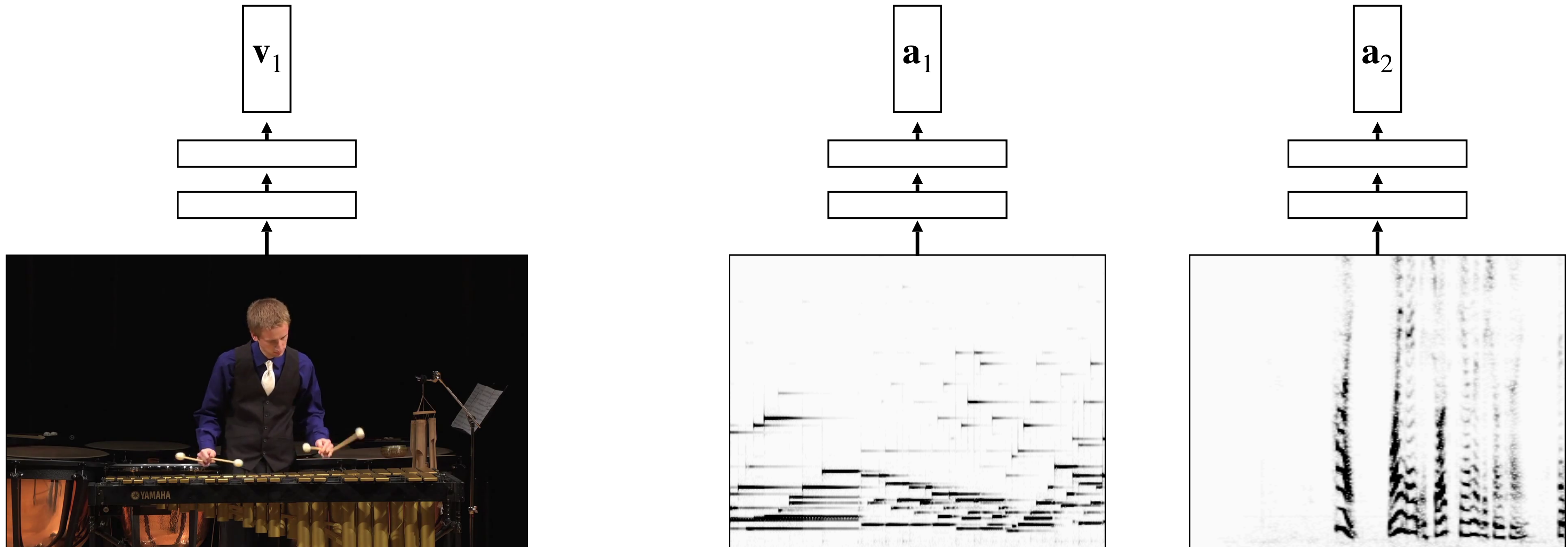


Contrastive audio-visual learning

Learn networks where:

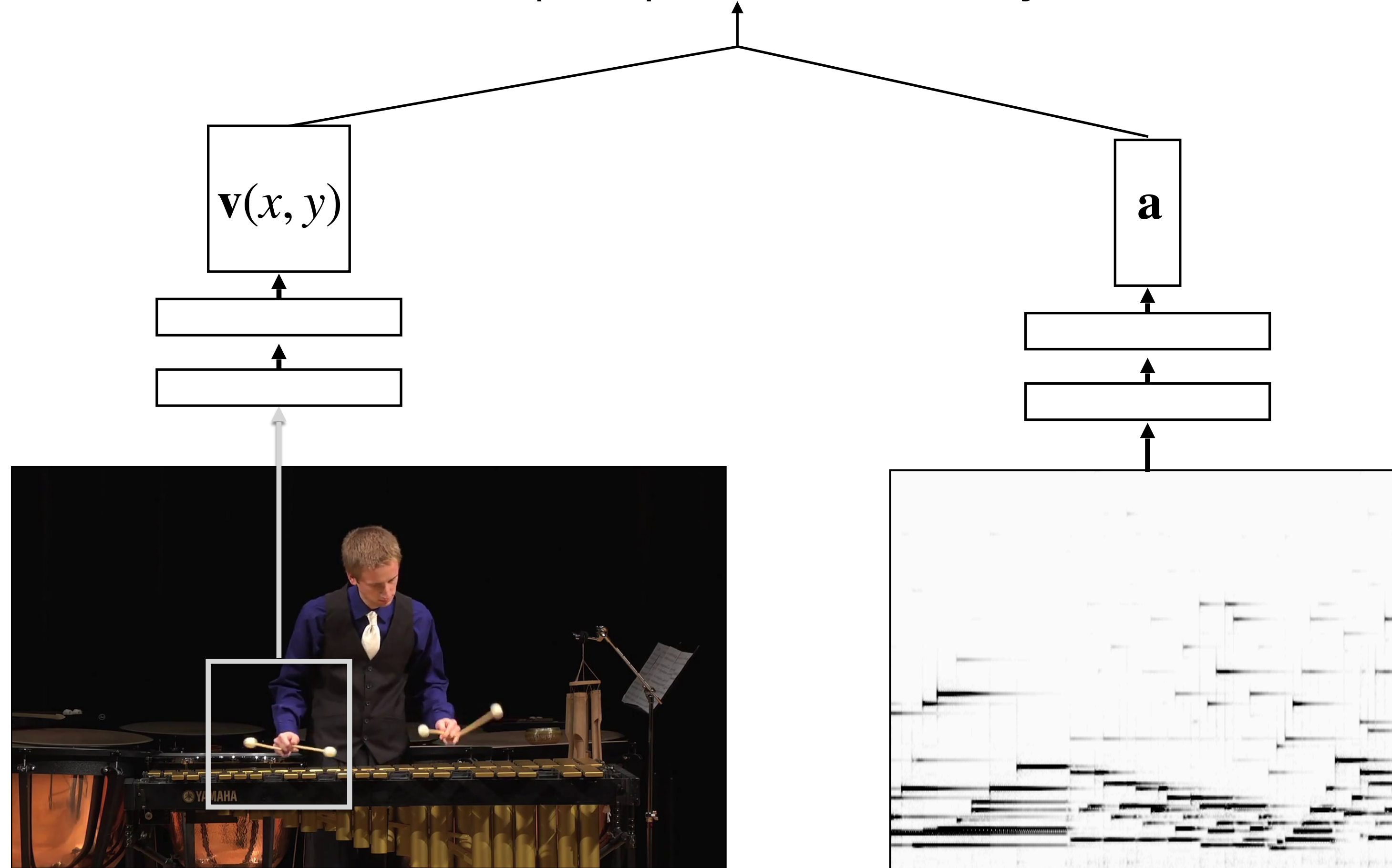
$\mathbf{v}_1^T \mathbf{a}_1 \longrightarrow$ High dot product with self

$\mathbf{v}_1^T \mathbf{a}_2 \longrightarrow$ Low dot product with others



Application: sound source localization

Combine per-patch similarity scores



During training, maximize:

$$\max_{x,y} \mathbf{v}(x, y)^{\top} \mathbf{a}$$

Underlying assumption:
there is one sound source.
You can ignore the rest of
the image.

Application: sound source localization

Objects that Sound

Relja Arandjelović¹, Andrew Zisserman^{1,2}

¹DeepMind ²University of Oxford

Frames are processed completely independently, motion information is not used, and there is no temporal smoothing

Input single
frame



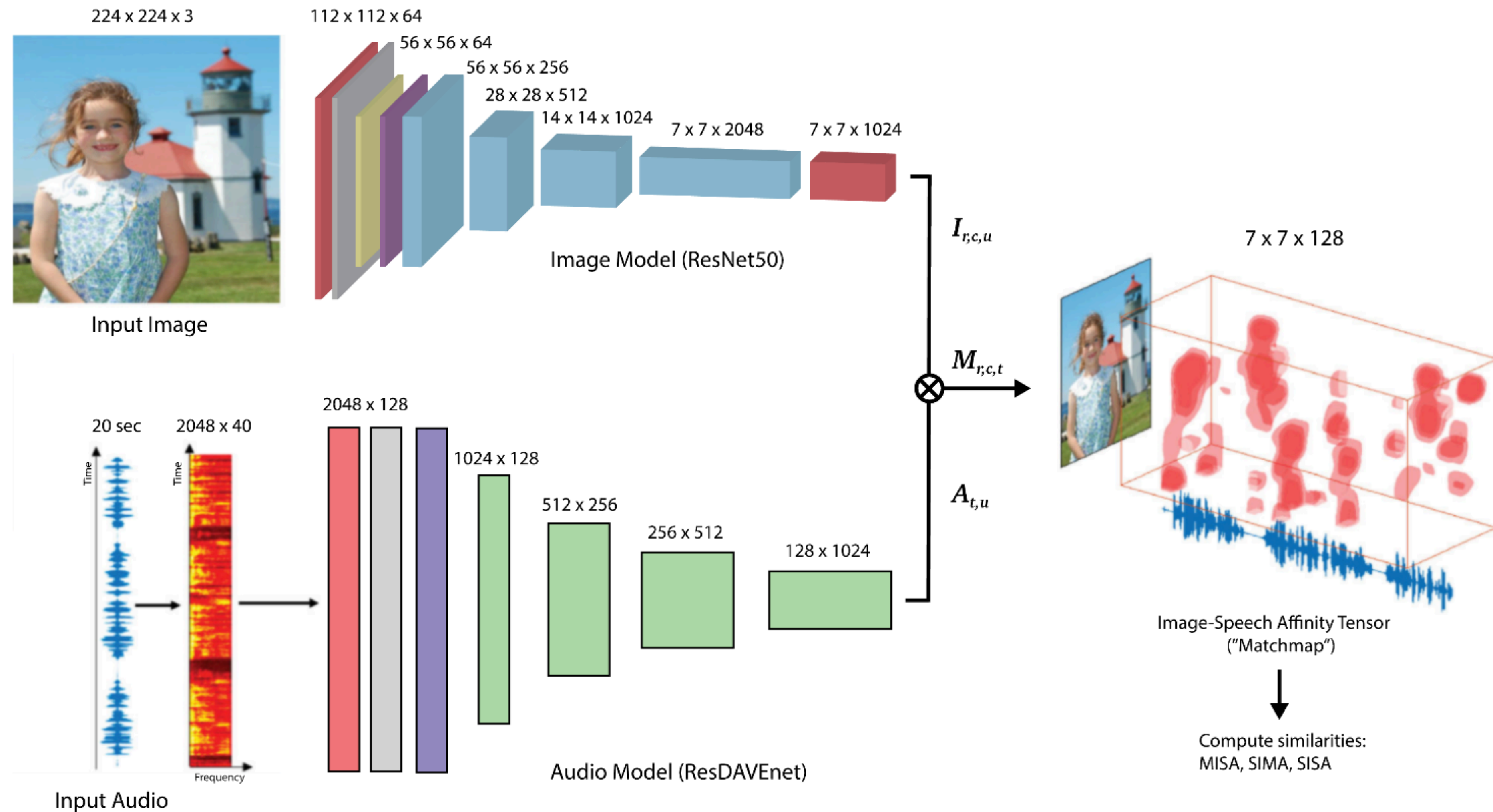
Frame/
Localization
overlaid



Localization



Application: associating image content with speech



Discovering words in raw audio-visual data

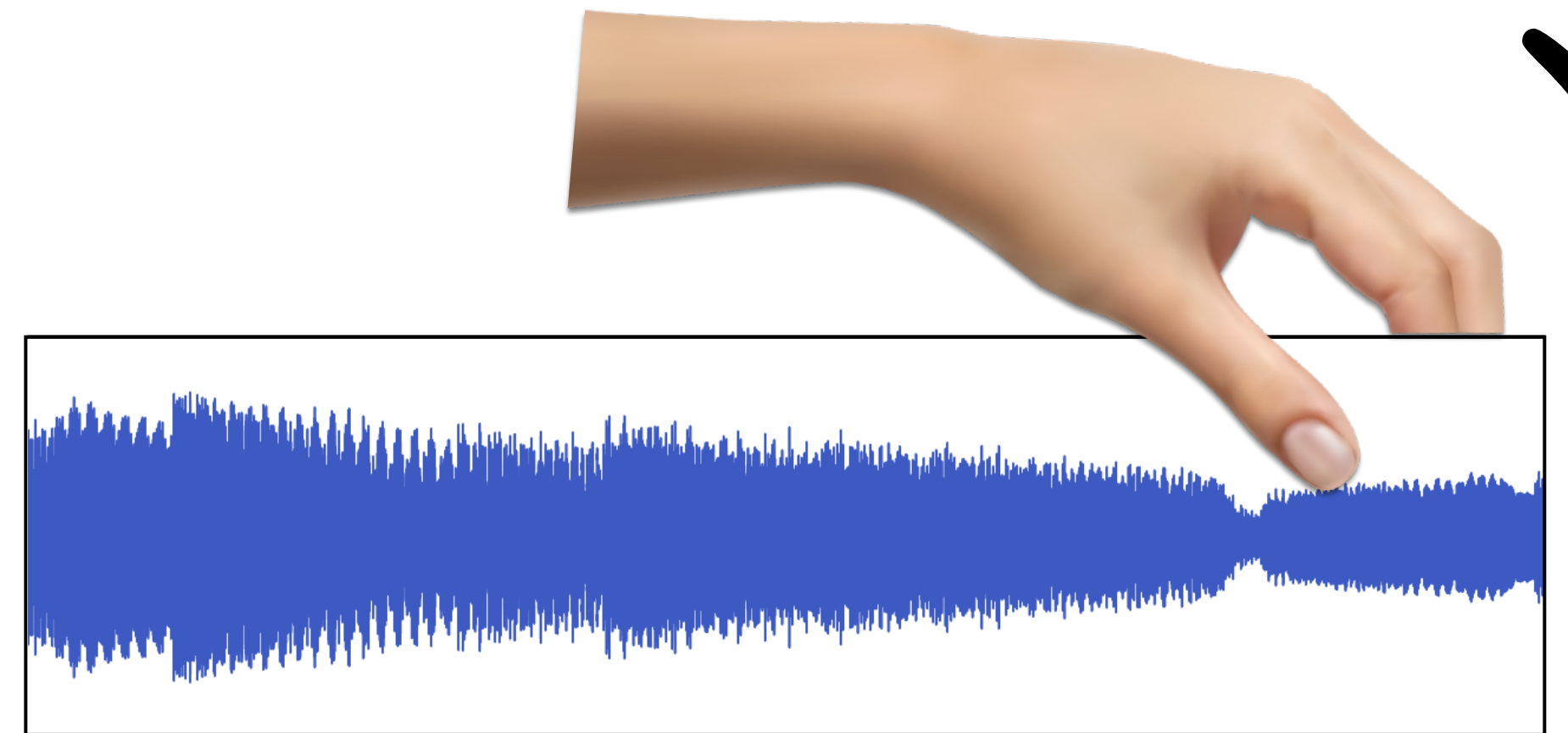


88

Is this task too easy?



