Penn
UNIVERSITY of PENNSYLVANIA

What does Embodied Intelligence mean?

Lessons Learned from Drone Racing

Antonio Loquercio

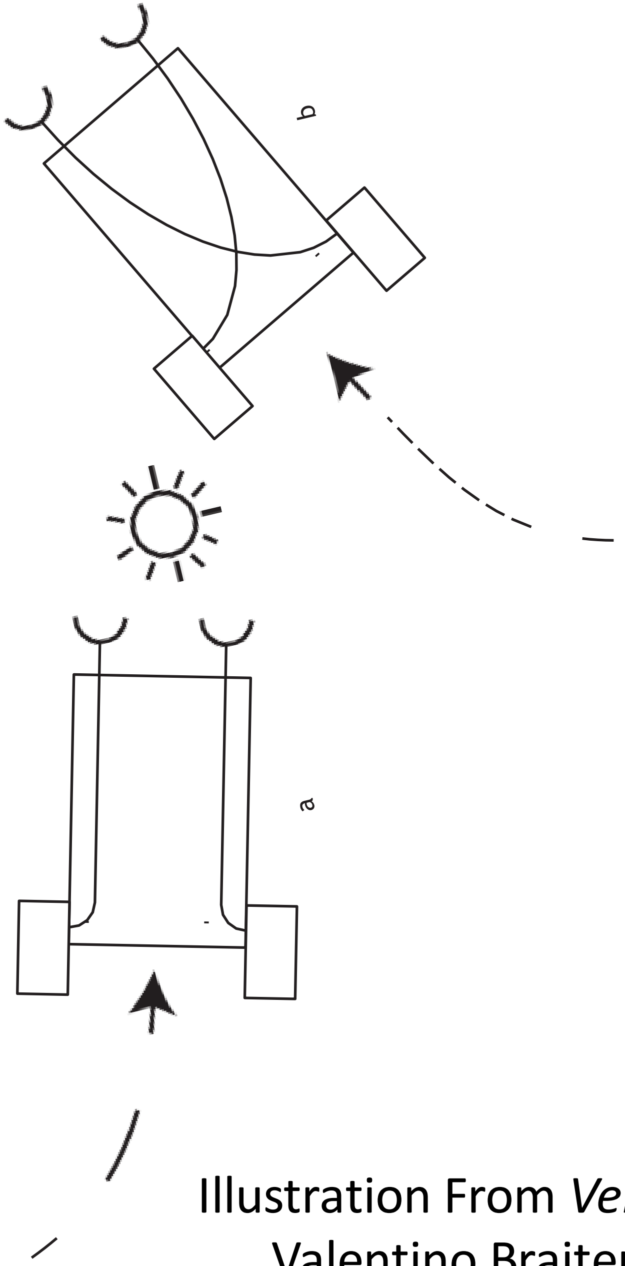
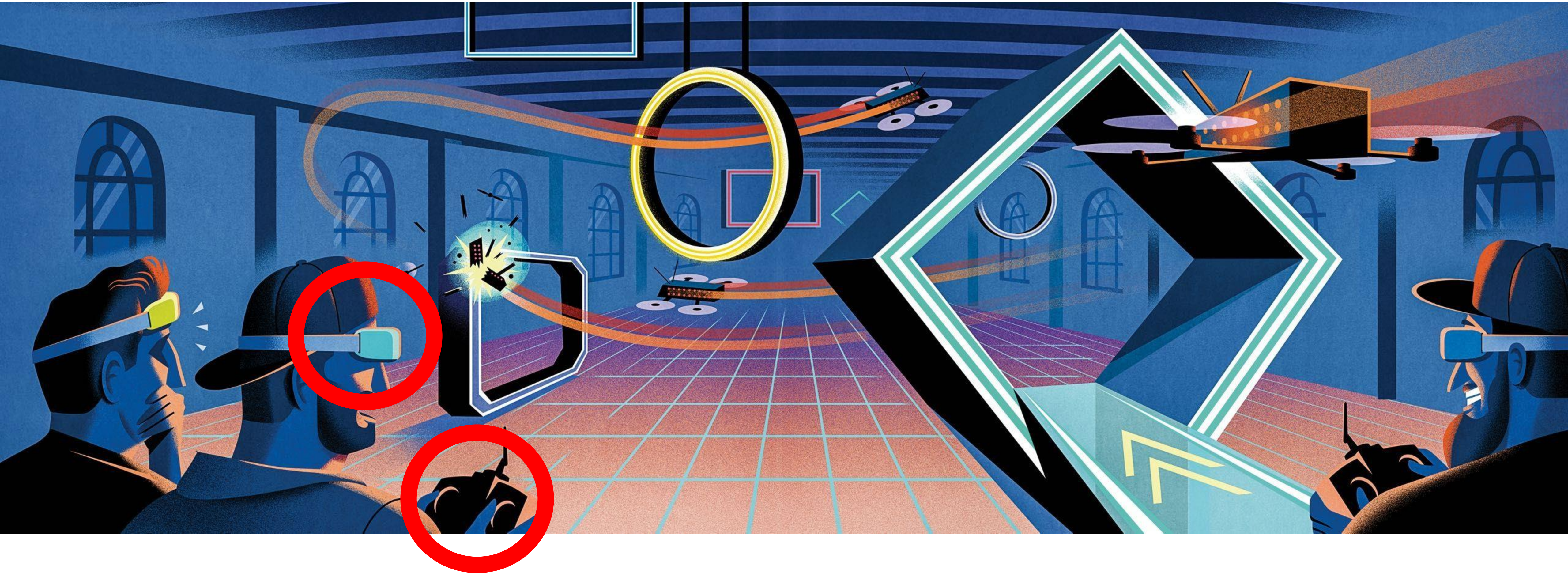


Illustration From *Vehicles* by Valentino Braitenberg



Source: The New Yorker

World Championship Qualifiers

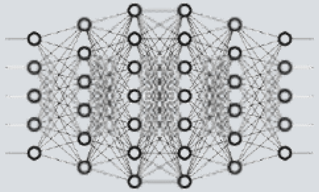
	Name	3 laps (seconds)
1	MinChan ‘MCKFPV’ Kim	27.057
2	Konstantin ‘KostaFPV’ Sonnentag	28.771
3	Levi ‘Leviathann’ Johnson	29.229056
4	Silas ‘Propsicle’ Aaron	29.329408
5	Marvin ‘MARV_FPV’ Schäpper	29.748
6	Mason ‘Hyper’ Lively	29.81888
7	Jacob ‘JakeHammer’ Capobres	30.010368
8	Evan ‘headsupfpv’ Turner	30.019584
9	Ashton ‘Drobotracer’ Gamble	30.400992
10	Sebastian ‘SebaFPV’ Espinal	30.44

