OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

面向机器人操作的视觉表征预训练

《端到端具身智能体》讲习班

2024.6.9

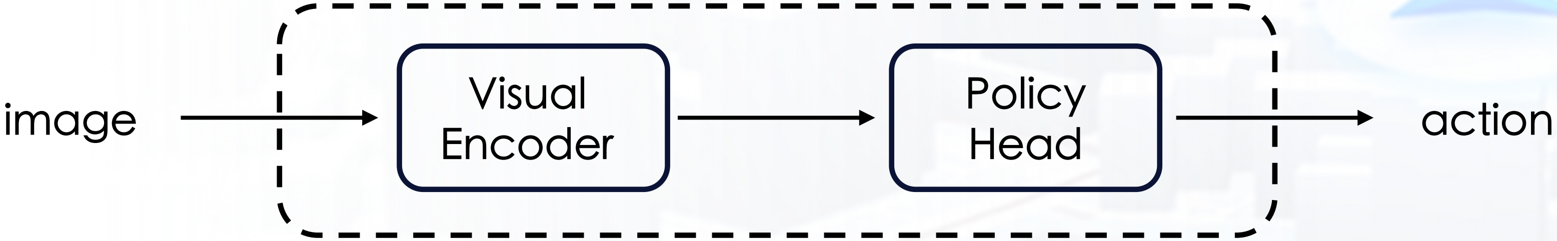
曾嘉 上海人工智能实验室

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

面向具身操作的视觉表征方法解析

Background: Robotic Manipulation



image

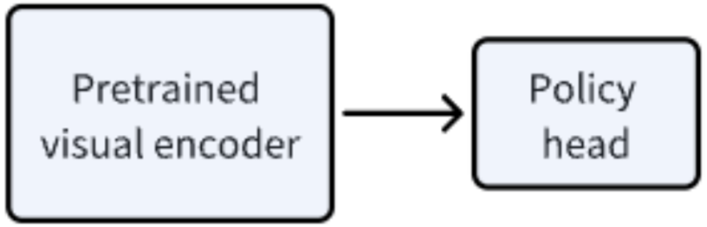
Visual Encoder

Policy Head

action

Background

In the field of visuomotor control, a number of approaches find that existing representations such as features from models trained on ImageNet, or features from CLIP enable more efficient learning for imitation learning and reinforcement learning

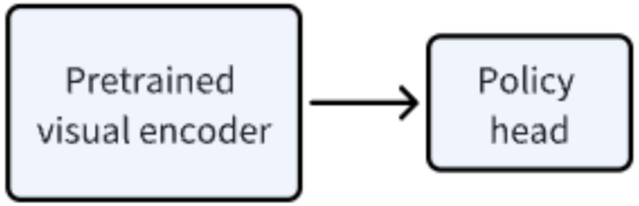


Pretrained visual encoder

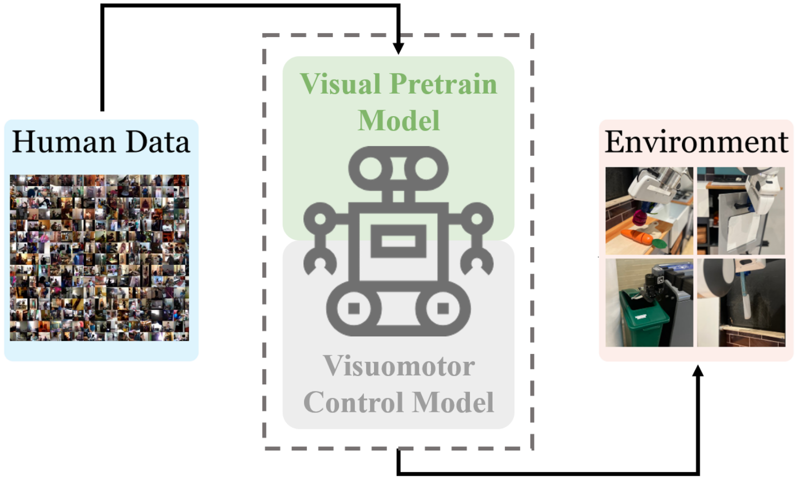
Policy head

Background

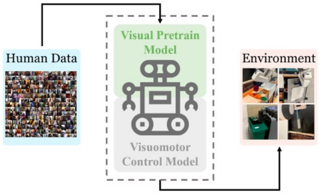
In the field of visuomotor control, a number of approaches find that existing representations such as features from models trained on ImageNet, or features from CLIP enable more efficient learning for imitation learning and reinforcement learning



However, in-domain robot data is scarce, whereas there is a much more abundant supply of human data performing daily tasks (e.g. Ego4D).



Representation Learning for Robotic Manipulation



CoRL 2022

R3M [2]

2022/03

2022/03

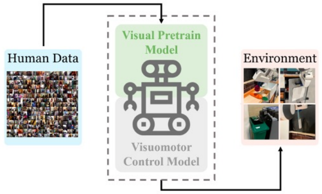
Time

MVP [1]

CoRL 2022 Oral

[1] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *CoRL*, 2022.
[2] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *CoRL*, 2022.
[3] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023.
[4] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *RSS*, 2023.
[5] Wu H, Jing Y, Cheang C, Chen G, Xu J, Li X, Liu M, Li H, Kong T. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In ICLR, 2024.

Representation Learning for Robotic Manipulation



2022/03

MVP [1]

CoRL

CoRL 2022 Oral

Berkeley
UNIVERSITY OF CALIFORNIA

Stanford University

CoRL 2022

R3M [2]

2022/03

CoRL

2022/09

VIP [3]

ICLR

ICLR 2023 Spotlight

Penn
UNIVERSITY of PENNSYLVANIA

Stanford University

TOYOTA RESEARCH INSTITUTE

RSS 2023 最佳论文提名

Voltron [4]

2023/02

ROBOTICS SCIENCE AND SYSTEMS

2023/12

GR-1 [5]

ICLR

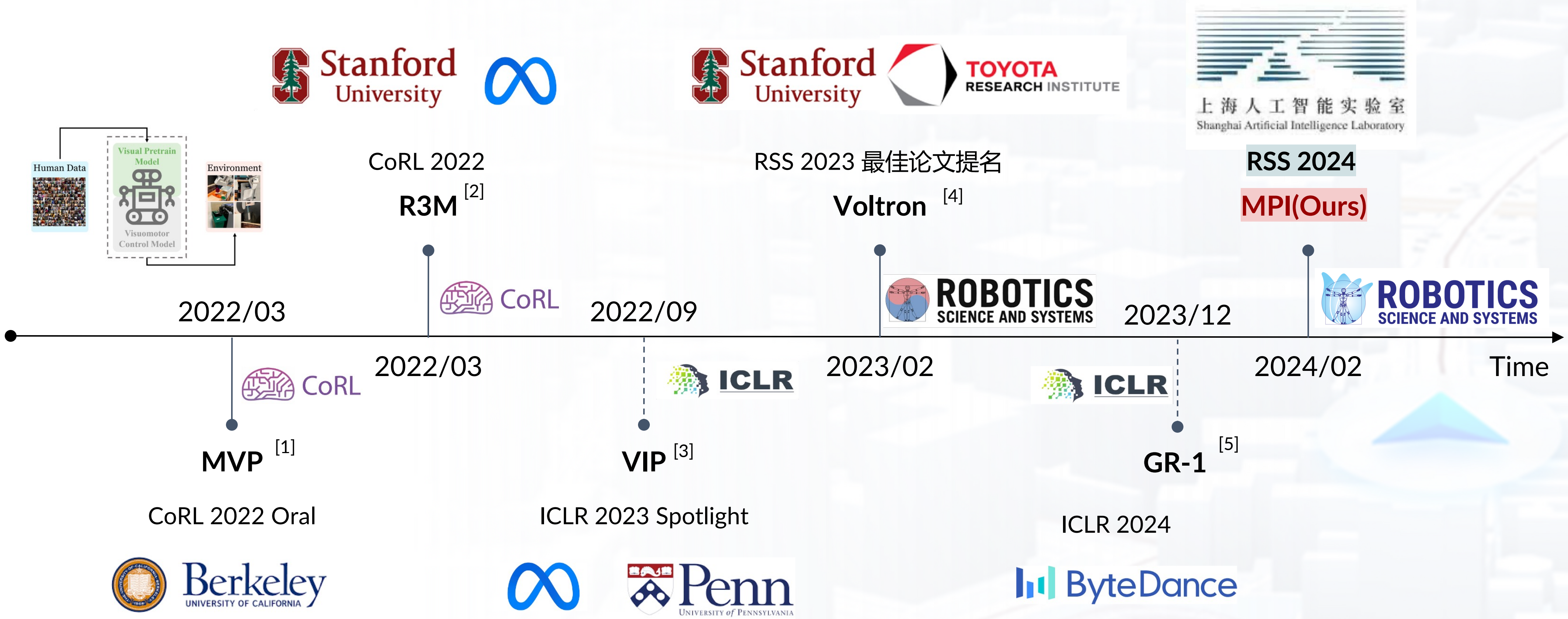
ICLR 2024

ByteDance

Time

[1] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *CoRL*, 2022.
[2] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *CoRL*, 2022.
[3] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023.
[4] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *RSS*, 2023.
[5] Wu H, Jing Y, Cheang C, Chen G, Xu J, Li X, Liu M, Li H, Kong T. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In ICLR, 2024.

Representation Learning for Robotic Manipulation



2022/03
MVP [1]
CoRL 2022 Oral
Berkeley University of California

2022/03
R3M [2]
CoRL 2022
Stanford University
Meta

2022/09
VIP [3]
ICLR 2023 Spotlight
Meta
Penn University of Pennsylvania

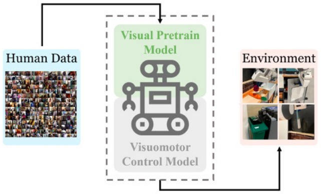
2023/02
Voltron [4]
RSS 2023 最佳论文提名
Stanford University
TOYOTA RESEARCH INSTITUTE

2023/12
GR-1 [5]
ICLR 2024
ByteDance

2024/02
RSS 2024
MPI(Ours)
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

[1] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *CoRL*, 2022.
[2] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *CoRL*, 2022.
[3] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023.
[4] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *RSS*, 2023.
[5] Wu H, Jing Y, Cheang C, Chen G, Xu J, Li X, Liu M, Li H, Kong T. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In ICLR, 2024.

Representation Learning for Robotic Manipulation



2022/03

MVP [1]

CoRL 2022 Oral

2022/03

CoRL 2022

R3M [2]

2022/09

VIP [3]

ICLR 2023 Spotlight

2023/02

RSS 2023 最佳论文提名

Voltron [4]

2023/12

GR-1 [5]

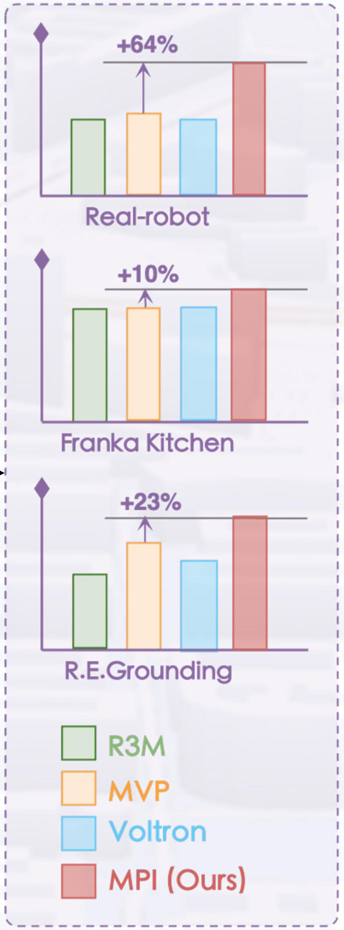
ICLR 2024

2024/02

RSS 2024

MPI(Ours)

Time



---

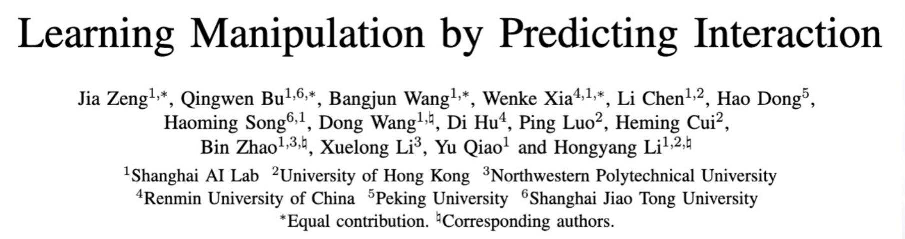
[1] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *CoRL*, 2022.
[2] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *CoRL*, 2022.
[3] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023.
[4] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *RSS*, 2023.
[5] Wu H, Jing Y, Cheang C, Chen G, Xu J, Li X, Liu M, Li H, Kong T. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In ICLR, 2024.

