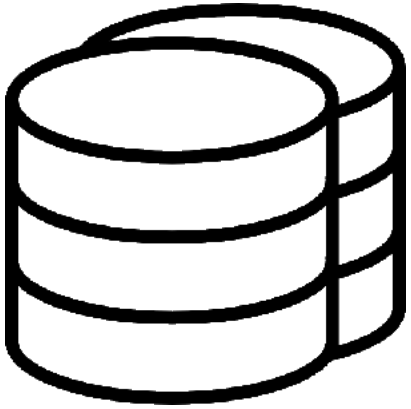
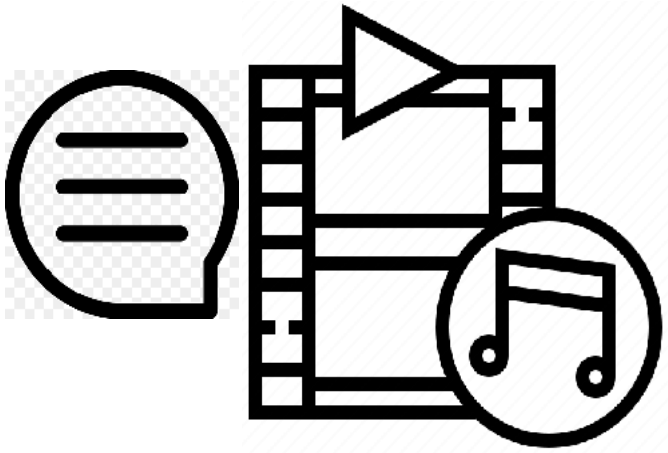
First-person vision: examples in research



