ELLIOT: European Large Open Multimodal Foundation Models for Scalable, Robust Generalization

Vladimir Petrik

IMPACT, CIIRC, CTU

May 22, 2025

Motivation

- LLMs shows remarkable performance
  - Generate human-like text
  - Translate languages
  - Write code
  - In depth analysis of related work

Motivation

- LLMs shows remarkable performance
  - Generate human-like text
  - Translate languages
  - Write code
  - In depth analysis of related work
- VLMs (Vision-Language Models) are also very powerful

openai.com/gpt-4

Motivation

- LLMs shows remarkable performance
  - Generate human-like text
  - Translate languages
  - Write code
  - In depth analysis of related work
- VLMs (Vision-Language Models) are also very powerful
- But the world is not just text or vision
  - Robot proprioception

openai.com/gpt-4

Motivation

- LLMs shows remarkable performance
  - Generate human-like text
  - Translate languages
  - Write code
  - In depth analysis of related work
- VLMs (Vision-Language Models) are also very powerful
- But the world is not just text or vision
  - Robot proprioception
  - RGBD cameras

openai.com/gpt-4

Motivation

- LLMs shows remarkable performance
  - Generate human-like text
  - Translate languages
  - Write code
  - In depth analysis of related work
- VLMs (Vision-Language Models) are also very powerful
- But the world is not just text or vision
  - Robot proprioception
  - RGBD cameras
  - Radar

openai.com/gpt-4

Motivation

- LLMs shows remarkable performance
  - Generate human-like text
  - Translate languages
  - Write code
  - In depth analysis of related work
- VLMs (Vision-Language Models) are also very powerful
- But the world is not just text or vision
  - Robot proprioception
  - RGBD cameras
  - Radar
  - **ELLIOT will develop the next generation of open Multimodal Generalist Foundation Models**

openai.com/gpt-4

ELLIOT: EUROPEAN LARGE OPEN MULTIMODAL FOUNDATION MODELS FOR SCALABLE ROBUST GENERALISATION

- ELLIOT is a project funded by the European Commission
- Call: Advancing Large AI Models: Integration of New Data Modalities and Expansion of Capabilities (RIA)
- Consortium: 32 partners across Europe
- Duration: 4 years (2025-2029)
- Budget: EUR 28.5M

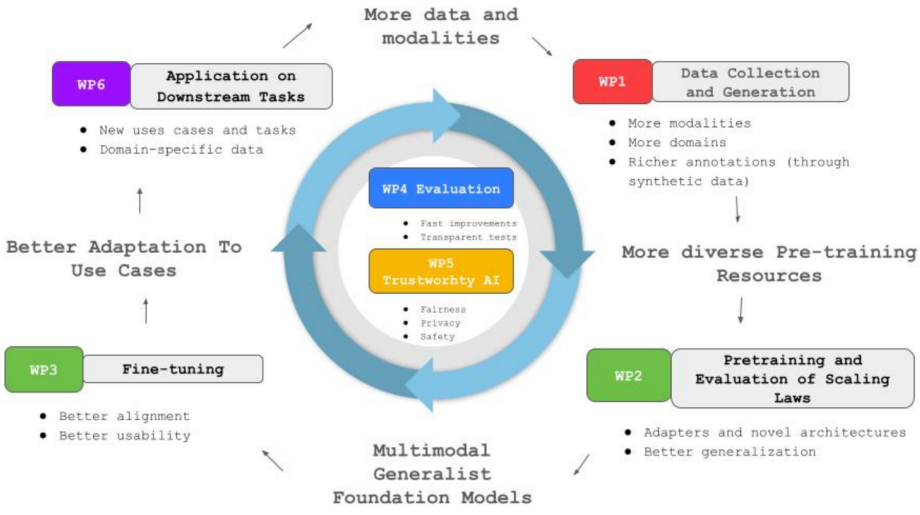
ELLIOT's objectives

- Strong, robust generalization
  - Key desirable property of foundation models
  - Robustness to distribution shifts
  - Generalization to unseen data and learning from them

ELLIOT's objectives

- Strong, robust generalization
  - Key desirable property of foundation models
  - Robustness to distribution shifts
  - Generalization to unseen data and learning from them
- Multimodality support
  - various input as well as output modalities
  - text, images, audio, video, 3D point clouds, proprioception
  - long sequences of multimodal data (temporal modality)
    - spatio-temporal audio-video data
    - watching educational videos
    - monitoring environmental changes

How to achieve it?



WP1: Data Collection and Generation

- ELLIOT's success relies on high-quality, diverse, and compliant data.

