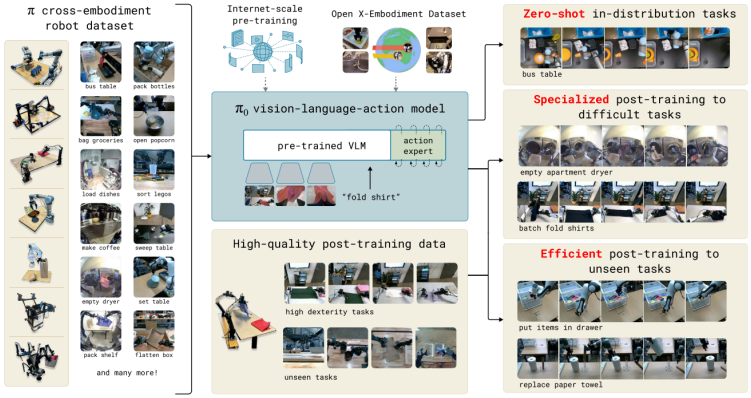
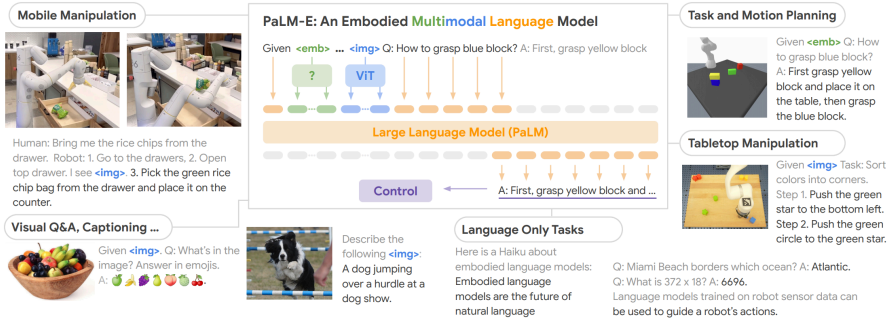
RobotArena ∞ : Scalable Robot Benchmarking via Real-to-Sim Translation

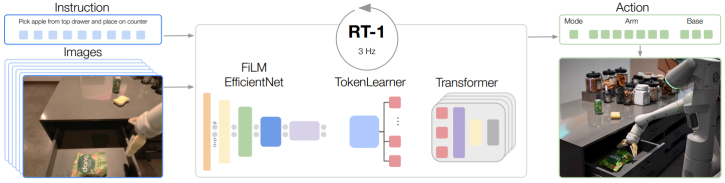
Robot Foundational Policies Need General Scalable Benchmarks



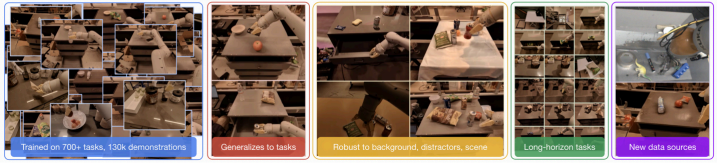
π0: A Vision-Language-Action Flow Model for General Robot Control



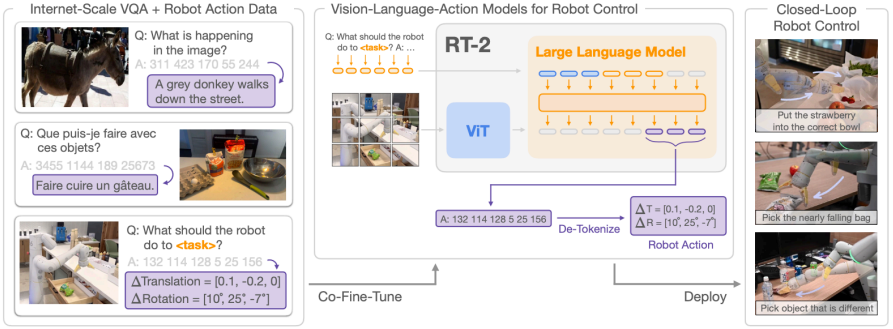
PALM-E, Driess et al., March 2023



(a) RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).

