Generative Embodied AI

Guest Lecture

Ruoshi Liu   Columbia University

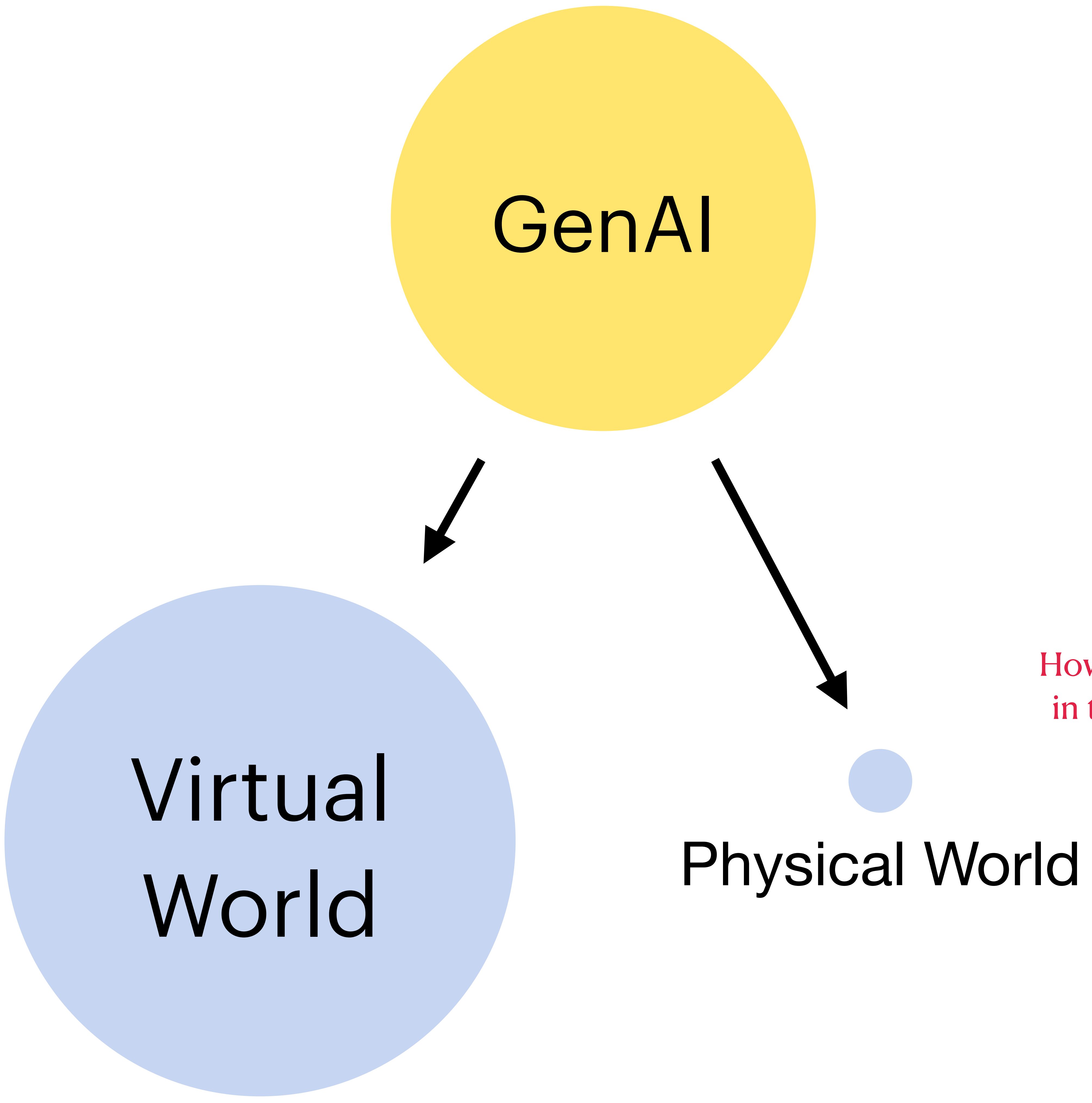
10-16-2024

Generative Embodied AI
(With the generous help of neural rendering)

Guest Lecture

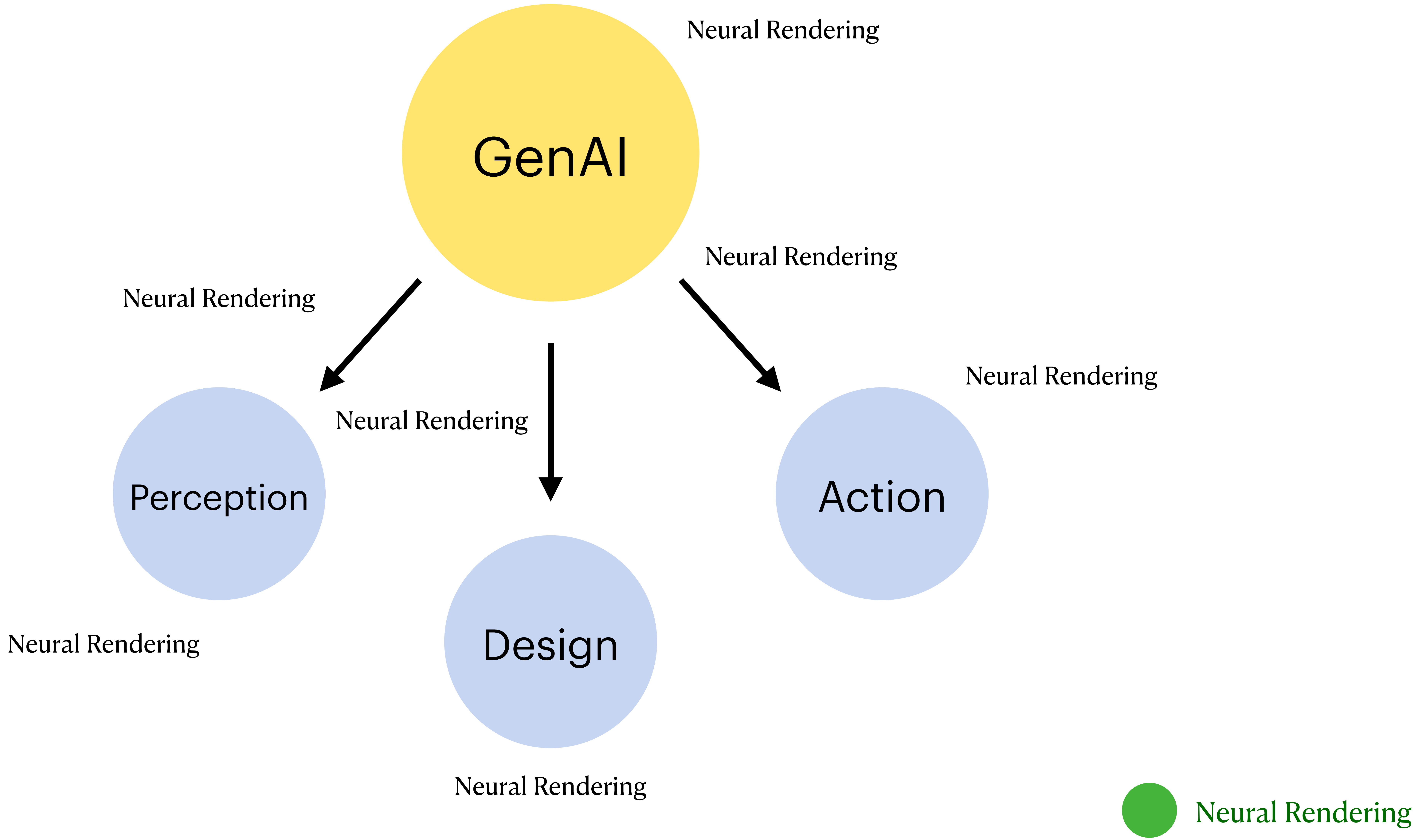
Ruoshi Liu   Columbia University
10-16-2024

Generative Models Today

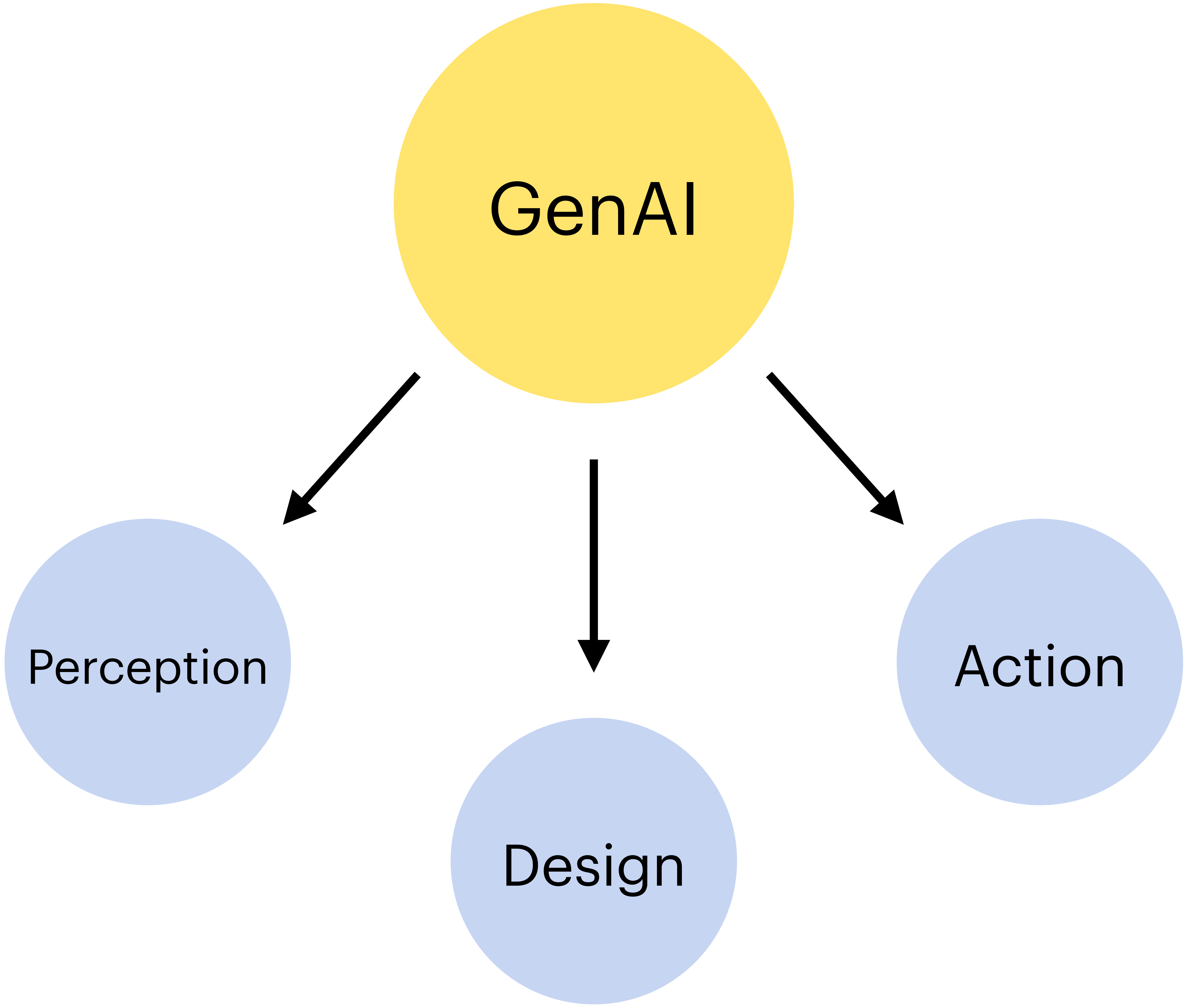


How can GenAI help us in the physical world?

Generative Embodied AI



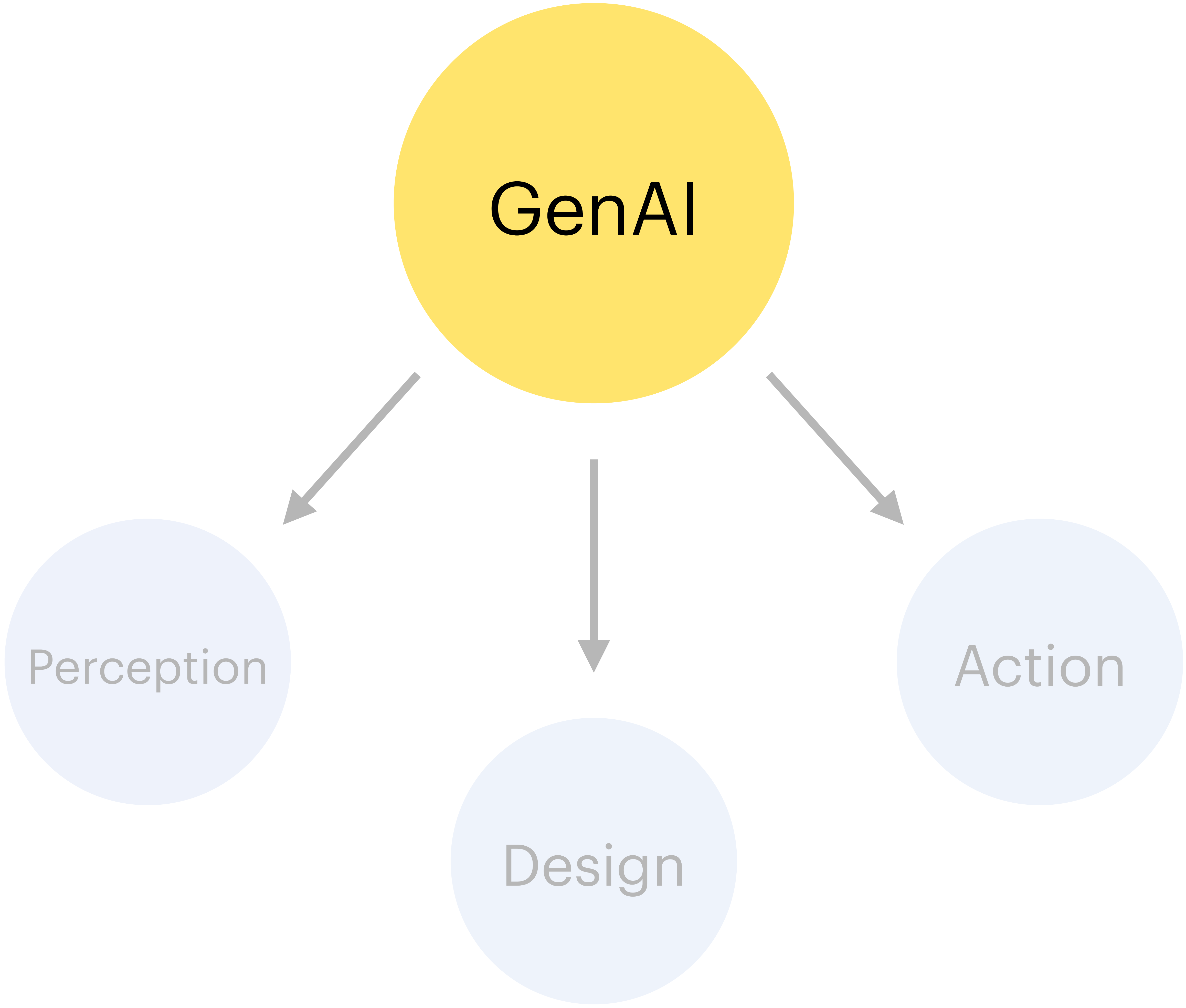
Generative Embodied AI



How do we build better 3D generative models?

How can 3d Gen help us in the physical world?

Generative Embodied AI



How do we build better 3D generative models?

Vision Generative Model Timeline

Vision Generative Model Timeline

Figure 6. Random samples drawn from the PCA face model.

Image Parsing: Unifying Segmentation, Detection, and Recognition

ZHUOWEN TU AND XIANGRONG CHEN
Departments of Statistics, University of California, Los Angeles, Los Angeles, CA 90095, USA
ztu@stat.ucla.edu
xrchen@stat.ucla.edu

ALAN L. YUILLE
Departments of 'Statistics' and 'Psychology', University of California, Los Angeles, Los Angeles, CA 90095, USA
yuille@stat.ucla.edu

SONG-CHUN ZHU
Departments of 'Statistics' and 'Computer Science' University of California, Los Angeles, Los Angeles, CA 90095, USA
sczhu@stat.ucla.edu

Vision Generative Model Timeline

Vision Generative Model Timeline

PCA (Tu et a
2006

NeurIPS 2010 DL Workshop:https://twitter.com/ethanCaballero/status/1544400983261954048?s=20&t=6ocUrsX7sOCYkONn6Zas3w

Vision Generative Model Timeline

PCA
(Tu et al.)
2006

DBN
(Vincent
et al.)
2010

VAE
(Rezende/
Kingma et al.)
2013

Vision Generative Model Timeline

Auto-Encoding Variational Bayes

Diederik P. Kingma
Machine Learning Group
Universiteit van Amsterdam
dpkingma@gmail.com

Max Welling
Machine Learning Group
Universiteit van Amsterdam
welling.max@gmail.com

Vision Generative Model Timeline

[Submitted on 16 Jan 2014 (v1), last revised 30 May 2014 (this version, v3)]

Stochastic Backpropagation and Approximate Inference in Deep Generative Models

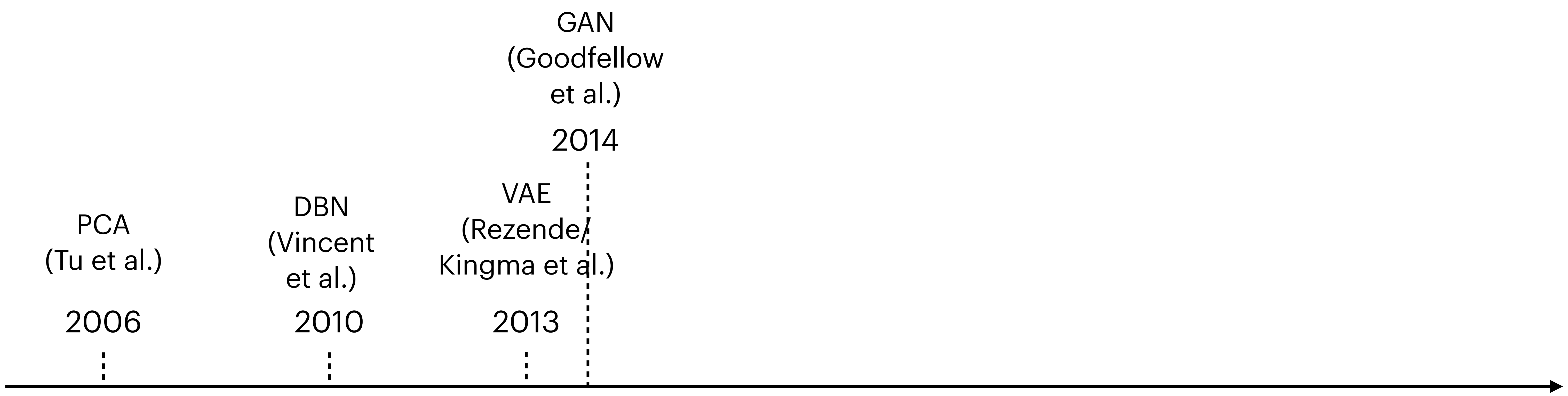
Danilo Jimenez Rezende, Shakir Mohamed, Daan Wierstra

[Submitted on 20 Dec 2013 (v1), last revised 10 Dec 2022 (this version, v11)]

Auto-Encoding Variational Bayes

Diederik P Kingma, Max Welling

Vision Generative Model Timeline

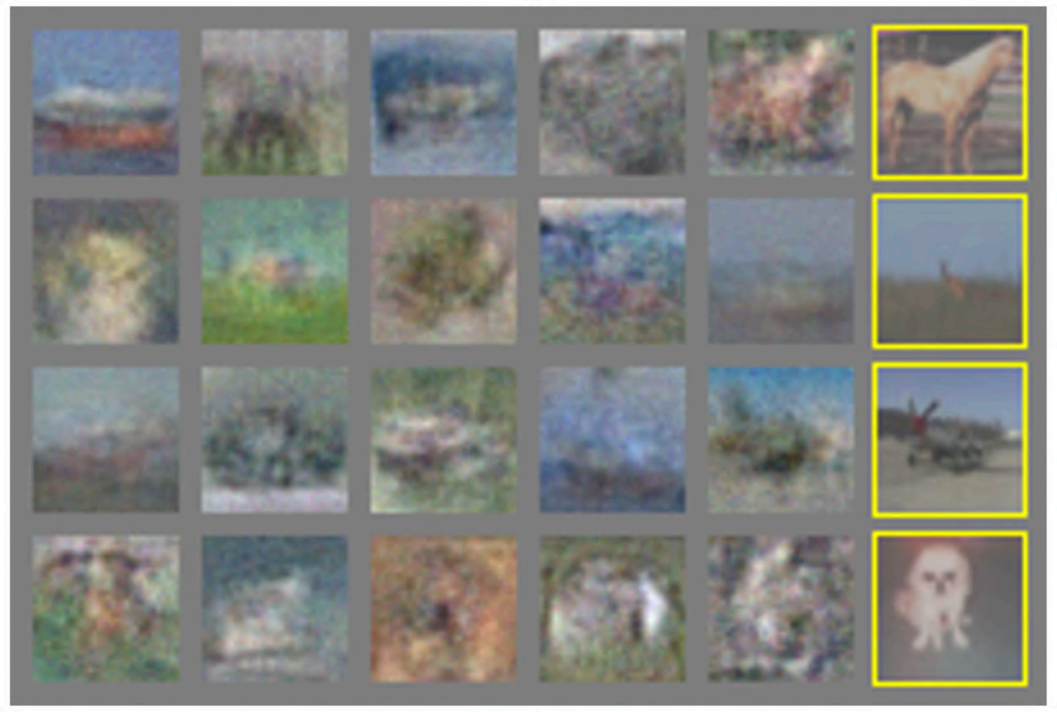


Generative Adversarial Nets

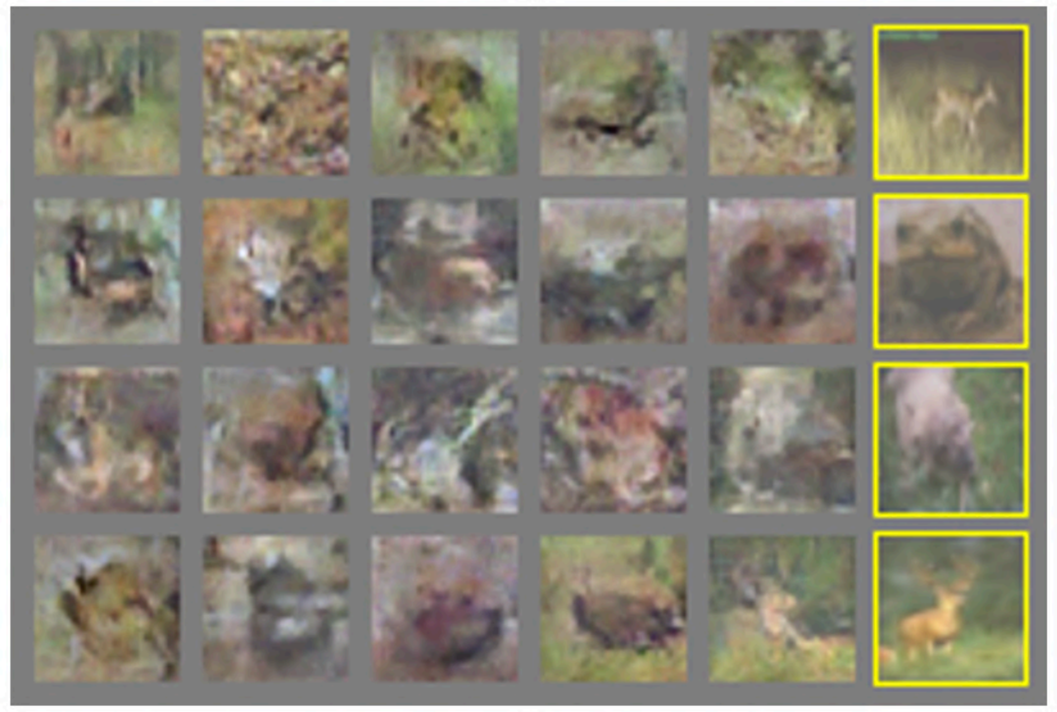
Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley,
Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

a)

b)

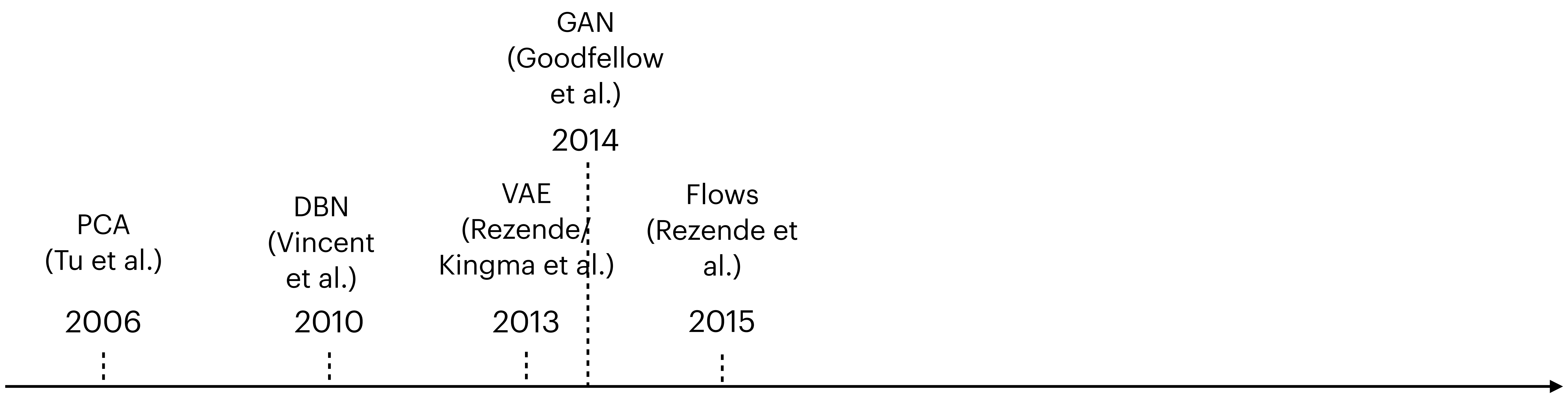


c)

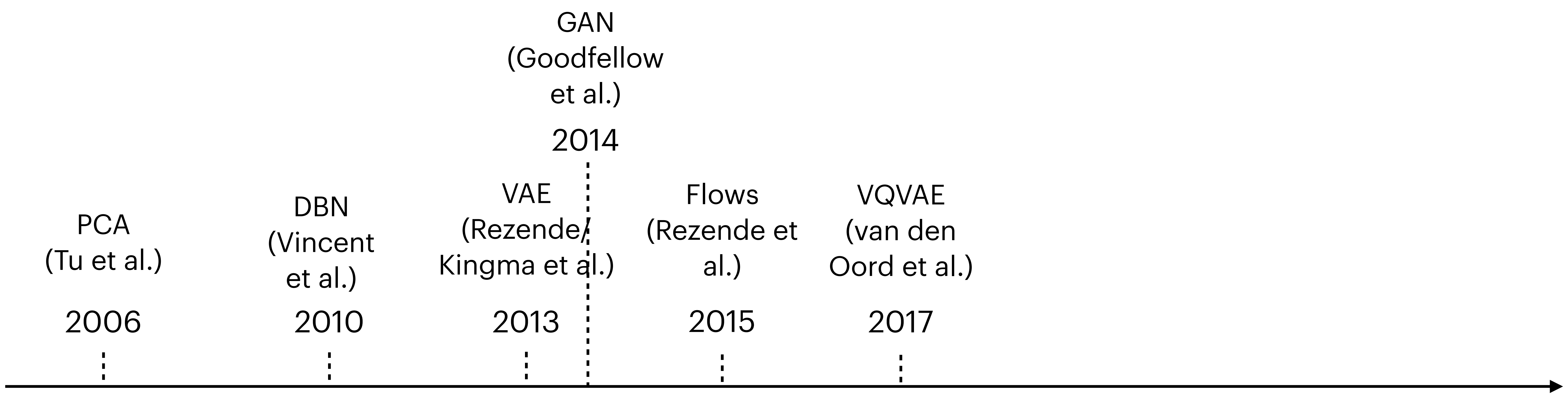


d)

Vision Generative Model Timeline



Vision Generative Model Timeline



Vision Generative Model Timeline

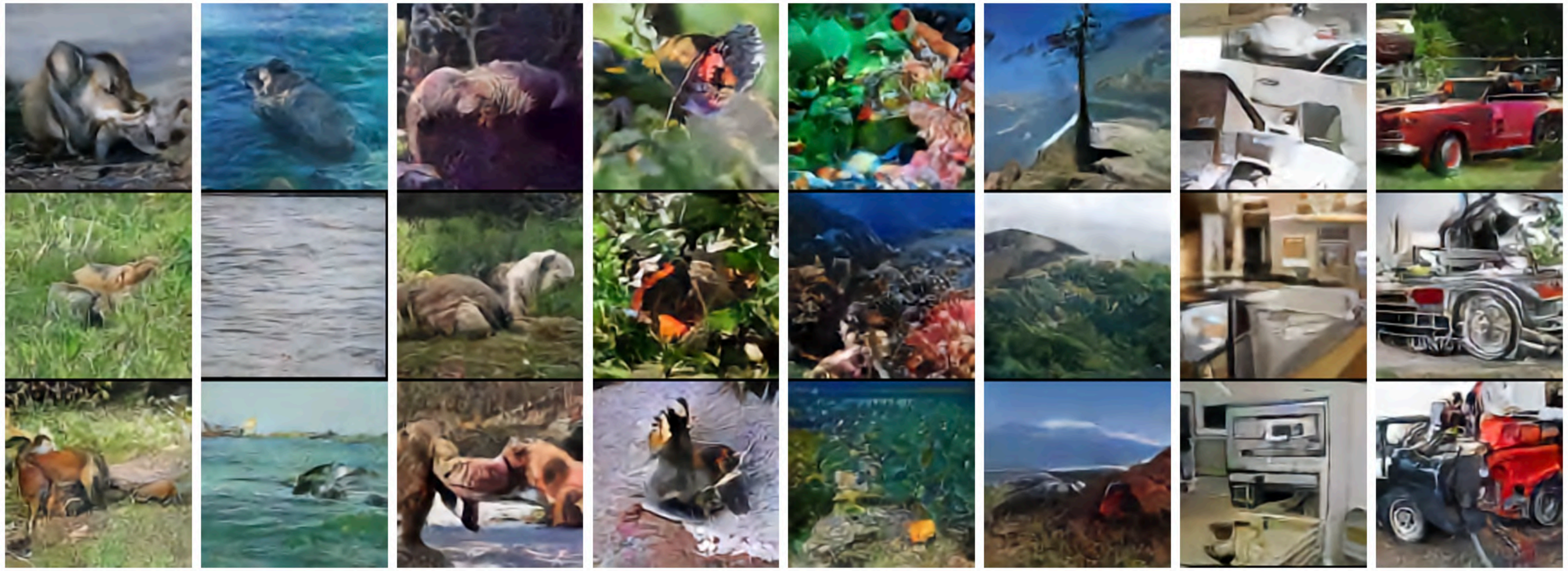
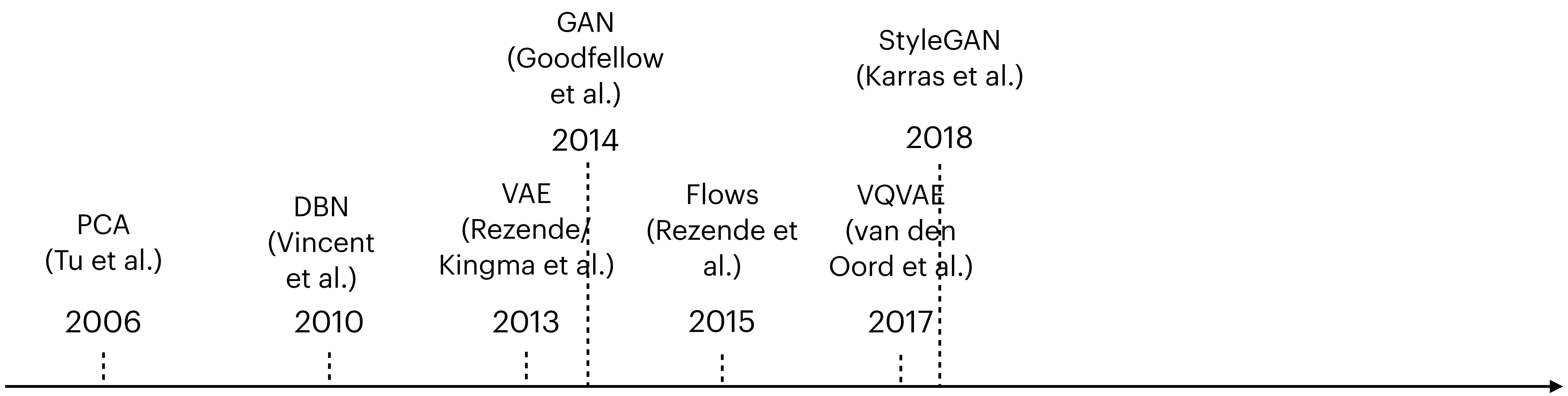


Figure 3: Samples (128x128) from a VQ-VAE with a PixelCNN prior trained on ImageNet images. From left to right: kit fox, gray whale, brown bear, admiral (butterfly), coral reef, alp, microwave, pickup.

Vision Generative Model Timeline



Vision Generative Model Timeline

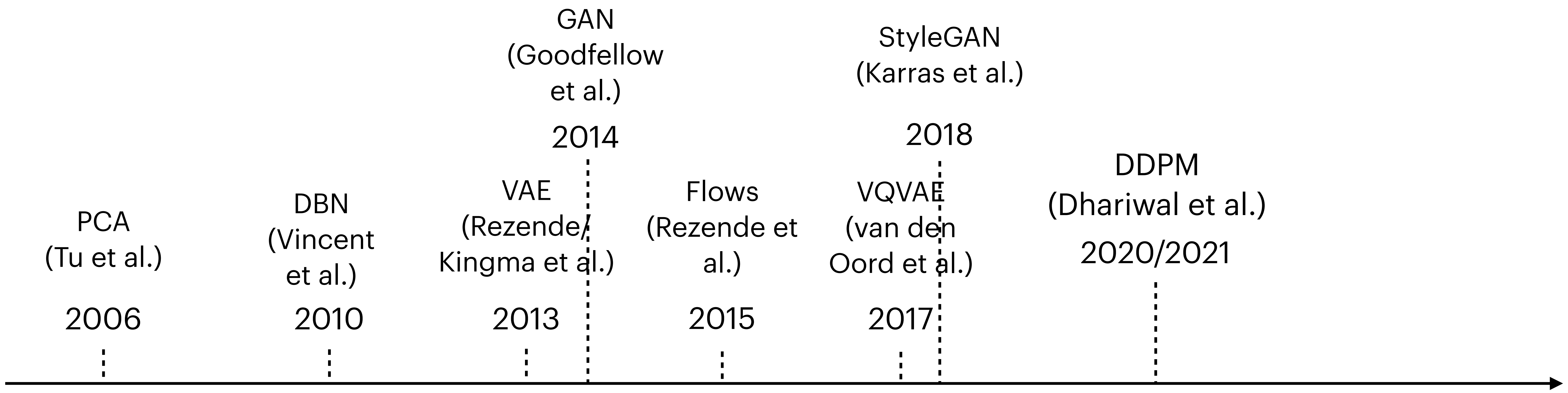
A Style-Based Generator Architecture for Generative Adversarial Networks

Tero Karras
NVIDIA
tkarras@nvidia.com

Samuli Laine
NVIDIA
slaine@nvidia.com

Timo Aila
NVIDIA
taila@nvidia.com

Vision Generative Model Timeline

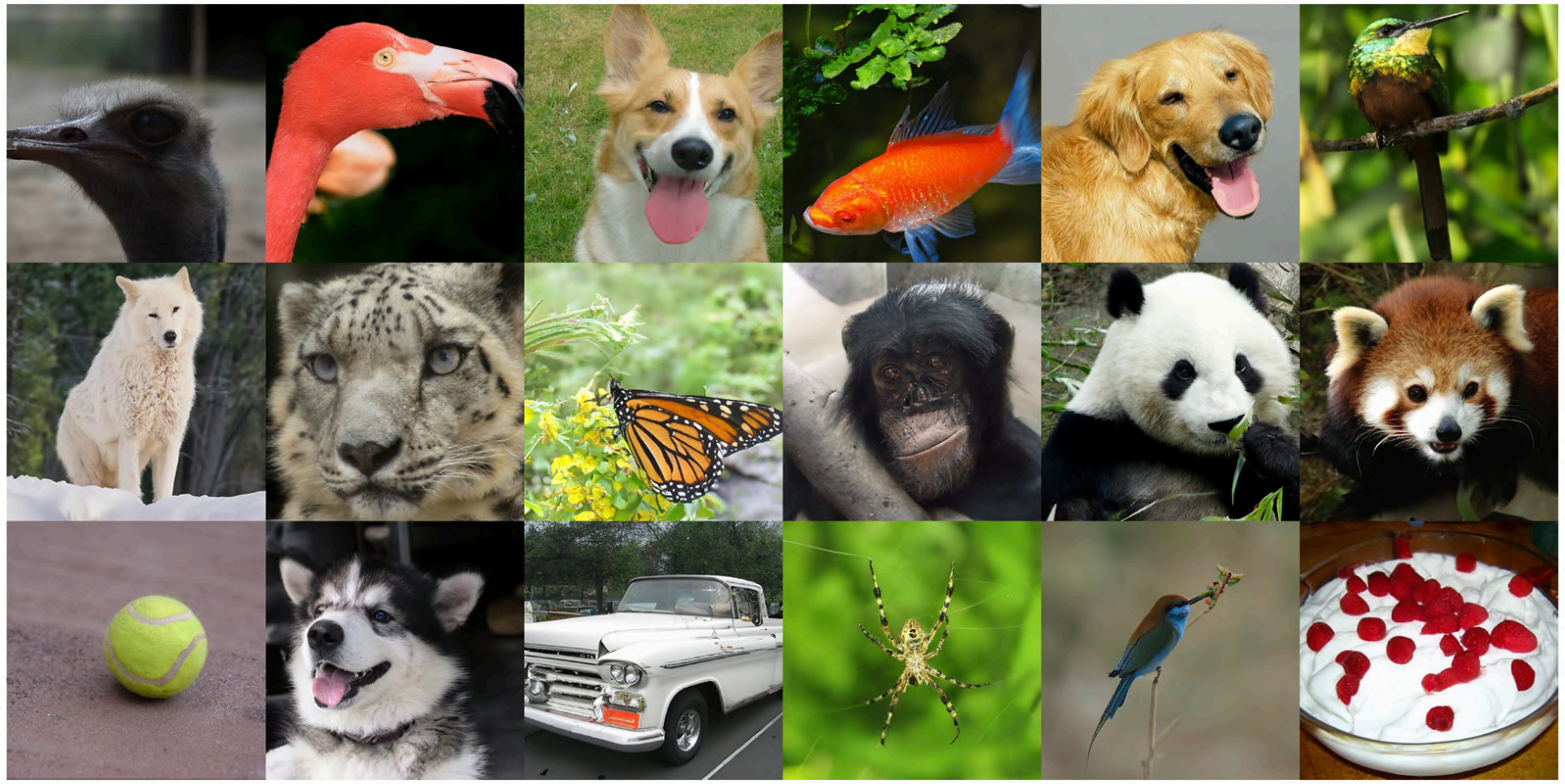


Vision Generative Model Timeline

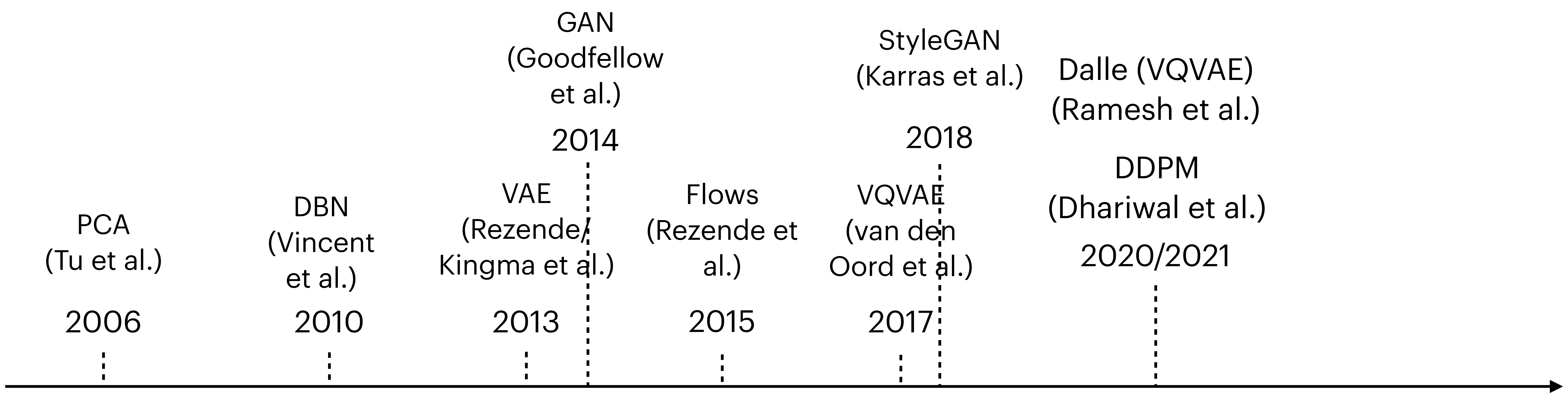
Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com



Vision Generative Model Timeline

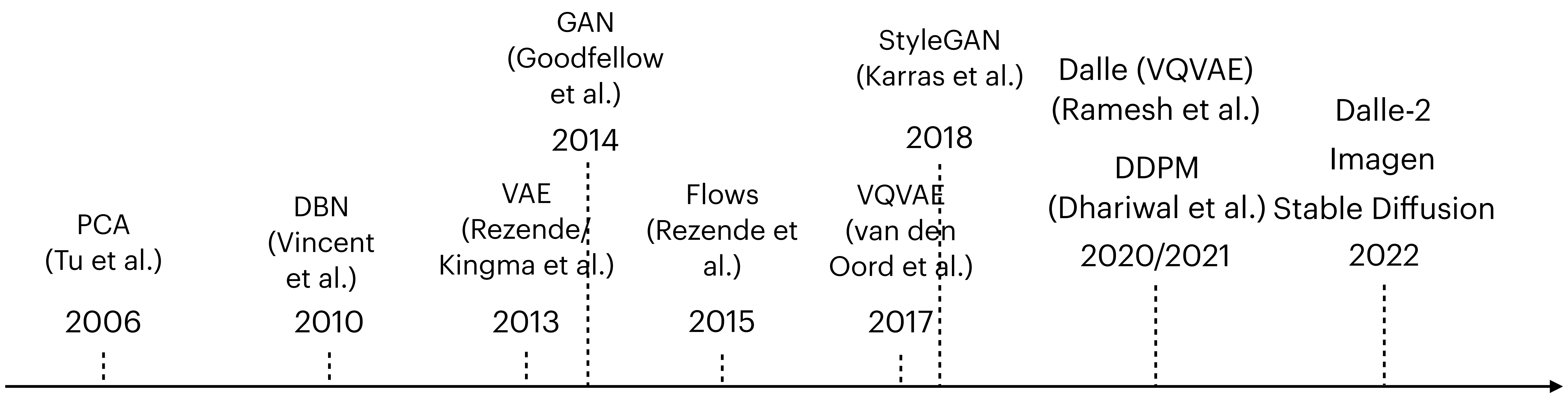


Vision Generative Model Timeline

Zero-Shot Text-to-Image Generation

Aditya Ramesh [1] Mikhail Pavlov [1] Gabriel Goh [1] Scott Gray [1]
Chelsea Voss [1] Alec Radford [1] Mark Chen [1] Ilya Sutskever [1]

Vision Generative Model Timeline



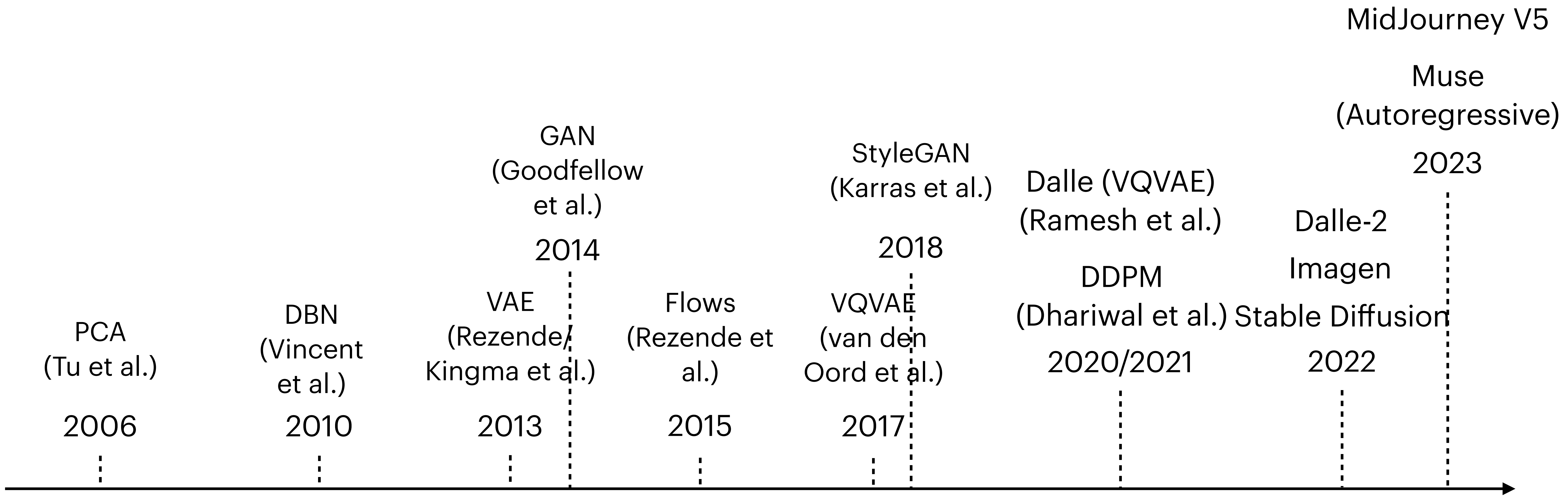
Vision Generative Model Timeline

Dalle-2

Stable Diffusion

Imagen

Vision Generative Model Timeline



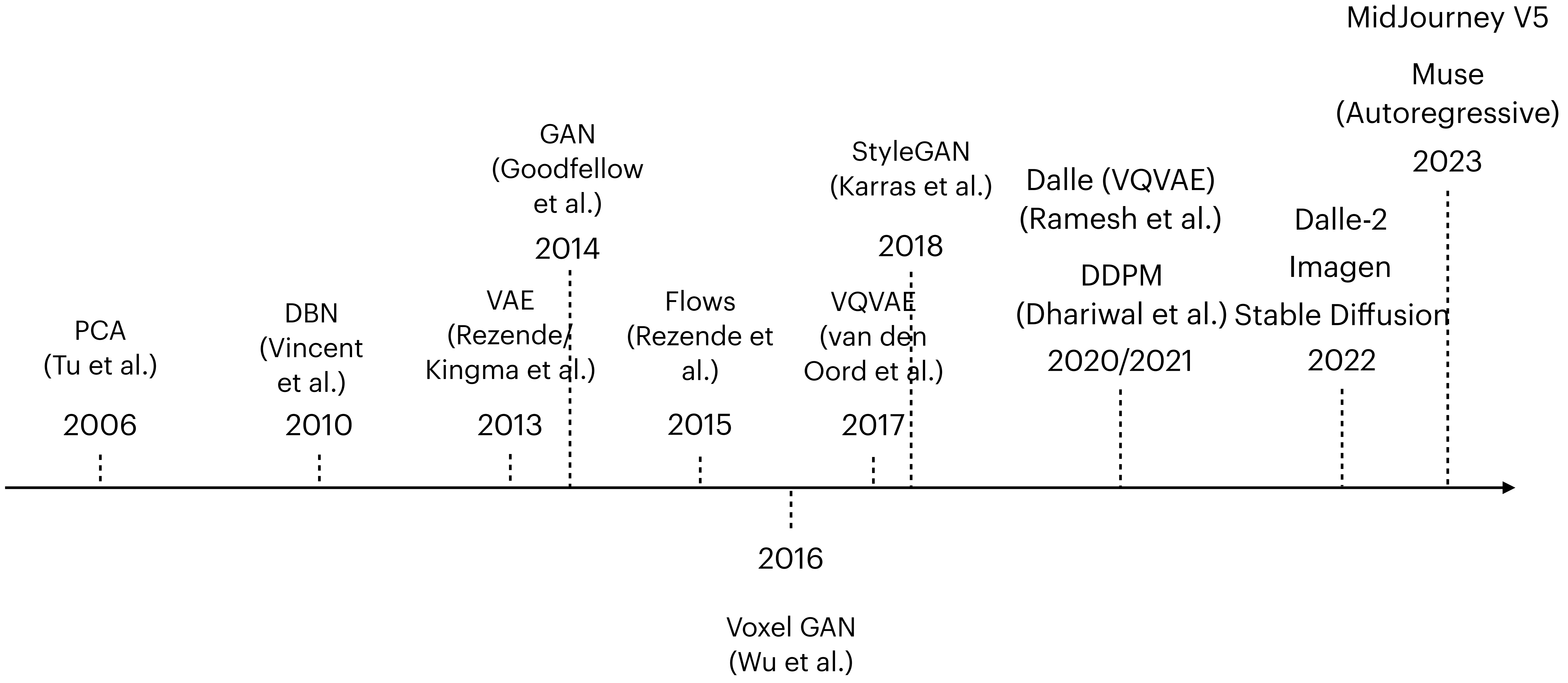
Vision Generative Model Timeline

2006

2018

2023

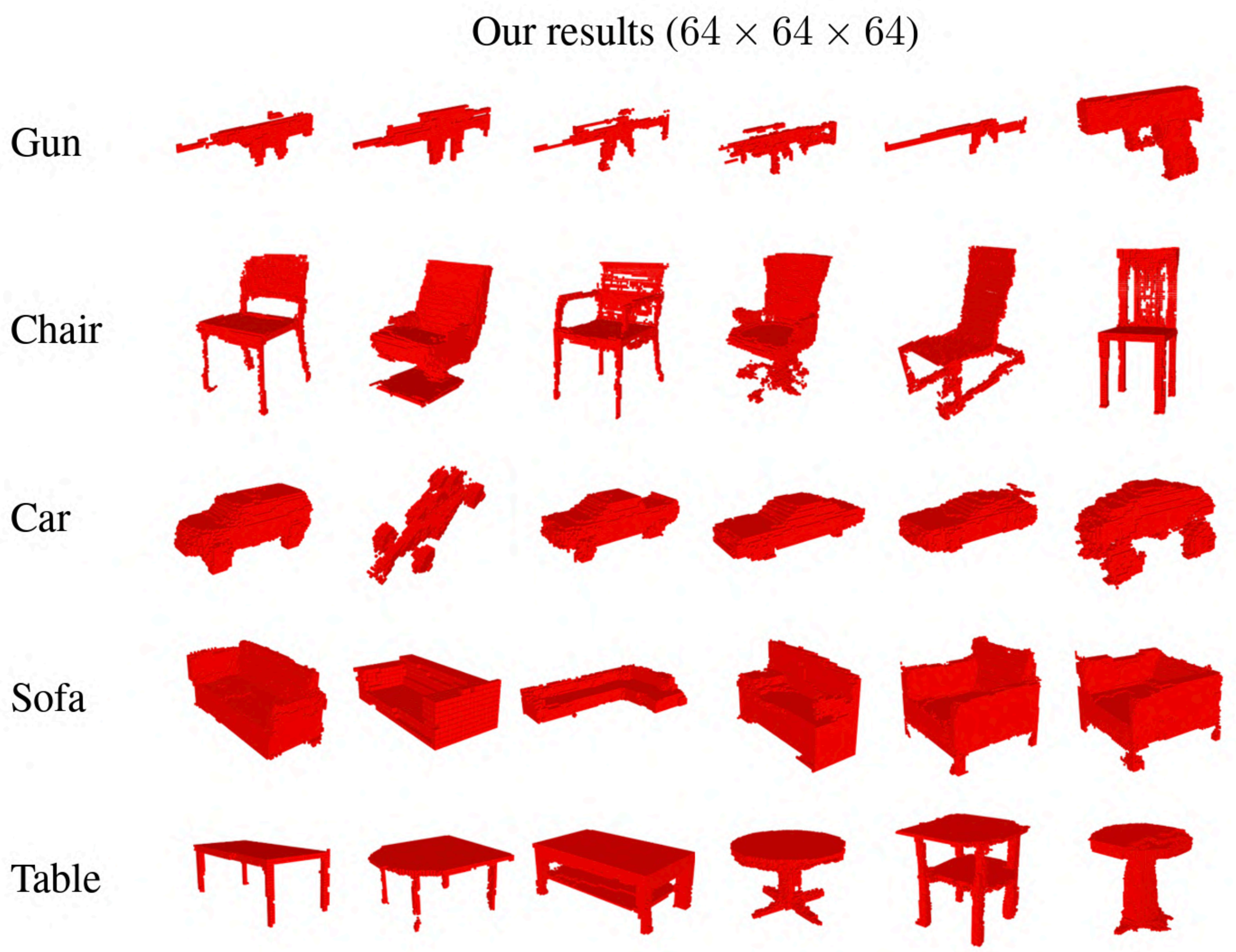
Vision Generative Model Timeline



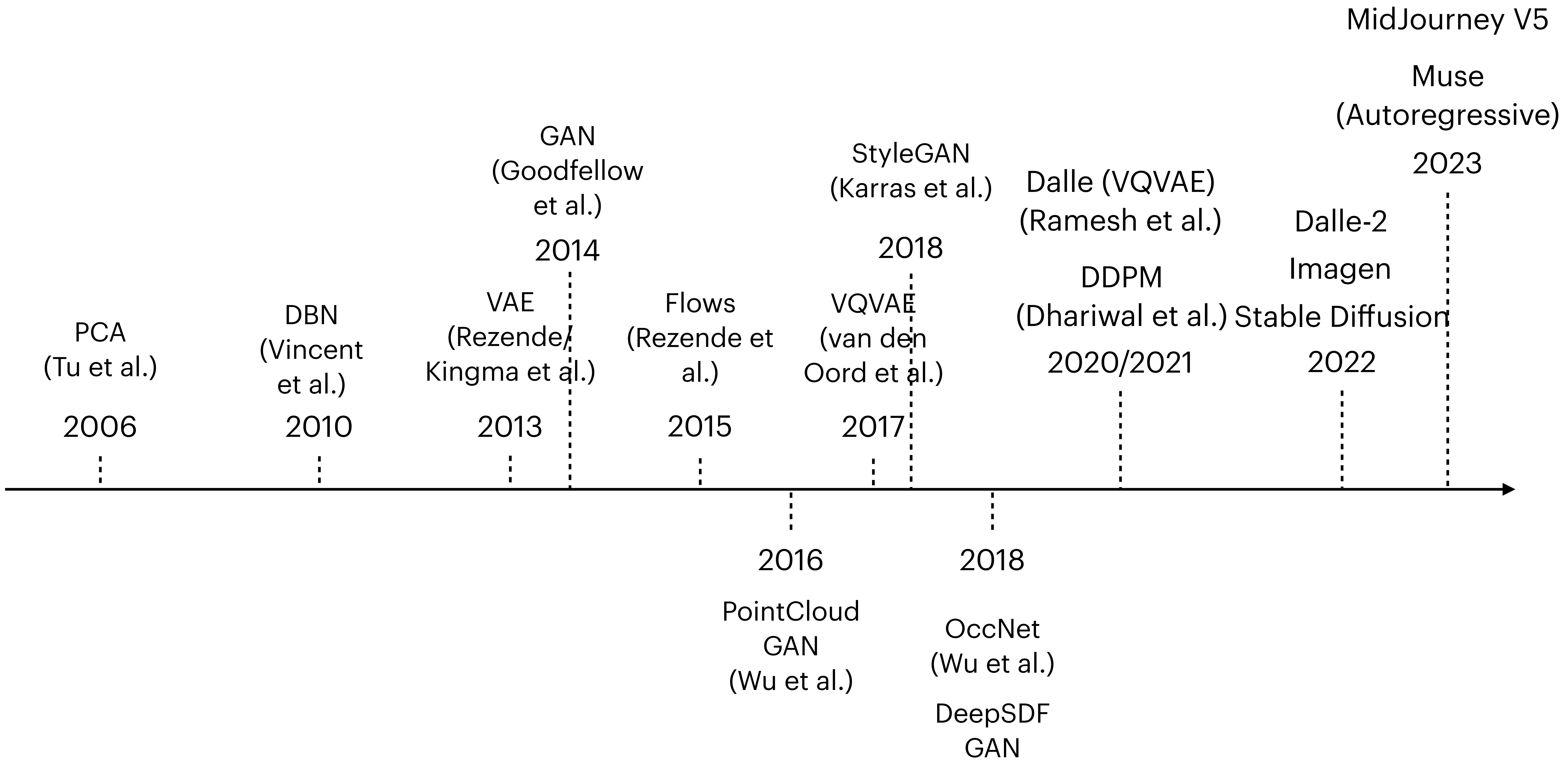
Vision Generative Model Timeline

Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling

Jiajun Wu* MIT CSAIL
Chengkai Zhang* MIT CSAIL
Tianfan Xue MIT CSAIL
William T. Freeman MIT CSAIL, Google Research
Joshua B. Tenenbaum MIT CSAIL