MPI相比于R3M、MVP、Voltron等工作，对下游任务的泛化能力更强

Representation Learning for Robo

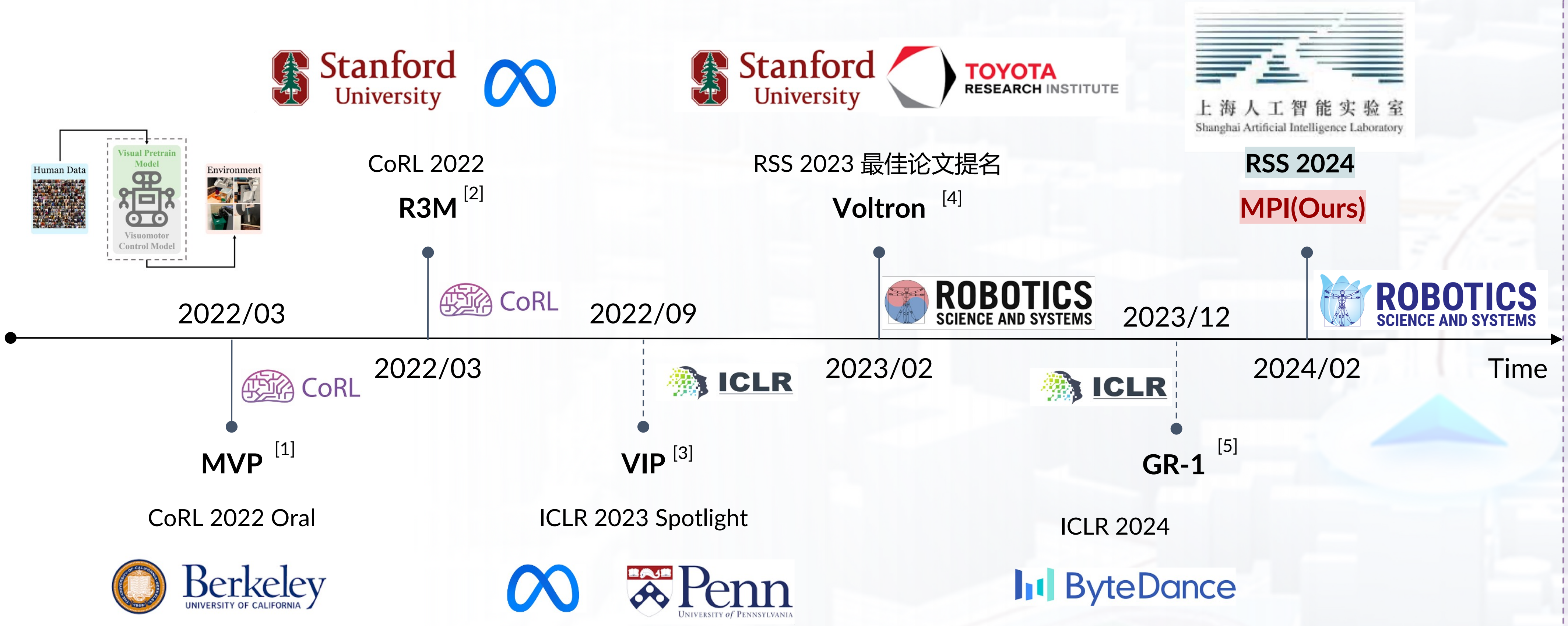
交互导向的机器人视觉表征预训练

中稿机器人最难会议 RSS 2024 年共收录135篇



Learning Manipulation by Predicting Interaction

Jia Zeng[1,*], Qingwen Bu[1,6,*], Bangjun Wang[1,*], Wenke Xia[4,1,*], Li Chen[1,2], Hao Dong[5],
Haoming Song[6,1], Dong Wang[1,♮], Di Hu[4], Ping Luo[2], Heming Cui[2],
Bin Zhao[1,3,♮], Xuelong Li[3], Yu Qiao[1] and Hongyang Li[1,2,♮]
[1]Shanghai AI Lab [2]University of Hong Kong [3]Northwestern Polytechnical University
[4]Renmin University of China [5]Peking University [6]Shanghai Jiao Tong University
*Equal contribution. ♮Corresponding authors.


Human Data
Visual Pretrain Model
Visuomotor Control Model
Environment
Stanford University
Stanford University
TOYOTA RESEARCH INSTITUTE
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory
CoRL
CoRL
ICLR
ICLR
ROBOTICS SCIENCE AND SYSTEMS
ROBOTICS SCIENCE AND SYSTEMS
Berkeley UNIVERSITY OF CALIFORNIA
Penn UNIVERSITY of PENNSYLVANIA
ByteDance

2022/03 — **MVP** [1] — CoRL 2022 Oral

2022/03 — CoRL 2022 — **R3M** [2]

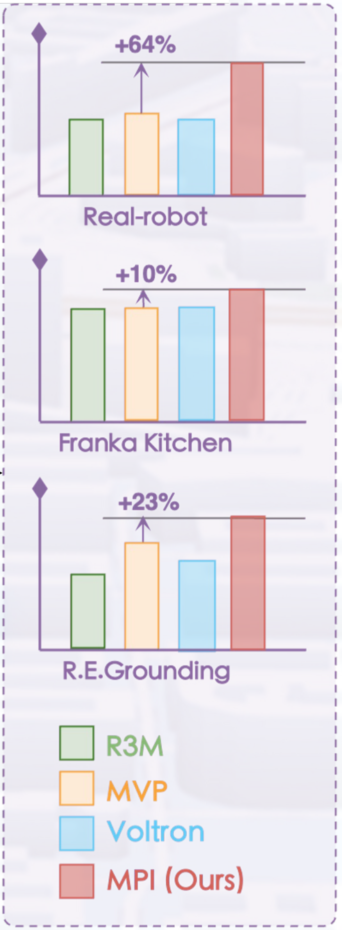
2022/09 — **VIP** [3] — ICLR 2023 Spotlight

2023/02 — RSS 2023 最佳论文提名 — **Voltron** [4]

2023/12 — **GR-1** [5] — ICLR 2024

2024/02 — **RSS 2024** — **MPI(Ours)**

Time



🟩 R3M
🟧 MVP
🟦 Voltron
🟥 MPI (Ours)

---

[1] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *CoRL*, 2022.
[2] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *CoRL*, 2022.
[3] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023.
[4] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *RSS*, 2023.
[5] Wu H, Jing Y, Cheang C, Chen G, Xu J, Li X, Liu M, Li H, Kong T. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In ICLR, 2024.

MPI相比于R3M、MVP、Voltron等工作，对下游任务的泛化能力更强

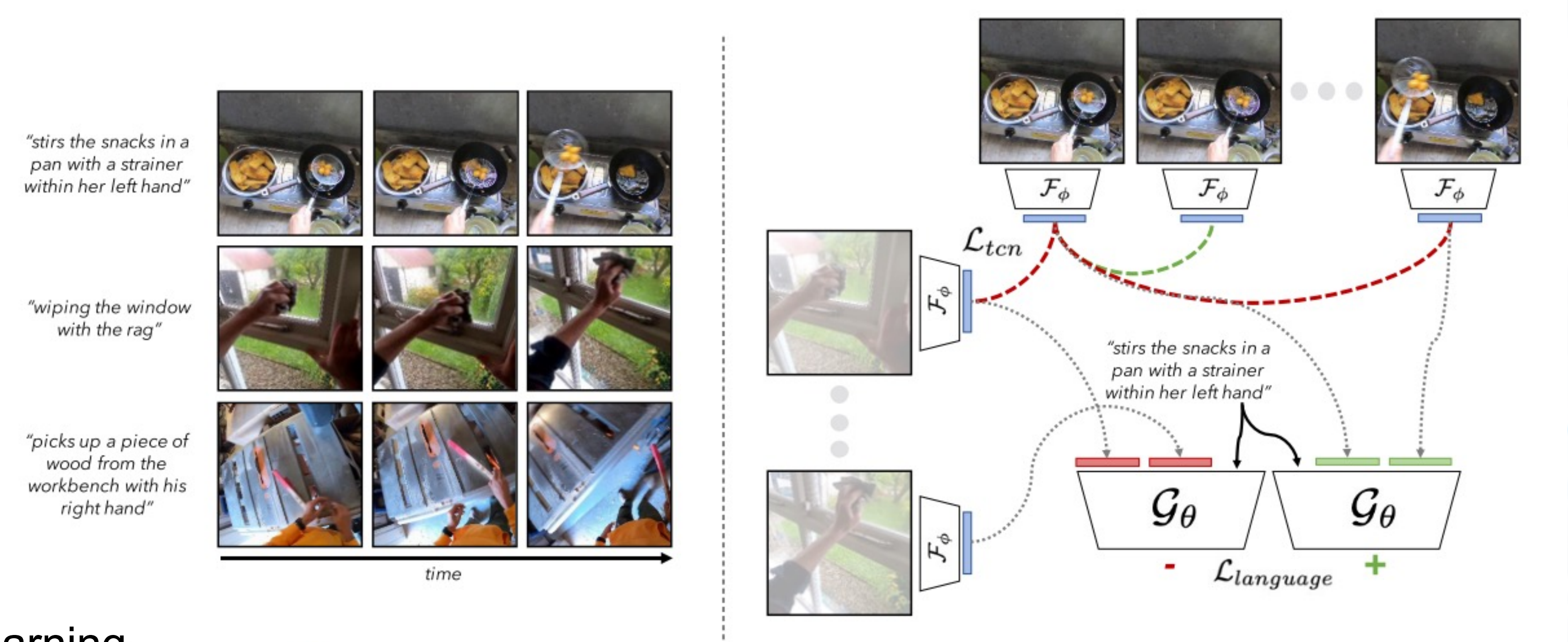
Introduction | Representation Learning for Robotic Manipulation

R3M : A Universal Visual Representation for Robot Manipulation

CoRL 2022

Project page | Paper | Code

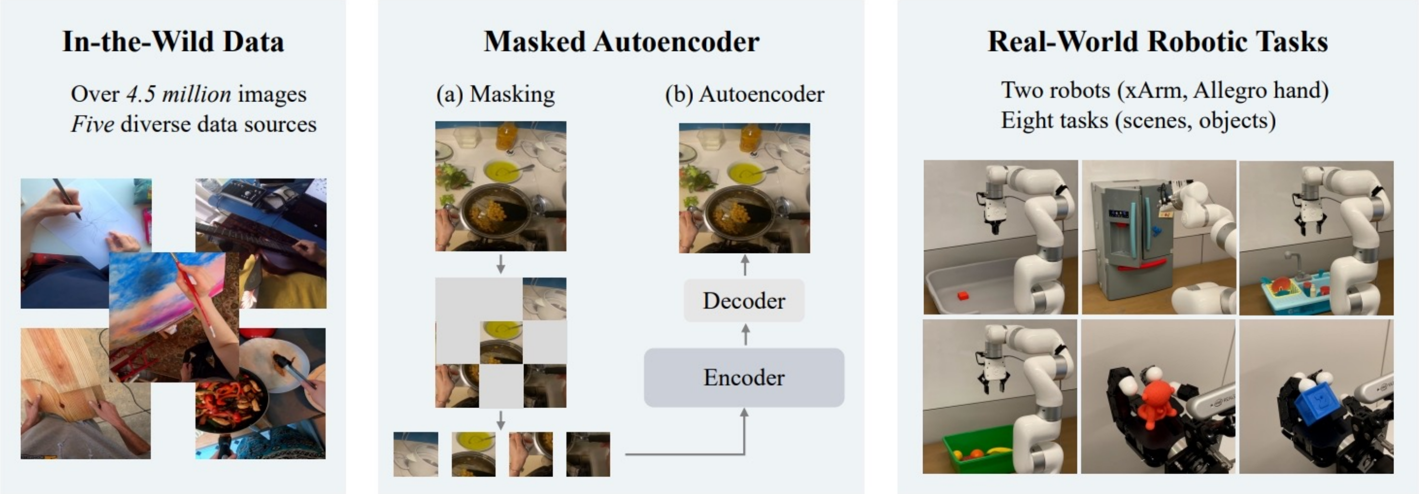
Research team: Stanford University， Meta AI

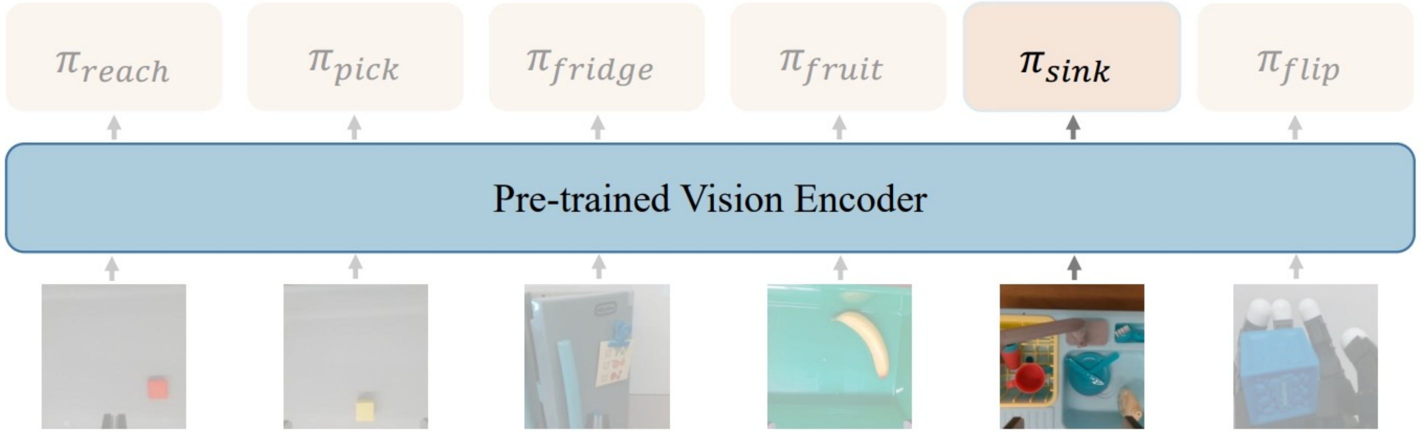

"stirs the snacks in a pan with a strainer within her left hand"
"wiping the window with the rag"
"picks up a piece of wood from the workbench with his right hand"
time
$\mathcal{F}_\phi$
$\mathcal{L}_{tcn}$
"stirs the snacks in a pan with a strainer within her left hand"
$\mathcal{G}_\theta$
-
$\mathcal{L}_{language}$
+

Contrastive learning

- Constructing positive and negative sample pairs from paired video and text
- Construct sample pairs within a video, where closely spaced frames are more similar than distant ones

Introduction | Representation Learning for Robotic Manipulation





Real-World Robot Learning with Masked Visual Pre-training (MVP)
CoRL 2022 oral
Project page | Paper | Code
Research team: University of California, Berkeley

Introducing the MAE-style image encoder pre-training into robotic manipulation tasks.

