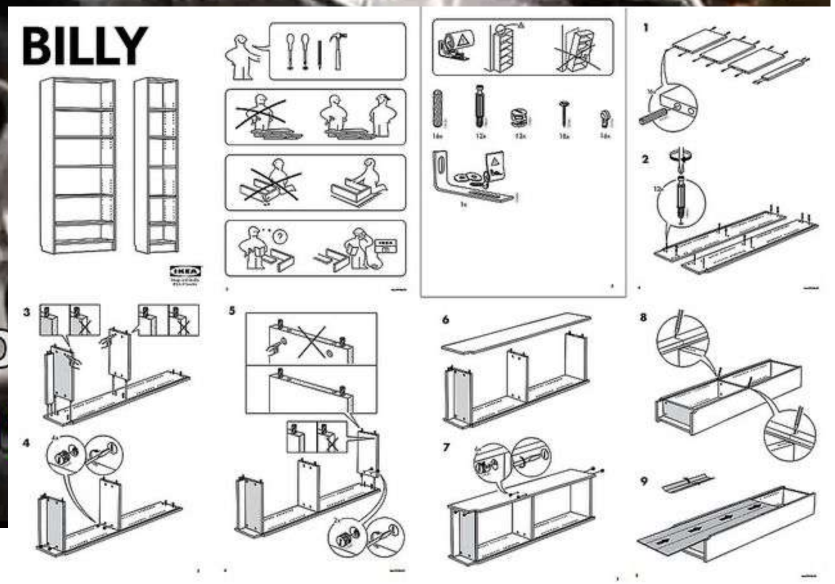
# Learning manipulation skills from instructional videos


Josef Sivic

CZECH TECHNICAL UNIVERSITY IN PRAGUE

ellis unit | PRAGUE

# Motivation: learning from instructional videos


[Alyarac et al., CVPR 2016]

# Motivation: object manipulation for assistance

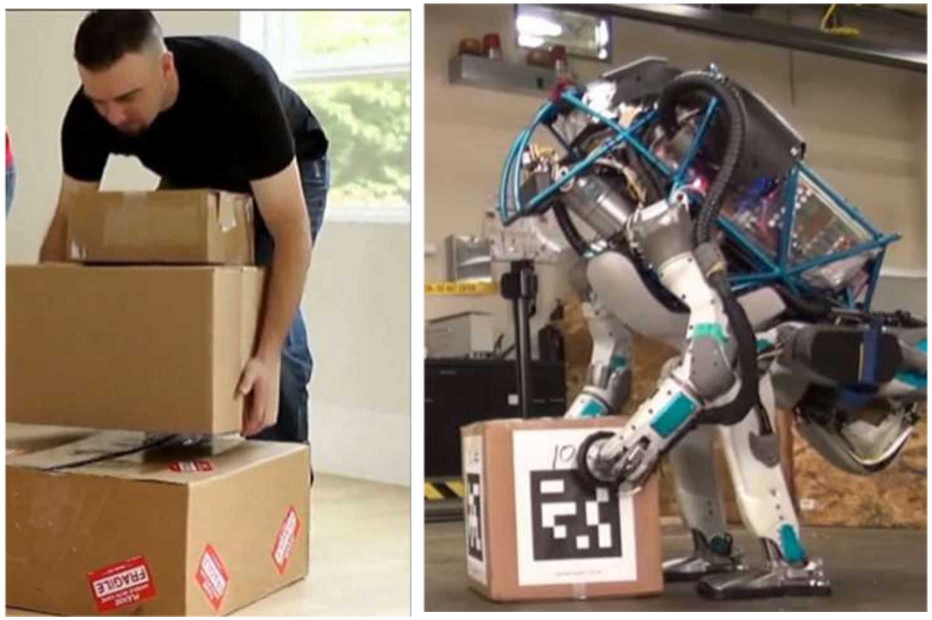
[Microsoft HoloLens]

Personal assistant

Konica Minolta AIRe Lens

Assistant for industrial environments

# Motivation: learning object manipulation skills



Moving goods

To operate in dangerous environments
[Darpa robot challenge 2015]

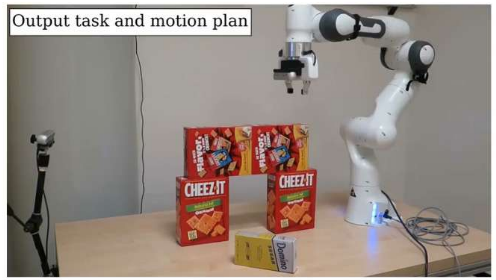
# Outline

## Learning manipulation skills from videos
[Zorina et al., IEEE RA-L 2022]

## Pre-training for visually guided manipulation
[Labbe et al., ECCV 2020, Labbe et al., CVPR 2021, Labbe et al., CoRL 2022, Fourmy et al., 2023]

**More data (and less 3D)**

## Multi-contact task and motion planning guided by video demonstration
[Zorina et al., ICRA 2023]




## Toward learning reward functions from videos
[Soucek et al., CVPR 2022], [Soucek et al., PAMI 2024], [Soucek et al., CVPR 2024]

# Outline

## Learning manipulation skills from videos

[Zorina et al., IEEE RA-L 2022]

## Pre-training for visually guided manipulation

[Labbe et al., ECCV 2020, Labbe et al., CVPR 2021, Labbe et al., CoRL 2022, Fourmy et al., 2023]

Kateryna Zorina

## Multi-contact task and motion planning guided by video demonstration

[Zorina et al., ICRA 2023]

## Toward learning reward functions from videos

[Soucek et al., CVPR 2022], [Soucek et al., PAMI 2024], [Soucek et al., CVPR 2024]

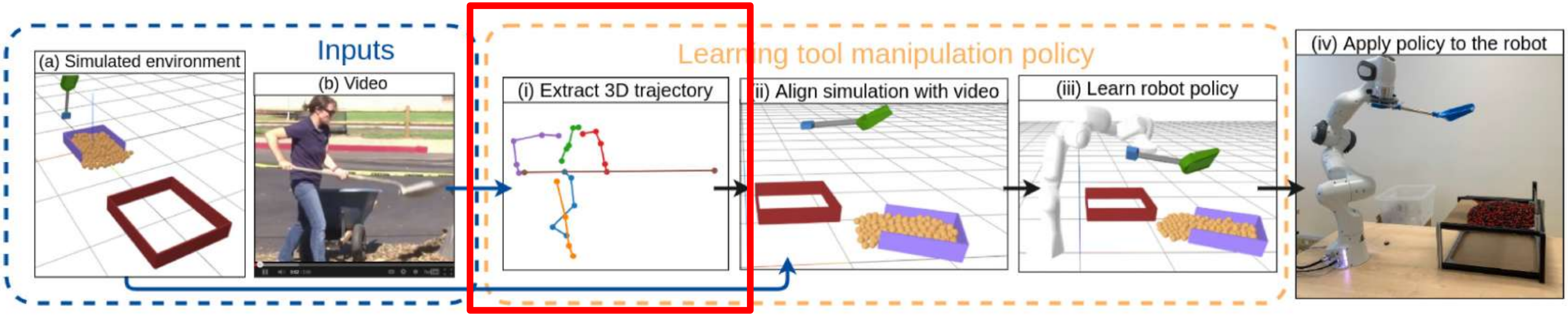
# Learning to Use Tools by Watching Videos

Input: instructional video from YouTube

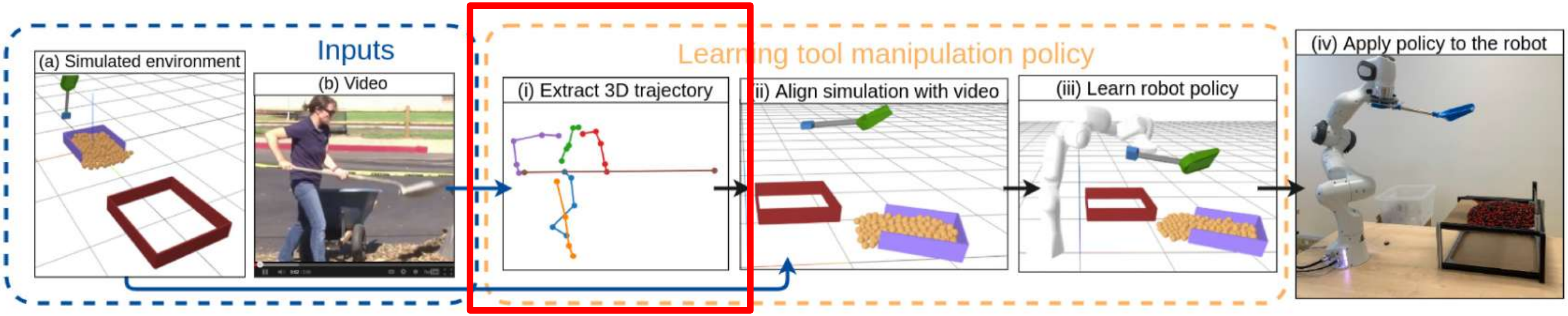
Output: tool manipulation skill transferred to a robot

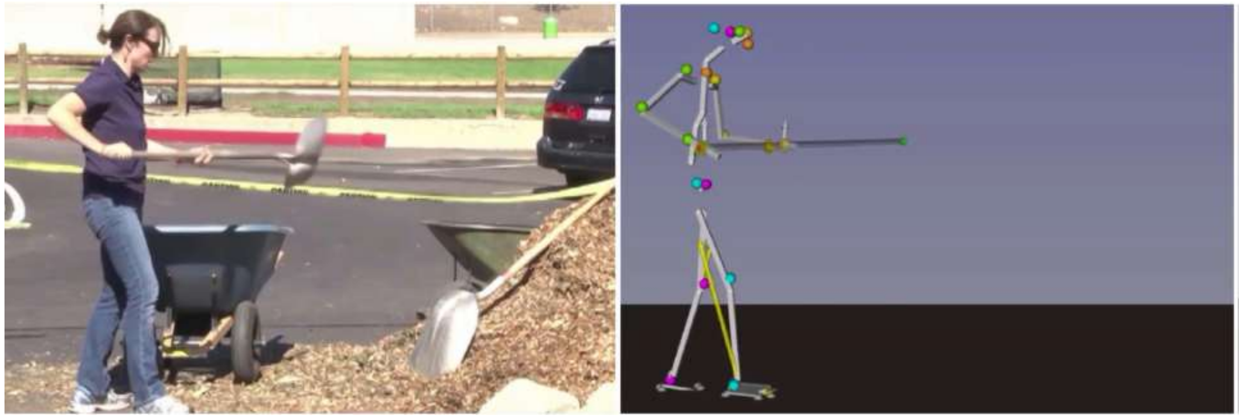
[Zorina et al., IEEE RA-L 2022]

# Approach overview



# Approach overview



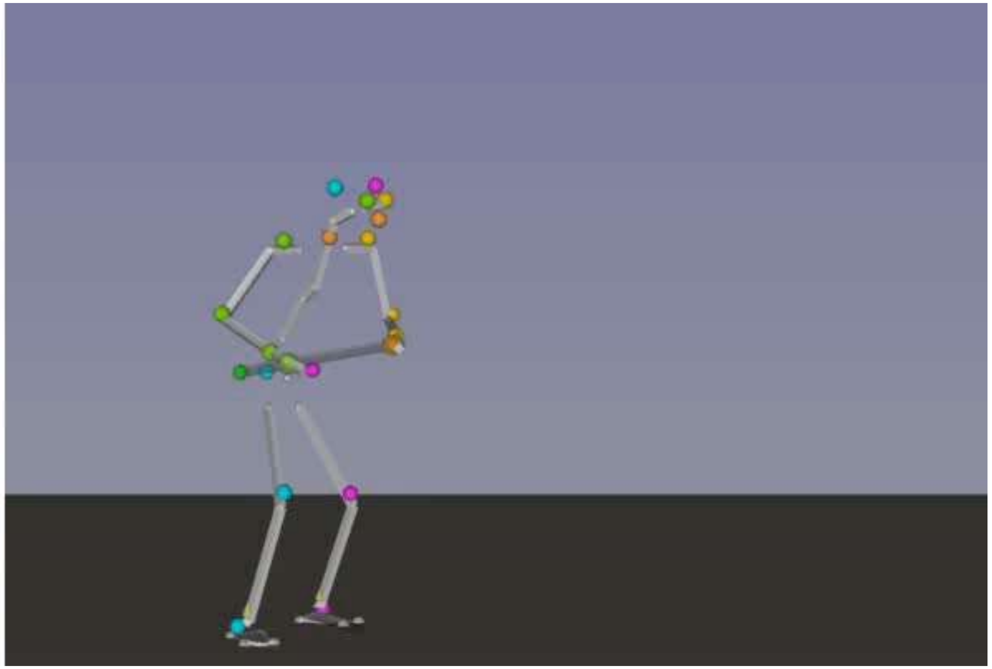



[Li et al., CVPR 2019, IJCV 2022]

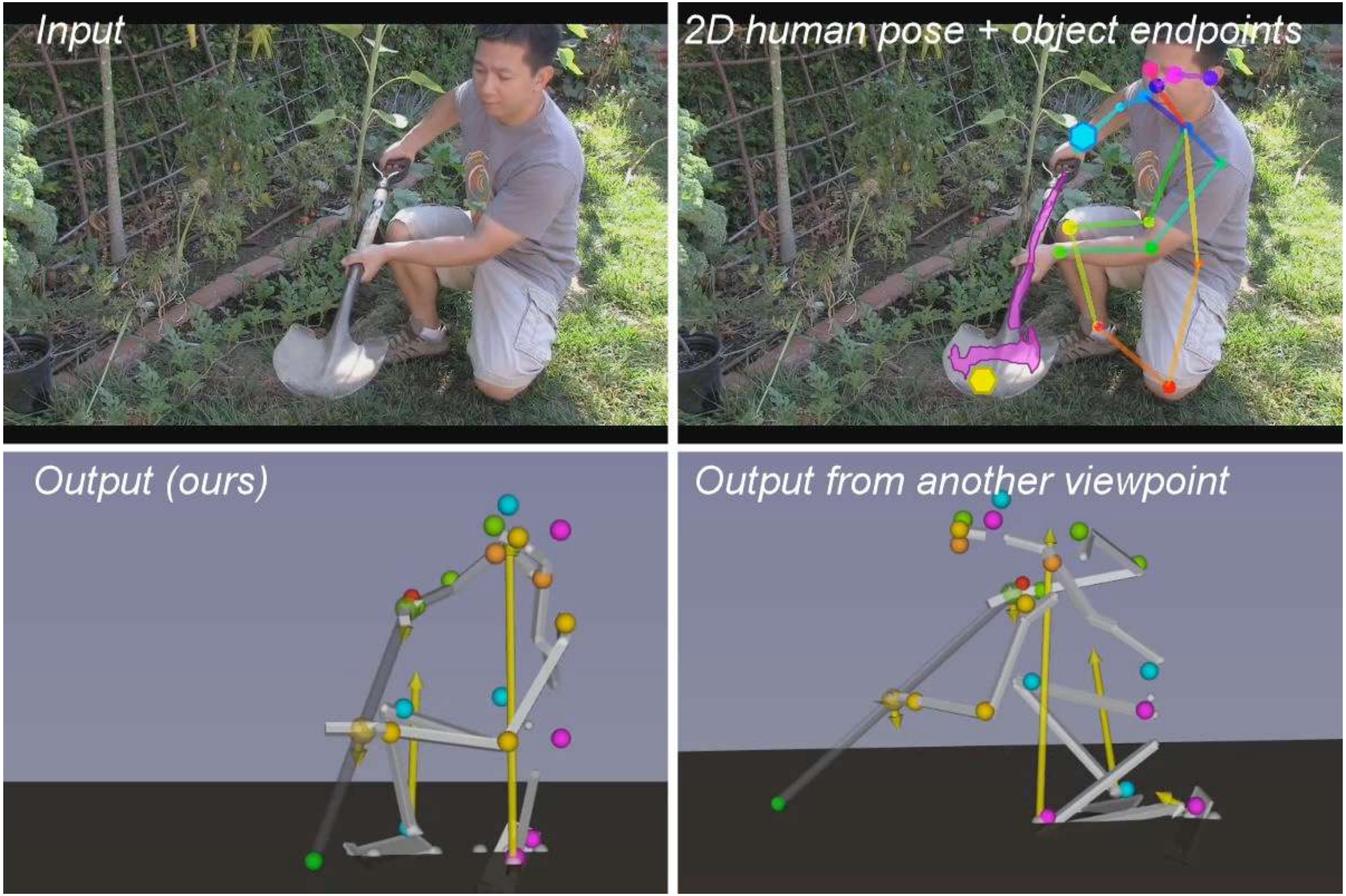
# The objective

**Input:**
- A monocular RGB video

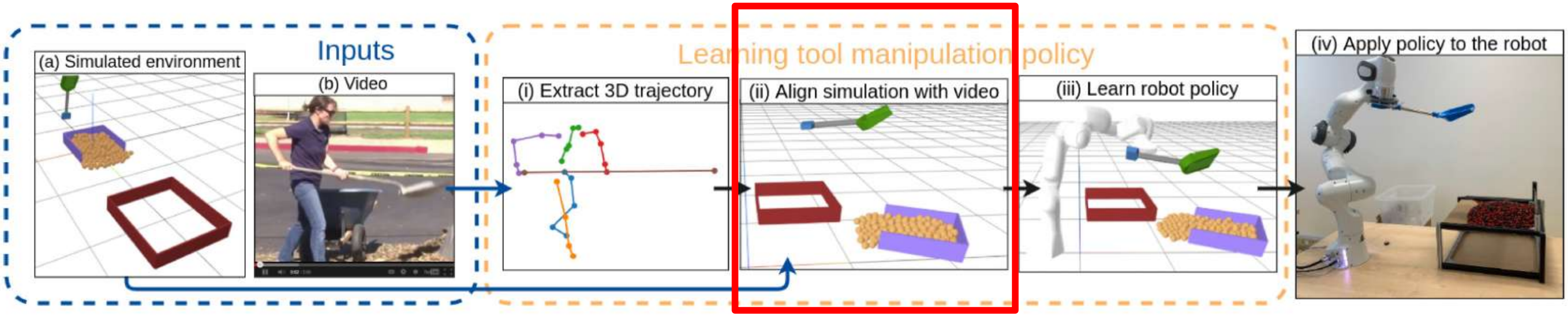
**Output:**
- Person & object 3D poses
- 3D contact forces



[Li, Sedlar, Carpentier, Mansard, Laptev, Sivic, Best paper finalist CVPR 2019; Extended version, IJCV 2021]


Input
2D human pose + object endpoints
Output (ours)
Output from another viewpoint

