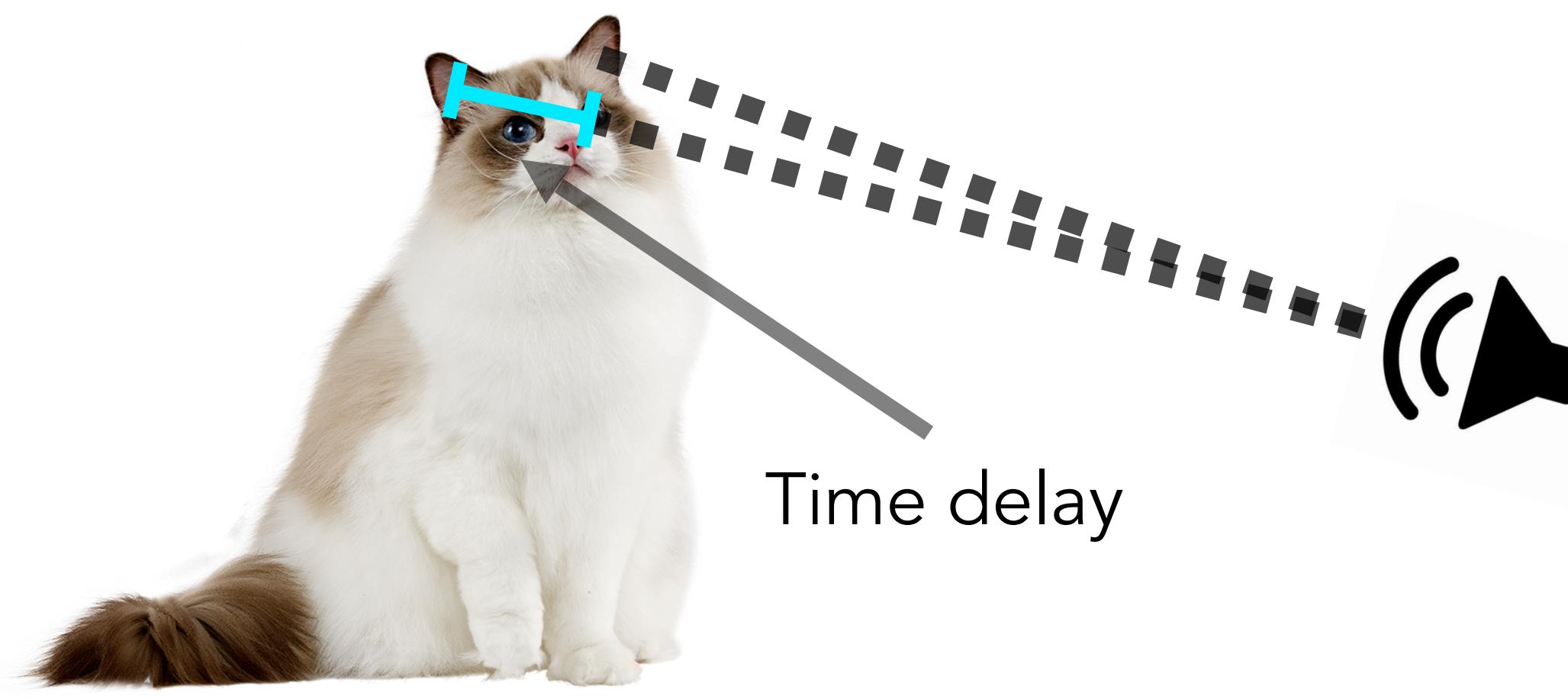


# Sound Localization by Self-Supervised Time Delay Estimation

Ziyang Chen, David F. Fouhey, Andrew Owens  
University of Michigan

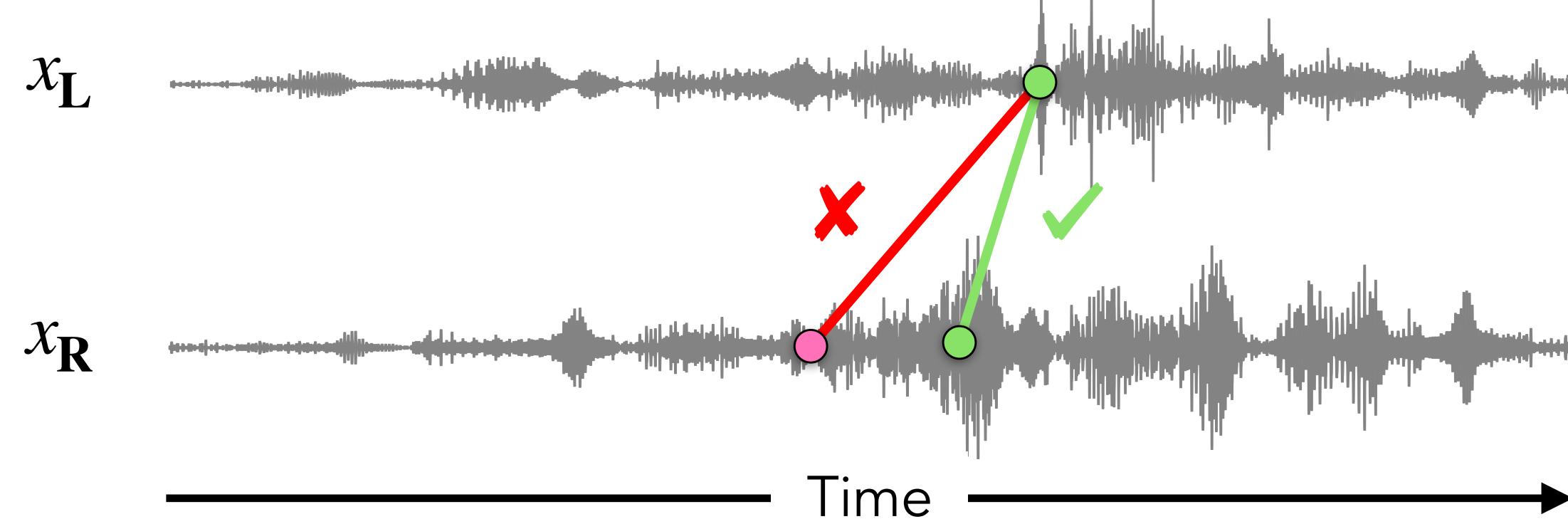
## Introduction

**Goal:** Localize sounds by estimating how much sooner they arrive at one ear than the other. We'll learn to do this using only unlabeled recordings.