Introduction | Representation Learning for Robotic Manipulation

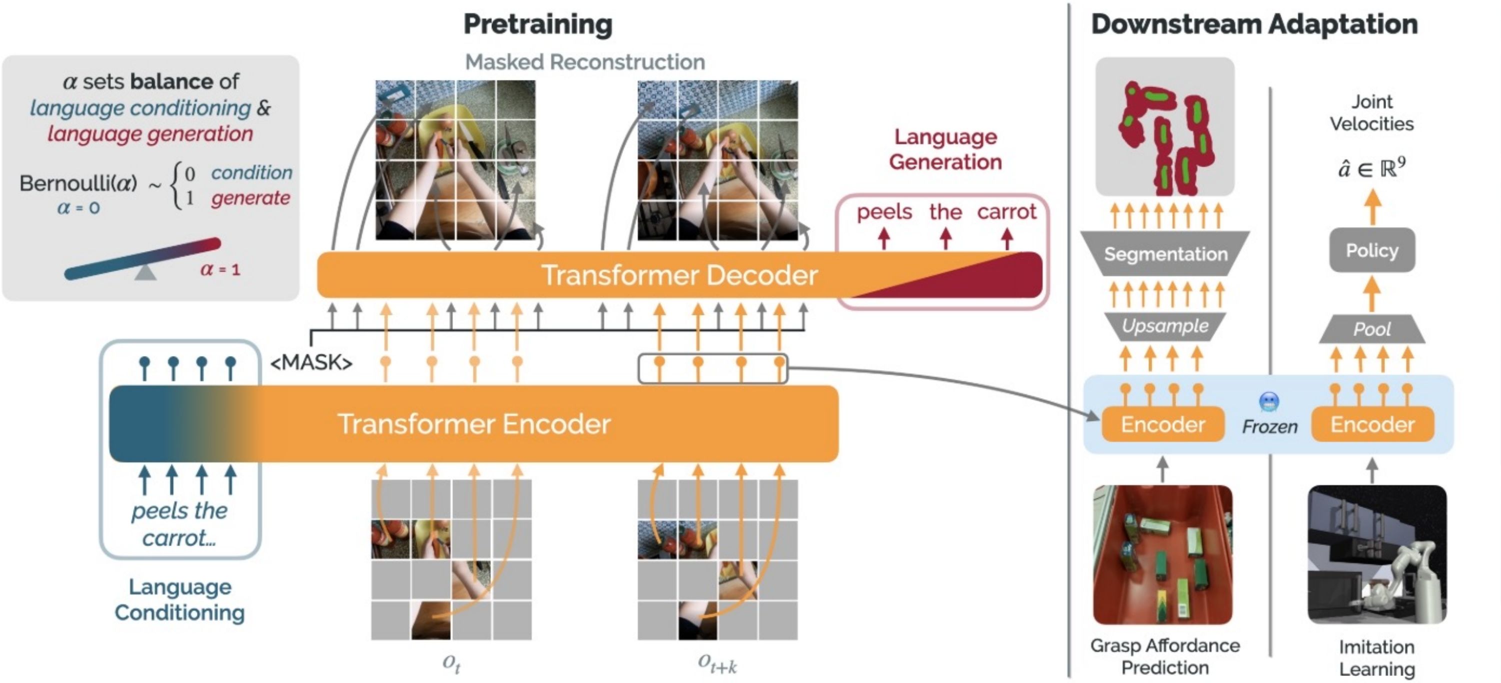
Language-Driven Representation Learning for Robotics

RSS 2023 Best Paper Award Finalists

Project page | Paper | Code

Research team: Stanford University， Toyota Research Institute

$$
\begin{aligned}
\mathcal{L}(\theta) &= \mathcal{L}_{\text{reconstruct}}(\theta) + \mathcal{L}_{\text{generate}}(\theta) \\
&= \begin{cases} \mathrm{MSE}(v_{\text{masked}}, \mathrm{R}_\theta(\mathrm{E}_\theta(v_{\text{visible}}, c))) & \text{if } z = 0 \\ \mathrm{MSE}(v_{\text{masked}}, \mathrm{R}_\theta(\mathrm{E}_\theta(v_{\text{visible}}, \texttt{<NULL>}))) & \text{if } z = 1 \\ \quad + \mathrm{NLL}(c, \mathrm{G}_\theta(\mathrm{E}_\theta(v_{\text{visible}}, \texttt{<NULL>}))) & \end{cases} \\
\text{and } \; & z \sim \mathrm{Bernoulli}(\alpha)
\end{aligned}
$$



Pre-train the visual encoder using a blend of **standard MAE**, **Language-conditioned MAE**, and **language.** This approach improves pixel-level detail recognition and high-level scene comprehension.

Motivation

R3M：Focus on high-level semantic information.

MVP：Focus on low-level pixel cues.

Voltron：By combining multiple pre-training tasks such as Language-conditioned MAE, vanilla MAE, and Language generation, the model focuses on both high-level semantic information and low-level pixel cues.

Motivation:
Existing pre-training methods lack the interactive-level features required for robot manipulation and do not adequately understand interactive dynamics, which refers to the patterns of behavior and physical interactions that occur between a robot and the environment.

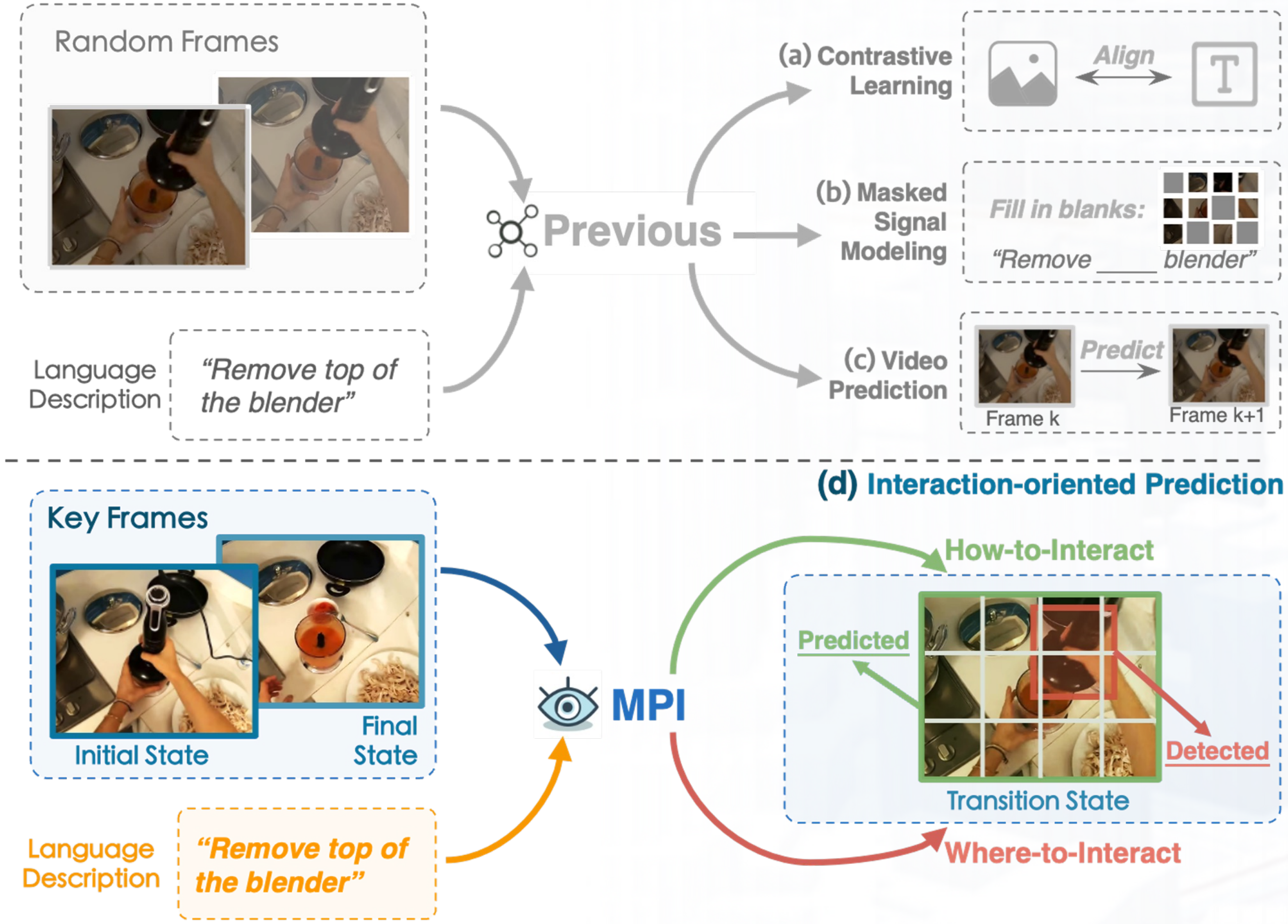
OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

基于预测交互过程的视觉表征预训练方法 MPI

Learning Manipulation by Predicting Interaction (MPI)

Method Comparison



lack explicit interaction modeling

Past

- (a) R3M: utilize contrastive learning, focus on high-level semantics.
- (b) MVP: apply MAE, mine low-level and fine-grained cues
- (c) GR-1: sequential video prediction, easy to introduce noise or redundant information

✗ *effectively capture the dynamic interactions*

Ours

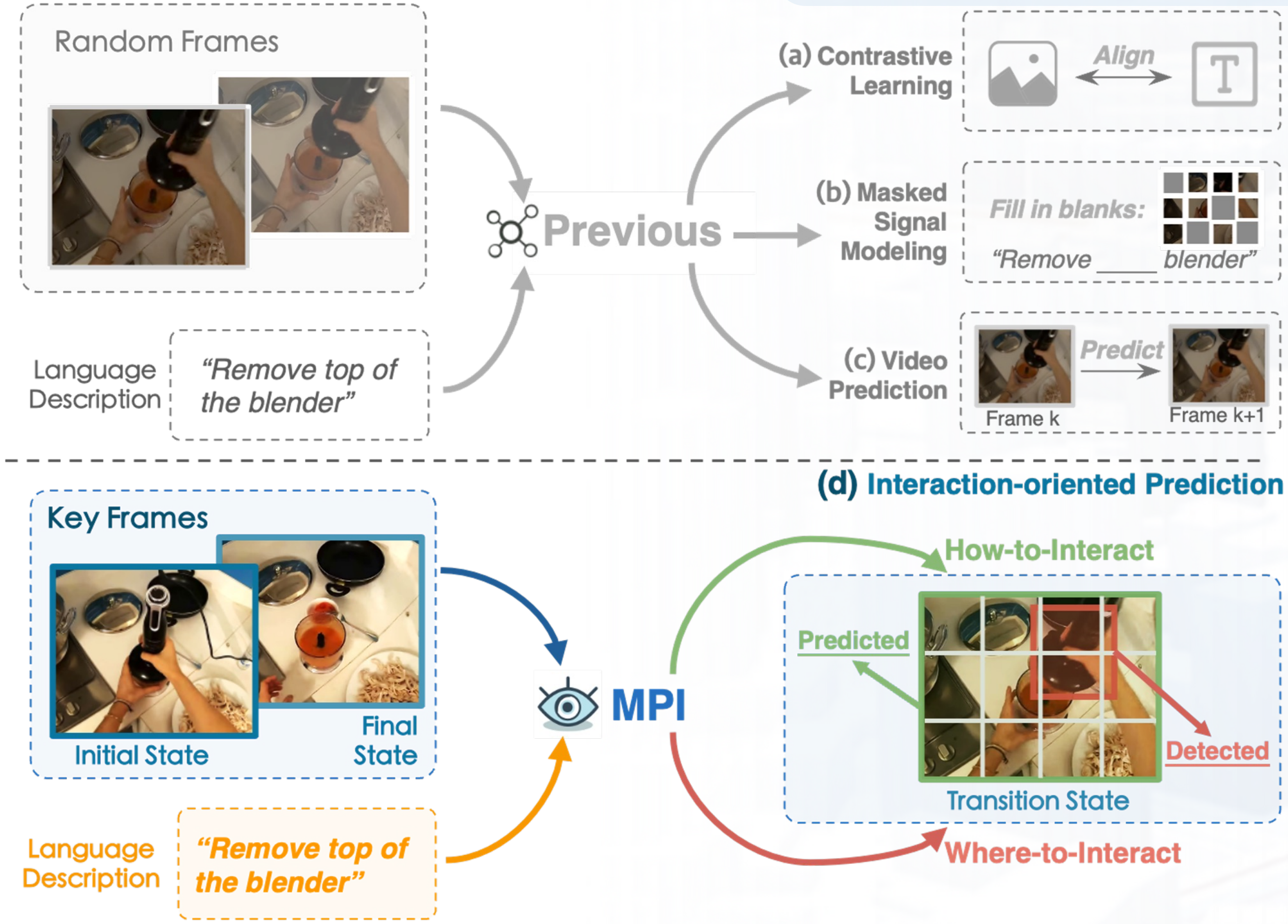
- Reflect upon the pre-training objectives
- Instill interactive dynamics by proposing an interaction-oriented prediction paradigm

**Paraphrase for interactive dynamics: the patterns of behavior and physical interactions that occur between a robot and the environment*

Learning Manipulation by Predicting Interaction (MPI)

- arXiv: https://arxiv.org/abs/2406.00439
- Project page: https://opendrivelab.com/MPI/
- code: https://github.com/OpenDriveLab/MPI

Method Comparison



lack explicit interaction modeling

Past

- (a) R3M: utilize contrastive learning, focus on high-level semantics.
- (b) MVP: apply MAE, mine low-level and fine-grained cues
- (c) GR-1: sequential video prediction, easy to introduce noise or redundant information

✗ *effectively capture the dynamic interactions*

Ours

- Reflect upon the pre-training objectives
- Instill interactive dynamics by proposing an interaction-oriented prediction paradigm

**Paraphrase for interactive dynamics: the patterns of behavior and physical interactions that occur between a robot and the environment*

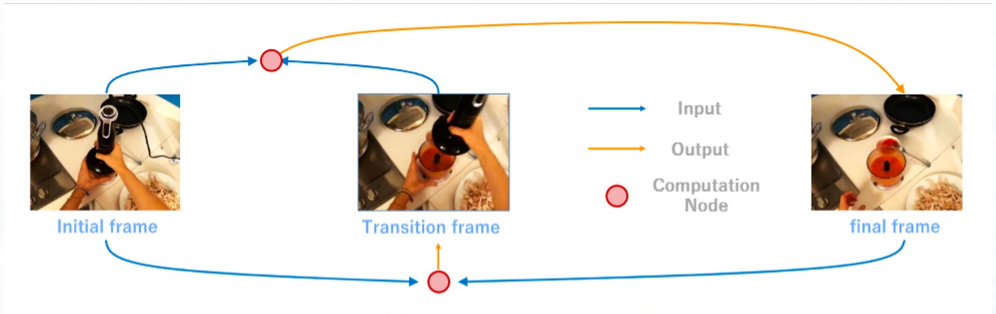
MPI | Pipeline and Framework

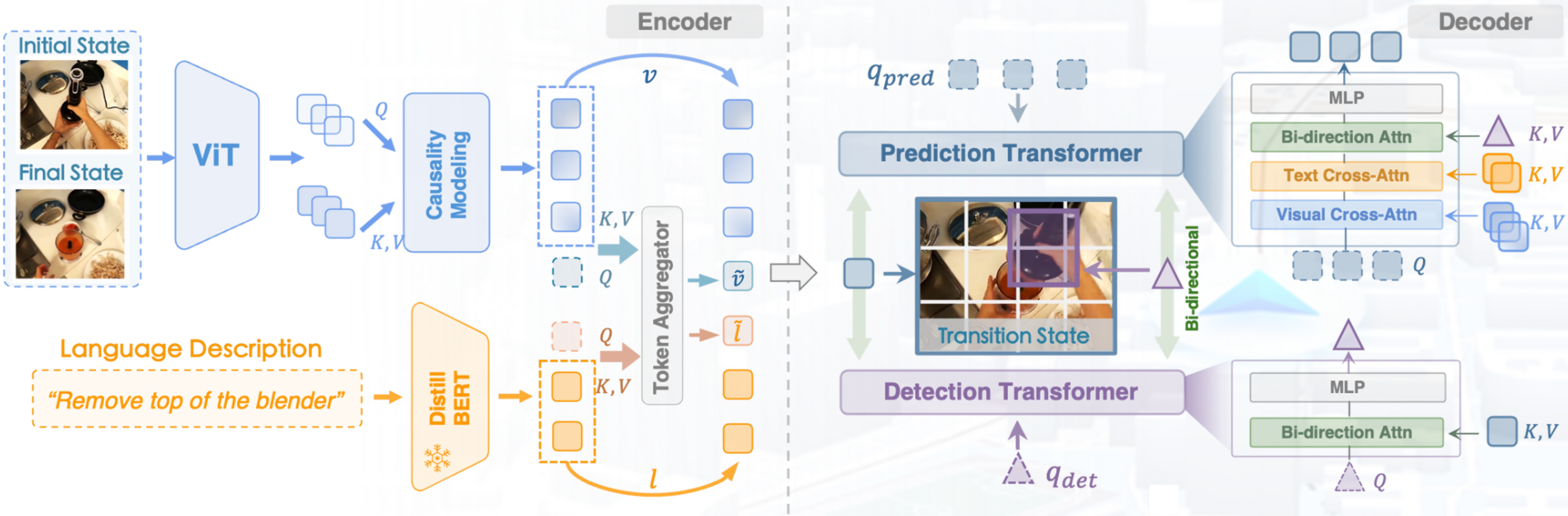
Pipeline

Two Training Objectives

“where to interact”

