Frontiers in Perceptual AI: First-Person Video and Multimodal Perception

Kristen Grauman

University of Texas at Austin

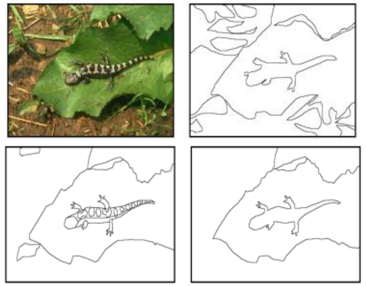
FAIR, Meta AI

THE UNIVERSITY OF TEXAS AT AUSTIN

Meta AI

The third-person Web perceptual experience

A curated “disembodied” moment in time
from a spectator’s perspective

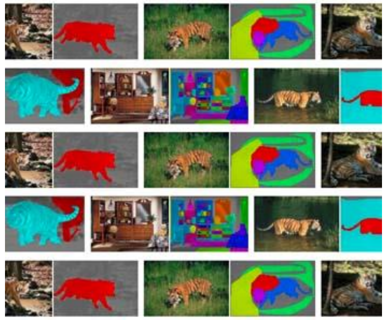


BSD (2001)

Caltech 101 (2004), Caltech 256 (2006)

PASCAL (2007-12)

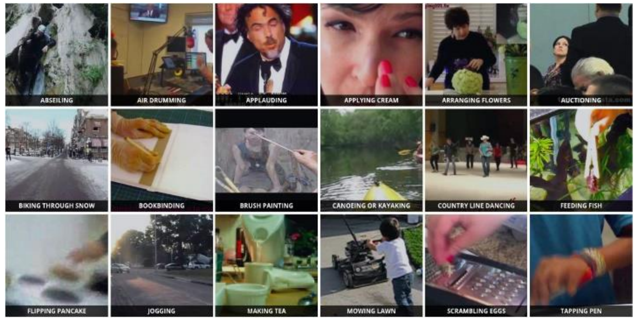
ActivityNet (2015)



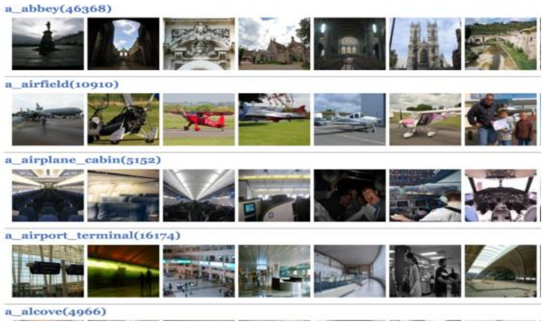
LabelMe (2007)

ImageNet (2009)

SUN (2010)

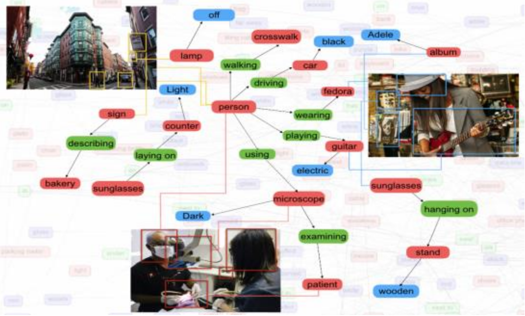


Kinetics (2017)



Places (2014)

MS COCO (2014)



Visual Genome (2016)

AVA (2018)

First-person “egocentric” perceptual experience

Uncurated long-form video stream driven by the agent’s goals, interactions, and attention

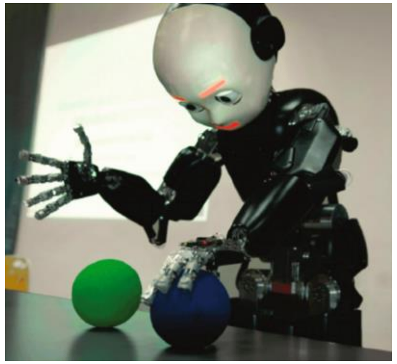
First-person perception and learning

Status quo:

Learning and inference with "disembodied" images/videos.

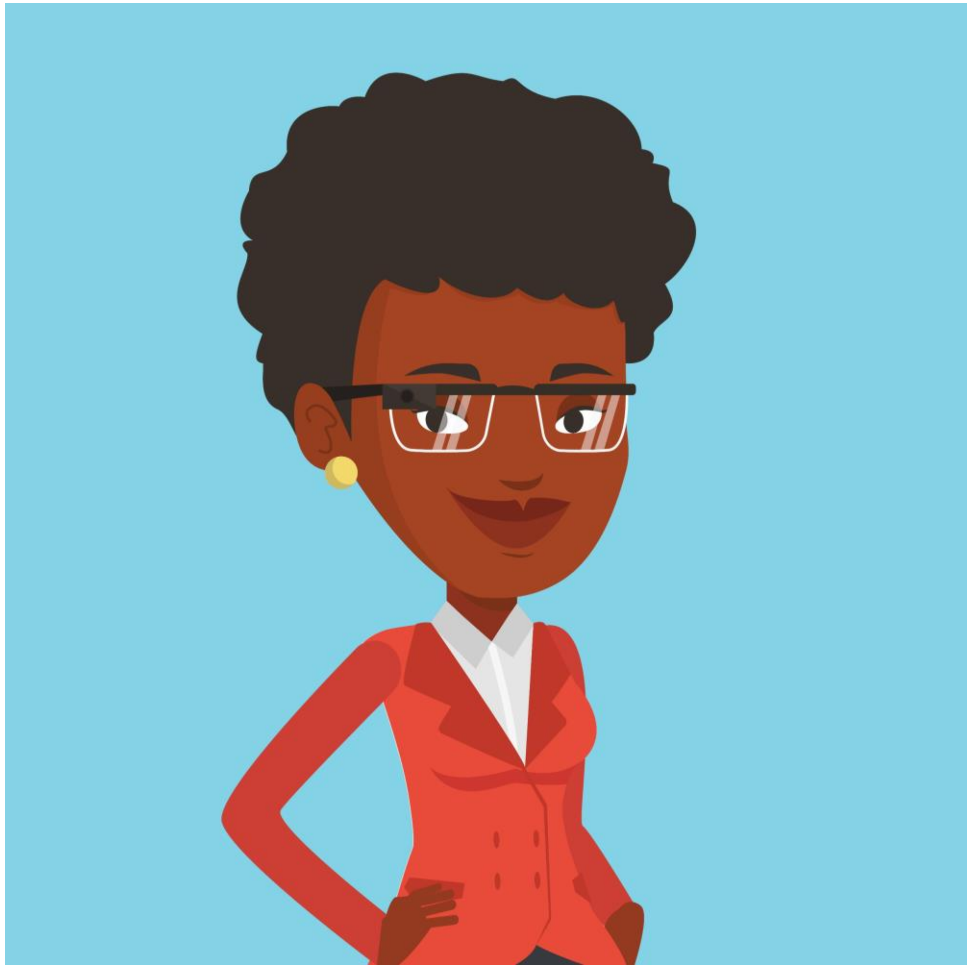
On the horizon:

Visual learning in the context of agent goals, interaction, and multi-sensory observations.


Why egocentric video?

Robot learning

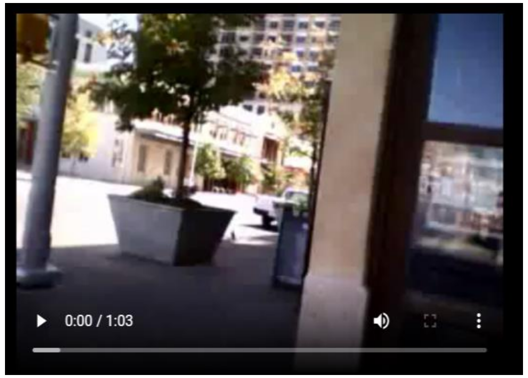


Augmented reality

Existing first-person video datasets

Inspire our effort, but call for greater scale, content, diversity

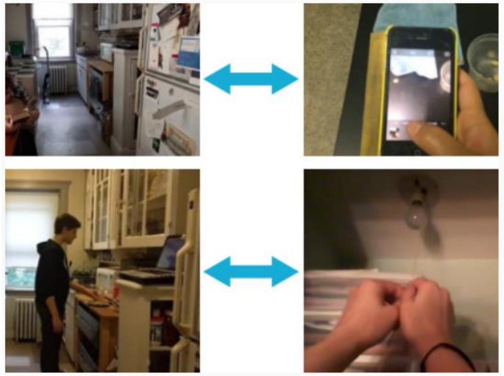
EPIC Kitchens
Damen et al. 2020
45 people, 100 hrs
kitchens only



UT Ego
Lee et al. 2012
4 people, 17 hrs
daily life, in/outdoors

EGTEA Gaze+
Li et al. 2018
32 people, 28 hrs
kitchens only

ADL
Pirsiavash 2012
20 people, 10 hrs
apartment



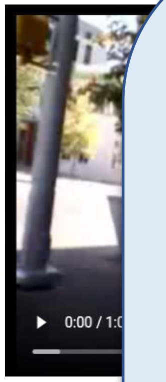
Charades-Ego
Sigurdsson 2018
71 people, 34 hrs
indoor

EGO 4D

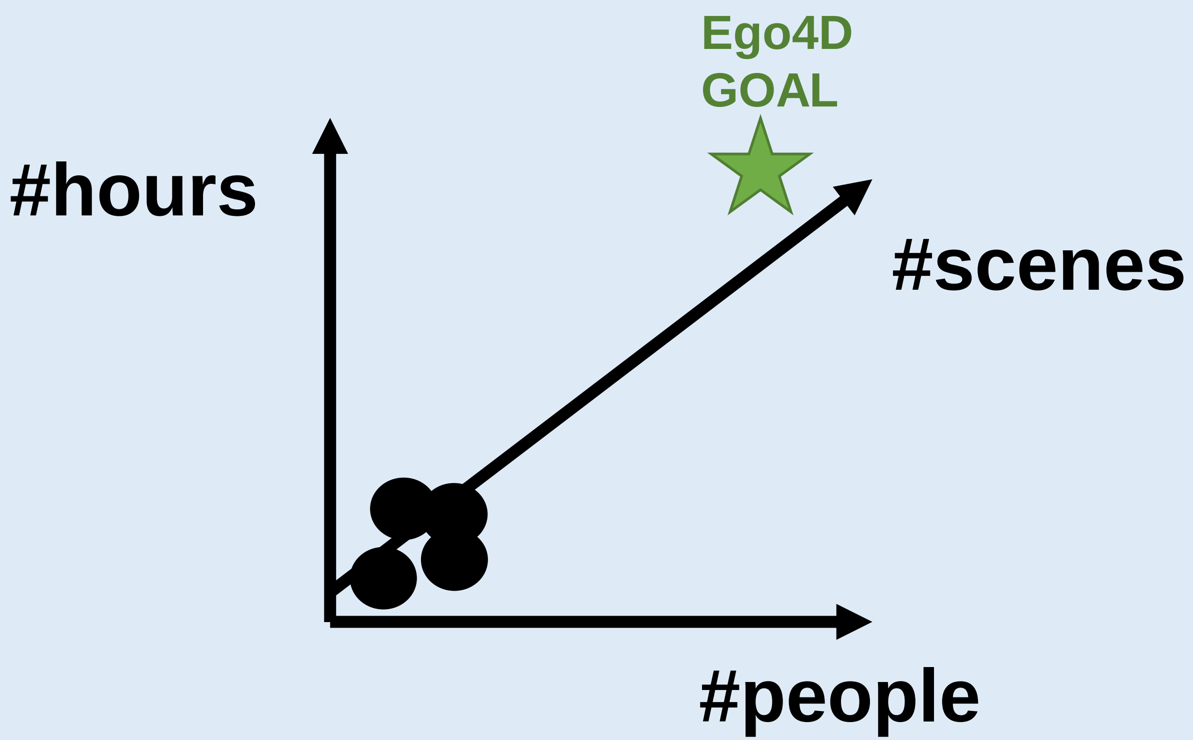
Existing first-person video datasets

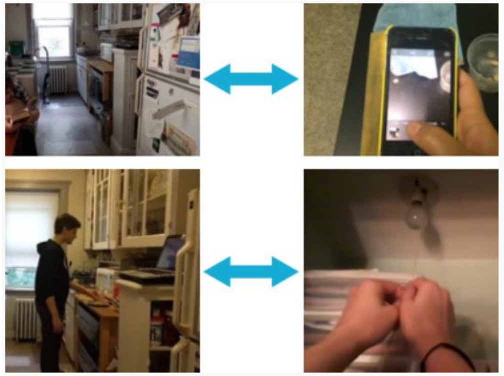
Inspire our effort, but call for greater scale, content, diversity

EPIC Kitchens
Damen et al. 2020
45 people, 100 hrs
kitchens only



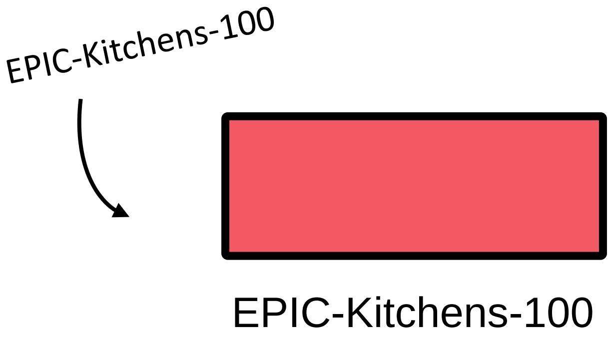
UT E
Lee
4 pe
daily





Charades-Ego
Sigurdsson 2018
71 people, 34 hrs
indoor

Existing first-person video datasets



Ego4D: A massive-scale egocentric dataset

3,670 hours of in-the-wild daily life activity

931 participants from 74 worldwide locations

Multimodal: audio, 3D scans, IMU, stereo, multi-camera

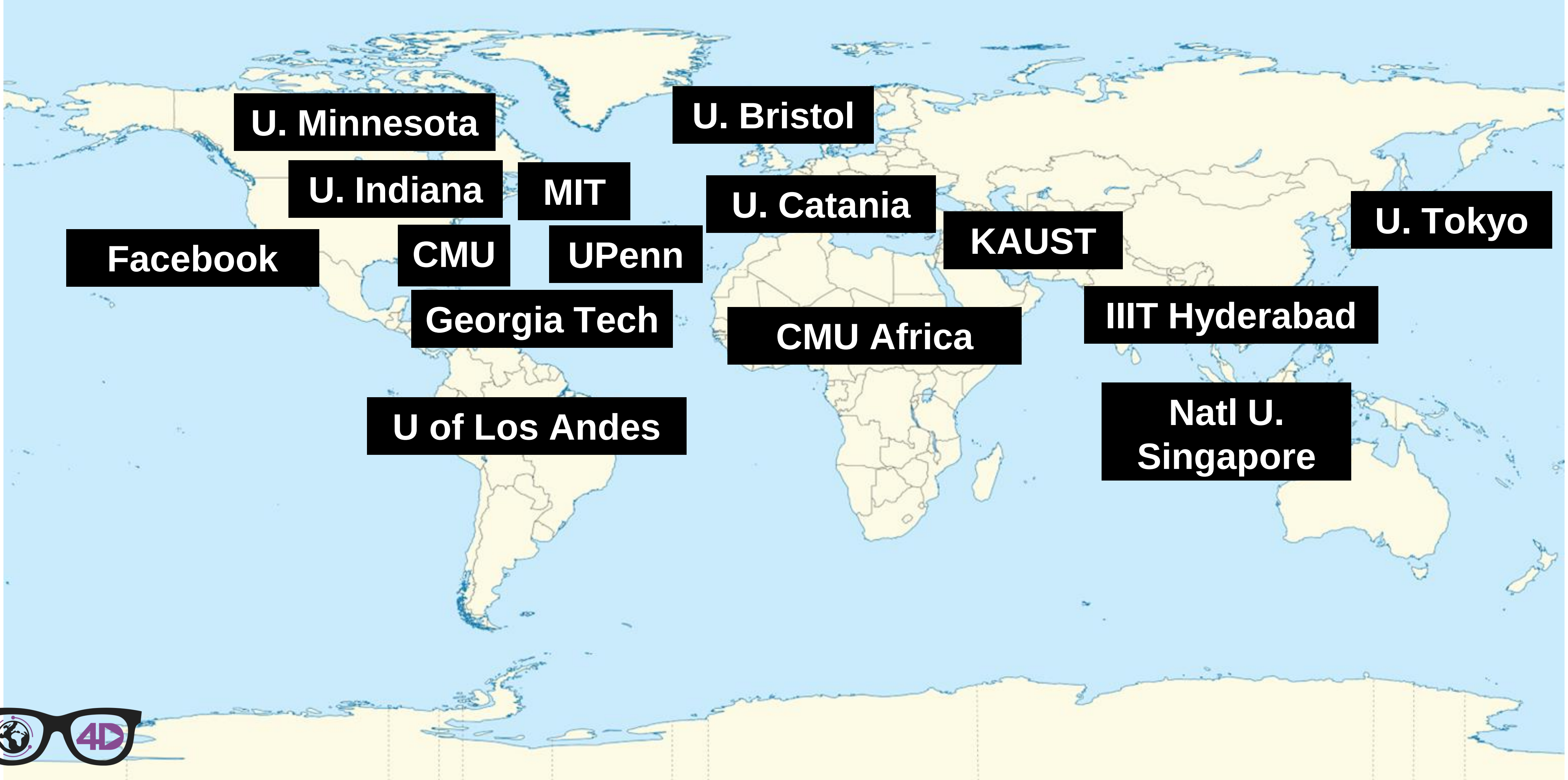
Benchmark tasks to catalyze research

[Grauman et al. CVPR 2022]