~2s difference

~1s difference

Racing is not a good fit for Imitation Learning

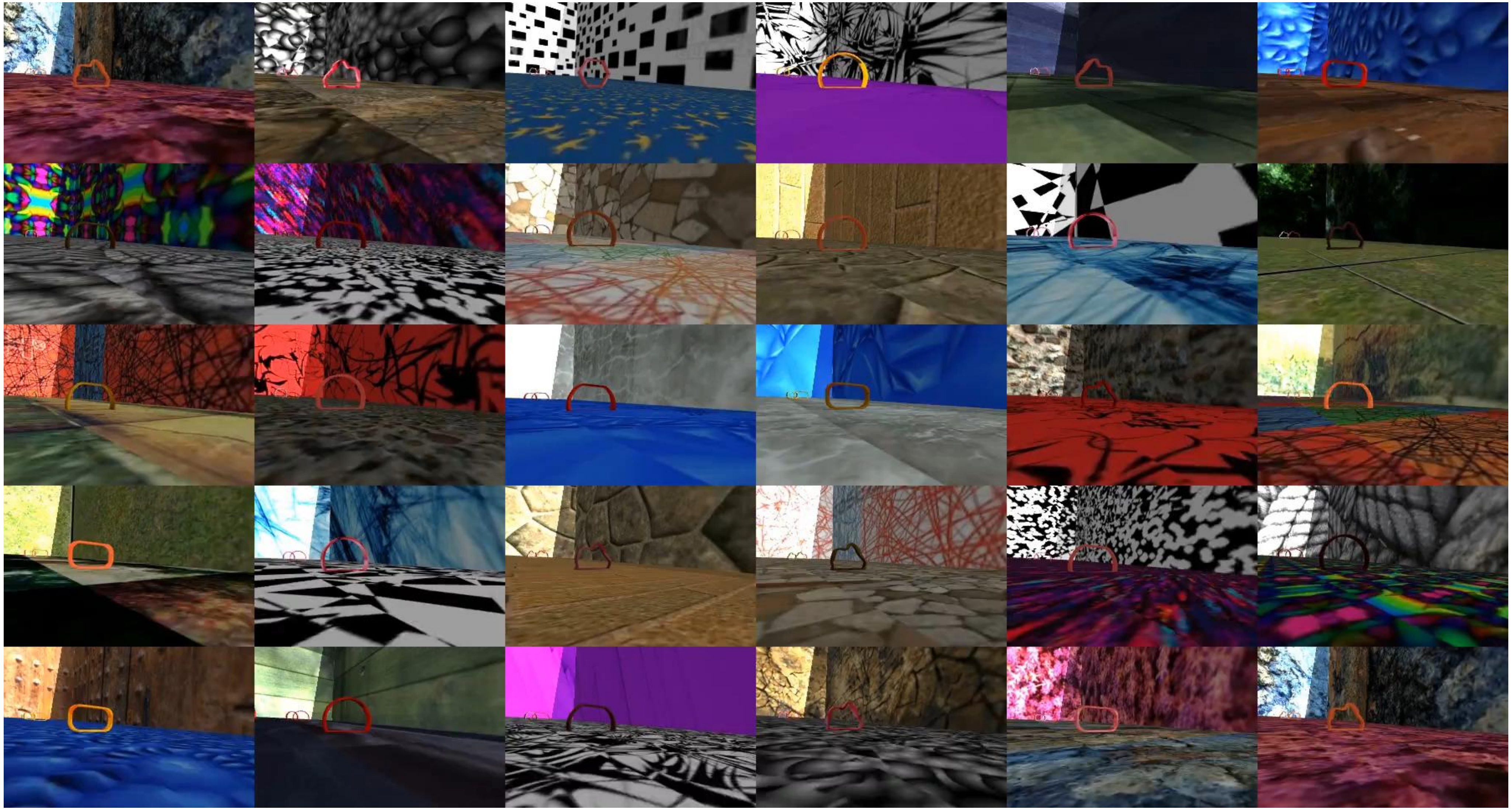
Body Rates
Thrust



Neural Networks

Body Rates
Thrust

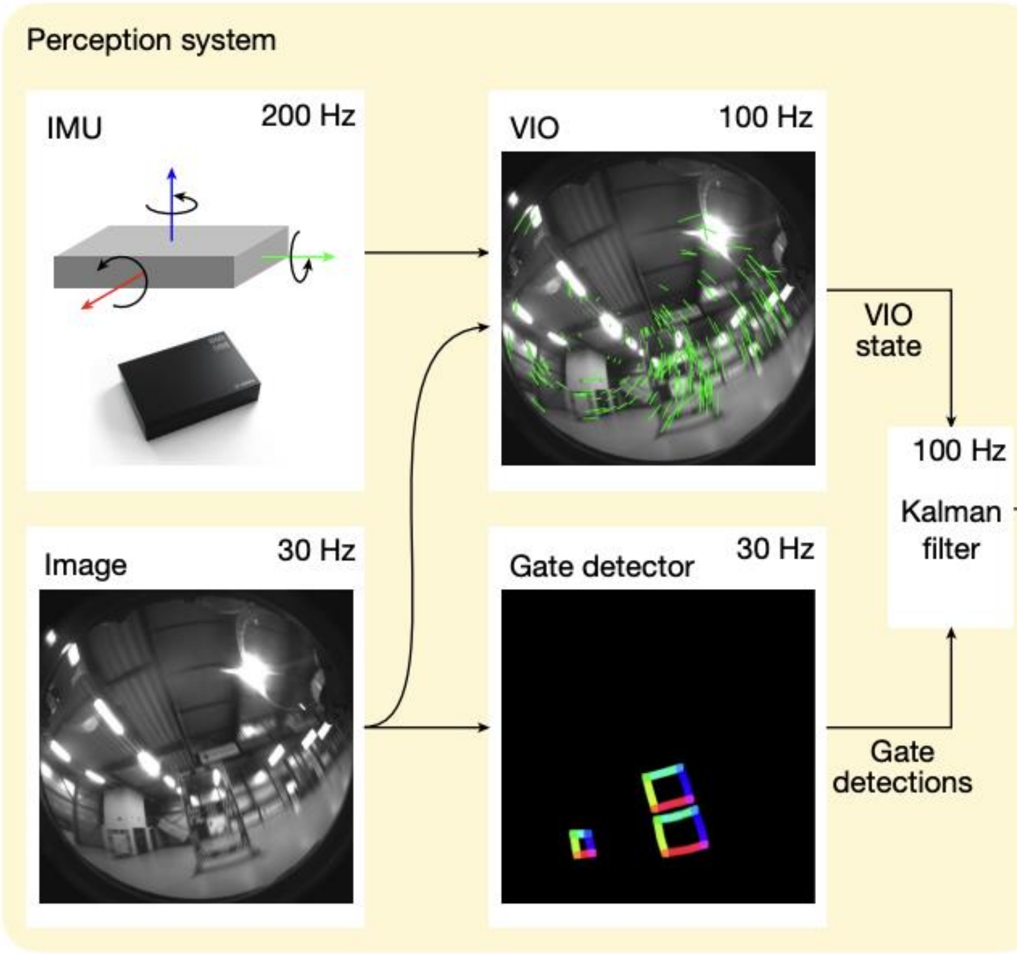
Learning End-To-End Control For Drone Racing



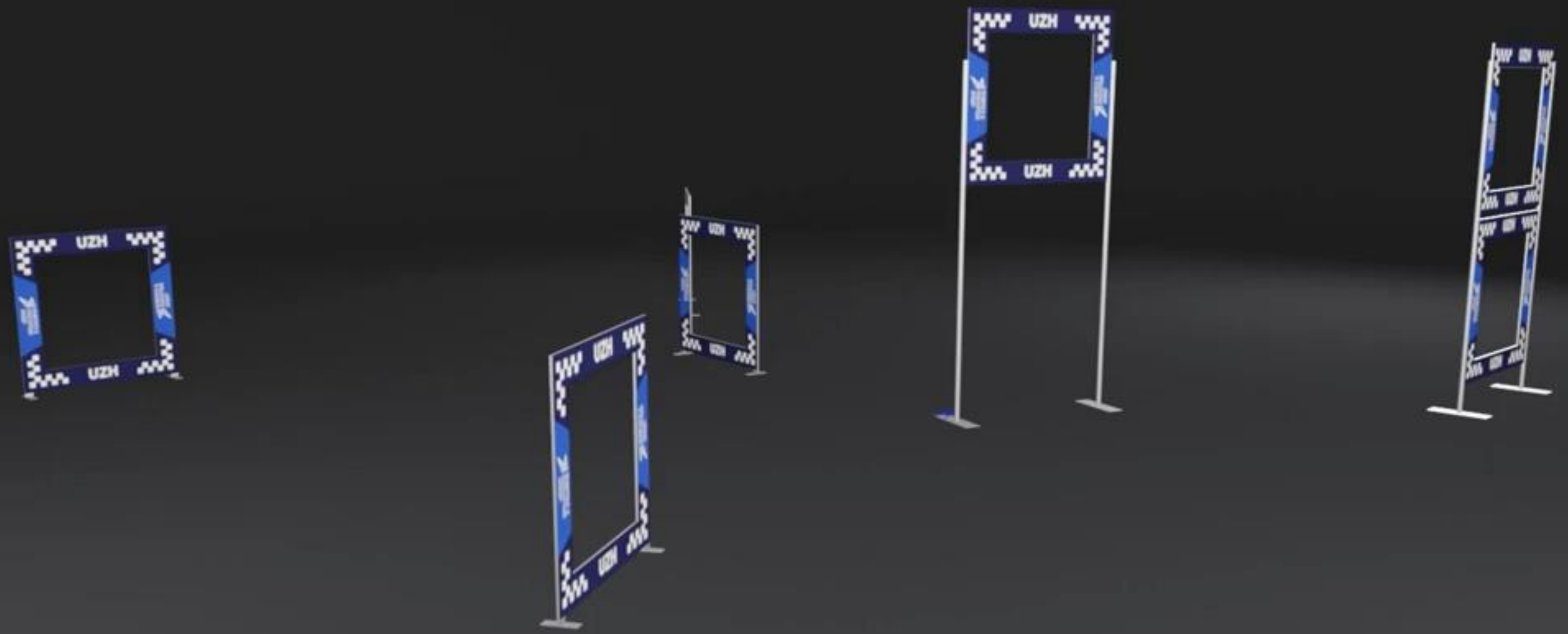
Deep Drone Racing: From Simulation to the Real World Using Domain Randomization. Loquercio et al.

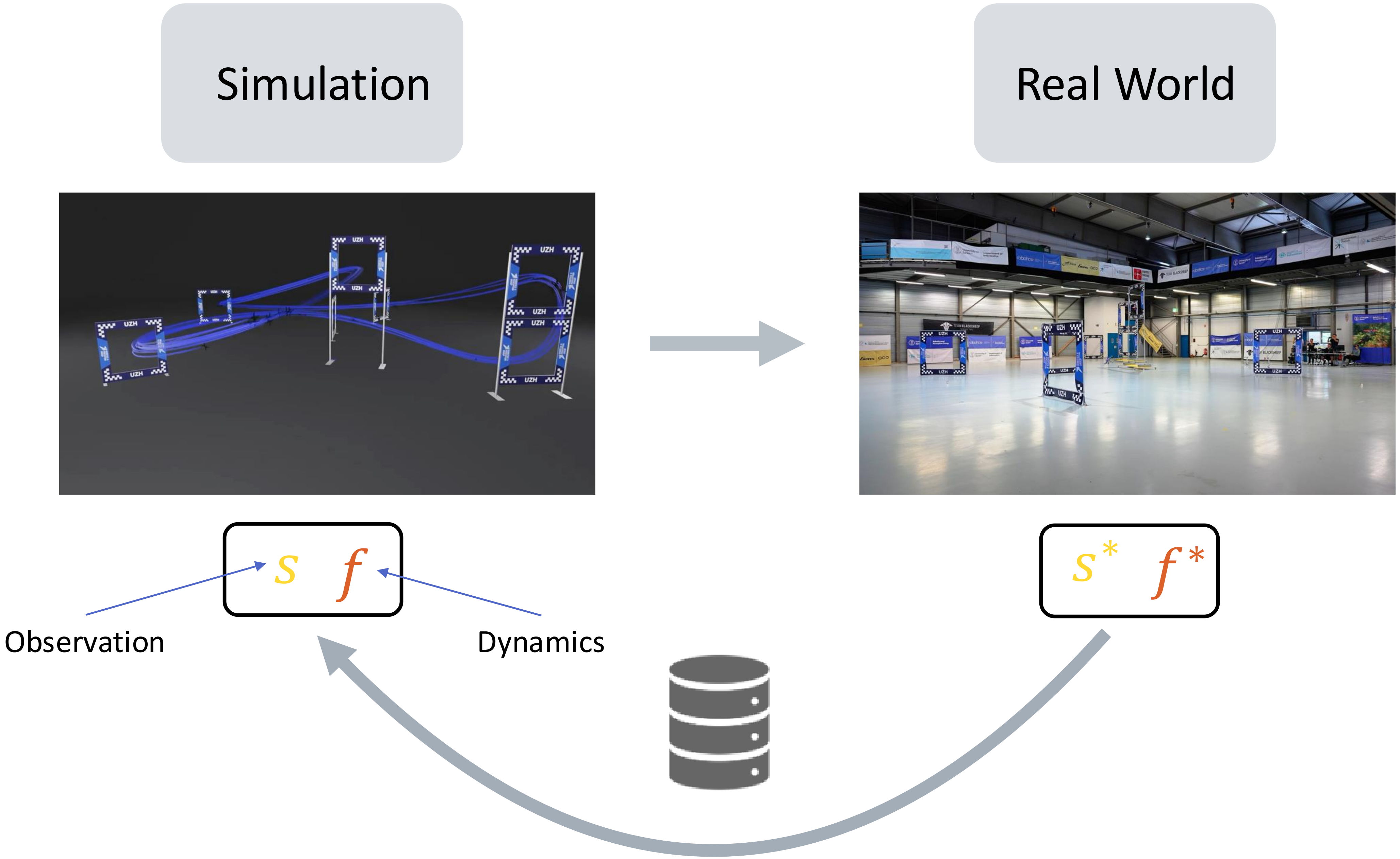
T-RO Best Paper Honorable Mention

A Modular Approach



Training




Simulation
Real World
s f
s^* f^*
Observation
Dynamics


ROBOTICS & PERCEPTION GROUP
Universität Zürich UZH
Robotics and Perception Group
robotics
1
2
Innovation Zurich
Canton of Zurich
Department for Economic Affairs
Office for Economy and Labour
UZH

Making the comparison as fair as possible

- The same drone.
- Compensation for human perception latency at the start.

But

- We use an onboard inertial measurement unit (IMU). But our camera updates only at 30Hz (120Hz for humans).
- We have lower latency (40ms vs ~200ms for humans). Unclear if that matters since the environment is predictable.