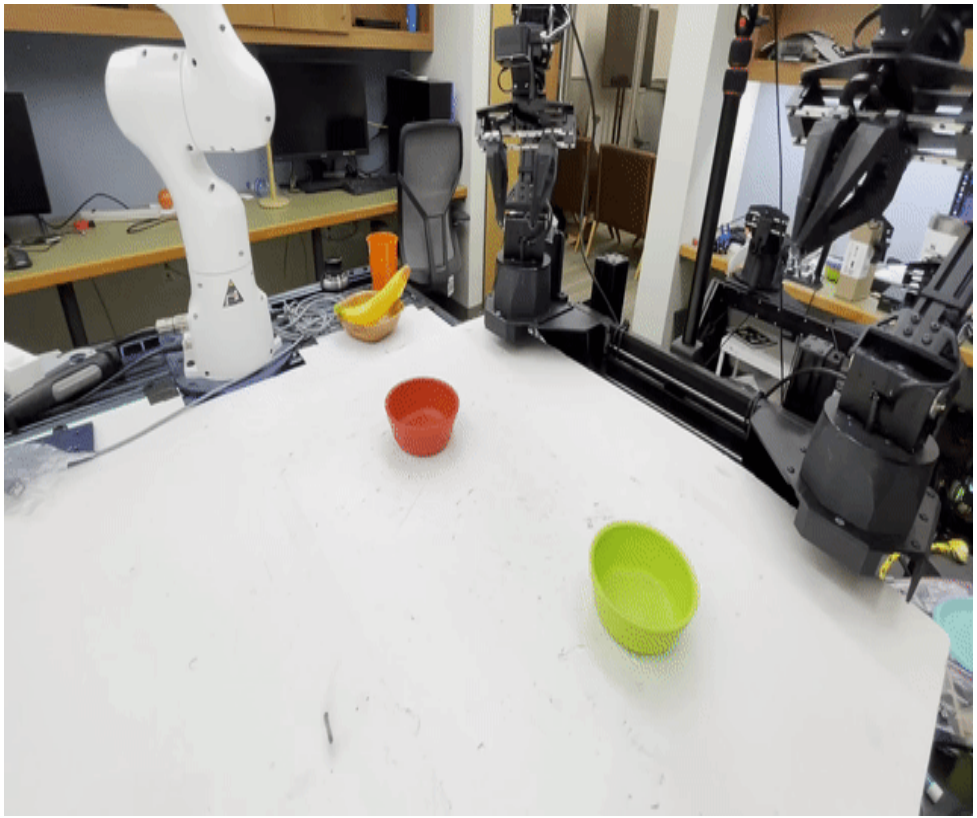


RT-1, Brohan et al., Dec. 2022



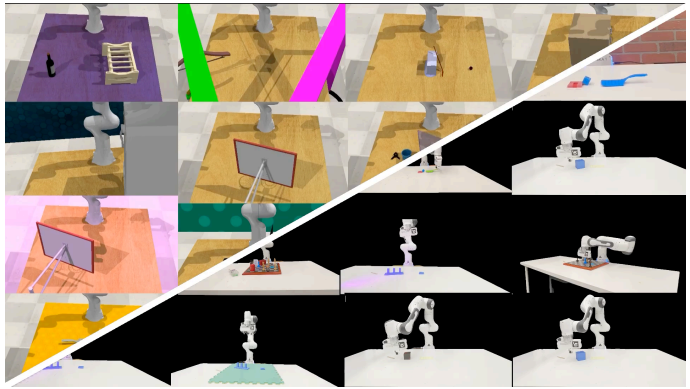
RT-2, Brohan et al., July. 2023

Evaluating robot policies in the real world does not scale

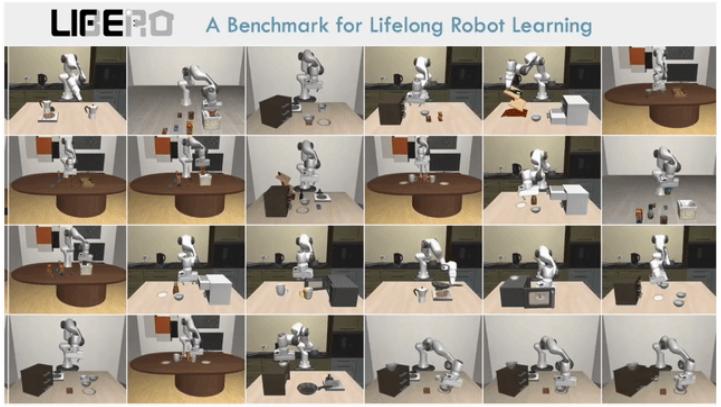


Simulation Benchmarks: Train and Test in Simulation

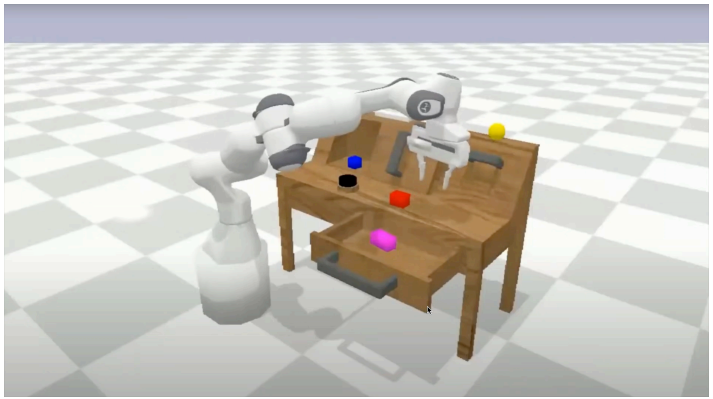
RL Bench



Colosseum

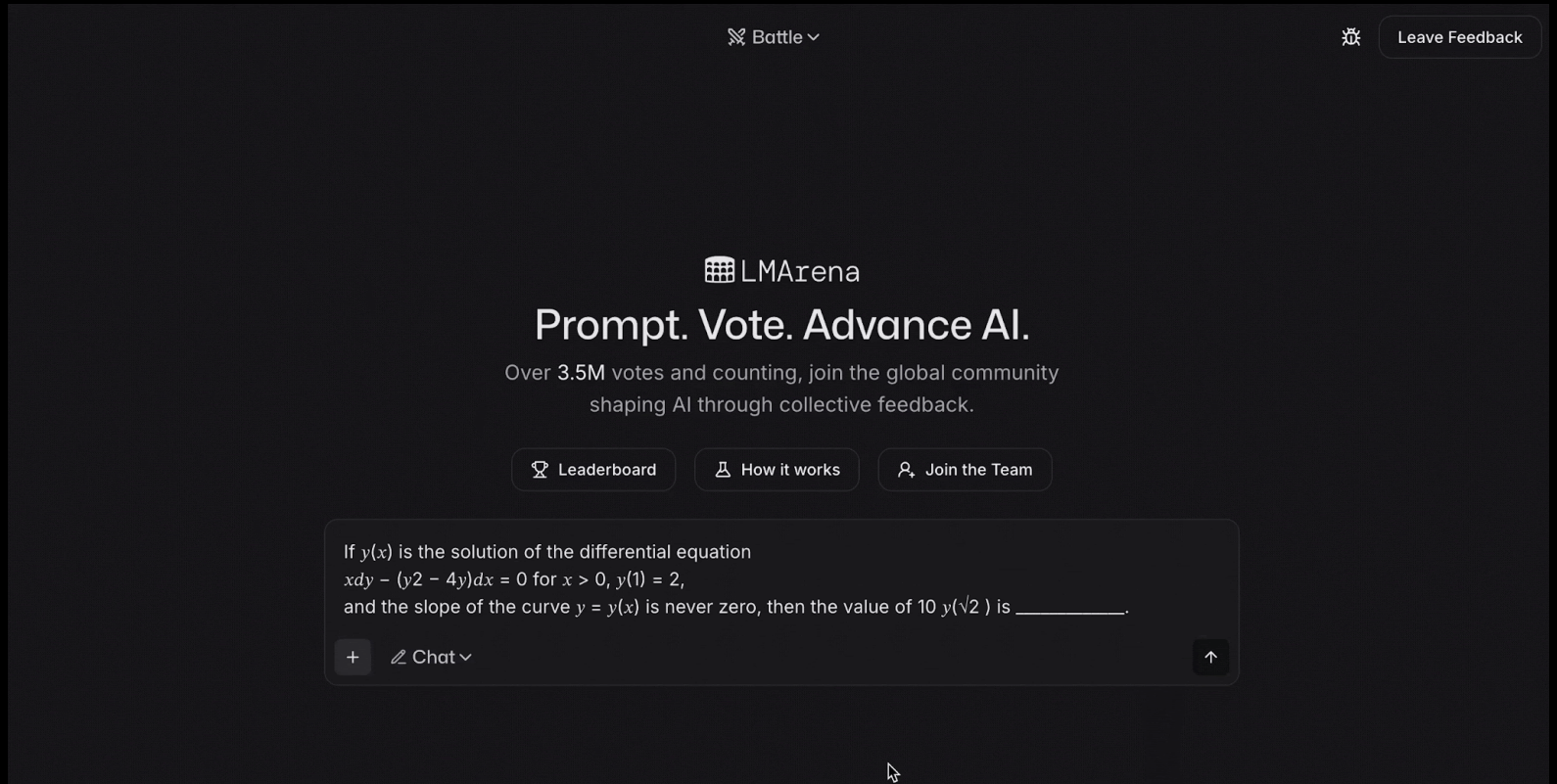


LIBERO



Calvin

LMArena: Evaluating LLMs trained on ANYTHING



Can we similarly benchmark robot policies trained on *anything*?

RobotArena ∞: Unlimited Robot Benchmarking via Real-to-Sim Translation

Robotic Datasets

DROID

"Make the cup lie down"

Bridge V2

"Put the tomato in the pot"

RH20T

"Plug the charger into the socket"