## Time Delay Estimation

**Idea:** Treat it like a self-supervised tracking problem! Find **interaural correspondences**, pairs of sounds from different channels that correspond to the same underlying events.

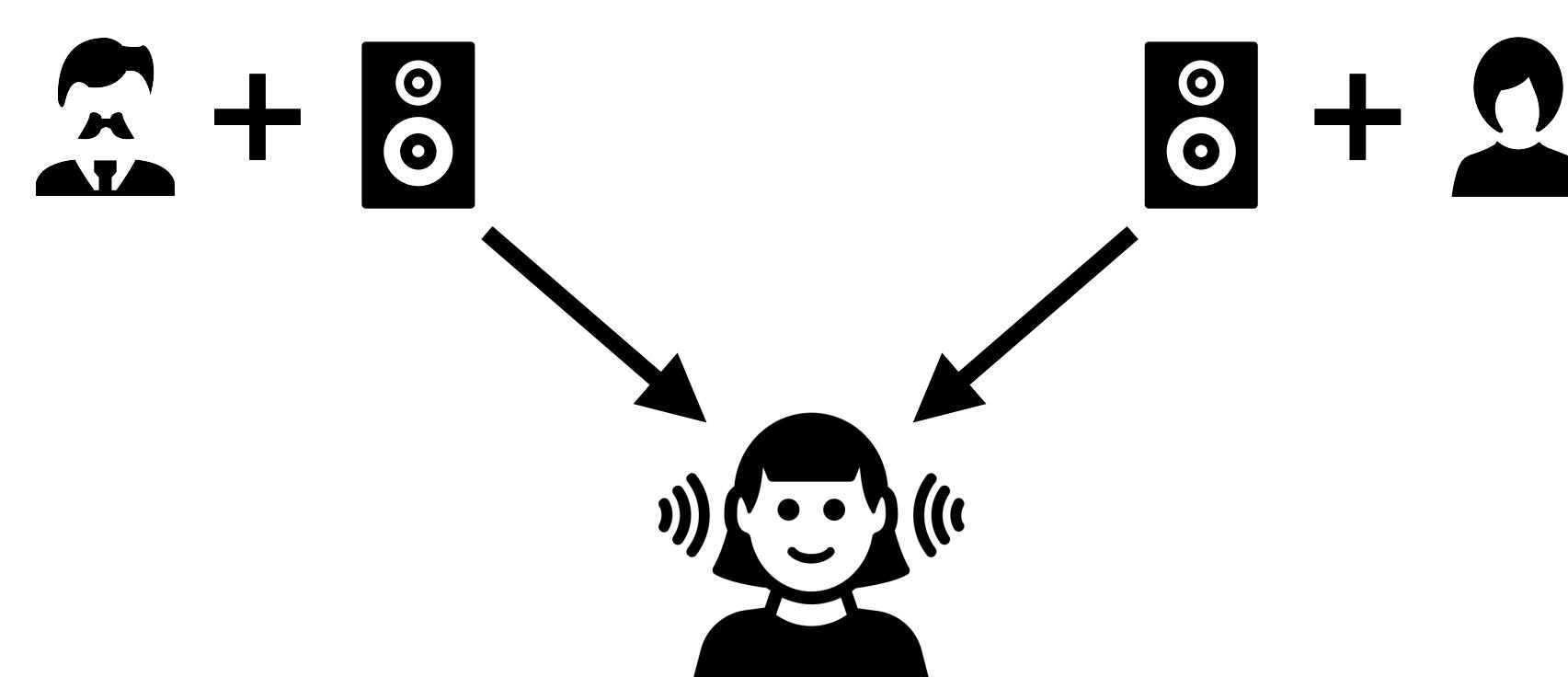


**Generalized cross-correlation:** find a time delay  $\tau$  that maximizes [Knapp & Carter, 1976]:

$$R_{\mathbf{x}_1, \mathbf{x}_2}(\tau) = \mathbb{E}_t [\mathbf{h}_1(t) \cdot \mathbf{h}_2(t - \tau)]$$

We'll learn audio (or audio-visual) features  $\mathbf{h}$ .

