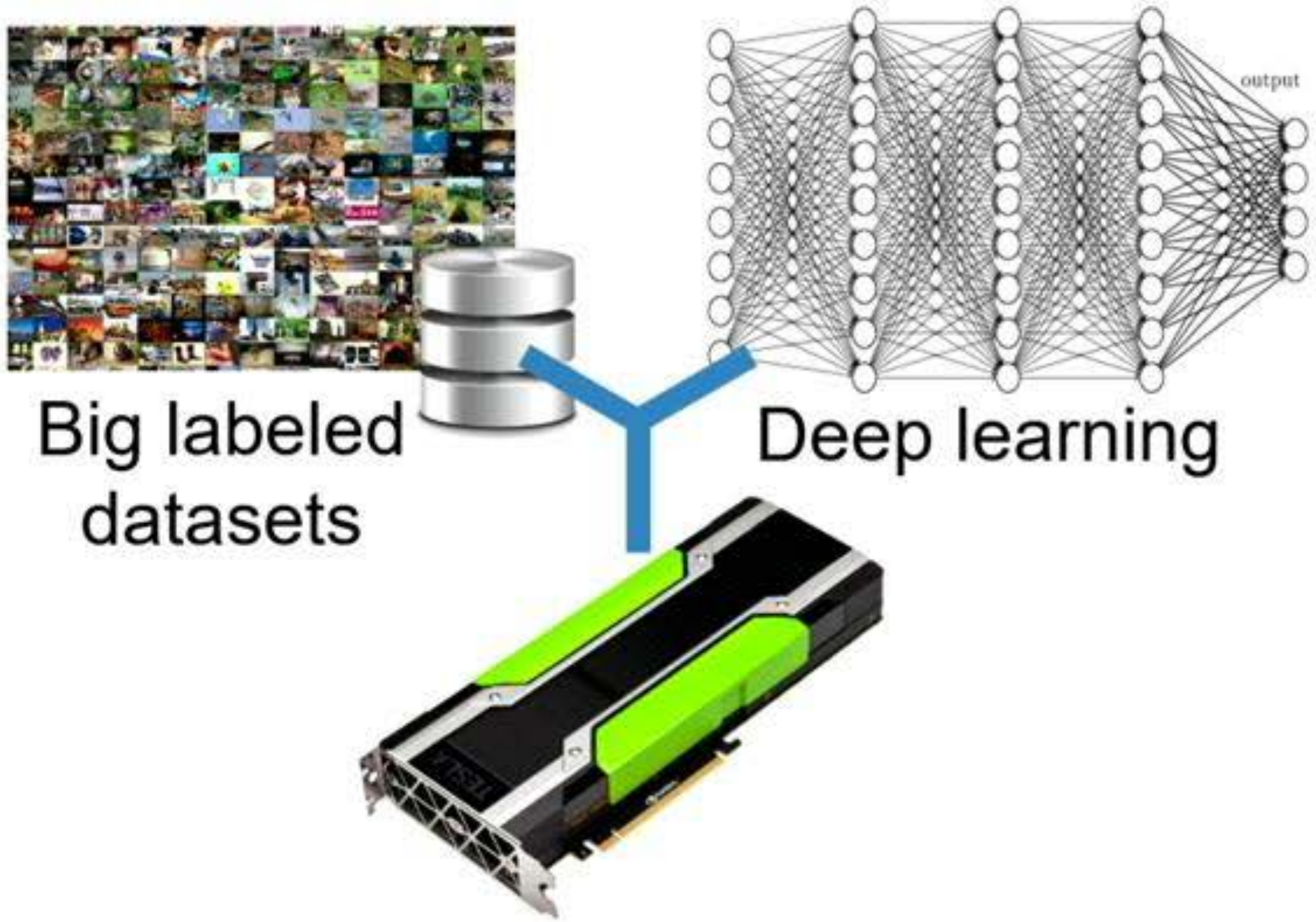
First-person perception and interaction

Kristen Grauman
University of Texas at Austin
Facebook AI Research

Visual recognition: significant progress

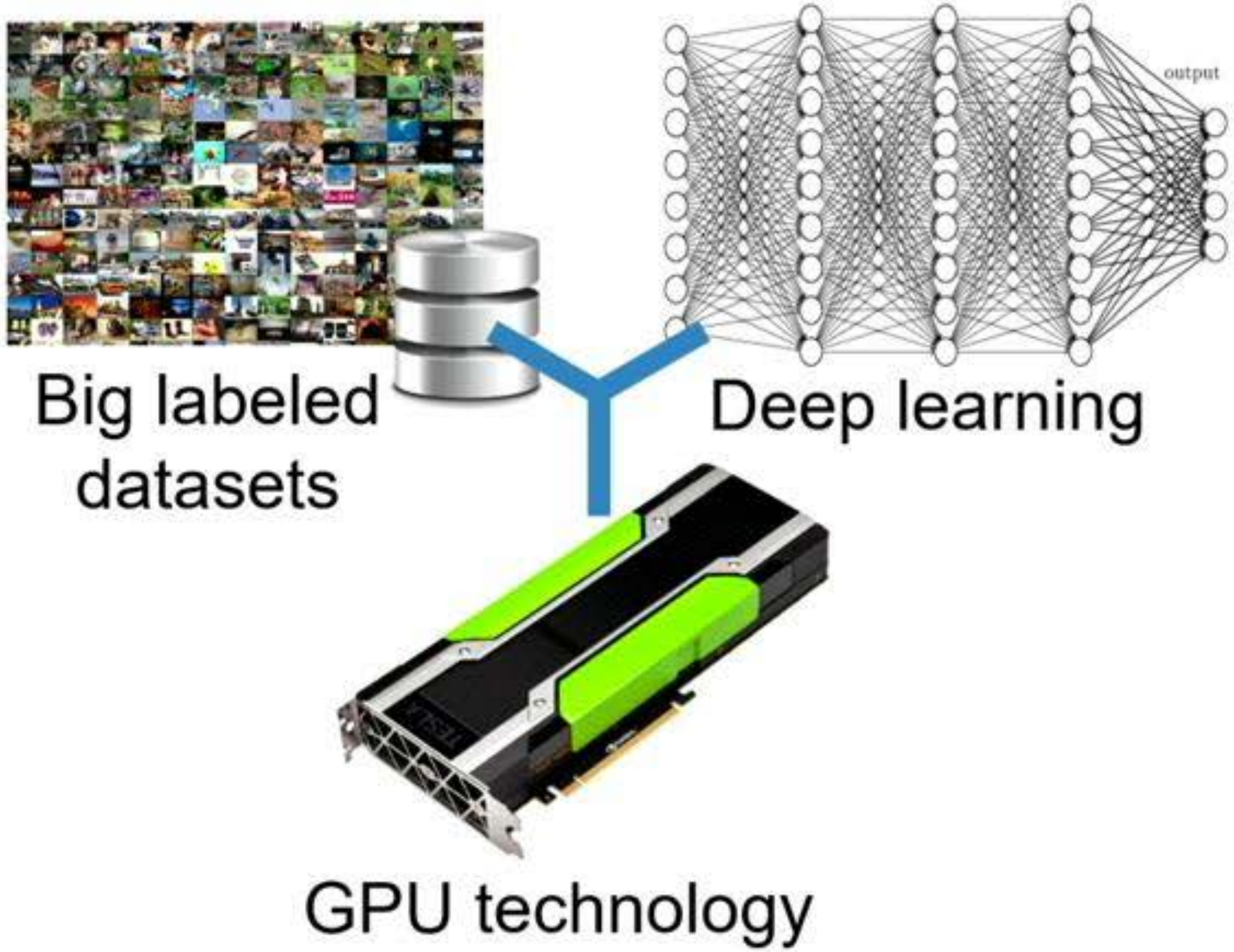


Big labeled datasets

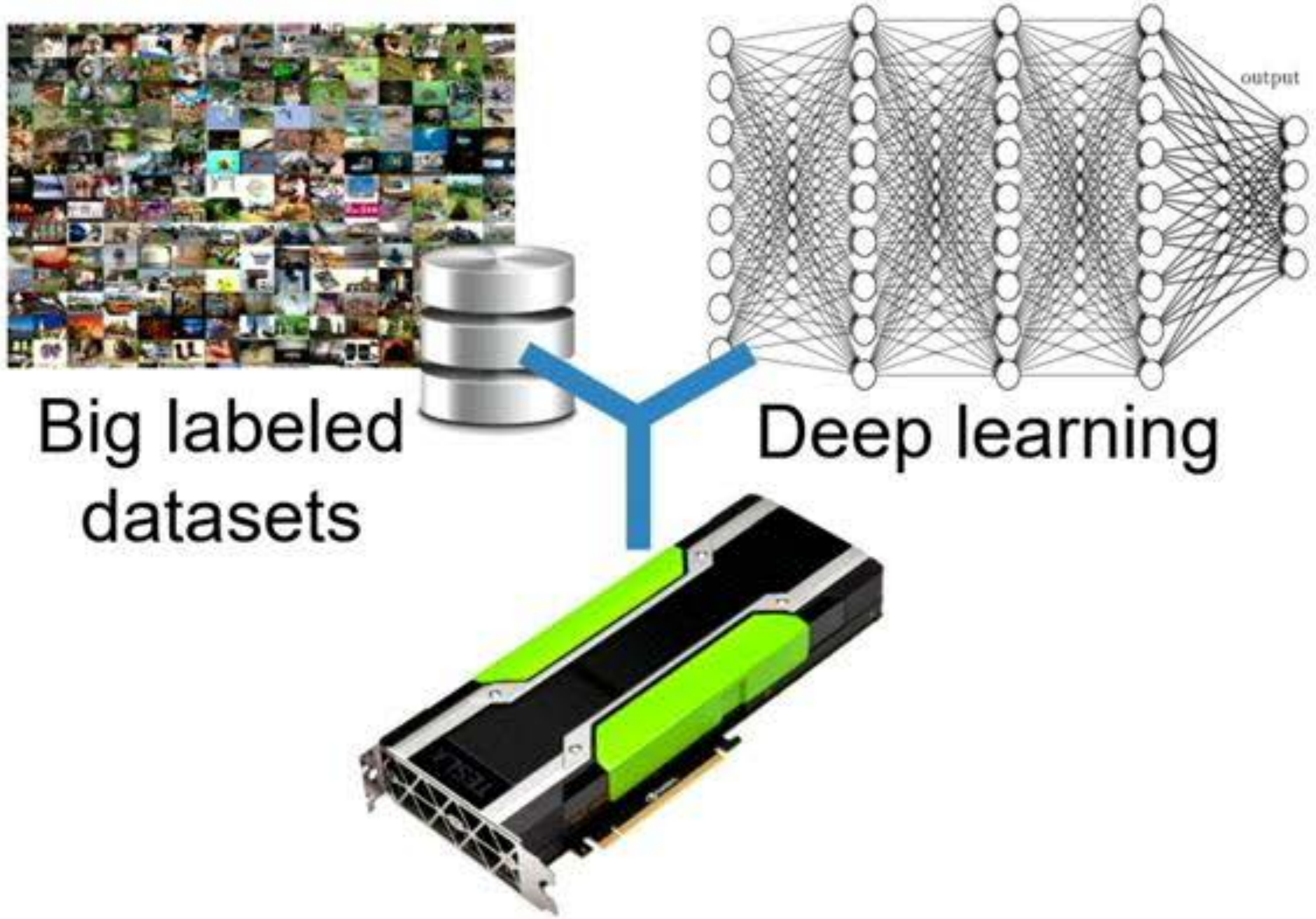
Deep learning

GPU technology

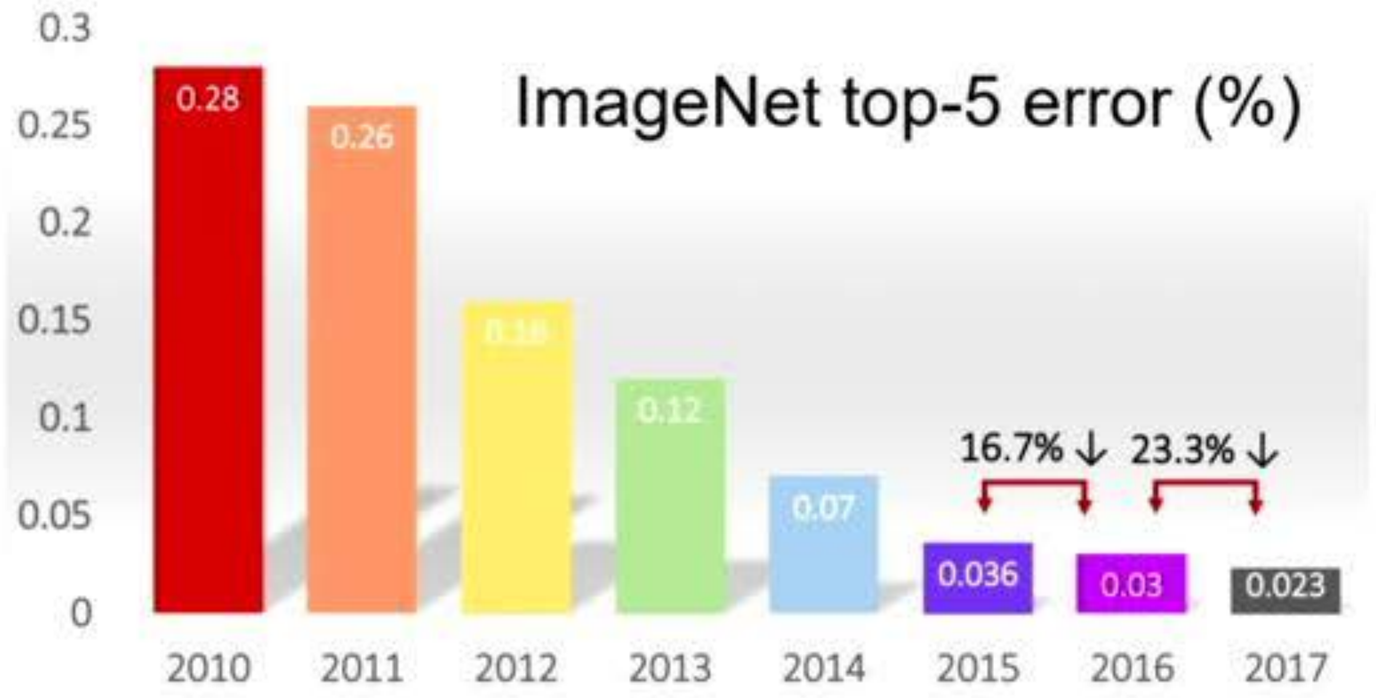
Visual recognition: significant progress



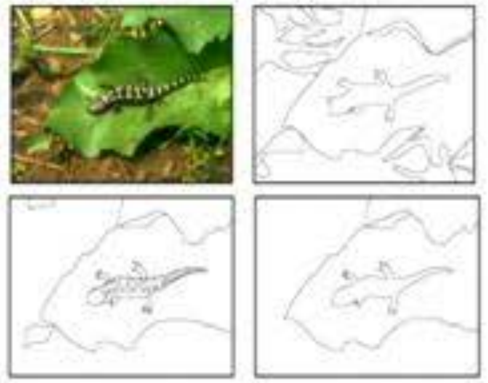
Visual recognition: significant progress



GPU technology

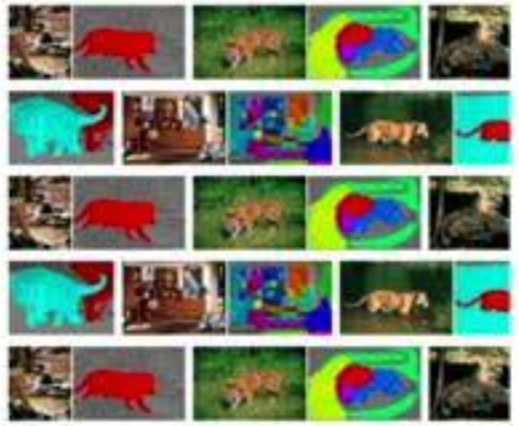


The Web photo perceptual experience


BSD (2001)


Caltech 101 (2004), Caltech 256 (2006)


PASCAL (2007-12)


LabelMe (2007)


ImageNet (2009)


SUN (2010)


Places (2014)


MS COCO (2014)


Visual Genome (2016)

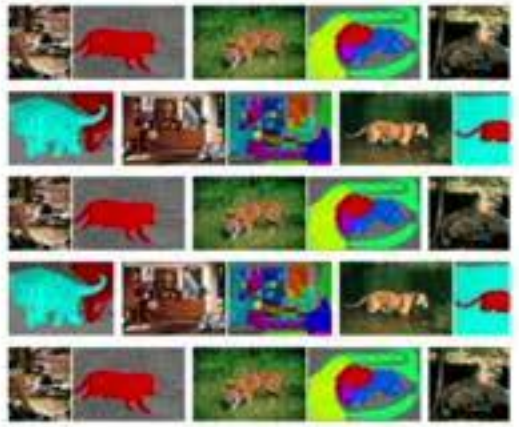
The Web photo perceptual experience

A "disembodied" well-curated moment in time

BSD (2001)

Caltech 101 (2004), Caltech 256 (2006)

PASCAL (2007-12)



LabelMe (2007)

ImageNet (2009)

SUN (2010)

Places (2014)

MS COCO (2014)

Visual Genome (2016)

First-person perception and learning

Status quo:

Learning and inference with “disembodied” photos.

First-person perception and learning

Status quo:

Learning and inference with "disembodied" photos.

On the horizon:

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

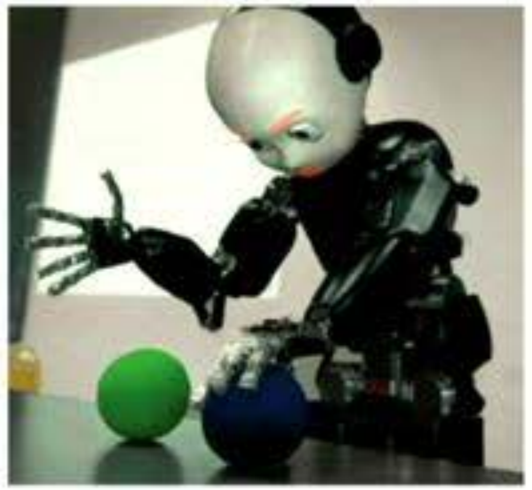
First-person perception and learning

Status quo:

Learning and inference with "disembodied" photos.