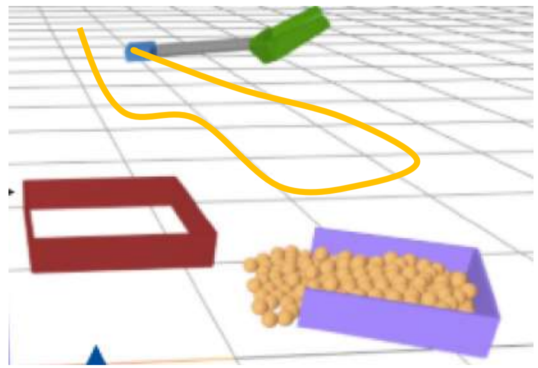
# Align simulation to video



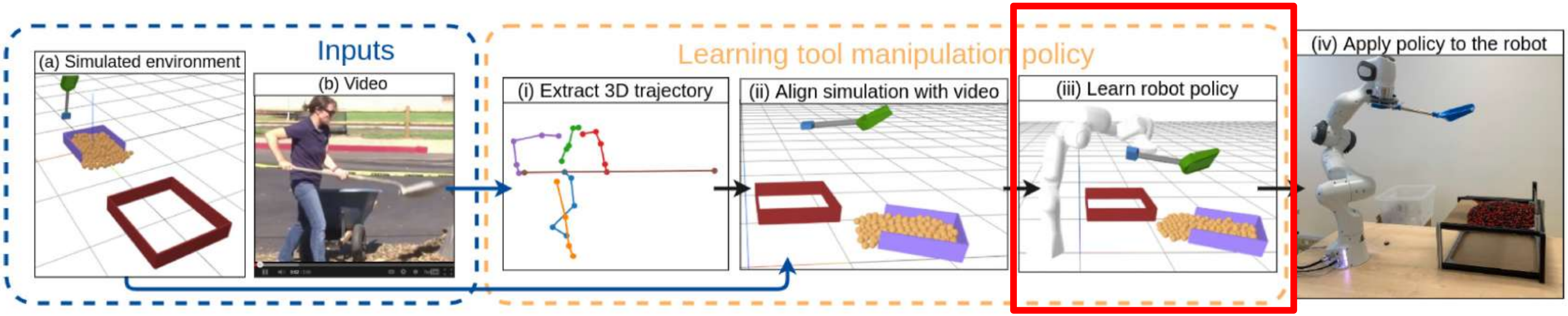

Goal: Find scene layout maximizing trajectory reward

Problem: large space of possible layouts

Approach: Guide sampling of 3D scene elements using the extracted tool trajectory



# Learn robot policy to maximize sparse reward

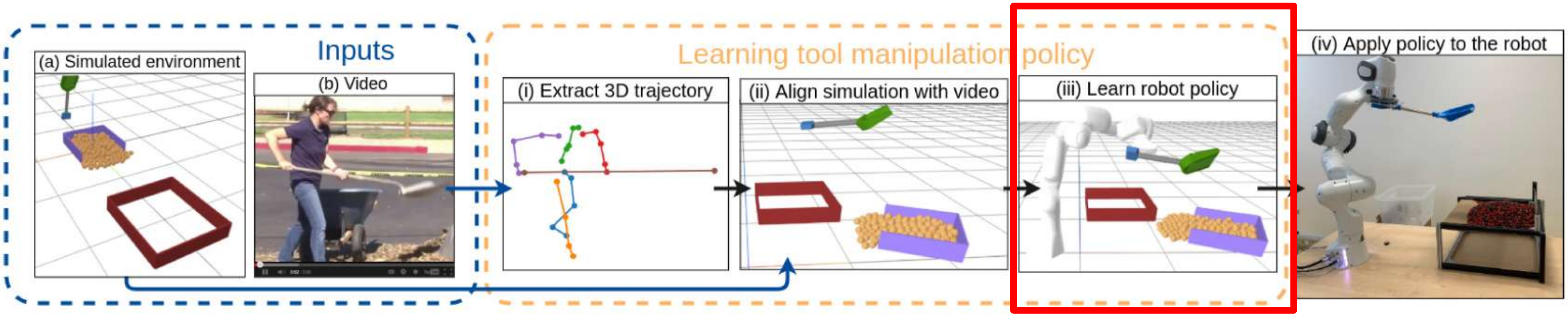



Goal: Learn **robot policy** to maximize sparse reward

Problem: RL directly applied on **sparse reward** is hard

Approach: **Imitate tool trajectory** via trajectory optimization followed by RL

# Learn robot policy




Initialize policy using **trajectory optimization**

Use trajectory optimization to find robot base position and initial robot trajectory:

$$\mathbf{b}^*, \mathbf{v}_0^*, \ldots, \mathbf{v}_T^* = \arg\min_{\mathbf{b}, \mathbf{v}_0, \ldots, \mathbf{v}_T} \sum_{t=0}^{T} \underbrace{d(\mathbf{b}, \mathbf{q}_t, \mathbf{p}_t)}_{\text{Distance to the demonstrated trajectory}} + \underbrace{w_v \mathbf{v}_t^\top \mathbf{v}_t}_{\text{Constrain speed}} + \underbrace{w_b c_b(\mathbf{q}_t)}_{\text{Respect joint limits}}$$

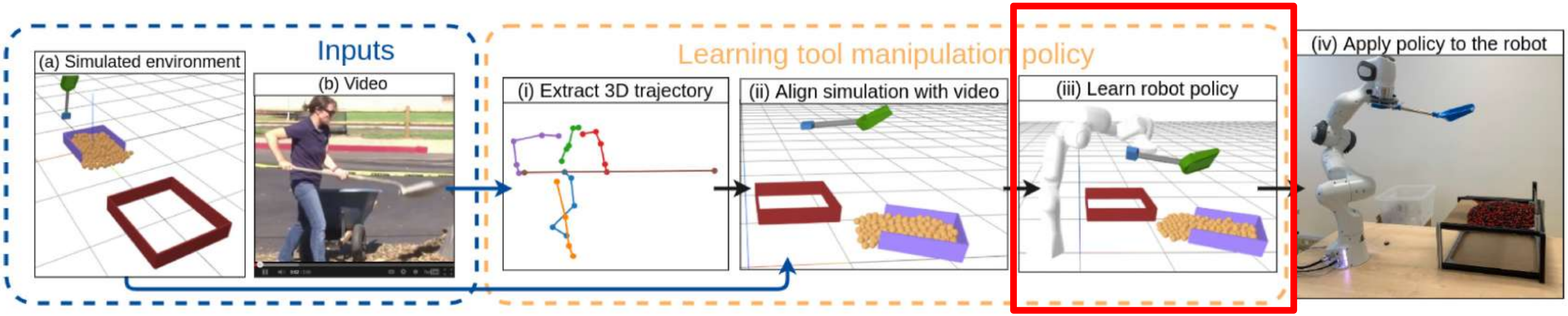
$$\text{s.t. } \mathbf{q}_t = \mathbf{q}_{t-1} + \mathbf{v}_t \Delta t$$

Target tool position

Joint position ( $q_0$ given)

Timestep

# Learn robot policy

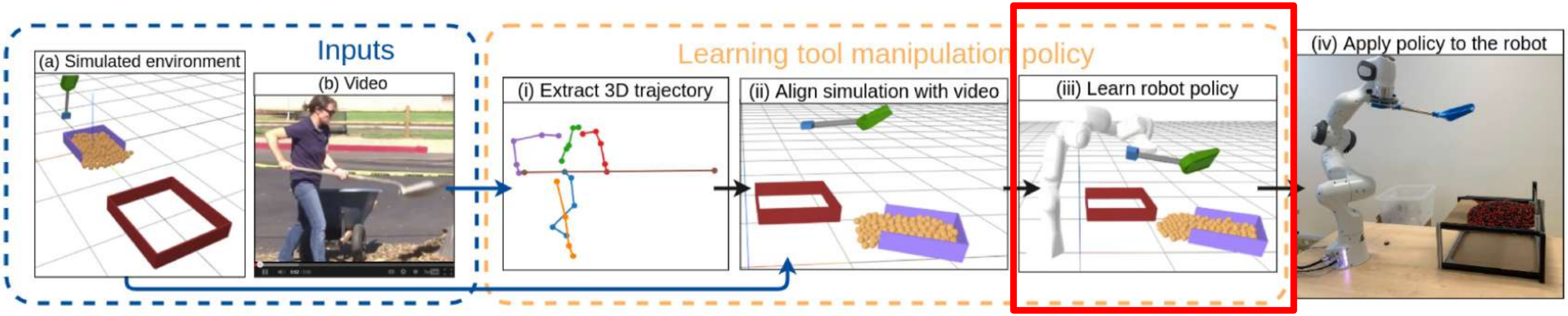



Pinocchio
Efficient and versatile rigid body dynamics algorithms

crocoddyl
Contact Robot Optimal Control by Differential Dynamic Library

https://github.com/stack-of-tasks/pinocchio
https://github.com/loco-3d/crocoddyl

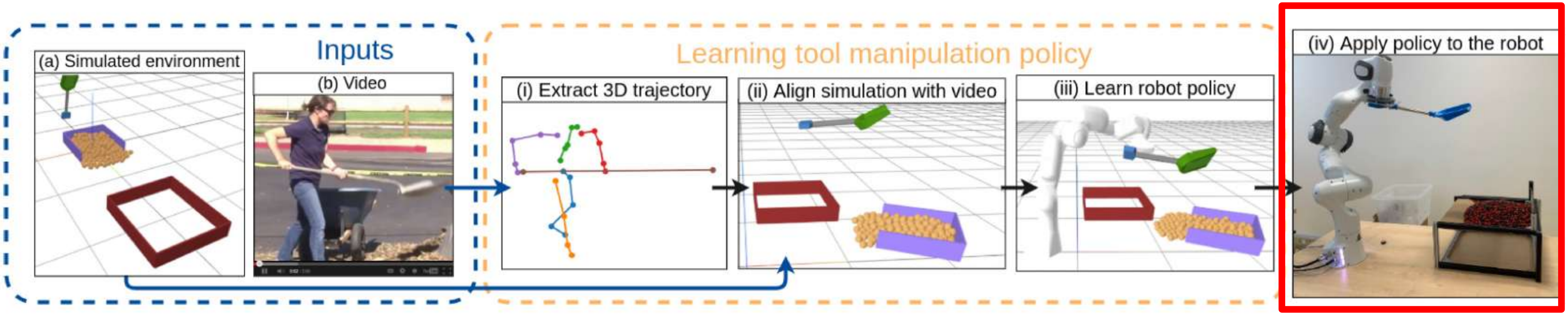
# Learn robot policy




Optimize the initial policy using **reinforcement learning**

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\tau} \sim \pi_{\boldsymbol{\theta}}} \left[ \sum_{t=0}^{H} \gamma^t r_t \right]$$