Statistics of Racing against Professional Pilots

Head-to-Head Racing Results

	Number of Races	Best Time-to-Finish	Wins	Losses	Win Ratio
A. Vanover vs. Swift	9	17.956 s	4	5	0.44
T. Bitmatta vs. Swift	7	18.746 s	3	4	0.43
M. Schaepper vs. Swift	9	21.160 s	3	6	0.33
Swift vs Human Pilots	25	**17.465 s**	15	10	**0.60**

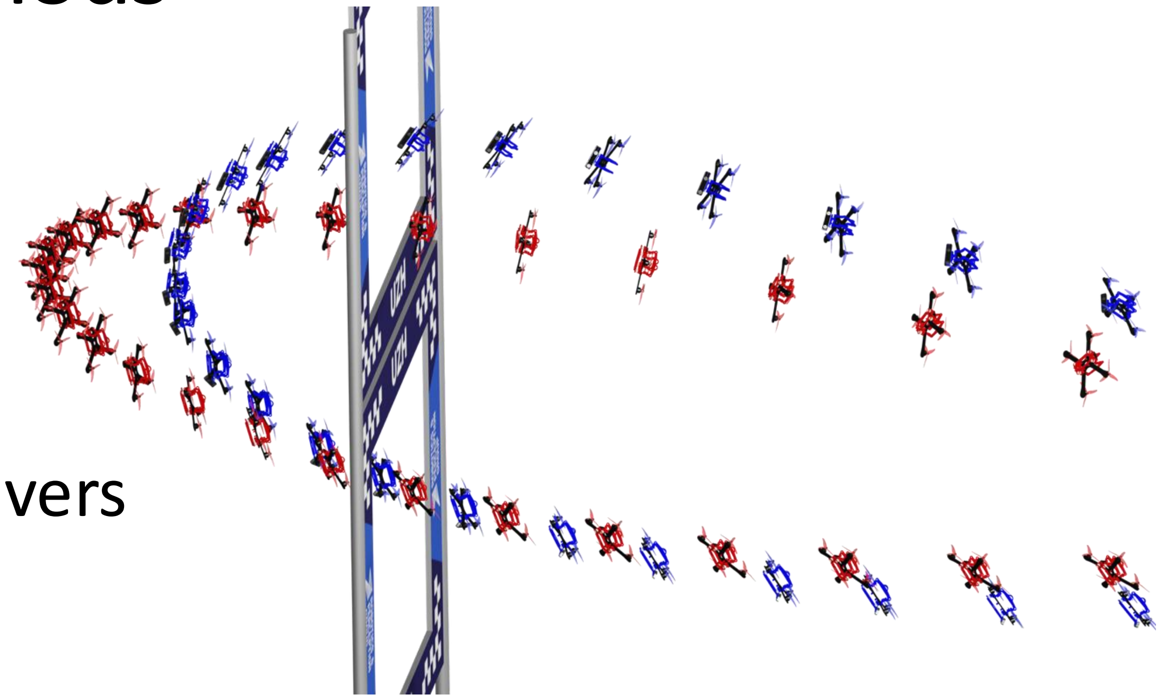
Differences Human vs. Autonomous

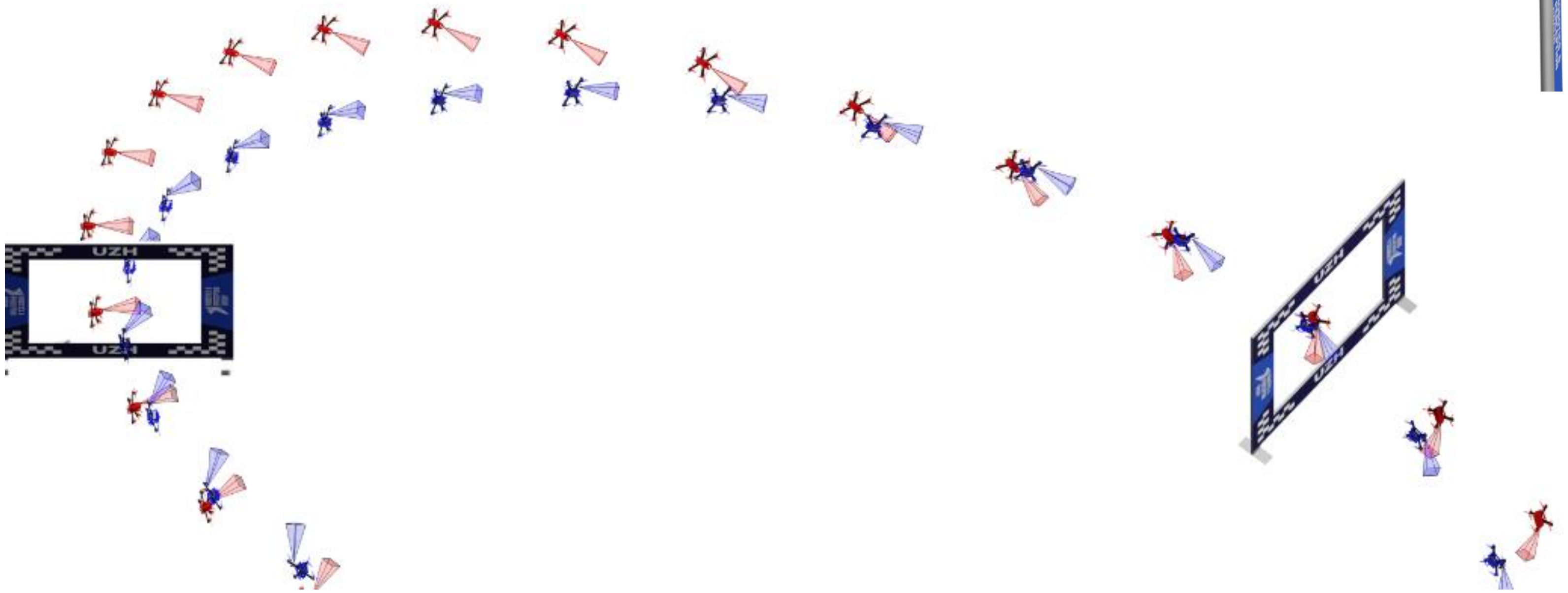
The Autonomous Drone …

… does not always fly faster

… is faster at the start

… takes a tighter path in difficult maneuvers





Human

Autonomous

Article | Open access | Published: 30 August 2023

Champion-level drone racing using deep reinforcement learning

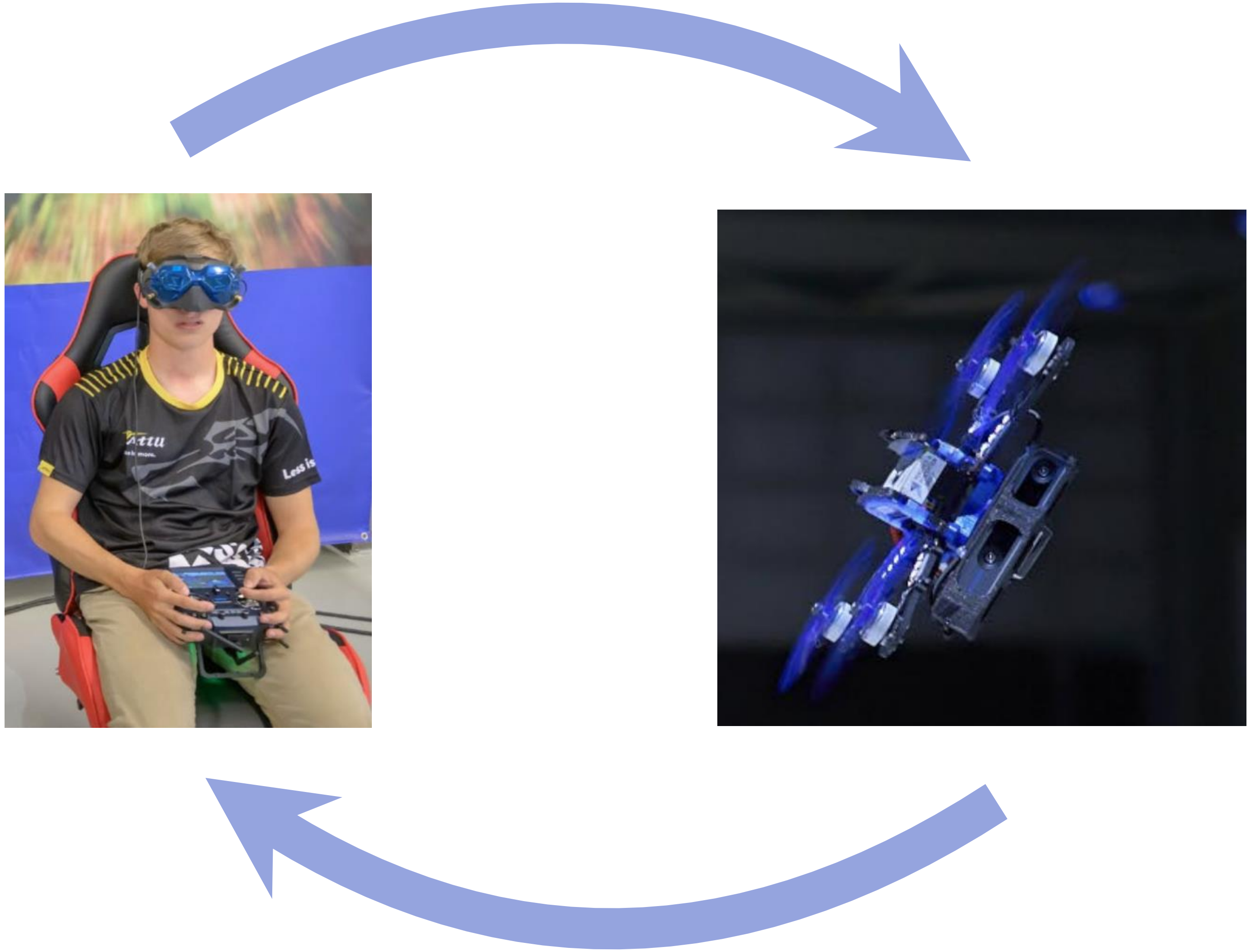
Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun & Davide Scaramuzza

Nature **620**, 982–987 (2023) | Cite this article

The Human Champions


Universität Zürich
Robotics and Perception Group

My Definition of Embodied Intelligence



How to get there?