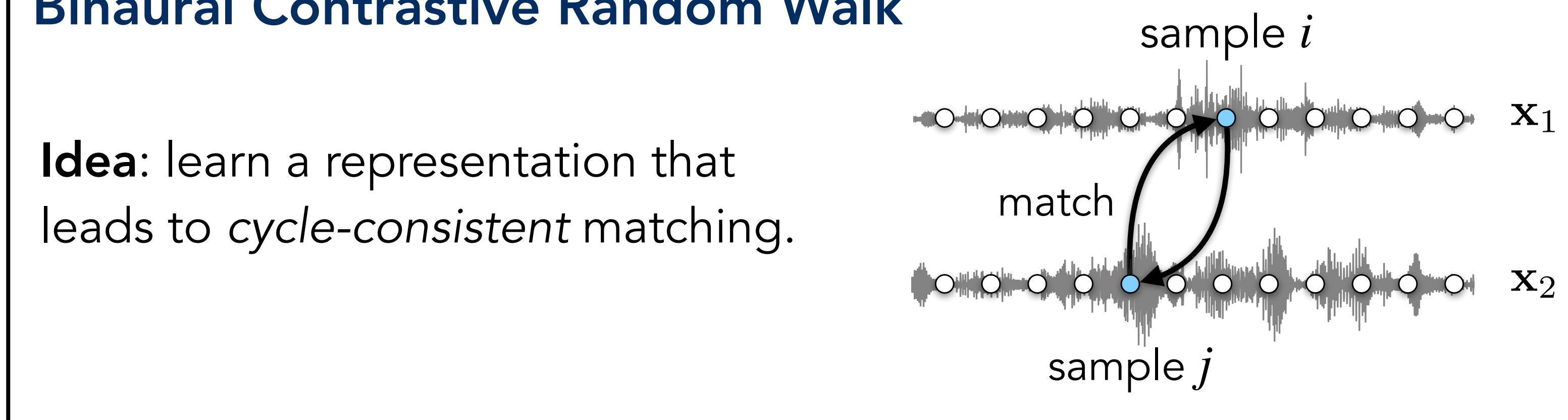
**Visually-guided sound localization:** localize a sound from a mixture, guided by the appearance of a speaker.



## Method

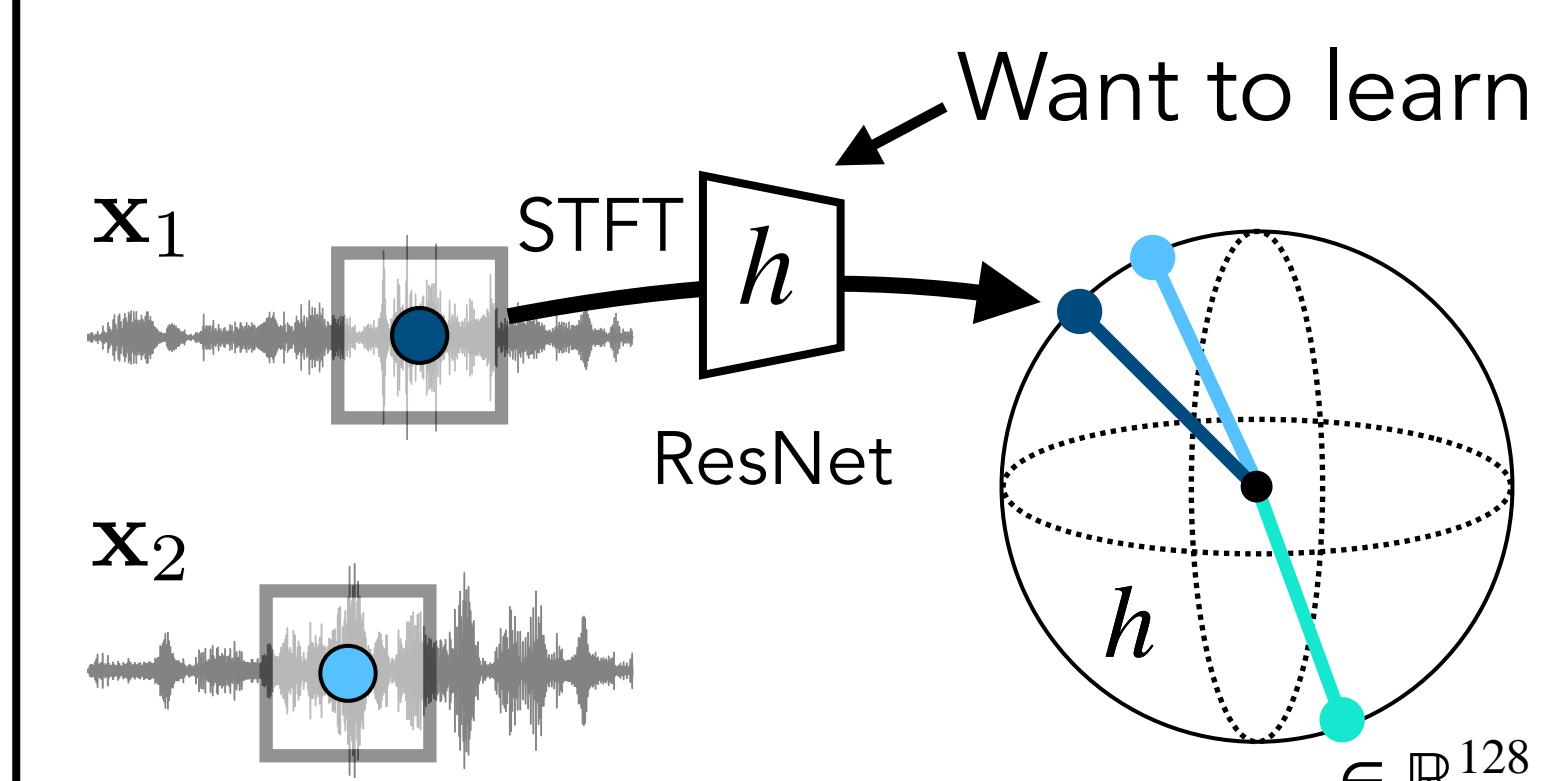
### Binaural Contrastive Random Walk

**Idea:** learn a representation that leads to cycle-consistent matching.



Transition probability from sample  $s$  in  $x_i$  to sample  $t$  in  $x_j$ :