# Transfer policy to the robot



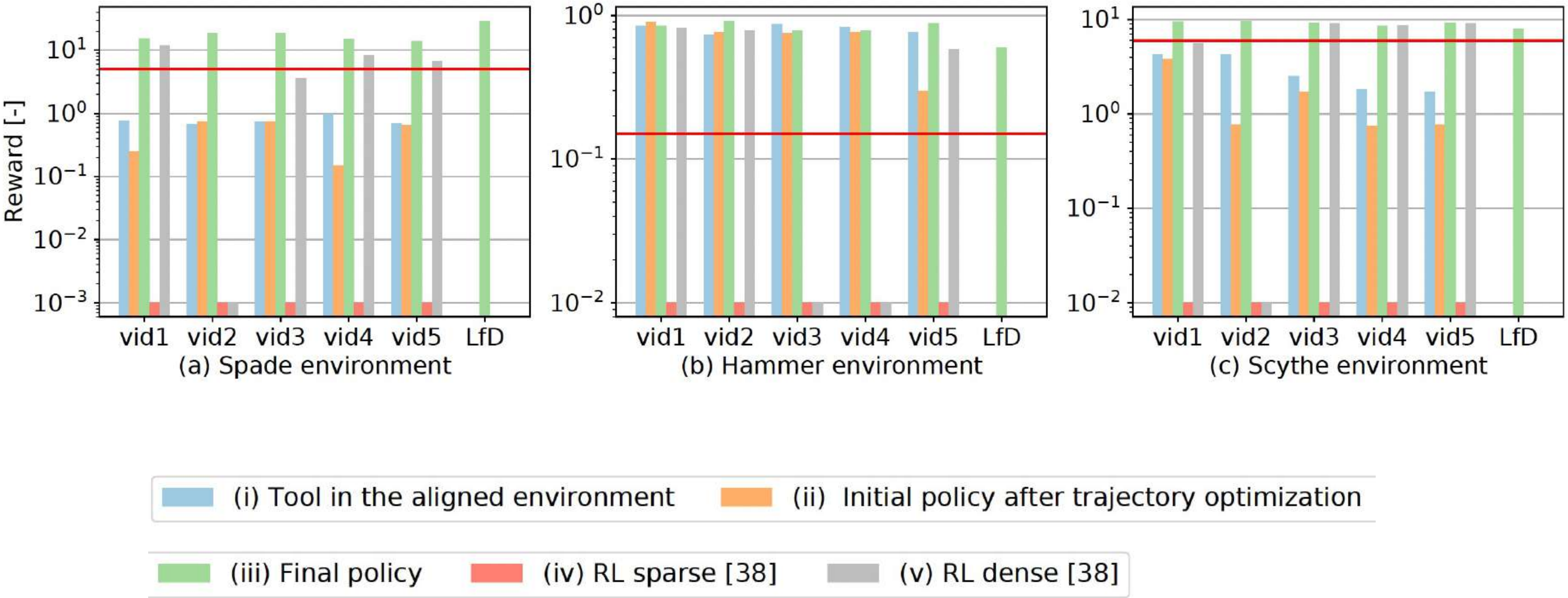
# Results

Dataset of 3 tasks (spade, hammer, scythe) and 5 videos for each task

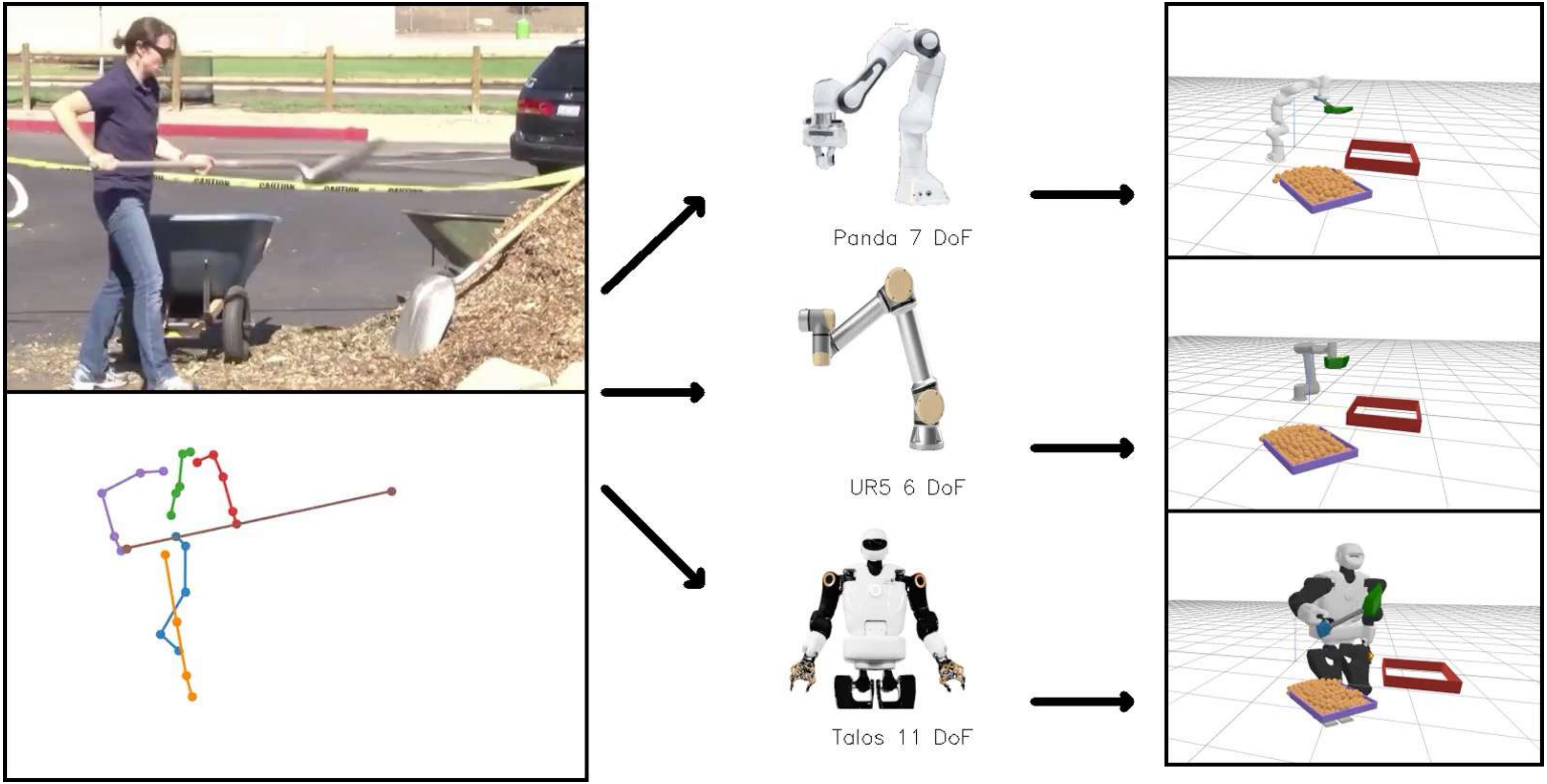
# Results

Dataset of 3 tasks (spade, hammer, scythe) and 5 videos for each task

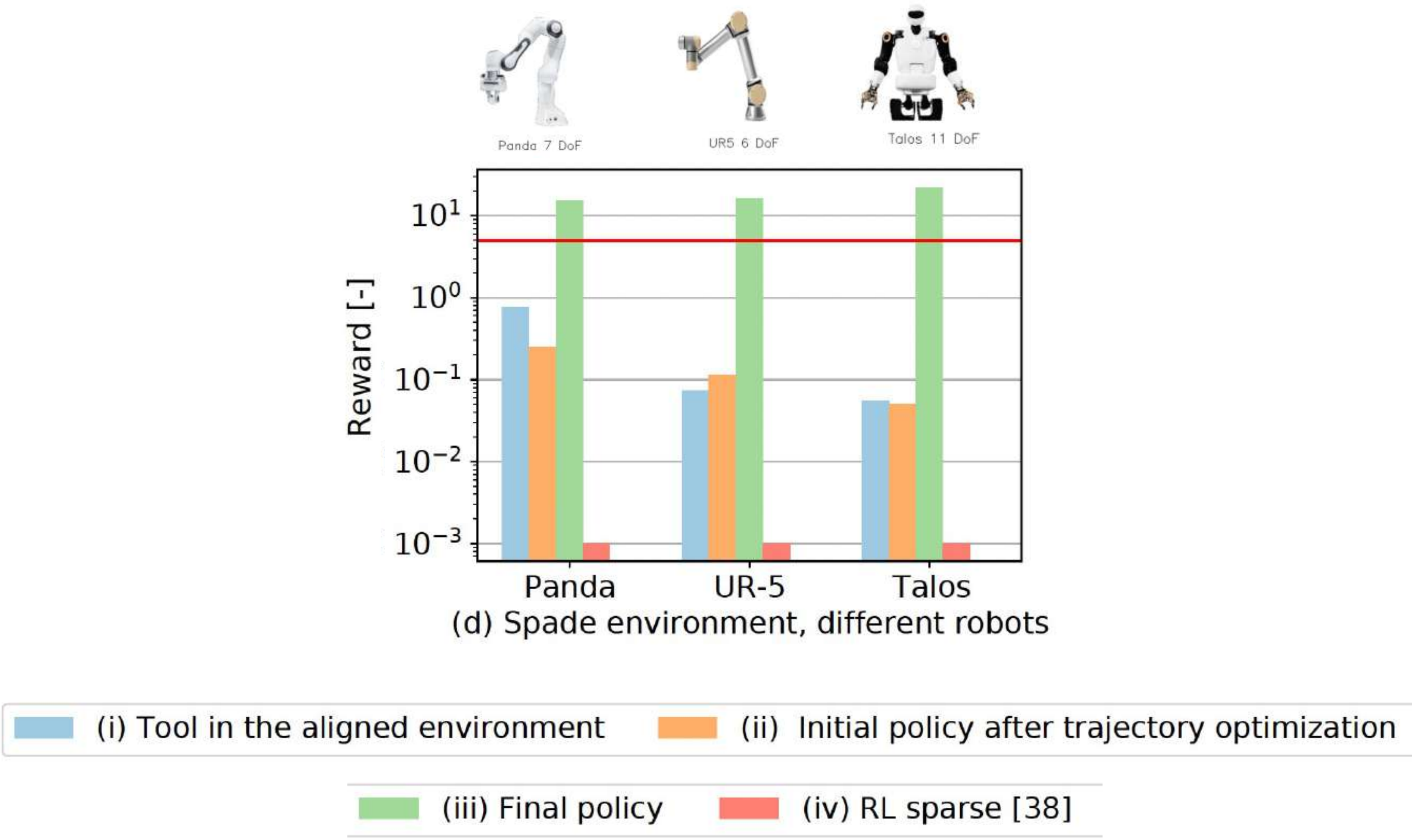
# Results: quantitative evaluation



# Results: different robots



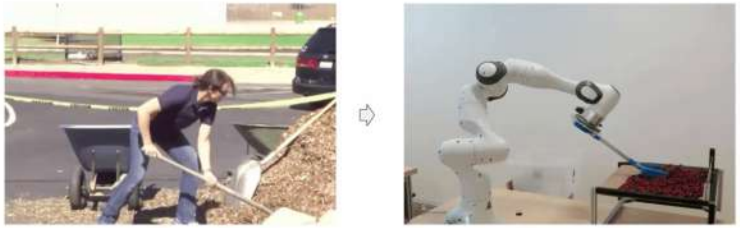
# Results: final reward for different robots




(d) Spade environment, different robots

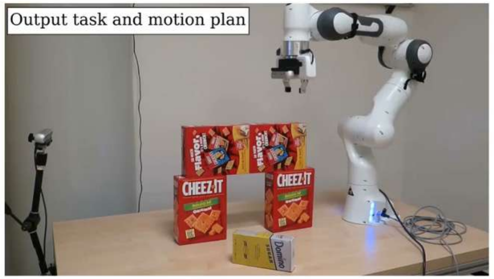
# Outline

## Learning manipulation skills from videos
[Zorina et al., IEEE RA-L 2022]



## Pre-training for visually guided manipulation
[Labbe et al., ECCV 2020, Labbe et al., CVPR 2021, Labbe et al., CoRL 2022, Fourmy et al., 2023]

## Multi-contact task and motion planning guided by video demonstration
[Zorina et al., ICRA 2023]




## Toward learning reward functions from videos
[Soucek et al., CVPR 2022], [Soucek et al., PAMI 2024], [Soucek et al., CVPR 2024]

# Outline

Learning manipulation skills from videos
[Zorina et al., IEEE RA-L 2022]

Pre-training for visually guided manipulation
[Labbe et al., ECCV 2020, Labbe et al., CVPR 2021, Labbe et al., CoRL 2022, Fourmy et al., 2023]

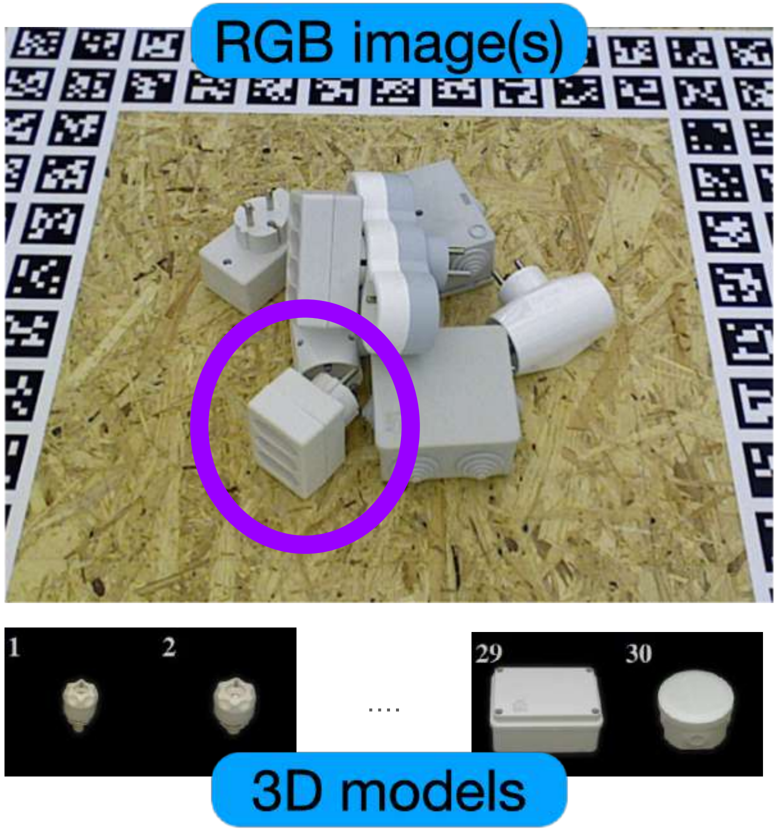
Multi-contact task and motion planning guided by video demonstration
[Zorina et al., ICRA 2023]

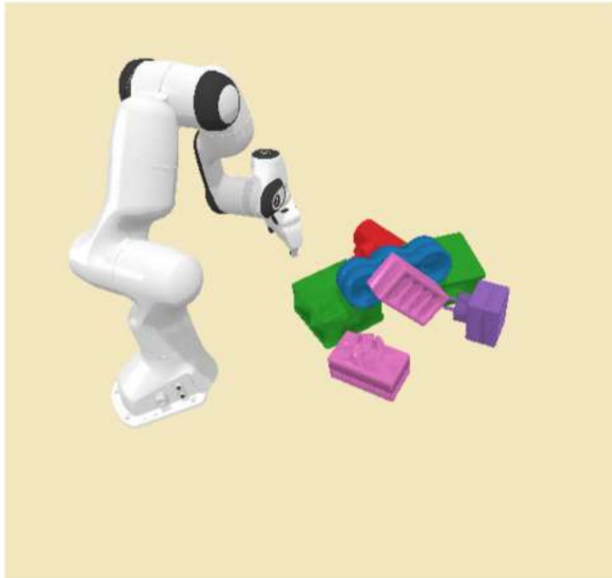
Yann Labbe

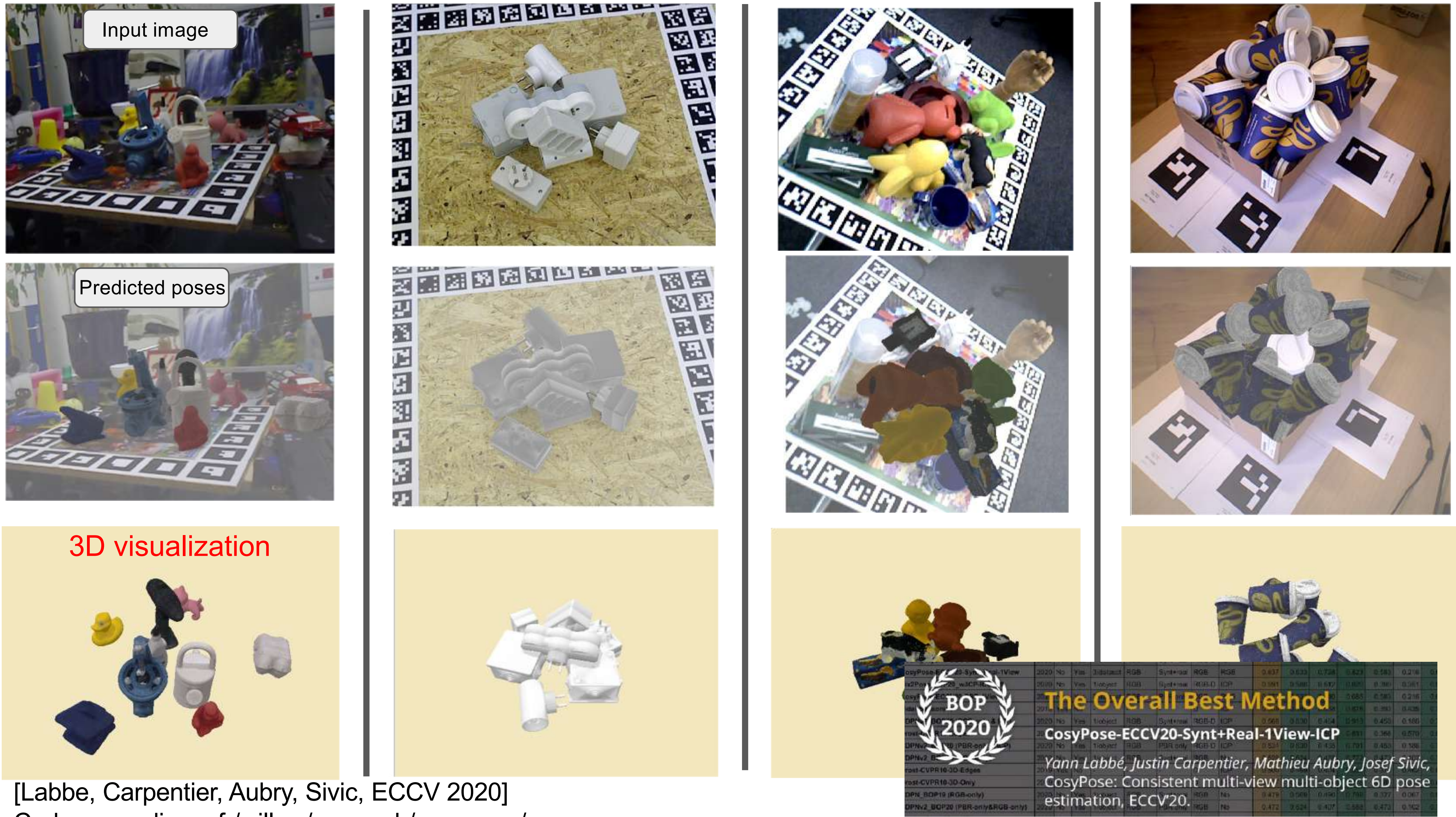
Toward learning reward functions from videos
[Soucek et al., CVPR 2022], [Soucek et al., PAMI 2024], [Soucek et al., CVPR 2024]

# Problem: 6D object pose estimation

Grasp object

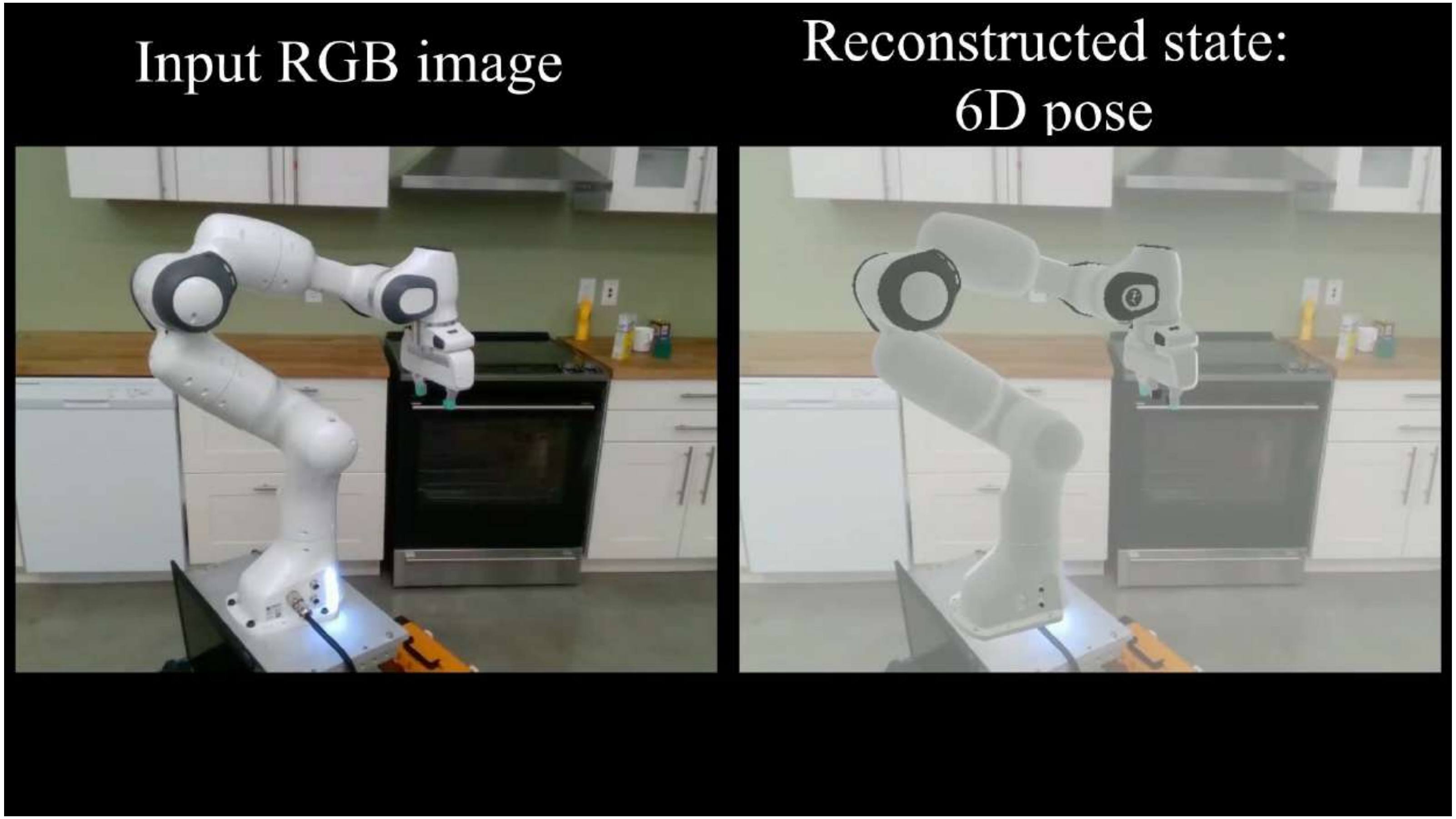




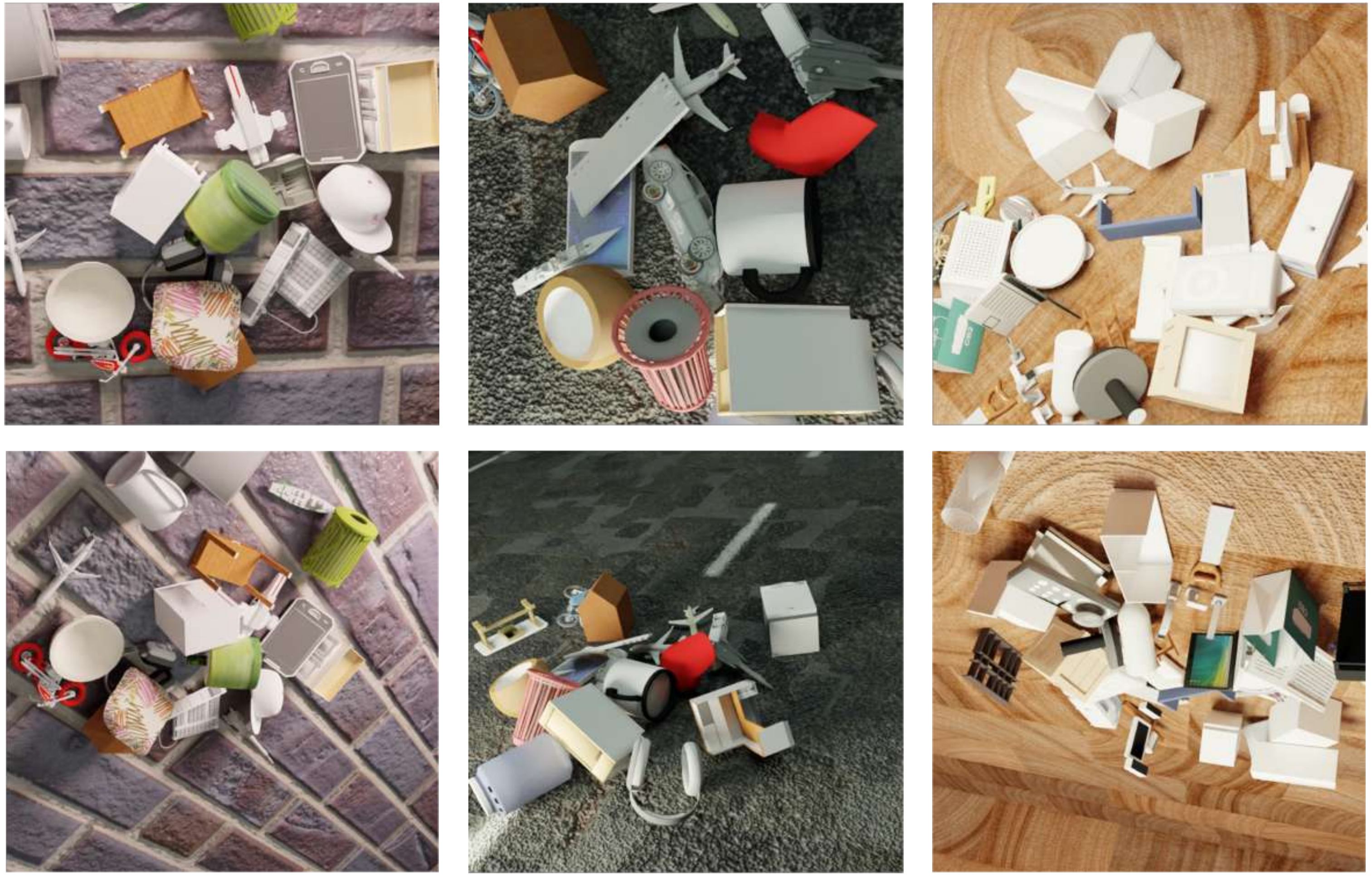

[Labbe, Carpentier, Aubry, Sivic, ECCV 2020]
Code: www.di.ens.fr/willow/research/cosypose/