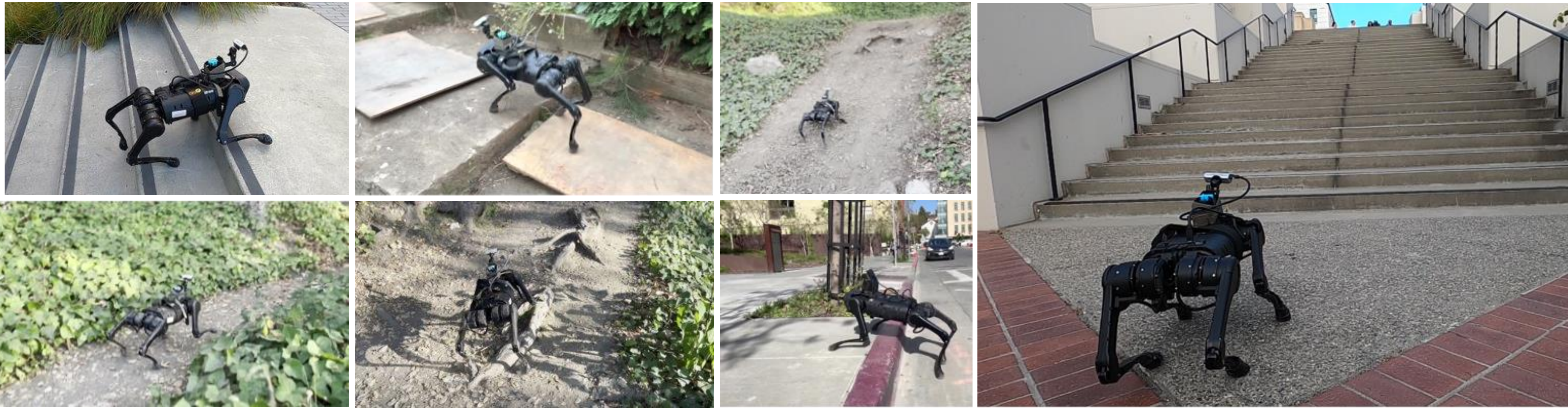
- “Collect a lot of teleoperation data”
- .
- .
- .
- .
- .
- “Tune costs/rewards”

How to get there?

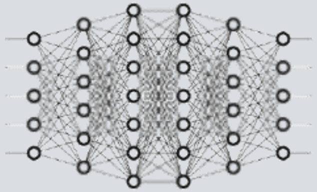
- “Collect a lot of teleoperation data.”
- .
- .
- “Learn to predict the world.” (akin to self-supervised learning)
- .
- .
- “Tune costs/rewards”

Learning Visual Locomotion with Cross-Modal Supervision

Loquercio A., Kumar A., Malik J.

