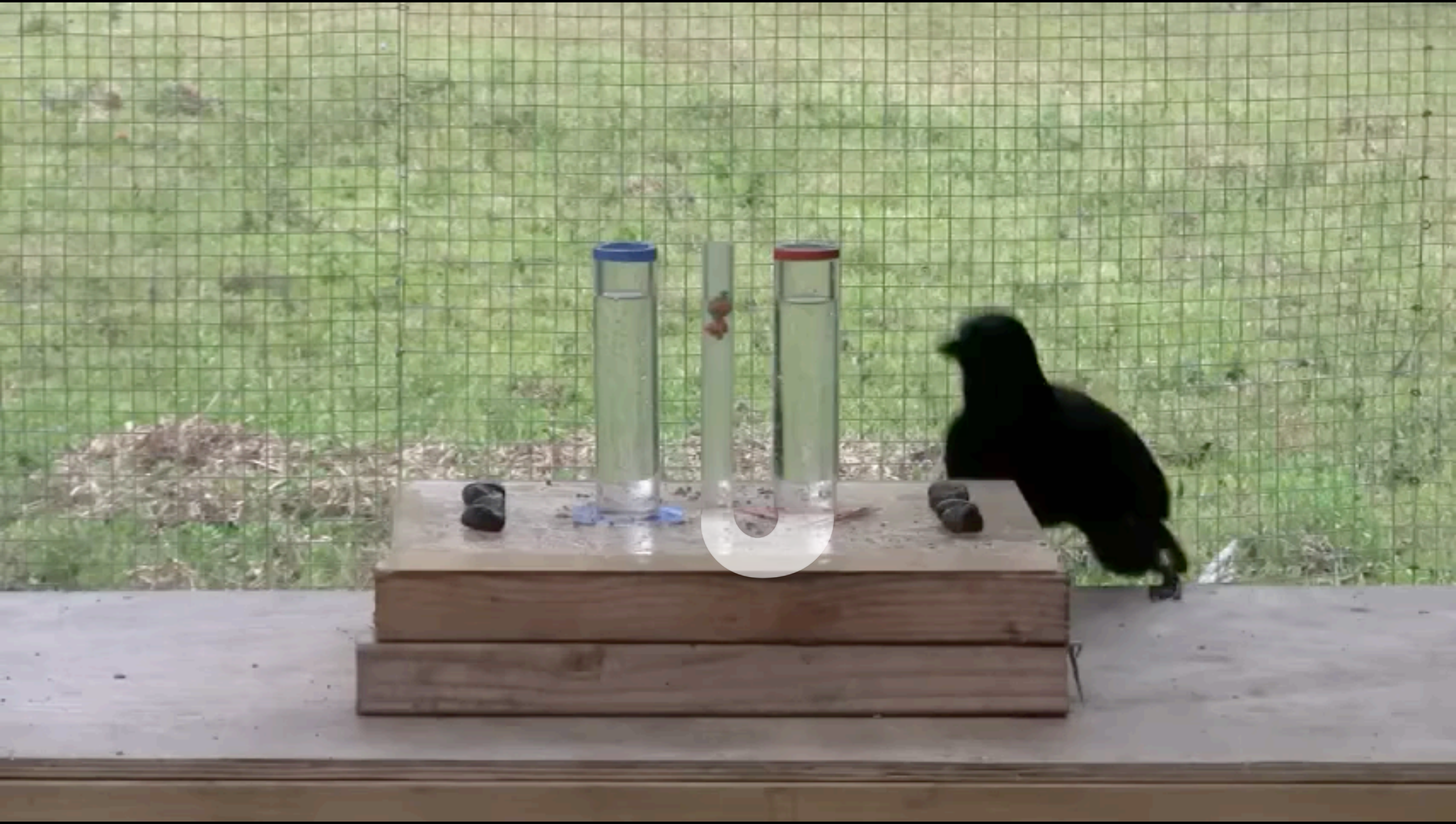
Generative Models for Computer Vision

Carl Vondrick
Columbia University



I ❤️ LLMs

- Nuclear Power Plants

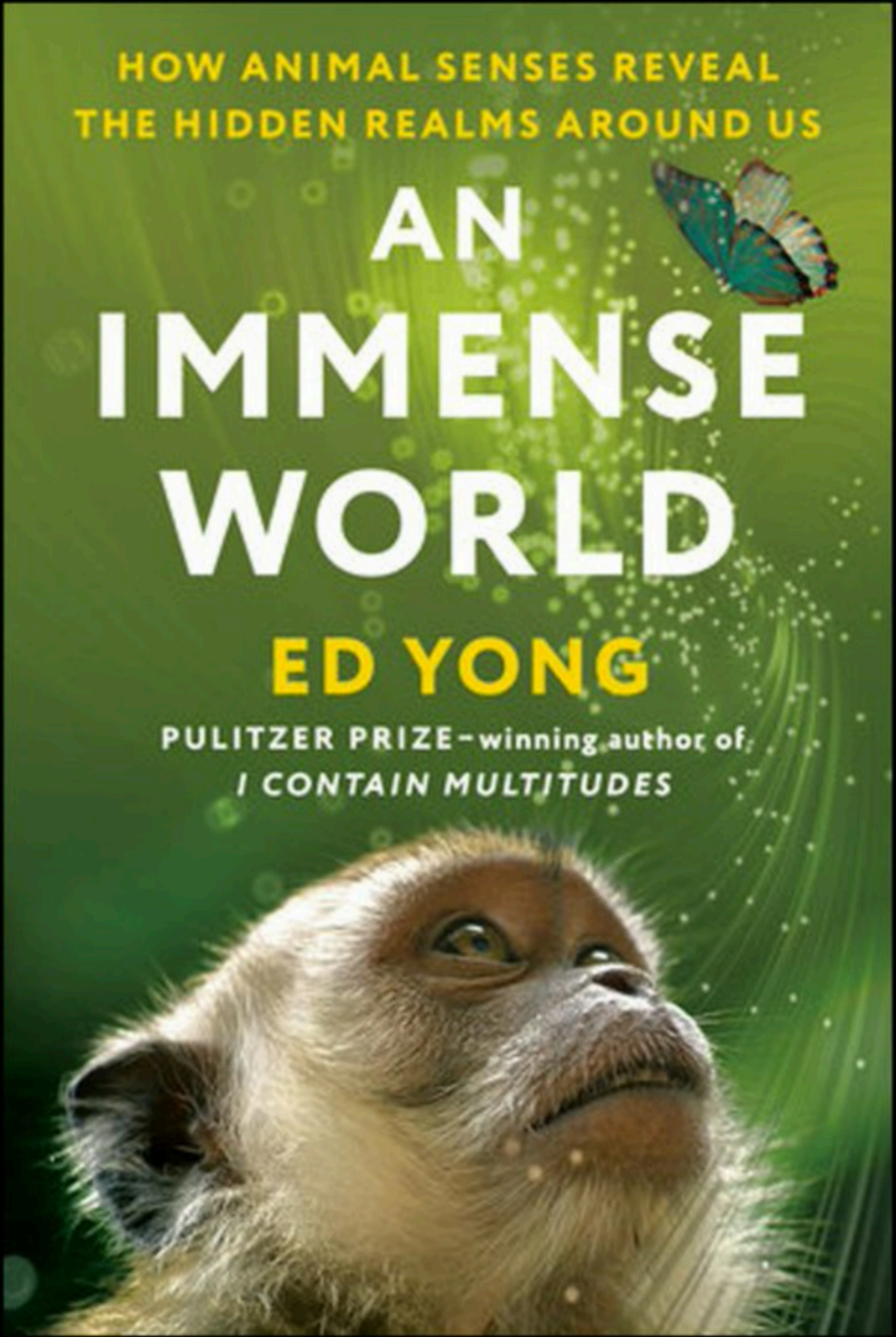
- Worms for Breakfast

- Nuclear Power Plants
- Hallucinates

- Worms for Breakfast
- Physical Consequences

- Nuclear Power Plants
- Hallucinates
- Learns from Text

- Worms for Breakfast
- Physical Consequences
- Learns from Surroundings

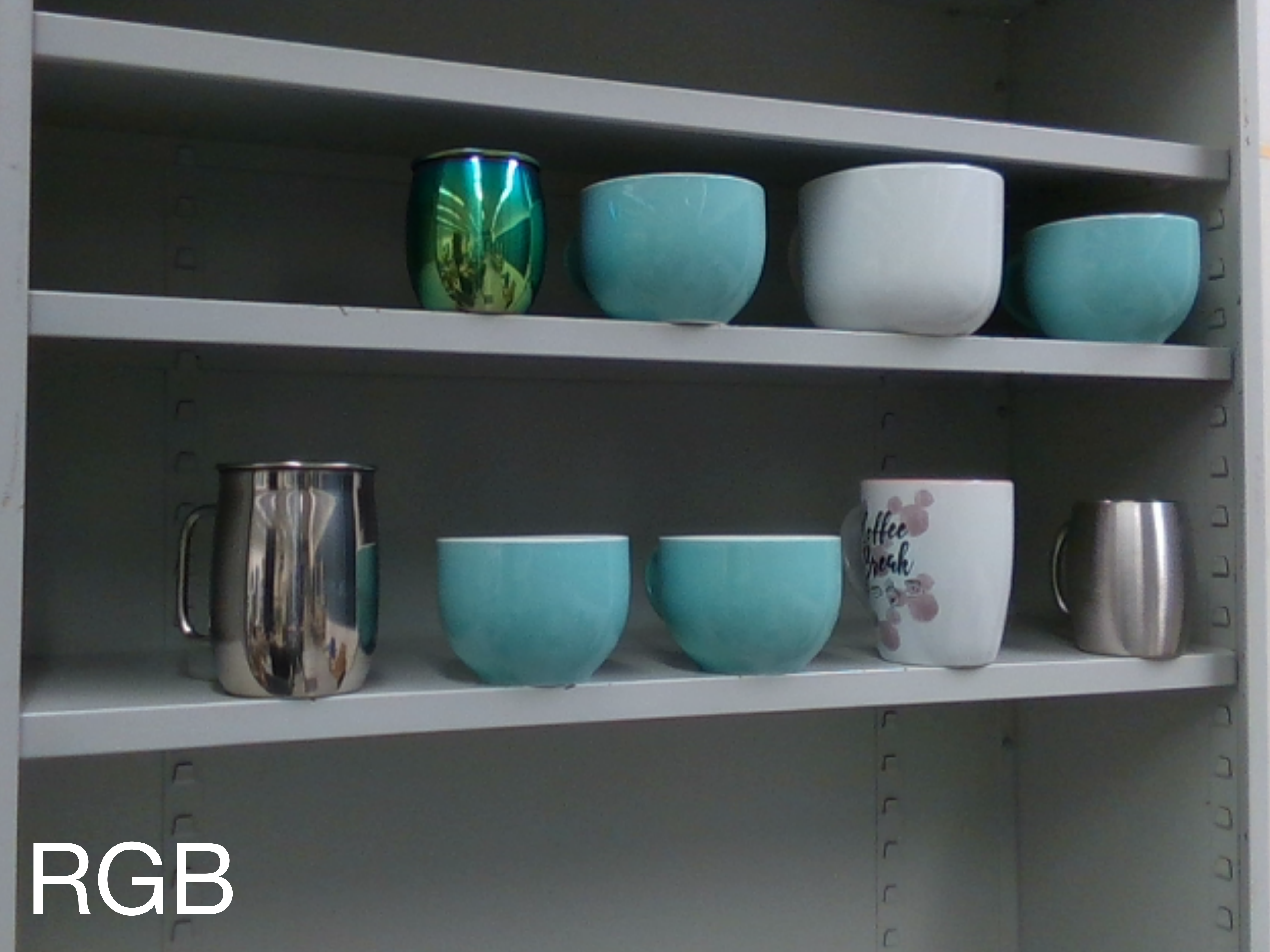


HOW ANIMAL SENSES REVEAL
THE HIDDEN REALMS AROUND US

AN IMMENSE WORLD

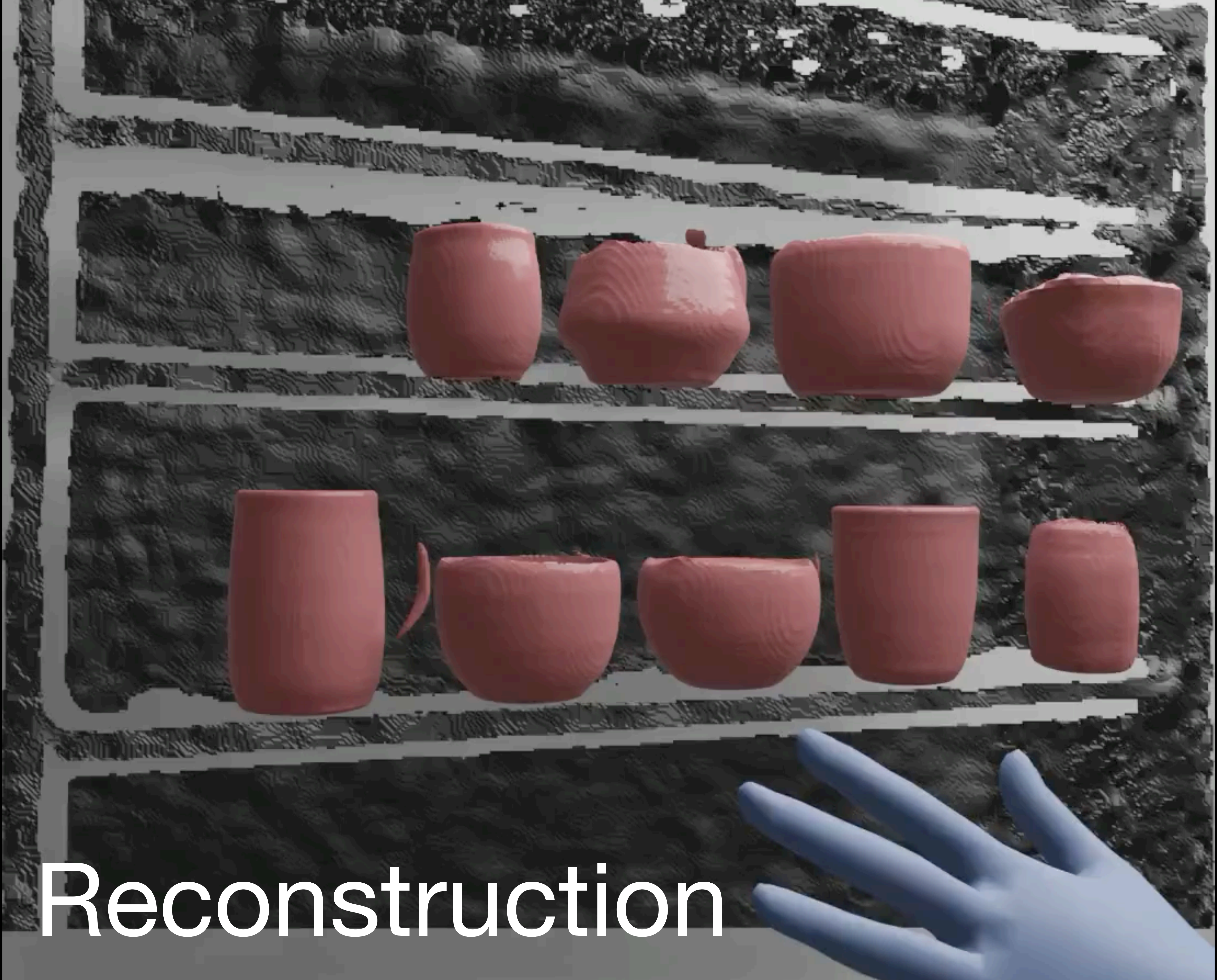
ED YONG

PULITZER PRIZE–winning author of
I CONTAIN MULTITUDES

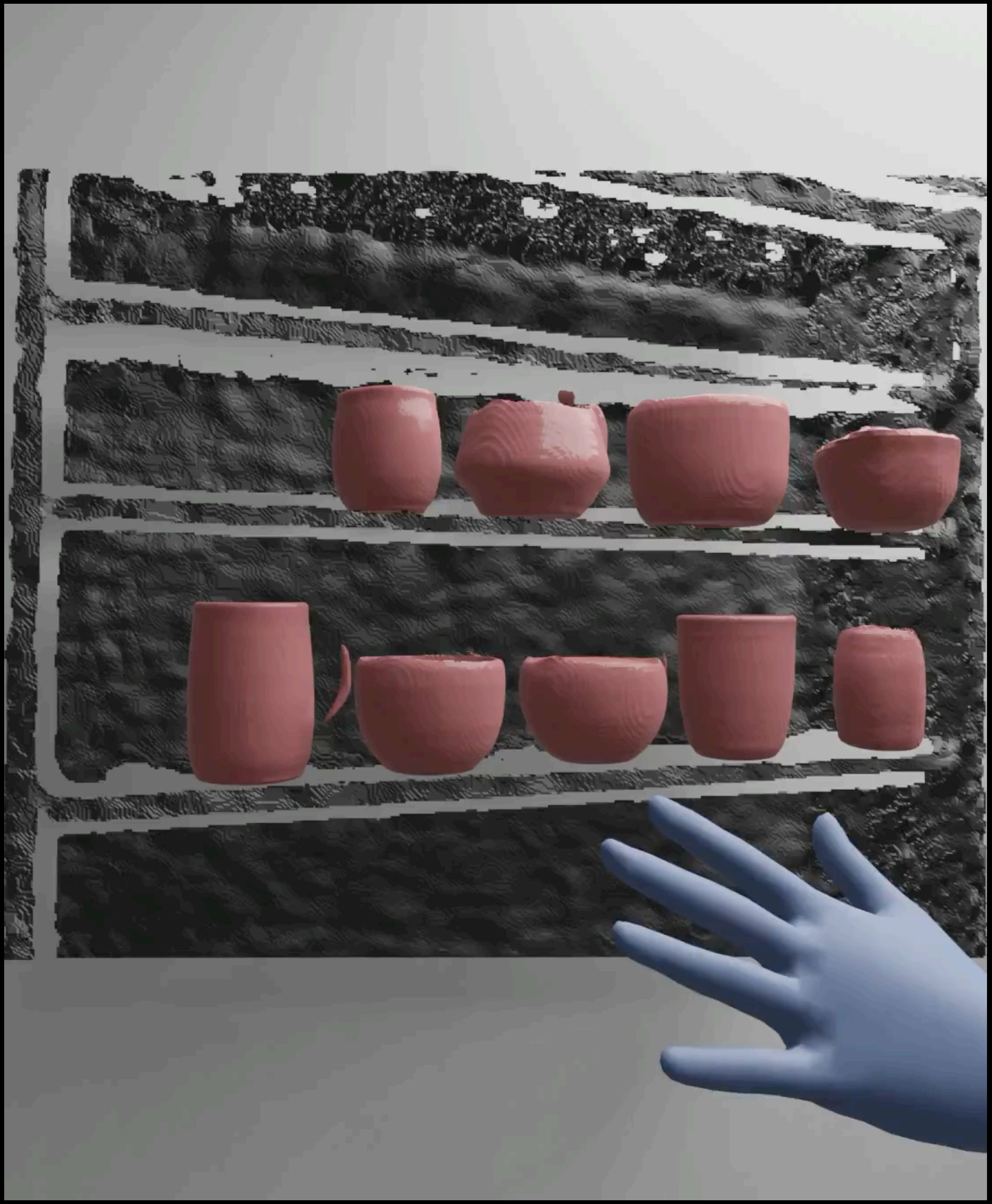


RGB

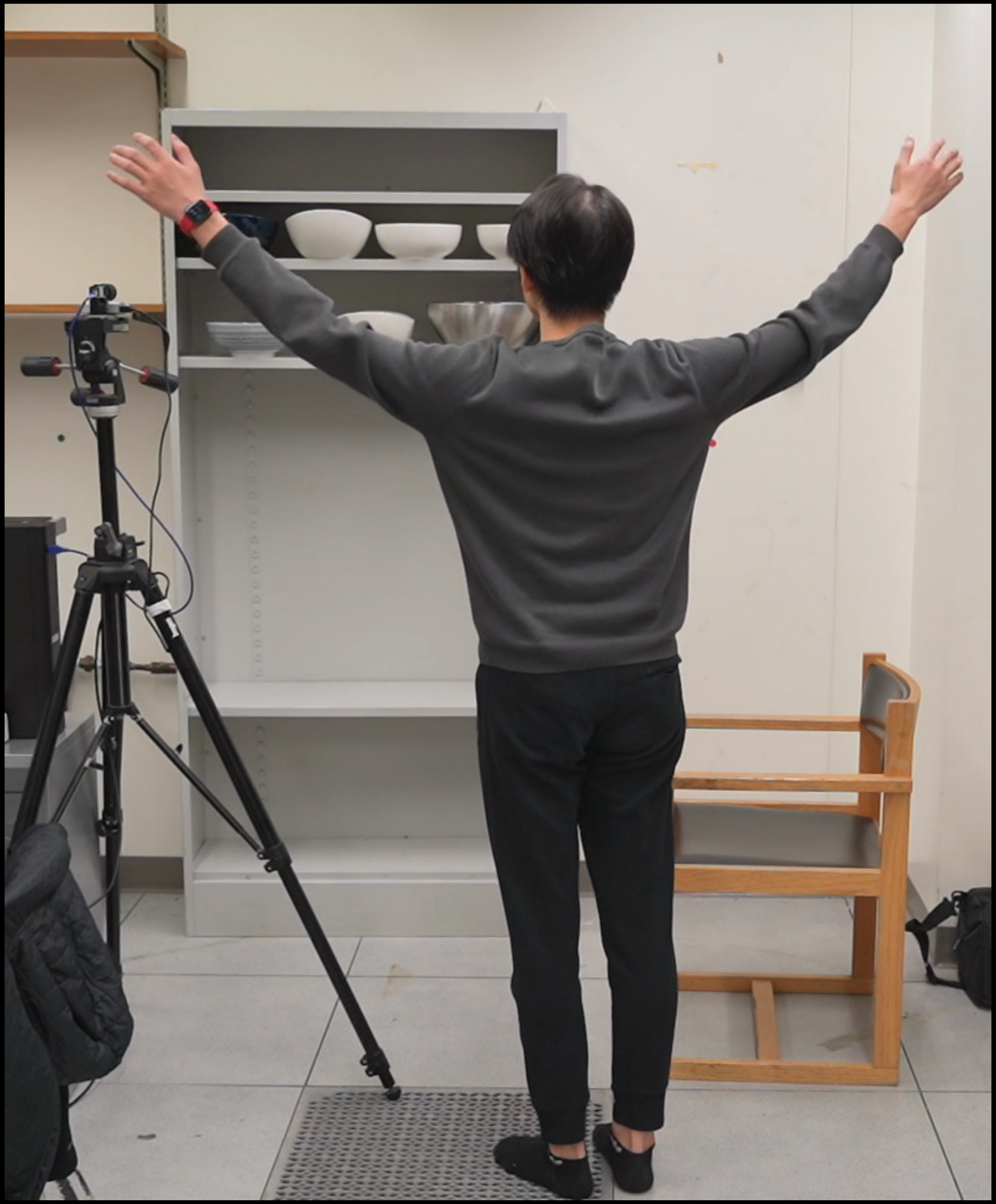
Thermal



Reconstruction