WP1: Data Collection and Generation

- ELLIOT's success relies on high-quality, diverse, and compliant data.
- Objectives of WP1:
  - Build a clean common crawl database
    - link to source data
    - licensing and regulatory content
    - transparency
    - methods for aligning data with AI/Data regulations

WP1: Data Collection and Generation

- ELLIOT's success relies on high-quality, diverse, and compliant data.
- Objectives of WP1:
  - Build a clean common crawl database
    - link to source data
    - licensing and regulatory content
    - transparency
    - methods for aligning data with AI/Data regulations
  - Framework for synthetic data generation at scale
    - physics-based simulations
    - generative models
    - multimodality and cross modal alignment

WP1: Data Collection and Generation

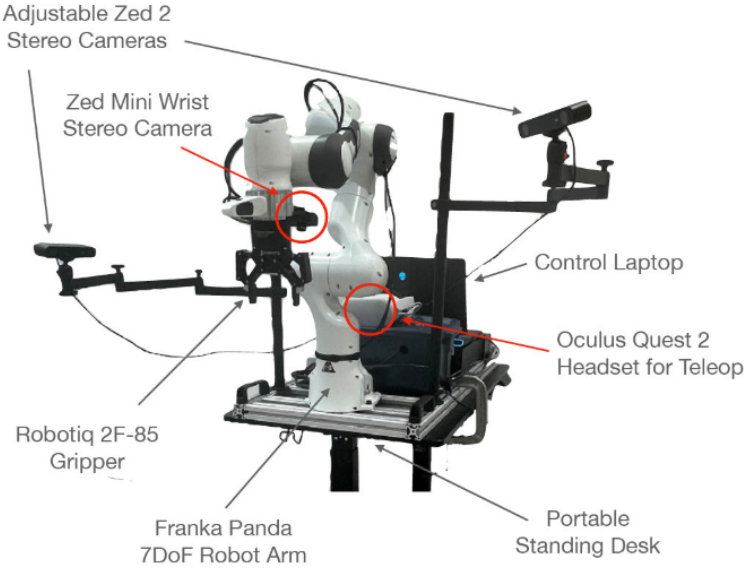
- ELLIOT's success relies on high-quality, diverse, and compliant data.
- Objectives of WP1:
  - Build a clean common crawl database
    - link to source data
    - licensing and regulatory content
    - transparency
    - methods for aligning data with AI/Data regulations
  - Framework for synthetic data generation at scale
    - physics-based simulations
    - generative models
    - multimodality and cross modal alignment
  - Proprietary data
    - domain-specific data

Data for most advanced robotics demos?

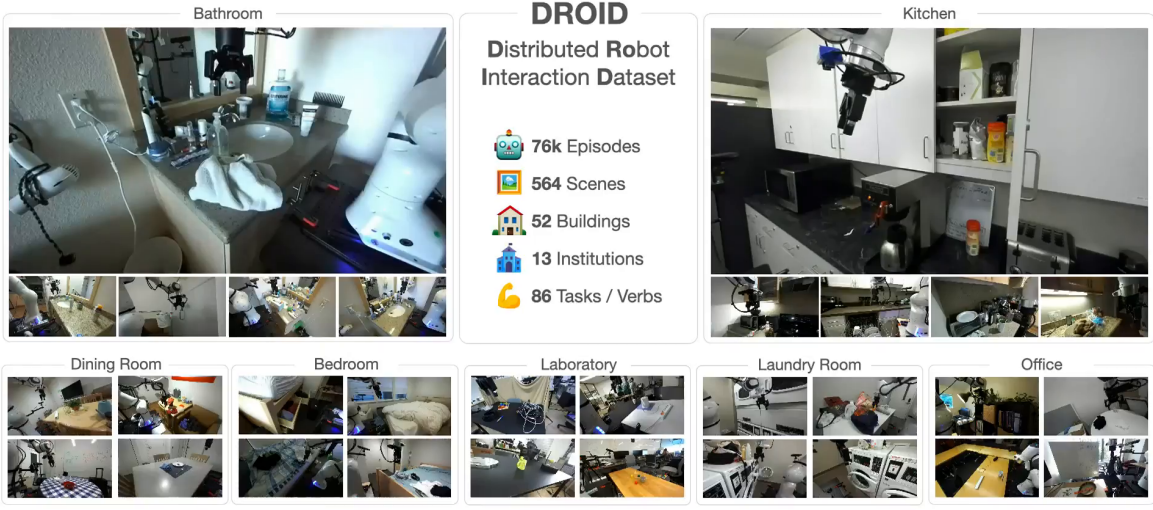
Aloha, Mobile Aloha (RSS23, CoRL24)

Droid Setup

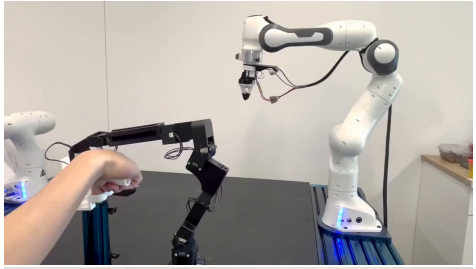
- Panda robot
- Robotiq gripper
- 3 Zed cameras
- Teleoperation by VR

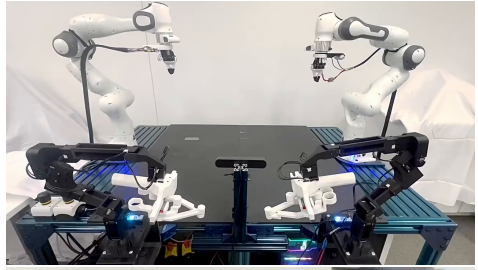


Droid data


Bathroom
DROID
Distributed Robot Interaction Dataset
76k Episodes
564 Scenes
52 Buildings
13 Institutions
86 Tasks / Verbs
Kitchen
Dining Room
Bedroom
Laboratory
Laundry Room
Office