Vision Generative Model Timeline



Vision Generative Model Timeline

Occupancy Networks: Learning 3D Reconstruction in Function Space

Lars Mescheder[1] Michael Oechsle[1,2] Michael Niemeyer[1] Sebastian Nowozin[3†] Andreas Geiger[1]
[1]Autonomous Vision Group, MPI for Intelligent Systems and University of Tübingen
[2]ETAS GmbH, Stuttgart
[3]Google AI Berlin
{firstname.lastname}@tue.mpg.de nowozin@gmail.com

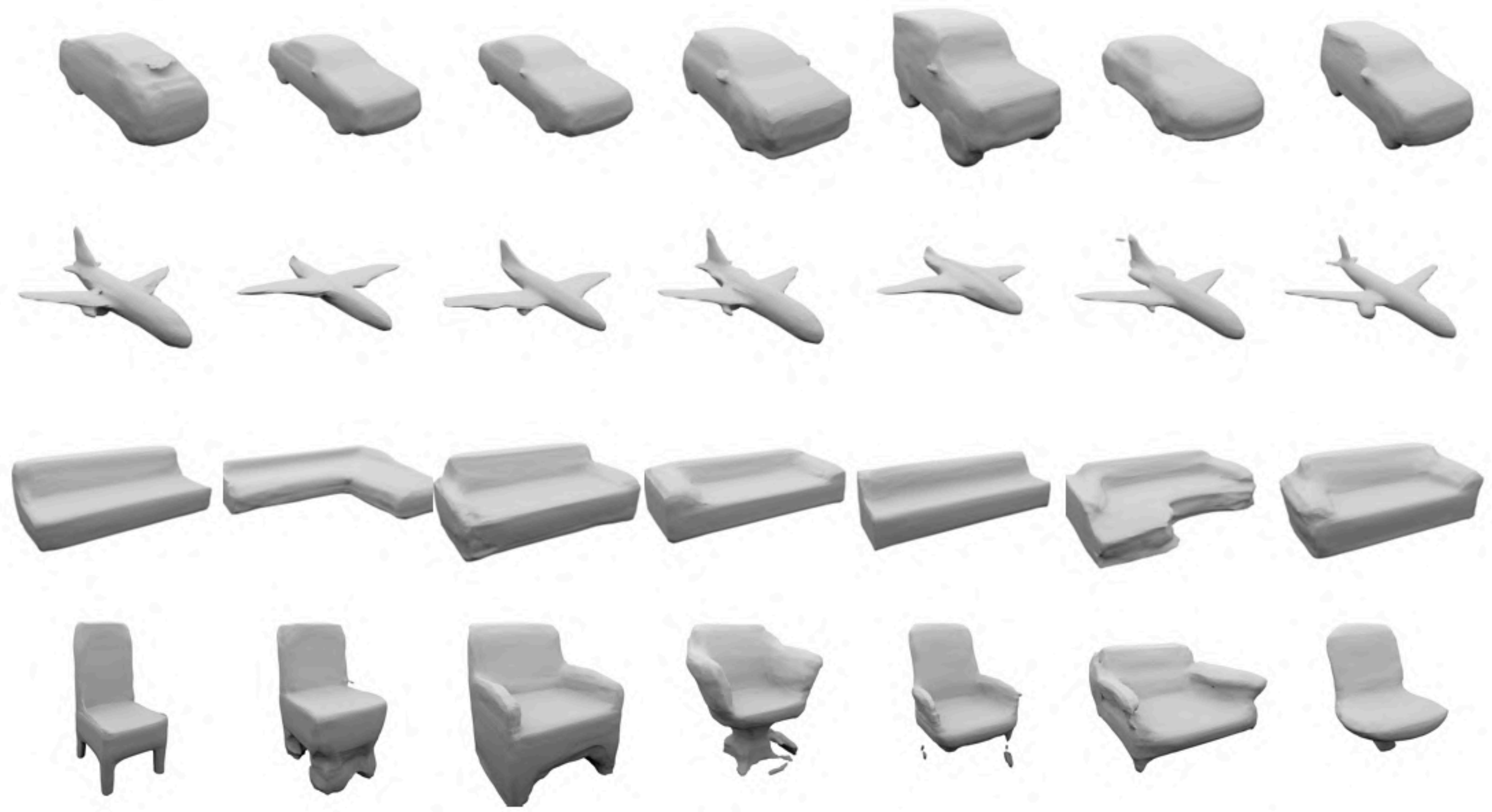
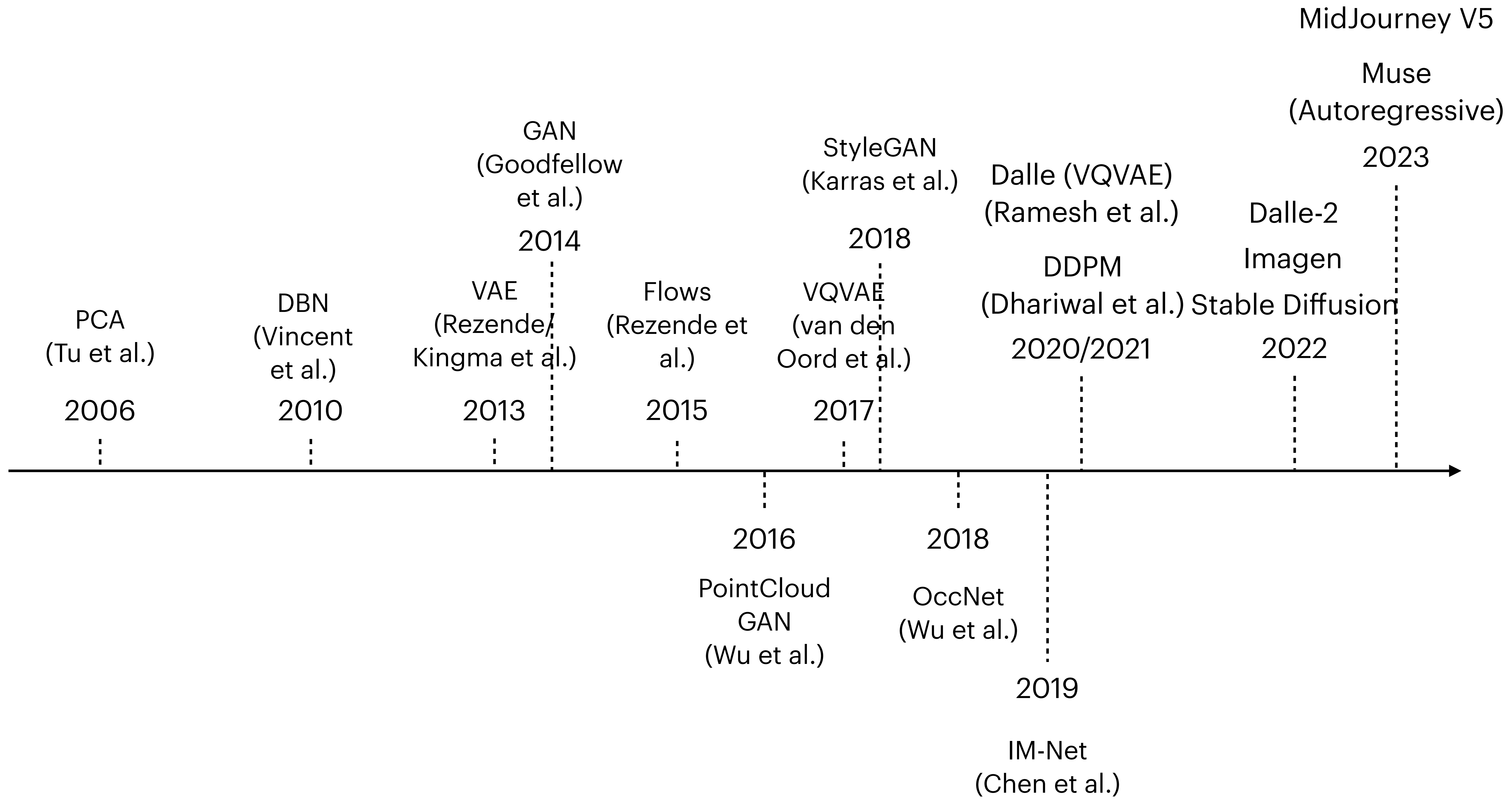


Figure 7: **Unconditional 3D Samples.** Random samples of our unsupervised models trained on the categories "car", "airplane", "sofa" and "chair" of the ShapeNet dataset. We see that our models are able to capture the distribution of 3D objects and produce compelling new samples.

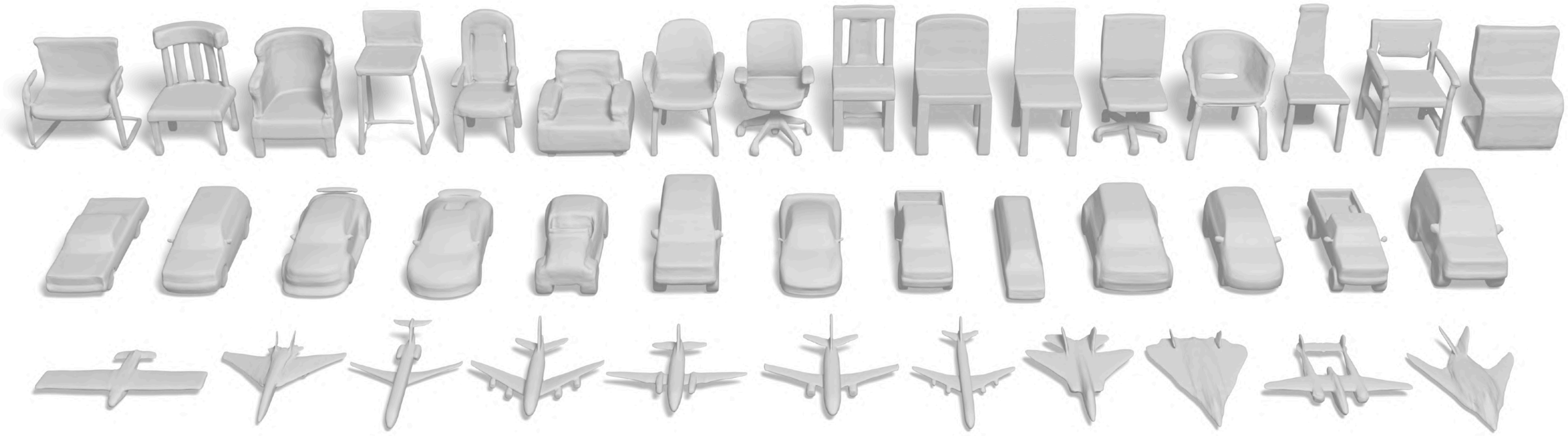
Vision Generative Model Timeline



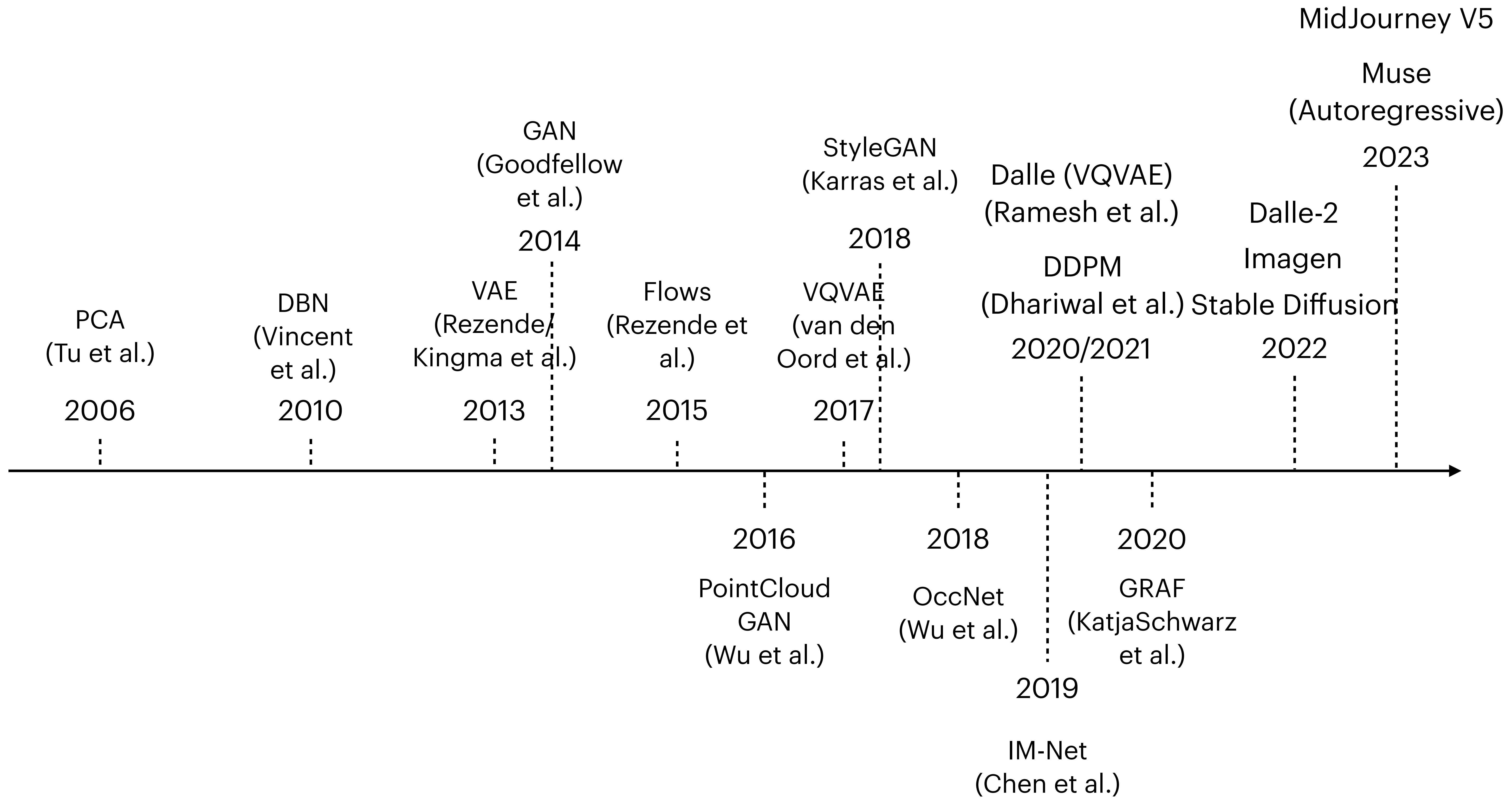
Learning Implicit Fields for Generative Shape Modeling

Zhiqin Chen
Simon Fraser University
zhiqinc@sfu.ca

Hao Zhang
Simon Fraser University
haoz@sfu.ca

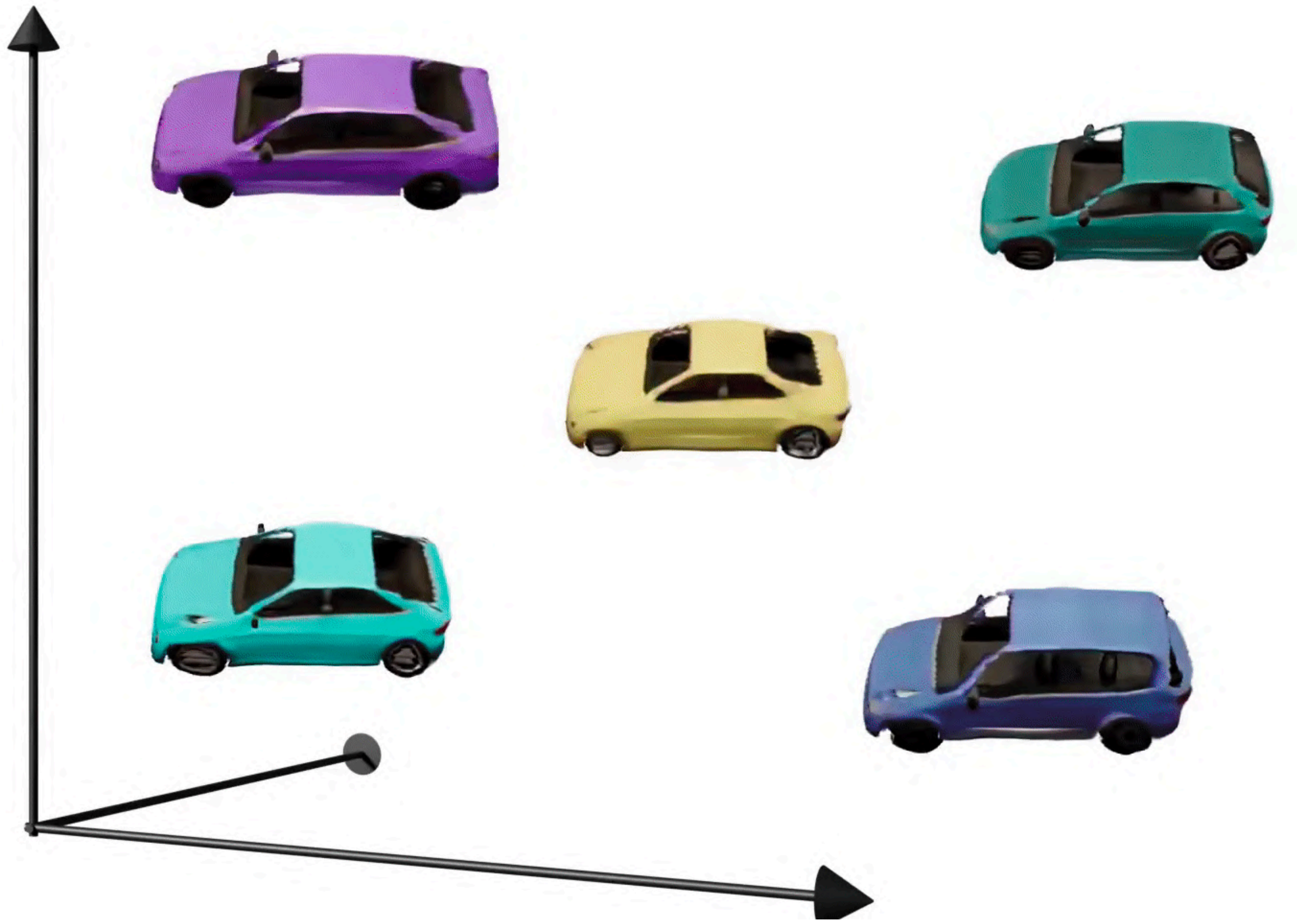


Vision Generative Model Timeline

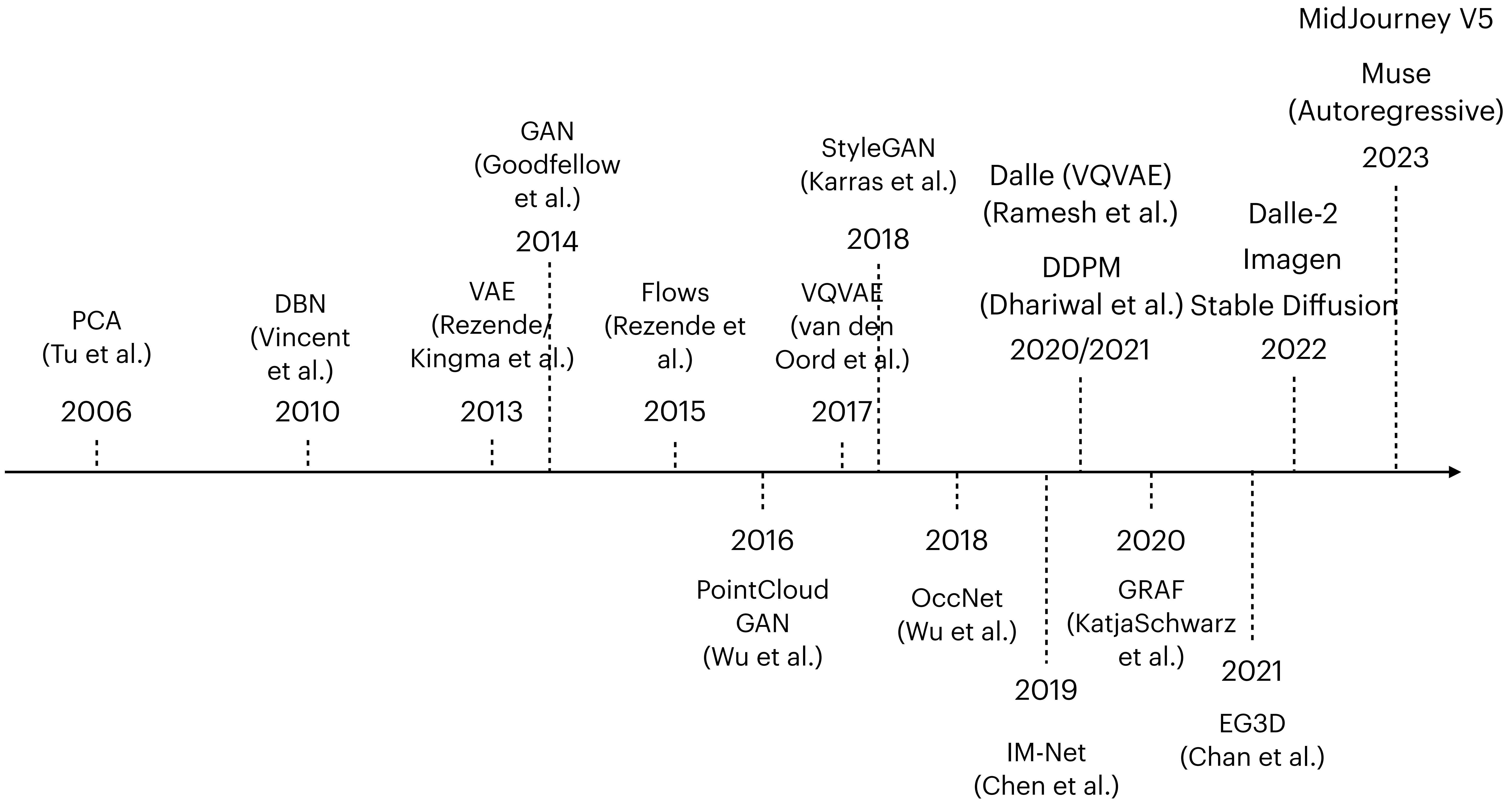


GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis

Katja Schwarz* **Yiyi Liao*** **Michael Niemeyer** **Andreas Geiger**
Autonomous Vision Group
MPI for Intelligent Systems and University of Tübingen
{firstname.lastname}@tue.mpg.de



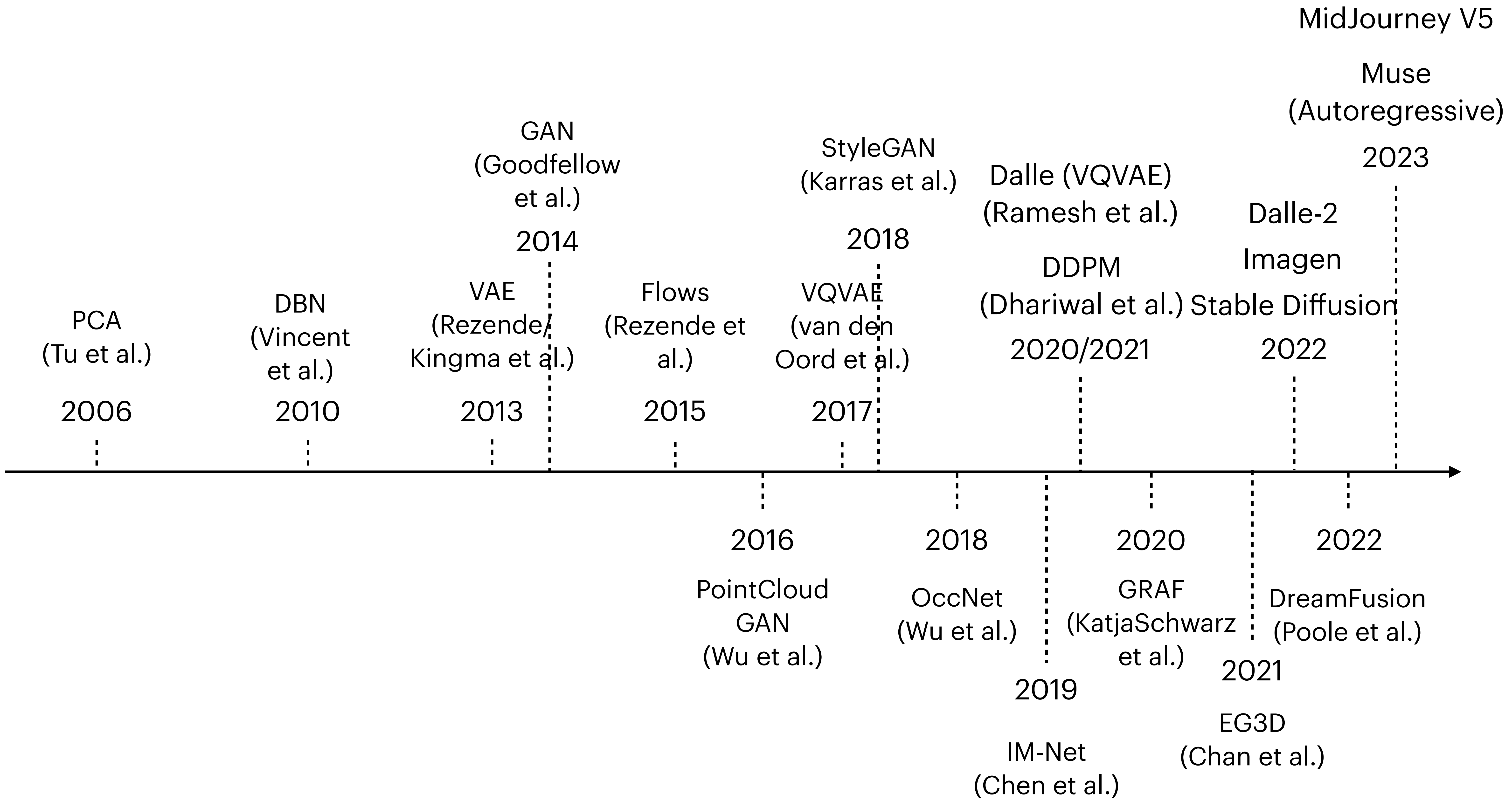
Vision Generative Model Timeline



Efficient Geometry-aware 3D Generative Adversarial Networks

Eric R. Chan *†[1,2], Connor Z. Lin*[1], Matthew A. Chan*[1], Koki Nagano*[2], Boxiao Pan[1], Shalini De Mello[2], Orazio Gallo[2], Leonidas Guibas[1], Jonathan Tremblay[2], Sameh Khamis[2], Tero Karras[2], and Gordon Wetzstein[1]

Vision Generative Model Timeline



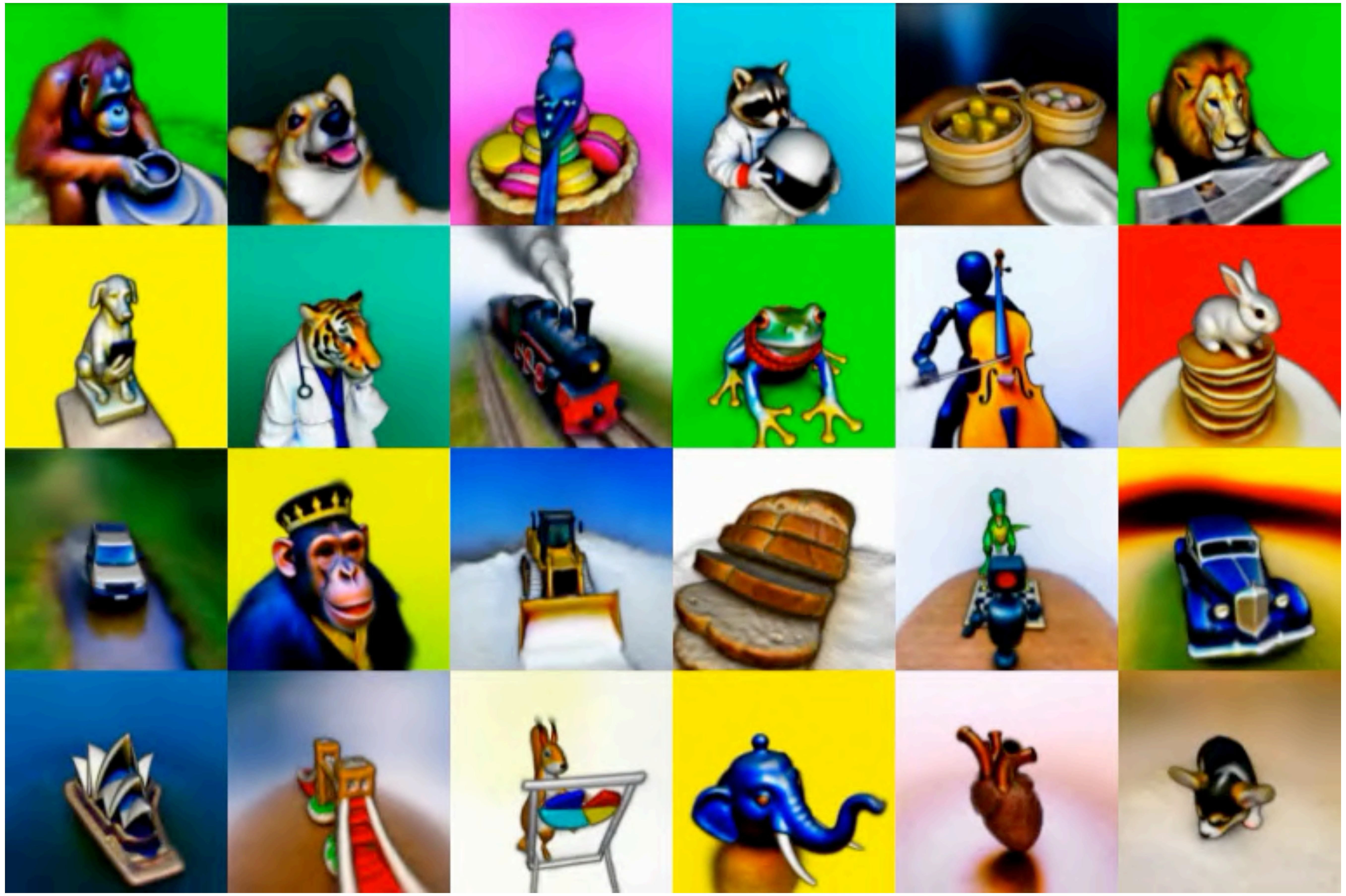
DreamFusion: Text-to-3D using 2D Diffusion

Ben Poole
Google Research

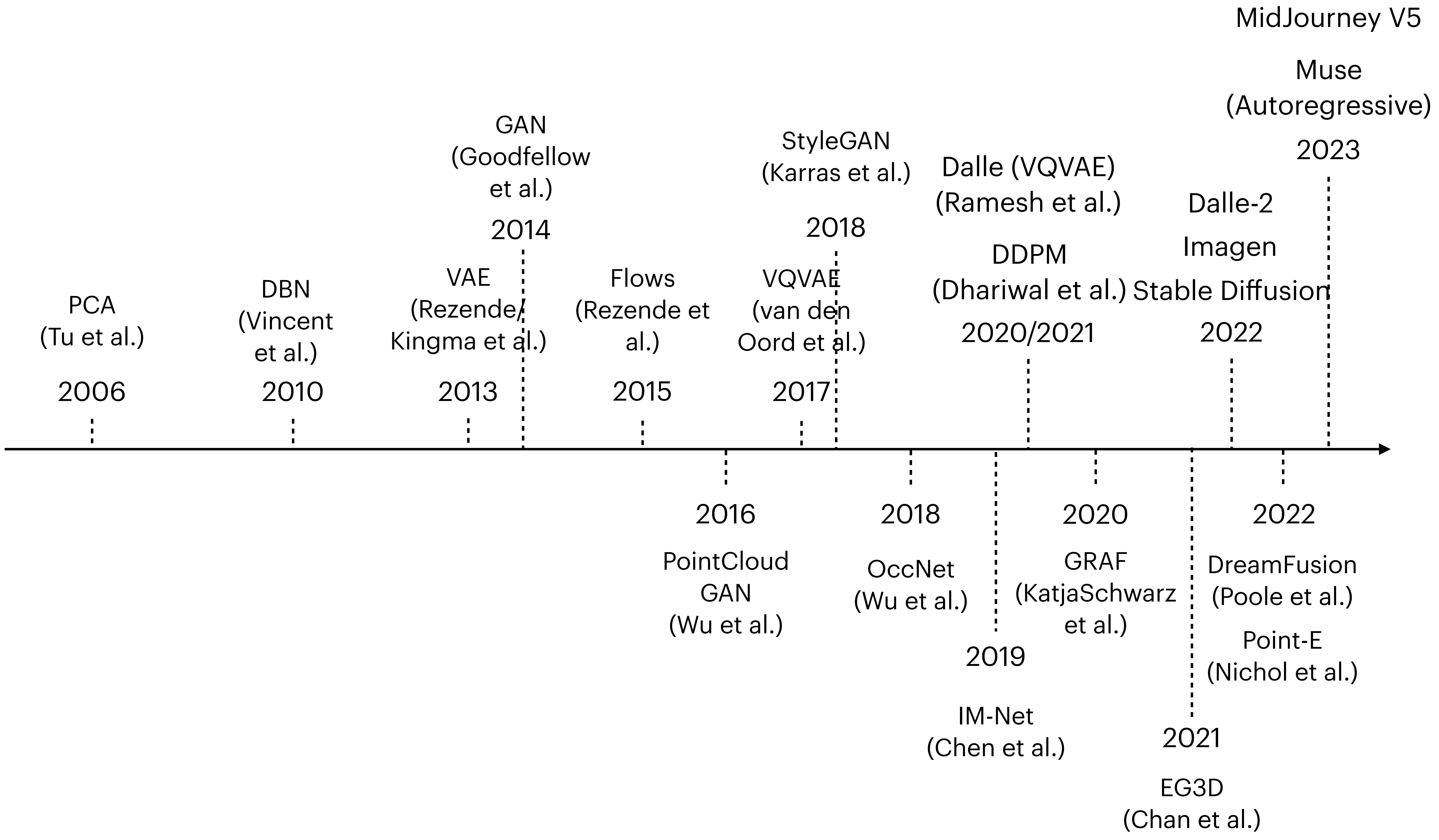
Ajay Jain
UC Berkeley

Jonathan T. Barron
Google Research

Ben Mildenhall
Google Research

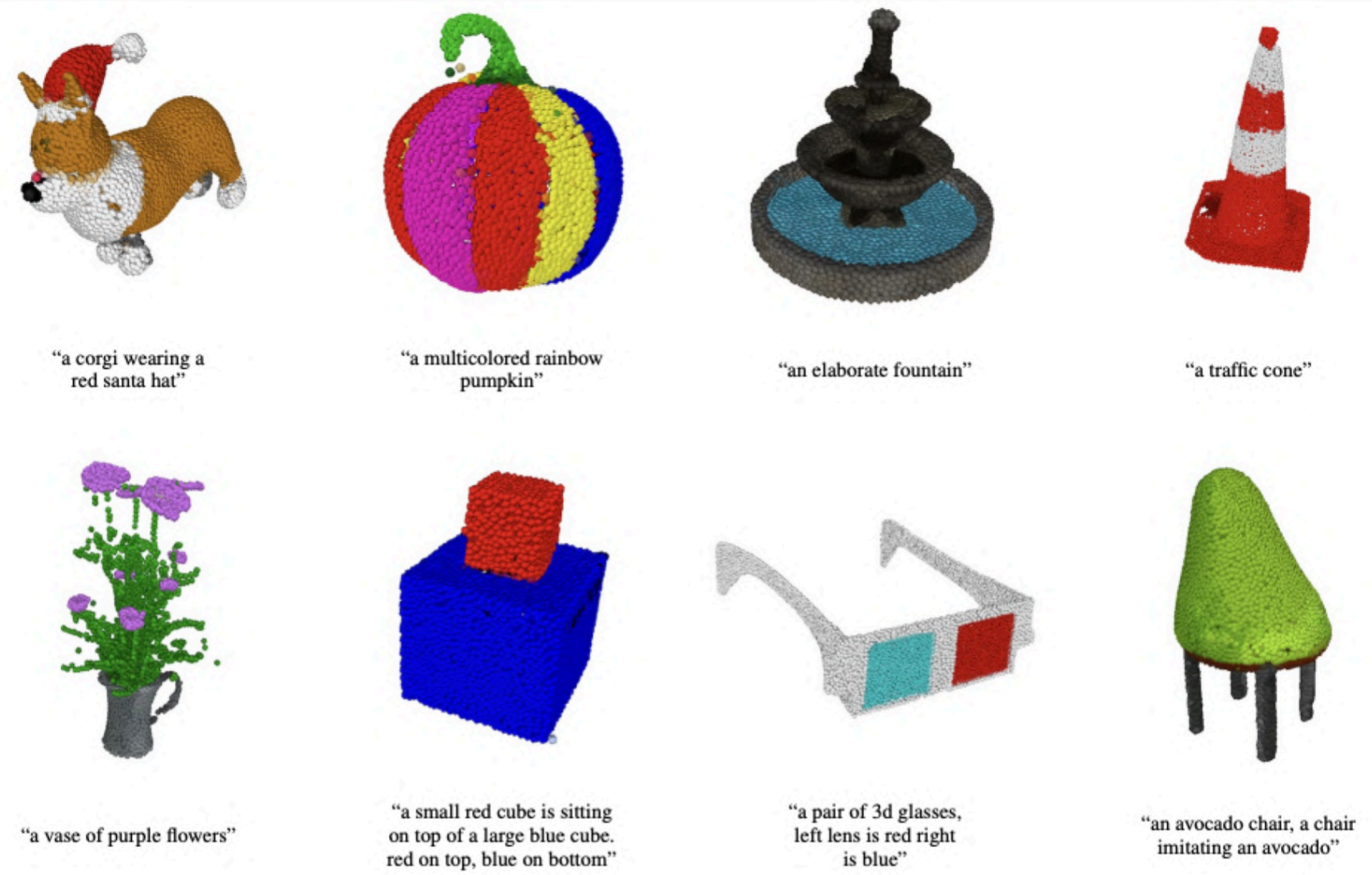


Vision Generative Model Timeline

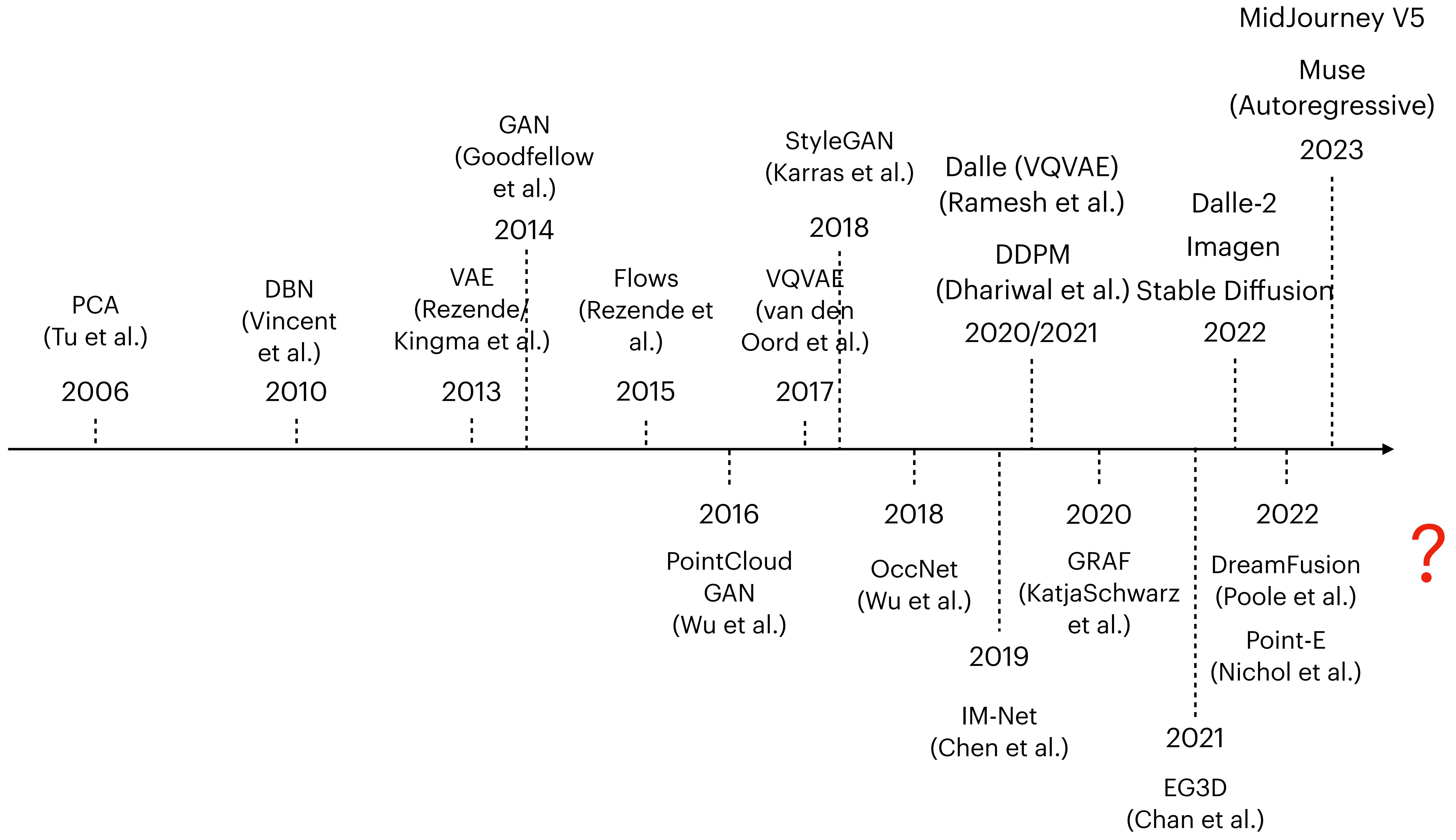


Point·E: A System for Generating 3D Point Clouds from Complex Prompts

Alex Nichol [*1] Heewoo Jun [*1] Prafulla Dhariwal [1] Pamela Mishkin [1] Mark Chen [1]

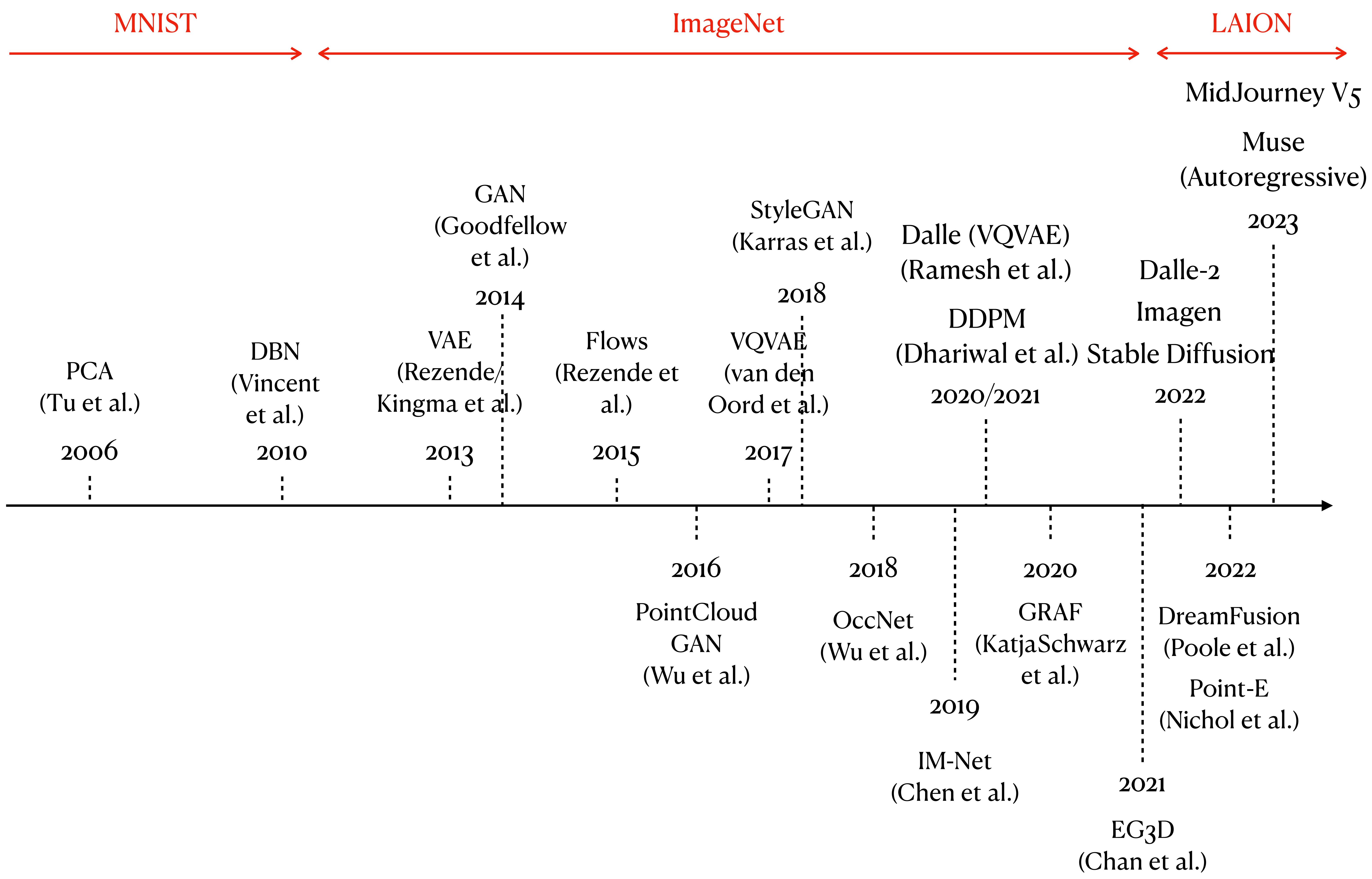


What's next?



PCA
(Tu et al.)
2006
DBN
(Vincent et al.)
2010
VAE
(Rezende/ Kingma et al.)
2013
GAN
(Goodfellow et al.)
2014
Flows
(Rezende et al.)
2015
VQVAE
(van den Oord et al.)
2017
StyleGAN
(Karras et al.)
2018
Dalle (VQVAE)
(Ramesh et al.)
DDPM
(Dhariwal et al.)
2020/2021
Dalle-2
Imagen
Stable Diffusion
2022
MidJourney V5
Muse
(Autoregressive)
2023
2016
PointCloud GAN
(Wu et al.)
2018
OccNet
(Wu et al.)
2019
IM-Net
(Chen et al.)
2020
GRAF
(KatjaSchwarz et al.)
2021
EG3D
(Chan et al.)
2022
DreamFusion
(Poole et al.)
Point-E
(Nichol et al.)
?