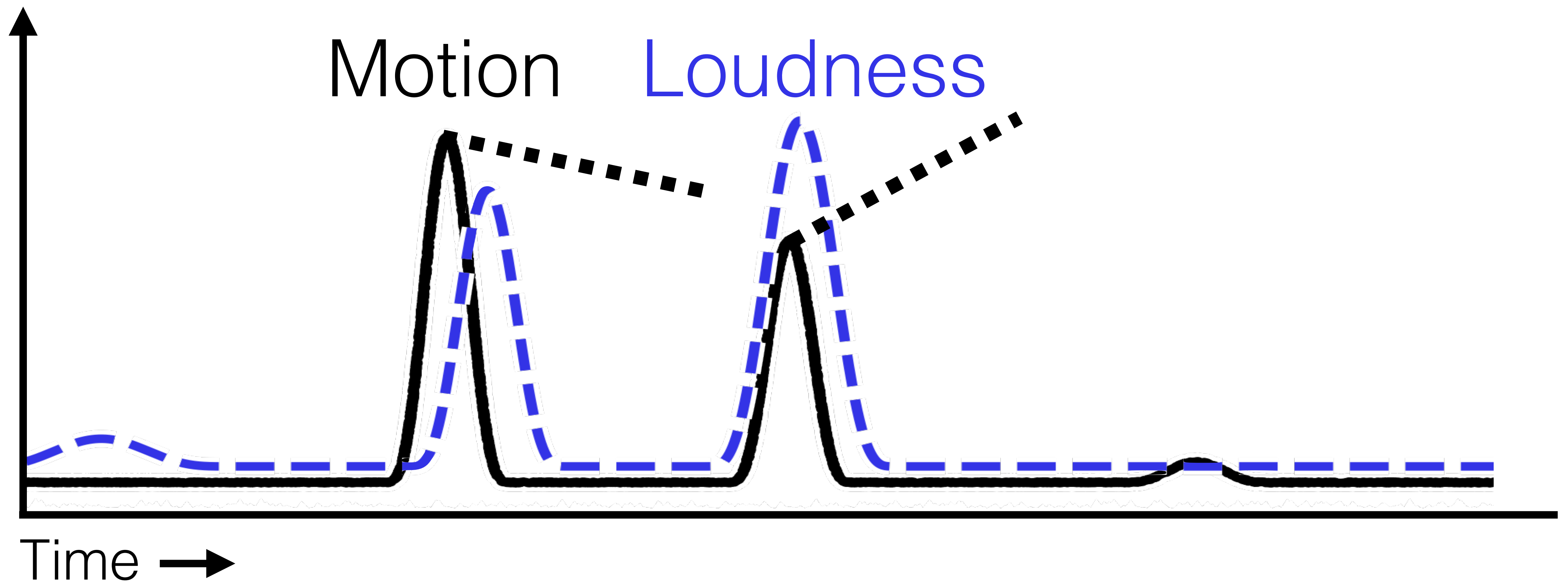
Idea #2: use time-shifted audio as negatives



Idea #2: use time-shifted audio as negatives



Idea #2: use time-shifted audio as negatives





Dribbling basketball

ALLERBOOTCAMP.COM

CLICK FOR A
Dribbling basketball



Dribbling basketball



Playing organ



Playing organ



Playing organ



Chopping wood



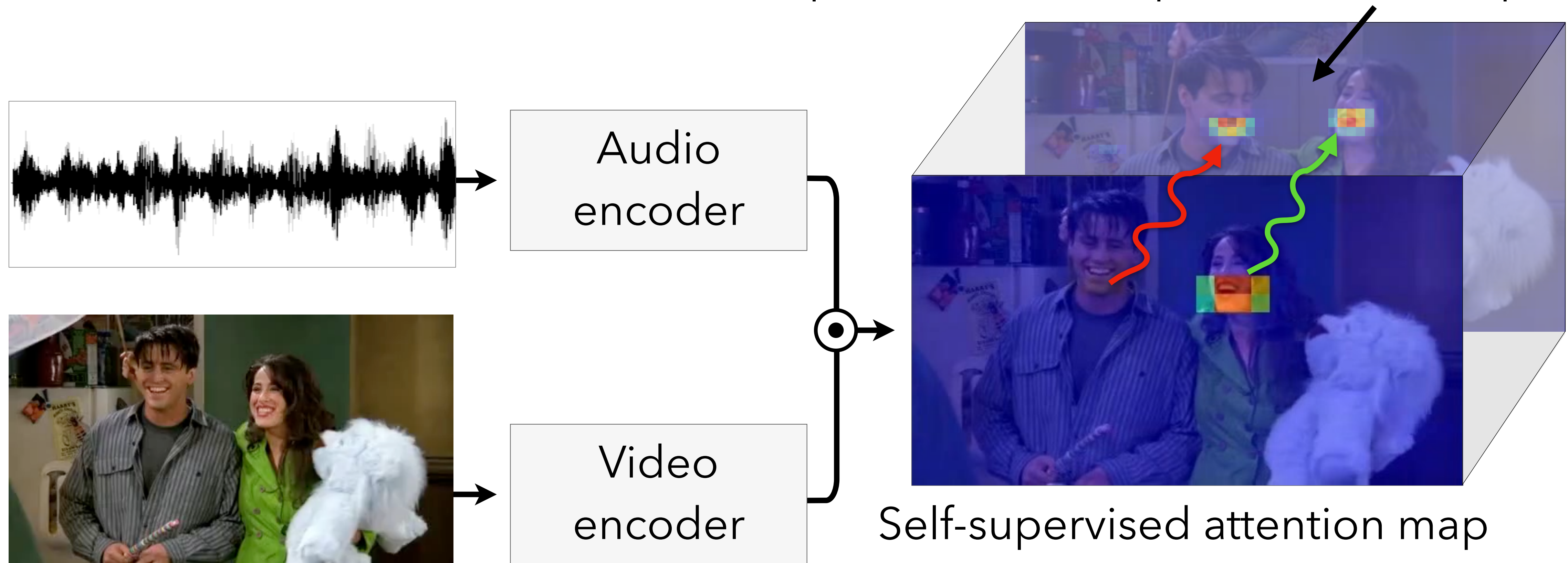
Chopping wood



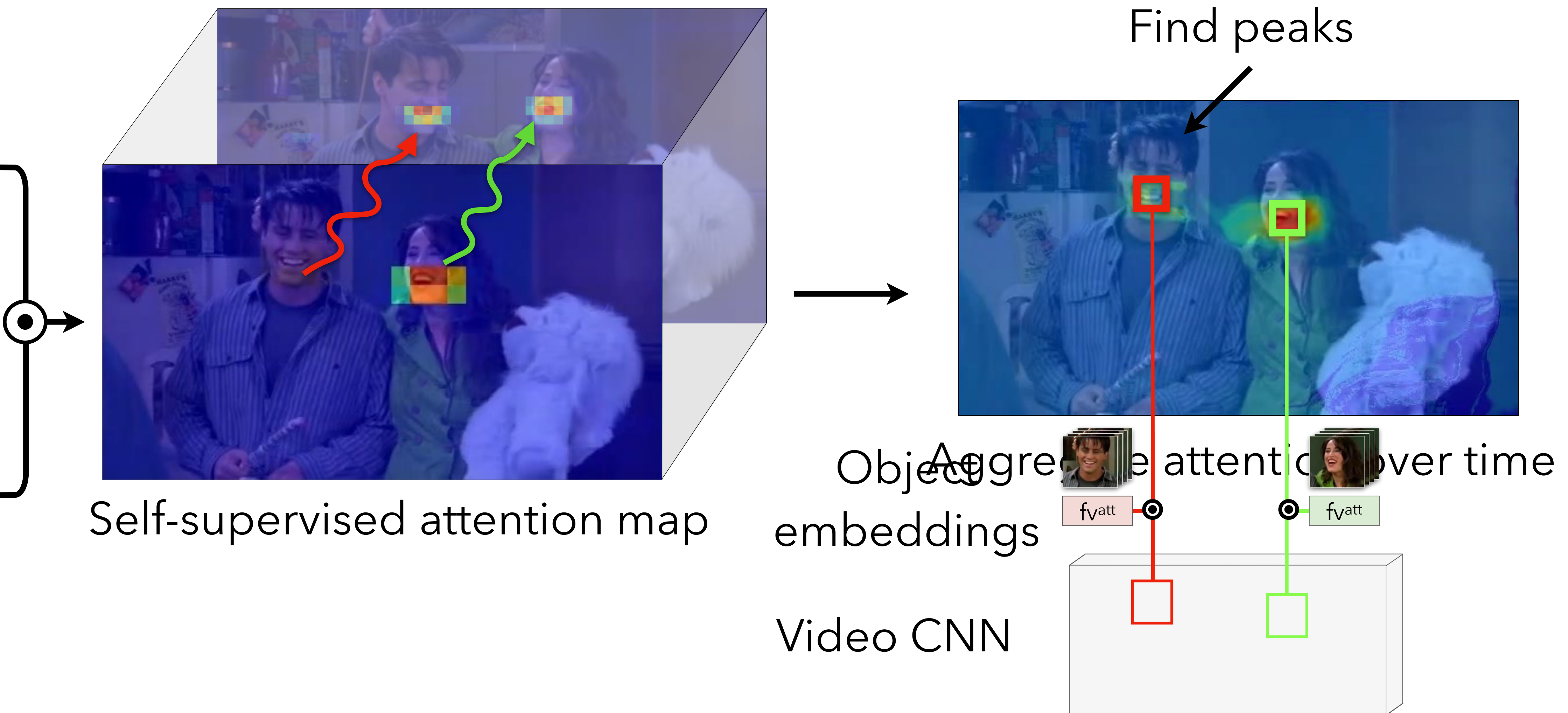
Chopping wood

Audio-visual sound localization

Space-time correspondence with optical flow



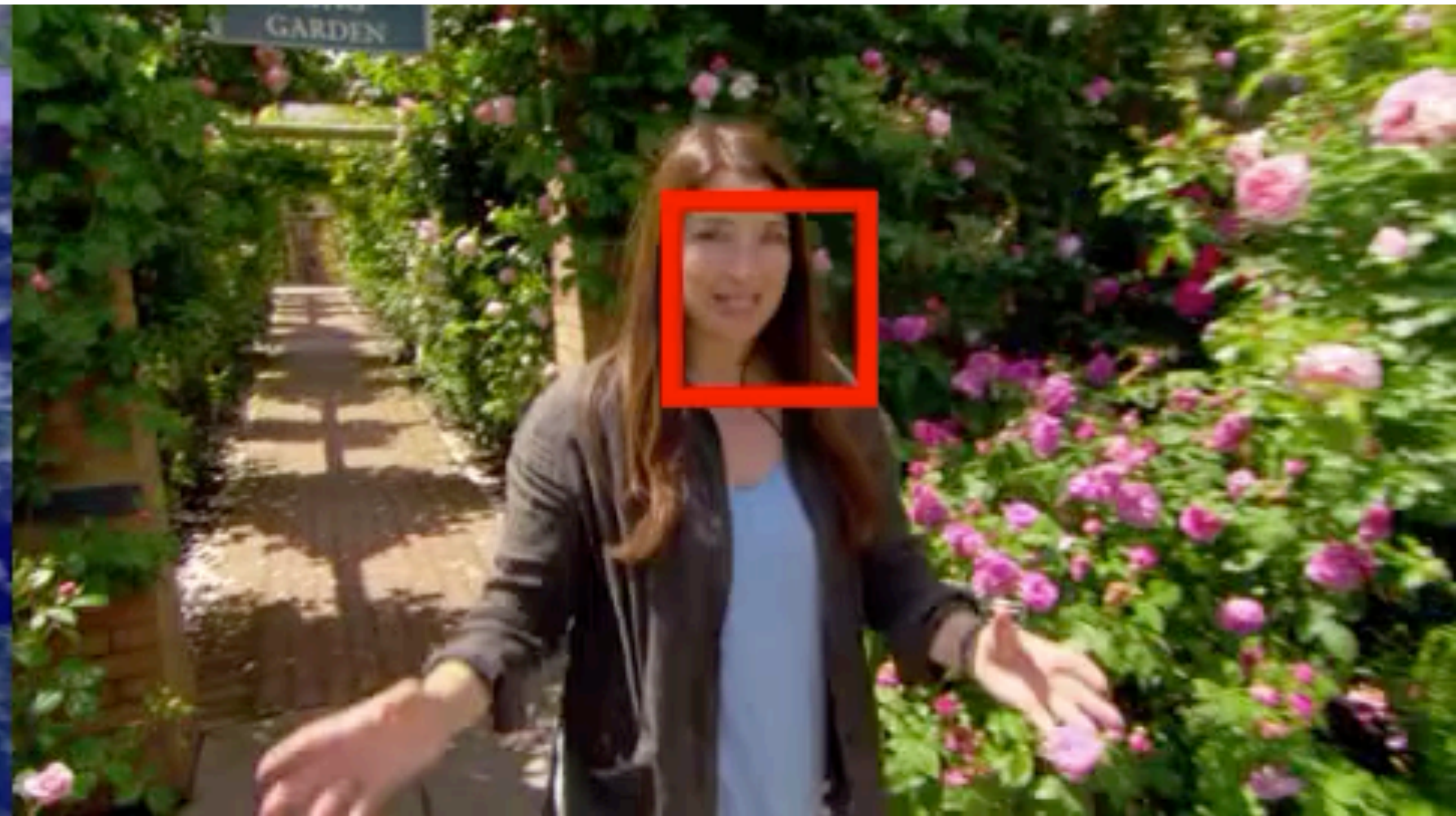
Grouping into distinct sound sources



Detecting sound sources



Attention map



Audio-visual object

Active Speaker Detection



Examples from the *Friends* series

Blue = active speaker
Red = inactive speaker

Other applications of audio-visual models

Application: on/off-screen source separation



Cocktail party problem



a

Sound mixture

=

b

On-screen

+

c

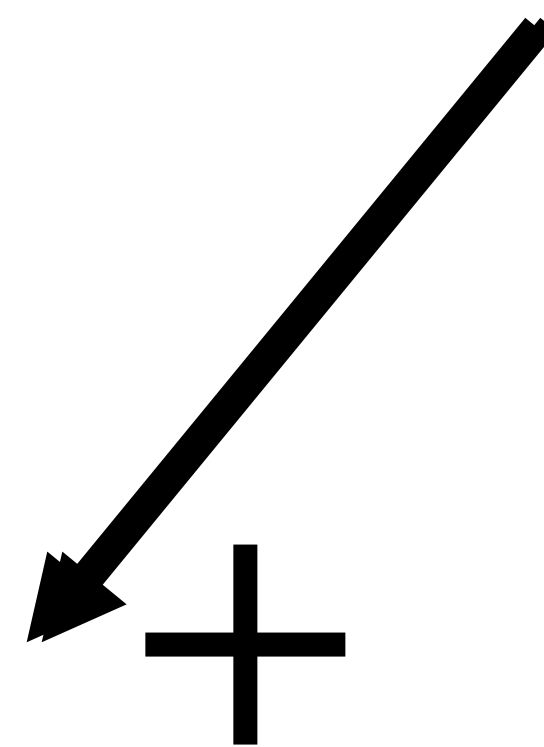
Off-screen

Underdetermined!