Force-feedback dual-arm teleoperation





FACTR, arXiv, 2025

DexWild

DexWild, arXiv, 2025

WP2: Pretraining of foundation models

- Dataset composition and preprocessing
  - pipeline to create generic multimodal datasets
  - sufficiently large and diverse

WP2: Pretraining of foundation models

- Dataset composition and preprocessing
  - pipeline to create generic multimodal datasets
  - sufficiently large and diverse
- Distributed training
  - develops a distributed training framework
  - various scale training including thousands of GPUs (for >34B parameters)

WP2: Pretraining of foundation models

- Dataset composition and preprocessing
  - pipeline to create generic multimodal datasets
  - sufficiently large and diverse
- Distributed training
  - develops a distributed training framework
  - various scale training including thousands of GPUs (for >34B parameters)
- Pre-training procedure exploration
  - dataset mixtures
  - model architecture
  - loss mixture
  - validation via scaling laws

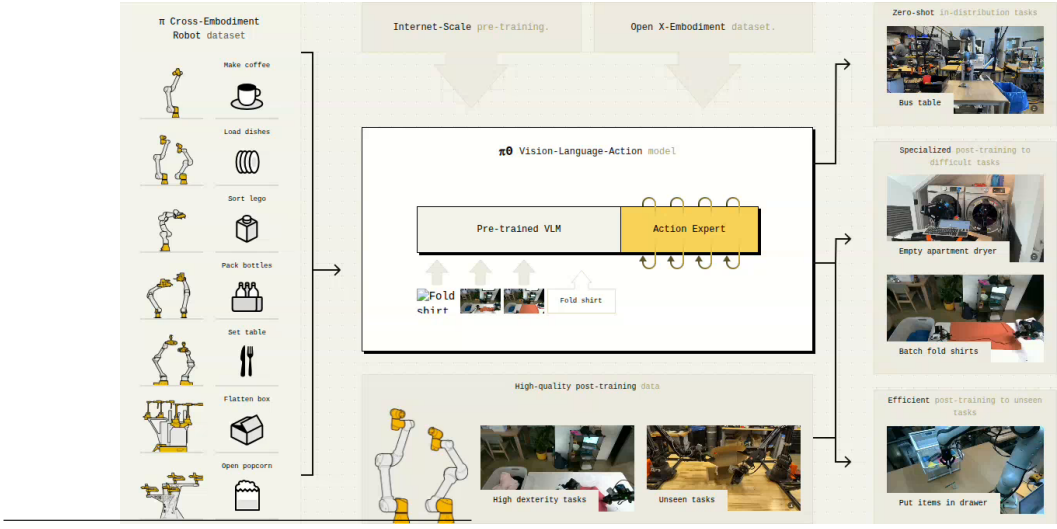
WP2: Pretraining of foundation models

- Dataset composition and preprocessing
  - pipeline to create generic multimodal datasets
  - sufficiently large and diverse
- Distributed training
  - develops a distributed training framework
  - various scale training including thousands of GPUs (for >34B parameters)
- Pre-training procedure exploration
  - dataset mixtures
  - model architecture
  - loss mixture
  - validation via scaling laws
- Evaluation and monitoring

WP2: Pretraining of foundation models

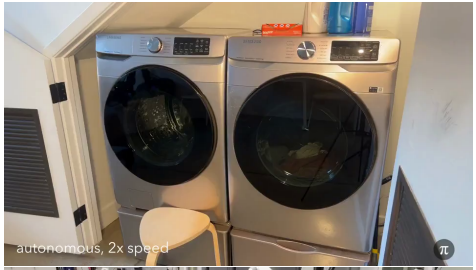
- Dataset composition and preprocessing
  - pipeline to create generic multimodal datasets
  - sufficiently large and diverse
- Distributed training
  - develops a distributed training framework
  - various scale training including thousands of GPUs (for >34B parameters)
- Pre-training procedure exploration
  - dataset mixtures
  - model architecture
  - loss mixture
  - validation via scaling laws
- Evaluation and monitoring
- Foundation model training
  - large scale training of multiple foundation models
  - models for various regulations
  - multiple weeks on thousands of GPUs

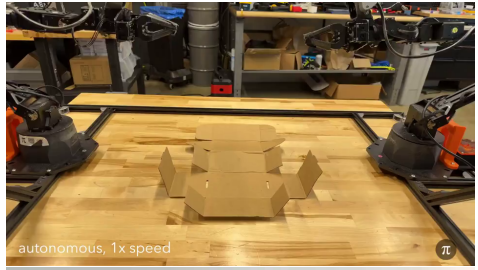
$\pi 0$ Vision Language Action model



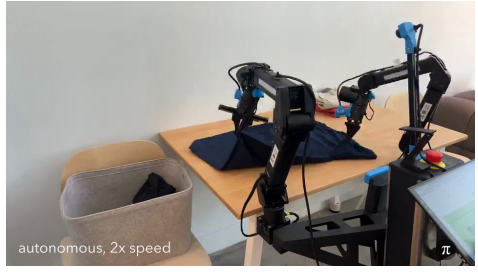
Physical Intelligence

$\pi 0$ results


autonomous, 2x speed
π


autonomous, 1x speed
π


autonomous, 2x speed
π


autonomous, 2x speed
π

WP3: Fine-tuning

- Human-aligned fine-tuning
  - European values and human goals
  - reinforcement learning from human feedback
  - preference optimization
  - trustworthy without hallucinations