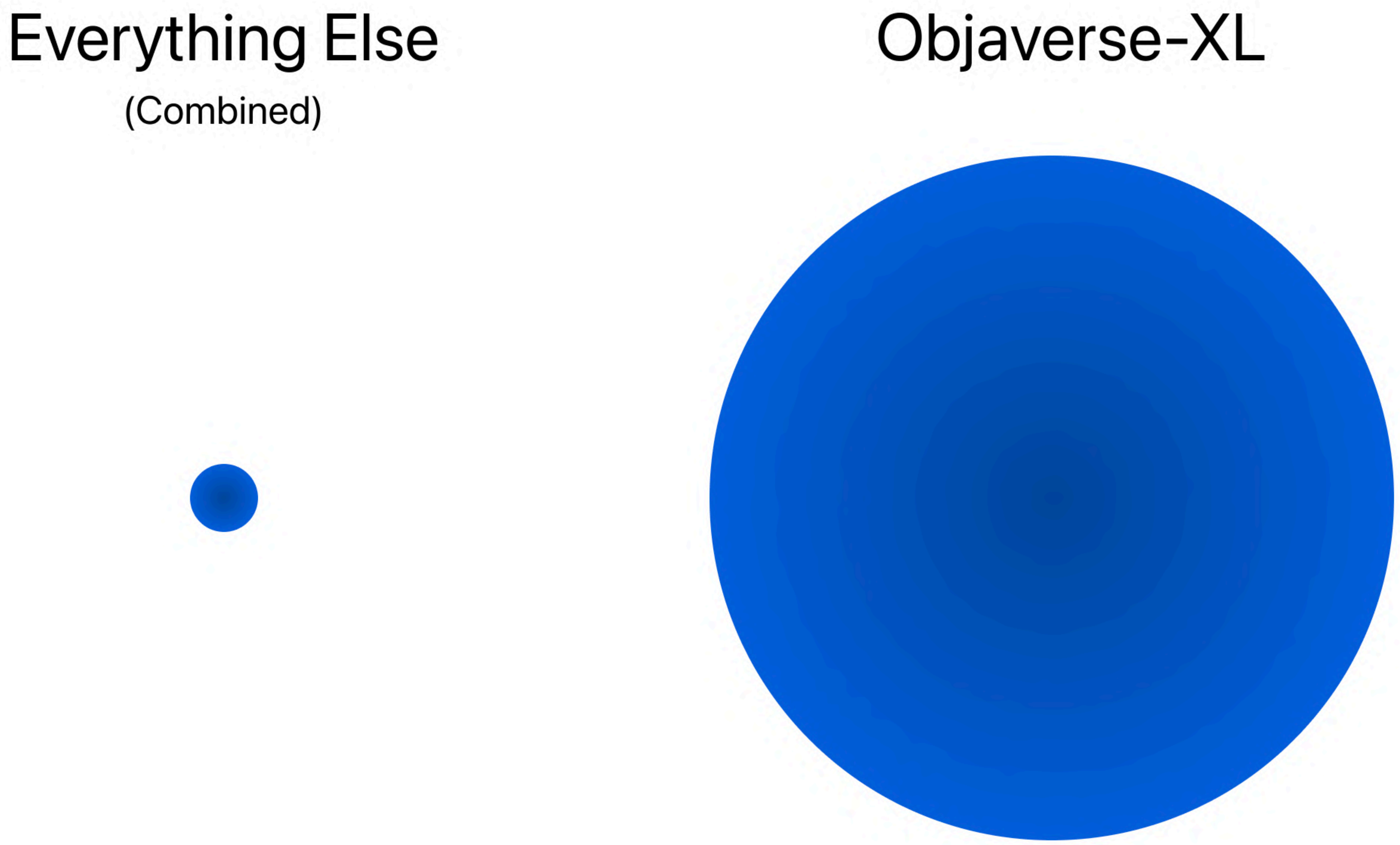
High-Quality Dataset is the Unseen Hero



Objaverse-XL: 10M+ high-quality 3D assets

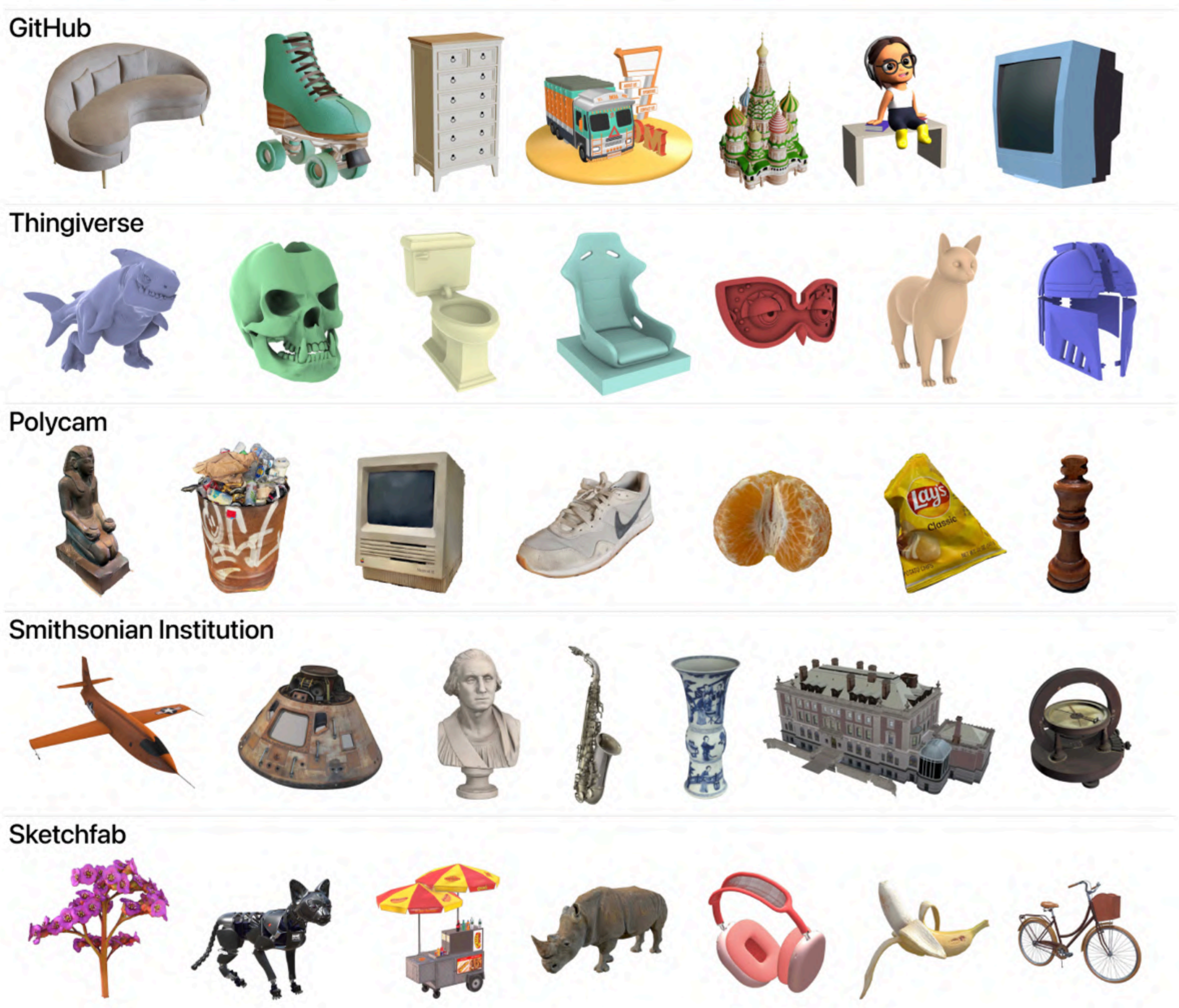
Deitke, **Liu**, et al. NeurIPS 2023.

Objaverse-XL: 10M+ high-quality 3D assets



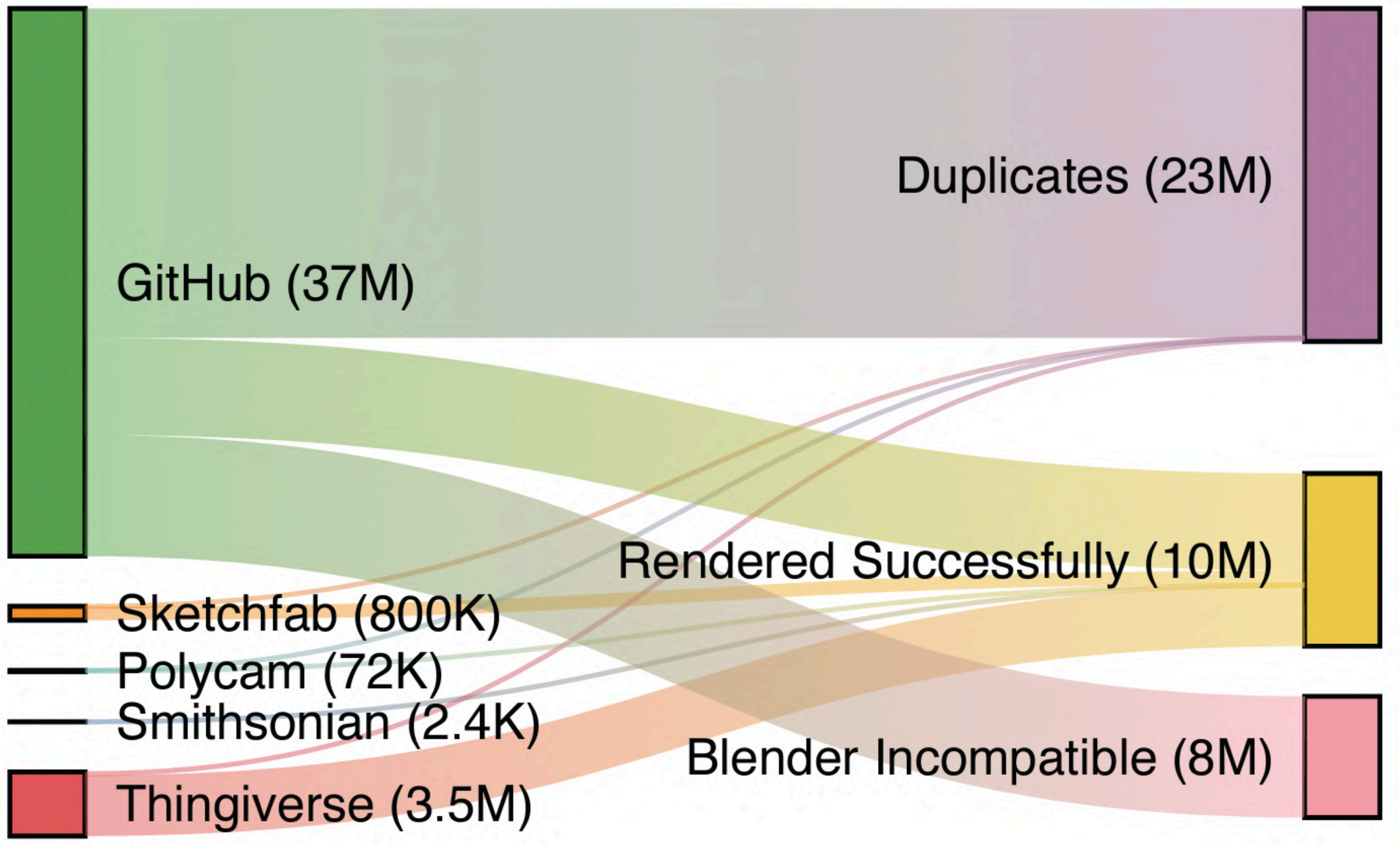
Deitke, **Liu**, et al. NeurIPS 2023.

Objaverse-XL: 10M+ high-quality 3D assets



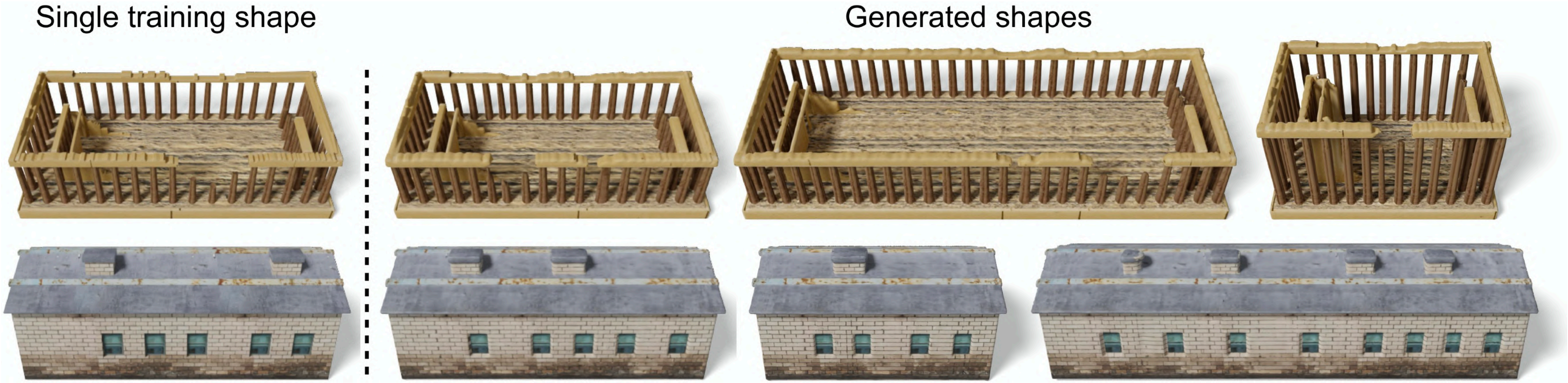
Deitke, **Liu**, et al. NeurIPS 2023.

Objaverse-XL: 10M+ high-quality 3D assets



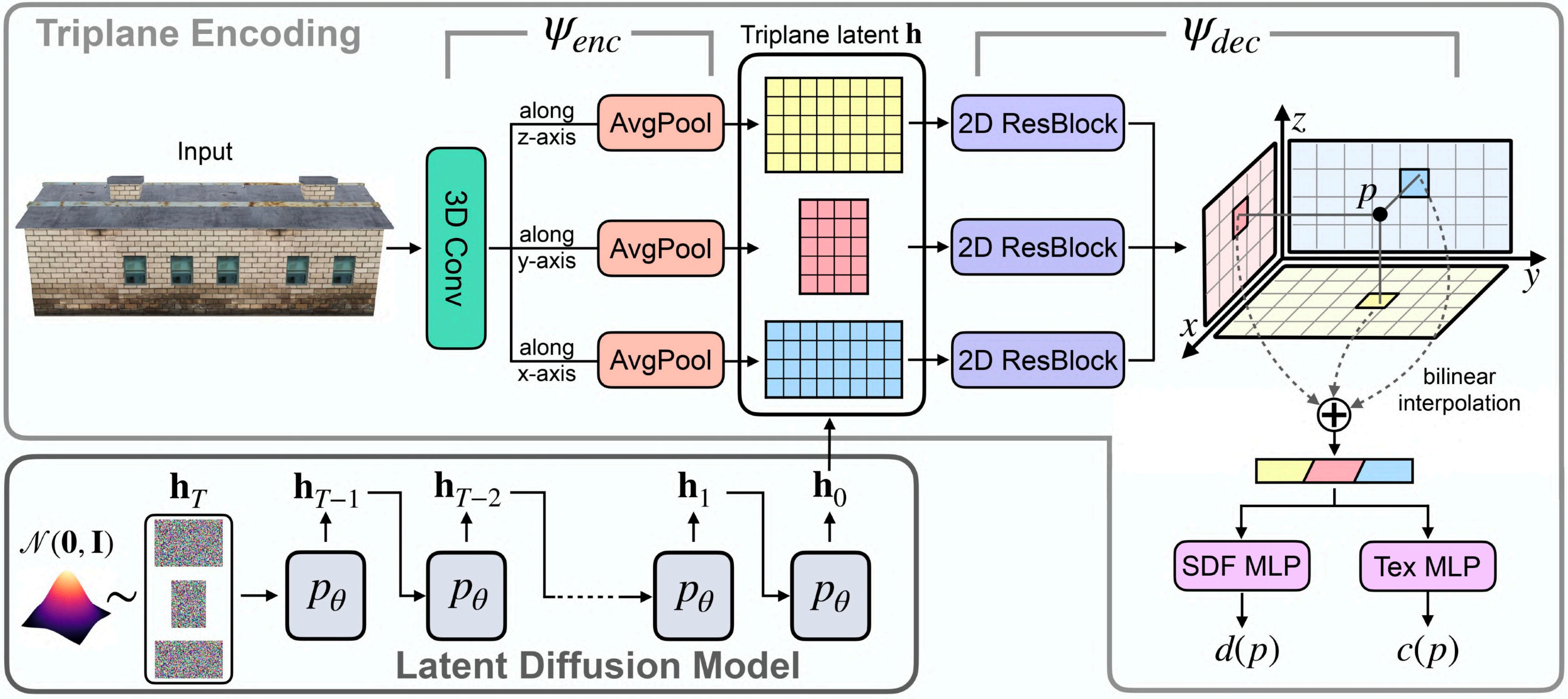
Deitke, **Liu**, et al. NeurIPS 2023.

Sin3DM: a method for 3D data augmentation



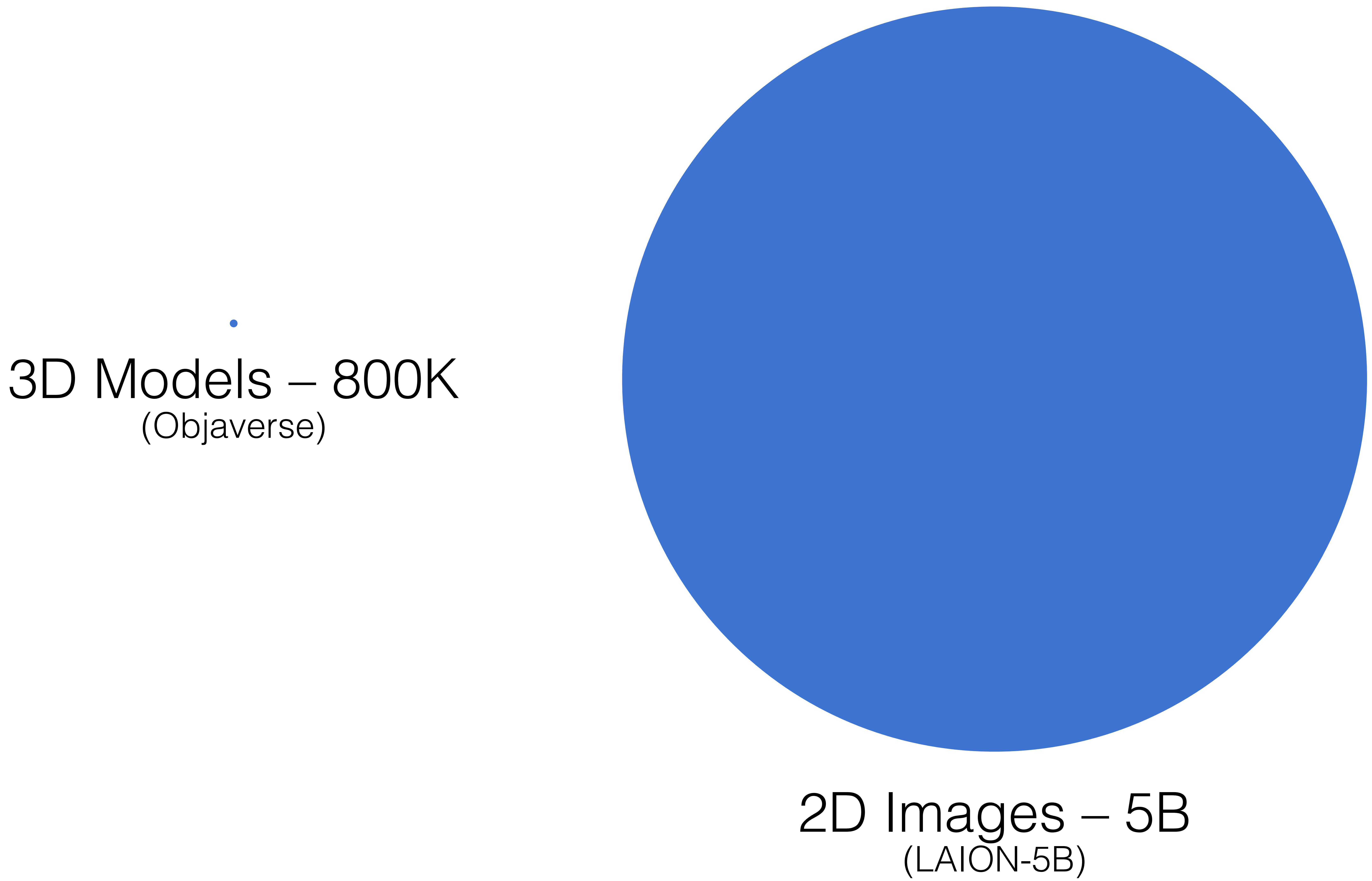
Wu, **Liu**, Vondrick, Zheng, ICLR 2024

Learning a Diffusion Model from a Single 3D Textured Shape

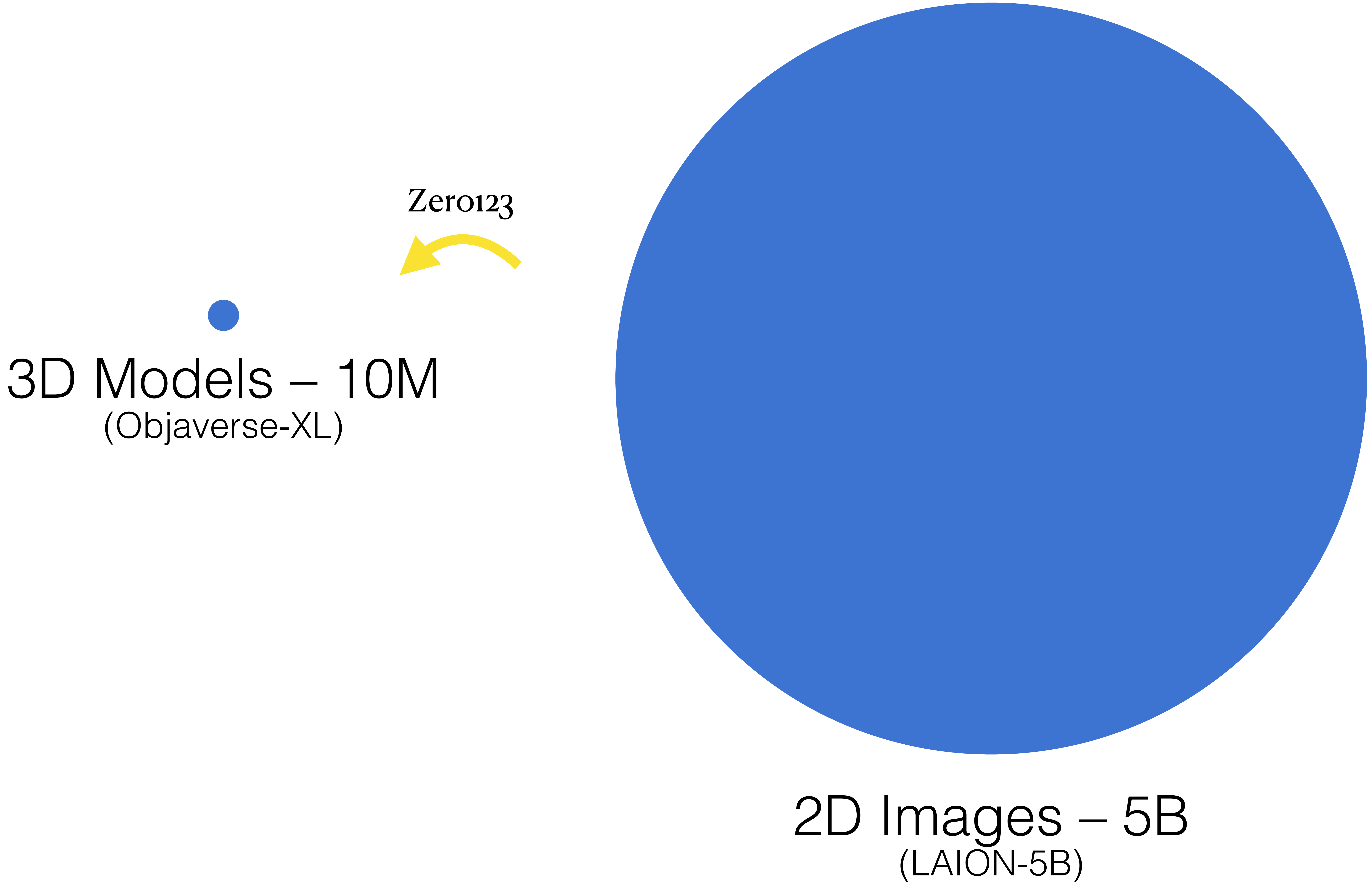


Wu, **Liu**, Vondrick, Zheng, ICLR 2024

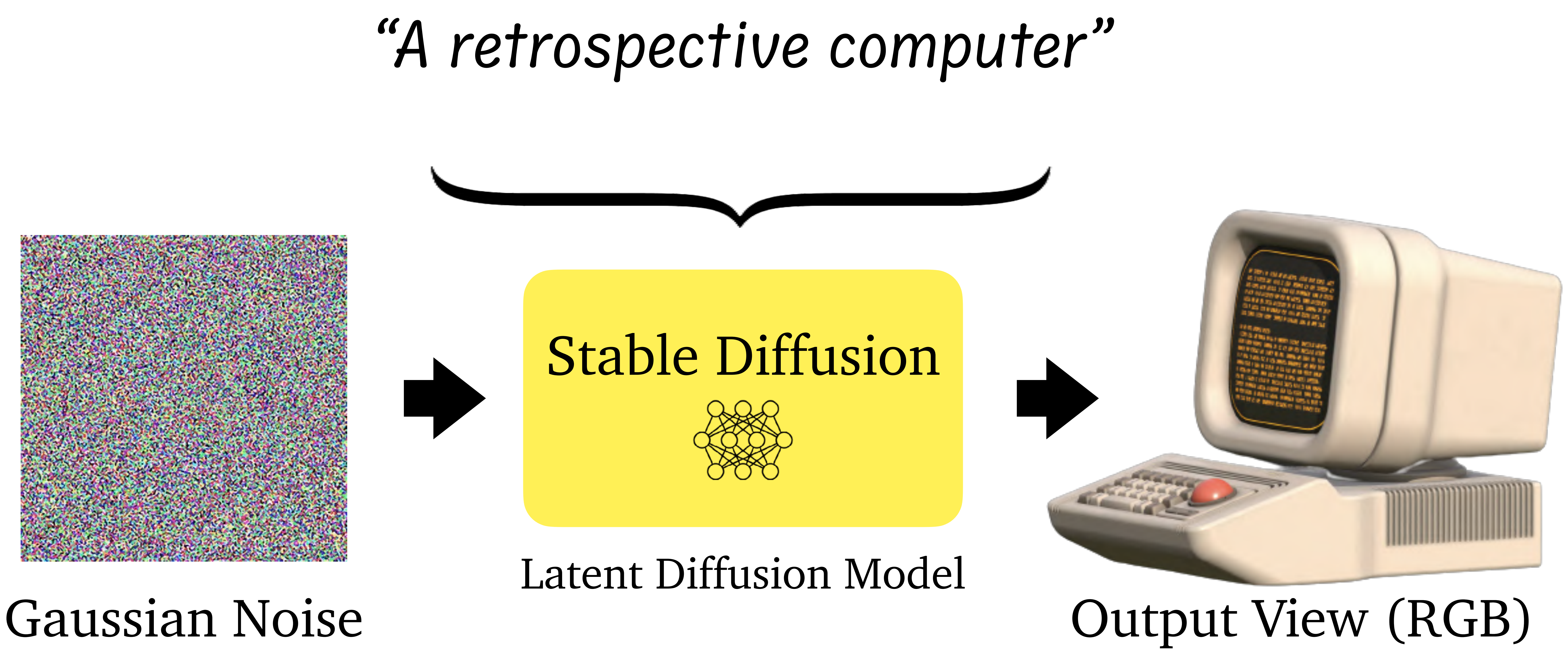
Is Dataset Scaling Enough for 3D Gen?



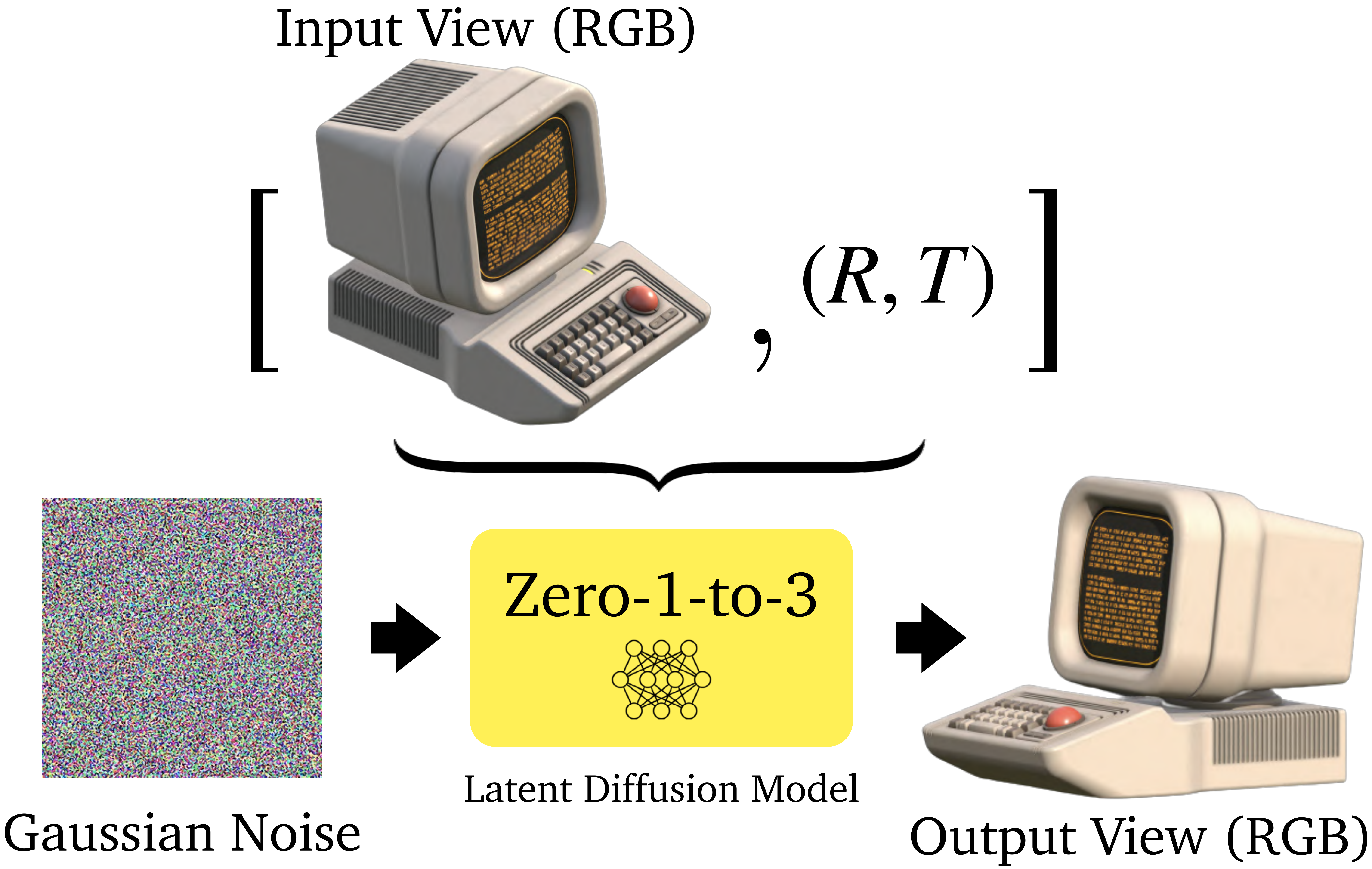
Is Dataset Scaling Enough for 3D Gen?



Stable Diffusion



Zero-1-to-3: Zero-Shot One Image to 3D Objects



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

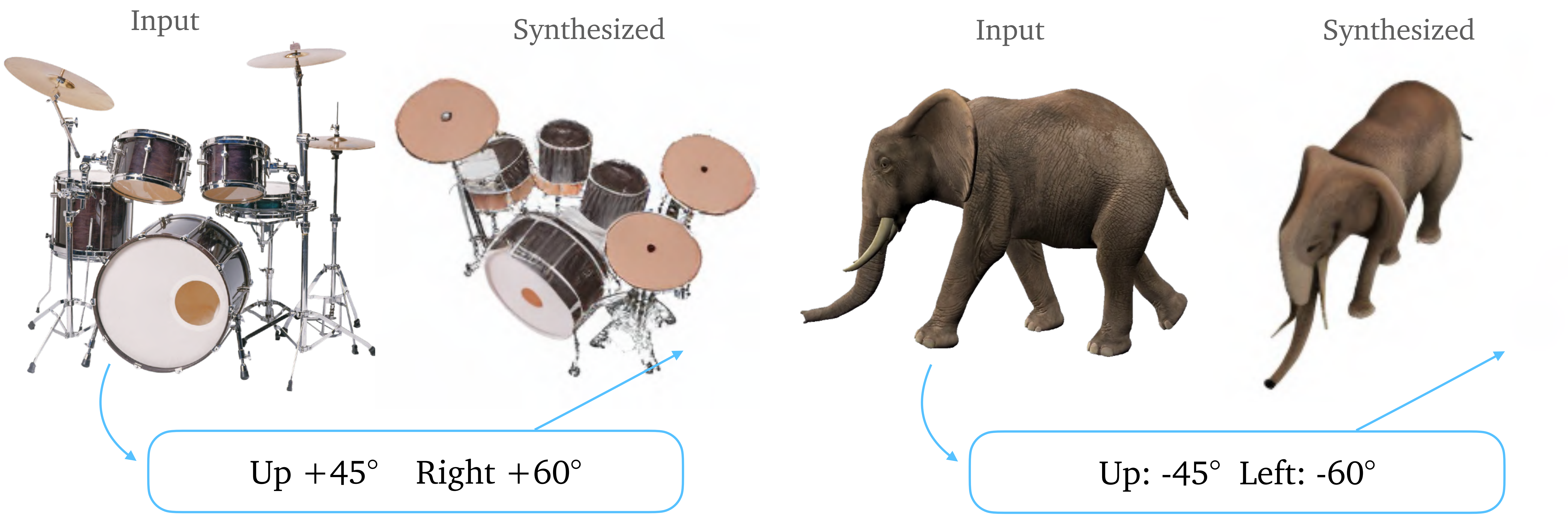
Dataset of Novel Views

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

Everyday Objects

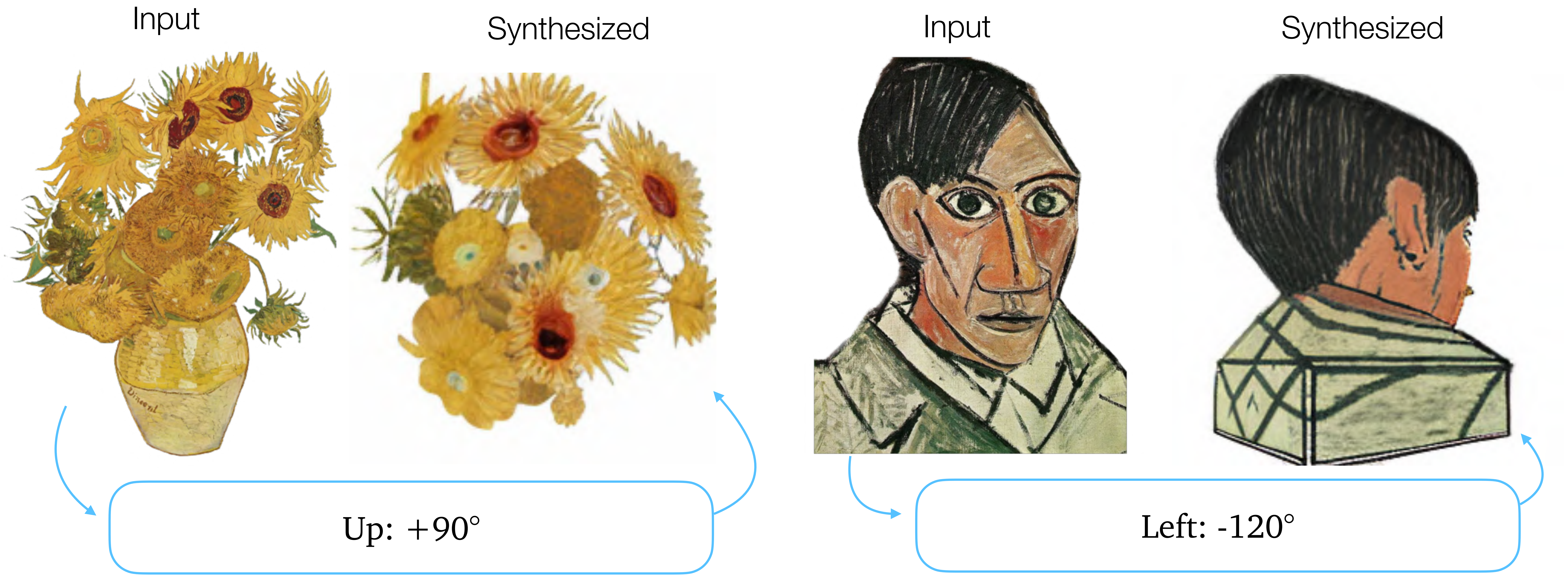
Liu, Wu, Van Hoorick, Tokmakov, Zakharov, Vondrick. ICCV 2023.

Complex/Deformable Objects



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

Abstract Paintings

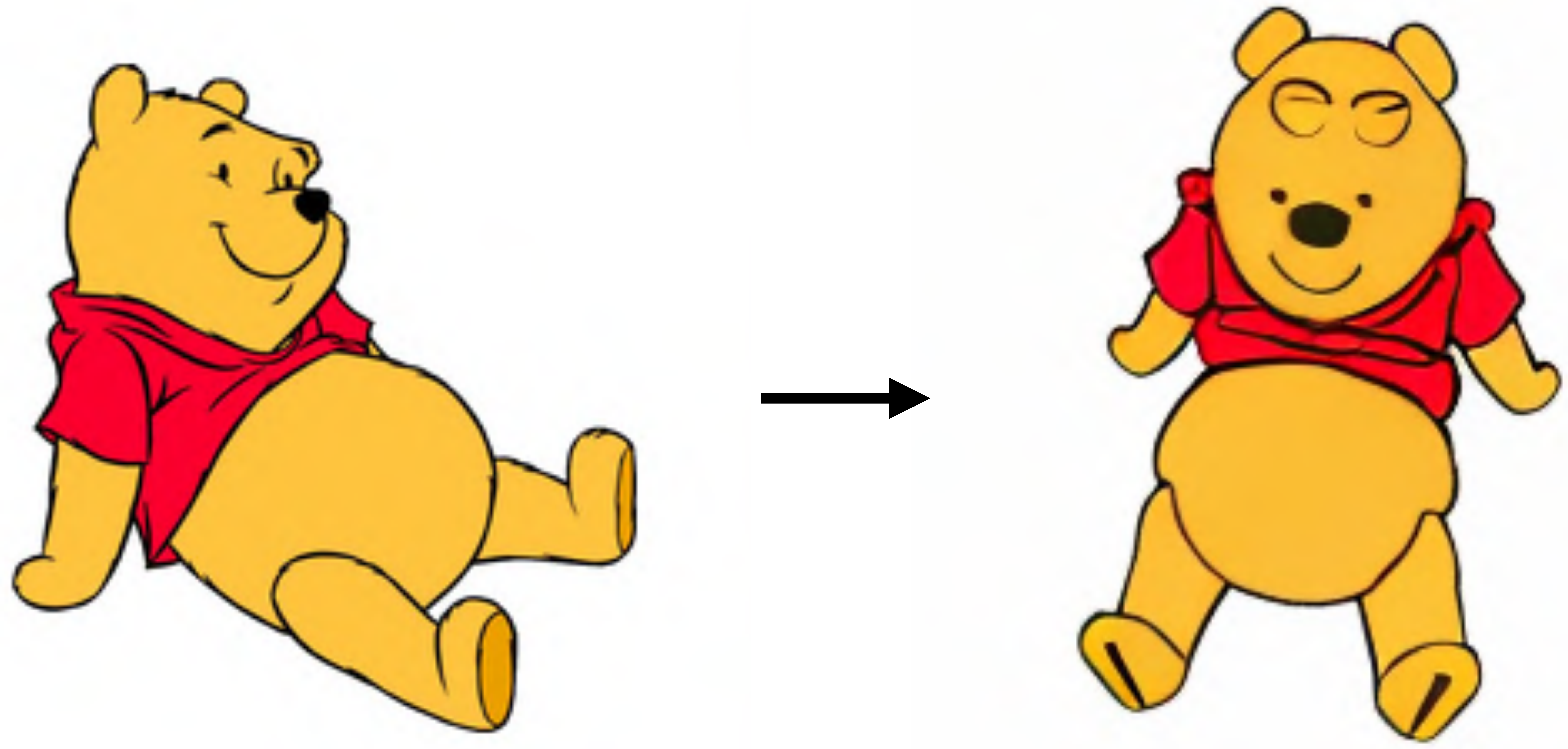


Liu, Wu, Van Hoorick, Tokmakov, Zakharov, Vondrick. ICCV 2023.

Oil Paintings

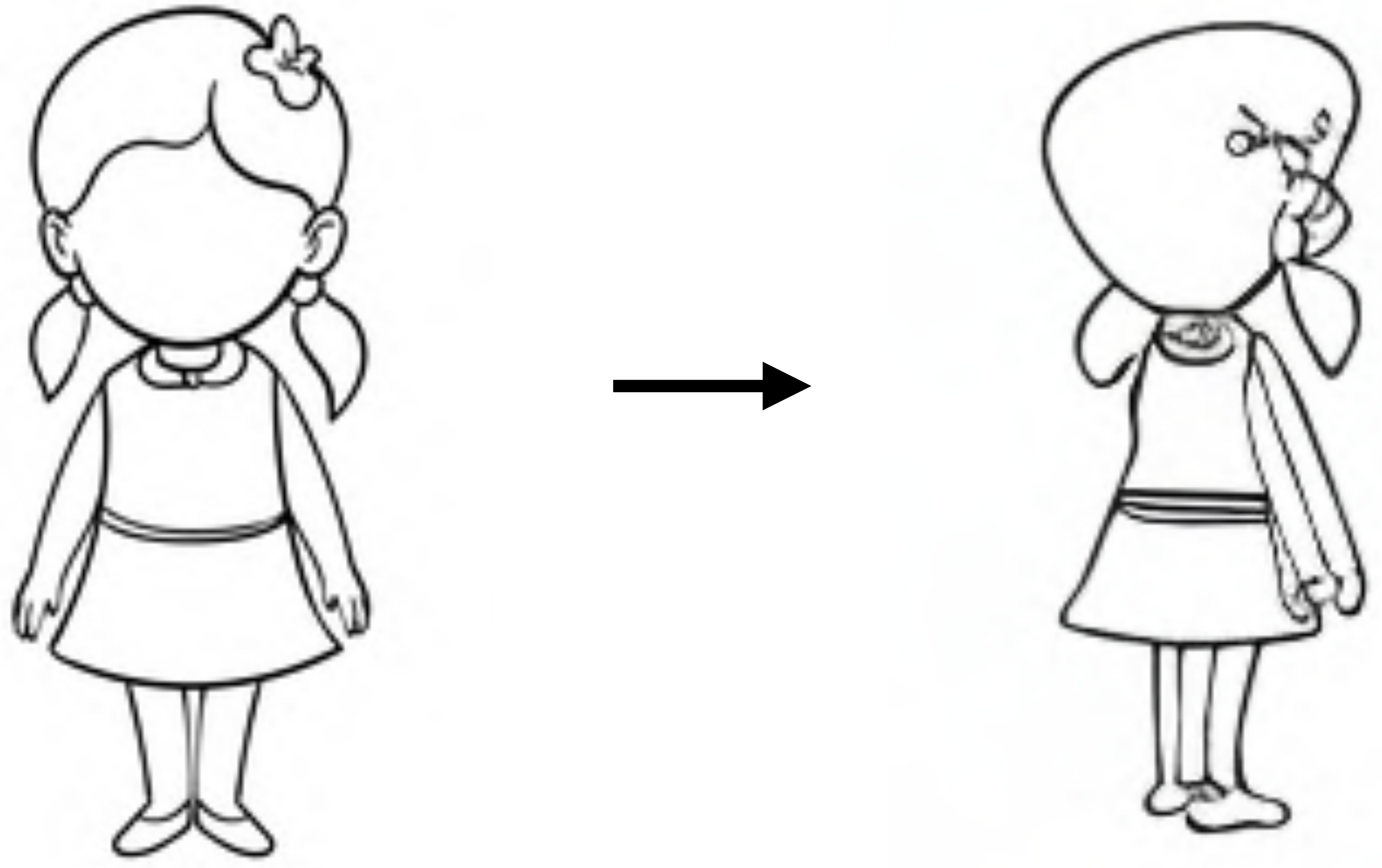
Deitke, **Liu**, et al. 2023.

Cartoons



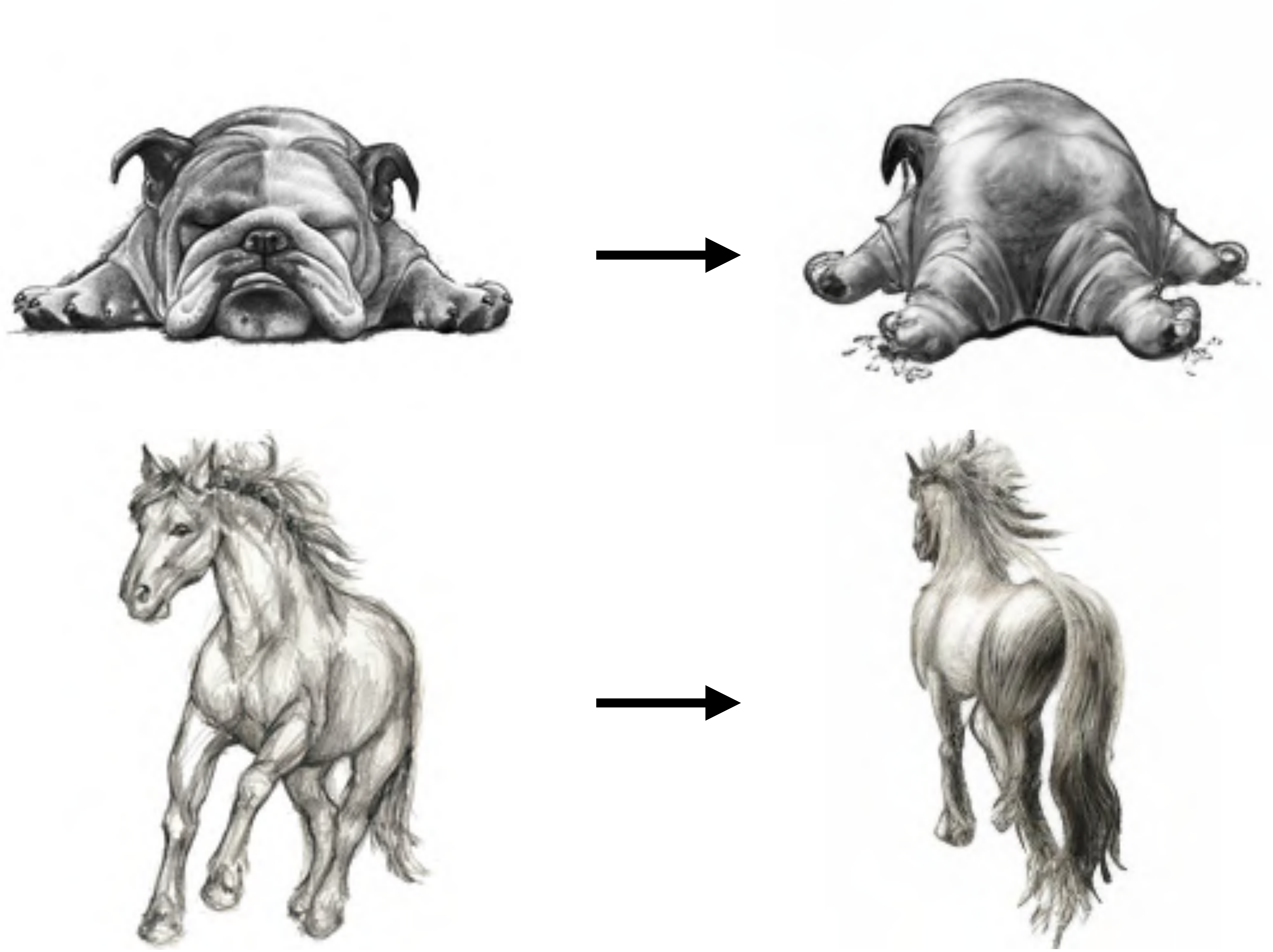
Deitke, **Liu**, et al. 2023.

Line Drawings



Deitke, **Liu**, et al. 2023.

Sketches



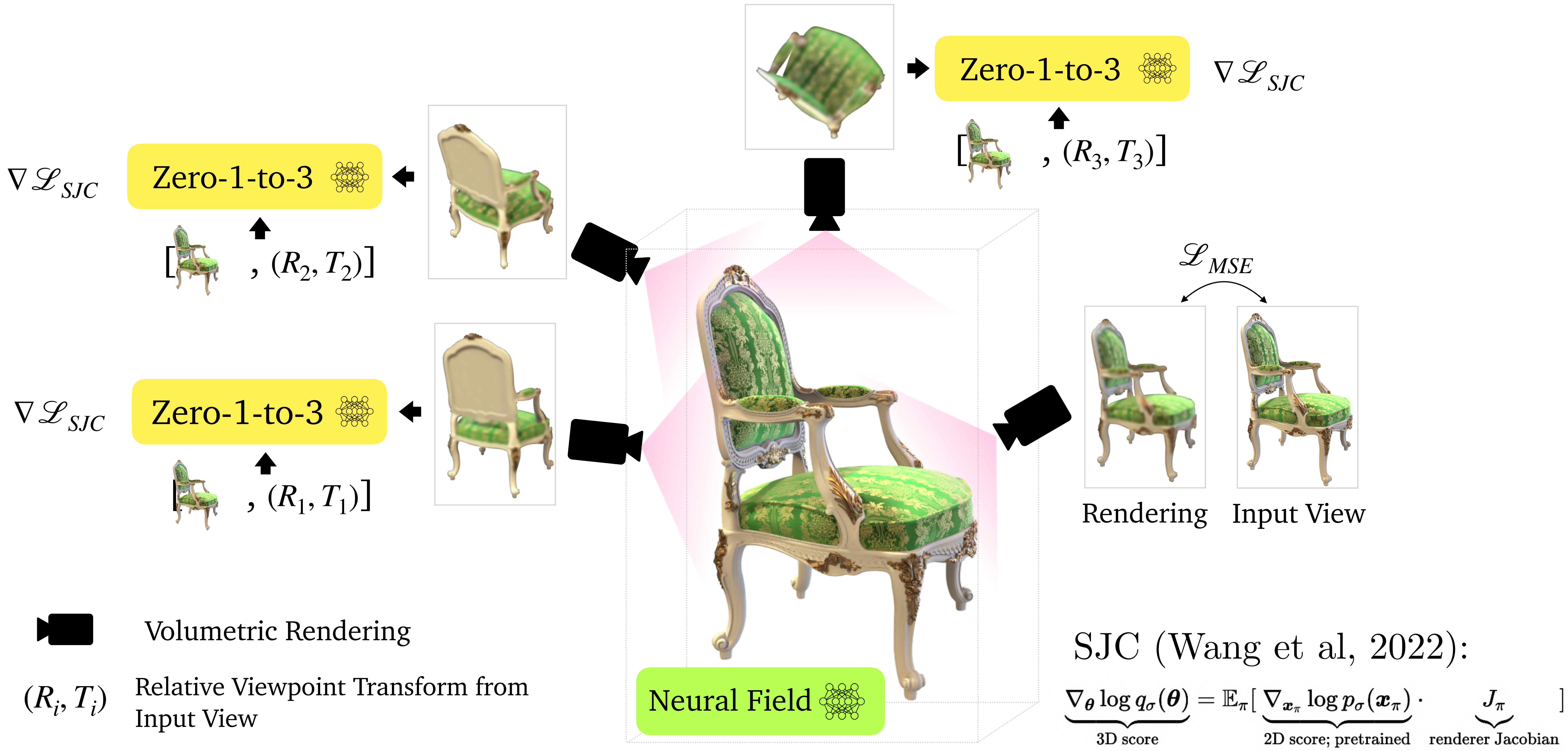
Deitke, **Liu**, et al. 2023.

