

Towards Safe and Aligned Embodied AI in the Era of Robotics Foundation Models

Thomas (Ran) Tian
rantian@berkeley.edu



Final-year Ph.D. student at UC Berkeley

Supervised by Prof. Masayoshi Tomizuka and Prof. Andrea Bajcsy

My Research Trajectory

Constrained
stochastic control

Motion planning

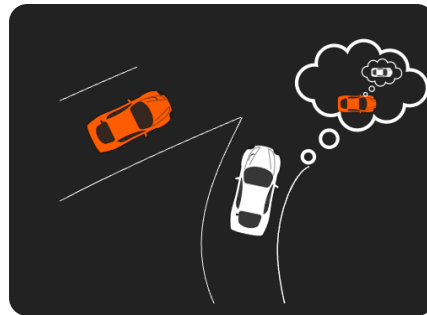
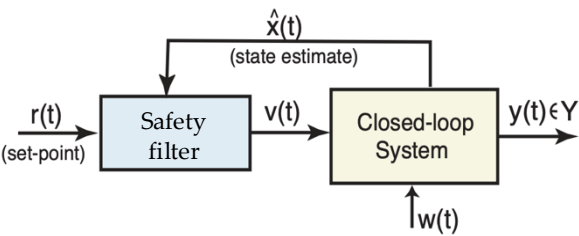


Foundation driving model

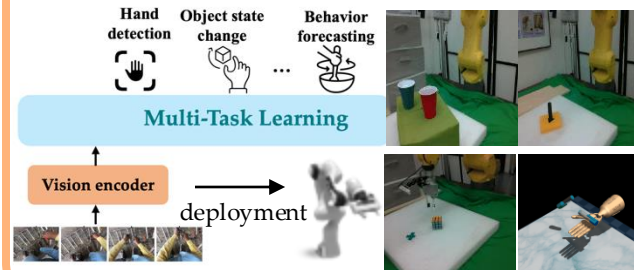
Pre-training Post-training Efficient deployment

Dynamical game

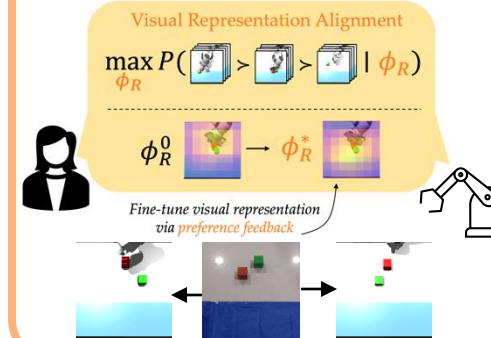
Safety and alignment



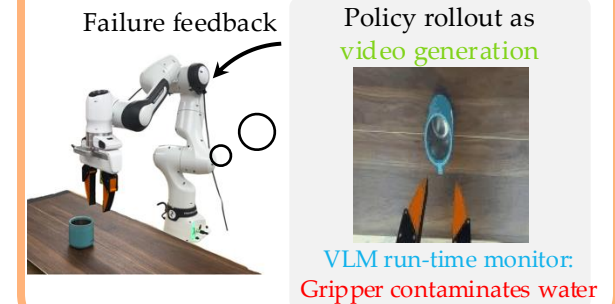
Representation & policy
pre-training



Post-training
preference alignment



Failure prediction and
mitigation beyond collision



A modern control view of robot autonomy

$$\max_{\pi} \mathbb{E} \left[\sum_{\tau=0}^{T-1} \underbrace{r_R}_{\text{Task reward}}(\underbrace{s_{t+\tau}, \pi(s_{t+\tau})}_{\text{State space representation of the environment}}) + \underbrace{V(s_{t+T})}_{\text{Terminal reward}} \right] \quad \text{s.t.} \quad \begin{aligned} &\underbrace{P(s_{t+1} | s_t, a_t)}_{\text{Dynamics}} \\ &\underbrace{P(s_{t+\tau} \in \text{safe set}) \geq \Delta}_{\text{Desired safety guarantee}} \end{aligned}$$

Model-predictive control

Dynamic programming

Reach-avoid game

Belief space planning

...

A modern control view of robot autonomy

$$\max_{\pi} \mathbb{E} \left[\sum_{\tau=0}^{T-1} r_R(s_{t+\tau}, \pi(s_{t+\tau})) + V(s_{t+T}) \right] \quad \text{s.t.} \quad \begin{aligned} &P(s_{t+1}|s_t, a_t) \\ &P(s_{t+\tau} \in \text{safe set}) \geq \Delta \end{aligned}$$

When this works great?

Constrained, parsed, and well modeled environment

Other “agents” are well defined

State is clearly defined

Can know how state evolves given actions of all “agents”

Have the tools to efficiently solve this problem

A modern control view of robot autonomy

$$\max_{\pi} \mathbb{E} \left[\sum_{\tau=0}^{T-1} r_R(s_{t+\tau}, \pi(s_{t+\tau})) + V(s_{t+T}) \right] \quad \text{s.t.} \quad \begin{aligned} &P(s_{t+1} | s_t, a_t) \\ &P(s_{t+\tau} \in \text{safe set}) \geq \Delta \end{aligned}$$

When this works great?

Constrained, parsed, and well modeled environment

Other “agents” are well defined

State is clearly defined

Can know how state evolves given actions of all “agents”

Have the tools to efficiently solve this problem

But in real world, ...



A modern control view of robot autonomy

$$\max_{\pi} \mathbb{E} \left[\sum_{\tau=0}^{T-1} r_R(s_{t+\tau}, \pi(s_{t+\tau})) + V(s_{t+T}) \right] \quad \text{s.t.} \quad \begin{aligned} &P(s_{t+1}|s_t, a_t) \\ &P(s_{t+\tau} \in \text{safe set}) \geq \Delta \end{aligned}$$

When this works great?

Constrained, parsed, and well modeled environment

Other “agents” are well defined

State is clearly defined

Can know how state evolves given actions of all “agents”

Have the tools to efficiently solve this problem

But in real world, ...



A modern control view of robot autonomy

$$\max_{\pi} \mathbb{E} \left[\sum_{\tau=0}^{T-1} r_R(s_{t+\tau}, \pi(s_{t+\tau})) + V(s_{t+T}) \right] \quad \text{s.t.} \quad \begin{aligned} &P(s_{t+1}|s_t, a_t) \\ &P(s_{t+\tau} \in \text{safe set}) \geq \Delta \end{aligned}$$

When this works great?

Constrained, parsed, and well modeled environment

Other “agents” are well defined

State is clearly defined

Can know how state evolves given actions of all “agents”

Have the tools to efficiently solve this problem

But in real world, ...



From modern control theory to autonomy 2.0 and beyond

Beginning:
X-MPC with symbolic representations
defined by engineers

State-space
representation

Model-
based approaches



Perception models

Structured information

- Agent property
 - Bounding box
 - Label
 - Obs status
 - Risk level
- Lane property
- ...