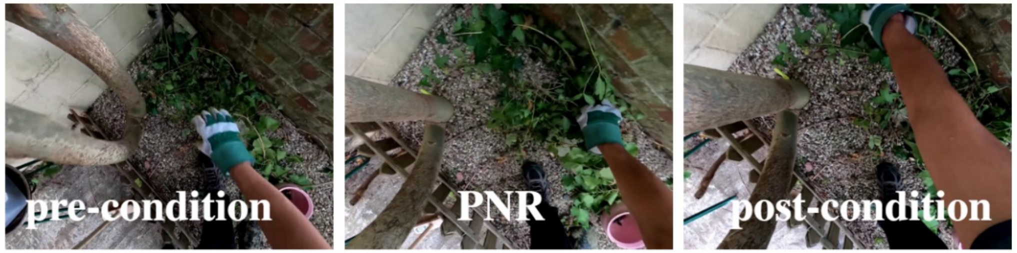
“how to interact”



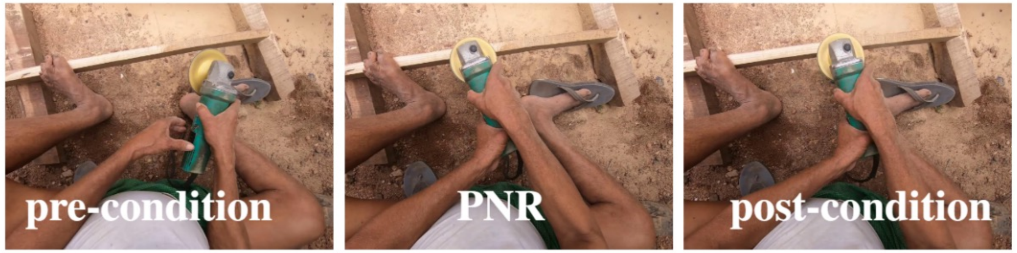


MPI | Dataset

Ego4D
Hand-and-Object subset

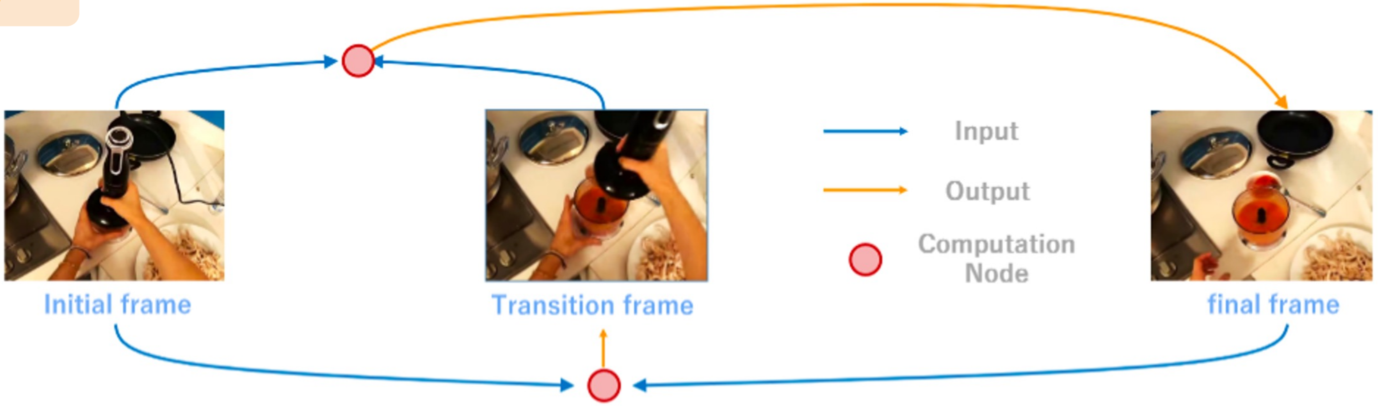


State-change: Plant removed from ground

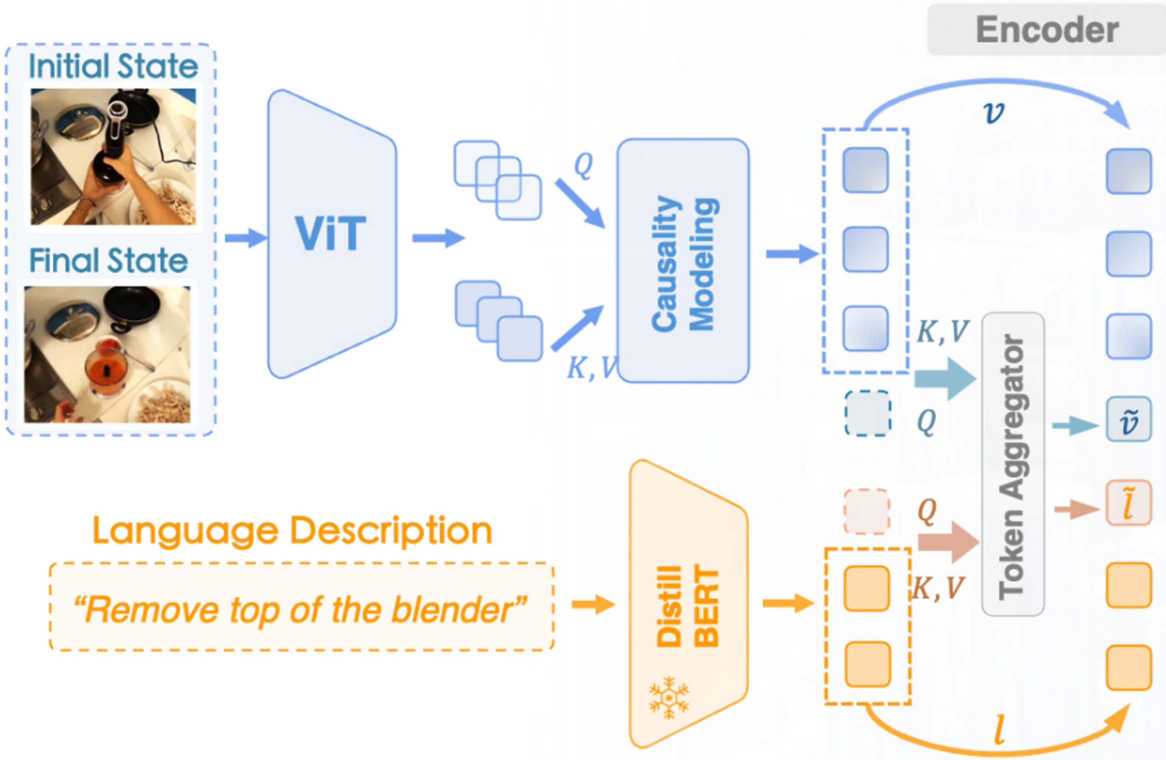


State-change: Wood smoothed

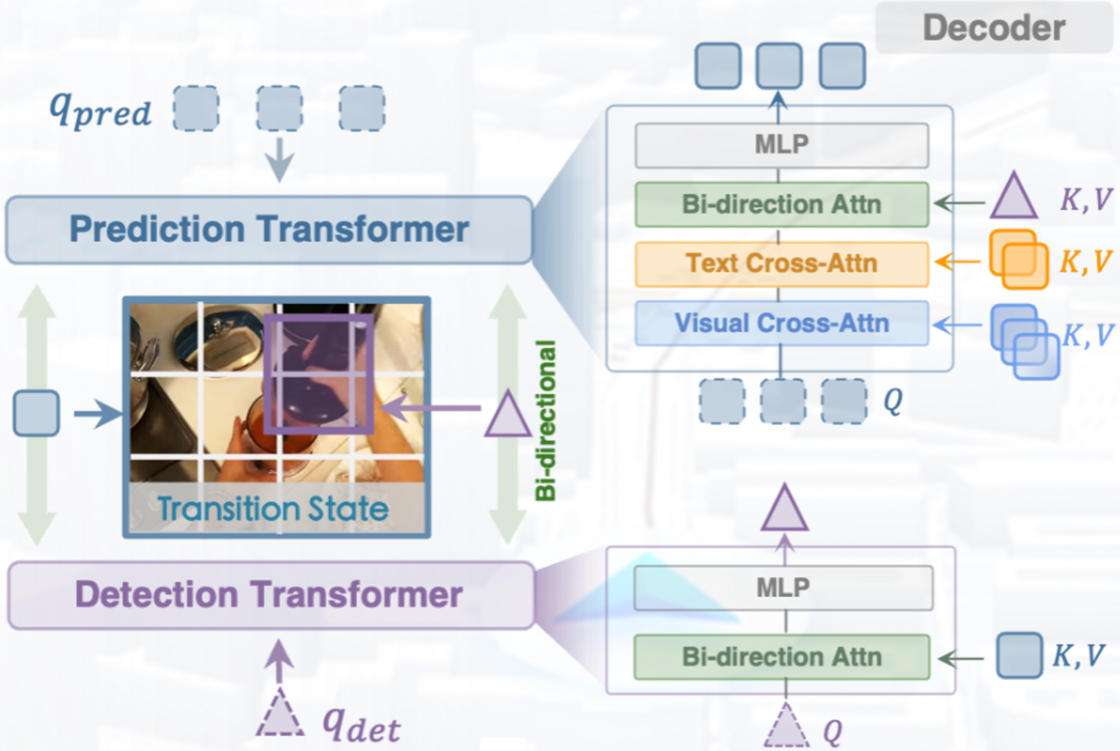
Keyframes



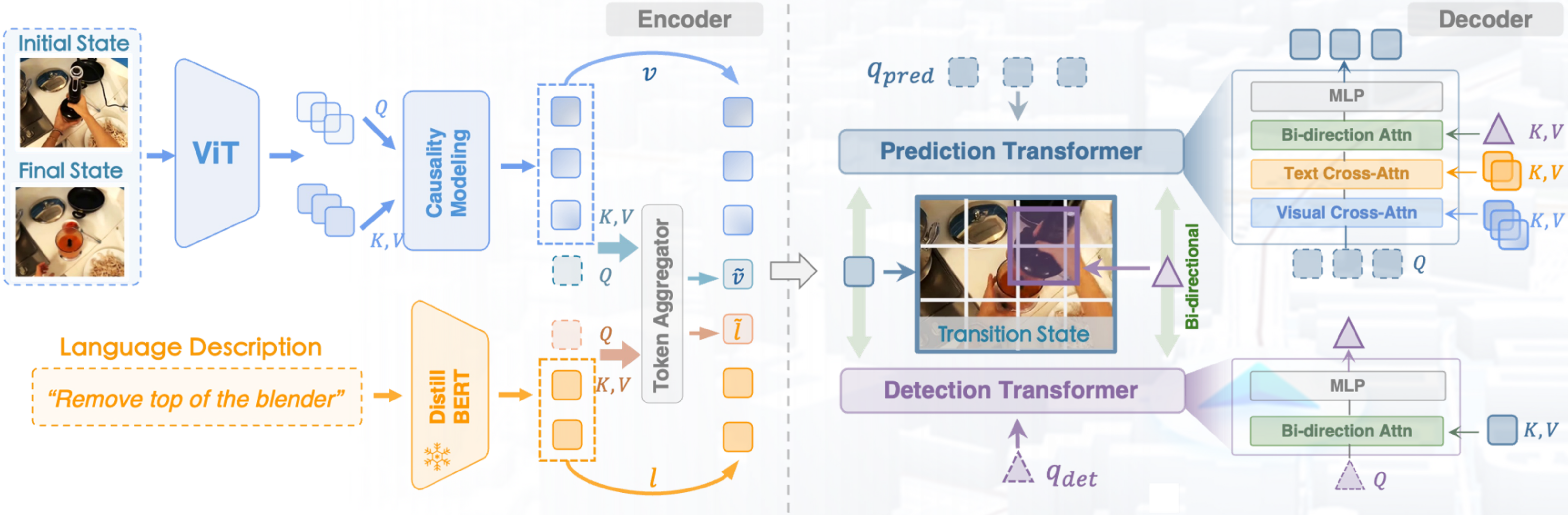
MPI | Network


Encoder
Initial State
Final State
ViT
Q
K, V
Causality Modeling
v
K, V
Q
Token Aggregator
ṽ
Q
l̃
K, V
l
Language Description
"Remove top of the blender"
Distill BERT