⋮

User-Recorded Video

User

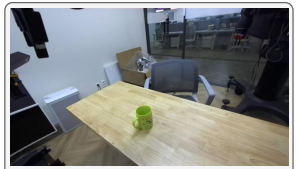
…

Real2Sim

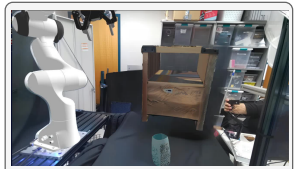
Large Scale Simulated Environments

Automated Environment Creation from Real Video for VLA Evaluation

Close a sliding window



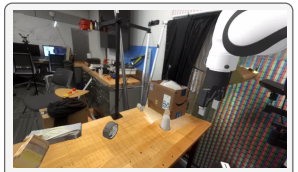
Move the green cup to the right



put the mug in the shelf

Clean the tabletop with a sponge

ROBOTARENA ∞



Move the grey tape to the right

put the knife on to the plate

Use cup to pour something granular

Put the tomato in the pot

Automatic Policy Evaluation

Policy	VLM Scores		
π_1	79.32	77.03	…
π_2	76.38	77.50	…
π_3	83.77	77.18	…

⋮

Base Env | Under Controlled Perturbation

Human Preference Evaluation

$\pi_4 > \pi_2$

$\pi_2 < \pi_1$

$\pi_3 > \pi_1$

⋮ ⋮

Pair-wise Preference

π_3

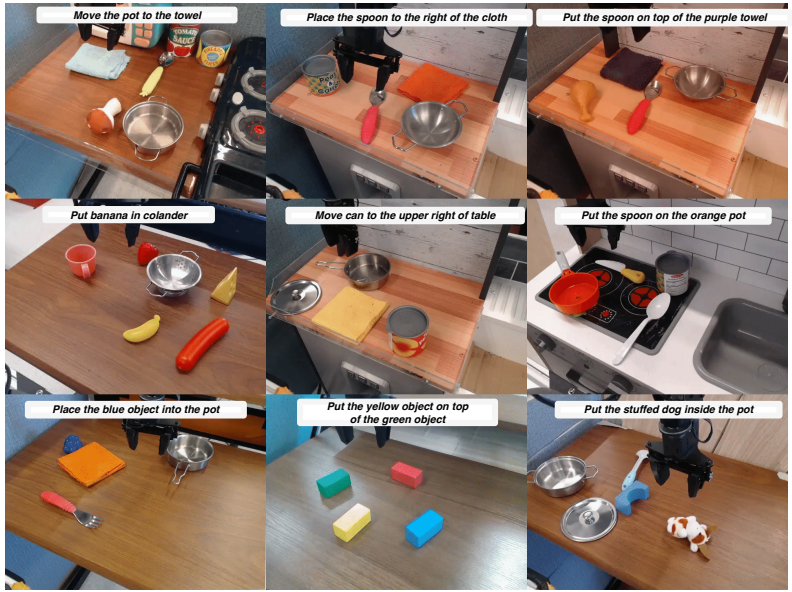
π_1

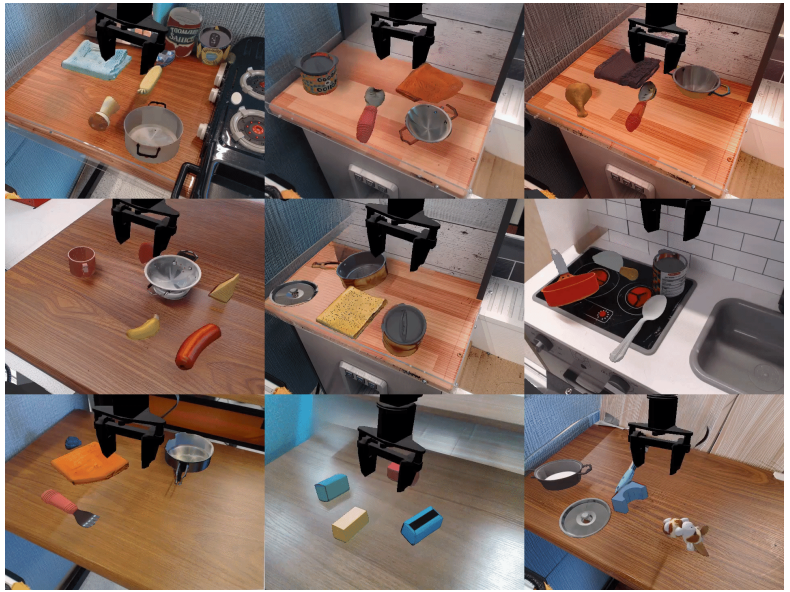
π_2

π_4

Policy Ranking

RobotArena ∞: Unlimited Robot Benchmarking via Real-to-Sim Translation



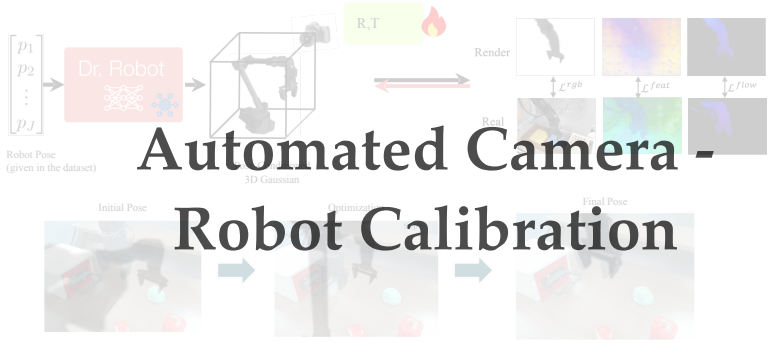


real-to-sim

No human involvement

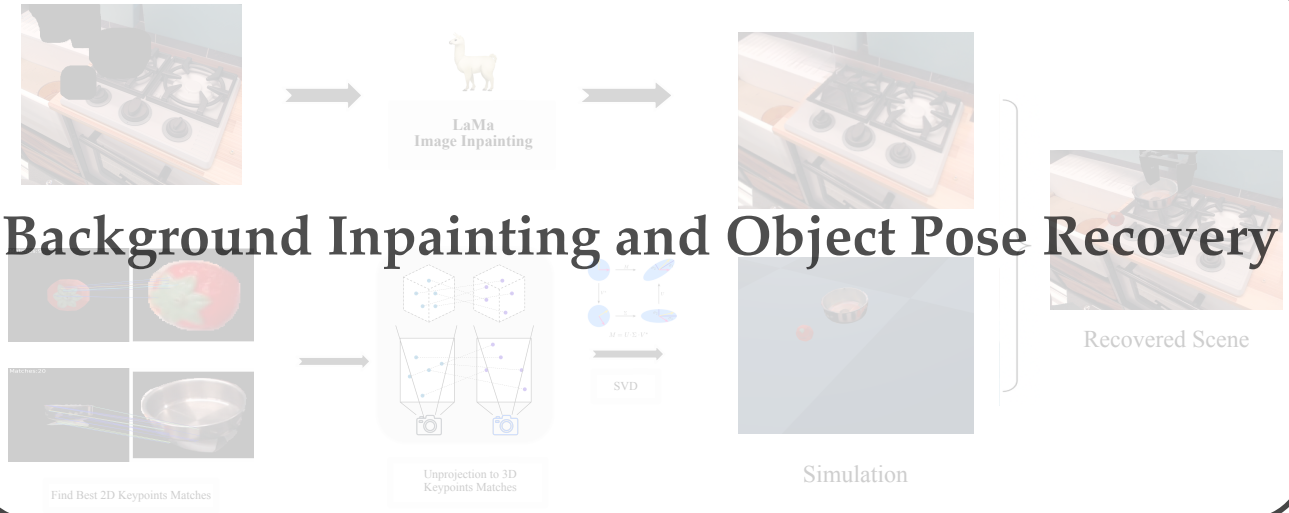
Real2Sim Translation