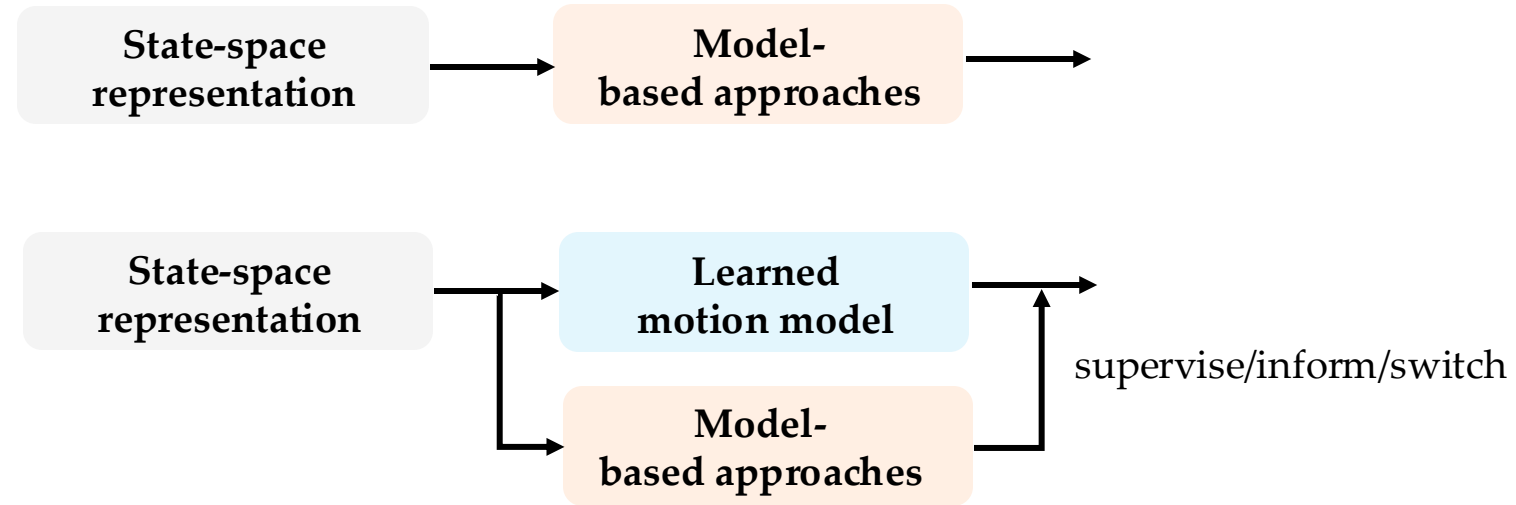
Formulate planning
optimization

Model-
based approaches

From modern control theory to autonomy 2.0 and beyond

Beginning:
X-MPC with symbolic representations
defined by engineers

Autonomy 1.0 :
[system 1, system 2] with symbolic
representations



Human behavior prediction



Neural dynamics



Control gain scheduling

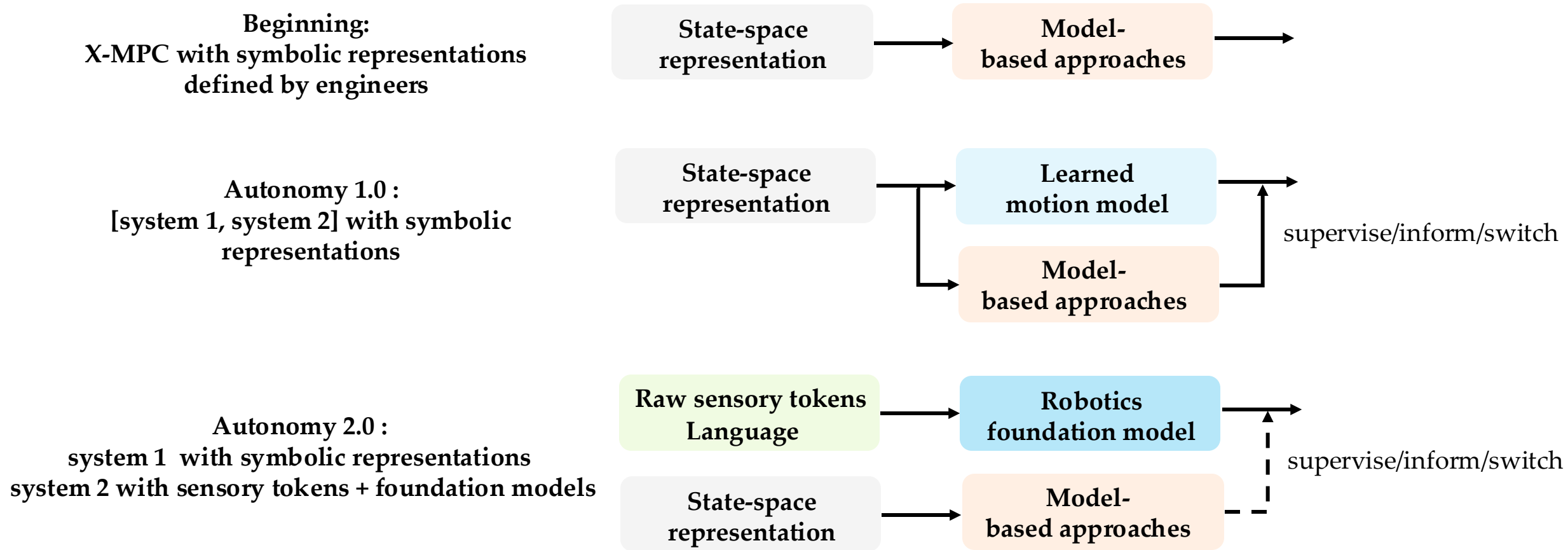
...



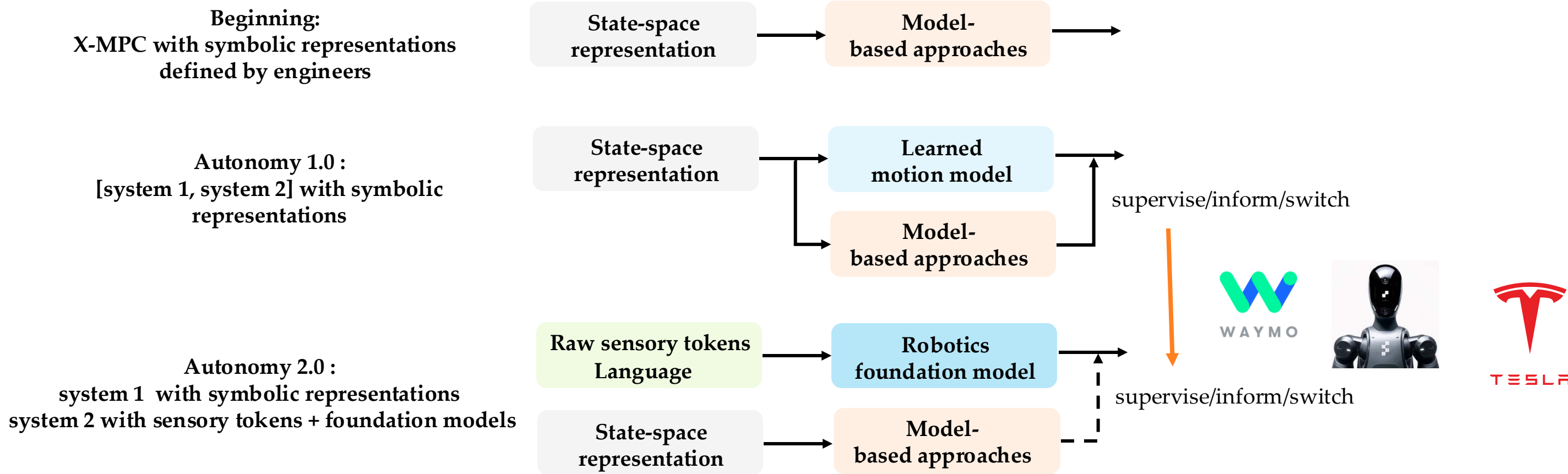
[Nayakanti, Nigamaa, et al. "Wayformer: Motion forecasting via simple & efficient attention networks." ICRA, 2023.]

[Wang, Changhao, et al. "Safe online gain optimization for cartesian space variable impedance control." CASE. IEEE, 2022.]

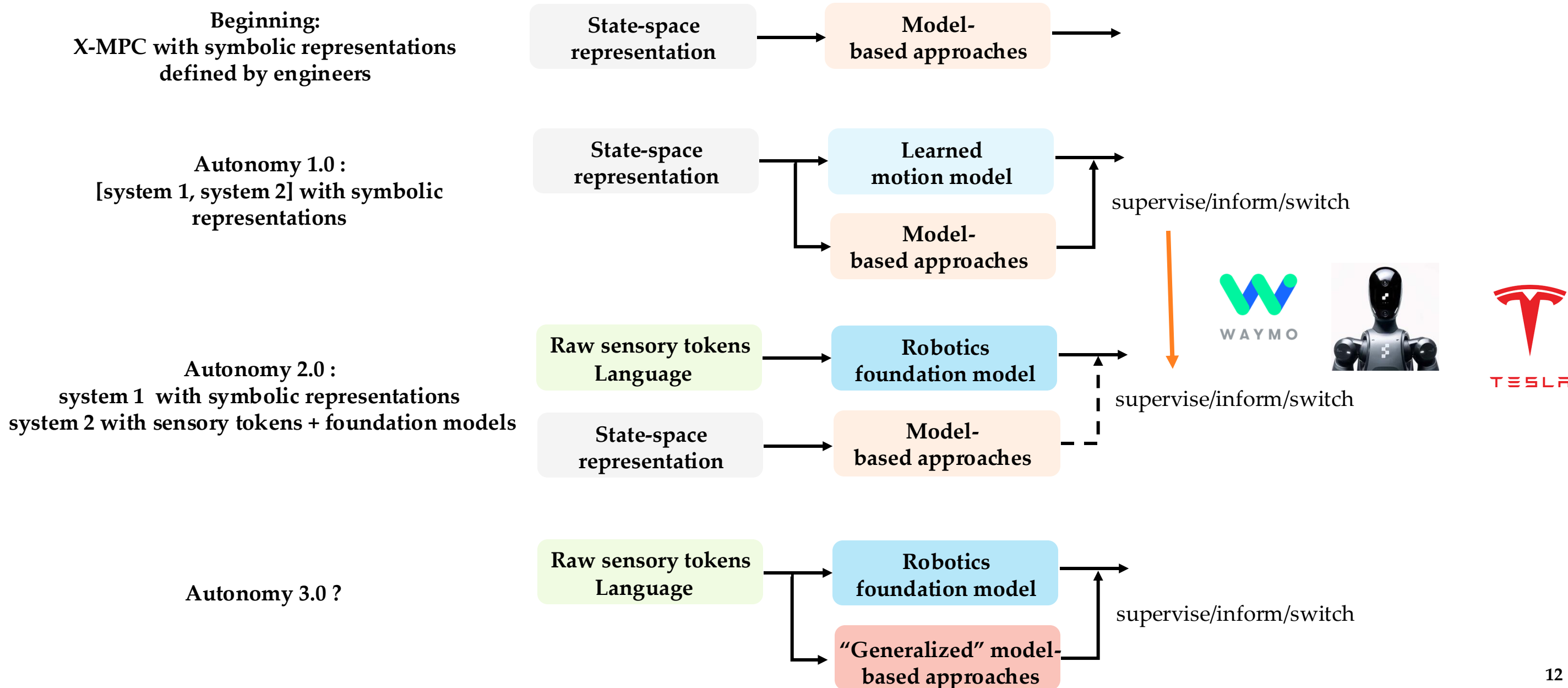
From modern control theory to autonomy 2.0 and beyond