Data foundations

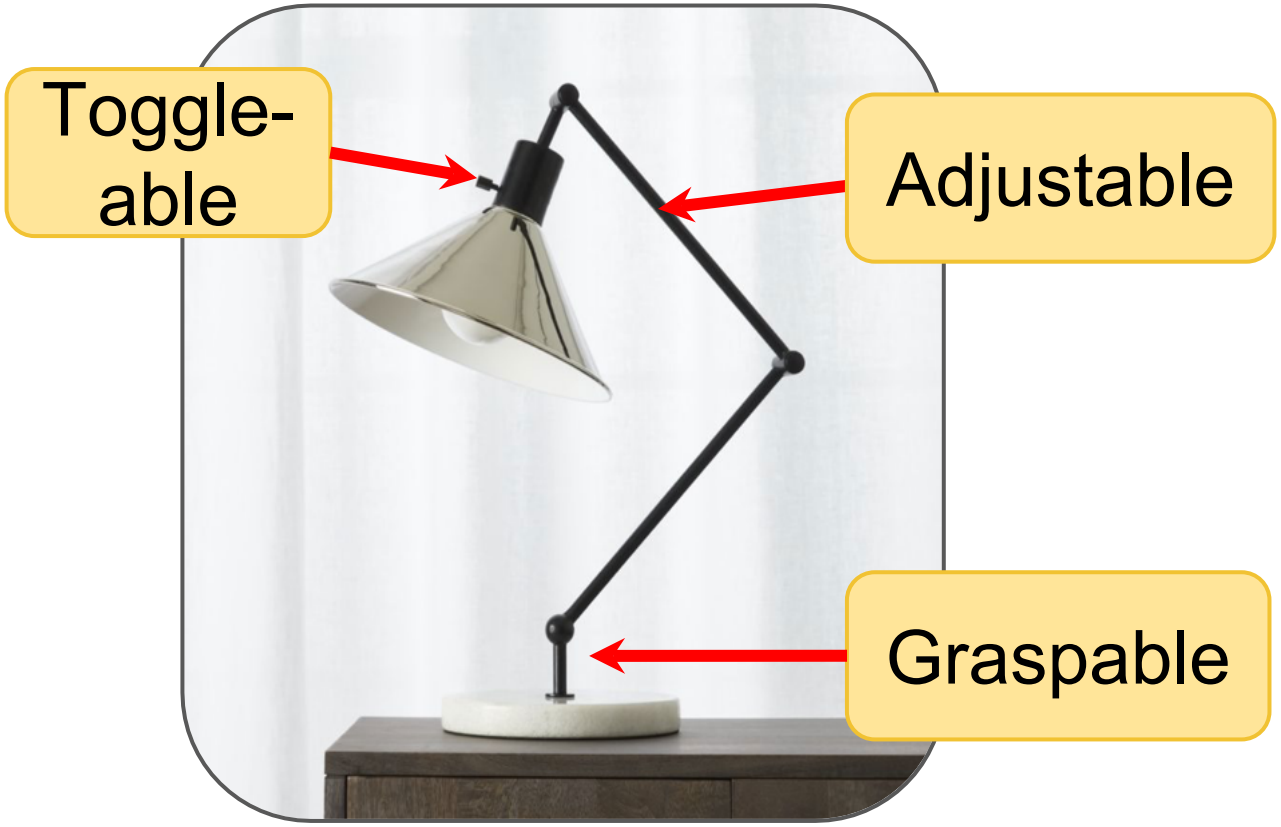
Affordances and manipulation



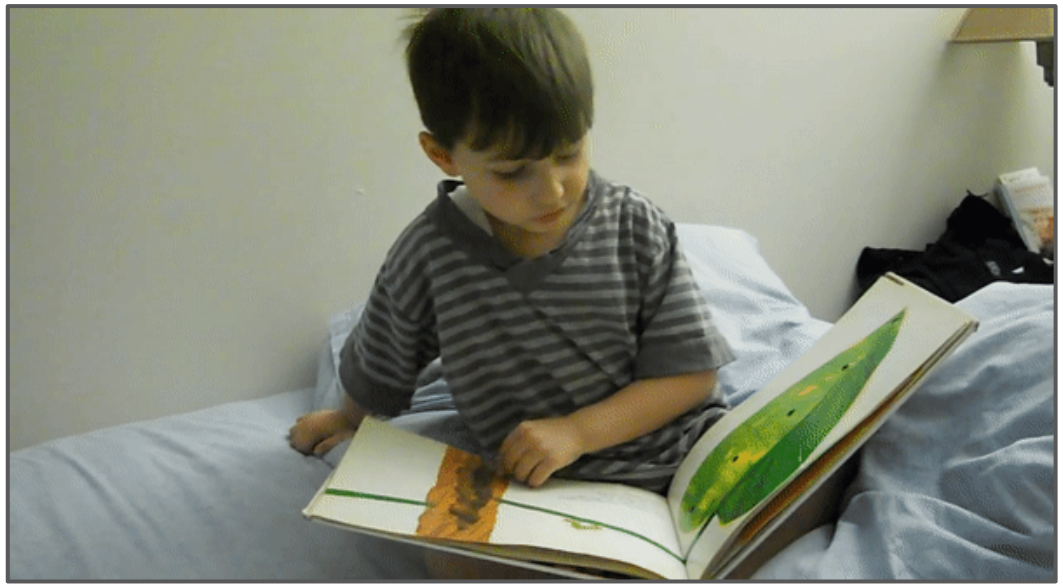
Multimodal learning

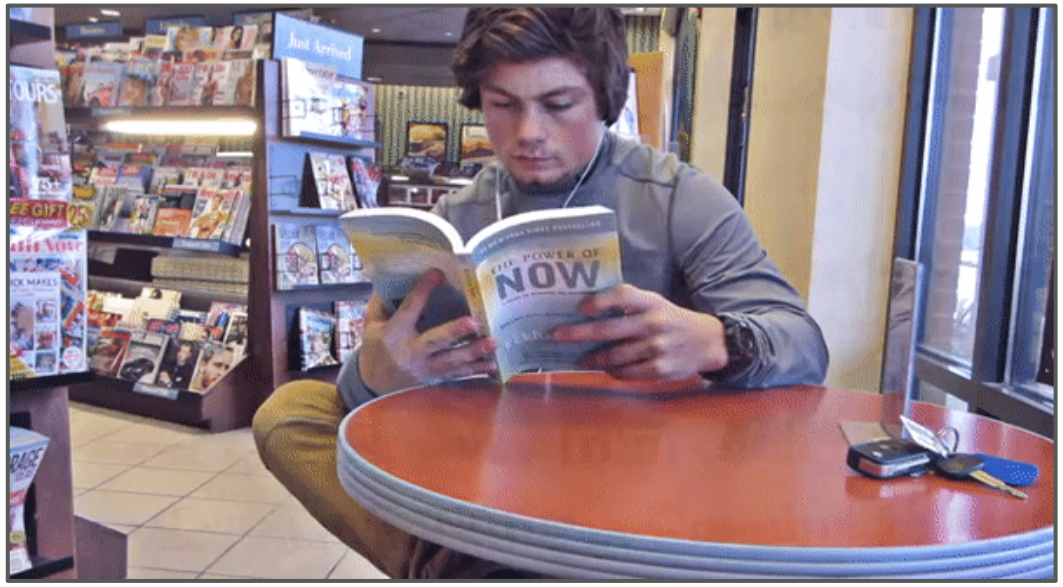
From *naming* objects to *using* them

Lamp!