Ambiguity from Occlusion

Input View

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

Method: 3D Reconstruction



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

Results: 3D Reconstruction

Input Image

Reconstruction

Voxel NeRF + SJC

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

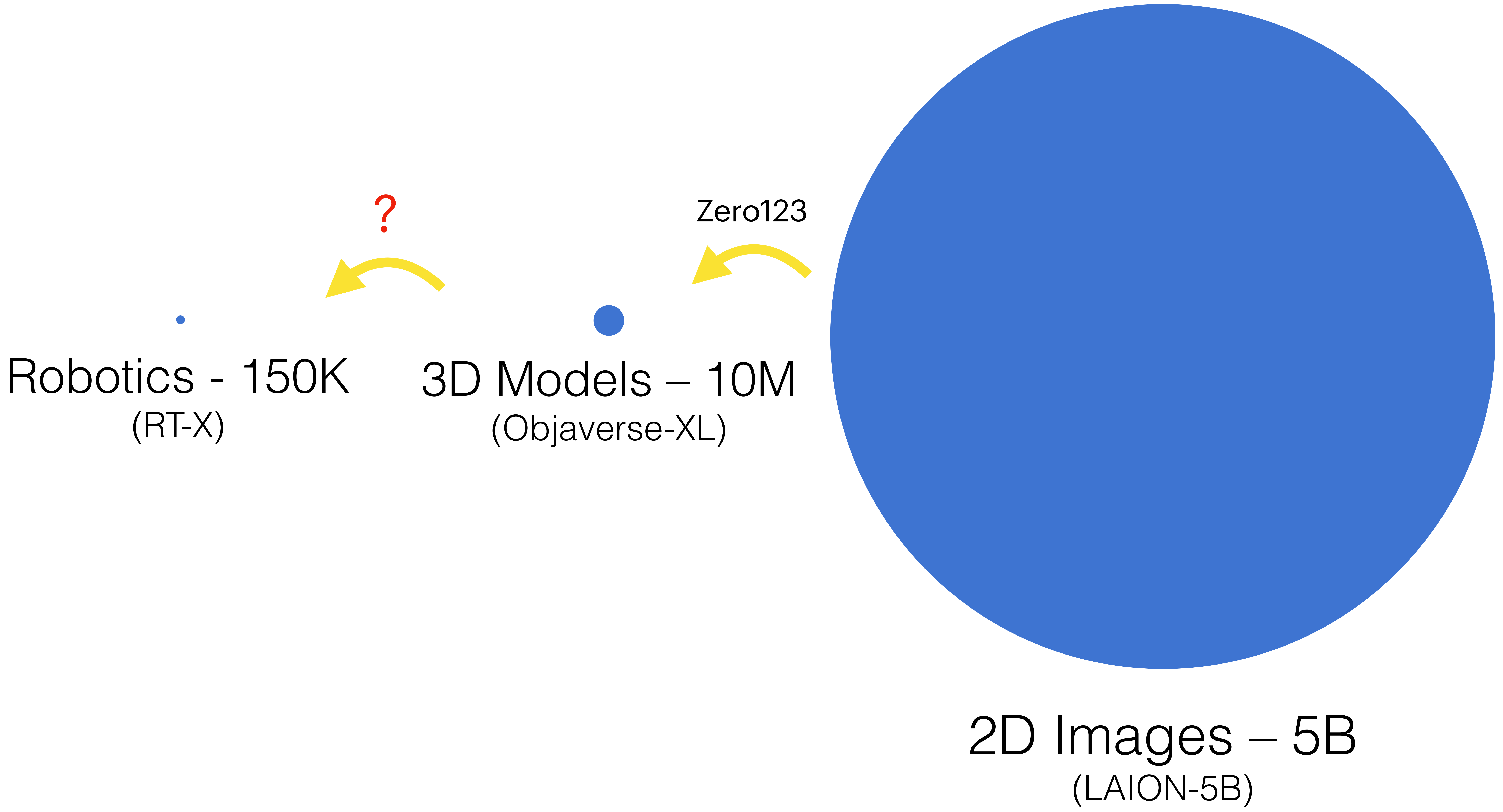
Results: 3D Reconstruction

Zero123 + Instant-NGP + SDS + DMTet

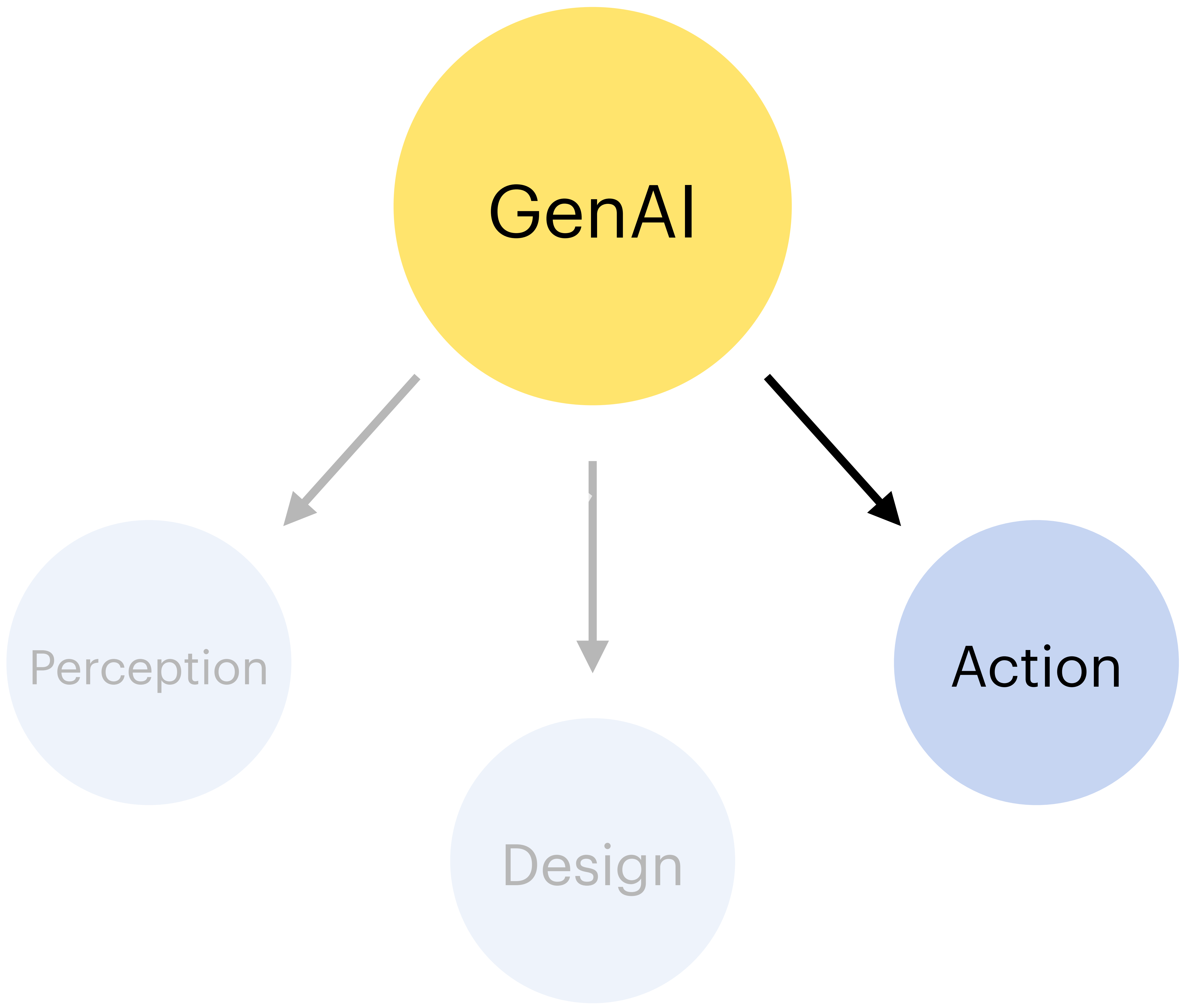
From Stability AI: https://github.com/threestudio-project/threestudio#zero-1-to-3-

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

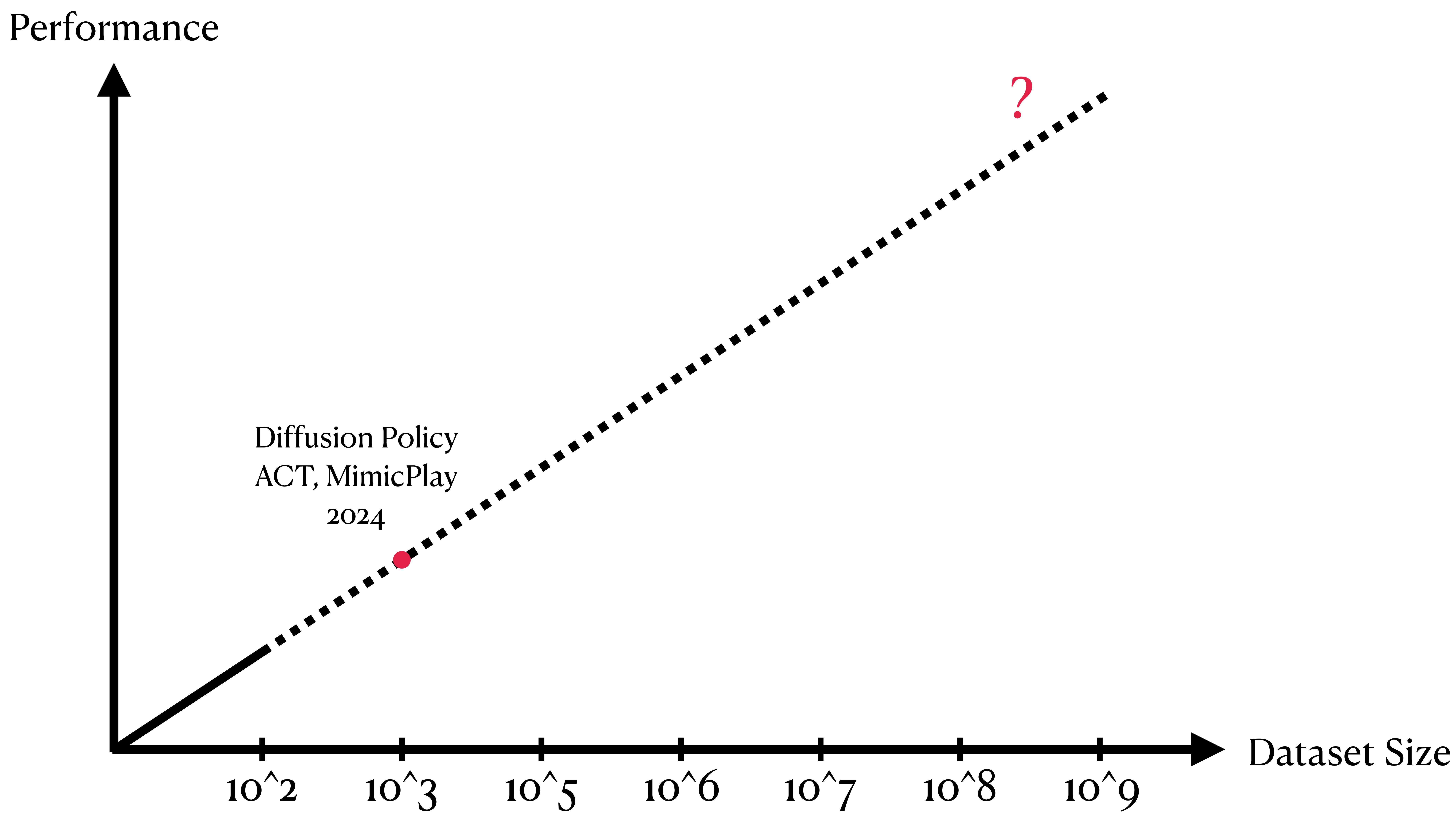
What about robotics?



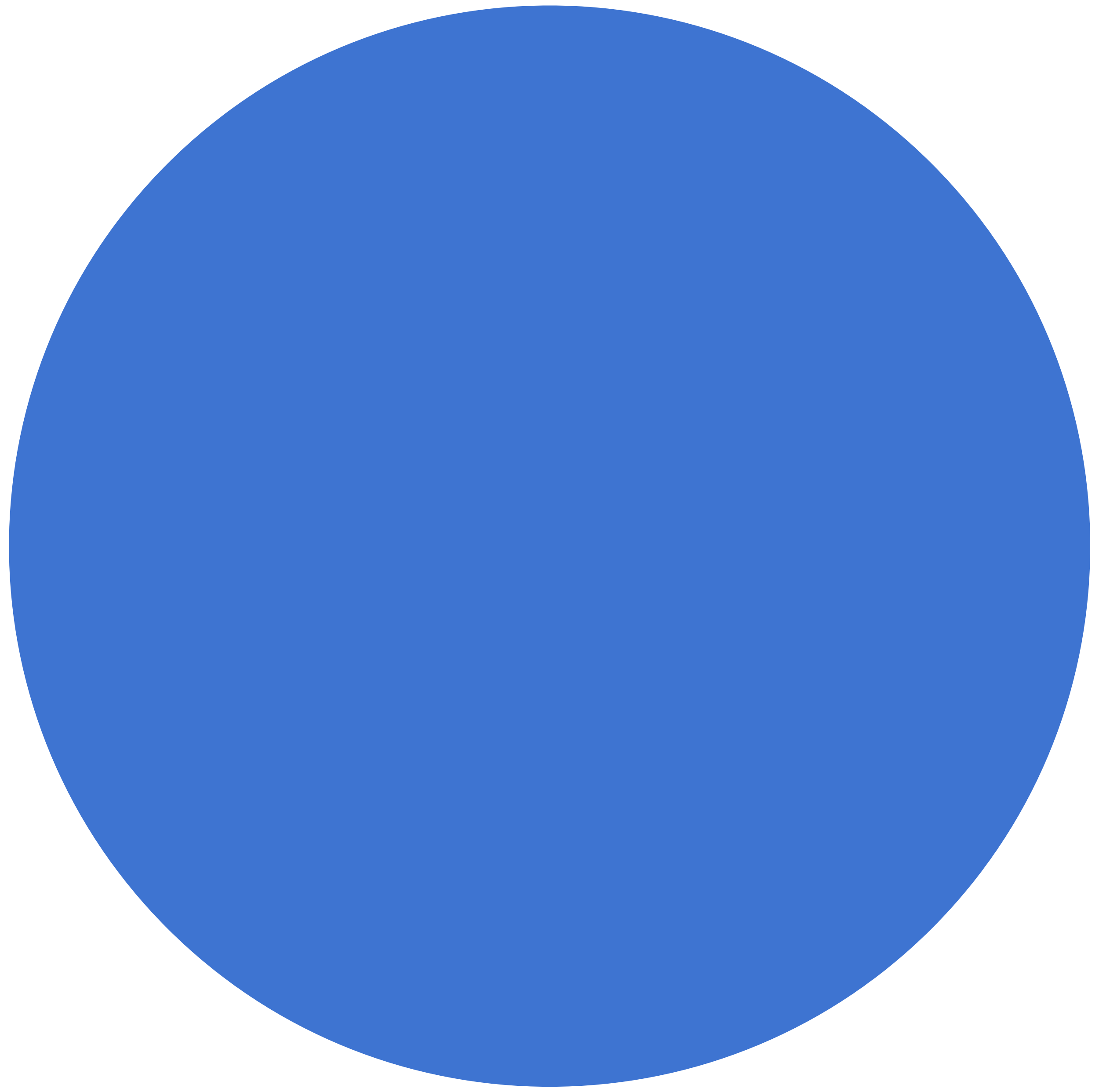
Generative Embodied AI



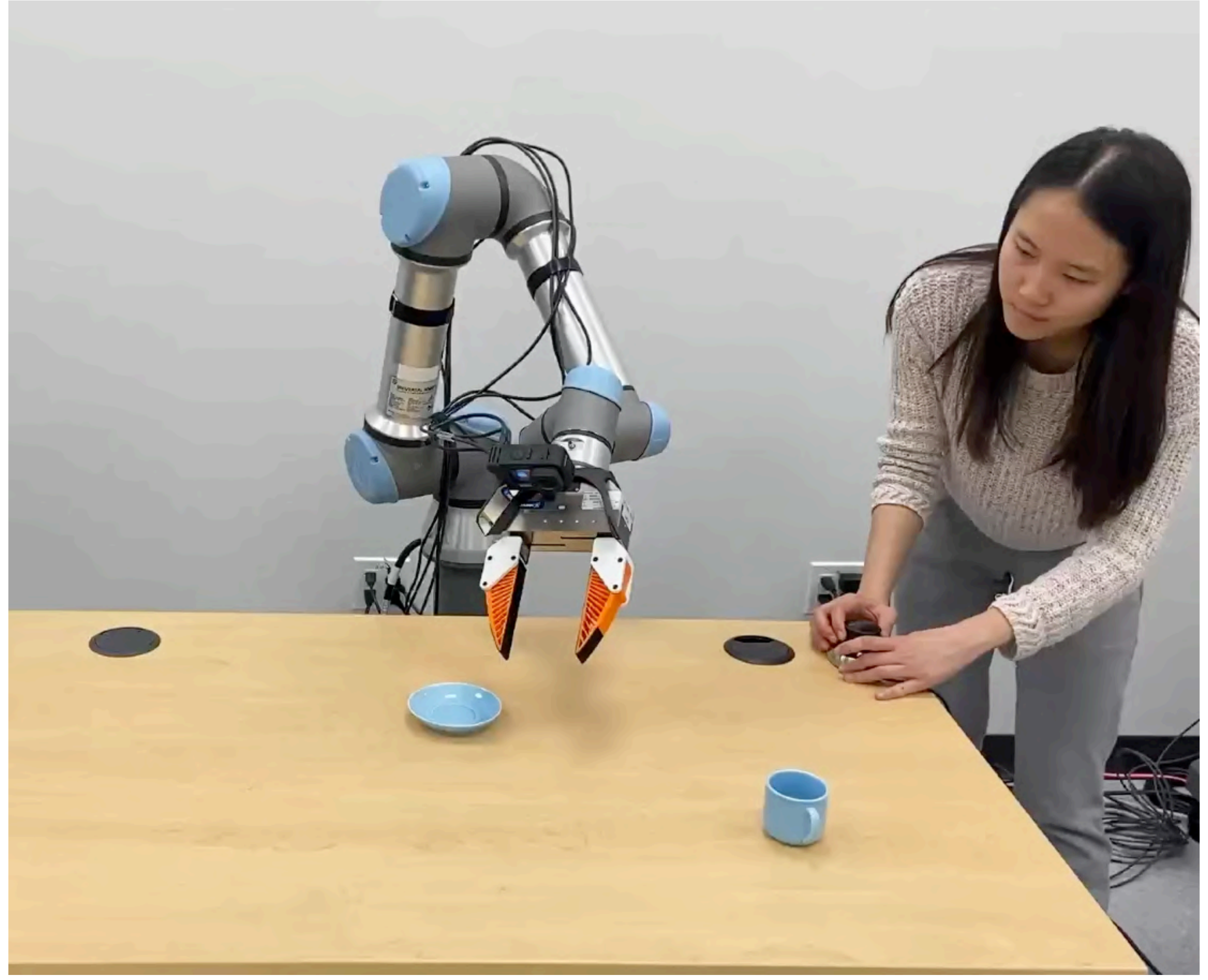
Behavior Cloning Puts Robot Learning on a Scaling Curve



Robot Learning Has a Data Problem



Robot Data



The Dilemma of Robot vs. Visual Data

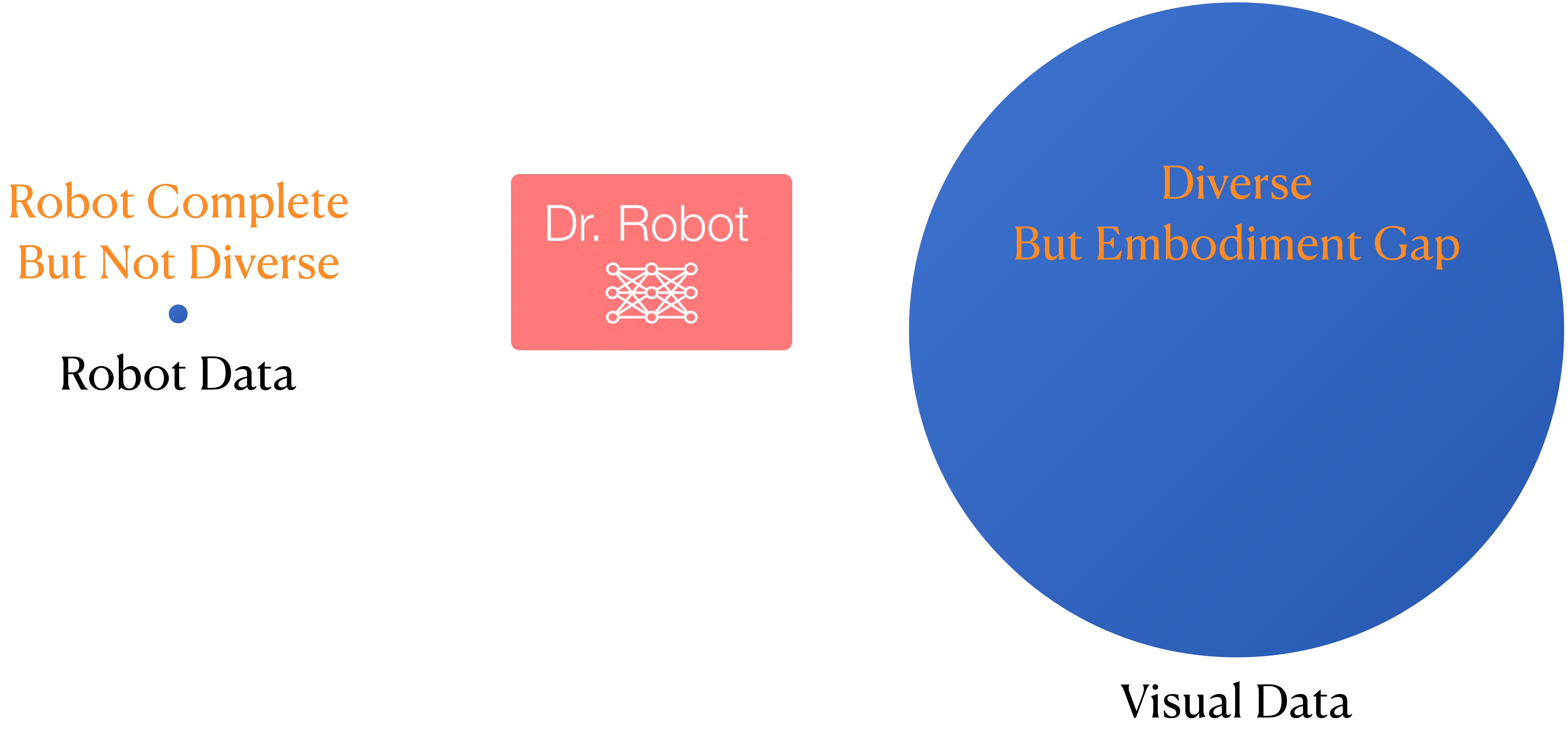
Limited
but Robot Complete

Robot Data

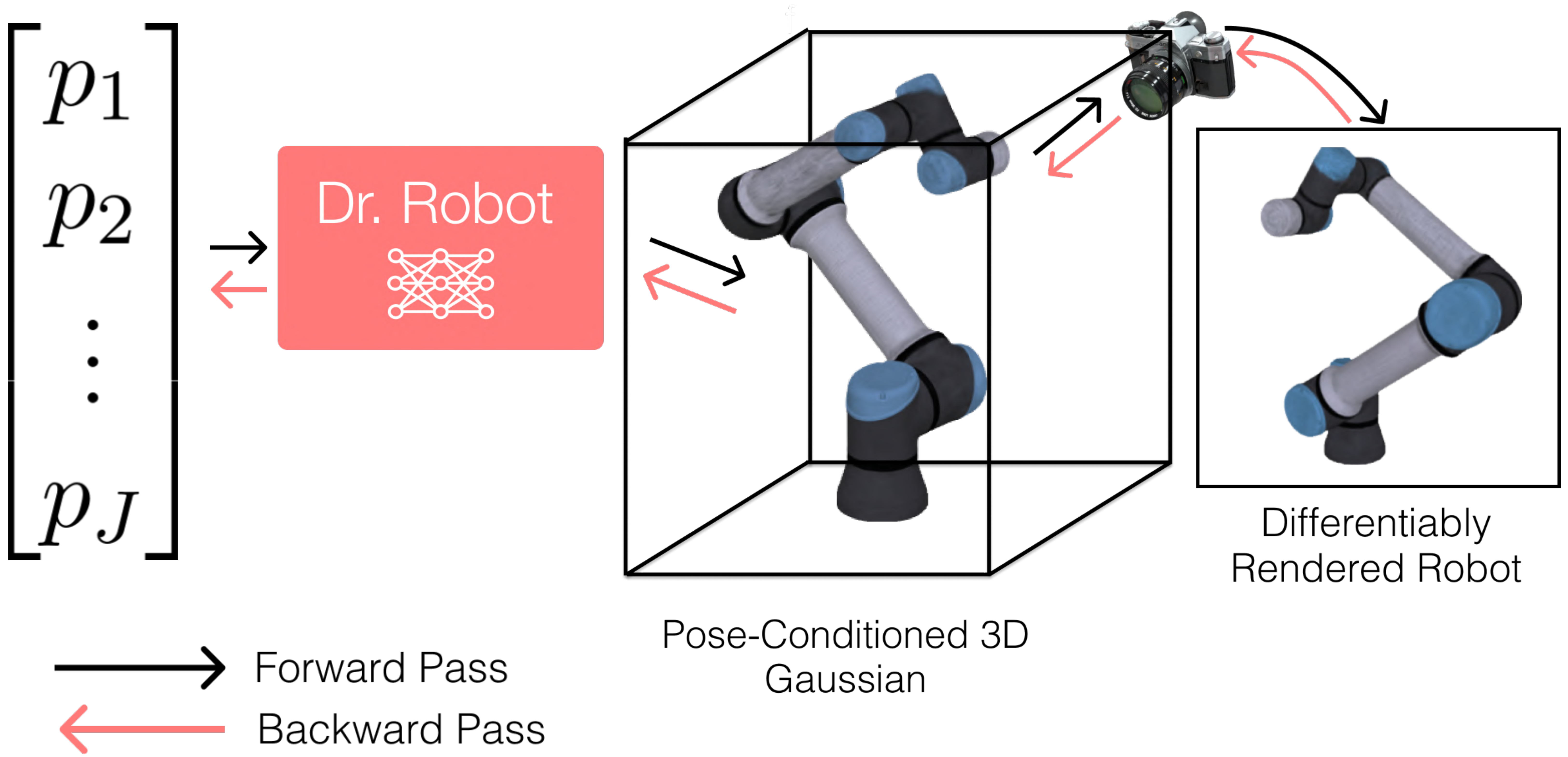
Diverse
but Embodiment Gap

Internet Image / Video

We need an interface between visual data and robot control!

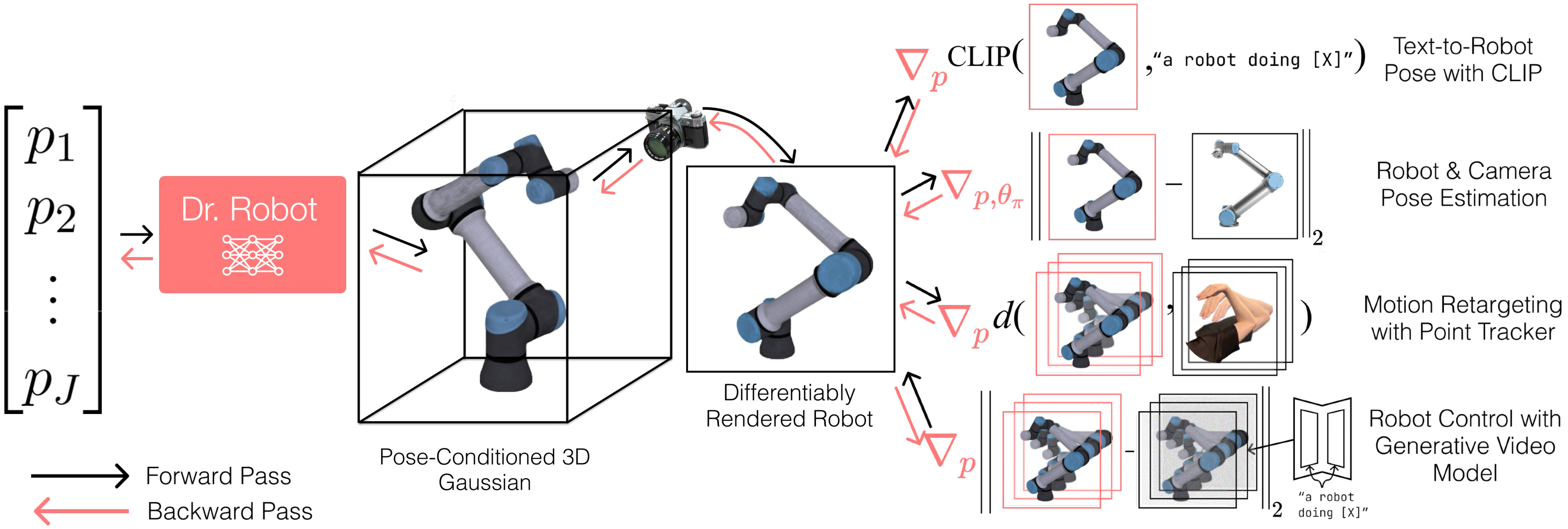


Differentiable Robot Rendering



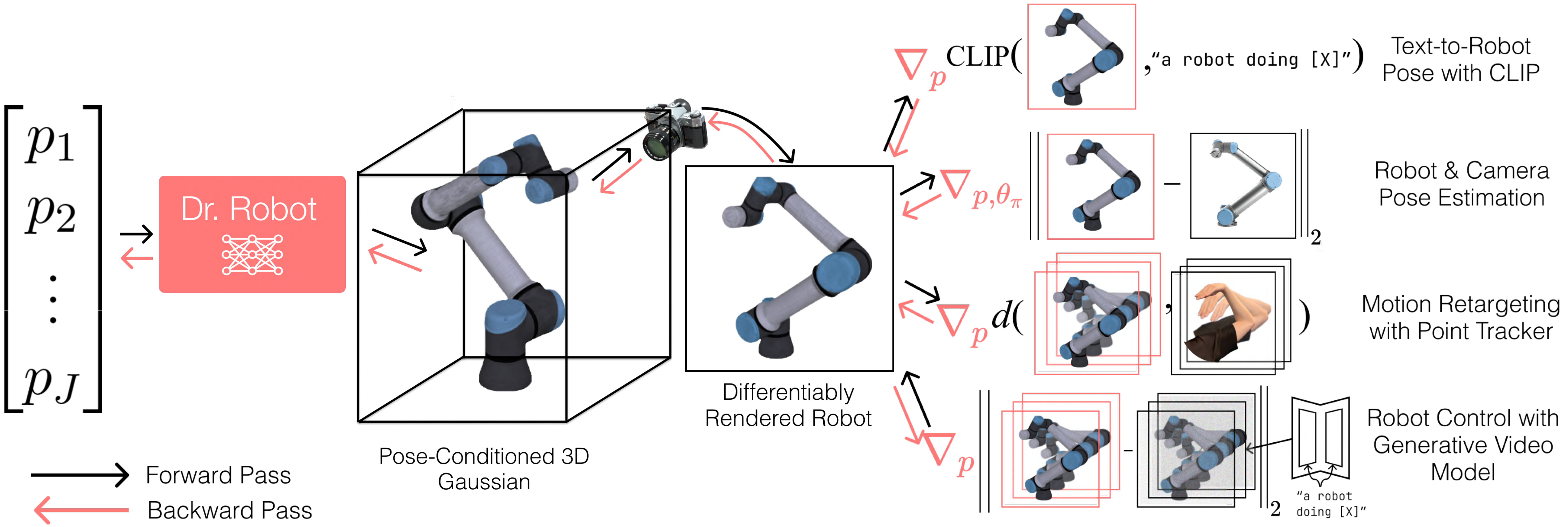
Liu, Canberk, Song, Vondrick. CoRL 2024

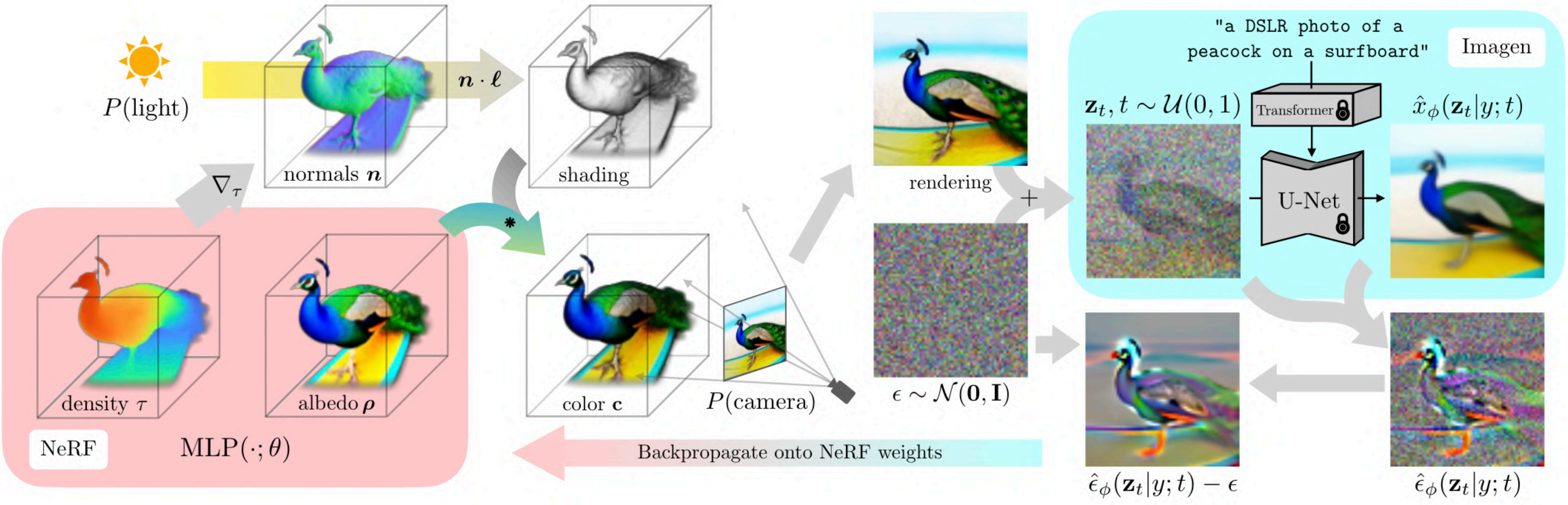
Dr. Robot Allows Visual Appearance of Robots to be Differentiable w.r.t. to Control Parameters



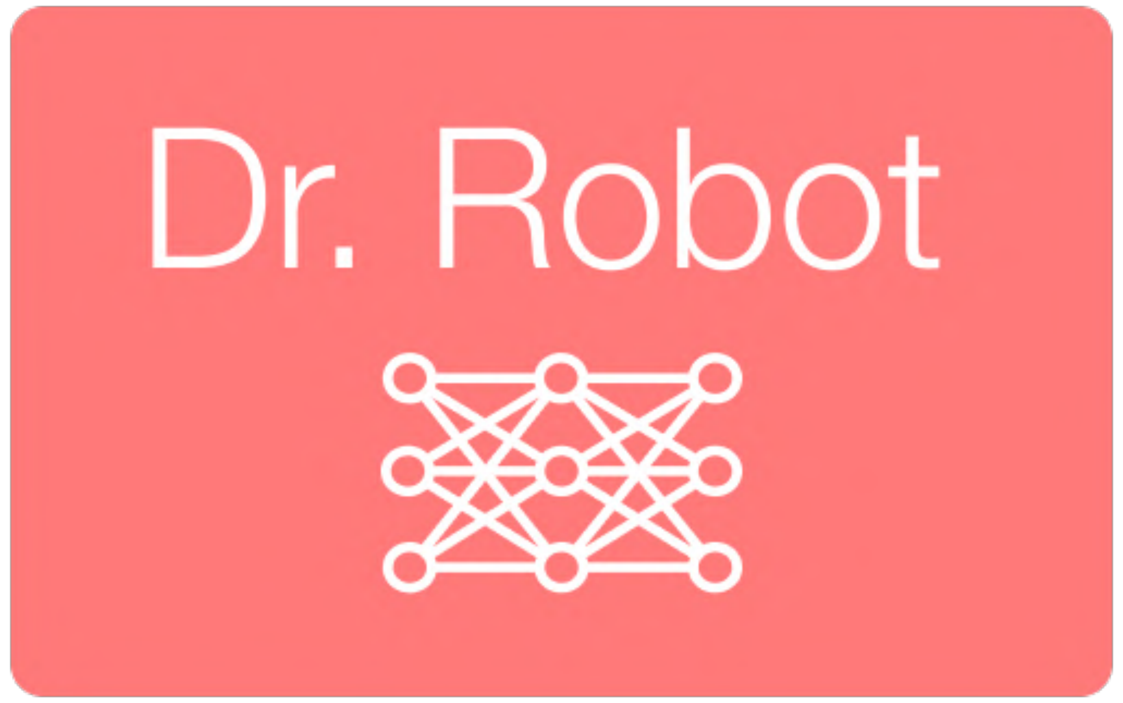
Liu, Canberk, Song, Vondrick. CoRL 2024

3D Gen, but for robot 🤖



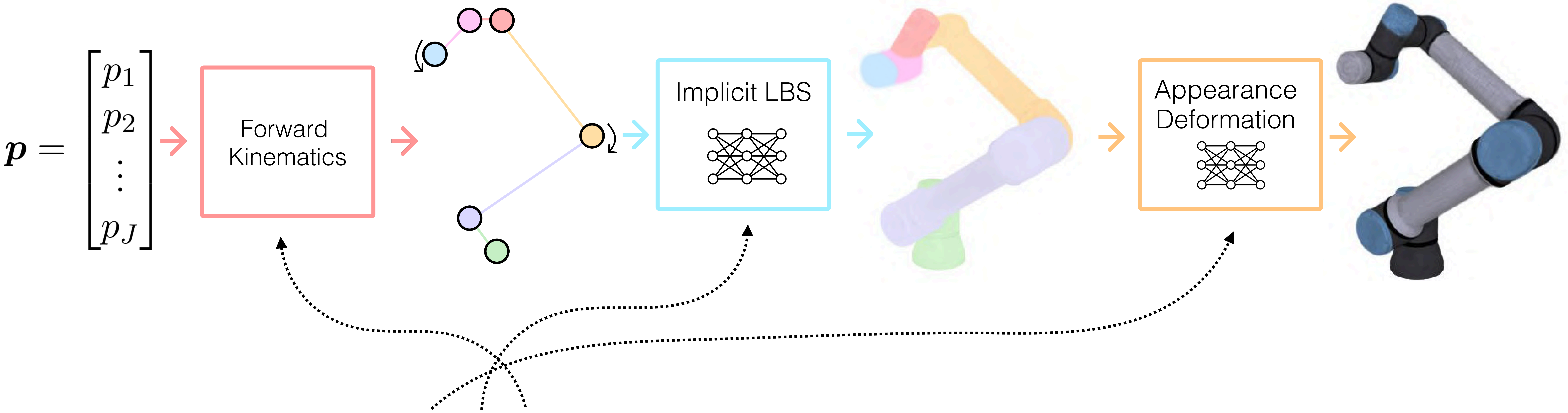


How did we achieve differentiability?



Liu, Canberk, Song, Vondrick. CoRL 2024

Differentiable Robot Rendering



$$\hat{I} = \pi(X(D(T(p))))$$

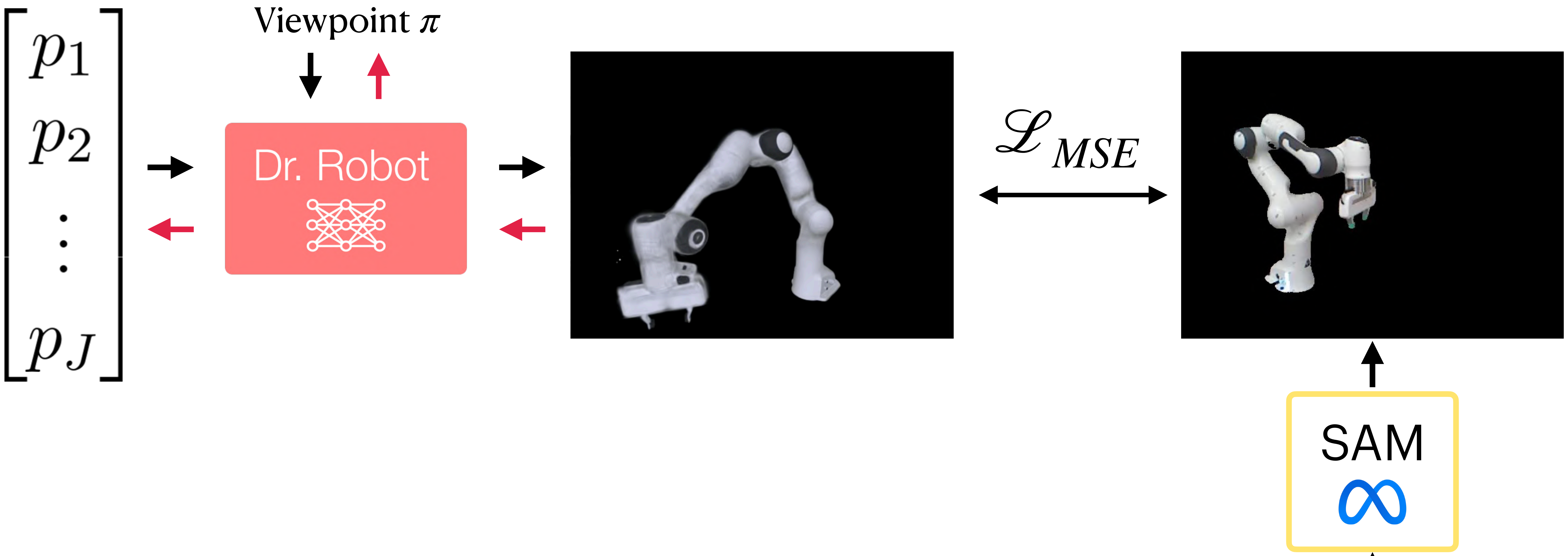
$$\min_{\theta_X, \theta_W, \theta_G} L = ||I - \hat{I}||_2$$

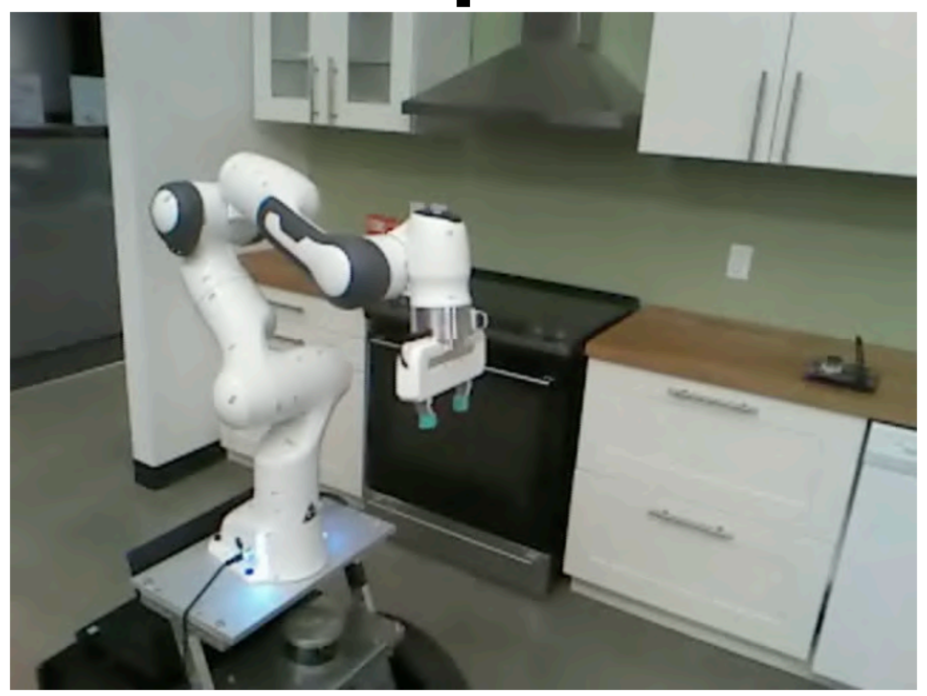
Liu, Canberk, Song, Vondrick. CoRL 2024

Built from any URDF

Liu, Canberk, Song, Vondrick. CoRL 2024

Robot Poses Reconstruction from Single Image Through Analysis-by-Synthesis



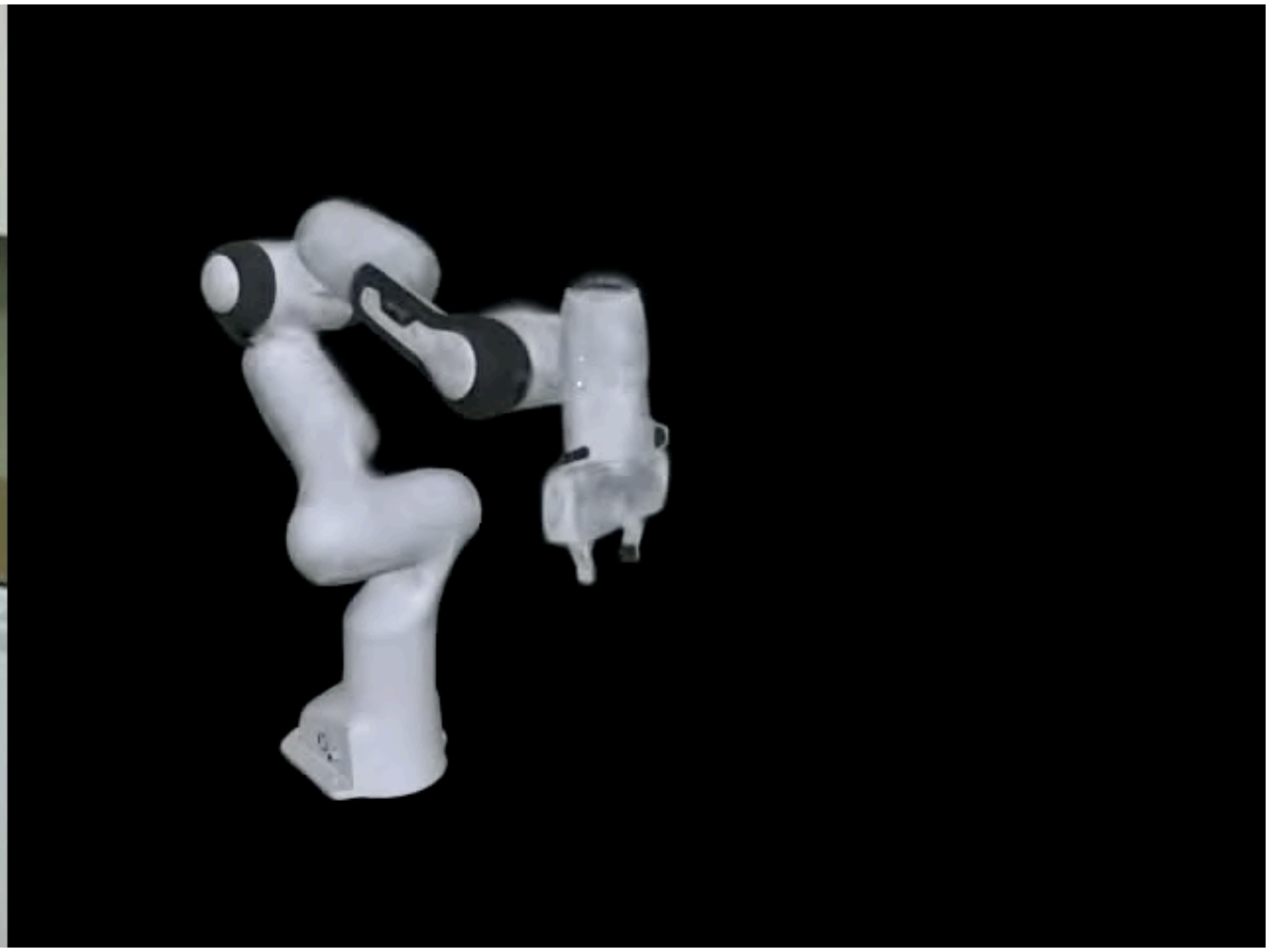


Optimization Objective: $\min_{p,\theta_\pi} \mathscr{L}_{MSE}(p, I) = ||I - \pi(f(p))||_2$

Liu, Canberk, Song, Vondrick. CoRL 2024

Robot Poses Reconstruction from Single Image Through Analysis-by-Synthesis

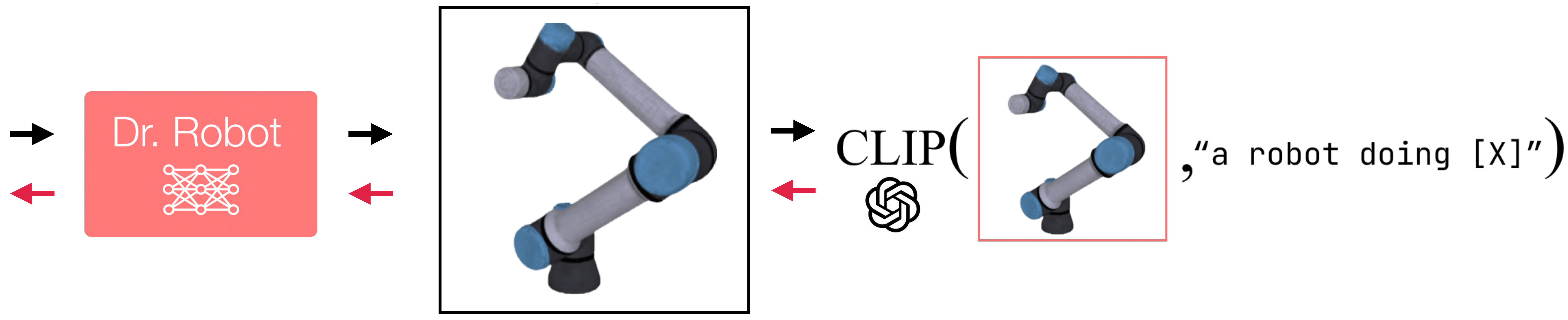
Original Video



Reconstructed Robot

Liu, Canberk, Song, Vondrick. CoRL 2024

Visual MPC with CLIP

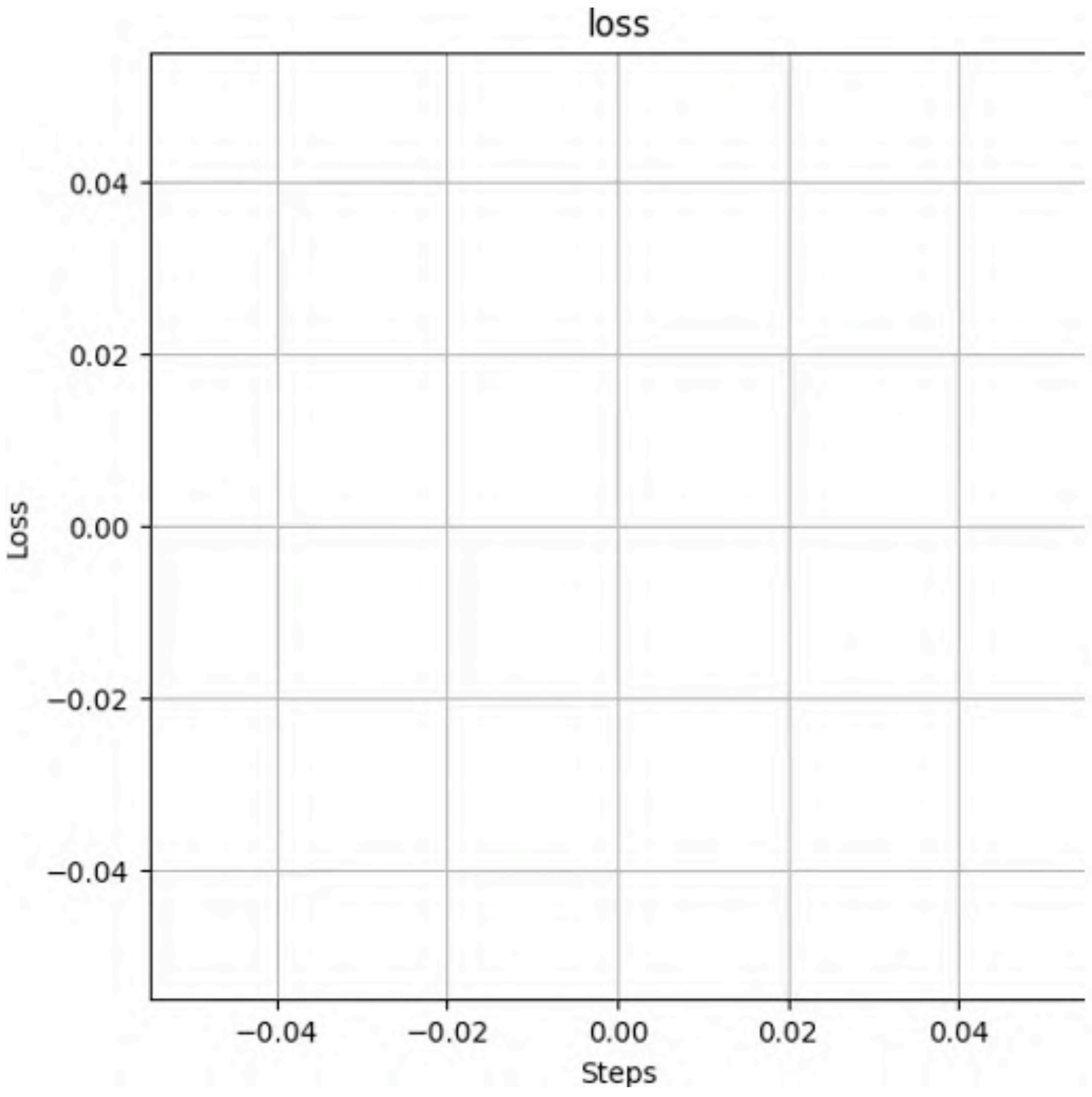


$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_J \end{bmatrix}$

Optimization Objective: $\min_p \mathscr{L}_{\text{CLIP}}(p; \text{prompt}) = \text{CLIP}(\text{prompt}, \pi(f(p)))$

Liu, Canberk, Song, Vondrick. CoRL 2024

Visual MPC with CLIP

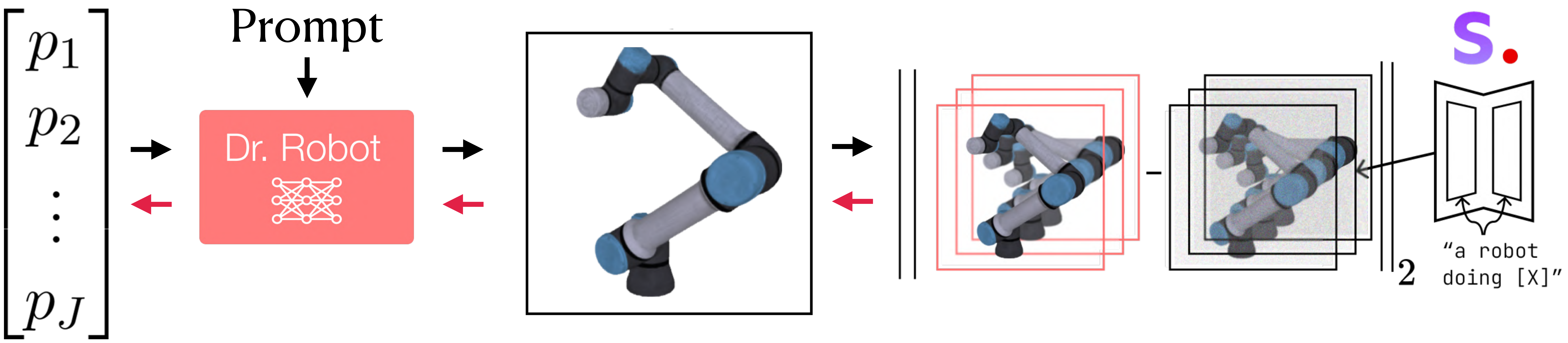


$\mathscr{L}_{\mathrm{CLIP}}(p; \mathrm{prompt})$

"Victory Sign"

Liu, Canberk, Song, Vondrick. CoRL 2024

Robot Control with Text2Video Model



Optimization Objective: $\min_{p,\theta_\pi} \sum_{i}^{T} \mathscr{L}_{MSE}(p_i, I_i) = ||I_i - \pi(f(p_i))||_2$

Liu, Canberk, Song, Vondrick. CoRL 2024

Robot Control with Text2Video Model

Current Observation

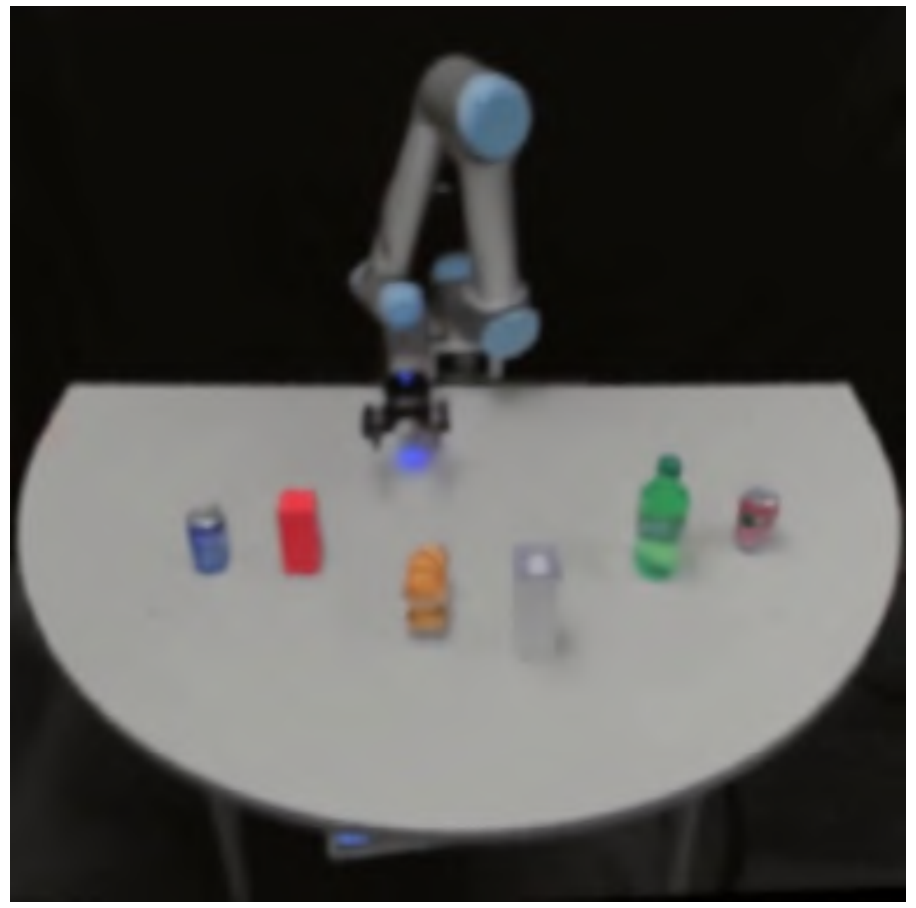
“Pick up the blue can.”