On the horizon:

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



This talk

Main idea:

Towards embodied perception

This talk

Main idea:

Towards embodied perception
via agents that learn to anticipate their perceptual
experience as a function of their own actions

This talk

Multi-sensory | **Motion** | **Interaction**

Towards embodied perception

This talk

Multi-sensory

Audio-visual learning

Motion

Interaction

Towards embodied perception

This talk

Multi-sensory

Audio-visual learning

Motion

Navigation policies

Interaction

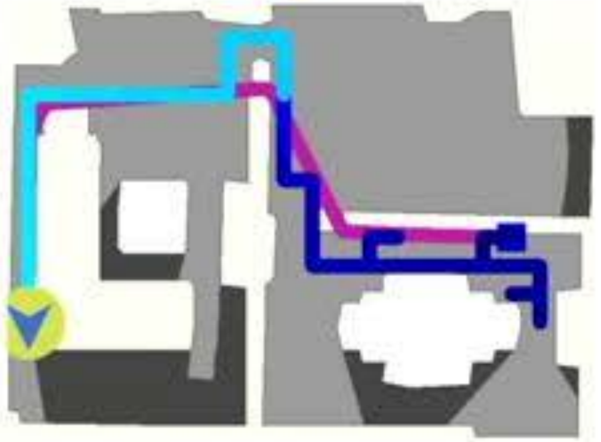
Towards embodied perception

This talk

Multi-sensory

Audio-visual learning

Motion



Navigation policies

Interaction

Affordance learning

Towards embodied perception

Spatial effects in audio

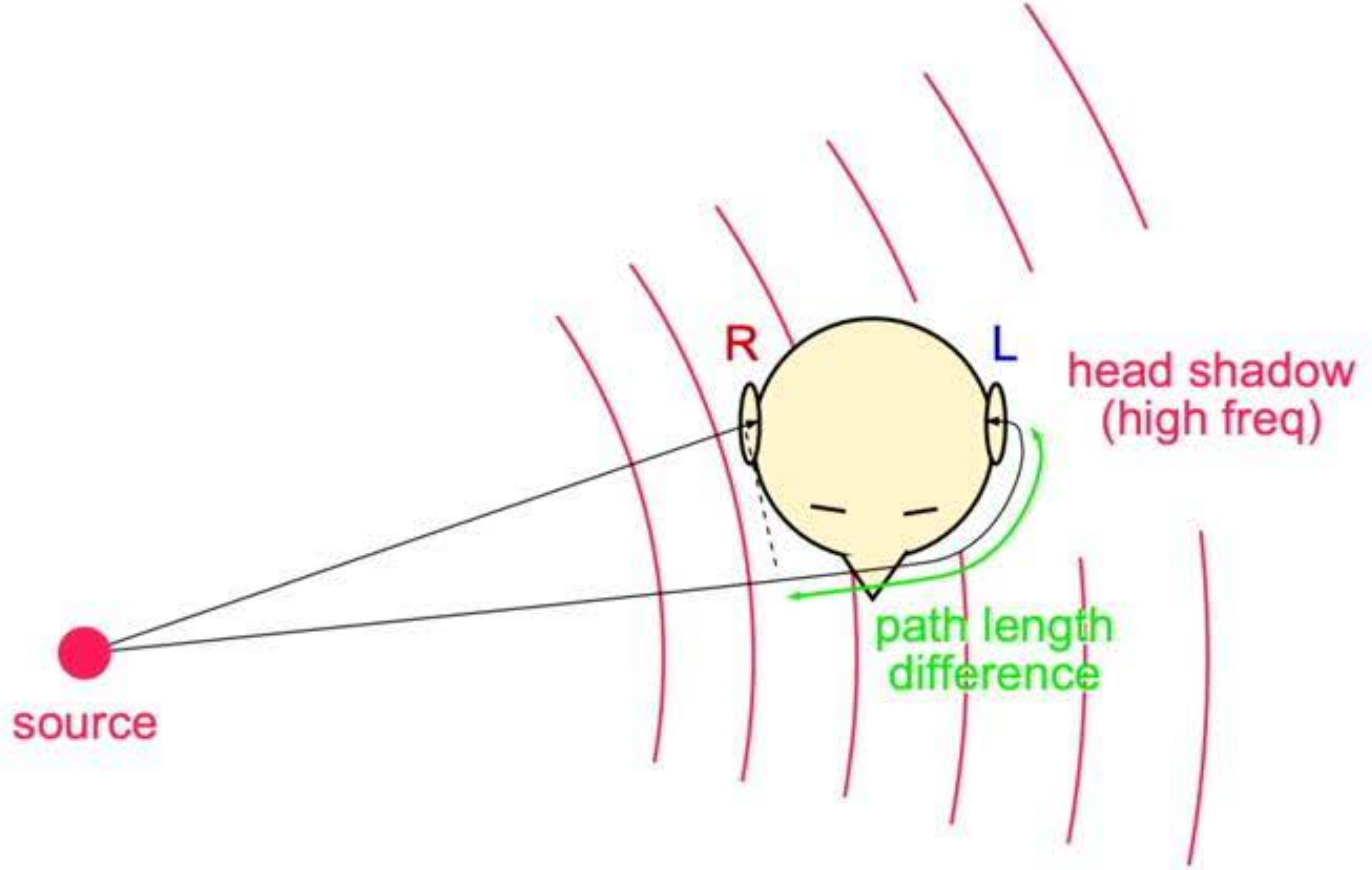
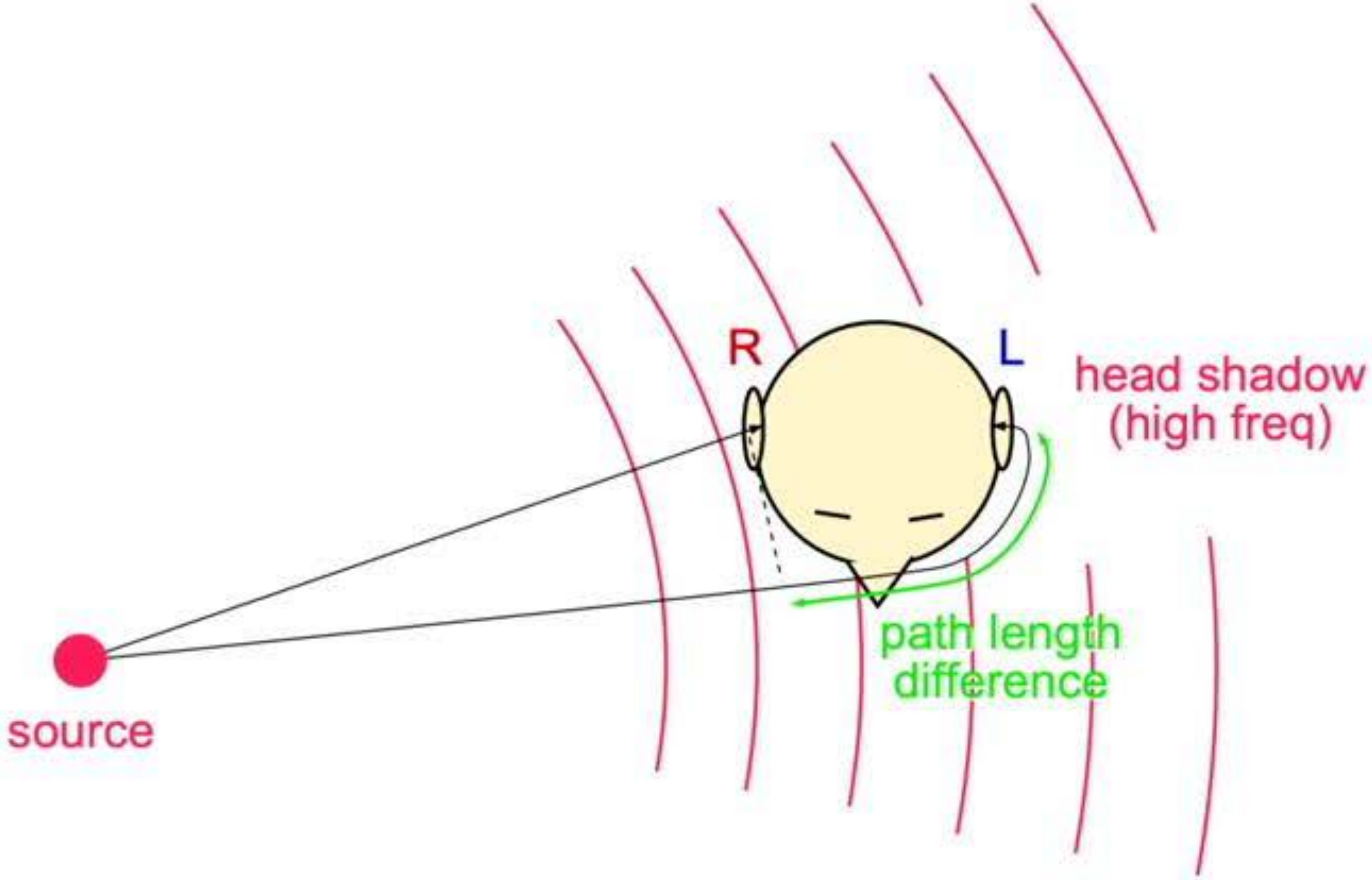


Image Credit: Michael Mandel

Spatial effects in audio



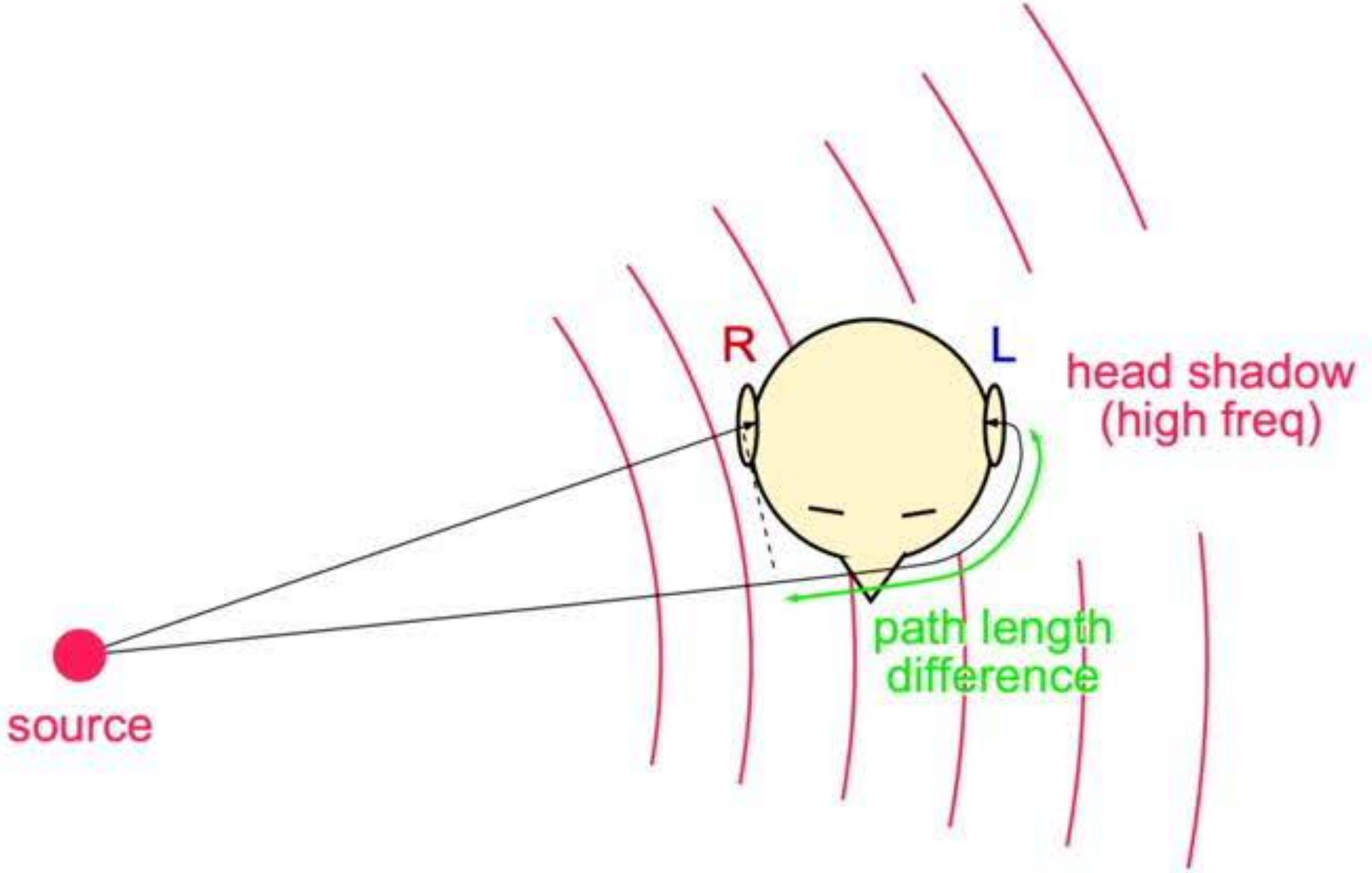
Cues for spatial hearing:

- Interaural time difference (ITD)
- Interaural level difference (ILD)
- Spectral detail (from pinna reflections)

Image Credit: Michael Mandel

Spatial effects in audio

Spatial effects absent in monaural audio



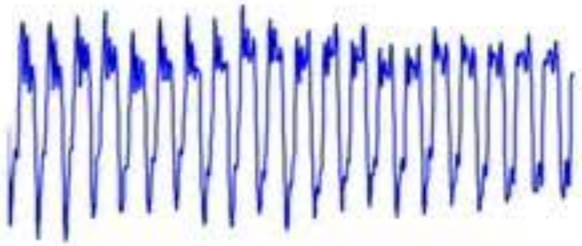
Cues for spatial hearing:
- Interaural time difference (ITD)
- Interaural level difference (ILD)
- Spectral detail (from pinna reflections)

Image Credit: Michael Mandel

Our idea: 2.5D visual sound

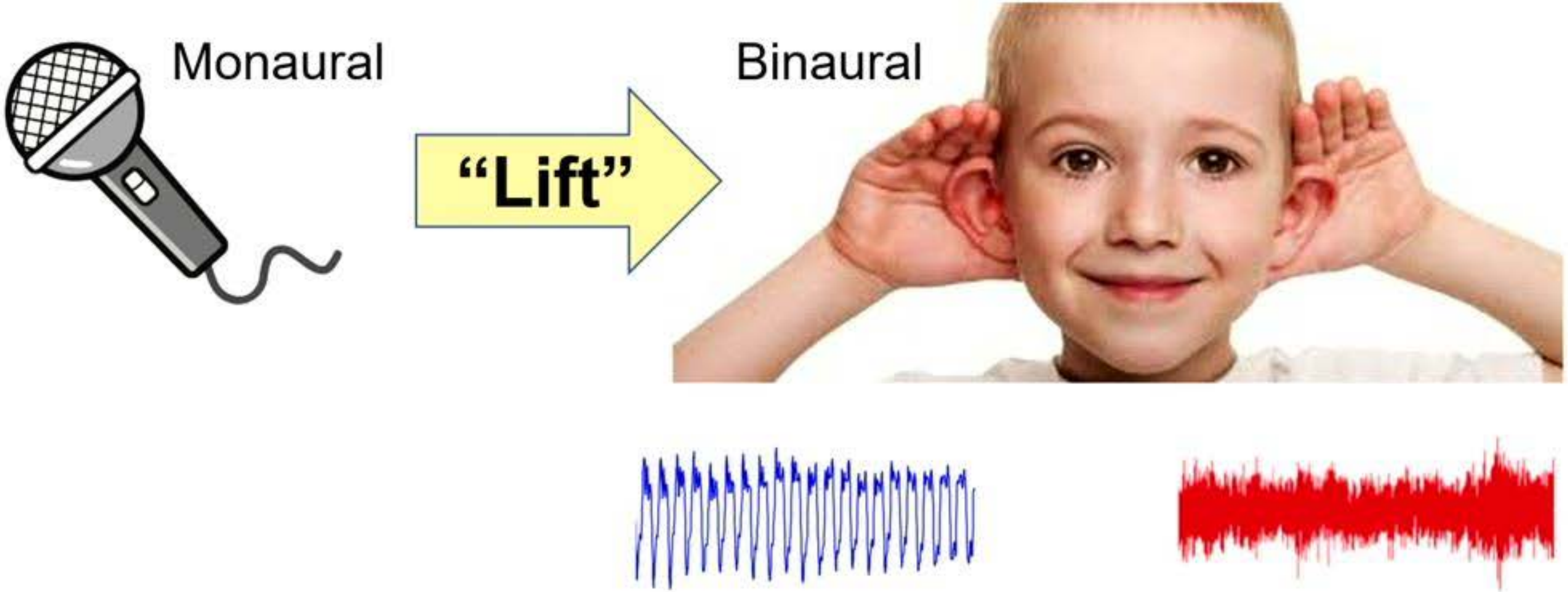
Monaural

Binaural



Gao & Grauman, CVPR 2019

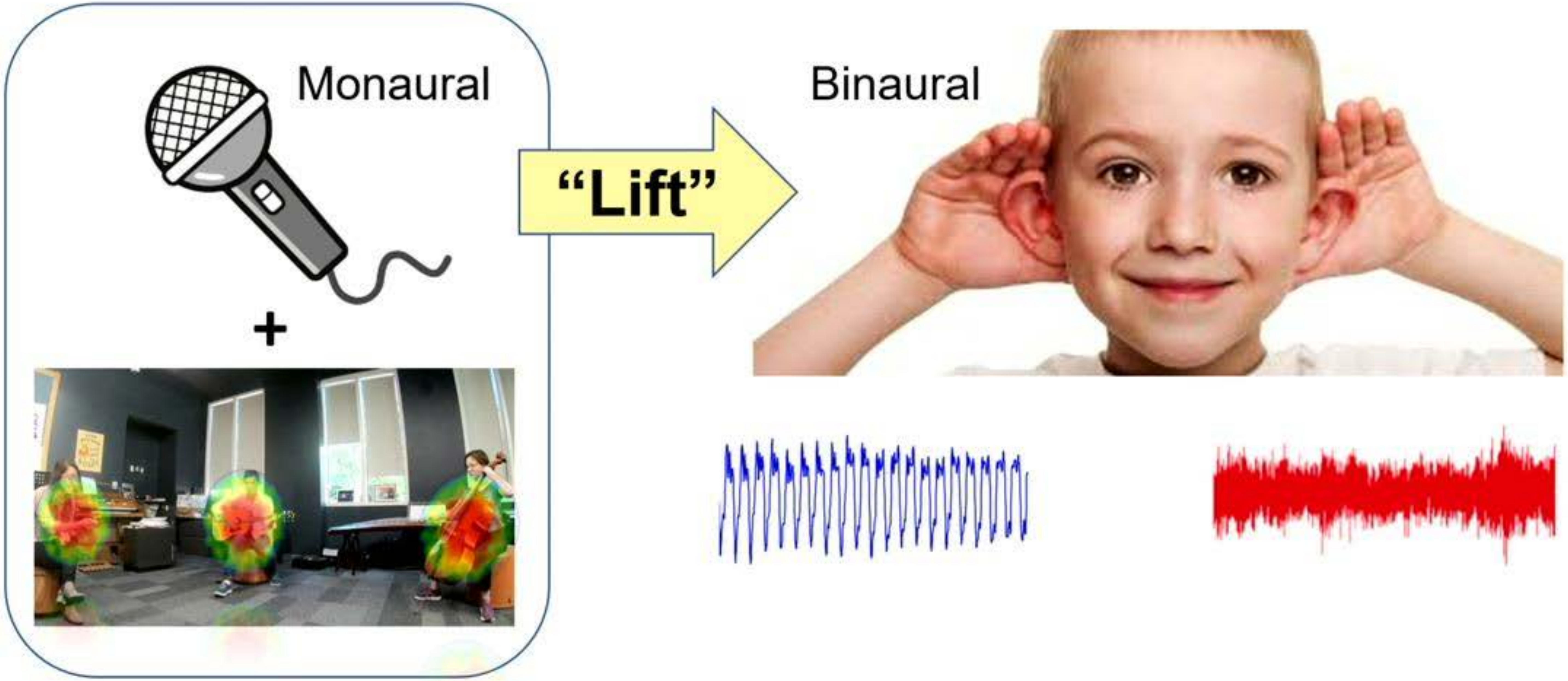
Our idea: 2.5D visual sound



Gao & Grauman, CVPR 2019

Our idea: 2.5D visual sound

"Lift" mono audio to spatial audio via visual cues

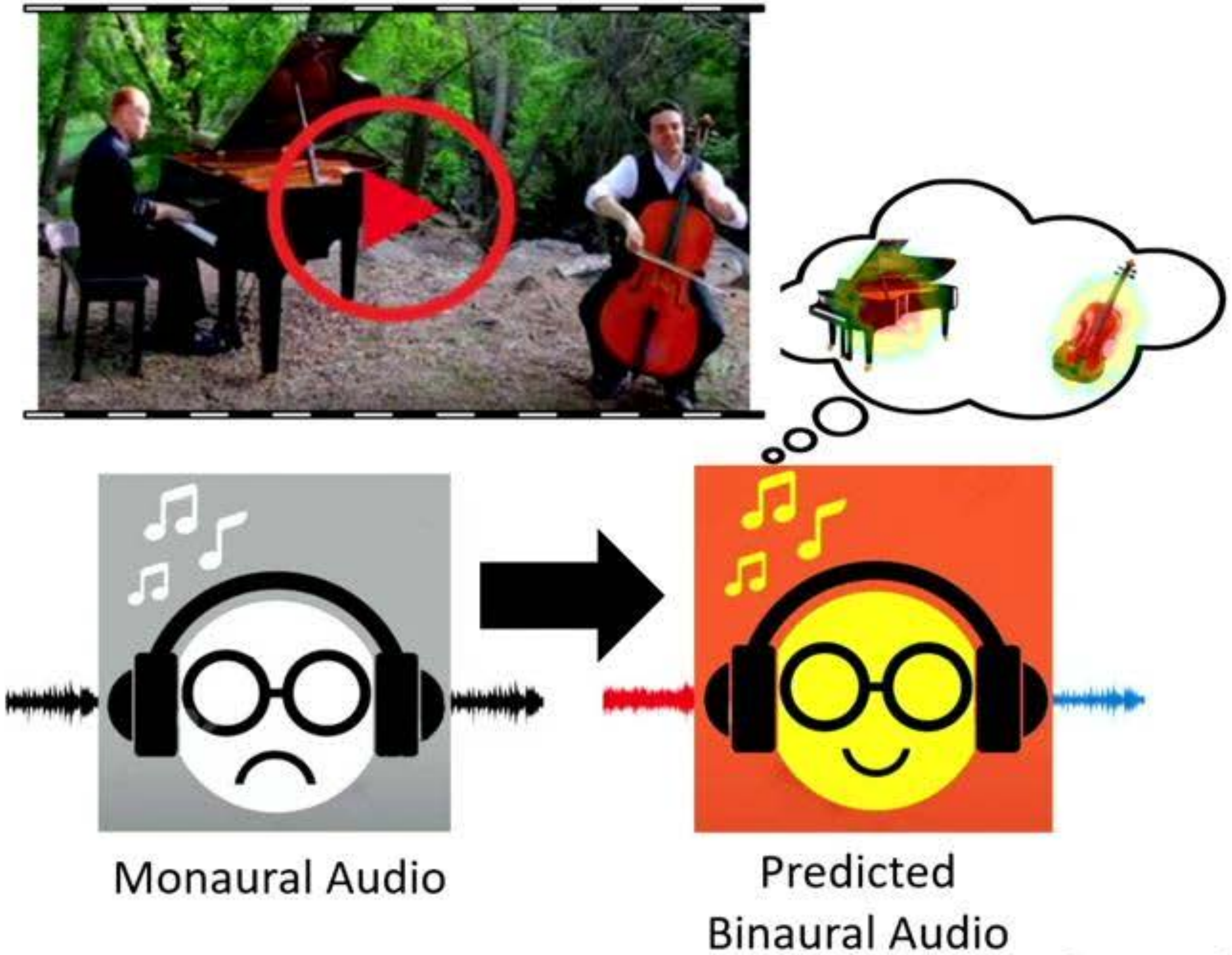


Gao & Grauman, CVPR 2019

Why infer binaural sound?