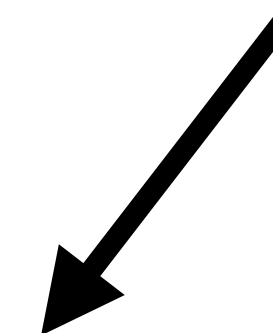
Learning from synthetic sound mixtures



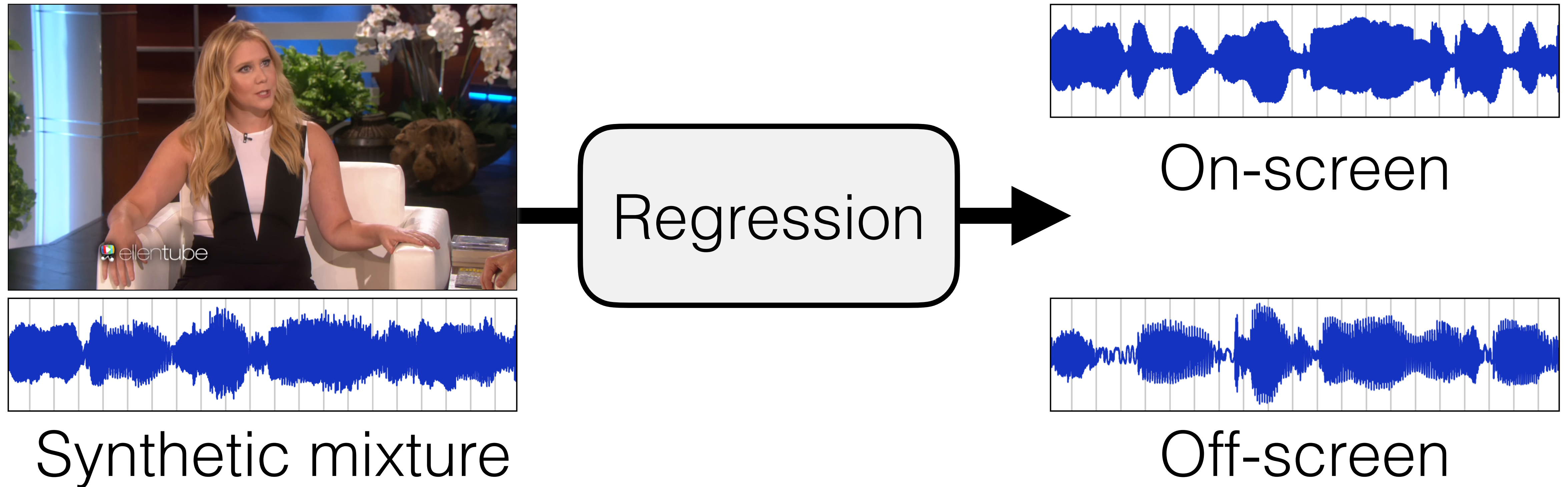
Synthetic mixtures



+



Learning from synthetic sound mixtures



Related work: audio-only separation (Hershey 2016)

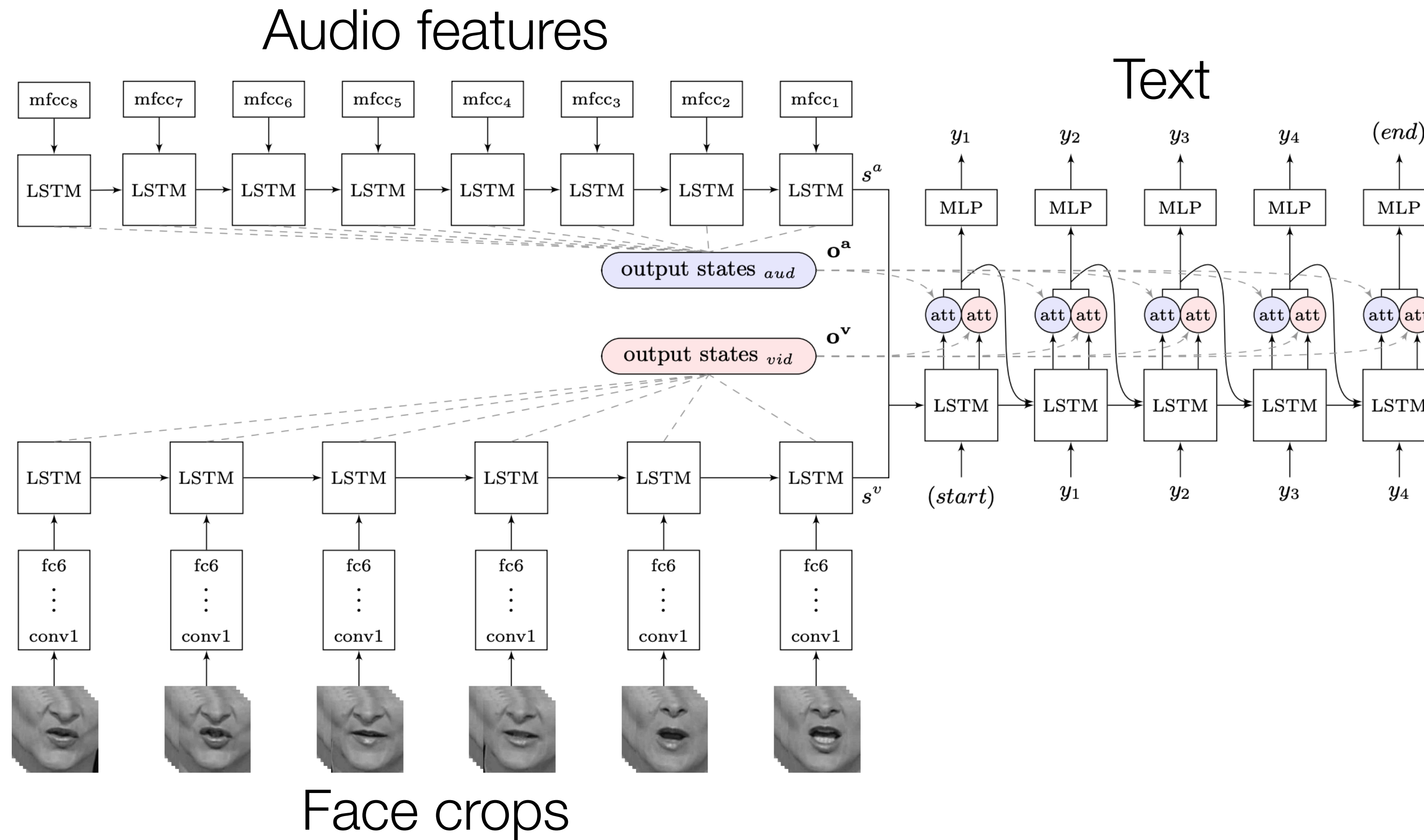


Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation

Ariel Ephrat Inbar Mosseri Oran Lang Tali Dekel Kevin Wilson
Avinatan Hassidim William T. Freeman Michael Rubinstein

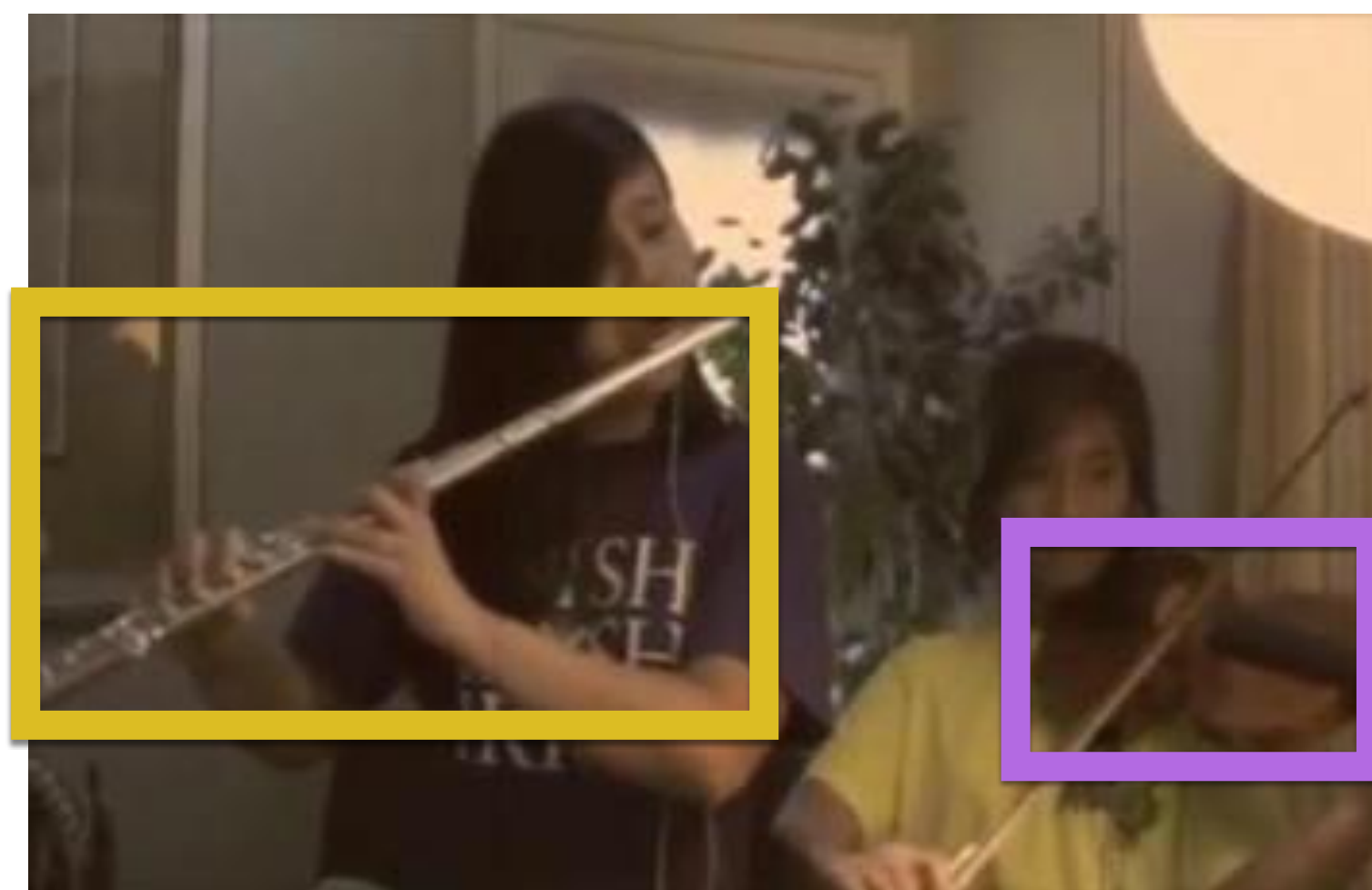
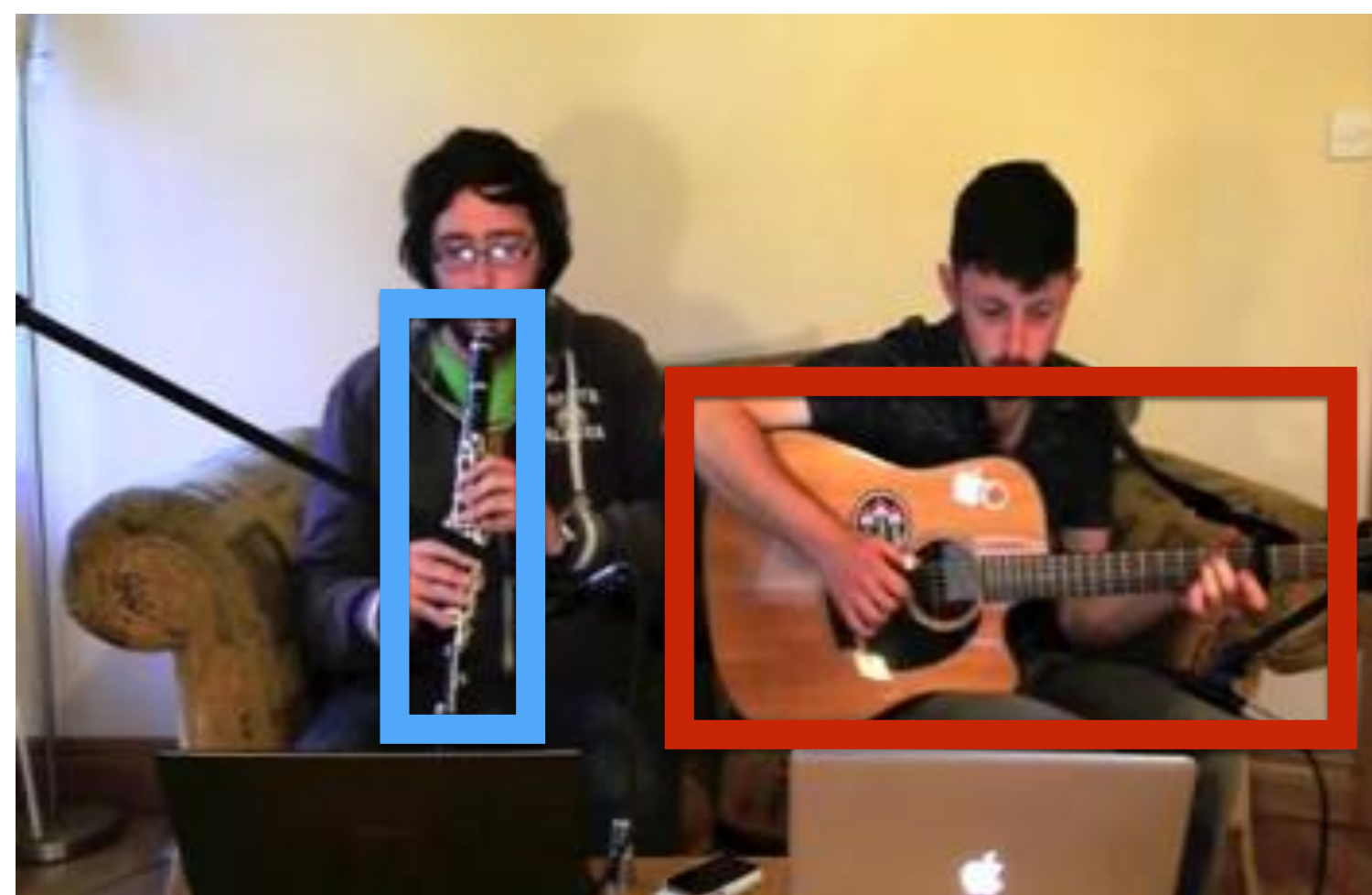
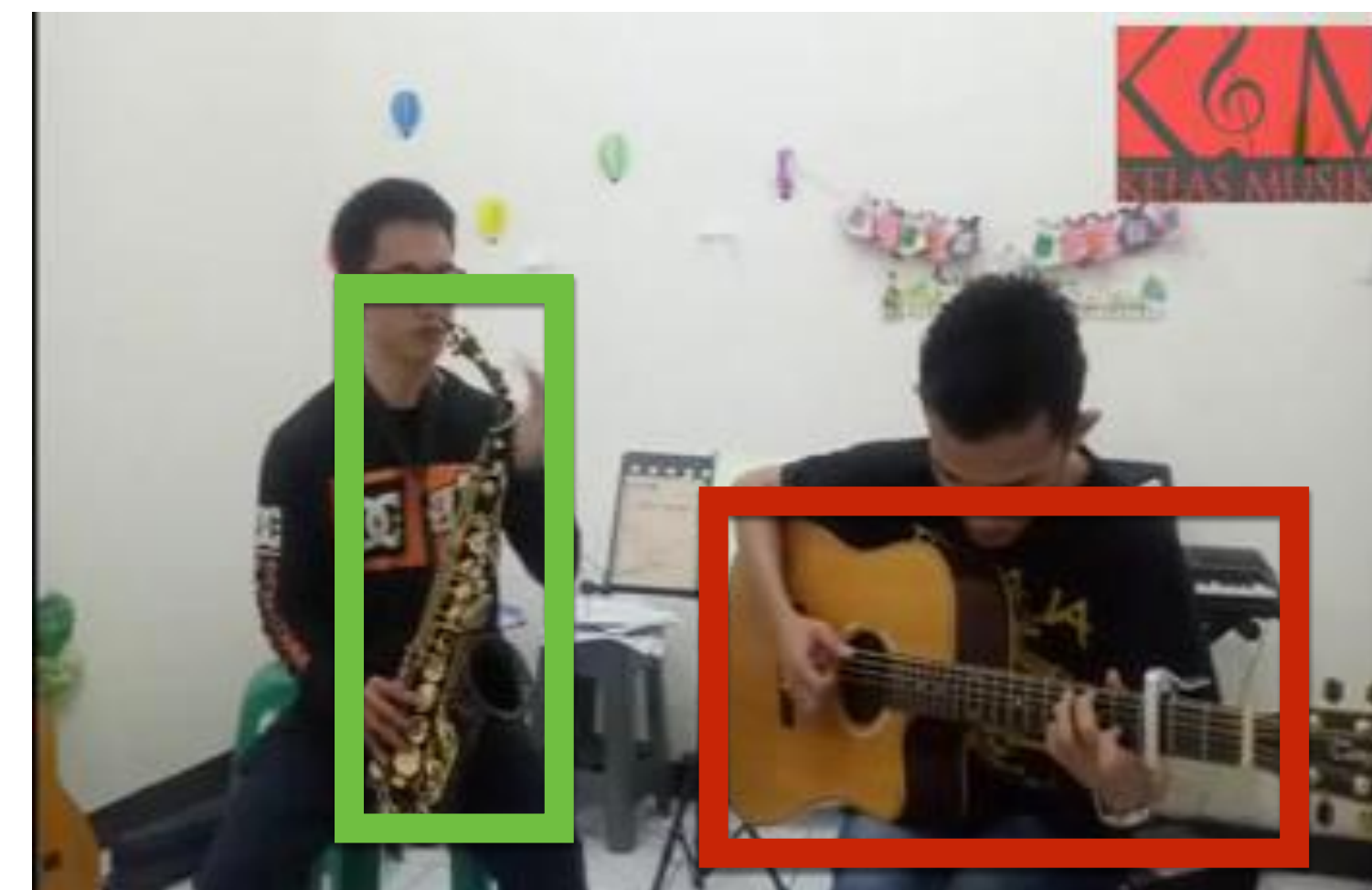
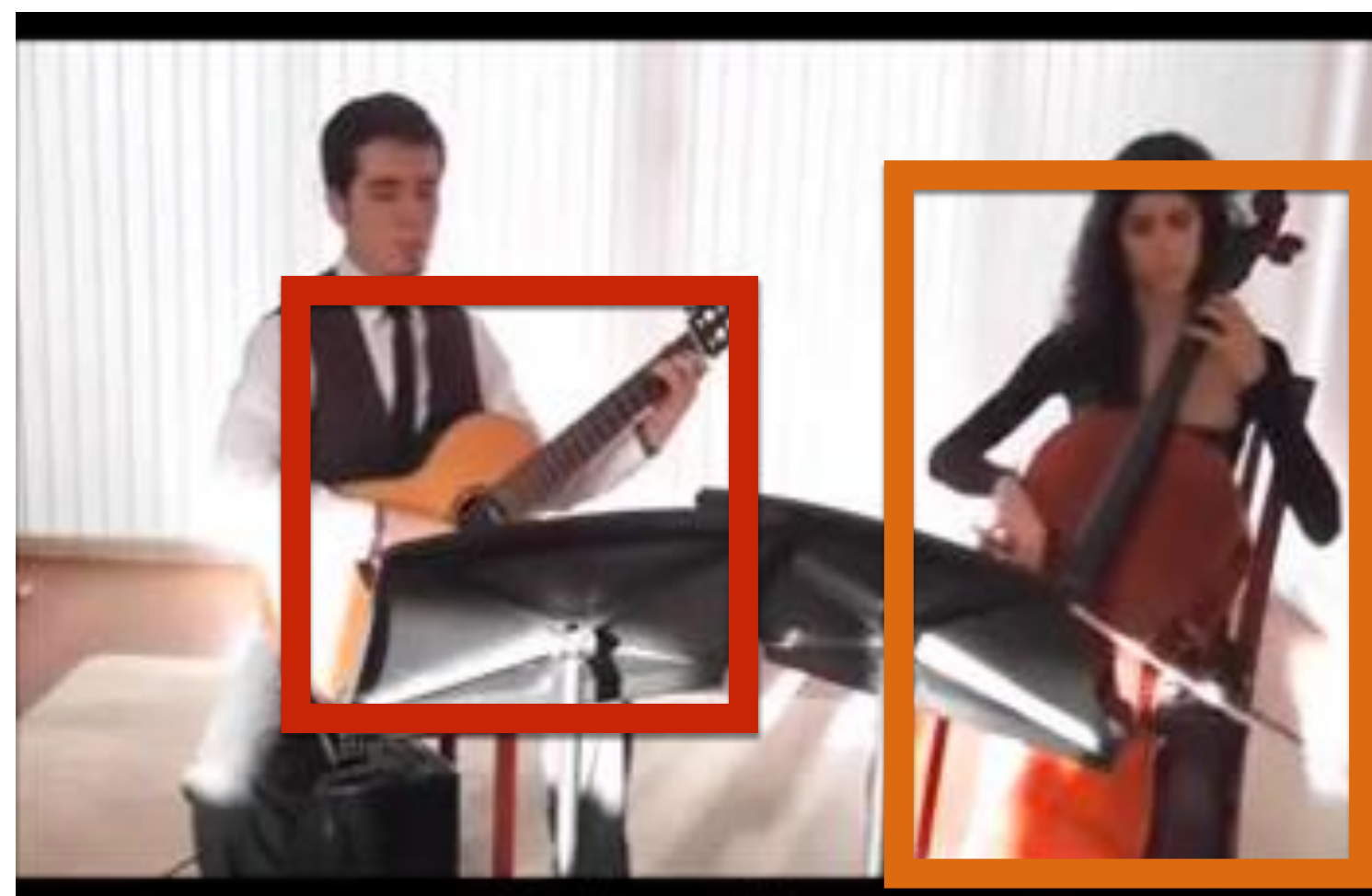
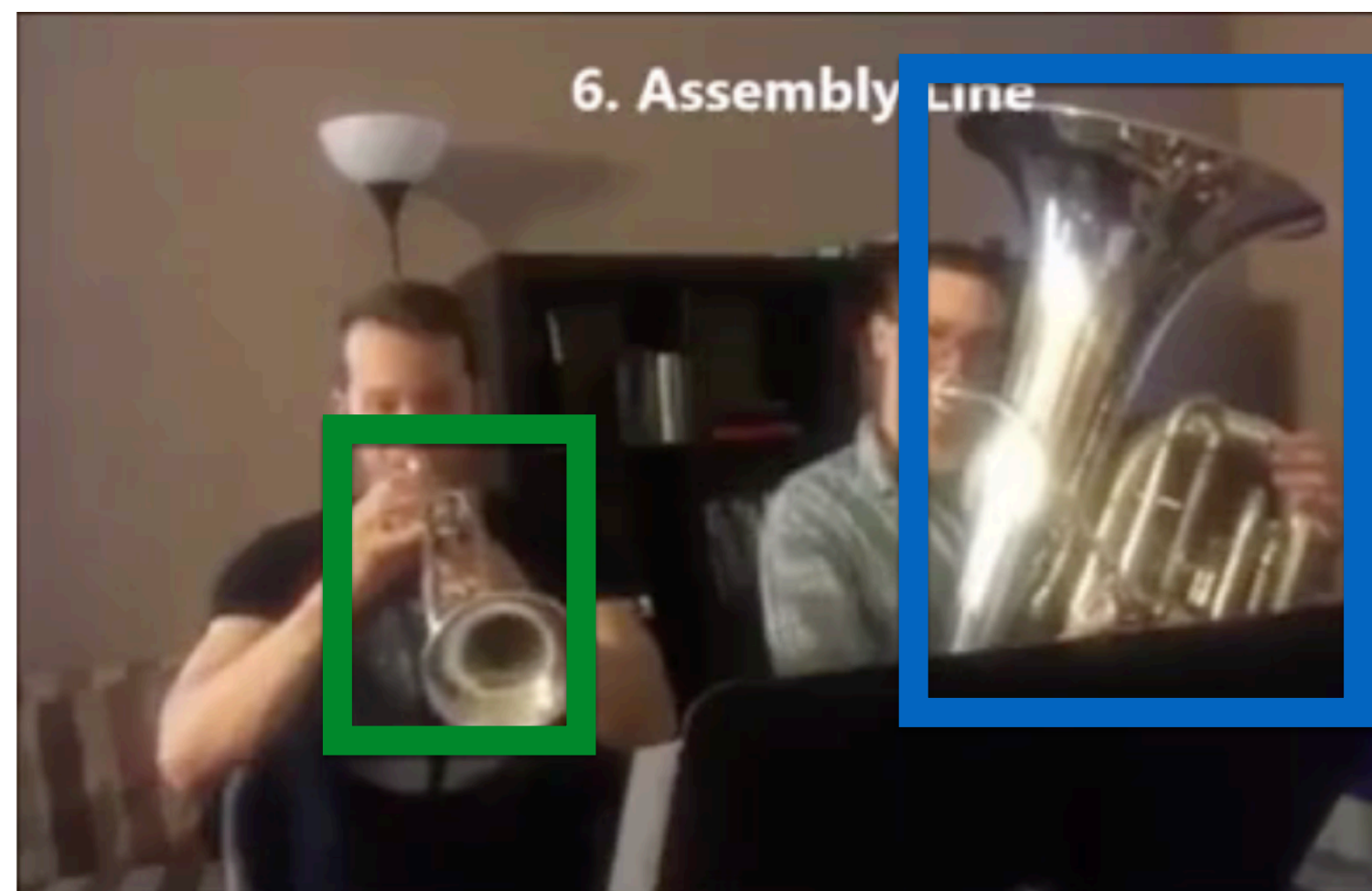
Sports debate