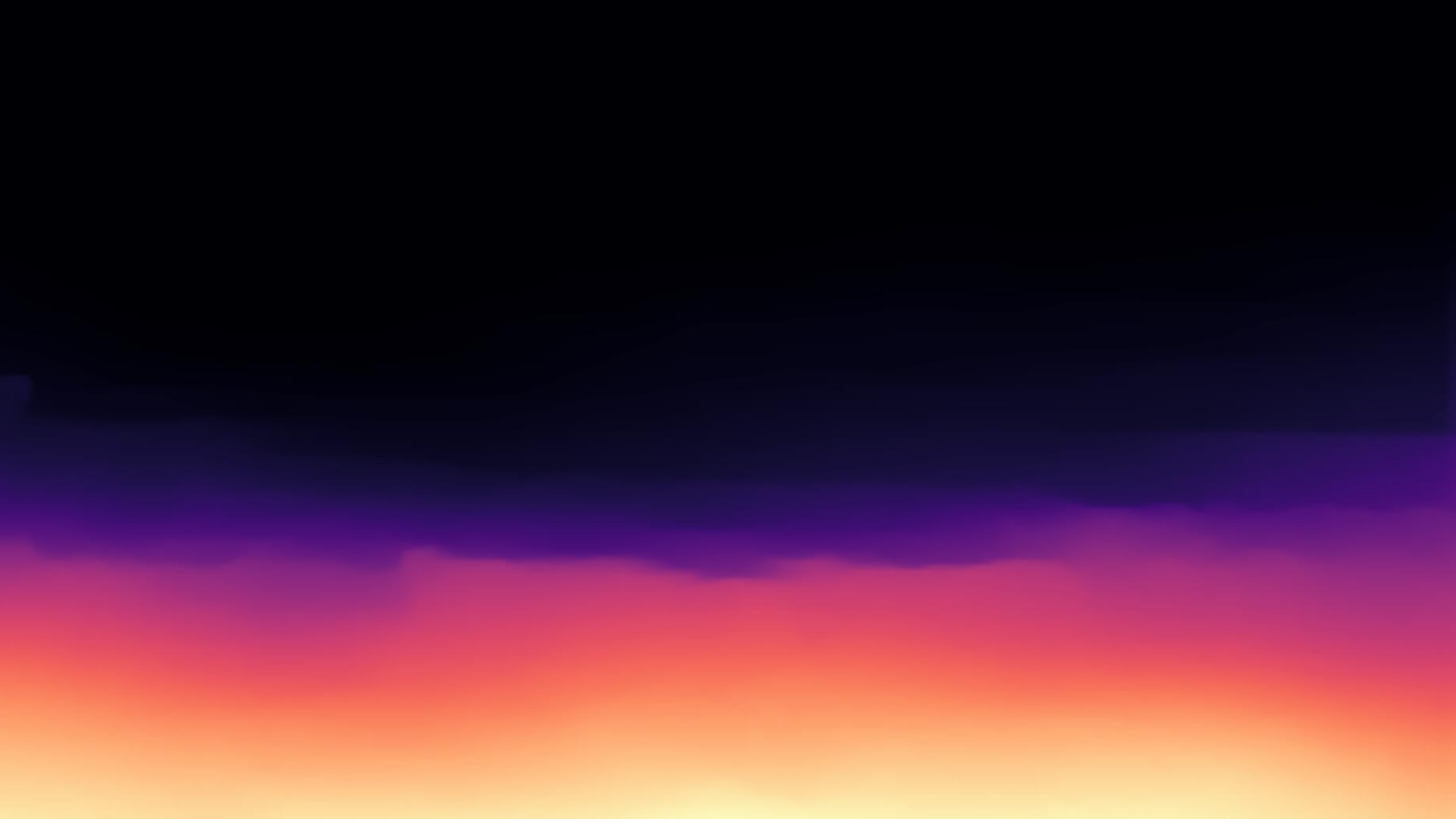


Previous Work on Vision-Based Locomotion

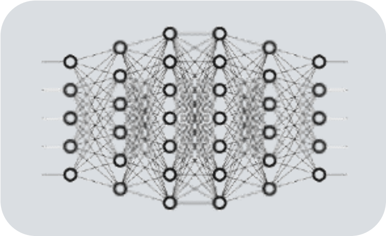


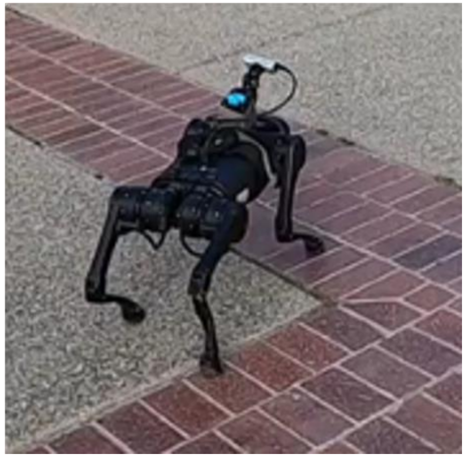
Neural Networks

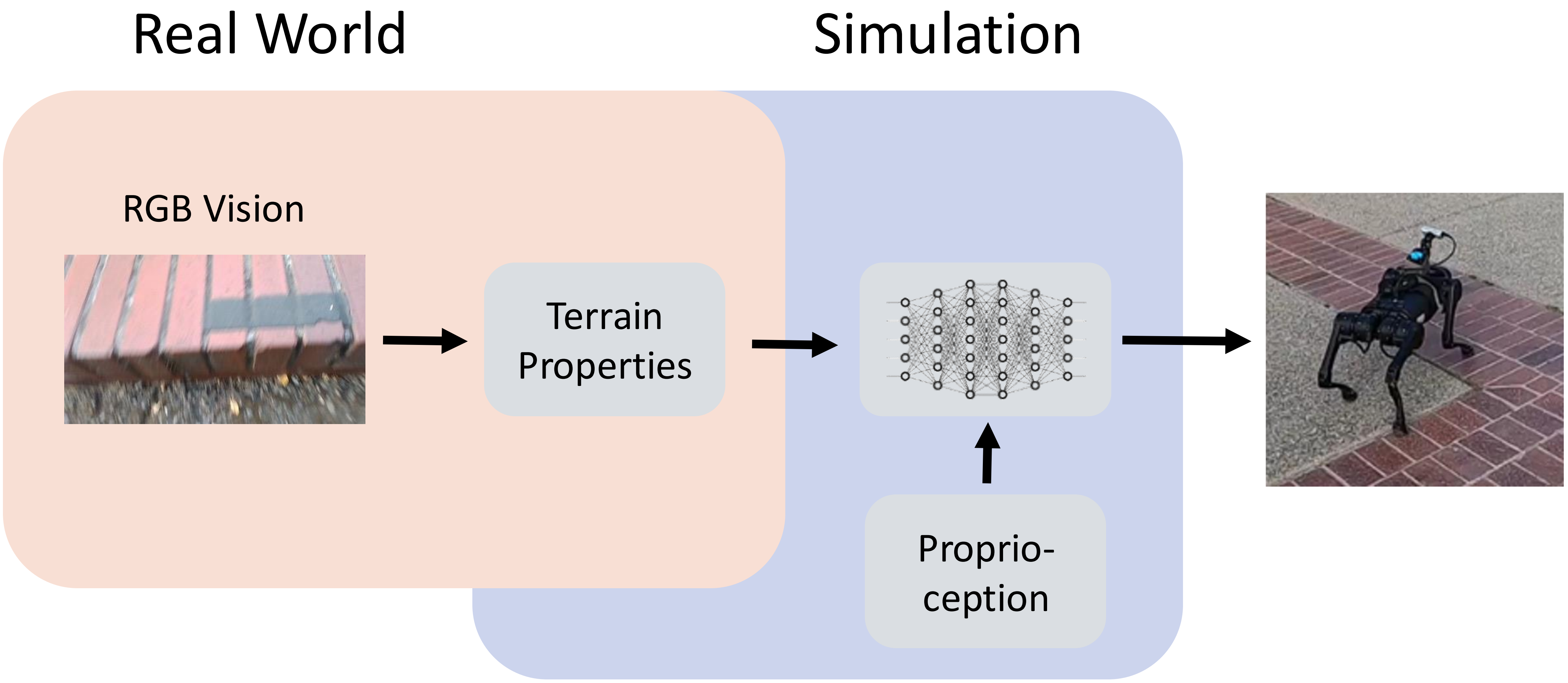
Actions



RGB Vision







Hwangbo et al., 2019
Lee et al., 2020
Kumar et al., 2020

RGB Vision

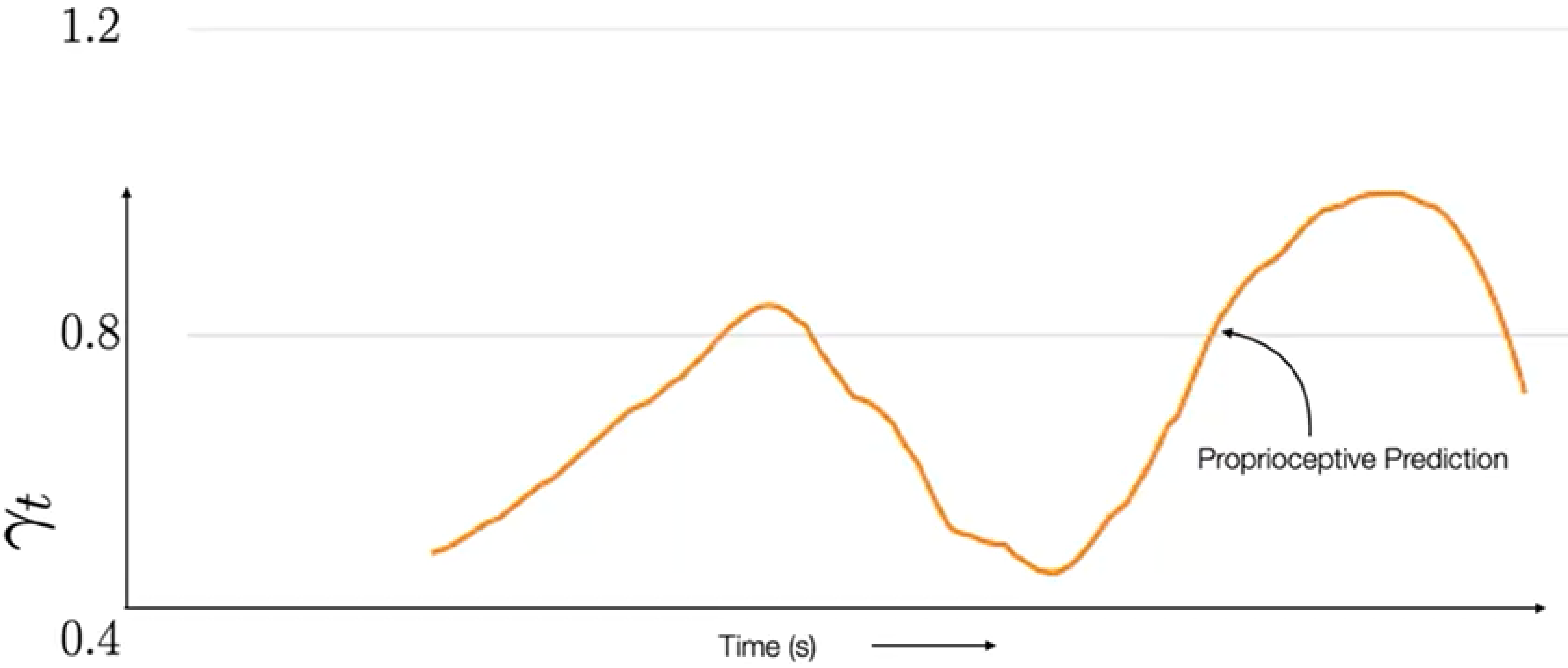
Terrain Properties

How do we train this estimator?

1. We can’t use existing datasets
2. Humans can’t provide annotations

Proprioception to Estimate Terrain Properties

Cross-Modal Supervision



Blind

Loquercio et. al, ICRA, 2023

Vision-Based

Loquercio et. al,
ICRA, 2023

Day 1 (2X)

Loquercio et. al,
ICRA, 2023

Loquercio et. al,
ICRA, 2023

Discrete Terrain

Loquercio et. al,
ICRA, 2023

Construction Zone

Loquercio et. al,
ICRA, 2023

Loquercio et. al,
ICRA, 2023

Visual Plasticity

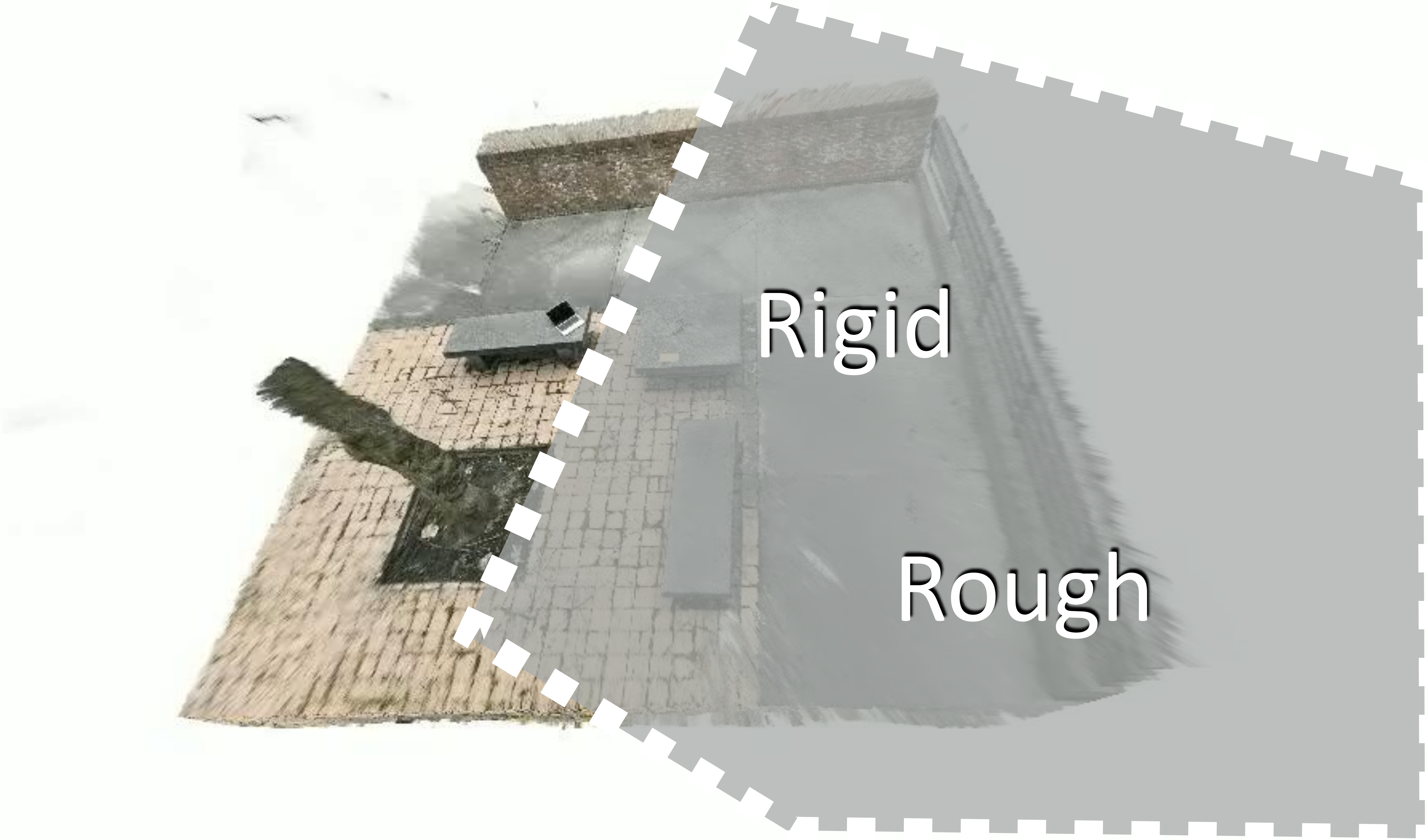
Before Adaptation

After 1min of data

Takeaways

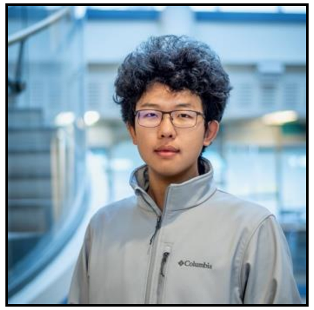
- Use a self-supervised loss (predict one sensor from the other) to recover from failures and/or adapt to novel conditions.

- Interaction is a tool to learn about the environment.