MPI | Network

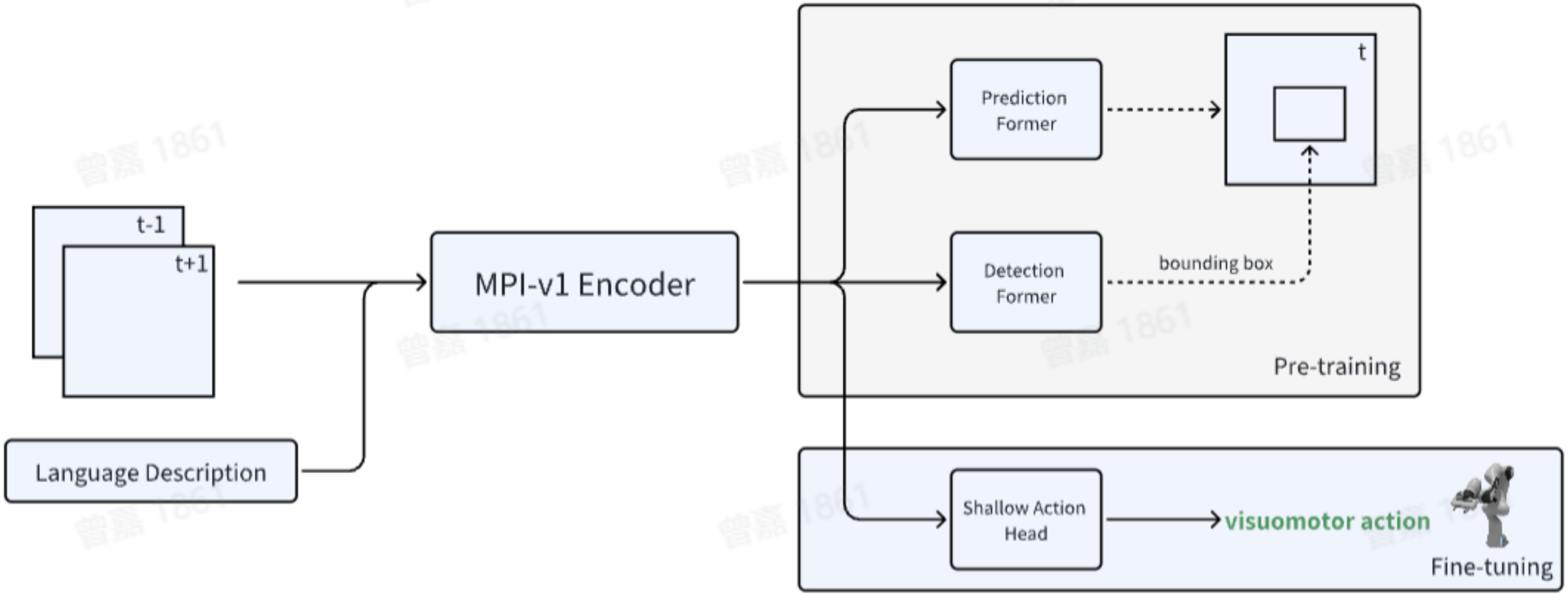

Decoder
q_{pred}
Prediction Transformer
Transition State
Bi-directional
Detection Transformer
q_{det}
MLP
Bi-direction Attn
Text Cross-Attn
Visual Cross-Attn
K,V
K,V
K,V
Q
MLP
Bi-direction Attn
K,V
Q

MPI | Network



Encoder
Decoder
Initial State
Final State
ViT
Q
K, V
Causality Modeling
v
K, V
Q
Token Aggregator
$\tilde{v}$
Q
$\tilde{l}$
K, V
l
Language Description
"Remove top of the blender"
Distill BERT
q_{pred}
Prediction Transformer
Transition State
Bi-directional
Detection Transformer
q_{det}
MLP
Bi-direction Attn
K, V
Text Cross-Attn
K, V
Visual Cross-Attn
K, V
Q
MLP
Bi-direction Attn
K, V
Q

MPI | Pipeline



t-1

t+1

Language Description

MPI-v1 Encoder

Prediction Former

t

Detection Former

bounding box

Pre-training

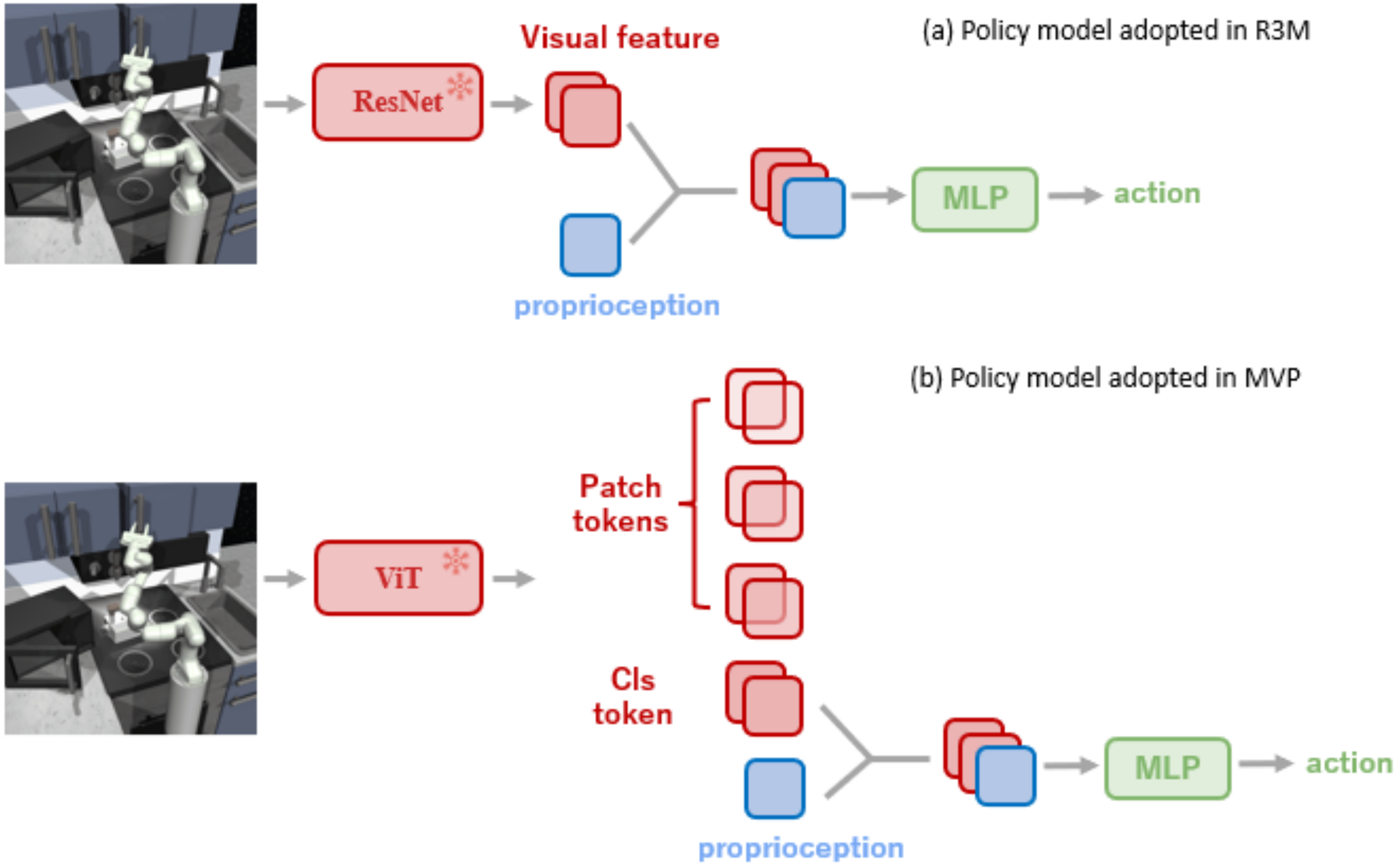
Shallow Action Head

visuomotor action

Fine-tuning

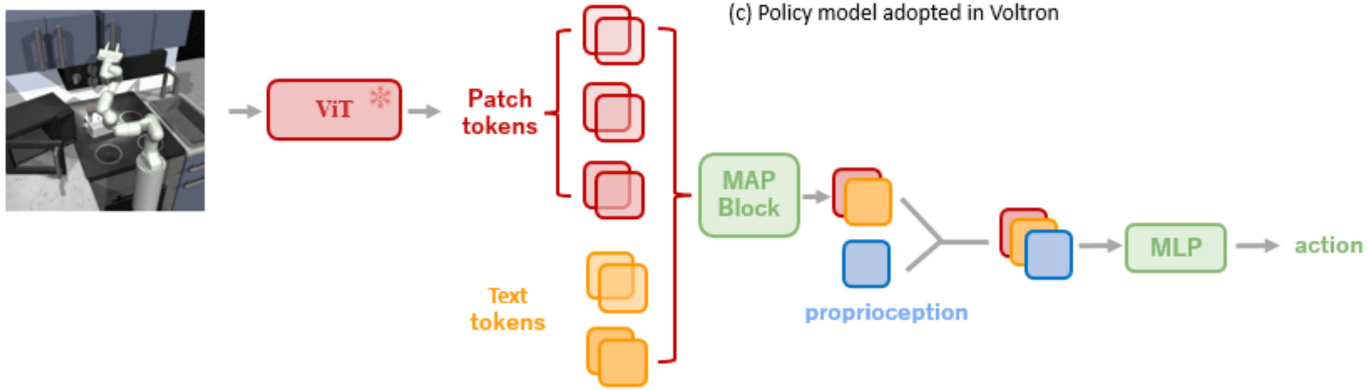
曾露 1861

MPI | Pipeline

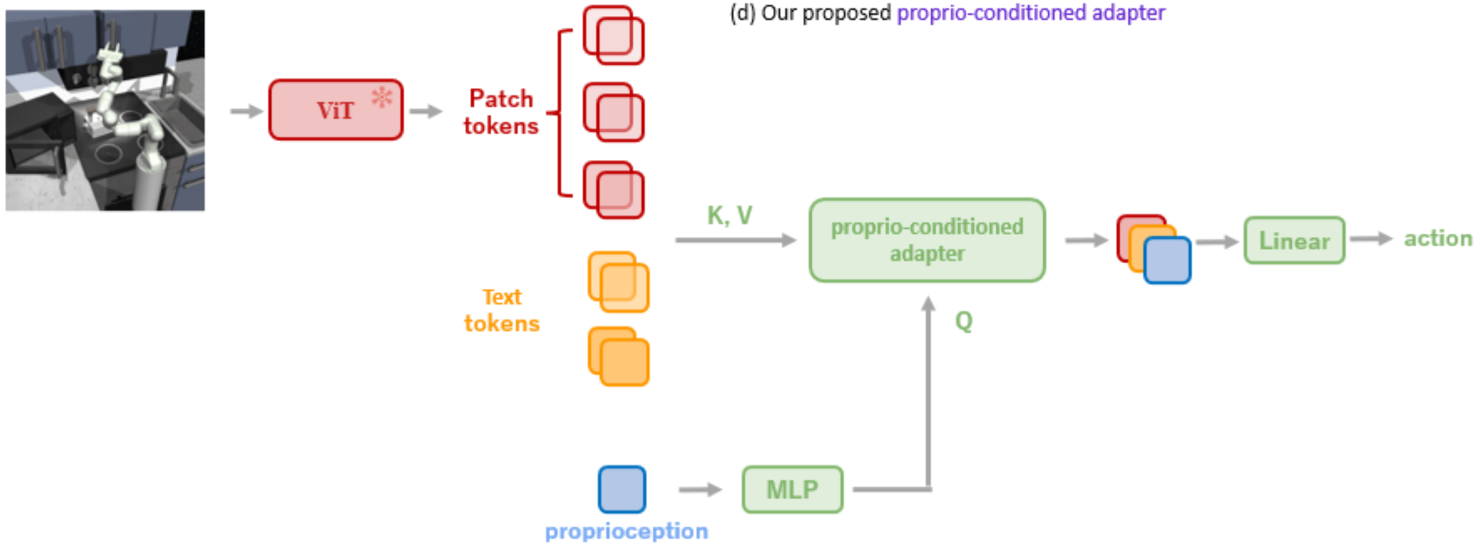


MPI | Pipeline

(c) Policy model adopted in Voltron



(d) Our proposed proprio-conditioned adapter

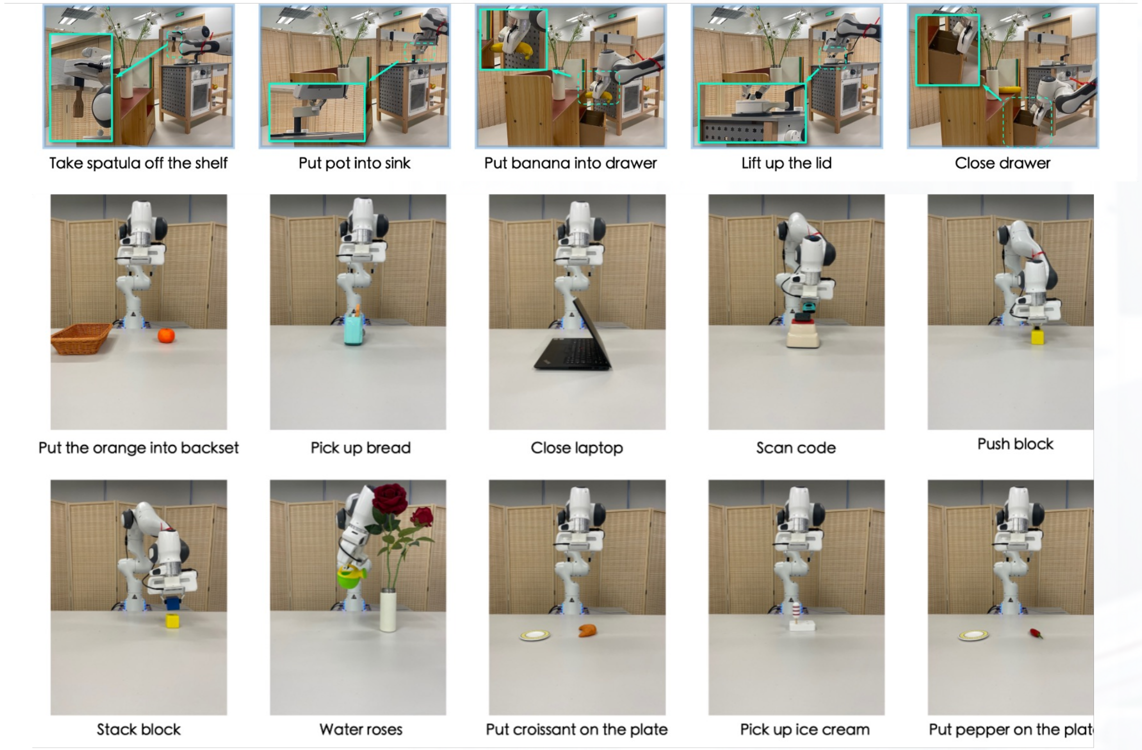


MPI | Experiments

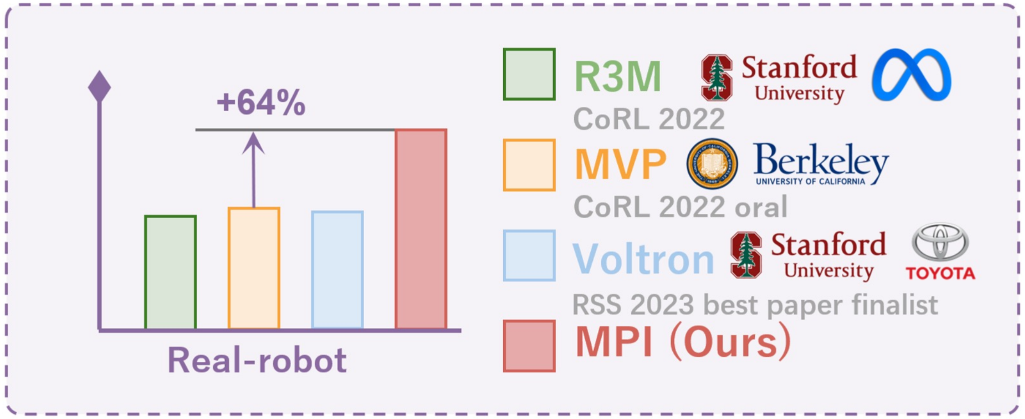
Real-robot Experiment Setting

5 complex kitchen environment

10 clean background


Take spatula off the shelf
Put pot into sink
Put banana into drawer
Lift up the lid
Close drawer
Put the orange into backset
Pick up bread
Close laptop
Scan code
Push block
Stack block
Water roses
Put croissant on the plate
Pick up ice cream
Put pepper on the plat