WP3: Fine-tuning

- Human-aligned fine-tuning
  - European values and human goals
  - reinforcement learning from human feedback
  - preference optimization
  - trustworthy without hallucinations
- Grounded fine-tuning
  - inserting modalities
  - space, time, causal, depth, object-3D-awareness

WP3: Fine-tuning

- Human-aligned fine-tuning
  - European values and human goals
  - reinforcement learning from human feedback
  - preference optimization
  - trustworthy without hallucinations
- Grounded fine-tuning
  - inserting modalities
  - space, time, causal, depth, object-3D-awareness
- Test-time fine-tuning
  - test-time adaptation
  - handle unseen distribution shifts

WP3: Fine-tuning

- Human-aligned fine-tuning
  - European values and human goals
  - reinforcement learning from human feedback
  - preference optimization
  - trustworthy without hallucinations
- Grounded fine-tuning
  - inserting modalities
  - space, time, causal, depth, object-3D-awareness
- Test-time fine-tuning
  - test-time adaptation
  - handle unseen distribution shifts
- Fine-tuning to combine multiple modalities
  - joint fine-tuning to align multiple models

WP3: Fine-tuning

- Human-aligned fine-tuning
  - European values and human goals
  - reinforcement learning from human feedback
  - preference optimization
  - trustworthy without hallucinations
- Grounded fine-tuning
  - inserting modalities
  - space, time, causal, depth, object-3D-awareness
- Test-time fine-tuning
  - test-time adaptation
  - handle unseen distribution shifts
- Fine-tuning to combine multiple modalities
  - joint fine-tuning to align multiple models
- Efficiency fine-tuning
  - distillation
  - pruning
  - architecture search

Robotics has specific modalities

- Measured quantities
  - joint angle measurements
  - force-torque measurements
  - multiview RGBD cameras

Tell and show, arXiv, 2024

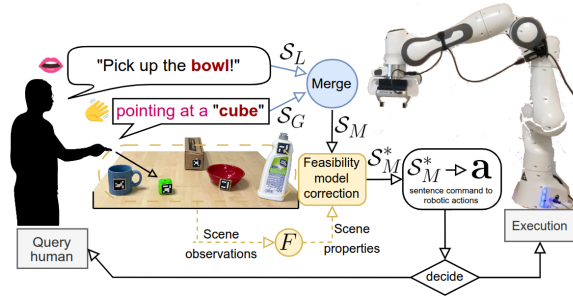
Robotics has specific modalities

- Measured quantities
  - joint angle measurements
  - force-torque measurements
  - multiview RGBD cameras
- Different actuation modes
  - continuous / discrete
  - torque / position / impedance control

Tell and show, arXiv, 2024

Robotics has specific modalities

- Measured quantities
  - joint angle measurements
  - force-torque measurements
  - multiview RGBD cameras
- Different actuation modes
  - continuous / discrete
  - torque / position / impedance control
- Human in the loop
  - feedback from human
  - life-long learning
  - catastrophic forgetting



Tell and show, arXiv, 2024

Testing and Evaluation

- Evaluation platform
  - develop open-source platform for evaluation
  - orchestration tooling, HPC integration, monitoring
  - automated evaluation via FM judges

Testing and Evaluation

- Evaluation platform
  - develop open-source platform for evaluation
  - orchestration tooling, HPC integration, monitoring
  - automated evaluation via FM judges
- Evaluation for new data modalities
  - new benchmark datasets
  - ability to reason about new data modalities

Testing and Evaluation

- Evaluation platform
  - develop open-source platform for evaluation
  - orchestration tooling, HPC integration, monitoring
  - automated evaluation via FM judges
- Evaluation for new data modalities
  - new benchmark datasets
  - ability to reason about new data modalities
- Evaluation for new AI capabilities
  - robot perception and autonomous driving
  - simulated and real-world benchmarks
  - generalization to unseen objects, tasks, and environments

Testing and Evaluation

- Evaluation platform
  - develop open-source platform for evaluation
  - orchestration tooling, HPC integration, monitoring
  - automated evaluation via FM judges
- Evaluation for new data modalities
  - new benchmark datasets
  - ability to reason about new data modalities
- Evaluation for new AI capabilities
  - robot perception and autonomous driving
  - simulated and real-world benchmarks
  - generalization to unseen objects, tasks, and environments
- Evaluation for AI safety
  - robustness against attacks
  - privacy issues
  - ethical and regulatory standards