# 6D pose estimation of articulated objects



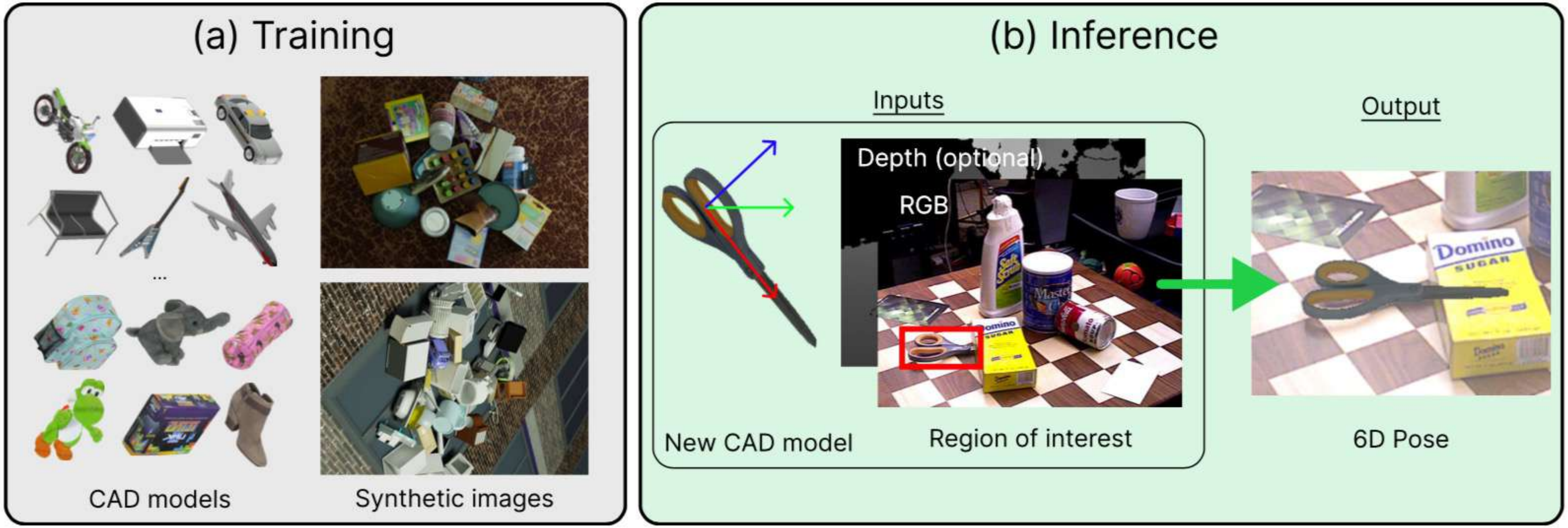

[Single-view robot pose and joint angle estimation via render&compare, Y. Labbé, J. Carpentier, M. Aubry, J.Sivic,  CVPR 2021].

# Generalization to objects unseen at training?



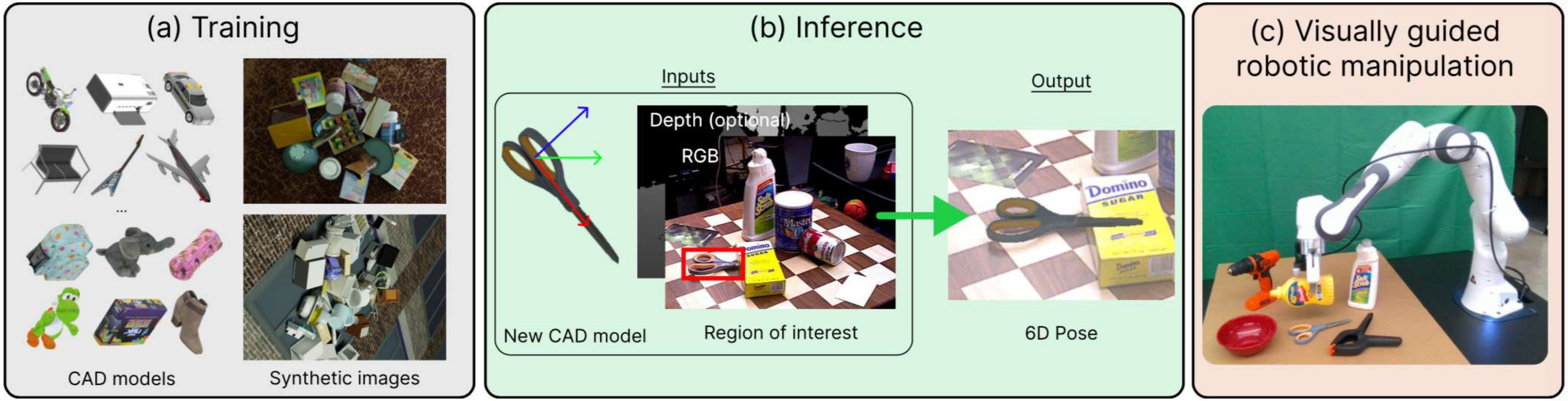
[Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, J. Sivic, CoRL 2022.].

# MegaPose: set-up
# Generalization to new objects to unseen at training




[Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, J. Sivic, CoRL 2022,].

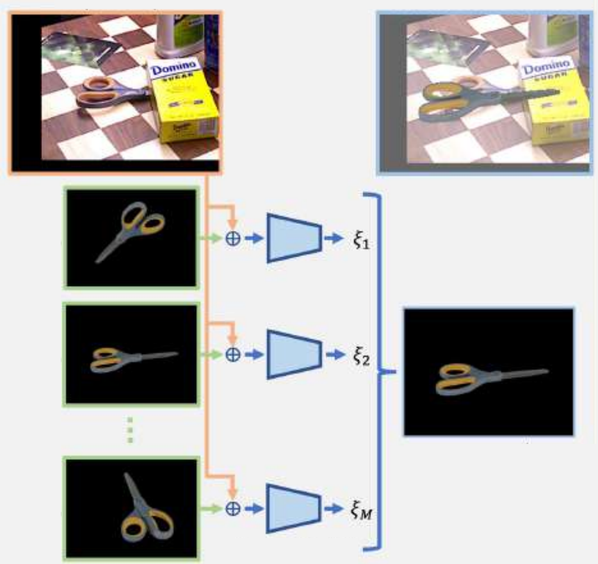
# Motivation: Visually guided manipulation



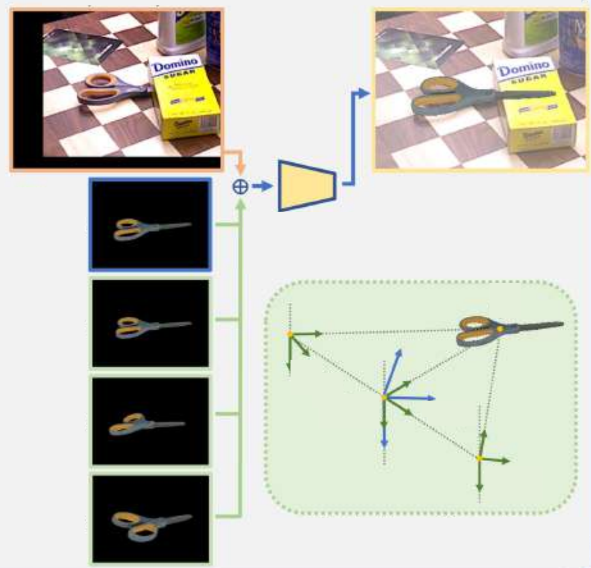
# Method overview

1. Large-scale synthetic training data

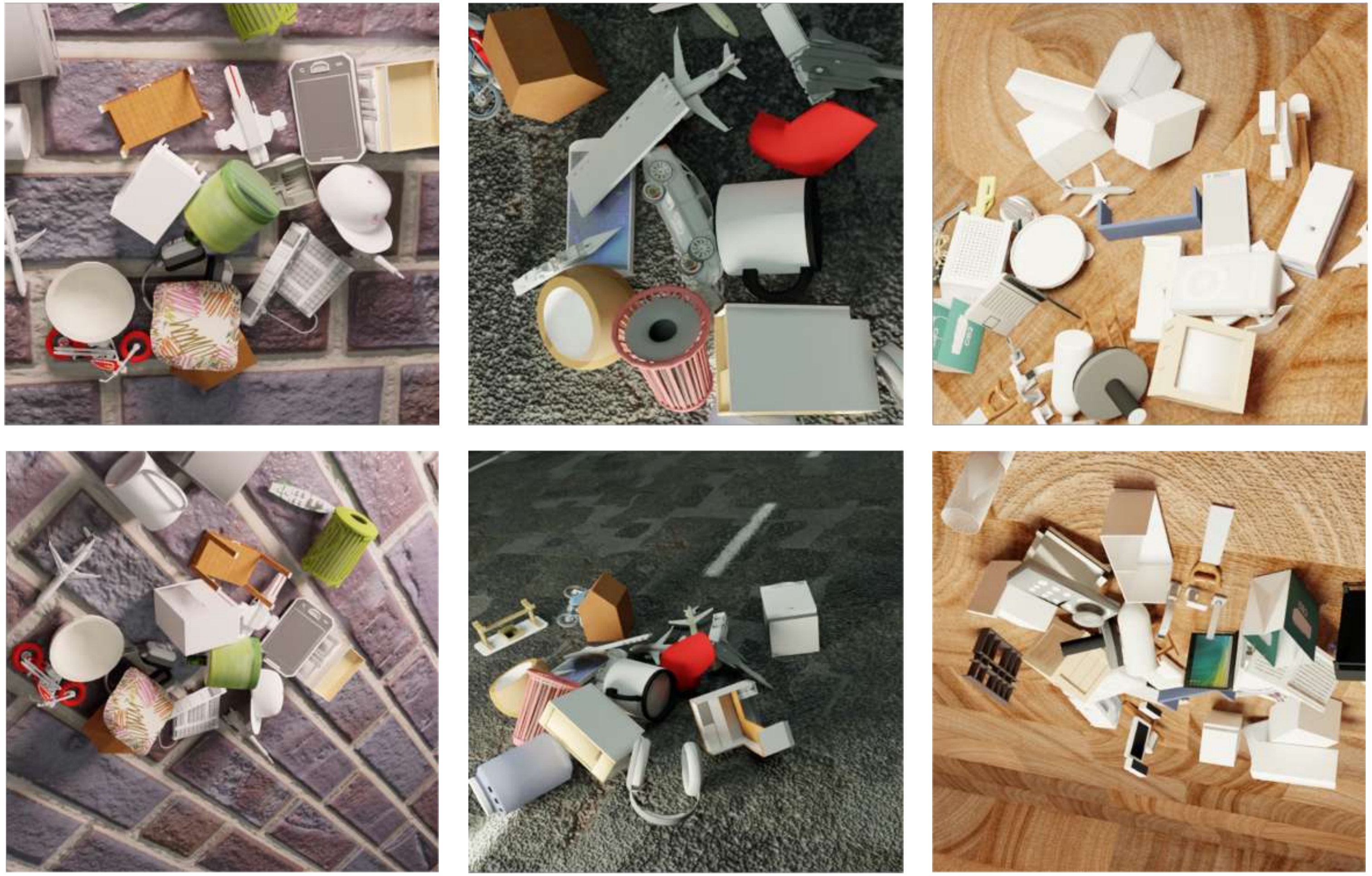
2. Coarse estimation




3. Refinement



# 1. Generate training dataset



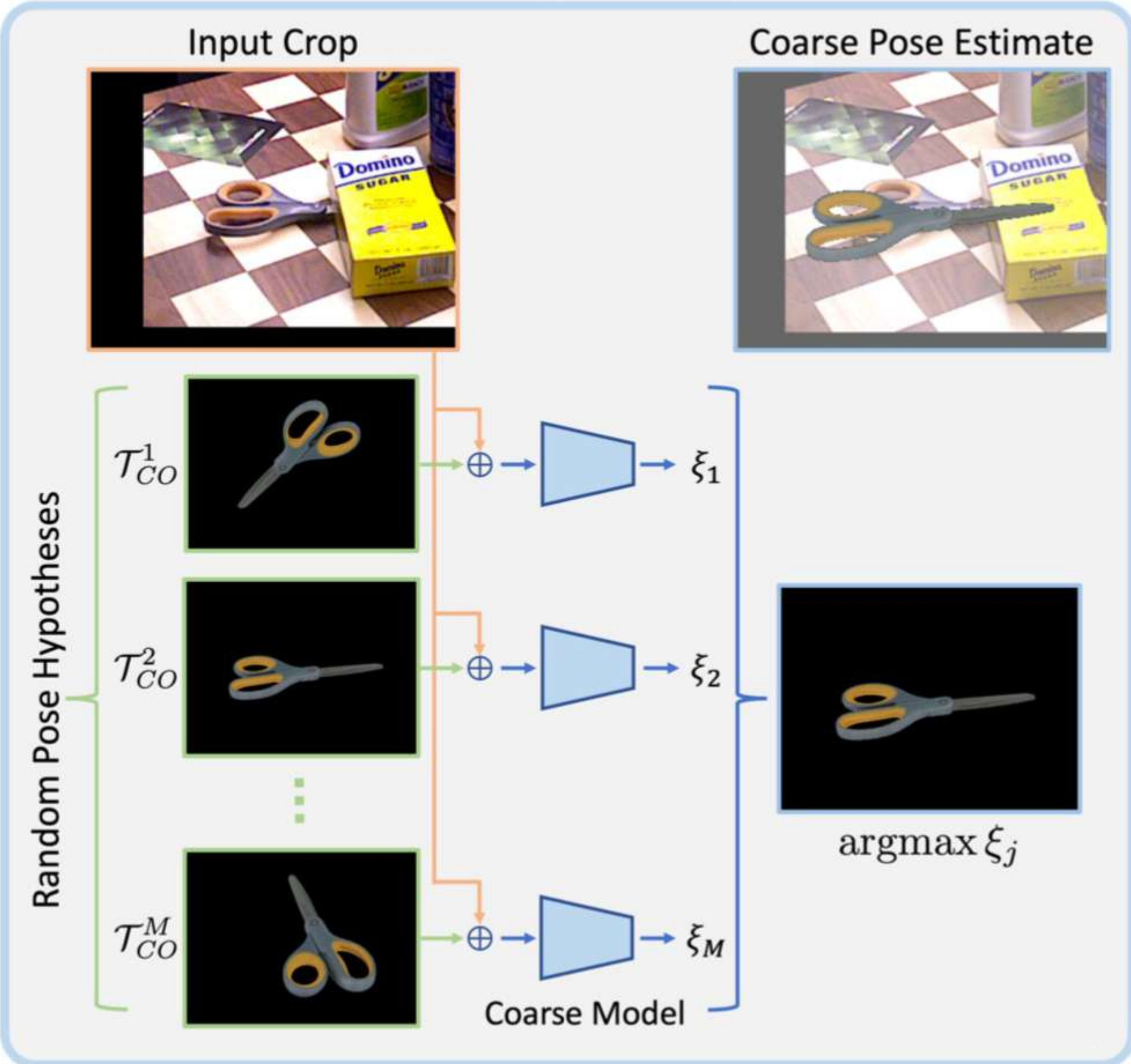
**Synthetic Dataset:**
- 2 million images,
- 50k scenes,
- 40 objects / scene,
- sampled from a database of 50,000 objects
- ShapeNet & GoogleScannedObjects
- Generated using BlenderProc2 renderer

[Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, J. Sivic, CoRL 2022.].

# 1. Generate training dataset


[Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, J. Sivic, CoRL 2022.].