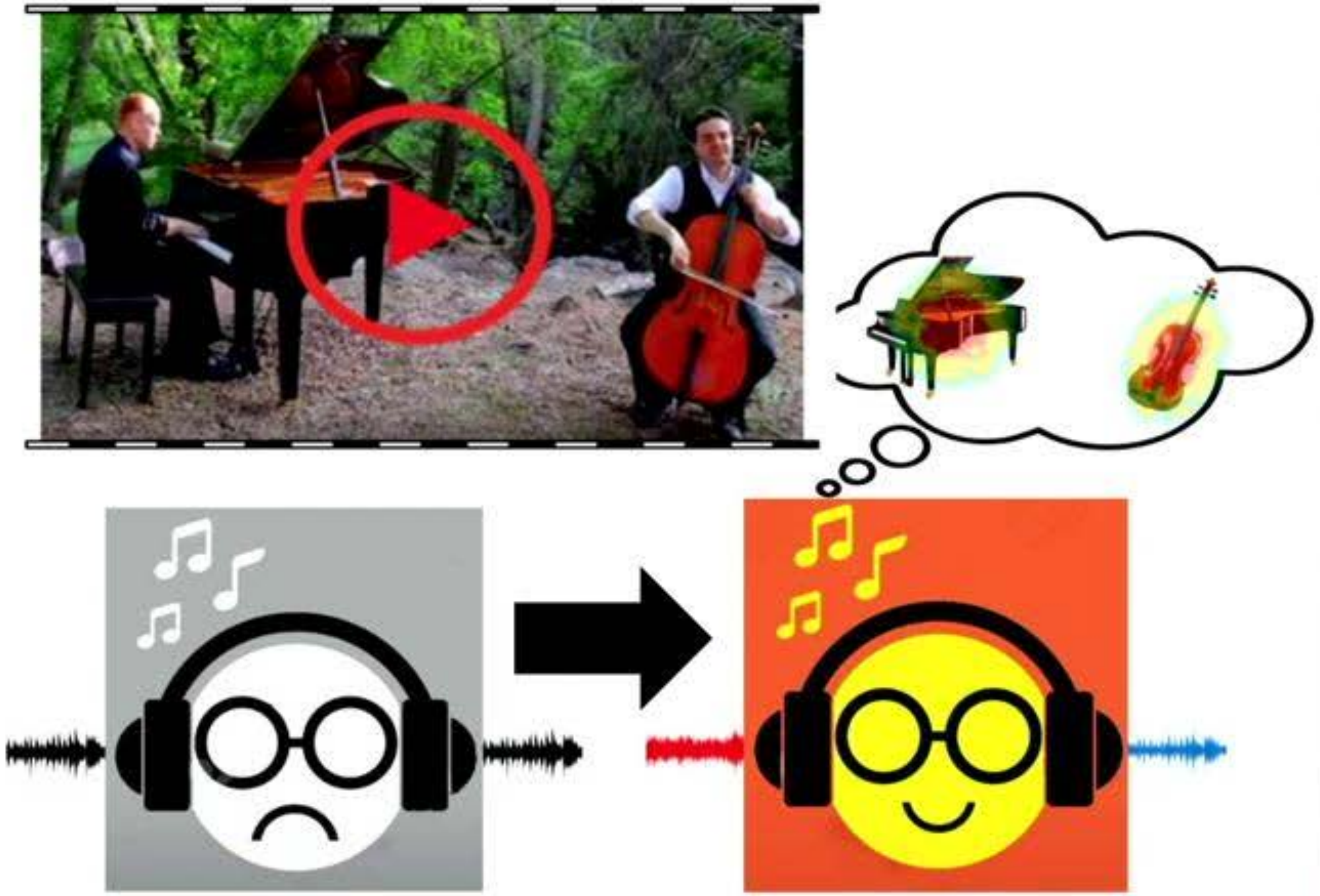
Why infer binaural sound?

Upgrade audio



Why infer binaural sound?

Upgrade audio



Monaural Audio

Predicted Binaural Audio

Improve separation

Our idea: 2.5D visual sound

“Lift” mono audio to spatial audio via visual cues

Gao & Grauman, CVPR 2019

Our idea: 2.5D visual sound

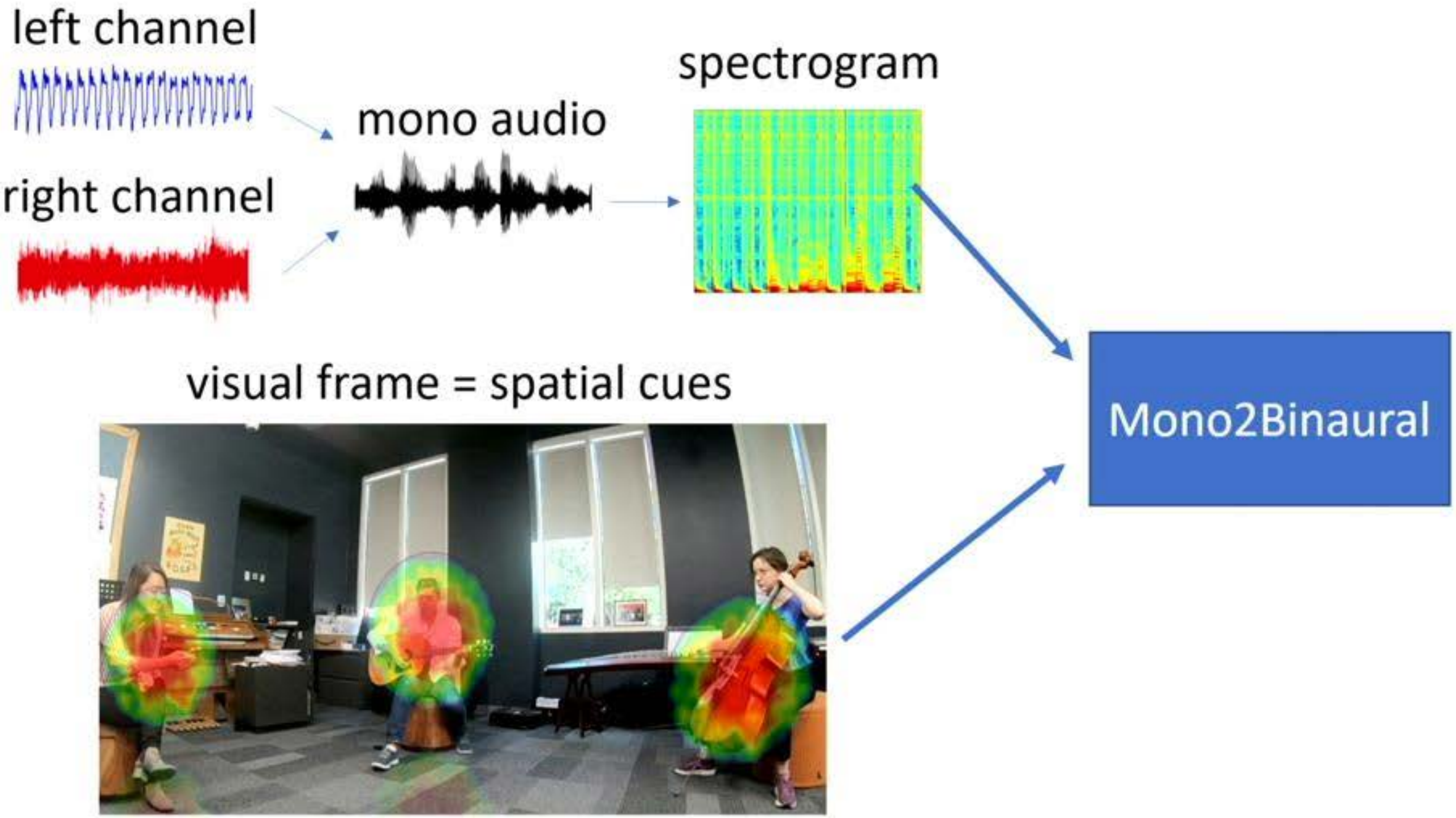
"Lift" mono audio to spatial audio via visual cues

visual frame = spatial cues

Gao & Grauman, CVPR 2019

Our idea: 2.5D visual sound

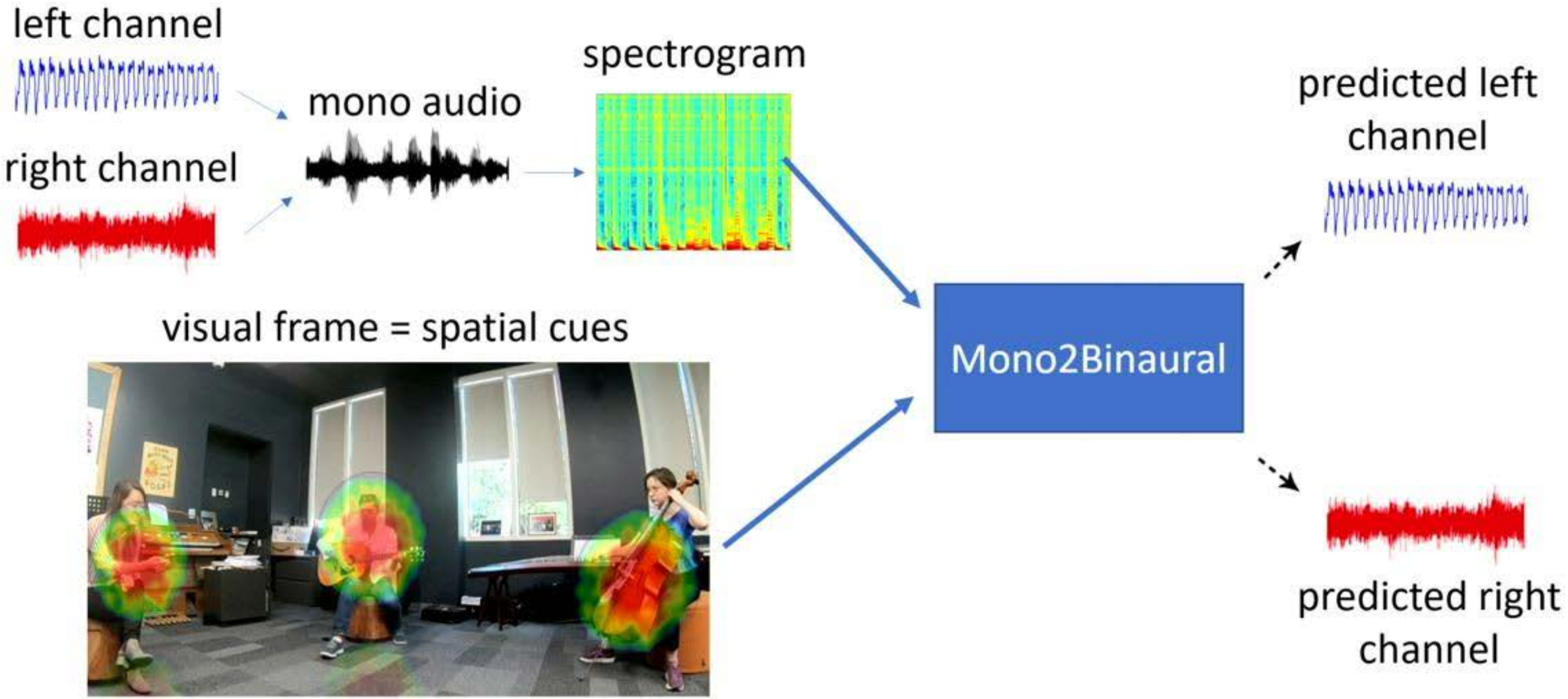
"Lift" mono audio to spatial audio via visual cues



Gao & Grauman, CVPR 2019

Our idea: 2.5D visual sound

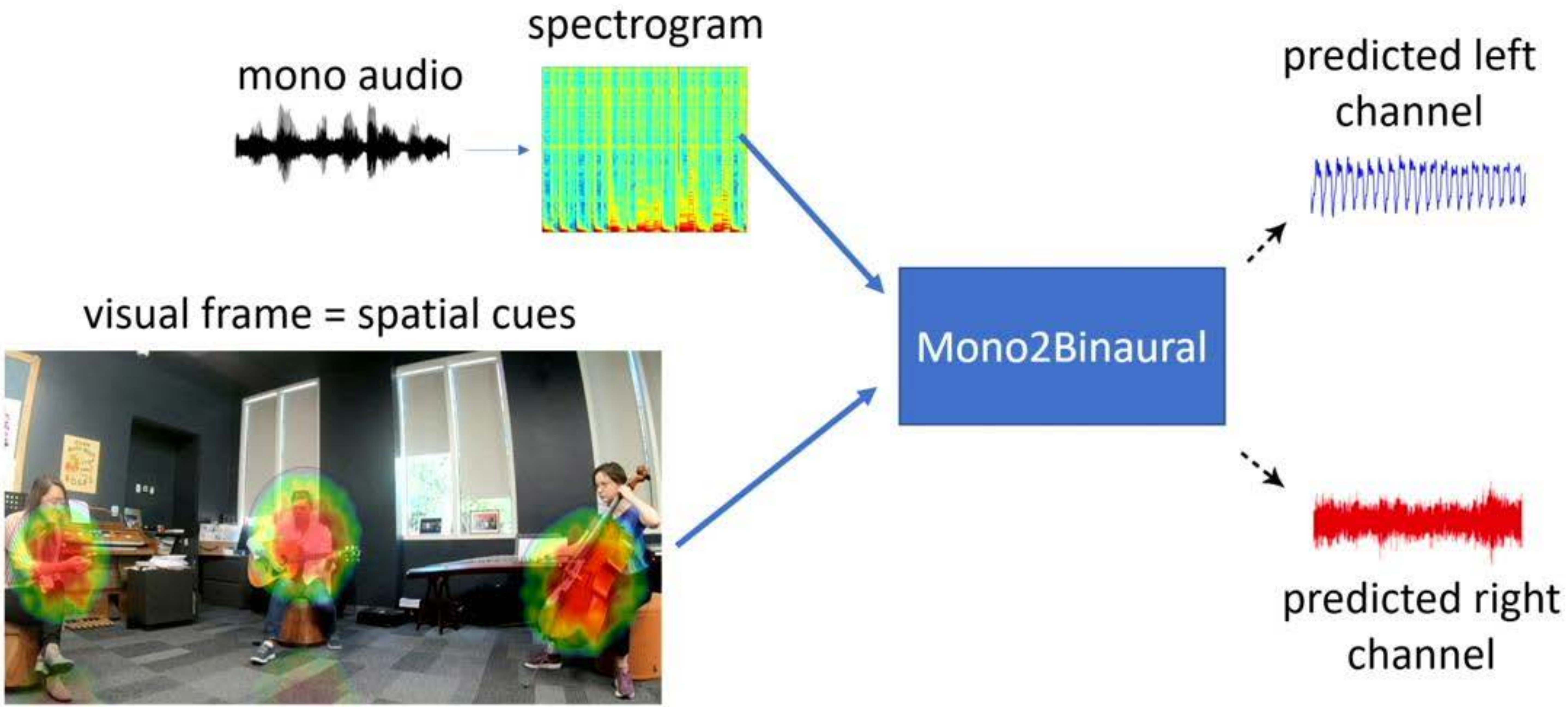
"Lift" mono audio to spatial audio via visual cues



Gao & Grauman, CVPR 2019

Our idea: 2.5D visual sound

"Lift" mono audio to spatial audio via visual cues



Gao & Grauman, CVPR 2019

FAIR-Play dataset

https://github.com/facebookresearch/FAIR-Play

Gao & Grauman, CVPR 2019

FAIR-Play dataset

https://github.com/facebookresearch/FAIR-Play

Data collection rig

Gao & Grauman, CVPR 2019

FAIR-Play dataset

https://github.com/facebookresearch/FAIR-Play

Binaural microphone rig linked to camera and monaural mic

Kristen Grauman, FAIR & UT Austin

Gao & Grauman, CVPR 2019

FAIR-Play dataset

https://github.com/facebookresearch/FAIR-Play

Binaural microphone rig linked to camera and monaural mic

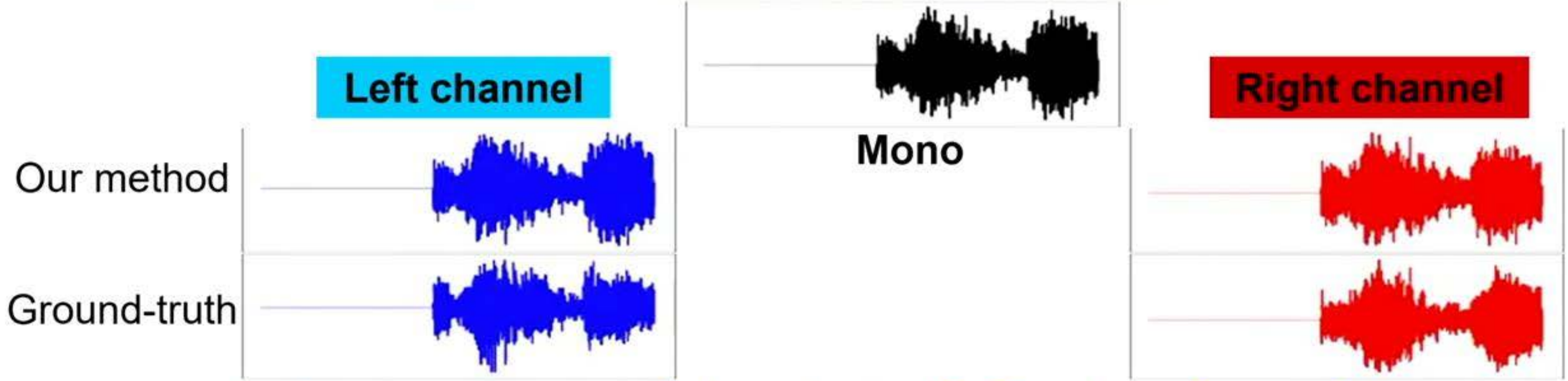
Capture ~5 hours video and binaural sound in a music room

Gao & Grauman, CVPR 2019

Results: 2.5D visual sound

Results: 2.5D visual sound

Input video



vision.cs.utexas.edu/projects/2.5D_visual_sound/

Ask listener: where is the drum/piano?