Evaluation in robotics

- Real-world evaluation is expensive

Evaluation in robotics

- Real-world evaluation is expensive
- Simulated evaluation has sim2real gap

Evaluation in robotics

- Real-world evaluation is expensive
- Simulated evaluation has sim2real gap
- We are developing a new evaluation platform
  - analyze the sim2real correlation
    - control gap
    - perception gap

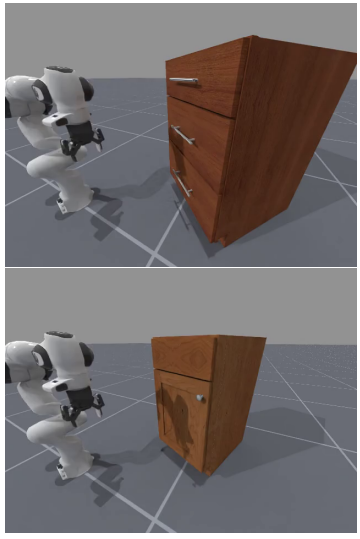
Evaluation in robotics

- Real-world evaluation is expensive
- Simulated evaluation has sim2real gap
- We are developing a new evaluation platform
  - analyze the sim2real correlation
    - control gap
    - perception gap
  - evaluate the performance of foundation models / finetuned models

Evaluation in robotics

- Real-world evaluation is expensive
- Simulated evaluation has sim2real gap
- We are developing a new evaluation platform
  - analyze the sim2real correlation
    - control gap
    - perception gap
  - evaluate the performance of foundation models / finetuned models
  - evaluate generalization capabilities
    - novel poses of known objects
    - novel objects
    - novel tasks

Simulated benchmark



WP6: Use Case Driven Model Transfer and Deployment

Application Domains	Use Case Owners		Data / Modalities
	Industry / Government	Academic	
Media Use Case 1.1: New media production Use Case 1.2: Live fact-checking	vrt, voxist	CERTH, Università di Trento	Audio / Speech | Video
Earth Modelling Use Case 2.1: Earth observation Use Case 2.2: Climate modelling	deimos	Universitat de València, Barcelona Supercomputing Center	Multispectral | Hyperspectral Thermal | ERA5 | CMIP6
Robot Perception Use Case 3.1: Robotic surface treatment	Robotwin	CTU, ETH zürich, Università di Trento	RGBD | Motion Proprioception | Language
Mobility Use Case 4.1: Autonomous driving Use Case 4.2: Infrastructure monitoring	Valeo, Generalitat de Catalunya	CVC, Università di Trento	RGBD | Lidar | RADAR | GPS
Computer Engineering Use Case 5.1: Code generation Use Case 5.2: Hardware design	openchip	Barcelona Supercomputing Center, CISPA	Source Code Formal Languages
Workflow Automation Use Case 6.1: Document understanding Use Case 6.2: Understanding tabular datasets	yooz	Jülich, CVC, TU/e, ellis Institute Tübingen	Language | Vision Tabular data | Layout

RoboTwin

RoboTwin

Compute

- Training requires large scale compute resources
- European HPC infrastructure involved in the consortium
  - BSC (4500 GPUs)
  - CINECA hosts HPC Leonardo (14000 GPUs)
  - CSC hosts LUMI (12000 GPUs)
  - FZJ (4000 GPUs)
  - Swiss AI initiative (10000 GPUs)
- 2M Eur allocated for compute resources and data collections

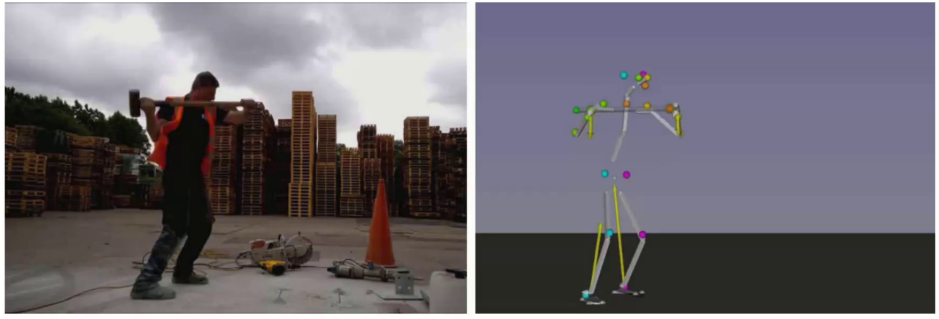
Our path toward ELLIOT's objectives

- Data collection for robotics is costly
- Simulation is not variable enough
- How we can use data from the internet?