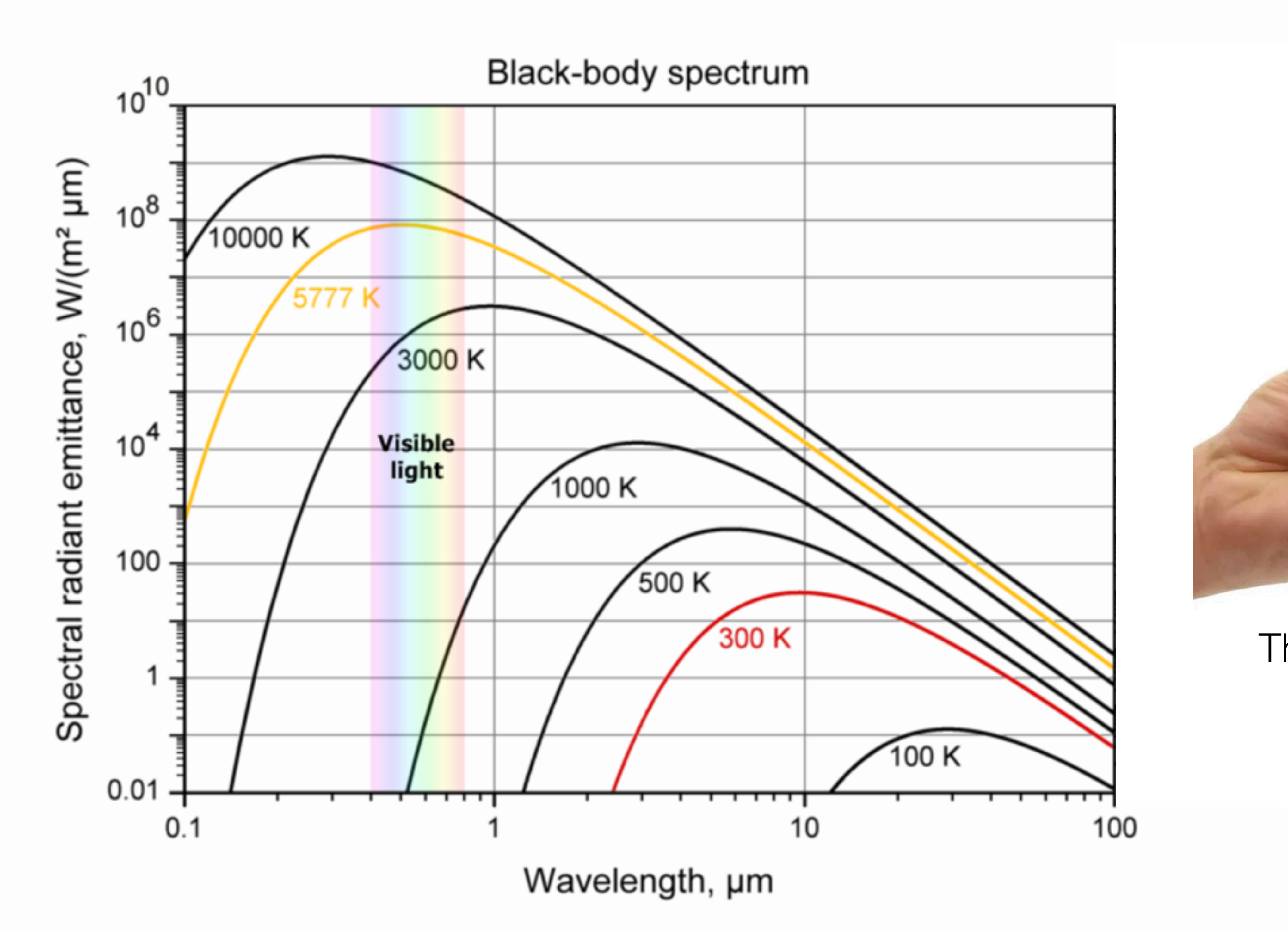


Reconstruction

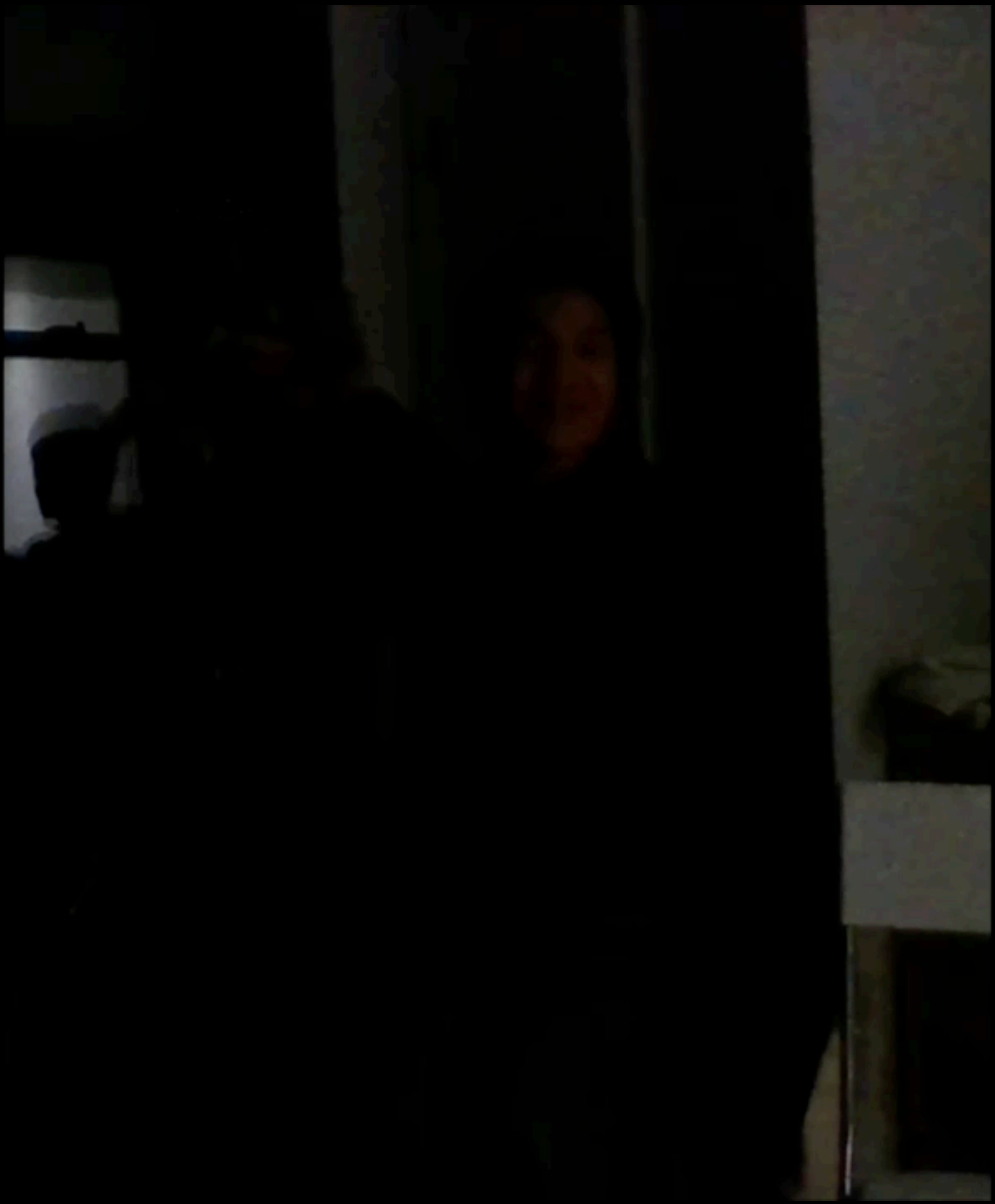


True Scene

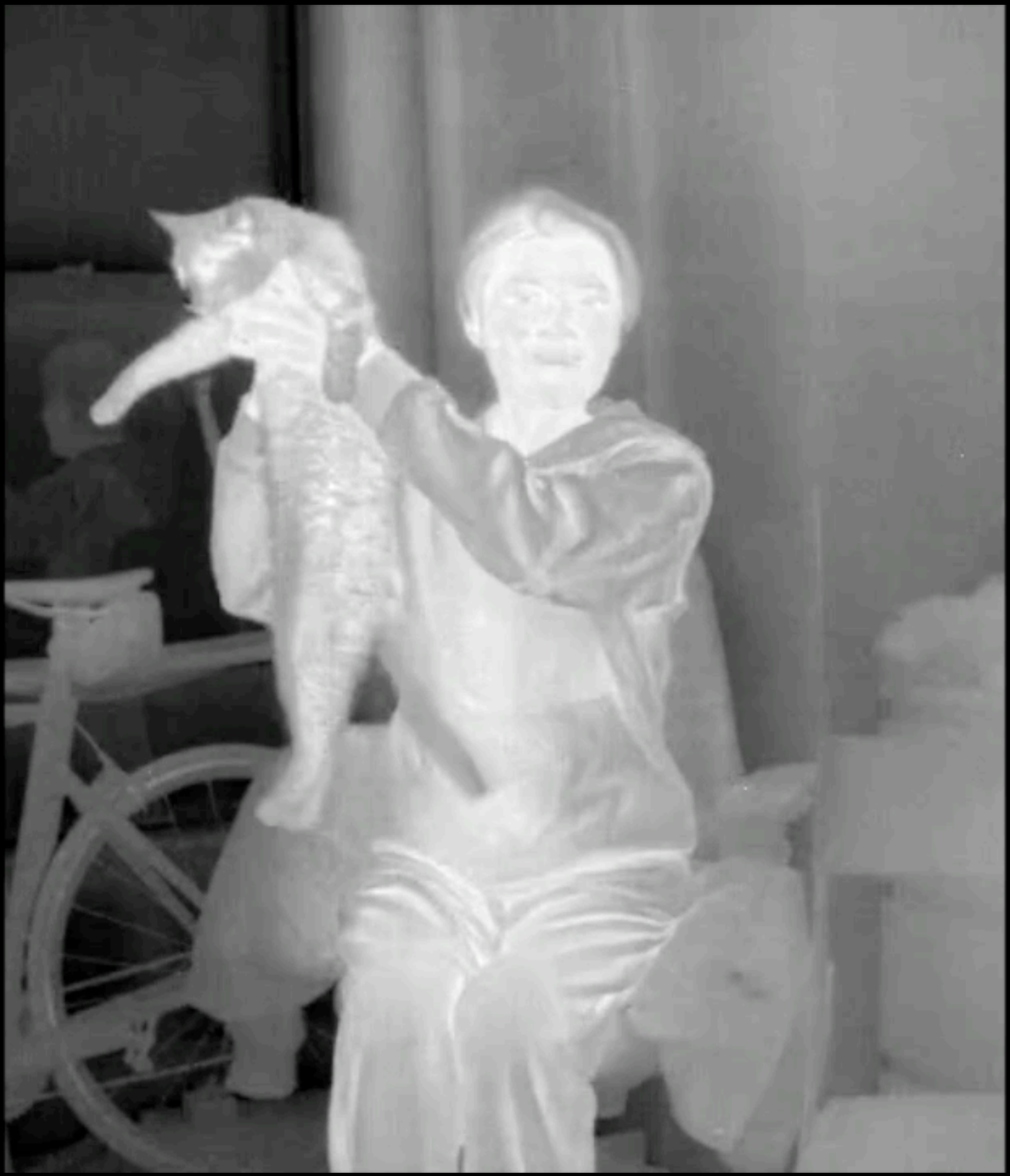
Black-body Radiation



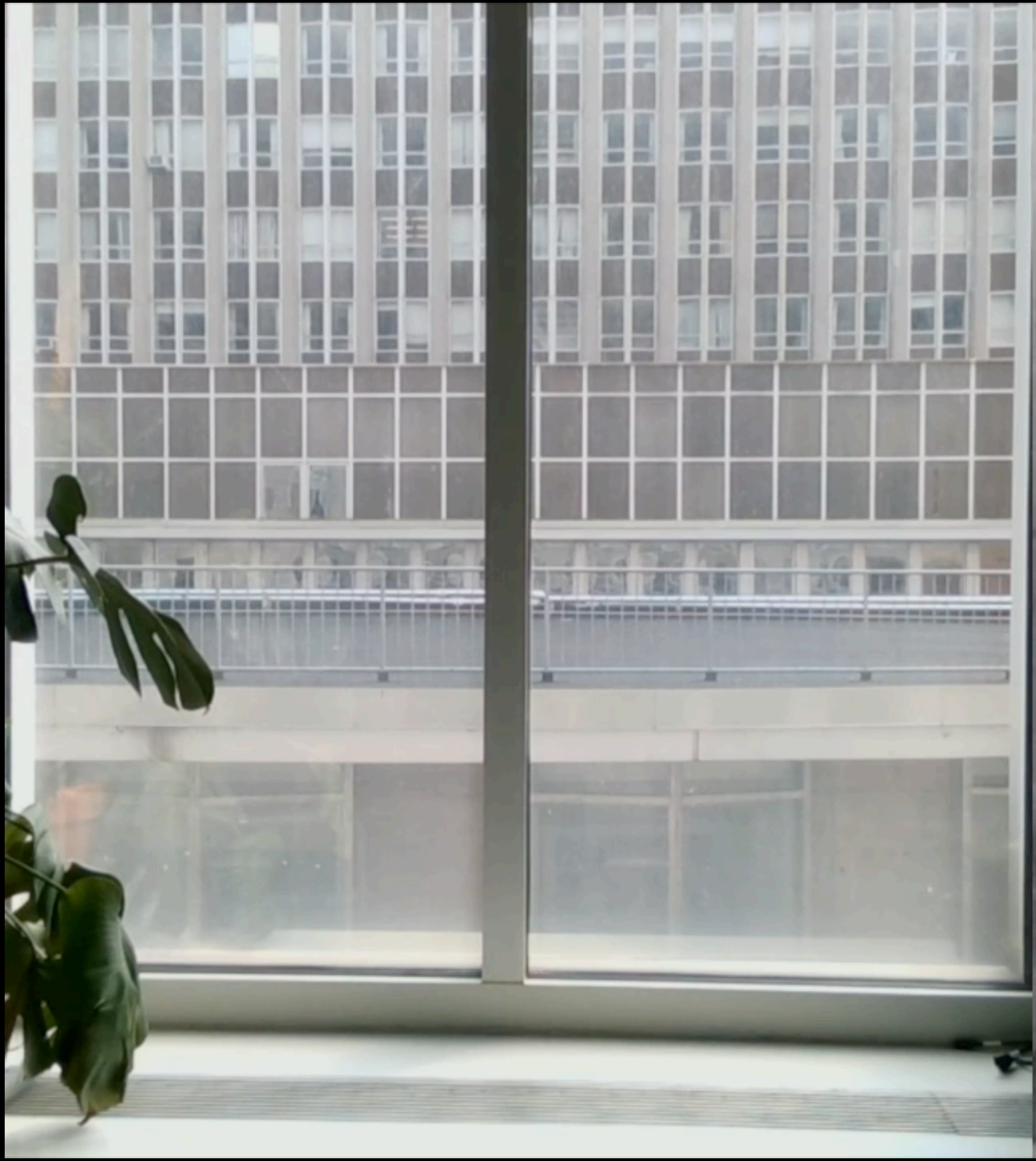
Thermal Camera
(7-14 µm)



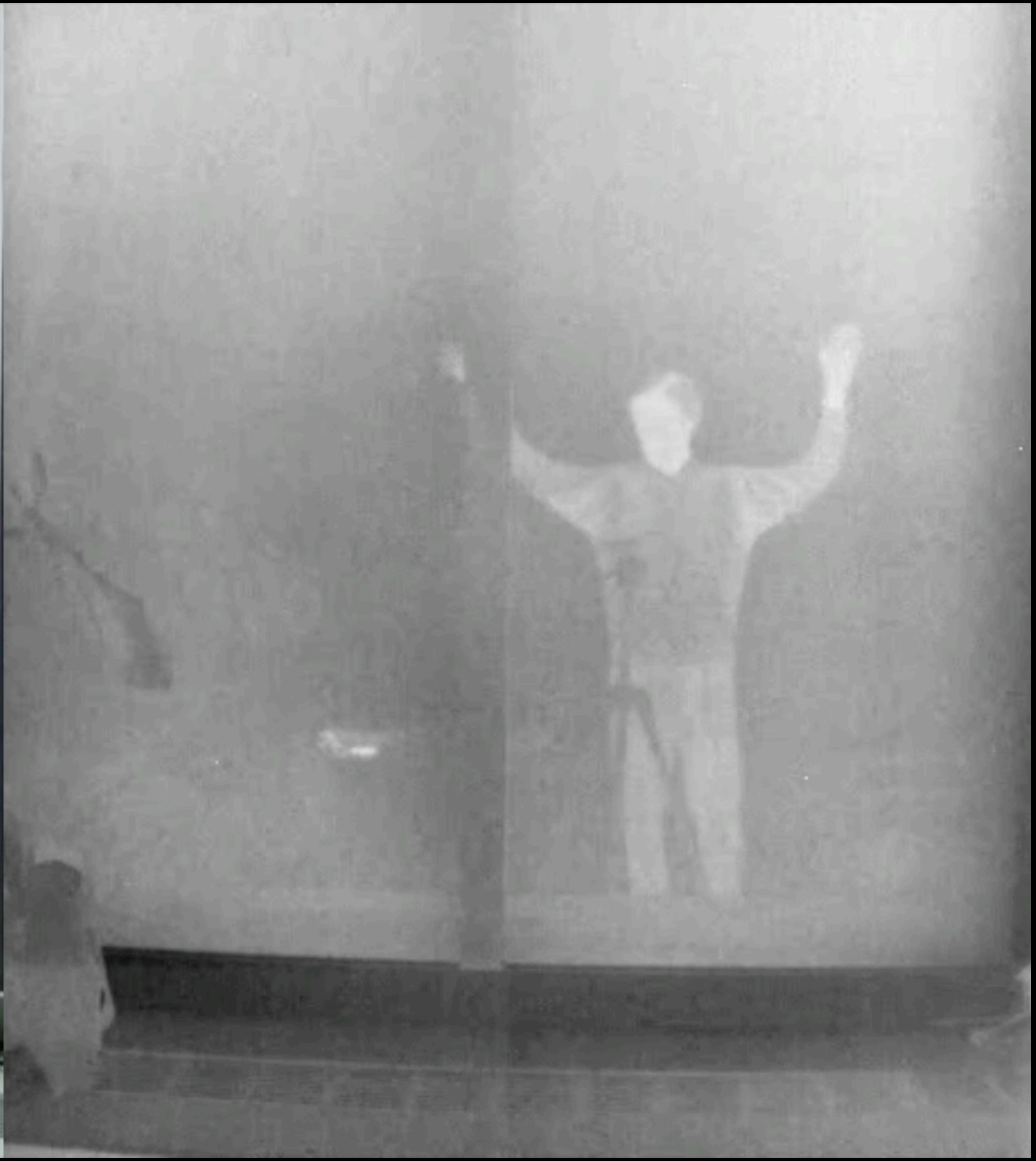
Normal Camera (0.4-0.7 µm)



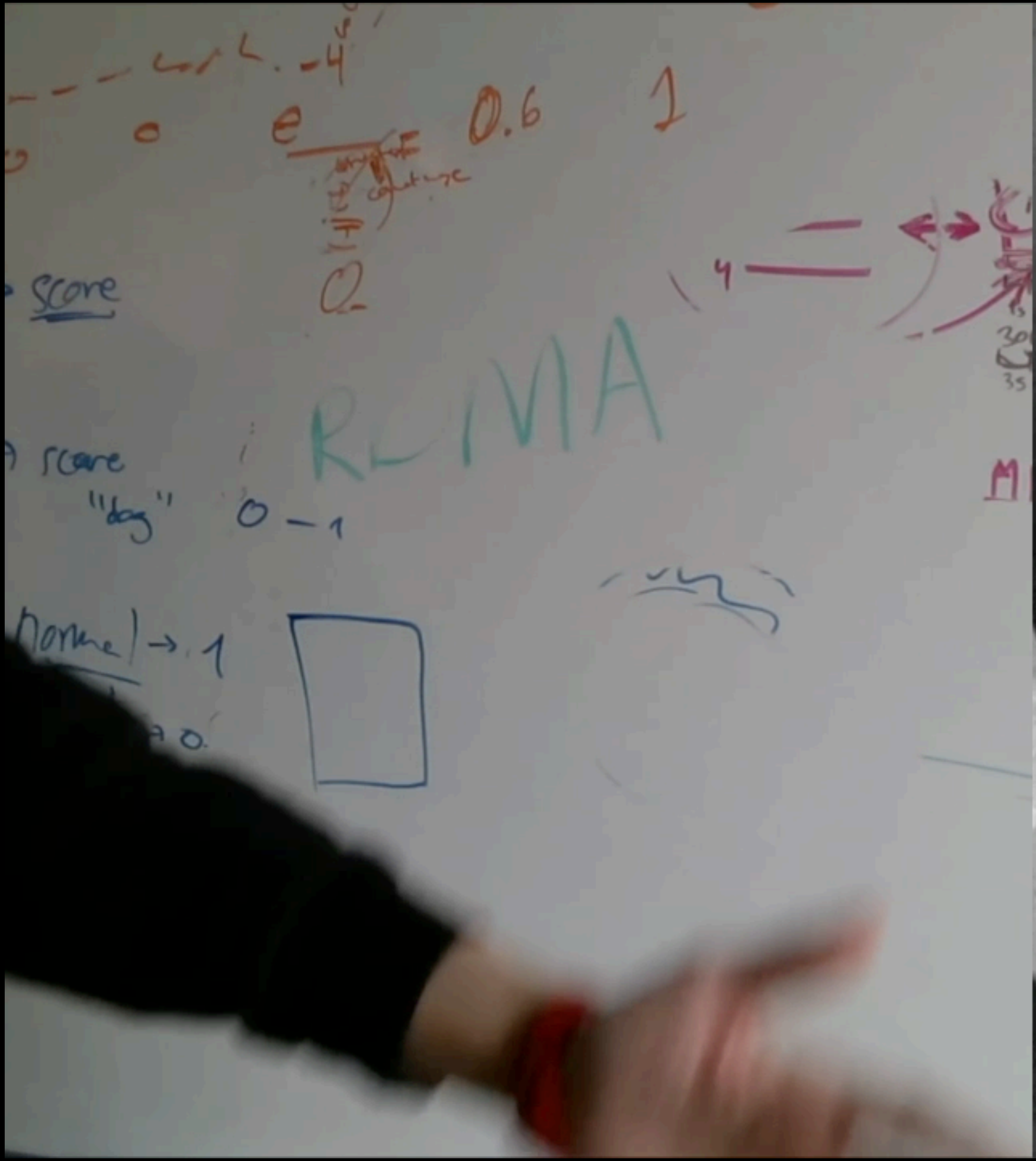
Thermal Camera (7-14 µm)



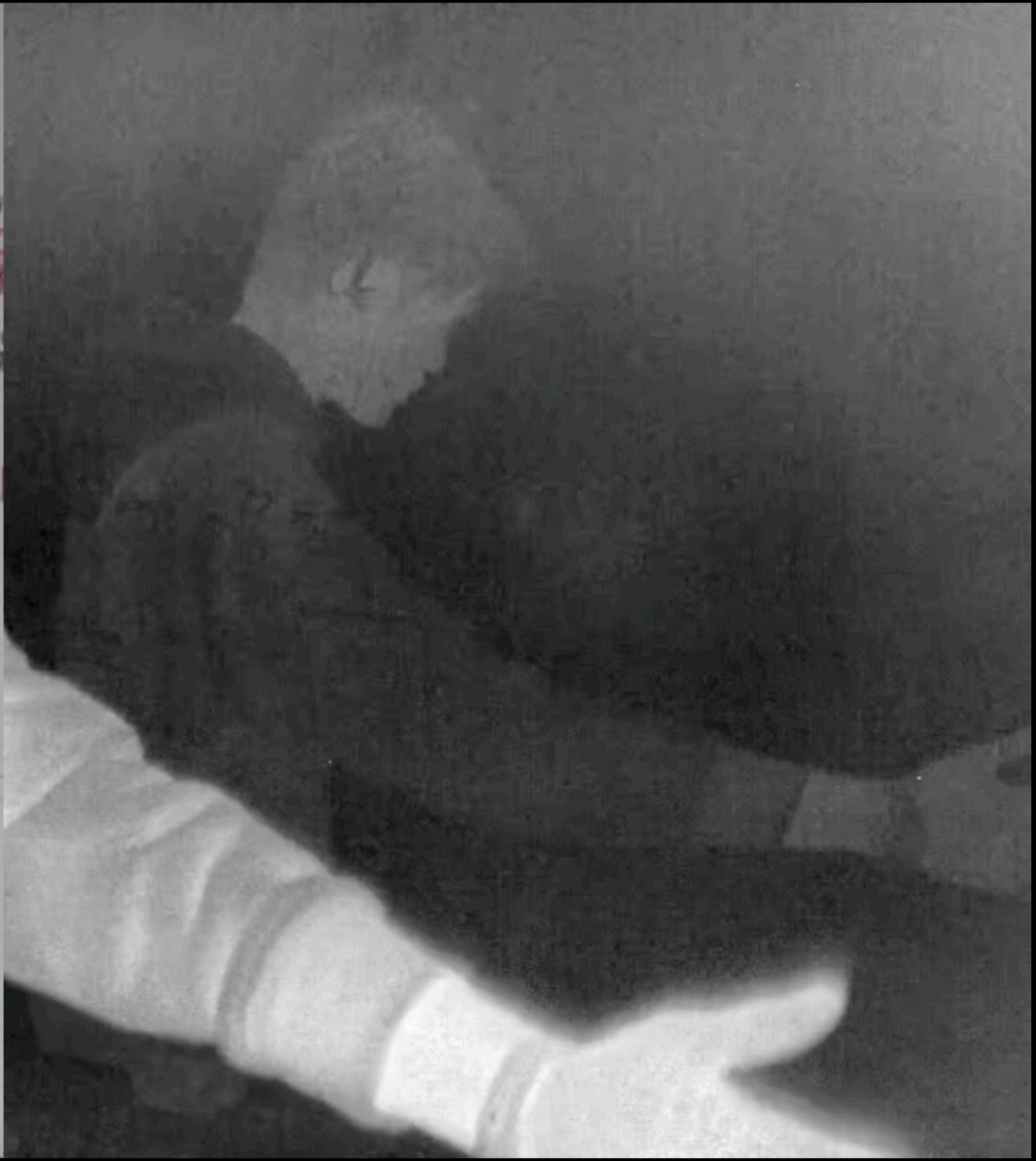
Normal Camera (0.4-0.7 μm)



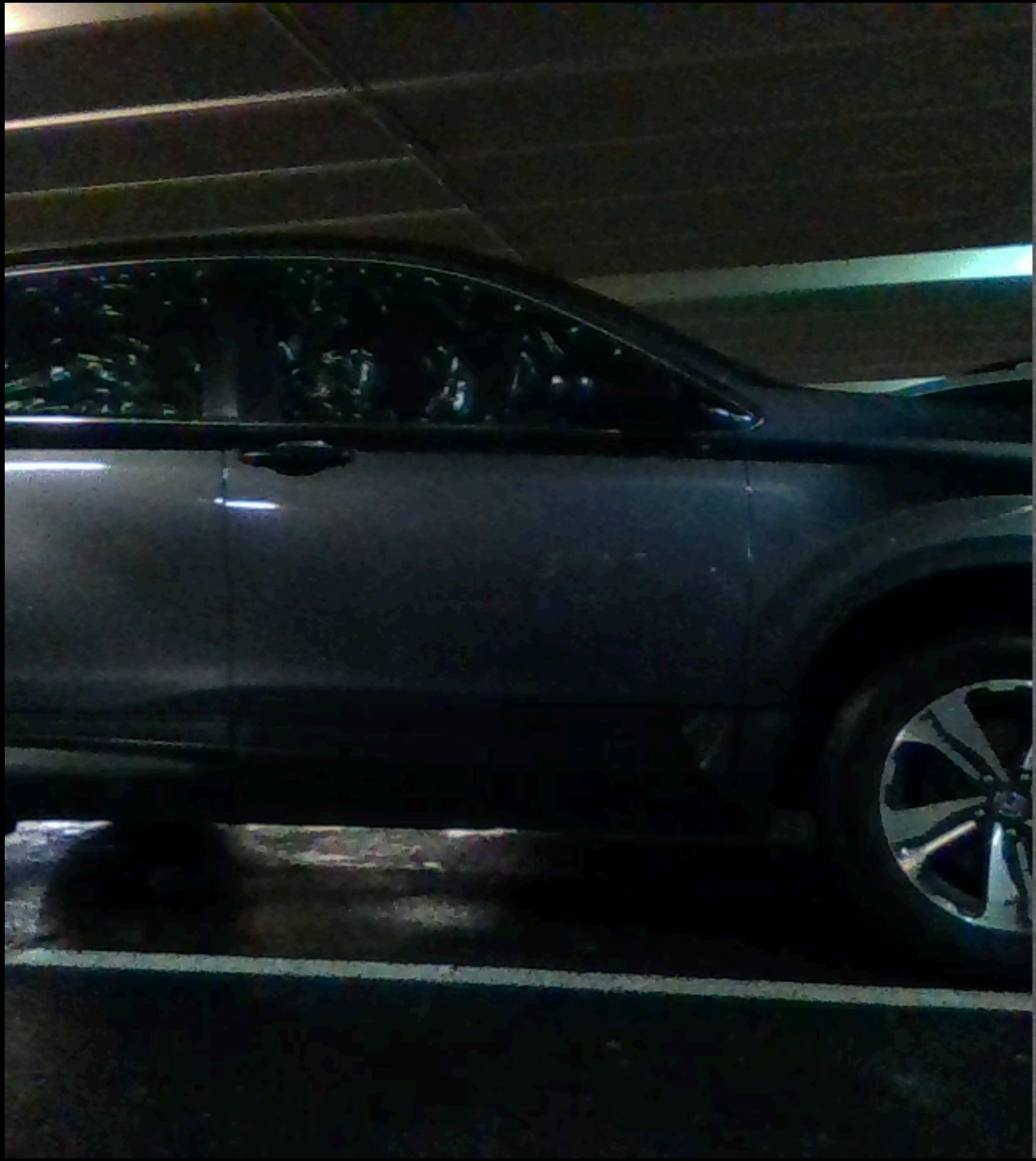
Thermal Camera (7-14 μm)



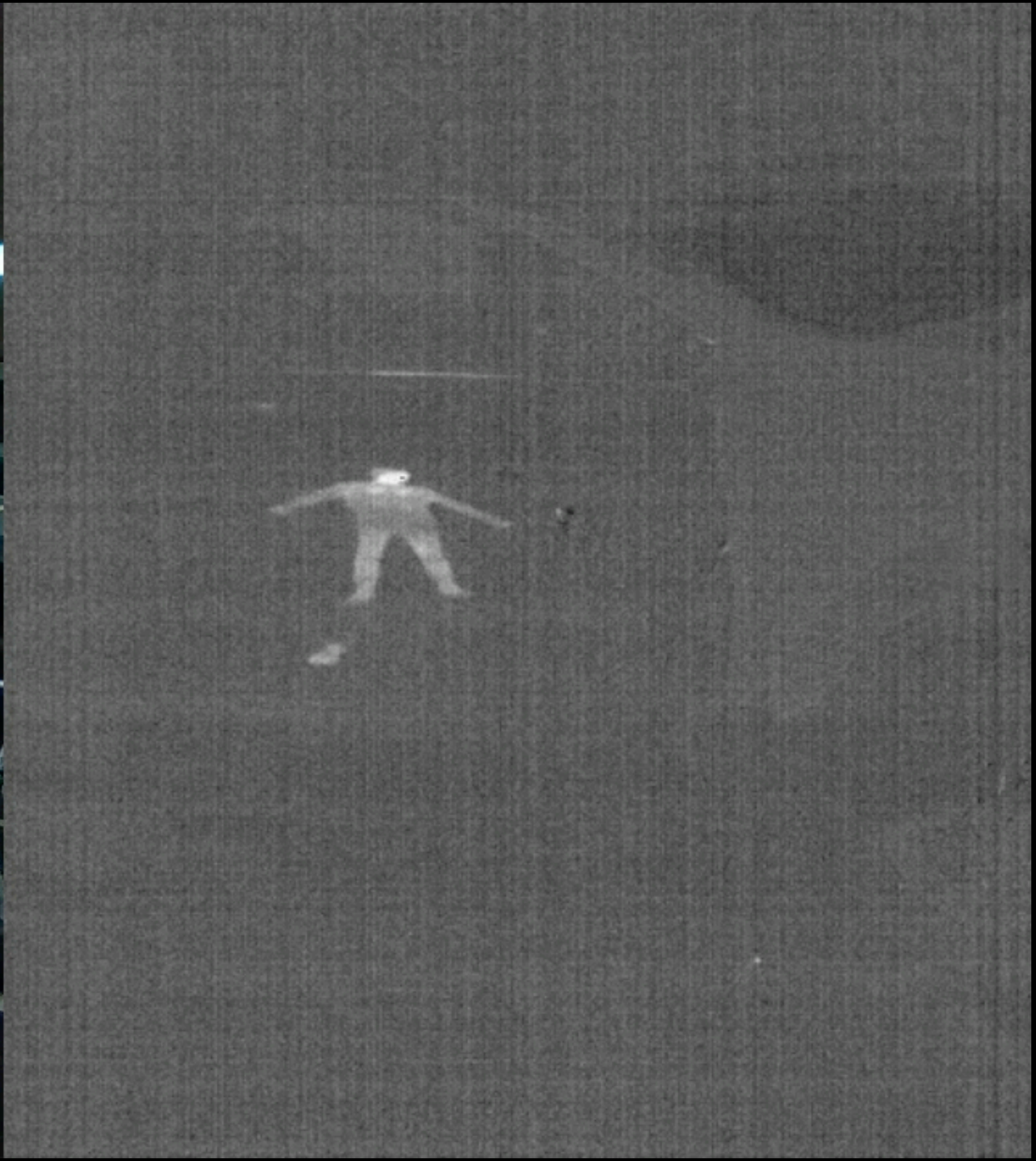
Normal Camera (0.4-0.7 µm)



Thermal Camera (7-14 µm)

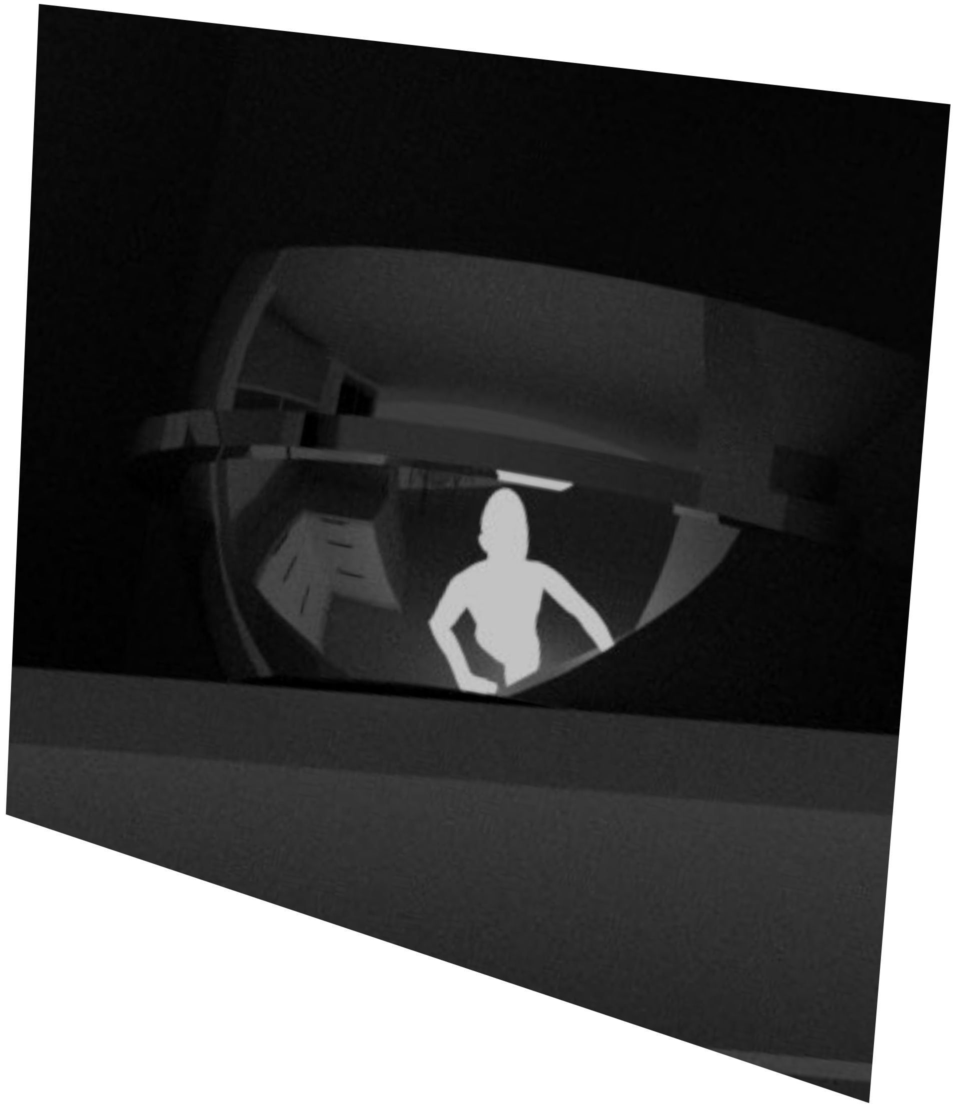


Normal Camera (0.4-0.7 µm)

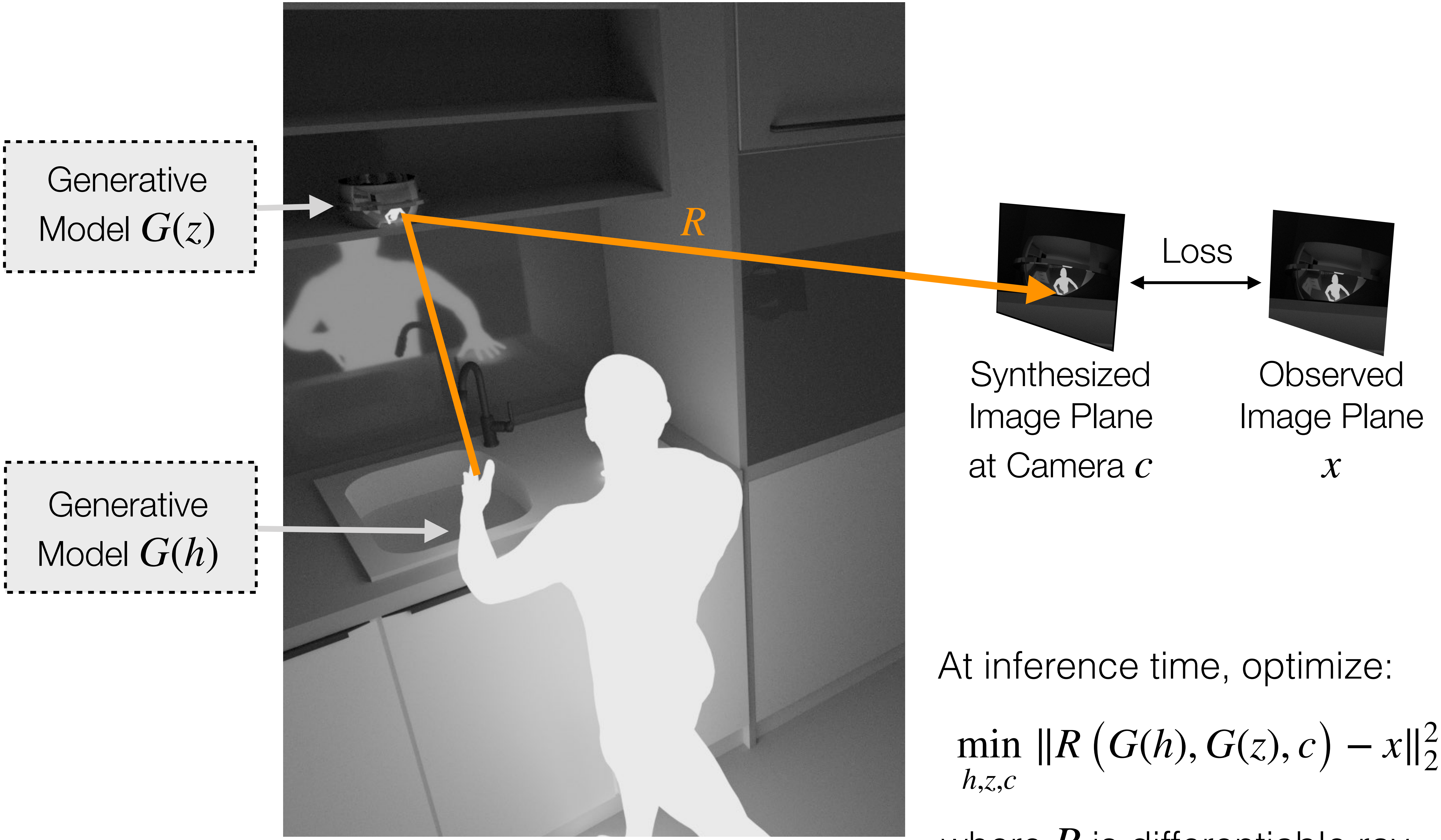


Thermal Camera (7-14 µm)

What 3D world created this reflection?