Text-to-Video

Dr. Robot

Generated Video

Robot Control



Current Observation

“Pick up the red block.”

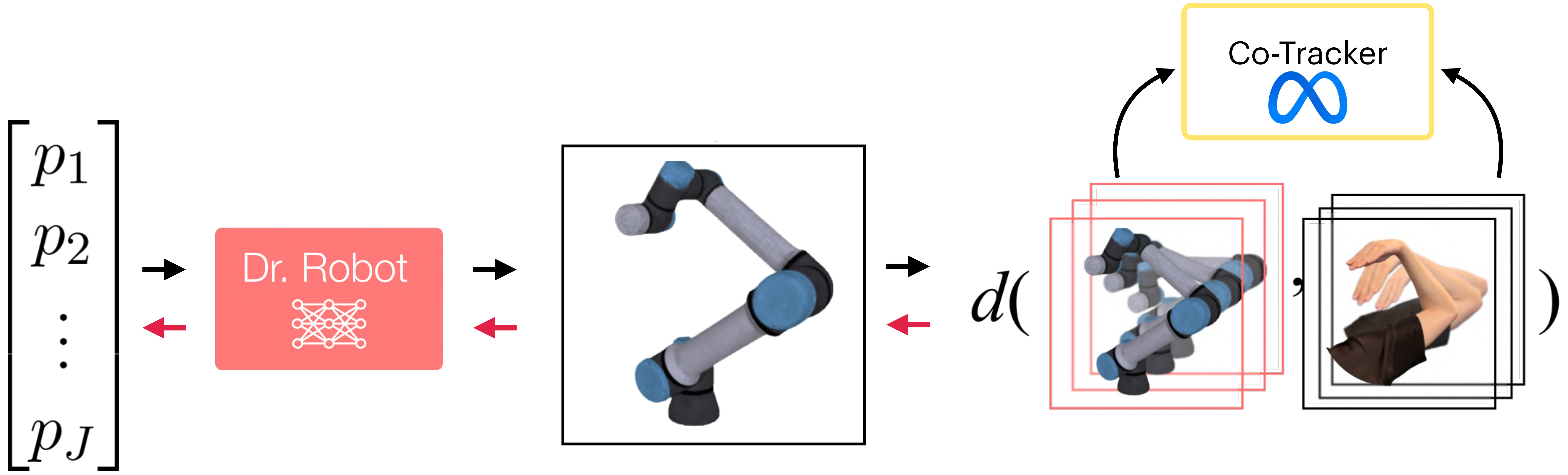
Text-to-Video

Dr. Robot

Generated Video

Robot Control

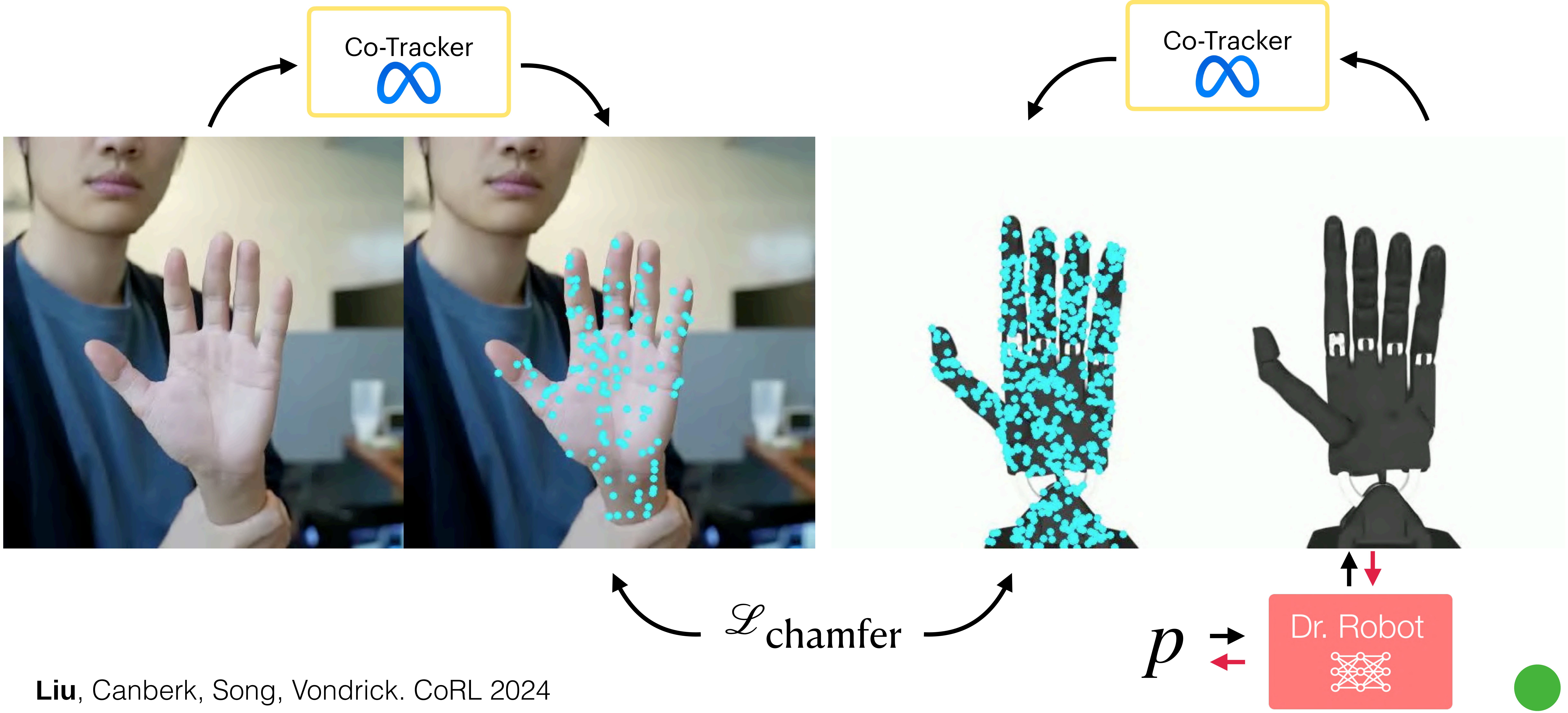
Motion Retargetting with Point Tracker



Optimization Objective: $\mathscr{L}_{\text{Track}}(p_{1:T}; c_0^r, c_{1:T}) = \sum_{1=t}^{T} d(c_t, c_t^r) = \sum_{1=t}^{T} d(c_t, \mathscr{P}(c_0^r, \pi(p_{1:T}))_t))$

Liu, Canberk, Song, Vondrick. CoRL 2024

Motion Retargetting with Point Tracker



Liu, Canberk, Song, Vondrick. CoRL 2024

What about policy learning?

Visuomotor Policy

Observation

Visuomotor Policy

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_J \end{bmatrix}$$

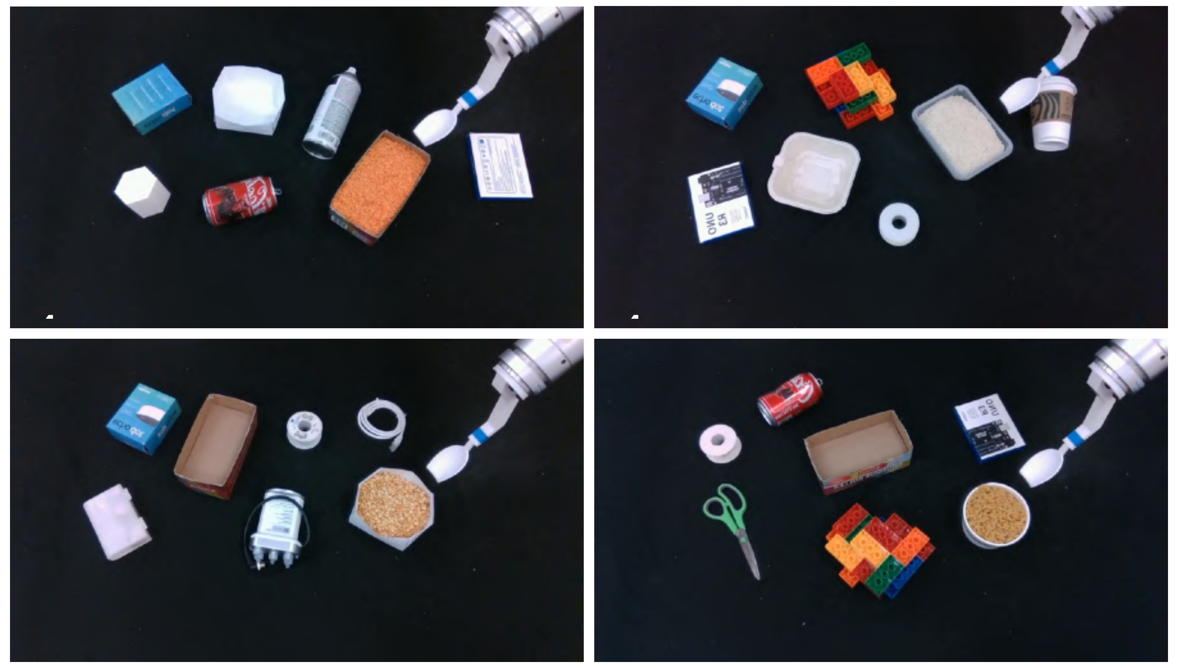
Motor Actions

Behavior Cloning is Supervised Learning of Human Behavior

Lots of Observation

Visuomotor Policy

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_J \end{bmatrix} \cdots \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_J \end{bmatrix}$$



Demonstrations

Video Generation as Intermediate Action Representation

Liang*, **Liu***, et al. CoRL 2024

Dreamitate

Visuomotor Policy = **Dream** + **Imitate**

Input Image

Generated Video (indicated by)

Robot Execution

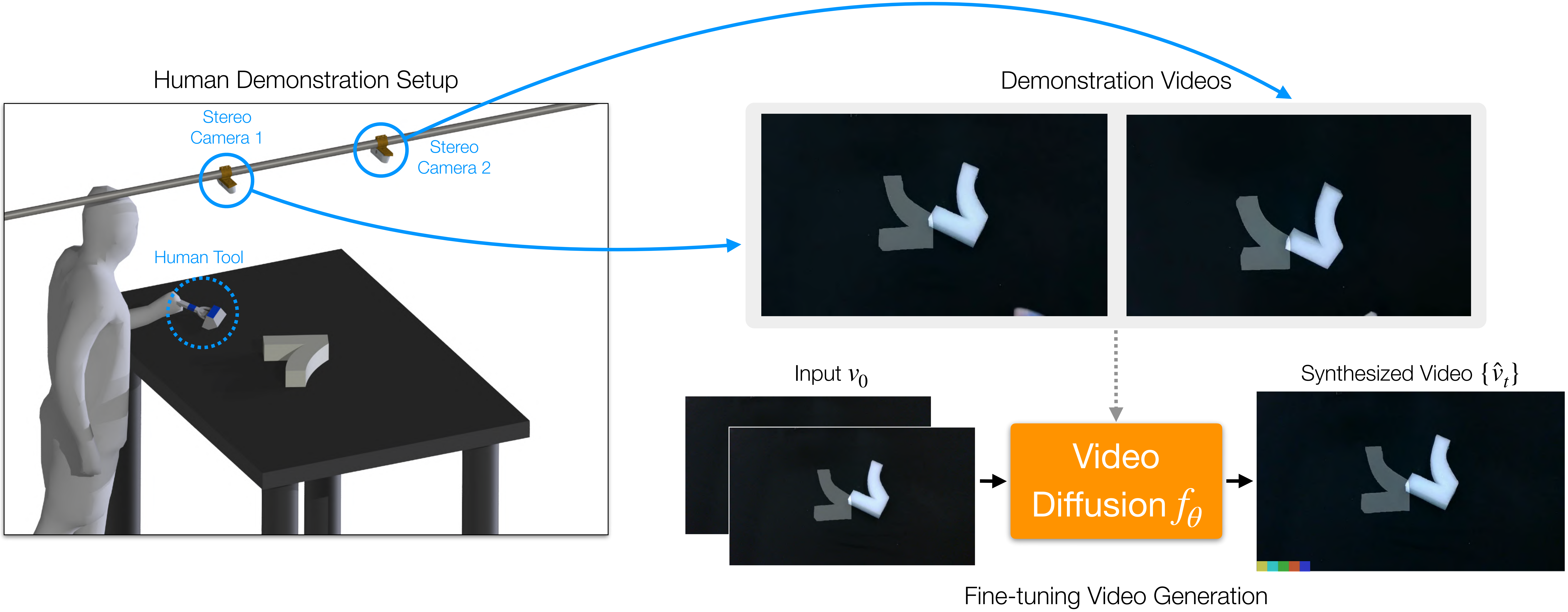
Liang*, **Liu***, et al. CoRL 2024

Aligning Video Model to Robot

OpenAI SORA

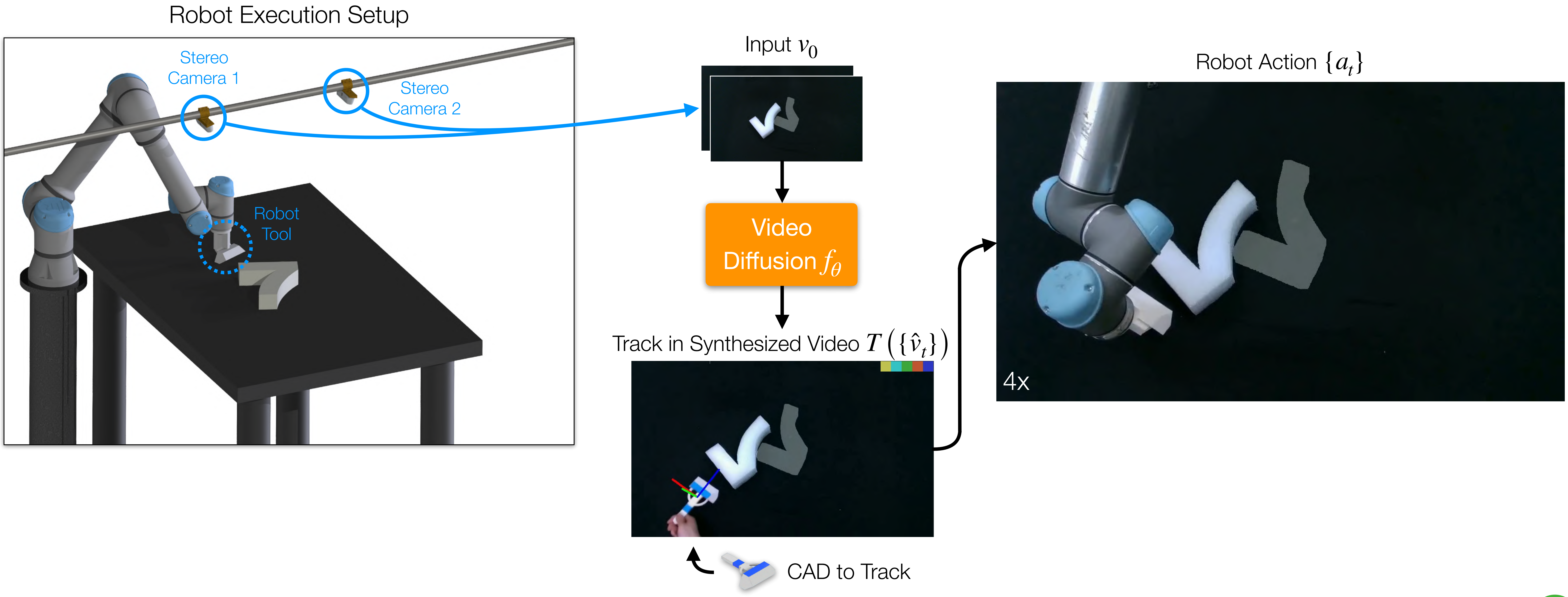
Luma Dream Machine

Data Collection



Liang*, **Liu***, et al. CoRL 2024

Robot Execution



Liang*, **Liu***, et al. CoRL 2024

Evaluation Tasks

Demonstration Videos

Liang*, **Liu***, et al. CoRL 2024

Rotation Task

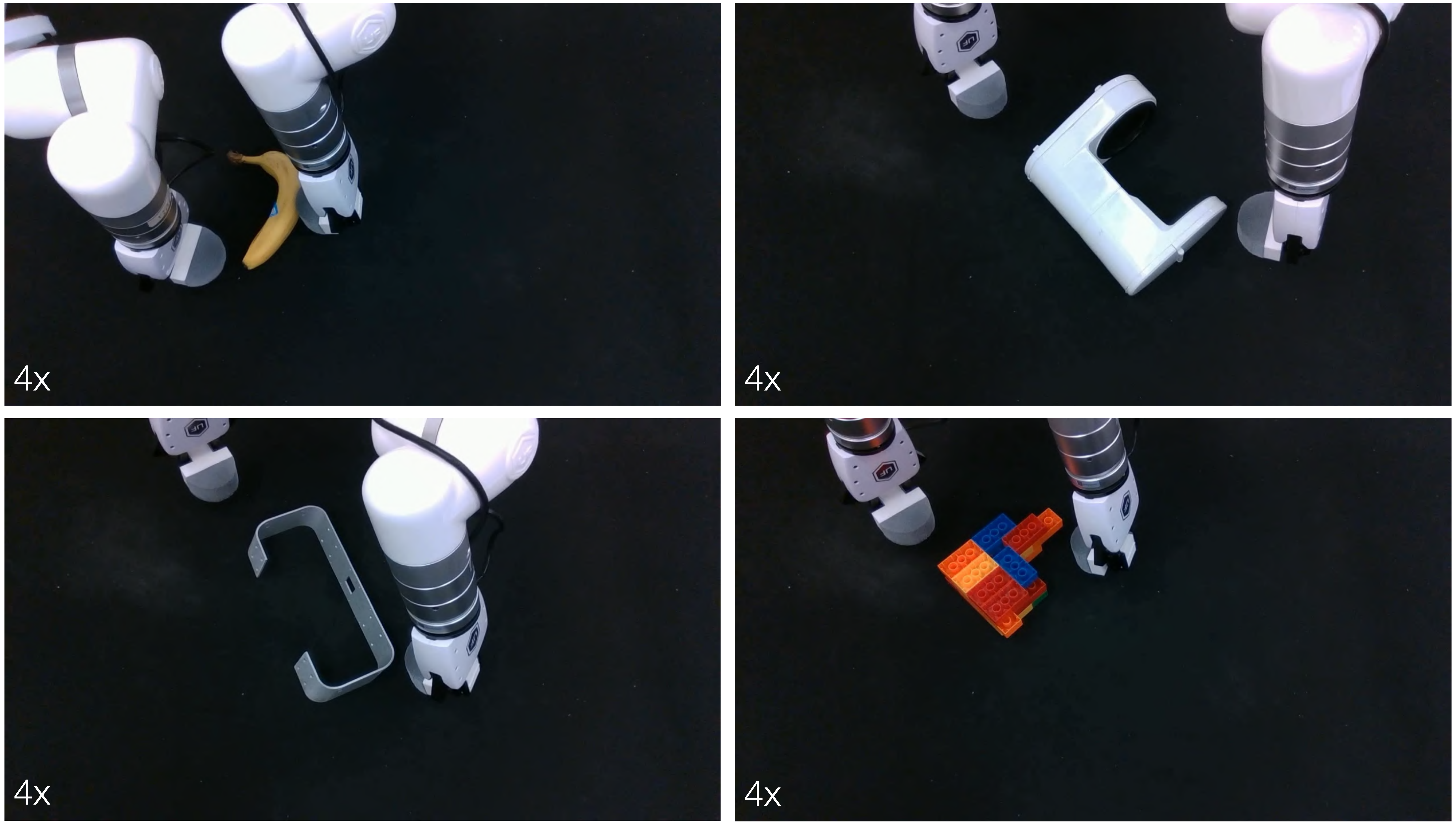
- Diffusion Policy: 22/40
- Ours: **37/40**

Test Generations

Liang*, **Liu***, et al. CoRL 2024

Rotation Task

Execution



Liang*, **Liu***, et al. CoRL 2024

Scooping Task

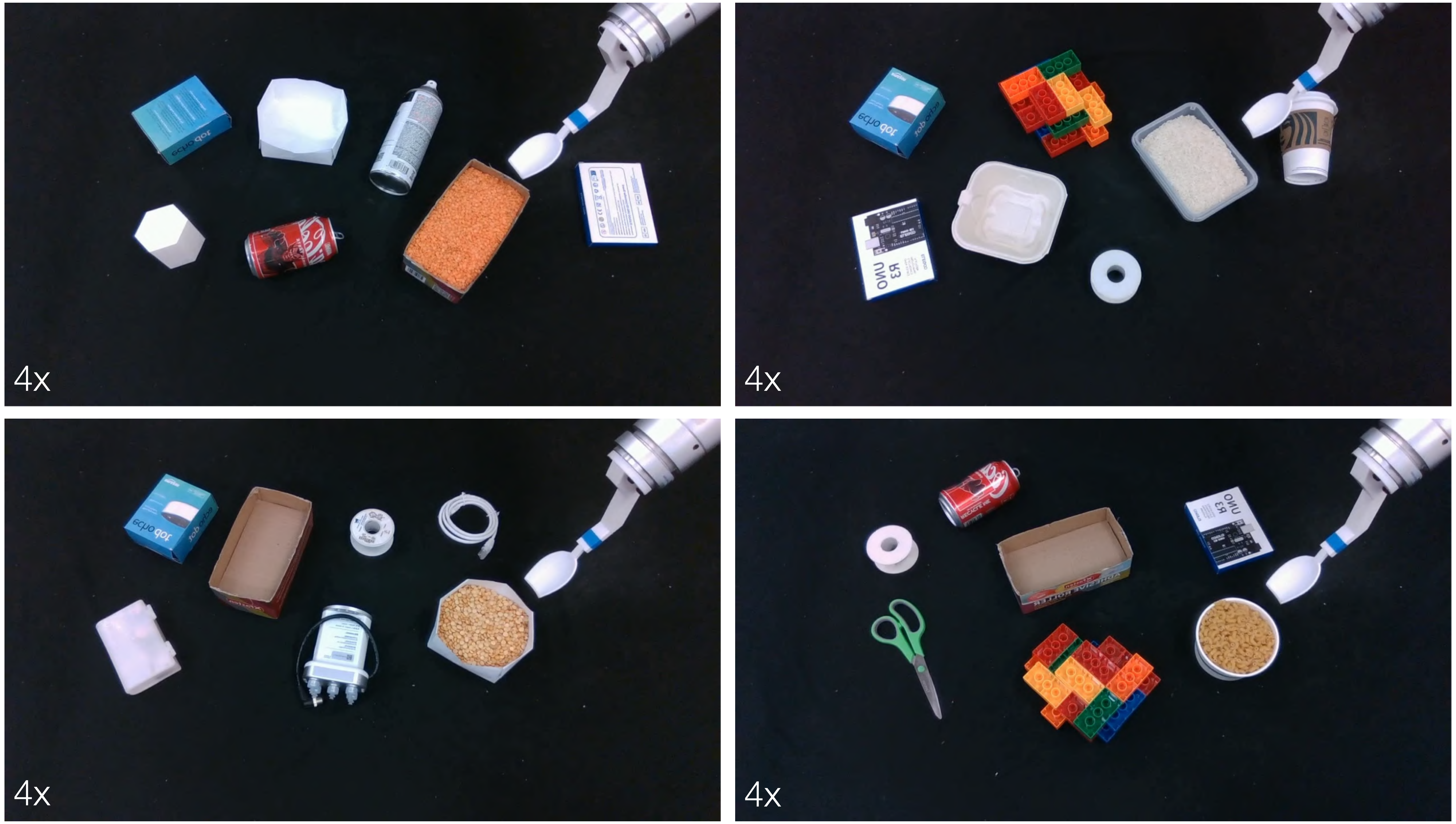
- Diffusion Policy: 22/40
- Ours: **36/40**

Test Generations

Liang*, **Liu***, et al. CoRL 2024

Scooping Task

Robot Execution

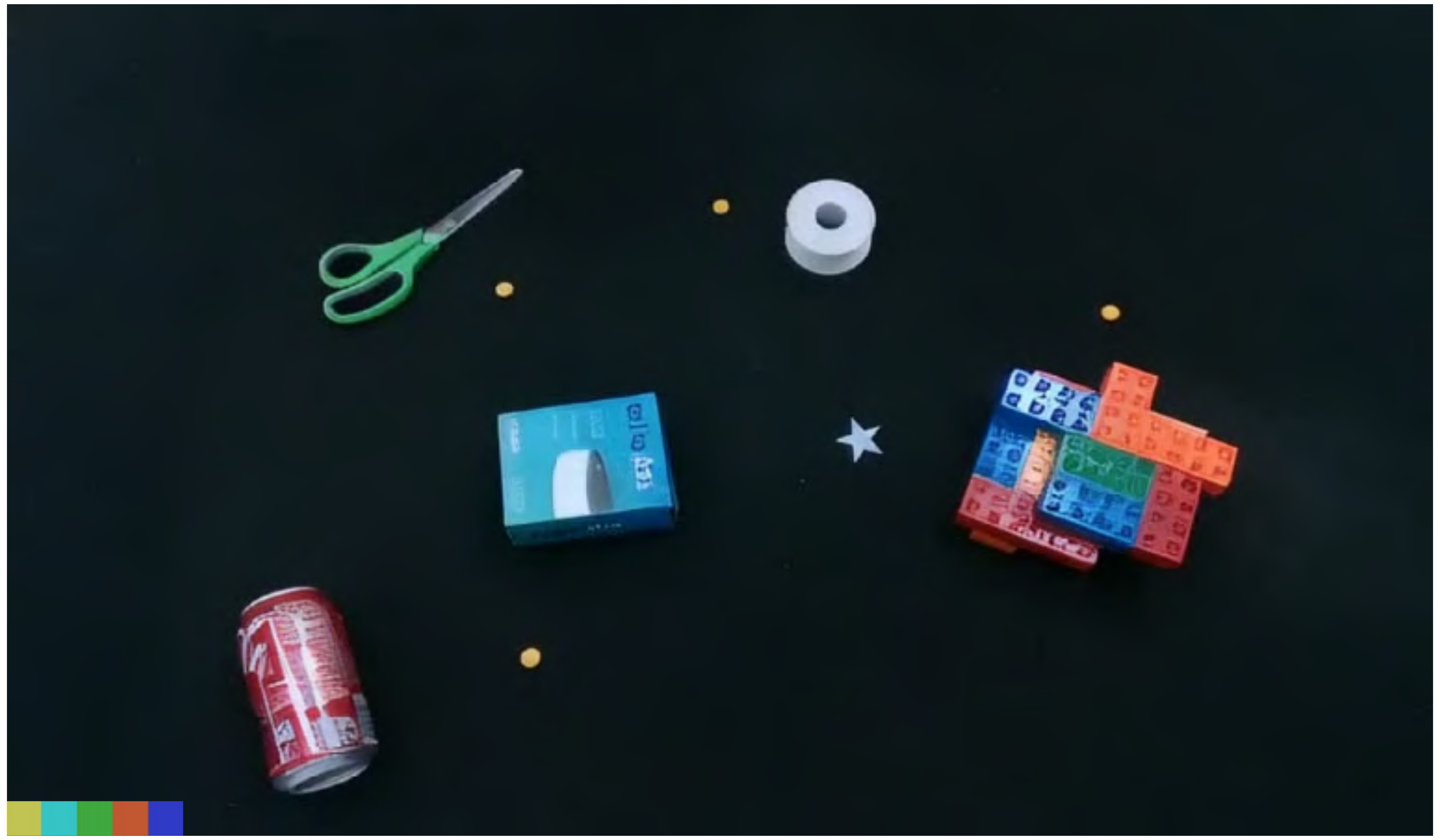


Liang*, **Liu***, et al. CoRL 2024

Sweeping Task

- Diffusion Policy: 5/40
- Ours: **37/40**

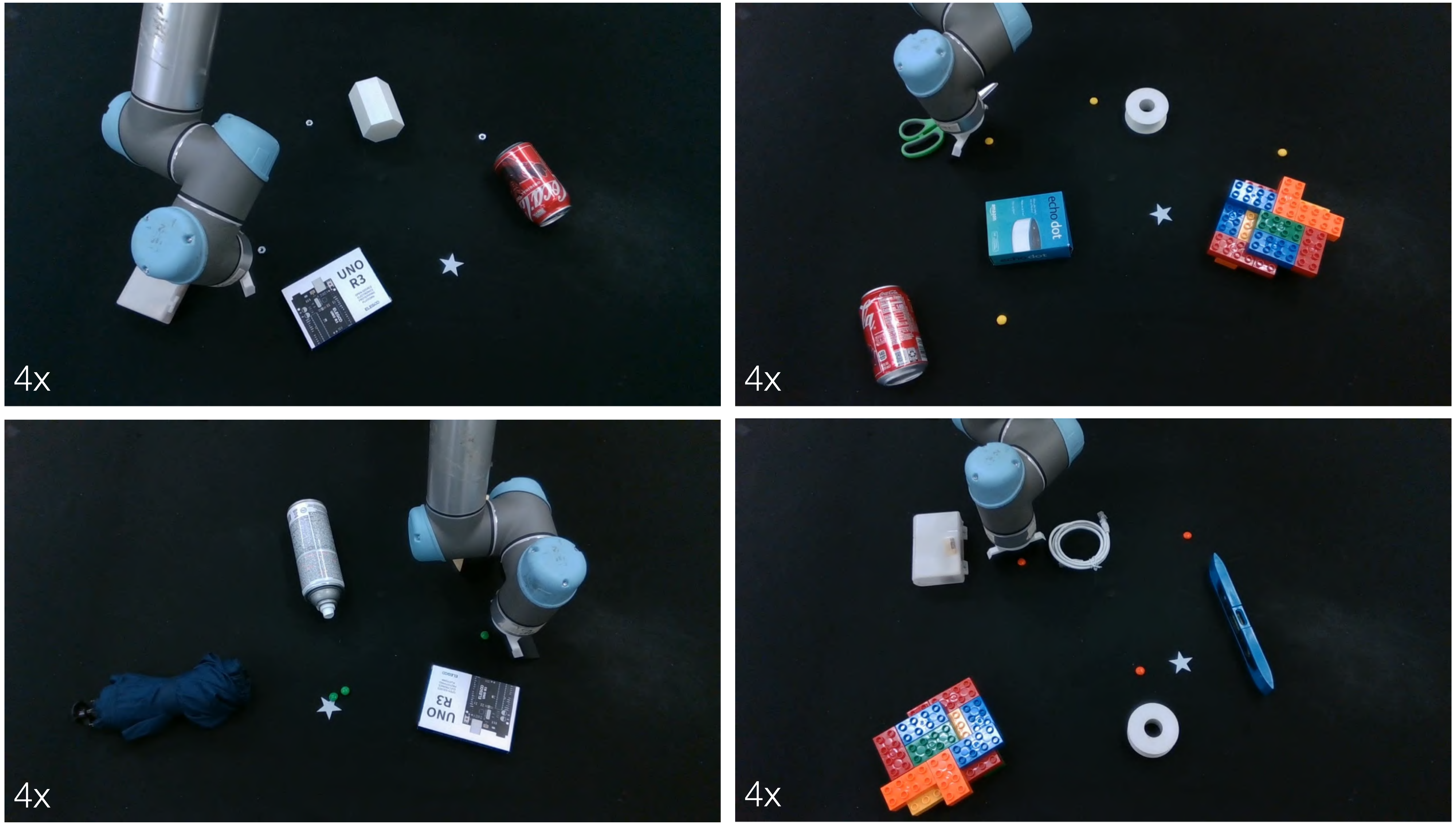
Test Generations



Liang*, **Liu***, et al. CoRL 2024

Sweeping Task

Robot Execution



Liang*, **Liu***, et al. CoRL 2024

Push-Shape Task

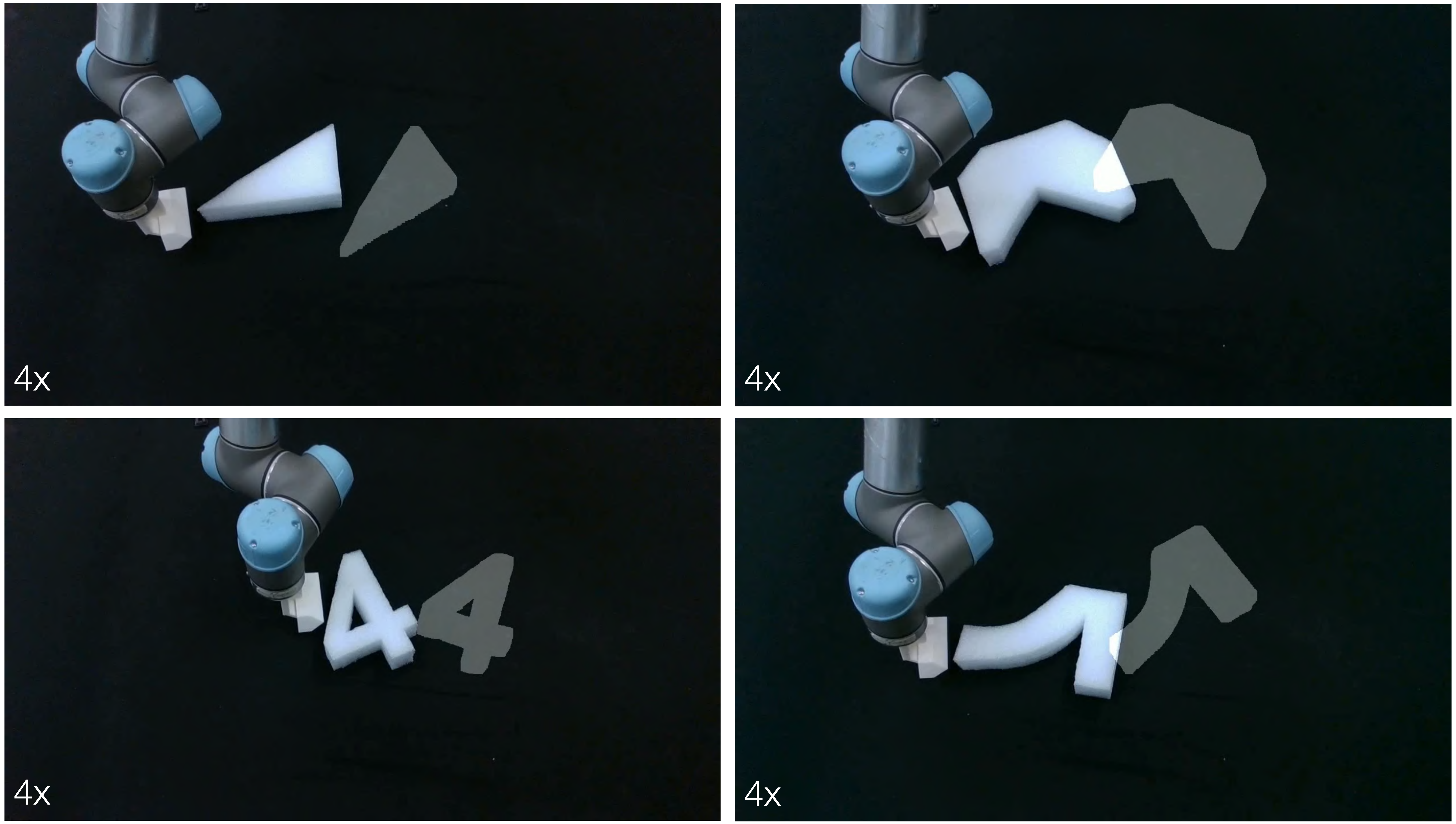
- Diffusion Policy: 0.550 mIoU, 48.2° avg. rotation error
- Ours: **0.731 mIoU, 8.0° avg. rotation error**