Automated Camera - Robot Calibration

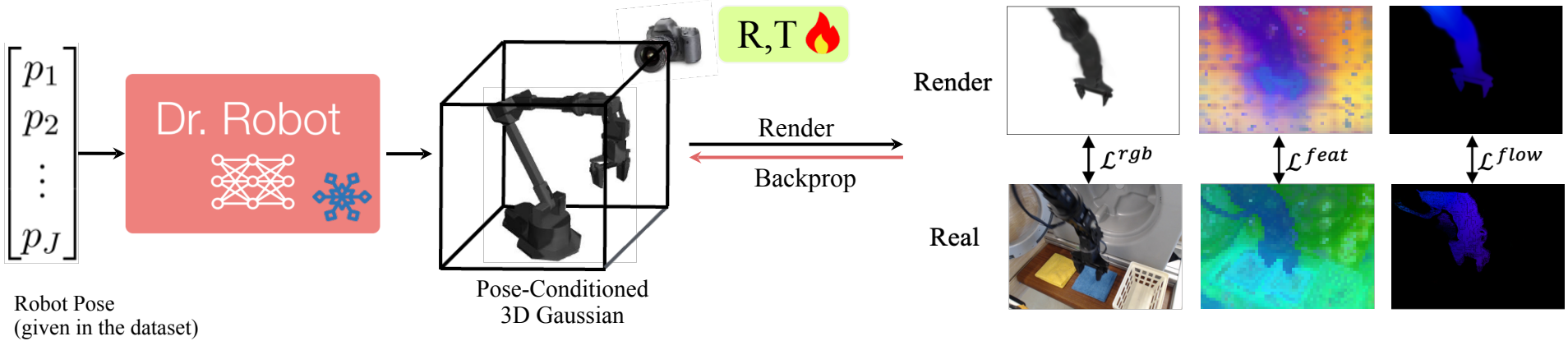


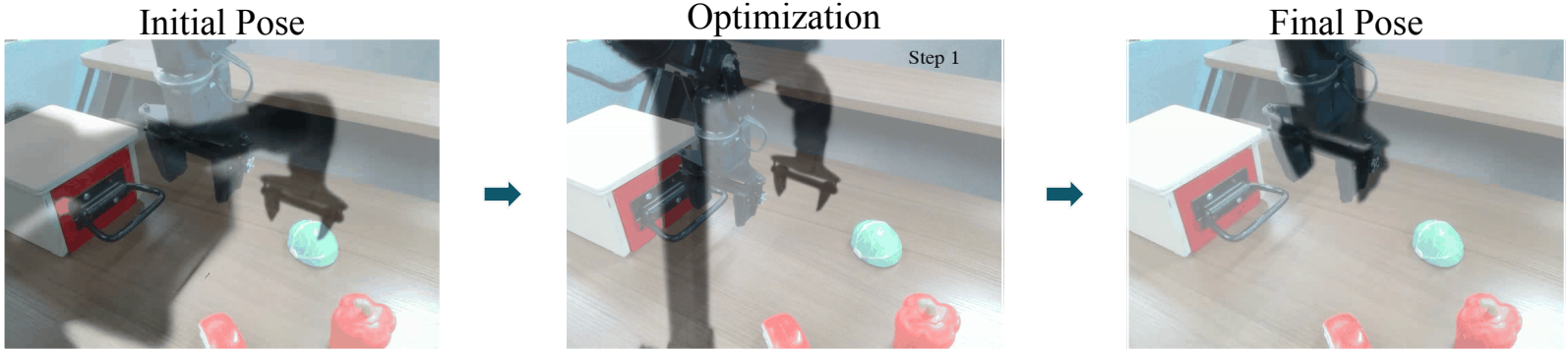
Relevant Object Reconstruction

Background Inpainting and Object Pose Recovery



Automated Camera - Robot Calibration

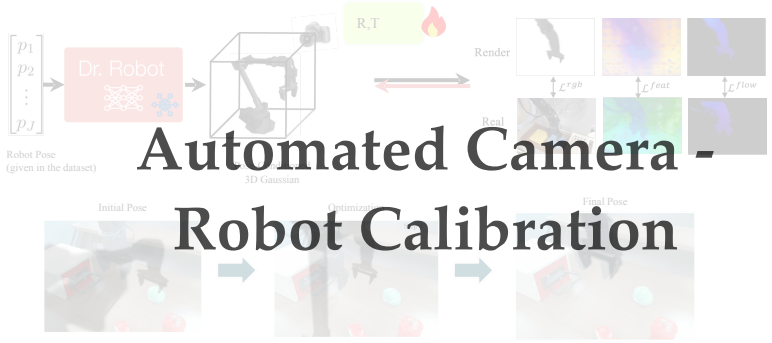




Liu et al, Differentiable Robot Rendering. CoRL 2024

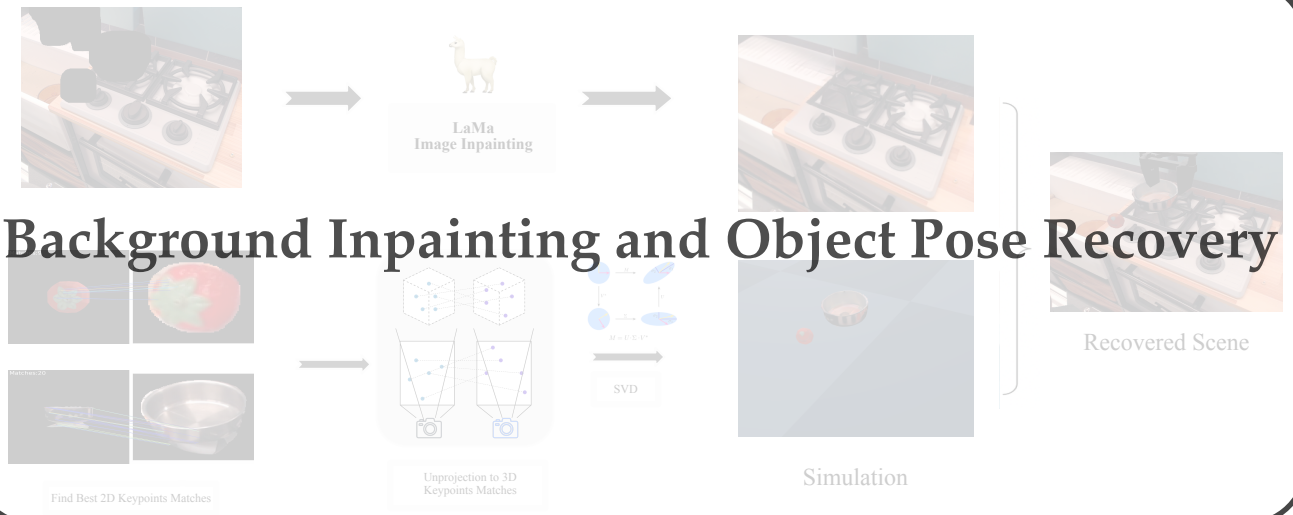
Real2Sim Translation

Automated Camera - Robot Calibration



Relevant Object Reconstruction

Background Inpainting and Object Pose Recovery



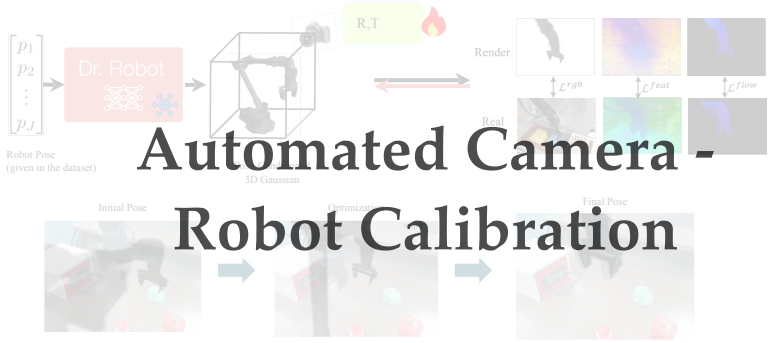
Segmenting and Labelling Objects

Prompt: "...Output a JSON list of segmentation masks where each entry contains the 2D bounding box in "box_2d", a descriptive text label in "label", and the mask in "mask" ".

Gemini Team. (2022), Gemini: A Family of Highly Capable Multimodal Models.

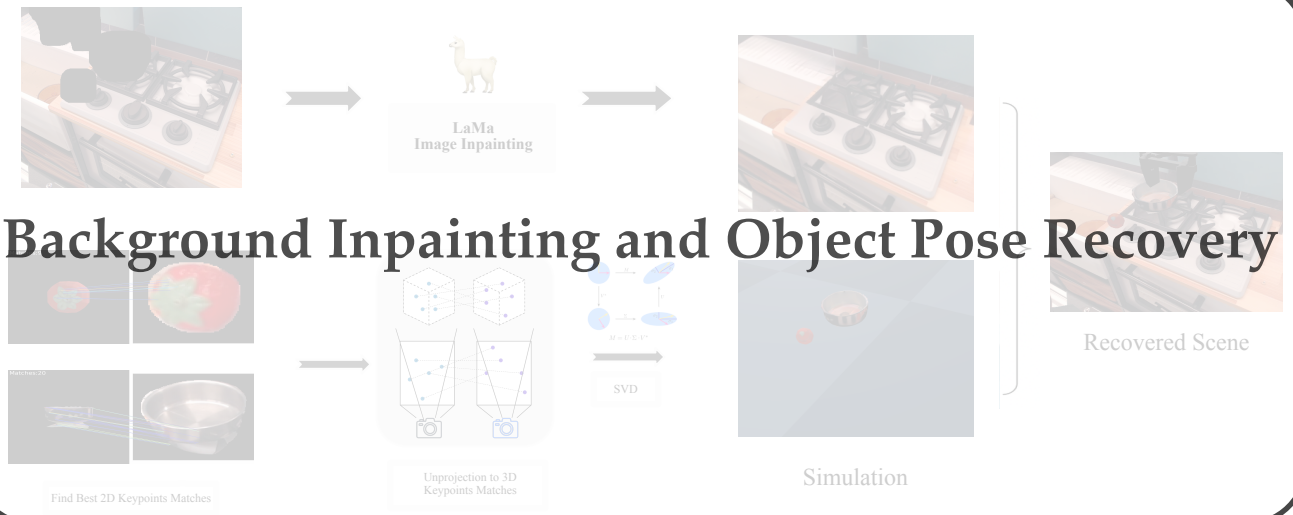
Real2Sim Translation

Automated Camera - Robot Calibration



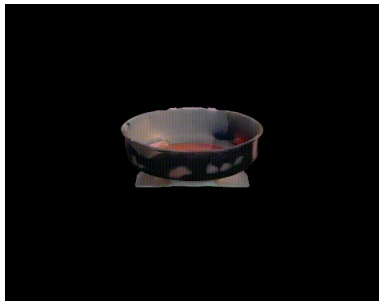
Relevant Object Reconstruction

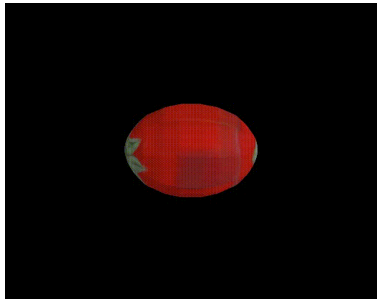
Background Inpainting and Object Pose Recovery