Liu, Vondrick. CVPR 2023.



Generative Model $G(z)$

Generative Model $G(h)$

R

Loss

Synthesized Image Plane at Camera c

Observed Image Plane x

3D World Scene

At inference time, optimize:

$$\min_{h,z,c} \|R\left(G(h), G(z), c\right) - x\|_2^2$$

where R is differentiable ray tracing with reflection.

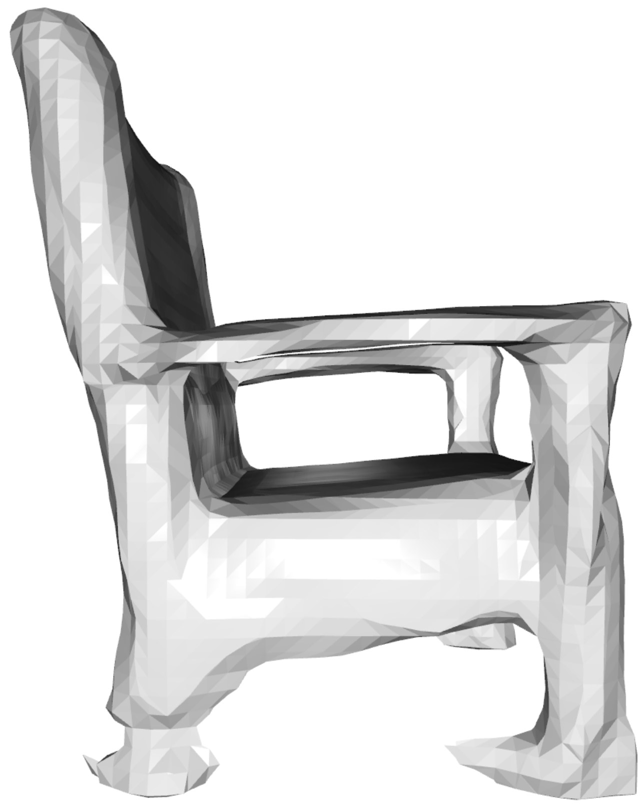
Liu, Vondrick. CVPR 2023.

RGB

RGB

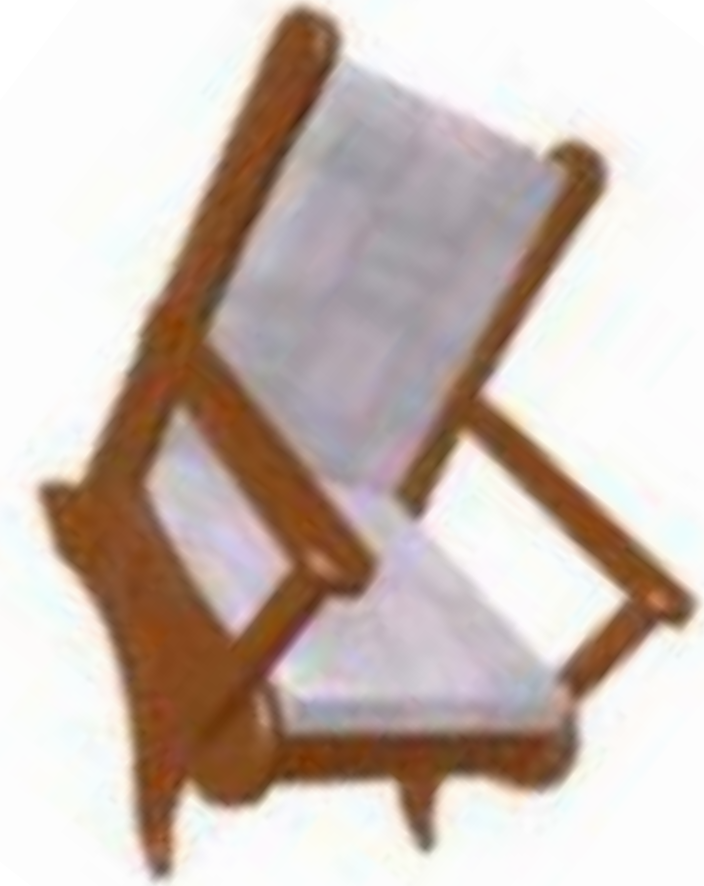
RGB

Single RGB Image

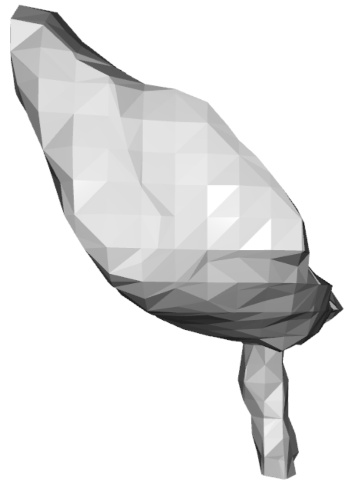


3D Reconstruction

Results from Occupancy Networks



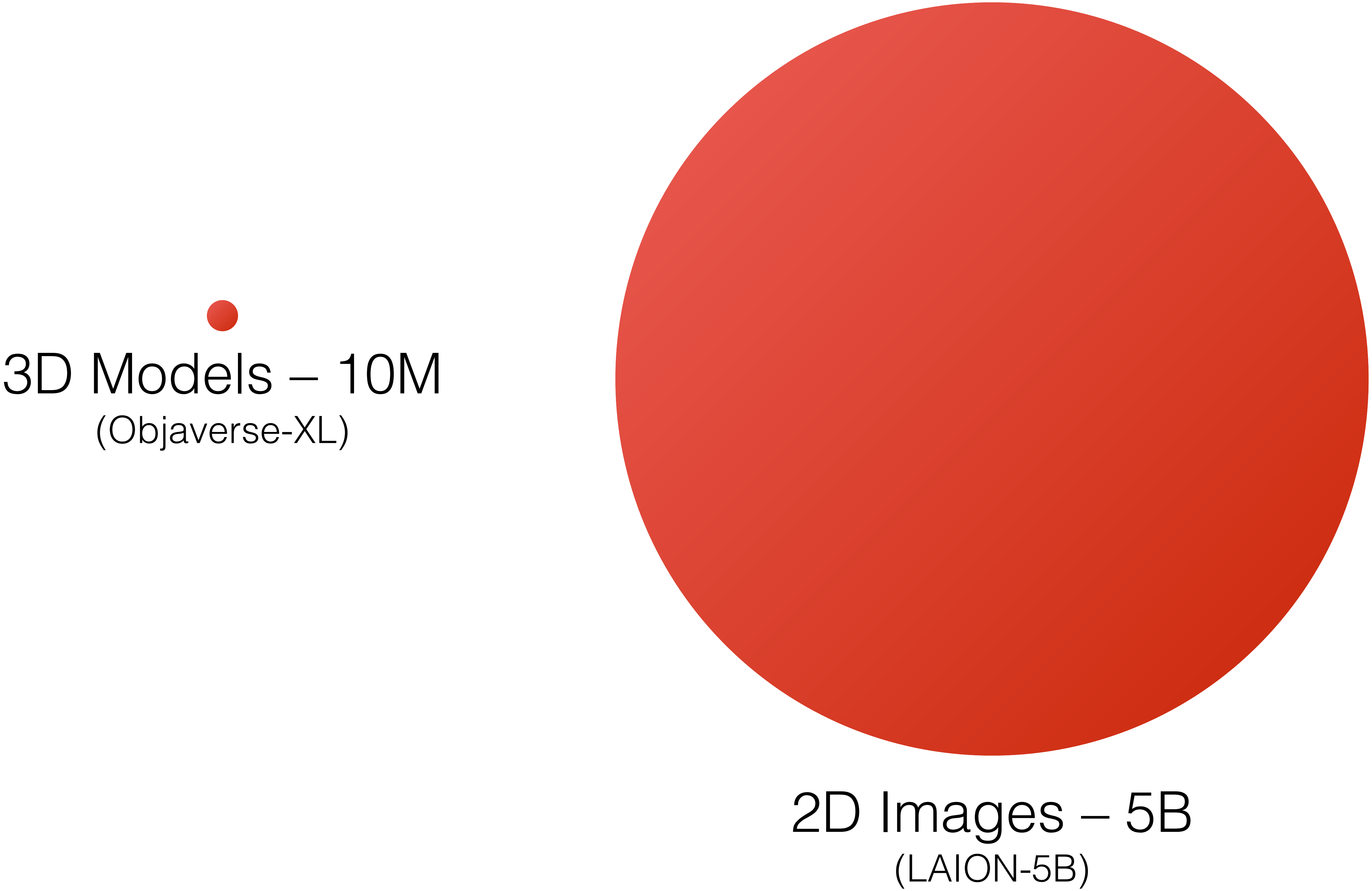
Single RGB Image



3D Reconstruction

Results from Occupancy Networks

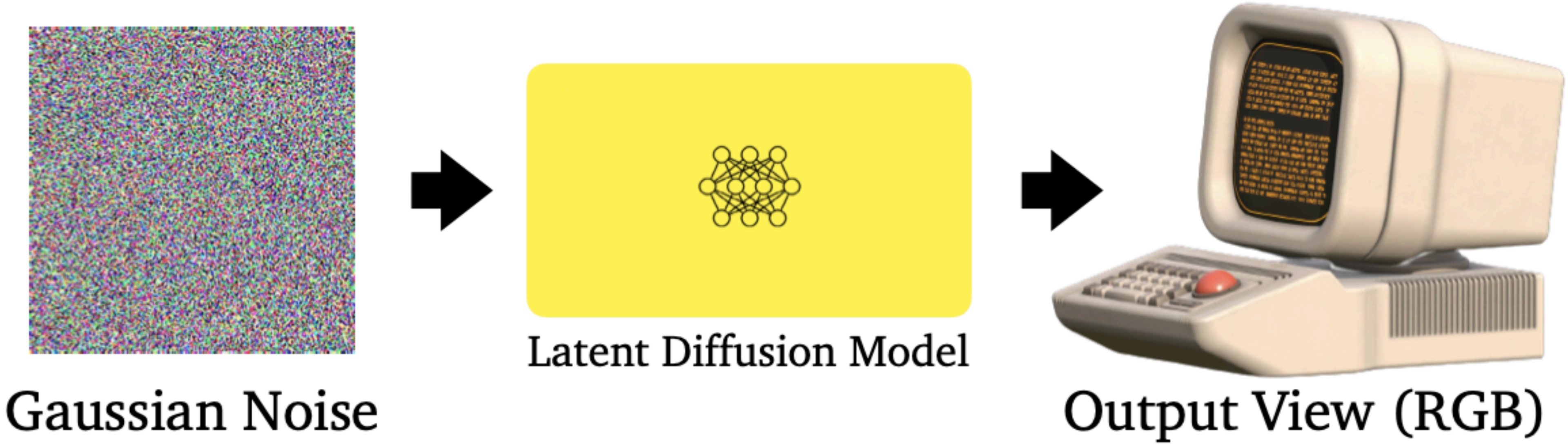
Size of Visual Datasets



3D Models – 10M
(Objaverse-XL)

2D Images – 5B
(LAION-5B)

Generated by Stable Diffusion

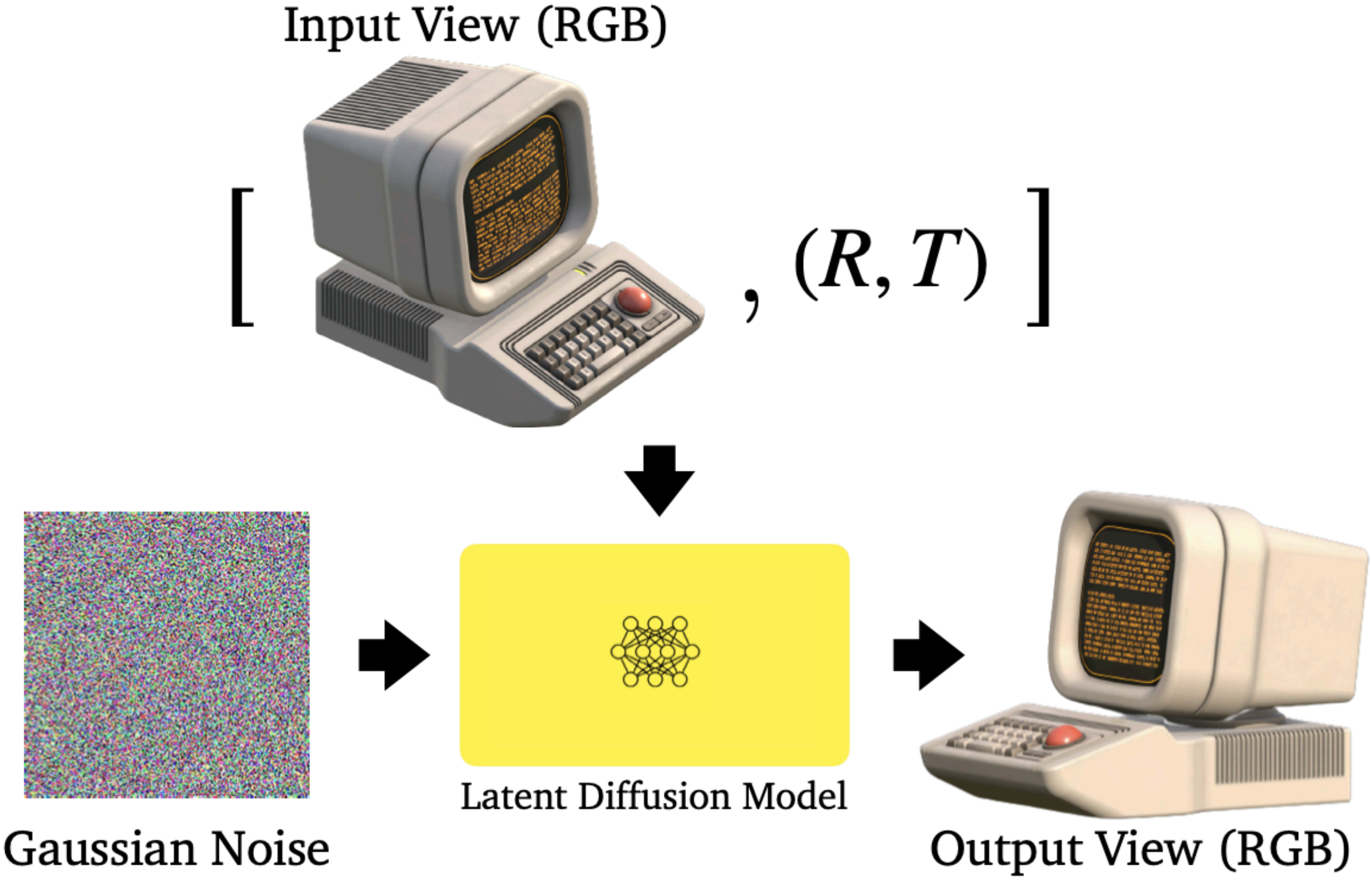


Gaussian Noise

Latent Diffusion Model

Output View (RGB)

Learning Camera Control



Input View (RGB)

$[\quad , (R, T)]$

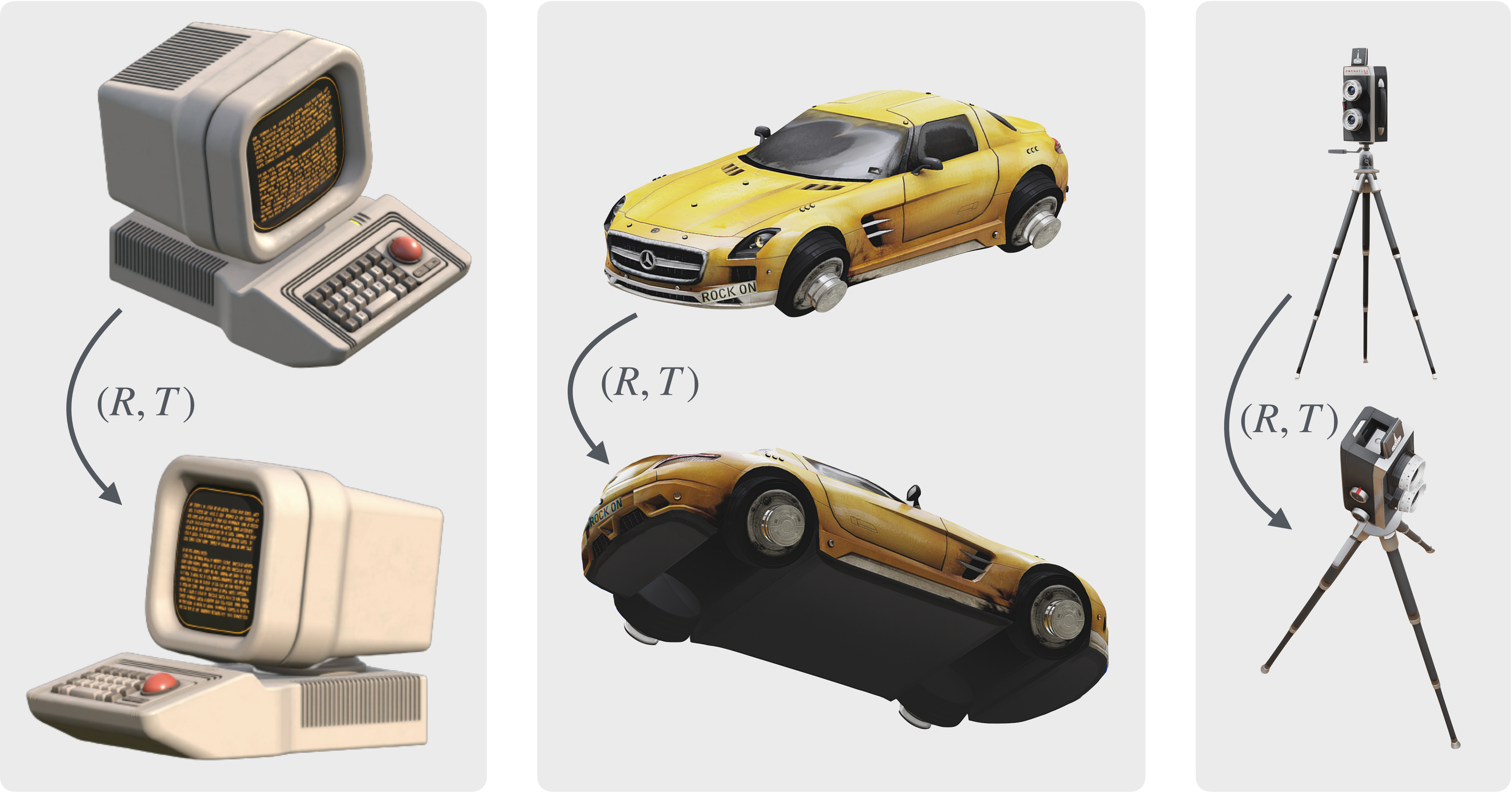
Gaussian Noise

Latent Diffusion Model

Output View (RGB)

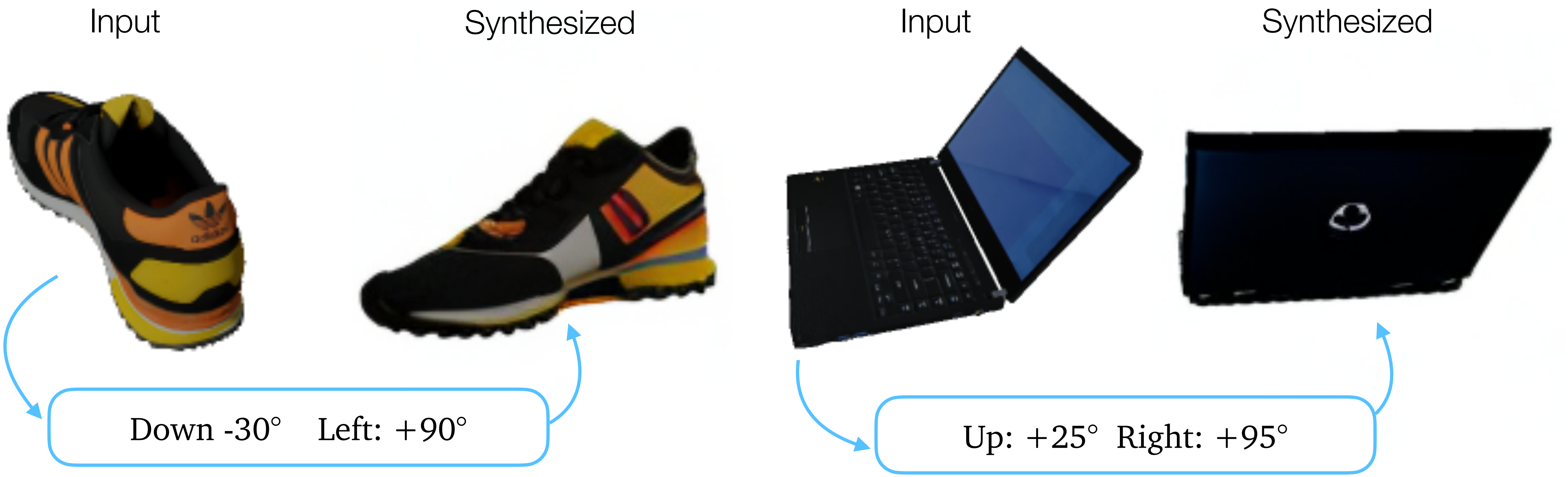
Liu, Wu, Van Hoorick, Tokmakov, Zakharov, Vondrick. ICCV 2023.

Dataset of Novel Views



Objaverse: 5,000 times smaller than LAION-5B

View Synthesis



Input

Synthesized

Input

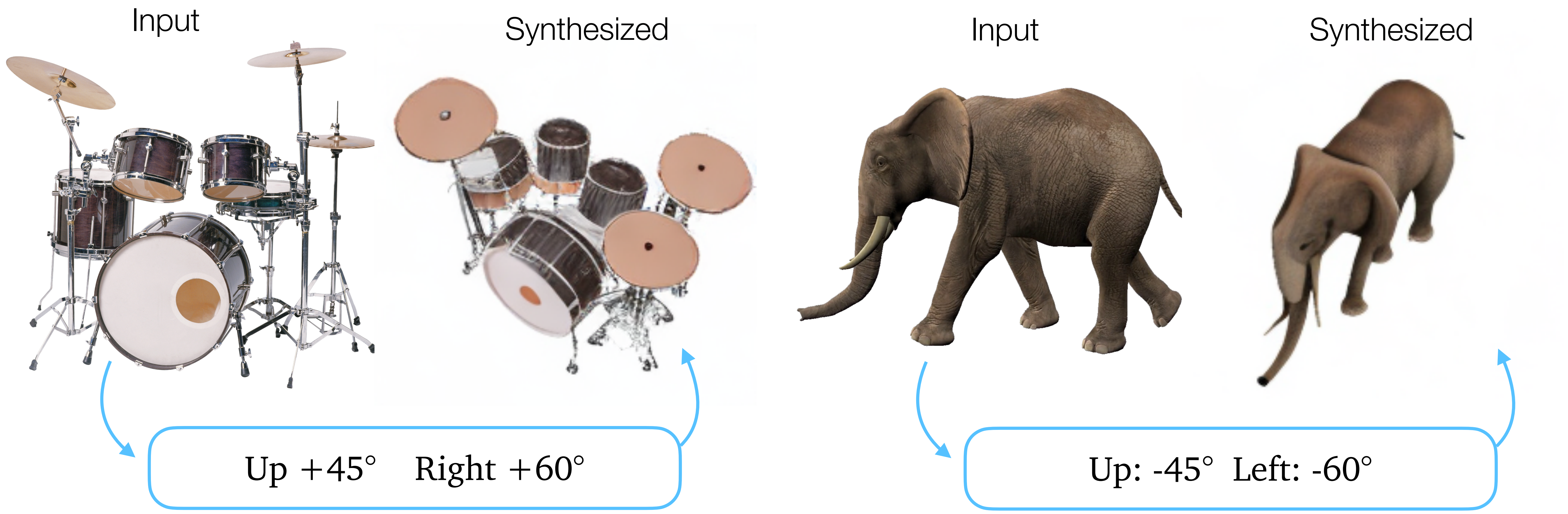
Synthesized

Down -30° Left: +90°

Up: +25° Right: +95°

Liu, Wu, Van Hoorick, Tokmakov, Zakharov, Vondrick. ICCV 2023.

View Synthesis



Input

Synthesized

Up +45° Right +60°

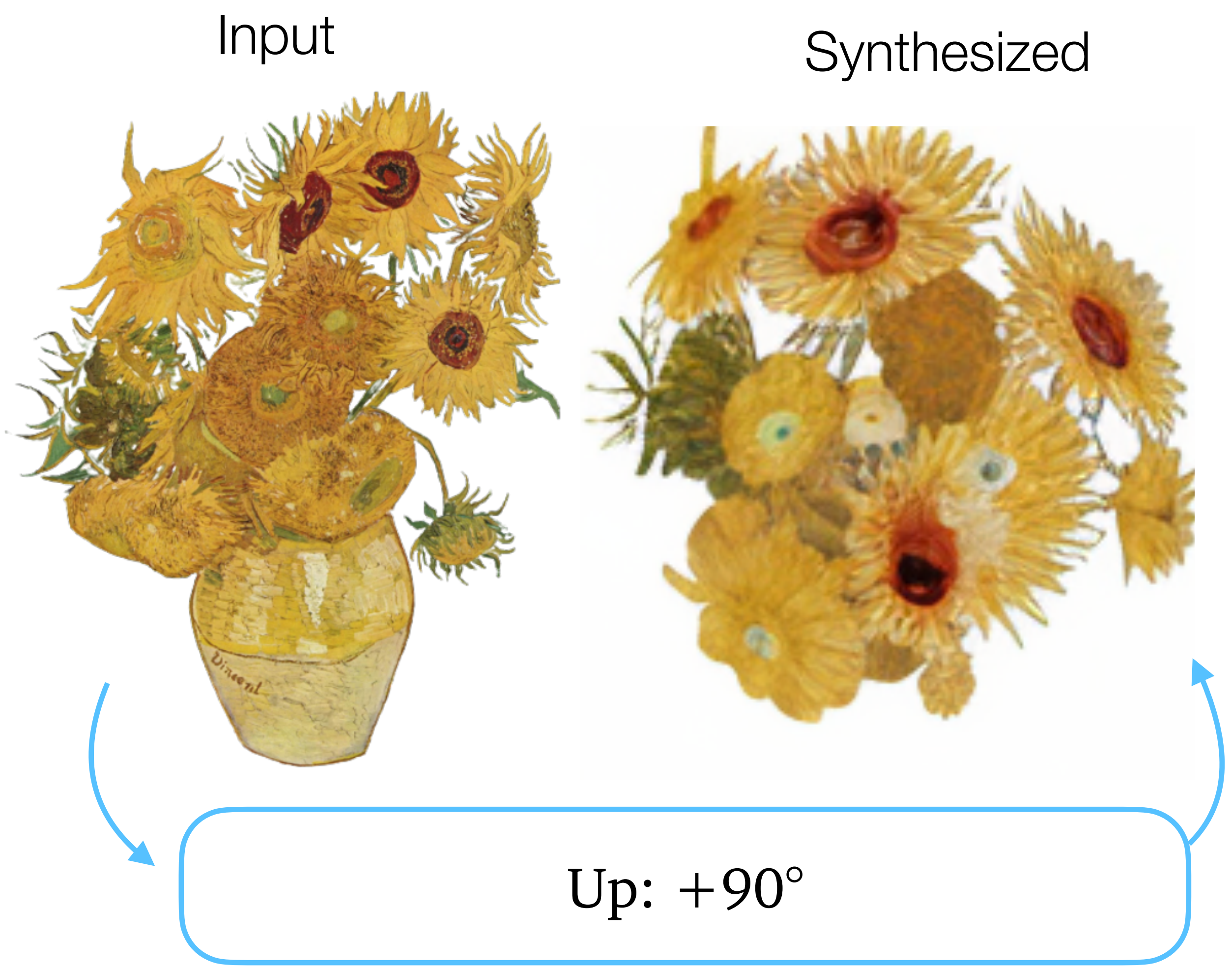
Input

Synthesized

Up: -45° Left: -60°

Liu, Wu, Van Hoorick, Tokmakov, Zakharov, Vondrick. ICCV 2023.