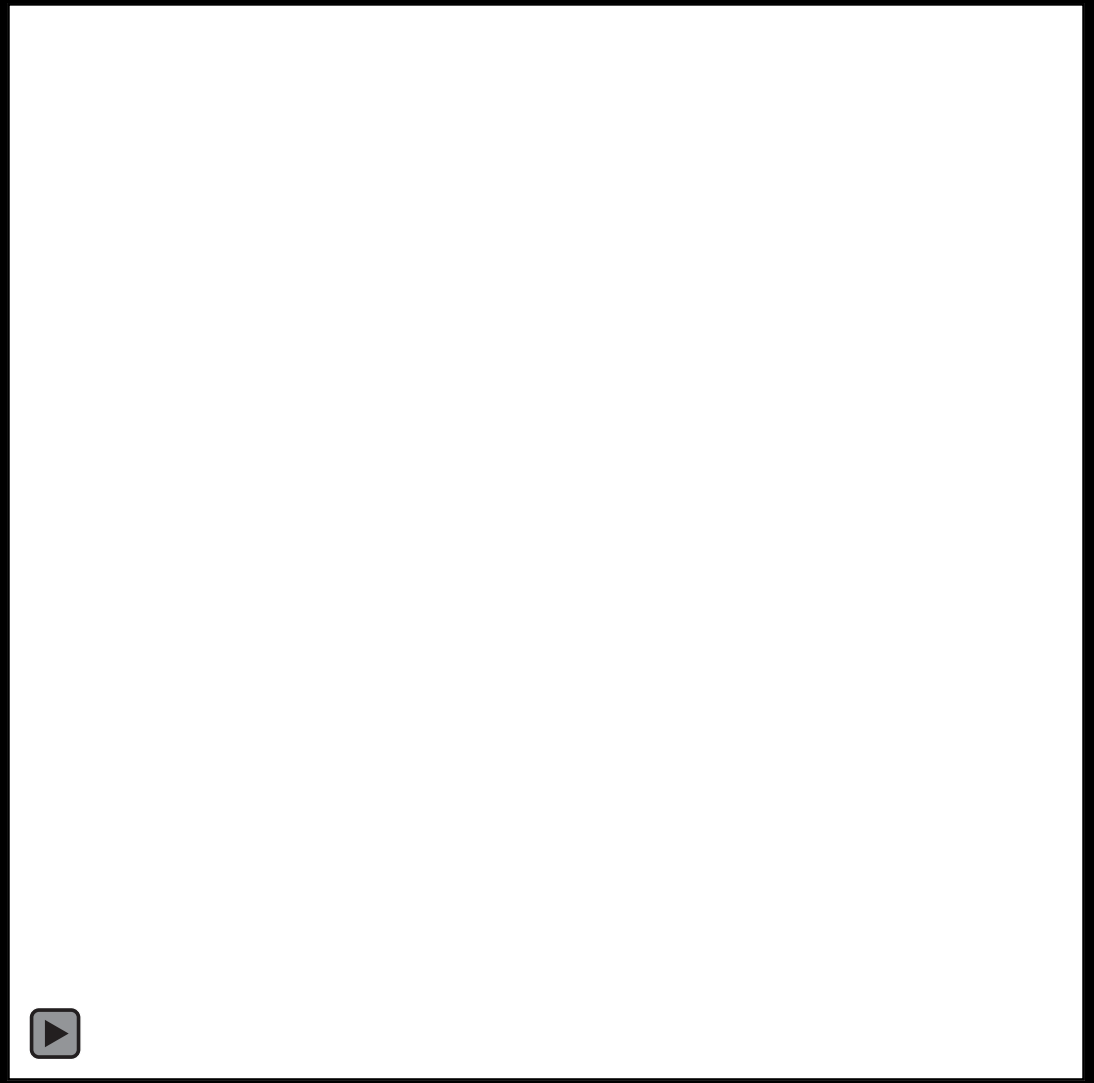


Our idea: Learn affordances from video





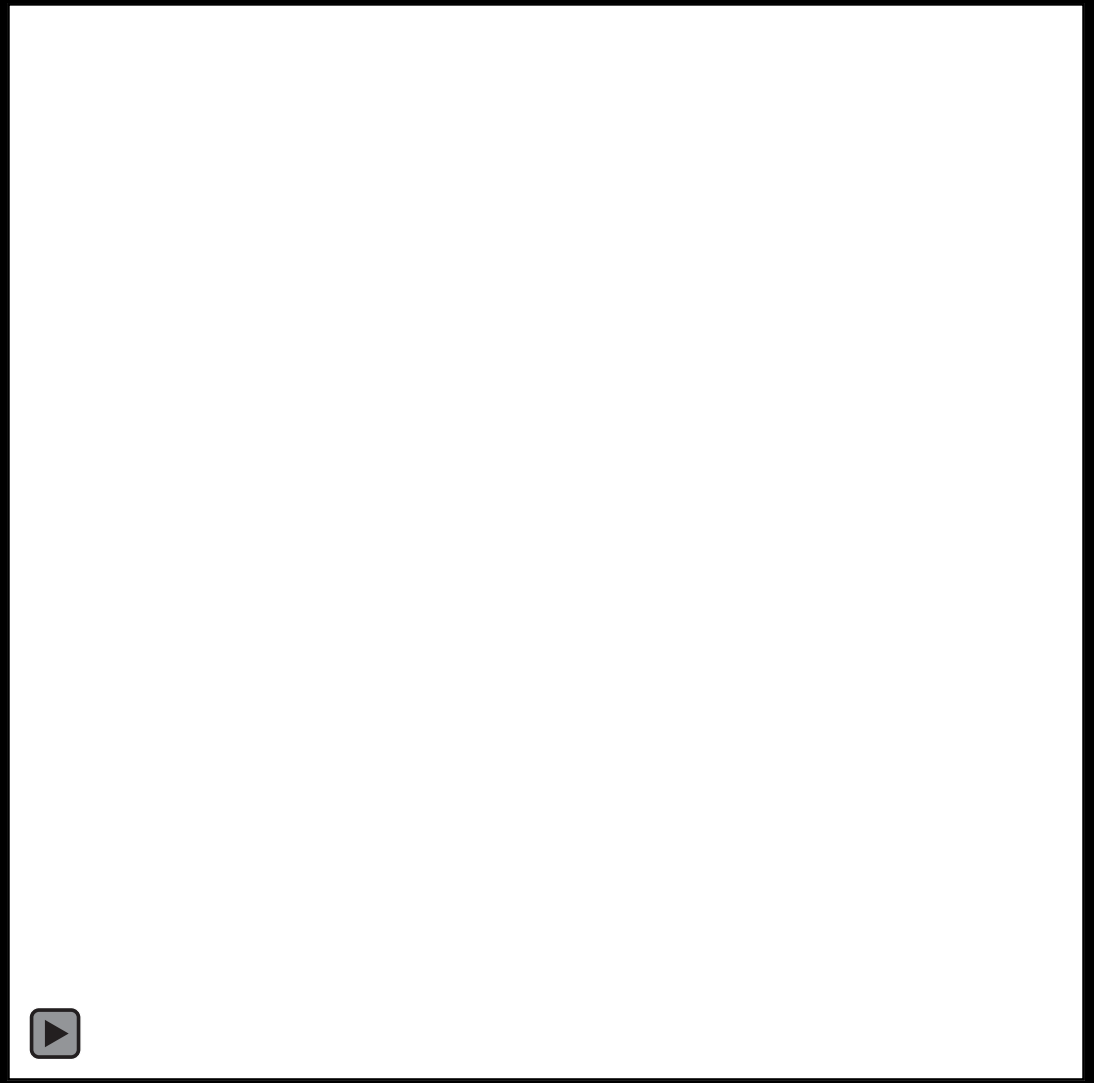
Results: interaction hotspots



cuttable mixable adjustable openable washable

[Nagarajan et al. ICCV 2019]

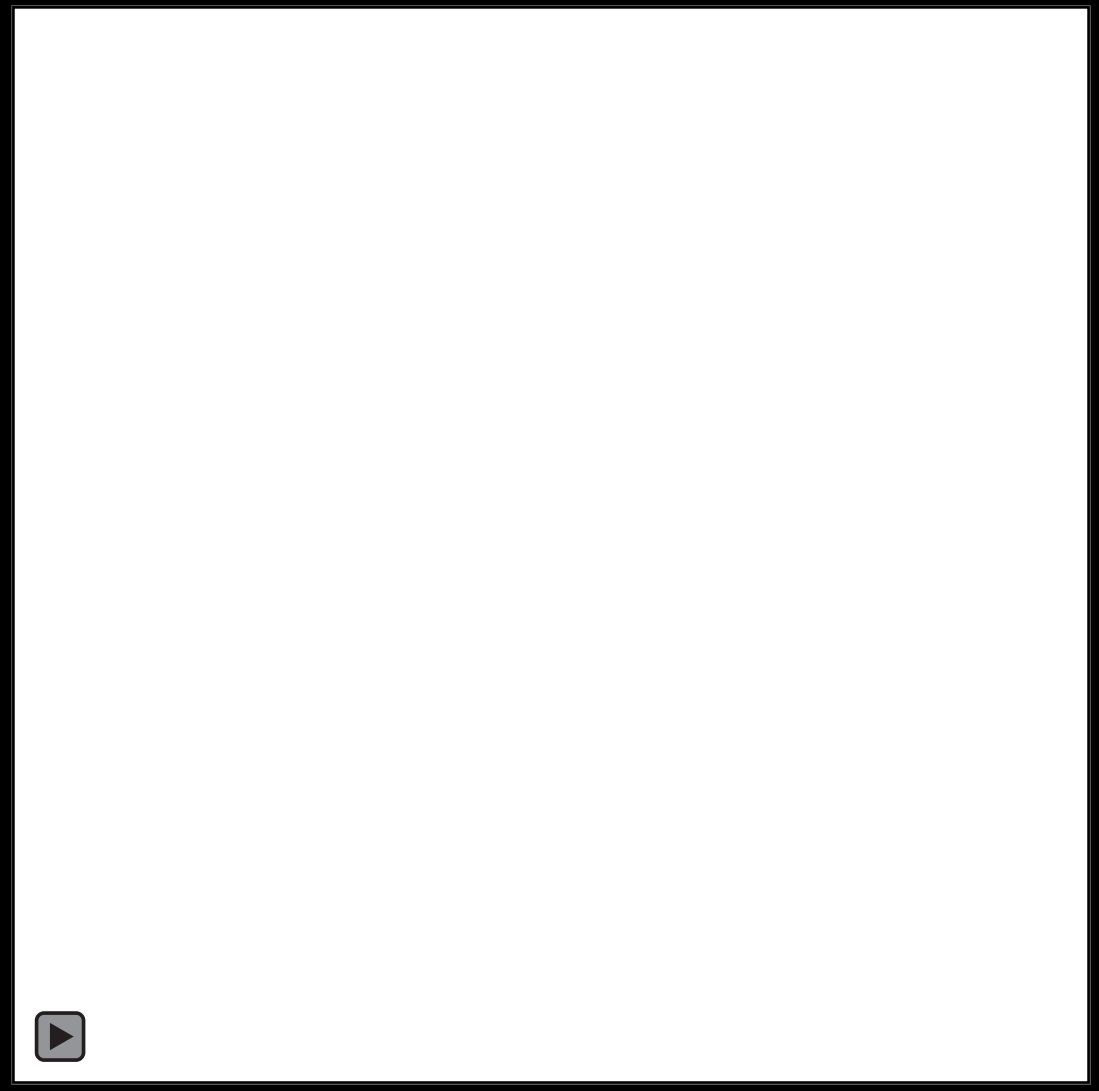
Results: interaction hotspots



cuttable mixable adjustable openable washable

[Nagarajan et al. ICCV 2019]