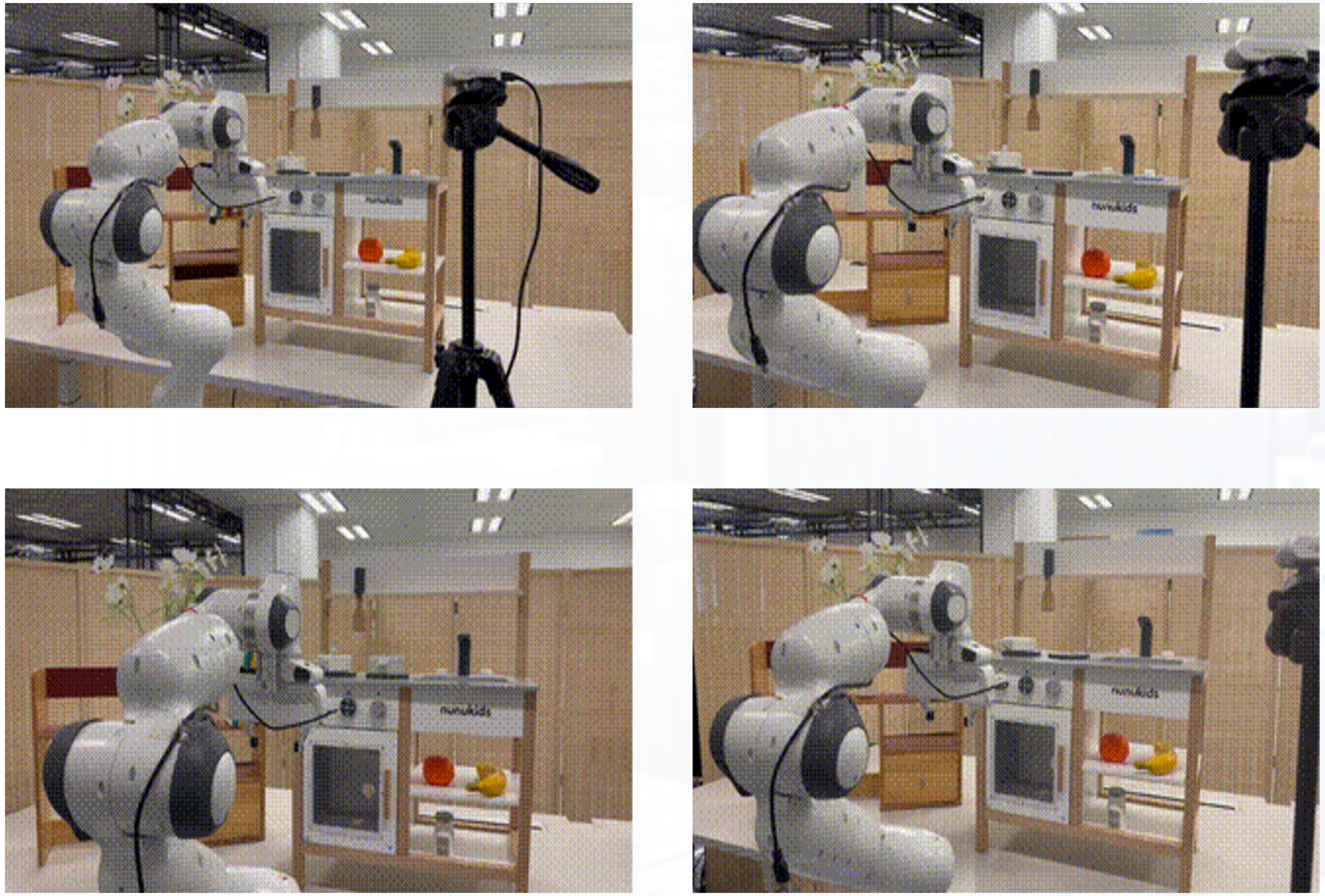
Performance Comparison



MPI - Testament on Real Robots

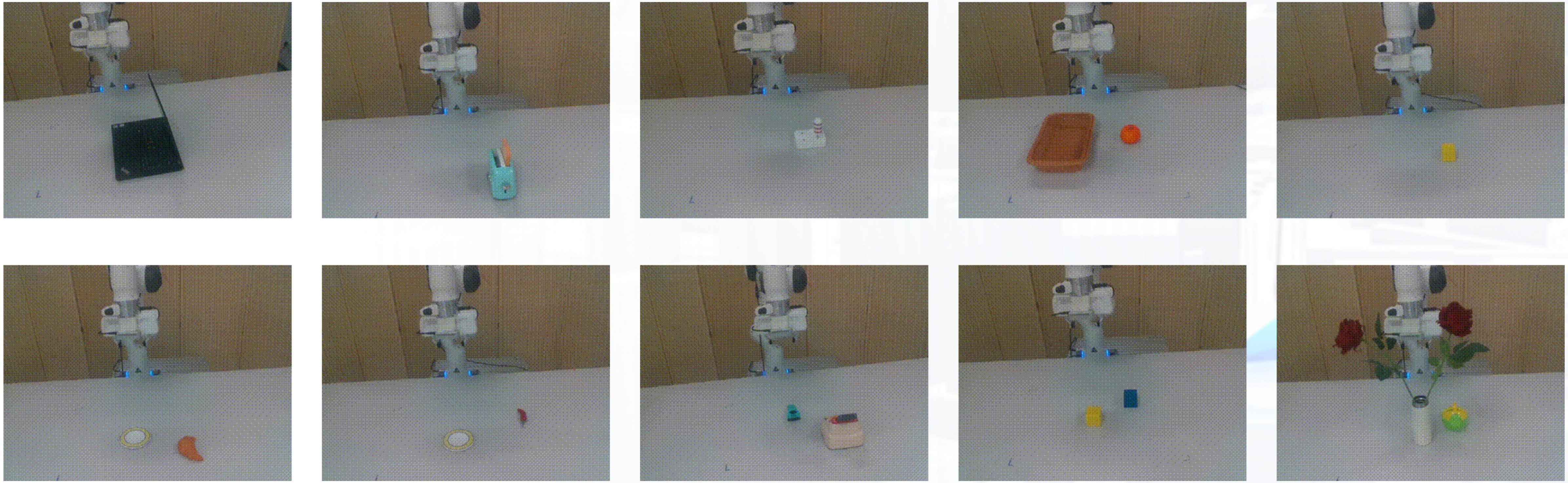
真机效果

Demos in kitchen environment



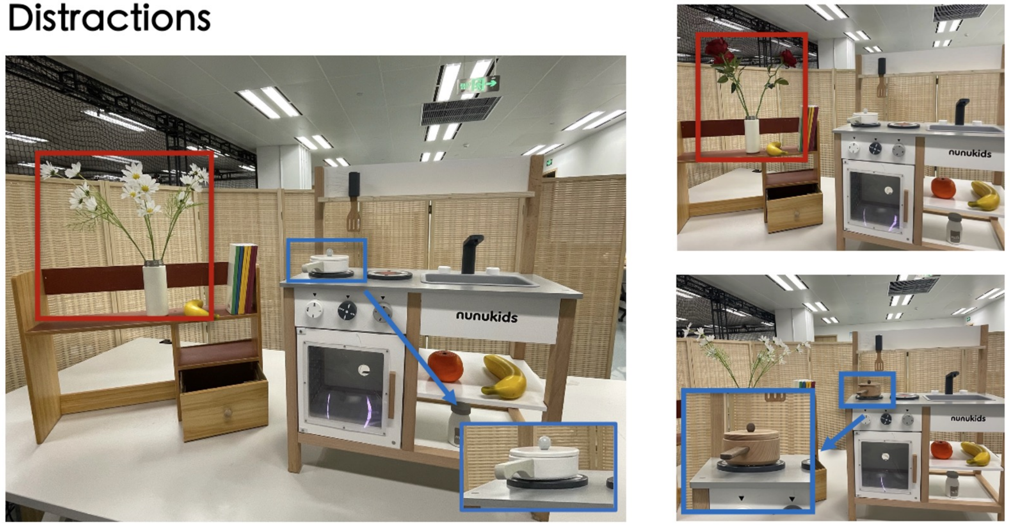
MPI - Testament on Real Robots 真机效果

Demos in clean background



MPI - Generalization

Generalization Validation

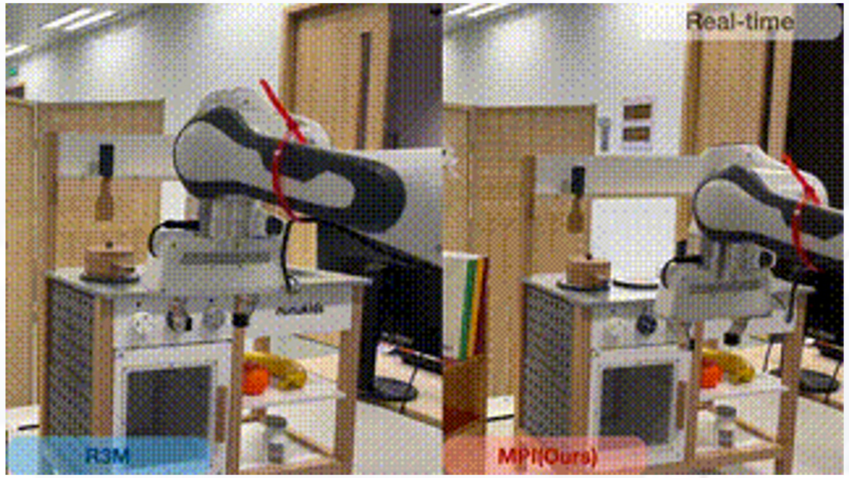

Robustness to Visual Distractions

(a) Original Setting

(b) BG. Distraction

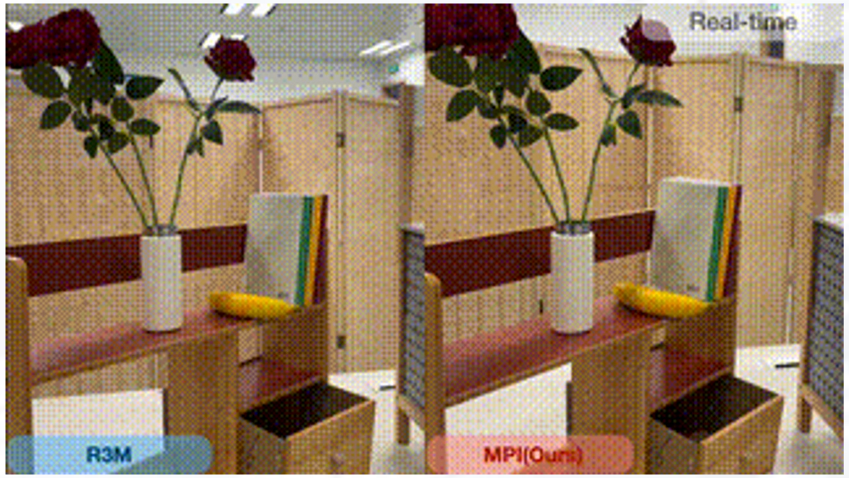
(c) Obj. Variation

Validation on generalization



Object Variation

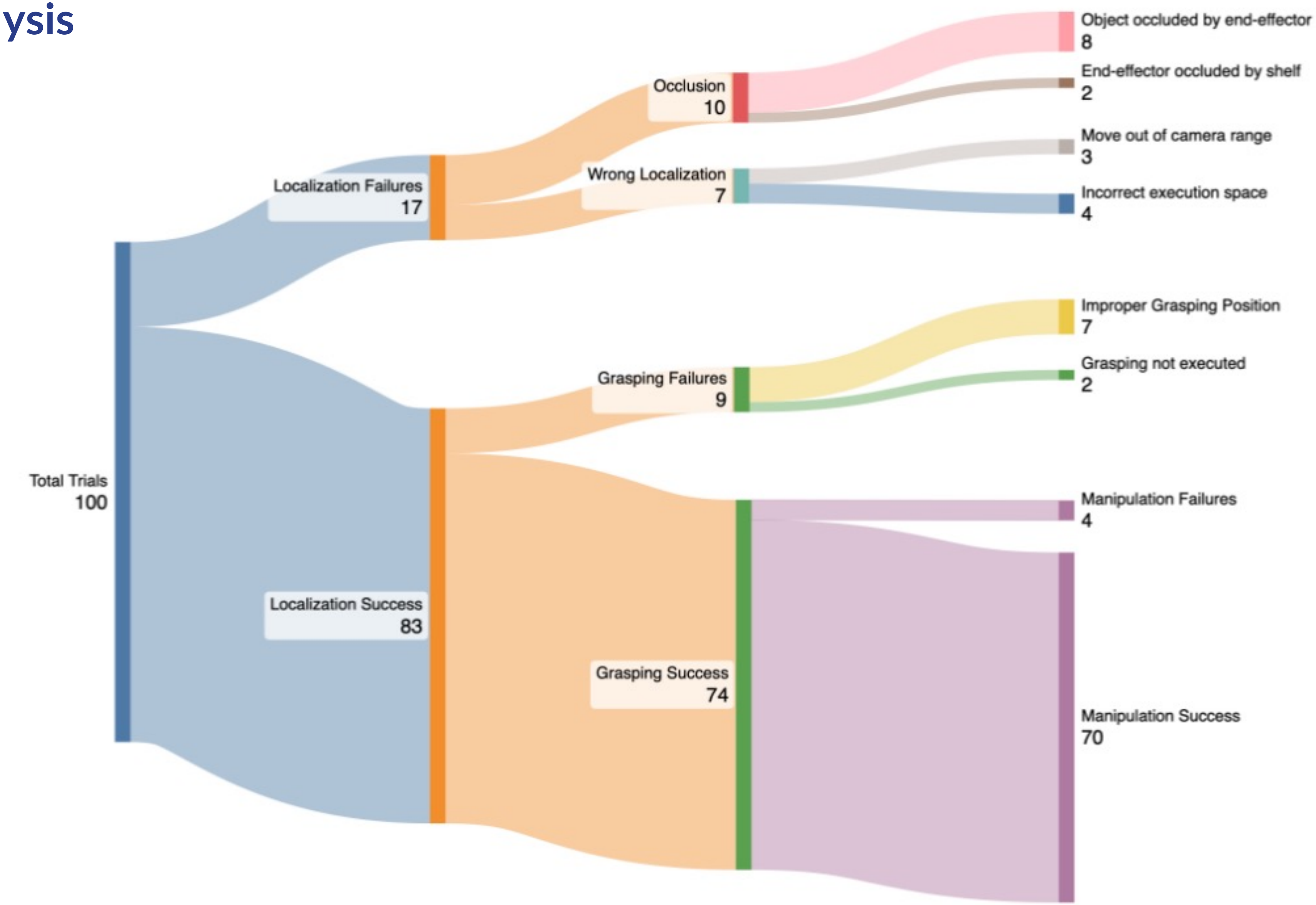
White plastic pot → Wooden pot



Background Distraction

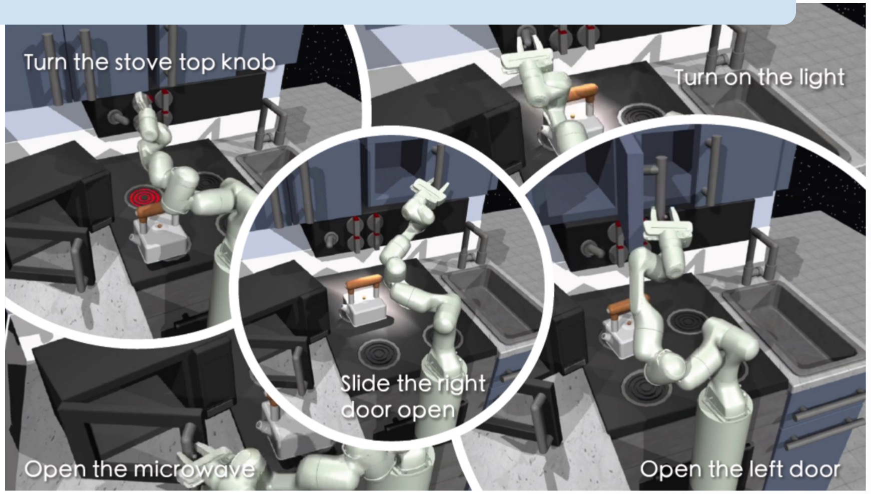
Daisies → Roses

Failure Analysis

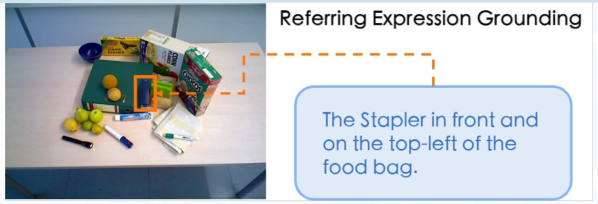


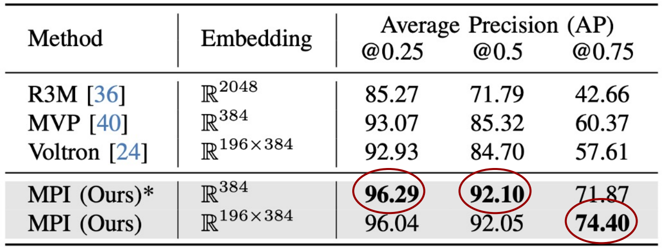
MPI | Experiments

Visuomotor Control in Simulation



Referring Expression Grounding

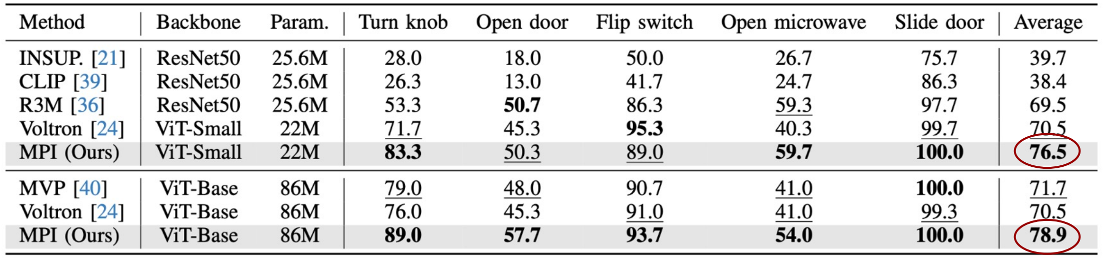

Referring Expression Grounding
The Stapler in front and on the top-left of the food bag.



Method	Embedding	Average Precision (AP) @0.25	@0.5	@0.75
R3M [36]	$\mathbb{R}^{2048}$	85.27	71.79	42.66
MVP [40]	$\mathbb{R}^{384}$	93.07	85.32	60.37
Voltron [24]	$\mathbb{R}^{196\times384}$	92.93	84.70	57.61
MPI (Ours)*	$\mathbb{R}^{384}$	**96.29**	**92.10**	71.87
MPI (Ours)	$\mathbb{R}^{196\times384}$	96.04	92.05	**74.40**

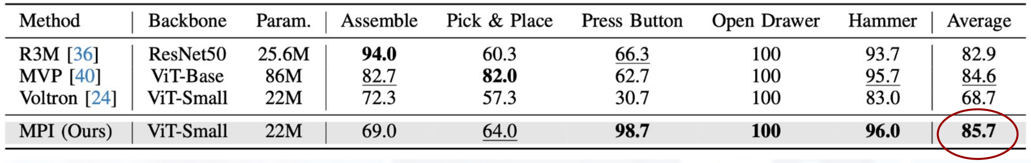
The experimental results reveal that MPI yields **state-of-the-art** performance on a broad 🏆 spectrum of downstream tasks.

TABLE II: **Results of single-task visuomotor control on Franka Kitchen.** We report the success rate (%) over 50 randomly sampled trajectories. We **bold** the best result for models with similar parameters and underline the second. "INSUP." represents classification-based supervised learning on ImageNet. MPI consistently exhibits superior performance across multiple tasks.



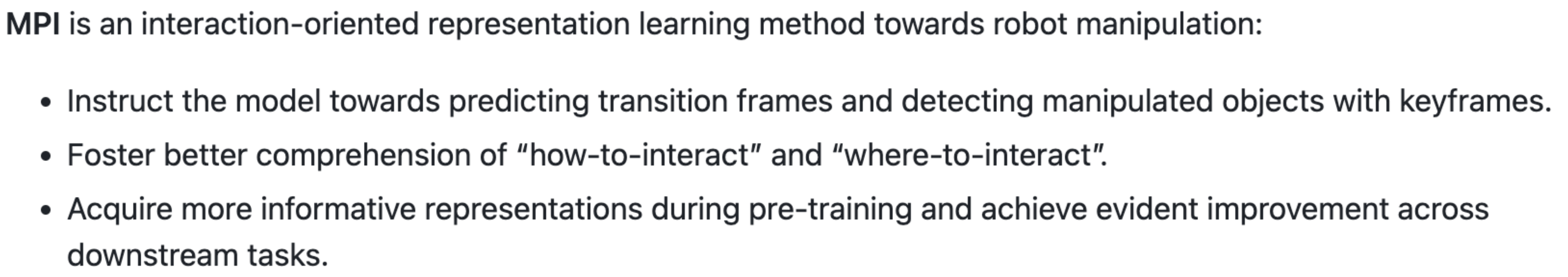
Method	Backbone	Param.	Turn knob	Open door	Flip switch	Open microwave	Slide door	Average
INSUP. [21]	ResNet50	25.6M	28.0	18.0	50.0	26.7	75.7	39.7
CLIP [39]	ResNet50	25.6M	26.3	13.0	41.7	24.7	86.3	38.4
R3M [36]	ResNet50	25.6M	53.3	**50.7**	86.3	59.3	97.7	69.5
Voltron [24]	ViT-Small	22M	71.7	45.3	**95.3**	40.3	99.7	70.5
MPI (Ours)	ViT-Small	22M	**83.3**	50.3	89.0	**59.7**	**100.0**	**76.5**
MVP [40]	ViT-Base	86M	79.0	48.0	90.7	41.0	**100.0**	71.7
Voltron [24]	ViT-Base	86M	76.0	45.3	91.0	41.0	99.3	70.5
MPI (Ours)	ViT-Base	86M	**89.0**	**57.7**	**93.7**	**54.0**	**100.0**	**78.9**

TABLE III: **Results of single-task visuomotor control on Meta-World simulation environment.** We report the success rate (%) over 50 randomly sampled trajectories. The best results are **bolded** and the second highest are underlined. MPI showcases exemplary performance across three tasks, exhibiting a superior average success rate in comparison to prior methods.