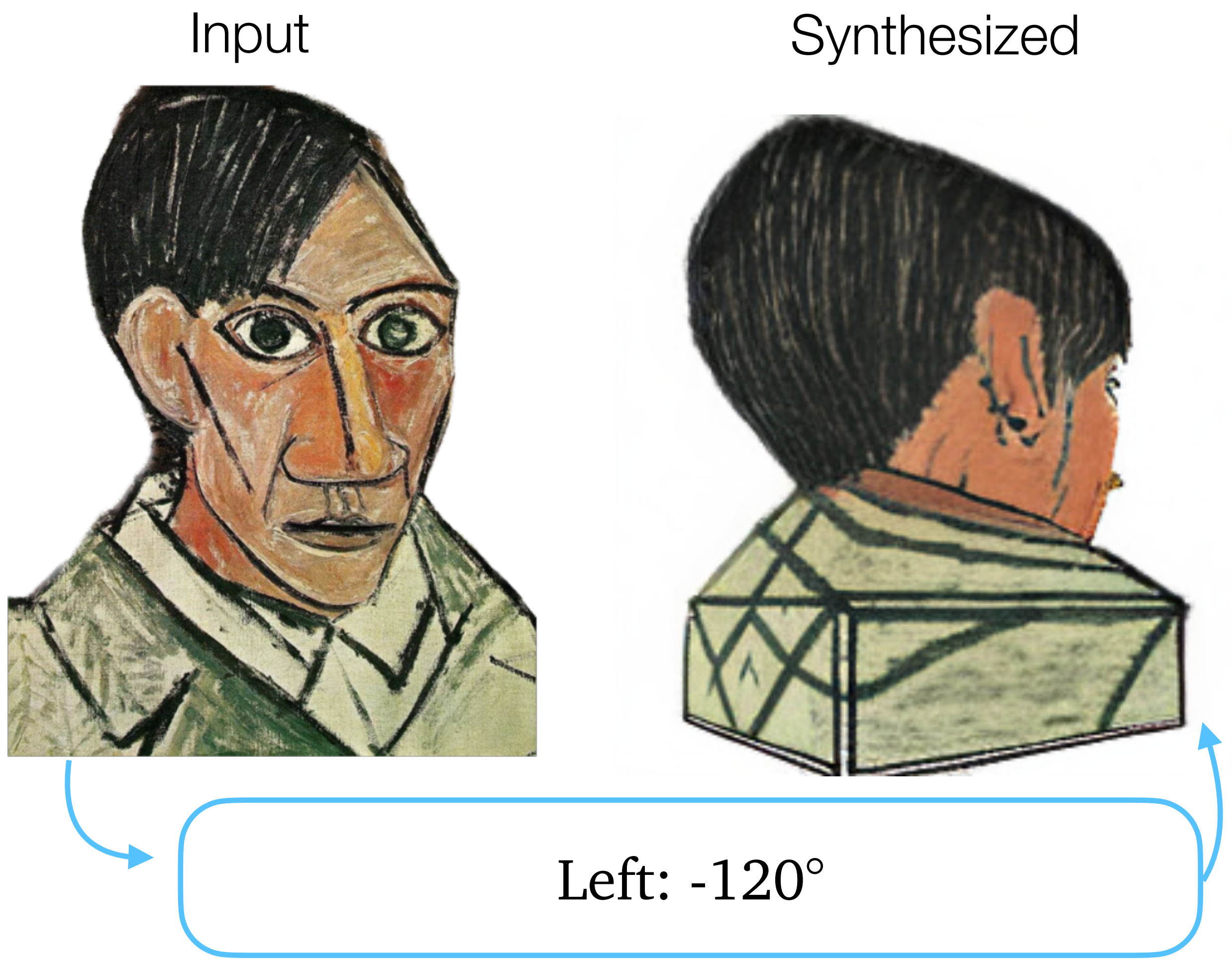
Zero-1-to-3



Input

Synthesized

Up: +90°



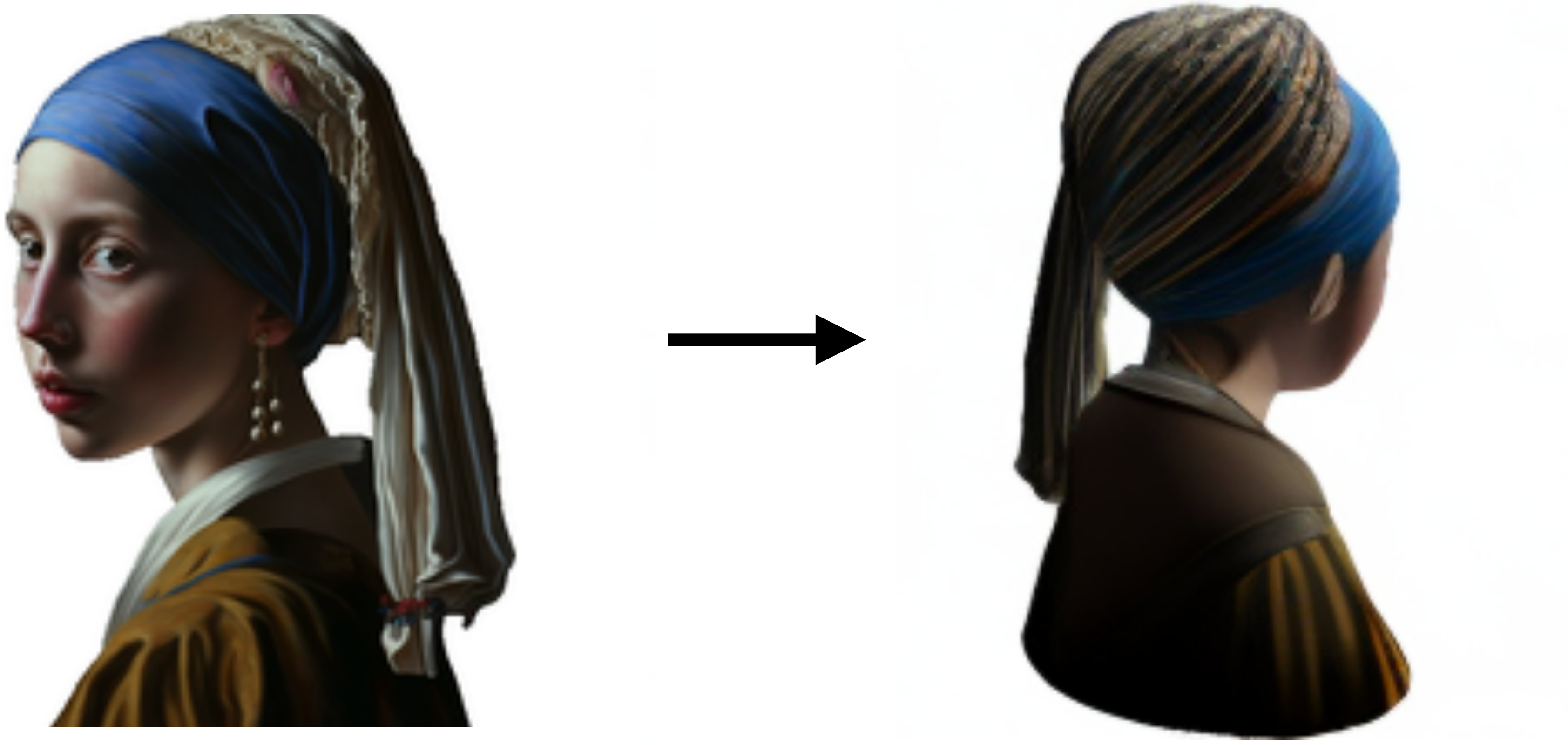
Input

Synthesized

Left: -120°

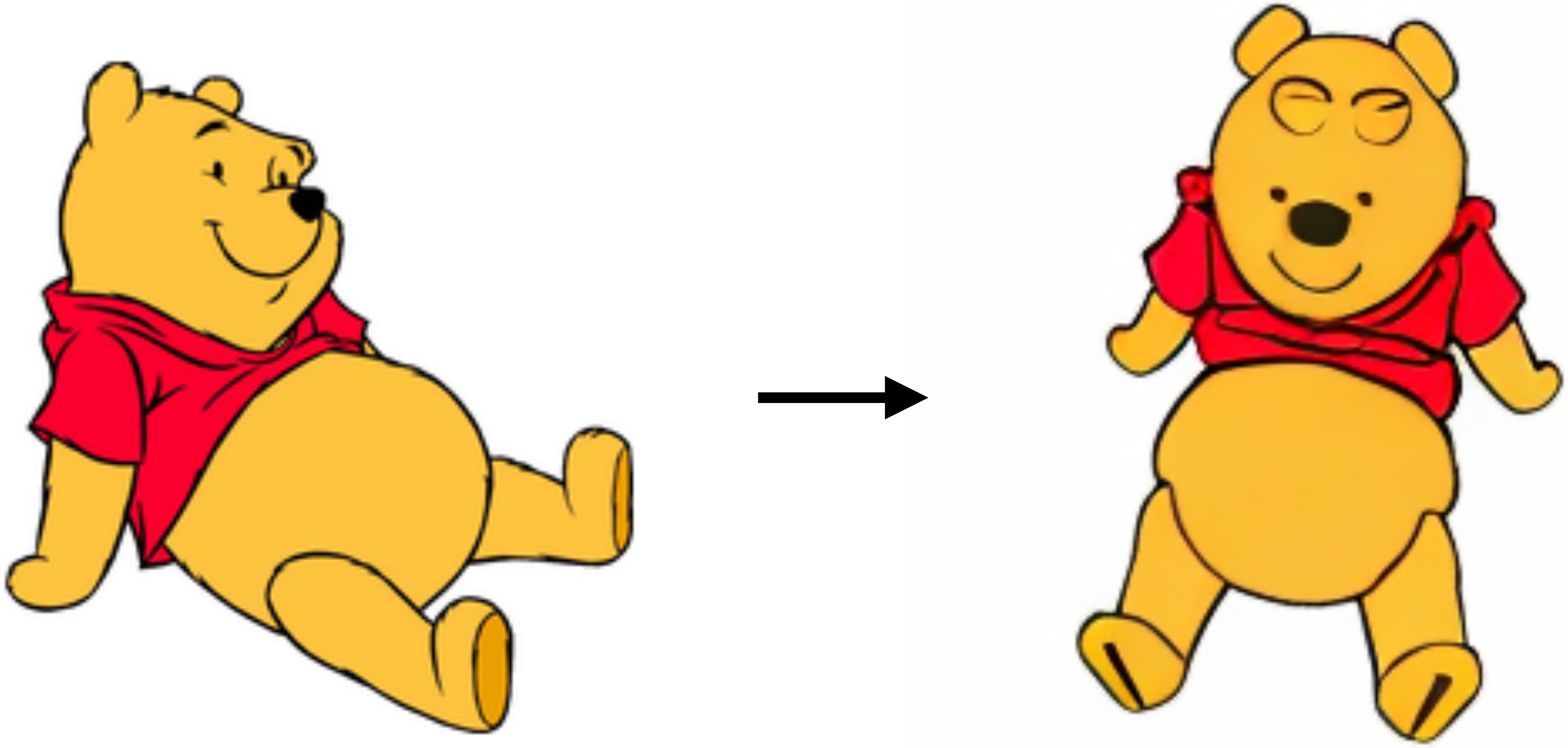
Liu, Wu, Van Hoorick, Tokmakov, Zakharov, Vondrick. ICCV 2023.

Oil Paintings



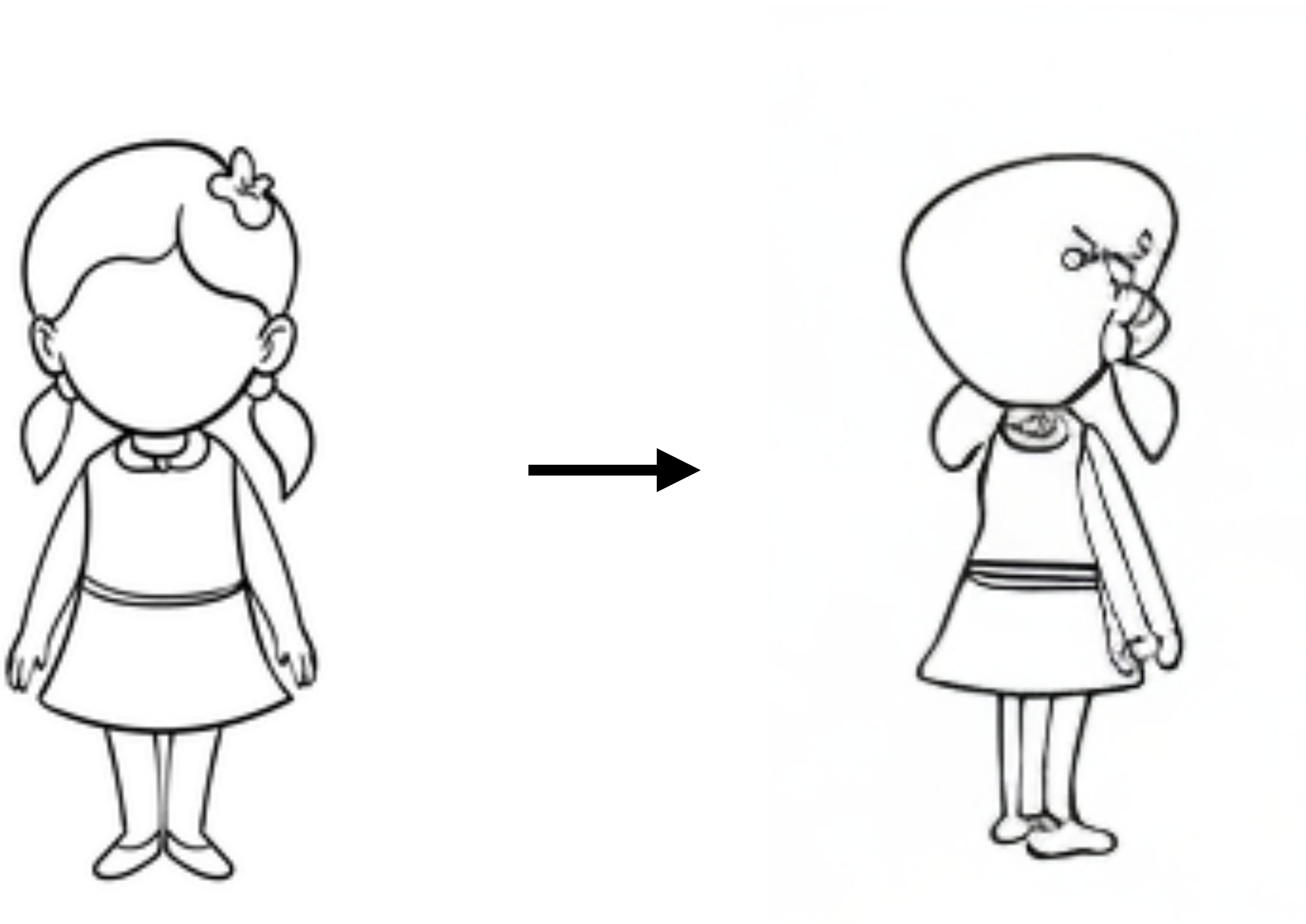
Deitke, et al. 2023.

Cartoons



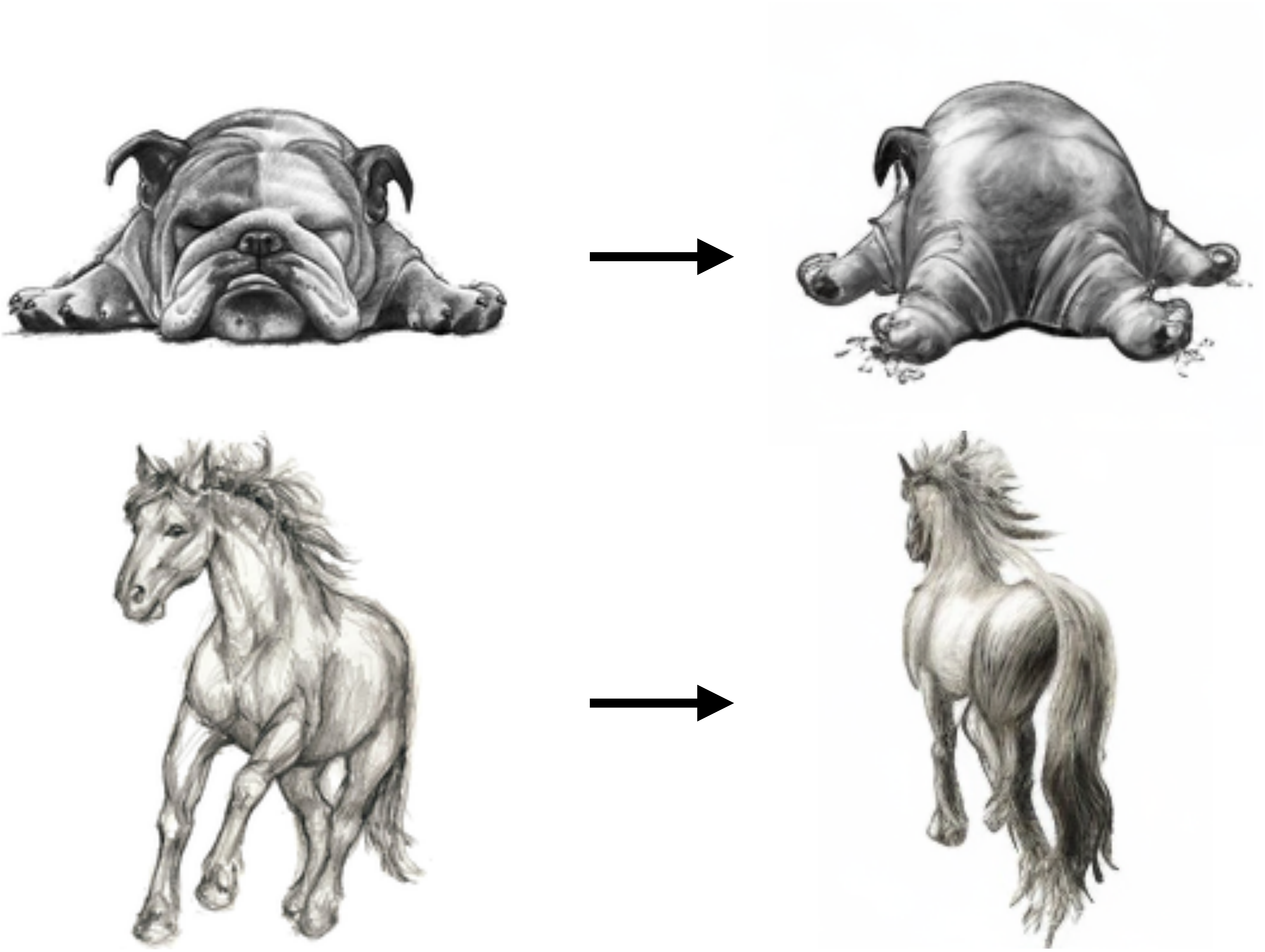
Deitke, et al. 2023.

Line Drawings



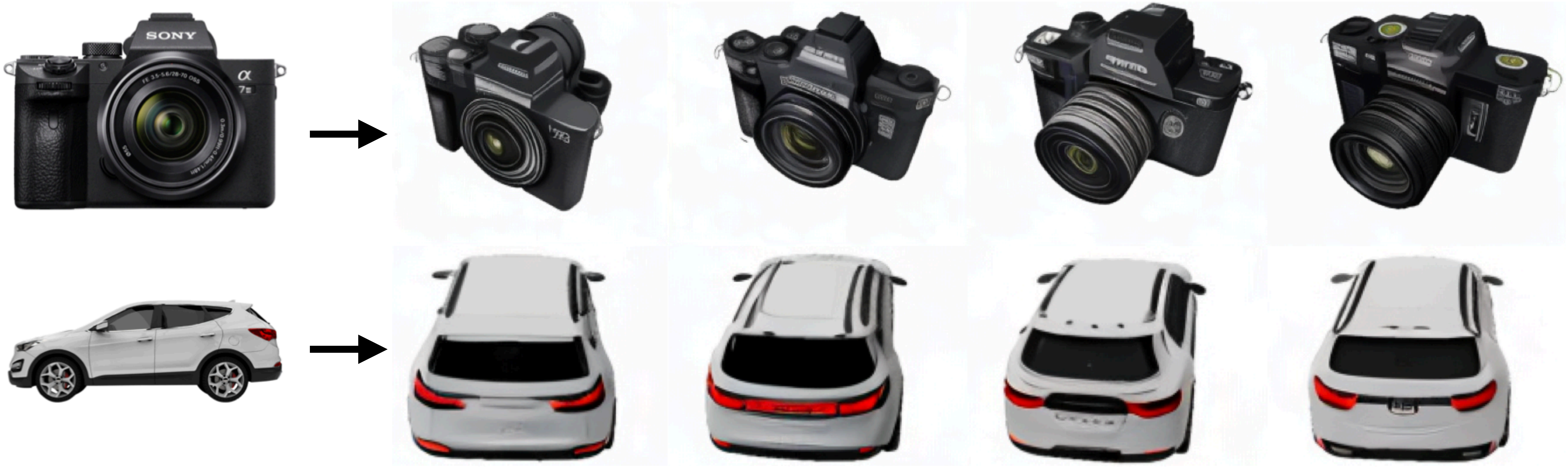
Deitke, et al. 2023.

Sketches



Deitke, et al. 2023.

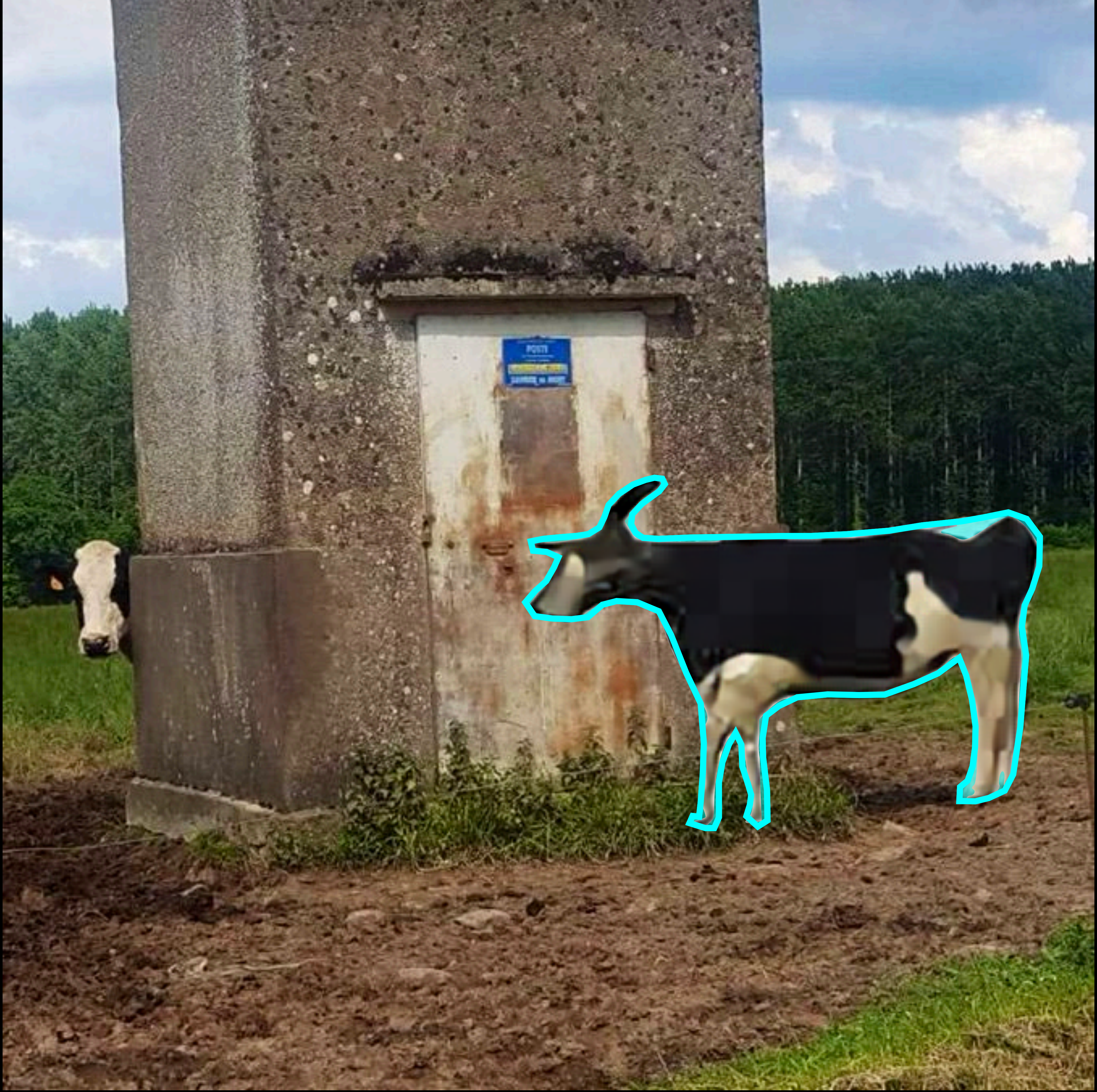
Diversity from Occlusions

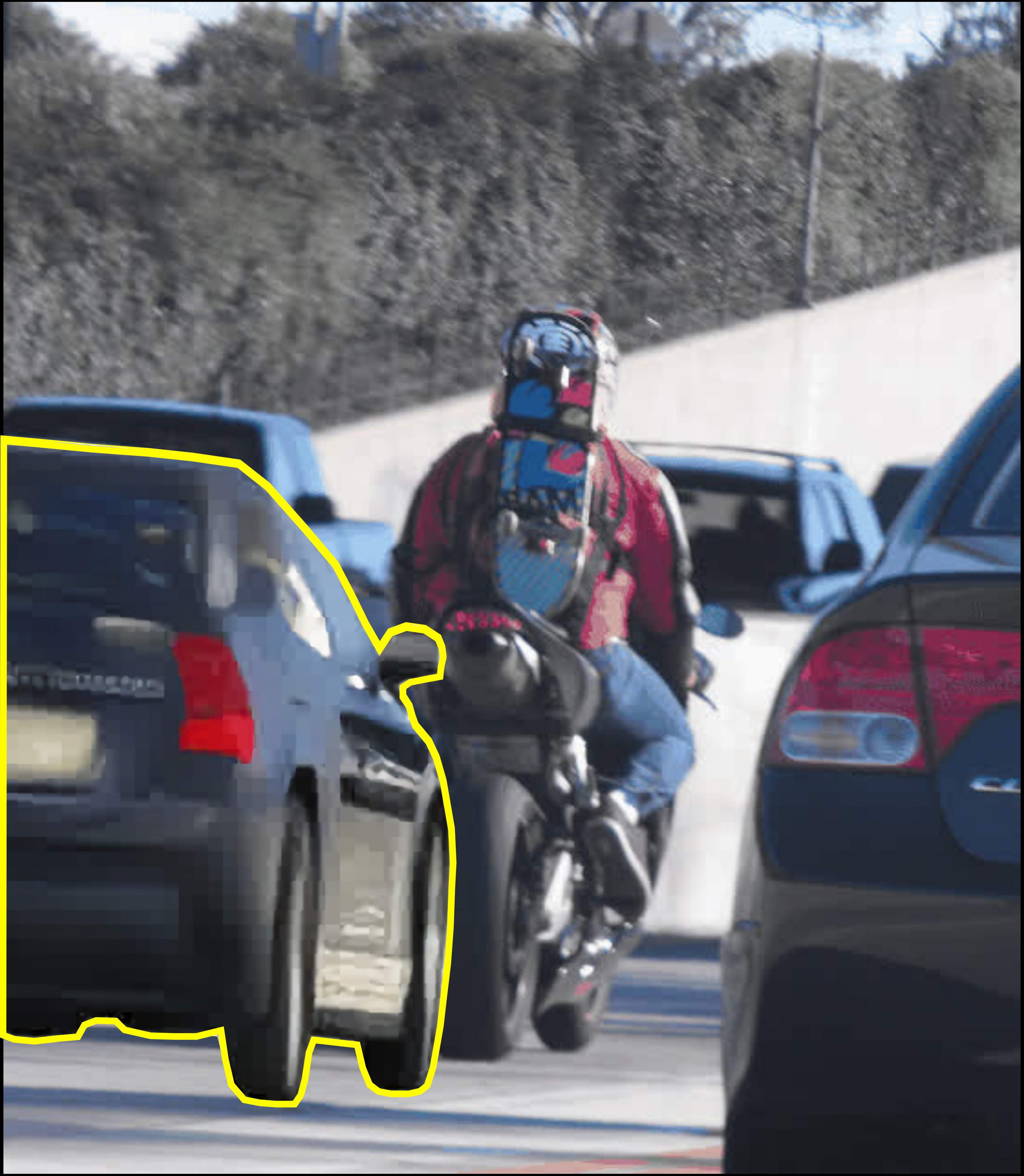


Input View

New View (Different Samples)

Liu, Wu, Van Hoorick, Tokmakov, Zakharov, Vondrick. ICCV 2023.

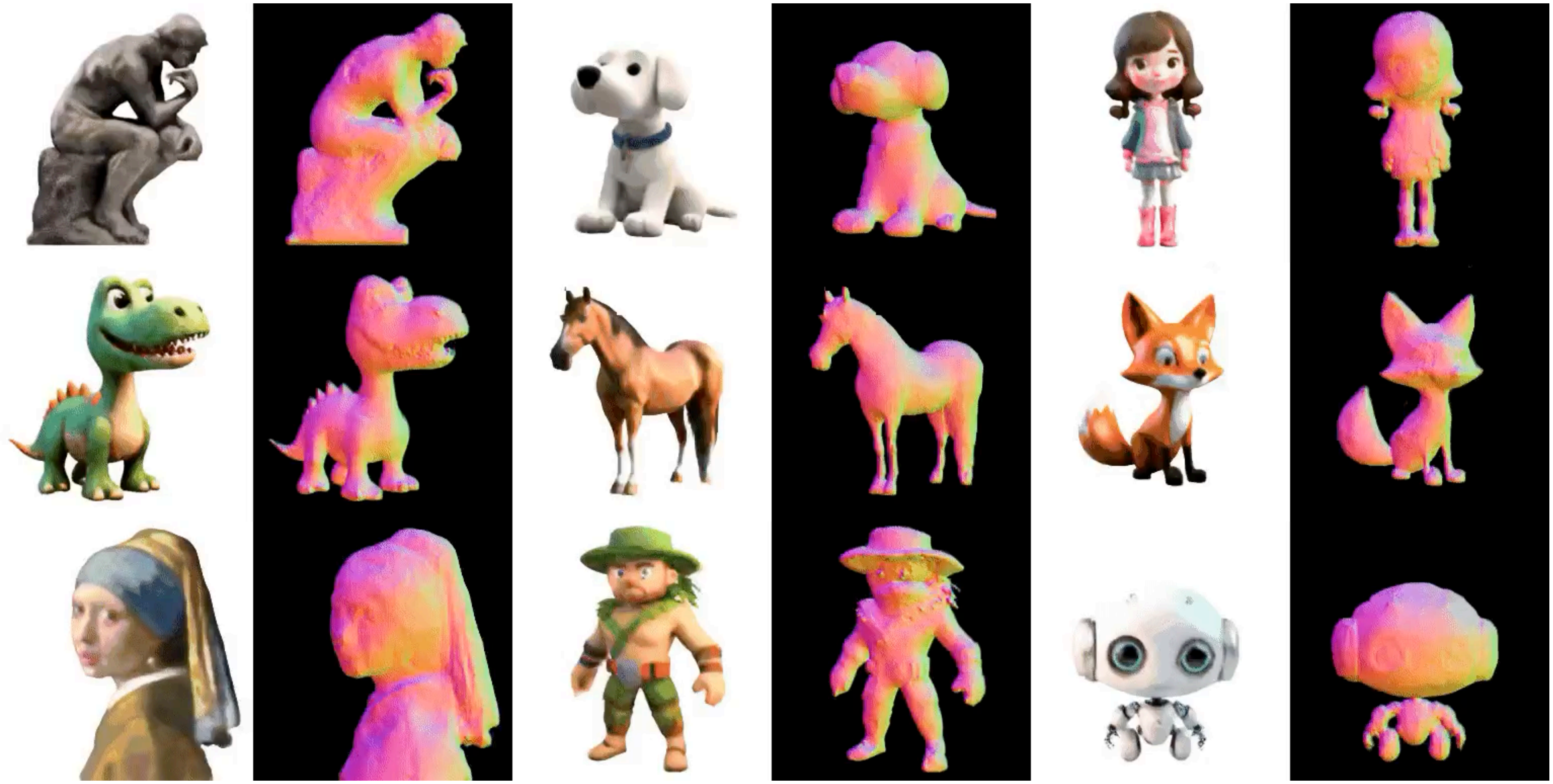

POSTE



Ozguroglu, Liu, Suris, Chen, Dave, Tokmakov, Vondrick. In Submission.

Ozguroglu, Liu, Suris, Chen, Dave, Tokmakov, Vondrick. In Submission.

3D Reconstruction



@threestudio-project on Github

Zero123... Go!