Rigid
Rough


Soft
Crumbly

Hearing Hands:
Generating Sounds from Physical Interactions in 3D Scenes


Yiming Dou


Wonseok Oh


Yuqing Luo

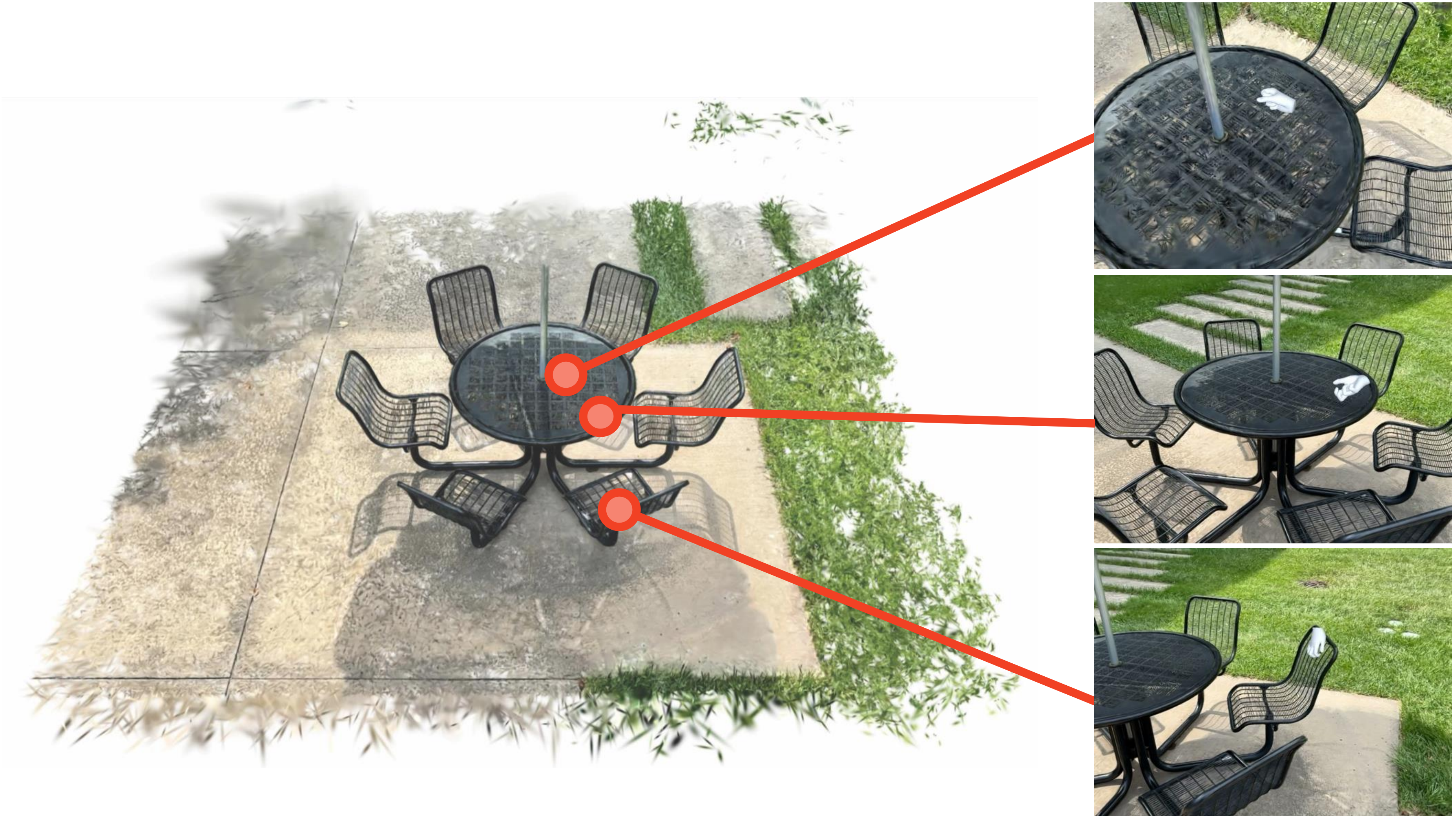

Antonio Loquercio


Andrew Owens

CVPR Nashville JUNE 11-15, 2025

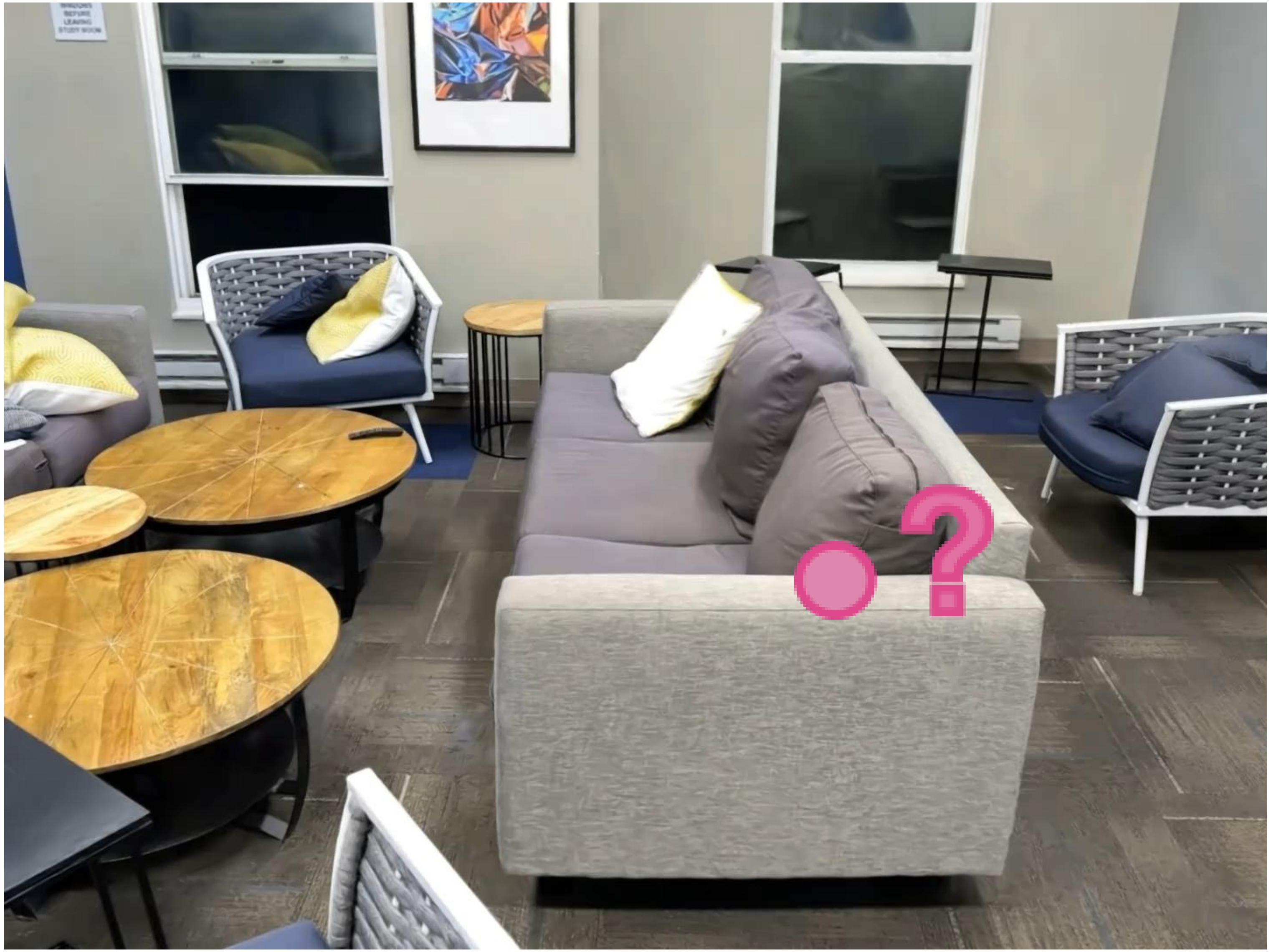
Poster #151, Fri 10:30-12:30
(poster session 1)

Predicting the sound of actions



Predicting the Sound of Actions

- **Step 1**: Pick a location to interact with in a 3D scene



Predicting the Sound of Actions

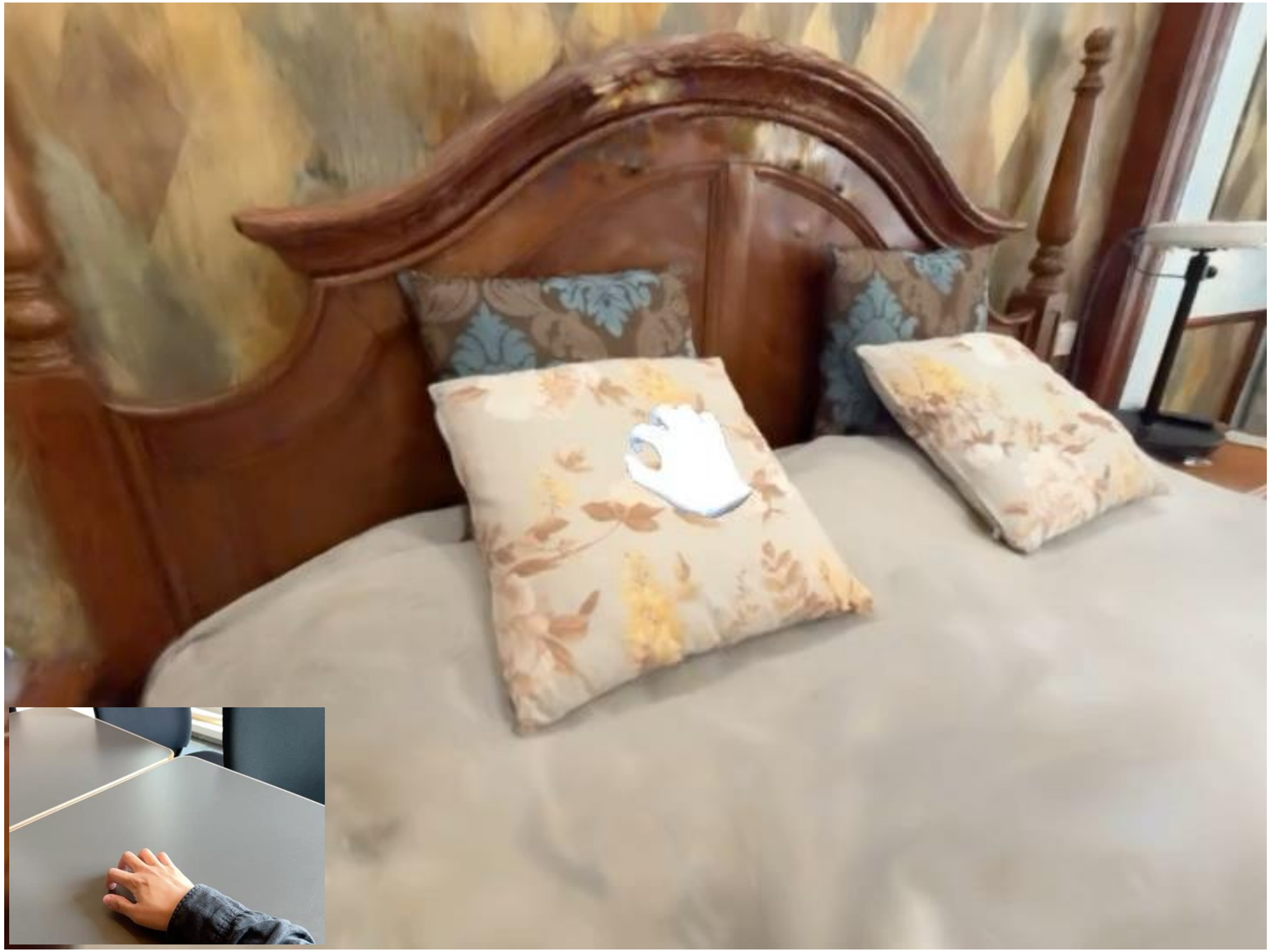
- **Step 1**: Pick a location to interact with in a 3D scene
- **Step 2**: Record the desired hand motion

Predicting the Sound of Actions

- **Step 1**: Pick a location to interact with in a 3D scene
- **Step 2**: Record the desired hand motion
- **Step 3**: Generate synthetic interaction sound