Method	Backbone	Param.	Assemble	Pick & Place	Press Button	Open Drawer	Hammer	Average
R3M [36]	ResNet50	25.6M	**94.0**	60.3	66.3	100	93.7	82.9
MVP [40]	ViT-Base	86M	82.7	**82.0**	62.7	100	95.7	84.6
Voltron [24]	ViT-Small	22M	72.3	57.3	30.7	100	83.0	68.7
MPI (Ours)	ViT-Small	22M	69.0	64.0	**98.7**	**100**	**96.0**	**85.7**

Conclusion and Limitation



MPI is an interaction-oriented representation learning method towards robot manipulation:

- Instruct the model towards predicting transition frames and detecting manipulated objects with keyframes.
- Foster better comprehension of “how-to-interact” and “where-to-interact”.
- Acquire more informative representations during pre-training and achieve evident improvement across downstream tasks.

Limitation:

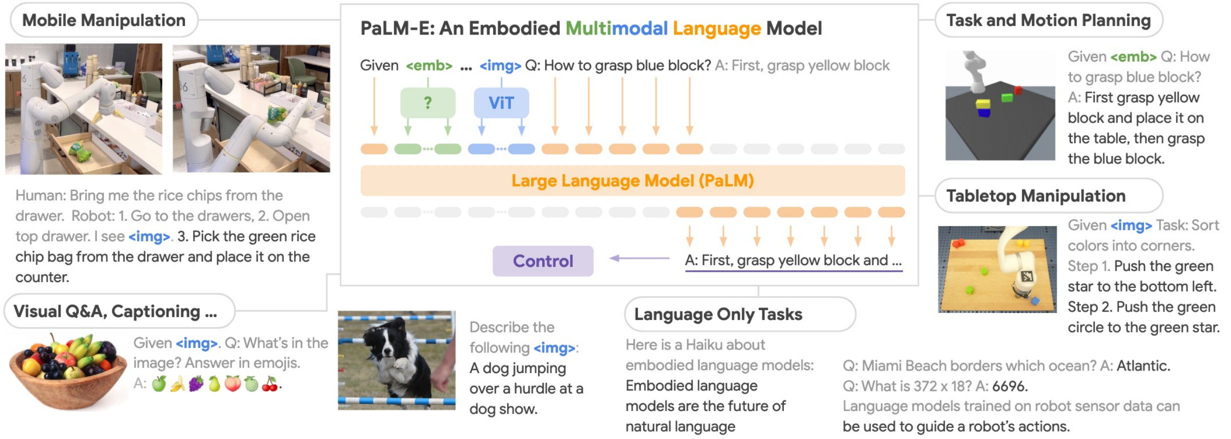
Our framework by far utilizes explicit annotations i.e. keyframes, languages, and bounding boxes for interaction object) provided in the Ego4D-HoI dataset. This could limit the applicability of our methods to broader datasets.

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

What’s Next?

Embodied Multimodal Language Model


Mobile Manipulation
PaLM-E: An Embodied Multimodal Language Model
Task and Motion Planning
?
ViT
Large Language Model (PaLM)
Control
Tabletop Manipulation
Visual Q&A, Captioning ...
Language Only Tasks

Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block

A: First, grasp yellow block and ...

Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see <img>. 3. Pick the green rice chip bag from the drawer and place it on the counter.

Given <emb> Q: How to grasp blue block? A: First grasp yellow block and place it on the table, then grasp the blue block.

Given <img> Task: Sort colors into corners. Step 1. Push the green star to the bottom left. Step 2. Push the green circle to the green star.

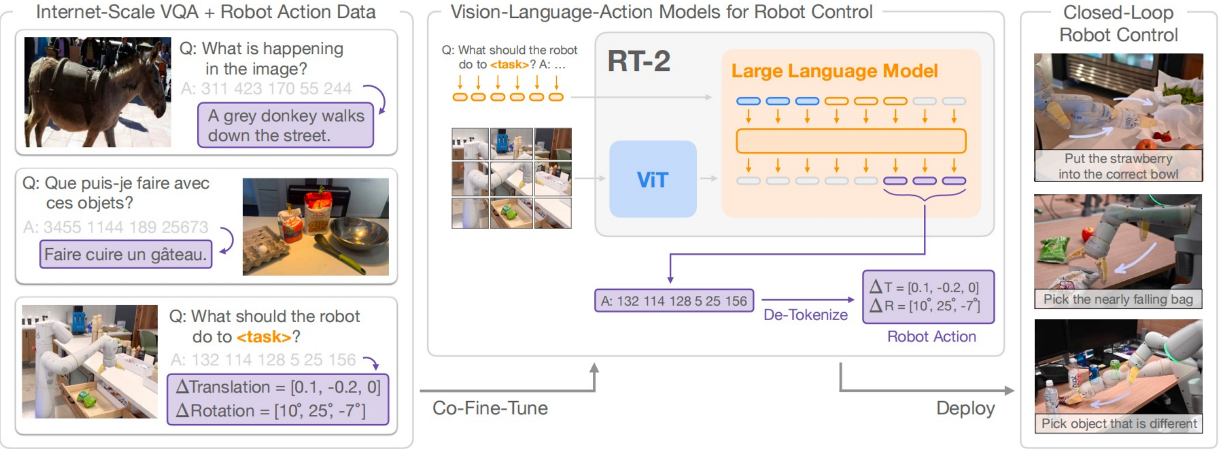
Given <img>. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍈🍒.

Describe the following <img>: A dog jumping over a hurdle at a dog show.

Here is a Haiku about embodied language models: Embodied language models are the future of natural language

Q: Miami Beach borders which ocean? A: Atlantic. Q: What is 372 x 18? A: 6696. Language models trained on robot sensor data can be used to guide a robot's actions.

PaLM-E


Internet-Scale VQA + Robot Action Data
Vision-Language-Action Models for Robot Control
Closed-Loop Robot Control
RT-2
Large Language Model
ViT
A: 132 114 128 5 25 156
De-Tokenize
ΔT = [0.1, -0.2, 0]
ΔR = [10°, 25°, -7°]
Robot Action
Co-Fine-Tune
Deploy
Put the strawberry into the correct bowl
Pick the nearly falling bag
Pick object that is different

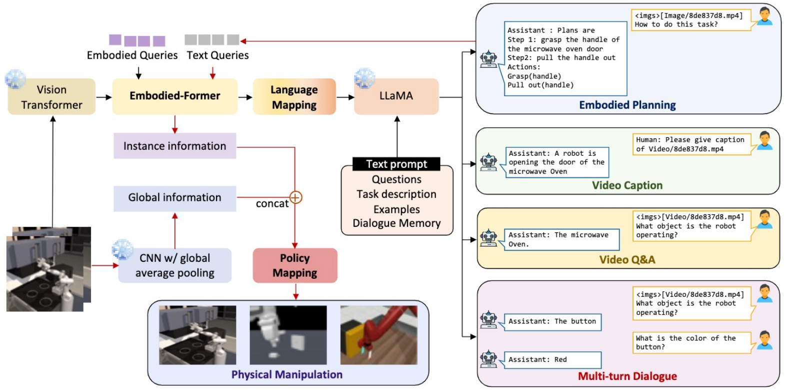
Q: What is happening in the image? A: 311 423 170 55 244 A grey donkey walks down the street.