Test Generations

Liang*, **Liu***, et al. CoRL 2024

Push-Shape Task

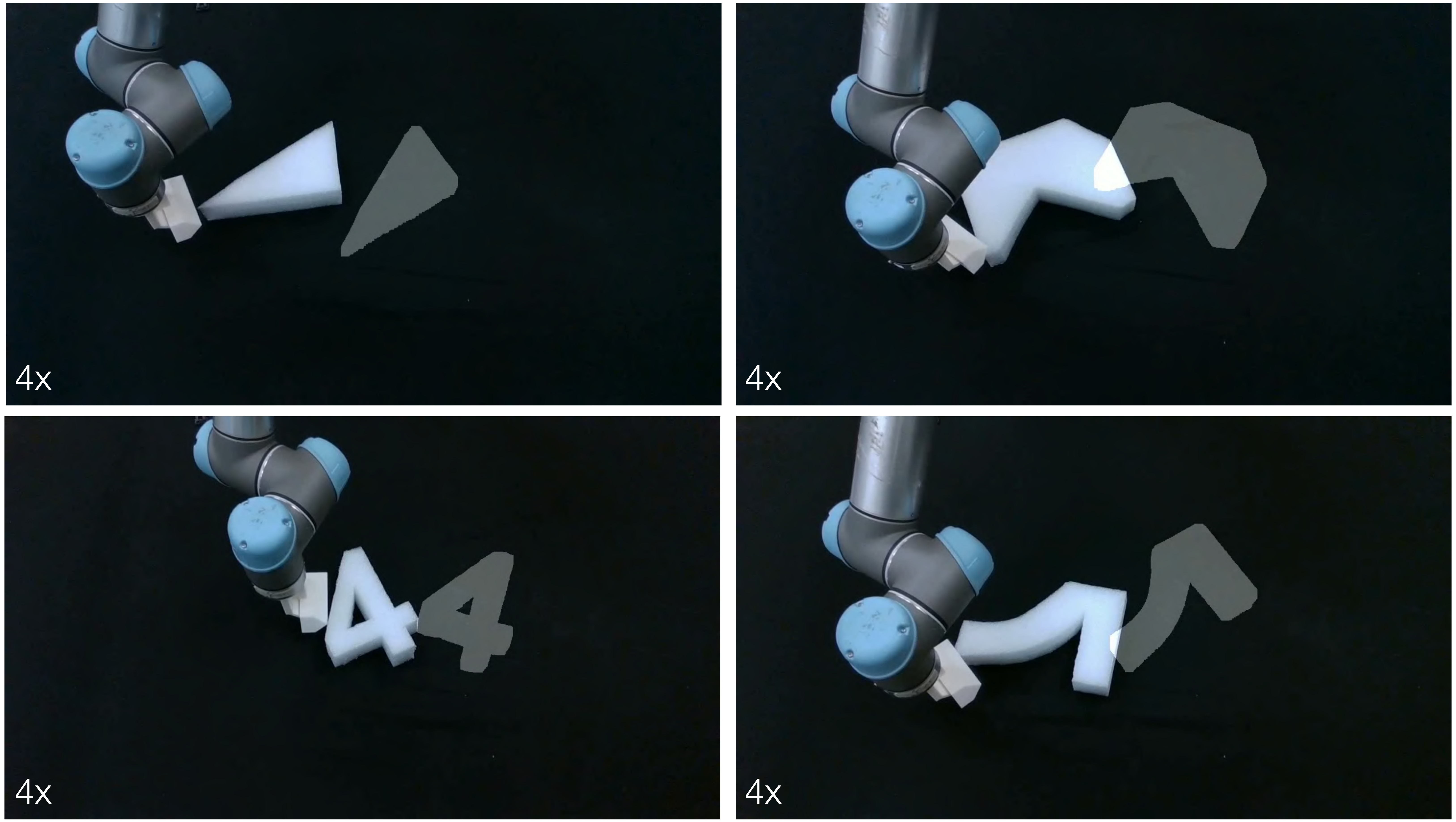
Robot Execution



Liang*, **Liu***, et al. CoRL 2024

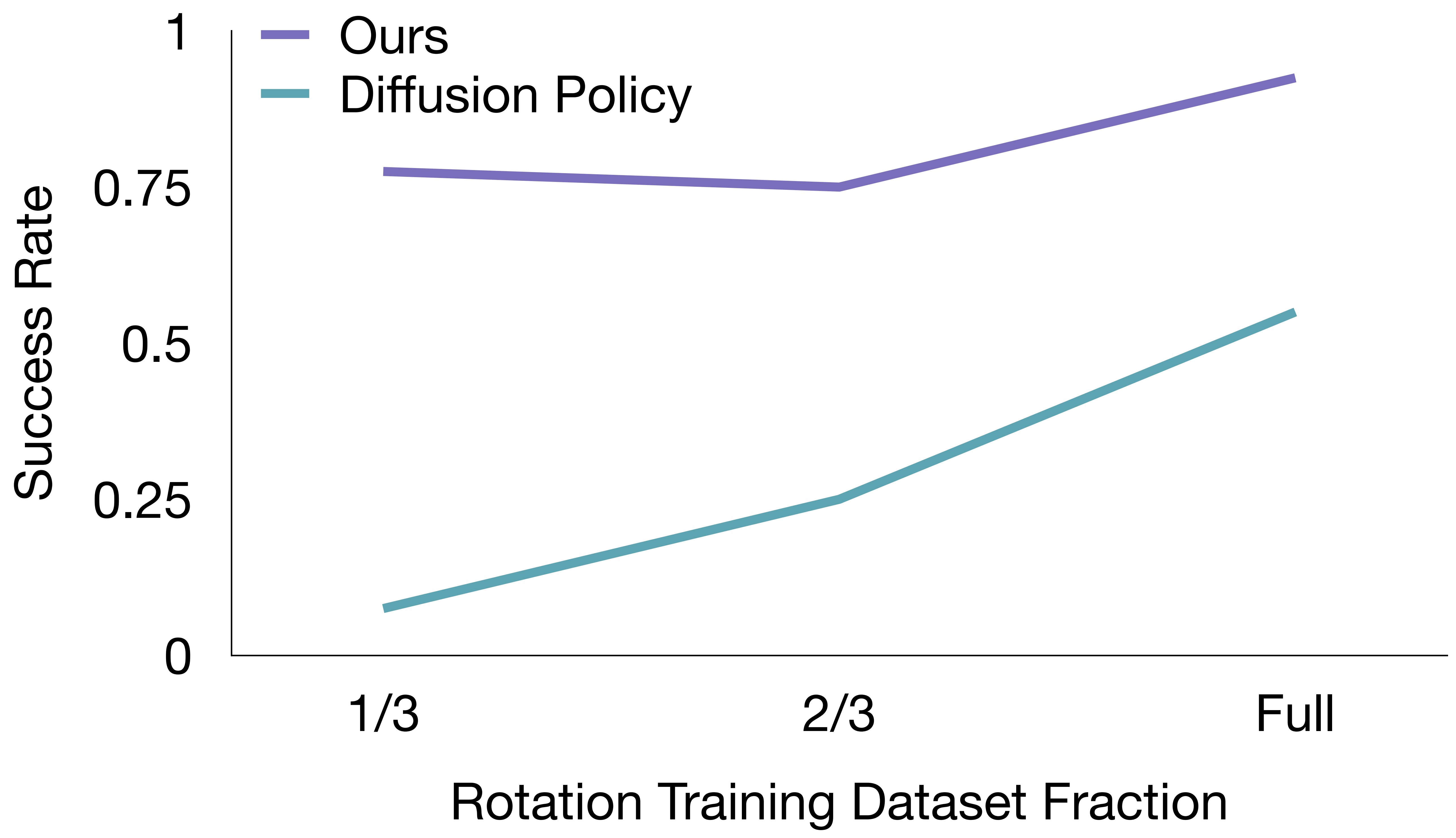
Push-Shape Task

Diffusion Policy Execution



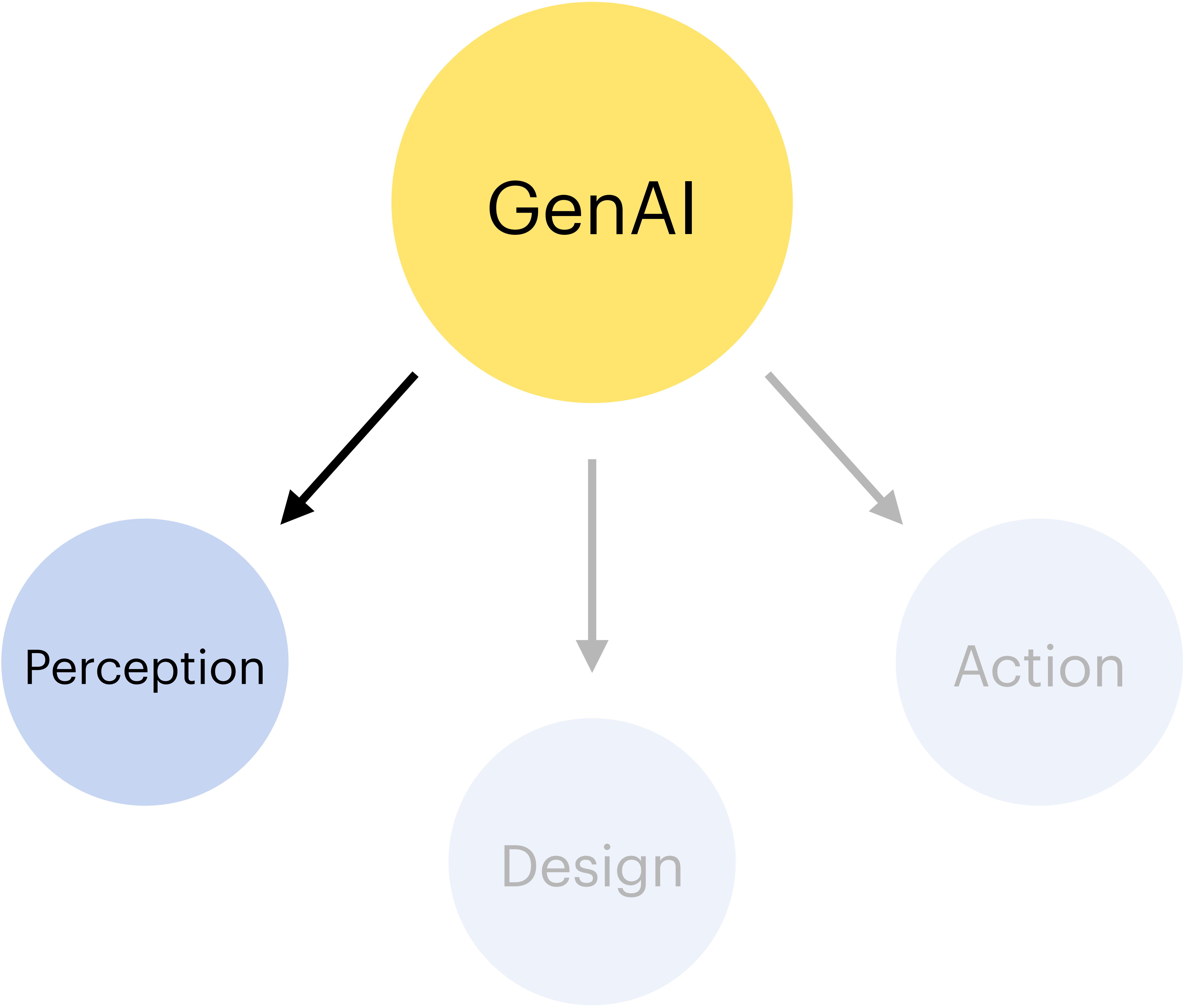
Liang*, **Liu***, et al. CoRL 2024

Performance Scaling Curve

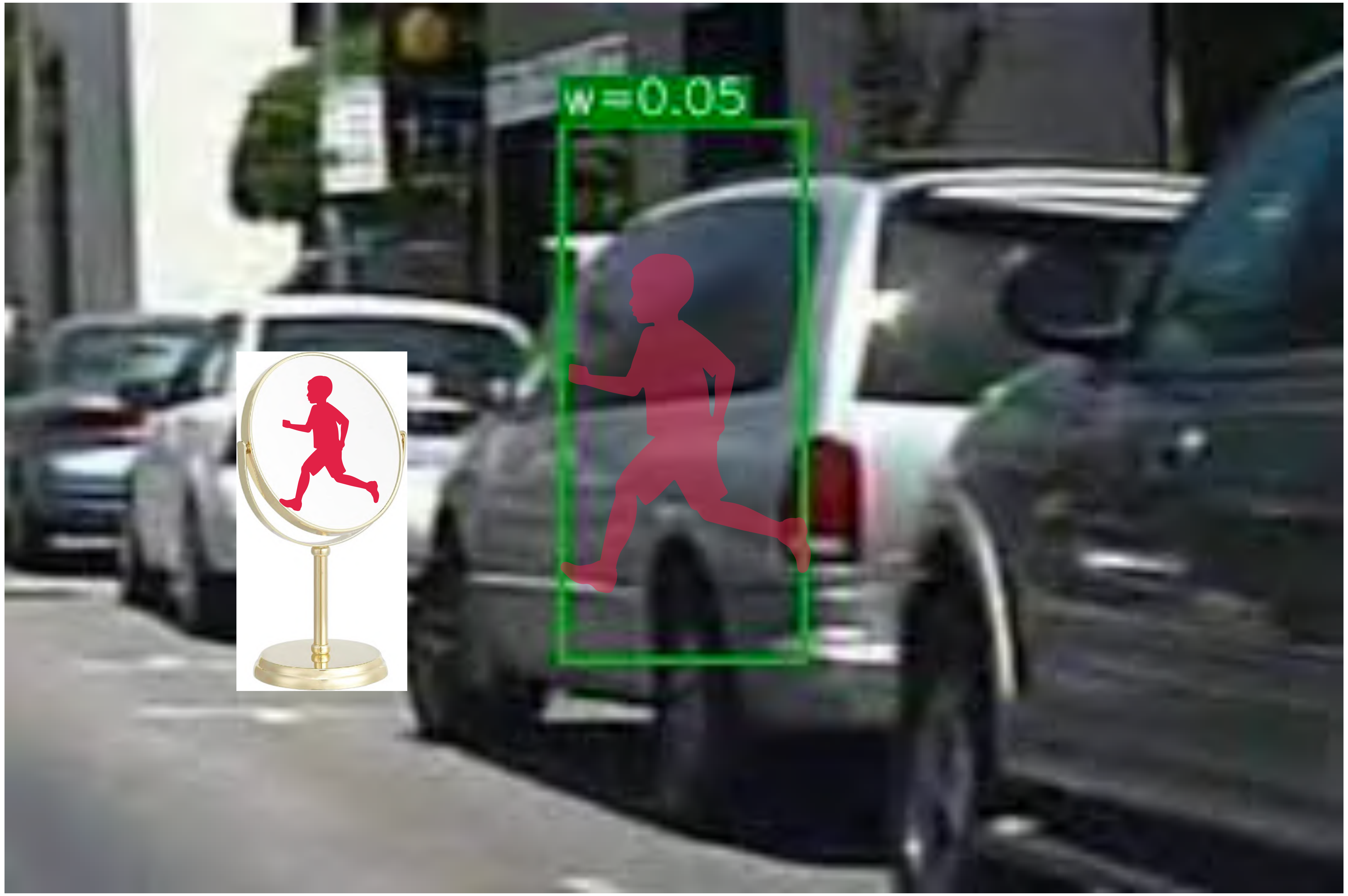


Liang*, **Liu***, et al. CoRL 2024

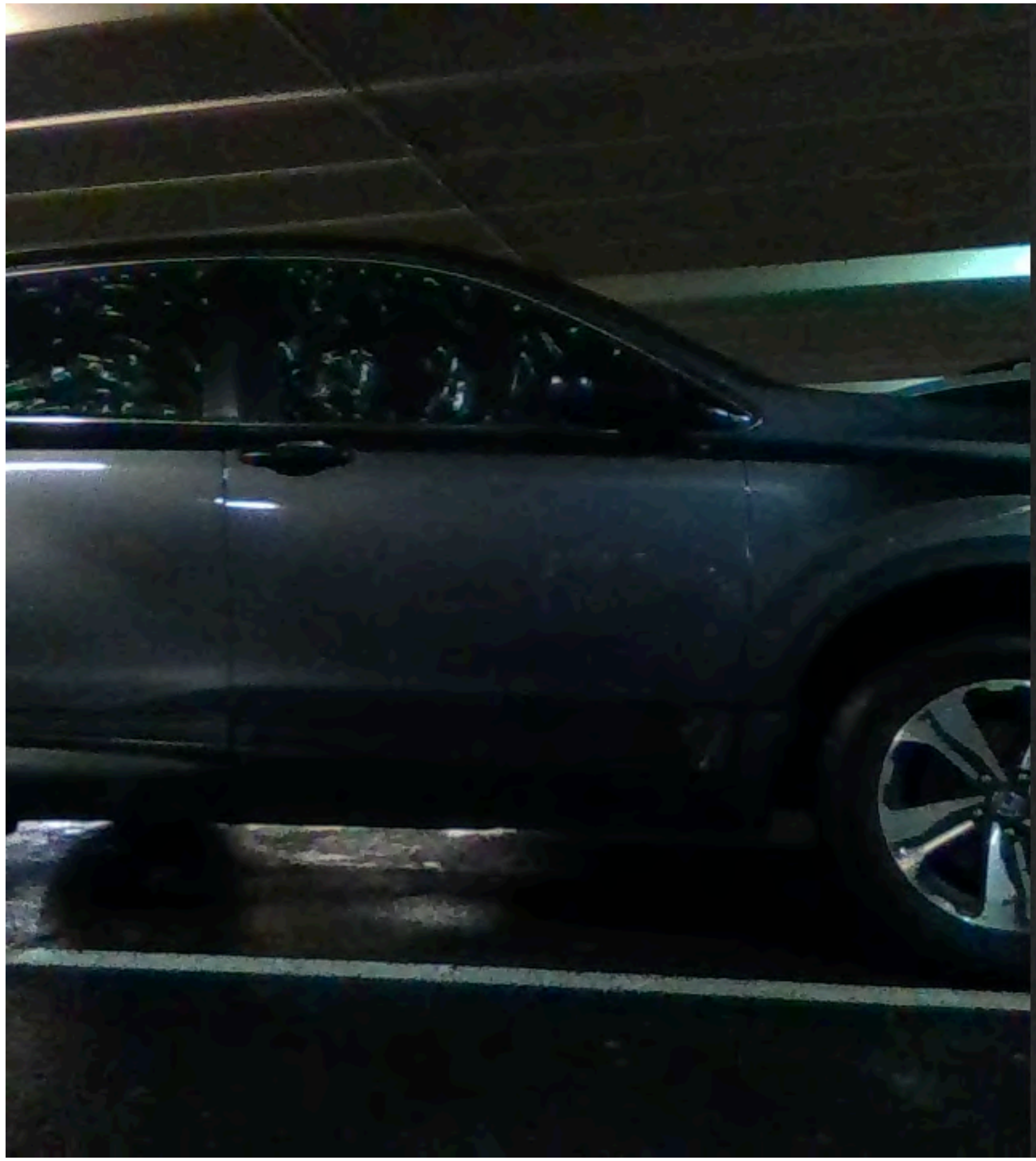
Generative Embodied AI



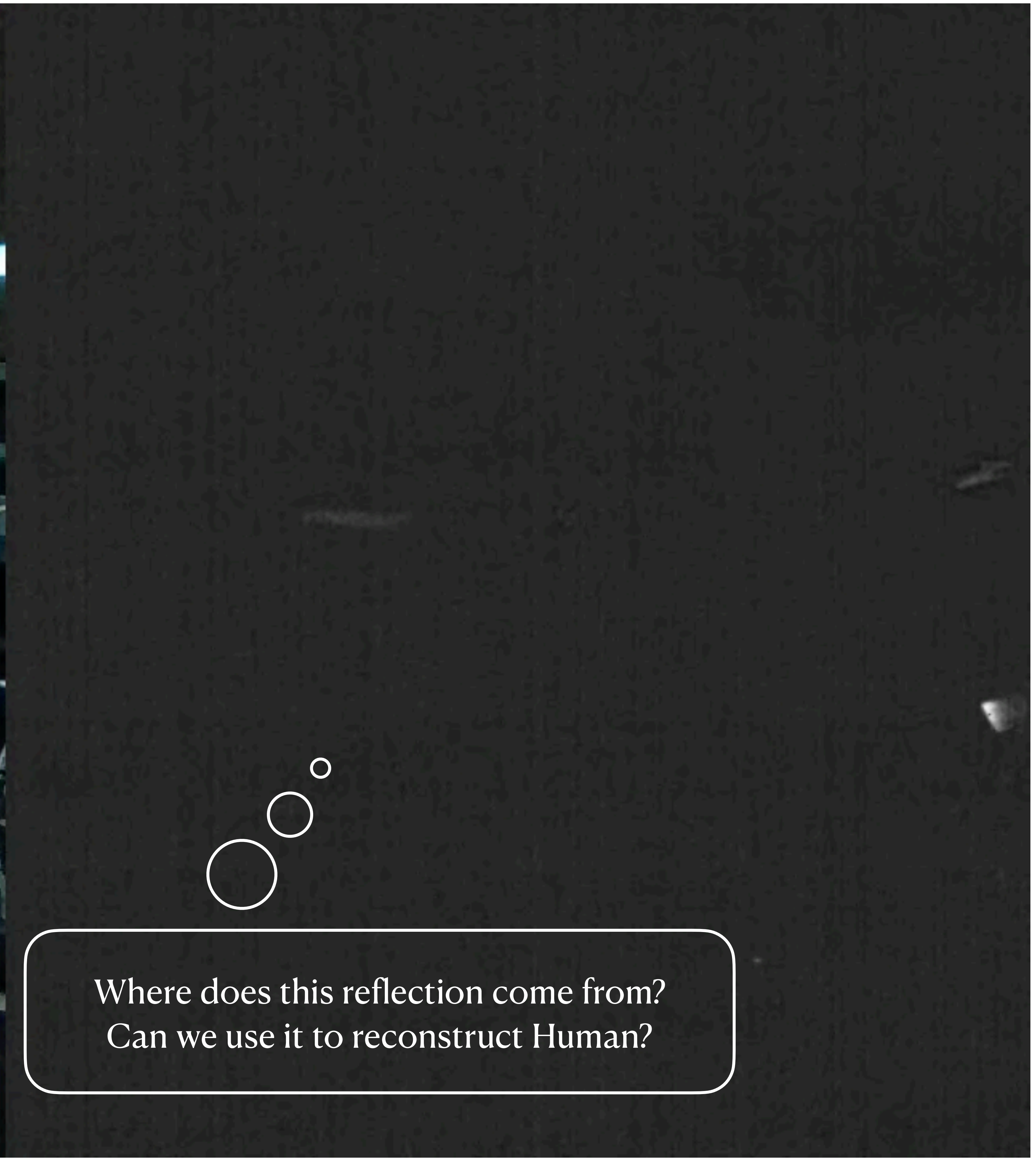
3D Reconstruction Occluded Human



Infrared Turns Cars into Mirrors!

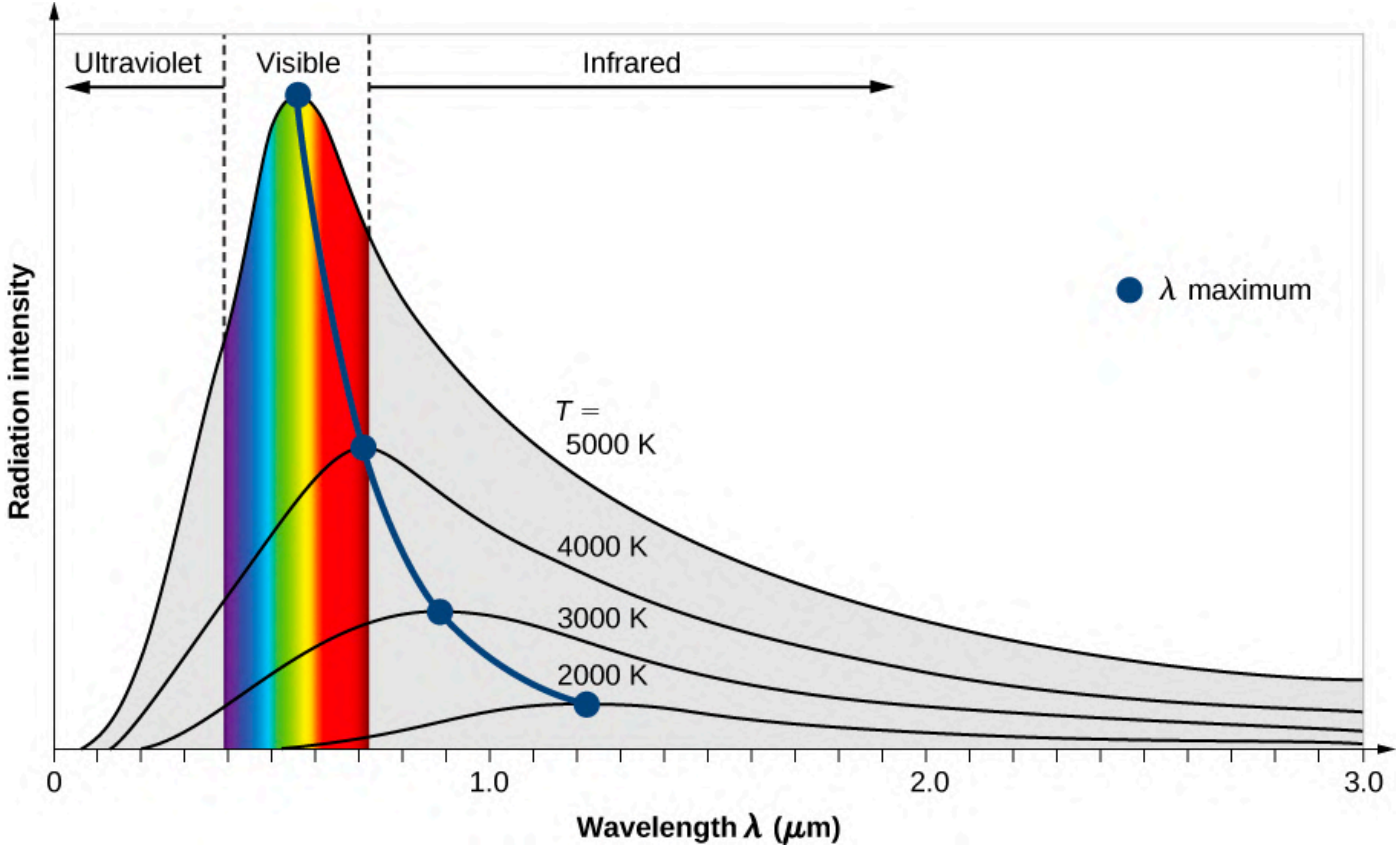


Normal Camera (0.4-0.7 µm)



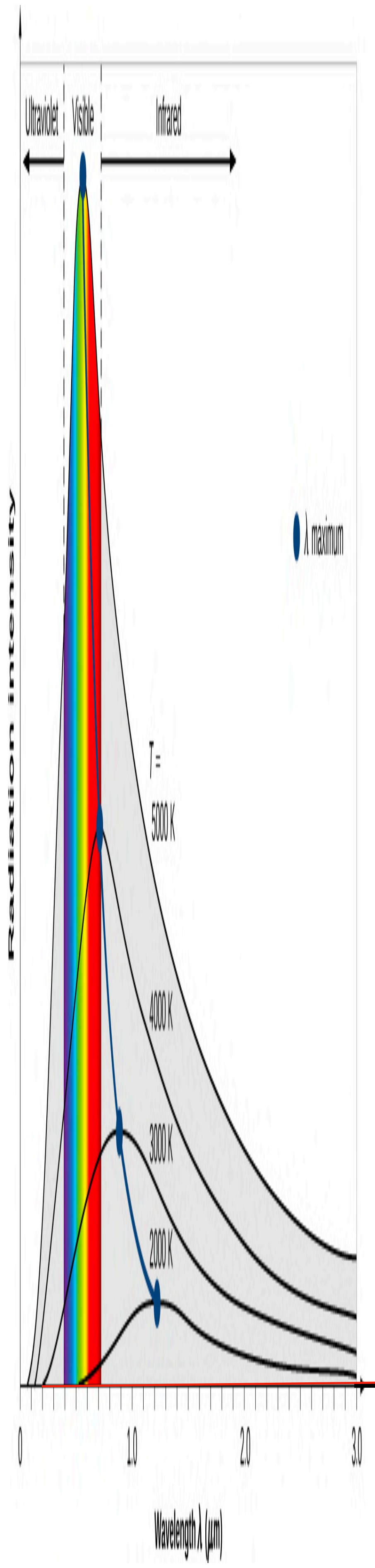
Thermal Camera (7-14 µm)

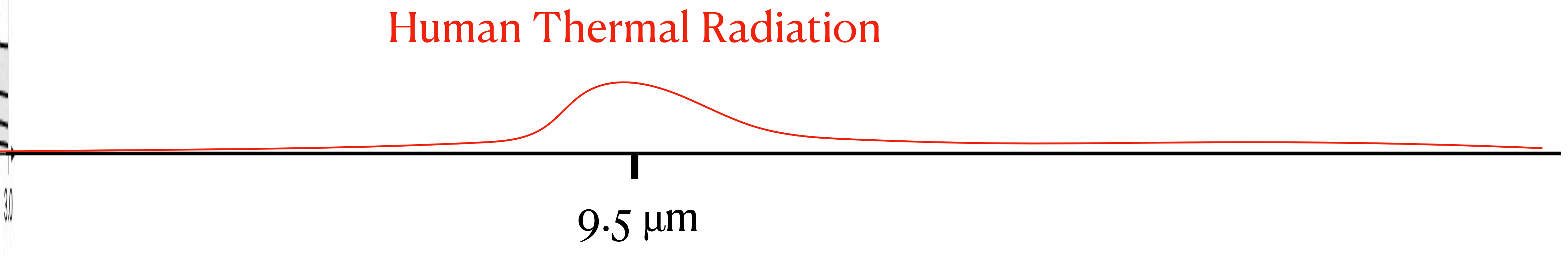
Black-body Radiation



Source: Wikipedia

Human bodies are long wavelength infrared light bulbs





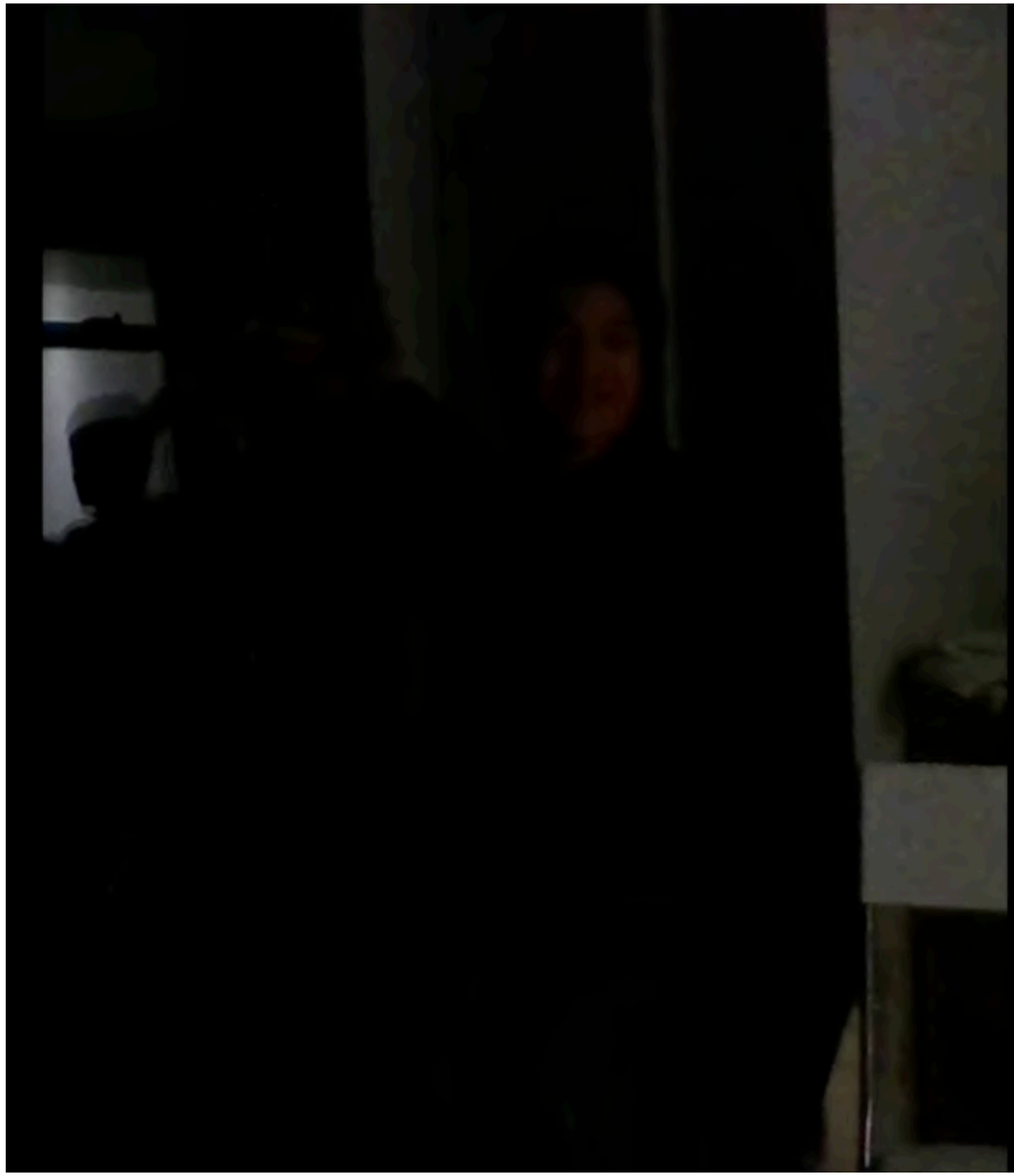
Source: Wikipedia

Humans are infrared light bulbs

Normal Camera (0.4-0.7 μm) | Thermal Camera (7-14 μm)

Source: Wikipedia

Cats are infrared light bulbs



Normal Camera (0.4-0.7 μm)

Thermal Camera (7-14 μm)

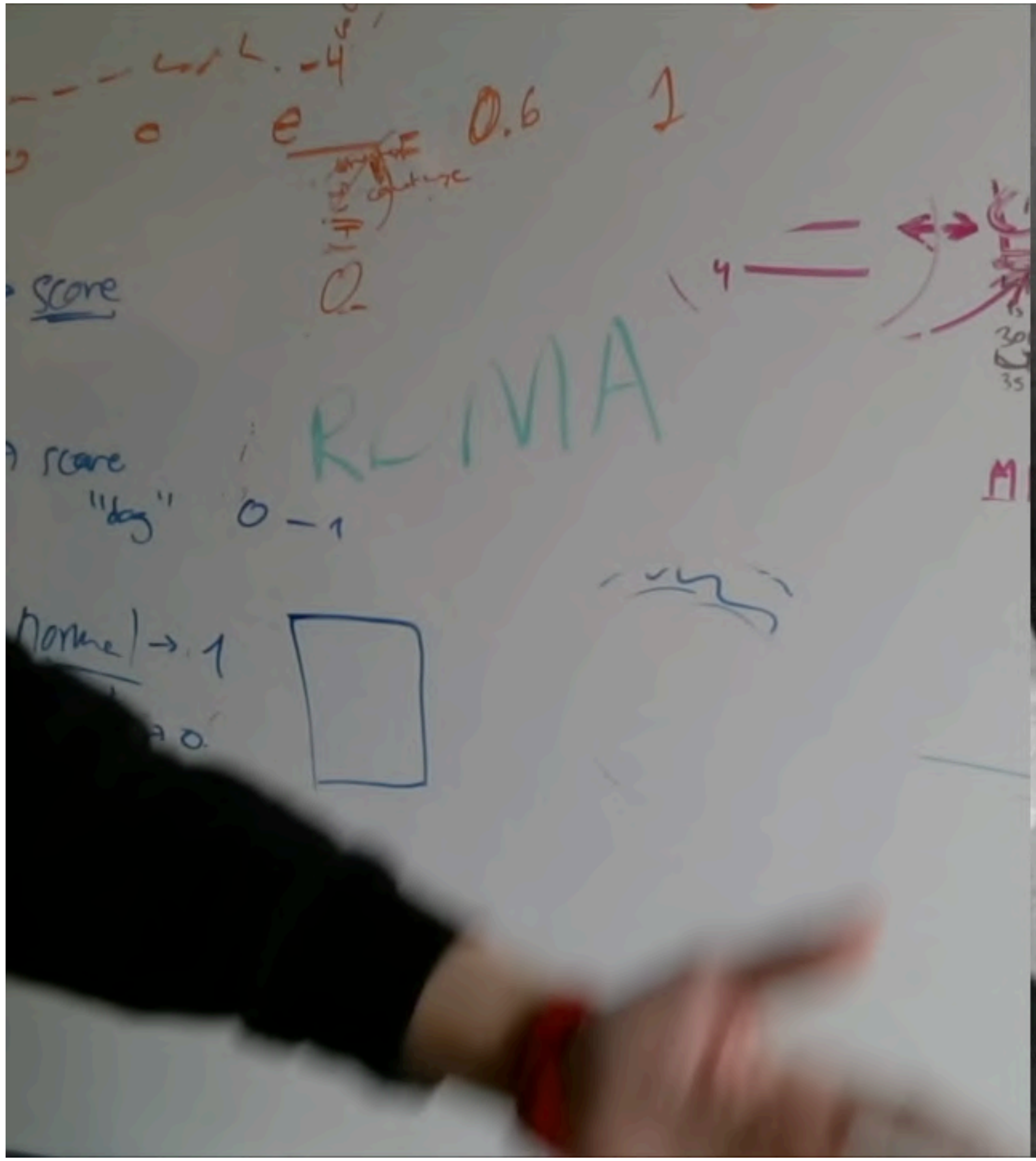
Source: Wikipedia

Many objects are more reflective in infrared

Normal Camera (0.4-0.7 μm)

Thermal Camera (7-14 μm)

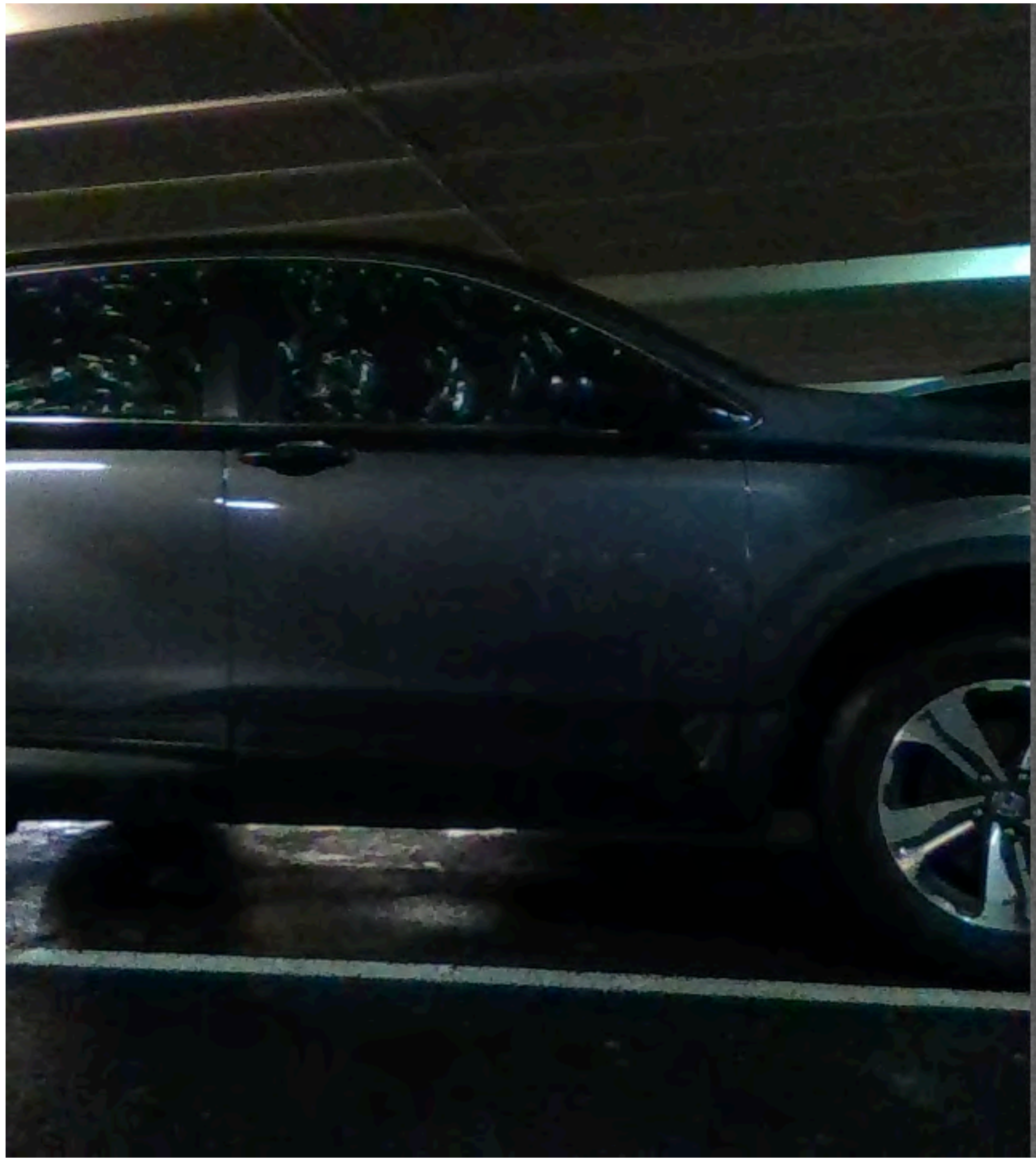
Many objects are more reflective in infrared



Normal Camera (0.4-0.7 μm)

Thermal Camera (7-14 μm)

Many objects are more reflective in infrared



Normal Camera (0.4-0.7 μm)

Thermal Camera (7-14 μm)

Many objects are more reflective in infrared

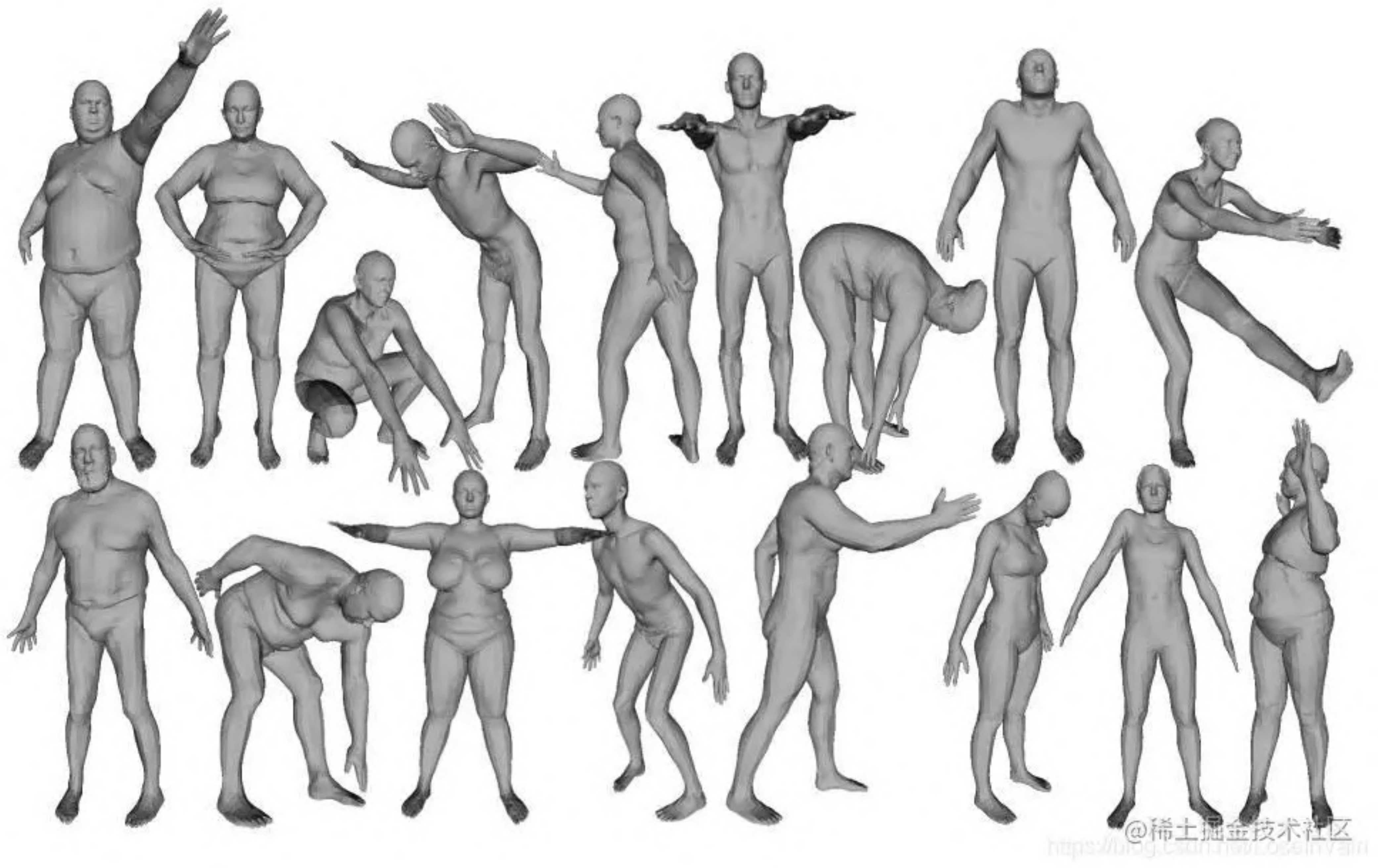
Normal Camera (0.4-0.7 μm) Thermal Camera (7-14 μm)

Many objects are more reflective in infrared

Normal Camera (0.4-0.7 µm)

Thermal Camera (7-14 µm)

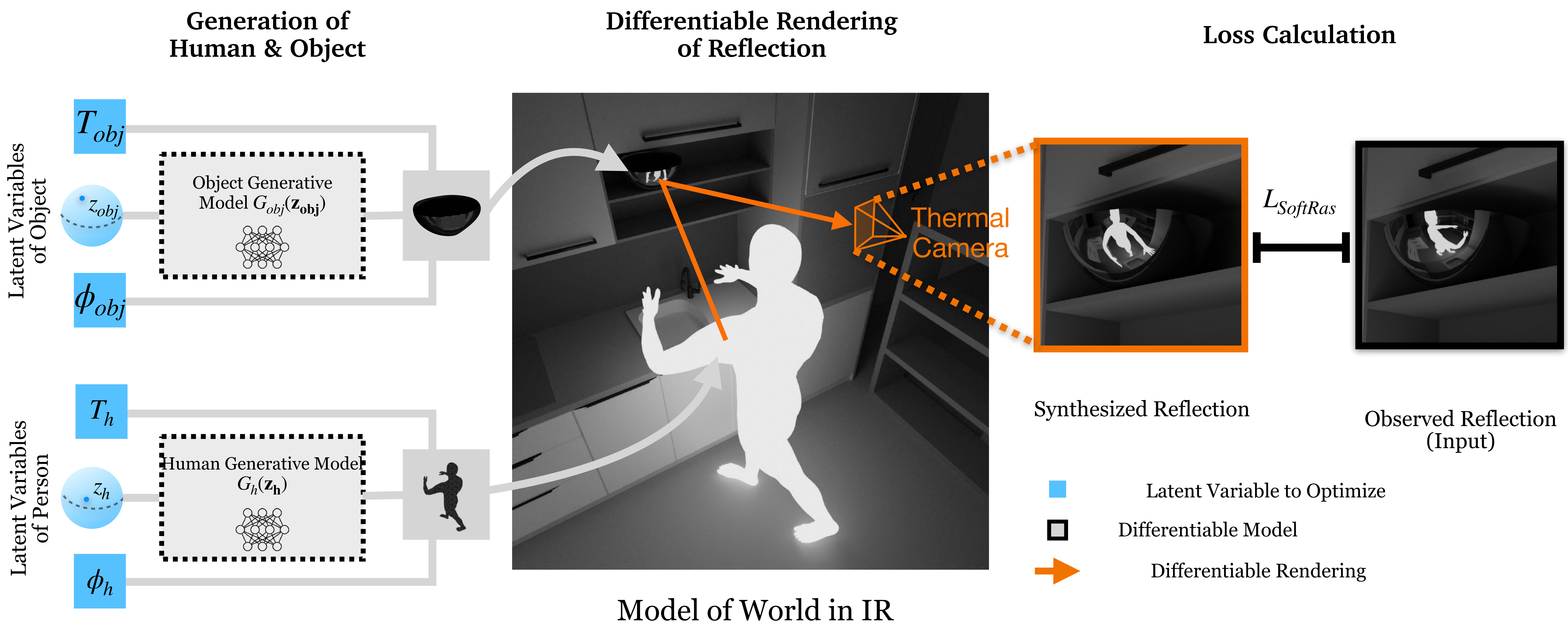
3D Generative Models



3D Human

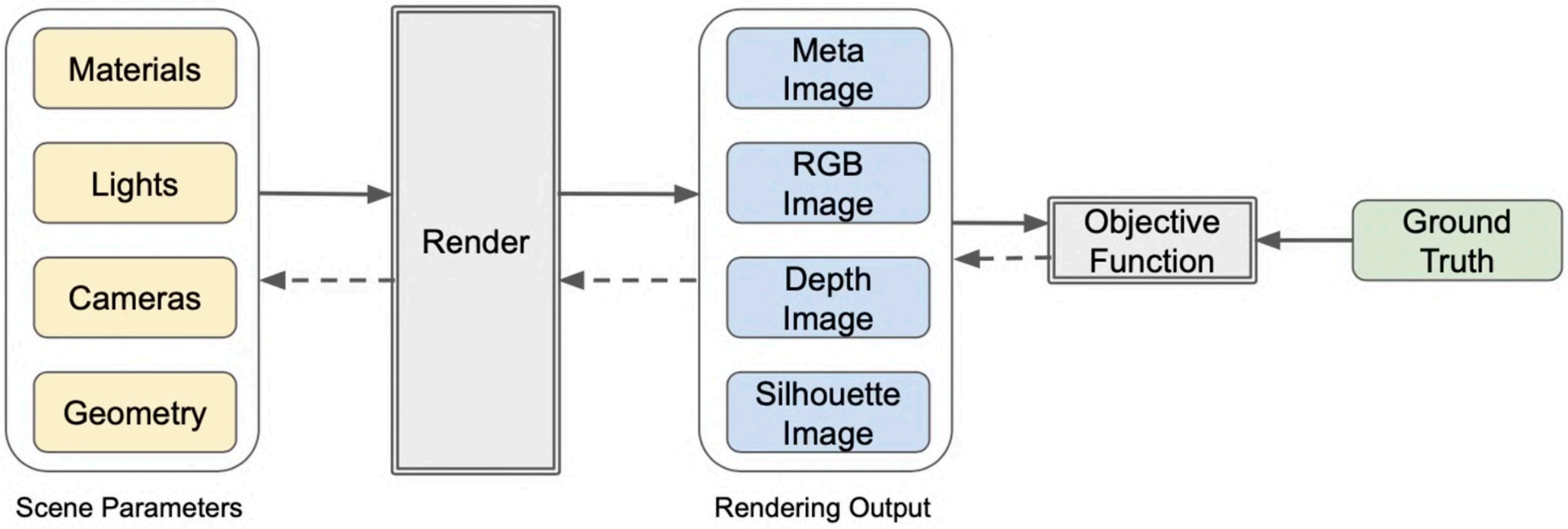
3D Objects

Method

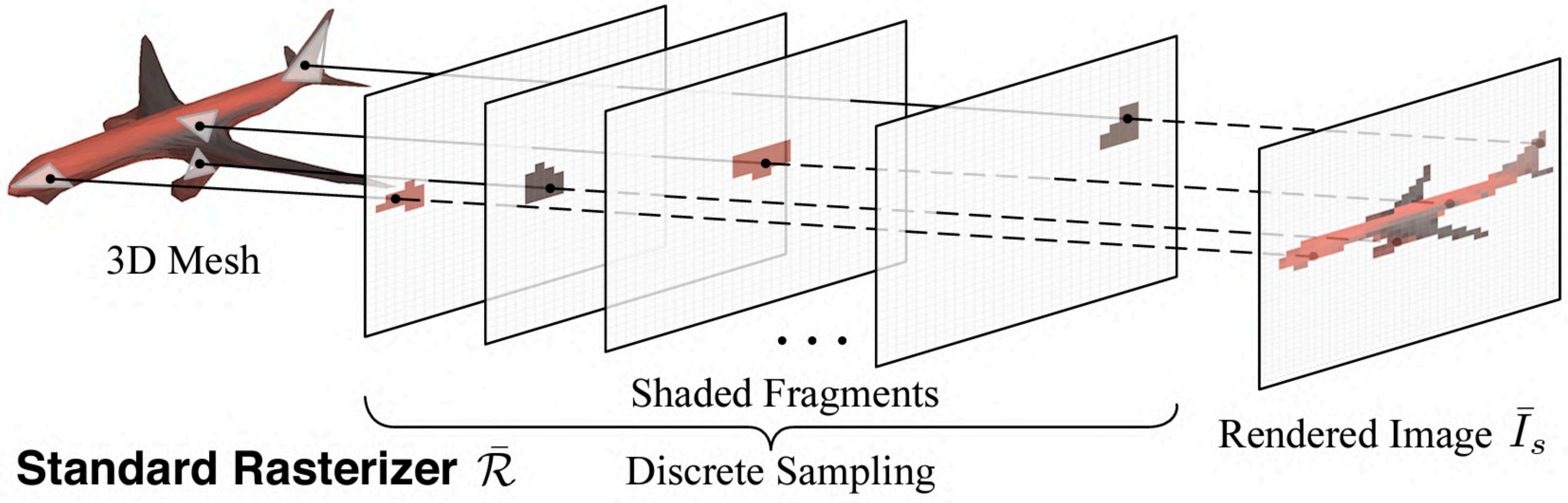


Liu, Vondrick. CVPR 2023.

Differentiable Rendering

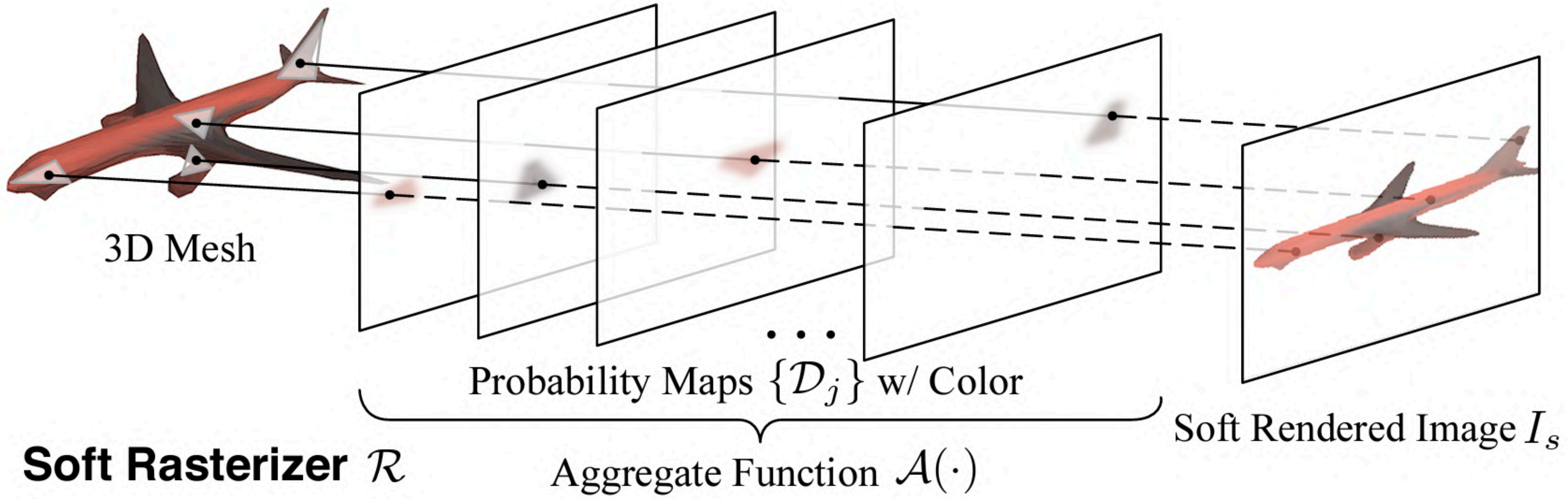


Non-differentiable Rendering



Liu, Li, Chen, Li. CVPR 2019.

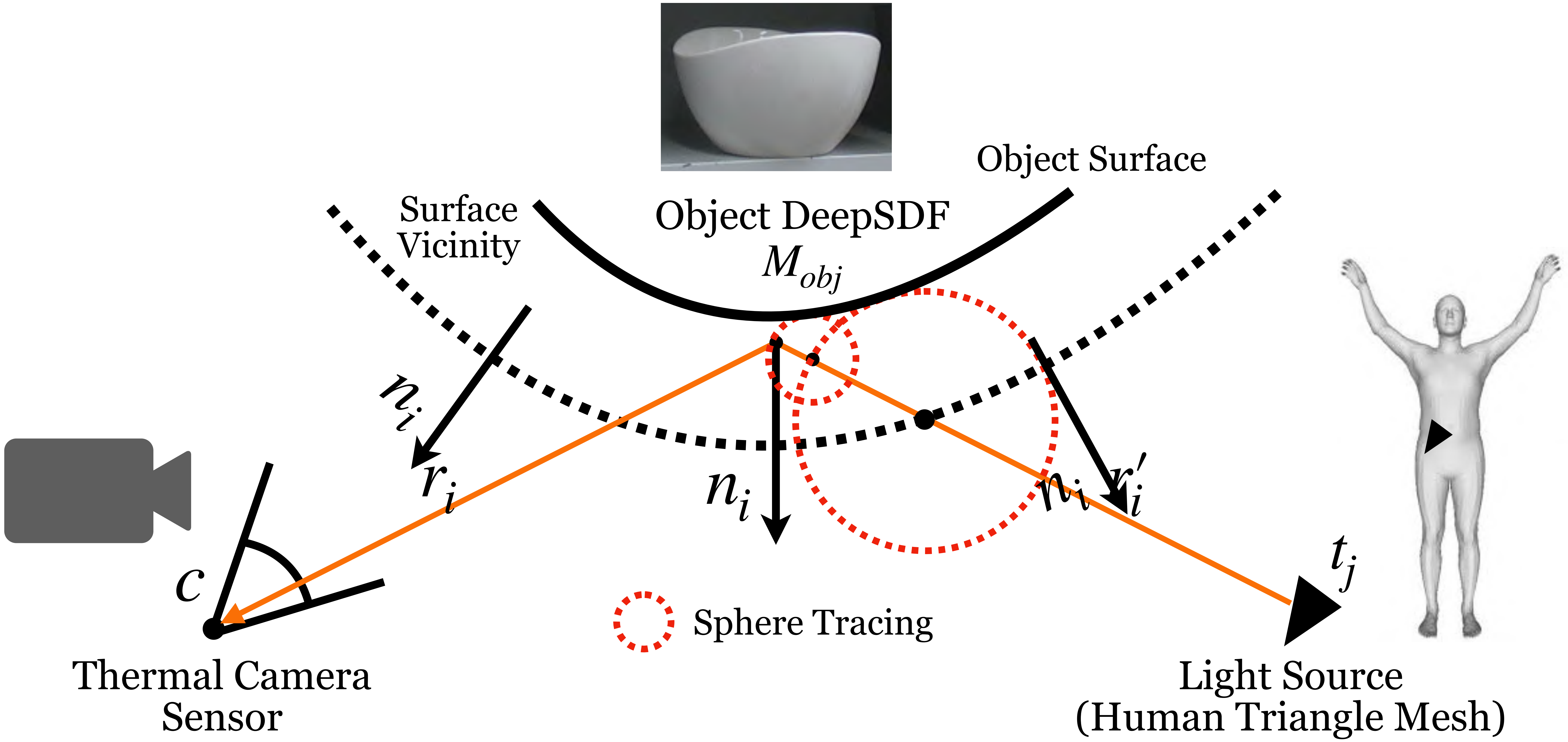
Differentiable Rendering



What about reflection?

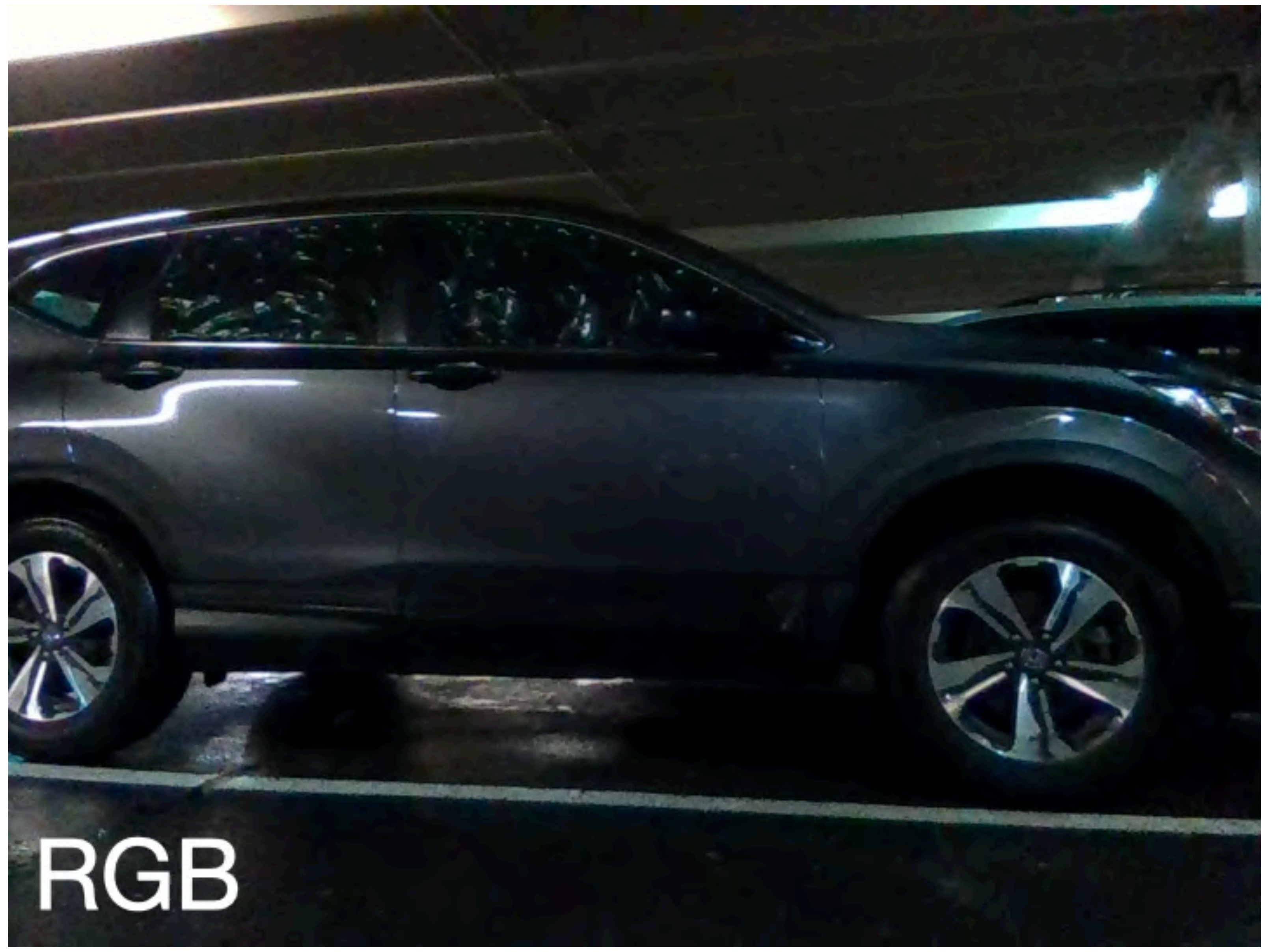
Liu, Li, Chen, Li. CVPR 2019.