Lip reading



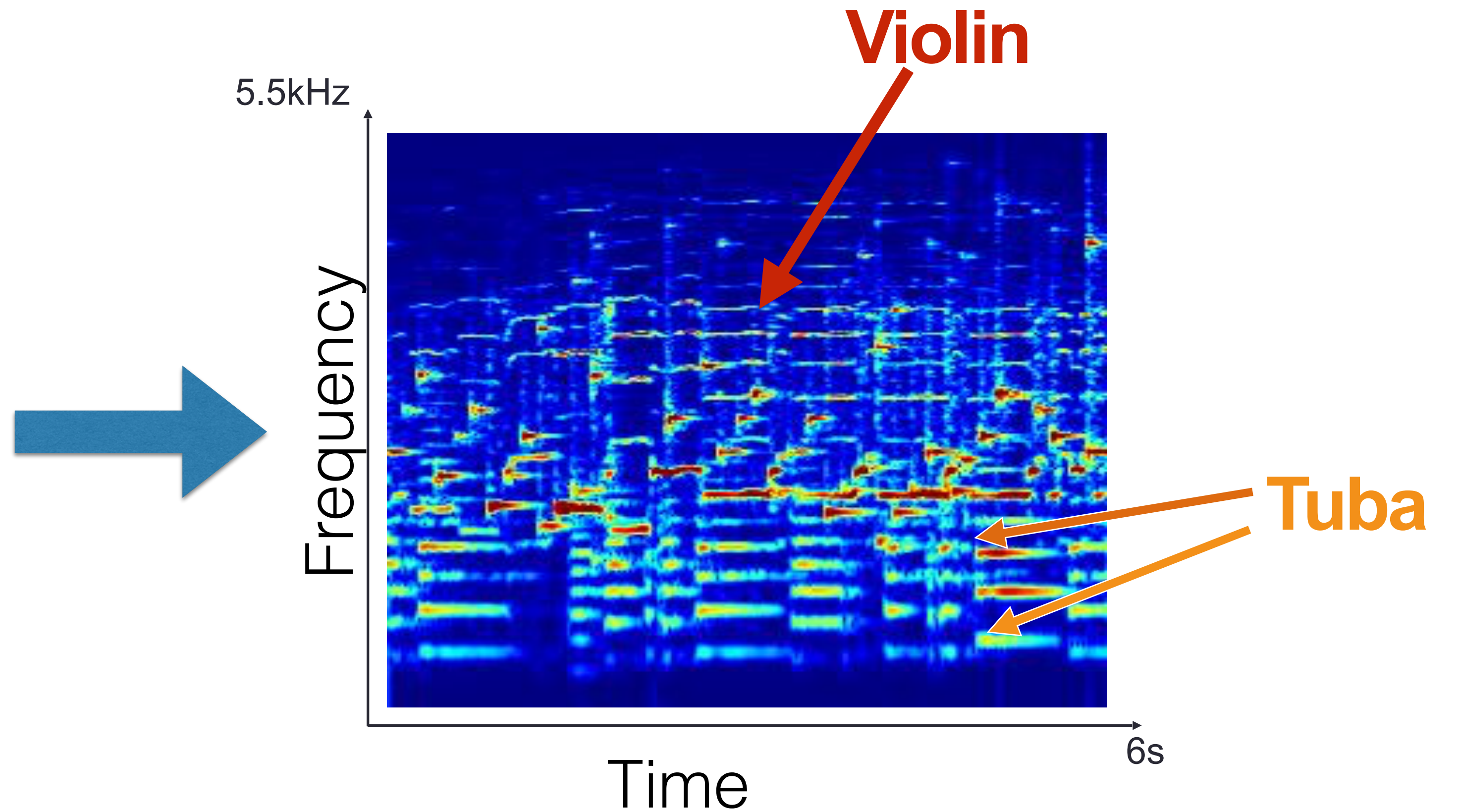
The video-to-text alignment is also
generated by the model.

Discovering musical instruments

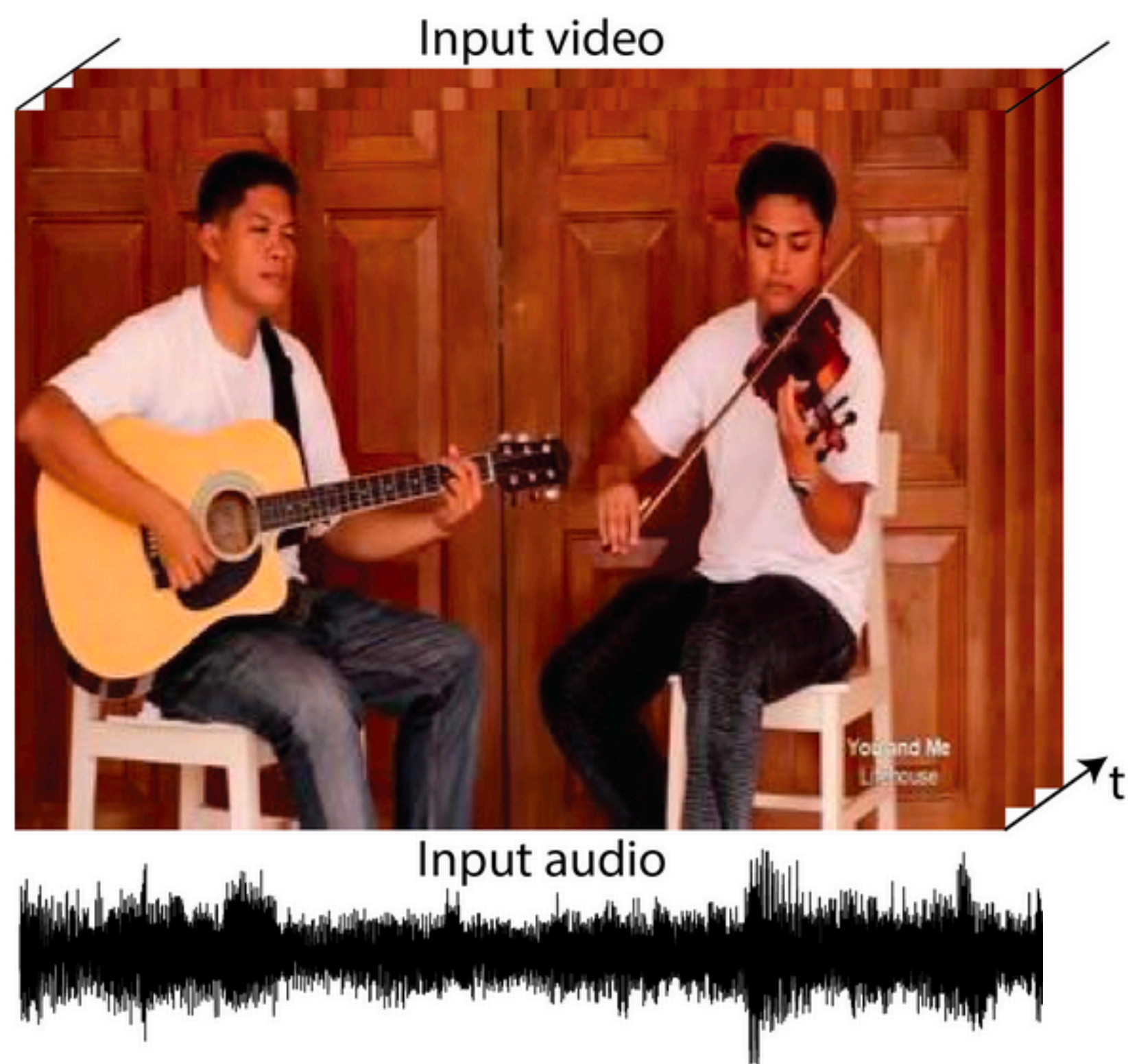


[Zhao et al., “The Sound of Pixels”, 2018]