# 2. Coarse estimation

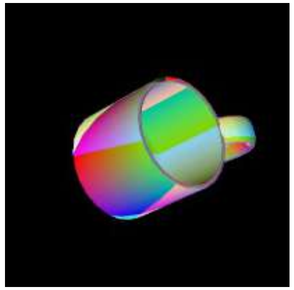


# 2. Coarse estimation: training

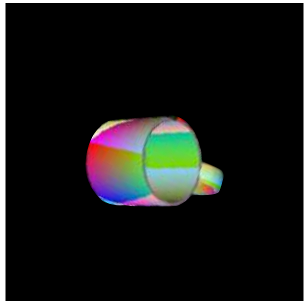
Image

Sample observation

Ground truth pose

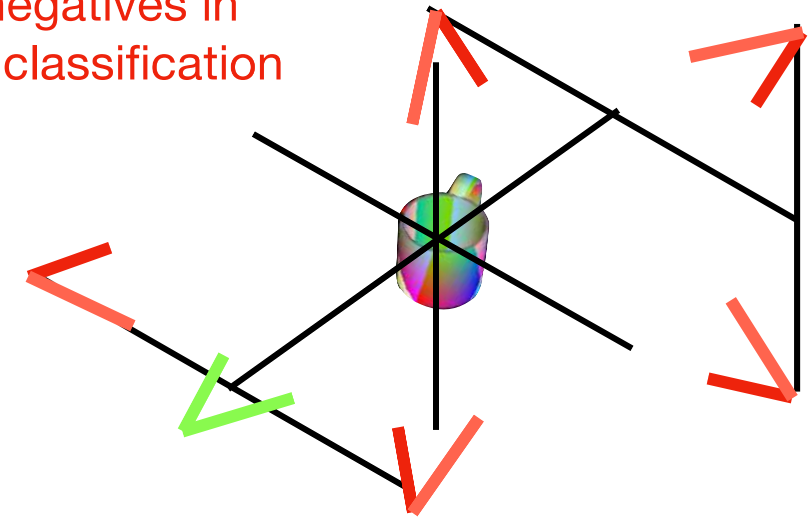


Ground truth + noise

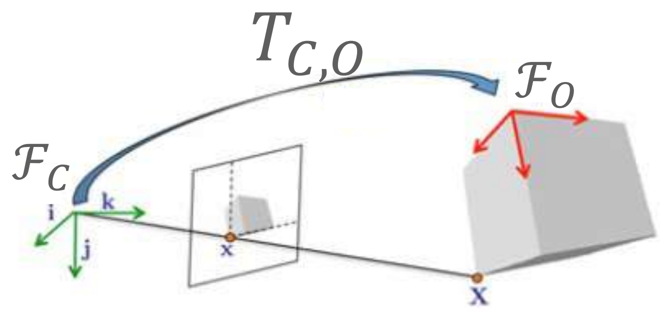


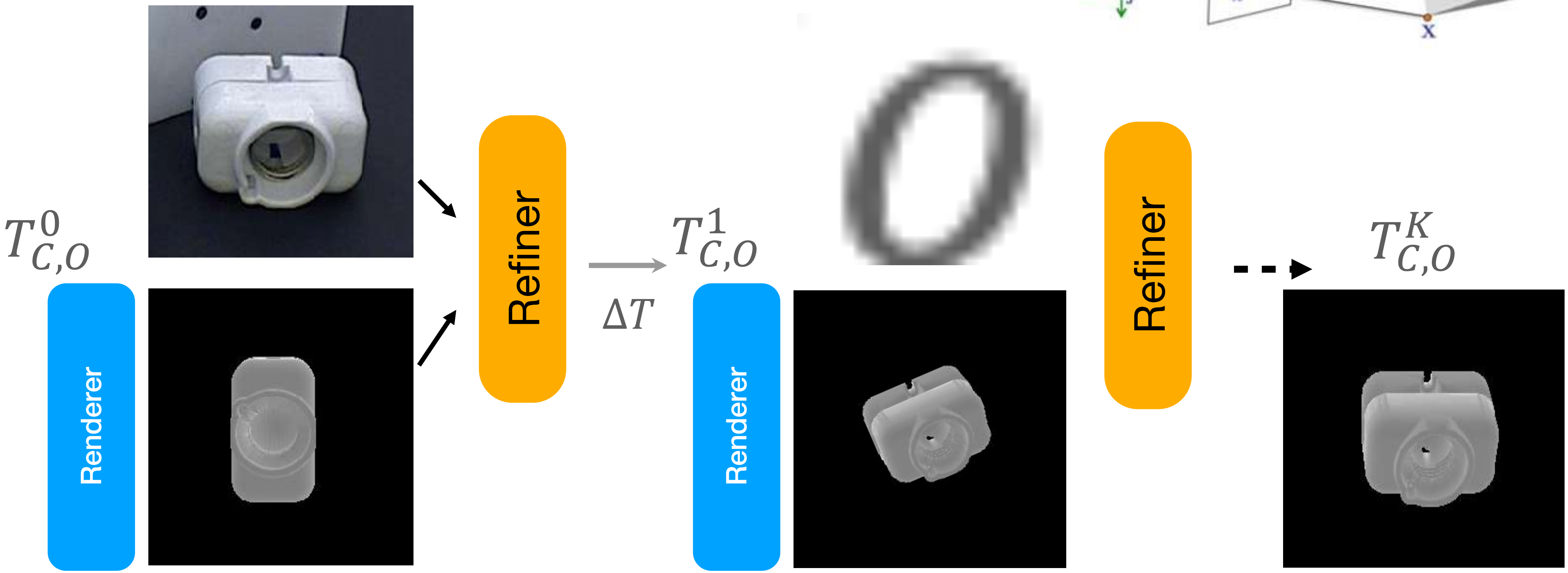
Positive in binary classification

(26) negatives in binary classification



# 3. Refiner: render & compare

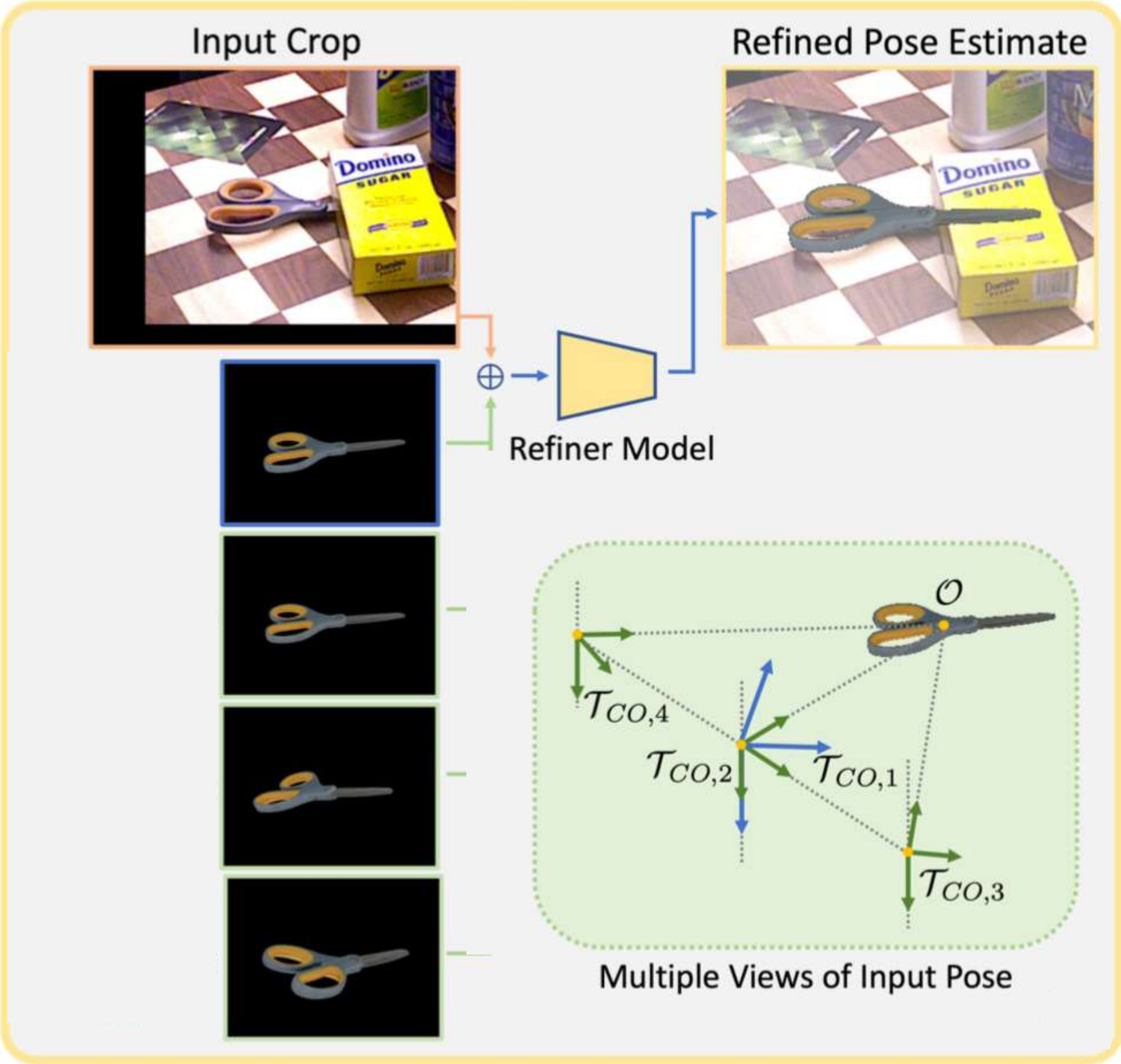



Training a Feedback Loop for Hand Pose Estimation, Oberweger et al, ICCV 2015
BB8, Rad et Lepetit, ICCV 2017
Deep model-based 6d pose refinement in rgb, Manhardt et al, ECCV 2018
DeepIM, Li et al, ECCV 2018
CosyPose, Labbé et al, ECCV 2020

# 3. Refiner: render & compare with **multiple views**

56



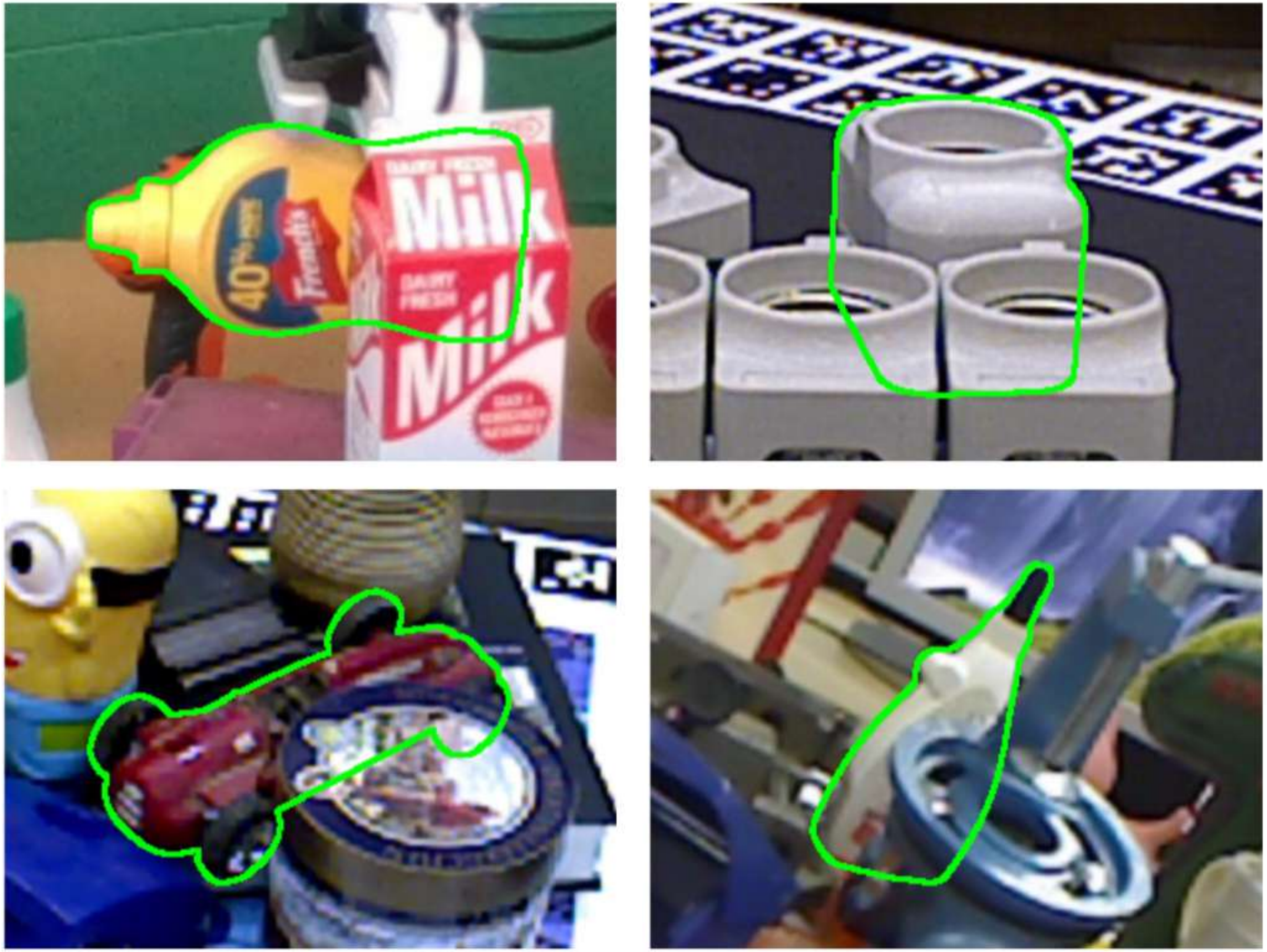
# Experiments:
# Benchmark for 6D Object Pose Estimation
# (BOP Challenge – 7 datasets)







# Example results




[Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, J. Sivic, CoRL 2022.]

# Experiments: Benchmark for 6D Object Pose Estimation (BOP Challenge – 7 datasets)

| | Pose Initialization | | Pose Refinement | | | BOP Datasets | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | Novel objects | Method | Novel objects | RGB-D Input | LM-O | T-LESS | TUD-L | IC-BIN | ITODD | HB | YCB-V | Mean |
| 4 | OSOP [16] | ✓ | Multi-Hyp. | ✓ | | 31.2 | - | - | - | - | 49.2 | 33.2 | - |
| 5 | OSOP [16] | ✓ | MH+ICP | ✓ | ✓ | 48.2 | - | - | - | - | 60.5 | 57.2 | - |
| 6 | (PPF, Sift) + Zephyr [20] | ✓ | - | ✓ | ✓ | **59.8** | - | - | - | - | - | 51.6 | - |
| 7 | (PPF, Sift) + Our coarse | ✓ | Our refiner | ✓ | ✓ | 57.0 | - | - | - | - | - | 62.3 | - |
| 12 | Ours | ✓ | Ours | ✓ | | 53.7 | **62.2** | 58.4 | **43.6** | 30.1 | 72.9 | 60.4 | 54.5 |
| 13 | Ours | ✓ | Ours | ✓ | ✓ | 58.3 | 54.3 | **71.2** | 37.1 | **40.4** | **75.7** | **63.3** | **57.2** |
| | | | | | RGB | **67.3** | **79.7** | **75.1** | **53.3** | | | **70.0** | **69.1** |
| | | | | | RGBD | **72.9** | **74.1** | **91.2** | **58.5** | | | **85.7** | **76.5** |

See also: https://bop.felk.cvut.cz/challenges/

**The Best Open-Source Method i BOP 2023.**

# Example results

# Example results

**Predicted 6D pose of the novel object: contour**

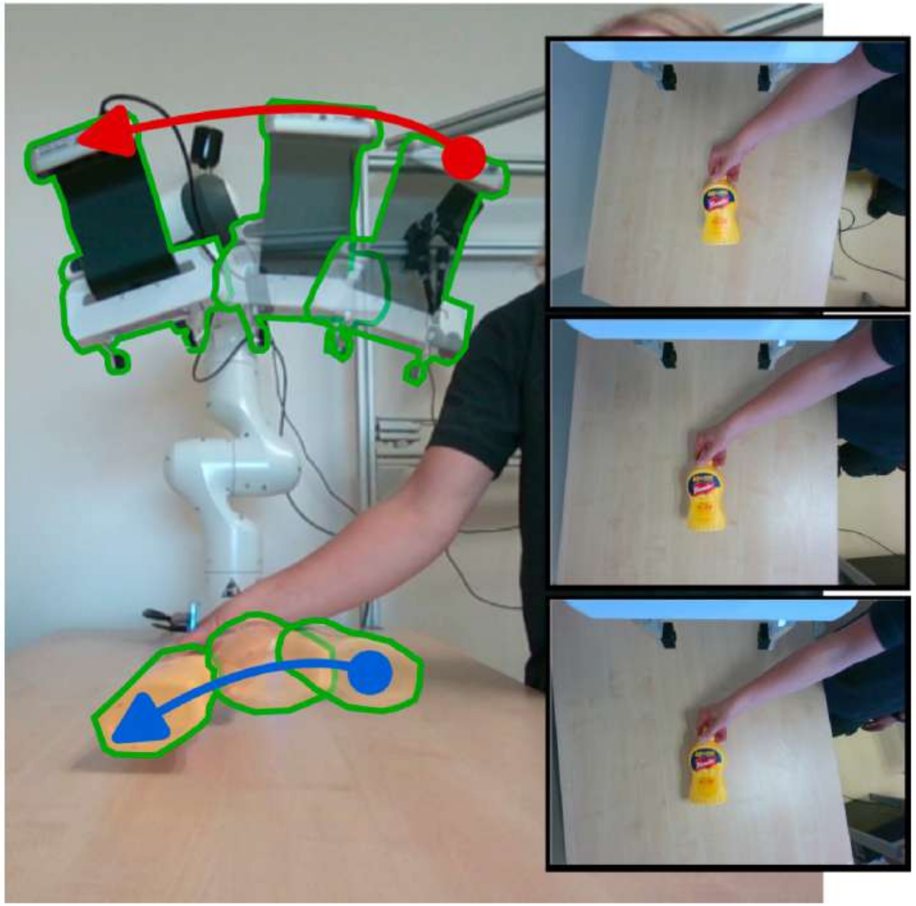
**Predicted 6D pose of the novel object: 3D model**


**MegaPose: 6d pose estimation of novel objects via render & compare**

Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, J. Sivic

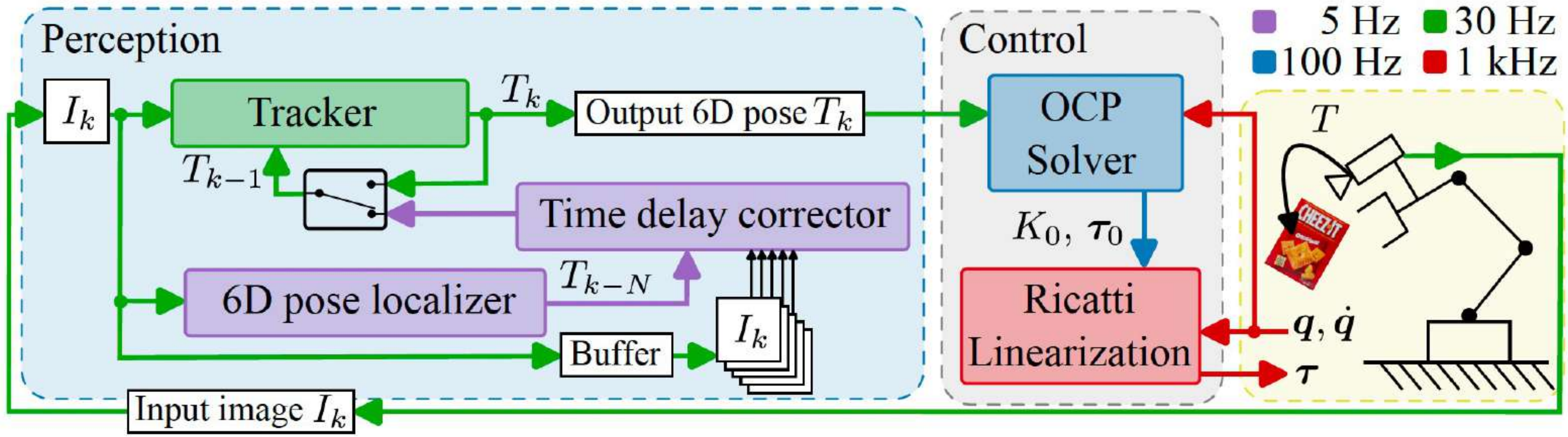
CoRL 2022

# Example application: visually guided control



Fourmy et al., Visually Guided Model Predictive Robot Control via 6D Object Pose Localization and Tracking
https://arxiv.org/pdf/2311.05344.pdf

# Example application: visually guided control



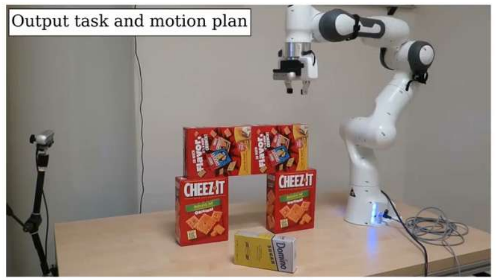

Fourmy et al., Visually Guided Model Predictive Robot Control via 6D Object Pose Localization and Tracking
https://arxiv.org/pdf/2311.05344.pdf