Listener does not see any video

Datasets

FAIR-Play

- 10 musical instruments, e.g., cello, guitar, harp, trumpet, etc.
- ~5 hours of performances

YouTube Datasets

[Morgado *et al*. NeurIPS 2018]

- **Streets, random YouTube**
- ~1000 360° video clips
- Converted to binaural audio using decoder

Gao & Grauman, CVPR 2019

Results: Binaural audio prediction

	FAIR-Play		REC-STREET		YT-CLEAN		YT-MUSIC	
	STFT	ENV	STFT	ENV	STFT	ENV	STFT	ENV
Ambisonics	-	-	0.744	0.126	1.435	0.155	1.885	0.183
Audio-Only	0.966	0.141	0.590	0.114	1.065	0.131	1.553	0.167
Flipped-Visual	1.145	0.149	0.658	0.123	1.095	0.132	1.590	0.165
Mono-Mono	1.155	0.153	0.774	0.136	1.369	0.153	1.853	0.184
MONO2BINAURAL (Ours)	**0.836**	**0.132**	**0.565**	**0.109**	**1.027**	**0.130**	**1.451**	**0.156**

Ambisonics: *Morgado et al. NeurIPS 2018*

Gao & Grauman, CVPR 2019

Results: Binaural audio prediction

	FAIR-Play		REC-STREET		YT-CLEAN		YT-MUSIC	
	STFT	ENV	STFT	ENV	STFT	ENV	STFT	ENV
Ambisonics	-	-	0.744	0.126	1.435	0.155	1.885	0.183
Audio-Only	0.966	0.141	0.590	0.114	1.065	0.131	1.553	0.167
Flipped-Visual	1.145	0.149	0.658	0.123	1.095	0.132	1.590	0.165
Mono-Mono	1.155	0.153	0.774	0.136	1.369	0.153	1.853	0.184
MONO2BINAURAL (Ours)	**0.836**	**0.132**	**0.565**	**0.109**	**1.027**	**0.130**	**1.451**	**0.156**

Best binaural prediction results on all four datasets

Ambisonics: *Morgado et al. NeurIPS 2018*

Gao & Grauman, CVPR 2019

Results: Audio-visual source separation

original video
(before separation)

visual predictions:
dog & violin

2.5d visual sound → better audio separation

[Gao, Feris, & Grauman, ECCV 2018]

Results: Binaural audio prediction

	FAIR-Play		REC-STREET		YT-CLEAN		YT-MUSIC	
	STFT	ENV	STFT	ENV	STFT	ENV	STFT	ENV
Ambisonics	-	-	0.744	0.126	1.435	0.155	1.885	0.183
Audio-Only	0.966	0.141	0.590	0.114	1.065	0.131	1.553	0.167
Flipped-Visual	1.145	0.149	0.658	0.123	1.095	0.132	1.590	0.165
Mono-Mono	1.155	0.153	0.774	0.136	1.369	0.153	1.853	0.184
MONO2BINAURAL (Ours)	**0.836**	**0.132**	**0.565**	**0.109**	**1.027**	**0.130**	**1.451**	**0.156**

Best binaural prediction results on all four datasets

Ambisonics: *Morgado et al. NeurIPS 2018*

Gao & Grauman, CVPR 2019

Datasets

FAIR-Play

- 10 musical instruments, e.g., cello, guitar, harp, trumpet, etc.
- ~5 hours of performances

YouTube Datasets

[Morgado *et al*. NeurIPS 2018]

- **Streets, random YouTube**
- ~1000 360° video clips
- Converted to binaural audio using decoder

Gao & Grauman, CVPR 2019

Results: Audio-visual source separation

original video
(before separation)

visual predictions:
dog & violin

2.5d visual sound → better audio separation

[Gao, Feris, & Grauman, ECCV 2018]

This talk

Multi-sensory

Audio-visual learning

Motion

Navigation policies

Interaction

Affordance learning

Towards embodied perception

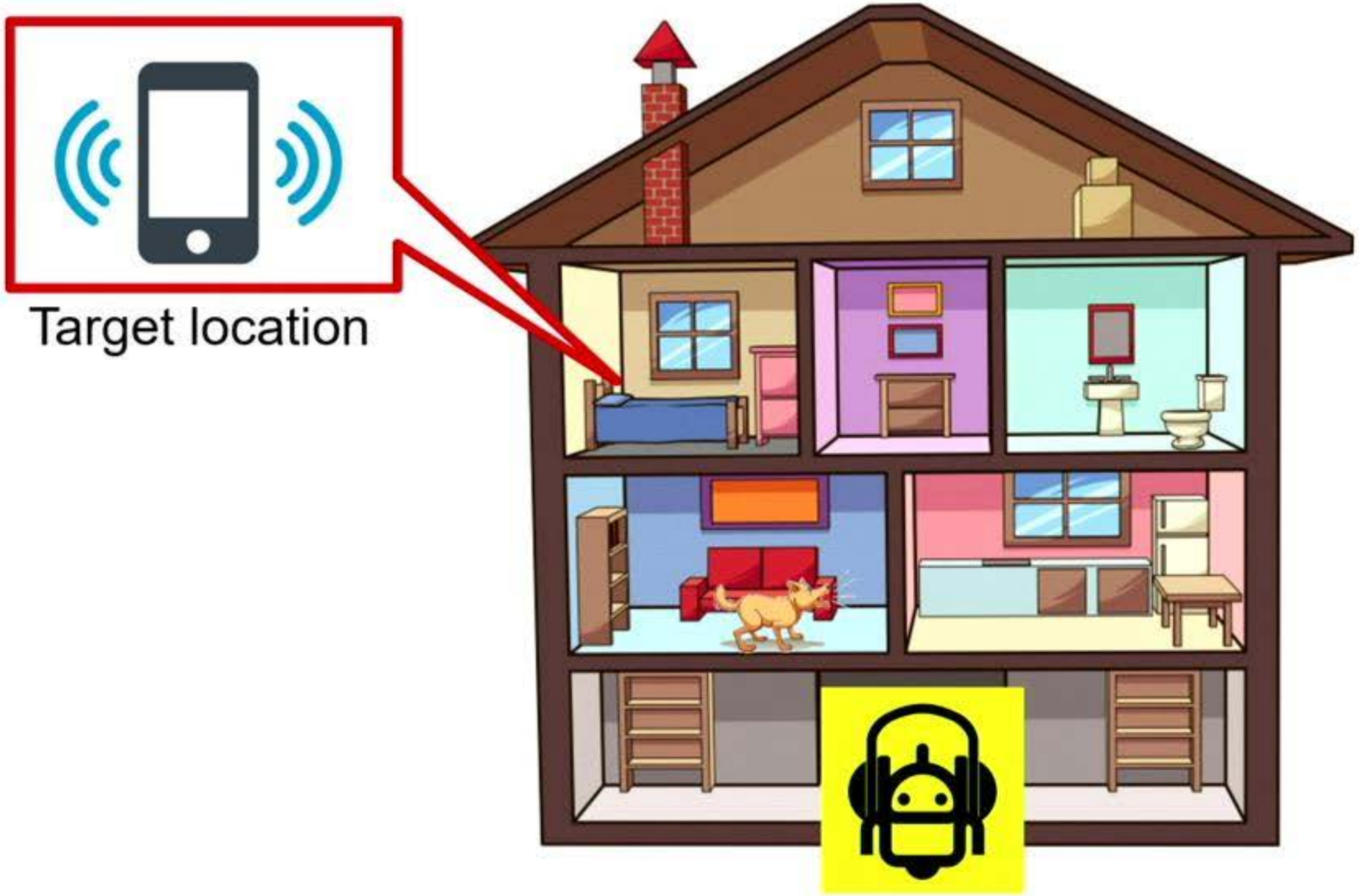
Visual navigation in novel unmapped environments

Where is the telephone?

Our idea: Audio-visual navigation

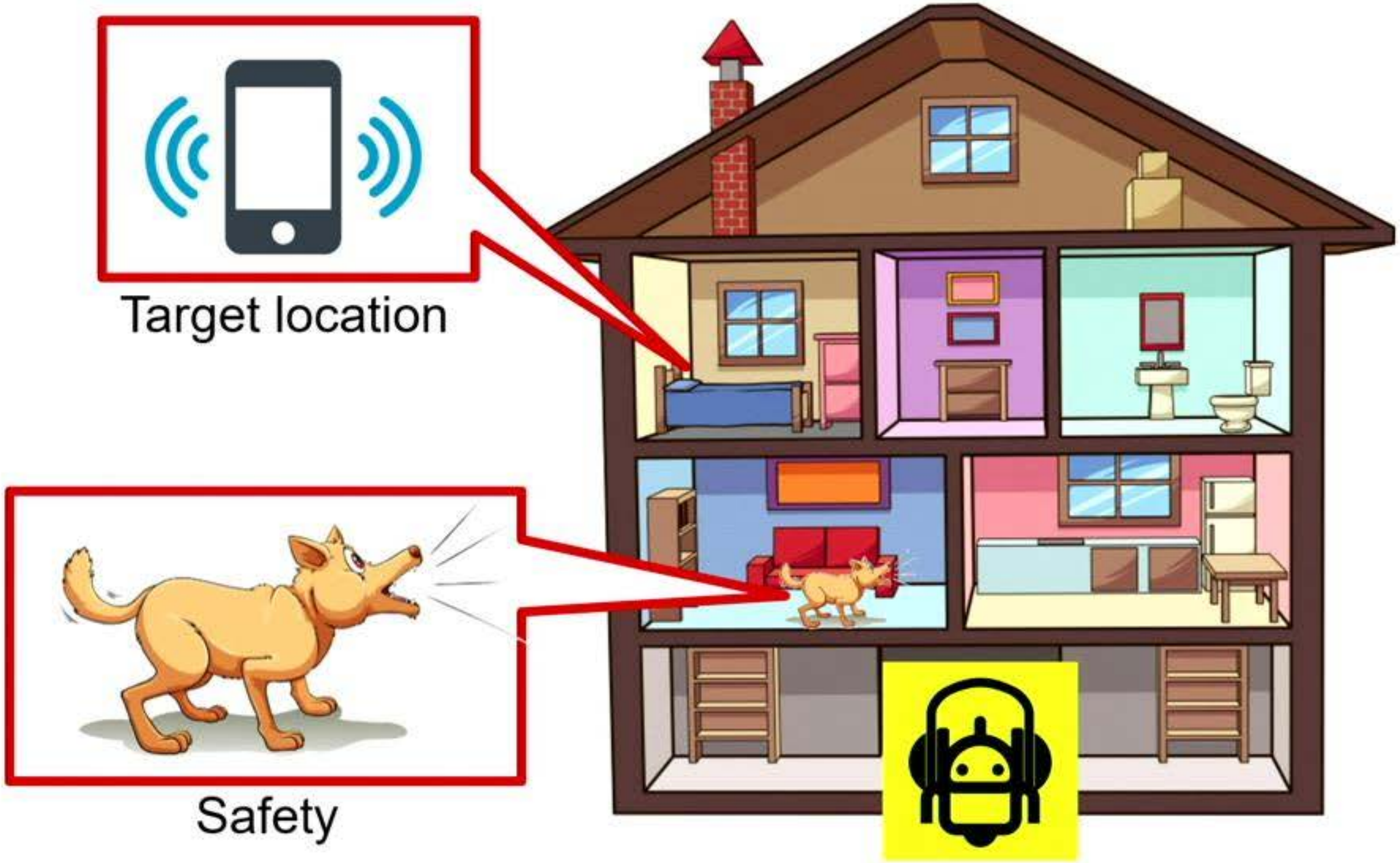
Our idea: Audio-visual navigation

Sound informs navigating agent about…



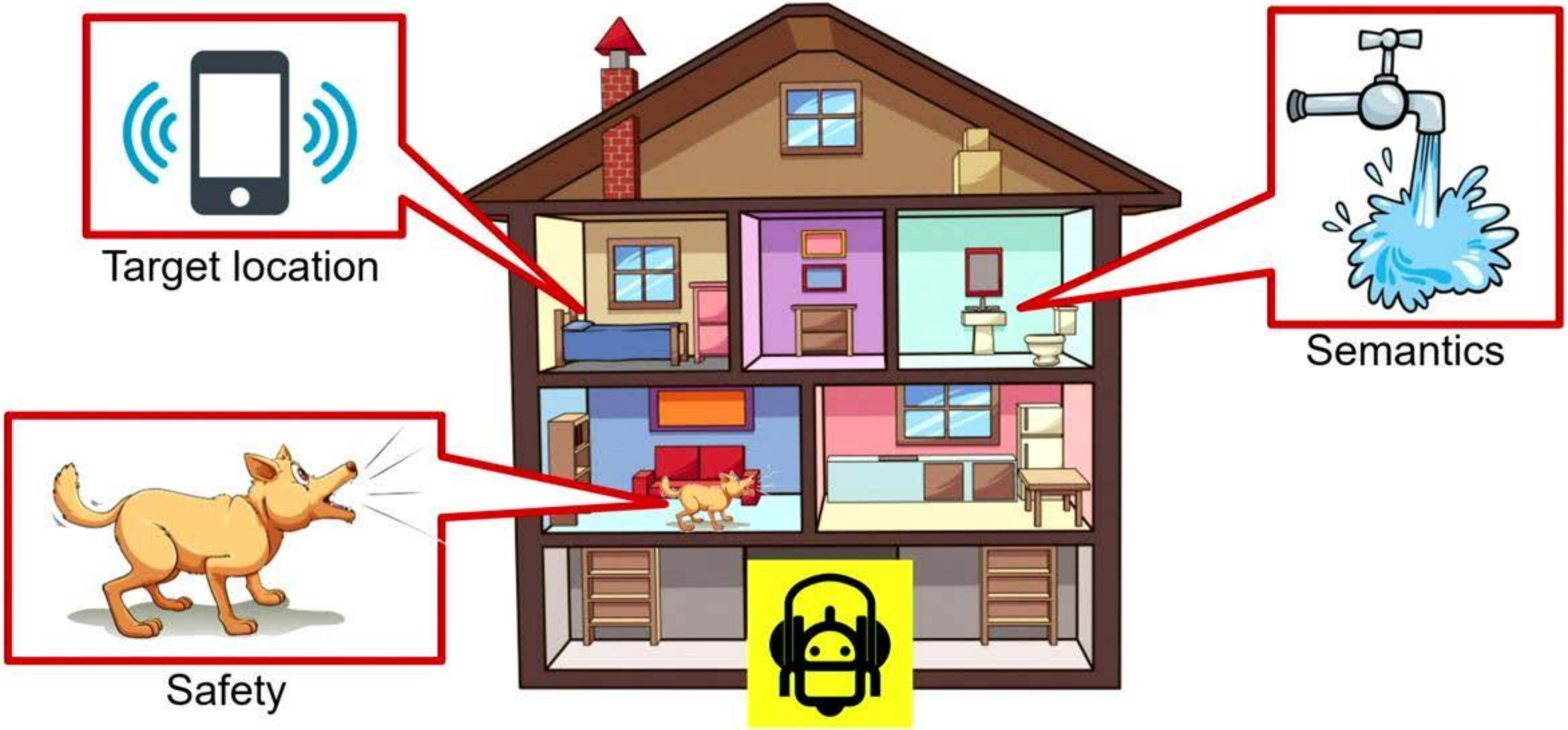
Our idea: Audio-visual navigation

Sound informs navigating agent about…



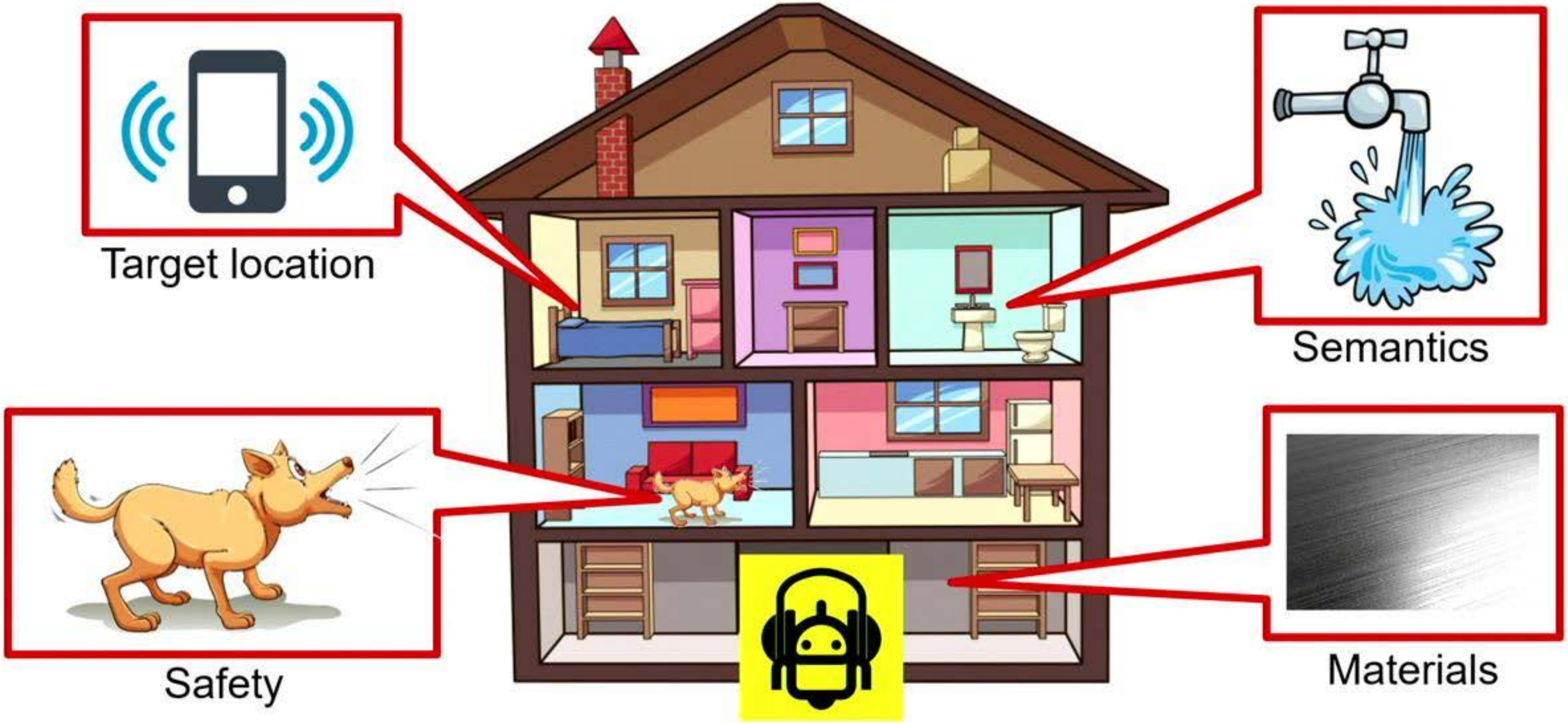
Our idea: Audio-visual navigation

Sound informs navigating agent about…

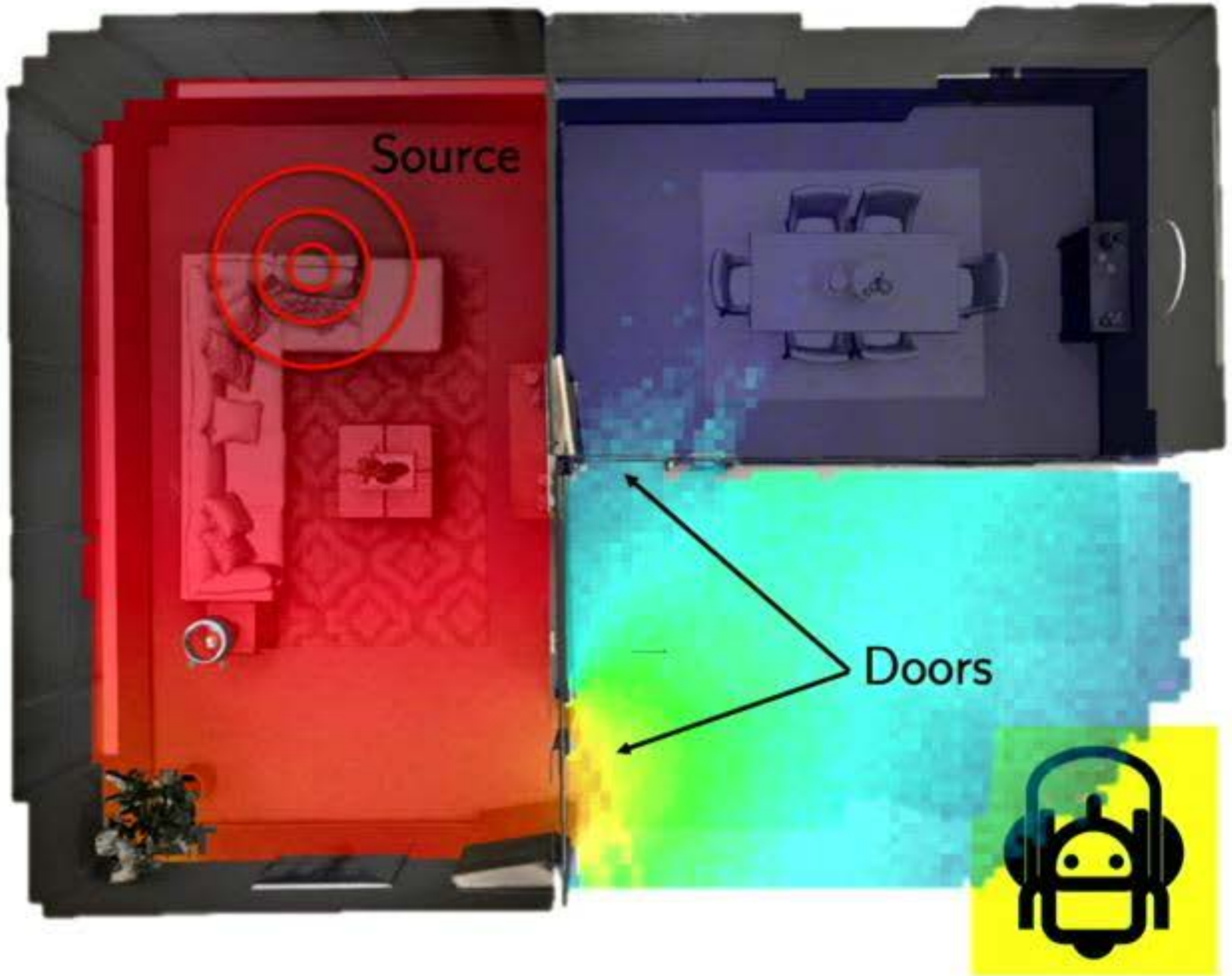


Our idea: Audio-visual navigation

Sound informs navigating agent about…



Our idea: Audio-visual navigation



[Chen et al., Audio-visual embodied navigation, arXiv 2019]

Audio simulation platform

We introduce audio simulation platform

[Chen et al., Audio-visual embodied navigation, arXiv 2019]