Musical instrument separation



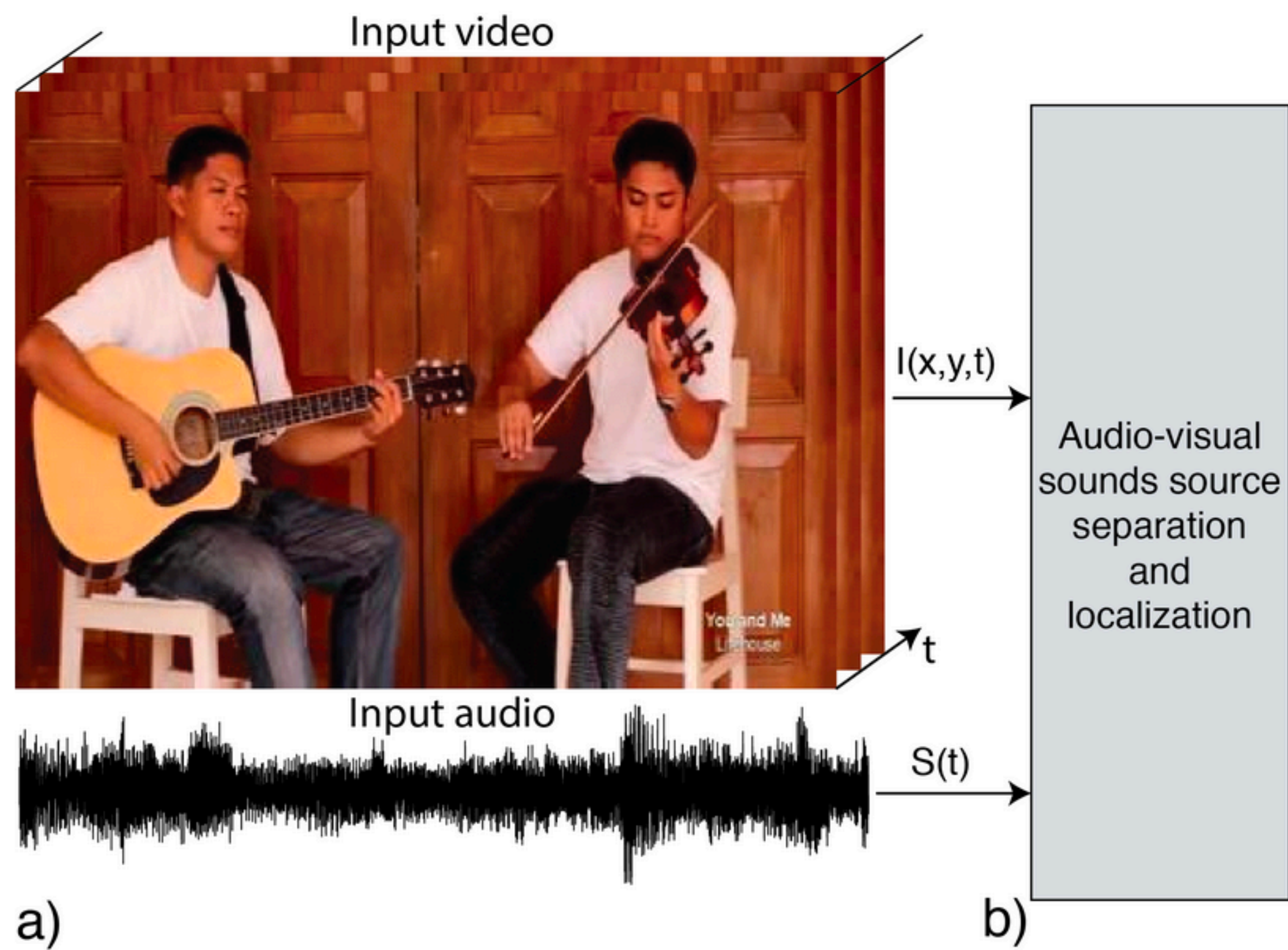
[Zhao et al., “The Sound of Pixels”, 2018]

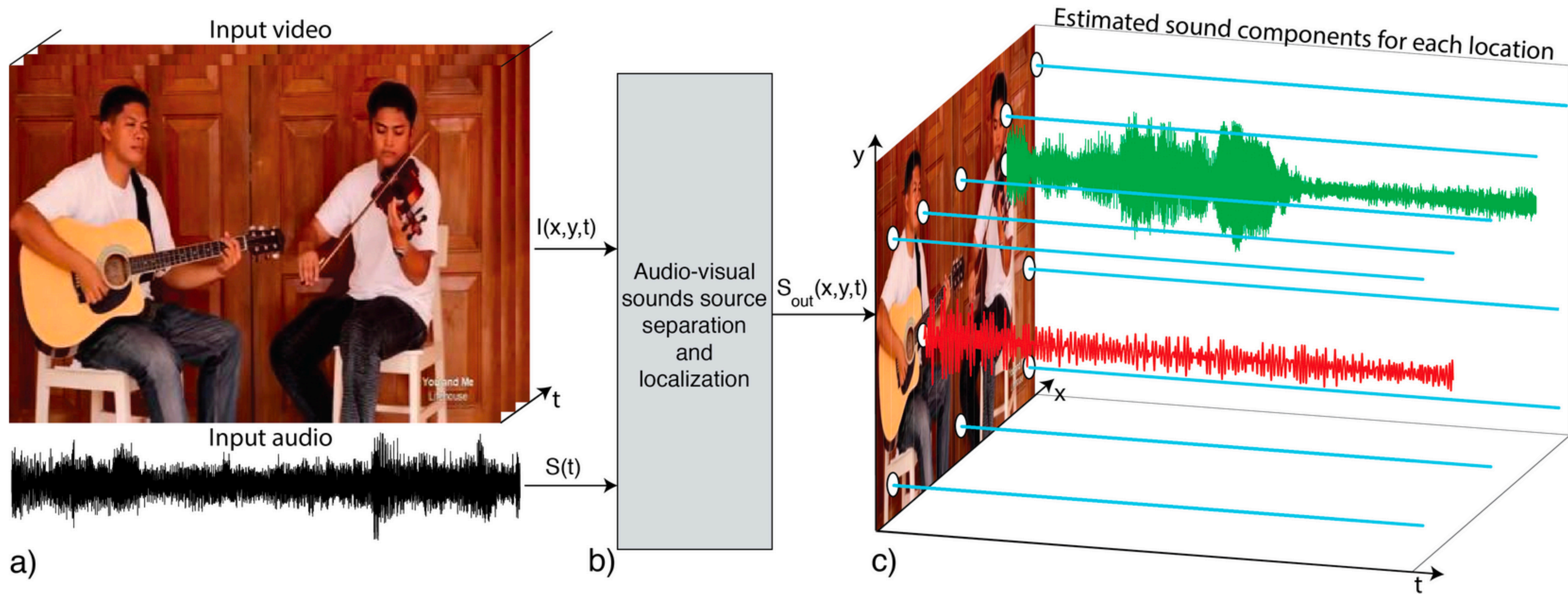


a)

<http://sound-of-pixels.csail.mit.edu/>

[Zhao et al., “The Sound of Pixels”, 2018]







What does sound tell you about scenes?