Ego4D: everyday activity around the world

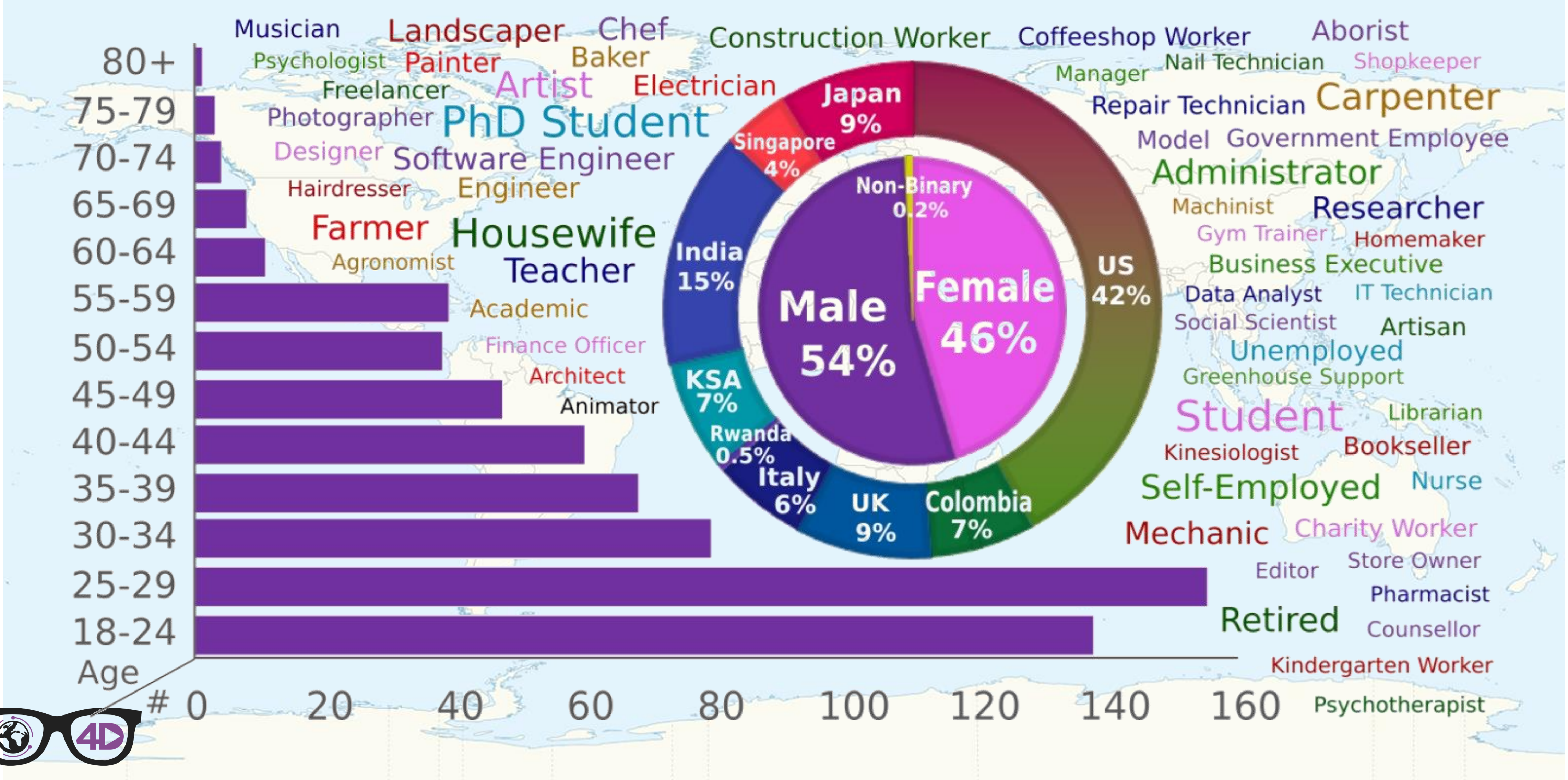
Ego4D: FAIR + university consortium

Towards diverse geographic coverage



931 unique camera wearers

Towards diverse demographic coverage



EGO 4D

EGO4D

Wearable cameras

GoPro

Vuzix Blade

Pupil Labs

ZShade

WeeView

We deploy a variety of head-mounted cameras.

Unscripted, daily-life scenarios

How people spend their days: US Bureau of Labor Statistics

Everyday activities in the home:

- Sleeping
- Daily hygiene
- Doing hair/make-up
- Cleaning / laundry
- Cooking
- Talking with family members
- Hosting a party
- Eating
- Yardwork / shoveling snow
- Household management - care for kids
- Fixing something in the home
- Playing with pets
- Crafting/knitting/sewing/drawing/painting/etc

Errands

- Grocery shopping
- Clothes, shopping
- Getting car fixed
- Going to the bank
- Walking the dog
- Washing the dog / pet, grooming horse
- Appointments: doctor, dentist, hair

Work

- Working at desk
- Participating in a meeting
- Attending a lecture/class
- Writing on whiteboard
- Video call
- Eating at the cafeteria
- Making coffee
- Talking to colleagues

Entertainment/Leisure

- Watching movies at cinema
- Watching tv
- Reading books
- Playing games / video games
- Attending sporting events - watching and participating in
- Attending play/ballet
- Attending concerts
- Hanging out with friends at a bar
- Eating at a restaurant
- Eating at a friend’s home
- Attending a party
- Talking on the phone
- Listening to music
- BBQ’ing/picnics
- Going to a salon (nail, hair, spa)
- Getting a tattoo / piercing
- Volunteering
- Practicing a musical instrument
- Attending a festival or fair
- Hanging out at a coffee shop

Exercise:

- Going to the gym
- Yoga practice
- Swimming in a pool/ocean
- Working out at home
- Cycling / jogging
- Dancing
- Working out outside
- Walking on street
- Going to the park
- Hiking
- Tourism

Transportation:

- Car - commuting, road trip
- Bus
- Train
- Airplane
- Bike
- Skateboard/scooter

https://www.bls.gov/news.release/atus.nr0.htm

Ego4D: everyday activity around the world

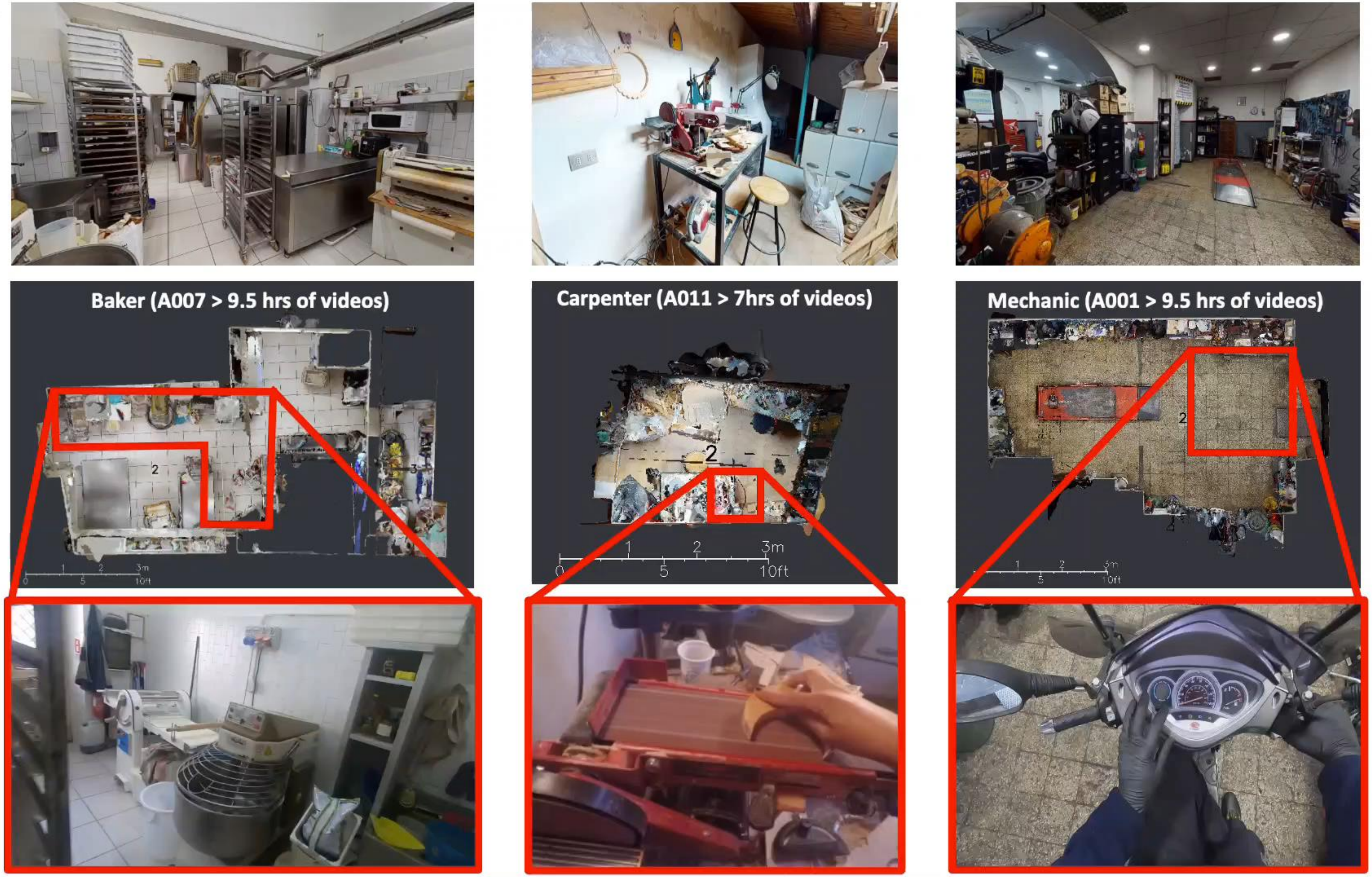
Ego4D data: 3D environment scans

EGO4D@UNICT Examples

3D

FloorPlan

EGO



Available for 491 hours of video

Ego4D data: multi-camera and eye gaze

Multiple simultaneous egocentric cameras

Eye gaze
(Indiana University)

Ego4D annotations: text narrations

Dense descriptive text of each camera wearer activity + clip-level summaries

13 sentences per minute

4M+ sentences

Privacy and ethics

- **Review**: Each partner underwent separate months-long IRB review process, overseeing ethical and privacy standards for data collection, management, and informed consent.
- **Consent**: Forms signed by all recorded people where relevant
- **De-identification**: State-of-the-art de-identification processes, featuring both automated and manual reviews for faces, screens, credit cards, and other identifiers

Ego4D benchmark suite

Past

Episodic Memory

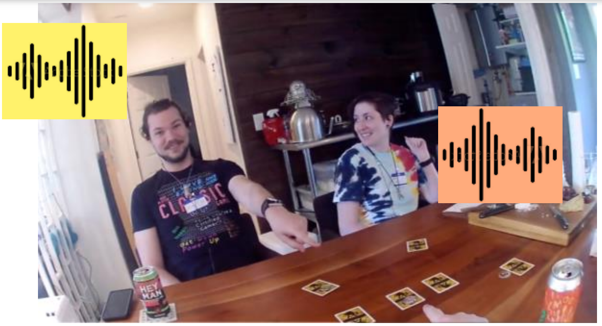
“where is my X?”

Present

Hands & Objects

“what am I doing and how?”

Present



Audio-visual Diarization

“who said what when?”

Social Interaction

“who is attending to whom?”

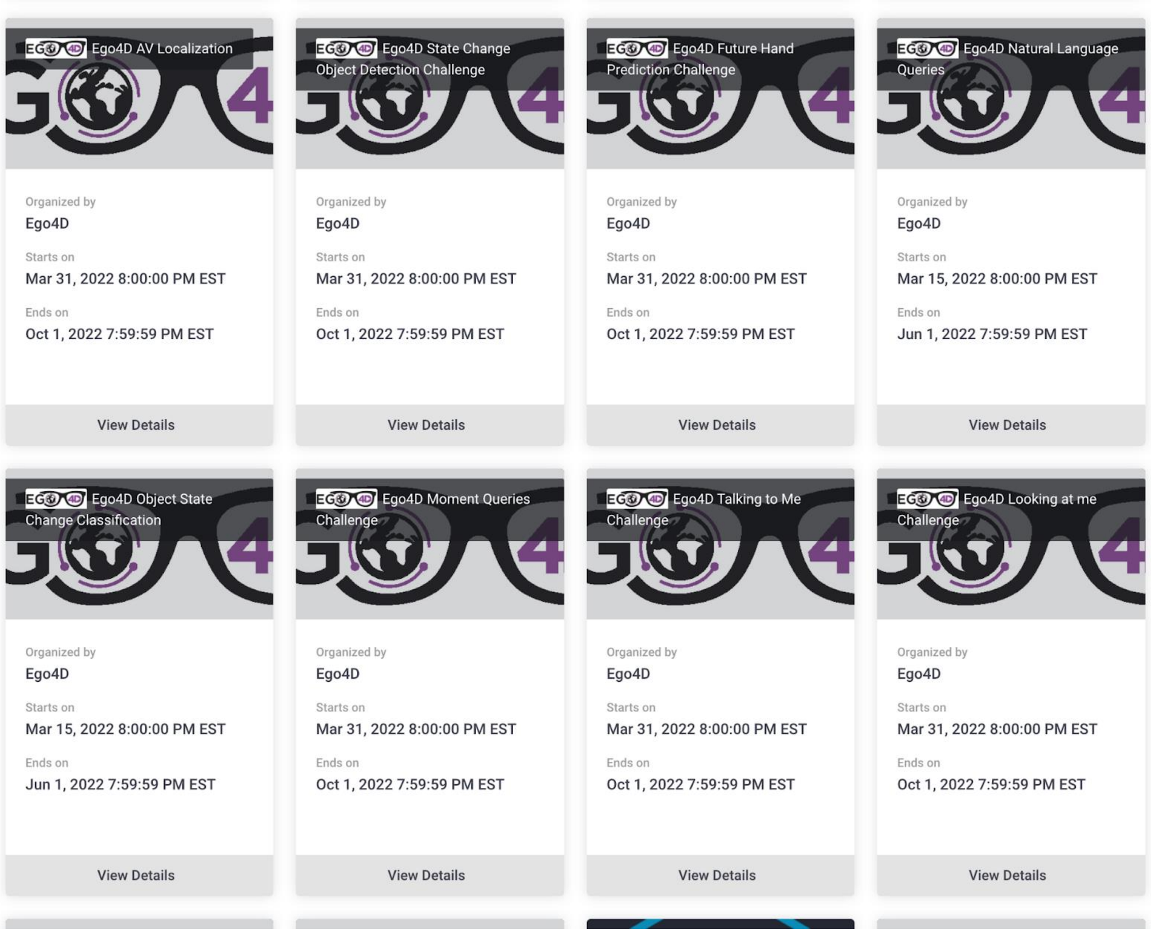
Future

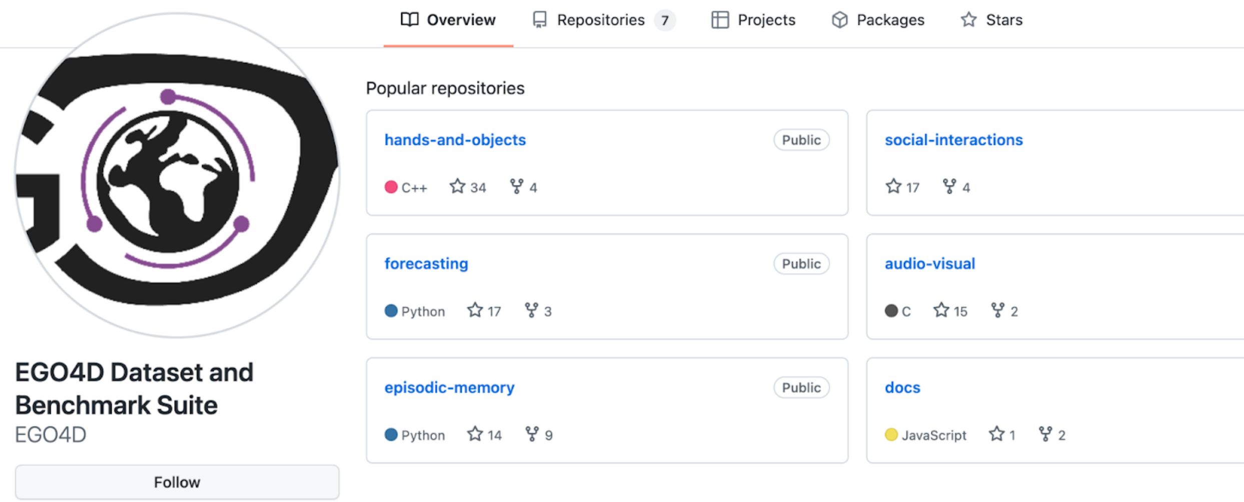
Forecasting

“what will I do next?”