Differentiable Rendering of Reflection

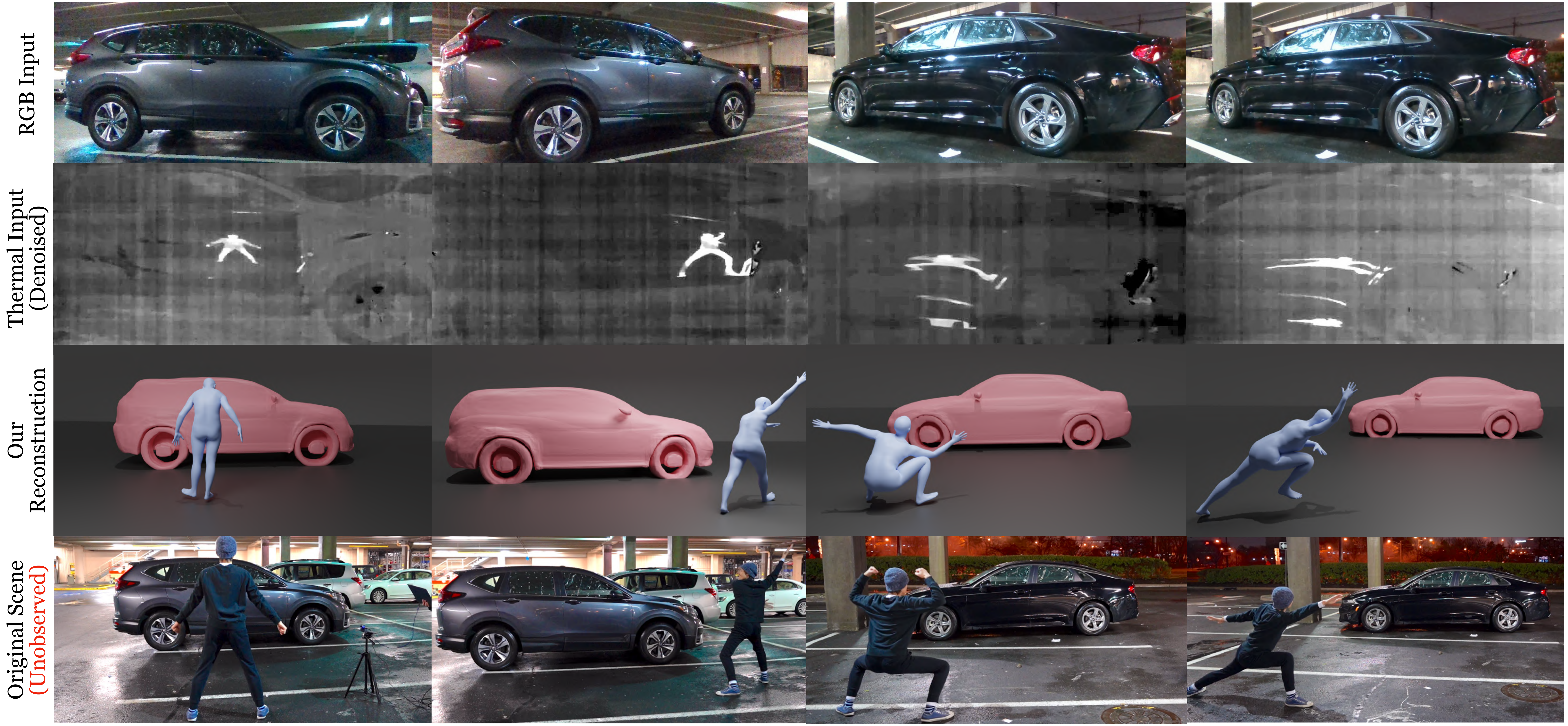


Liu, Vondrick. CVPR 2023.

Results



Results



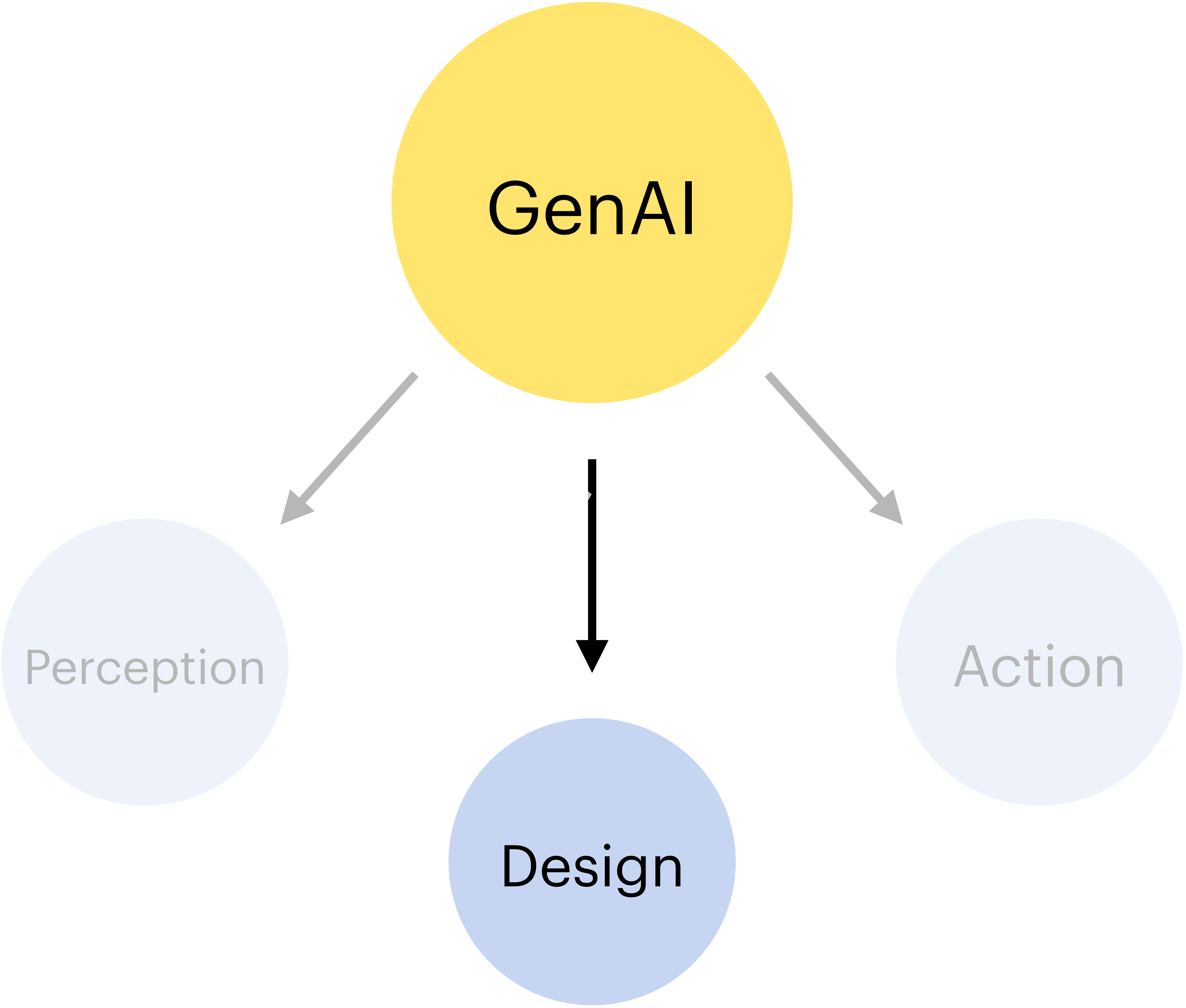
Liu, Vondrick. CVPR 2023.

RGB Input

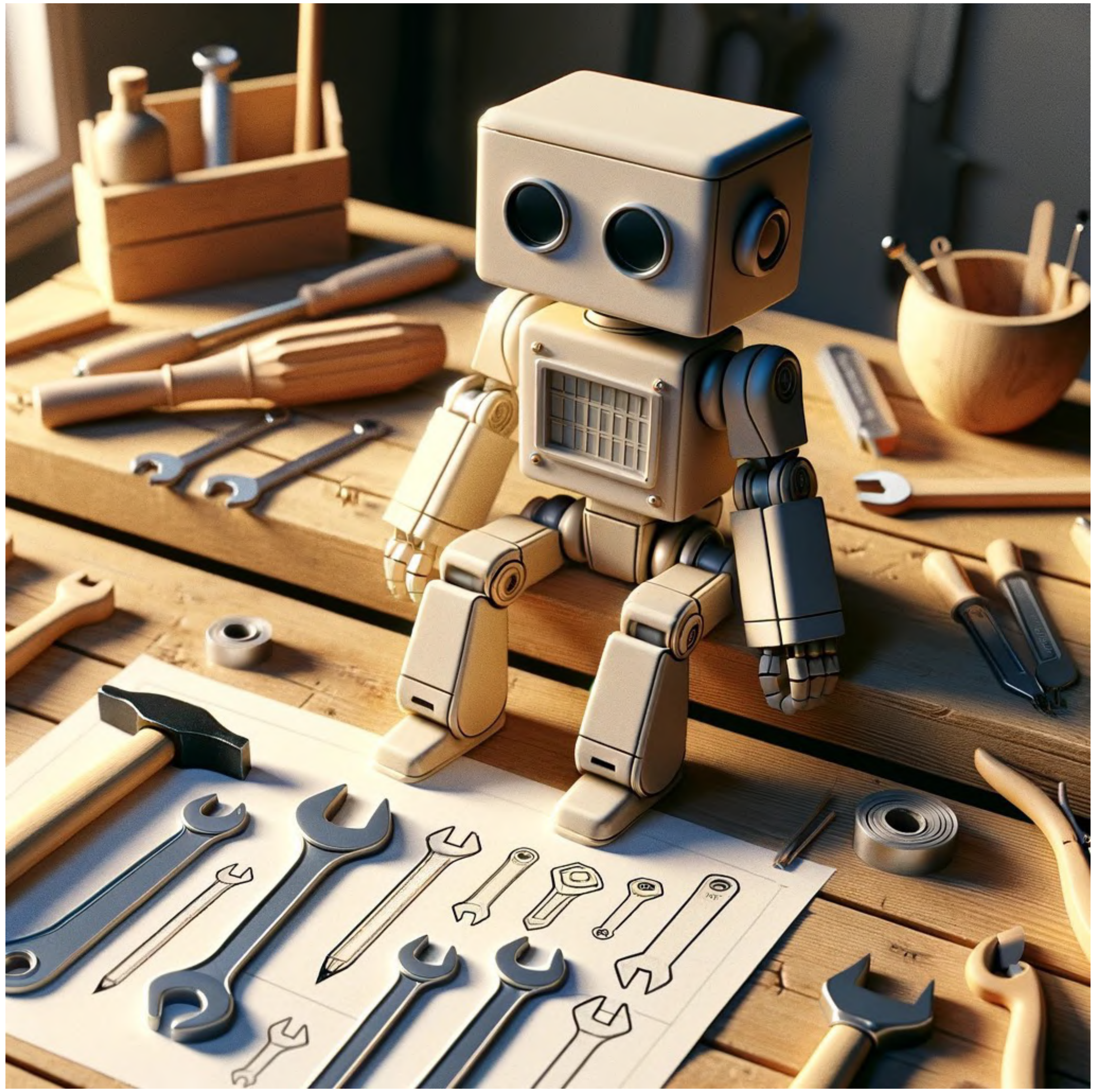
RGB Input

Liu, Vondrick. CVPR 2023.

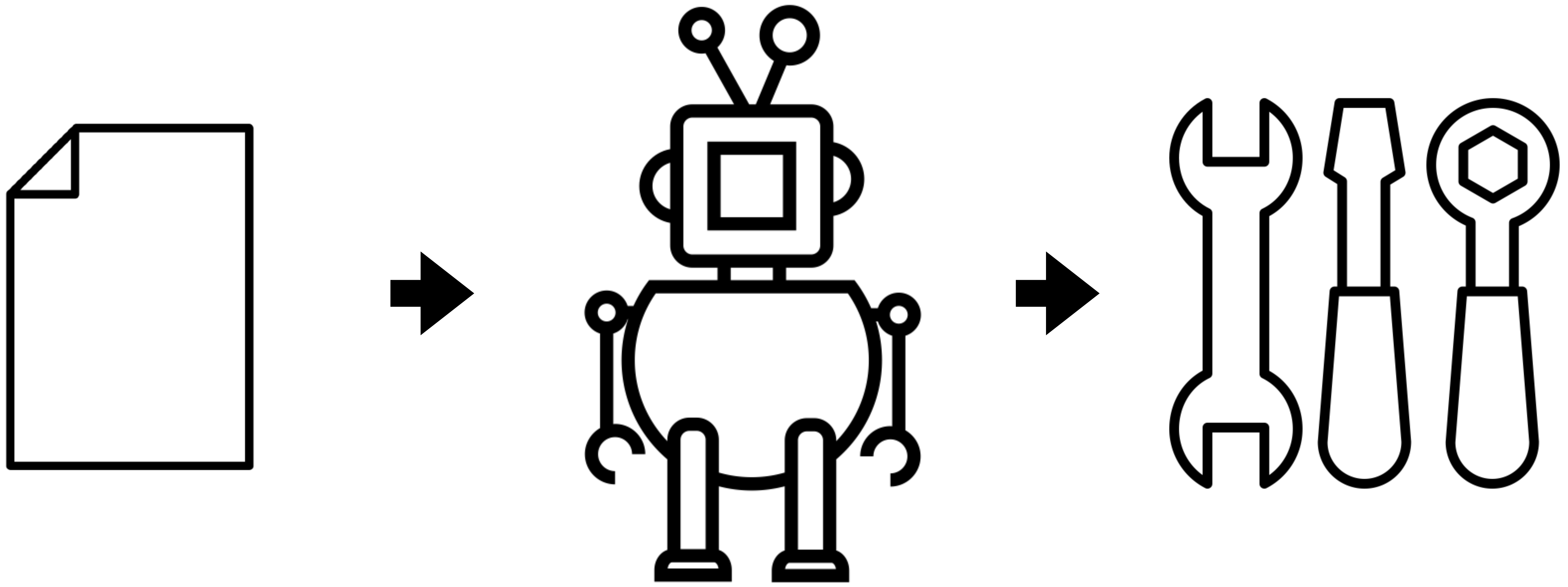
Generative Embodied AI



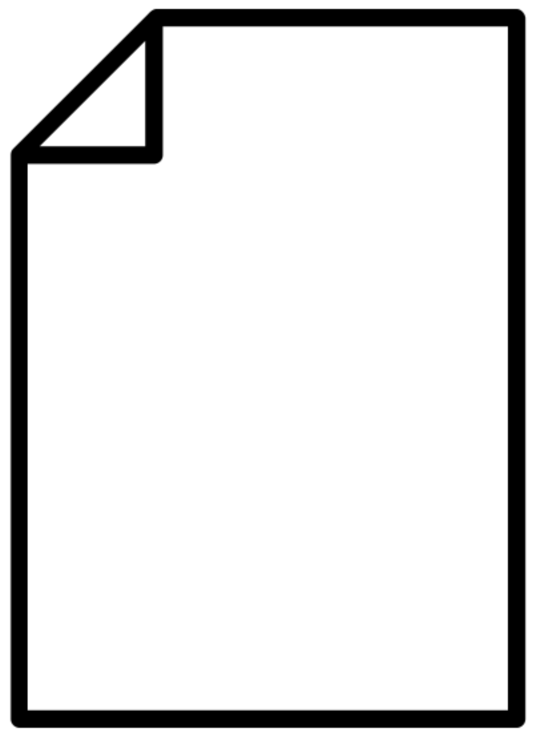
Robotic Tool Design



Designing Paper Tools

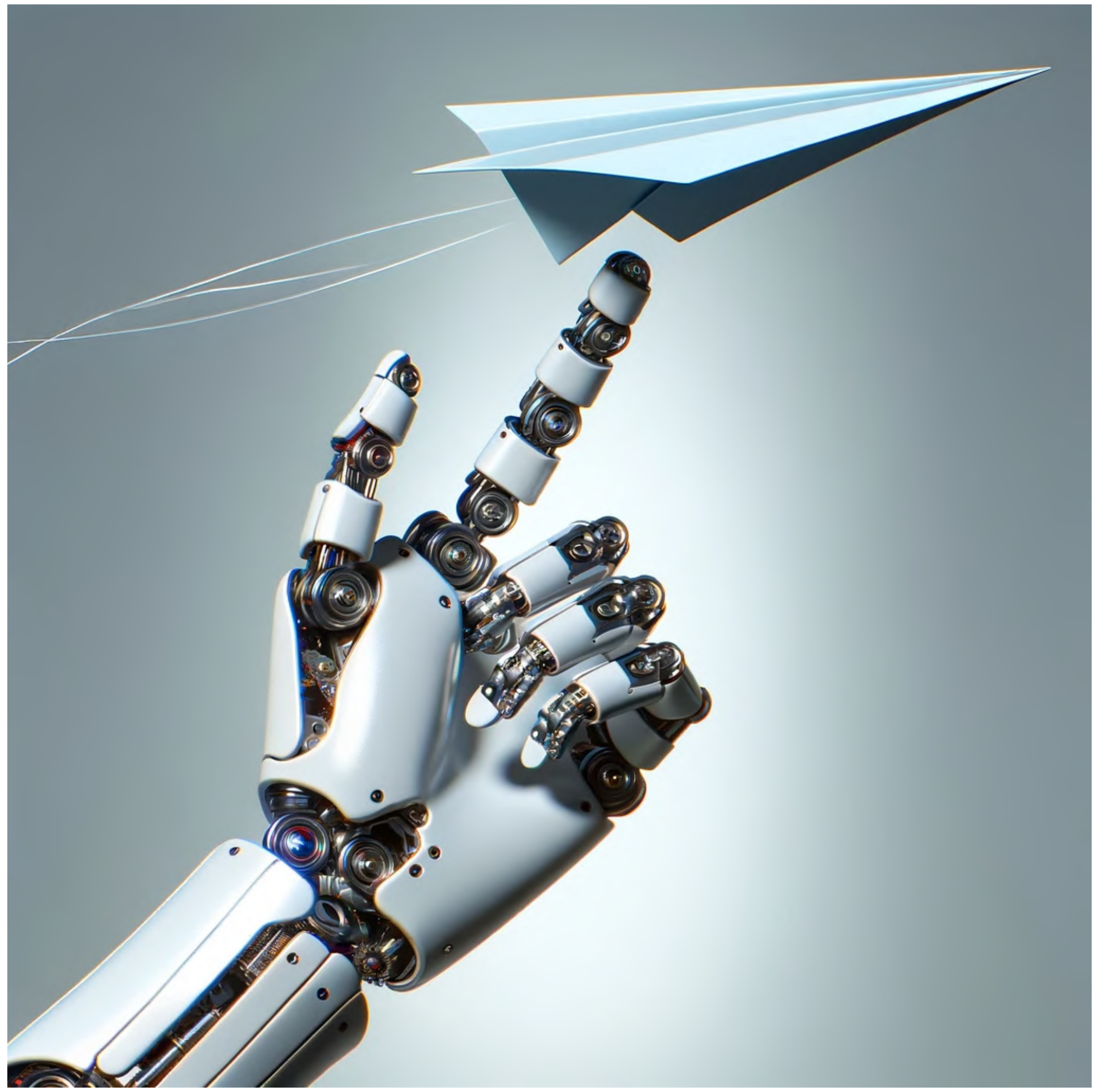


Why Paper?



Paper Makes Practical Recyclable Tools!

Case Study: Paper Airplane



Design

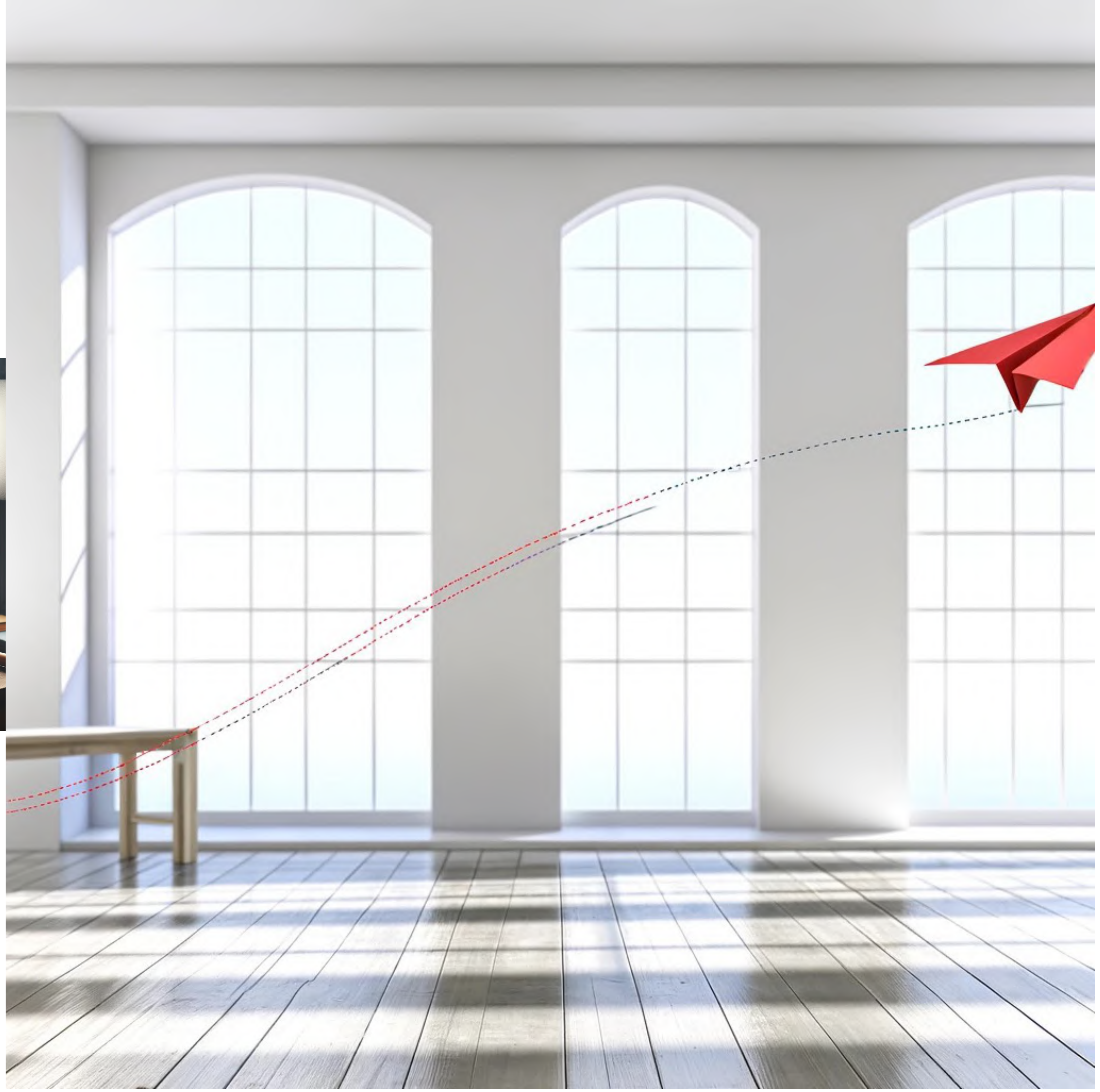
Design

Build

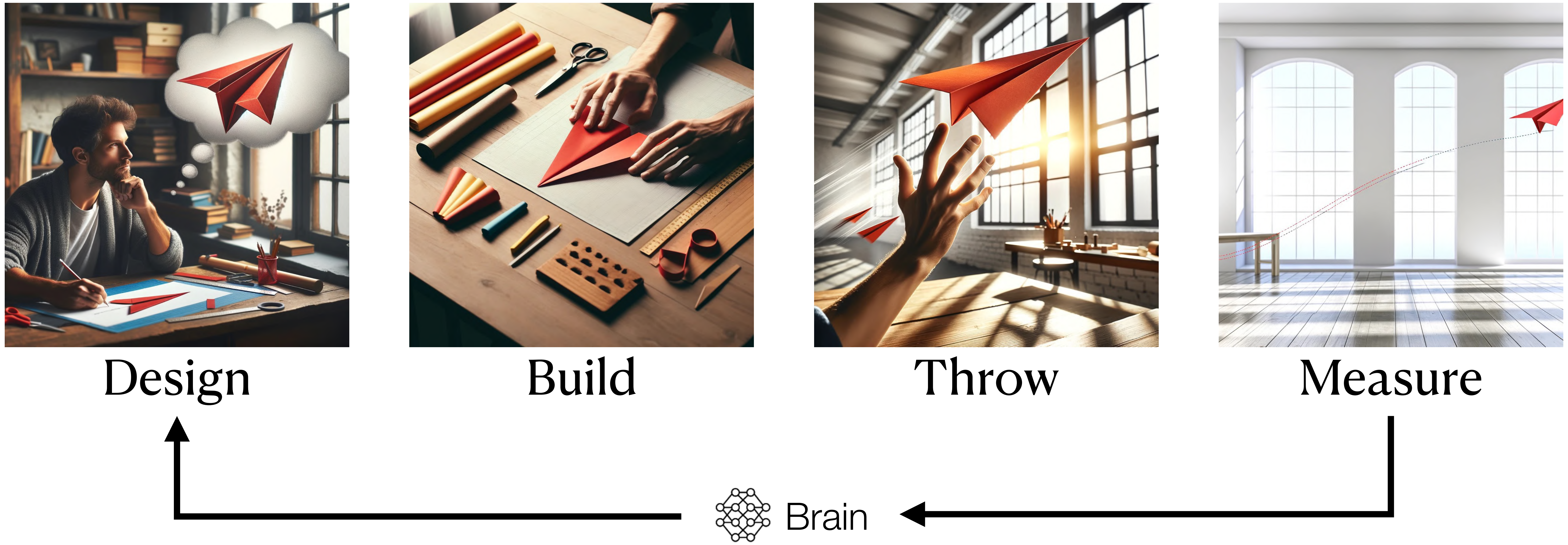

Design

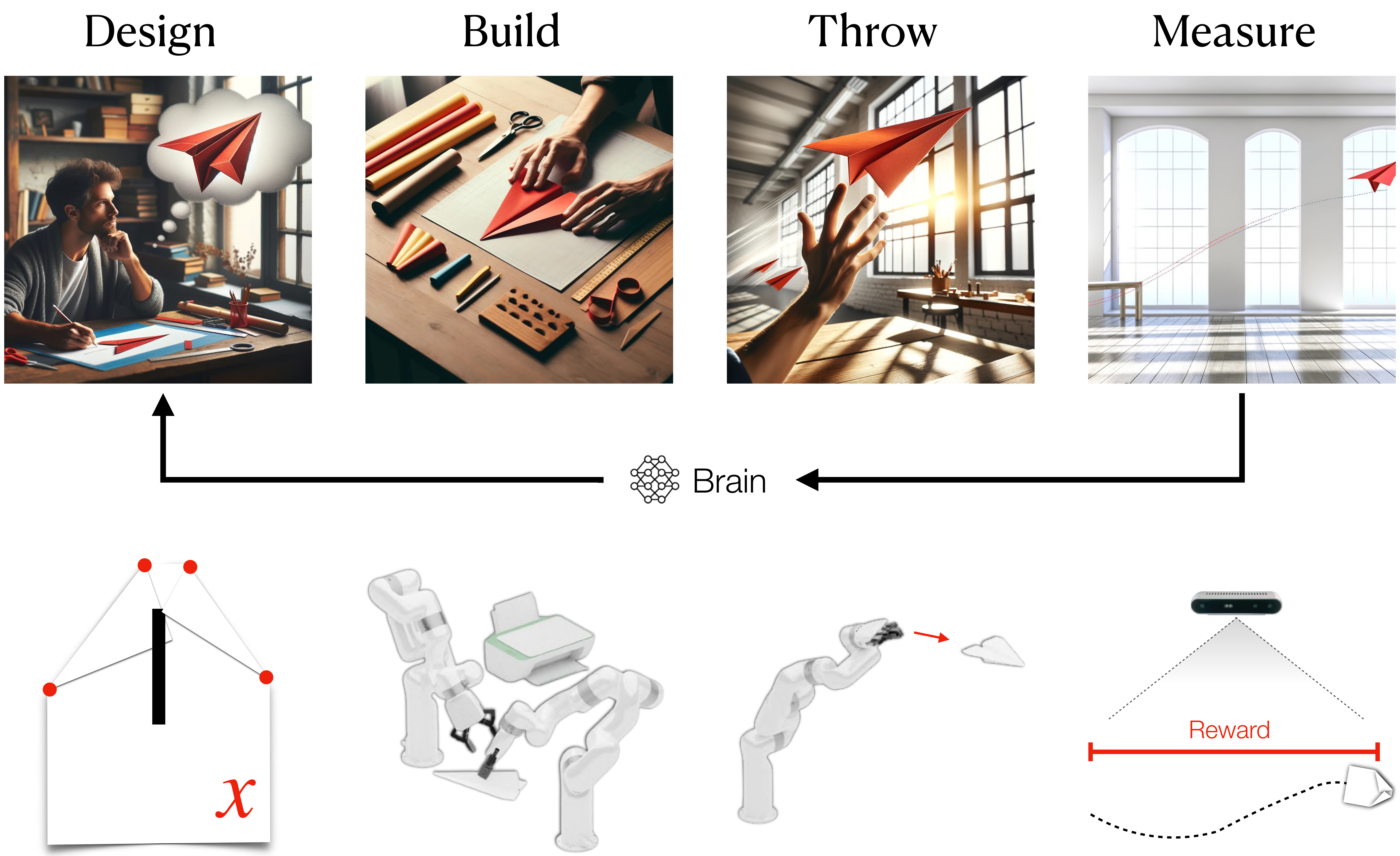

Throw

Design

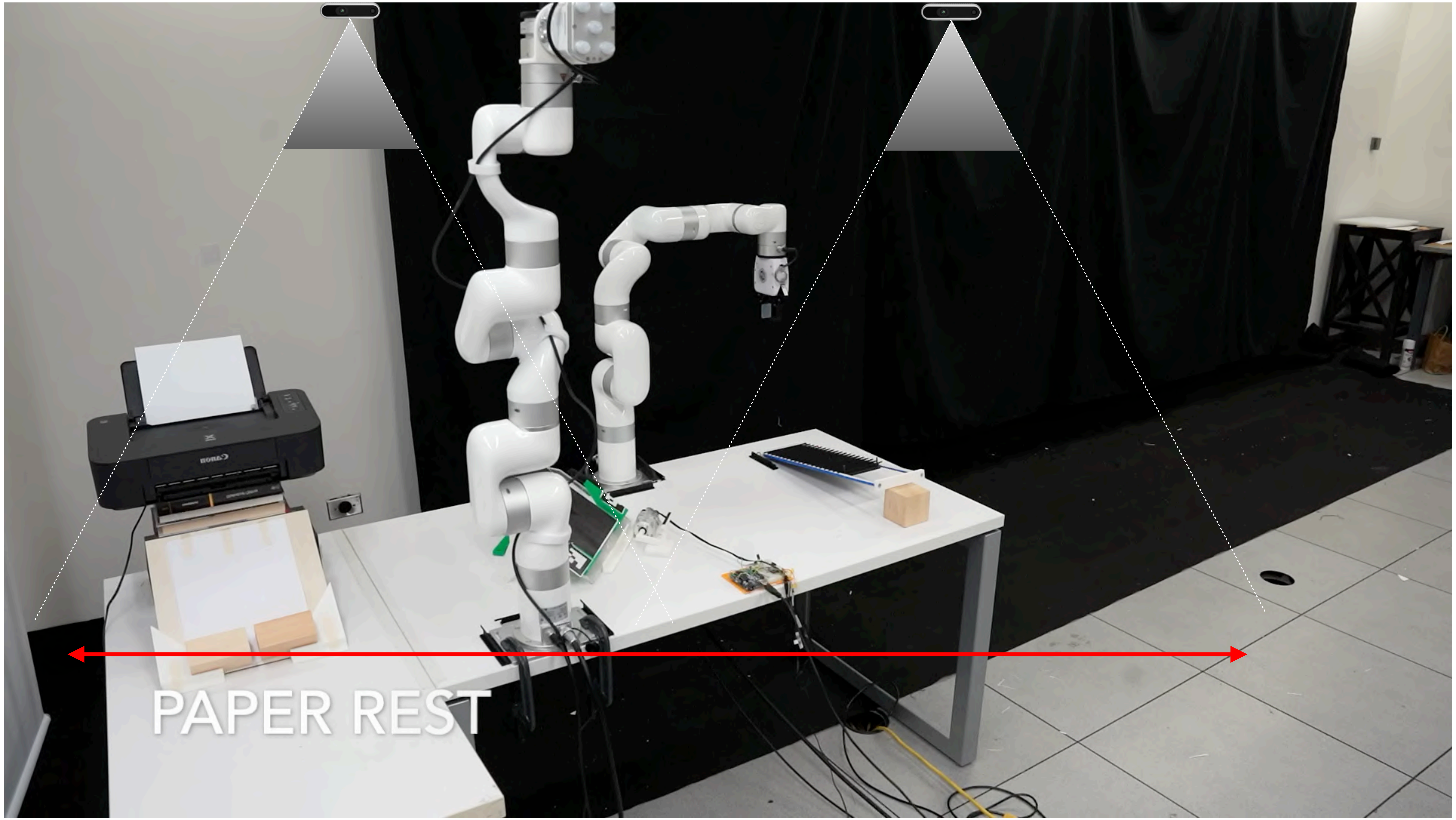


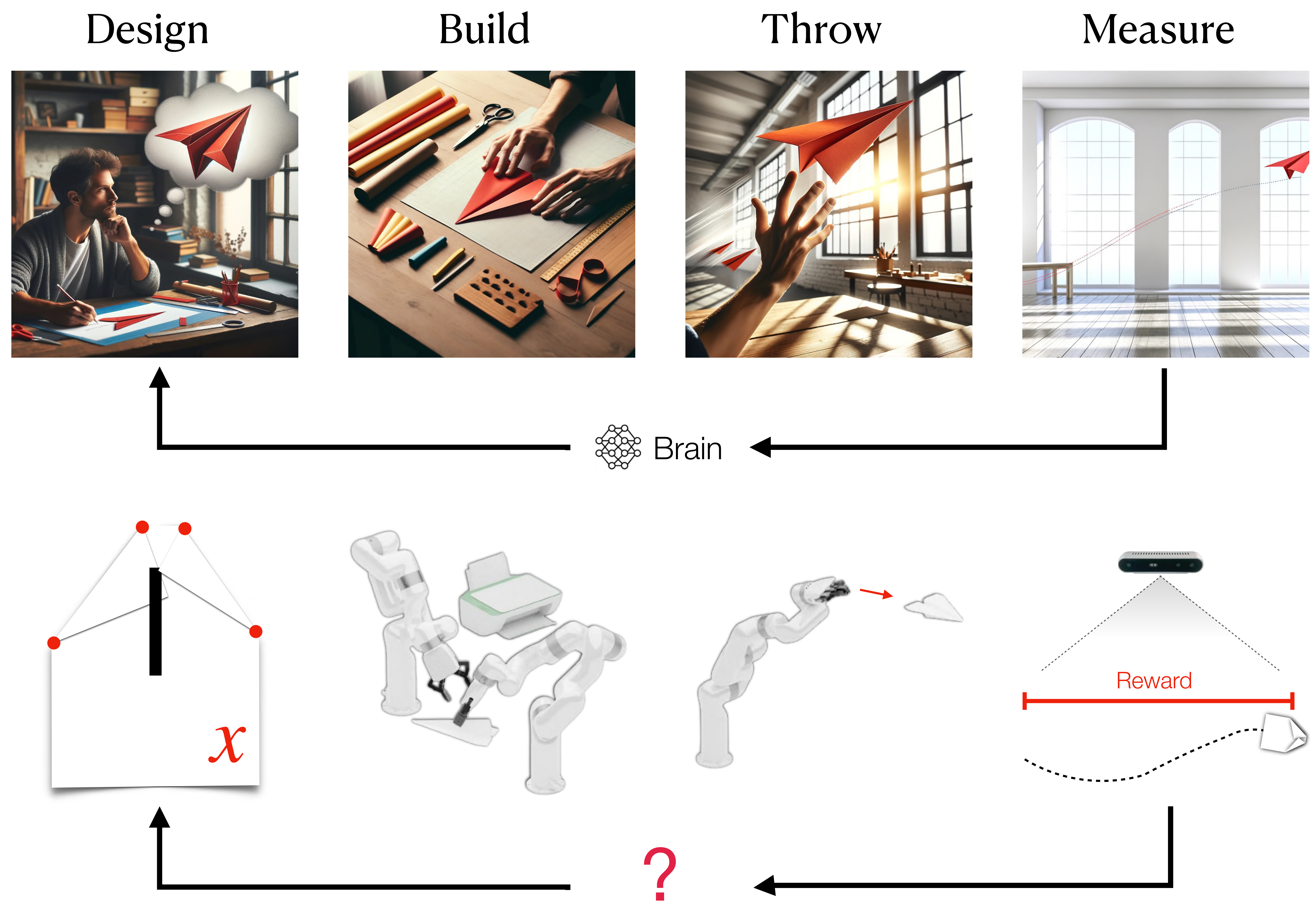
Measure

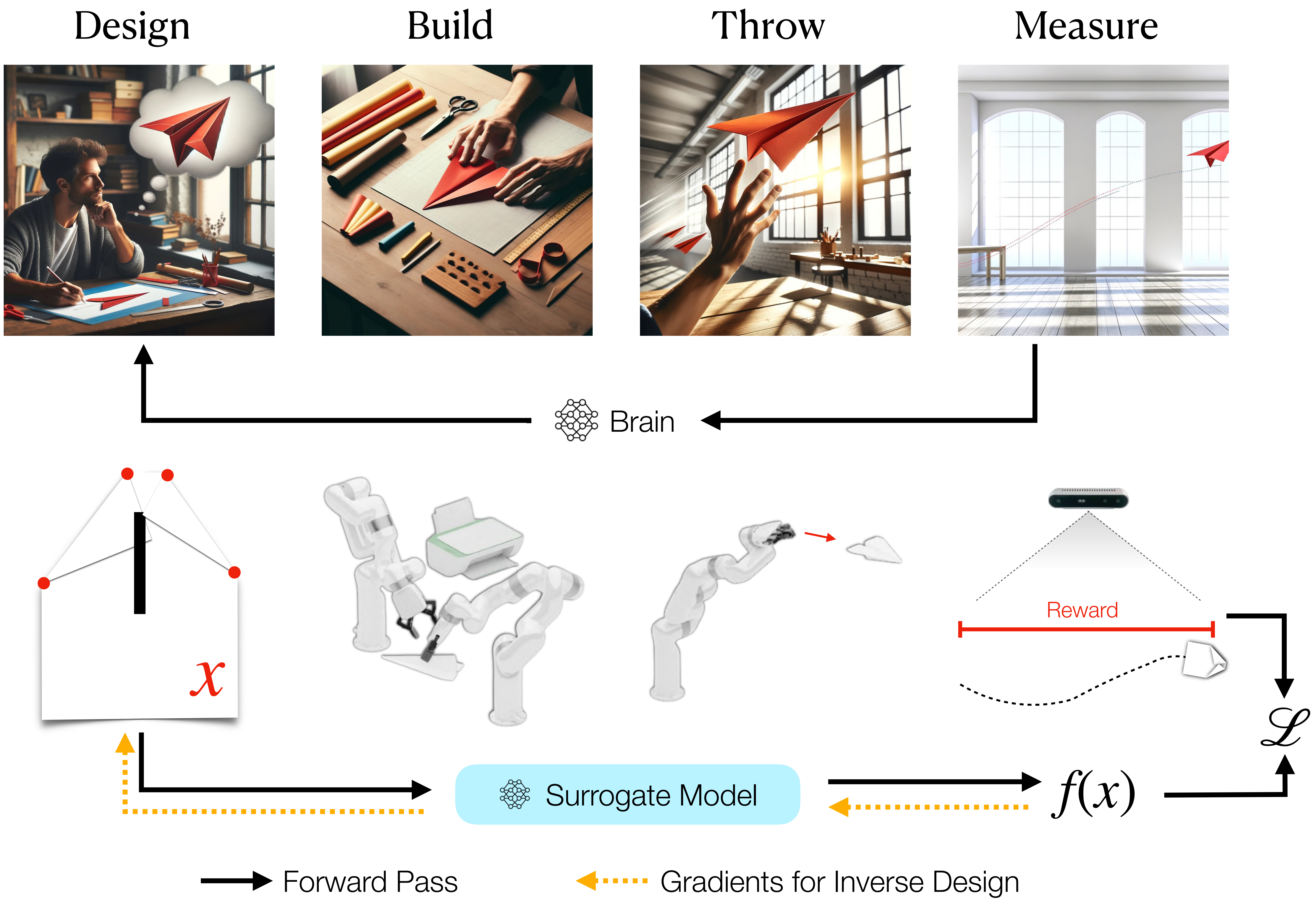

Design
Build
Throw
Measure
Brain


Design
Build
Throw
Measure
Brain
x
Reward

Automation

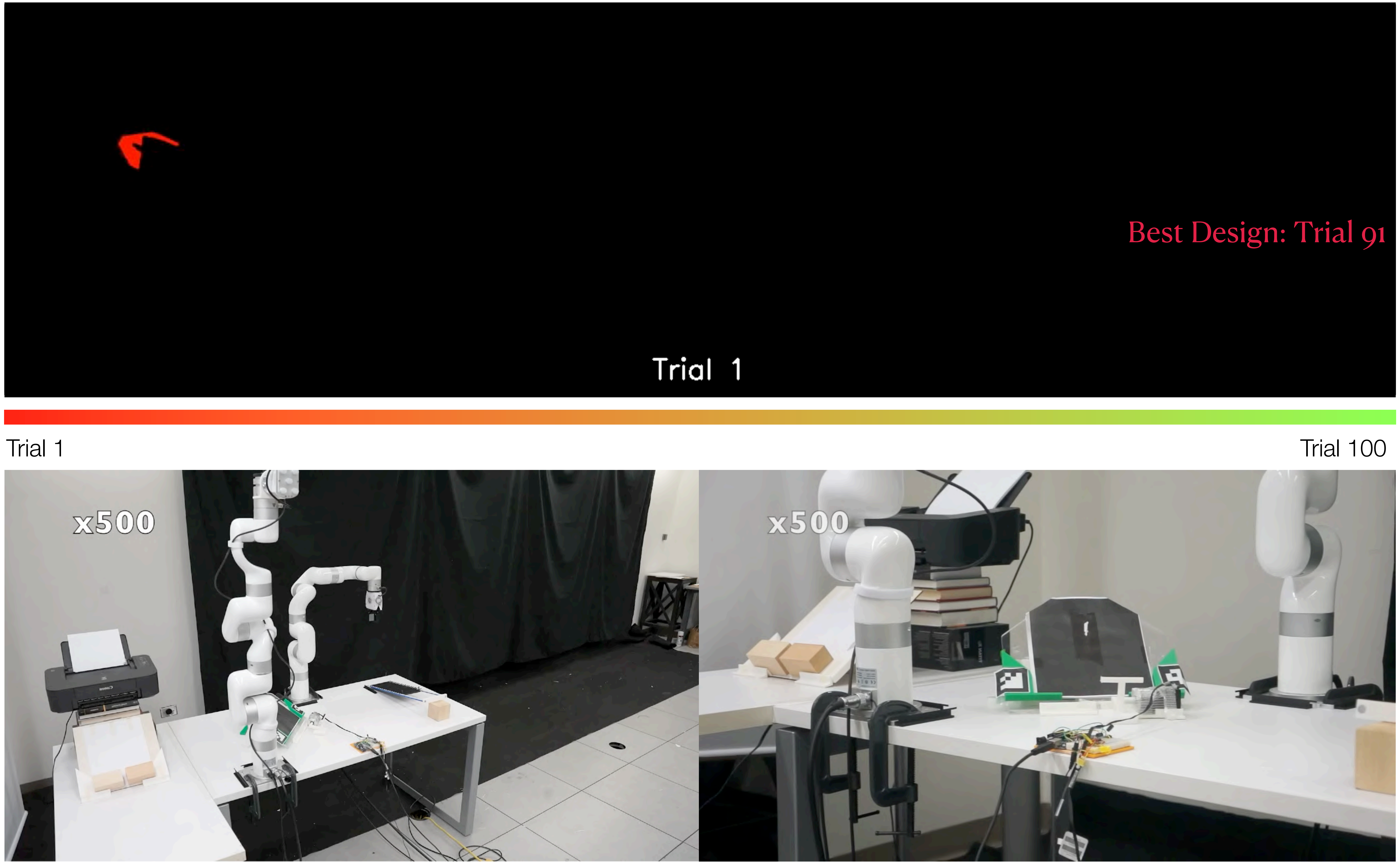



Design
Build
Throw
Measure
Brain
x
Reward
?