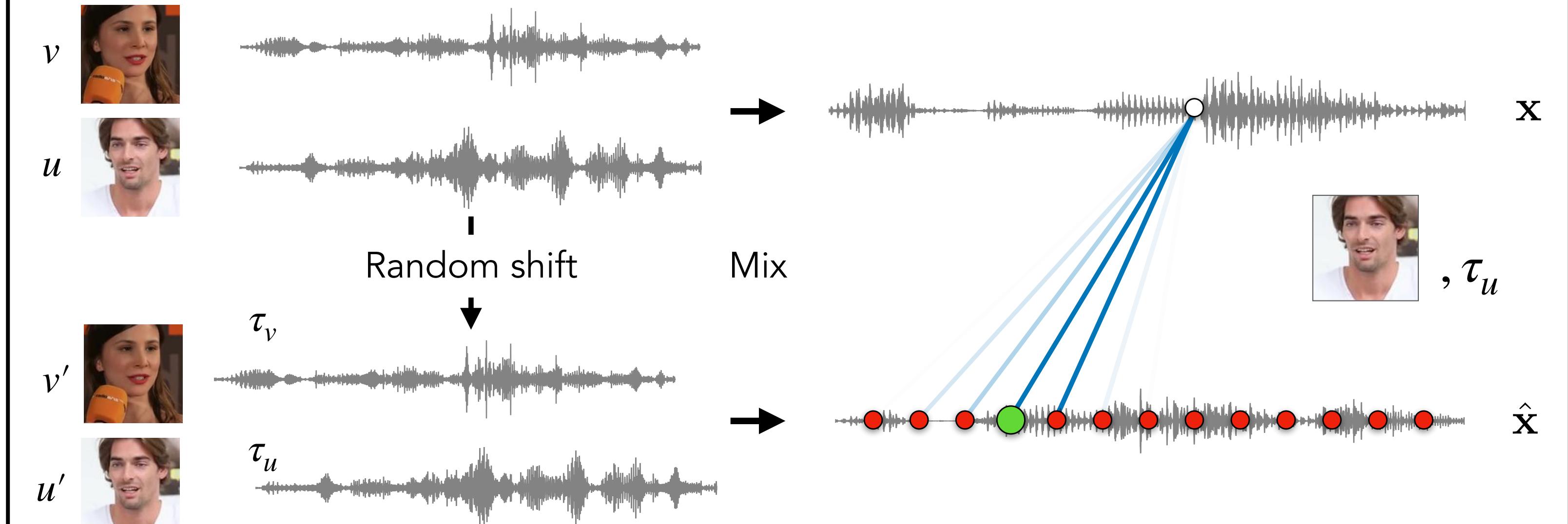
$$A_{ij}(s, t) = \frac{\exp(\mathbf{h}_i(s) \cdot \mathbf{h}_j(t)/c)}{\sum_{k=1}^n \exp(\mathbf{h}_i(s) \cdot \mathbf{h}_j(k)/c)}$$



**Objective:** maximize the return probability of a walk moves between channels [Jabri et al. 2020]:

$$\mathcal{L}_{\text{crw}} = -\text{tr}(\log(A_{12}A_{21}))$$

### Visually-Guided Time Delay Estimation

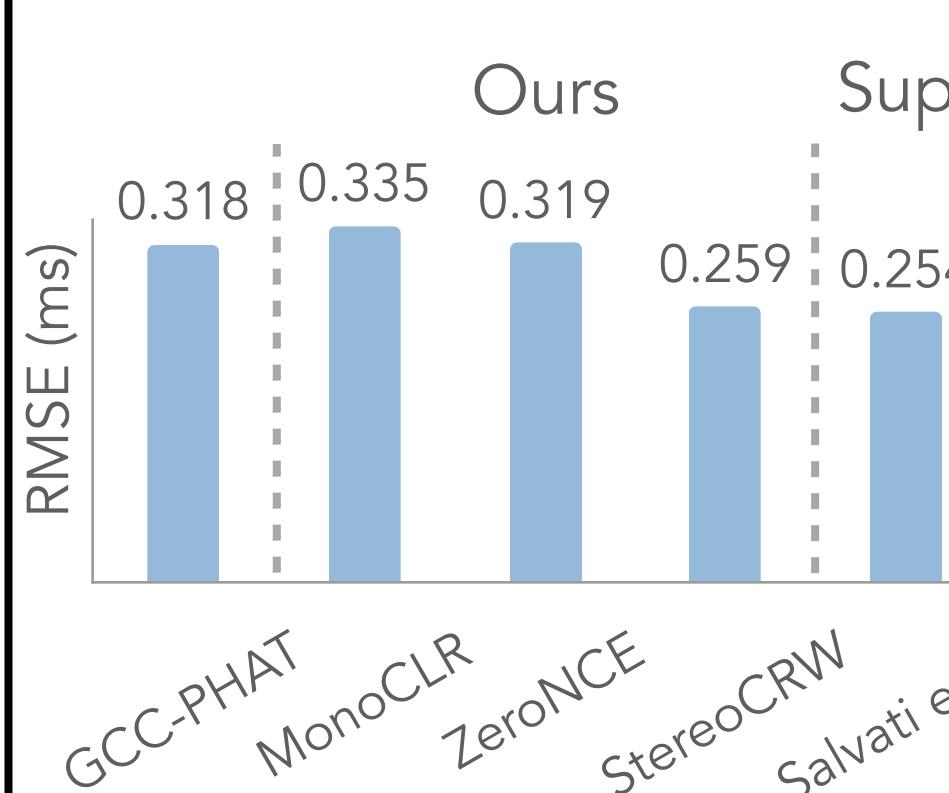


Learning audio-visual features  $\mathbf{g}$  by optimizing:

$$\mathcal{L}_{\text{av}} = -\log \frac{\exp(\mathbf{g}_1(t) \cdot \mathbf{g}_2(t - \tau_w)/c)}{\sum_{k=1}^n \exp(\mathbf{g}_1(t) \cdot \mathbf{g}_2(k)/c)}$$