Reconstructing 3D Objects

Image to 3D Mesh Generation

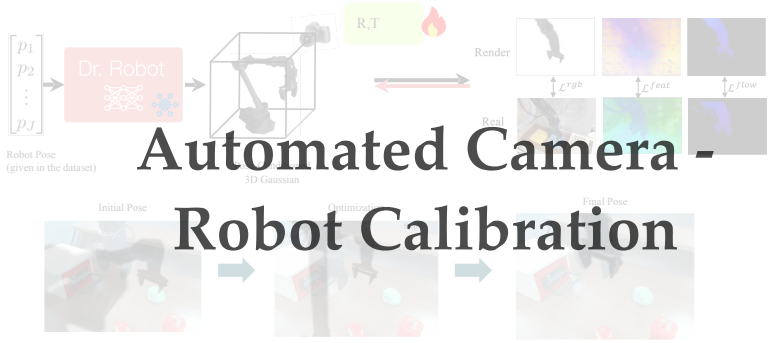




Tencent Hunyuan3D Team, Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation. CVPR 2025

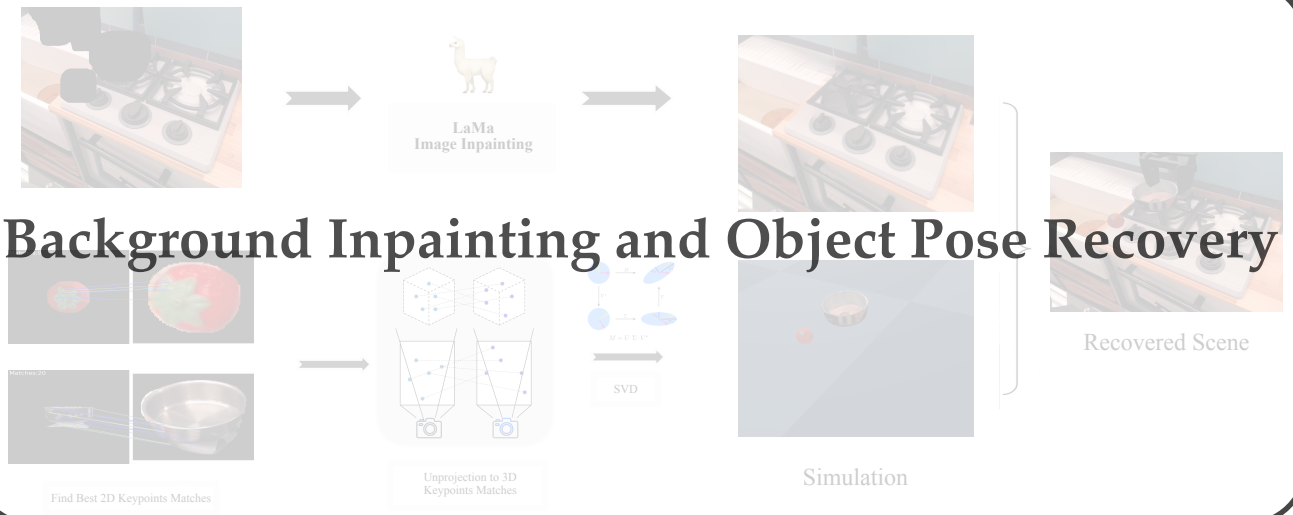
Real2Sim Translation

Automated Camera - Robot Calibration



Relevant Object Reconstruction

Background Inpainting and Object Pose Recovery



Querying VLMs for Object Physics Properties

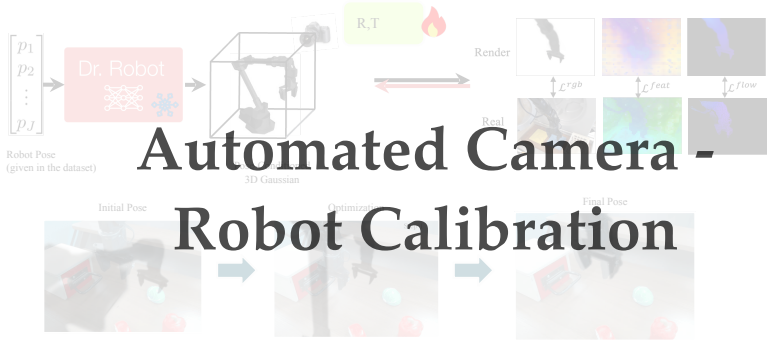
Red Tomato Toy | Metal Pot

Prompt: *I have a {class_label}. Estimate and return a JSON object describing its possible physical properties. For each entry, include one plausible material the object is most likely made of,, along with their estimated physics attributes. Do not hallucinate incorrect material properties or non-existent material types. Do not include duplicate materials. Please return it as a JSON object that can be parsed automatically in the following format. The format is:*

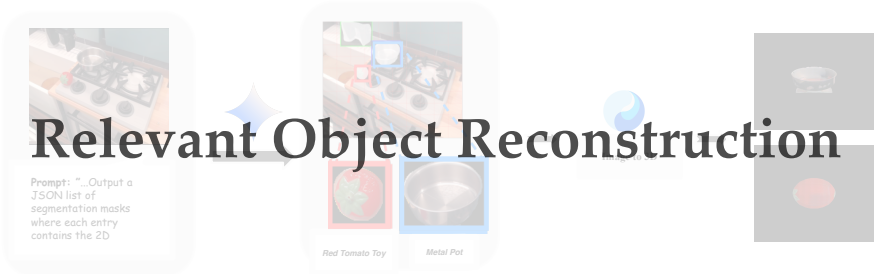
```
json_format = f"{{ \
    '{class_label}': [ \
        'Mass (kg)': ..., \
        'Friction Coefficient': ..., \
        'Material': one of ['Glass', 'Water', 'Emission', 'Plastic', 'Rough',
        'Smooth', 'Reflective', 'Metal', 'Iron', 'Aluminium', 'Copper', 'Gold'] \
    ] \
}} \"
```