Ego4D challenges

Broad and growing participation at CVPR 2022, ECCV 2022, CVPR 2023



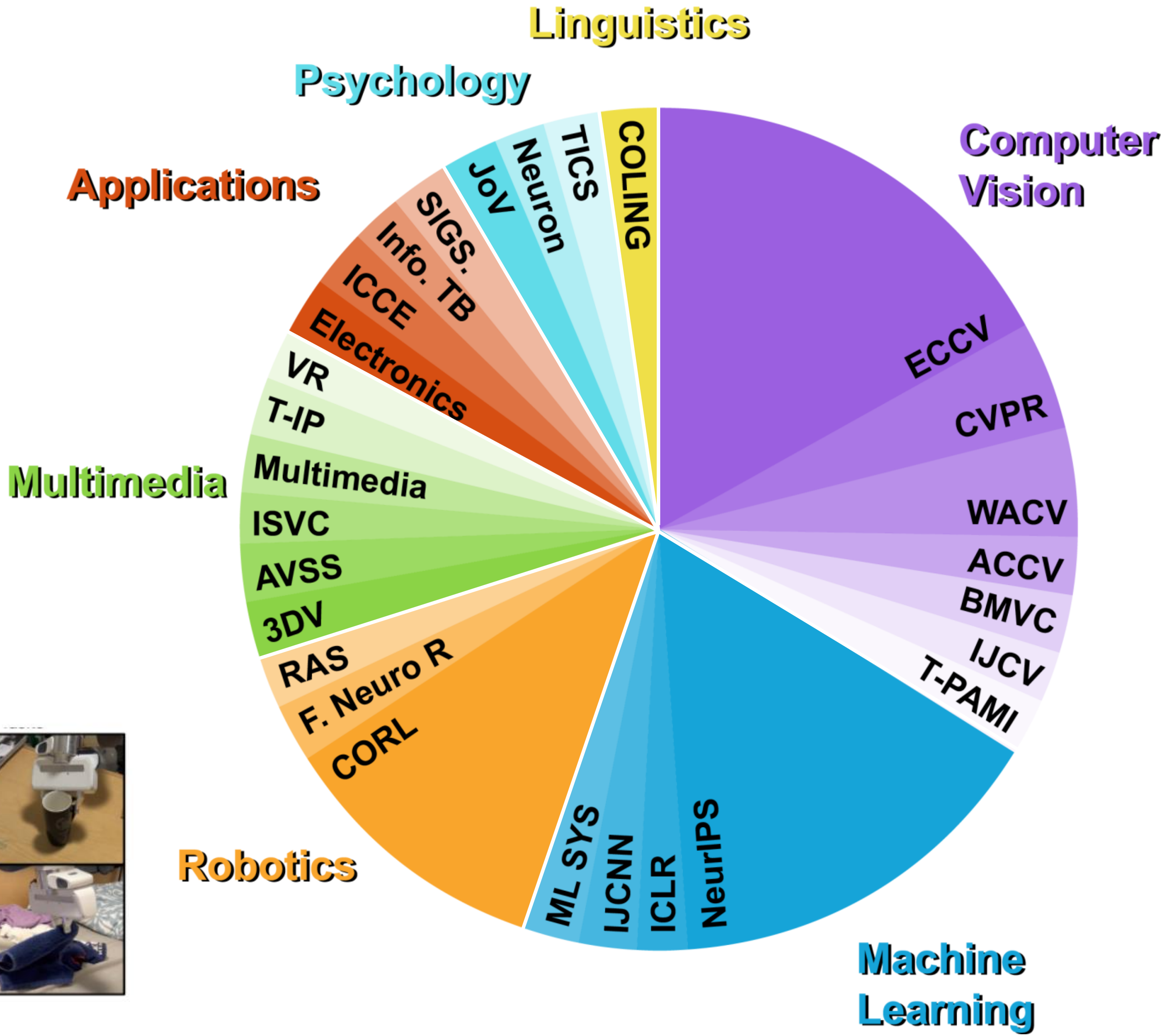


Hosted on EvalAI

https://ego4d-data.org/docs/challenge/

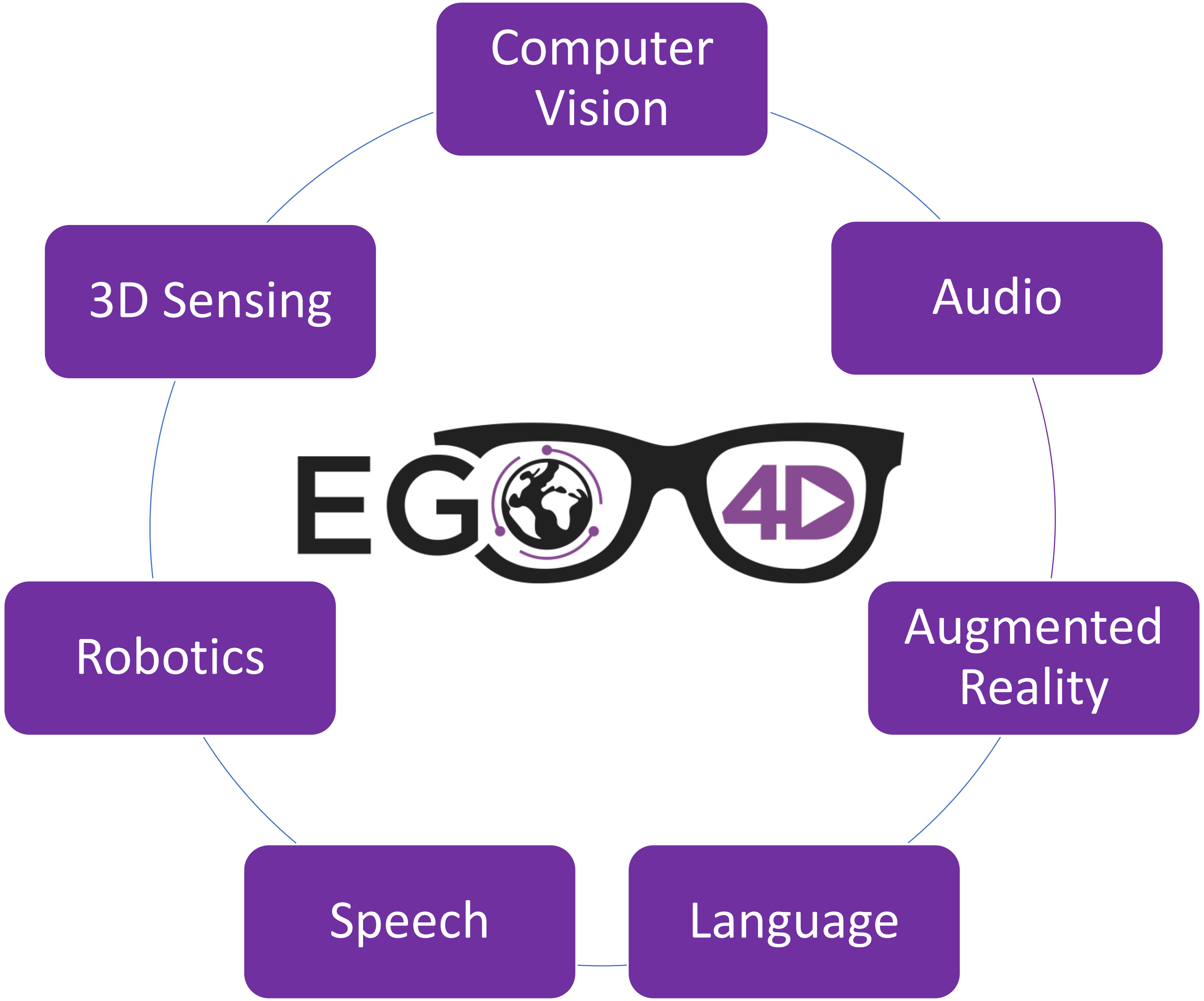
Ego4D is enabling new research in the broader community

Papers citing Ego4D



Source: Google Scholar
Slide credit: David Crandall

Ego4D: computer vision and beyond



Ego4D team

Kristen Grauman, Andrew Westbury, Eugene Byrne*, Zachary Chavis*, Antonino Furnari*, Rohit Girdhar*, Jackson Hamburger*, Hao Jiang*, Miao Liu*, Xingyu Liu*, Miguel Martin*, Tushar Nagarajan*, Ilija Radosavovic*, Santhosh Kumar Ramakrishnan*, Fiona Ryan*, Jayant Sharma*, Michael Wray*, Mengmeng Xu*, Eric Zhongcong Xu*, Chen Zhao*, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, Jitendra Malik

Carnegie Mellon University
Università di Catania
NUS National University of Singapore
King Abdullah University of Science and Technology
東京大学 The University of Tokyo
University of Bristol
Indiana University Bloomington
University of Minnesota
MIT
Penn University of Pennsylvania
Georgia Institute of Technology
International Institute of Information Technology Hyderabad
Carnegie Mellon University Africa
Universidad de los Andes Colombia
Facebook AI

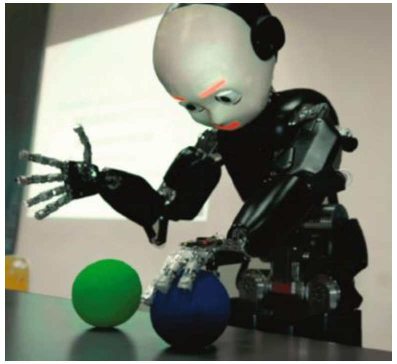
First-person perception and learning

Status quo:

Learning and inference with “disembodied” images/videos.

On the horizon:

Visual learning in the context of agent goals, interaction, and multi-sensory observations.

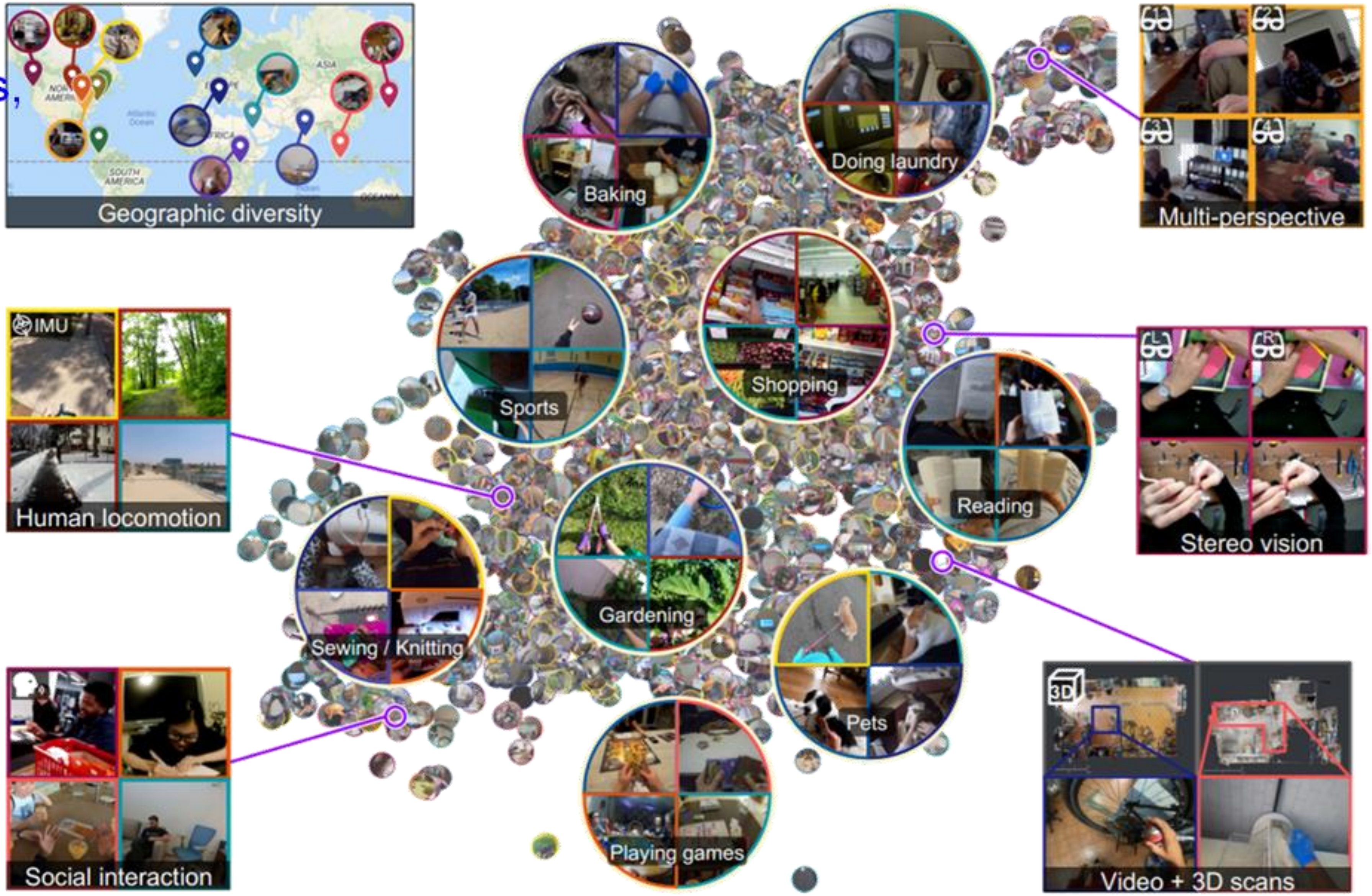


Ego4D: large-scale multimodal dataset

9 countries, 74 cities

IMU
836 hrs

Audio
2,207 hrs



Multiple synchronized ego-cameras
224 hrs

Stereo
80 hrs

3D environment scans
491 hrs

Grauman et al. CVPR 2022

Recall: Ego4D text narrations

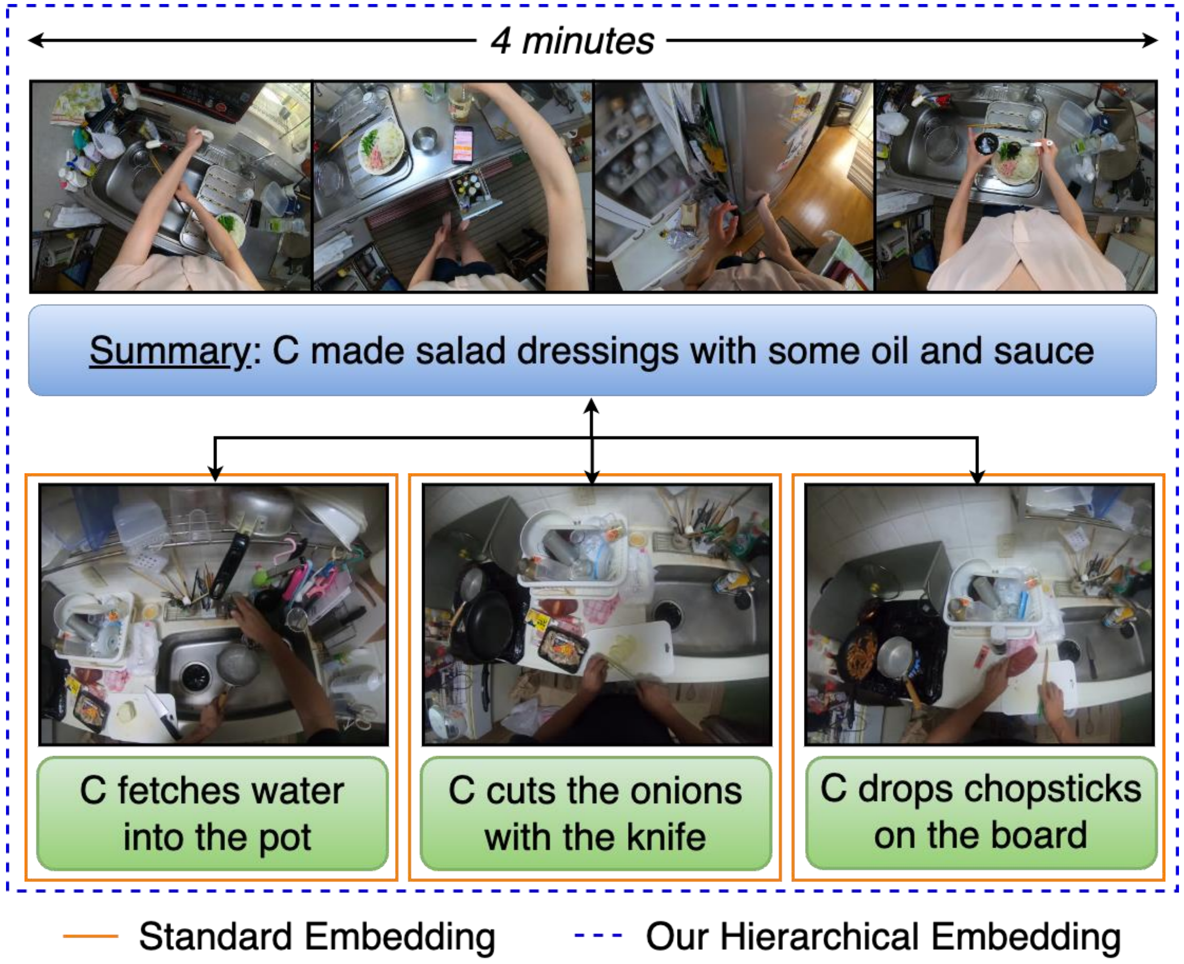
Dense
descriptive text
of each camera
wearer activity
+ clip-level
summaries

13 sentences
per minute

4M+ sentences

Hierarchical video-language learning

Our idea: video-language embedding learning, representing the hierarchical relationship between action descriptions (what) and higher-level summaries (why)



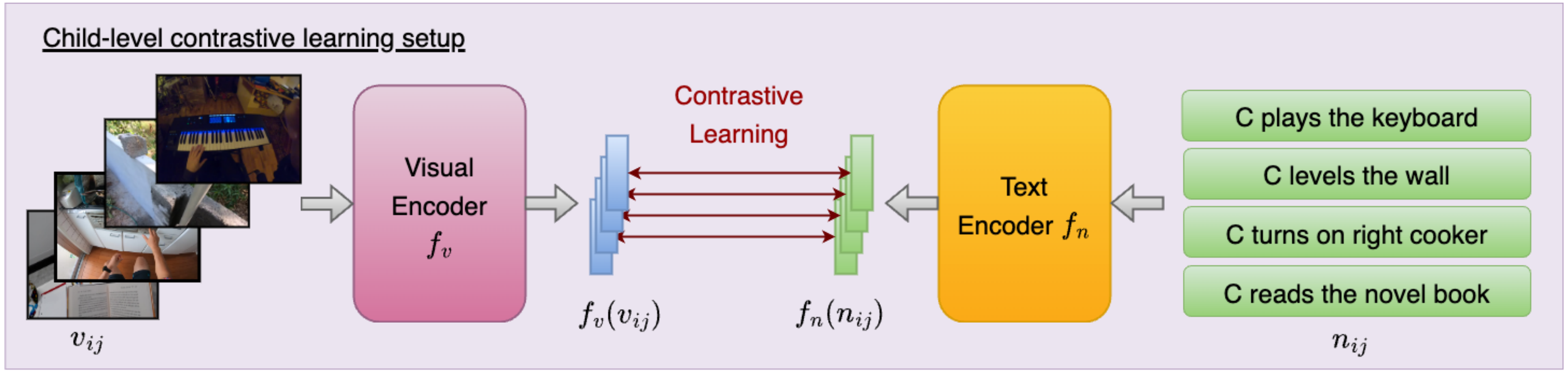
Existing methods: match short clips to corresponding narrations (Lin et al. 2022; Miech et al. 2020; Bain et al. 2021….)