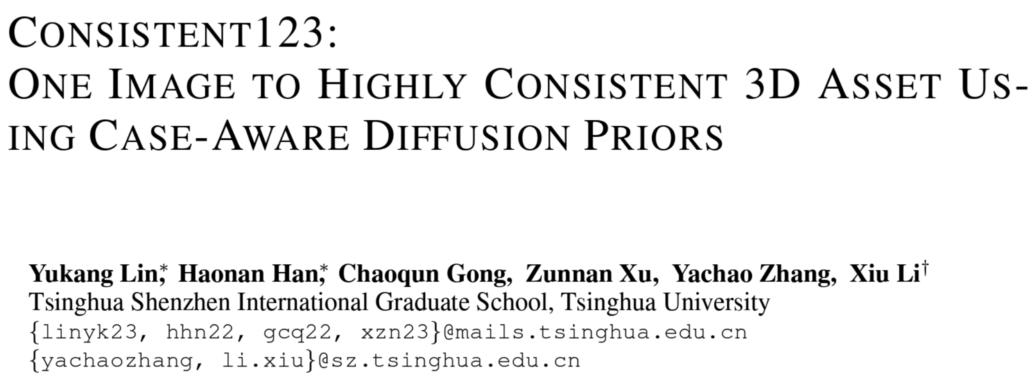
At launch, 140,000 downloads on HuggingFace

One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization

Minghua Liu[1*], Chao Xu[2*], Haian Jin[3,4*], Linghao Chen[1,4*], Mukund Varma T[5], Zexiang Xu[6], Hao Su[1]

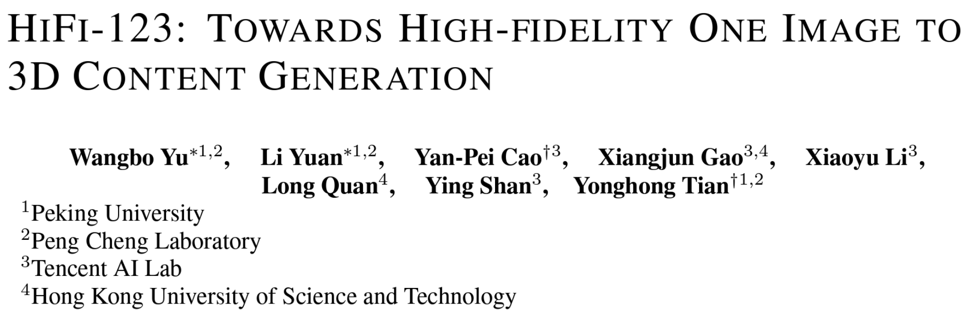
[1]UC San Diego, [2]UCLA, [3]Cornell University, [4]Zhejiang University, [5]IIT Madras, [6]Adobe Research

* Equal contribution



CONSISTENT123: ONE IMAGE TO HIGHLY CONSISTENT 3D ASSET USING CASE-AWARE DIFFUSION PRIORS

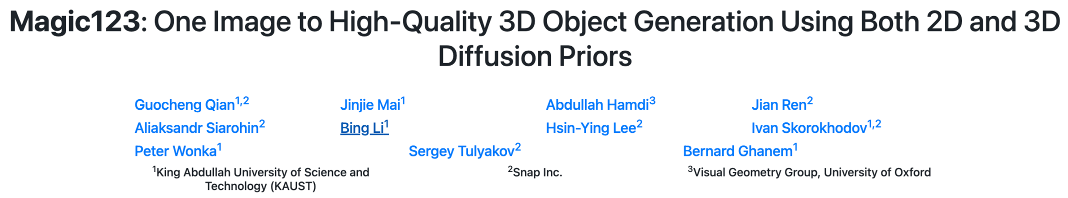
Yukang Lin*, Haonan Han*, Chaoqun Gong, Zunnan Xu, Yachao Zhang, Xiu Li†
Tsinghua Shenzhen International Graduate School, Tsinghua University
{linyk23, hhn22, gcq22, xzn23}@mails.tsinghua.edu.cn
{yachaozhang, li.xiu}@sz.tsinghua.edu.cn



HIFI-123: TOWARDS HIGH-FIDELITY ONE IMAGE TO 3D CONTENT GENERATION

Wangbo Yu[*1,2], Li Yuan[*1,2], Yan-Pei Cao[†3], Xiangjun Gao[3,4], Xiaoyu Li[3], Long Quan[4], Ying Shan[3], Yonghong Tian[†1,2]

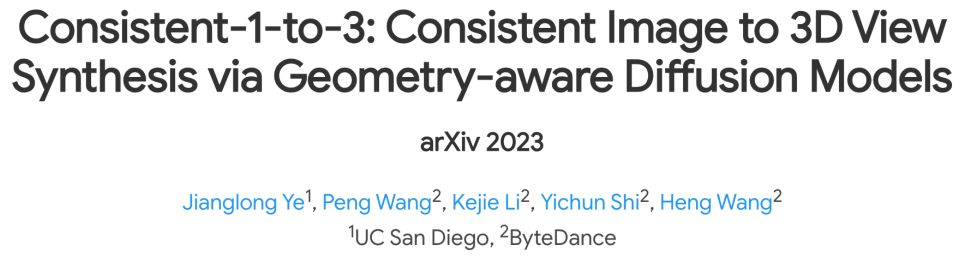
[1]Peking University
[2]Peng Cheng Laboratory
[3]Tencent AI Lab
[4]Hong Kong University of Science and Technology



Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors

Guocheng Qian[1,2] Jinjie Mai[1] Abdullah Hamdi[3] Jian Ren[2]
Aliaksandr Siarohin[2] Bing Li[1] Hsin-Ying Lee[2] Ivan Skorokhodov[1,2]
Peter Wonka[1] Sergey Tulyakov[2] Bernard Ghanem[1]

[1]King Abdullah University of Science and Technology (KAUST) [2]Snap Inc. [3]Visual Geometry Group, University of Oxford

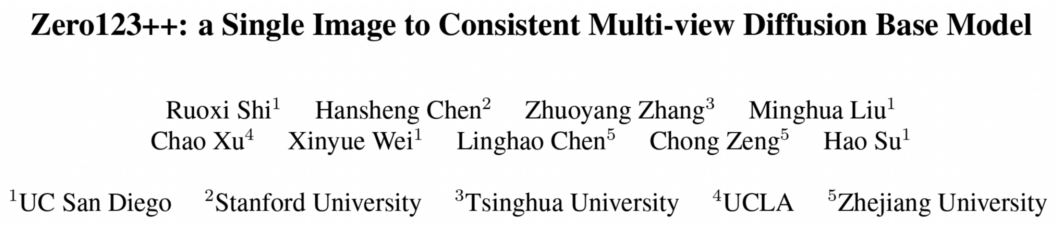


Consistent-1-to-3: Consistent Image to 3D View Synthesis via Geometry-aware Diffusion Models

arXiv 2023

Jianglong Ye[1], Peng Wang[2], Kejie Li[2], Yichun Shi[2], Heng Wang[2]

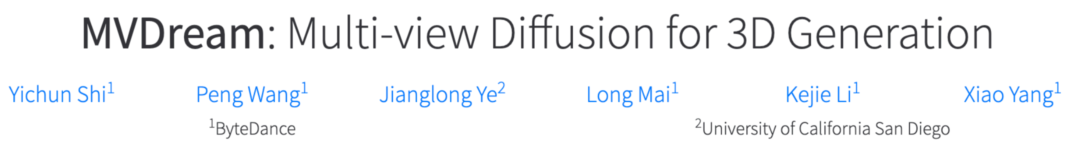
[1]UC San Diego, [2]ByteDance



Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model

Ruoxi Shi[1] Hansheng Chen[2] Zhuoyang Zhang[3] Minghua Liu[1]
Chao Xu[4] Xinyue Wei[1] Linghao Chen[5] Chong Zeng[5] Hao Su[1]

[1]UC San Diego [2]Stanford University [3]Tsinghua University [4]UCLA [5]Zhejiang University



MVDream: Multi-view Diffusion for 3D Generation

Yichun Shi[1] Peng Wang[1] Jianglong Ye[2] Long Mai[1] Kejie Li[1] Xiao Yang[1]

[1]ByteDance [2]University of California San Diego

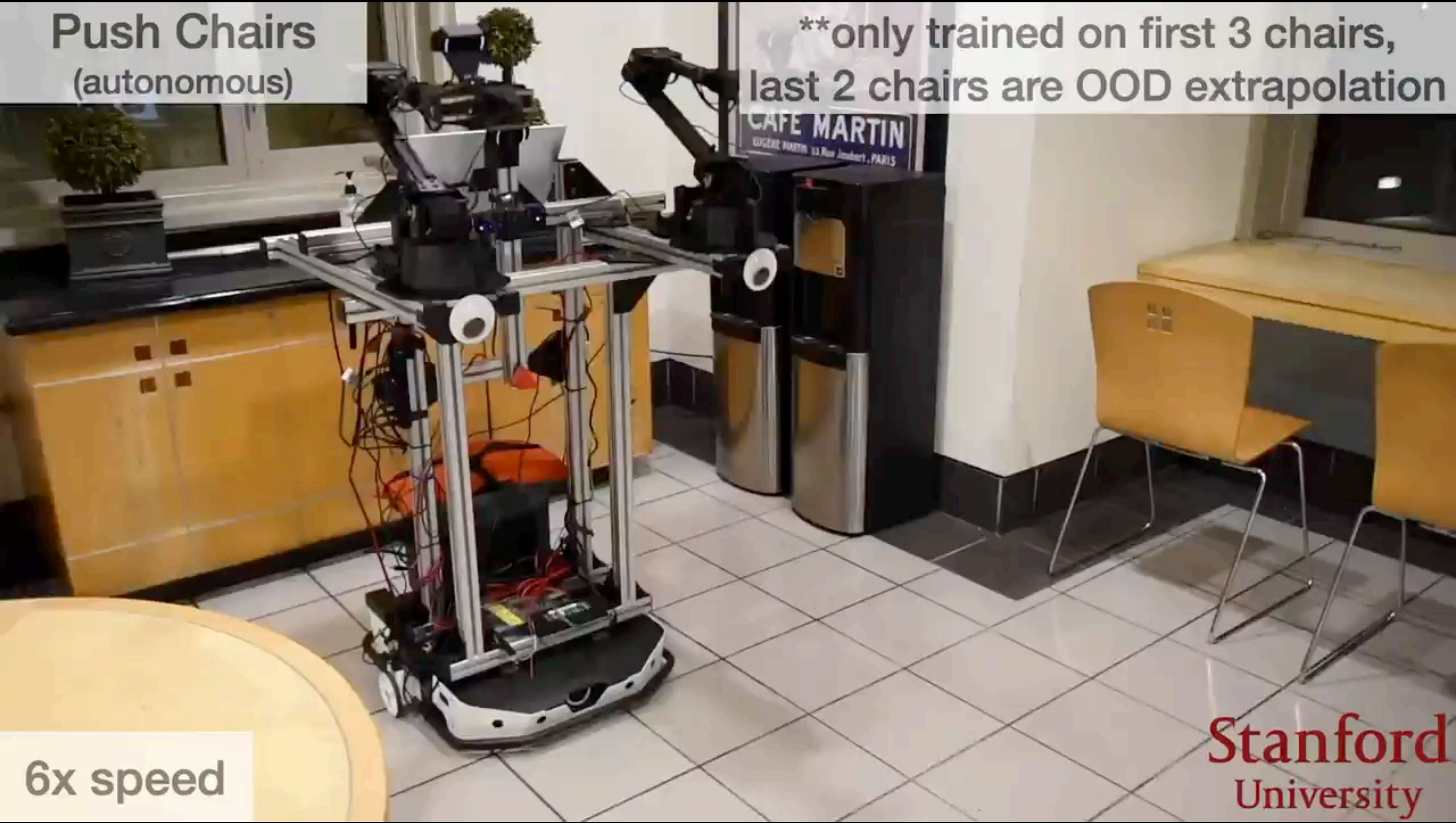
DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation

Arxiv 2023

Jiaxiang Tang[1], Jiawei Ren[2], Hang Zhou[3], Ziwei Liu[2], Gang Zeng[1]

[1] Peking University [2] S-Lab, Nanyang Technological University [3] Baidu

Generalization in Robotics



Push Chairs
(autonomous)

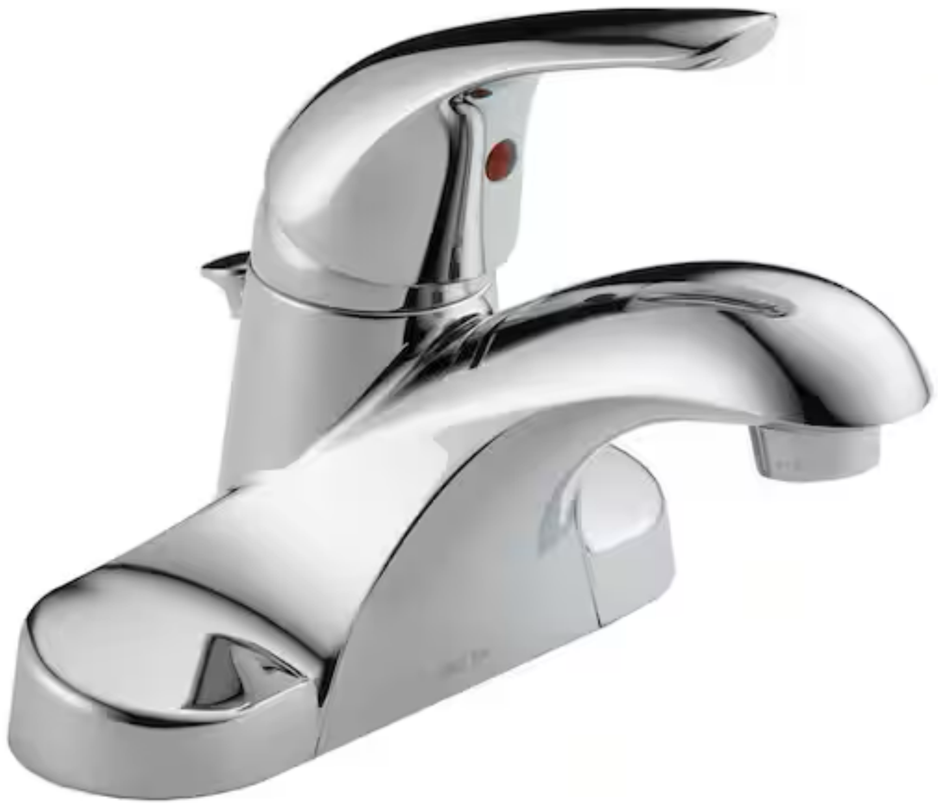
**only trained on first 3 chairs, last 2 chairs are OOD extrapolation