[redacted]

# Outline

Learning manipulation skills from videos
[Zorina et al., IEEE RA-L 2022]

Pre-training for visually guided manipulation
[Labbe et al., ECCV 2020, Labbe et al., CVPR 2021, Labbe et al., CoRL 2022, Fourmy et al., 2023]




Multi-contact task and motion planning guided by video demonstration
[Zorina et al., ICRA 2023]

Toward learning reward functions from videos
[Soucek et al., CVPR 2022], [Soucek et al., PAMI 2024], [Soucek et al., CVPR 2024]

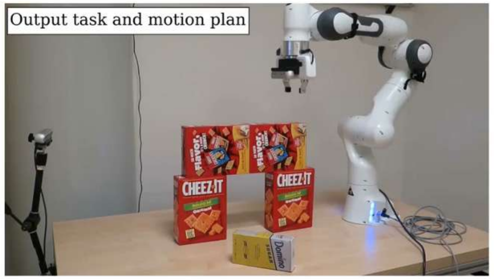
# Outline

Learning manipulation skills from videos
[Zorina et al., IEEE RA-L 2022]

Pre-training for visually guided manipulation
[Labbe et al., ECCV 2020, Labbe et al., CVPR 2021, Labbe et al., CoRL 2022, Fourmy et al., 2023]

Kateryna Zorina

Multi-contact task and motion planning guided by video demonstration
[Zorina et al., ICRA 2023]




Toward learning reward functions from videos
[Soucek et al., CVPR 2022], [Soucek et al., PAMI 2024],
[Soucek et al., CVPR 2024]

# Multi-Contact Task and Motion Planning Guided by Video Demonstration

Kateryna Zorina ♣ David Kovar ♣ Florent Lamiraux ◊ Nicolas Mansard ◊
Justin Carpentier ♥ Josef Sivic ♣ Vladimir Petrik ♣


[Zorina et al., ICRA 2023]

# Multi-Contact Task and Motion Planning Guided by Video Demonstration

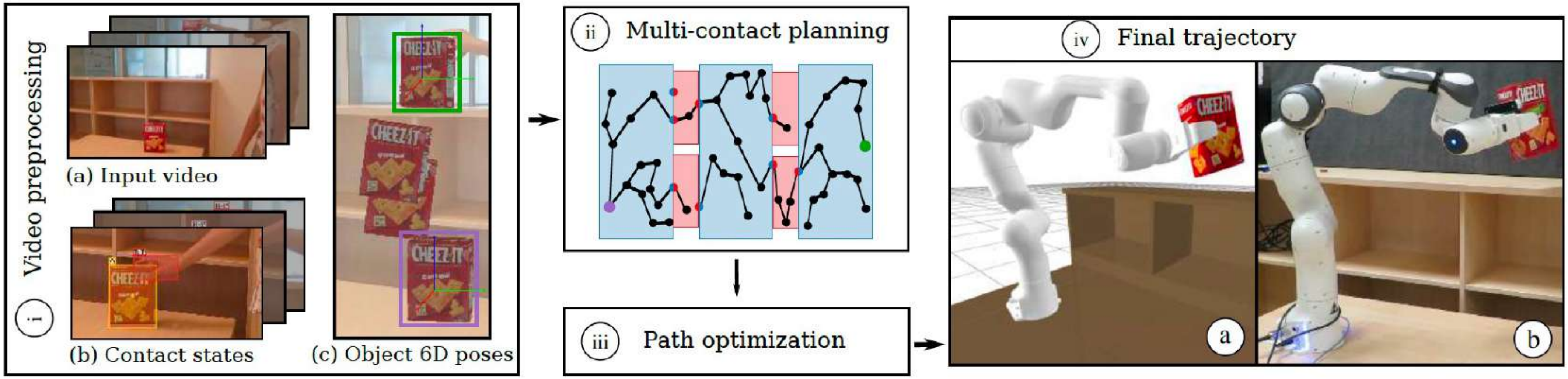
Kateryna Zorina ♣ David Kovar ♣ Florent Lamiraux ♢ Nicolas Mansard ♢
Justin Carpentier ♥ Josef Sivic ♣ Vladimir Petrik ♣



♣ CIIRC, Czech Technical University in Prague
♢ LAAS-CNRS, Universite de Toulouse, CNRS, Toulouse
♥ INRIA, Paris

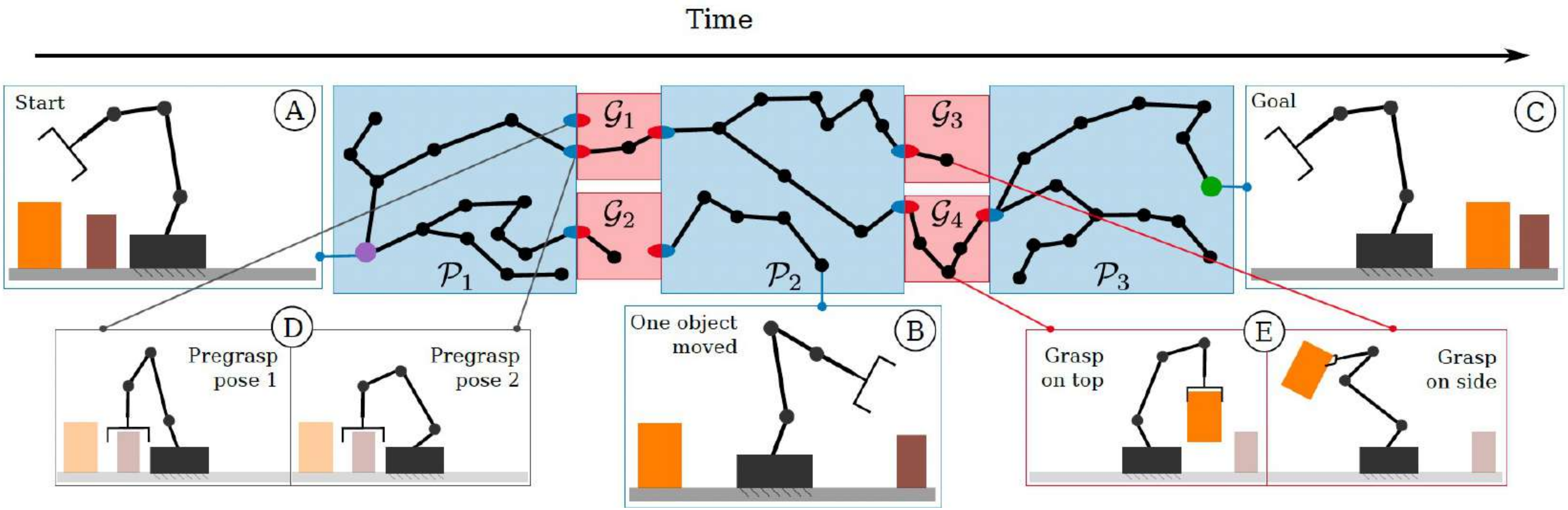
**[Zorina et al., ICRA 2023]**

# Approach overview



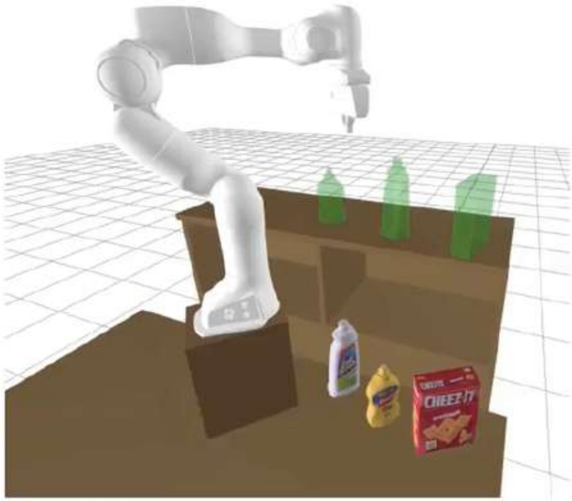
# Multi-contact planning guided by video demonstration

- Extension of **Rapidly-Exploring Random Tree (RRT)** planner
- Simultaneously **grow multiple trees** around **grasp and release states** extracted from the guiding video.

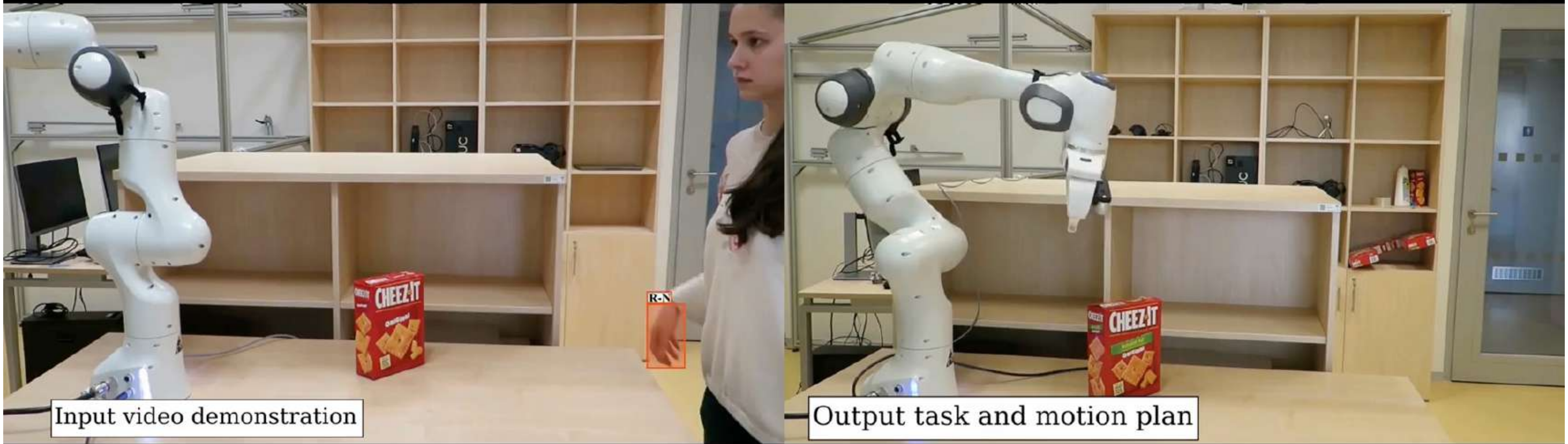


# Benchmark

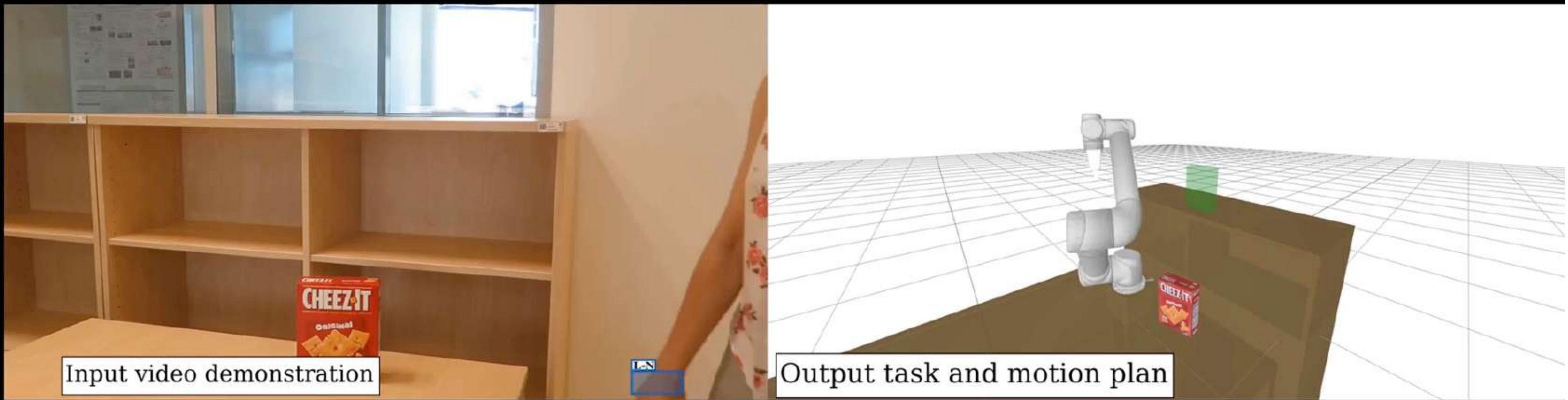
## Shelf task



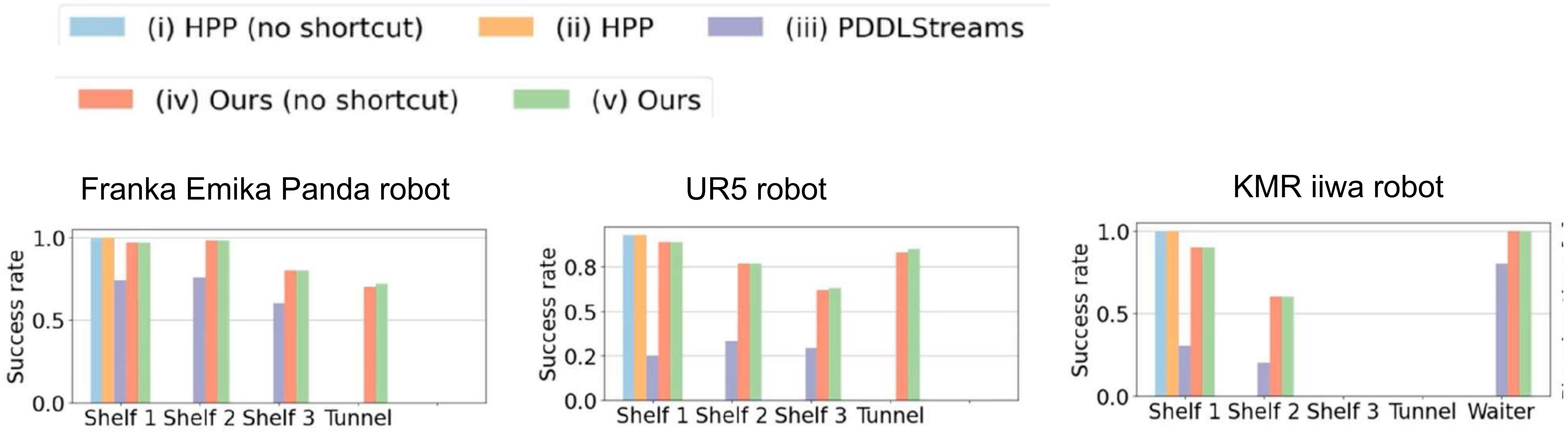
# More results with the Panda robot



# Results with the other robots (in simulation)

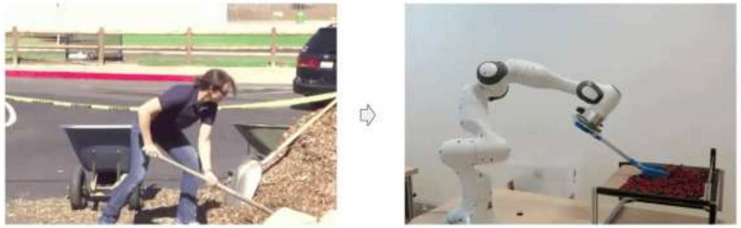


# Quantitative results



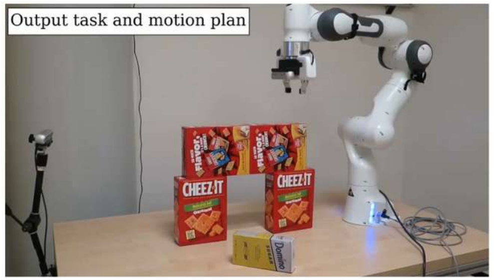
# Outline

## Learning manipulation skill from videos
[Zorina et al., IEEE RA-L 2022]



## Pre-training for visually guided manipulation
[Labbe et al., ECCV 2020, Labbe et al., CVPR 2021, Labbe et al., CoRL 2022]

## Multi-contact task and motion planning guided by video demonstration
[Zorina et al., ICRA 2023]




## Toward learning reward functions from videos
[Soucek et al., CVPR 2022], [Soucek et al., PAMI 2024], [Soucek et al., CVPR 2024]

# Outline

Learning manipulation skill from videos
[Zorina et al., IEEE RA-L 2022]

Pre-training for visually guided manipulation
[Labbe et al., ECCV 2020, Labbe et al., CVPR 2021, Labbe et al., CoRL 2022]

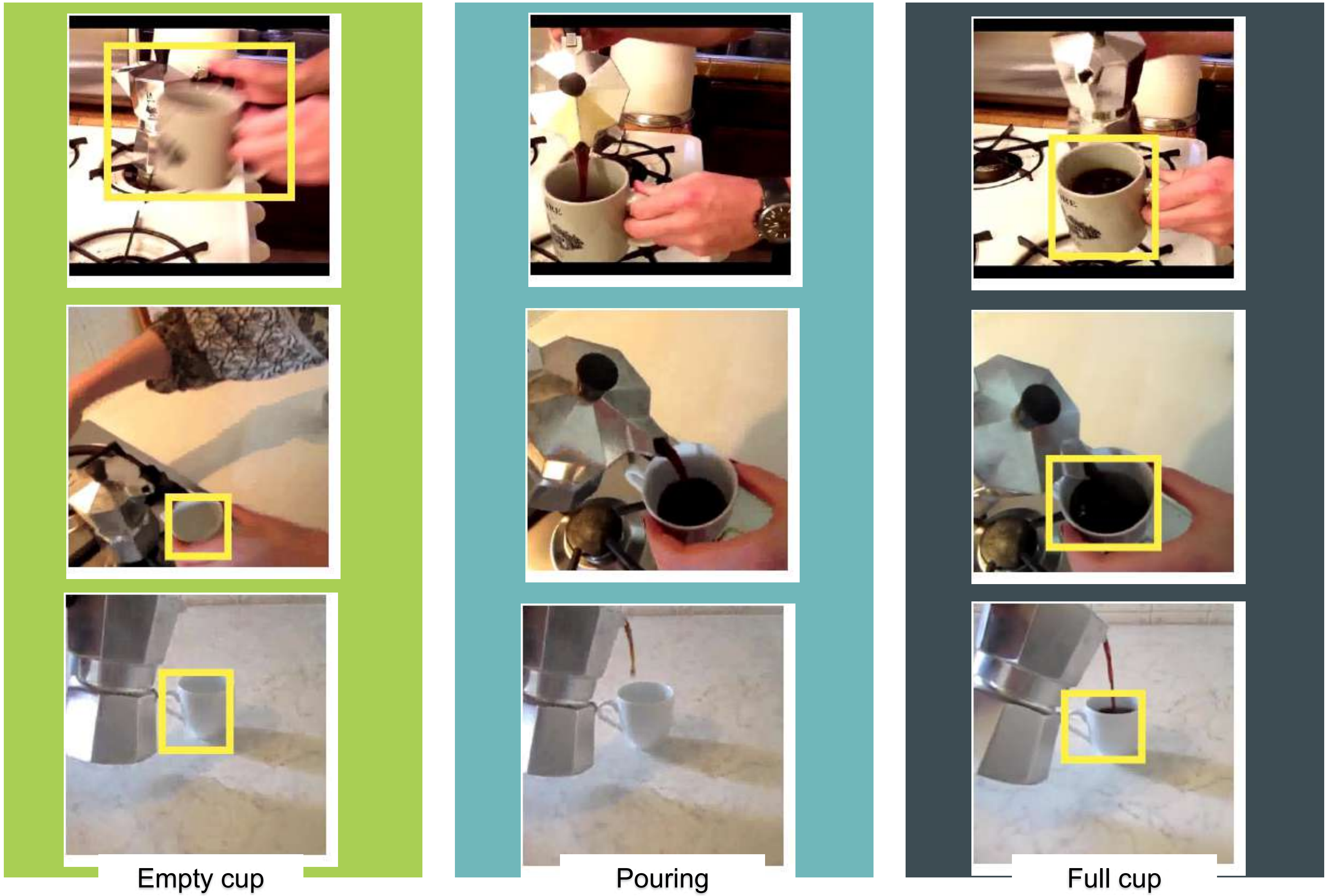
Multi-contact task and motion planning guided by video demonstration
[Zorina et al., ICRA 2023]