Results: interaction hotspots

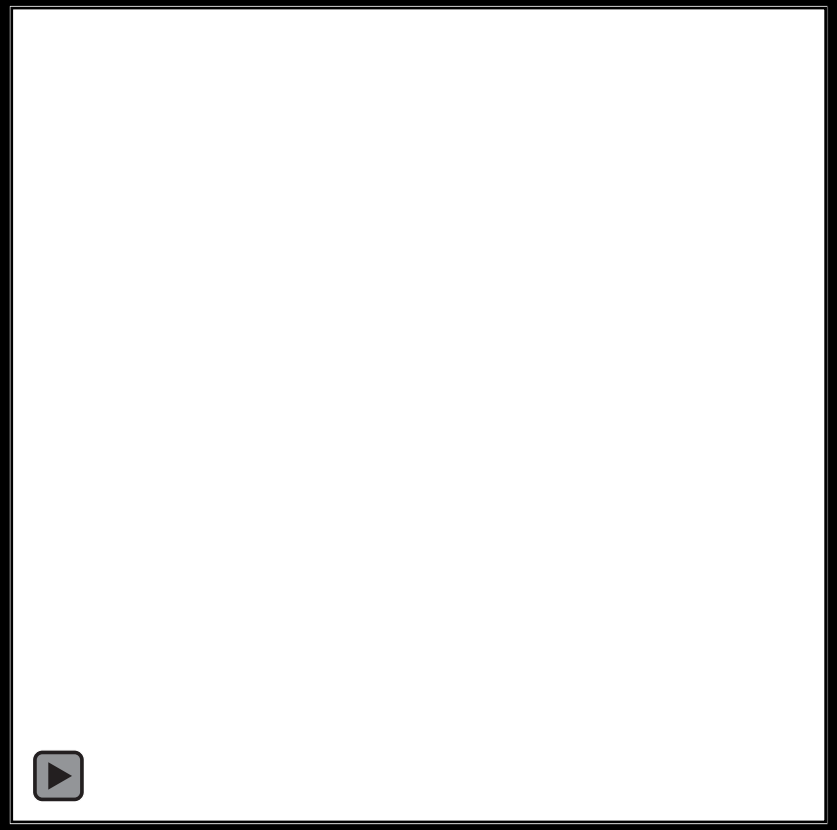


cuttable mixable adjustable openable washable

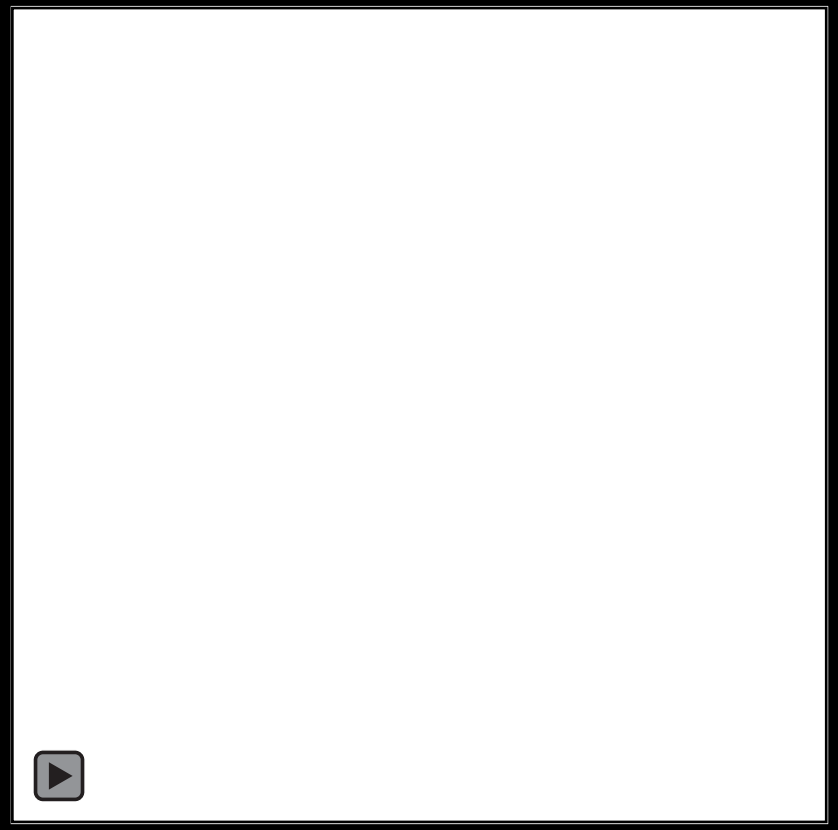
[Nagarajan et al. ICCV 2019]

Results: hotspots vs. saliency

Pan et al. 2017



Ours



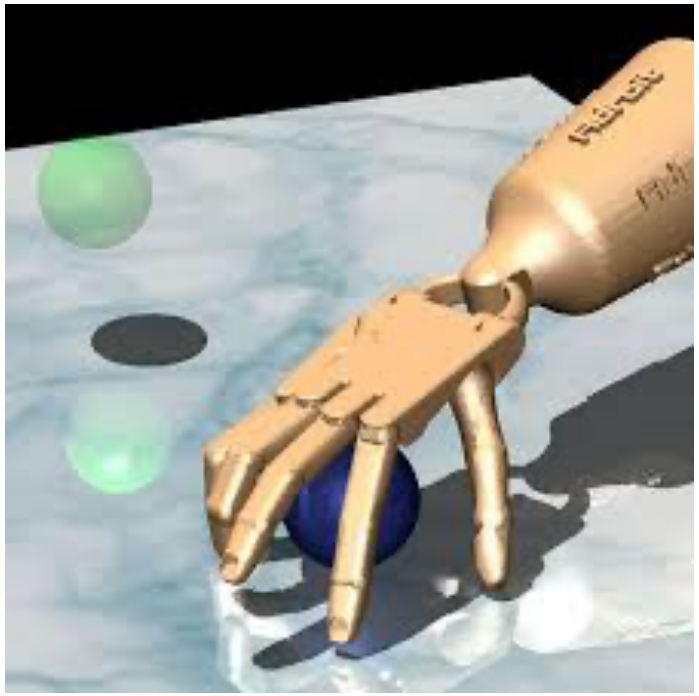
saliency

cuttable mixable adjustable

openable washable

[Nagarajan et al. ICCV 2019]