6x speed

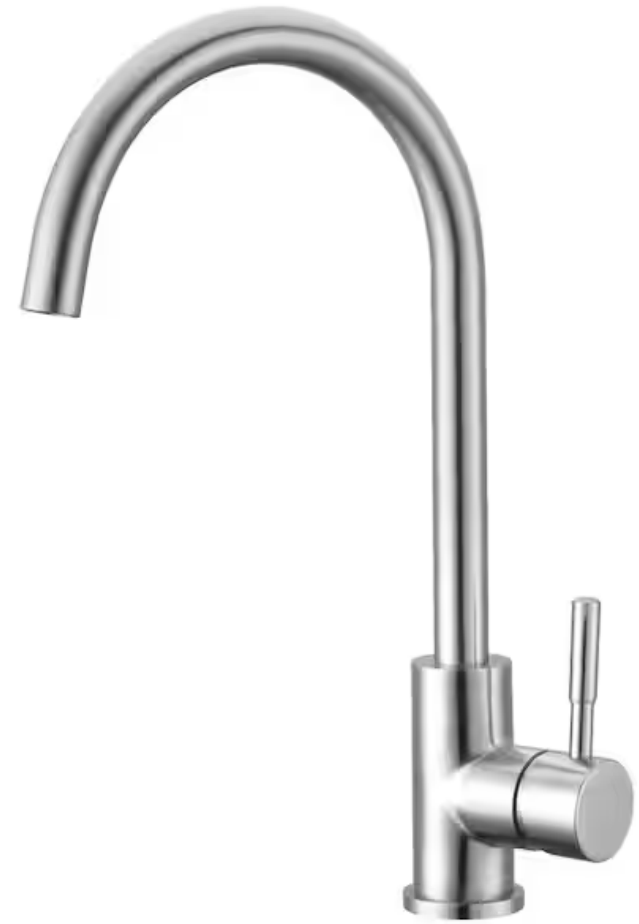
Stanford University

Generalization in Robotics

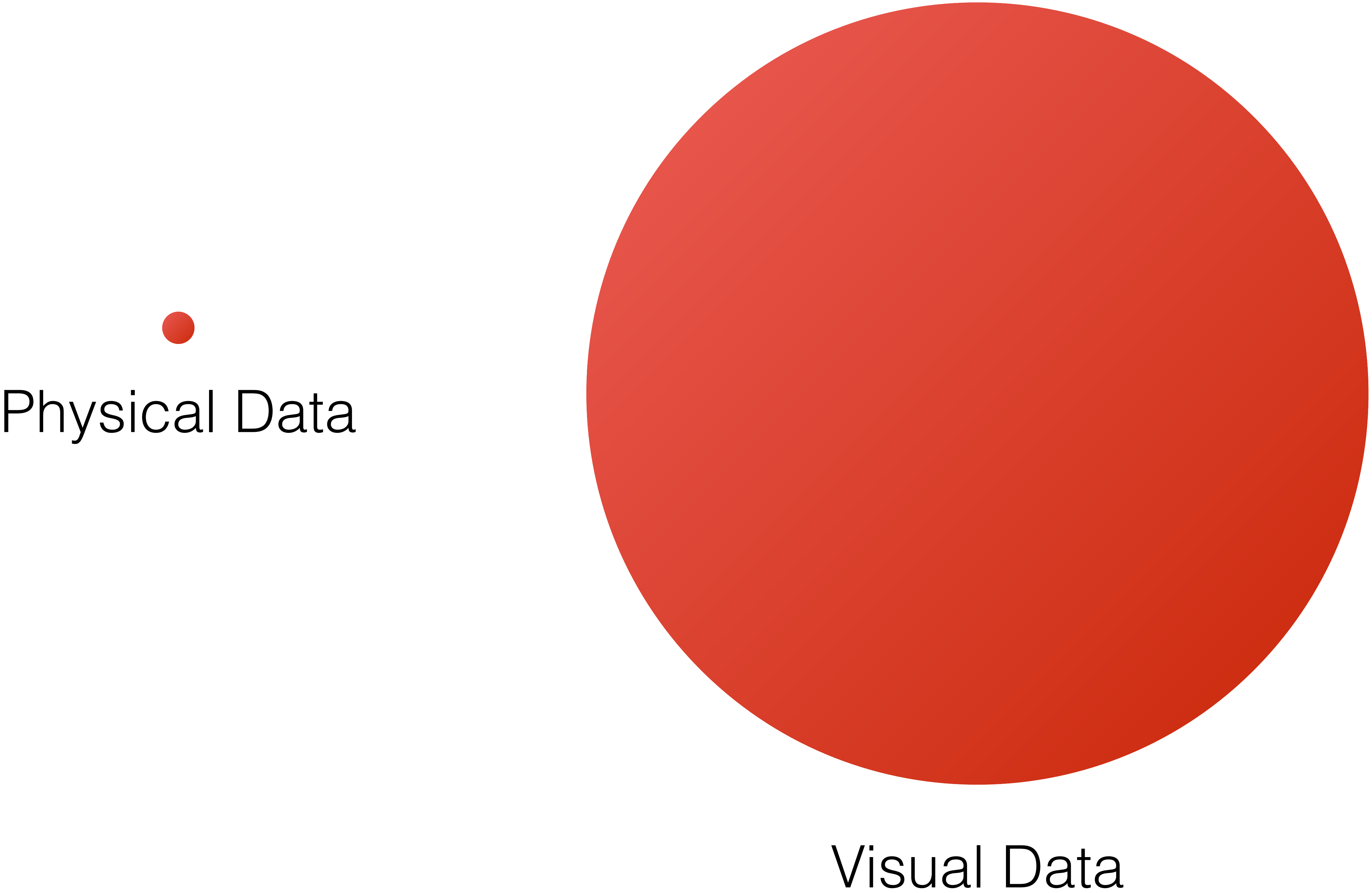
Training



Testing



Size of Data

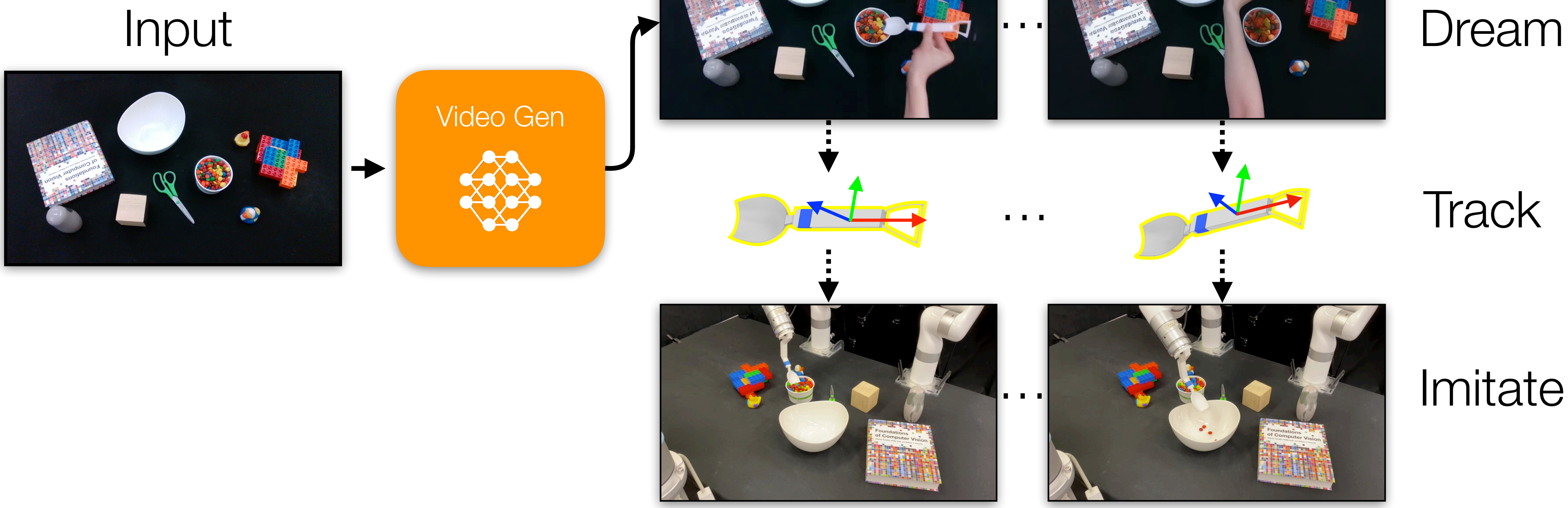


Physical Data

Visual Data

OpenAI Sora

Behavior via Video Generation



Input

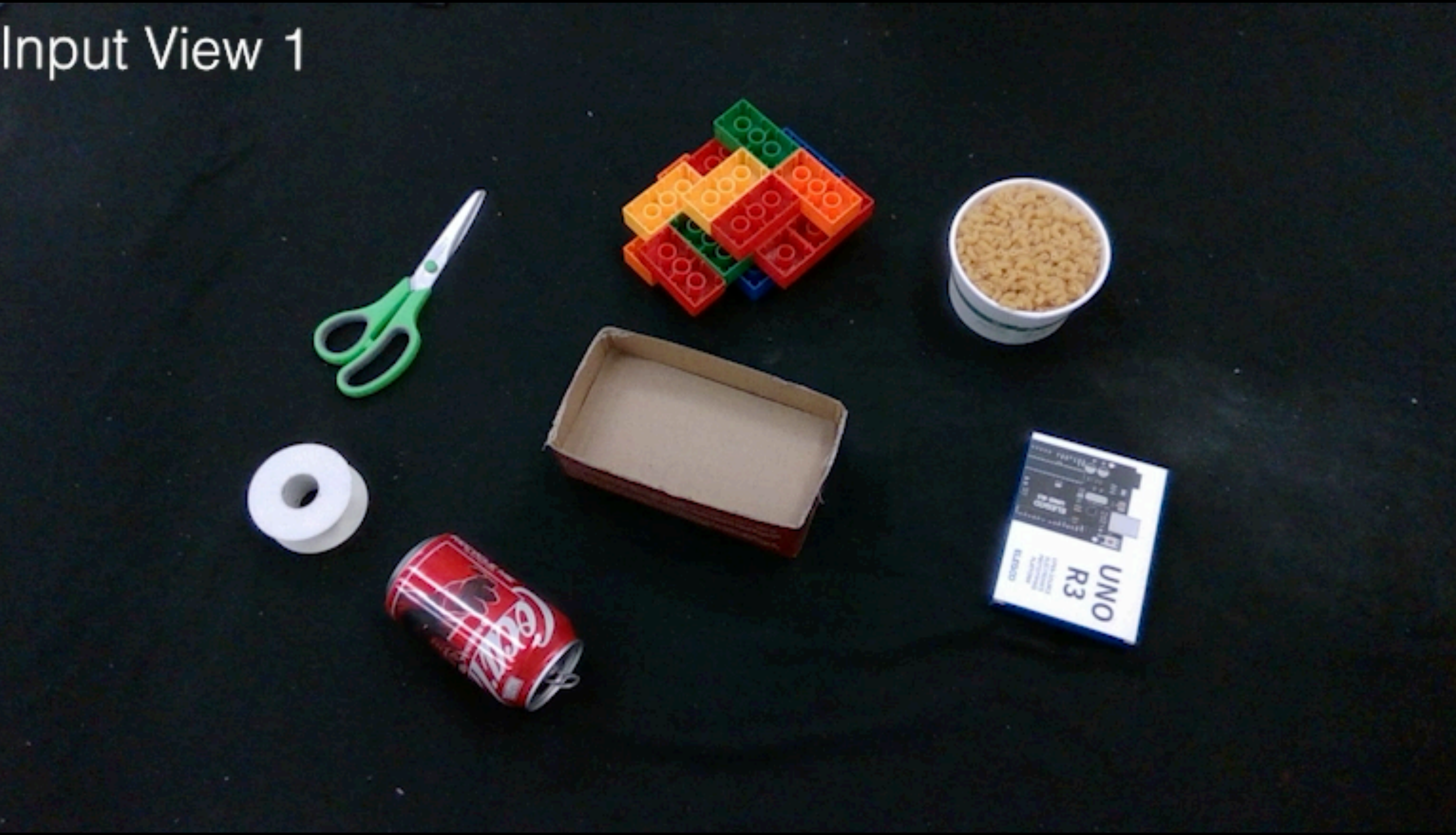
Video Gen

Dream

Track

Imitate

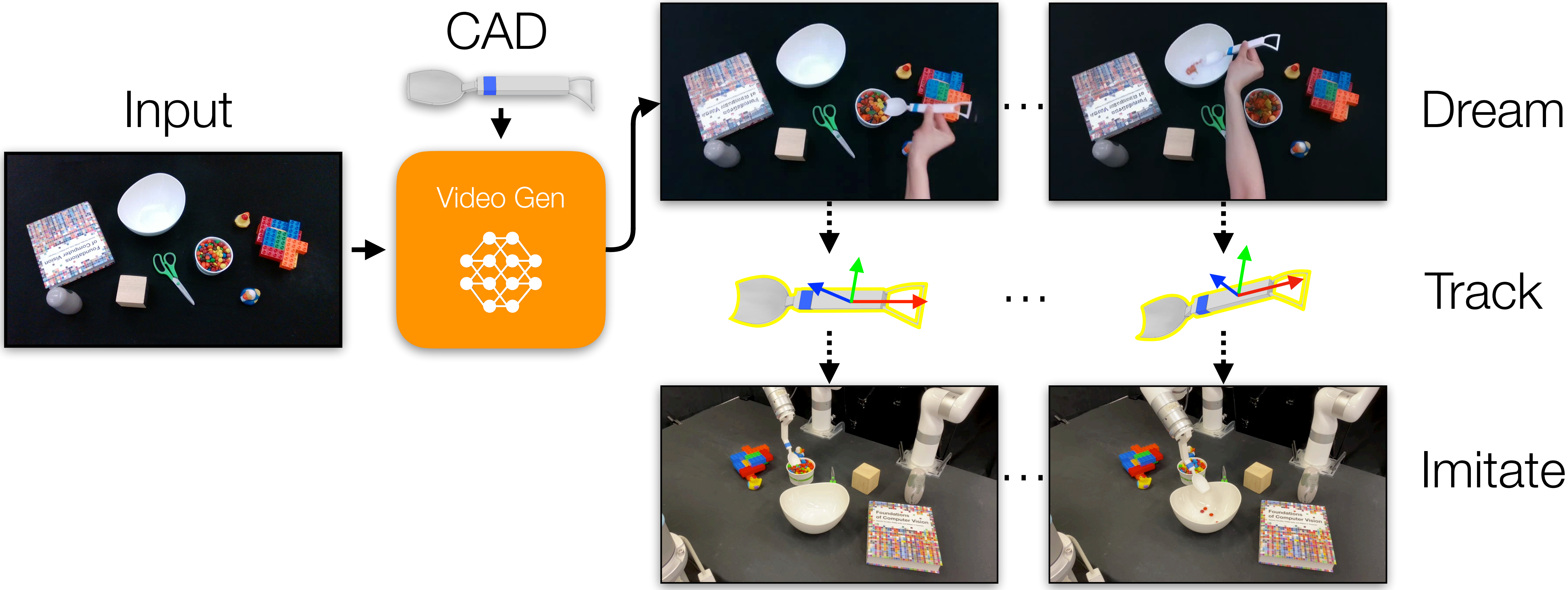
Perception → Predict → Act



Perception → Predict → Act

Perception → Predict → Act

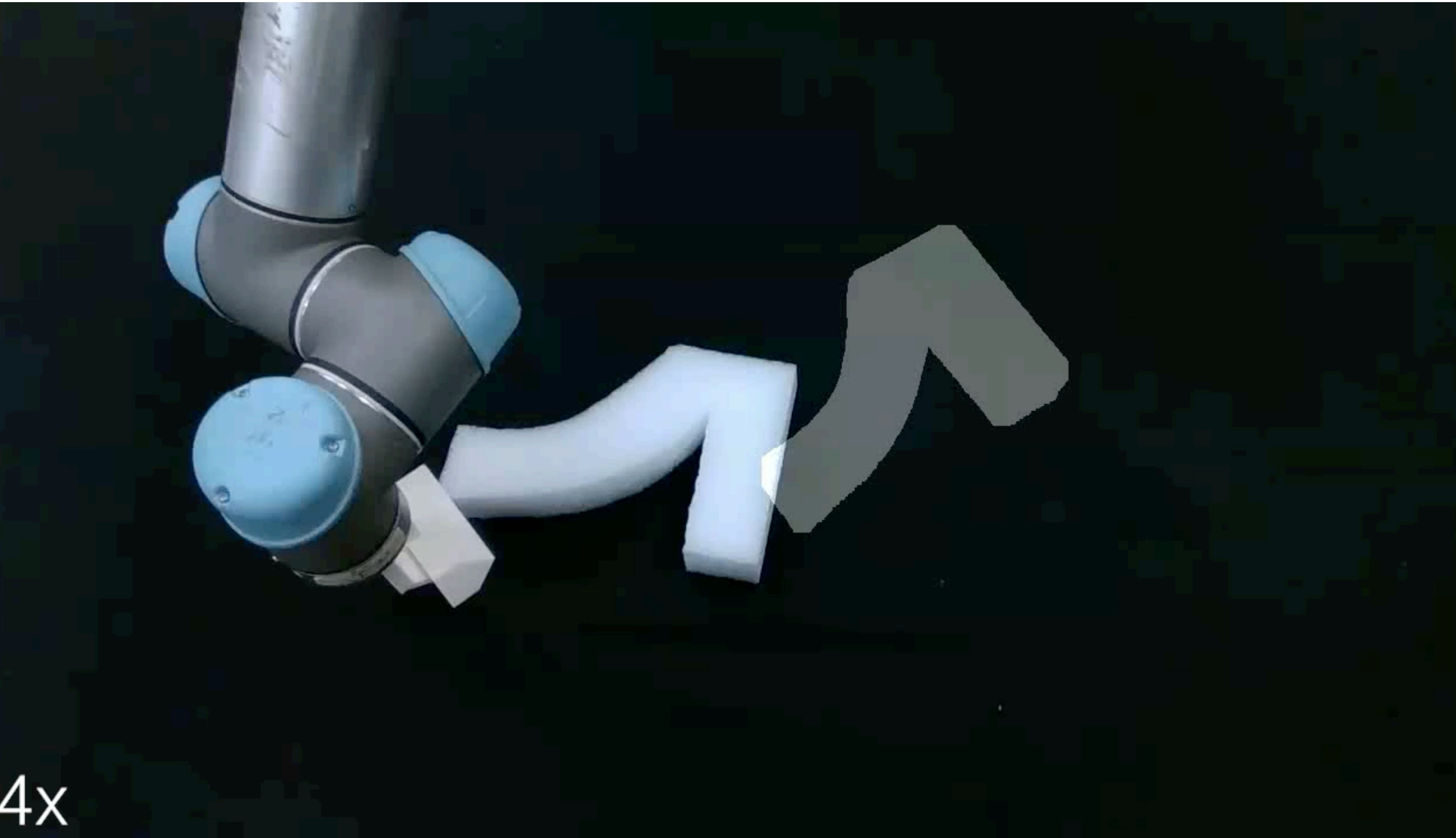
Behavior via Video Generation



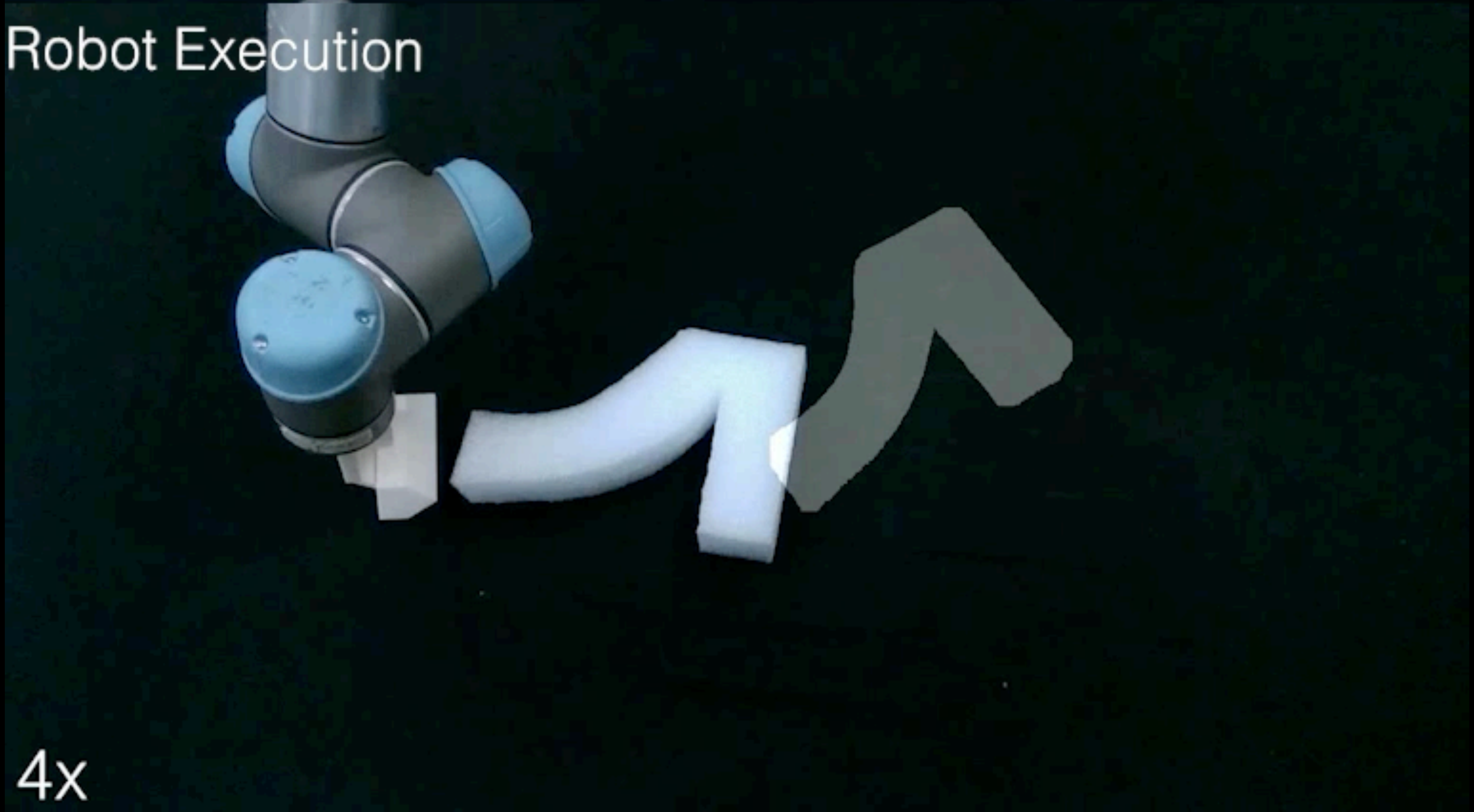
DreaImitate

Dreamitate

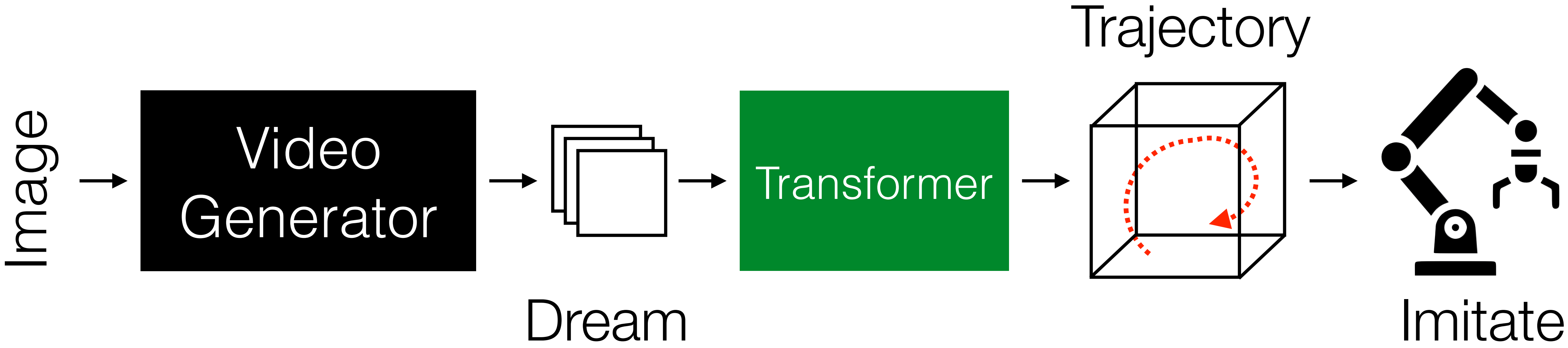
Behavior Cloning Only



Generated Video


Robot Execution
4x

Dreamitate v2



Image

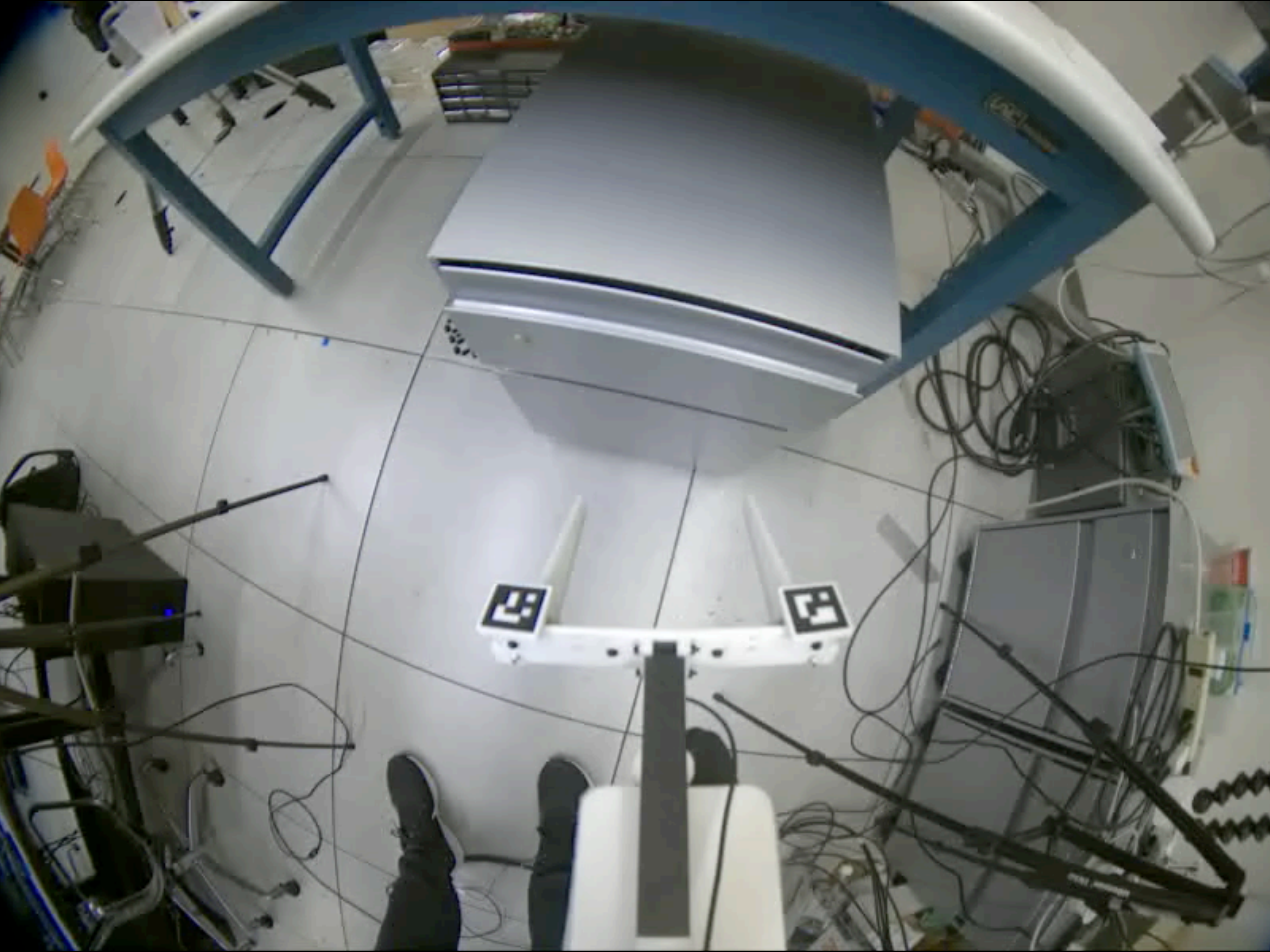
Video Generator

Dream

Transformer

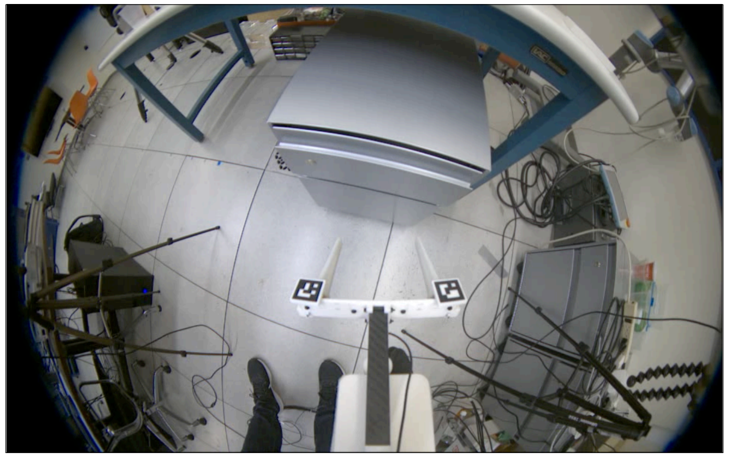
Trajectory

Imitate



Human Demonstrations

Align video generator by fine-tuning with just ~300 human demonstrations



Open Drawer

Pick and Place

M&Ms to Cup

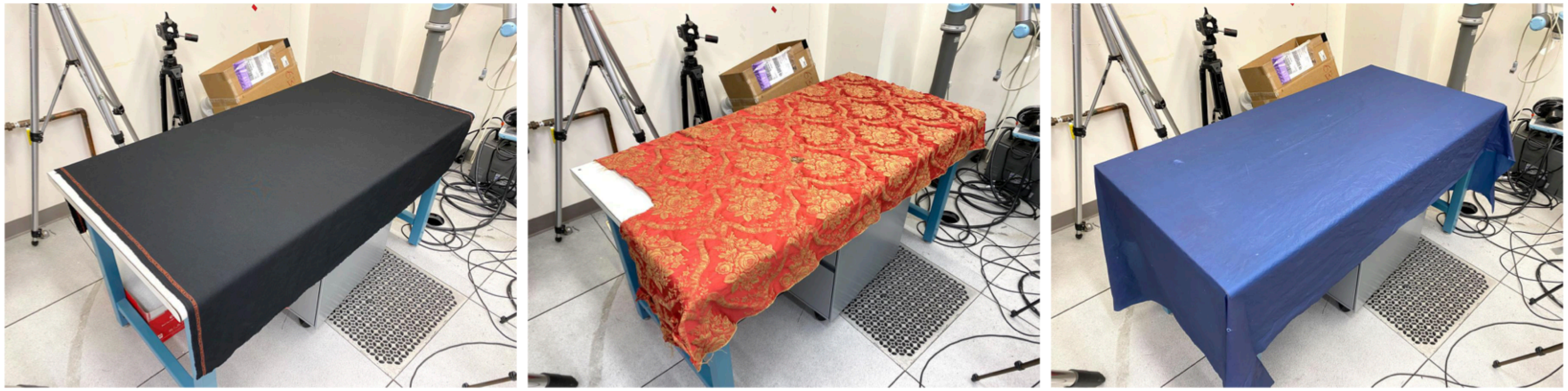
Upright Object

Stack Cups

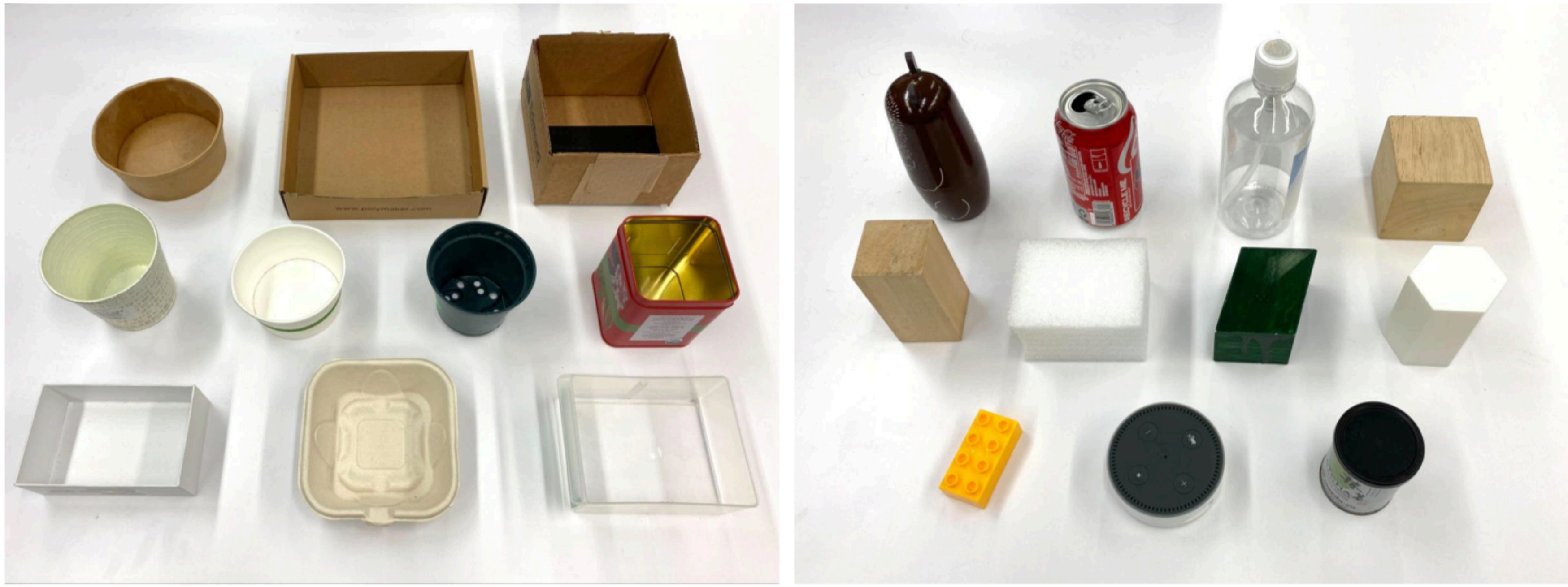
Generalization

Training Background

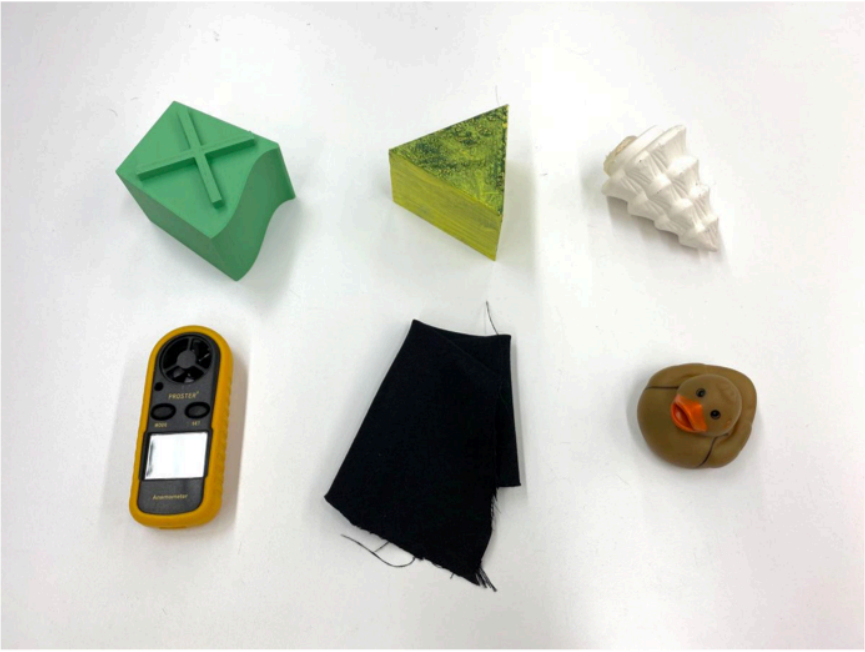
Unseen Backgrounds



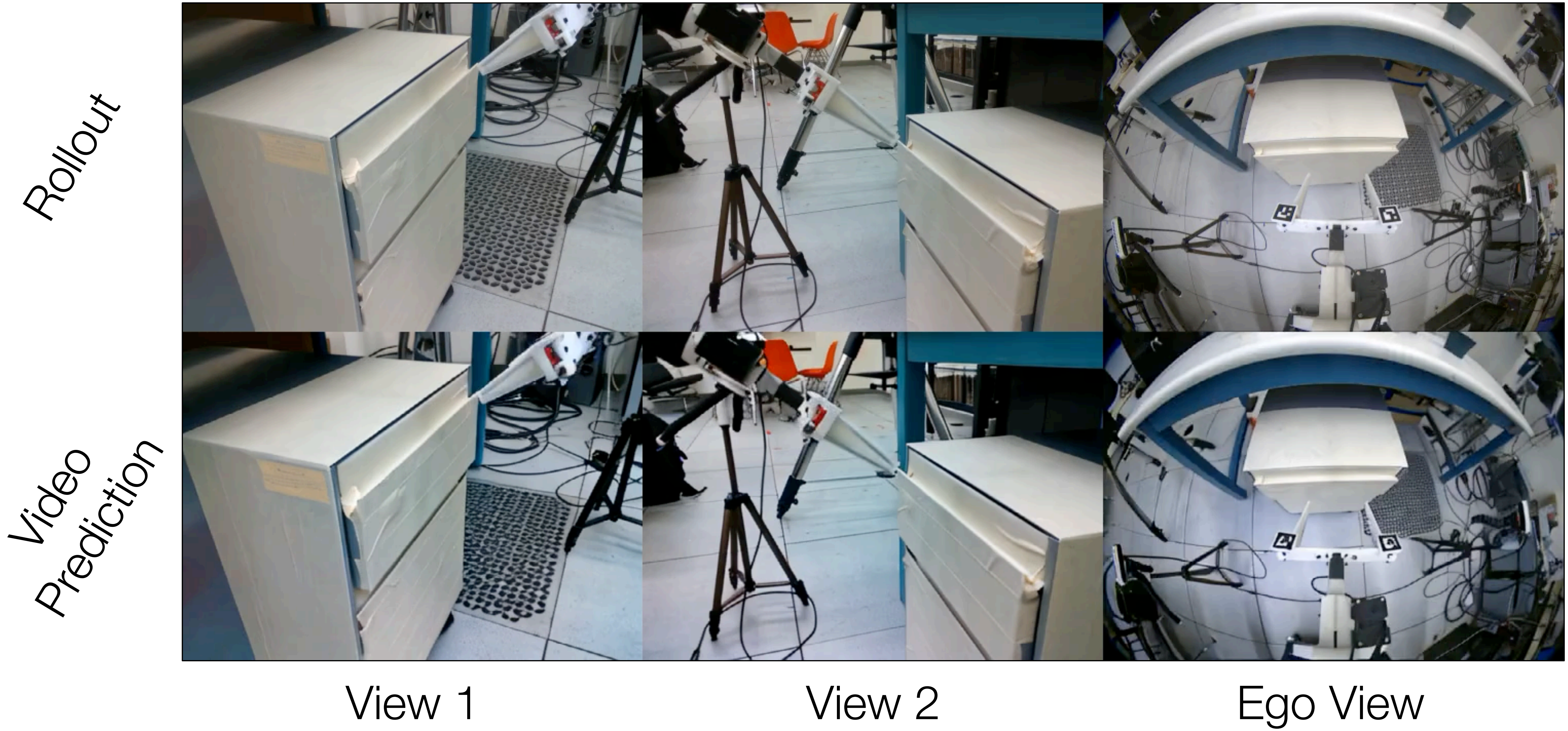
Training Set /
Vary Object Location Set



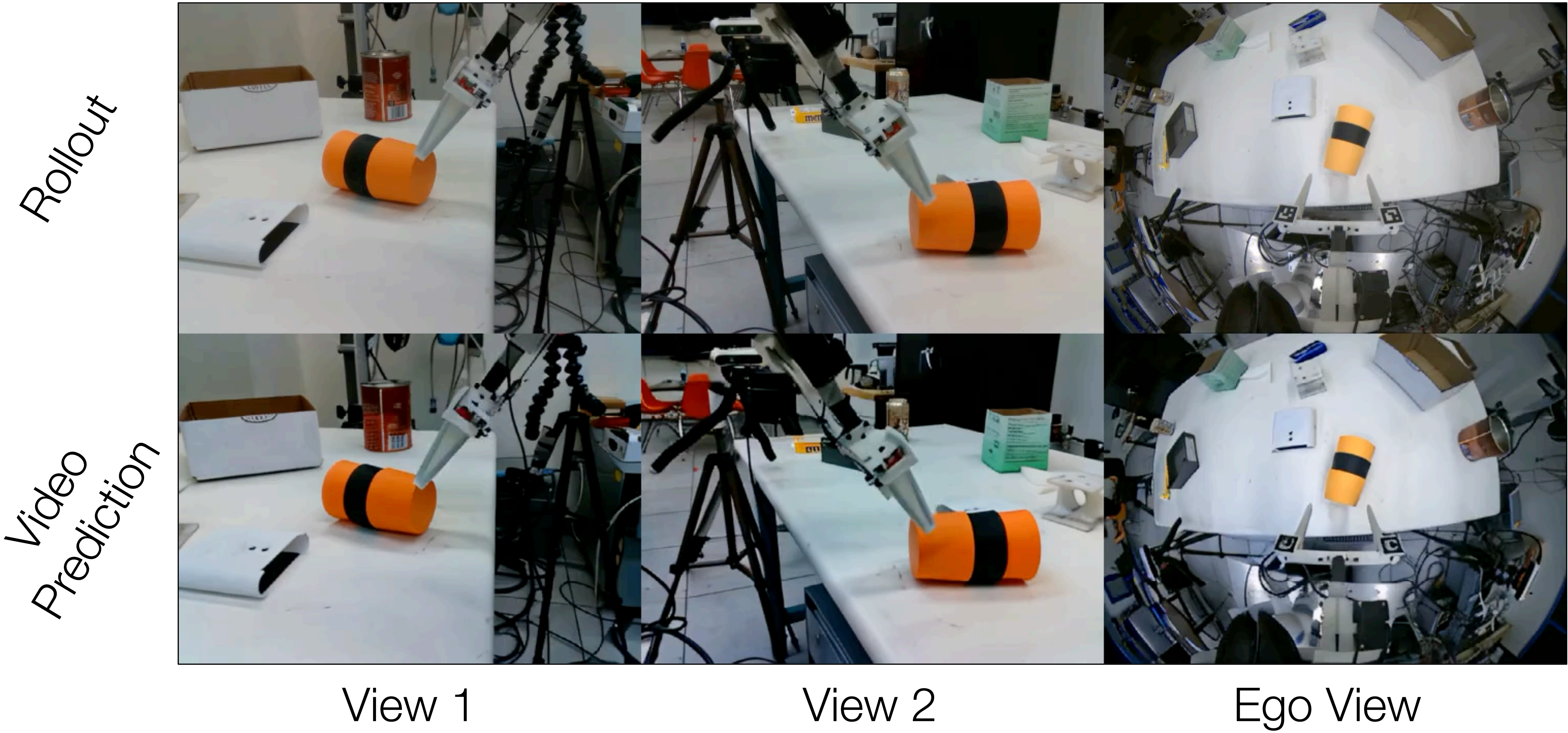
Unseen Objects Set



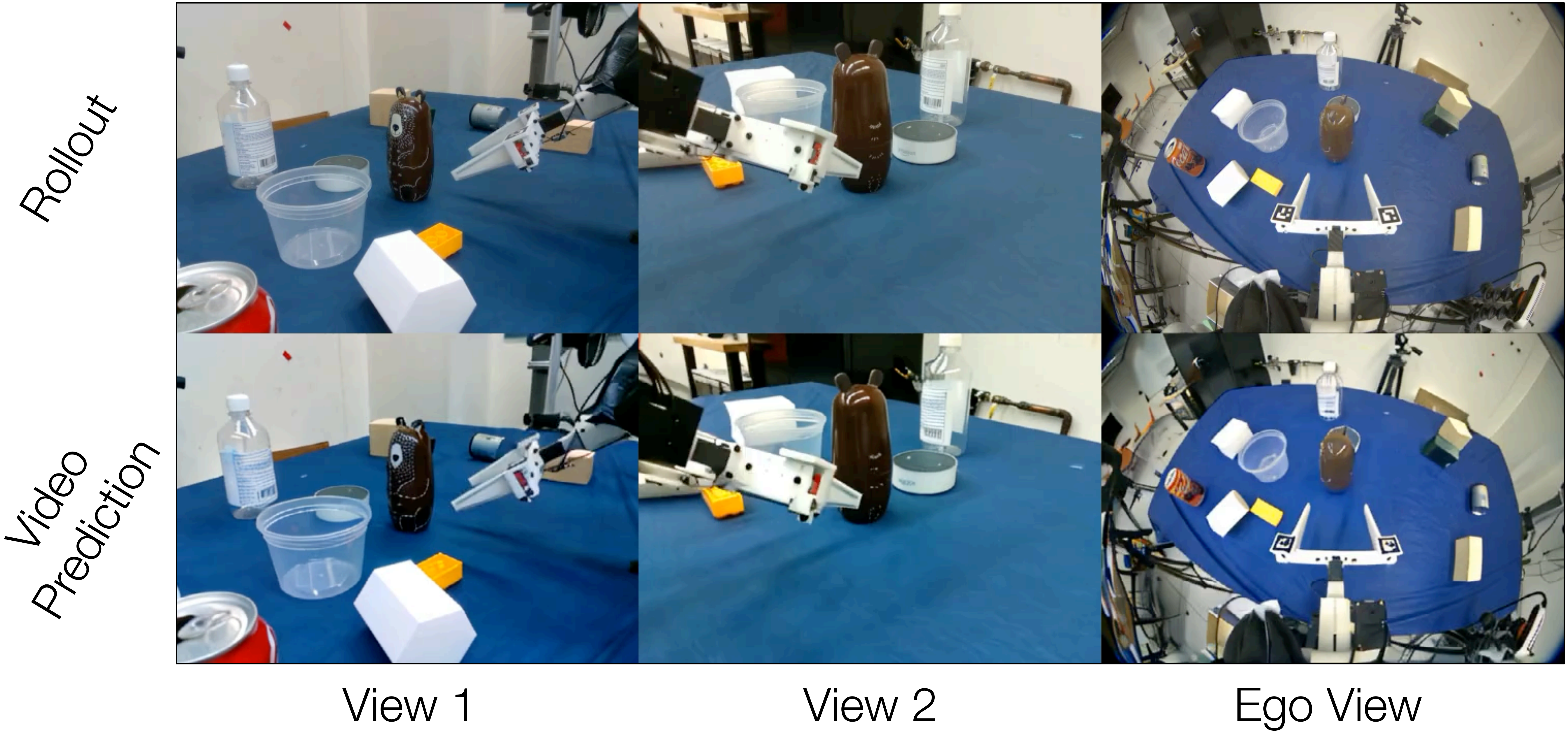
Opening Novel Drawers



Uprighting Novel Objects



Pick-n-Place Novel Objects

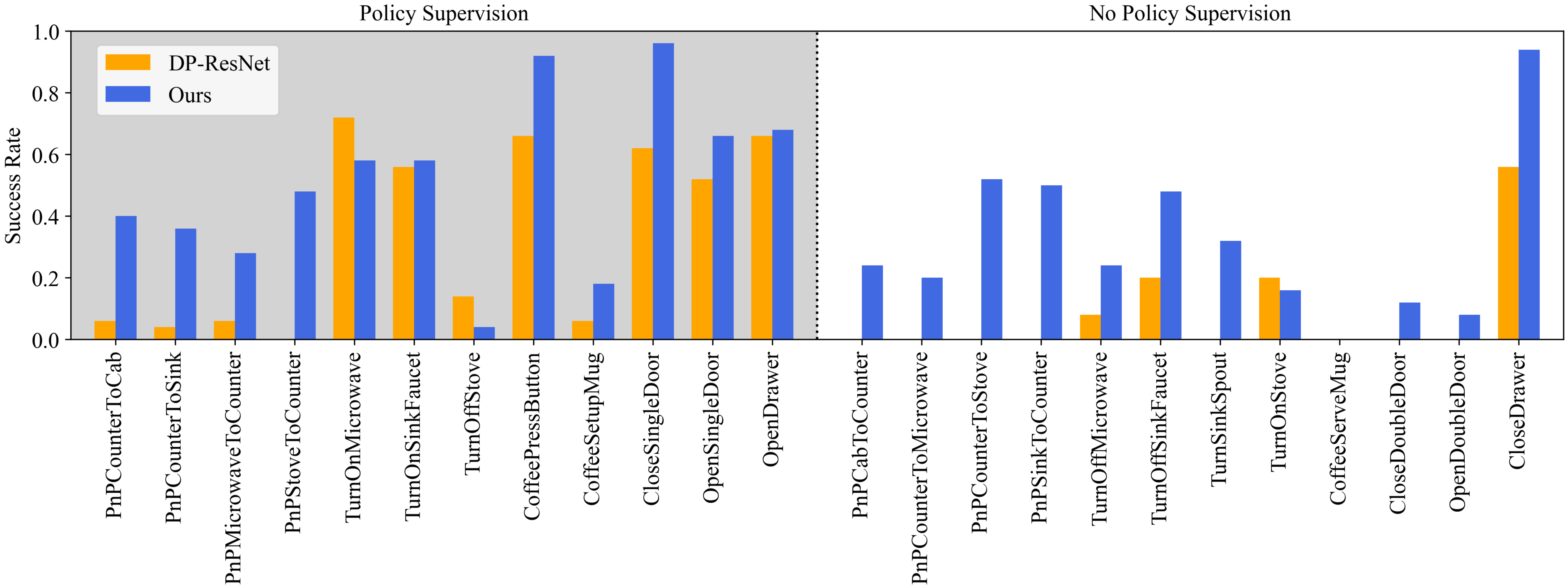


Policy Performance

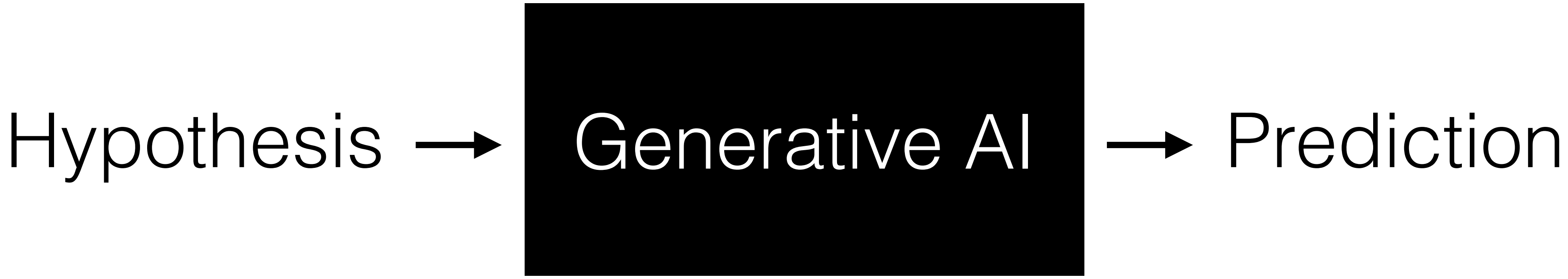
Model	**Average Success Rate**
3DA	0.06
DP3	0.23
DP-ResNet	0.41
DP-CLIP	0.43
GR00T	0.50
FPV	0.51
DP-VLA	0.57
Ours	**0.63**