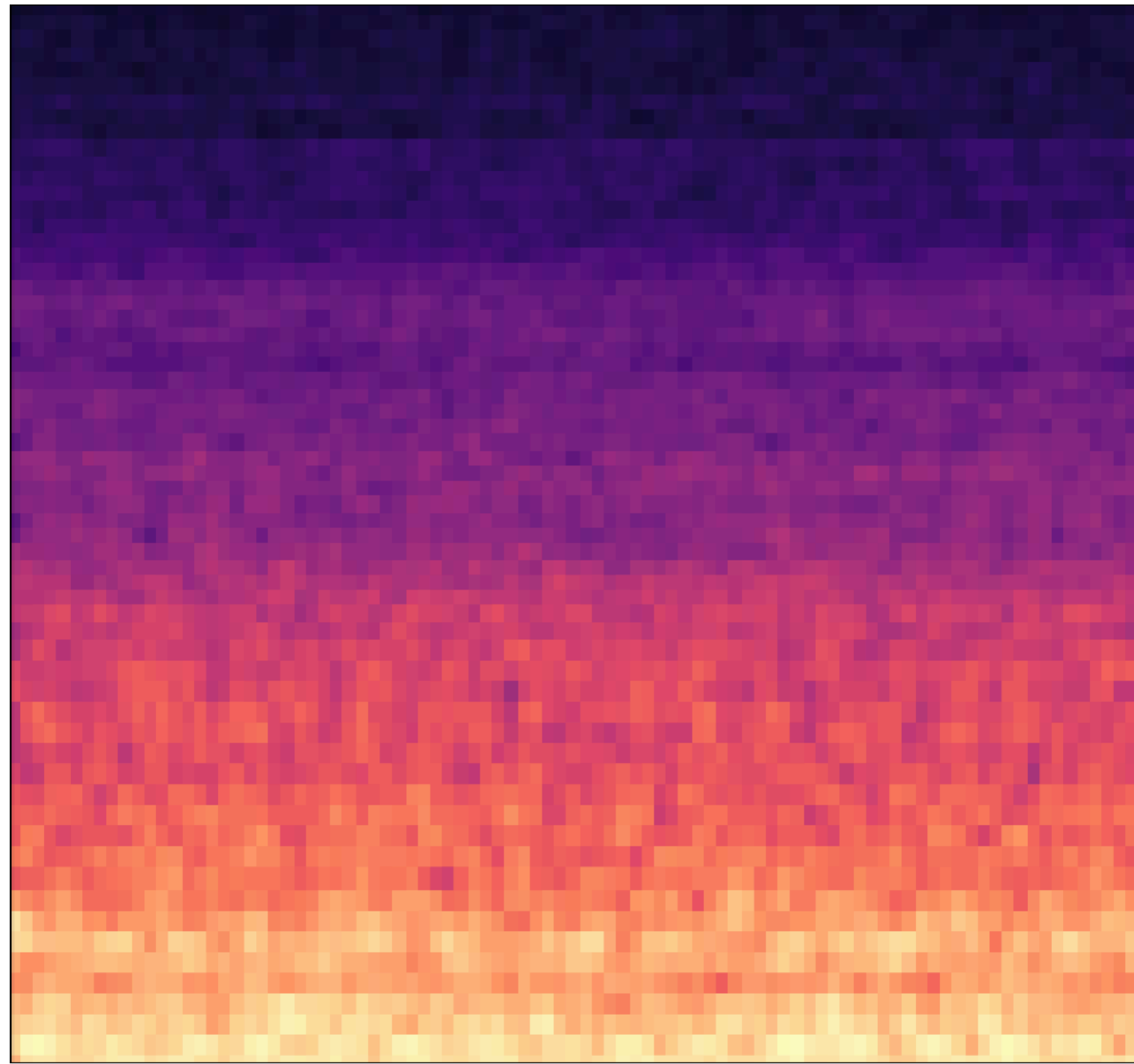
Sound conveys scene structure

Kitchen

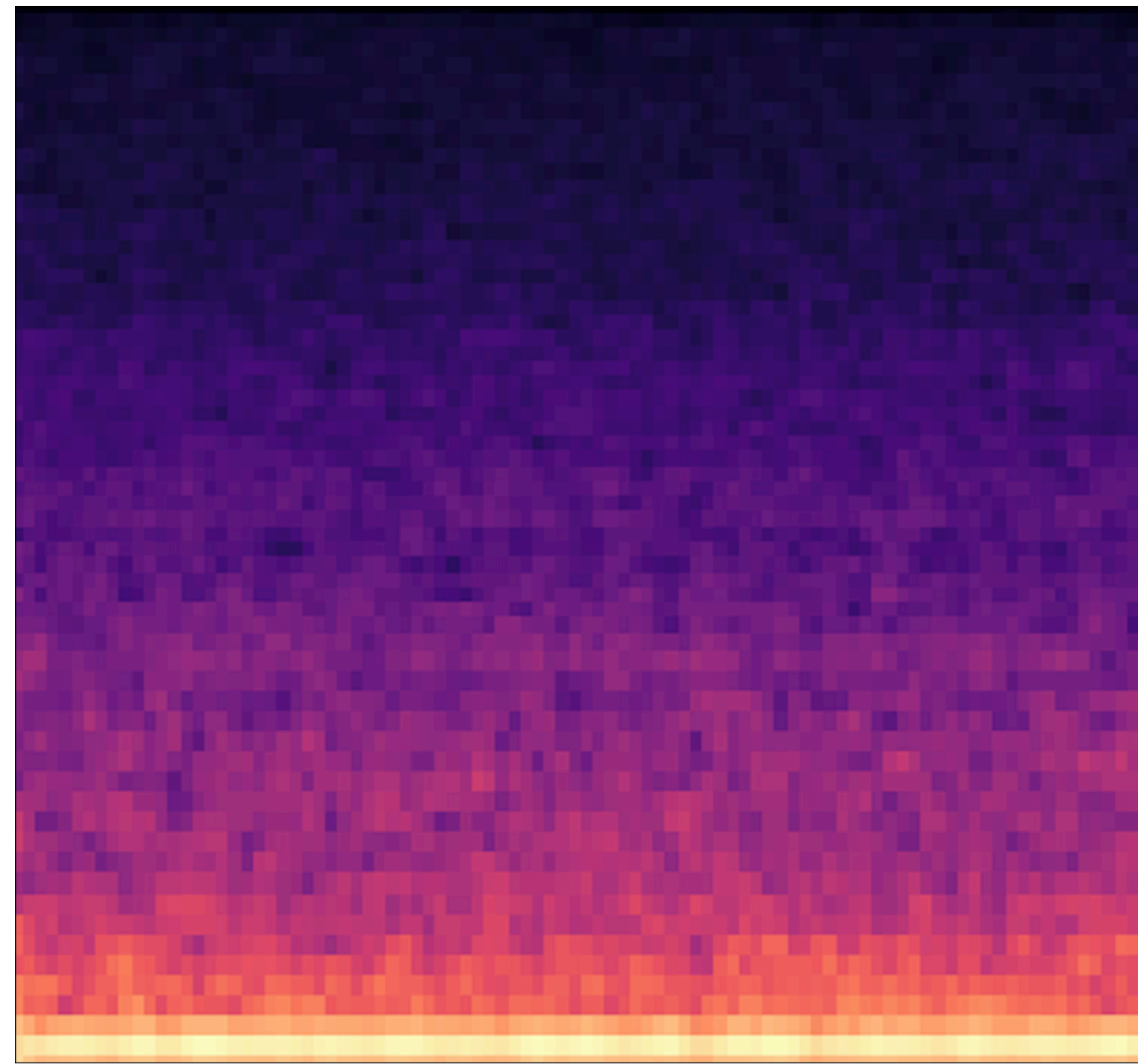
Vent

Window

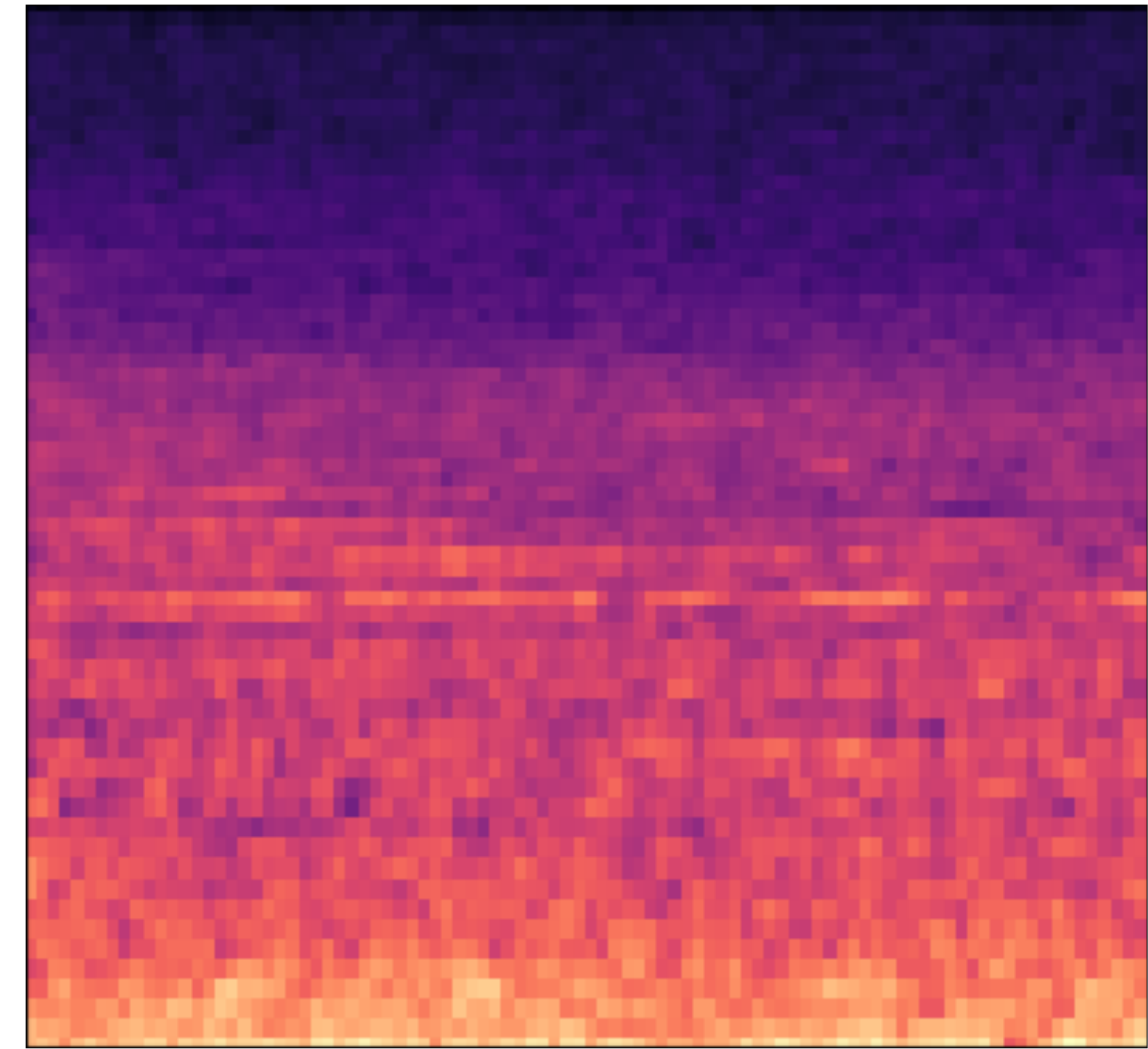
Frequency



Time



Time



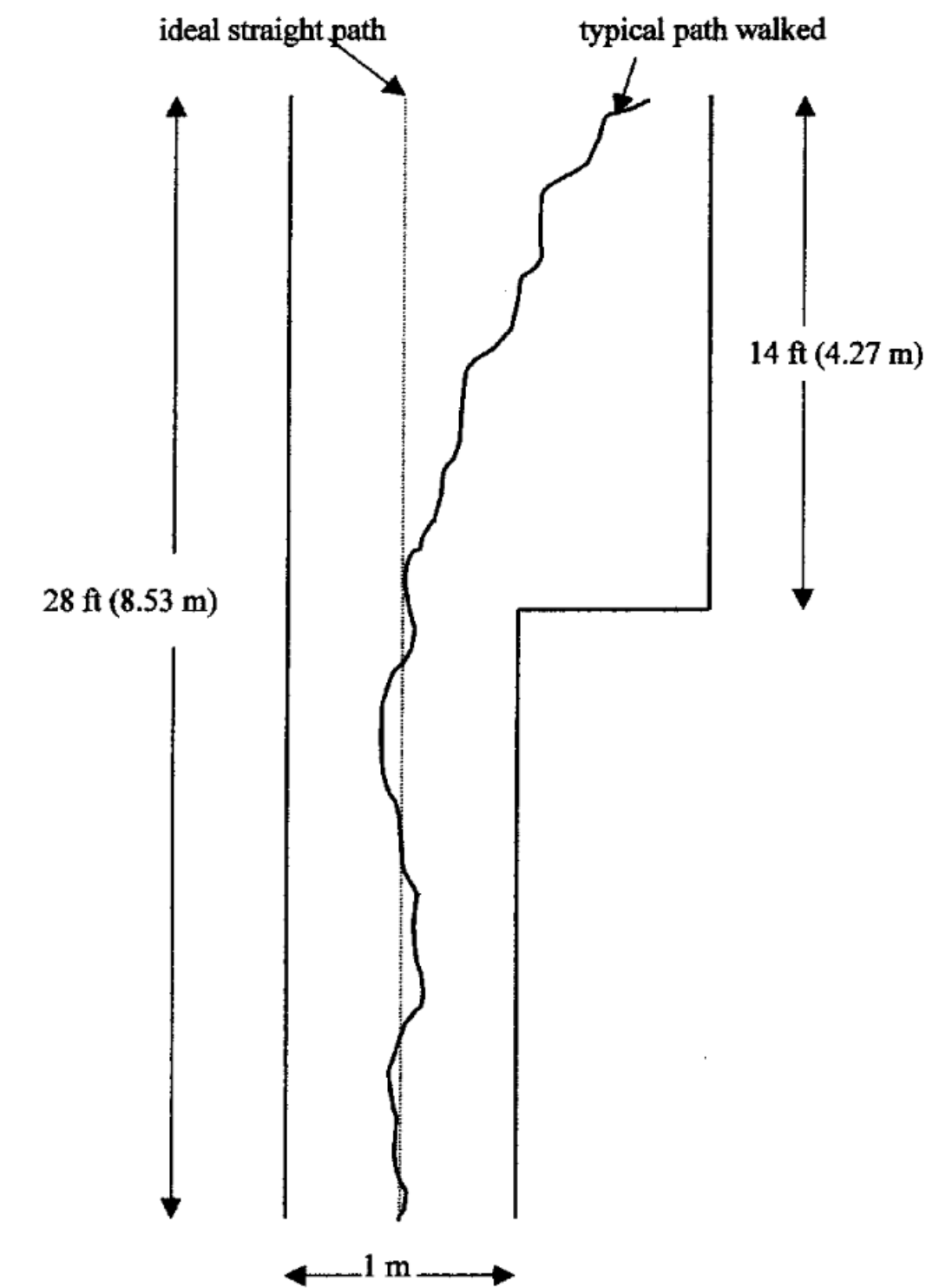
Time

Audio-based navigation

Echolocation Reconsidered: Using Spatial Variations in the Ambient Sound Field to Guide Locomotion

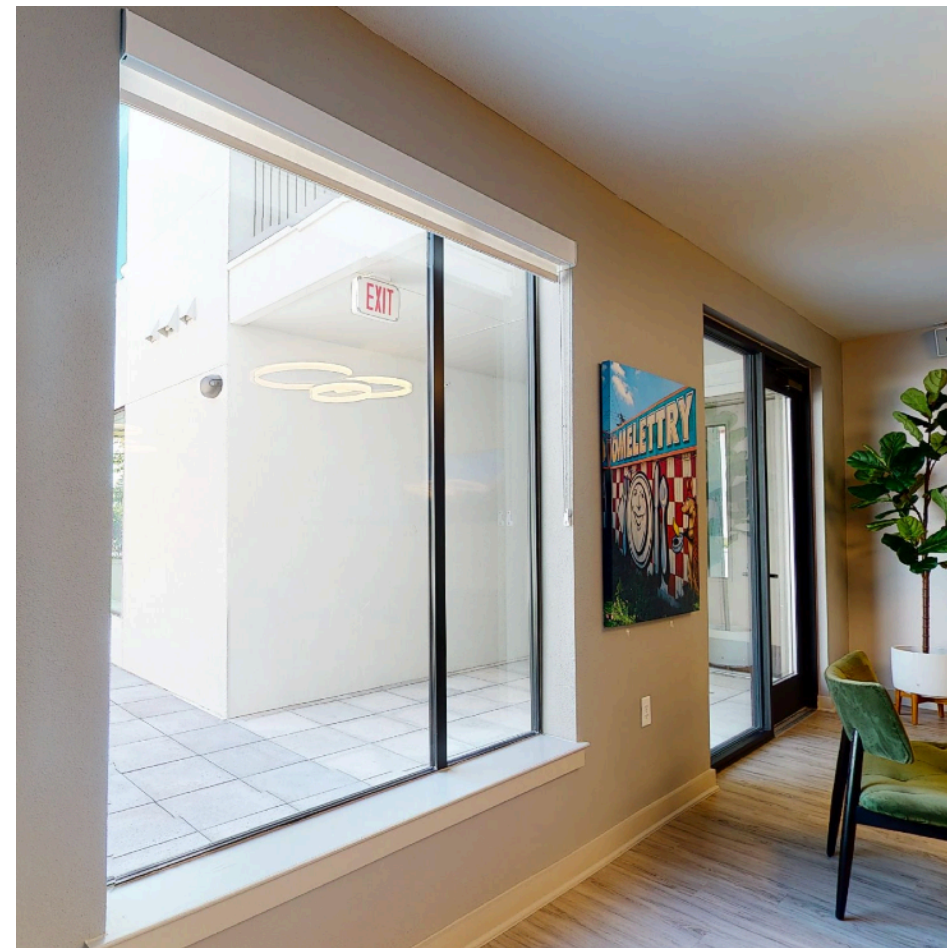
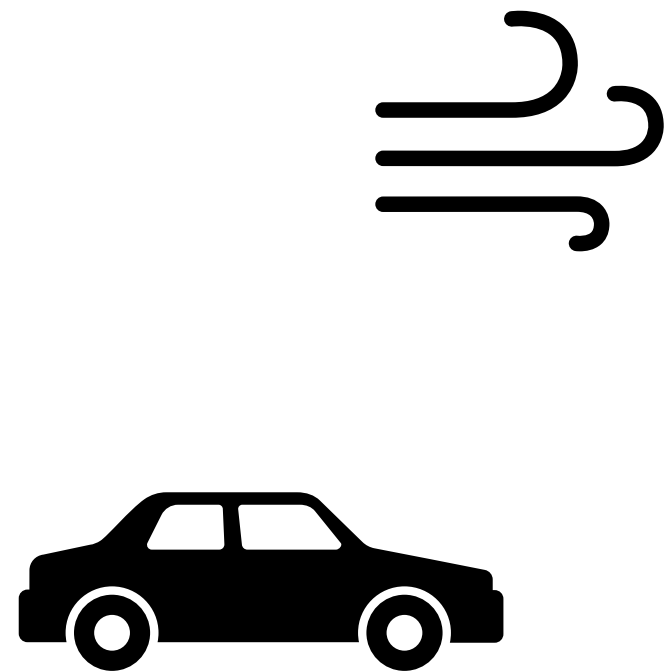
Daniel H. Ashmead, Robert S. Wall, Susan B. Eaton, Kiara A. Ebinger, Mary-Maureen Snook-Hill, David A. Guth, and Xuefeng Yang

Abstract: This article presents an acoustical model and evidence from four experiments that children with visual impairments use the buildup of low-frequency sound along walls to guide locomotion. The model differs from the concept of echolocation by emphasizing sound that is ambient, rather than self-produced, and of low, rather than high, frequency.



Plan of hallway with widened section and example of typical path walked

Ambient sounds convey 3D scene structure

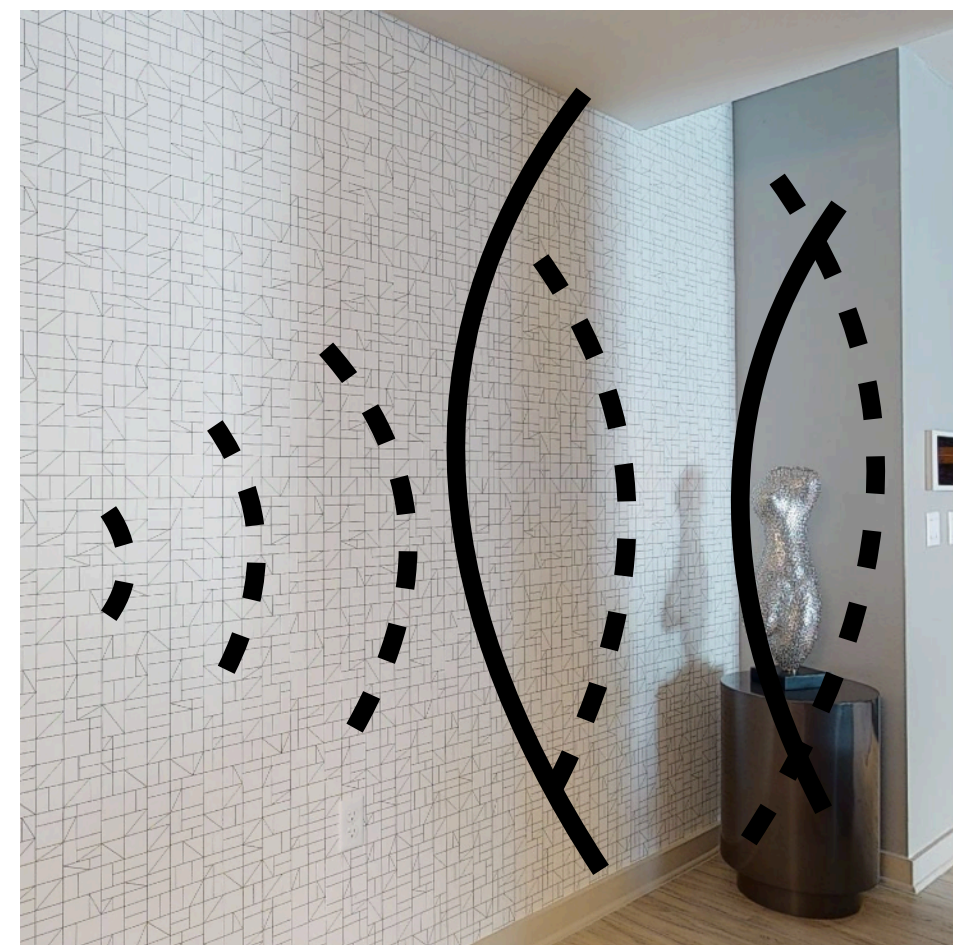


Context



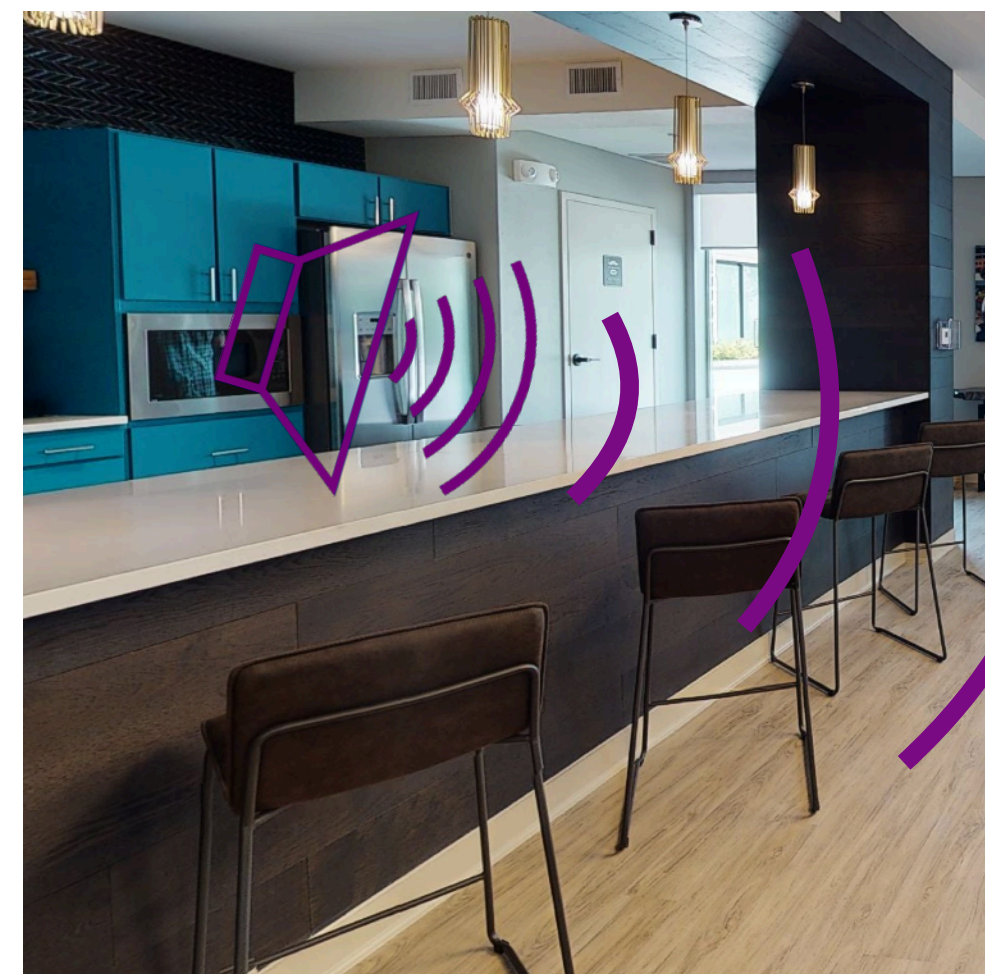
[Traer et al. 2016]