Our path toward ELLIOT's objectives

- Data collection for robotics is costly
- Simulation is not variable enough
- How we can use data from the internet?
- **YouTube instructional videos**

Extracting human and tool motion from video - IJCV 2022

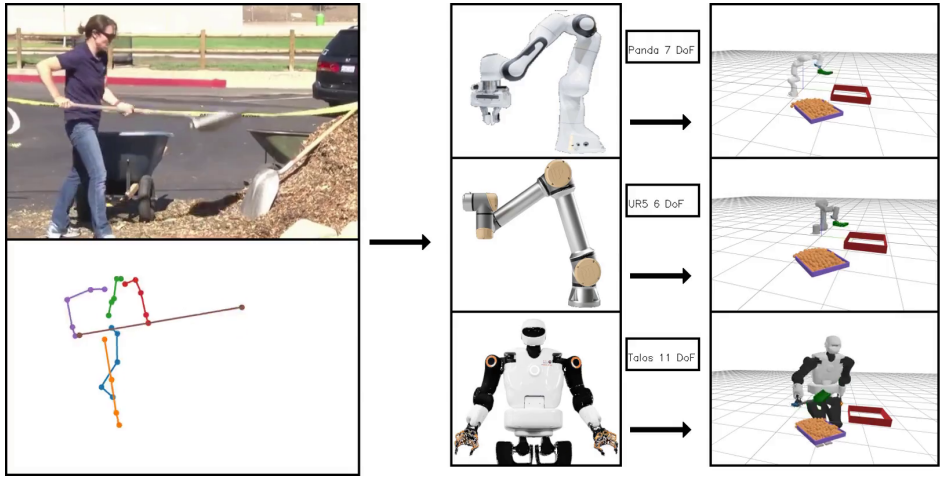


Learning to Use Tools by Watching Videos

Input: instructional video from YouTube

Output: tool manipulation skill transferred to a robot

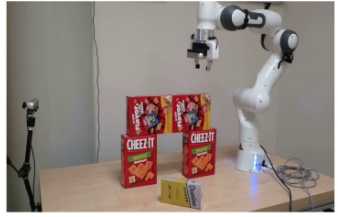
Learning tool manipulation - RAL 2022



TAMP guided by video - ICRA 2023

Multi-Contact Task and Motion Planning Guided by Video Demonstration

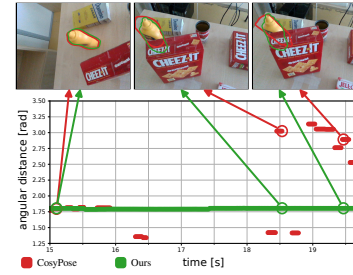
Kateryna Zorina ♣ David Kovar ♣ Florent Lamiraux ♢ Nicolas Mansard ♢
Justin Carpentier ♥ Josef Sivic ♣ Vladimir Petrik ♣

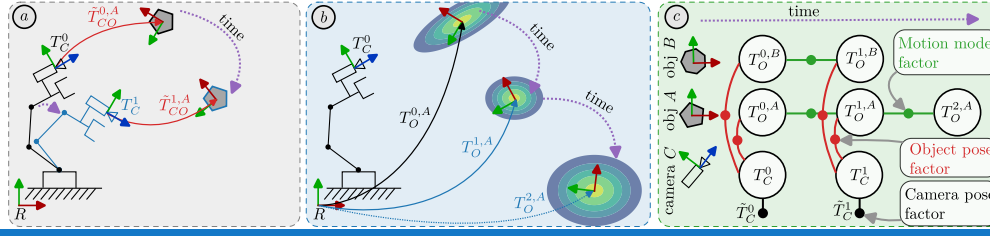


♣ CIIRC, Czech Technical University in Prague
♢ LAAS-CNRS, Universite de Toulouse, CNRS, Toulouse
♥ INRIA, Paris

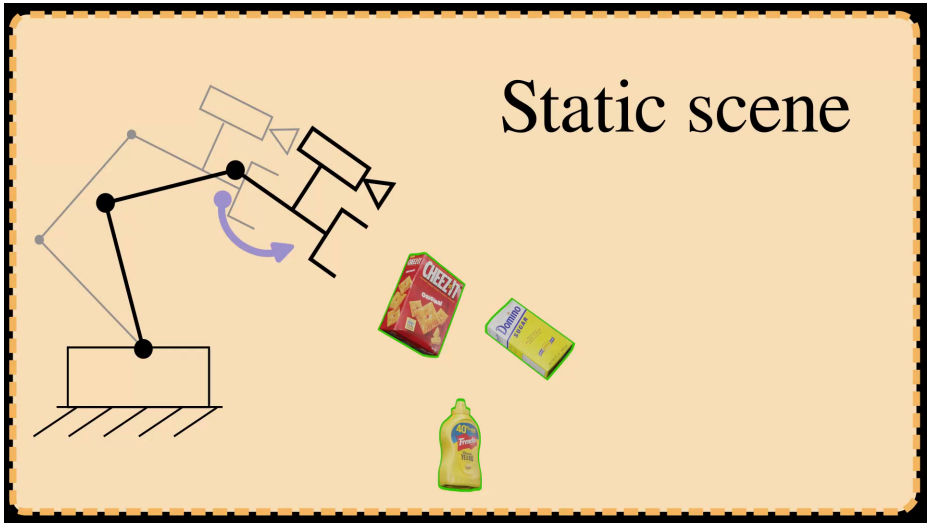
Temporal consistency for object pose estimation - RAL 2025

- Object pose estimation unstable under occlusions
- We need to ensure temporal consistency for control
- Using smoothing and mapping for temporal consistency