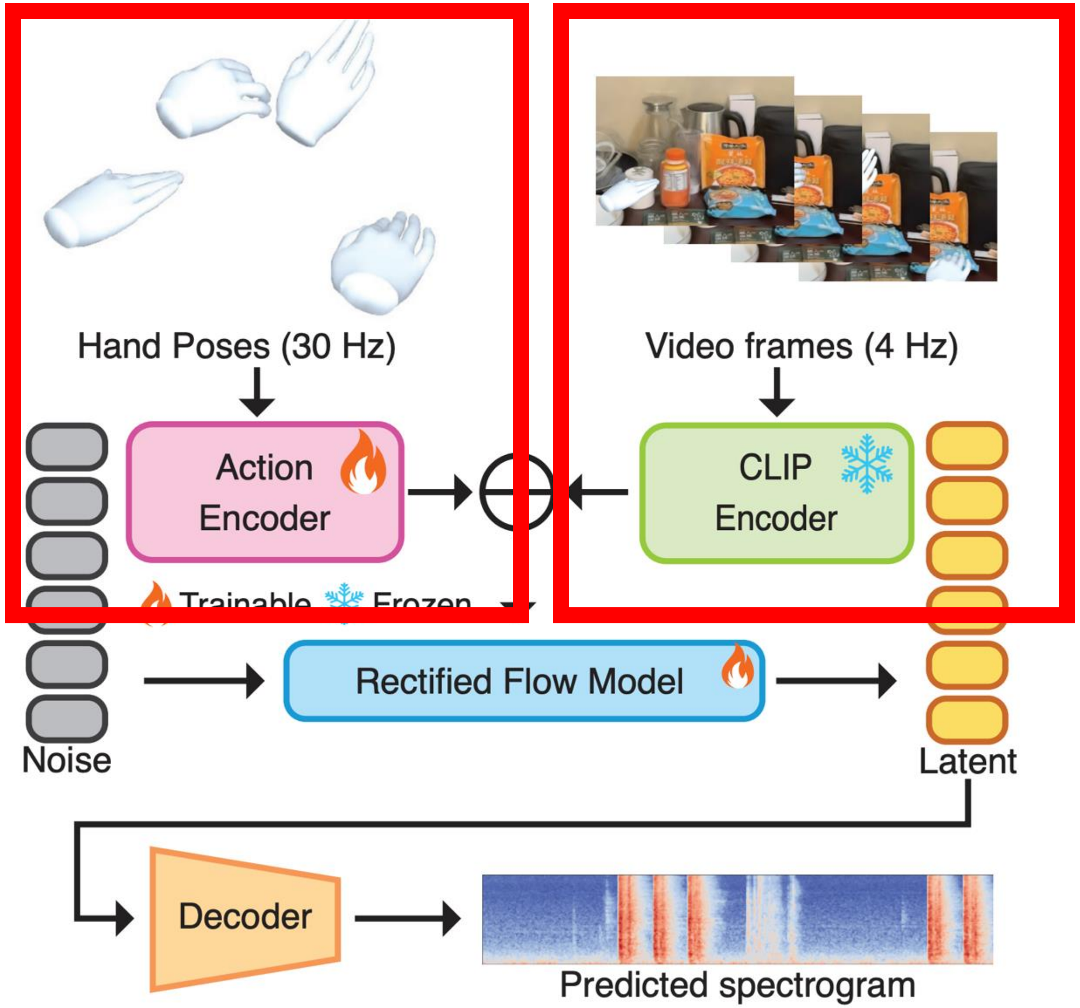
Predicting the Sound of Actions

Predicting the Sound of Actions



Predicting the Sound of Actions

Sound generation model



A Dataset of Hand-Generated Sounds

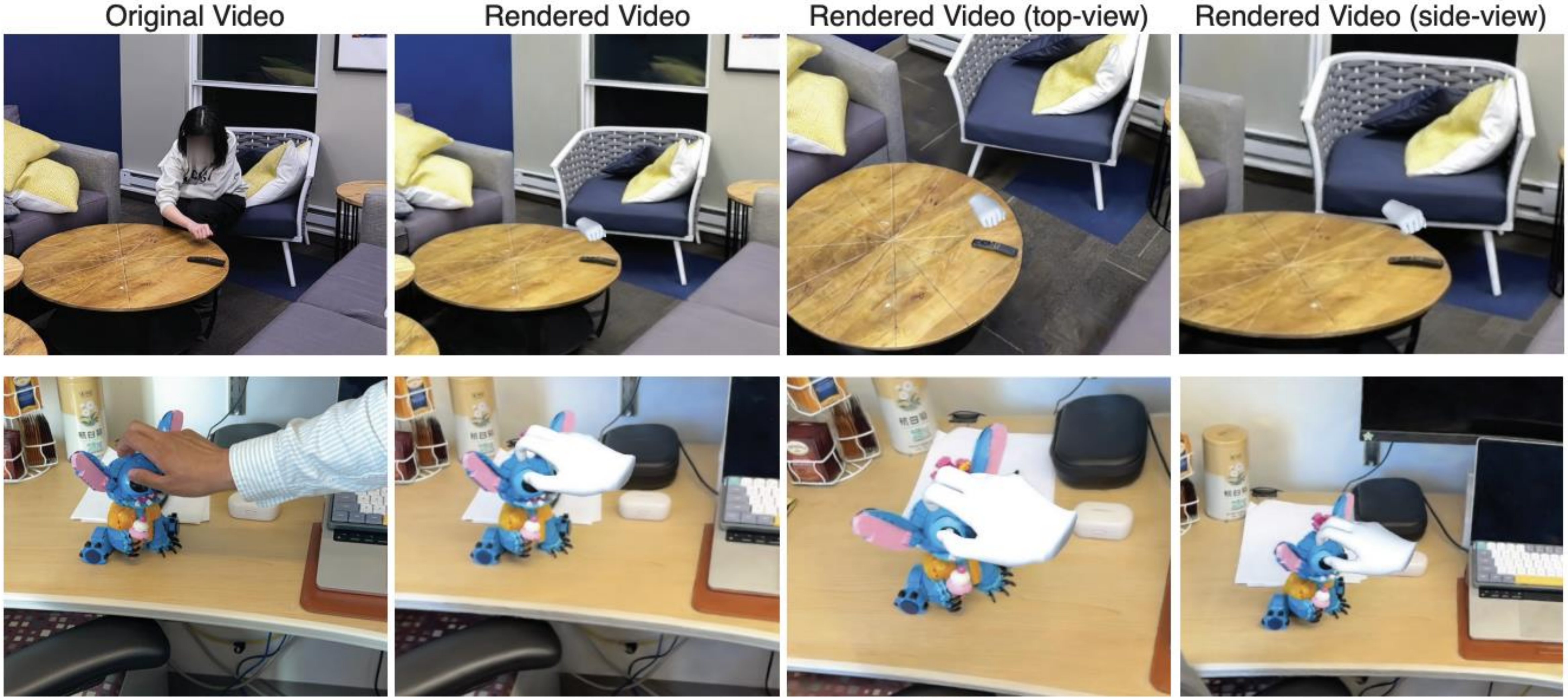
Register to the existing reconstruction

A Dataset of Hand-Generated Sounds

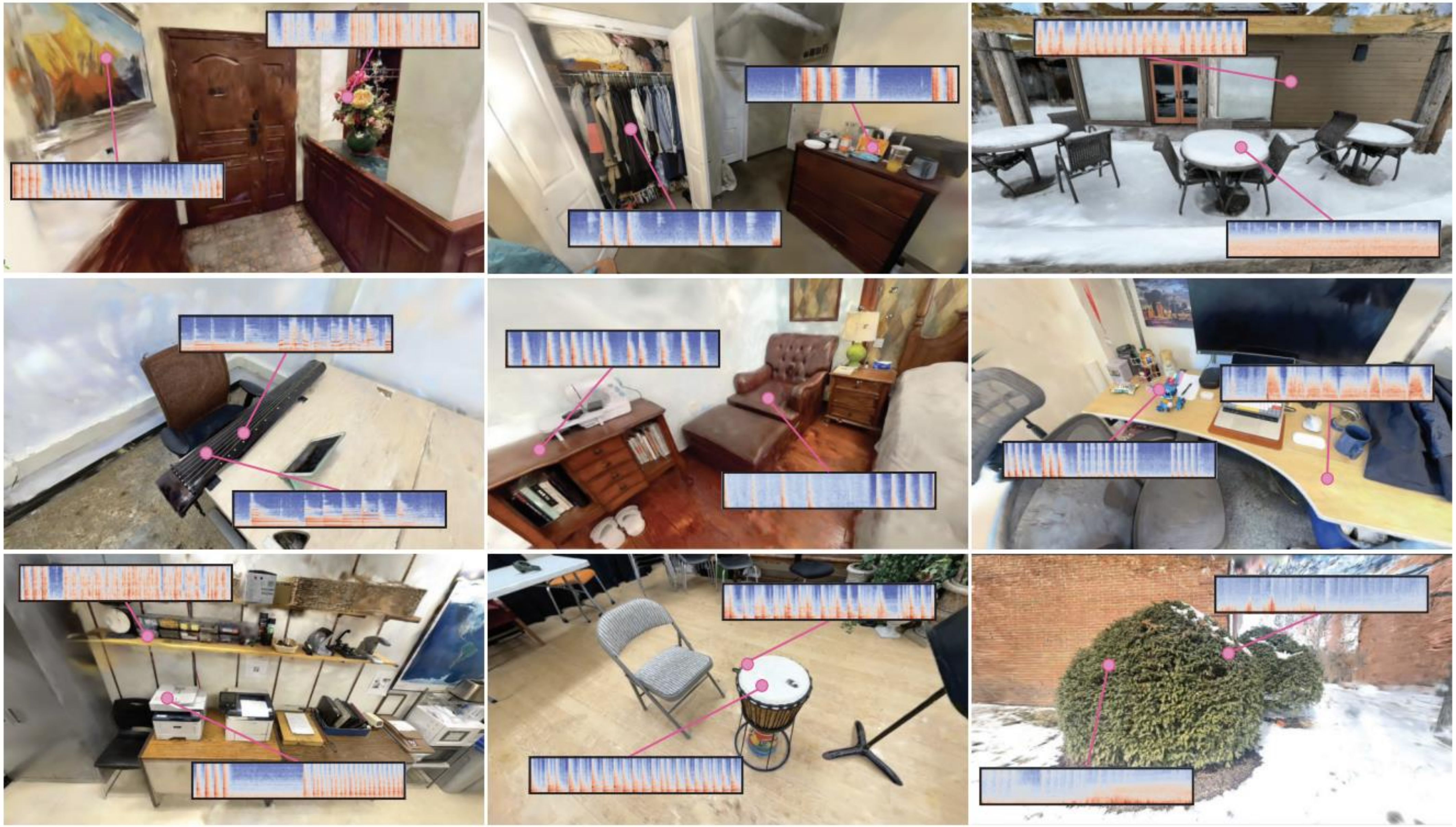
Register to the
existing reconstruction

A Dataset of Hand-Generated Sounds

A Dataset of Hand-Generated Sounds



A Dataset of Hand-Generated Sounds



Let’s Play a Game

Which one is generated?

Real

Generated

Which one is generated?