Kumar et al. CVPR 2023

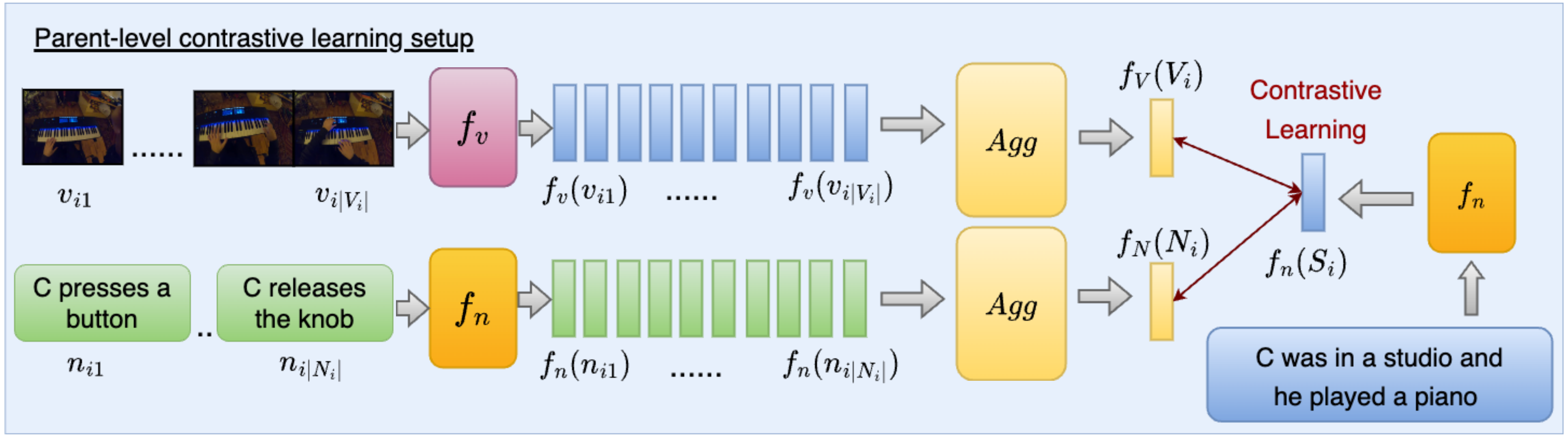
HierVL

Our idea: video-language embedding learning, representing the hierarchical relationship between action descriptions (what) and higher-level summaries (why)

Child-level (what)



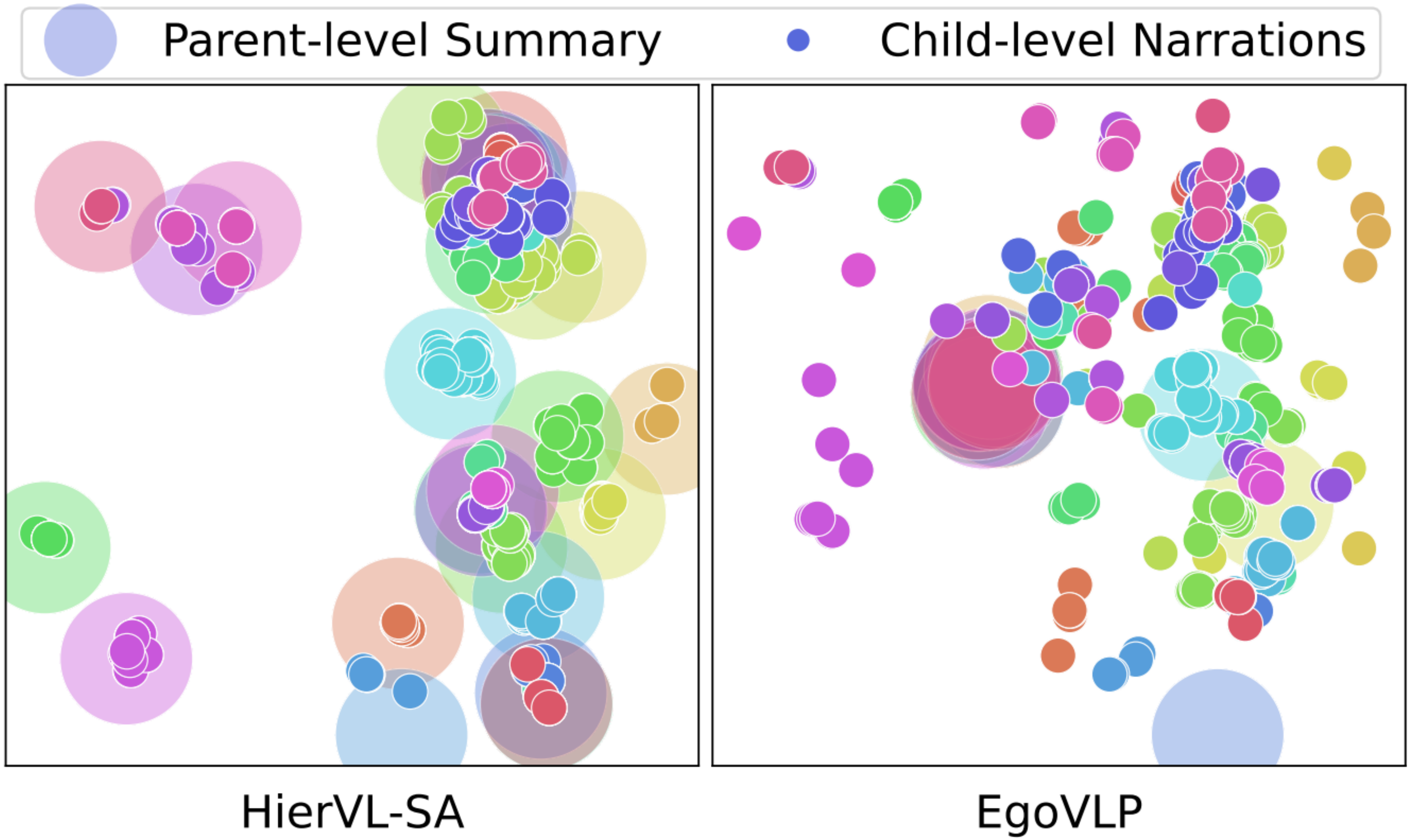
Parent-level (why)



Kumar et al. CVPR 2023

HierVL

T-SNE



Kumar et al. CVPR 2023

Lin et al. Neurips 2022

HierVL on downstream tasks

Pretrained HierVL features provide strong performance on multiple downstream video tasks

Method	mAP
Actor [69]	20.0
SSDA [12]	23.1
I3D [12]	25.8
Ego-Exo [49]	30.1
EgoVLP [52]	32.1
HierVL-w/o Hier	<u>32.6</u>
HierVL-Avg (Ours)	<u>32.6</u>
HierVL-SA (Ours)	**33.8**

CharadesEgo Action Recognition

Method	Verb ED ↓	Noun ED ↓	Act. ED ↓
Ego4D baseline [32]	0.7389	0.7800	0.9432
Robovision [16]	0.7389	0.7688	0.9412
I-CVAE [57]	0.7526	0.7489	<u>0.9308</u>
HierVL-w/o Hier	0.7691	<u>0.7454</u>	0.9451
HierVL-Avg (Ours)	**0.7223**	0.7527	0.9401
HierVL-SA (Ours)	<u>0.7239</u>	**0.7349**	**0.9275**

Ego4D Long Term Anticipation

Zero-shot		
Method	mAP Avg	nDCG Avg
EgoVLP [52]	16.6	23.1
HierVL-w/o Hier	<u>17.8</u>	<u>24.1</u>
HierVL-Avg (Ours)	16.7	23.5
HierVL-SA (Ours)	**18.9**	**24.7**
Fine-tuned		
Method	mAP Avg	nDCG Avg
MI-MM w/ S3D [84]	29.2	44.7
MME [79] w/ TBN [38]	38.5	48.5
JPoSE [79] w/ TBN [38]	44.0	53.5
EgoVLP [52]	<u>45.0</u>	59.4
HierVL-w/o Hier	44.7	<u>59.8</u>
HierVL-Avg (Ours)	44.9	<u>59.8</u>
HierVL-SA (Ours)	**46.7**	**61.1**

EPIC-KITCHENS Multi-Instance Retrieval

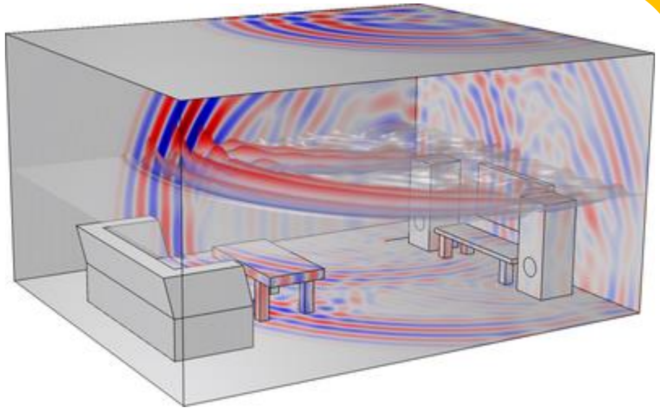
Kumar et al. CVPR 2023

Listening to learn about the visual world

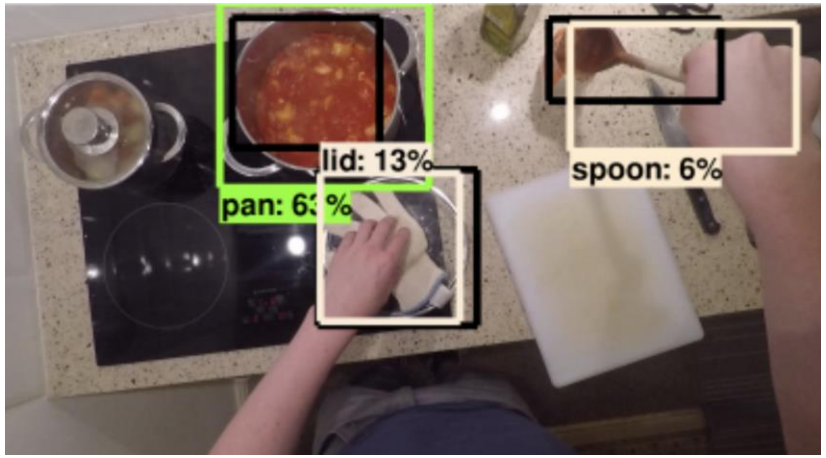

Object identity


Material properties


Emotion


3D space

Conversation



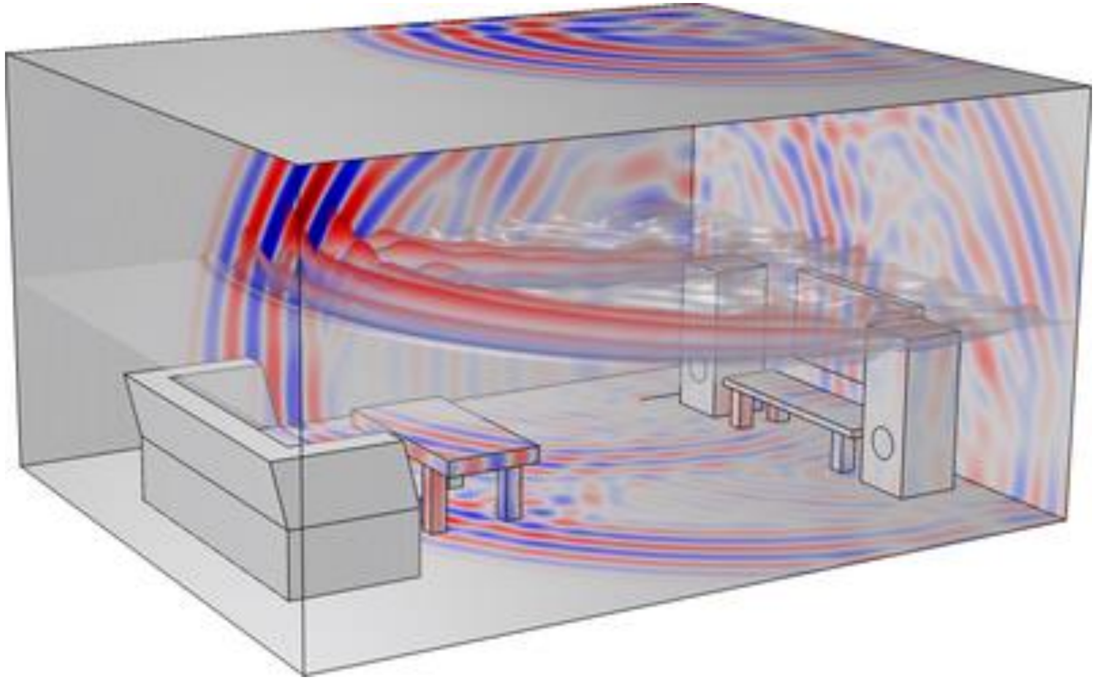
Egocentric activities

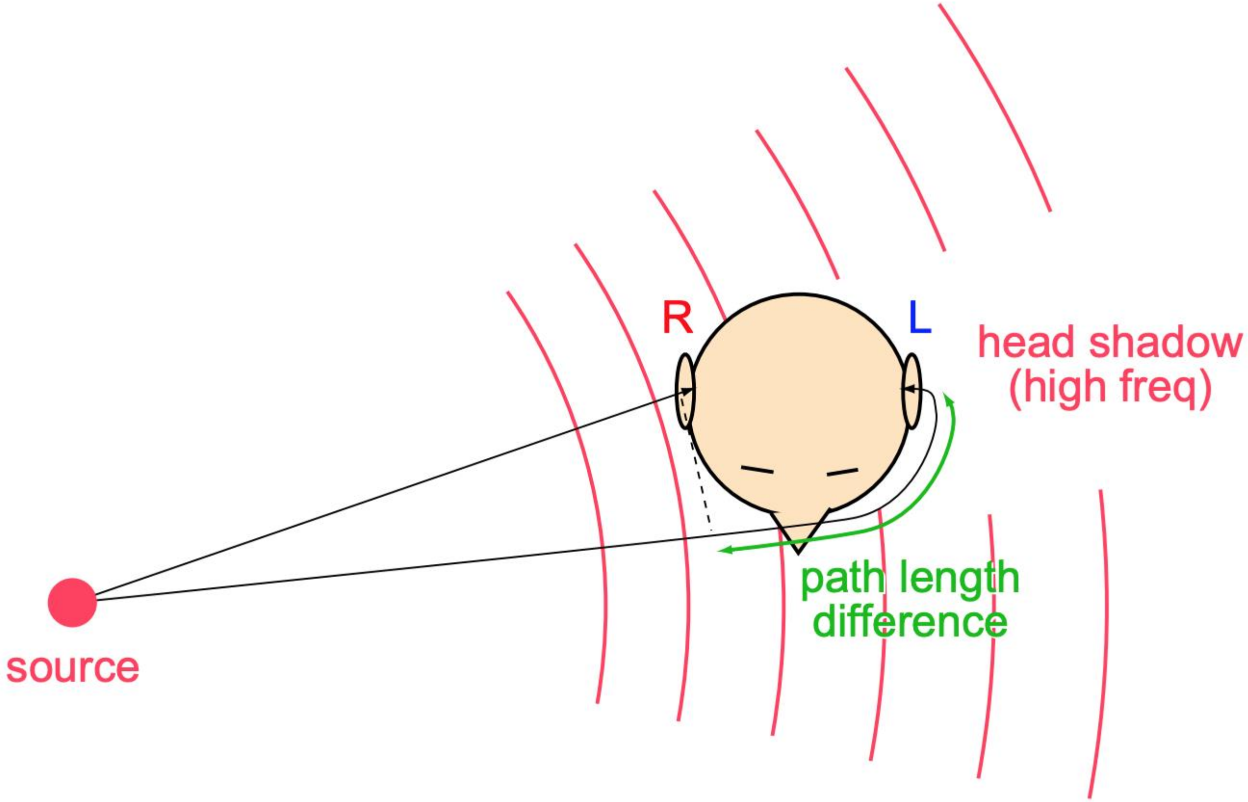

Ambient scene

1 drum kit, 5 different spaces

Source: Shred Shed Studio

Spatial effects in audio





Factors from 3D environment:
- Geometry of the space
- Materials in the room
- Position of source and receiver