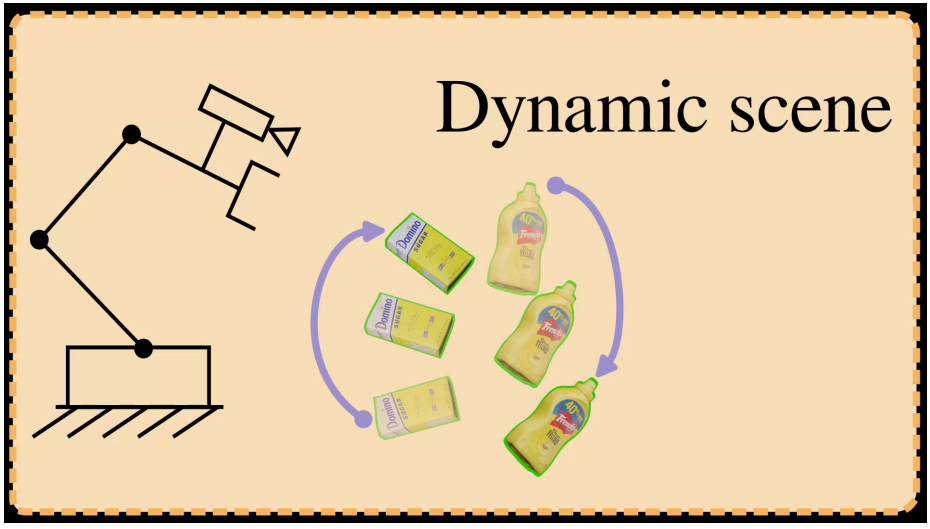




Temporal consistency results

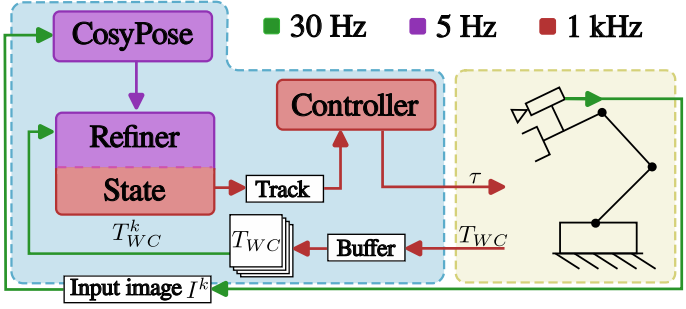


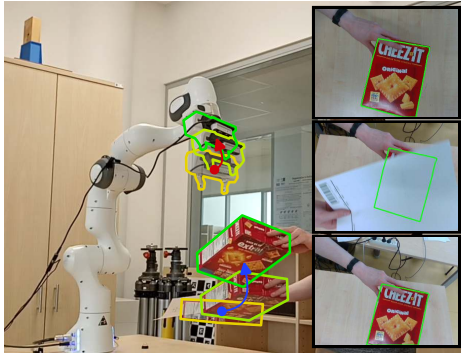
Temporal consistency results



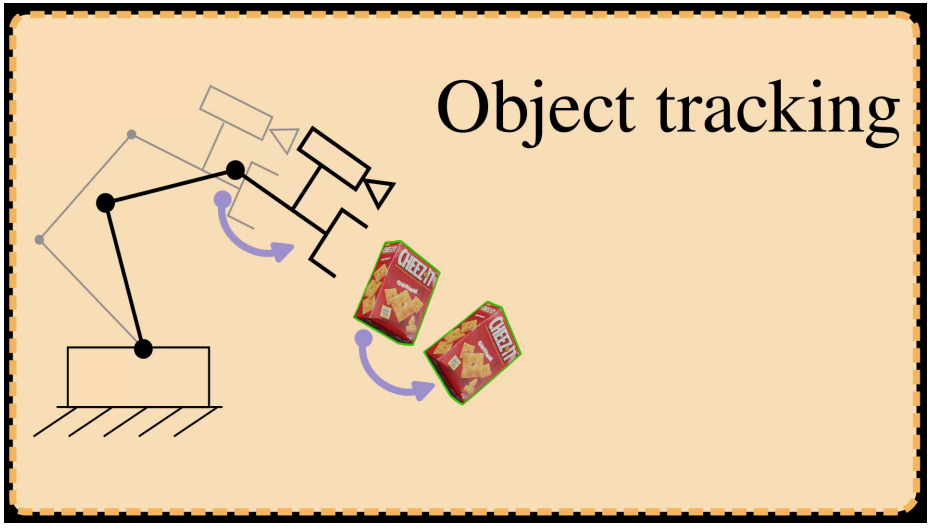
Temporal consistency for robot tracking

- Combine tracking with robot control

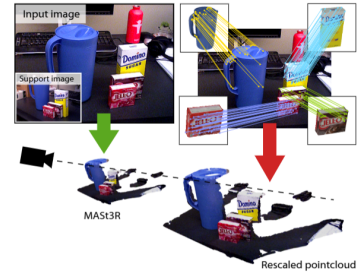




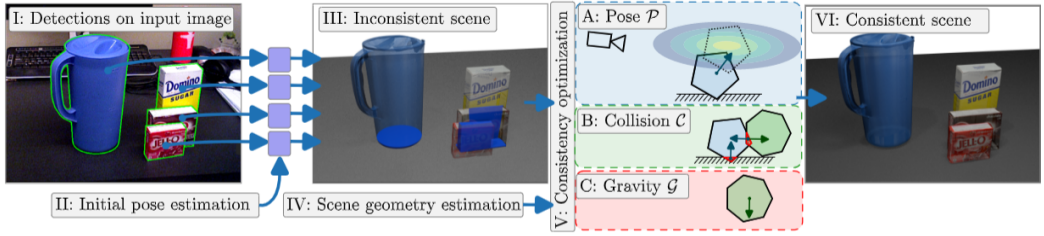
Temporal consistency results



PhysPose - Physical consistency



- Differentiable collision distance
- Estimated support table from two views
  - Dust3r/Mast3r multi-view depth (CVPR 2024)
  - Rescaled based on the reference objects sizes