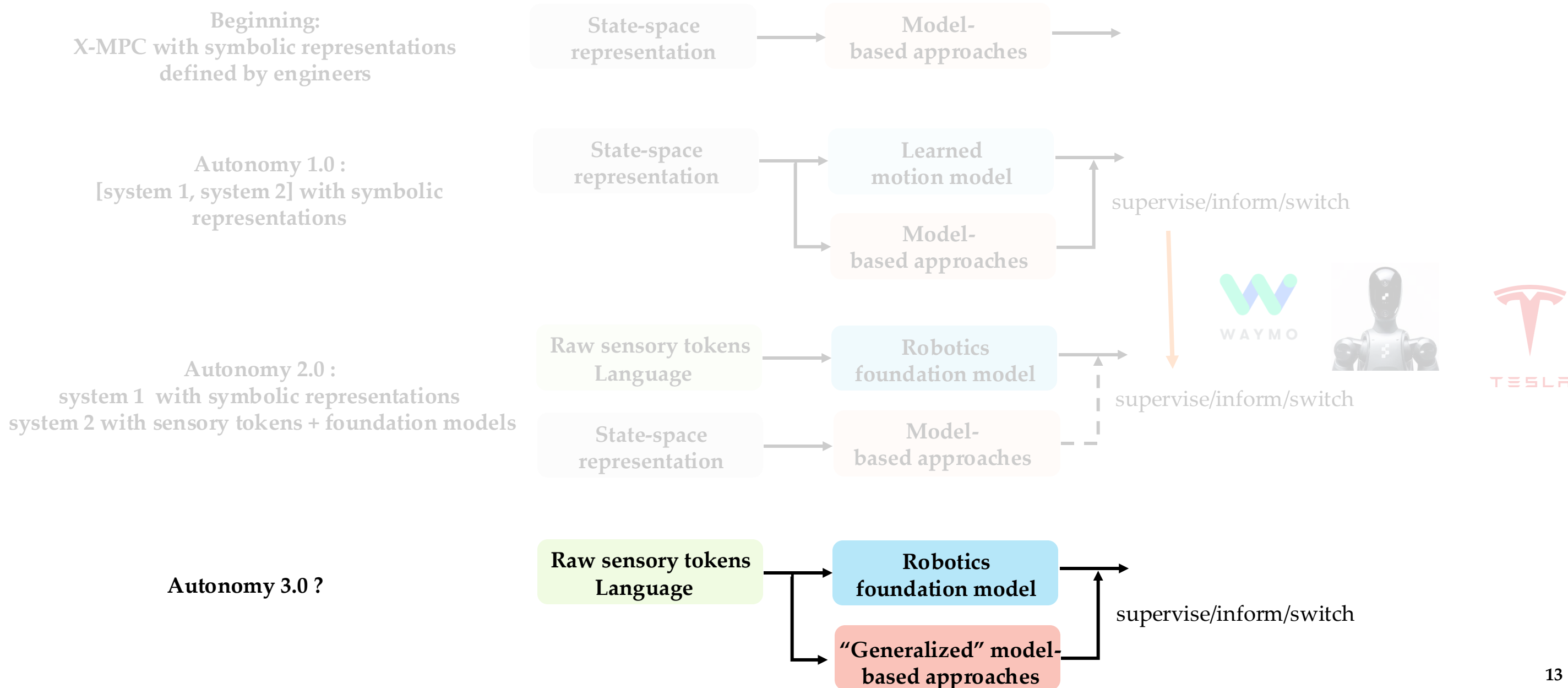
From modern control theory to autonomy 2.0 and beyond



From modern control theory to autonomy 2.0 and beyond



From modern control theory to autonomy 2.0 and beyond



Behavior cloning for robot learning

Let's remember what the expert did and copy them!



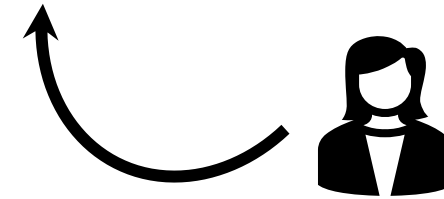
Behavior cloning for robot learning

Let's remember what the expert
did and copy them!

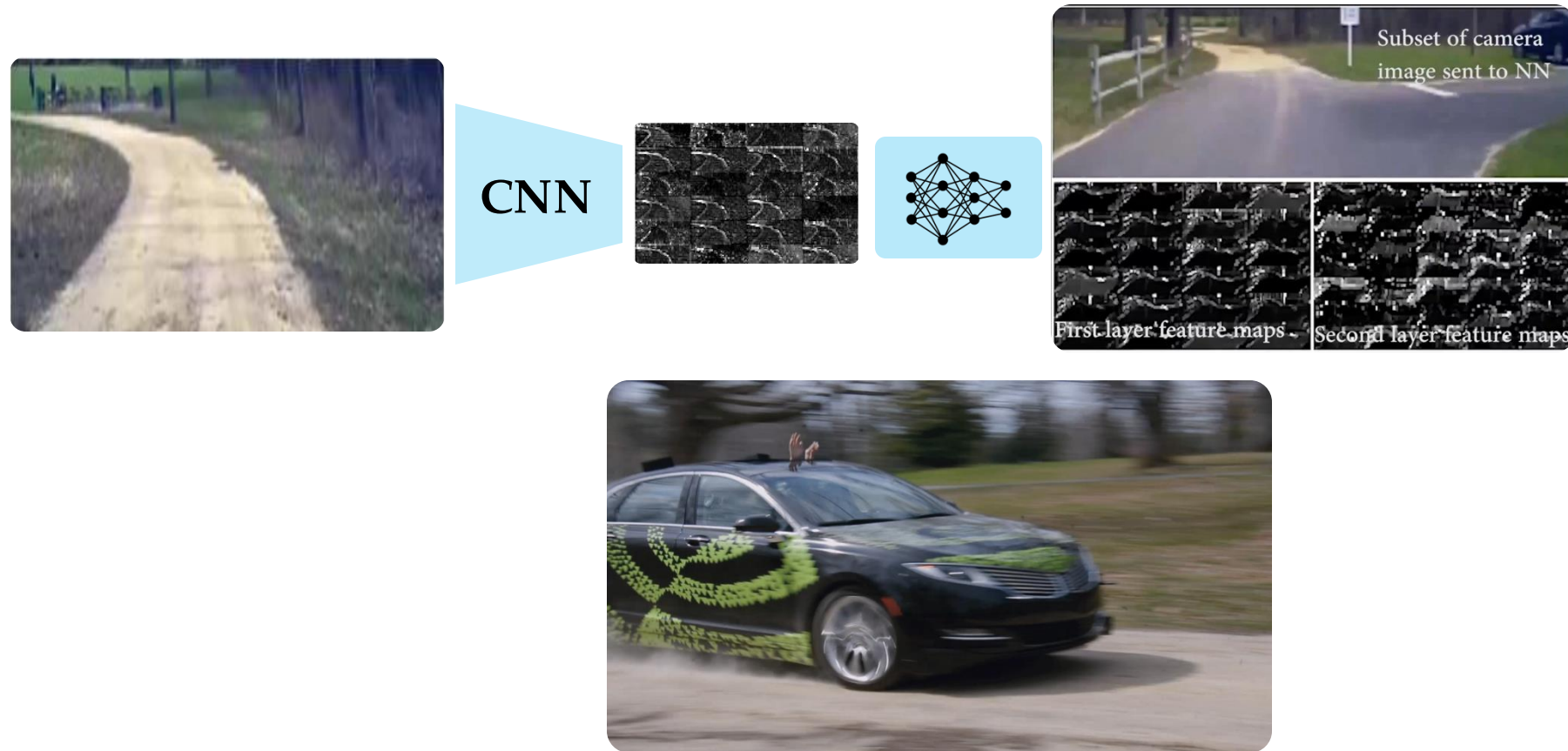
$$\max_{\theta} \mathbb{E} [\mathbb{P}_{\theta}(\mathbf{a}_{0:T}^* | \mathbf{o}_0; \text{context})]$$



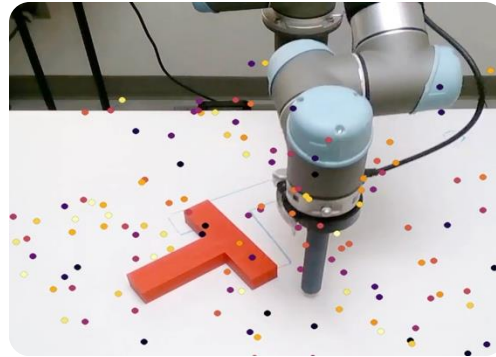
$$\{(\text{context}, \mathbf{o}_0, \mathbf{a}_{0:T}^*)^1, \dots\}$$