Audio simulation platform

We introduce audio simulation platform

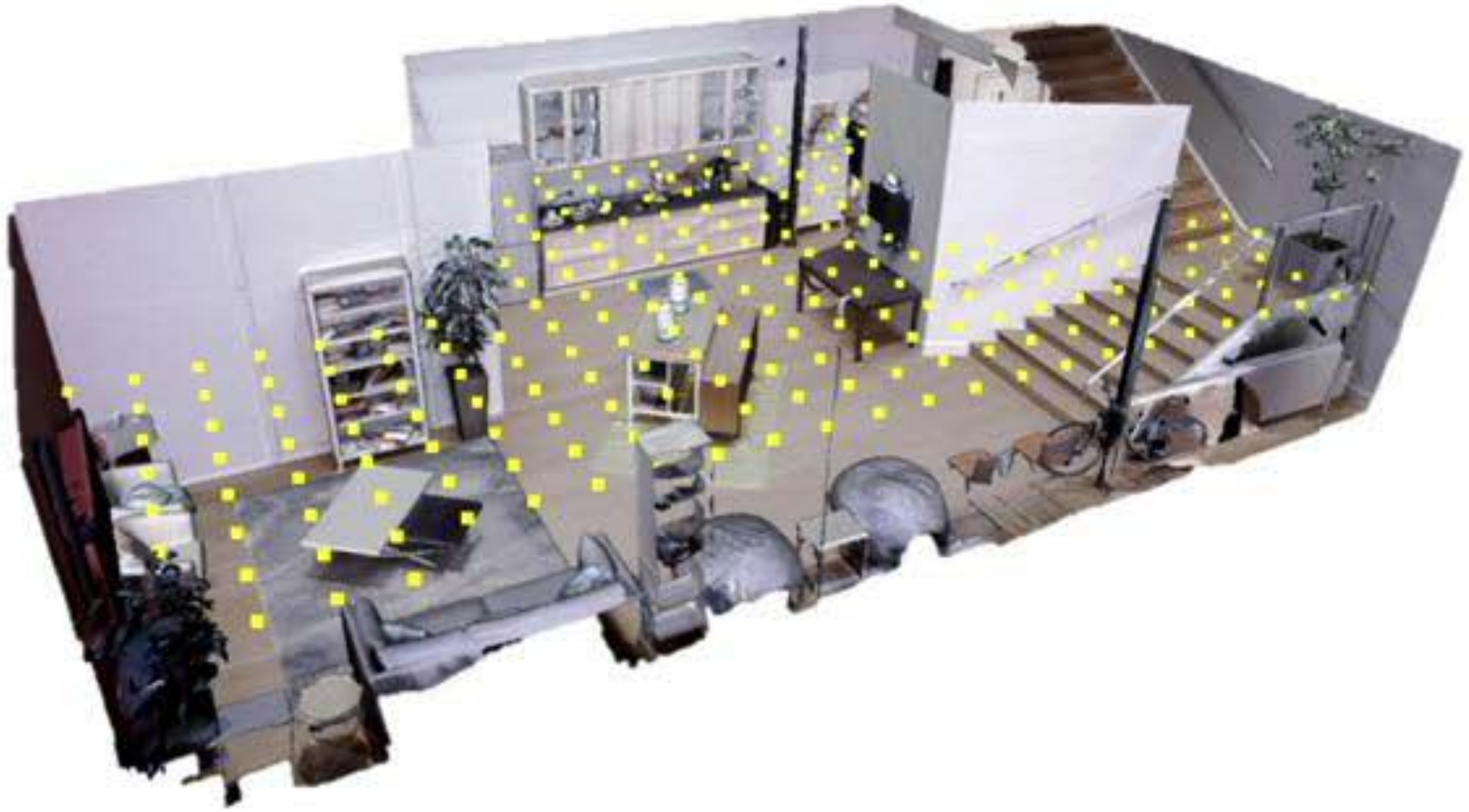
- Visually realistic 3D environments (Facebook Replica scenes)

[Chen et al., Audio-visual embodied navigation, arXiv 2019]

Audio simulation platform

We introduce audio simulation platform

- Visually realistic 3D environments (Facebook Replica scenes)
- Room impulse response (RIR) for all source *x* receiver locs

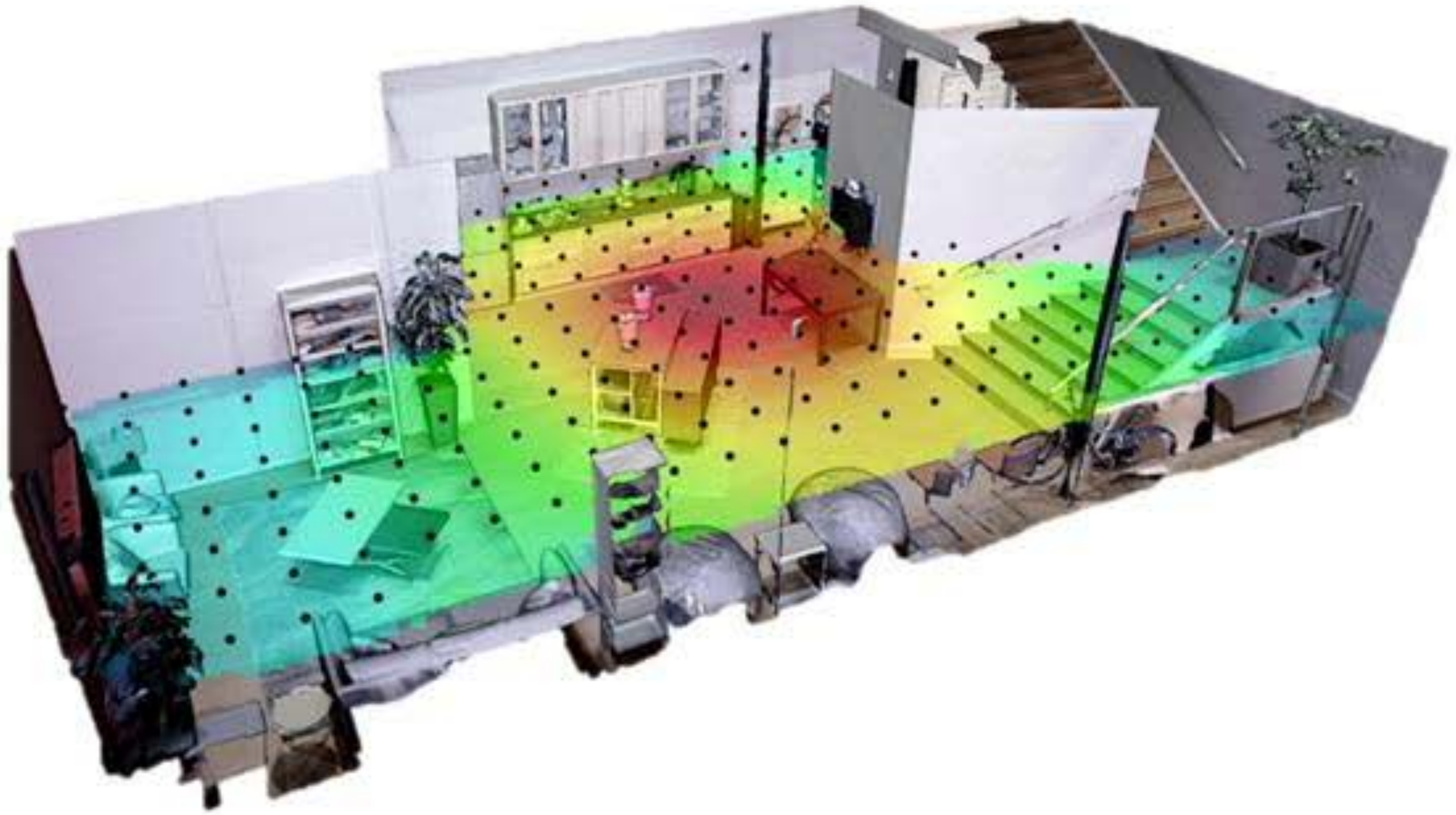


[Chen et al., Audio-visual embodied navigation, arXiv 2019]

Audio simulation platform

We introduce audio simulation platform

- Visually realistic 3D environments (Facebook Replica scenes)
- Room impulse response (RIR) for all source *x* receiver locs
- Convolve with arbitrary waveform to render binaural sound heard by agent

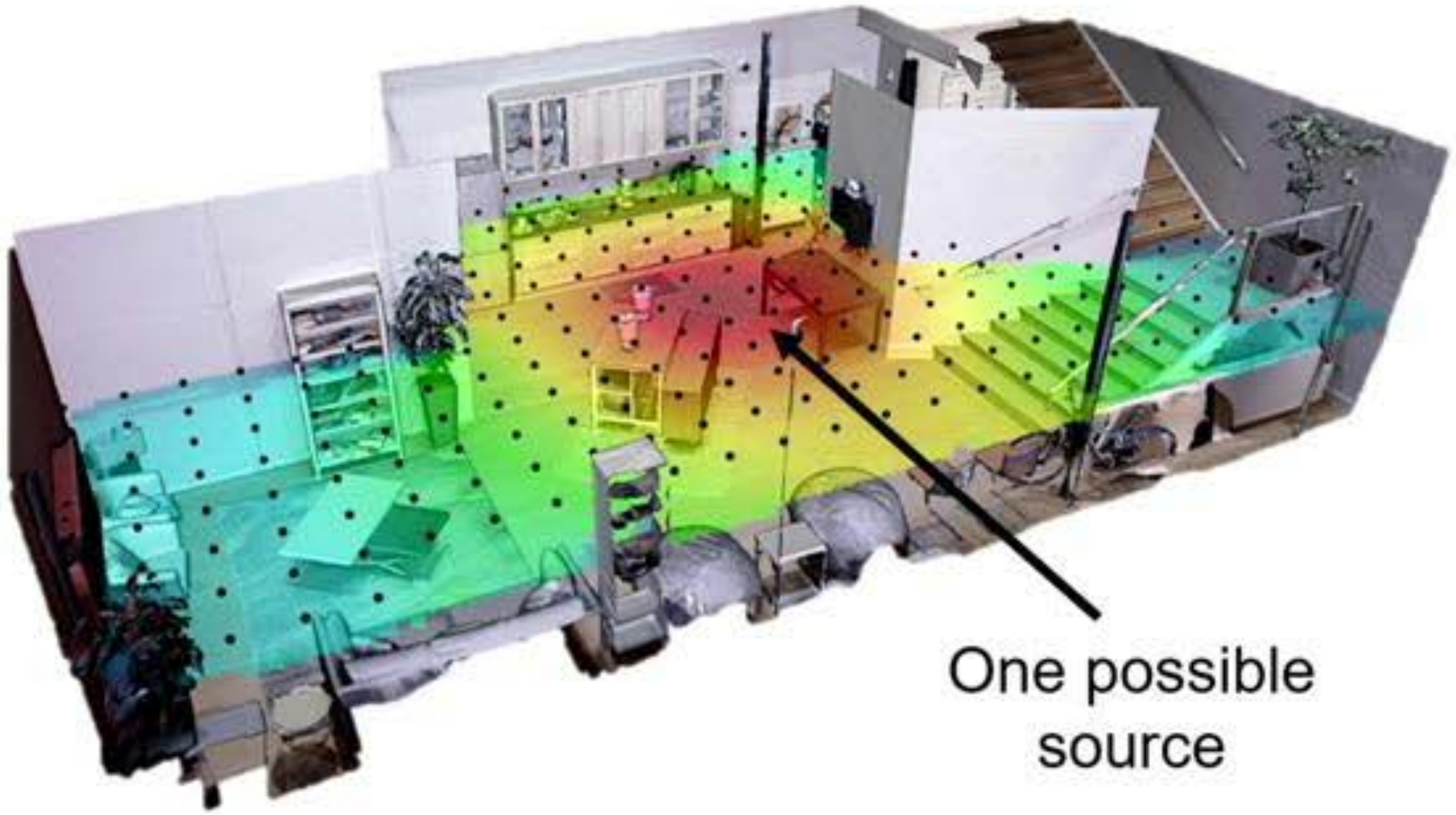


[Chen et al., Audio-visual embodied navigation, arXiv 2019]

Audio simulation platform

We introduce audio simulation platform

- Visually realistic 3D environments (Facebook Replica scenes)
- Room impulse response (RIR) for all source *x* receiver locs
- Convolve with arbitrary waveform to render binaural sound heard by agent



[Chen et al., Audio-visual embodied navigation, arXiv 2019]

Audio-visual navigation task

Navigate to an audio-emitting goal (e.g., phone ringing)

Agent

Seen/Unseen area

Occupied area

[Chen et al., Audio-visual embodied navigation, arXiv 2019]

Audio-visual navigation model

Reinforcement learning for agent's motion policy from multi-modal inputs

[Chen et al., Audio-visual embodied navigation, arXiv 2019]

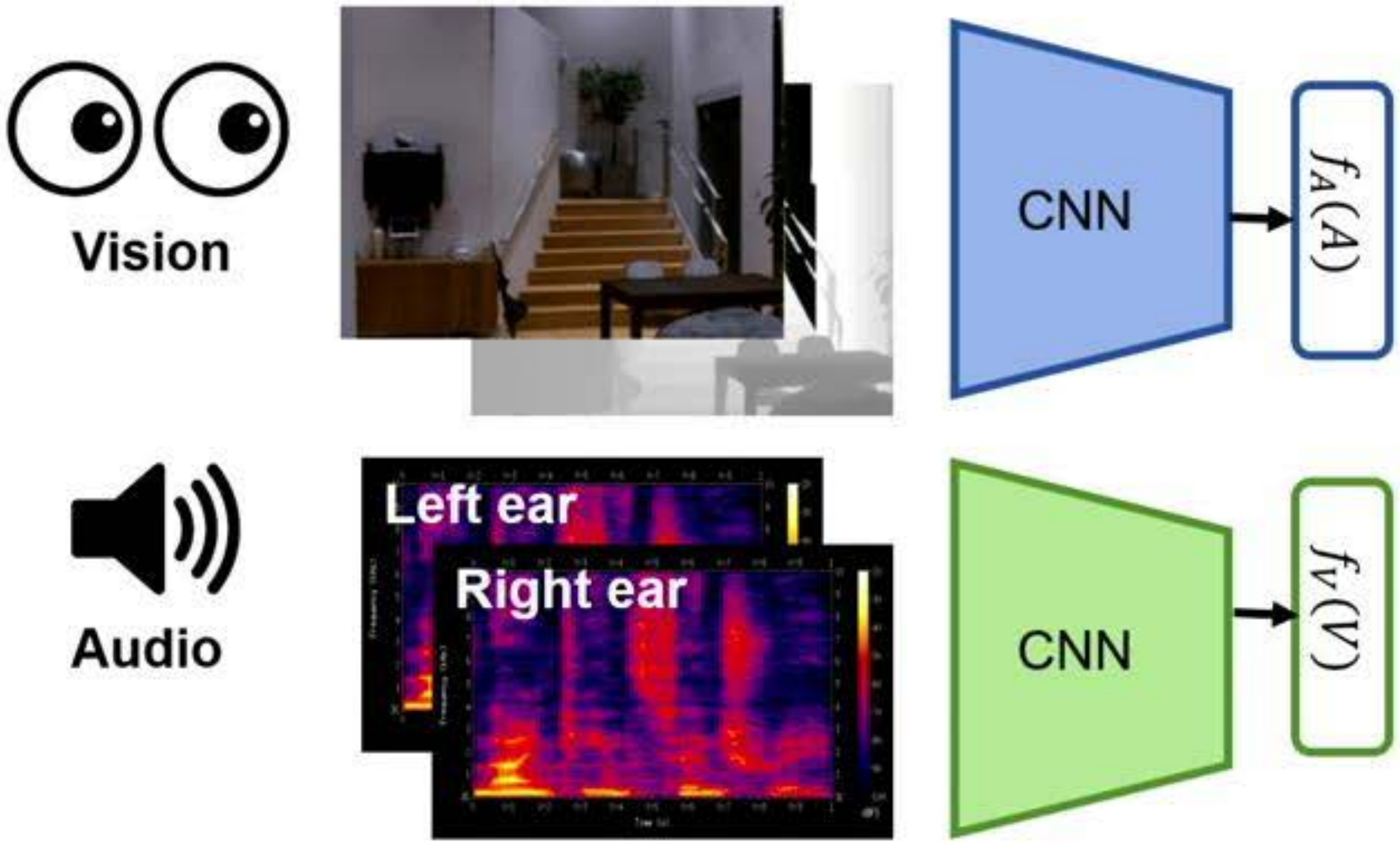
Audio-visual navigation model

Reinforcement learning for agent's motion policy from multi-modal inputs

[Chen et al., Audio-visual embodied navigation, arXiv 2019]

Audio-visual navigation model

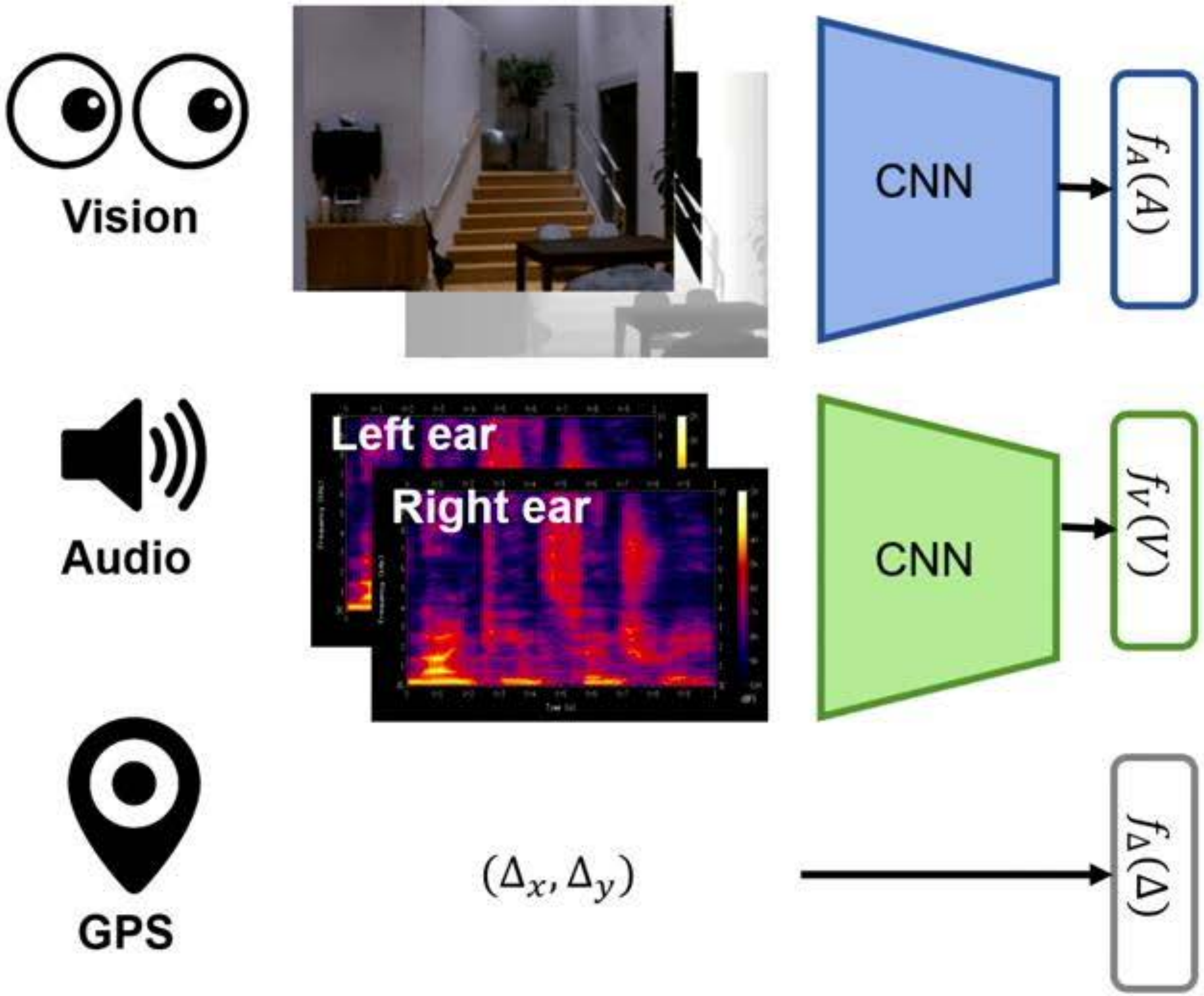
Reinforcement learning for agent's motion policy from multi-modal inputs



[Chen et al., Audio-visual embodied navigation, arXiv 2019]