Real

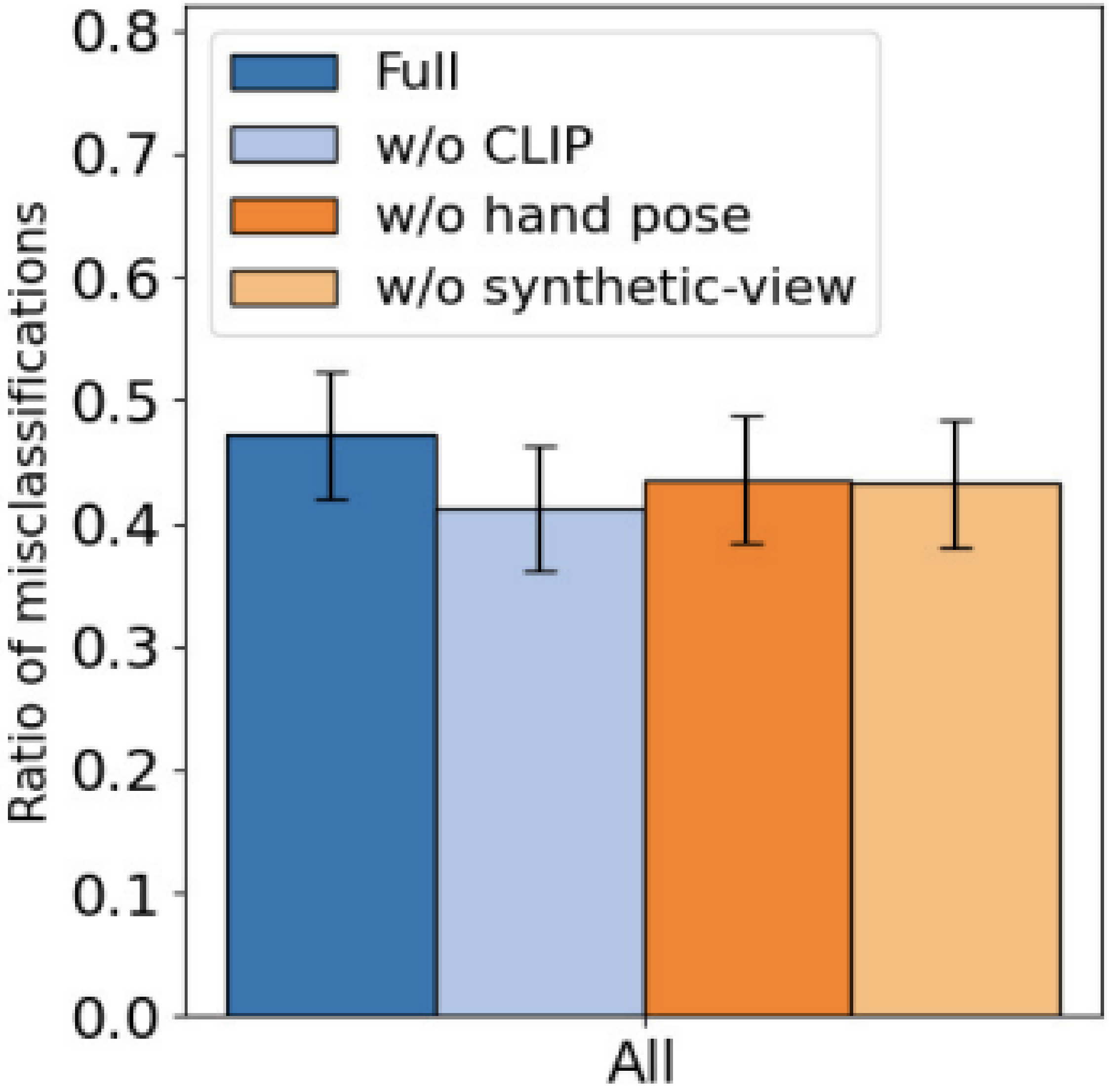
Generated

Which one is generated?

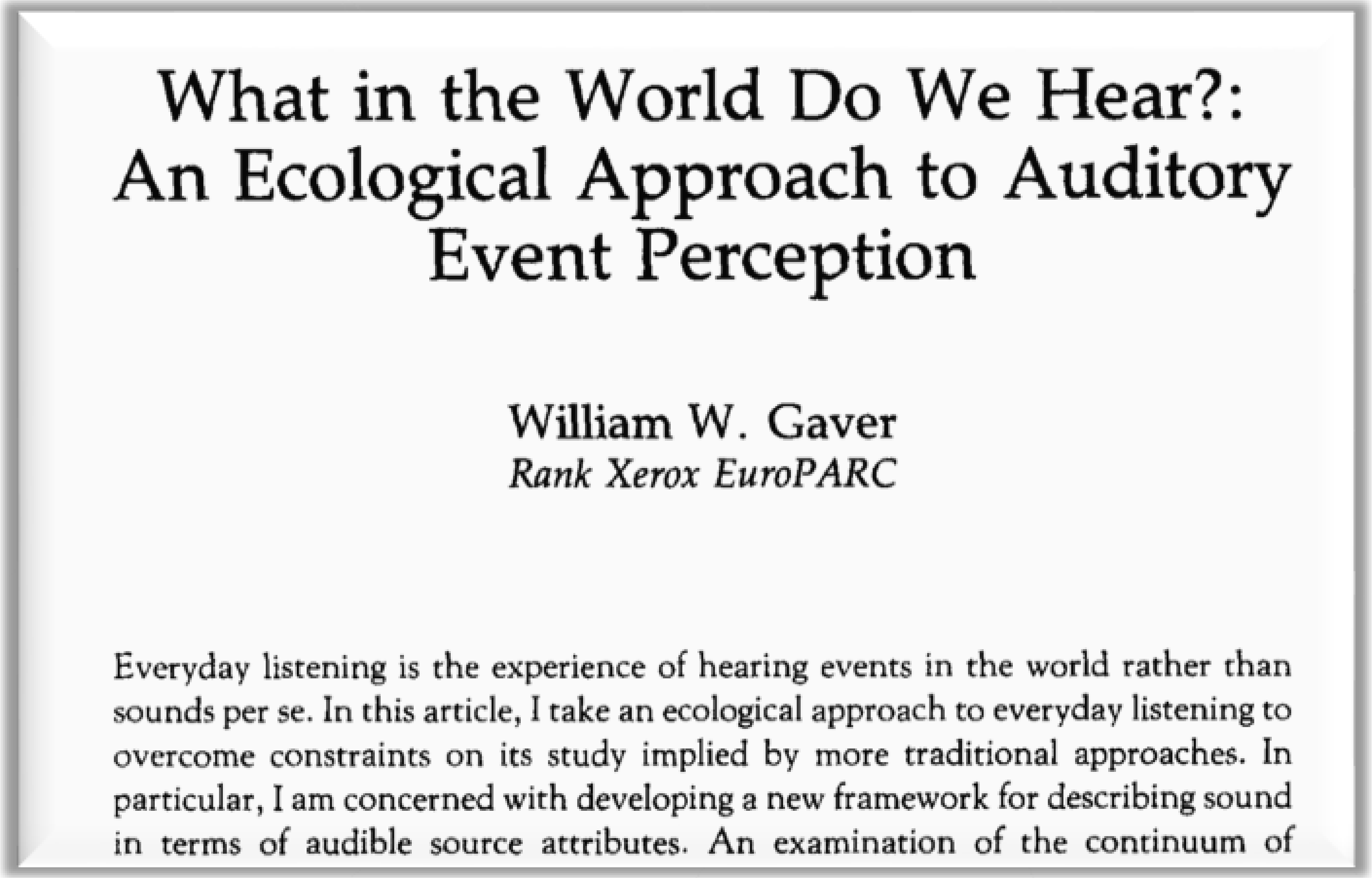
Real

Generated

User Study


Ratio of misclassifications
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Full
w/o CLIP
w/o hand pose
w/o synthetic-view
All

Human Perception of Sound



What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception

William W. Gaver
Rank Xerox EuroPARC

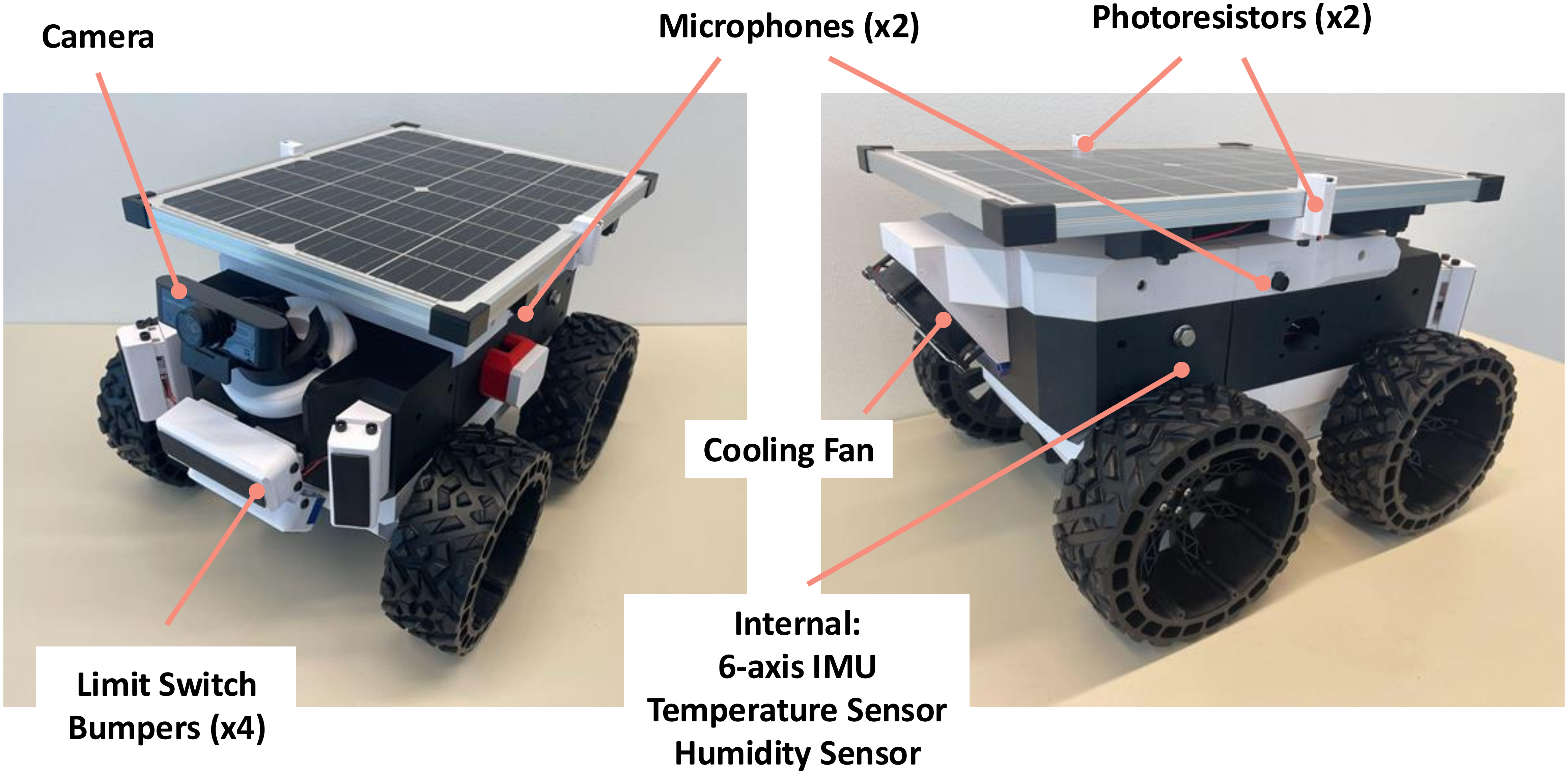
Everyday listening is the experience of hearing events in the world rather than sounds per se. In this article, I take an ecological approach to everyday listening to overcome constraints on its study implied by more traditional approaches. In particular, I am concerned with developing a new framework for describing sound in terms of audible source attributes. An examination of the continuum of

Two types of sound perception:

1. Musical Listening
2. Everyday Listening

The Survival Bot

A Diverse Array of Sensors



The Beauty of Real World

Next Steps: Month-Long Learning

Takeaways

- Embodied intelligence is the ability to deal with novelty, failure, and uncertainty.
- Interaction gives an agent the opportunity to learn about themselves and the environment.
- Get out of the lab!

Thank you!