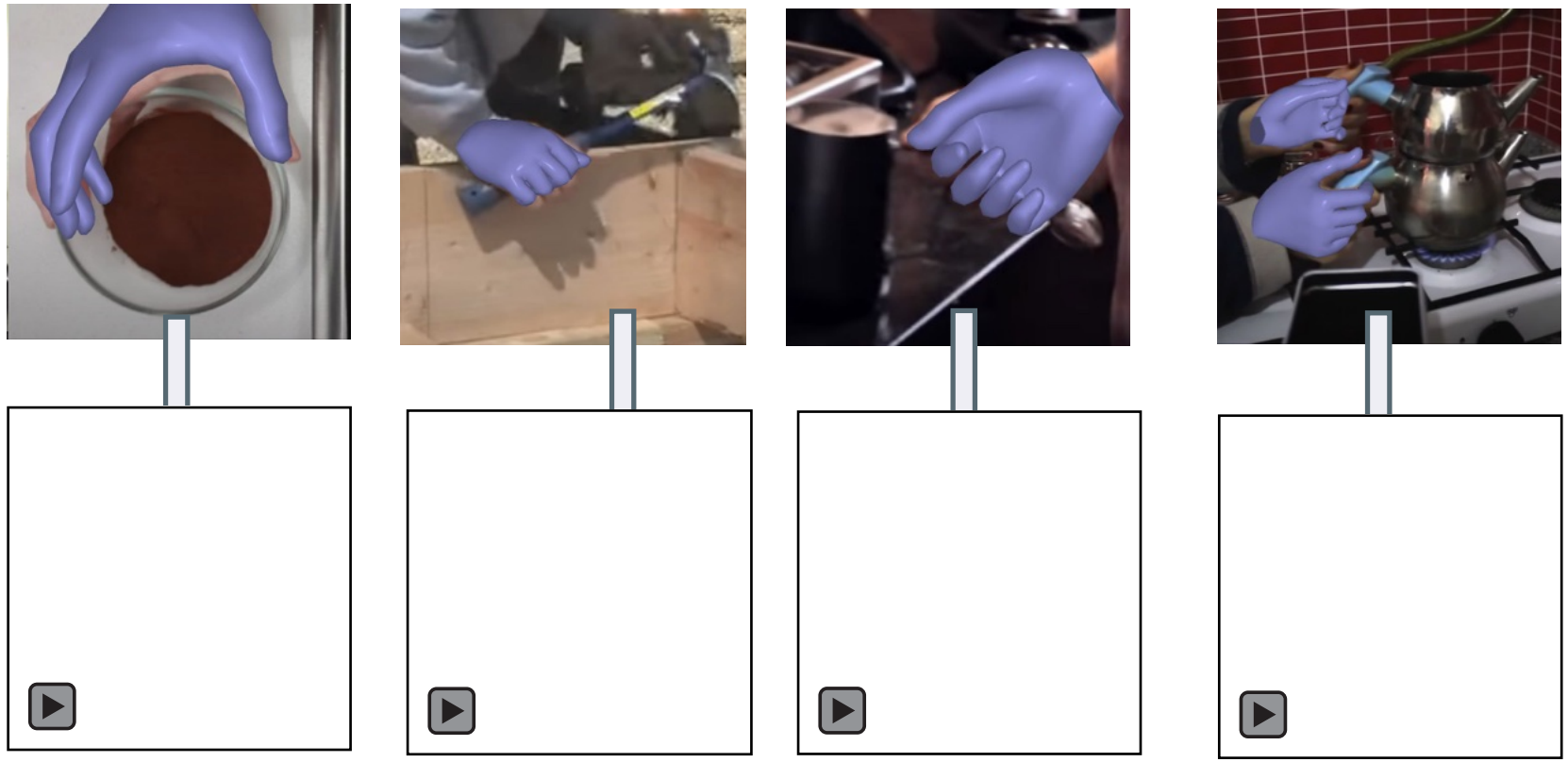
Learning dexterous grasping



Dexterity required, but highly complex action space!

Learning dexterous grasping

Idea: Can agent learn faster by preferring grasping at human affordance regions and with human-like hand poses?