Tomas Soucek

Toward learning reward functions from videos
[Soucek et al., CVPR 2022], [Soucek et al., PAMI 2024], [Soucek et al., CVPR 2024]

# Learn how actions change states of objects

Pour coffee

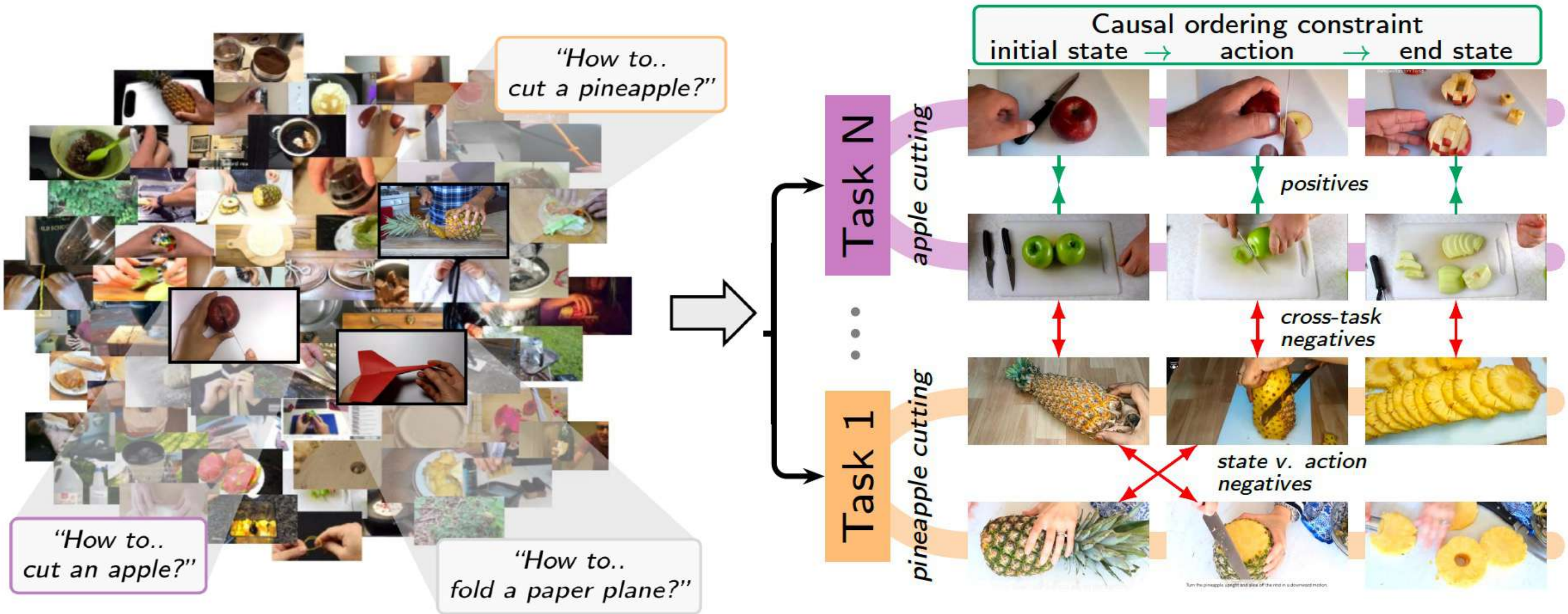



[Alayrac et al., ICCV 2017]

# Goal

**Input:** videos with noisy video-level labels

**Output:** temporal localization of object states and state-modifying actions

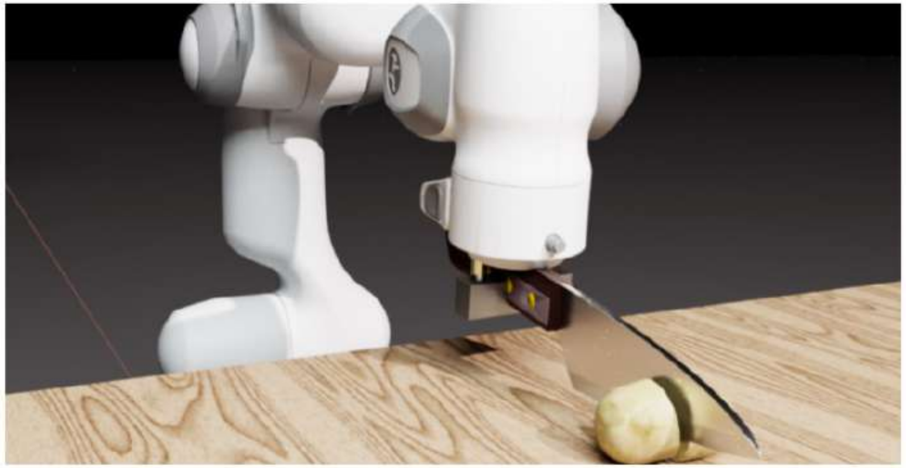



[**Tomas Soucek**, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic
Multi-Task Learning of Object States and State-Modifying Actions from Web Videos, CVPR 2022, PAMI 2022]

# Motivation: embodied perception

**Video demonstration**

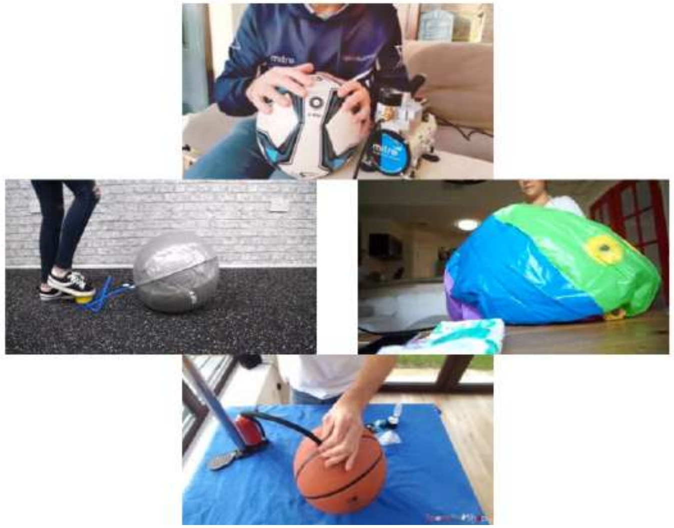
**Robot performing the action in a new environment**



See e.g., [E. Heiden et al. Disect: A differentiable simulation engine for autonomous robotic cutting. In Robotics: Science and Systems, 2021.]

# Challenges

**Visual variability**

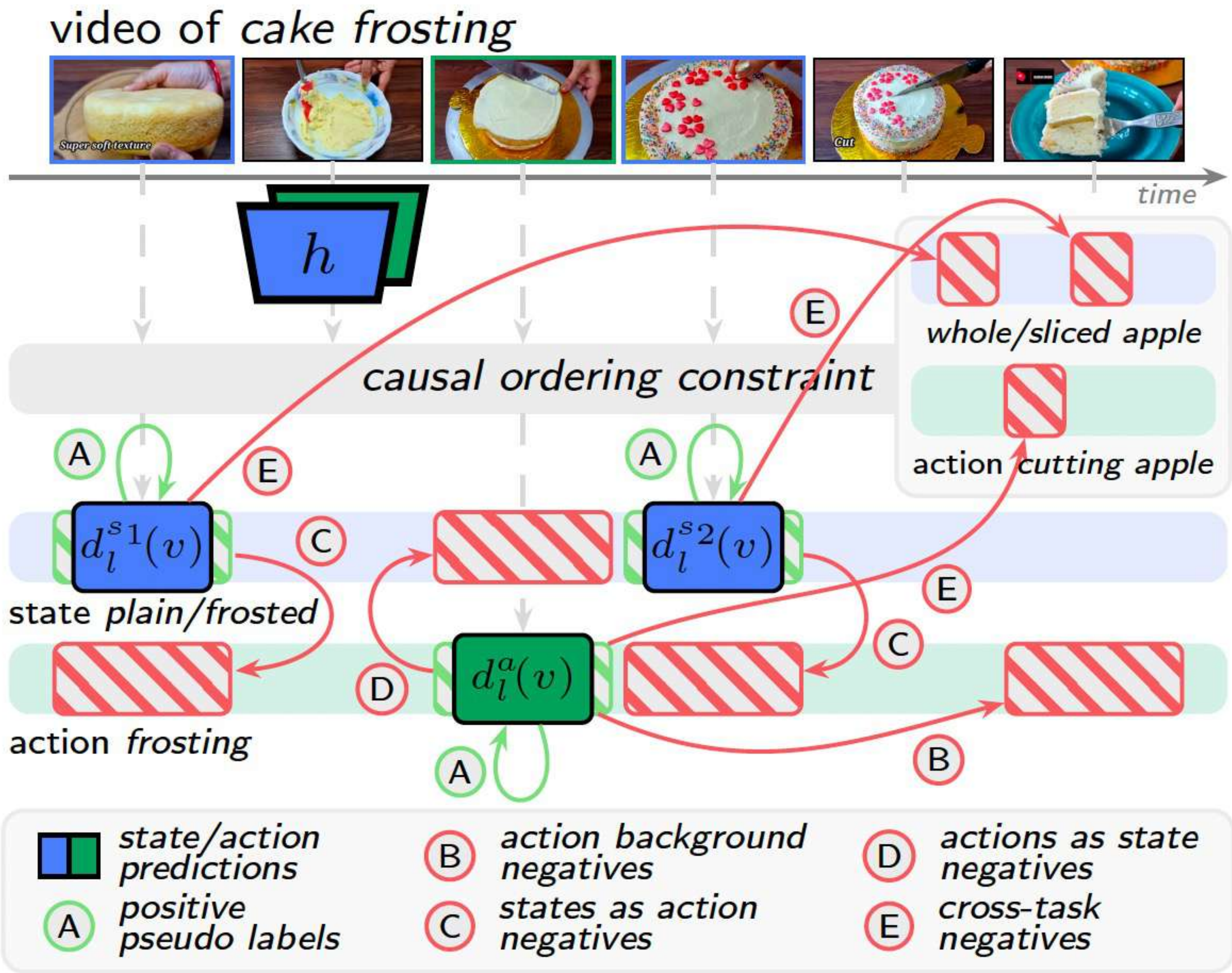


**Long videos**

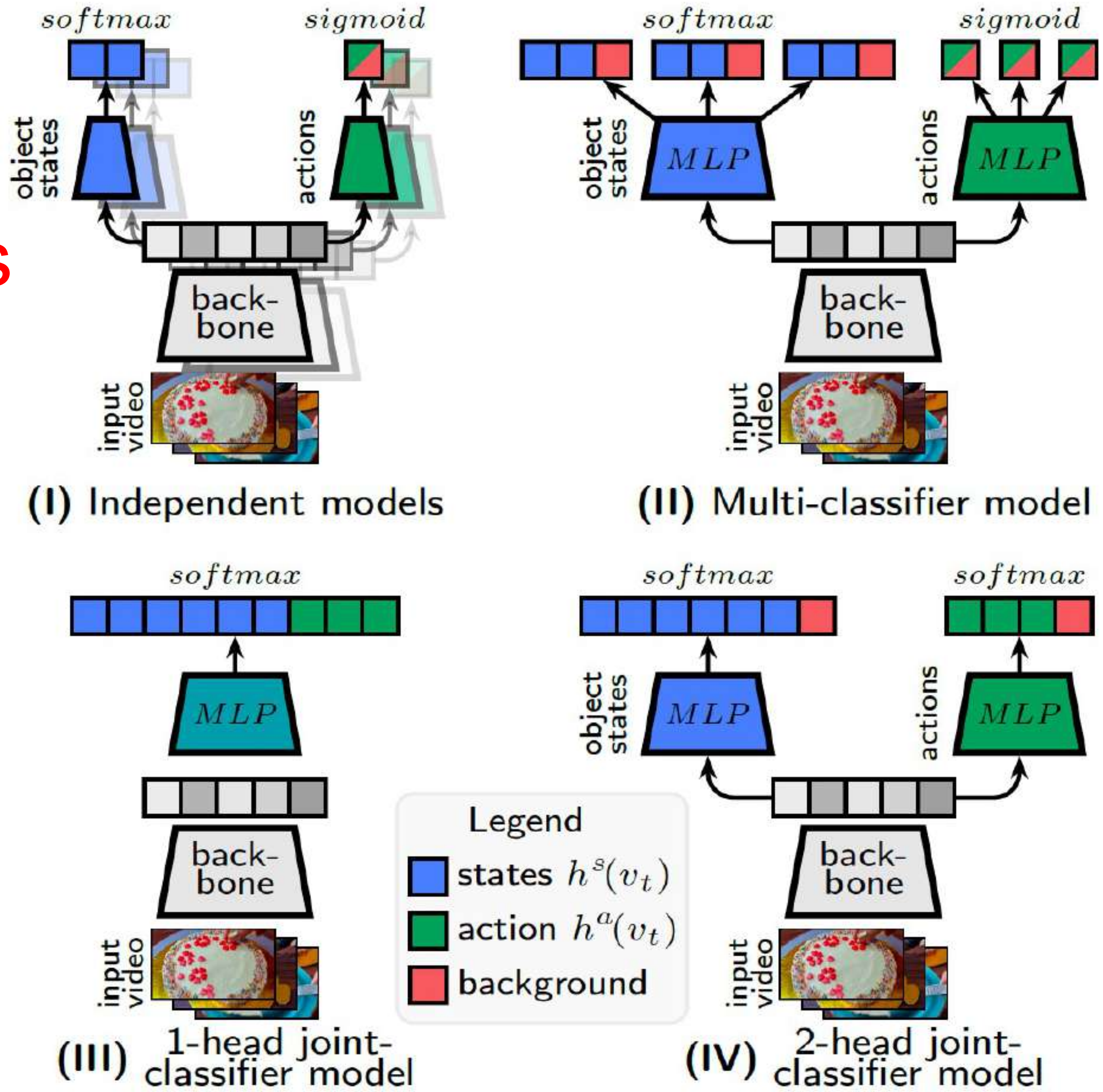

**In-the-wild, uncurated, noisy data**

See e.g., [E. Heiden et al. Disect: A differentiable simulation engine for autonomous robotic cutting. In Robotics: Science and Systems, 2021.]

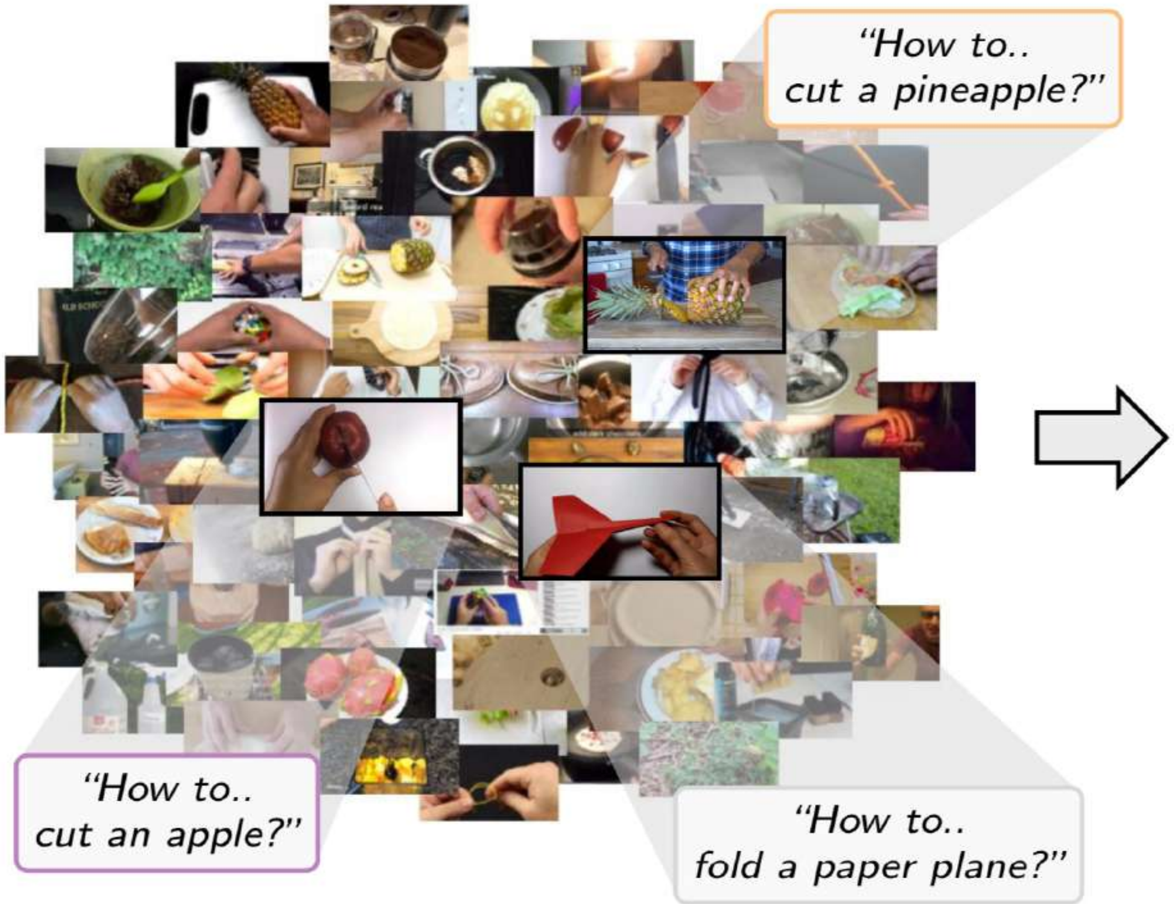
Contribution 1: Constraints for self-supervised learning