## Experiments

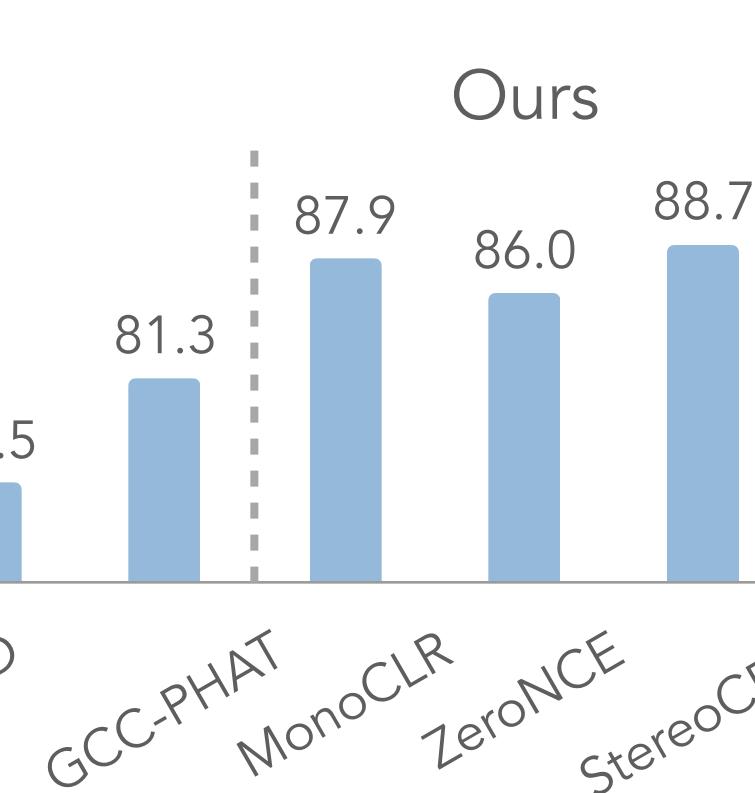
Simulated data



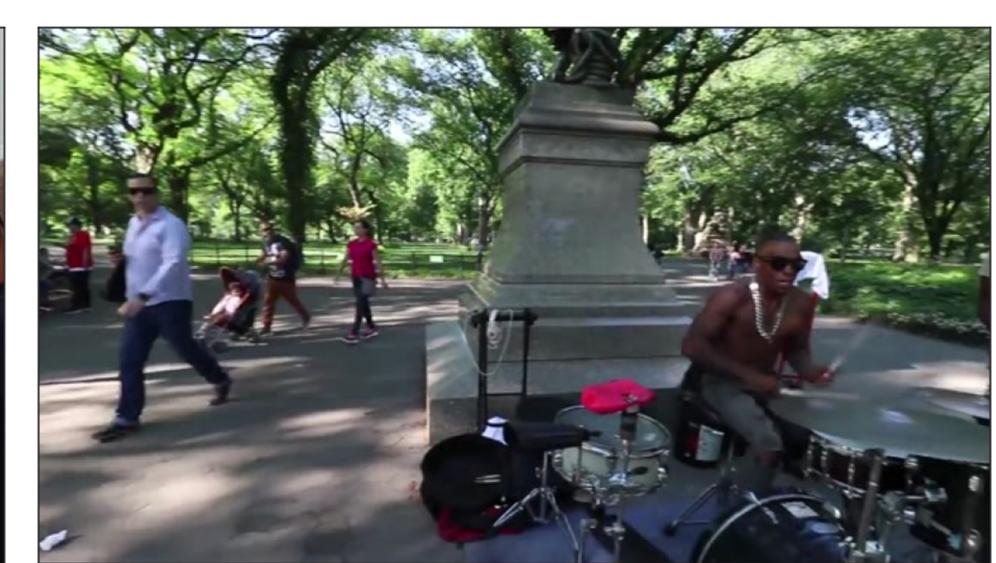
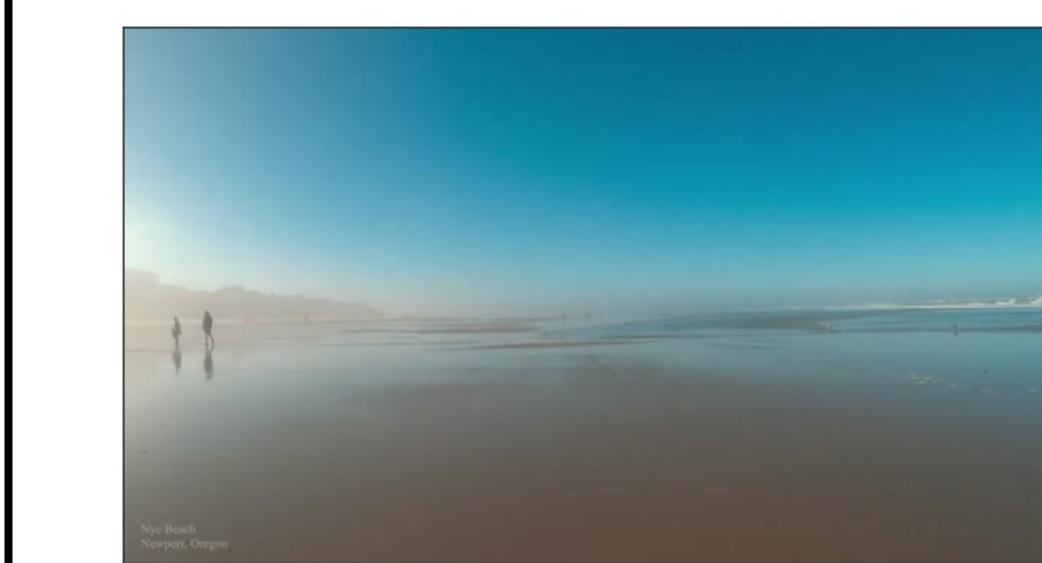
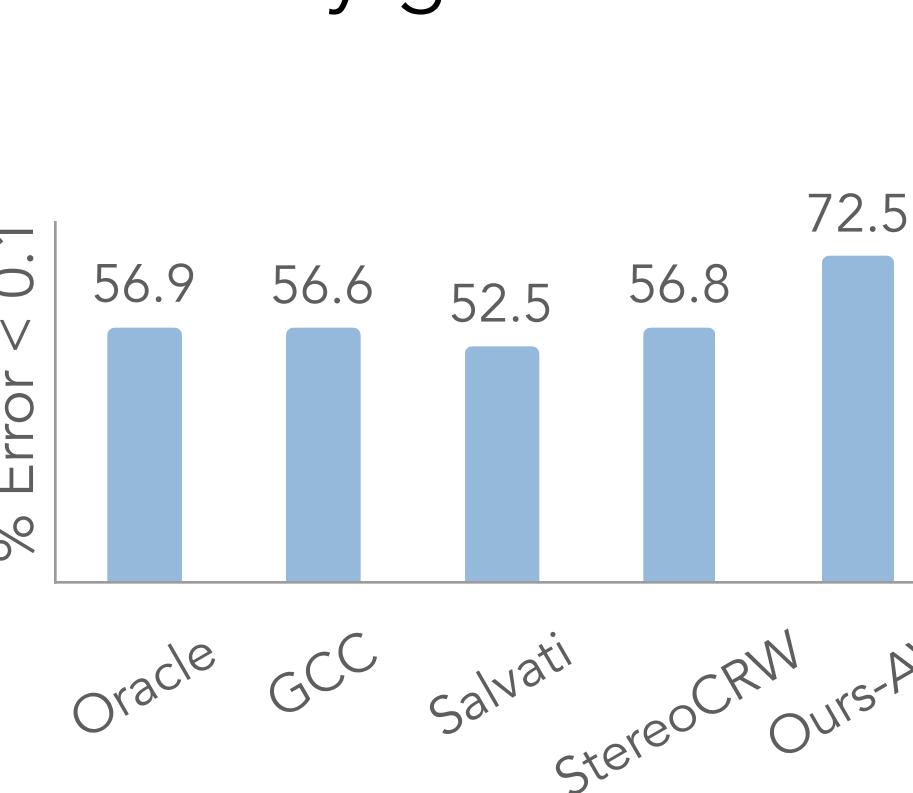
Acc