In-the wild YouTube videos

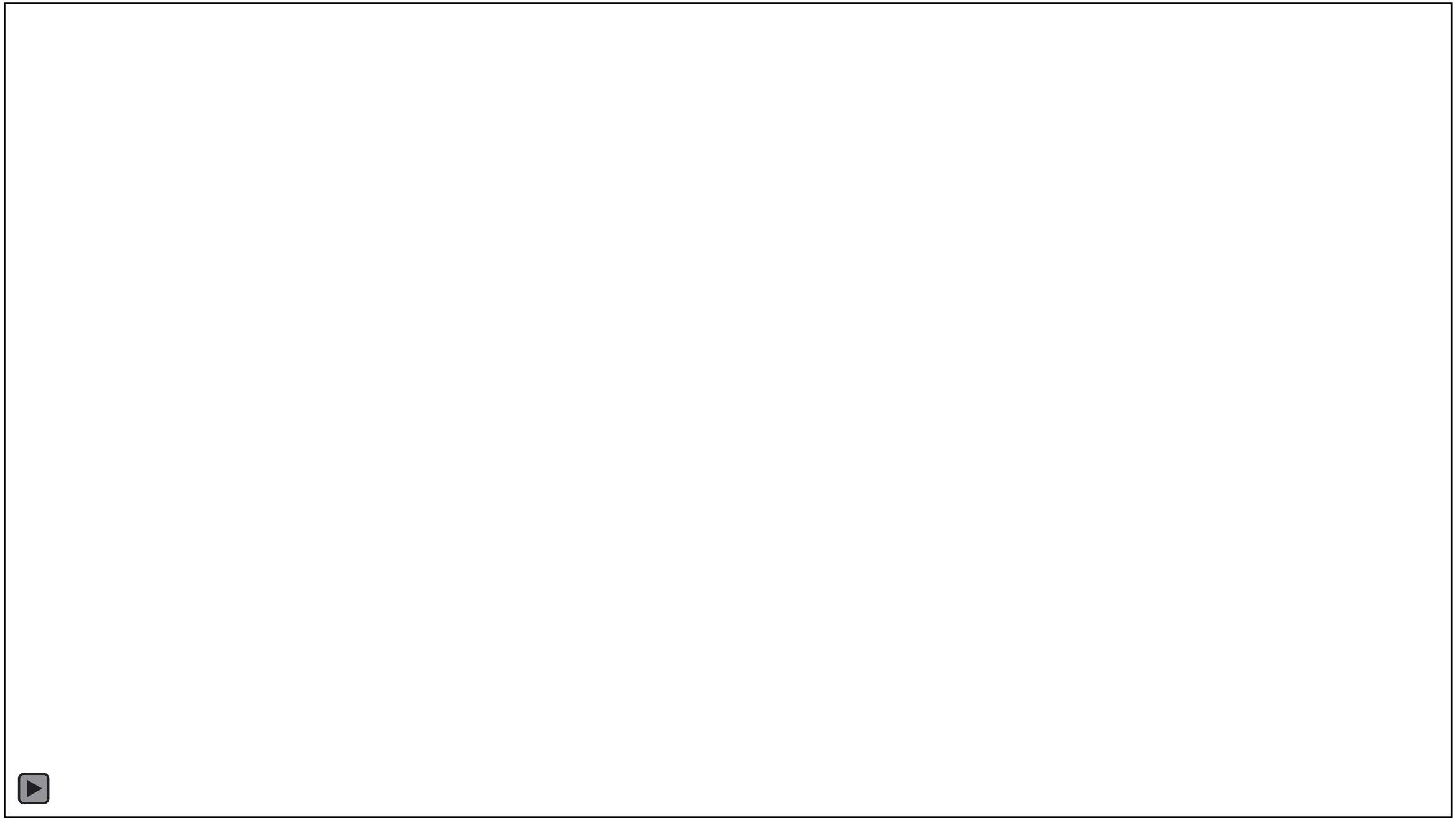
↓

Dexterous robotic grasping

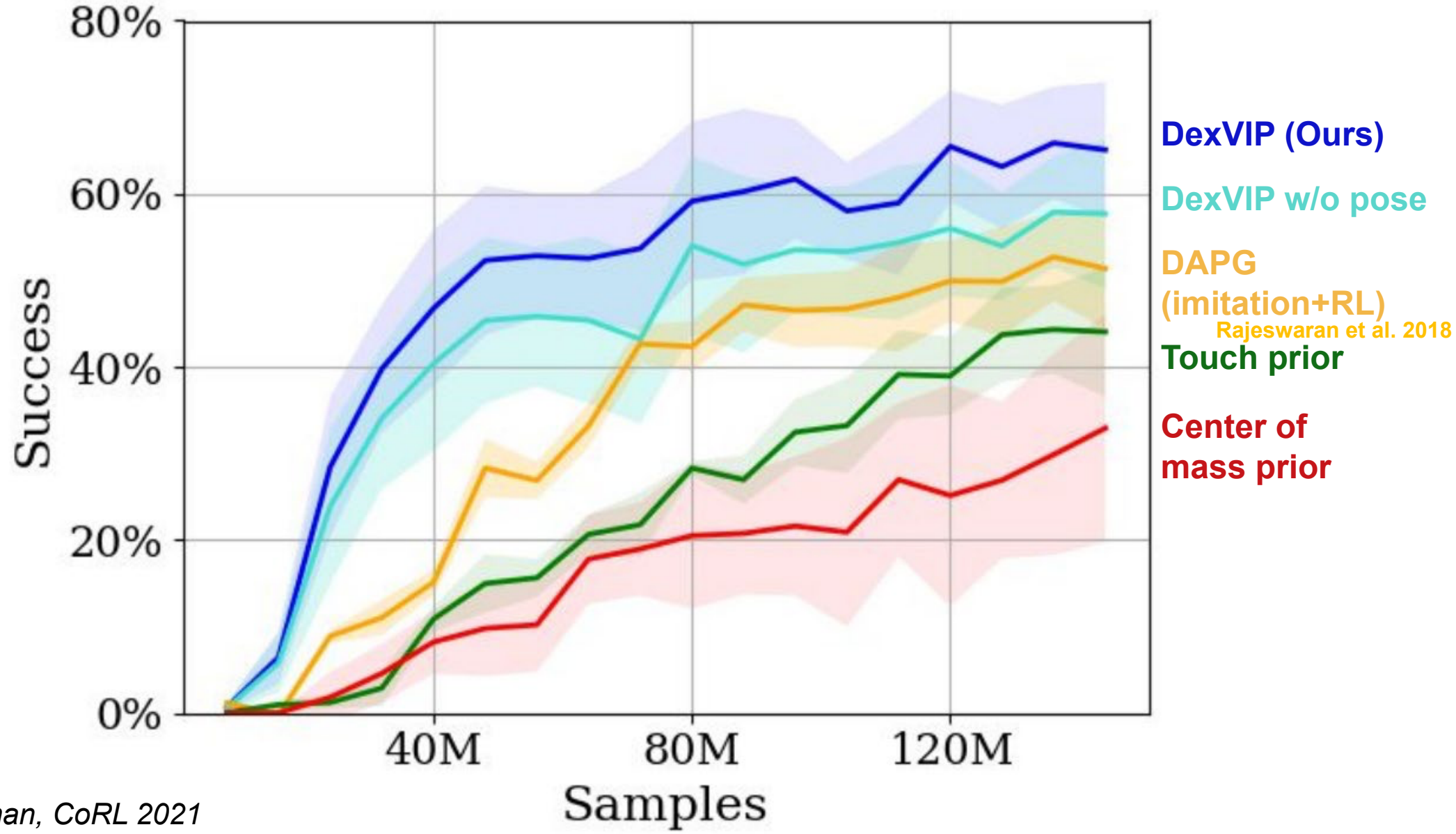
Mandikal & Grauman, ICRA 2021, CoRL 2021

Learning to grasp

Mandikal & Grauman, ICRA 2021



“People prior” from video accelerates learning



Mandikal & Grauman, CoRL 2021

First-person body pose

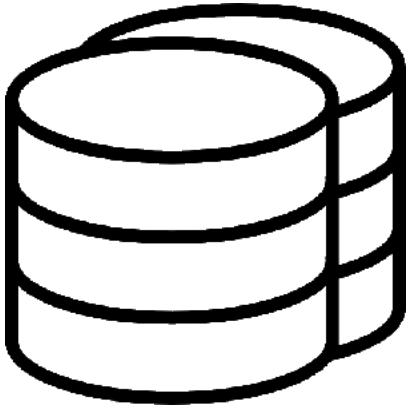
Infer 3D human body pose from the *outward* looking video