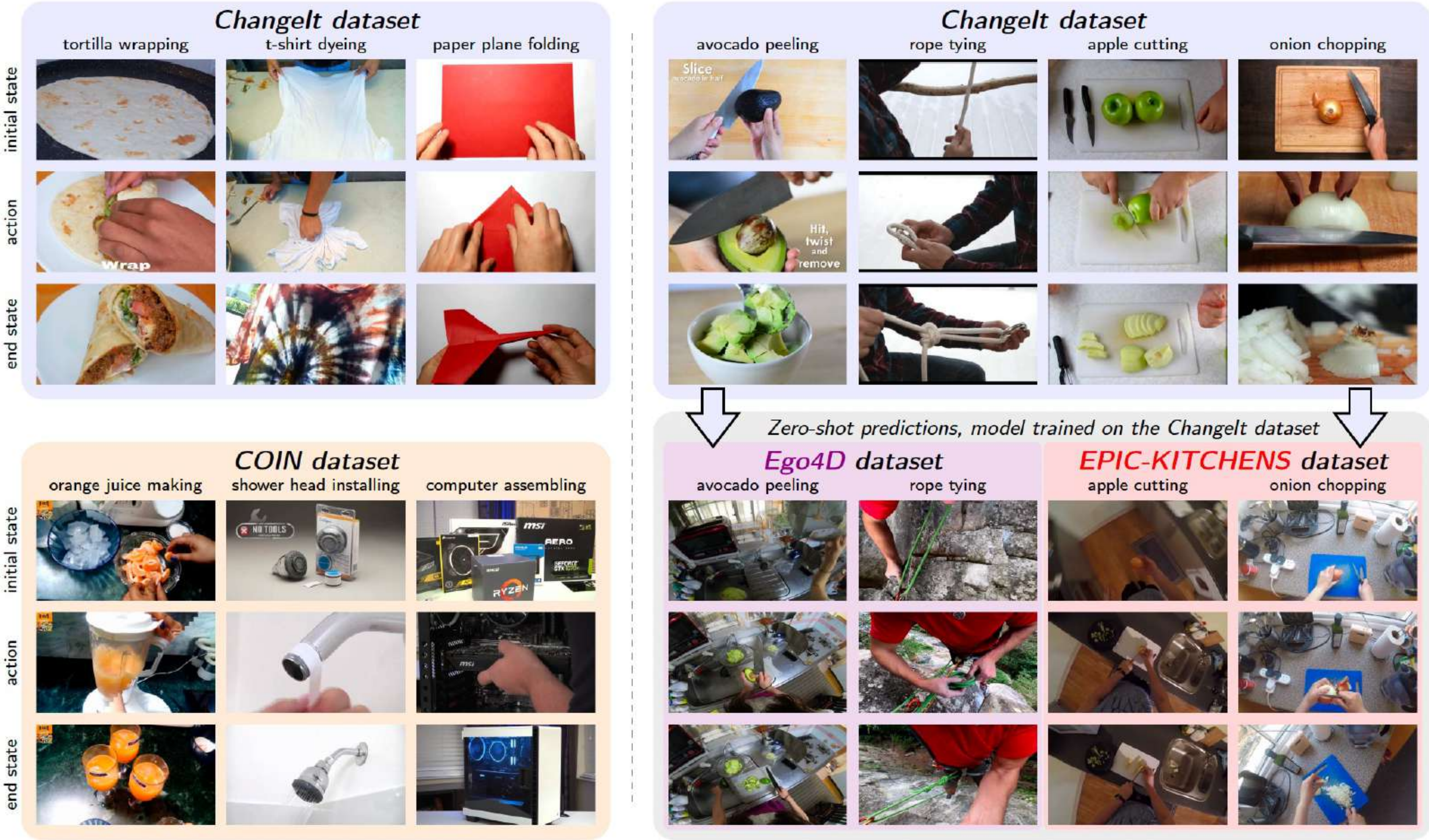
# Contribution 2: Investigate multi-task architectures



# ChangeIt dataset




- **44 interactions** such as "How to cut an apple?"
- **34,000+** videos, **2600+** hours
- Up to **15mins** long, **4.6mins** on average
- Auto-annotated with the **noisy video-level** category label
- **667** videos manually annotated with **temporal labels**.


ChangeIt dataset
tortilla wrapping
t-shirt dyeing
paper plane folding
initial state
action
end state
Wrap
ChangeIt dataset
avocado peeling
rope tying
apple cutting
onion chopping
Slice
Hit, twist and remove
COIN dataset
orange juice making
shower head installing
computer assembling
initial state
action
end state
NO TOOLS
msi
AERO
RYZEN
Zero-shot predictions, model trained on the ChangeIt dataset
Ego4D dataset
avocado peeling
rope tying
EPIC-KITCHENS dataset
apple cutting
onion chopping

# Results

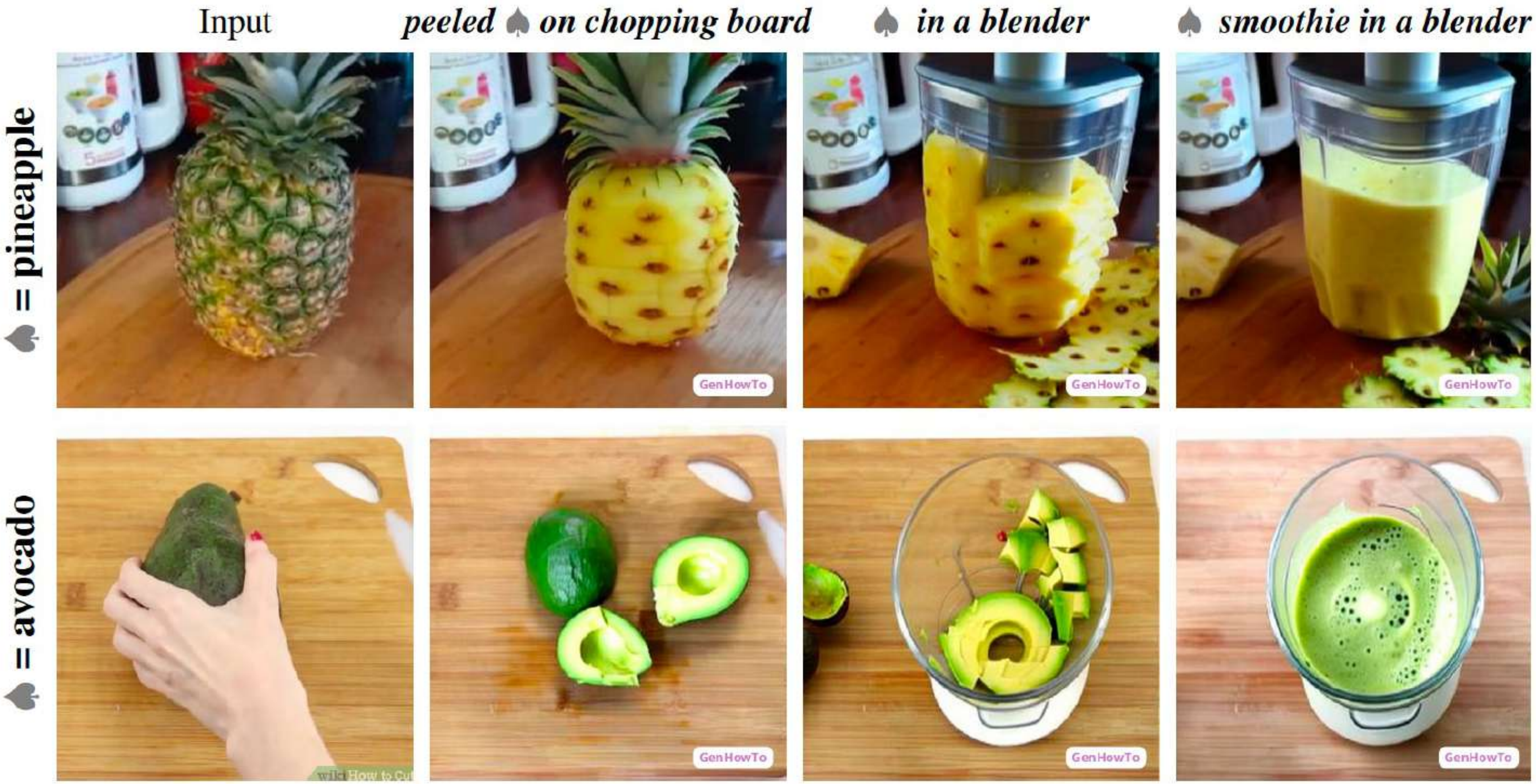
**Performance metric**: percentage of correctly (temporally) localized actions and object states

| Method | ChangeIt St prec. | ChangeIt Ac prec. | COIN [79] Ac prec. |
|---|---|---|---|
| Random | 0.15 | 0.41 | 0.42 |
| Merlot Reserve [93] | 0.27 | 0.57 | 0.69 |
| CLIP ViT-L/14 [66] | 0.30 | <u>0.63</u> | 0.65 |
| VideoCLIP [87] | <u>0.33</u> | 0.59 | <u>0.72</u> |
| Alayrac *et al.* [2] | 0.30 | 0.59 | 0.57 |
| Look for the Change [77] | <u>0.35</u> | <u>0.68</u> | <u>0.73</u> |
| **Ours** (backbone from [77]) | 0.47 | 0.77 | 0.79 |
| **Ours** (ViT-L/14 frozen) | 0.47 | 0.75 | 0.77 |
| **Ours** (ViT-L/14 finetuned) | **0.49** | **0.80** | **0.83** |

| Method | EPIC-K. [16] St mAP | EPIC-K. [16] Ac mAP | Ego4D [33] St mAP | Ego4D [33] Ac mAP |
|---|---|---|---|---|
| Random | 0.09 | 0.07 | 0.13 | 0.12 |
| Merlot Reserve [93] | <u>0.31</u> | 0.36 | <u>0.25</u> | 0.45 |
| CLIP ViT-L/14 [66] | 0.23 | 0.35 | 0.23 | 0.42 |
| VideoCLIP [87] | 0.25 | <u>0.44</u> | 0.23 | **0.49** |
| Look for the Ch. [77]$^\dagger$ | 0.12 | 0.15 | 0.20 | 0.17 |
| **Ours** (ViT-L/14 frozen)$^\dagger$ | **0.38** | <u>0.47</u> | <u>0.33</u> | <u>0.46</u> |
| **Ours** (ViT-L/14)$^\dagger$ | **0.38** | **0.51** | **0.37** | <u>0.48</u> |

$^\dagger$ Trained on the ChangeIt dataset, zero-shot evaluation.

# GenHowTo: Generate changes of object states



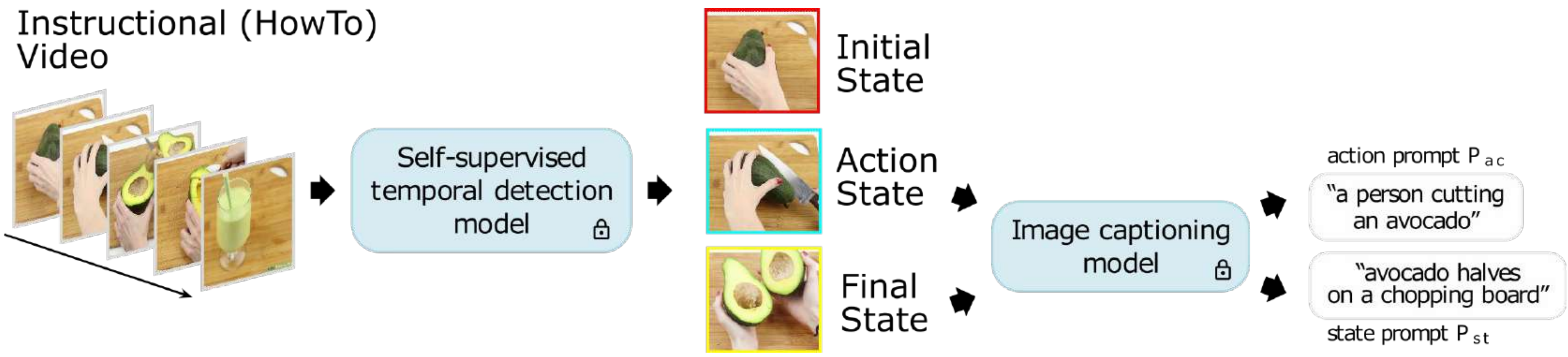

[**Tomas Soucek**, Dima Damen, Michael Wray, Ivan Laptev and Josef Sivic
GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos, CVPR 2024]

Challenges:

1. Change the object

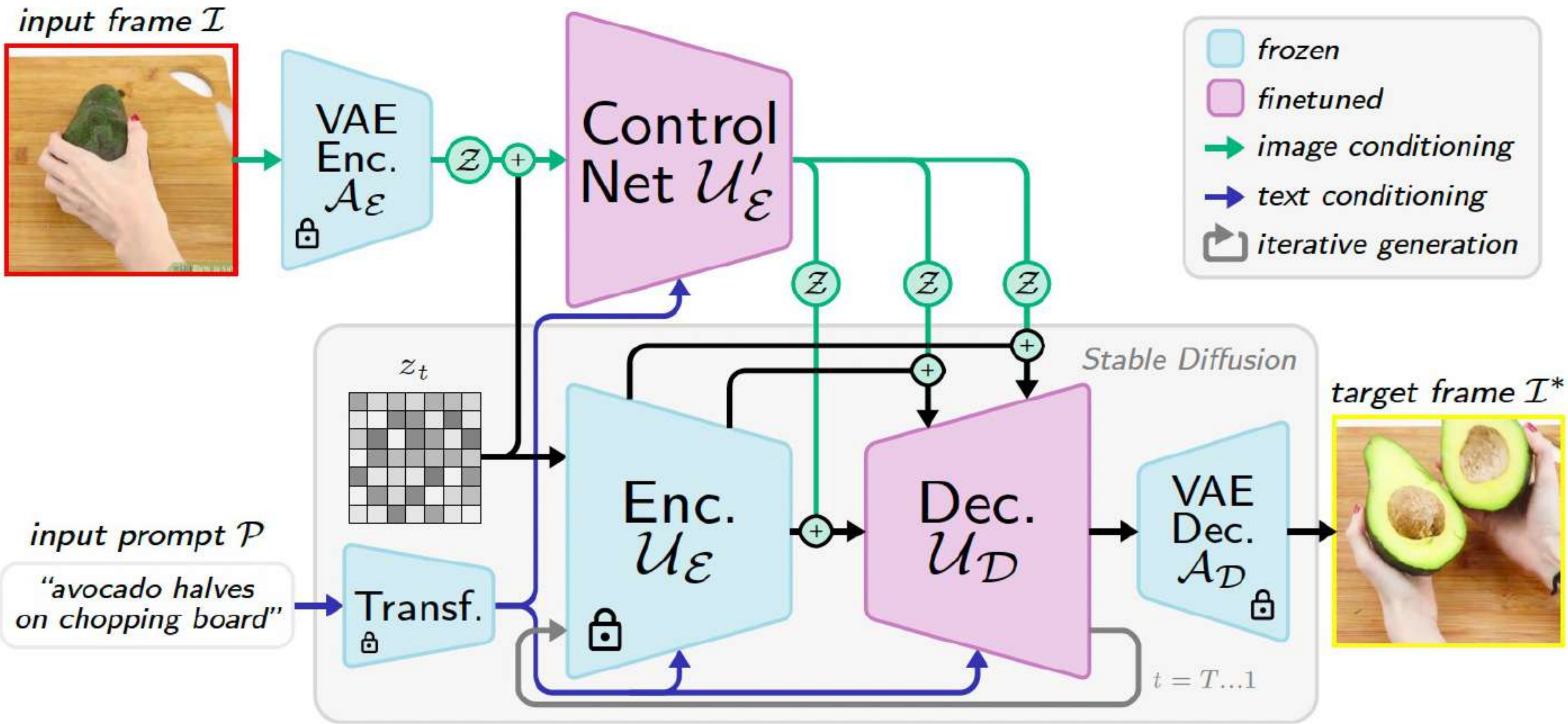
2. Keep the scene context

# Contribution 1: Dataset of annotated image triplets




[Tomas Soucek, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Multi-task learning of object state changes from uncurated videos, PAMI 2024.]

# Contribution 2: Method



# Contribution 2: Method

Preserves the scene while changing the object state

# Experiments: quantitative evaluation

| | Method | $\text{Acc}_{\text{ac}} \uparrow$ | $\text{Acc}_{\text{st}} \uparrow$ |
|---|---|---|---|
| | *test set categories unseen during training* | | |
| **(a)** | Stable Diffusion | 0.51 | 0.50 |
| **(b)** | Edit Friendly DDPM | 0.60 | 0.61 |
| **(c)** | InstructPix2Pix | 0.55 | 0.63 |
| **(d)** | *CLIP (manual prompts)* | 0.52 | 0.62 |
| **(e)** | **GenHowTo** | **0.66** | **0.74** |
| | *test set categories seen during training* | | |
| **(f)** | Edit Friendly DDPM$^{\dagger}$ | 0.69 | 0.80 |
| **(g)** | **GenHowTo**$^{\dagger}$ | **0.77** | **0.88** |
| **(h)** | *Real images* | 0.96 | 0.97 |

$^{\dagger}$ Models trained also on the test set *categories*.

# Experiments: qualitative results

Generated action

*a person is wrapping a tortilla on a plate*


Generated object state

*a plate with two burritos on it*


Generated action

*a man pouring beer into a glass*

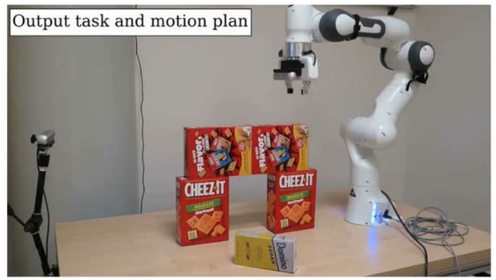

Generated object state

*a man sitting at a table holding a glass of beer*

# Summary

## Learning manipulation skills from videos
[Zorina et al., IEEE RA-L 2022]

## Pre-training for visually guided manipulation
[Labbe et al., ECCV 2020, Labbe et al., CVPR 2021, Labbe et al., CoRL 2022, Fourmy et al., 2024]




## Multi-contact task and motion planning guided by video demonstration
[Zorina et al., ICRA 2023]

## Toward learning reward functions from videos
[Soucek et al., CVPR 2022], [Soucek et al., PAMI 2024], [Soucek et al., CVPR 2024]