Audio-visual navigation model

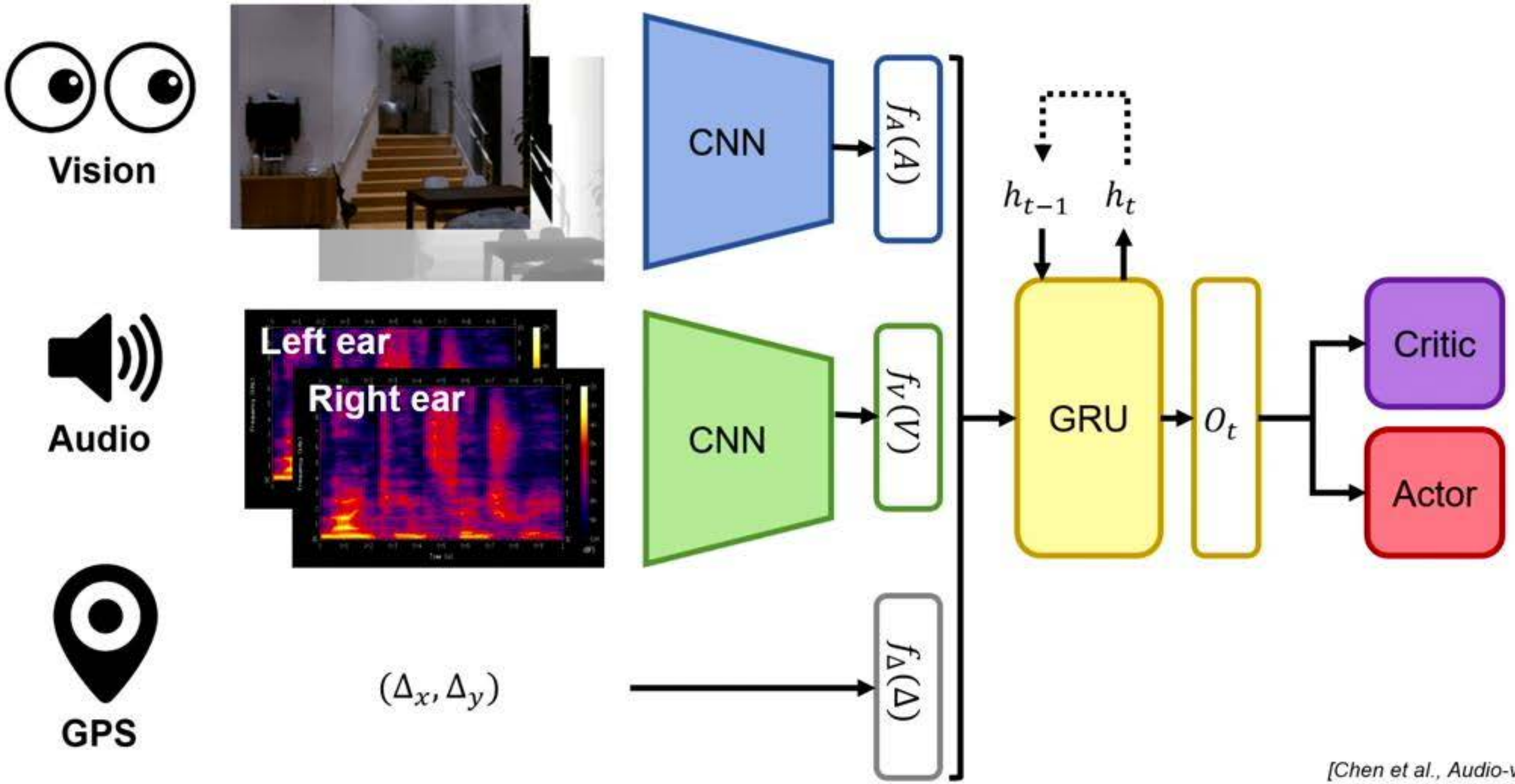
Reinforcement learning for agent's motion policy from multi-modal inputs



[Chen et al., Audio-visual embodied navigation, arXiv 2019]

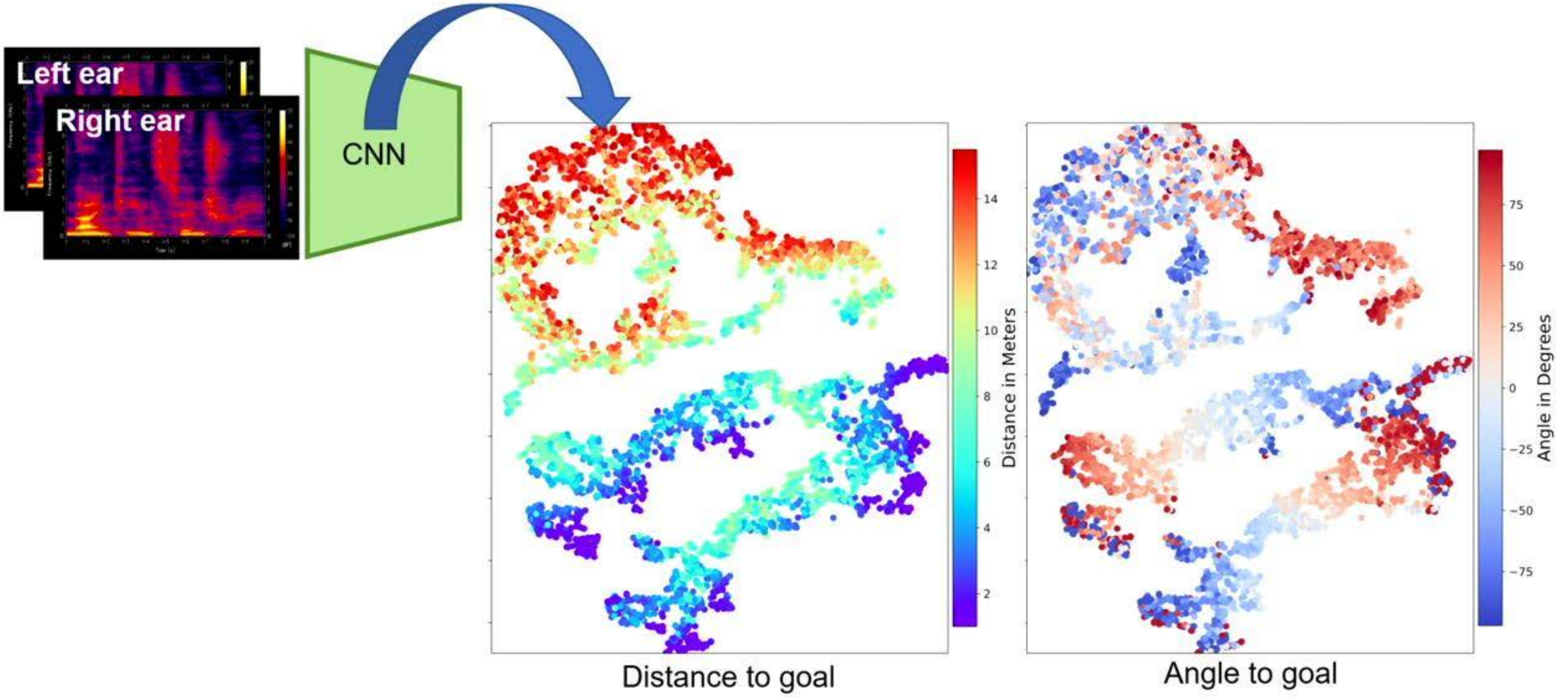
Audio-visual navigation model

Reinforcement learning for agent's motion policy from multi-modal inputs



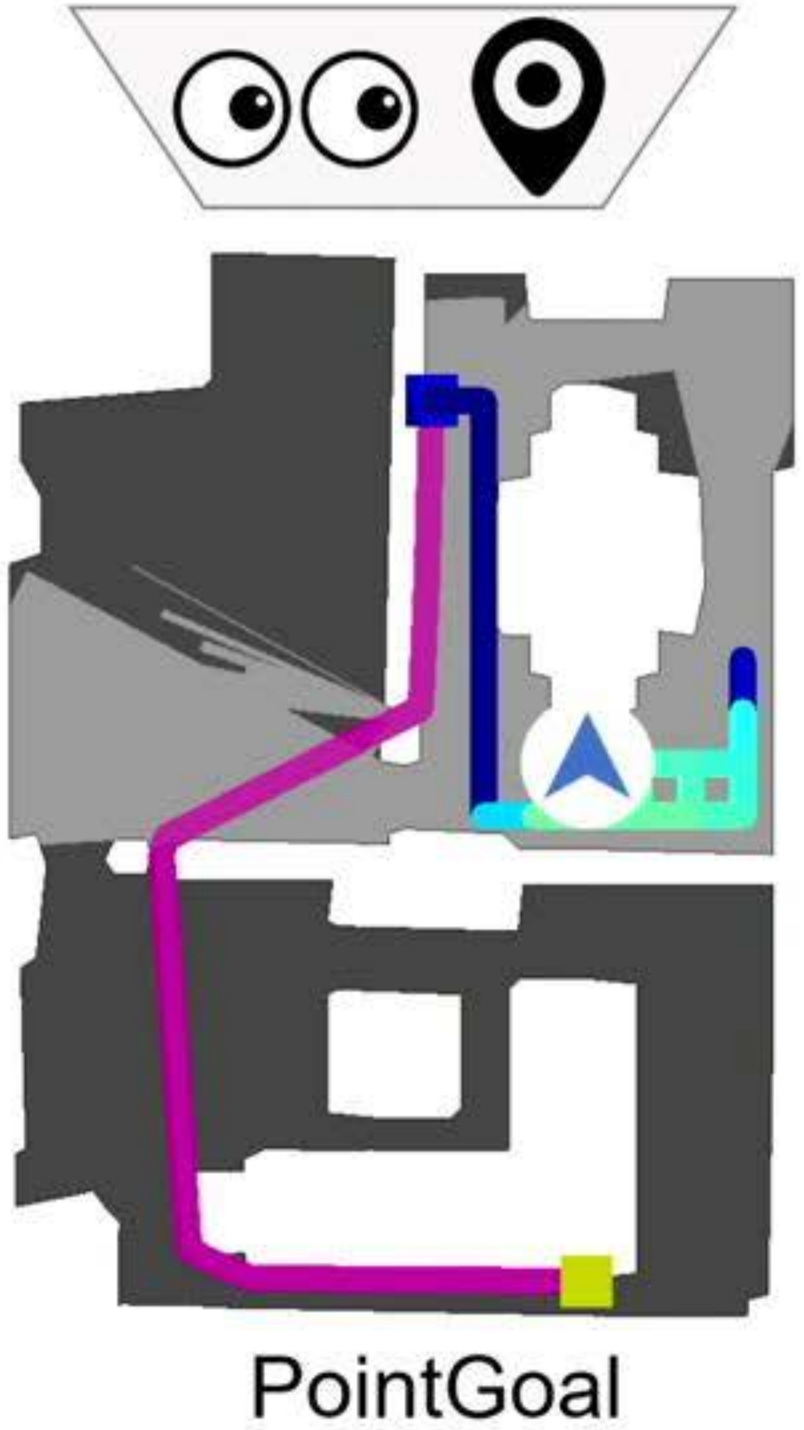
[Chen et al., Audio-visual embodied navigation, arXiv 2019]

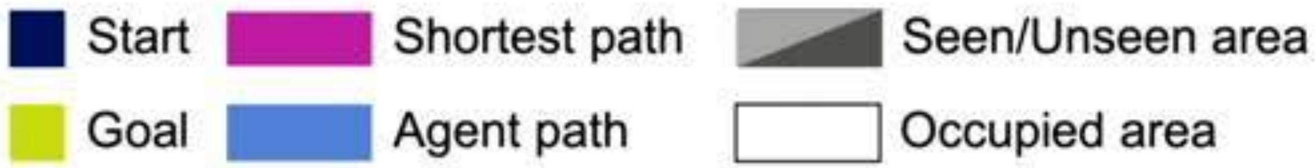
Does audio help navigation?



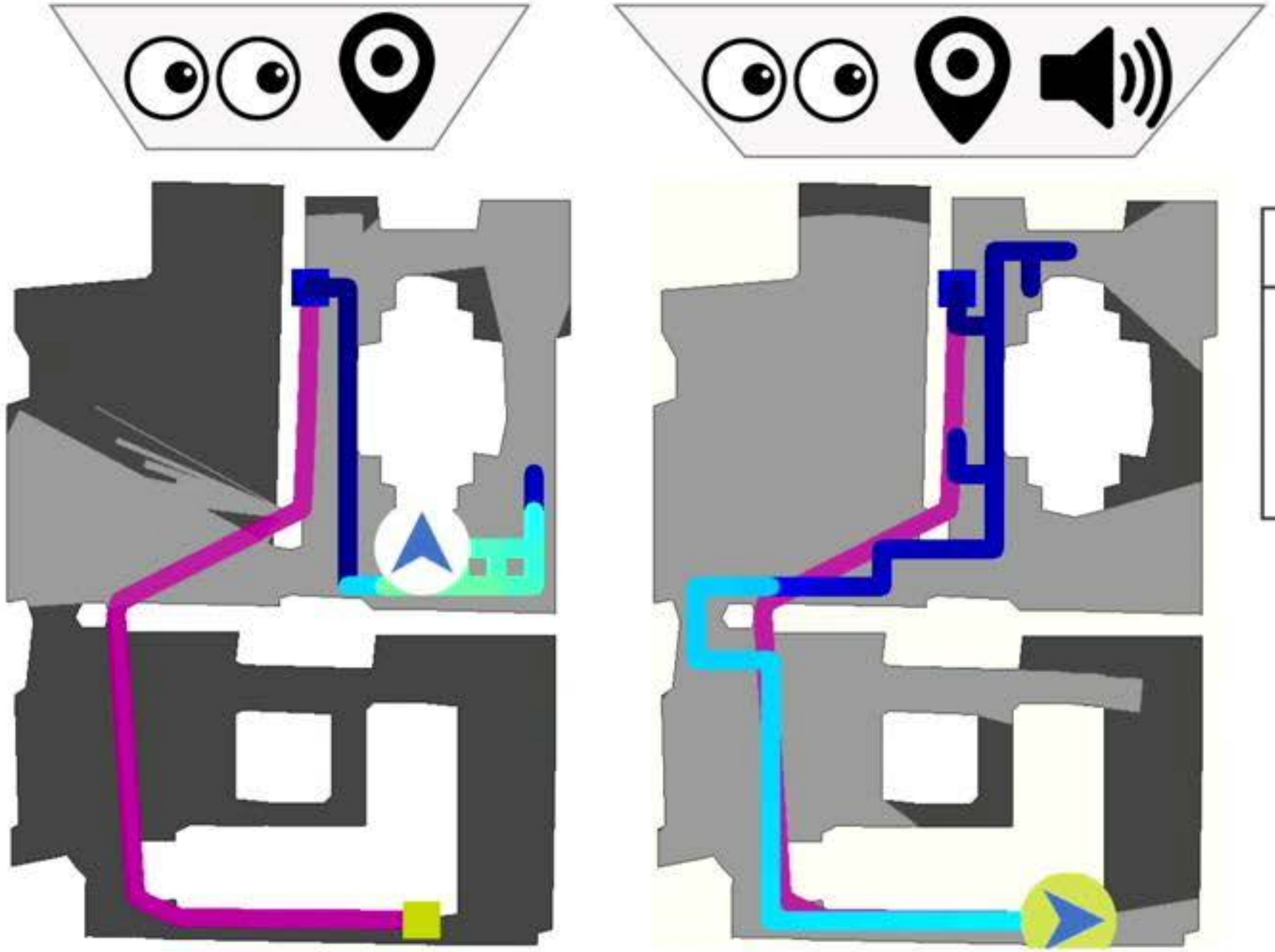
2D t-SNE projection of audio features learned by our agent

Does audio help navigation?





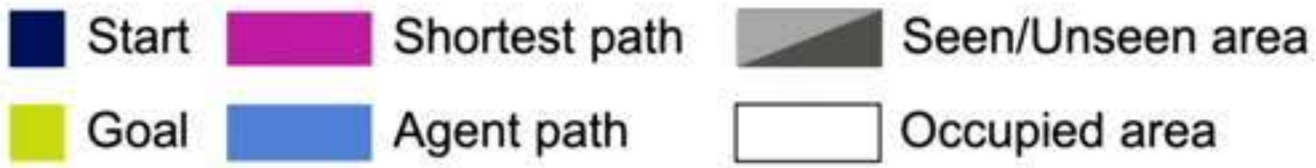
Does audio help navigation?



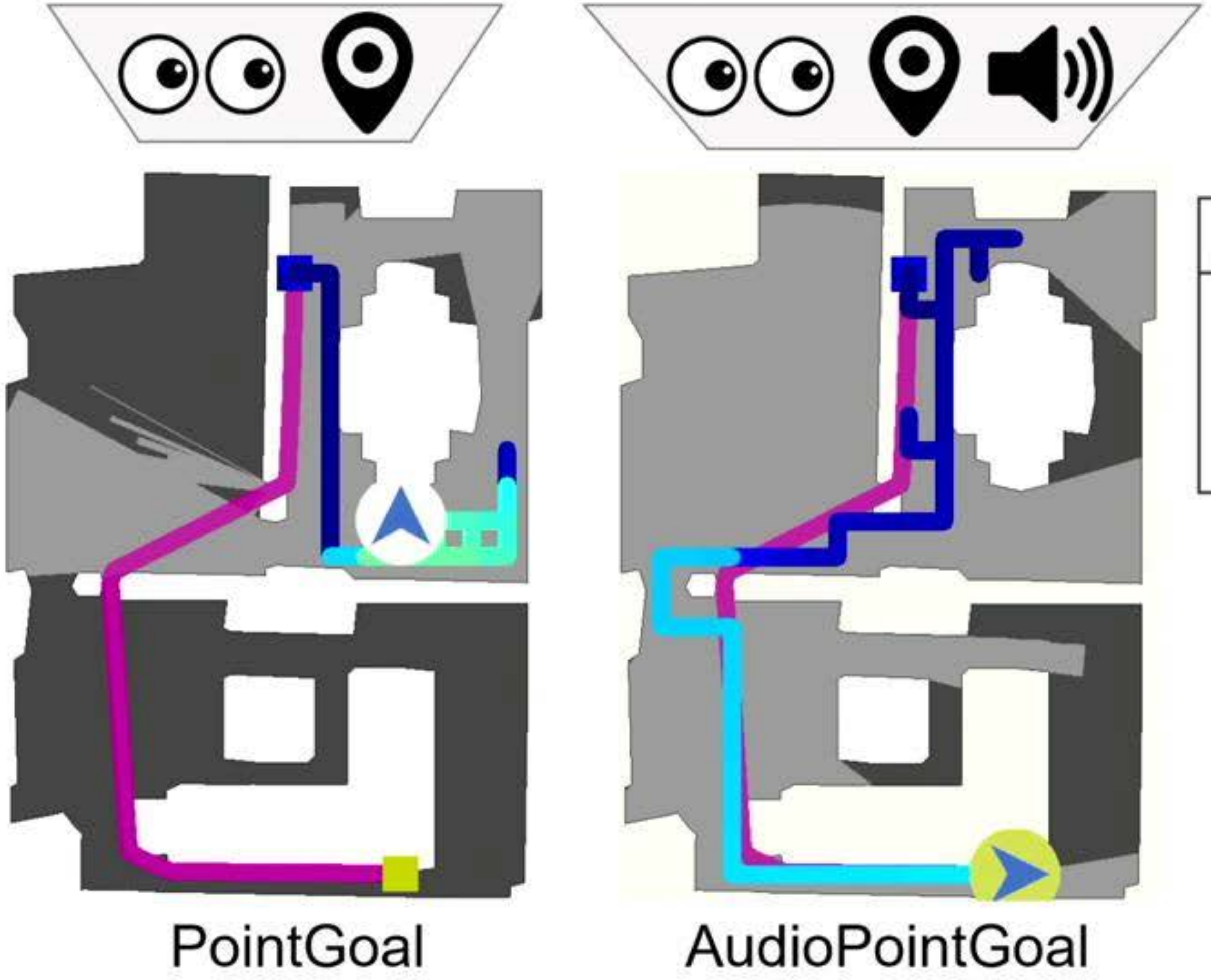
PointGoal AudioPointGoal

	PointGoal	
Blind	0.451	
RGB	0.465	
Depth	0.592	

success rate normalized by path length (SPL)

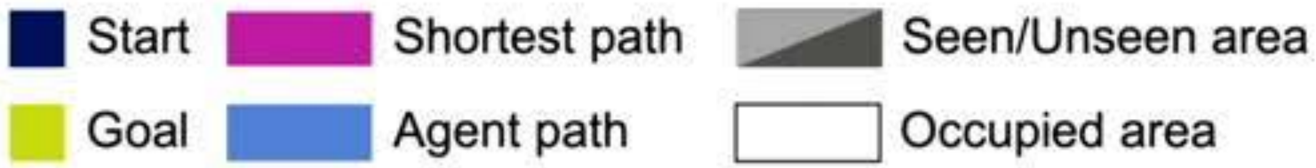


Does audio help navigation?