Average task success rate against baselines
across 24 RoboCasa atomic tasks

Task Generalization

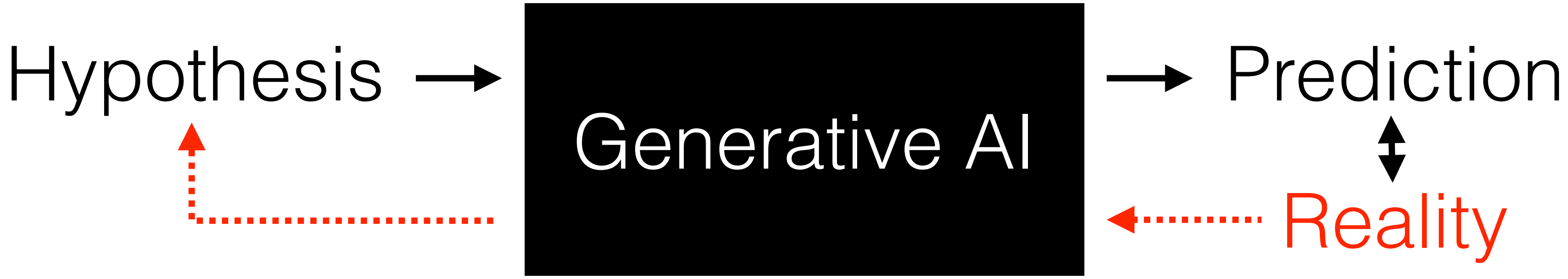


Not Always Realizable



Hypothesis → Generative AI → Prediction

Scientific Method

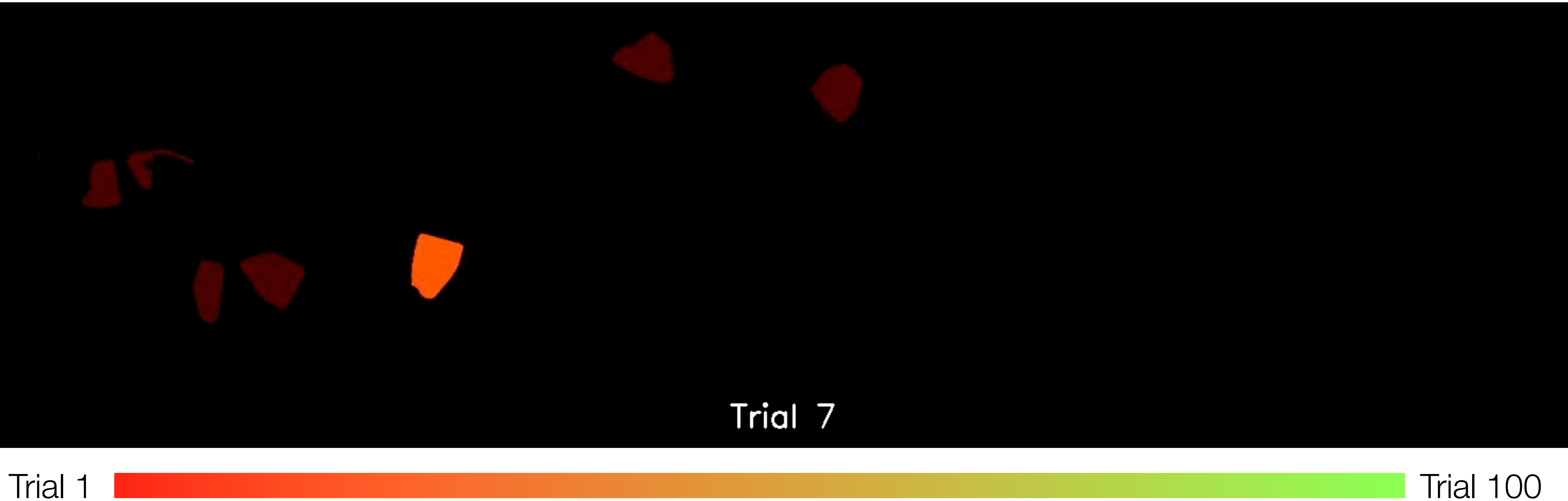


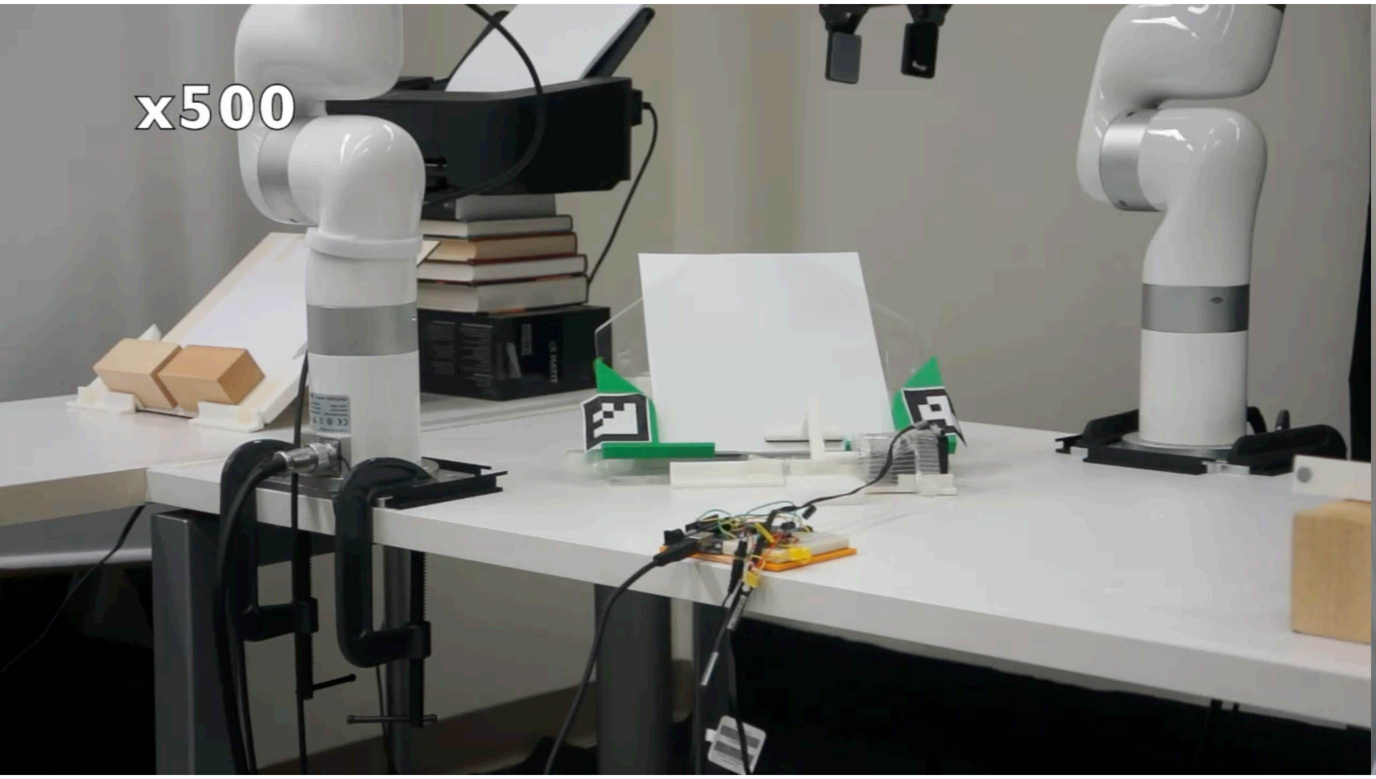
Hypothesis → Generative AI → Prediction

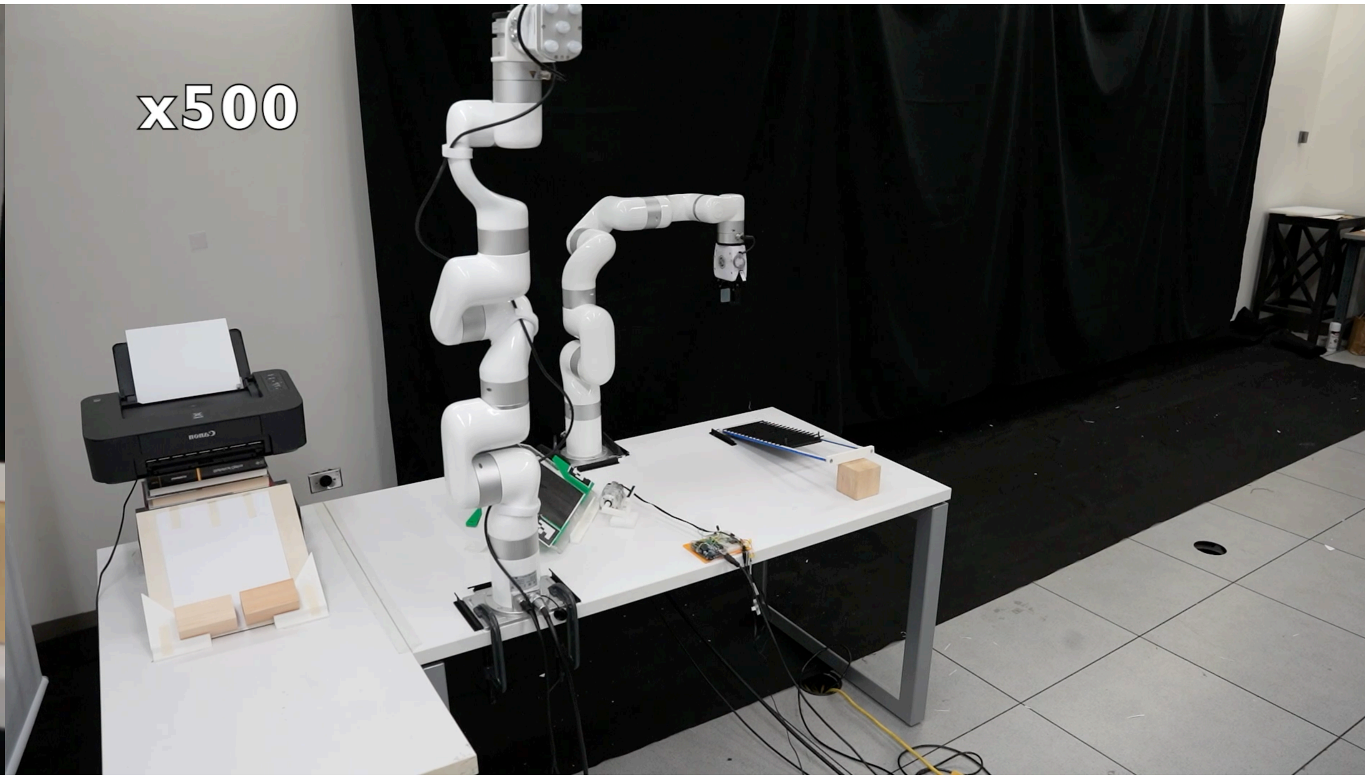
Prediction ↕ Reality

Reality → Generative AI → Hypothesis

Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.

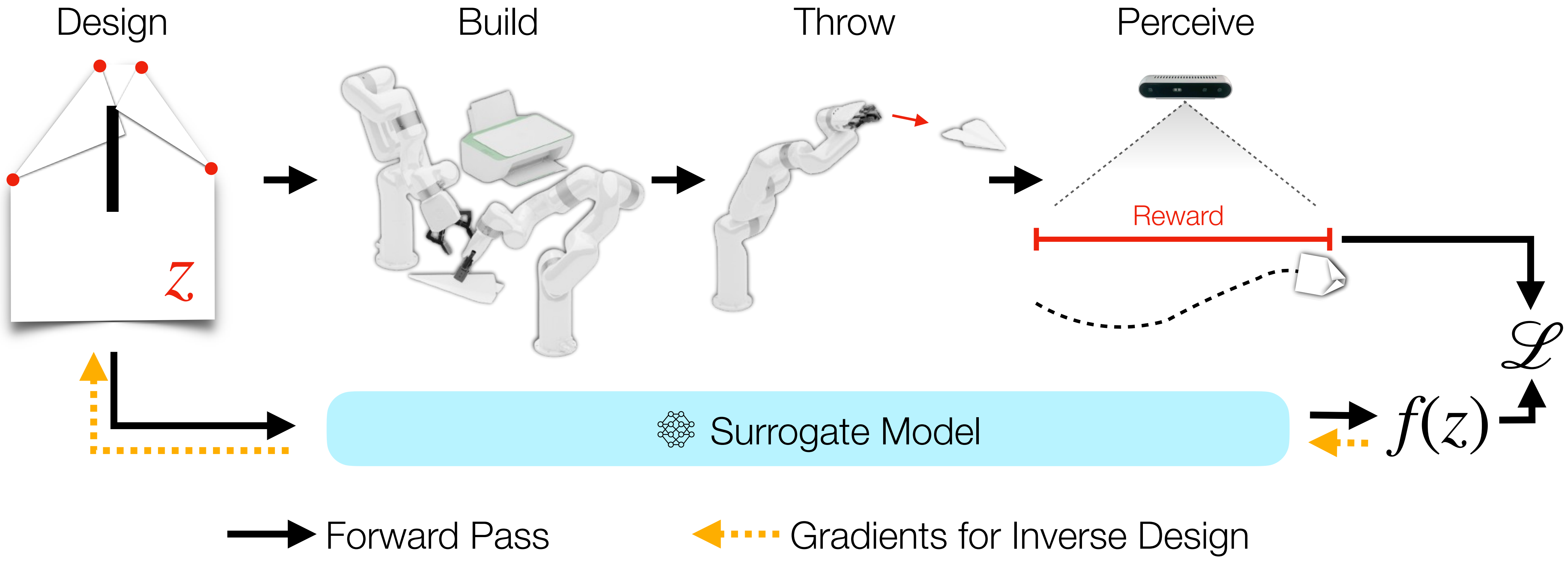






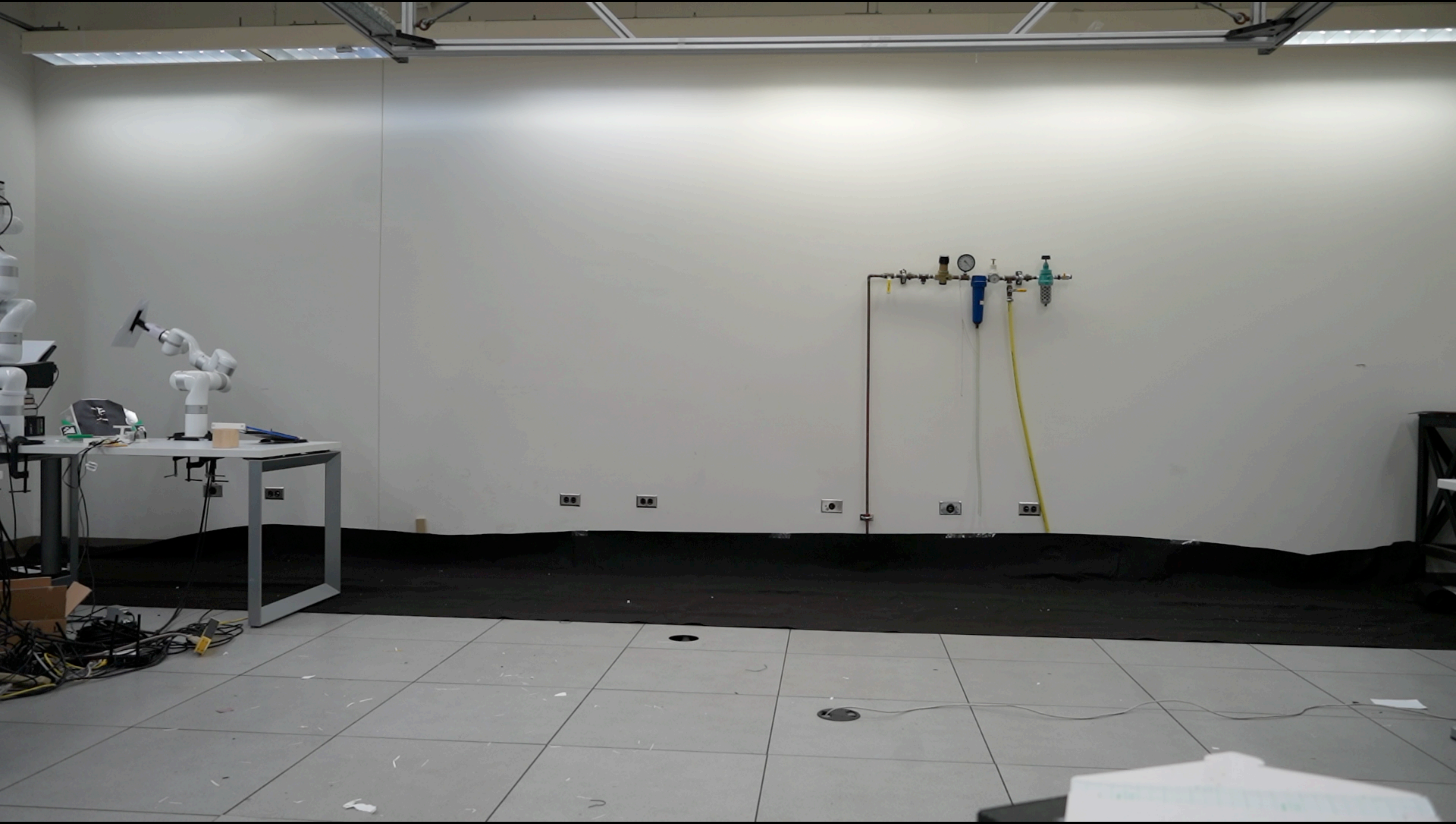
Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.

Paper Tool Design



Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.

Iteration 1



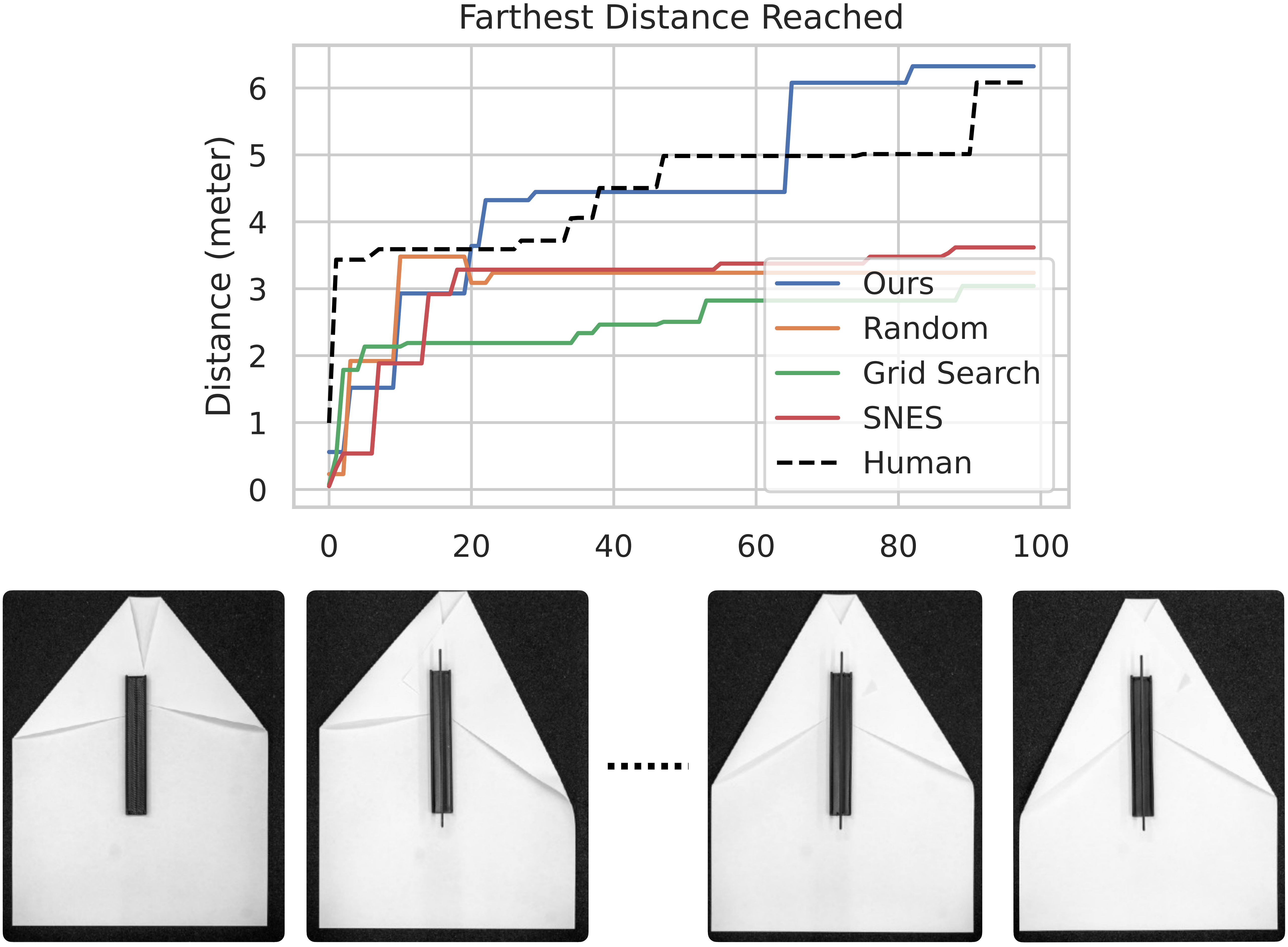
Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.

Iteration 56

Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.

Iteration 91

Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.



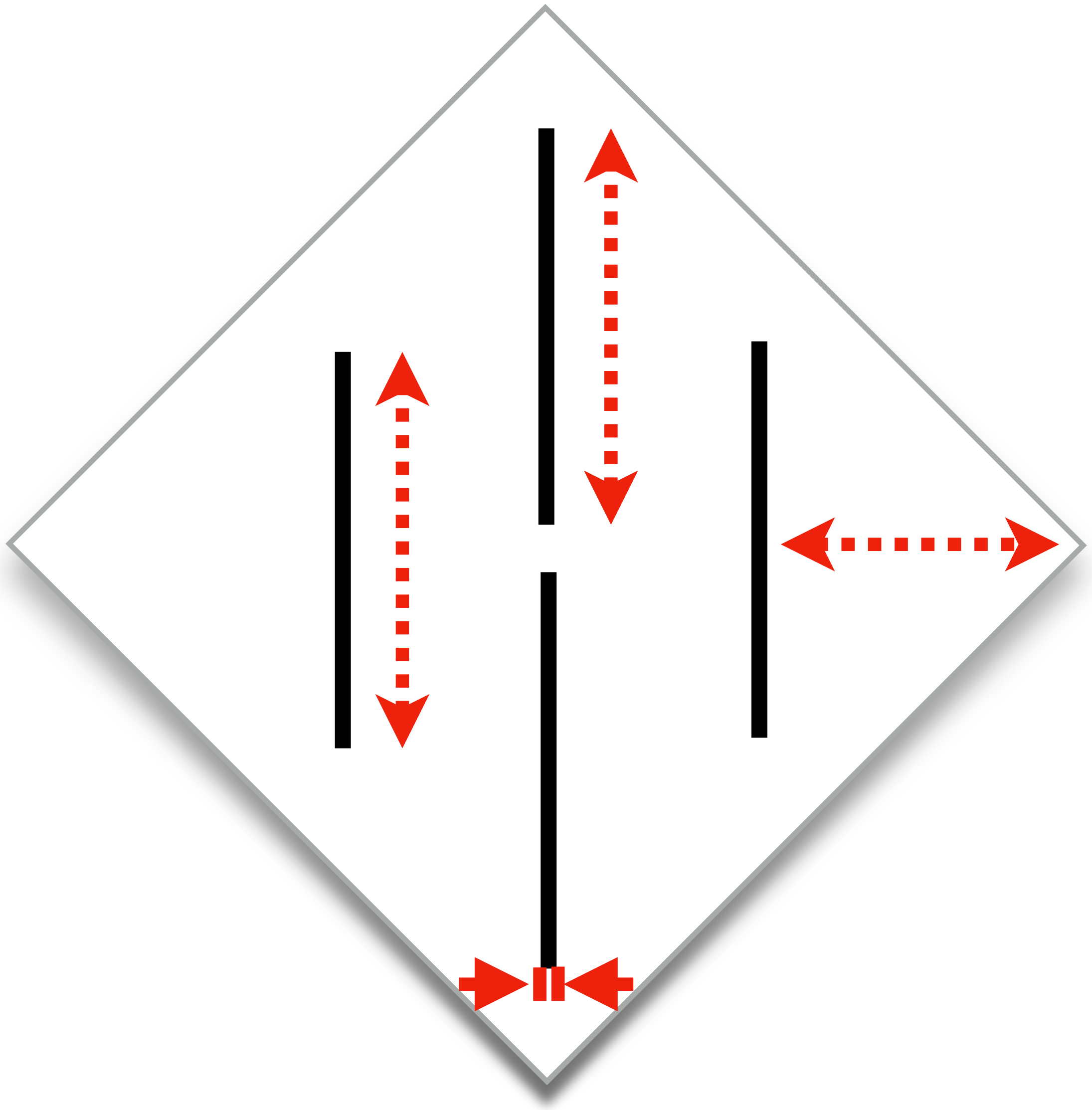
Iteration 3

Iteration 20

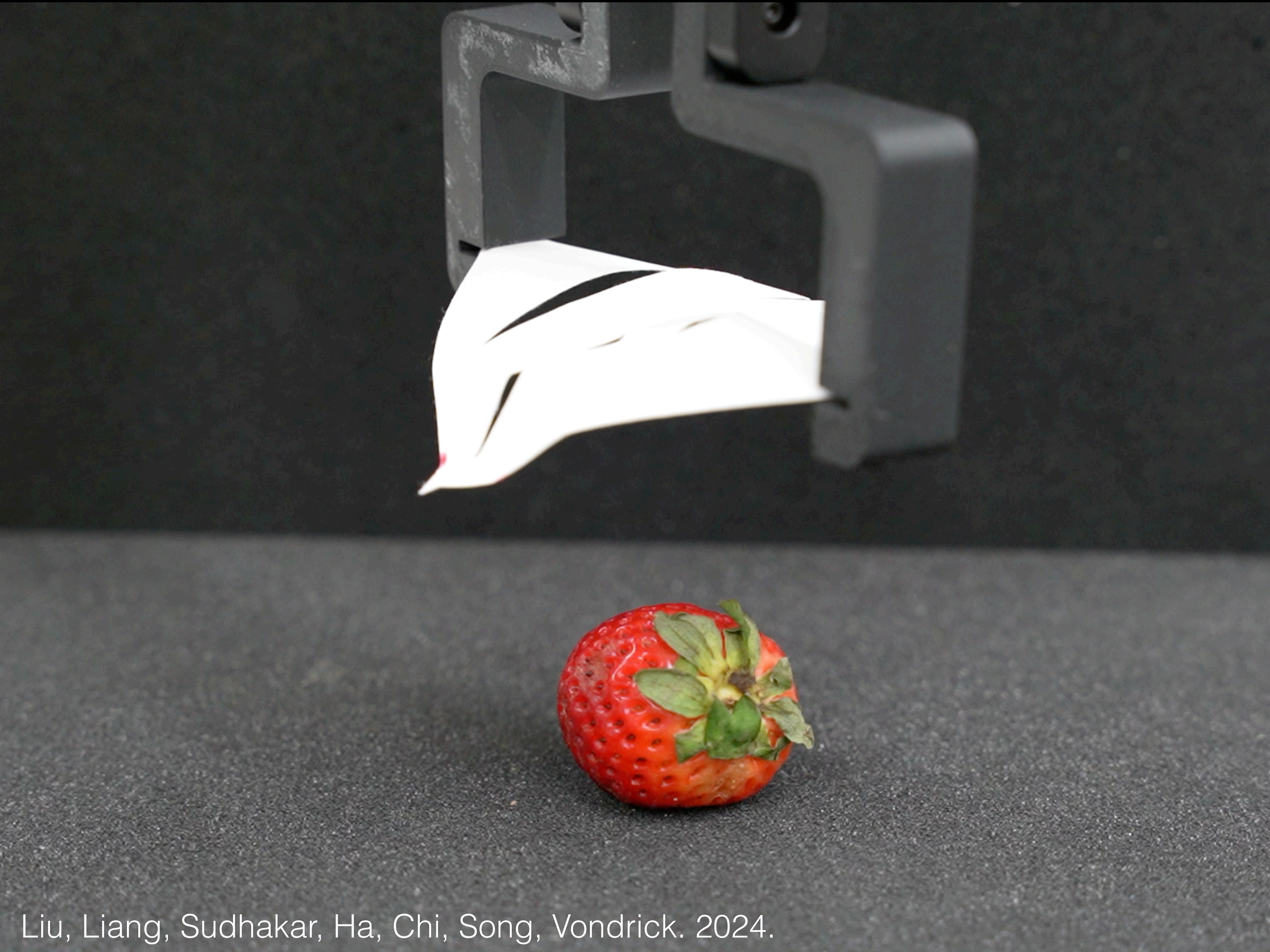
Iteration 70

Iteration 91

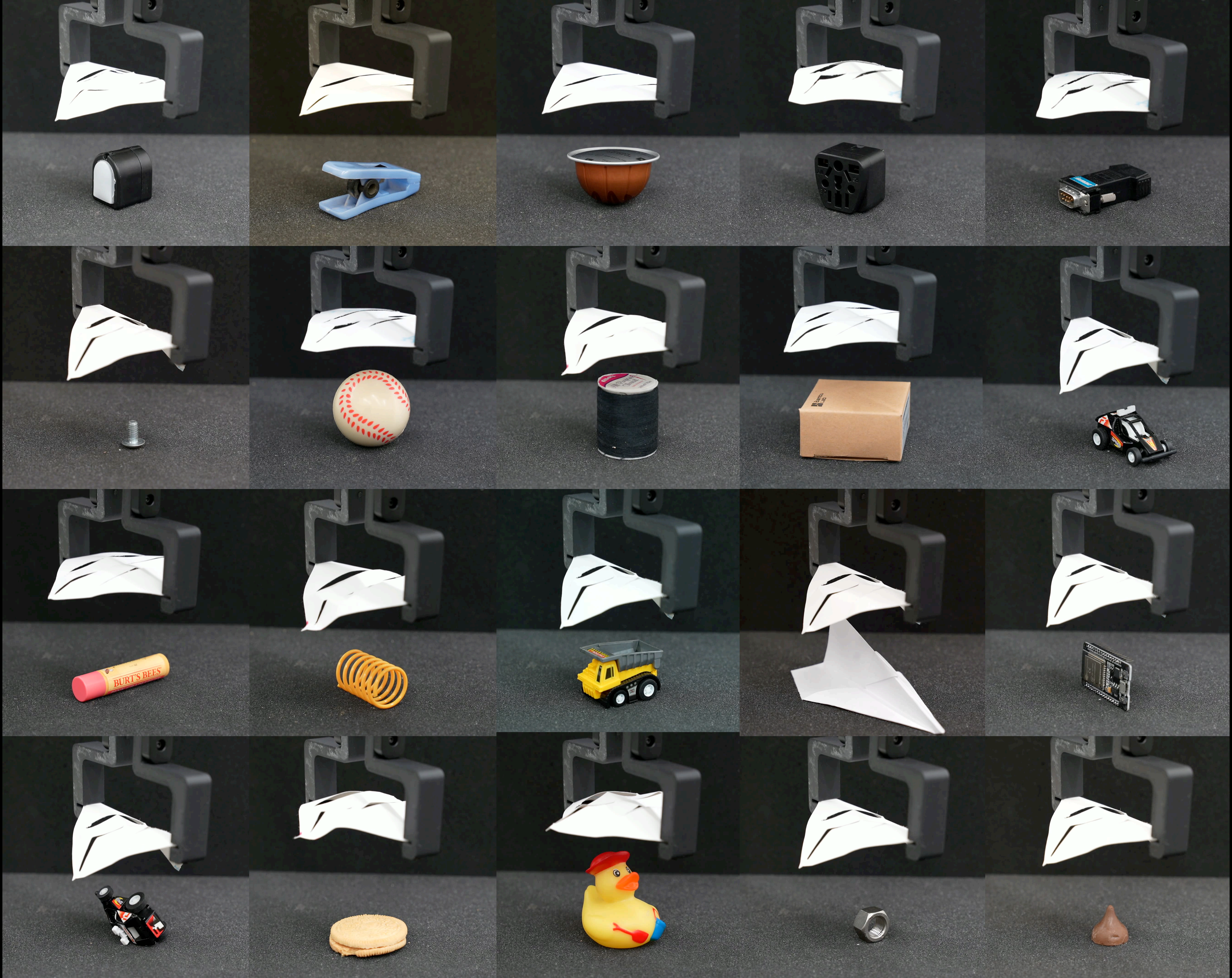
Kirigami Gripper

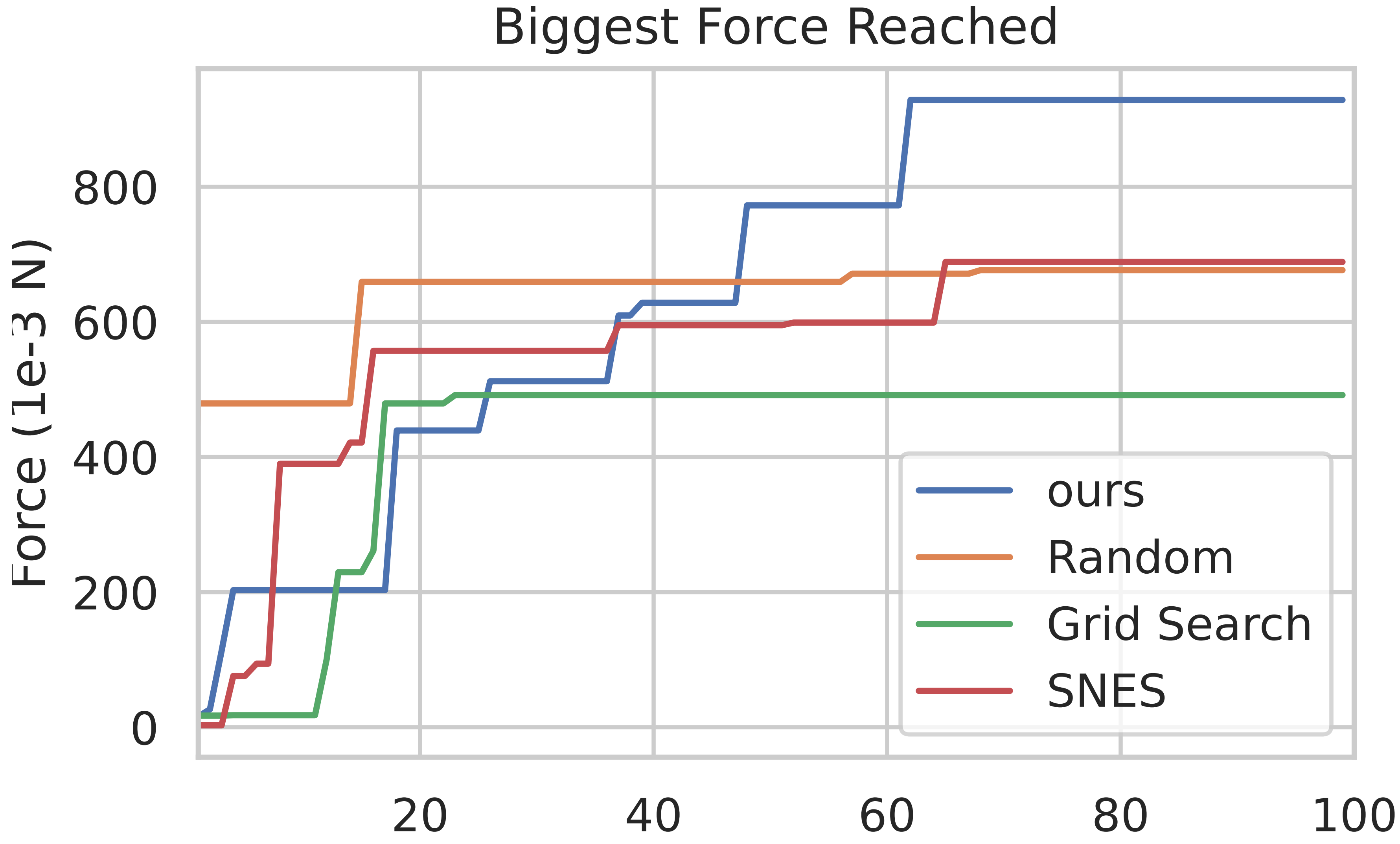


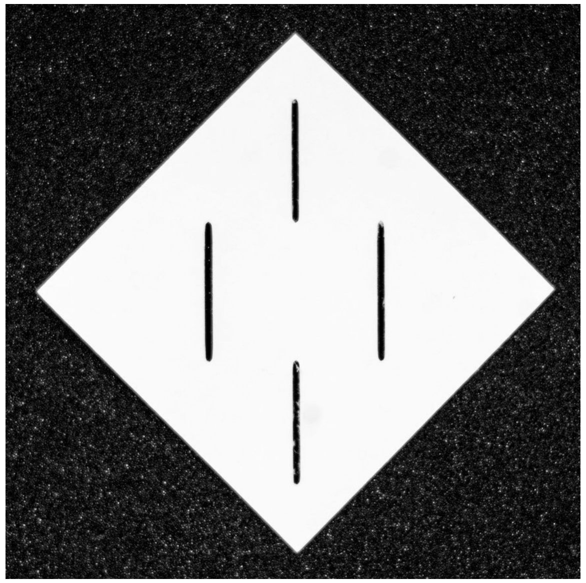
Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.