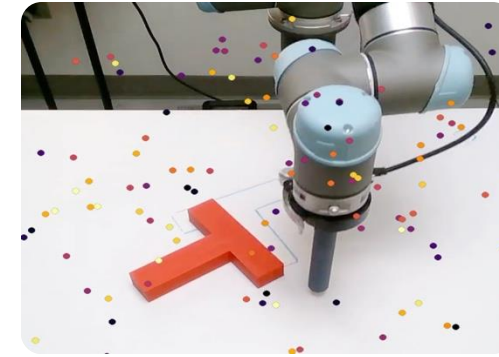
10 years ago – CNN based Policy Model



2 years ago – Diffusion Policy Model



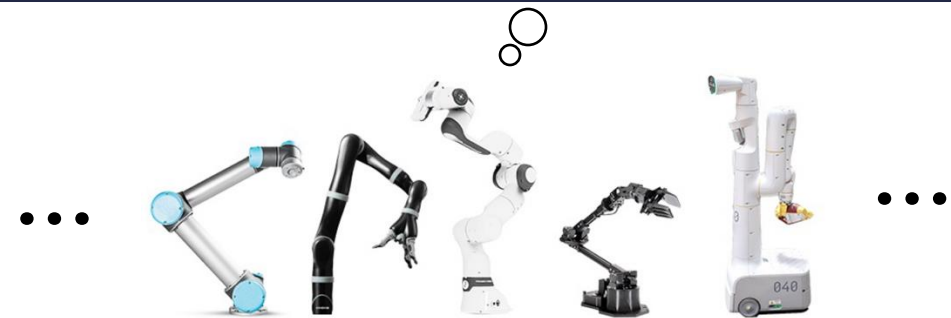
Diffusion Model



Large-scale data is a key factor for robotics foundation models

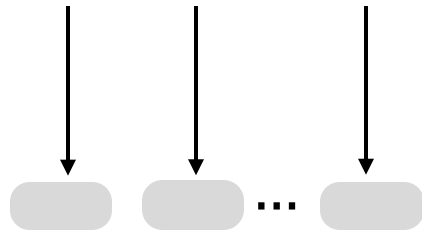


$\{(\text{observation}, \text{action}, \text{task spec})_t\}$



Now – Vision-Language-Action Robotics Foundation Model

*Q: Wha should the robot do
to pick up the chip bag?*

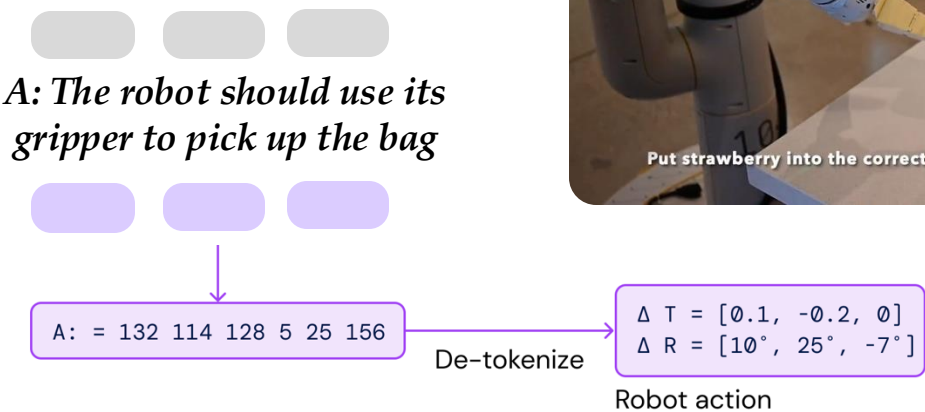
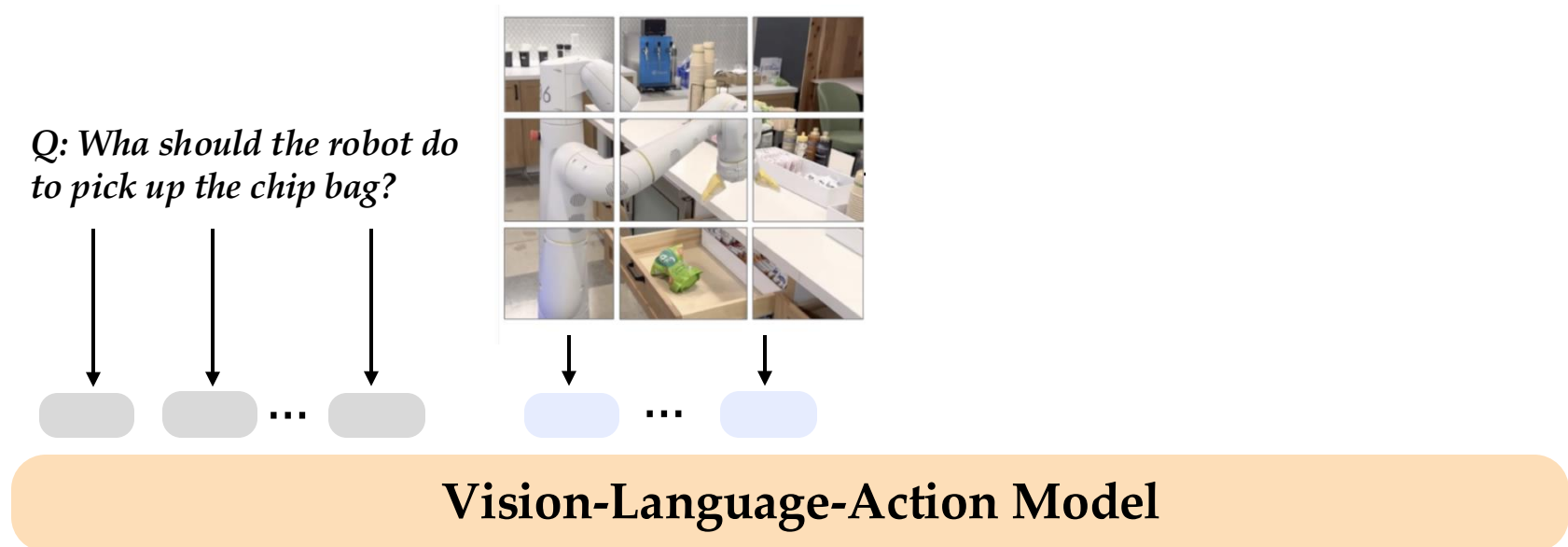