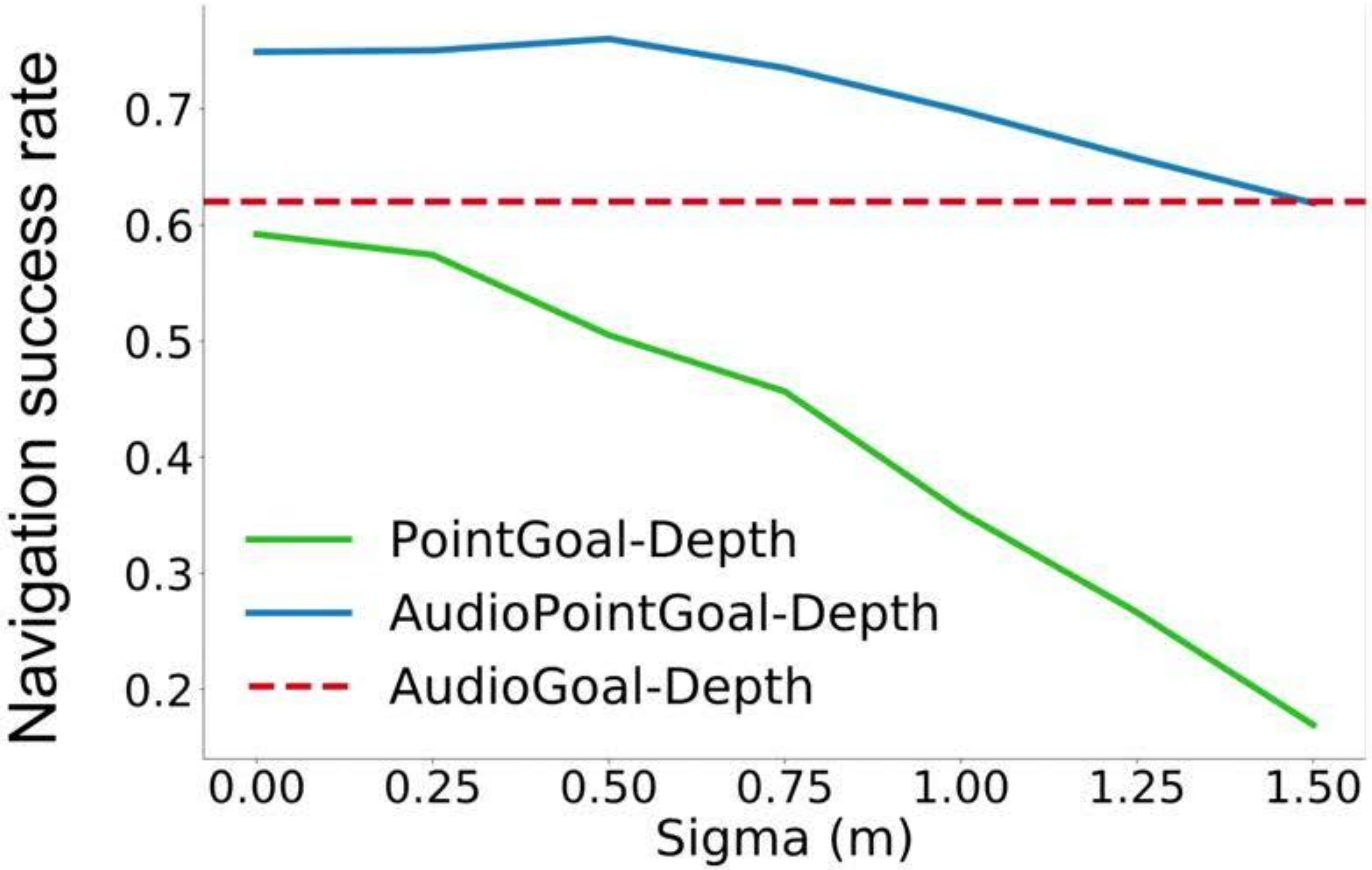


	PointGoal	AudioPointGoal
Blind	0.451	**0.647**
RGB	0.465	**0.735**
Depth	0.592	**0.749**

success rate normalized by path length (SPL)

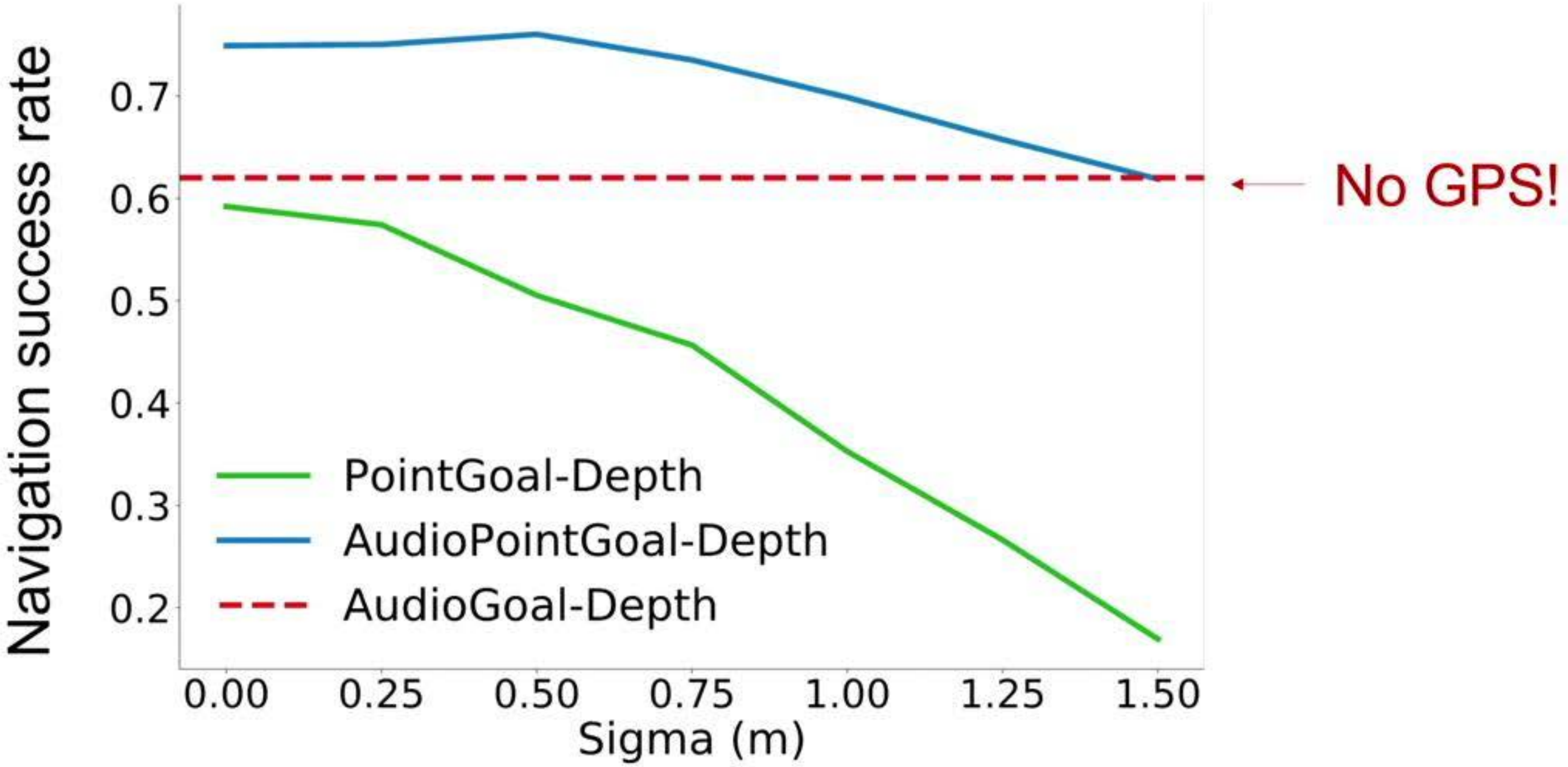


Can audio supplant GPS?



Perfect GPS → Noisier GPS

Can audio supplant GPS?



Perfect GPS ············→ Noisier GPS

This talk

Multi-sensory

Audio-visual learning

Motion

Navigation policies

Interaction

Affordance learning

Towards embodied perception

From *naming* objects to *using* them

Embodied perception system

Object manipulation

Turn on

Increase height

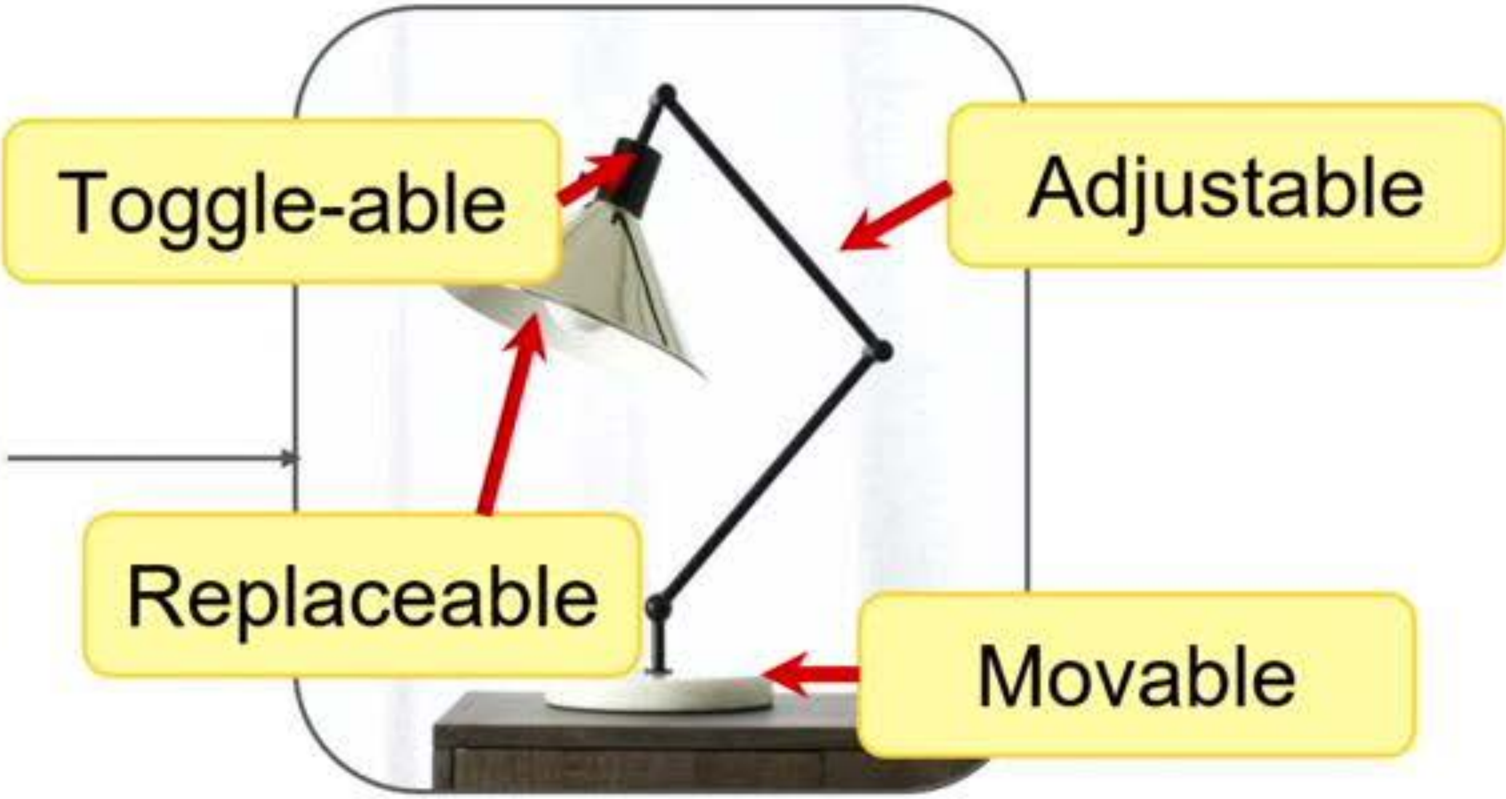
Move lamp

Replace lightbulb

From *naming* objects to *using* them

Embodied
perception system

Affordances



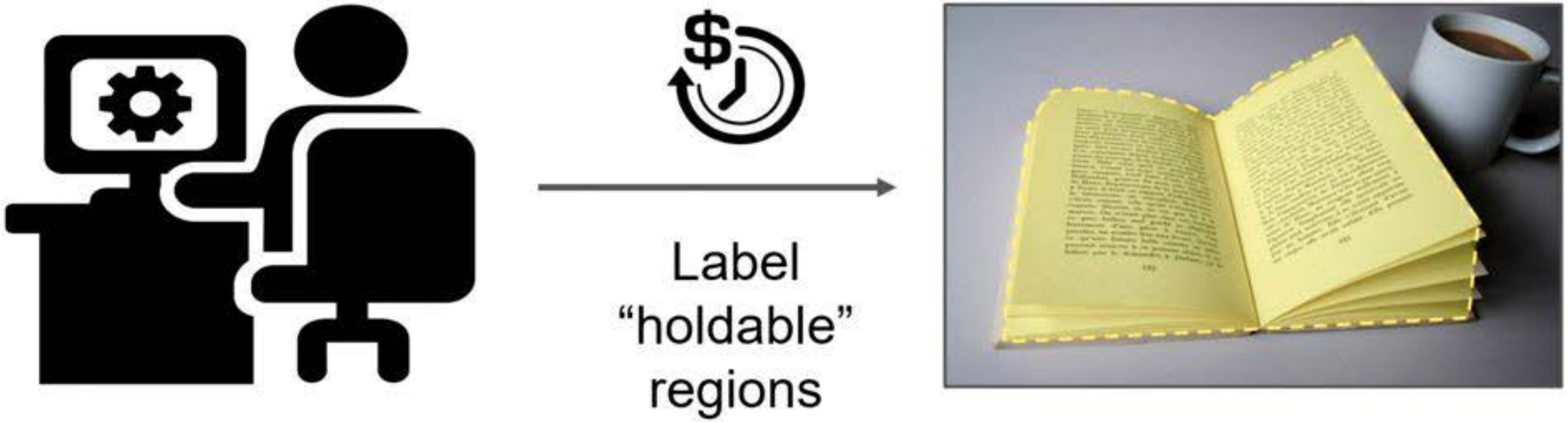
Object
manipulation

Current approaches: affordance as semantic segmentation

Label
“holdable”
regions

Sawatzky et al. (CVPR 17), Nguyen et al. (IROS 17), Roy et al. (ECCV 16), Myers et al. (ICRA 15), …

Current approaches: affordance as semantic segmentation



Sawatzky et al. (CVPR 17), Nguyen et al. (IROS 17), Roy et al. (ECCV 16), Myers et al. (ICRA 15), …

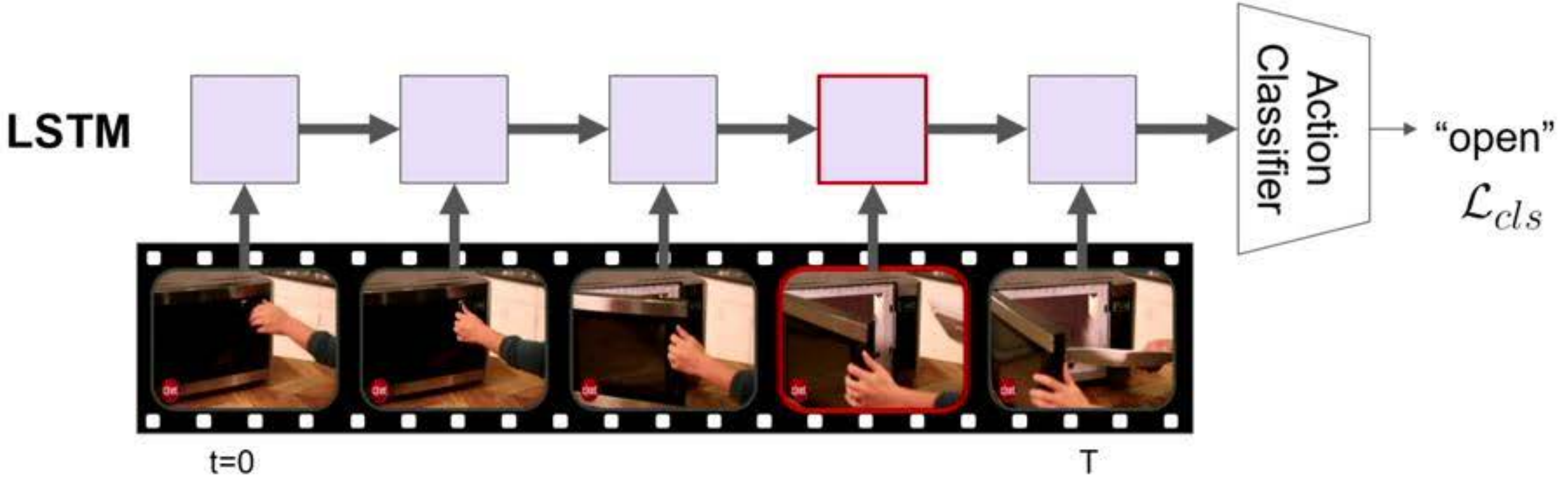
Current approaches: affordance as semantic segmentation

Label
"holdable"
regions

Captures annotators' expectations of what is important

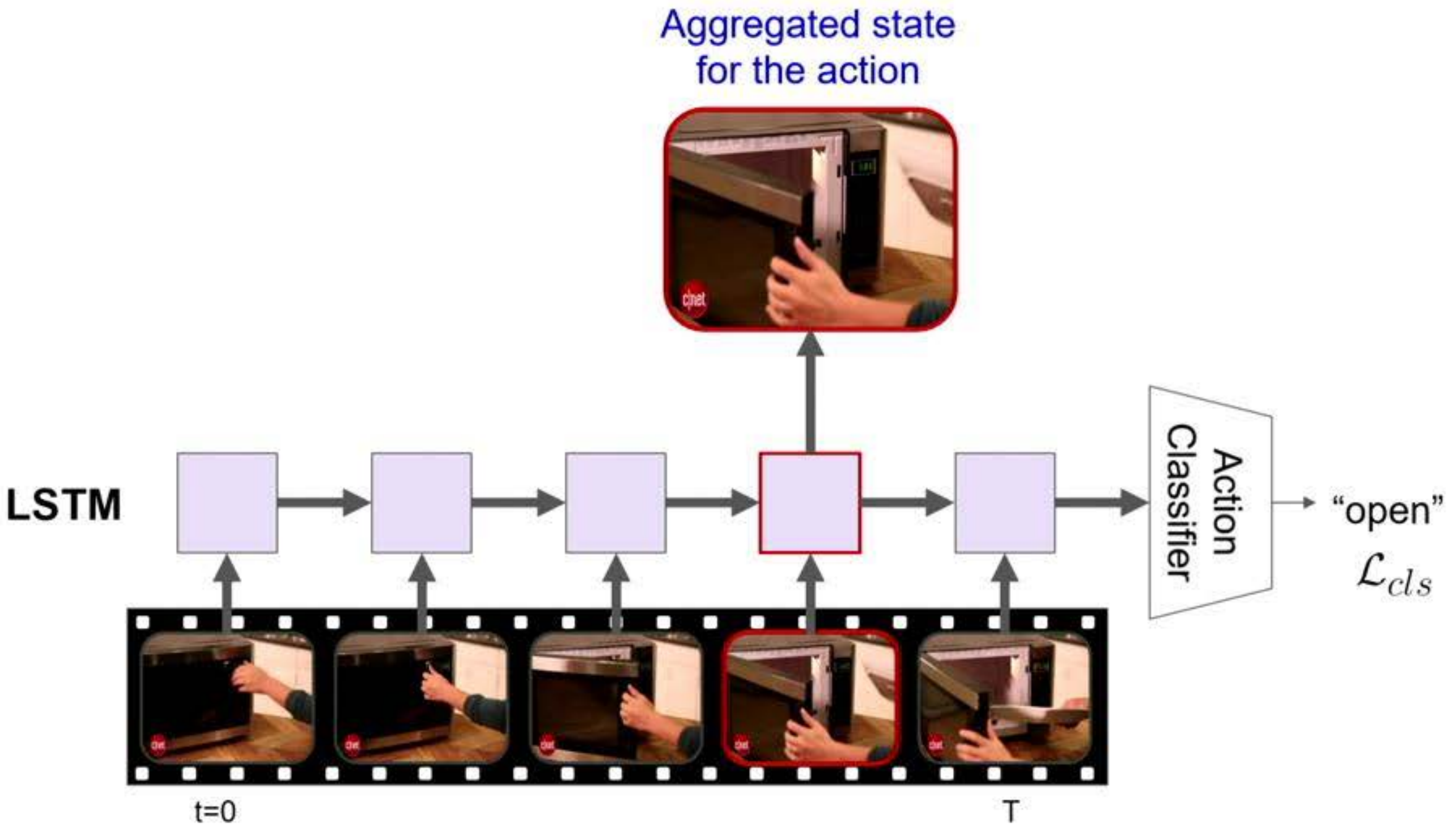
Sawatzky et al. (CVPR 17), Nguyen et al. (IROS 17), Roy et al. (ECCV 16), Myers et al. (ICRA 15), …

Learning affordances from video



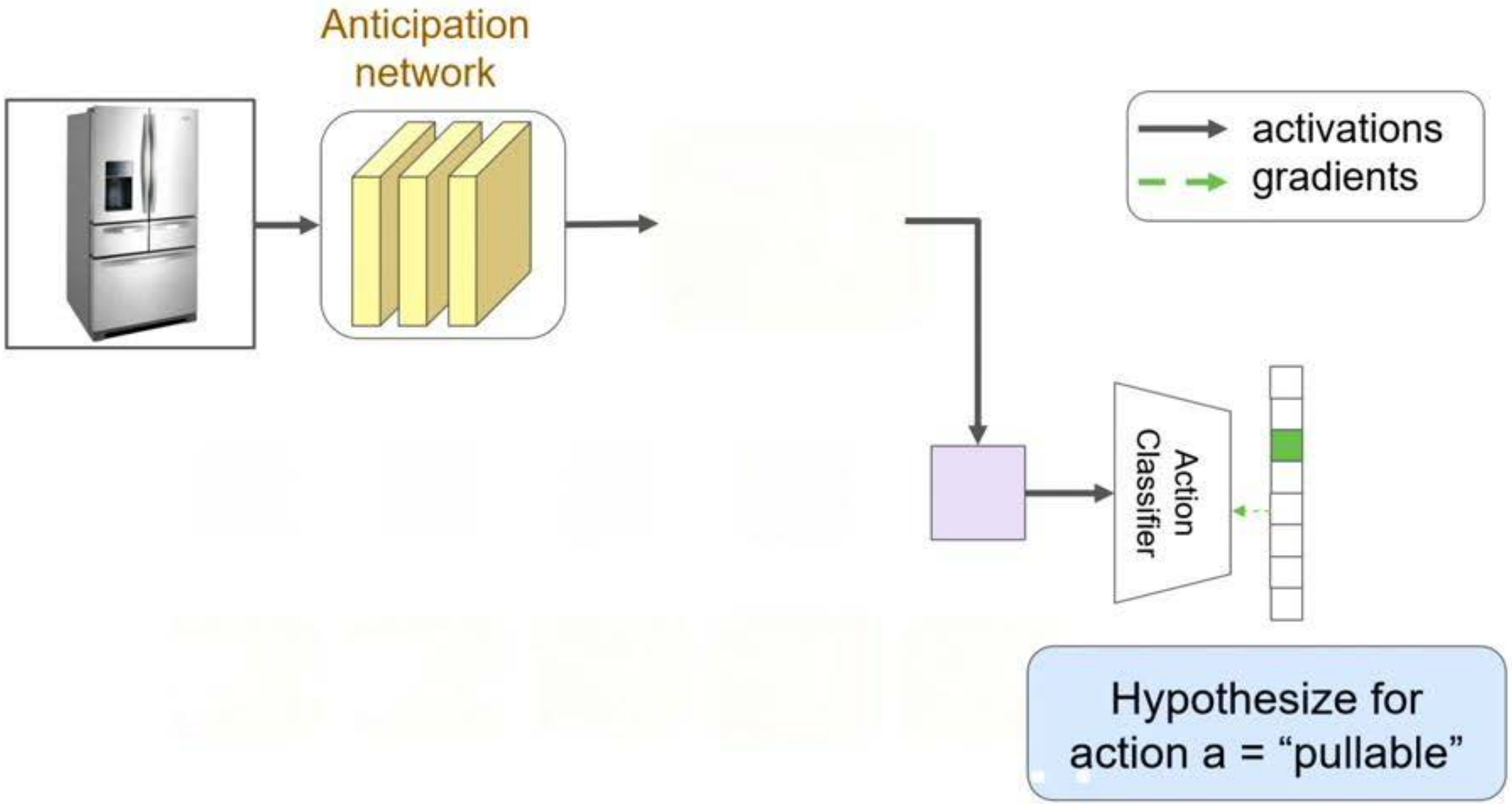
[Nagarajan et al. ICCV 2019]