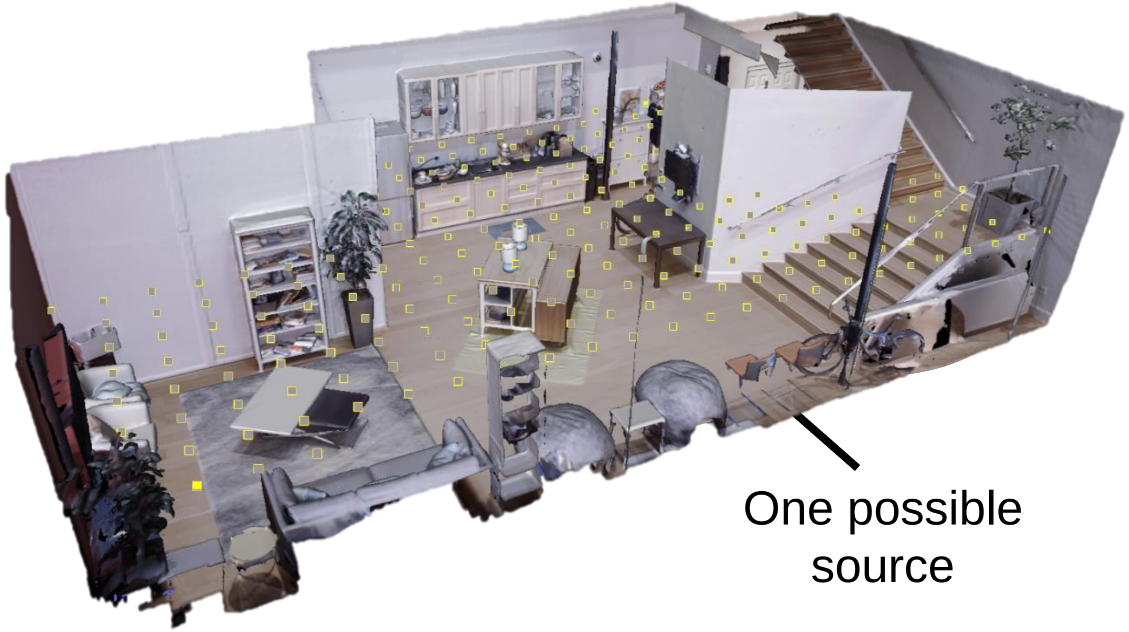
Agent’s spatial hearing cues:
- Interaural time difference (ITD),
- Interaural level difference (ILD)
- Spectral detail (from pinna reflections)

SoundSpaces audio simulation platform

C. Chen*, U. Jain*, et al., SoundSpaces, ECCV 2020; C. Chen et al. SoundSpaces 2.0, NeurIPS 2022

We introduce the *SoundSpaces* audio simulation platform

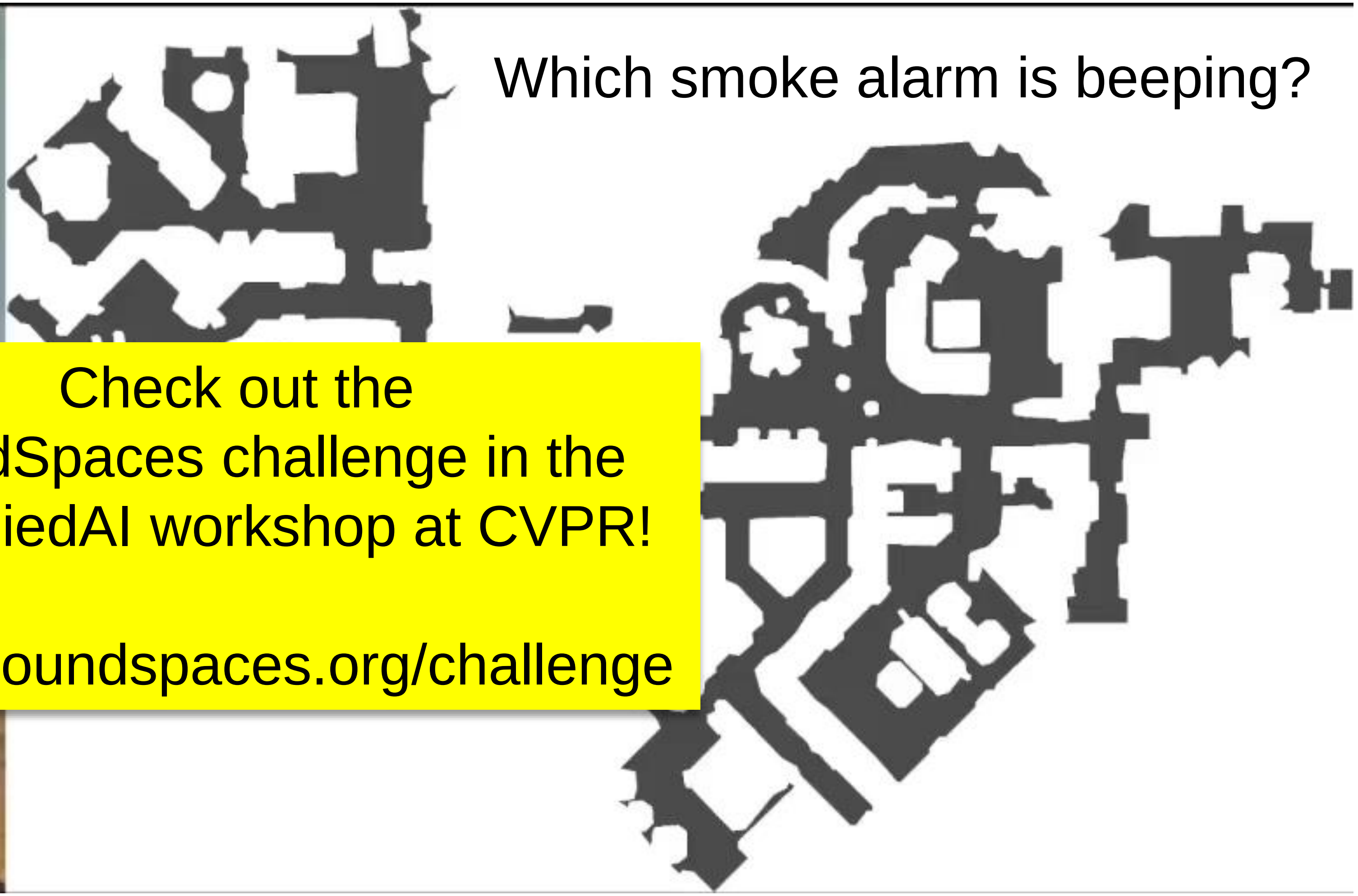
- Visually realistic real-world 3D environments (Matterport3D, Replica, Gibson, HM3D…)
- Acoustically realistic (geometry, materials, source location) binaural sound in real-time, for waveform of your choice
- Room impulse response (RIR) for any source *x* receiver location
- Habitat-compatible



SoundSpaces: https://github.com/facebookresearch/sound-spaces

SoundSpaces audio simulation

C. Chen*, U. Jain*, et al., SoundSpaces, ECCV 2020 & SoundSpaces 2.0, NeurIPS 2022



Which smoke alarm is beeping?

Check out the SoundSpaces challenge in the EmbodiedAI workshop at CVPR!

https://soundspaces.org/challenge

Agent view

Top-down map (unknown to the agent)

Listen with headphones online for spatial sound experience http://vision.cs.utexas.edu/projects/audio_visual_navigation/

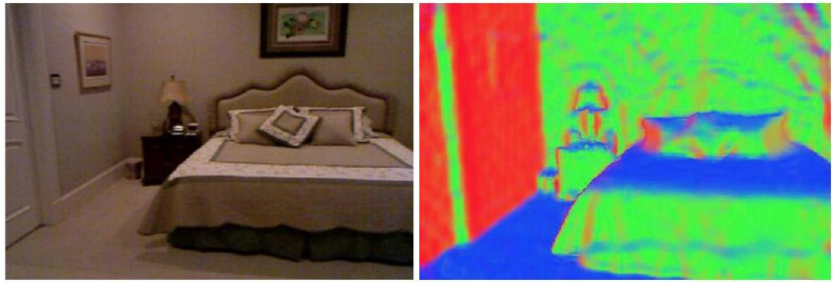
Recovering the shape of the scene

Daredevil (2003), Character Matt Murdock “sees” by listening

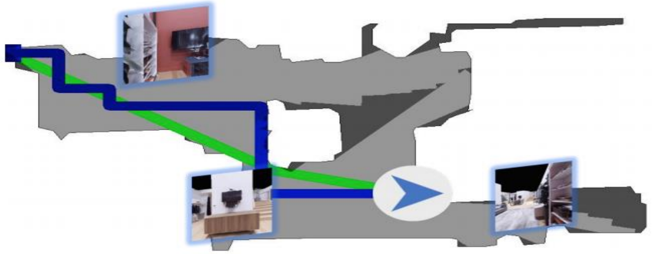
Our idea: VisualEchoes feature learning via echolocation

Goal: Learn image representation via echolocation to benefit downstream (visual-only) spatial tasks

Monocular depth prediction



Surface normal estimation



Visual navigation

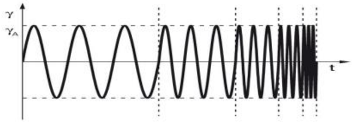
Key insight: supervision from acoustically interacting with the physical world.

Gao et al. ECCV 2020

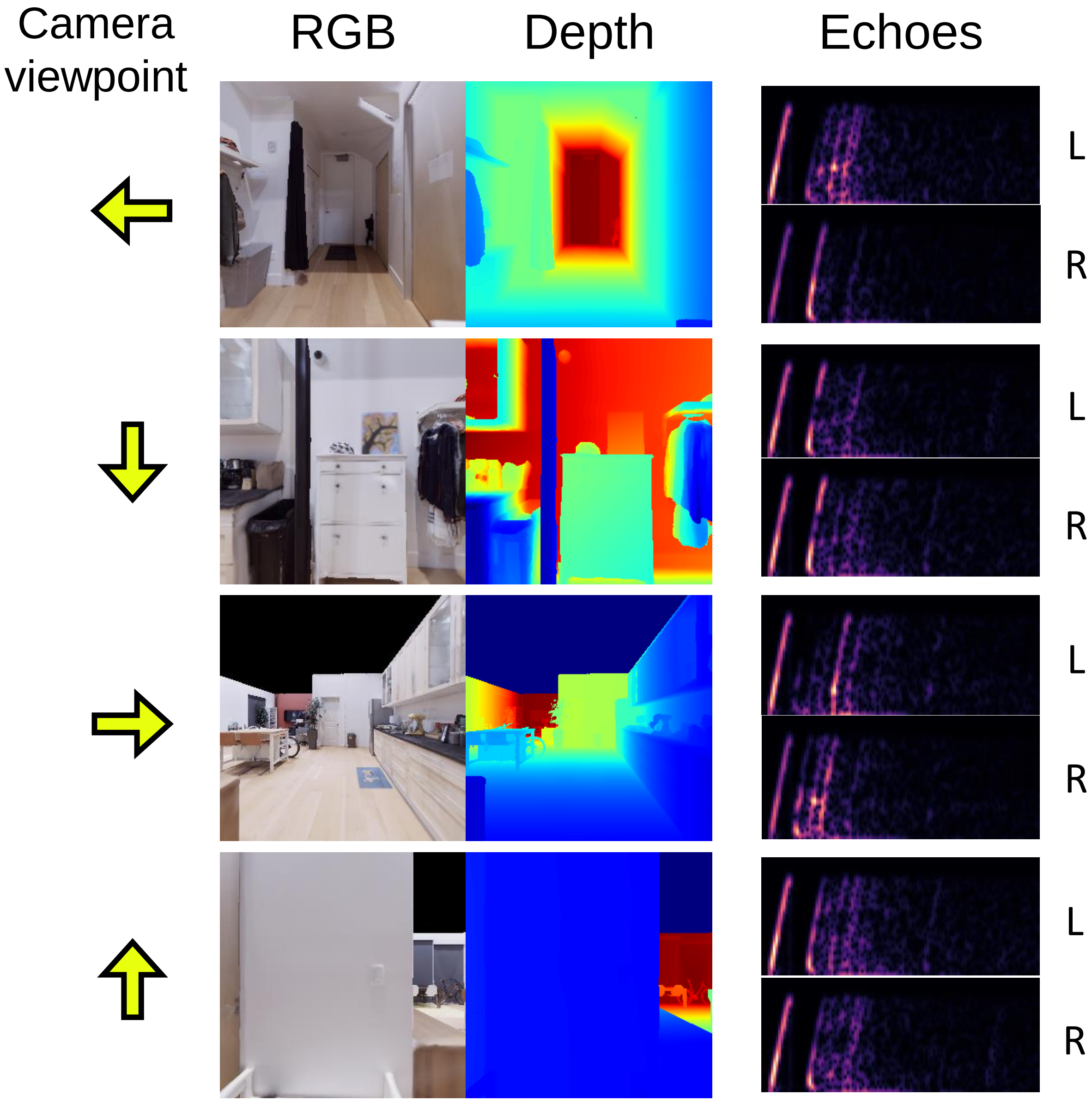
Echolocation in SoundSpaces

Emit a chirp at the receiver position and capture the resulting echoes

Top-down view of a Replica scene



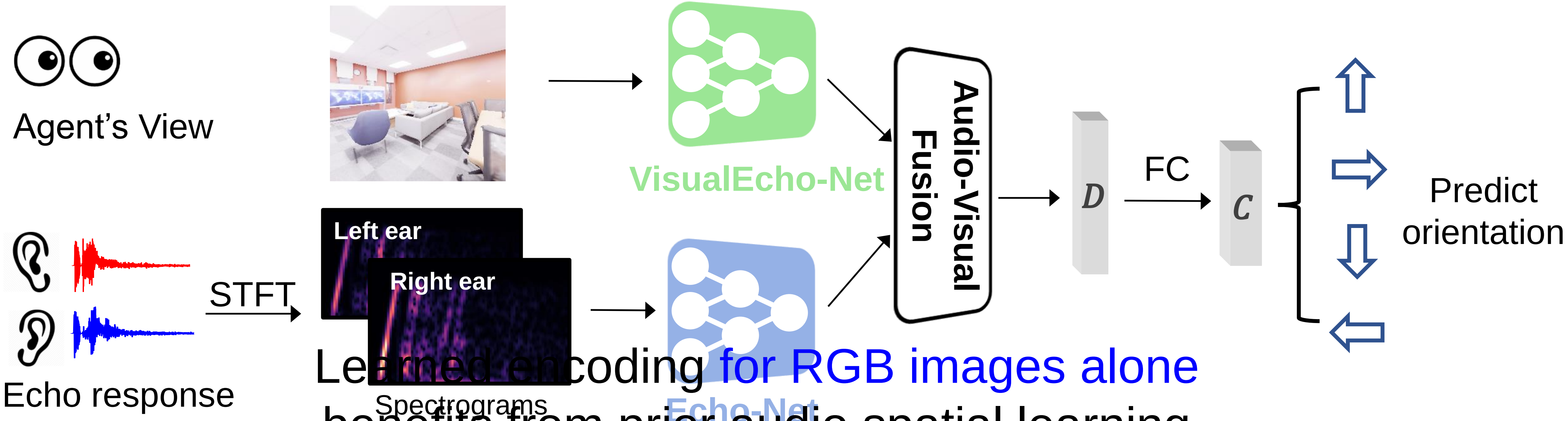
Freq Sweep



Gao et al. ECCV 2020

VisualEchoes approach

Learn *visual* representation from (in)congruence of echo and view



Learned encoding for RGB images alone benefits from prior audio spatial learning

Echo received from one of:

⇧ : the same orientation as the agent's current view

⇨ : the orientation if the agent turns right by 90°

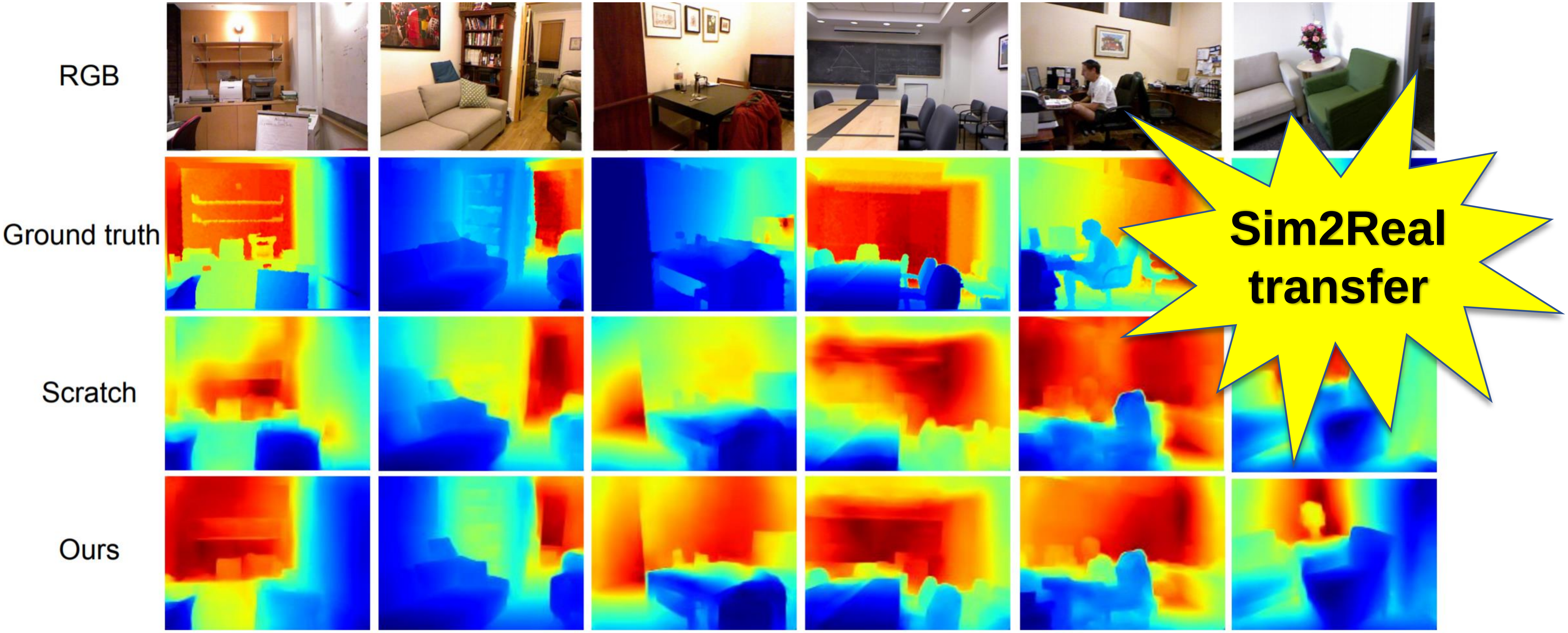
⇩ : the opposite orientation to the agent's current view

⇦ : the orientation if the agent turns left by 90°

VisualEchoes for downstream tasks

Pre-train monocular depth prediction CNN with VisualEchoes

No audio input, test on real images (NYU-V2 dataset)

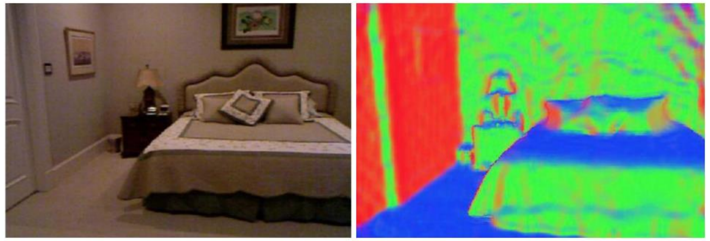


[Gao et al., ECCV 2020]

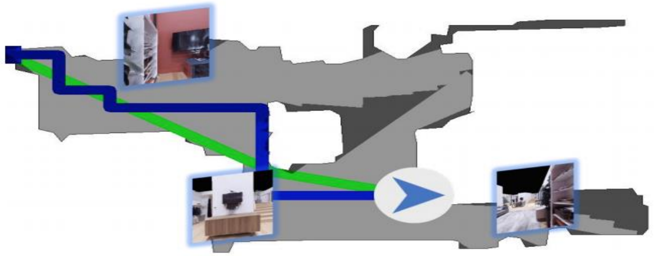
Hu et al., Revisiting single image depth estimation, WACV 2019

VisualEchoes for downstream tasks

Monocular depth prediction



Surface normal estimation



Visual navigation

		RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Sup	ImageNet Pre-trained	**0.555**	**0.126**	**0.054**	**0.843**	**0.968**	**0.991**
Sup	MIT Indoor Scene Pre-trained	0.711	0.180	0.075	0.730	0.925	0.979
Unsup	Scratch	0.804	0.209	0.086	0.676	0.897	0.967
Unsup	VISUALECHOES (Ours)	0.683	0.165	0.069	0.762	0.934	**0.981**

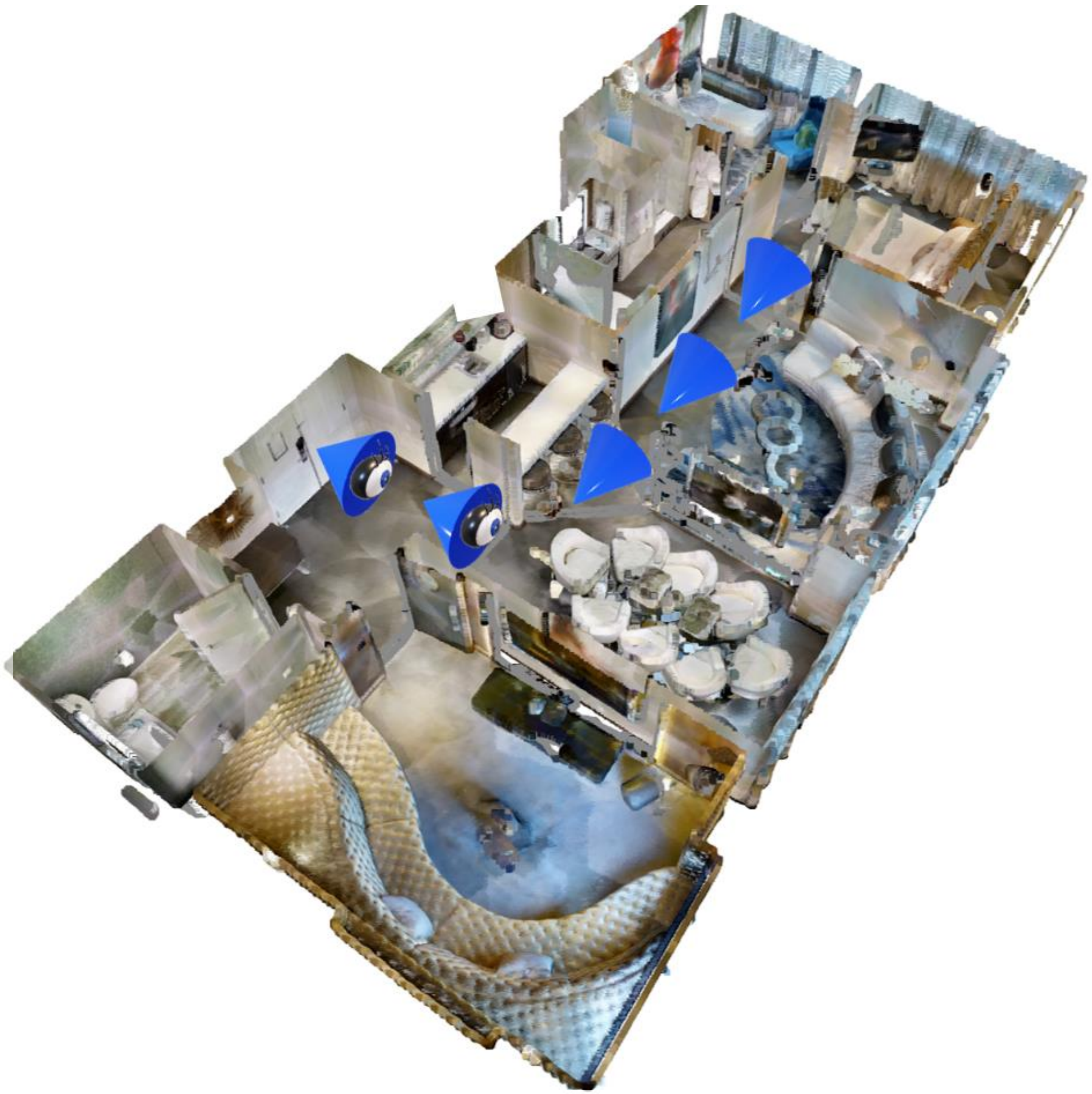
		Mean Dist. ↓	Median Dist. ↓	$t < 11.25°$ ↑	$t < 22.5°$ ↑	$t < 30°$ ↑
Sup	ImageNet Pre-trained	26.4	17.1	36.1	59.2	68.5
Sup	MIT Indoor Scene Pre-trained	25.2	17.5	36.5	57.8	67.2
Unsup	Scratch	26.3	16.1	37.9	60.6	69.0
Unsup	VISUALECHOES (Ours)	**22.9**	**14.1**	**42.7**	**64.1**	**72.4**

		SPL ↑	Distance to Goal ↓	Normalized Distance to Goal ↓
Sup	ImageNet Pre-trained	0.833	0.663	0.081
Sup	MIT Indoor Scene Pre-trained	0.798	1.05	0.124
Unsup	Scratch	0.830	0.728	0.096
Unsup	VISUALECHOES (Ours)	**0.856**	**0.476**	**0.061**

Competitive with (or even better than) supervised pre-training!

[Gao et al., ECCV 2020]

Audio-visual floorplans

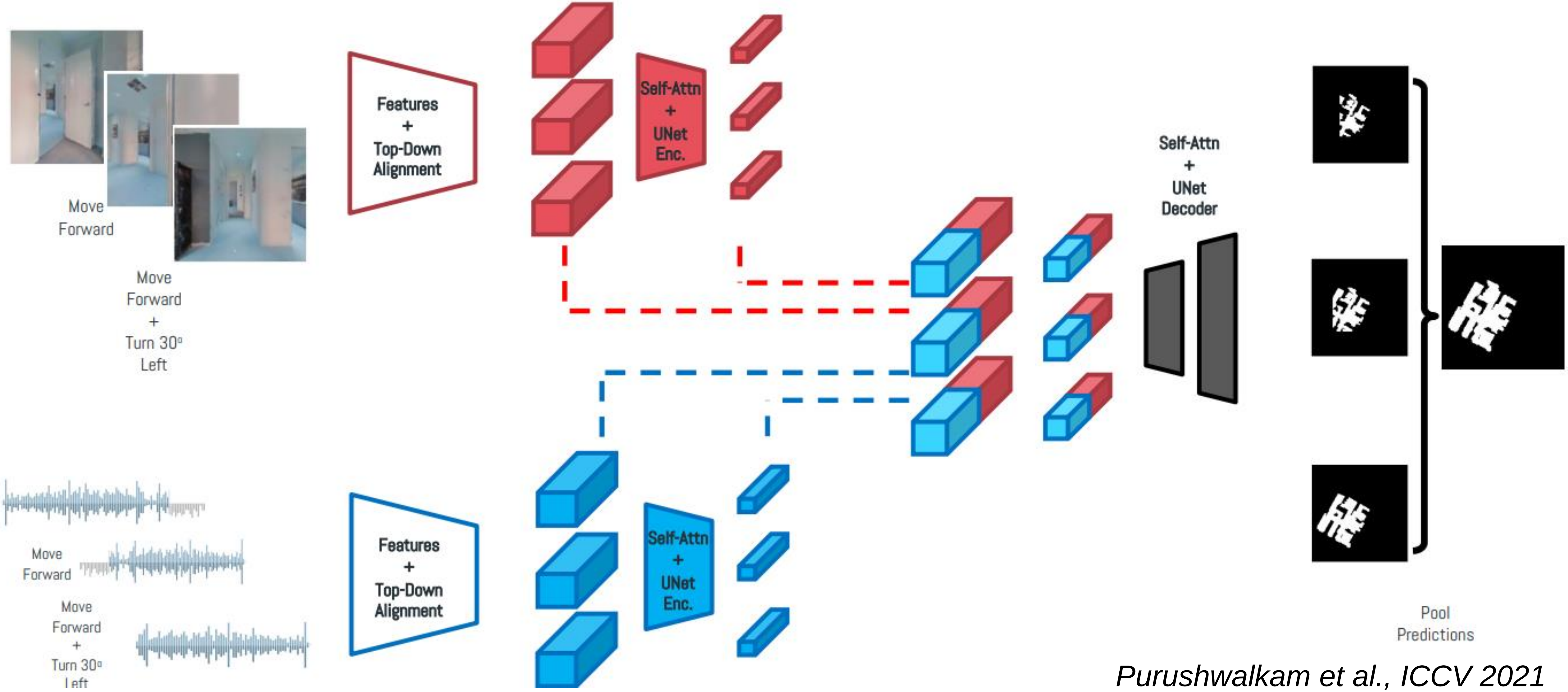


Given a short video, can we infer the layout of the entire home?