Wearable camera video

Inferred pose of camera wearer

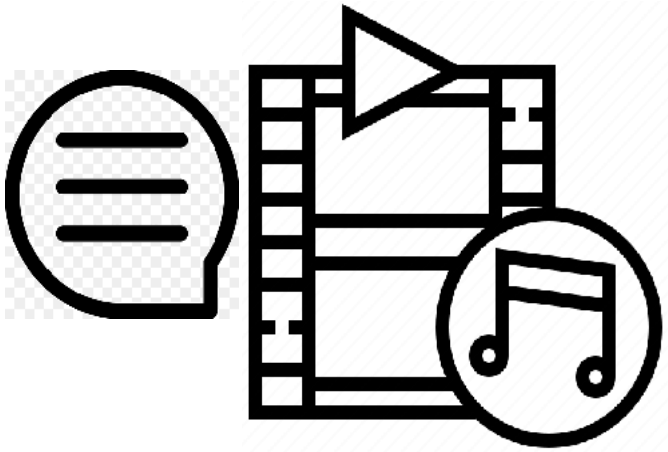
[Jiang & Grauman, CVPR 2017; Ng et al. CVPR 2020]

First-person vision: examples in research

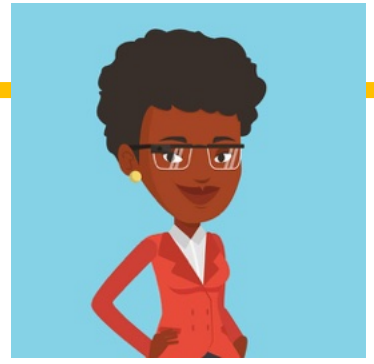


Data foundations

Affordances and manipulation

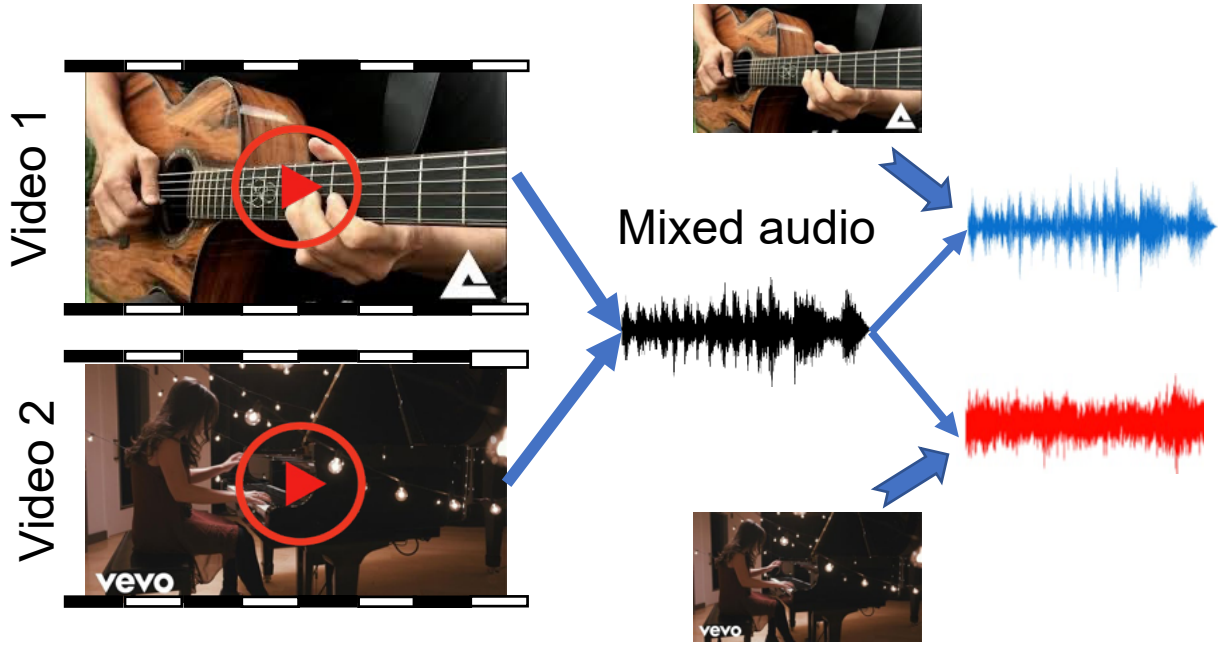


Multimodal learning



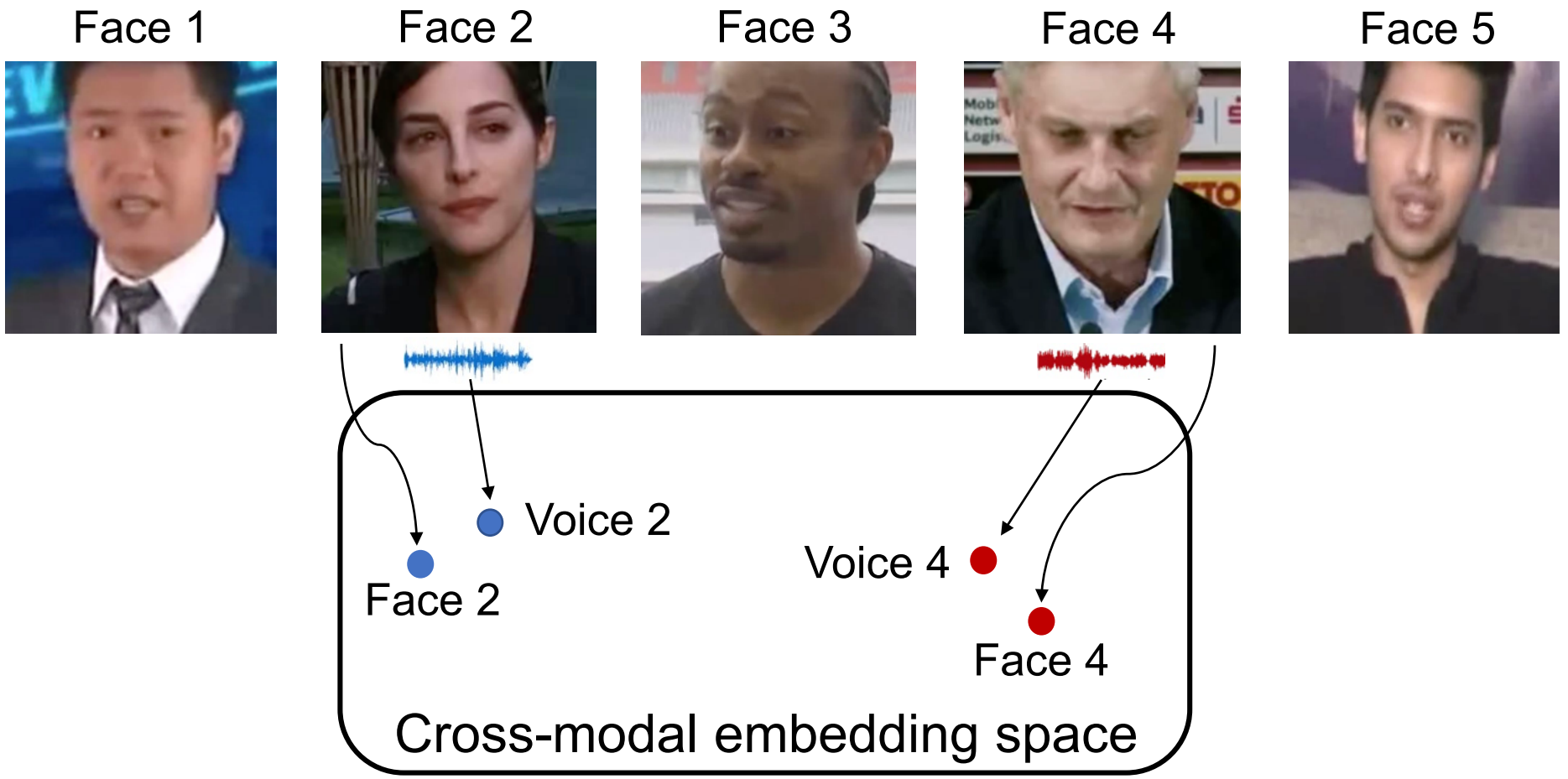
Cocktail-party problem

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

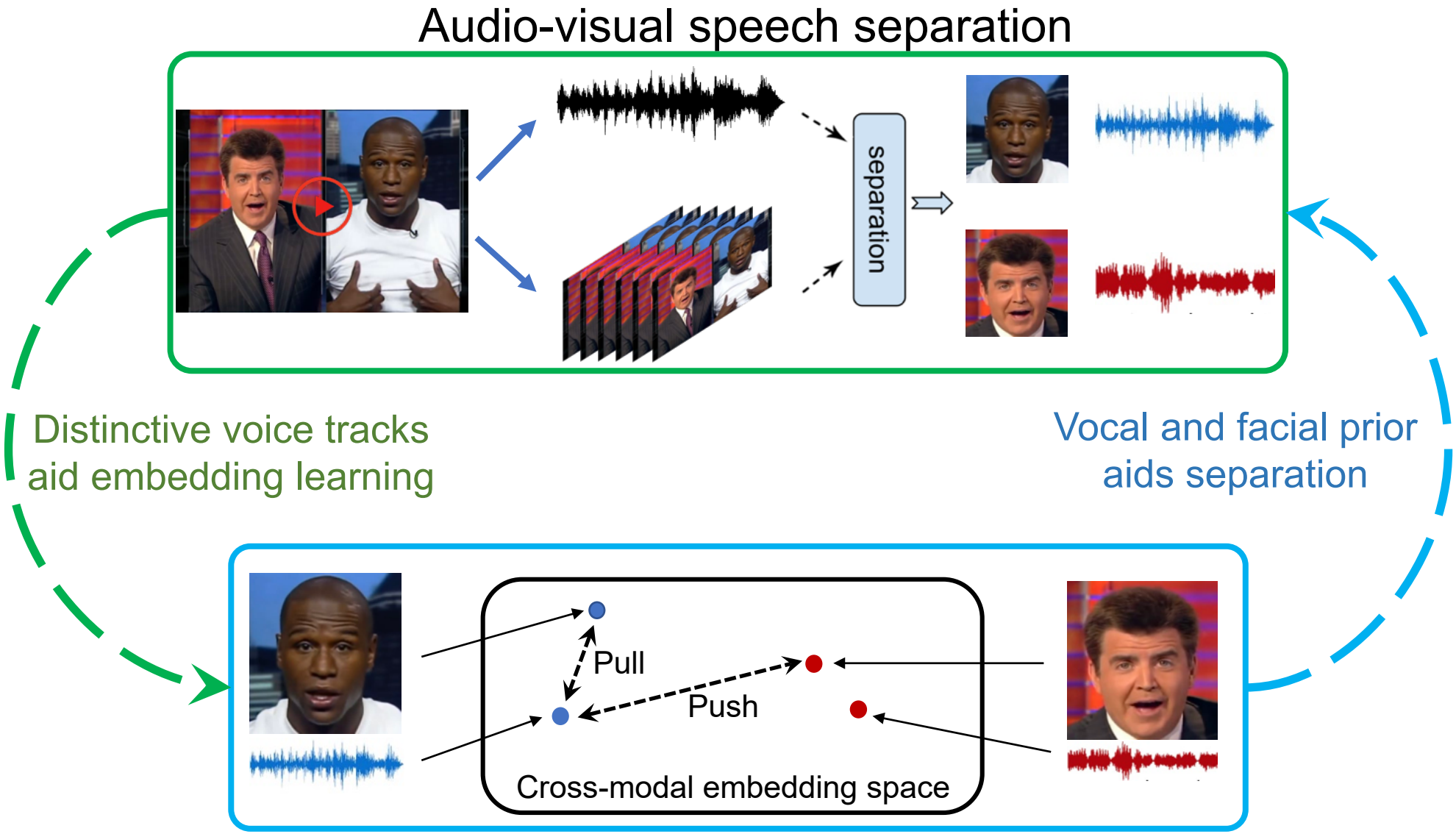
Facial appearance hints at voice qualities



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]


Enhanced speech
[VisualVoice, Gao et al. CVPR 2021]


Separated voice for the right speaker
[VisualVoice, Gao et al. CVPR 2021]


Separated voice for the left speaker
[VisualVoice, Gao et al. CVPR 2021]

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
grauman@cs.utexas.edu

Major successes in Computer Vision over the last decade, and

Major challenges ahead!

Increasingly blurred boundaries between vision, robotics, audio, language…

Embodied and first-person vision as a key frontier

Towards embodied multimodal first-person perception

- Ego4D: massive multimodal first-person data and benchmark
- Object affordances from video and influencing robot learning
- Audio-visual source separation to listen to voice of interest

Questions

1. **What is the difference between computer vision and computer graphics?**

A) Computer vision focuses on understanding the structure of the human visual system, whereas graphics focuses on understanding how the human visual system behaves.

B) Computer vision is the inverse of graphics. In vision, we observe the images but need to infer the underlying 3D content or semantics. In graphics, we have the 3D content but need to generate the image.

C) Computer vision attempts to build AI algorithms and computer graphics attempts to fool AI algorithms.

Questions

2. What are key factors that drove major advances in visual recognition over the last decade?

A) GPUs, large-scale labeled image datasets, large-scale annotation workforces

B) Faster networks, faster algorithms

C) Rapid new advances in brain modeling that inform new machine learning architectures

Questions

3. What is a first-person camera?

A) A camera that has never been used before

B) A camera worn on your head

C) A camera at close-field in surveillance settings