Large Language Model

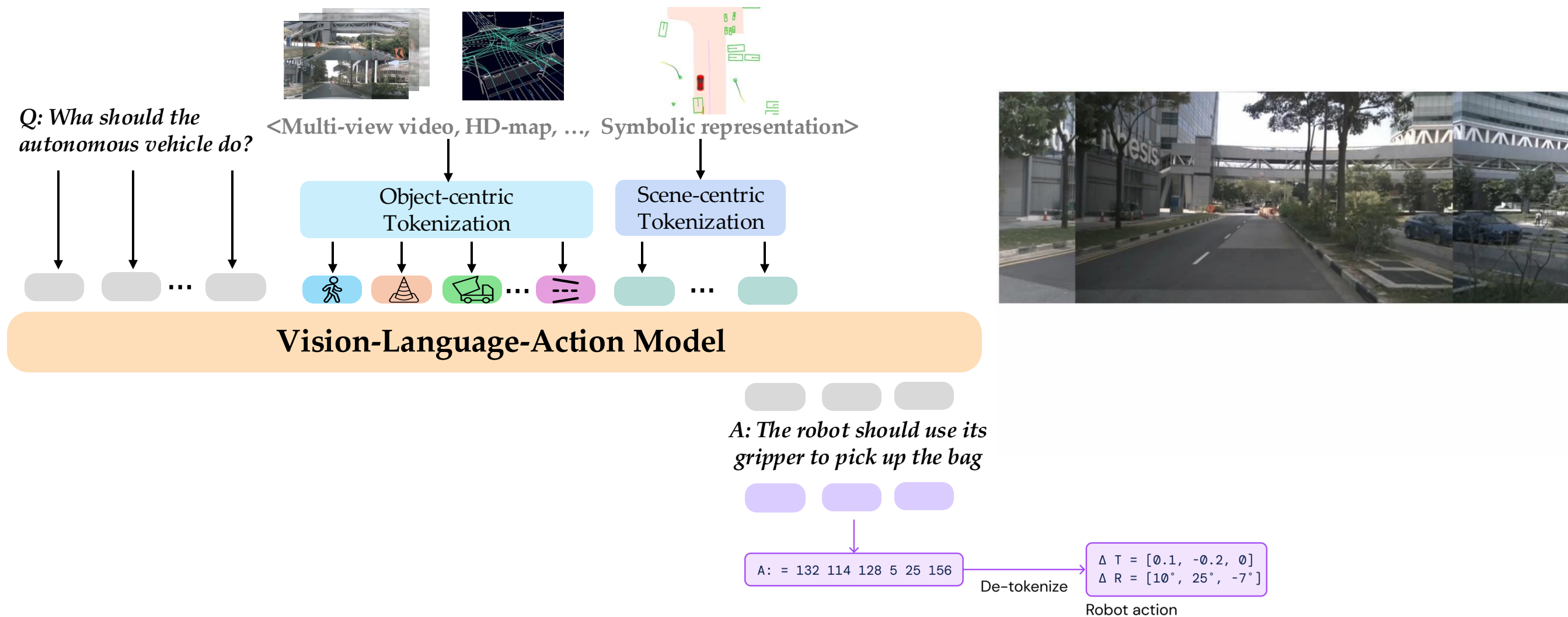


*A: The robot should use its
gripper to pick up the bag*

Now – Vision-Language-Action Robotics Foundation Model

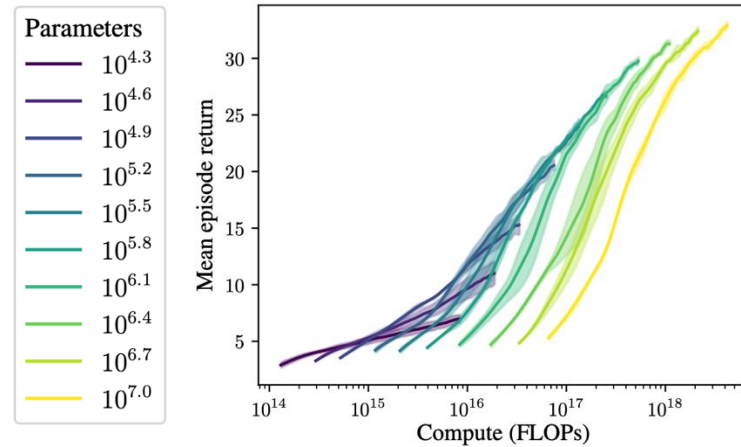


Now – Vision-Language-Action Robotics Foundation Model

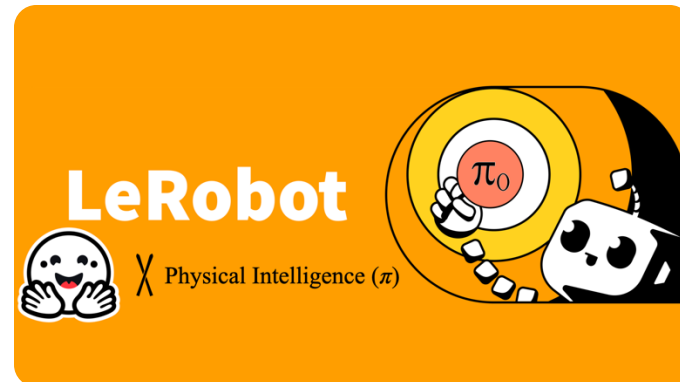
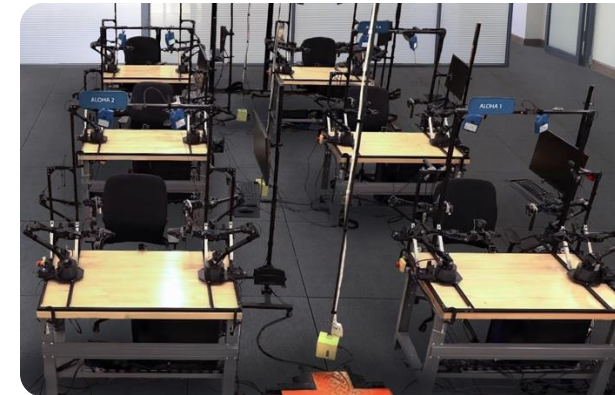


Shortest Path to Generalist Robots?

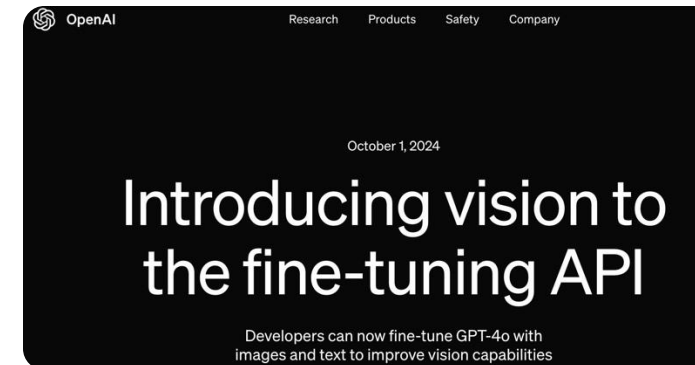
Scaling law seems to work in robotics



Commercialization of large-scale robot data collection



Open-sourced model / training pipeline



More affordable “one-click” fine-tuning API



Imitation is a “proxy” of the true training objective

Let's remember what the expert
did and copy them!

$$\max_{\theta} \mathbb{E} [\mathbb{P}_{\theta}(\mathbf{a}_{0:T}^* | \mathbf{o}_0; \text{context})]$$



\neq

Safety > comfort, progress, etc

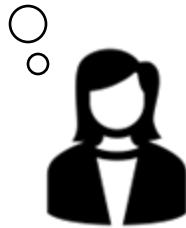
Miss-Alignment

By optimizing an **incomplete** or **mis-specified** objective,
these models lead to **undesirable** behaviors at best
and **safety hazards** at worst!

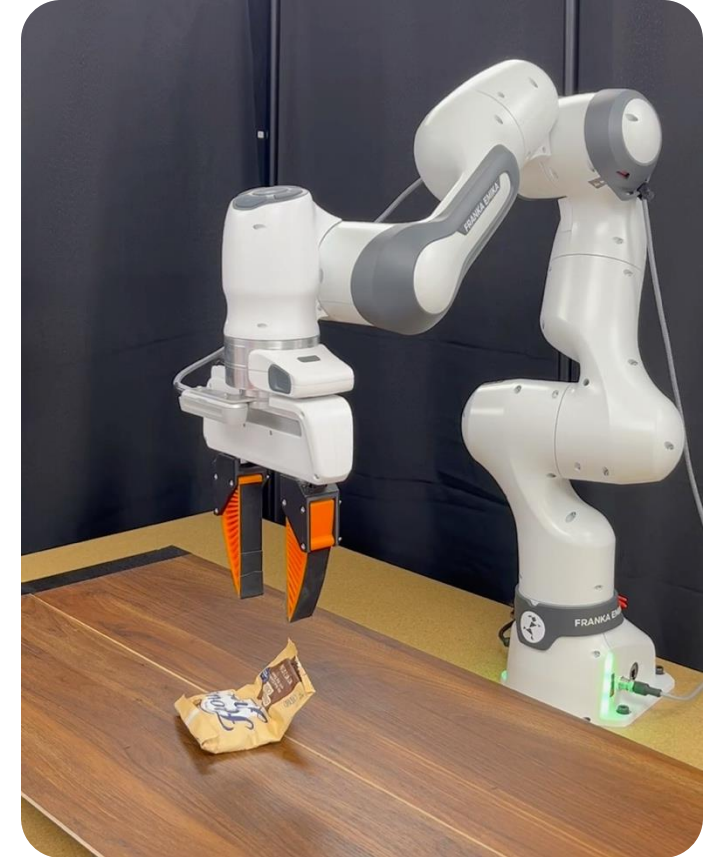


Hand me a bag of chips

*I don't want my chips
crashed...*



$$\max_{\theta} \mathbb{E} [\mathbb{P}_{\theta}(\mathbf{a}_{0:T}^* | \mathbf{o}_0; \text{context})]$$



Reinforcement Learning from Human Feedback for Post-training Preference Alignment

Step 2: Sample generations from the model

Step 5: Train the model to max. the reward

Model-based principles come back!

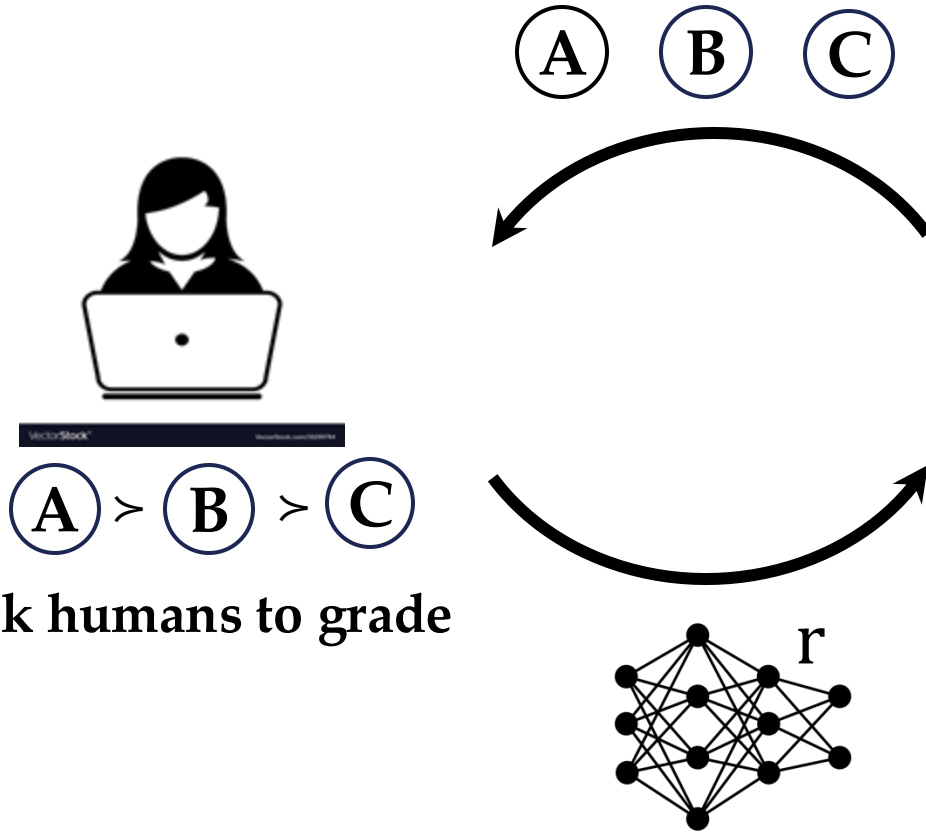
max r

Generative Model

Step 1: Train a generative model via BC

Step 4: Learn a reward function

Step 3: Ask humans to grade



Reinforcement Learning from Human Feedback for Post-training Preference Alignment

Predominant alignment mechanism in *non-embodied* domains

Text-image gen. : A cake in the city.