75.5

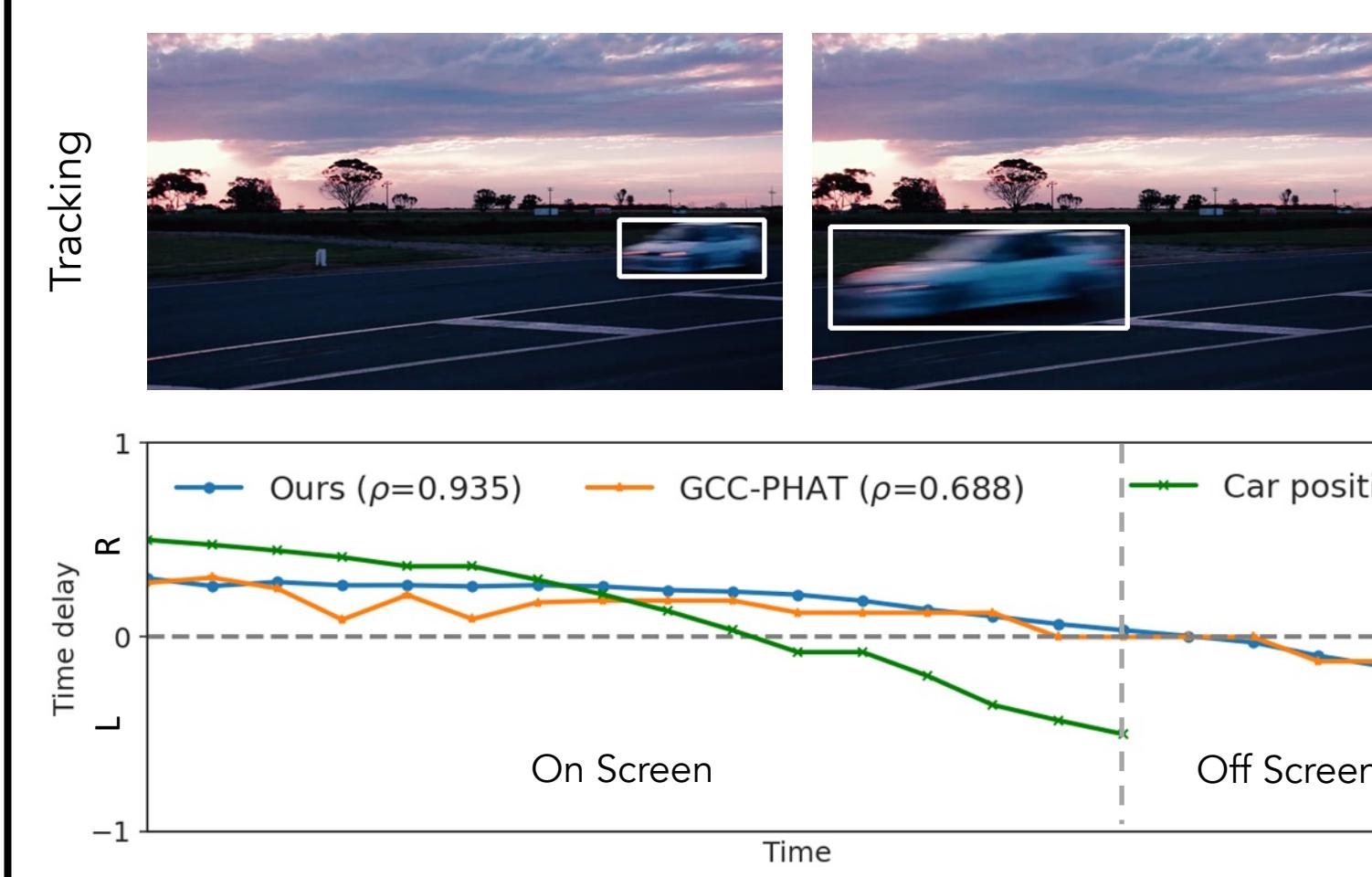
In-the-wild audio



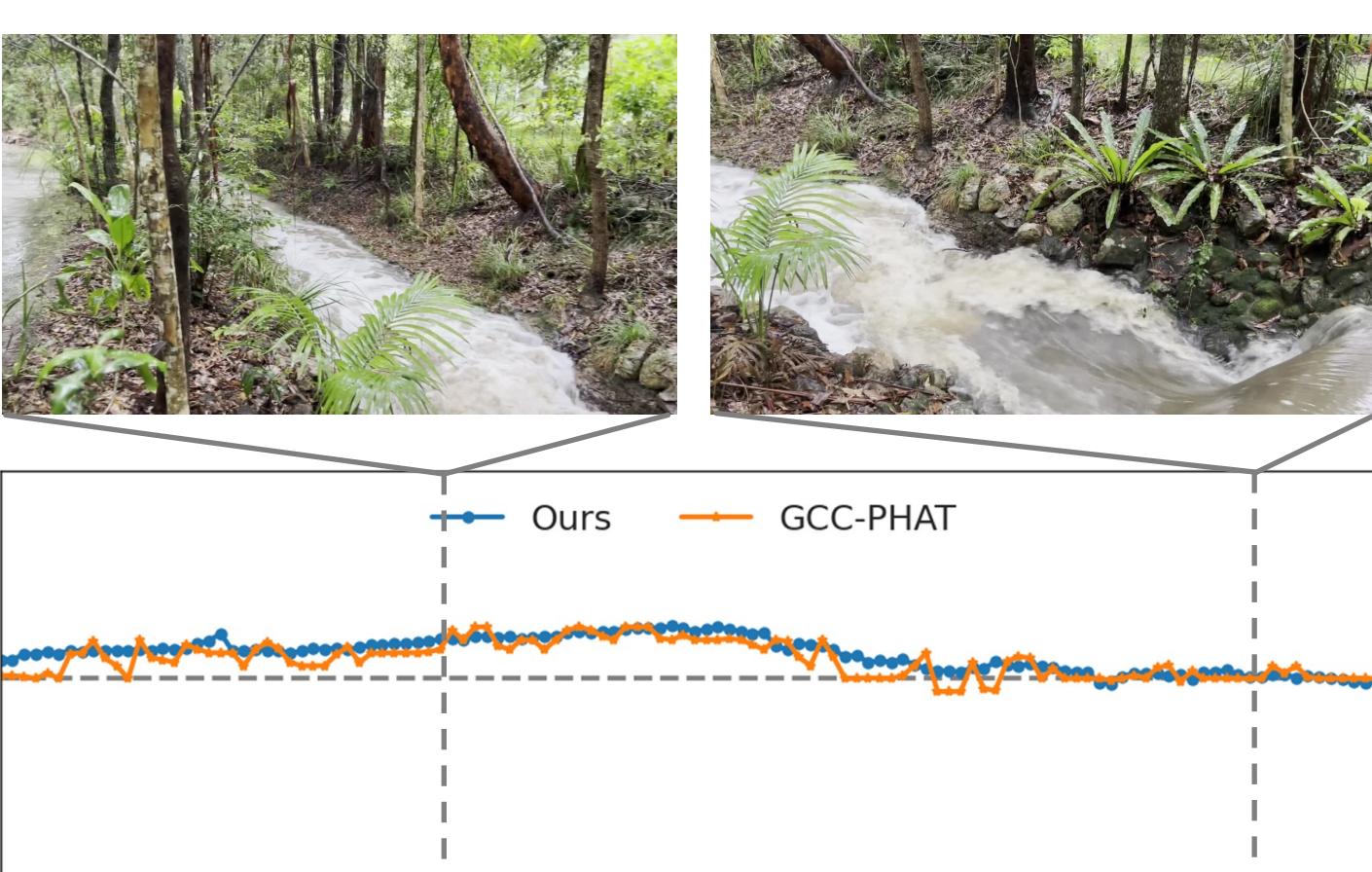
Visually-guided ITD



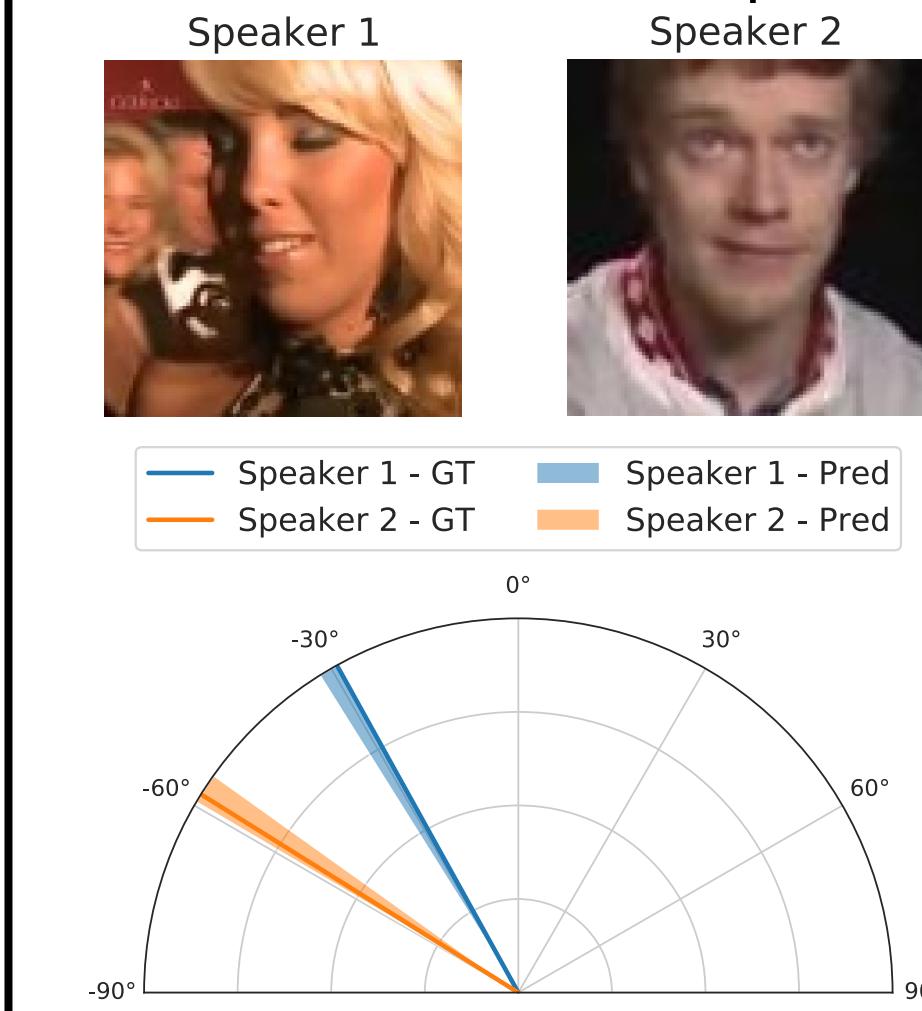
In-the-wild video