Purushwalkam et al., ICCV 2021

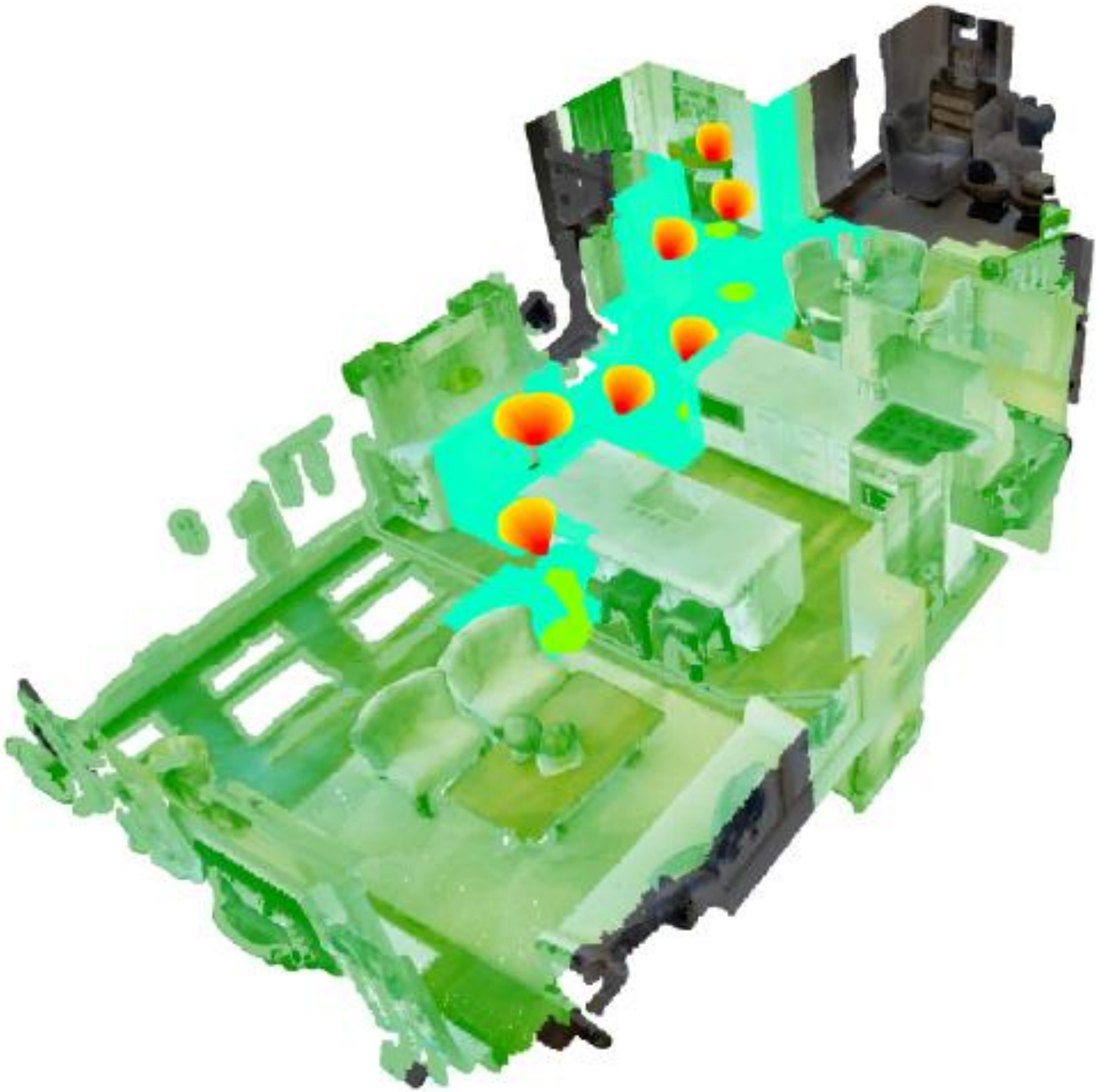
Audio-visual floorplans

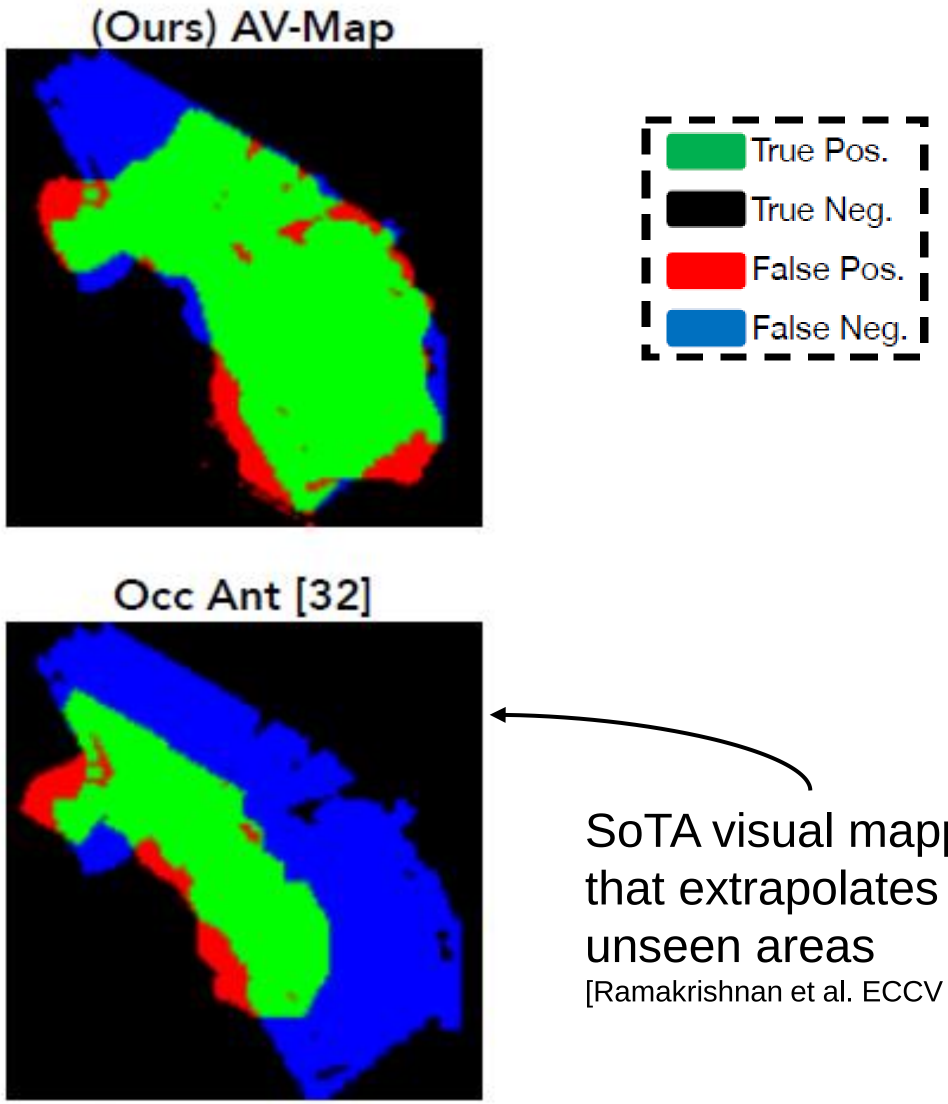
Supervised training of a multi-modal encoder-decoder network



Purushwalkam et al., ICCV 2021

Audio-visual floorplans

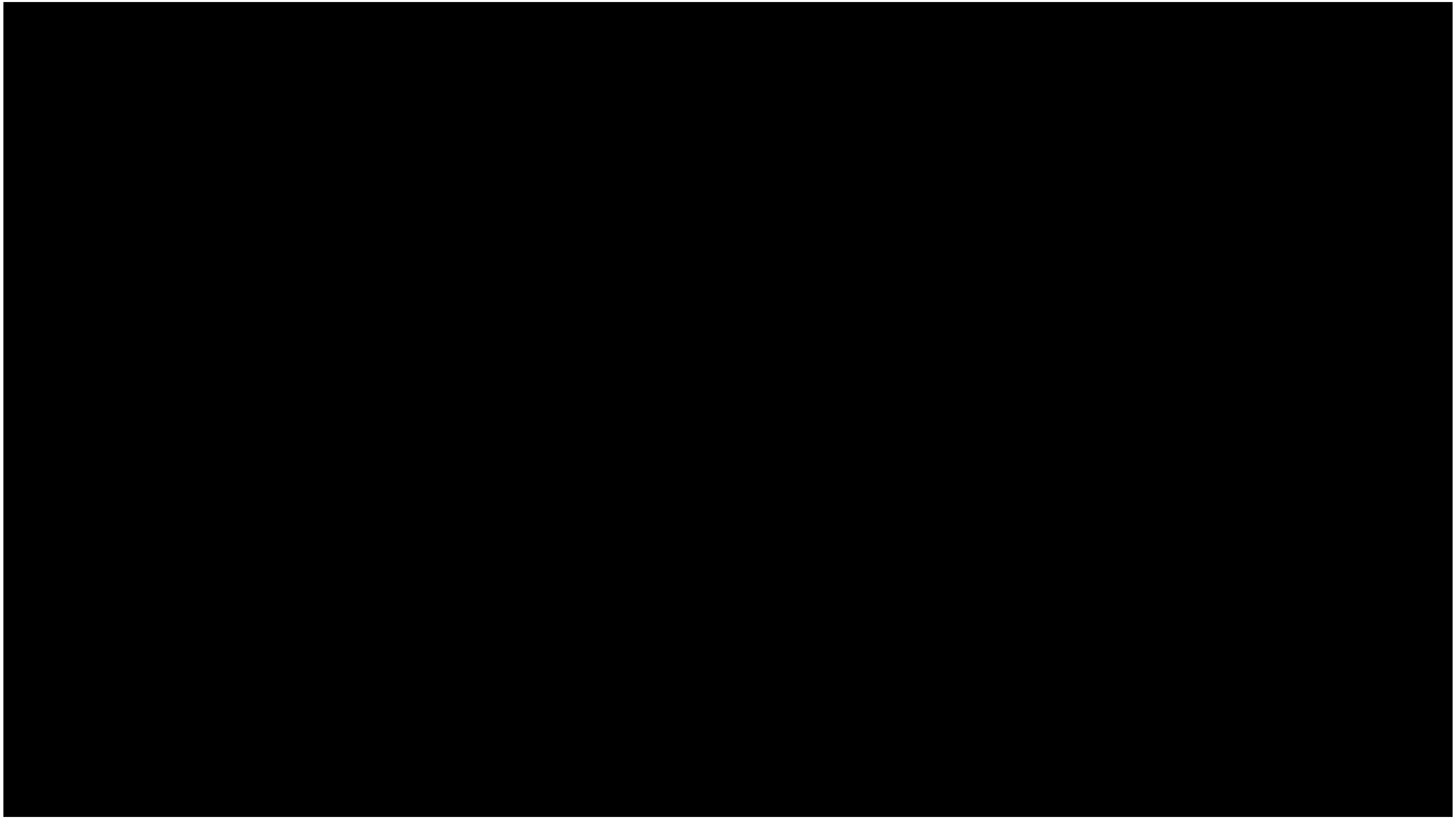




SoTA visual mapping that extrapolates to unseen areas
[Ramakrishnan et al. ECCV 2020]

Purushwalkam et al., ICCV 2021

Audio-visual floorplans



Kristen Grauman, FAIR & UT Austin

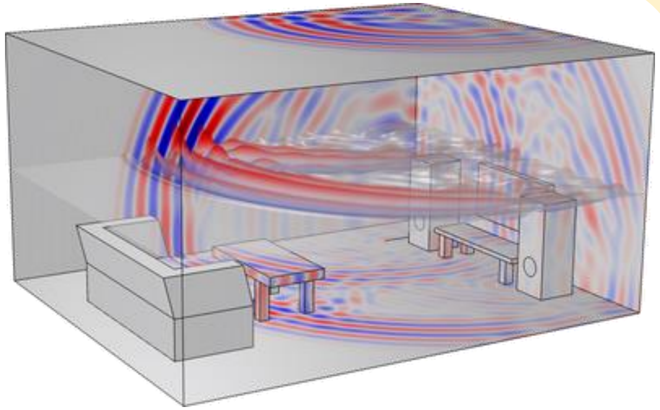
Purushwalkam et al., ICCV 2021

Listening to learn about the visual world

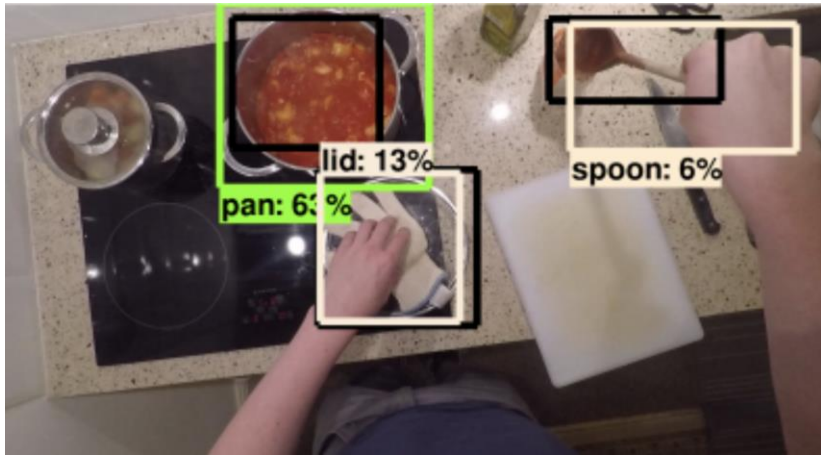

Object identity


Material properties


Emotion


3D space

Conversation

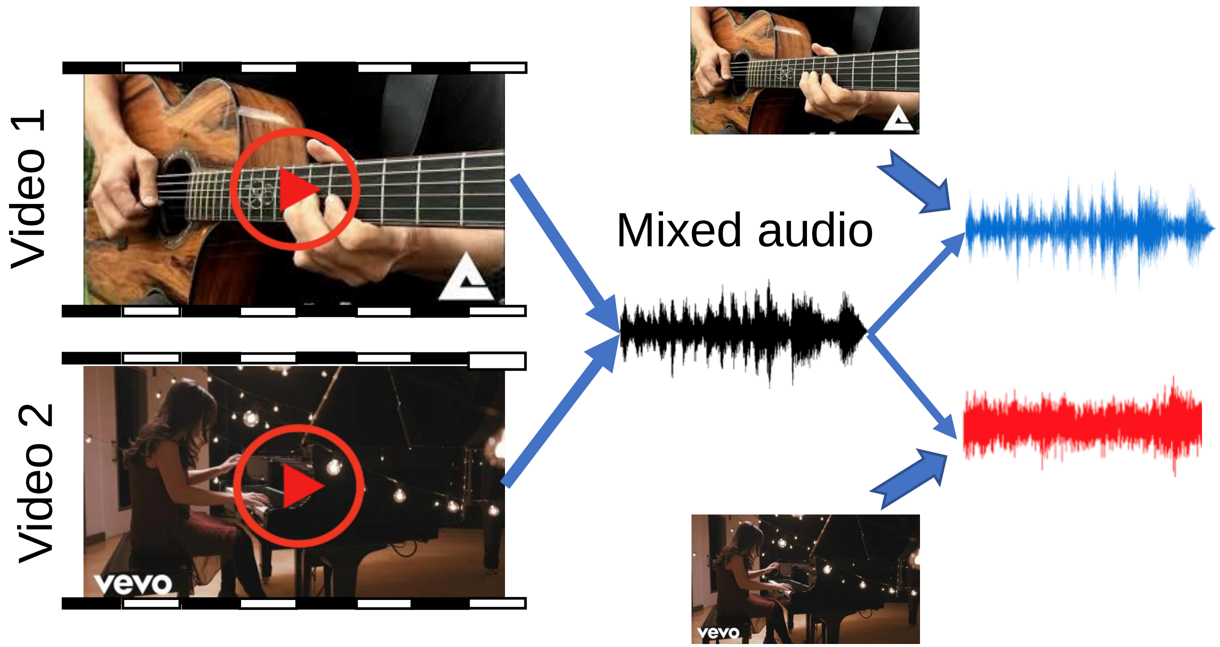


Egocentric activities


Ambient scene

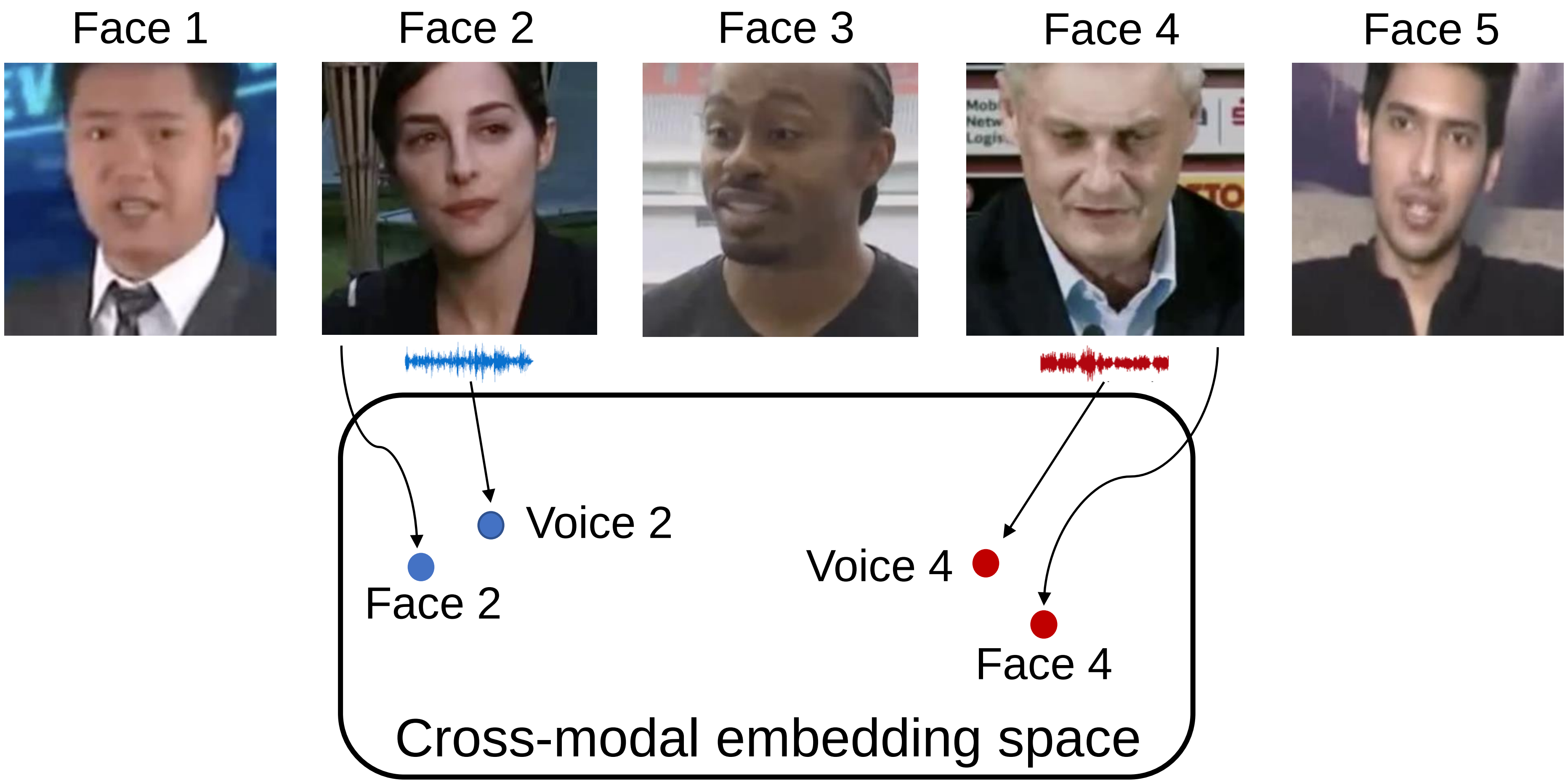
Cocktail-party problem

Self-supervising audio source separation: "Mix-and-Separate"



Simpson et al. 2015; Huang et al. 2015; Yu et al. 2017; Ephrat et al. 2018; Owens & Efros 2018; Zhao et al. 2018; Afouras et al. 2018; Zhao et al. 2019; Gao et al. 2019