Pre-trained



After alignment

Text-video gen. : An animated sneaker is playing basketball.



Pre-trained



After alignment

RLHF has yet to achieve the same impact in aligning
robotics motion generation models

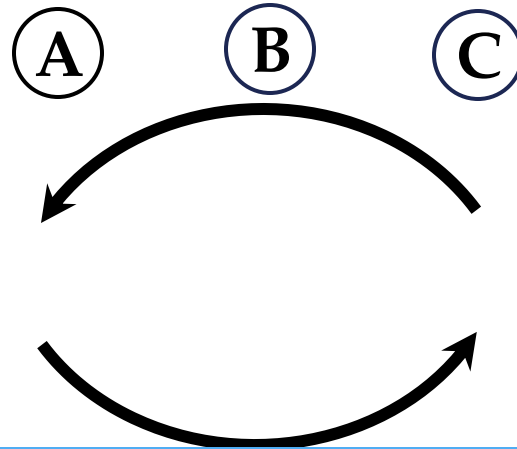
Reinforcement Learning from Human Feedback for Post-training Preference Alignment

Step 2: Sample generations from
the model



$\textcircled{A} > \textcircled{B} > \textcircled{C}$

Step 3: Ask humans to grade

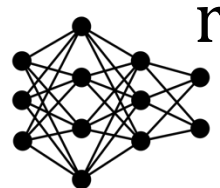


Step 5: Train the model to max. the reward

$\max r_R$

Motion Generation
Model

Step 1: Train a generative model via BC



Step 4: Learn a reward function

Reinforcement Learning from Human Feedback for Post-training Preference Alignment

Step 2: Sample generations from
the model



(A) (B) (C)

Step 5: Train the model to max. the reward

$$\max r_R(\text{image}, \text{task})$$

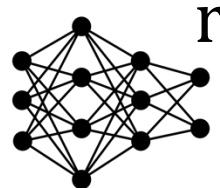
Motion Generation
Model

Step 1: Train a generative model via BC



(A) > (B) > (C)

Step 3: Ask humans to grade



Step 4: Learn a reward function

Learning a high-quality visual reward function requires an
impractically large amount of human preference feedback

Reinforcement Learning from Human Feedback for Post-training Preference Alignment