Q: Que puis-je faire avec ces objets? A: 3455 1144 189 25673 Faire cuire un gâteau.

Q: What should the robot do to <task>? A: 132 114 128 5 25 156 ΔTranslation = [0.1, -0.2, 0] ΔRotation = [10°, 25°, -7°]

Q: What should the robot do to <task>? A: ...

RT-2



<imgs>[Image/8de837d8.mp4] How to do this task?

Assistant : Plans are Step 1: grasp the handle of the microwave oven door Step2: pull the handle out Actions: Grasp(handle) Pull out(handle)

Human: Please give caption of Video/8de837d8.mp4

Assistant: A robot is opening the door of the microwave Oven

<imgs>[Video/8de837d8.mp4] What object is the robot operating?

Assistant: The microwave Oven.

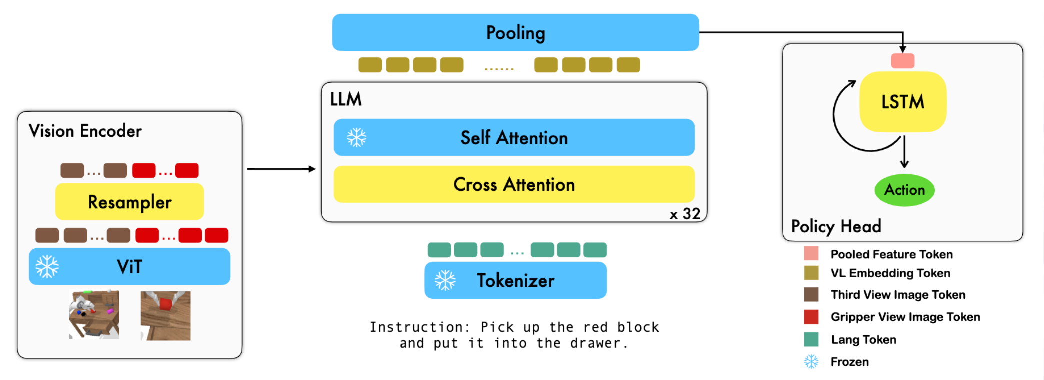
<imgs>[Video/8de837d8.mp4] What object is the robot operating?

Assistant: The button

What is the color of the button?

Assistant: Red

EmbodiedGPT

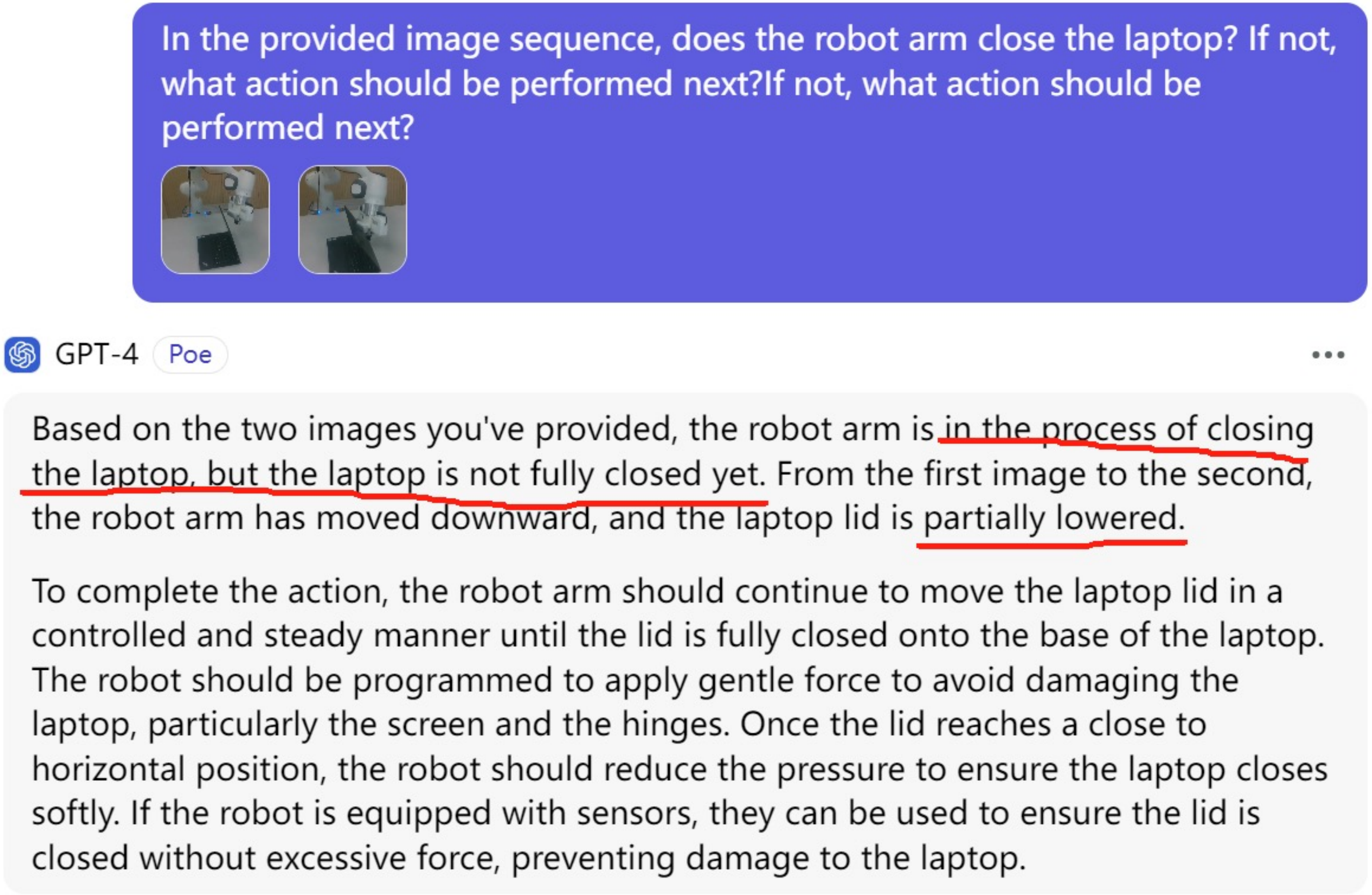


Instruction: Pick up the red block and put it into the drawer.

RoboFlamingo

- Vision-language models trained on Internet-scale data can be incorporated directly into **end-to-end robotic control**
- Goal: to **boost generalization** and enable emergent semantic reasoning
- Robotic tasks naturally fits into language at dissecting tasks step by step using language (prompt).
- Is it the **right way** to open the language tool box in Robotics?

Visual Pre-training for Multimodal Model

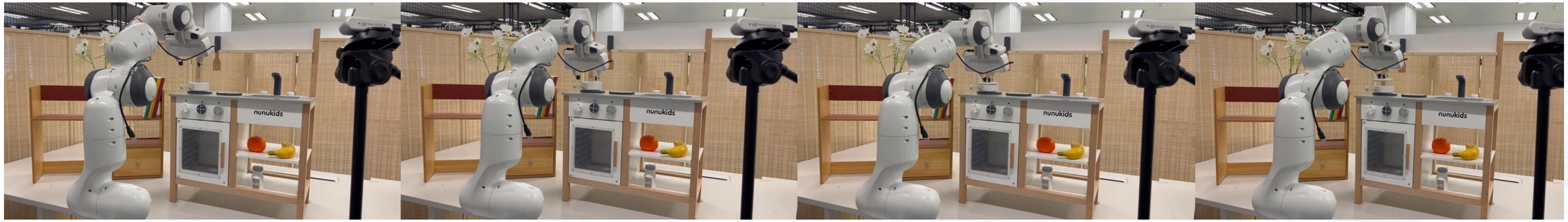


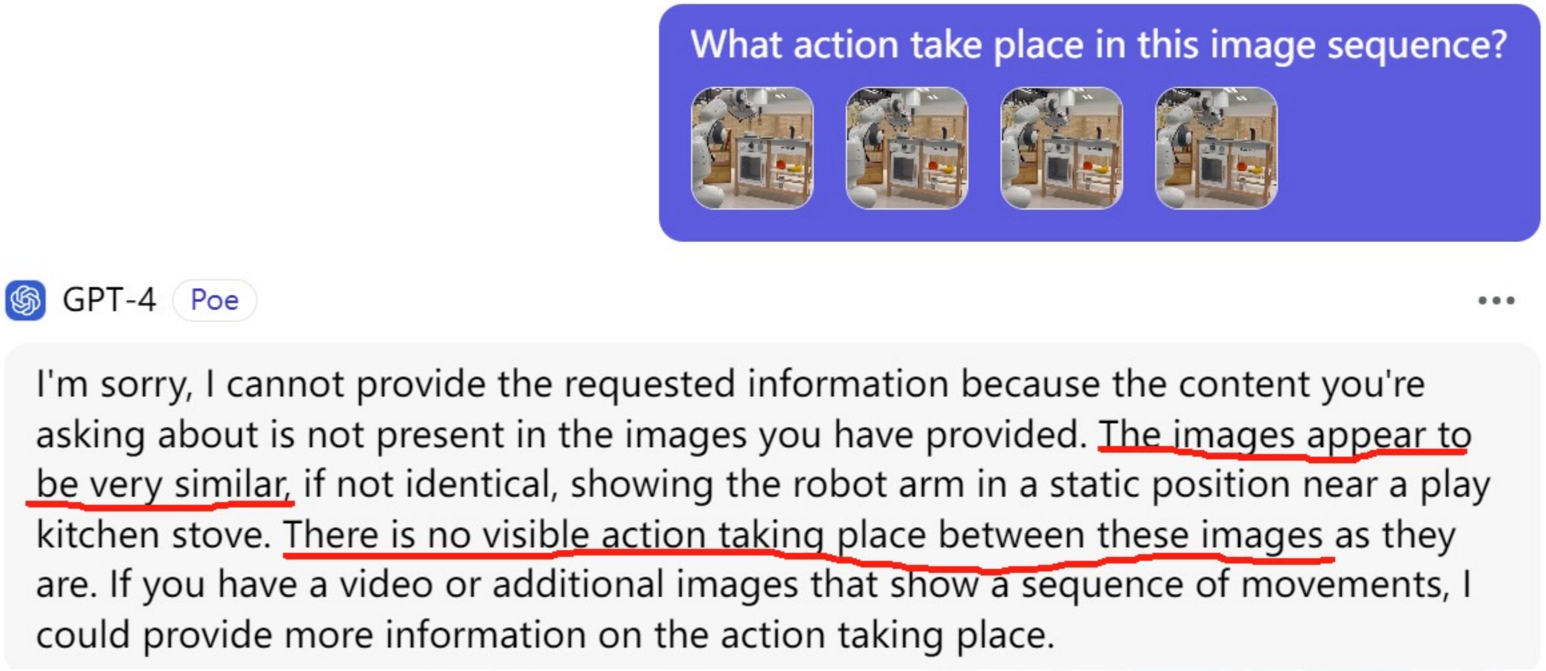
In the provided image sequence, does the robot arm close the laptop? If not, what action should be performed next?If not, what action should be performed next?

Based on the two images you've provided, the robot arm is in the process of closing the laptop, but the laptop is not fully closed yet. From the first image to the second, the robot arm has moved downward, and the laptop lid is partially lowered.

To complete the action, the robot arm should continue to move the laptop lid in a controlled and steady manner until the lid is fully closed onto the base of the laptop. The robot should be programmed to apply gentle force to avoid damaging the laptop, particularly the screen and the hinges. Once the lid reaches a close to horizontal position, the robot should reduce the pressure to ensure the laptop closes softly. If the robot is equipped with sensors, they can be used to ensure the lid is closed without excessive force, preventing damage to the laptop.

Visual Pre-training for MultiModal Model





What action take place in this image sequence?

I'm sorry, I cannot provide the requested information because the content you're asking about is not present in the images you have provided. The images appear to be very similar, if not identical, showing the robot arm in a static position near a play kitchen stove. There is no visible action taking place between these images as they are. If you have a video or additional images that show a sequence of movements, I could provide more information on the action taking place.

Existing Vision-Language Models struggle to capture such fine-grained visual information.

Visual pre-training can equip a visual encoder with enhanced sensitivity to fine-grained information within interactive environments.

Foundation model for long-horizon tasks

大方向

- **Introducing Vision-Language Models and Diffusion models can significantly enhance generalization capabilities.**

待突破问题

- Unable to handle tasks with **large temporal spans** and **high complexity**.
- Lacks **self-evaluation** and **self-correction** capabilities.

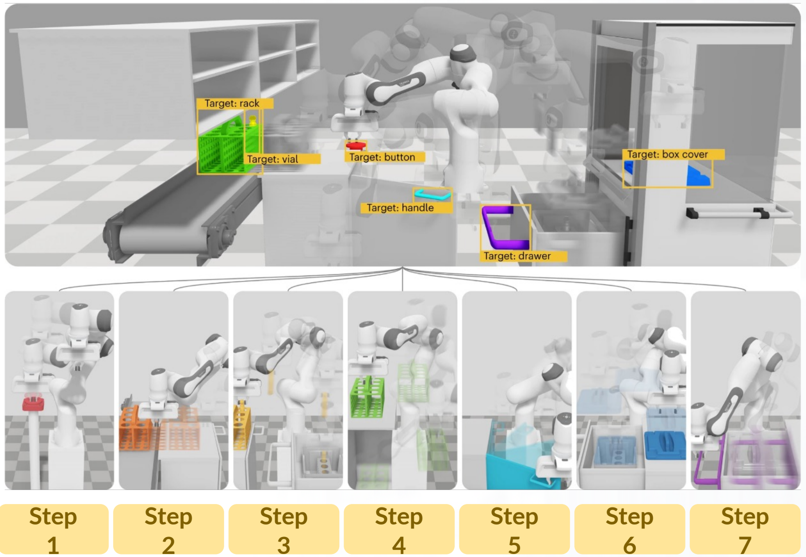


photo credit from "Hybrid hierarchical learning for solving complex sequential tasks using the robotic manipulation network ROMAN", Nature Machine Intelligence, 2023

Objective: Strengthen the spatiotemporal perception and causal reasoning capabilities of embodied agents

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Thanks