Real2Sim Translation

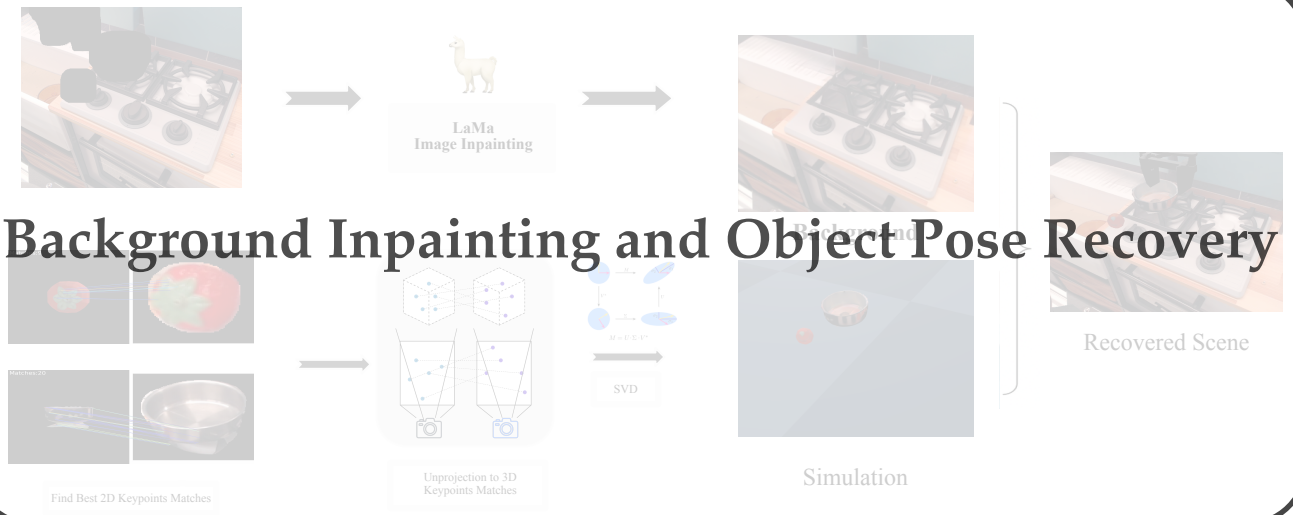
Automated Camera - Robot Calibration



Relevant Object Reconstruction



Background Inpainting and Object Pose Recovery



Background Inpainting

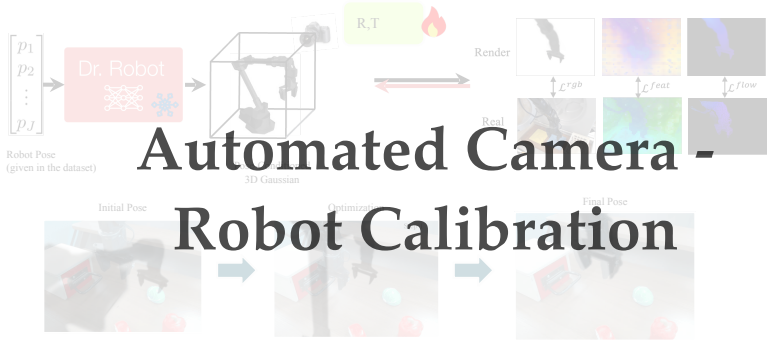
LaMa
Image Inpainting

Background

Suvorov et al., LaMa: Resolution-robust Large Mask Inpainting with Fourier Convolutions WACV 2022.

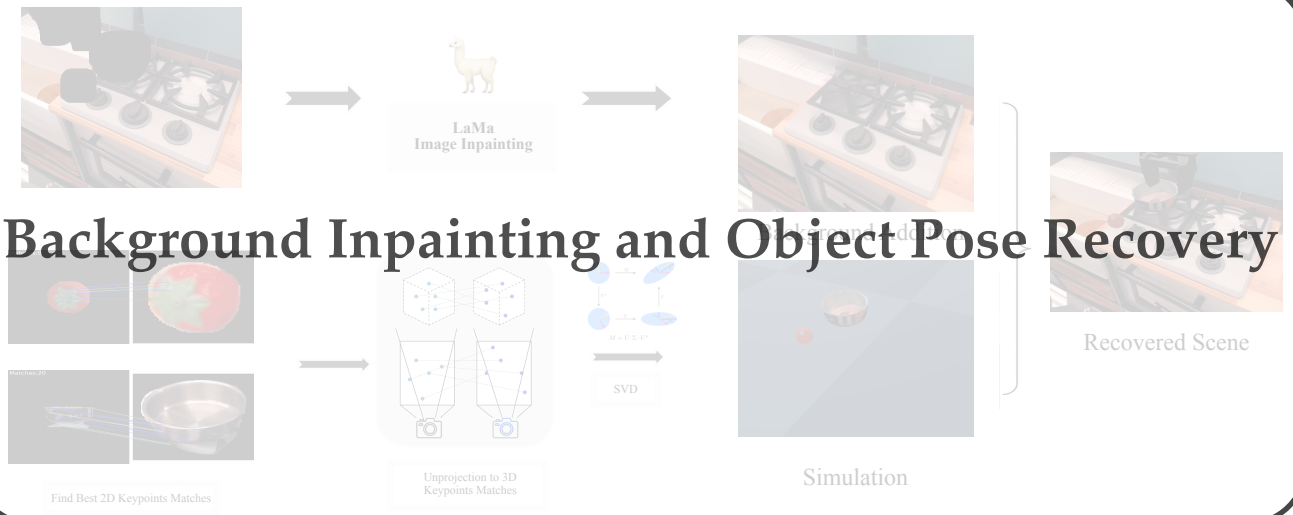
Real2Sim Translation

Automated Camera - Robot Calibration

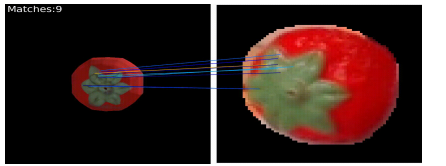


Relevant Object Reconstruction

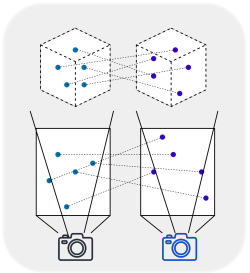
Background Inpainting and Object Pose Recovery



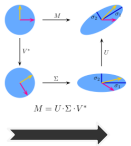
Estimating Object Scale, Orientation and Translation with Render-and-Compare



Find Best 2D Keypoints Matches

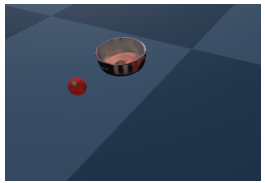


Unprojection to 3D Keypoints Matches



SVD

Background Addition



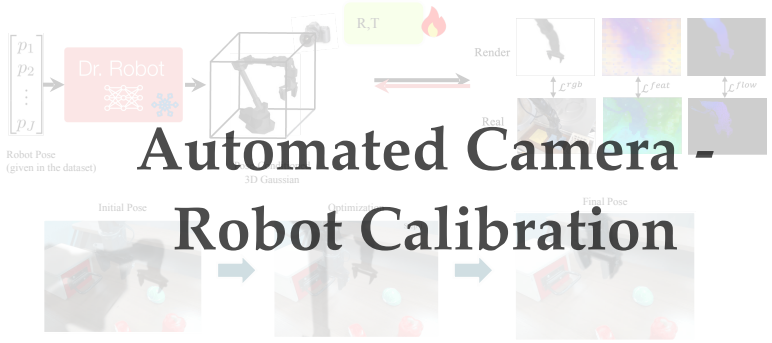
Simulation

Recovered Scene

Ren et al., MINIMA: Modality Invariant Image Matching CVPR 2025.