Reverberation



[Ashmead et al. 1998]

Interference

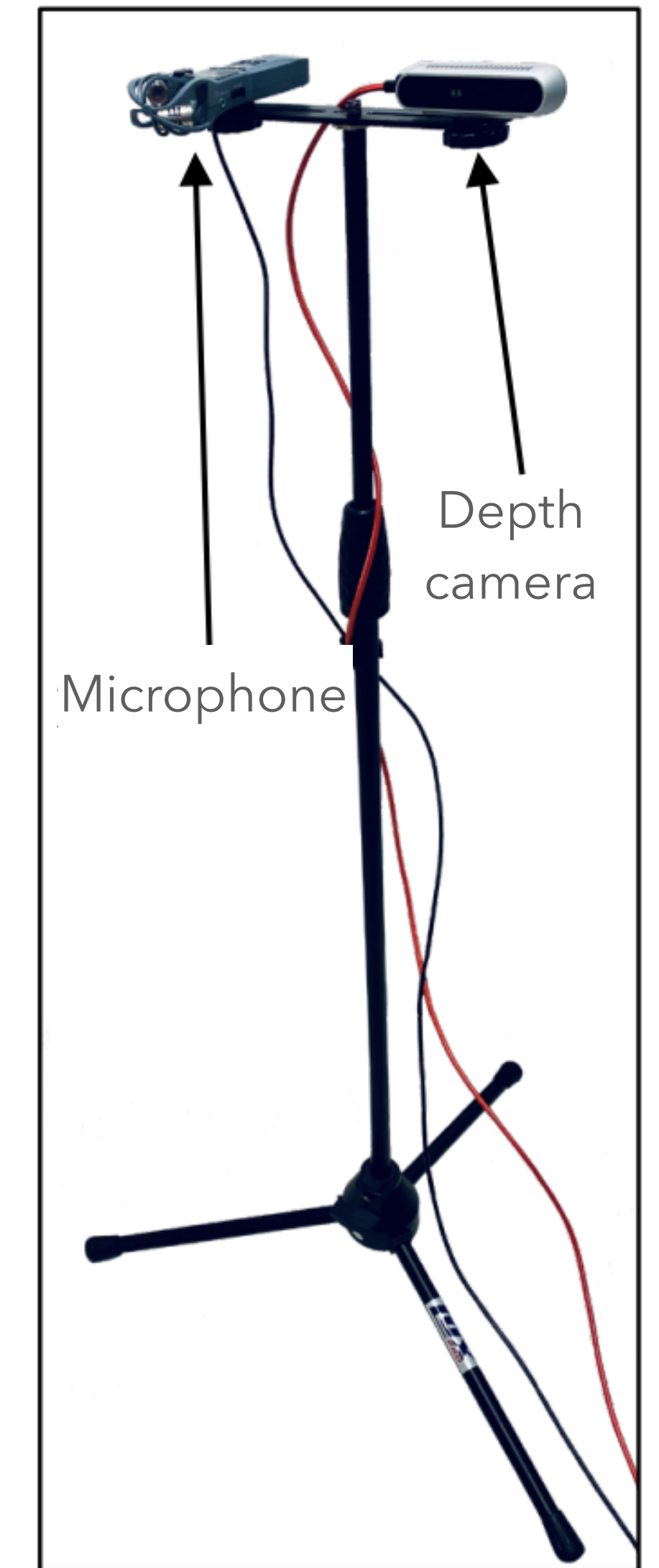
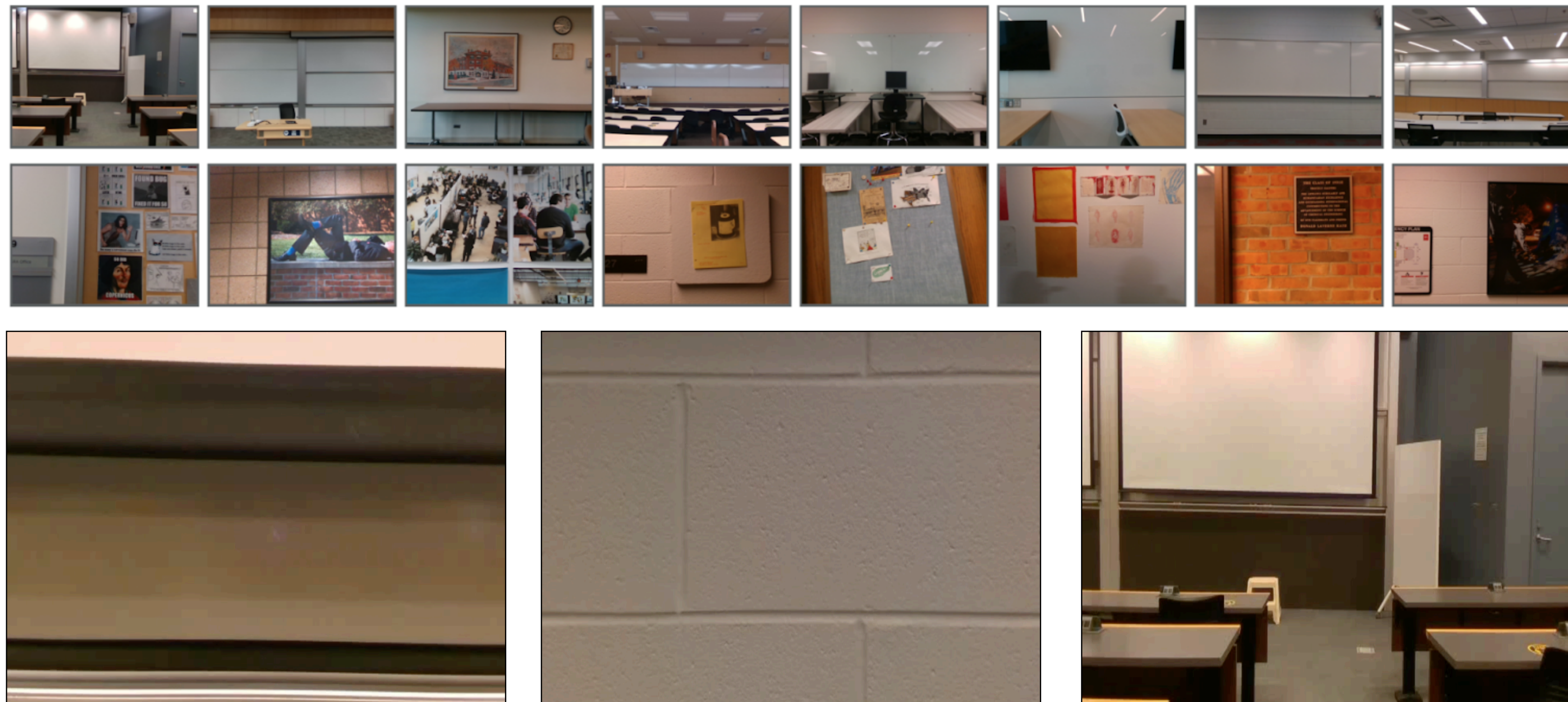


[Rosenblum et al. 2007]

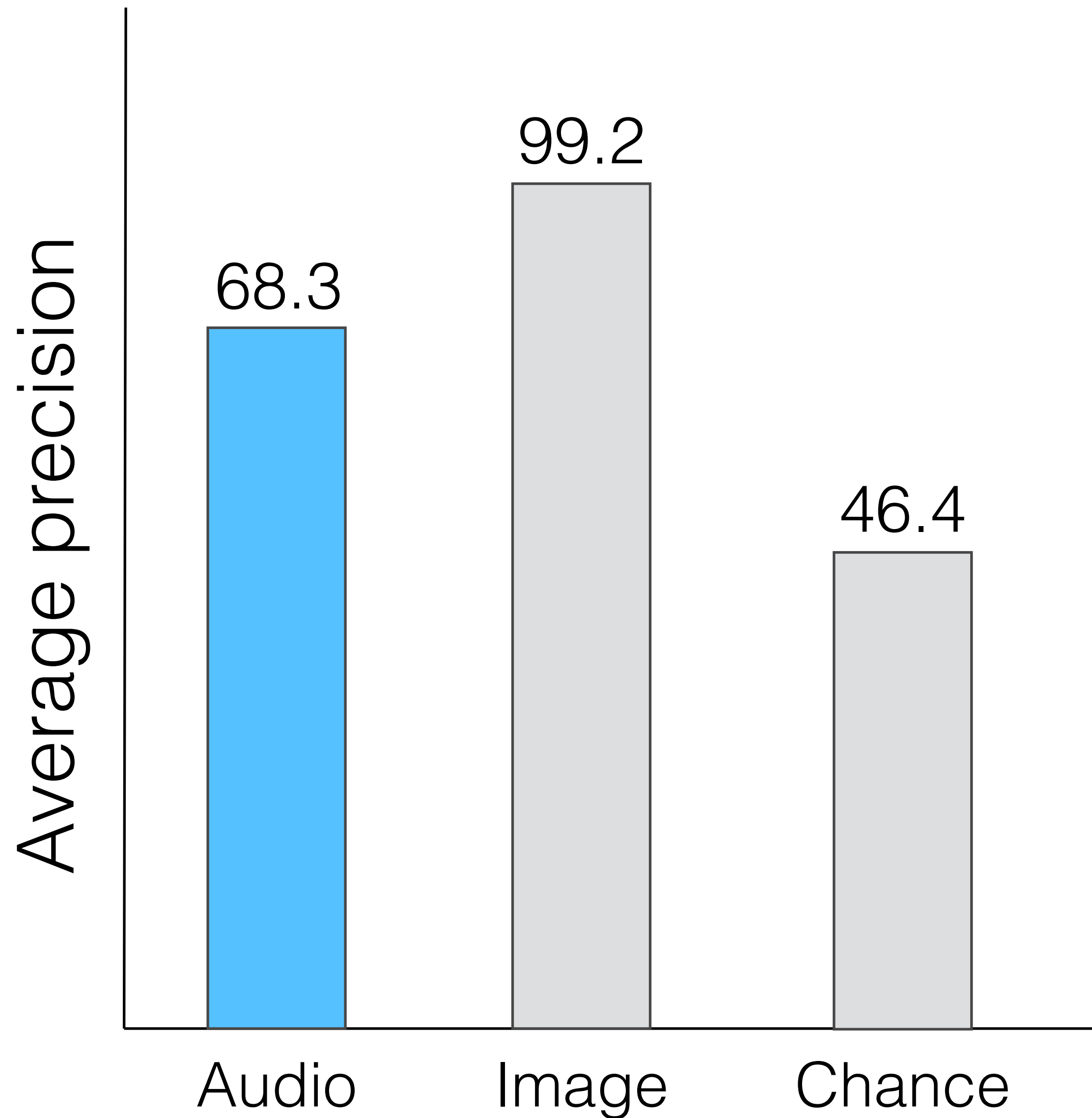
Sound shadow

The *Quiet Campus* Dataset

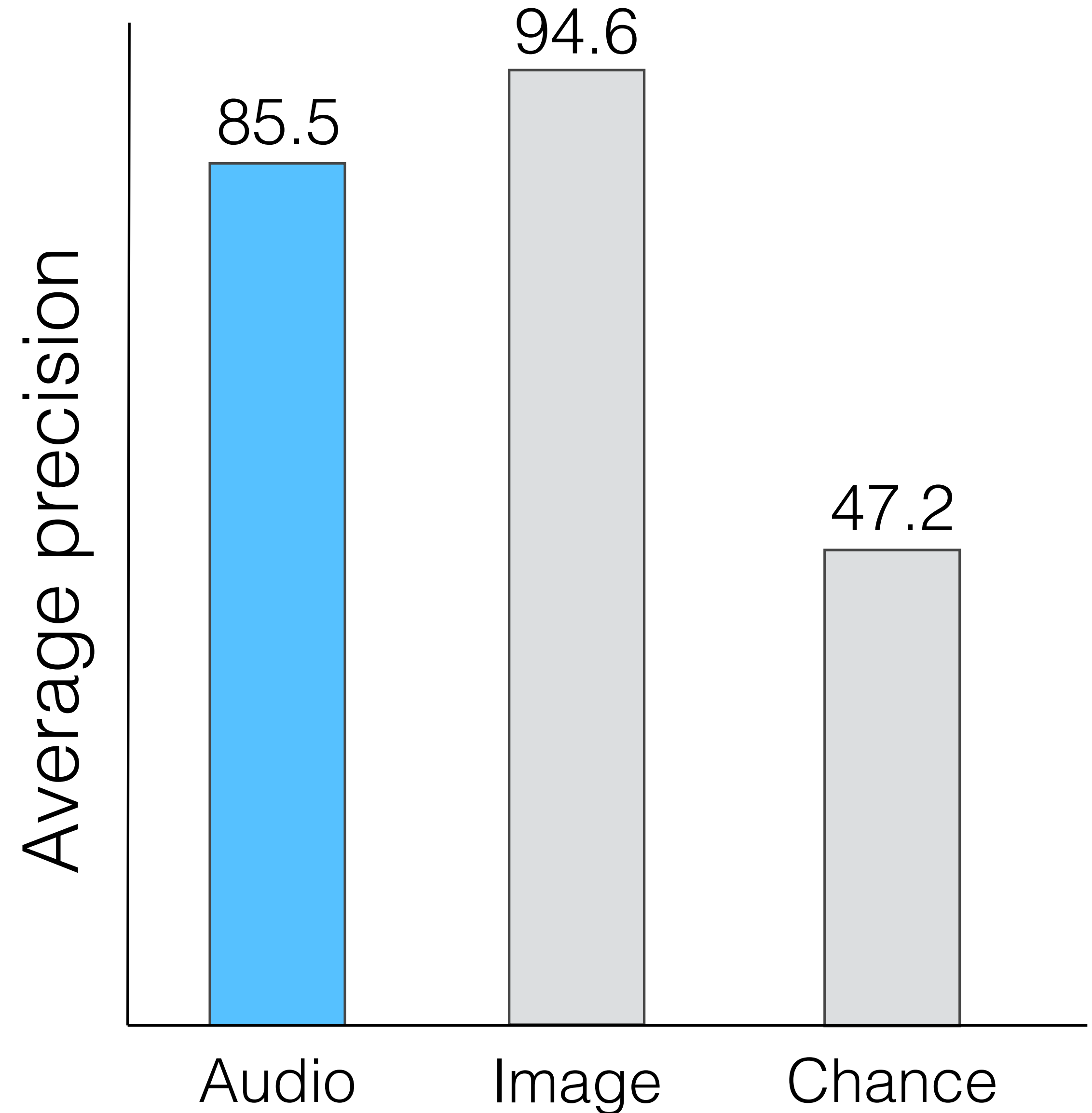
- Sound + RGB-D
- 46 rooms, 12 buildings
- 14 hours