Step 2: Sample generations from the model



A B C



A > B > C

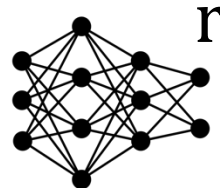
Step 3: Ask humans to grade

Step 5: Train the model to max. the reward

$$\max r_R(\succ, \text{task})$$

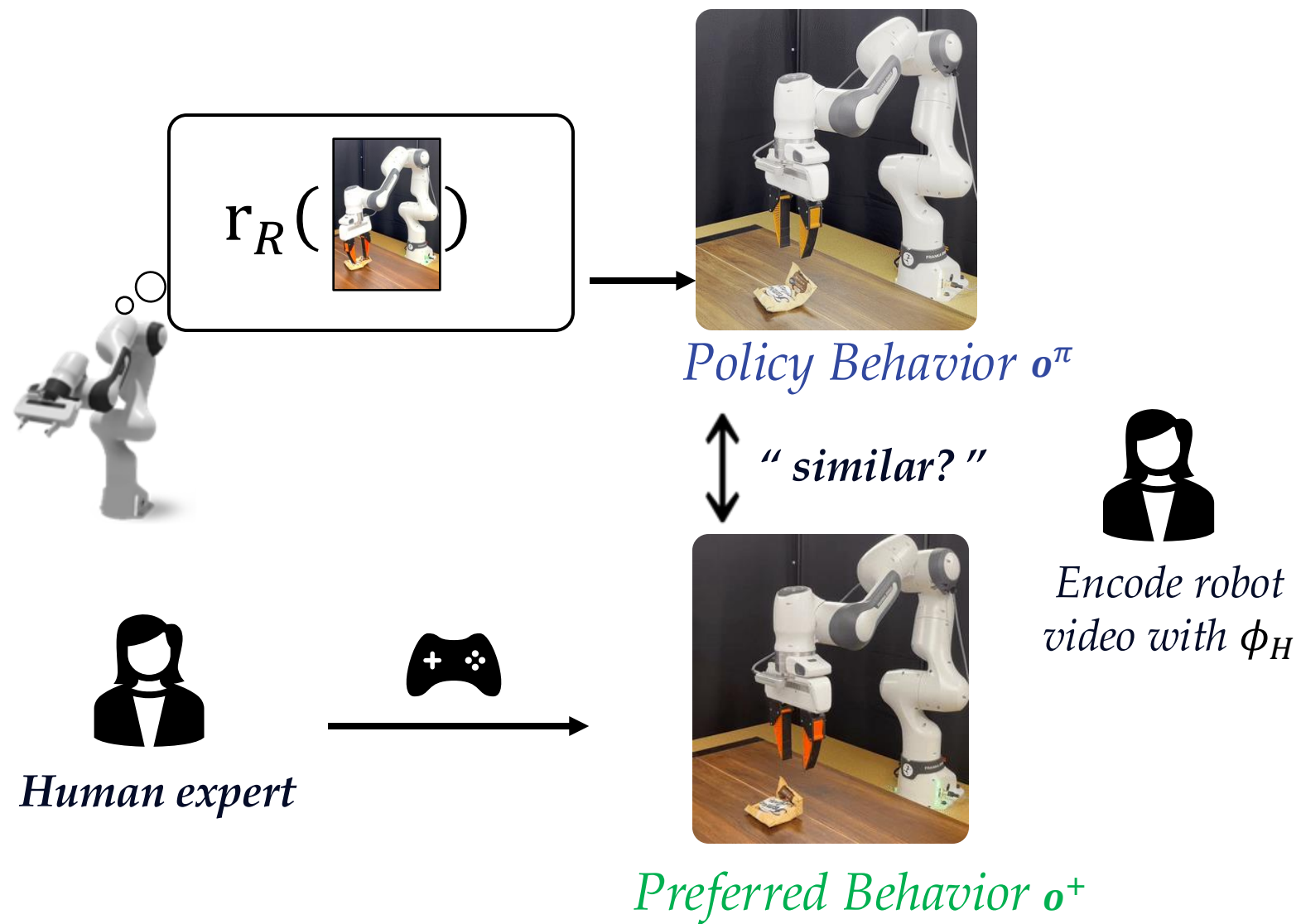
Motion Generation
Model

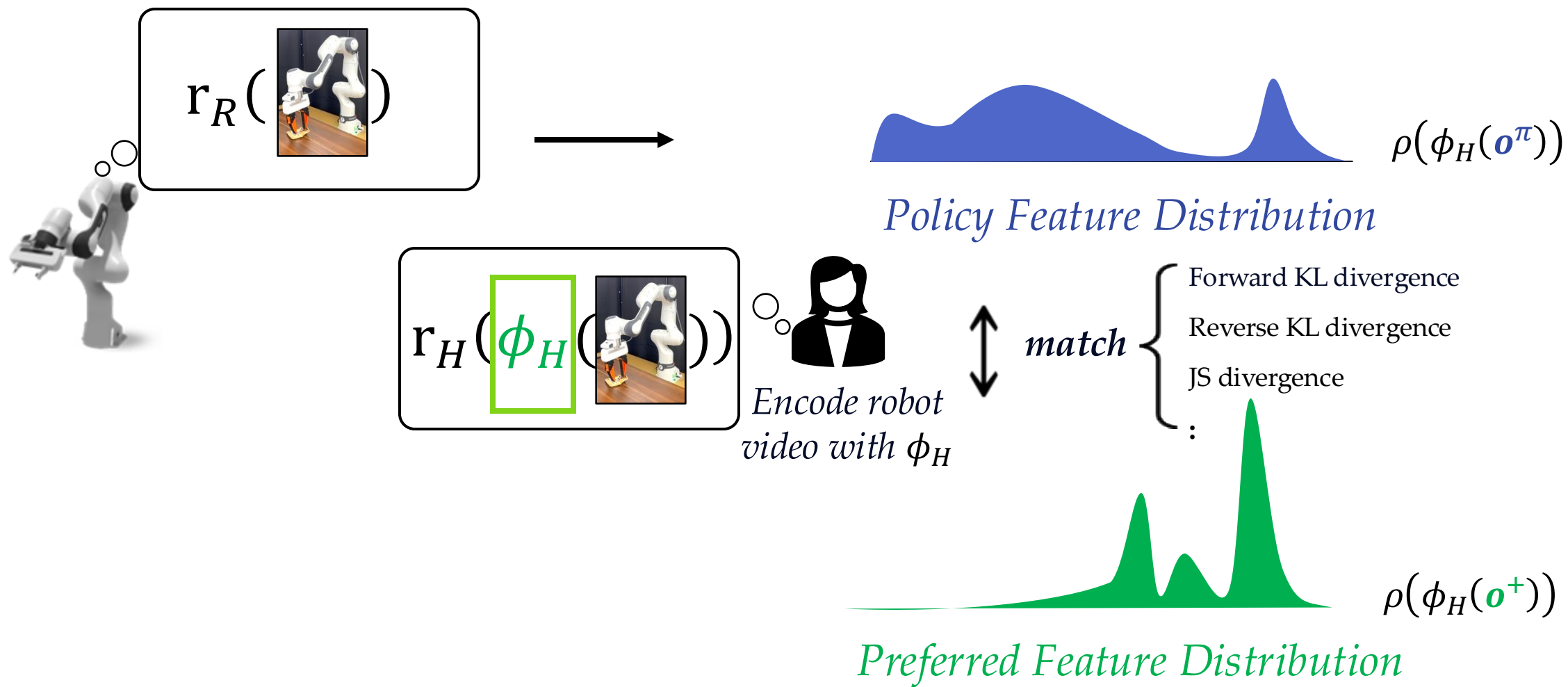
Step 1: Train a generative model via BC

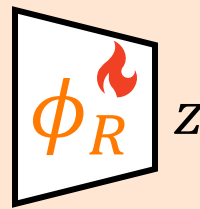
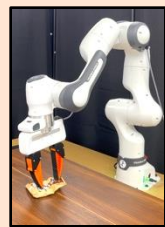


Step 4: Learn a reward function

Goal: Maximizing Alignment with Minimal Feedback!