Learning affordances from video



[Nagarajan et al. ICCV 2019]

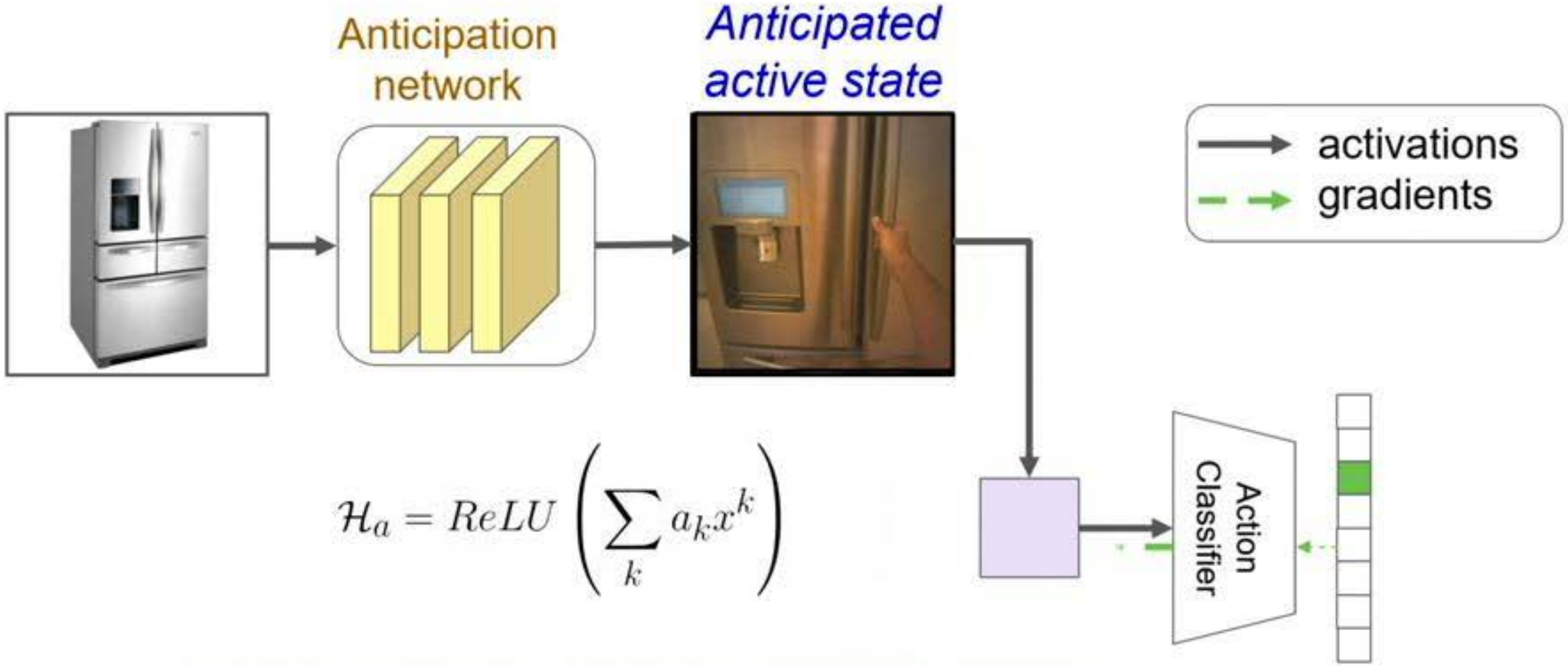
Extracting interaction hotspot maps



Activation mapping to identify responsible spatial regions

[Nagarajan et al. ICCV 2019]

Extracting interaction hotspot maps



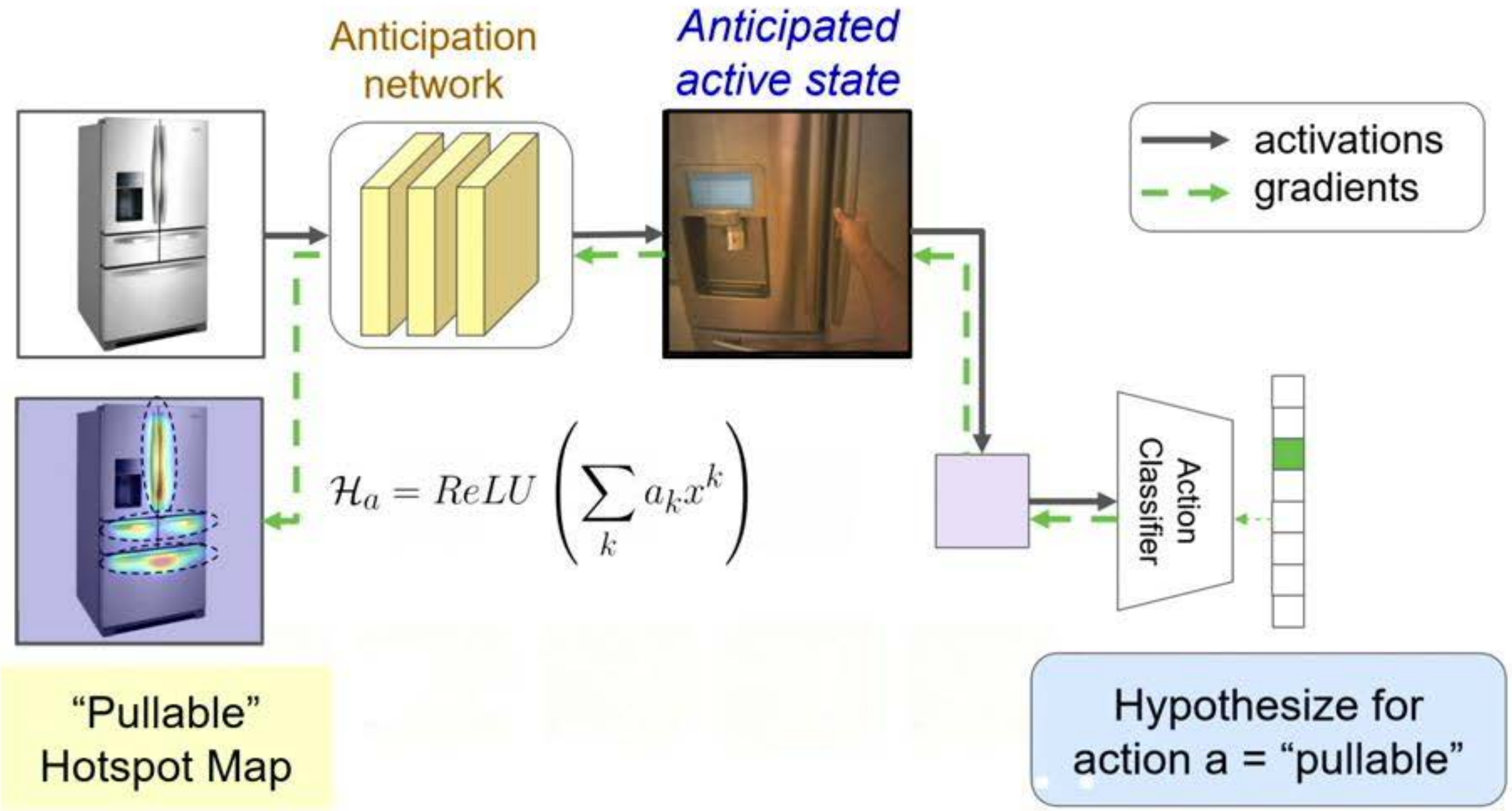
$$\mathcal{H}_a = ReLU\left(\sum_k a_k x^k\right)$$

Hypothesize for
action a = “pullable”

Activation mapping to identify responsible spatial regions

[Nagarajan et al. ICCV 2019]

Extracting interaction hotspot maps



Activation mapping to identify responsible spatial regions

[Nagarajan et al. ICCV 2019]

Evaluating interaction hotspots

OPRA
(Fang et al., CVPR 18)

EPIC Kitchens
(Damen et al., ECCV 18)

Evaluating interaction hotspots

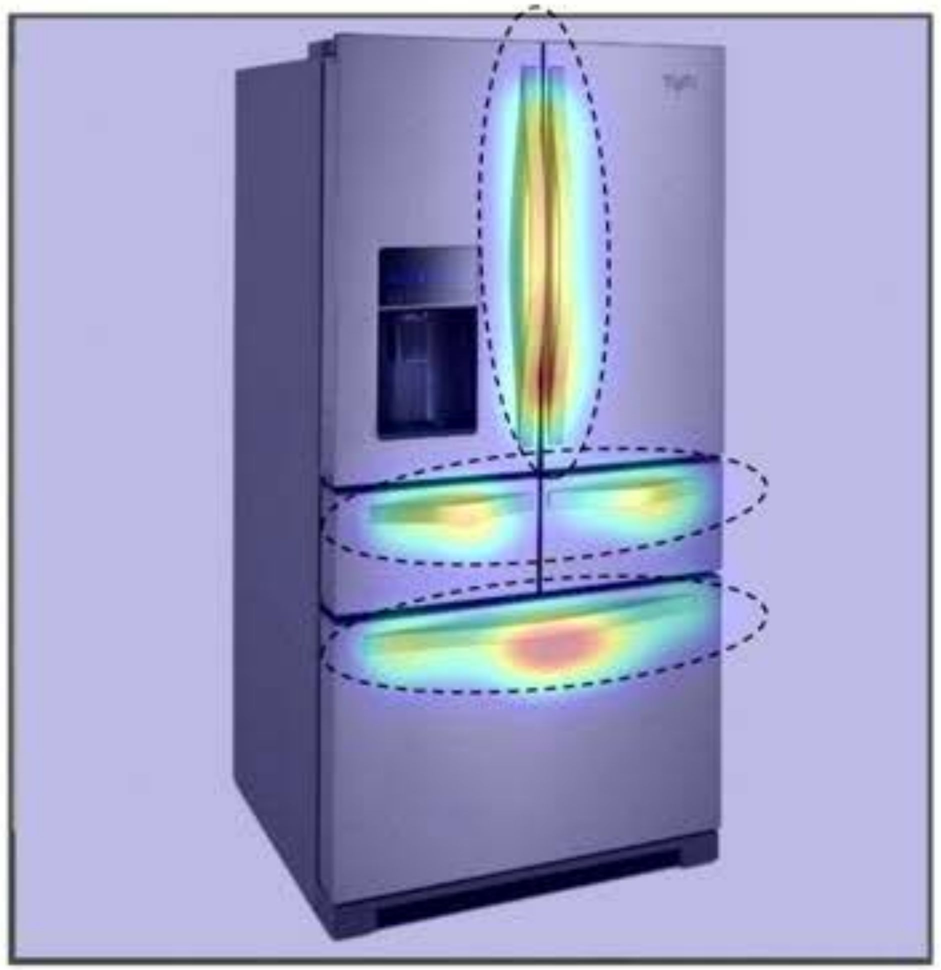
OPRA
(Fang et al., CVPR 18)

EPIC Kitchens
(Damen et al., ECCV 18)

Train on video datasets, generate heatmaps on novel images---
even from unseen categories

Results: interaction hotspots

Given static image of object at rest, infer affordance regions



[Nagarajan et al. ICCV 2019]

Results: interaction hotspots

Given static image of object at rest, infer affordance regions

	OPRA data			EPIC data		
	KLD ↓	SIM ↑	AUC-J ↑	KLD ↓	SIM ↑	AUC-J ↑
CENTER BIAS	11.132	0.205	0.625	10.660	0.222	0.634
LSTM+GRAD-CAM	8.573	0.209	0.620	6.470	0.257	0.626
EGOGAZE [27]	2.428	0.245	0.646	2.241	0.273	0.614
MLNET [6]	4.022	0.284	0.763	6.116	0.318	0.746
DEEPGAZEII [33]	1.897	0.296	0.720	1.352	0.394	0.751
SALGAN [40]	2.116	0.309	0.769	1.508	0.395	0.774
OURS	**1.427**	**0.362**	**0.806**	**1.258**	**0.404**	**0.785**
IMG2HEATMAP	1.473	0.355	0.821	1.400	0.359	0.794
DEMO2VEC [12]	1.197	0.482	0.847	–	–	–

[Nagarajan et al. ICCV 2019]

Results: interaction hotspots

Given static image of object at rest, infer affordance regions

Weakly Supervised

	OPRA data			EPIC data		
	KLD ↓	SIM ↑	AUC-J ↑	KLD ↓	SIM ↑	AUC-J ↑
CENTER BIAS	11.132	0.205	0.625	10.660	0.222	0.634
LSTM+GRAD-CAM	8.573	0.209	0.620	6.470	0.257	0.626
EGOGAZE [27]	2.428	0.245	0.646	2.241	0.273	0.614
MLNET [6]	4.022	0.284	0.763	6.116	0.318	0.746
DEEPGAZEII [33]	1.897	0.296	0.720	1.352	0.394	0.751
SALGAN [40]	2.116	0.309	0.769	1.508	0.395	0.774
OURS	**1.427**	**0.362**	**0.806**	**1.258**	**0.404**	**0.785**
IMG2HEATMAP	1.473	0.355	0.821	1.400	0.359	0.794
DEMO2VEC [12]	1.197	0.482	0.847	–	–	–

[Nagarajan et al. ICCV 2019]

Interaction hotspots for object recognition

Interaction hotspots for object recognition

ResNet-50 predictions on COCO objects

refrigerator - 0.997
dishwasher - 0.001

refrigerator - 0.454
ATM machine - 0.210

mailbox - 0.404
refrigerator - 0.139

bookstore - 0.747
refrigerator - 0.009

switchbox - 0.511
refrigerator - 0.005

Interaction hotspots for object recognition

N →	COCO			
	5	25	100	3300 (all)
VANILLA	44.3 ± 0.3	56.6 ± 0.2	**65.6 ± 0.4**	**75.2 ± 0.1**
AUTOENCODER	39.4 ± 0.4	51.2 ± 0.2	59.1 ± 0.2	72.8 ± 0.3
OURS	**46.8 ± 0.3**	**57.9 ± 0.1**	63.2 ± 0.2	73.9 ± 0.3

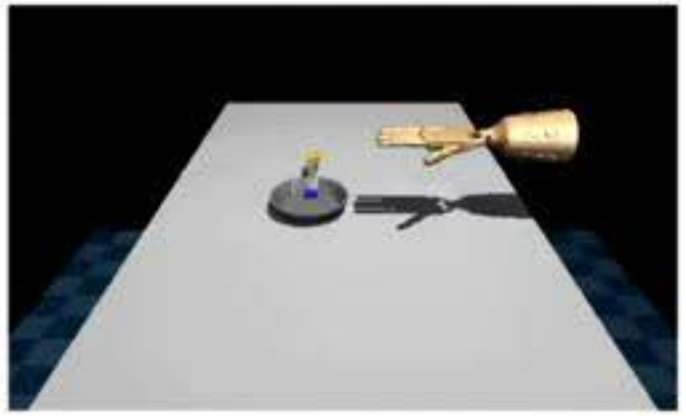
Interaction hotspots for object recognition

	COCO			
N →	5	25	100	3300 (all)
VANILLA	44.3 ± 0.3	56.6 ± 0.2	**65.6 ± 0.4**	**75.2 ± 0.1**
AUTOENCODER	39.4 ± 0.4	51.2 ± 0.2	59.1 ± 0.2	72.8 ± 0.3
OURS	**46.8 ± 0.3**	**57.9 ± 0.1**	63.2 ± 0.2	73.9 ± 0.3

Better low-shot object recognition by anticipating object function

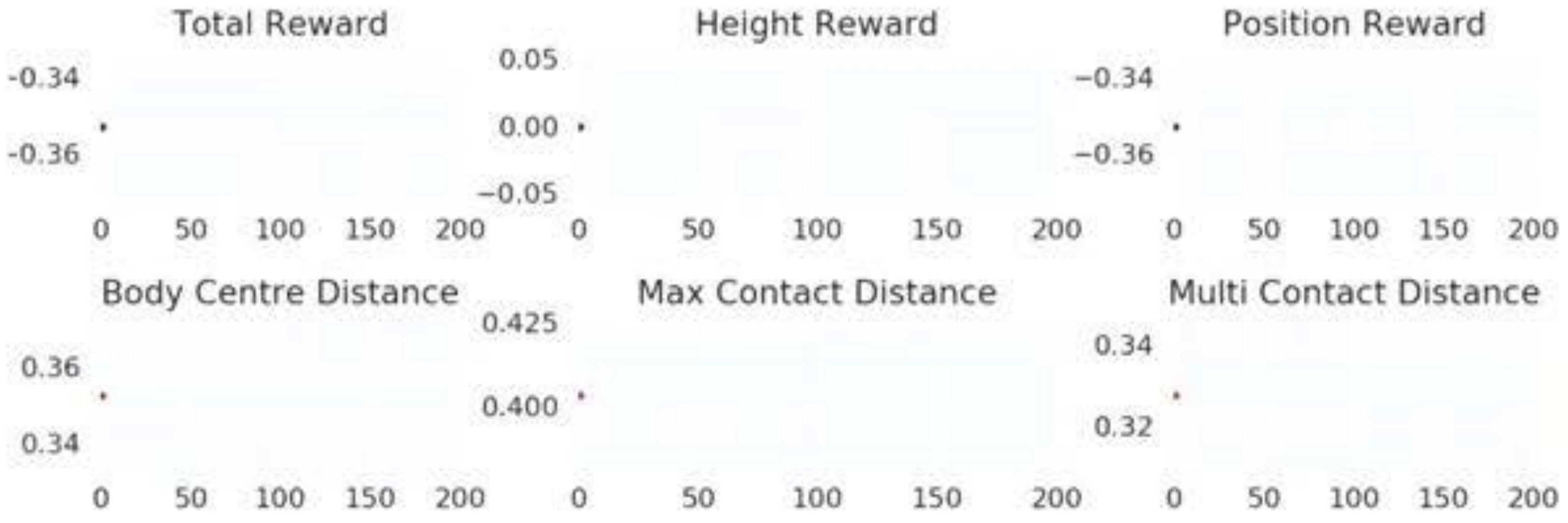
Interaction hotspots for robot grasping

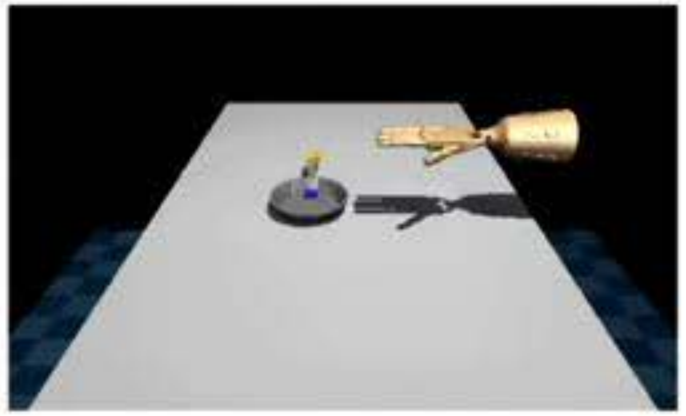
Without watching people



Interaction hotspots for robot grasping

Without watching people





Learn grasping policy for 24 DoF dexterous hand
that rewards closeness to hotspots

Summary

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

Towards first-person perception
- self-supervised learning via anticipation
- learning to autonomously direct the camera
- multi-sensory observations (audio, motion, visual)
- object interaction from video

Ruohan Gao

Tushar Nagarajan

Changan Chen

Unnat Jain

Christoph Feichtenhofer

Carl Schissler

Sebastià V. Amengual Garí