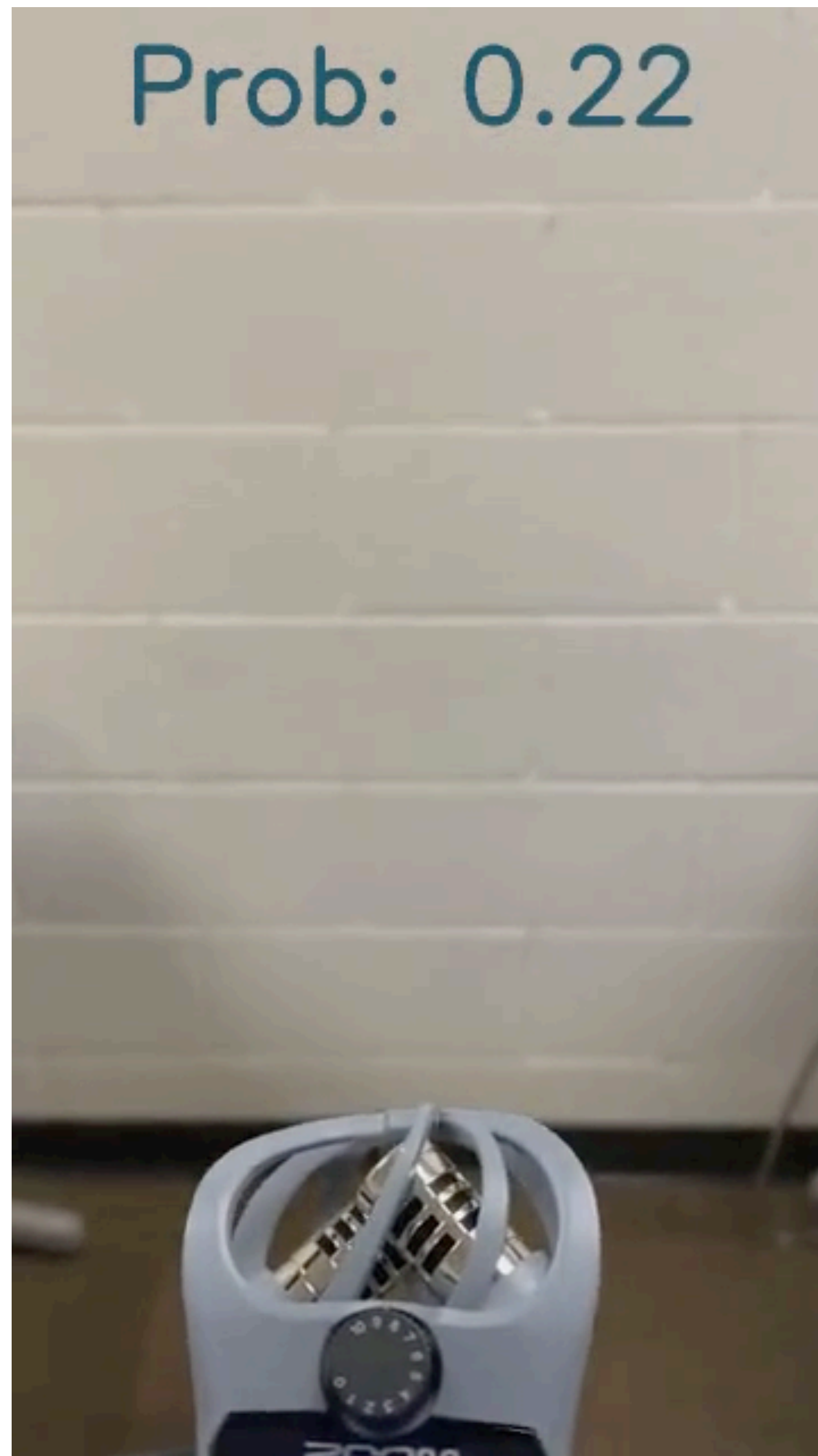
Depth estimation



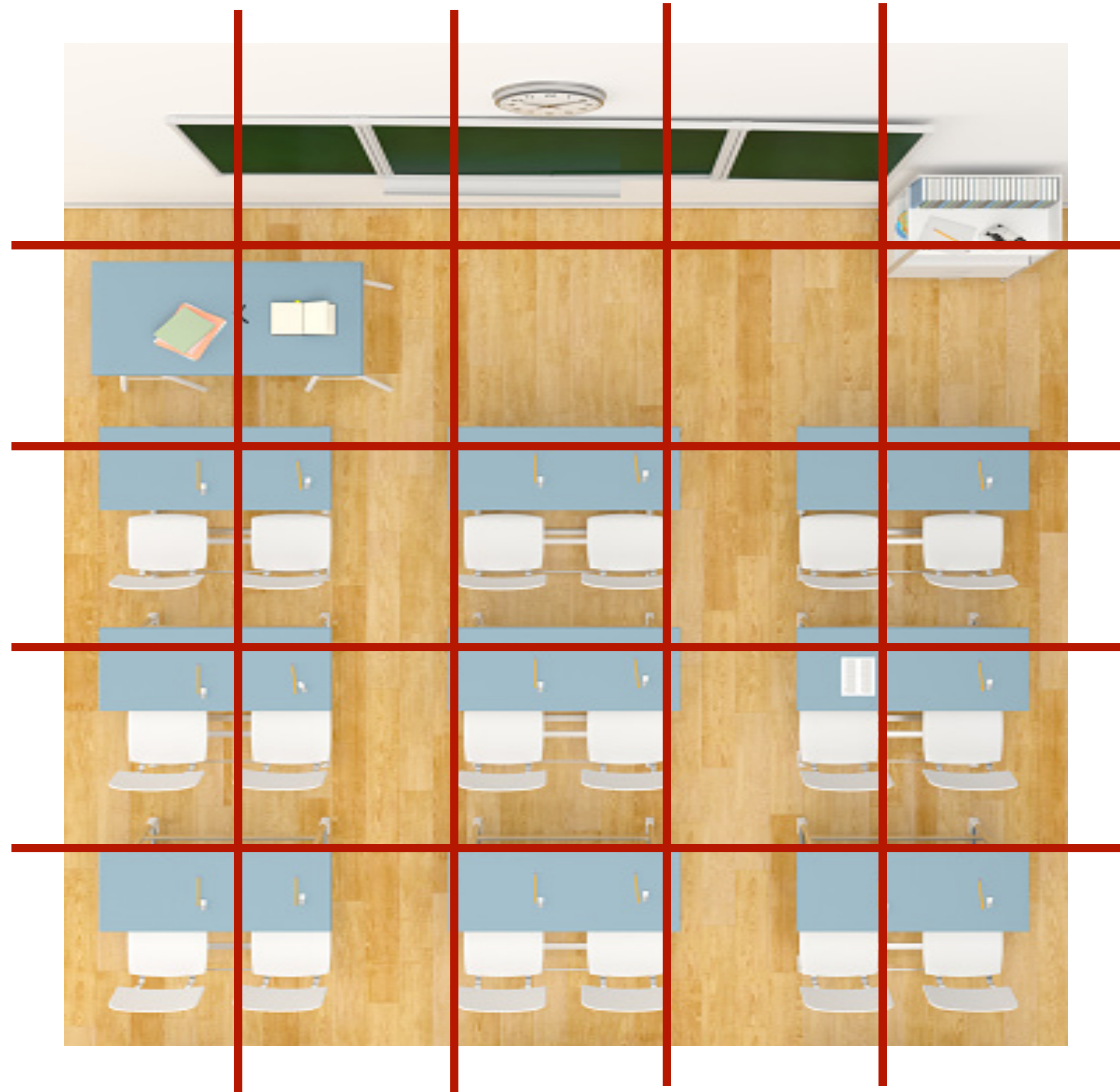
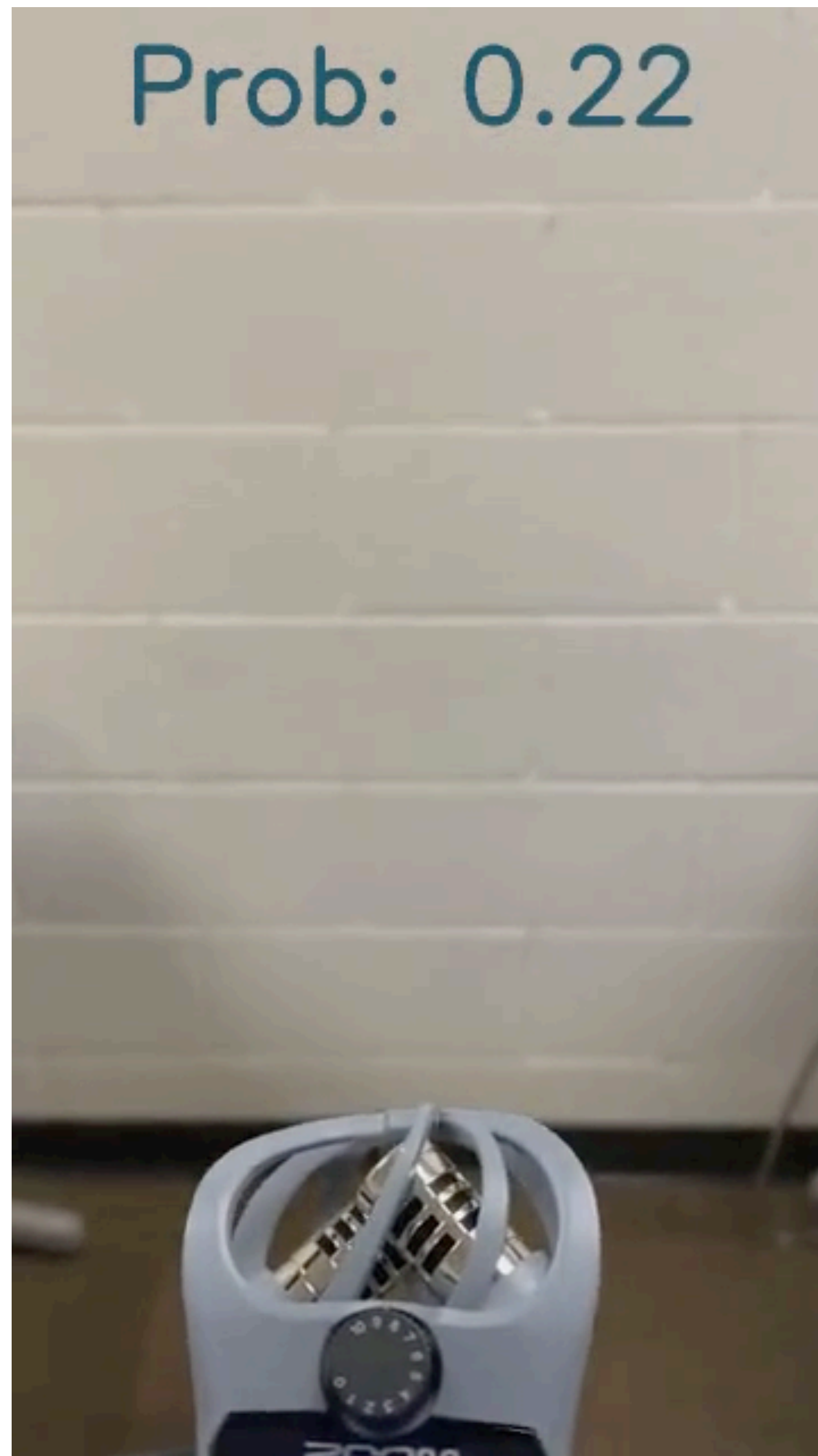
Relative depth estimation



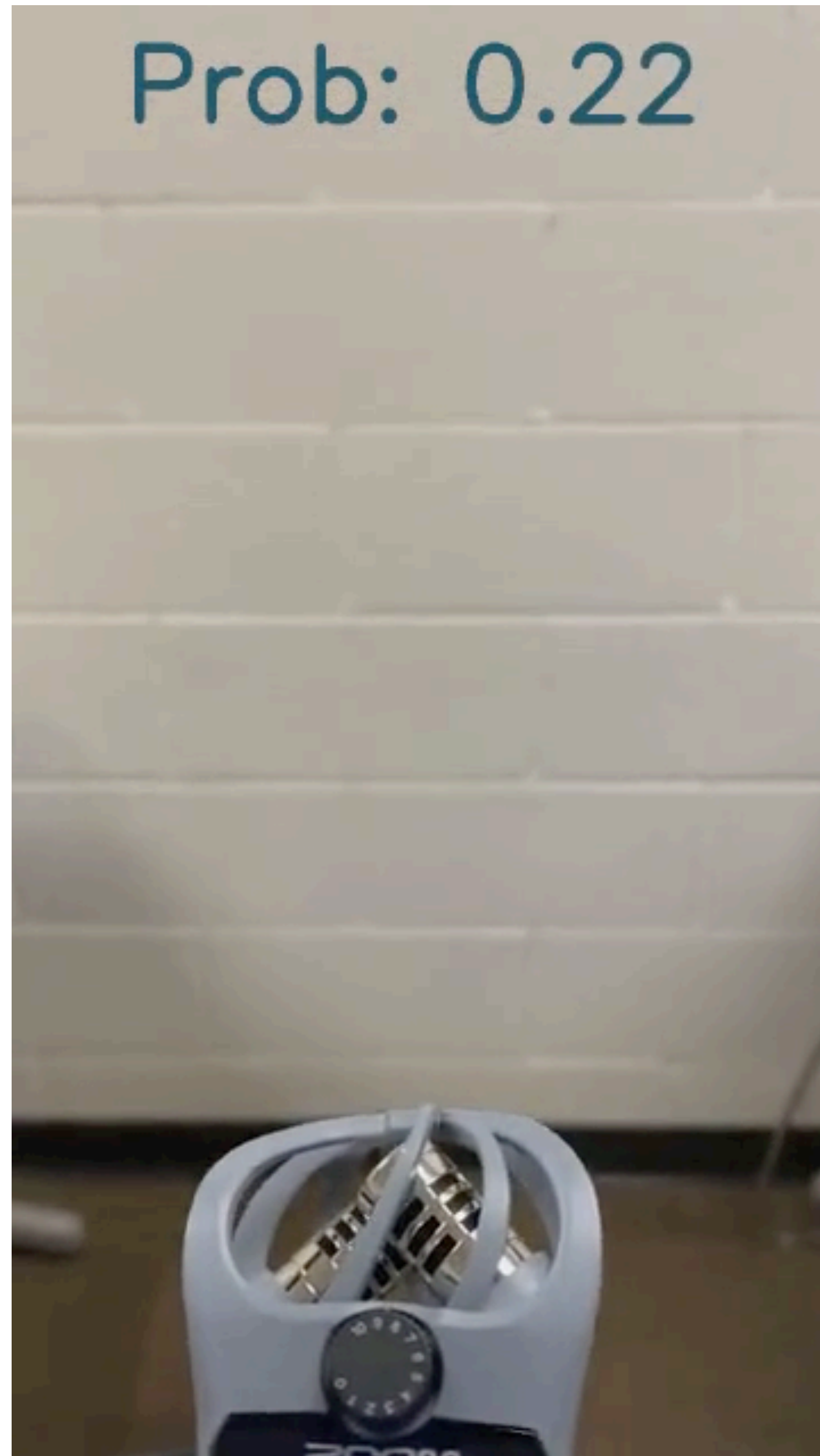
“Near a wall” prediction



"Near a wall" prediction



“Near a wall” prediction



0.91	0.78	0.88	0.86	0.76
0.97	0.4	0.43	0.57	0.86
0.89	0.35	0.26	0.35	0.71
0.79	0.25	0.31	0.18	0.86
0.92	0.92	0.96	0.92	0.87

Navigating with sound