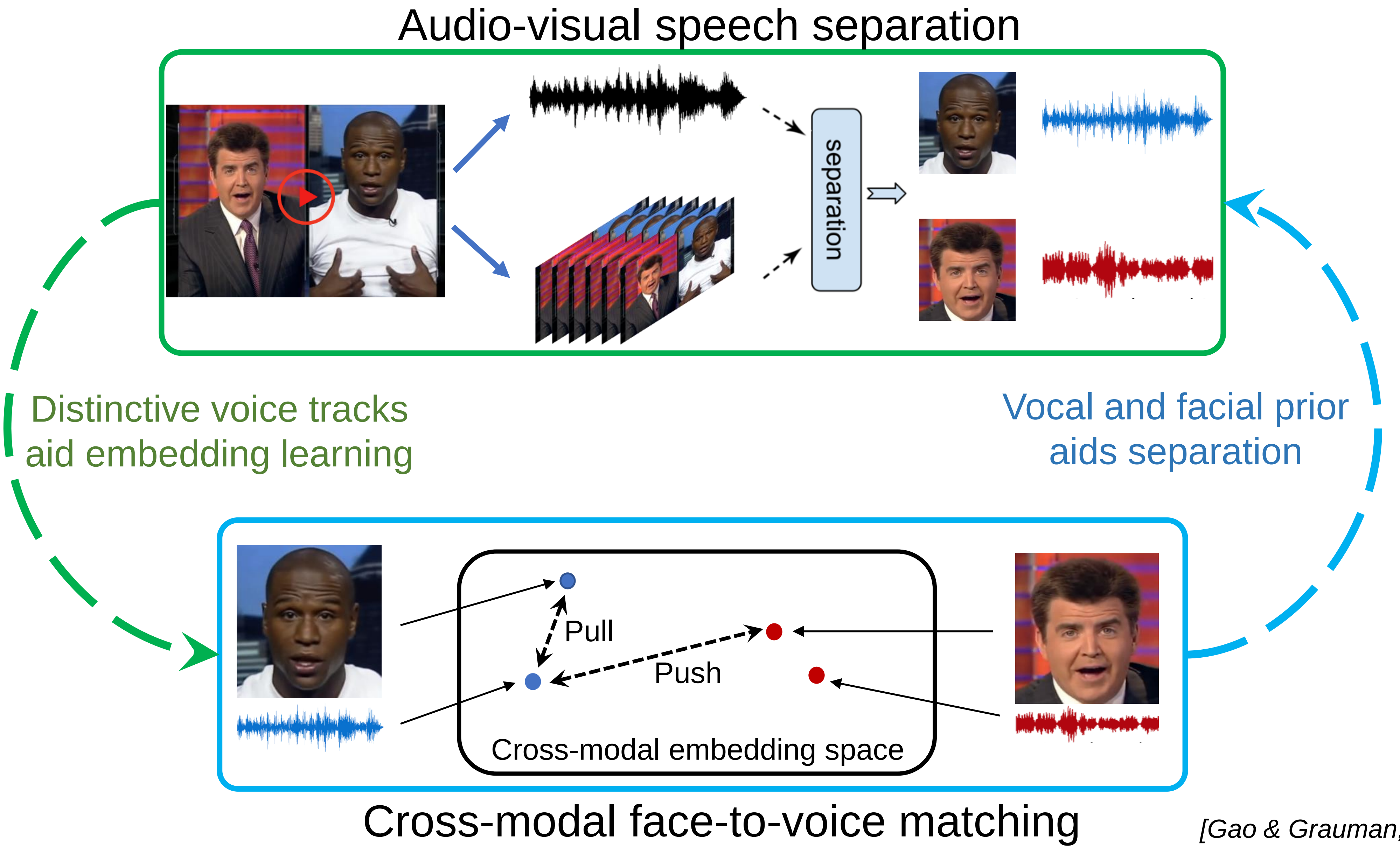
Facial appearance hints at voice qualities



Prior work learns cross-modal face-voice embeddings for person identification.

[Nagrani et al. ECCV’18, Nagrani et al. CVPR’18, Kim et al. ACCV’18, Chung et al. ICASSP’19, Wen et al. ICLR’19]

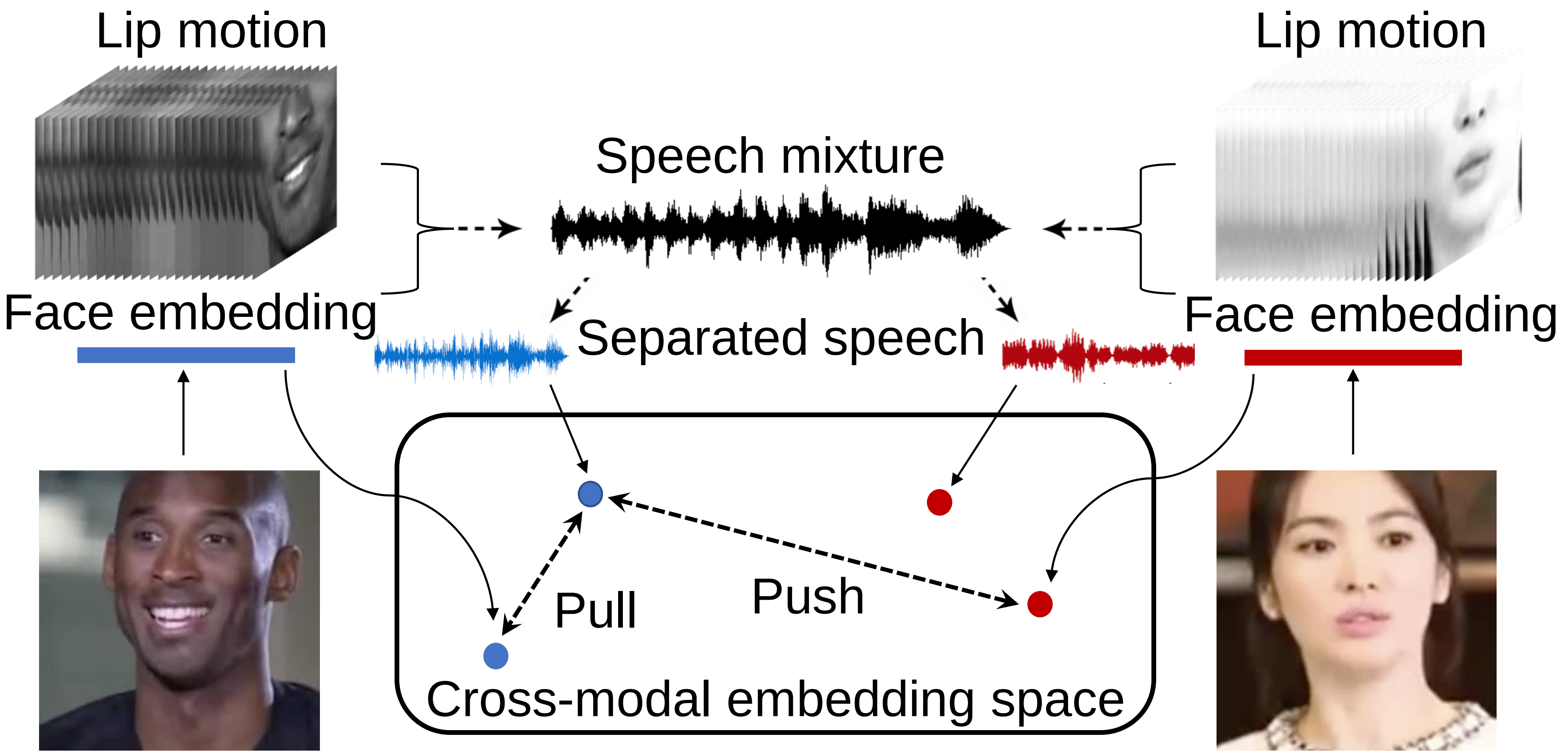
Our idea: Mutually beneficial tasks!



[Gao & Grauman, CVPR 2021]

VisualVoice

Jointly learn audio-visual speech separation and cross-modal face-voice embeddings

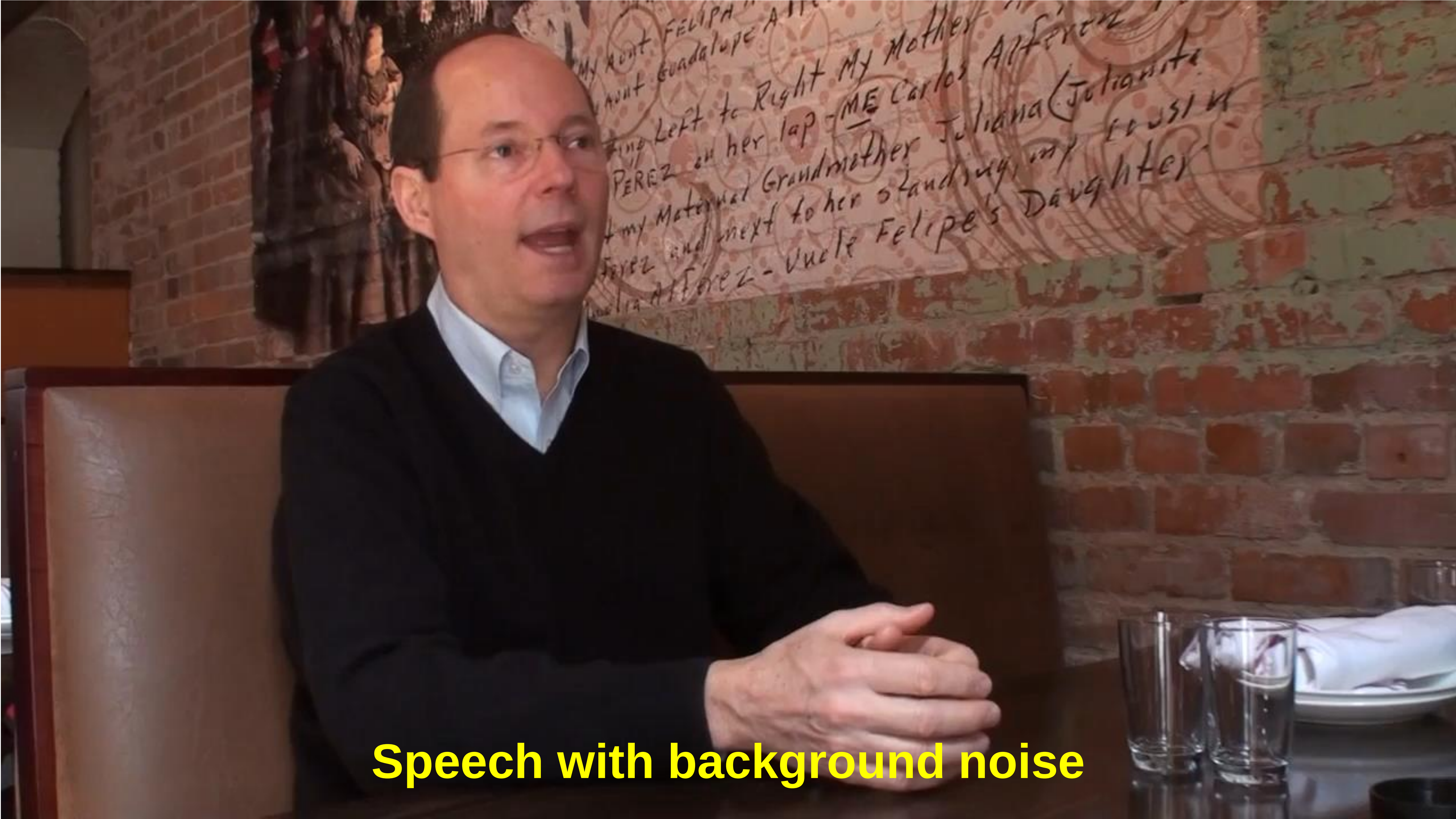


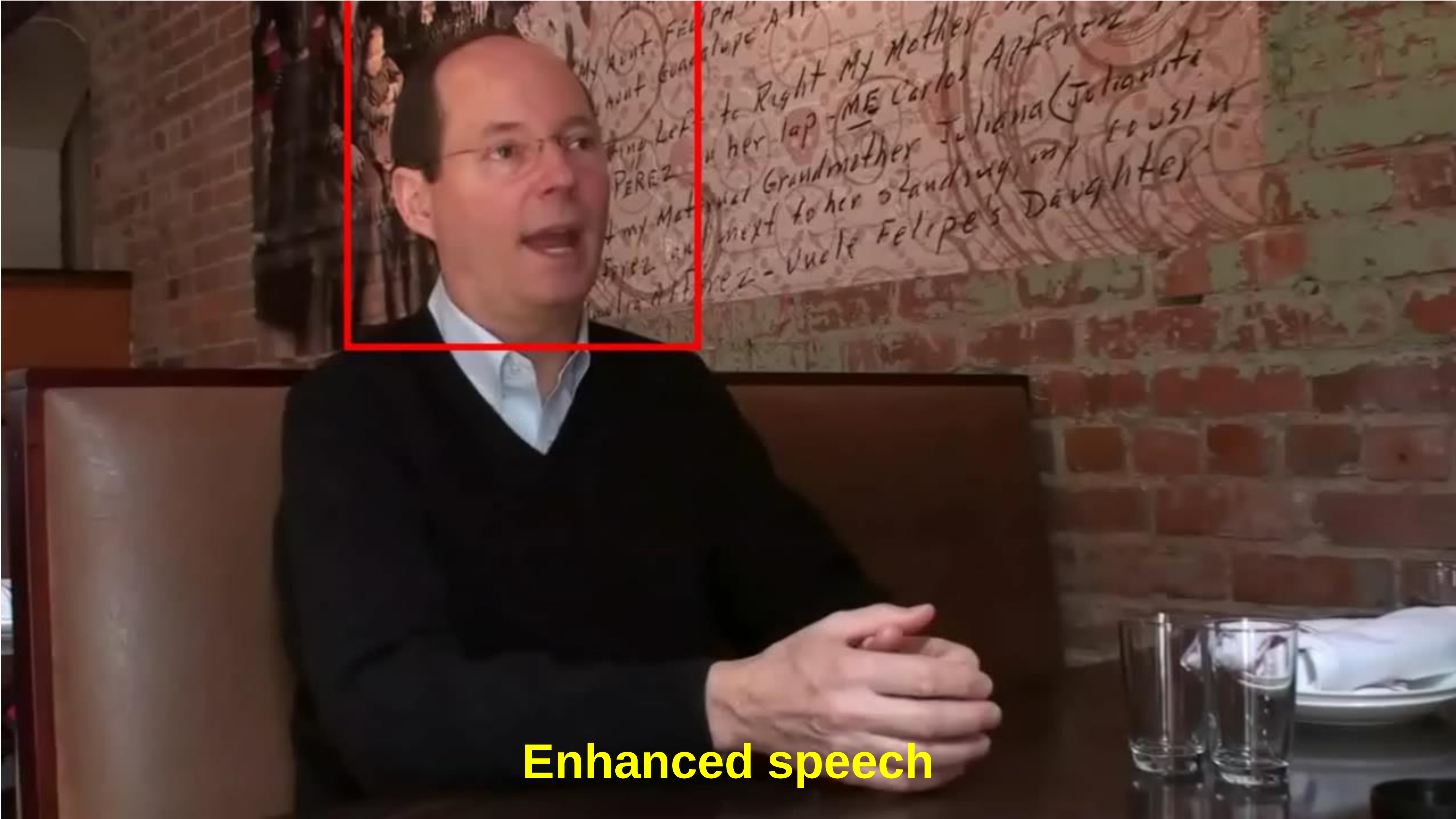
[Gao & Grauman, CVPR 2021]

Speech mixture


Separated voice for the right speaker


Separated voice for the left speaker


Speech with background noise


Enhanced speech

Speech with background noise


BBC
NEWS
Enhanced speech

VisualVoice vs. prior state-of-the-art methods

	Gabbay *et al.*	Hou *et al.*	Ephrat *et al.*	Ours
PESQ	2.25	2.42	2.50	**2.51**
STOI	–	0.66	0.71	**0.75**
SDR	–	2.80	6.10	**6.69**

(a) Results on Mandarin dataset.

	Gabbay *et al.*	Ephrat *et al.*	Ours
SDR	0.40	4.10	**10.9**
PESQ	2.03	2.42	**2.91**

(b) Results on TCD-TIMIT dataset.

	Casanovas *et al.*	Pu *et al.*	Ephrat *et al.*	Ours
SDR	7.0	6.2	12.6	**13.3**

(c) Results on CUAVE dataset.

	Afouras *et al.*	Afouras *et al.*	Ours
SDR	11.3	10.8	**11.8**
PESQ	3.0	3.0	**3.0**

(d) Results on LRS2 dataset.

	Chung *et al.*	Ours (static face)	Ours
SDR	2.53	7.21	**10.2**

(e) Results on VoxCeleb2 dataset.

Our method improves the state-of-the-art on all five datasets.

Summary

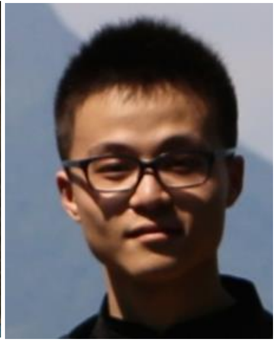
Kristen Grauman
UT Austin and FAIR
grauman@cs.utexas.edu

Towards embodied multimodal first-person perception

- Ego4D: massive multimodal first-person data and benchmark
- Hierarchical vision-language embedding to capture goals with actions
- Inferring the shape of a scene with echoes, sounds, and vision
- Audio-visual source separation to listen to voice of interest

Consortium


Changan Chen

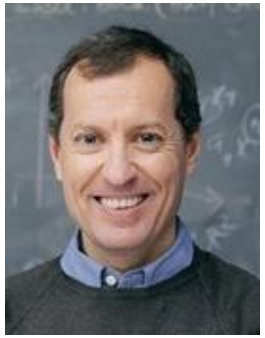

Ruohan Gao


Senthil Purushwalkam


Ashutosh Kumar


Ziad Al-Halah


Rohit Girdhar


Lorenzo Torresani