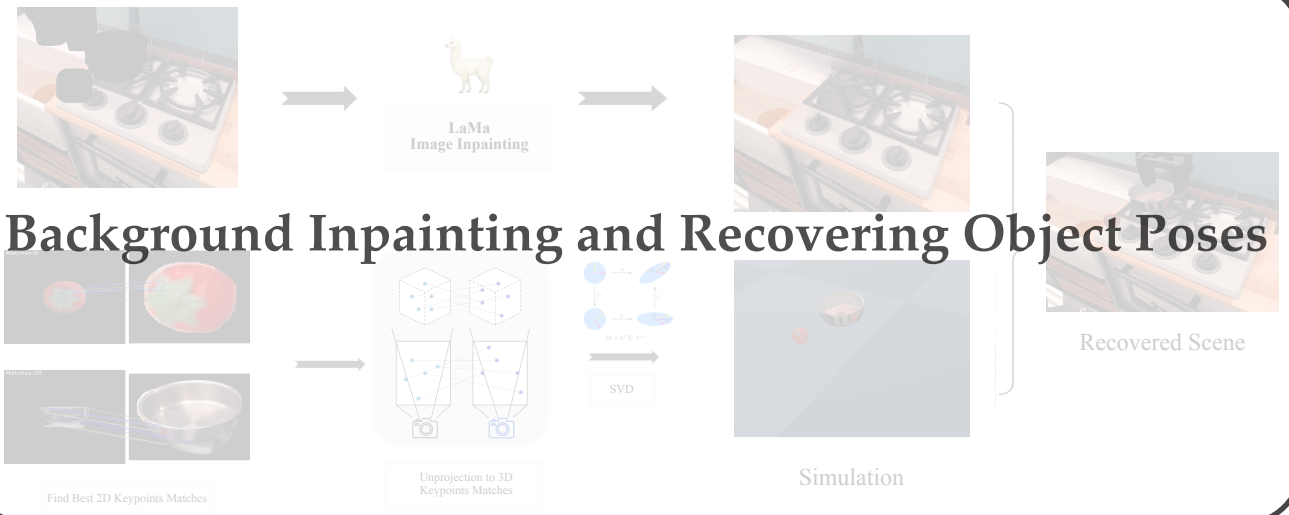
Real2Sim Translation

Automated Camera - Robot Calibration

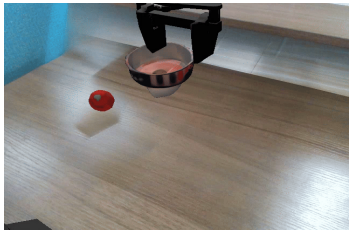


Relevant Object Reconstruction

Background Inpainting and Recovering Object Poses



Scene Perturbations



Background Change

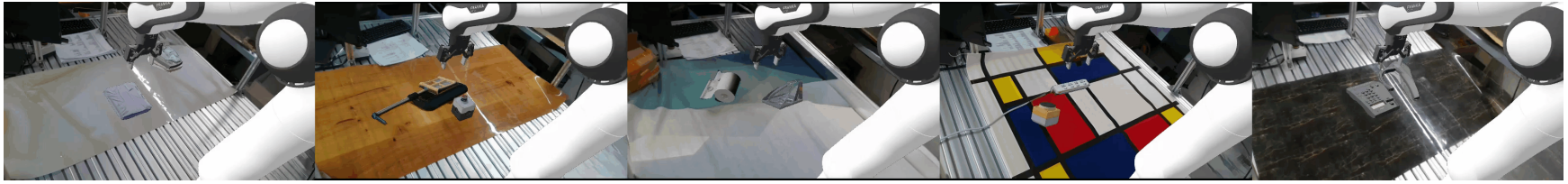
Base Scene

Color Shift

Object Pose Change

BridgeSim

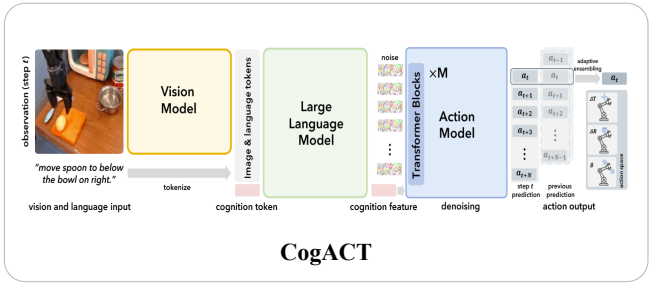
DROIDSim



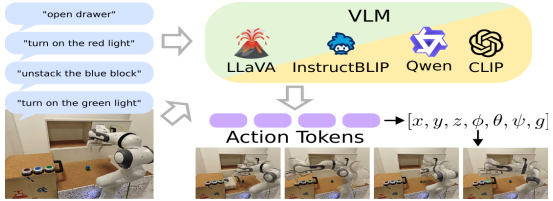
Rh20TSim

Walke et al. (2023). BridgeData V2: A Dataset for Robot Learning at Scale. In Conference on Robot Learning (CoRL).
Khazatsky et al. (2024a). DROID: A Large-Scale In-the-Wild Robot Manipulation Dataset. arXiv:2403.12945.
Fang et al. (2023). RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot. arXiv:2307.00595.

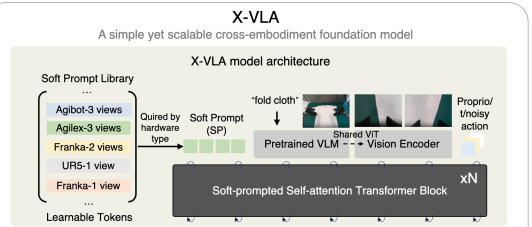
RobotArena ∞: Deploying VLAs



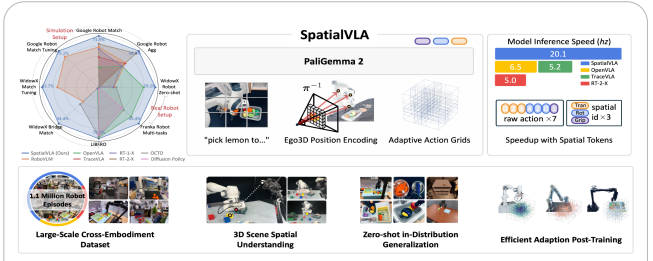
CogACT



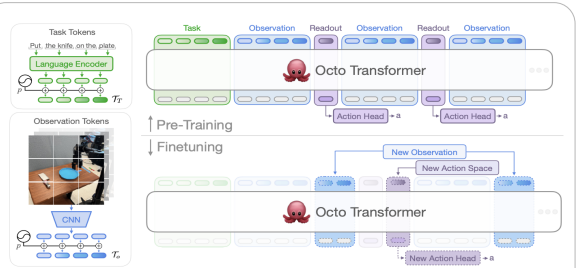
RoboVLMs



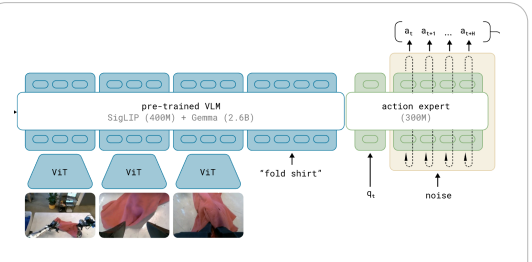
X-VLA



SpatialVLA



Octo



Open-pi-zero

Li et al. (2024), CogACT: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation. arXiv:2411.19650
Li et al. (2024), Towards Generalist Robot Policies: What Matters in Building Vision-Language-Action Models. arXiv:2412.14058
Qu et al. (2024), SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models. RSS 2025
Octo Model Team and Ghosh et al. (2023), Octo: An Open-Source Generalist Robot Policy. RSS 2024
The X-VLA Team and Chen et al. (2025), X-VLA: Soft-Prompted Transformer as Scalable Cross-Embodiment Vision-Language-Action Model. arXiv:2510.10247
Physical Intelligence and Black et al. (2024), π_0: A Vision-Language-Action Flow Model for General Robot Control. RSS 2025
Ren et al. (2024), Open-Pi-Zero: An Open-Source Re-implementation of Pi-0. GitHub Repository:allenzren/open-pi-zero

RobotArena ∞: Deploying VLAs

"Put the red object in the pot"

CogACT

RoboVLM

X-VLA

SpatialVLA

Octo

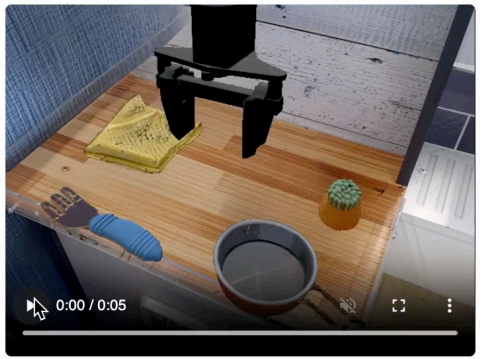
Open-pi-zero

Which policy is better?

Task: Place the fork near the yellow cloth

Video 7 of 150

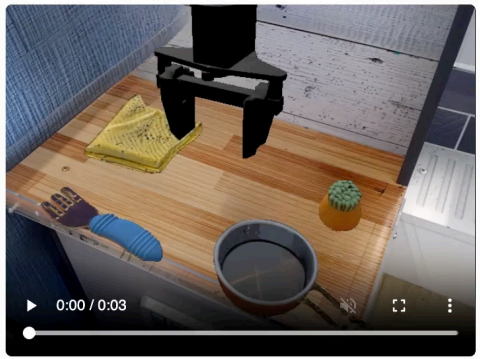
Policy A



Describe the robot's attempt to complete the task:

e.g., The robot reaches to the cup, picks it up and...

Policy B



Describe the robot's attempt to complete the task:

e.g., The robot made no meaningful move...