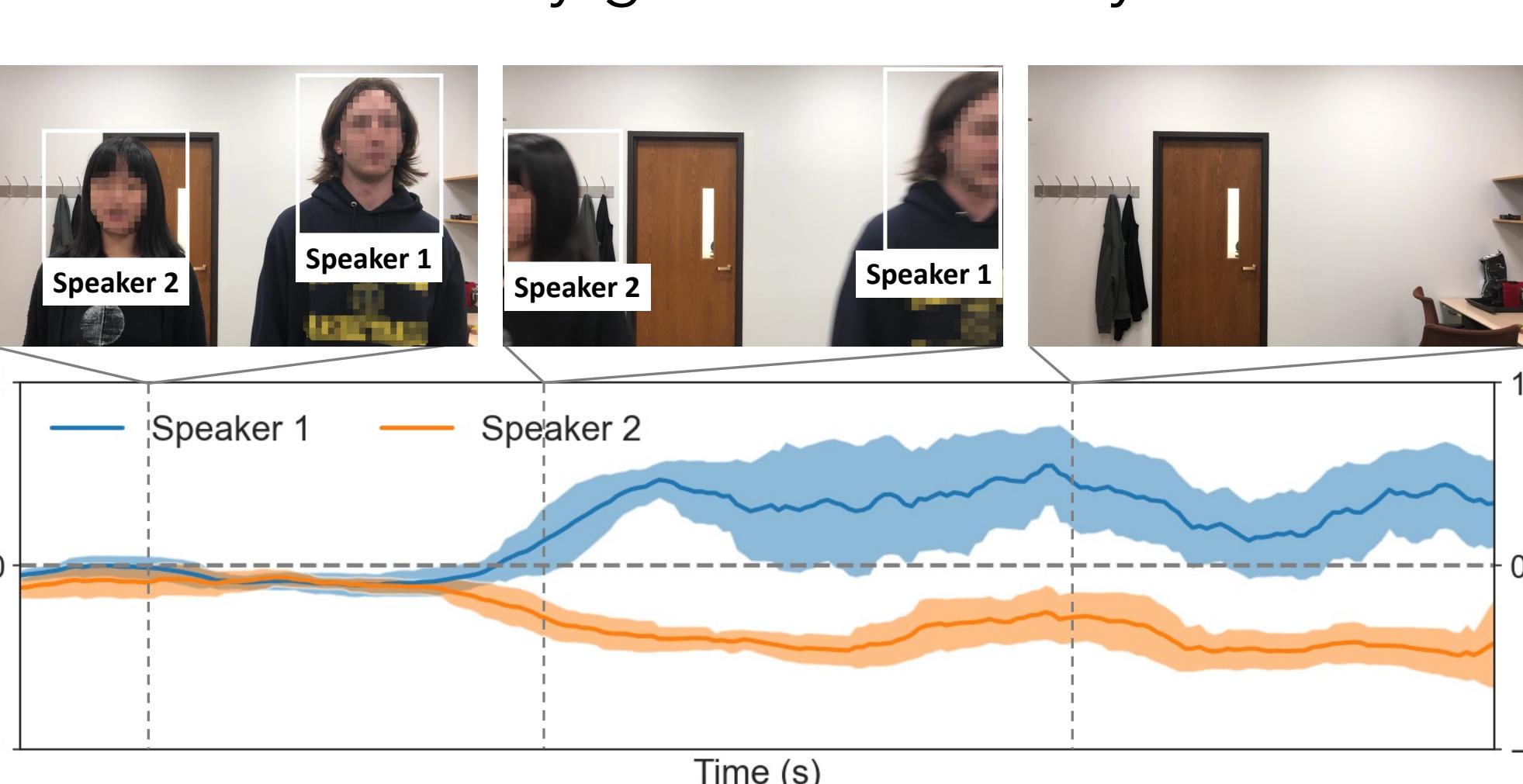
iPhone video



Simulated examples



Real-world visually-guided time delay estimation



## Conclusions

- We propose a simple, **self-supervised** method for learning to estimate interaural time delay via cycle-consistent random walks.
- Visual signals allow our models to localize specific speakers within mixtures.