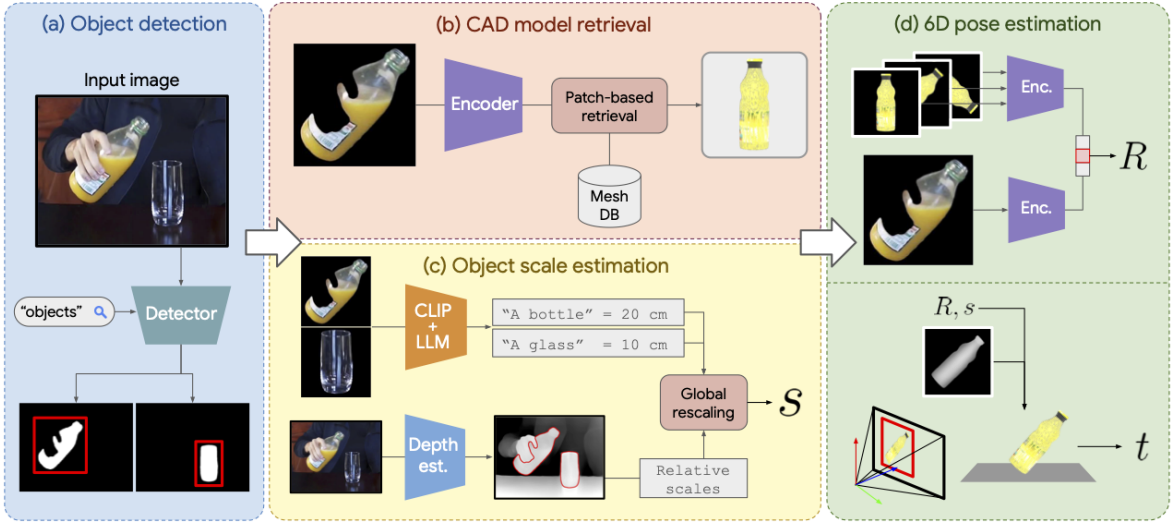


Physical consistency visualization

Supplementary Material for
PhyPose: Refining 6D Object Poses with Physical Constraints

Paper ID 14085

FreePose - ICLR 2025



FreePose

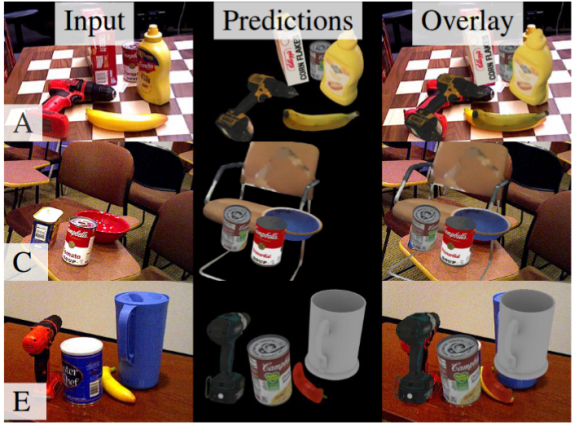
(a) input video

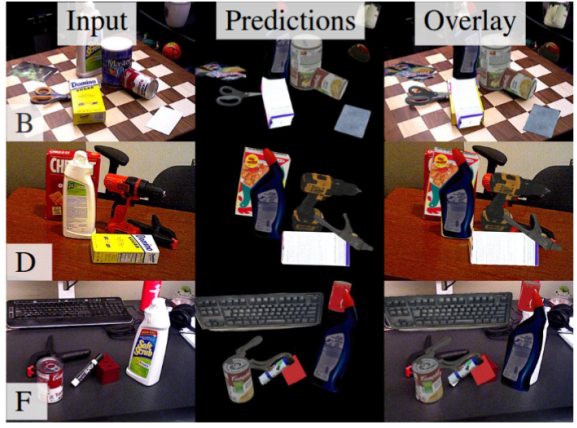
(b) retrieved mesh

(c) 6D pose trajectory

(d) robot trajectory

FreePose


Input
Predictions
Overlay
A
C
E


Input
Predictions
Overlay
B
D
F

FreePose

6D Object Pose Tracking in Internet Videos for Robotic Manipulation

Anonymous authors

Supplementary video for submission #8215
ICLR 2025

Thank you for your attention.