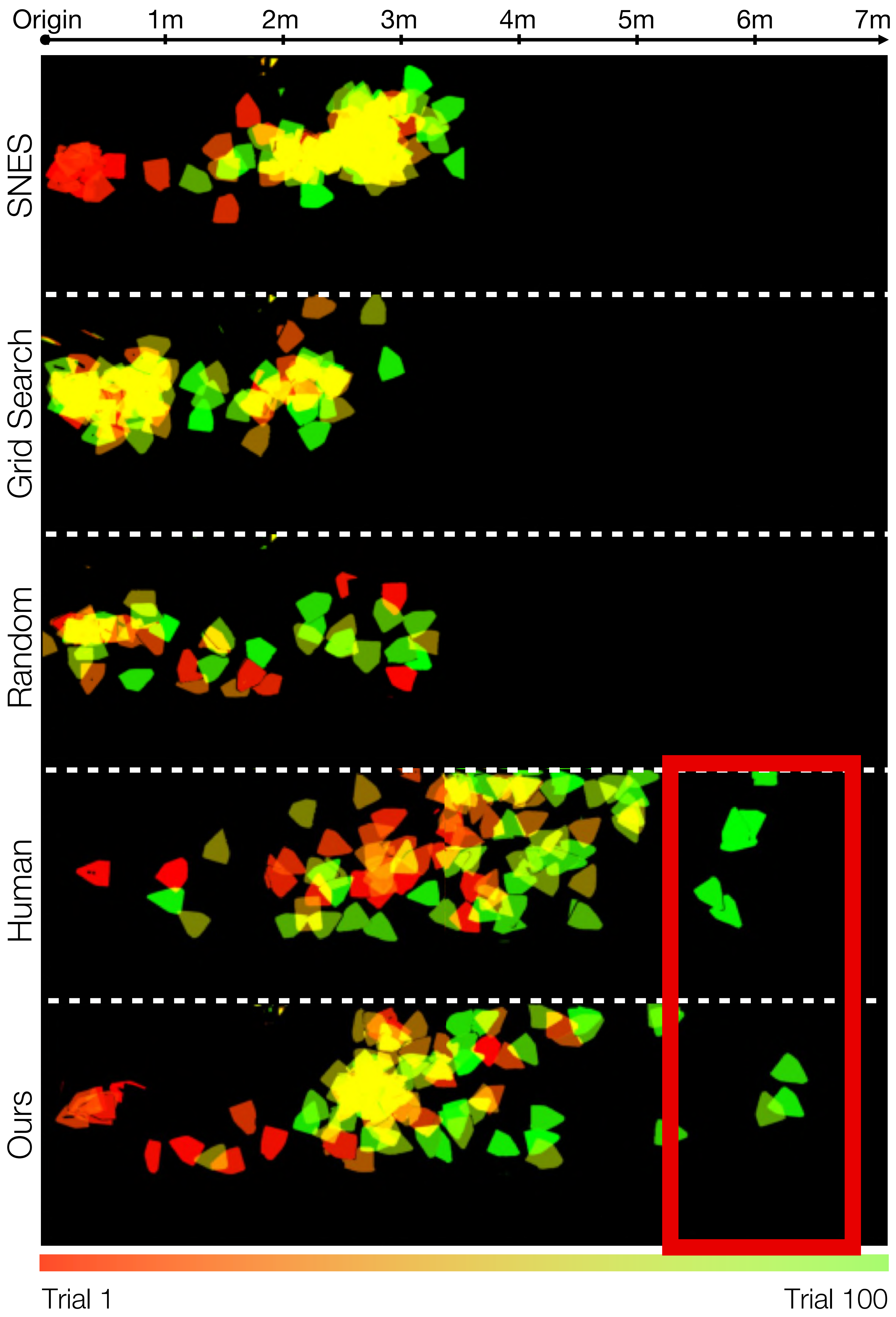
Liu, et al. In submission 2024

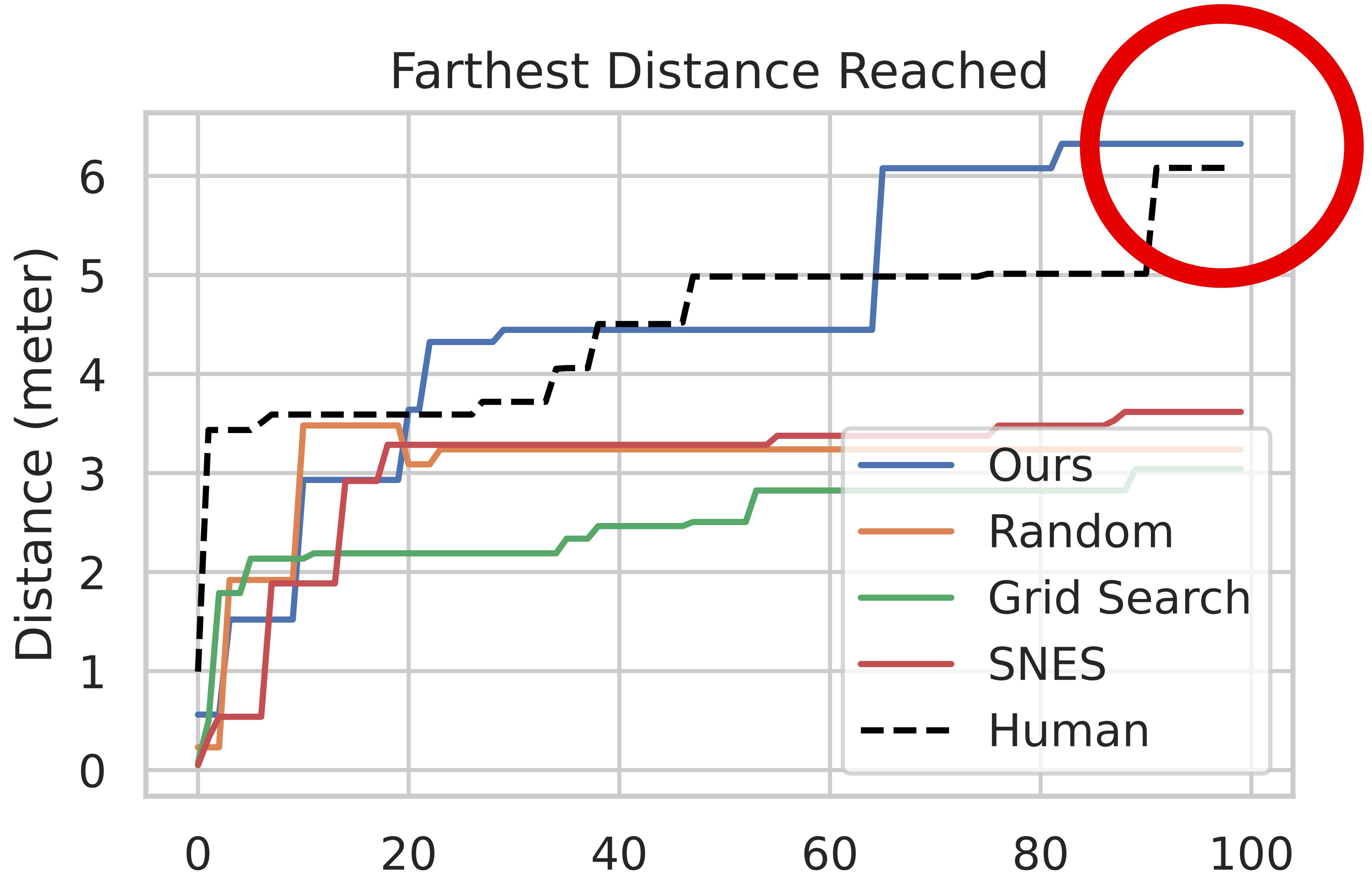
Learning



Liu, et al. In submission 2024

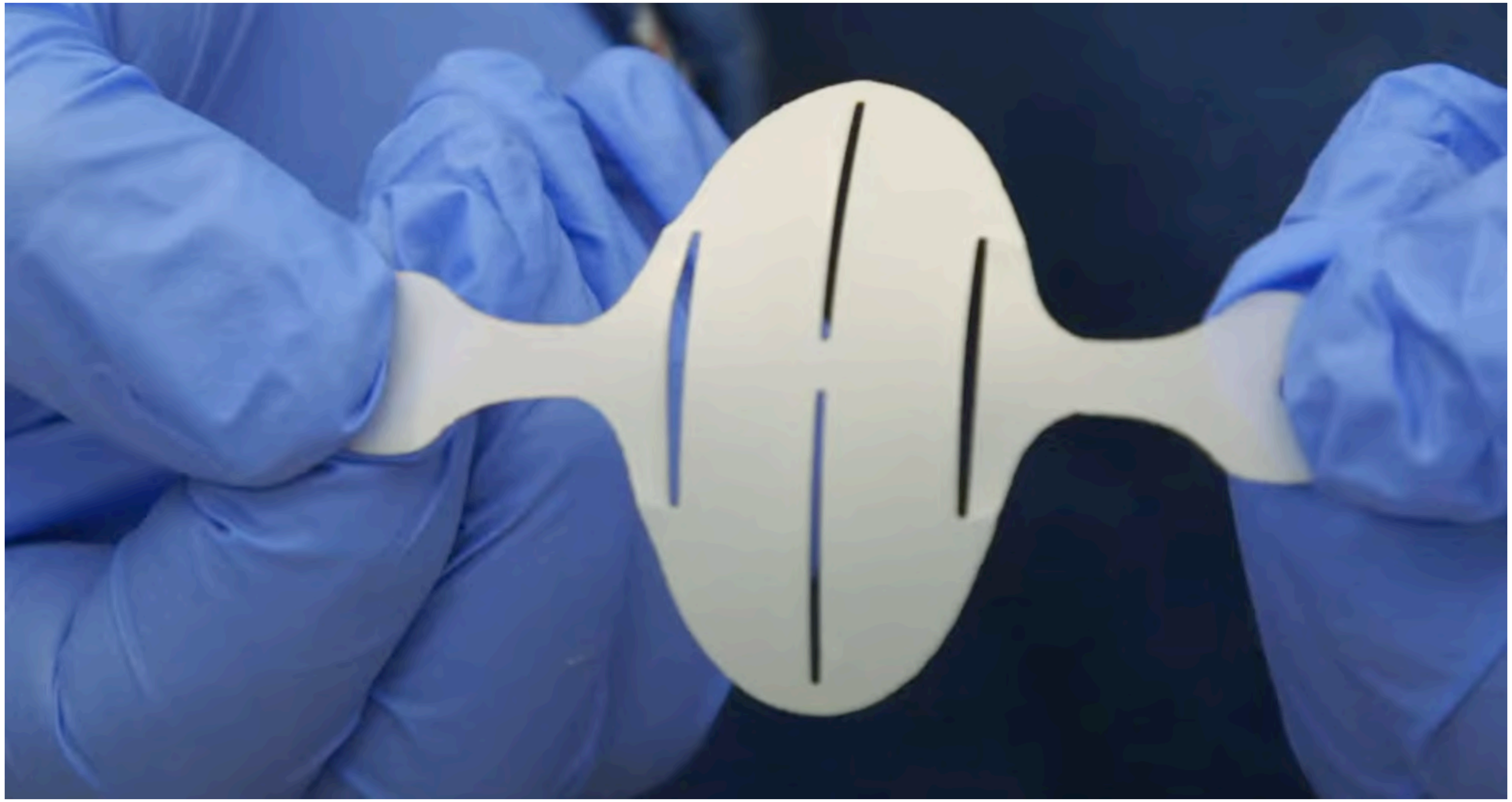
Comparison Against Baselines





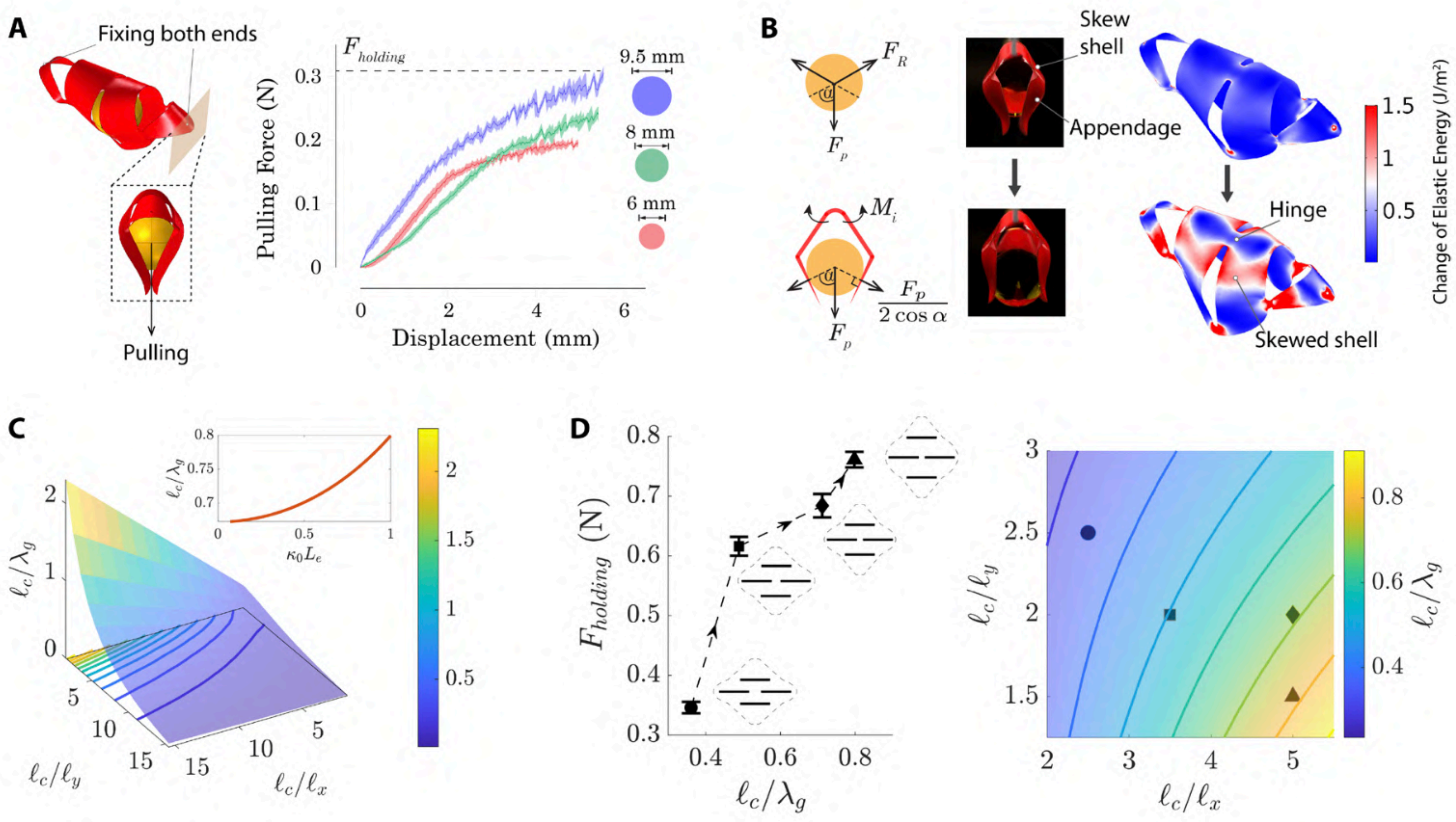
Liu, et al. In submission 2024

Case Study 2: Kirigami Gripper



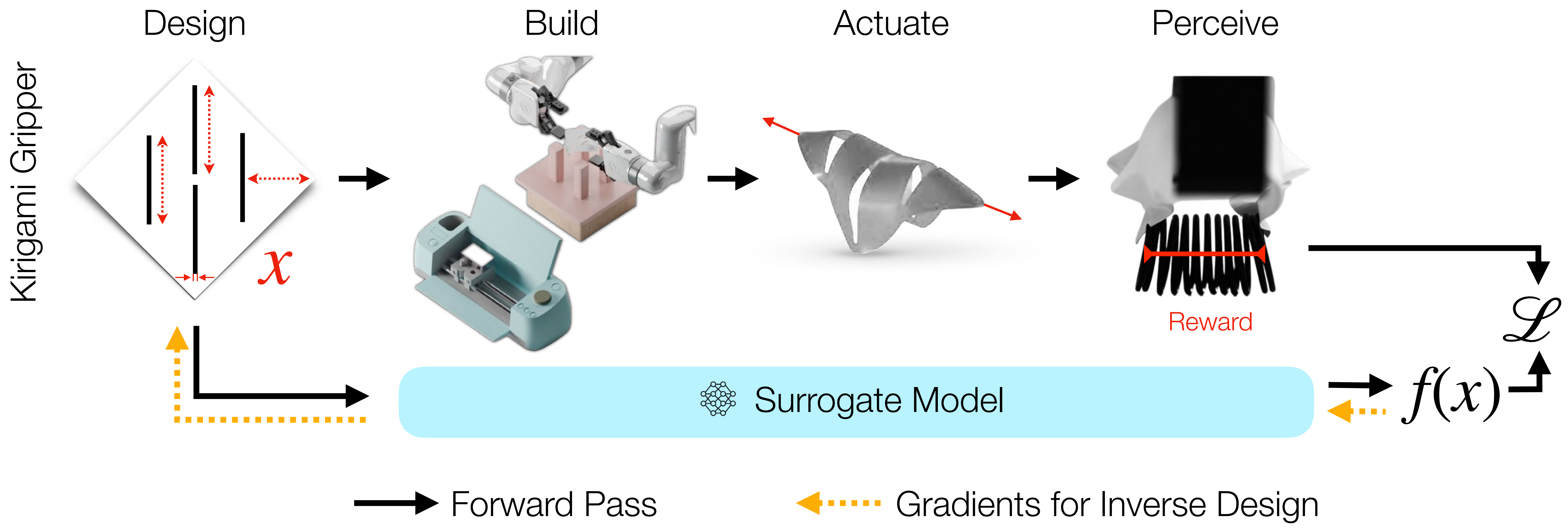
Grasping with kirigami shells. Yi Yang, Katherine Vella, Douglas P. Holmes
Science Robotics 2021

Prior Work Relies on Physical Analysis



Task 2: Kirigami Gripper

Learning to cut paper into grippers that exert maximum gripping force



Liu, et al. In submission 2024

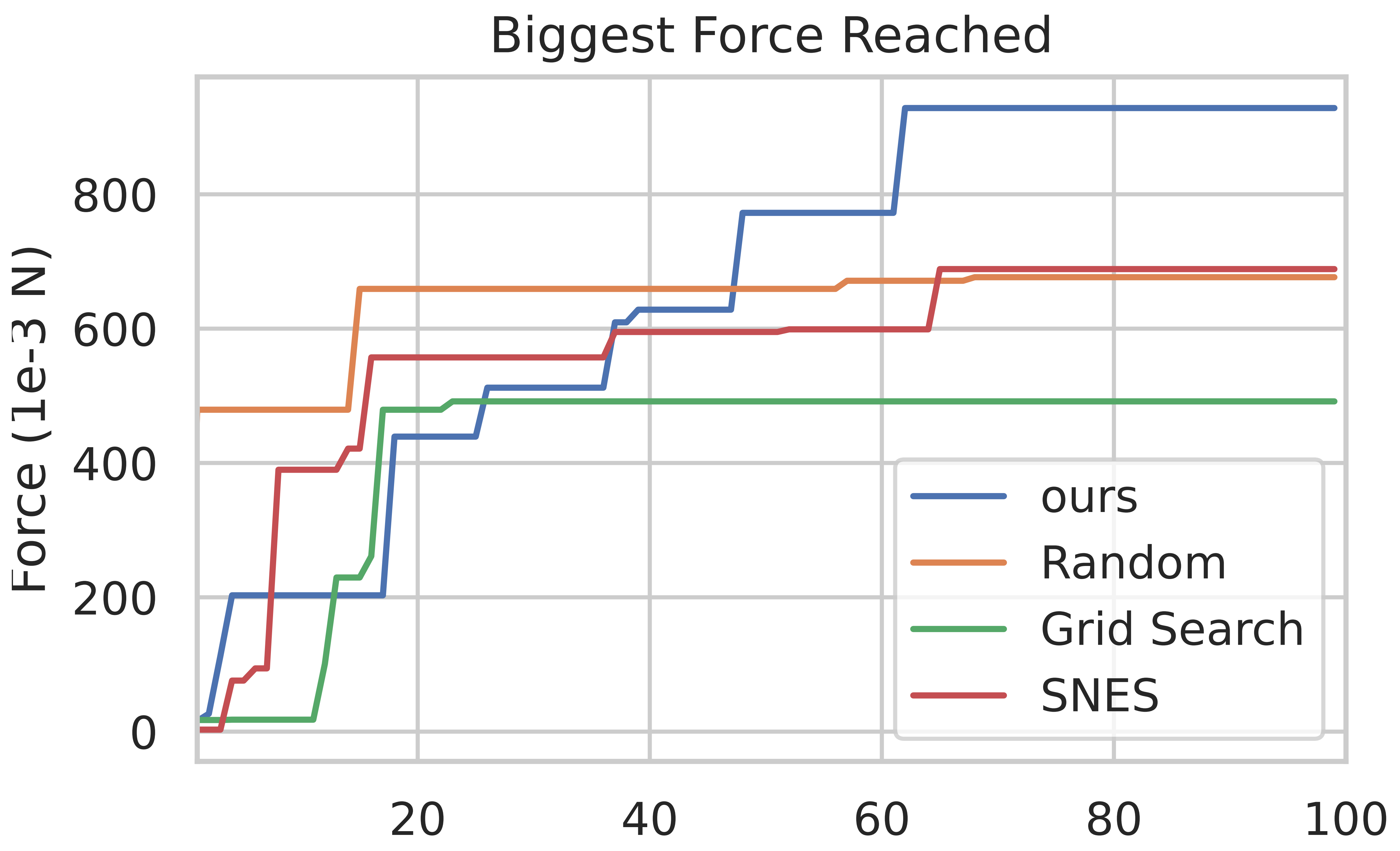
System Setup

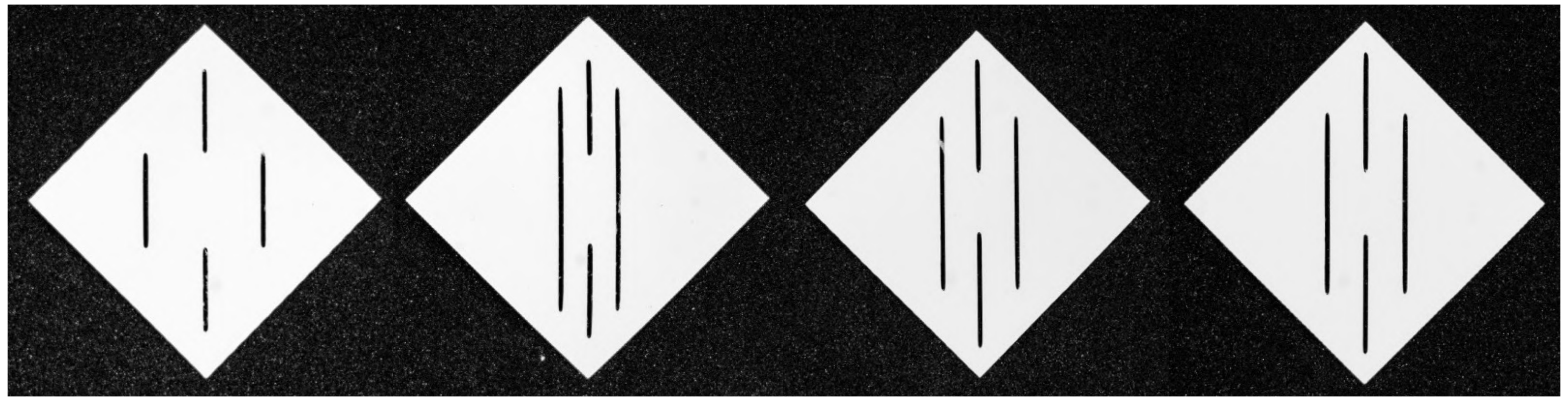
Cricut Maker 3

Load Cells

Liu, et al. In submission 2024

Experiments

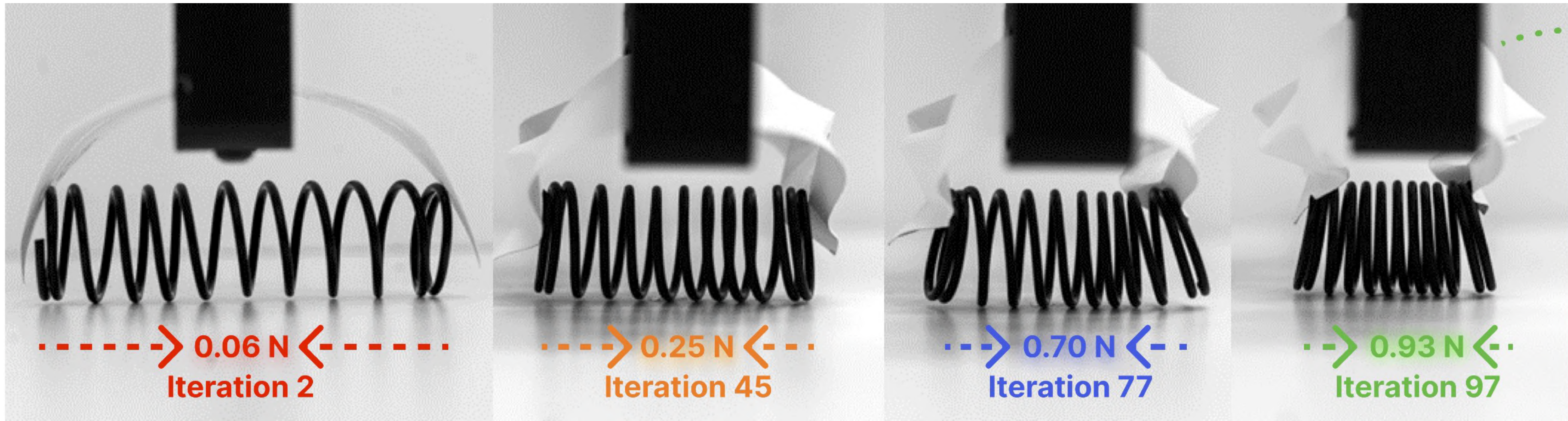




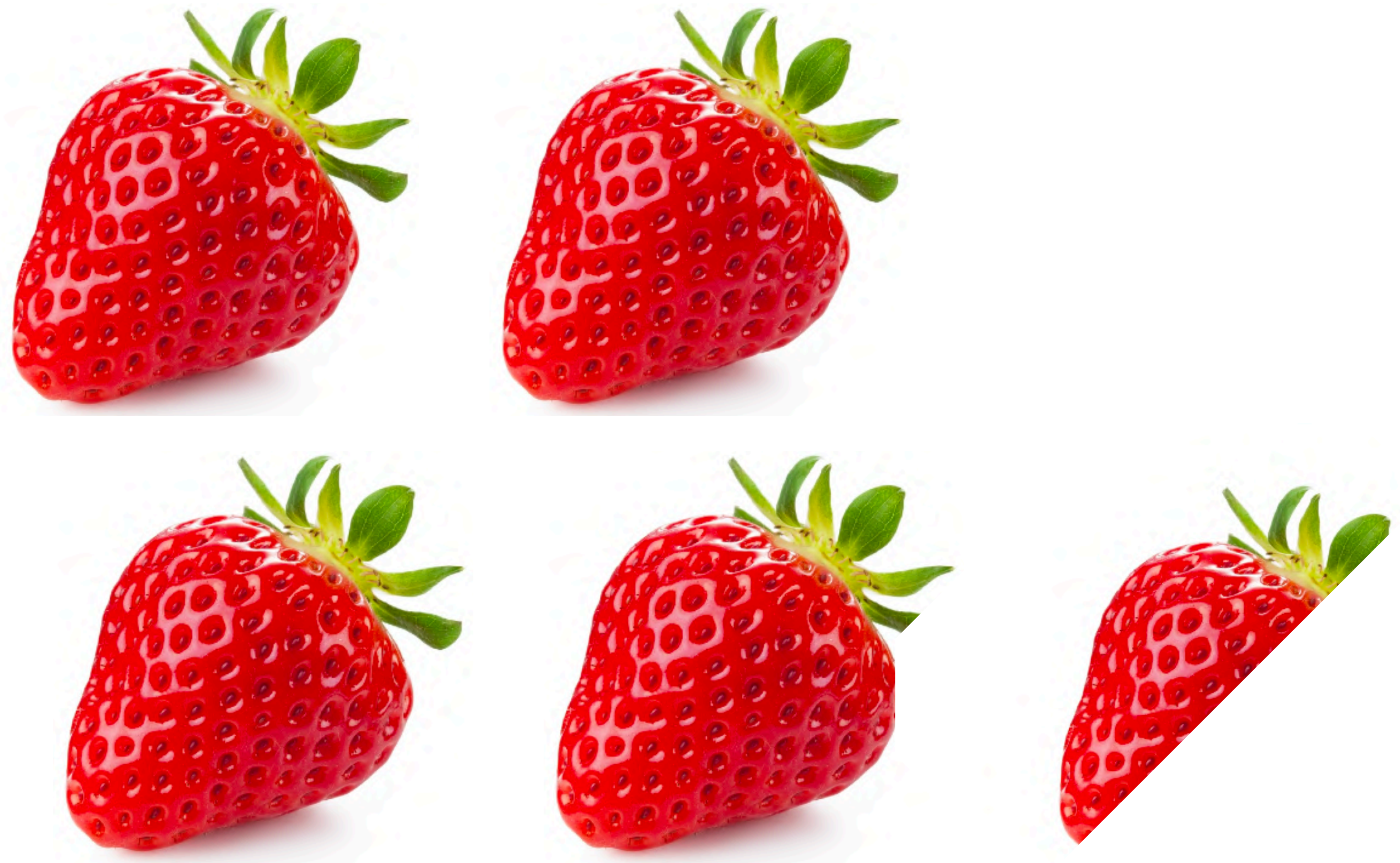
Iteration 2 Iteration 45 Iteration 77 Iteration 97

Liu, et al. In submission 2024

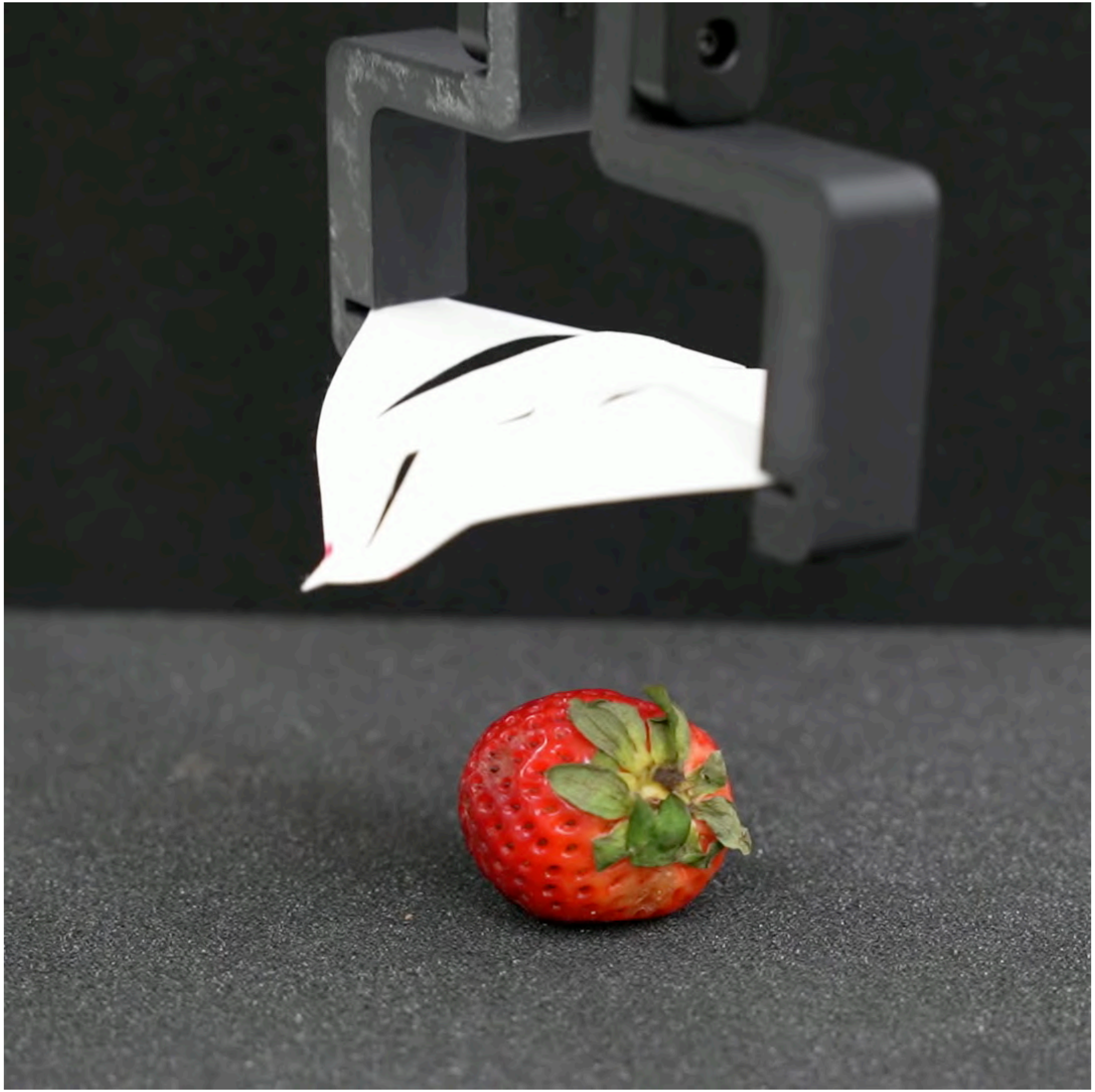
Results



≈



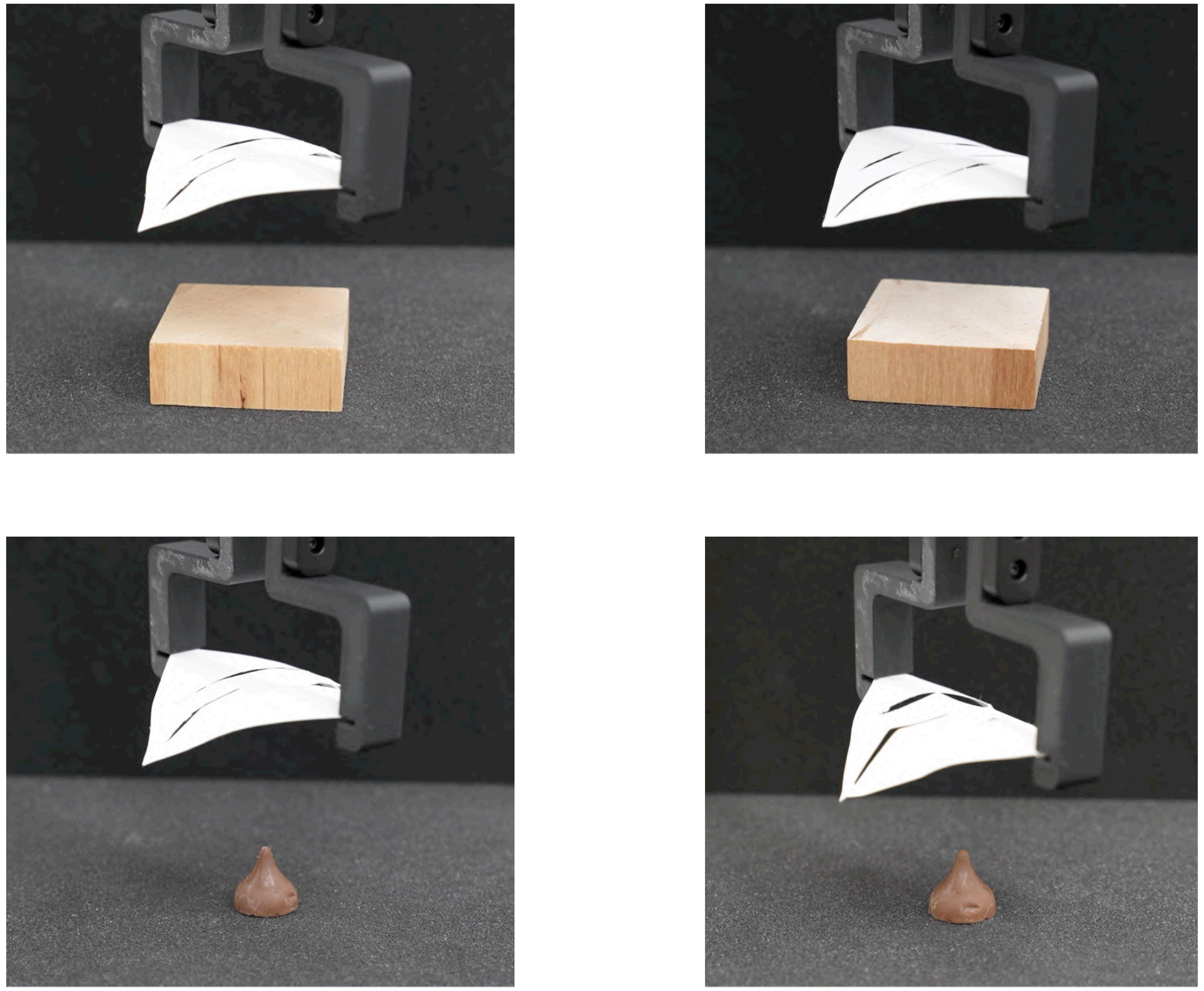
Liu, et al. In submission 2024




BURT'S BEES

Adaptation

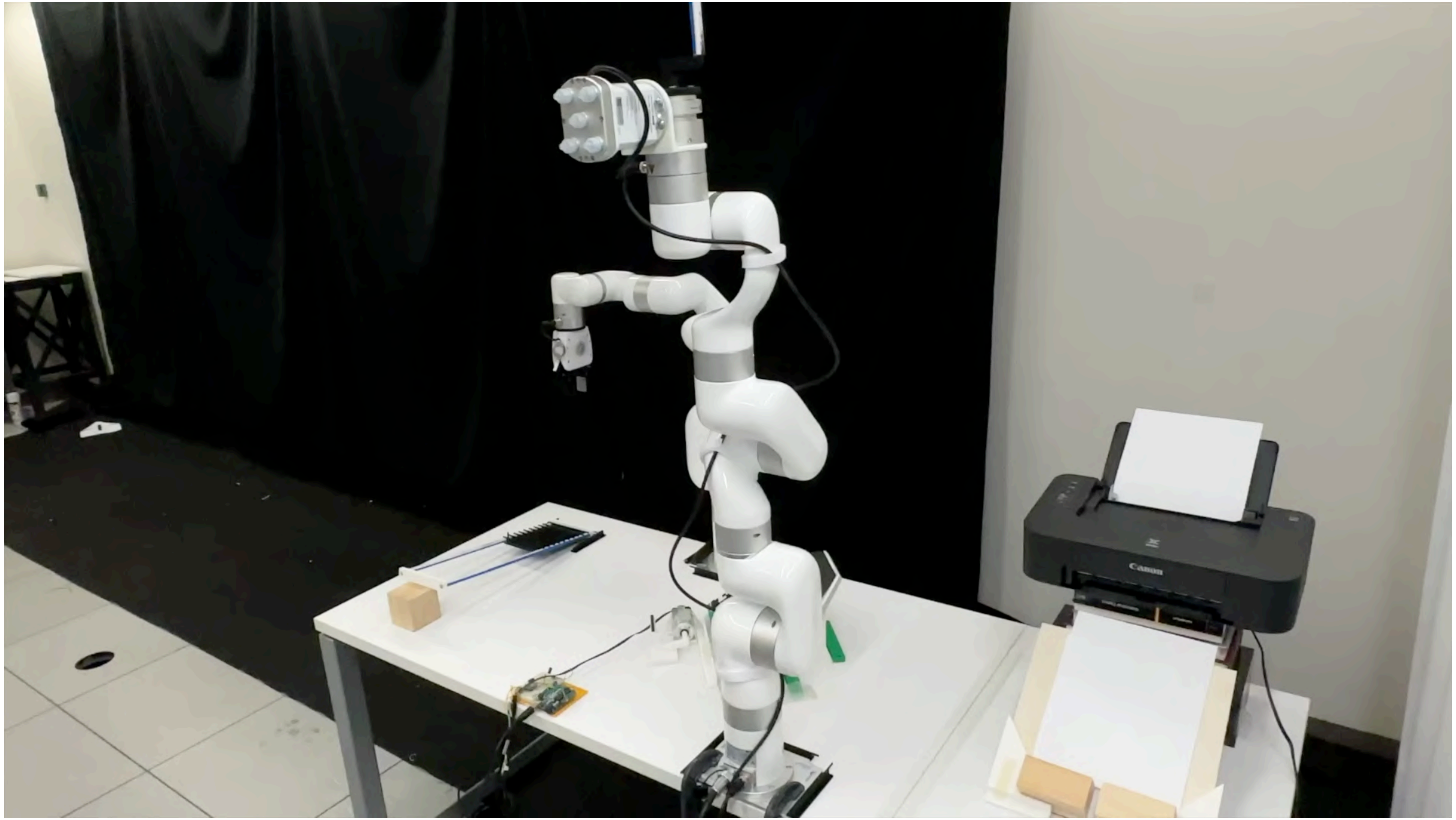
	Small (1.5cm)	Large (8cm)
Original (optimized for 5cm)	0.302 ± 0.012	0.055 ± 0.027
Adapted Gripper	0.442 ± 0.080	1.131 ± 0.235

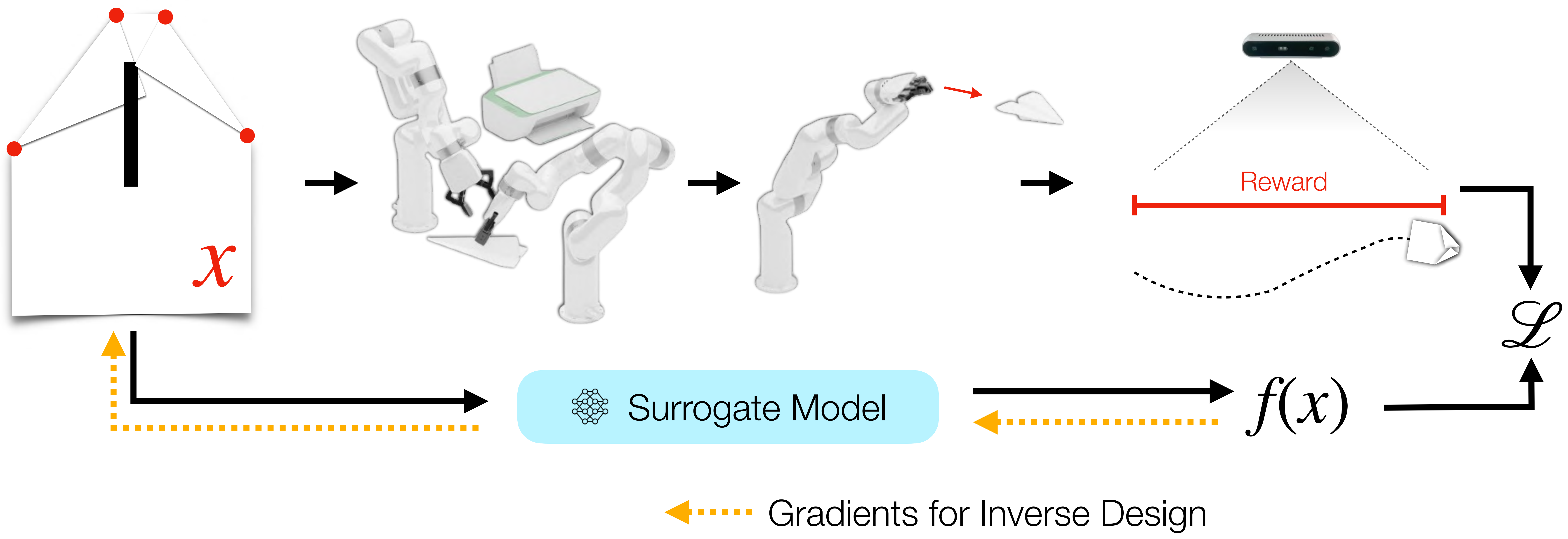


Before Adaptation After Adaptation

Liu, et al. In submission 2024

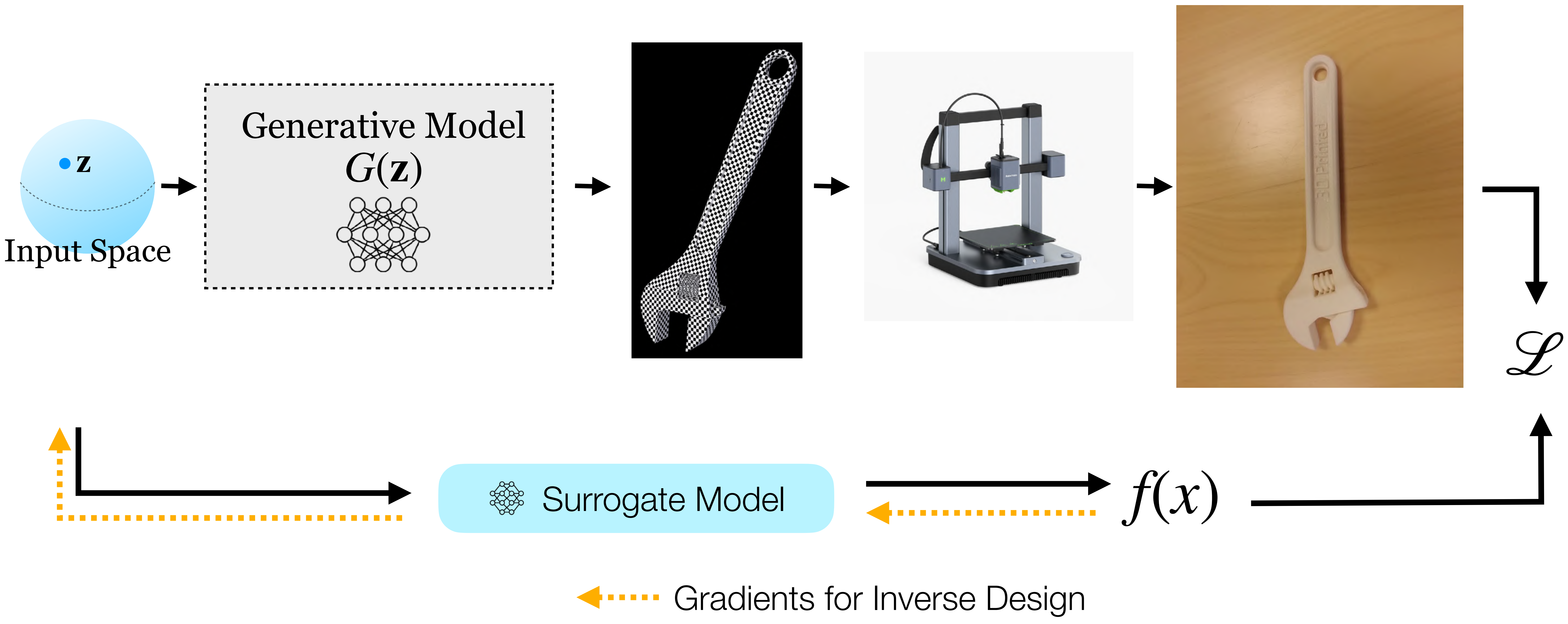
paperbot.cs.columbia.edu



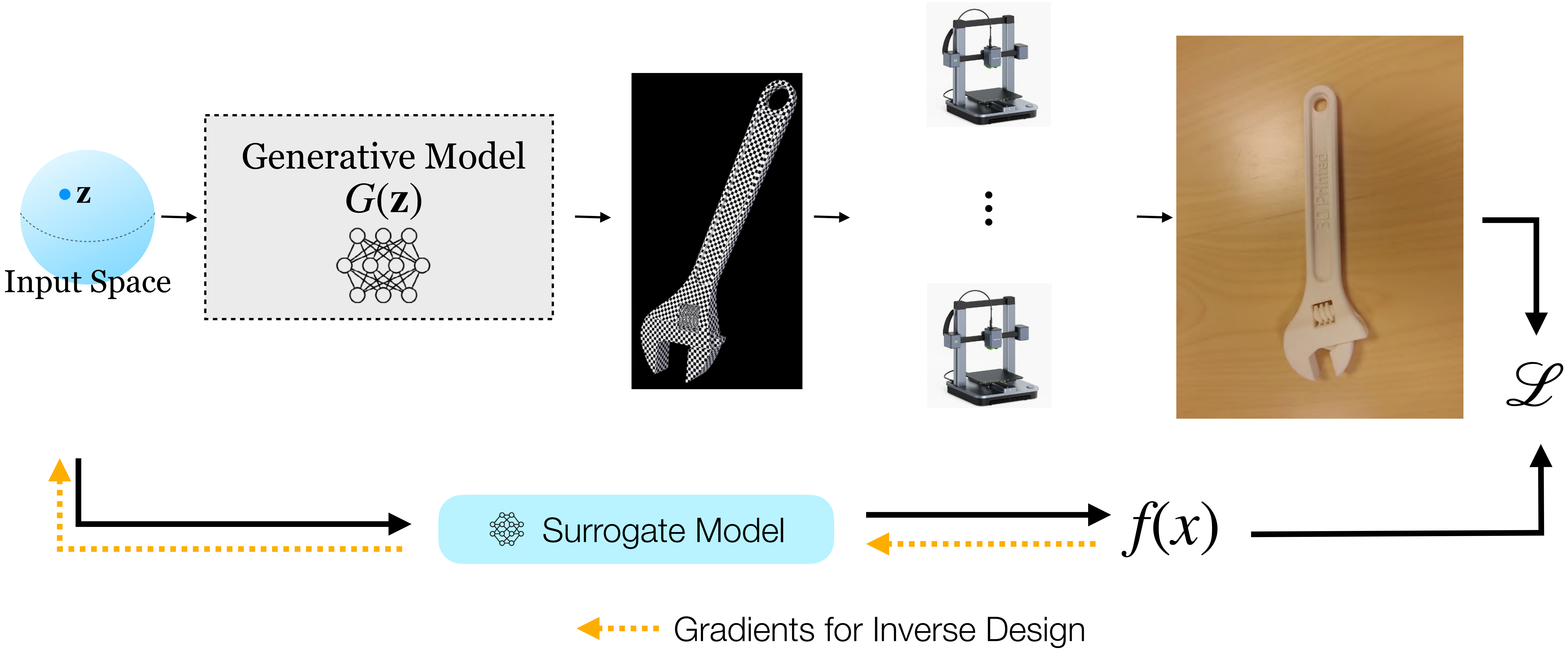


x
Reward
$\mathscr{L}$
Surrogate Model
$f(x)$
Gradients for Inverse Design

3D Generation for Tool Design

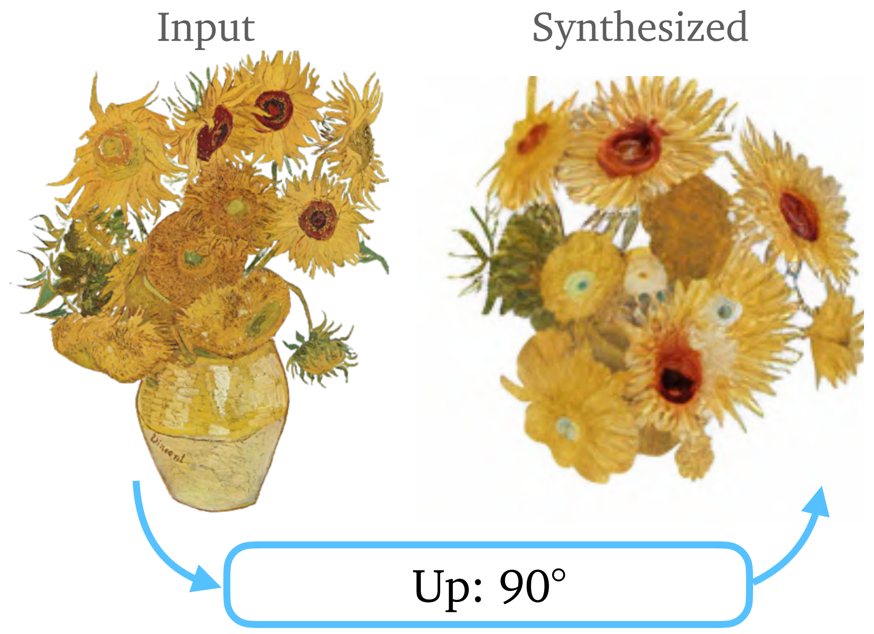


3D Generation for Tool Design



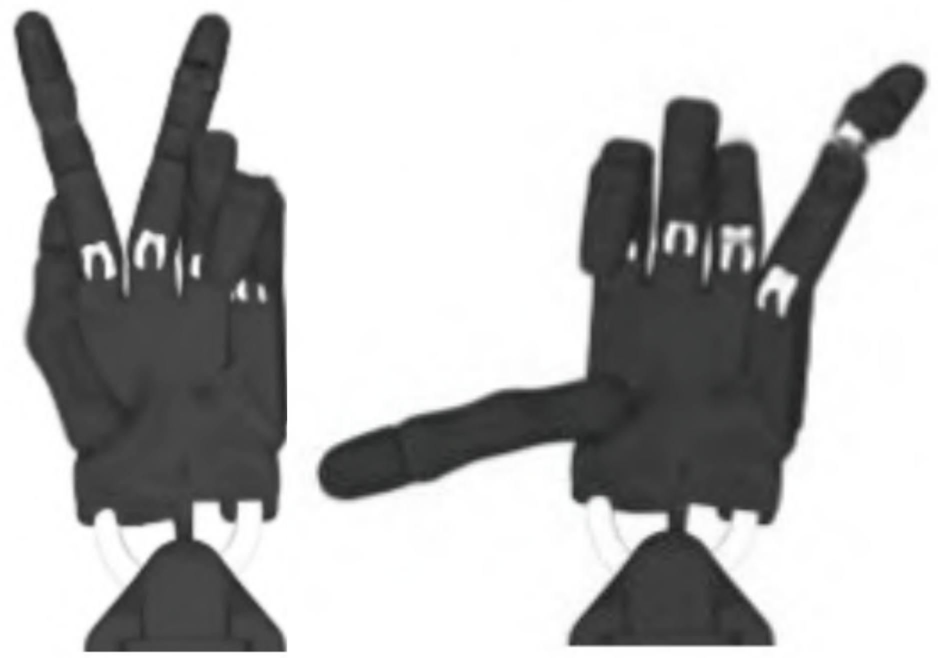
Generative Embodied AI

3D Generation

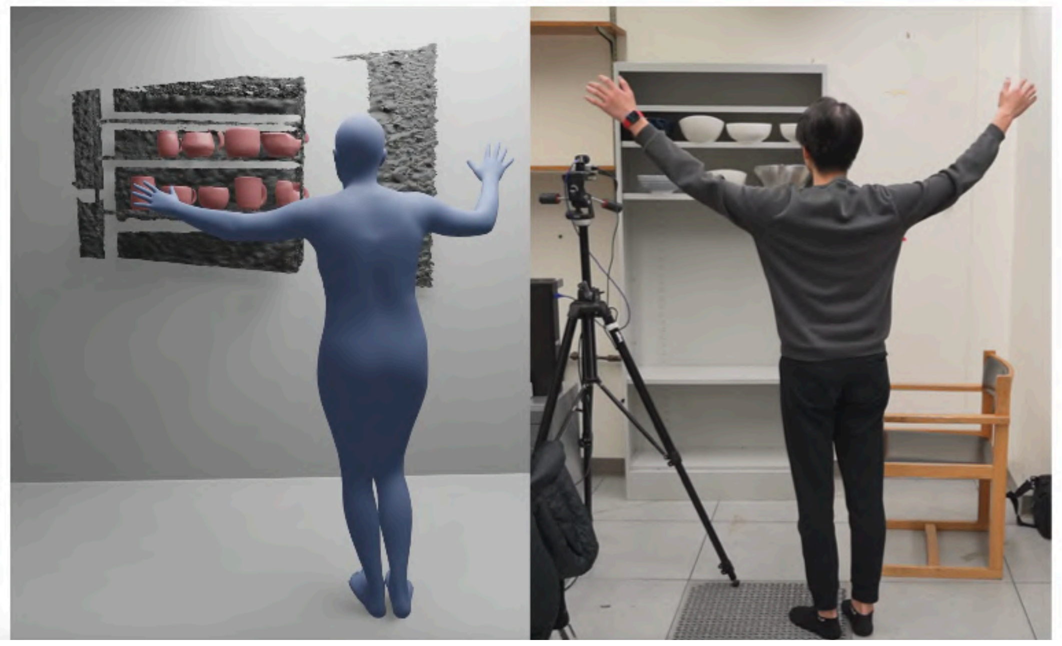


Zero123
ICCV 2023
Liu et al.

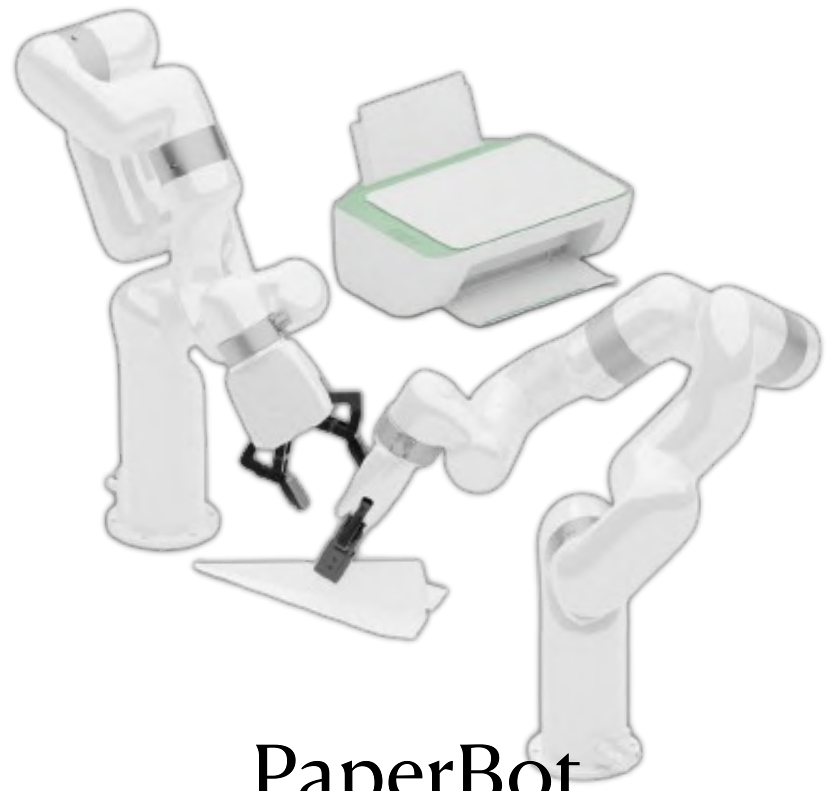
Objaverse-XL
NeurIPS 2023
Deitke. **Liu** et al.



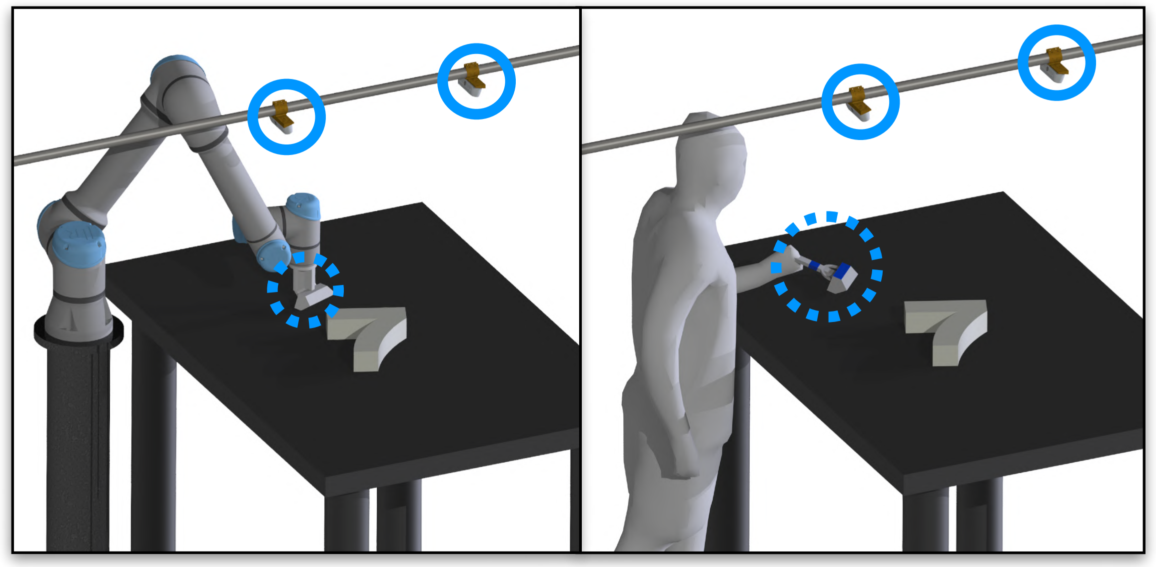
Dr. Robot
CoRL 2024
Liu*. Canberk* et al.



Humans as Light Bulbs
CVPR 2023
Liu et al.



PaperBot
RSS 2024
Liu et al.



Dreamitate
CoRL 2024
Liang*, **Liu*** et al.

Physical Reconstruction

Physical Design

Physical Interaction