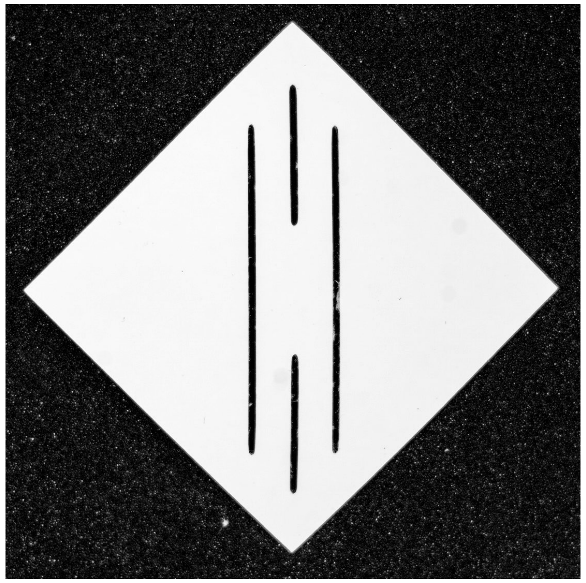
Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.


BURT'S BEES

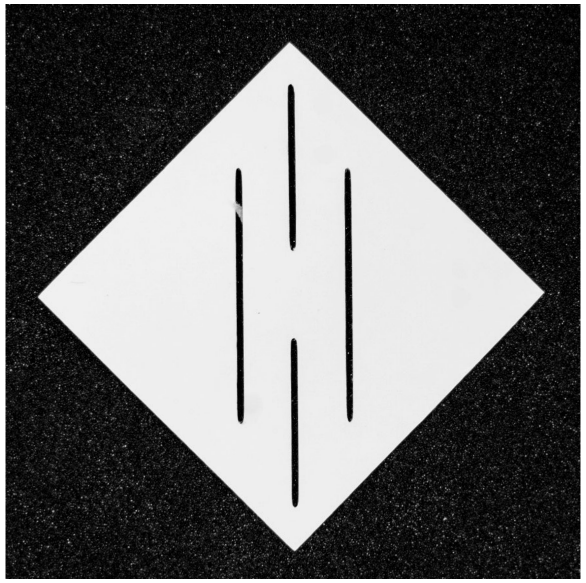




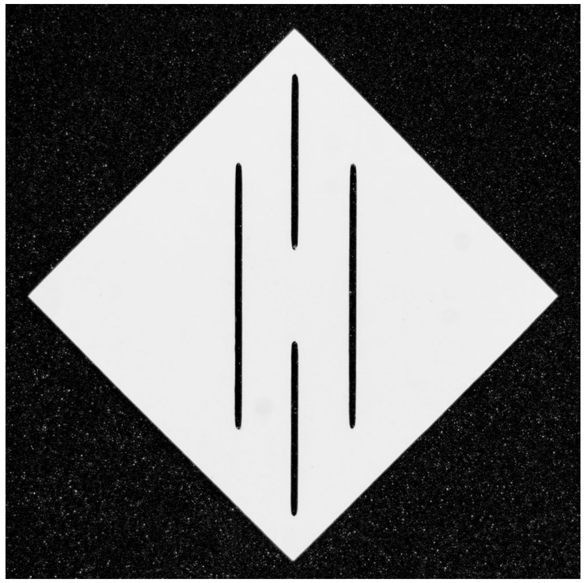
Iteration 2



Iteration 45



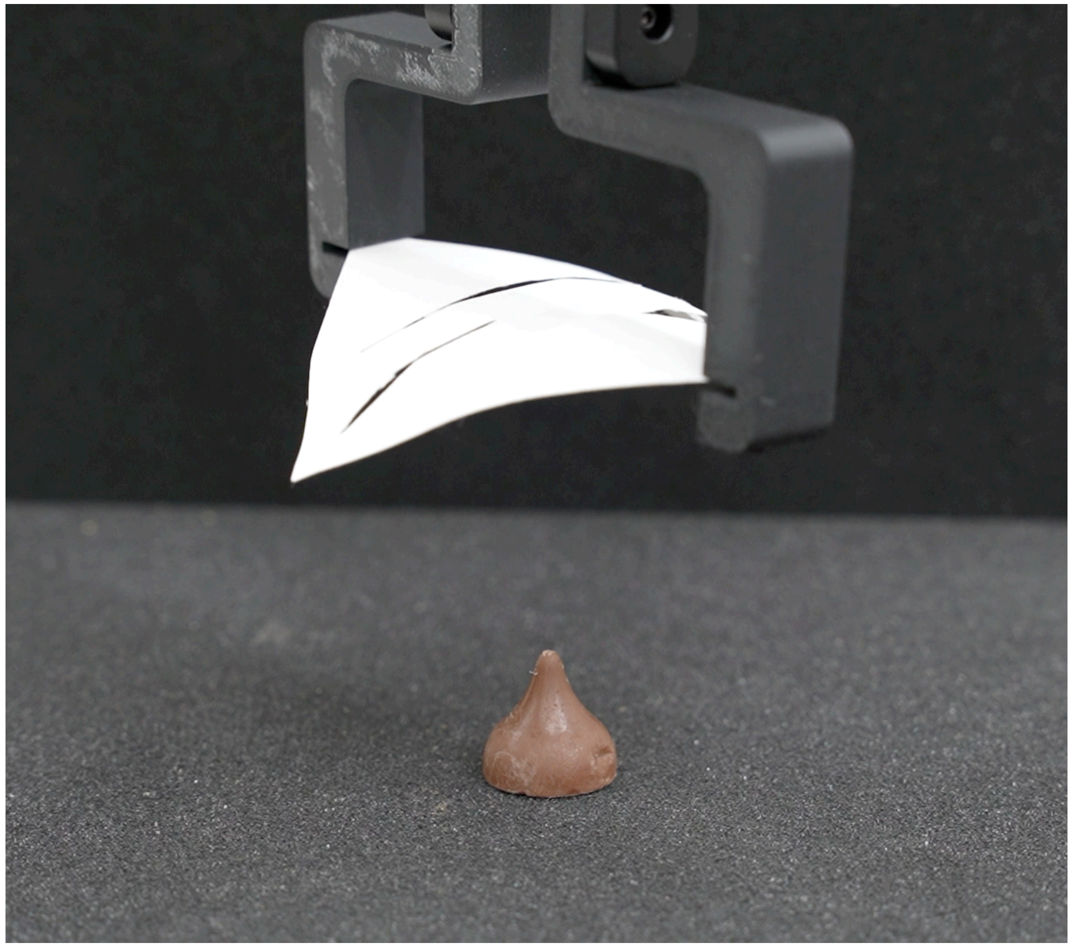
Iteration 77



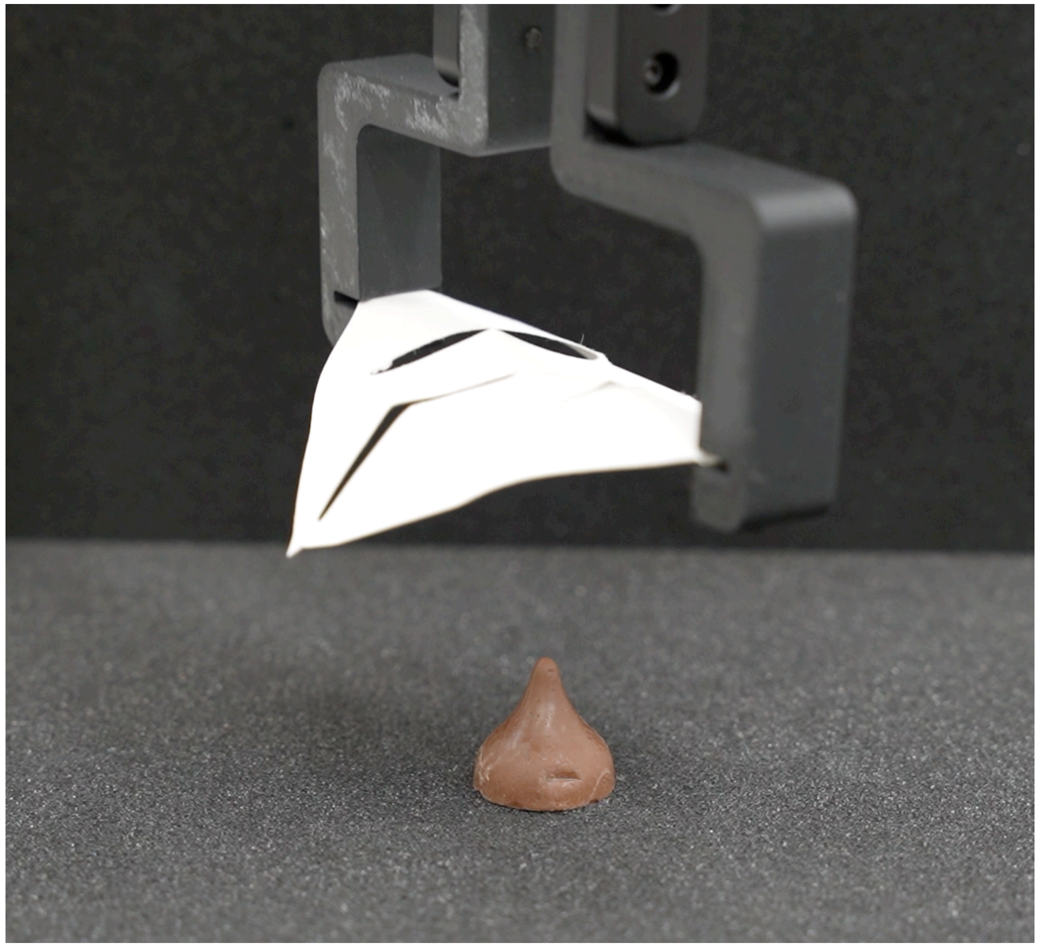
Iteration 97

Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.

Adapting to New Objects



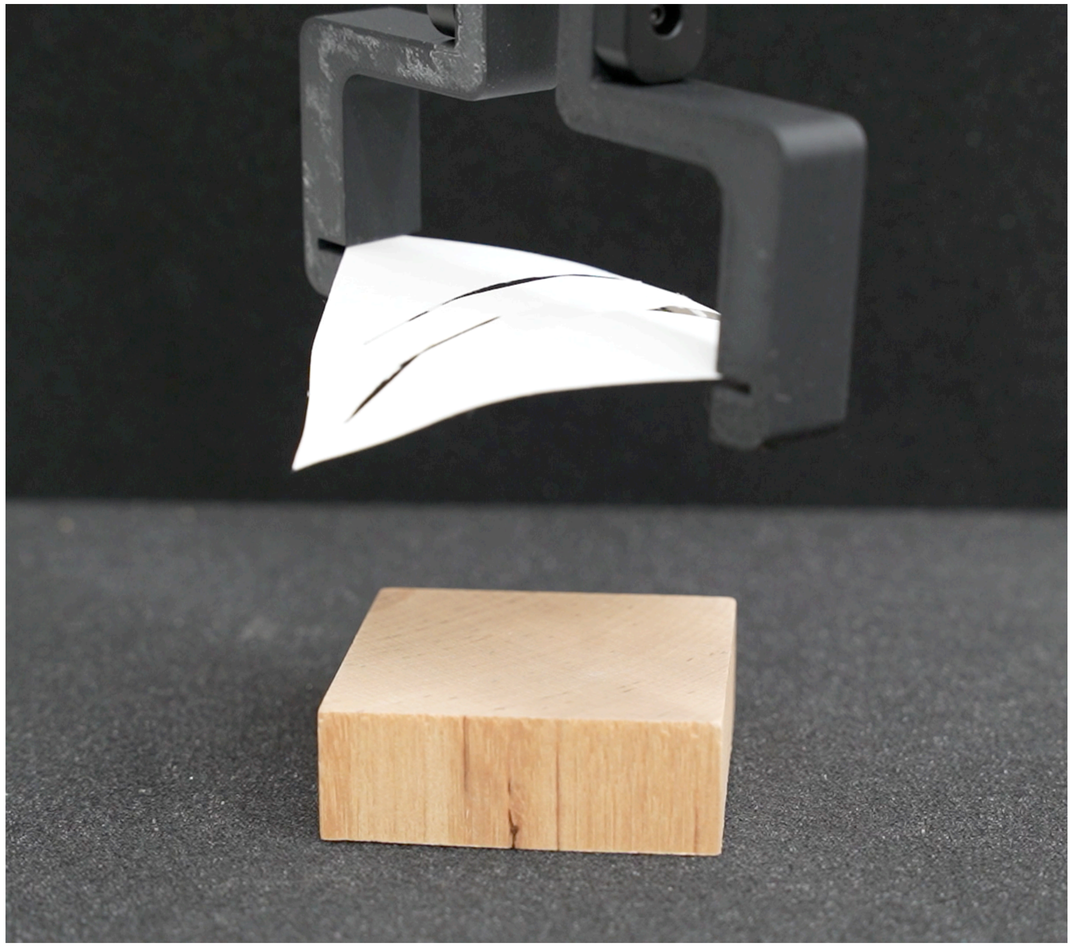
Before (0.30 N)



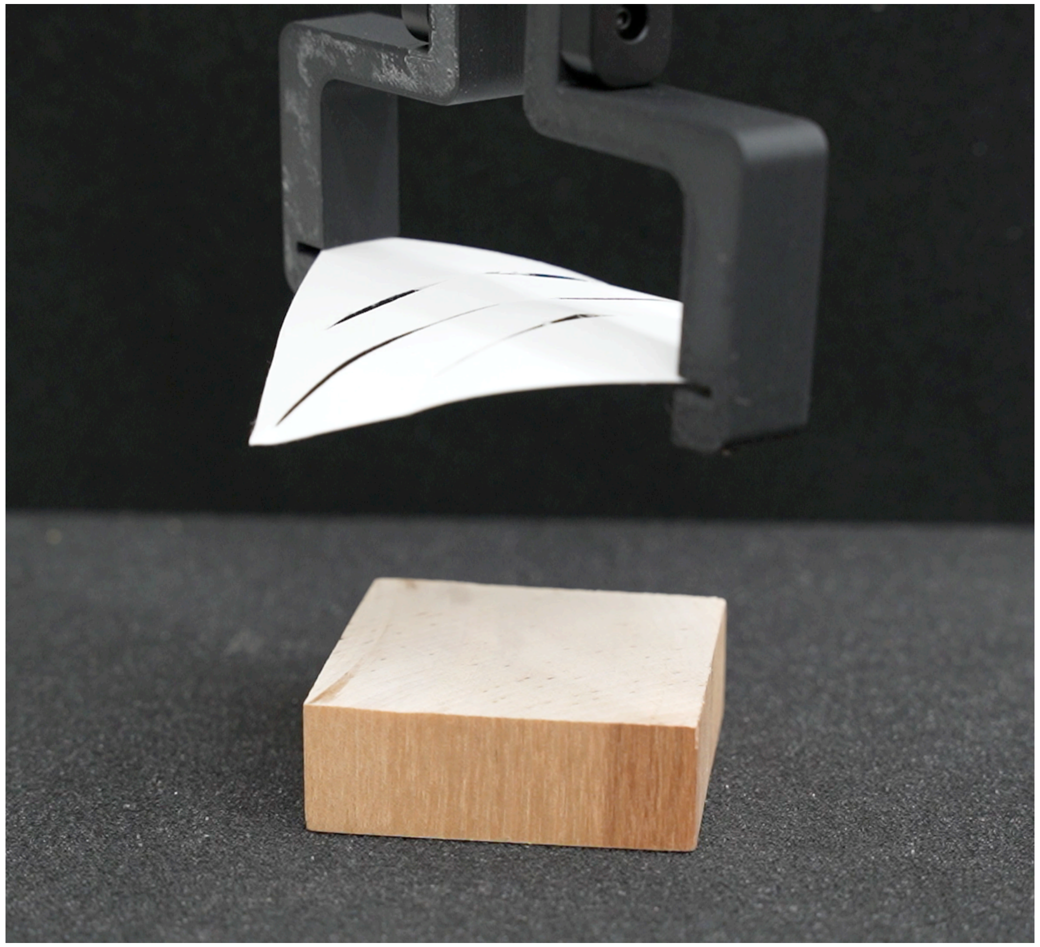
After (.44 N)

Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.

Adapting to New Objects



Before (0.05 N)



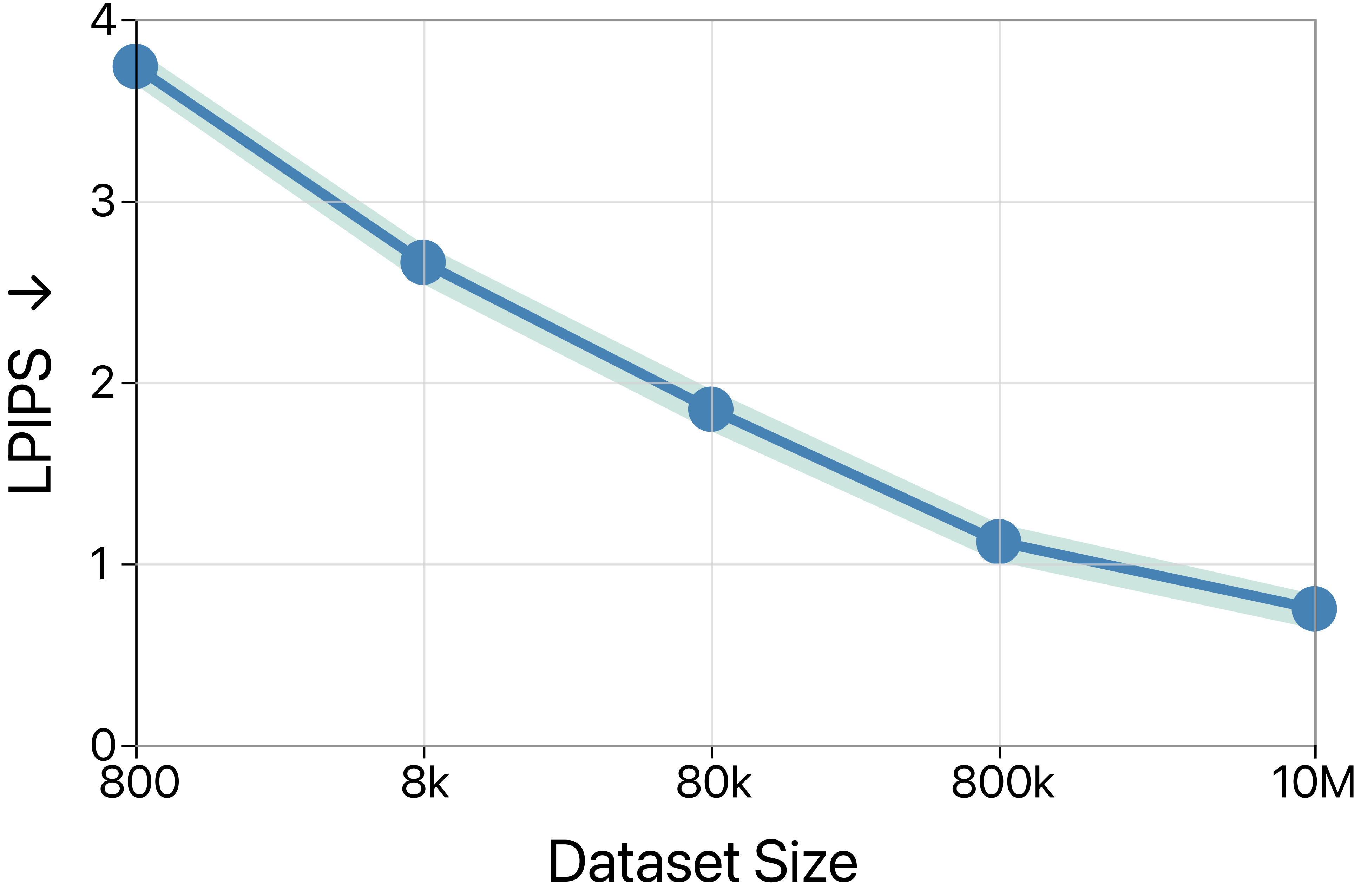
After (1.13 N)

Liu, Liang, Sudhakar, Ha, Chi, Song, Vondrick. 2024.

Generative Models for Computer Vision

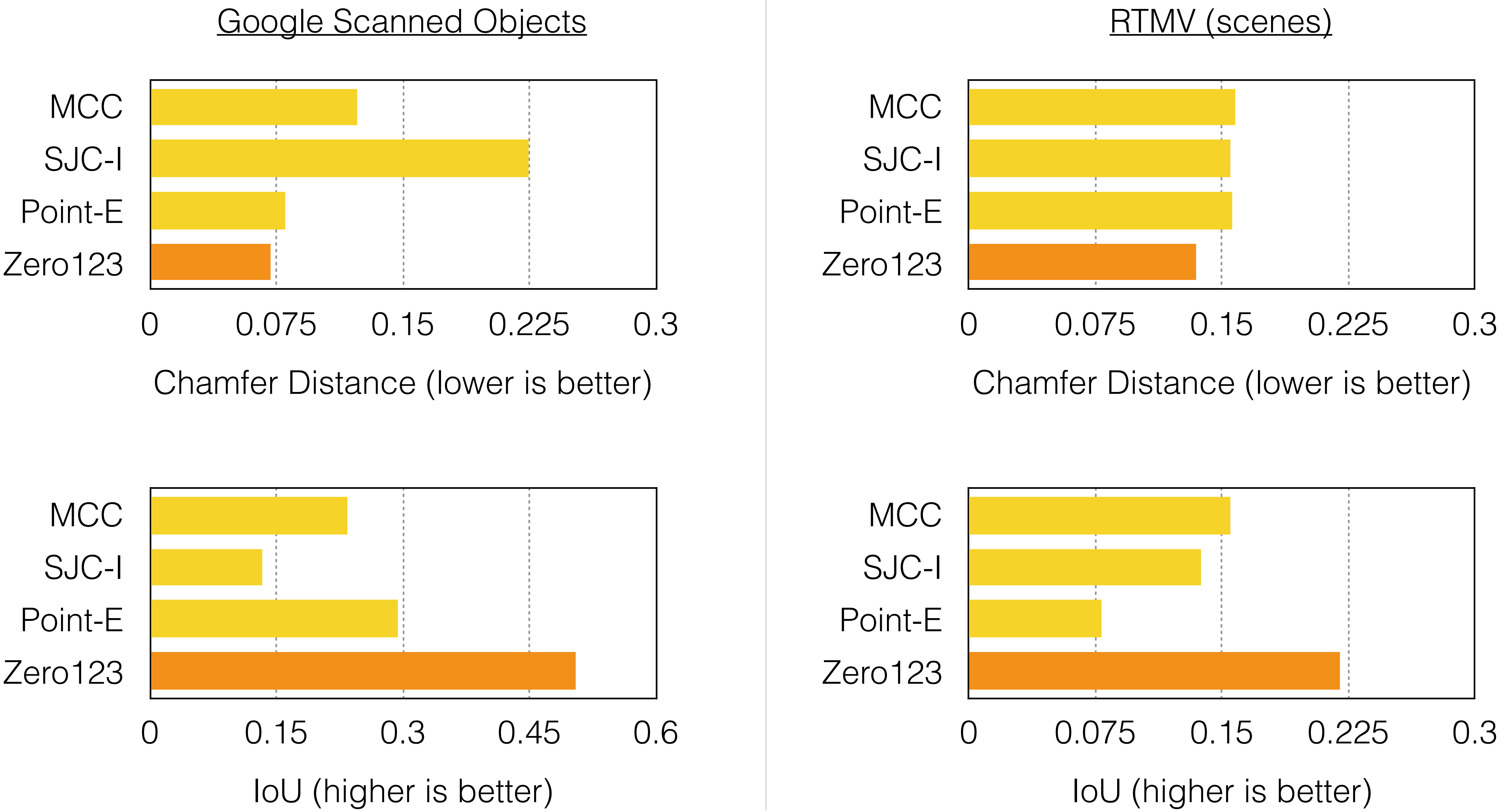
Carl Vondrick
Columbia University

Zero123 at Scale



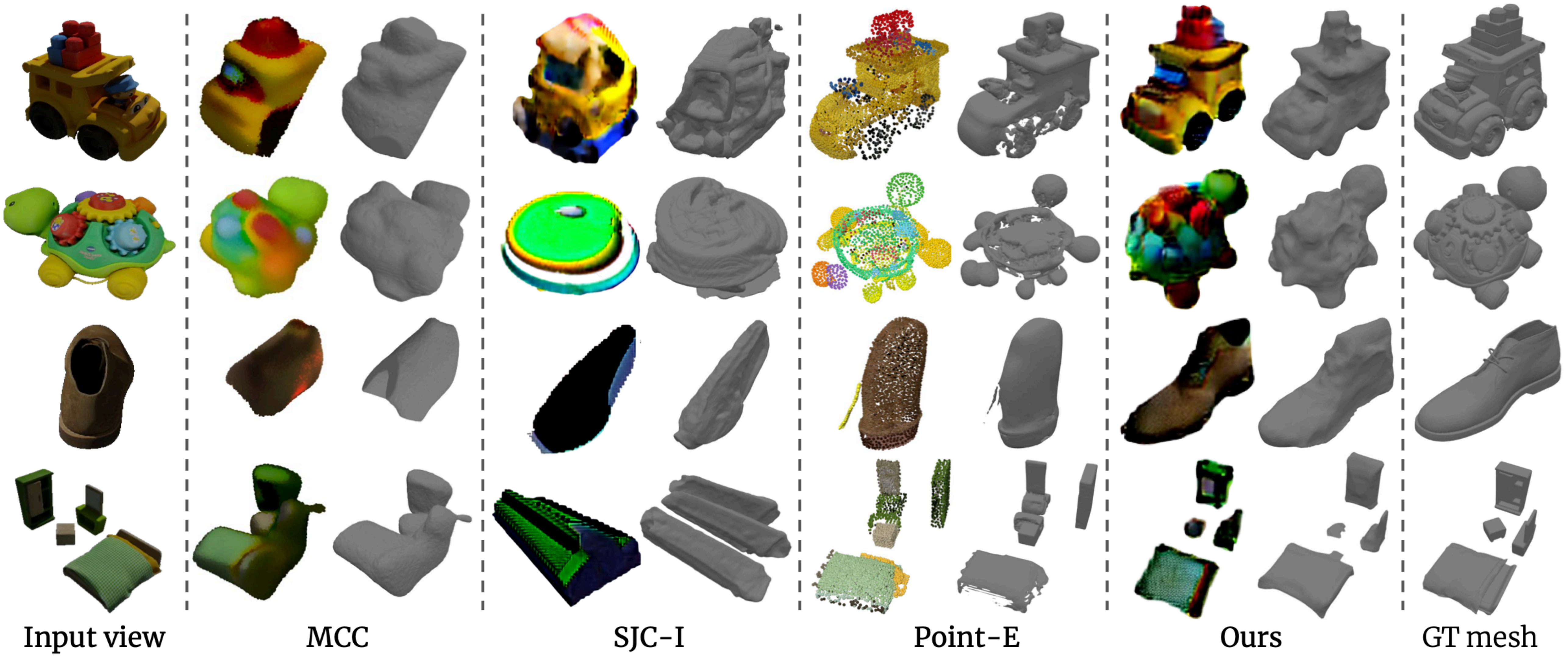
Deitke, et al. 2023.

3D Reconstruction from Single-view



Liu, Wu, Van Hoorick, Tokmakov, Zakharov, Vondrick. ICCV 2023.

3D Reconstruction from Single-view



Liu, Wu, Van Hoorick, Tokmakov, Zakharov, Vondrick. ICCV 2023.