Submit Descriptions

Please provide descriptions for both videos to continue.

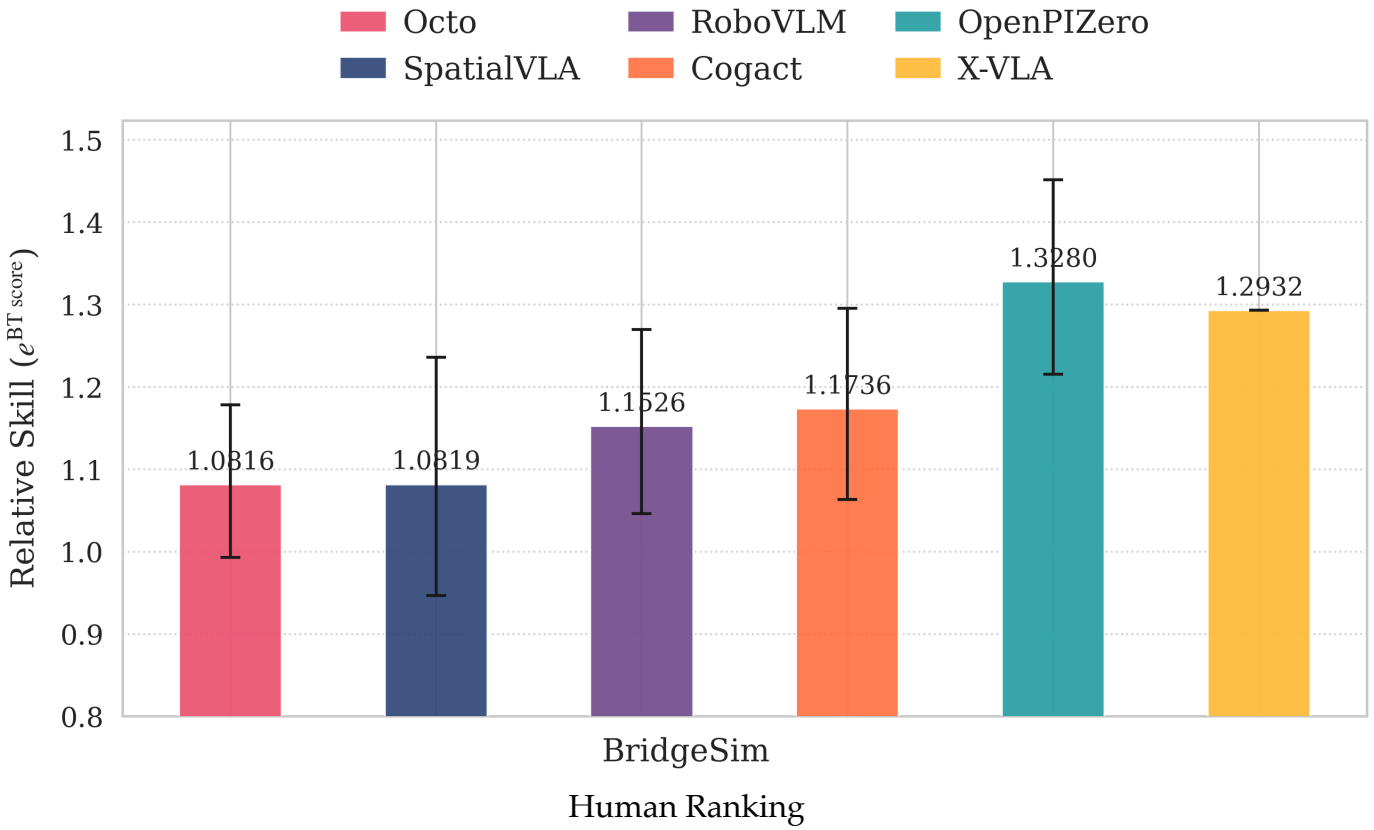
Your evaluation:

👈 Left is Better | 🤝 Tie | Right is Better 👉

(equally good or equally bad)

Next

RobotArena ∞: Policy Rankings at Scale



Task Progress Estimation with VLMs

"Put the tomato in the pot"

Task Prompt

Frames

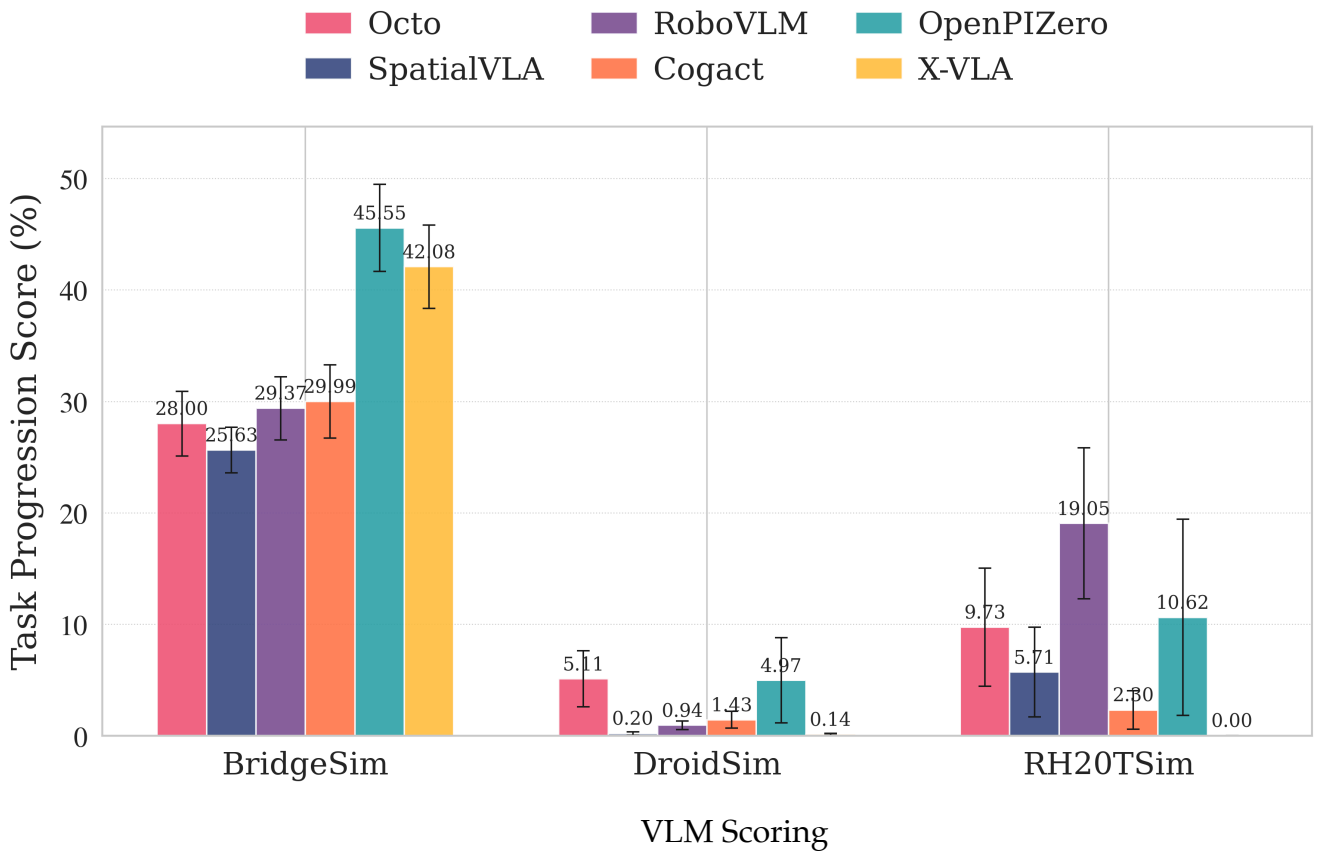
Robot and Object State

Vision Language Model

Autoregressive Prediction

100% → 79% → 24% → 0% → 52%

RobotArena ∞: Policy Rankings at Scale



Limitations and Ongoing Work

- 📽️ **Static Camera Assumption**
  The image background superimposition assume a static camera

- ⚙️ **Going beyond rigid objects:** toward fluids, gases, and deformable materials

Thank you!