z

$\underset{\phi_R}{\operatorname{argmin}} \operatorname{diff}(\phi_R, \phi_H)$

Visual Representation Alignment

$r_R(\phi_R(\text{robot image}))$



$\rho(\phi_H(\mathbf{o}^\pi))$

Policy Feature Distribution

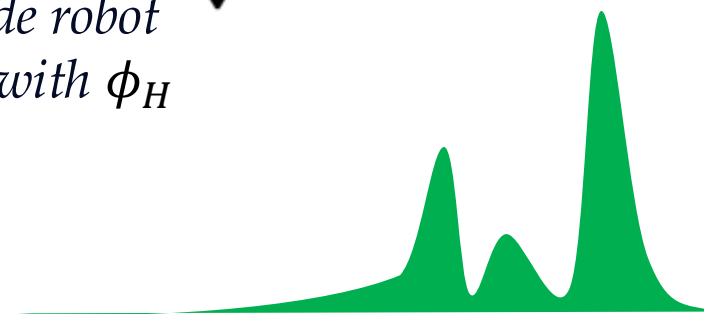
$r_H(\phi_H(\text{robot image}))$



Encode robot
video with ϕ_H

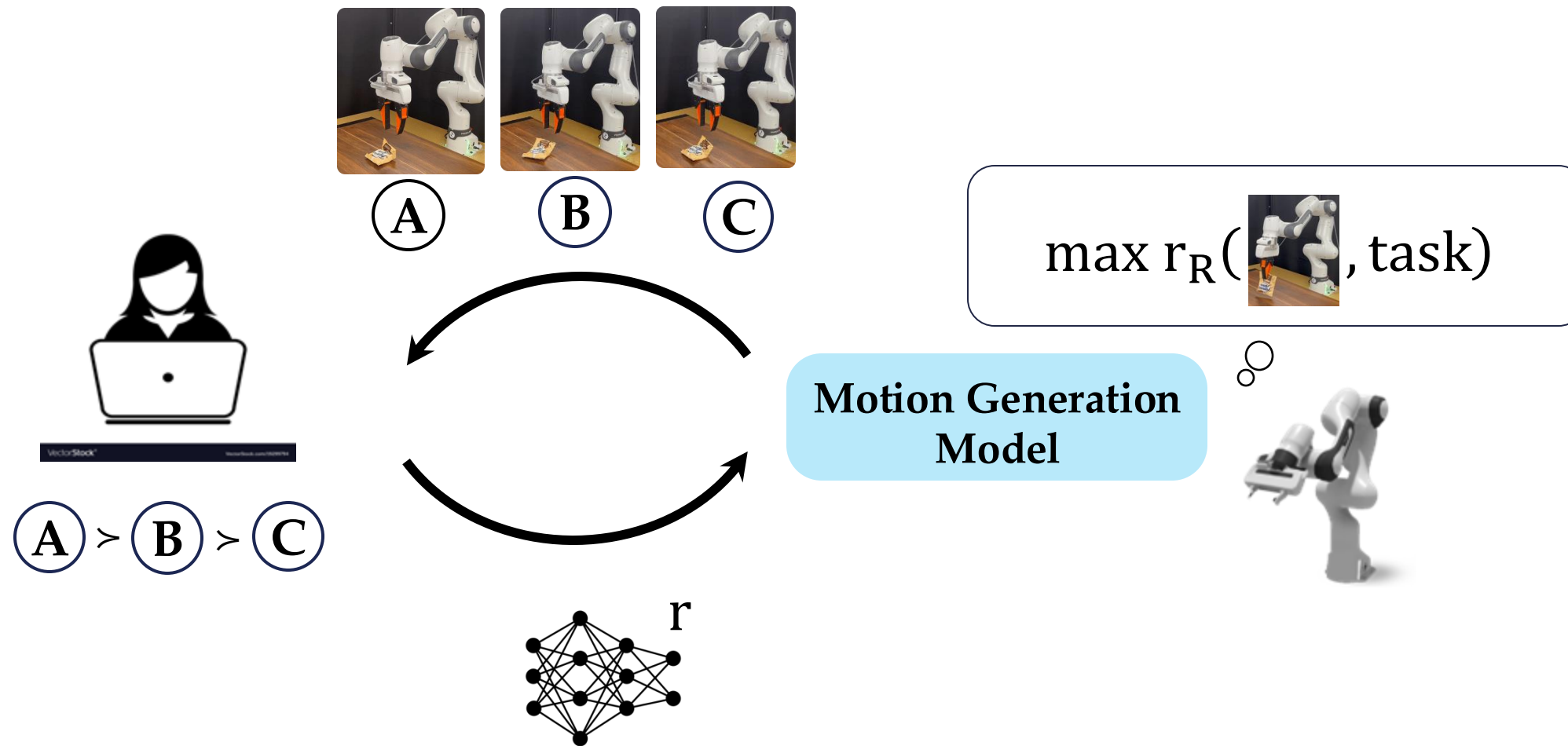


match

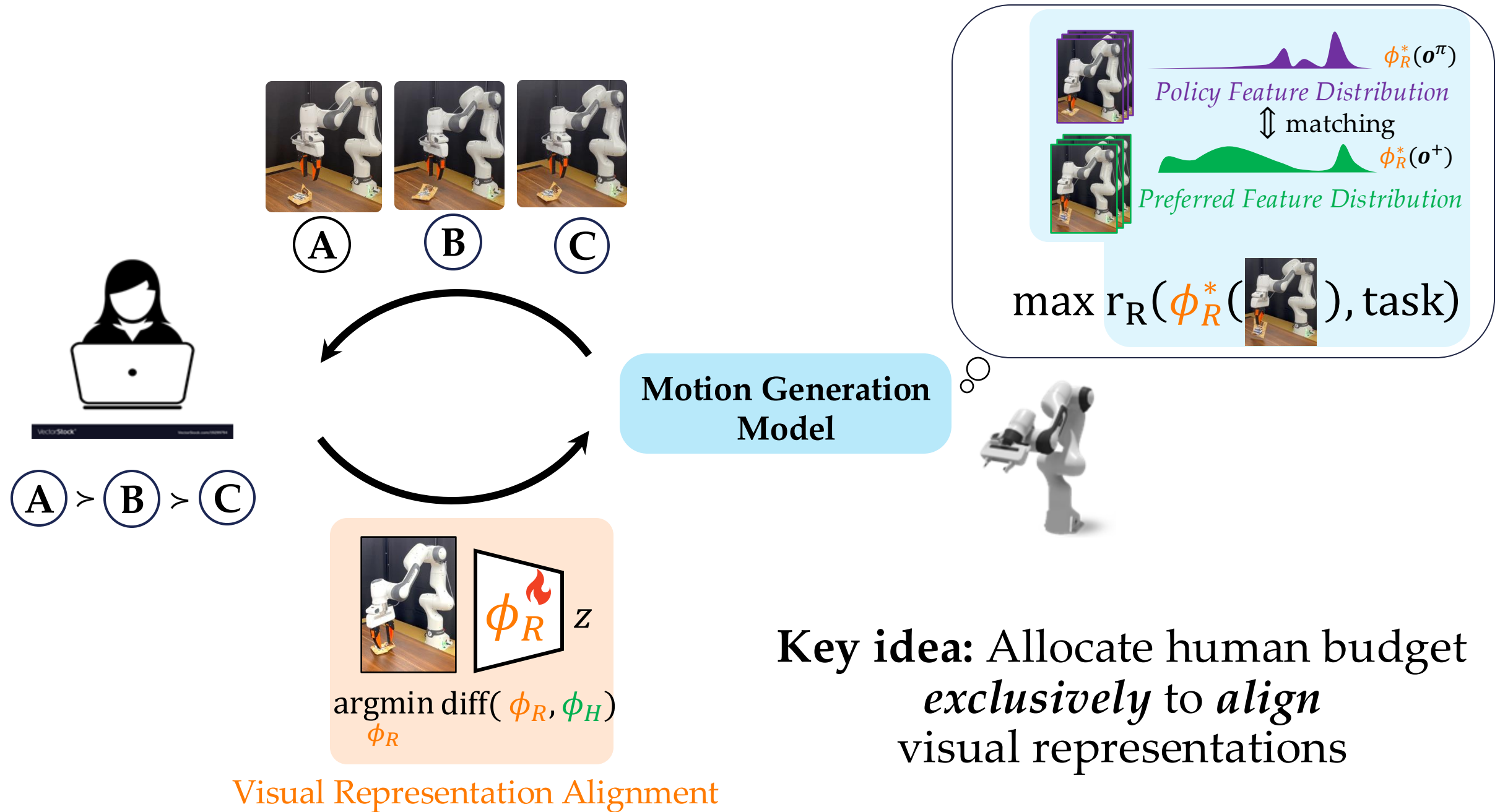


$\rho(\phi_H(\mathbf{o}^+))$

Preferred Feature Distribution

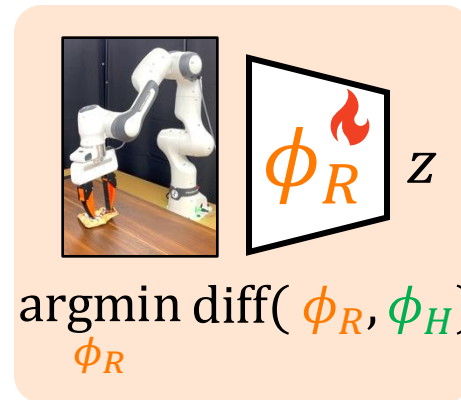


Learning reward end-to-end



Formalizing Visual Representation Alignment

How to formally describe and compare two encoders?



Formalizing Visual Representation Alignment

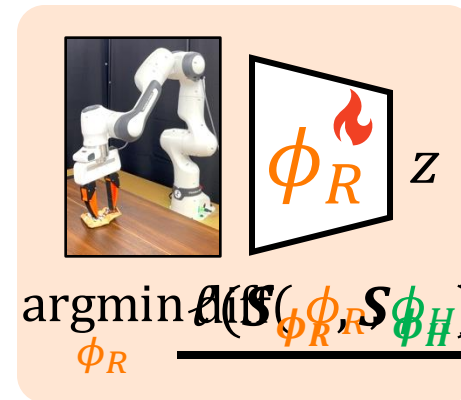
How to formally describe and compare two encoders?

Triplet-based representation space [Sucholutsky & Griffiths, NeurIPS 2023]

$$\phi \longrightarrow S_\phi := \left\{ \left(\begin{array}{ccc} \text{img}_i & \text{img}_j & \text{img}_k \\ \mathbf{o}^i & \mathbf{o}^j & \mathbf{o}^k \\ \text{anchor} & \text{positive} & \text{negative} \end{array} \right) : d(\phi(\mathbf{o}^i), \phi(\mathbf{o}^j)) < d(\phi(\mathbf{o}^i), \phi(\mathbf{o}^k)), \mathbf{o}^{i,j,k} \in \mathbb{E} \right\}$$

\mathbf{o}^i is closer to \mathbf{o}^j than to \mathbf{o}^k

learning ϕ_R which minimizes the difference between two agents' representation spaces (S)



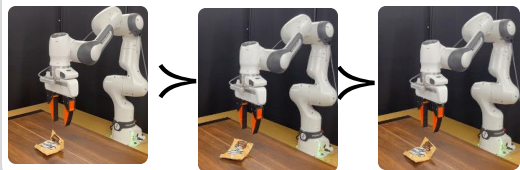
- 1) Don't have direct access to ϕ_H
- 2) S_{ϕ_H} is extremely large in space of videos

$$\min_{\phi_R} \ell(S_{\phi_R}, S_{\phi_H})$$

$$\min_{\phi_R} \ell(S_{\phi_R}, \tilde{S}_{\phi_H})$$

Query the end user

Preference Feedback



*I don't want
my chips
crushed!*



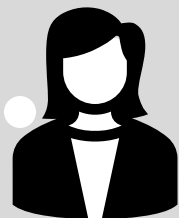
End-User

$$\min_{\phi_R} \ell(S_{\phi_R}, \tilde{S}_{\phi_H})$$

Preference Feedback



*I don't want
my chips
crushed!*



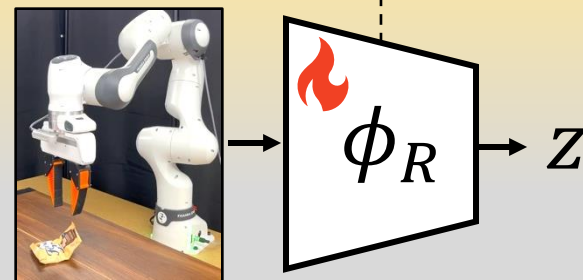
End-User

\tilde{S}_{ϕ_H}

Visual Representation Alignment

$$\max_{\phi_R} \mathbb{P}(\text{better} \mid \text{worse} \mid \text{most preferred} \mid \phi_R)$$

$\mathbf{o}^j \quad \mathbf{o}^k \quad \mathbf{o}^i$



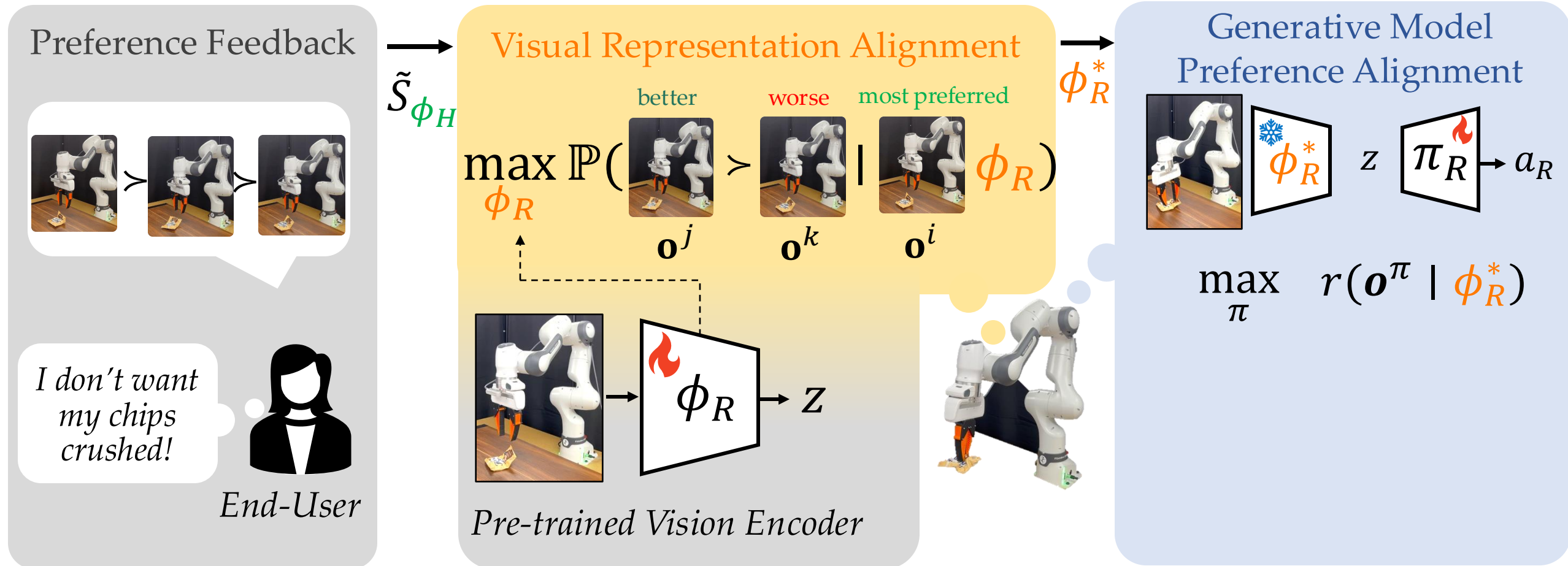
Pre-trained Vision Encoder

Bradley-Terry model

$$\frac{e^{-d(\phi_R(\mathbf{o}^i), \phi_R(\mathbf{o}^j))}}{e^{-d(\phi_R(\mathbf{o}^i), \phi_R(\mathbf{o}^j))} + e^{-d(\phi_R(\mathbf{o}^i), \phi_R(\mathbf{o}^k))}}$$

Two equally **preferred** behaviors should have **similar** feature representations.



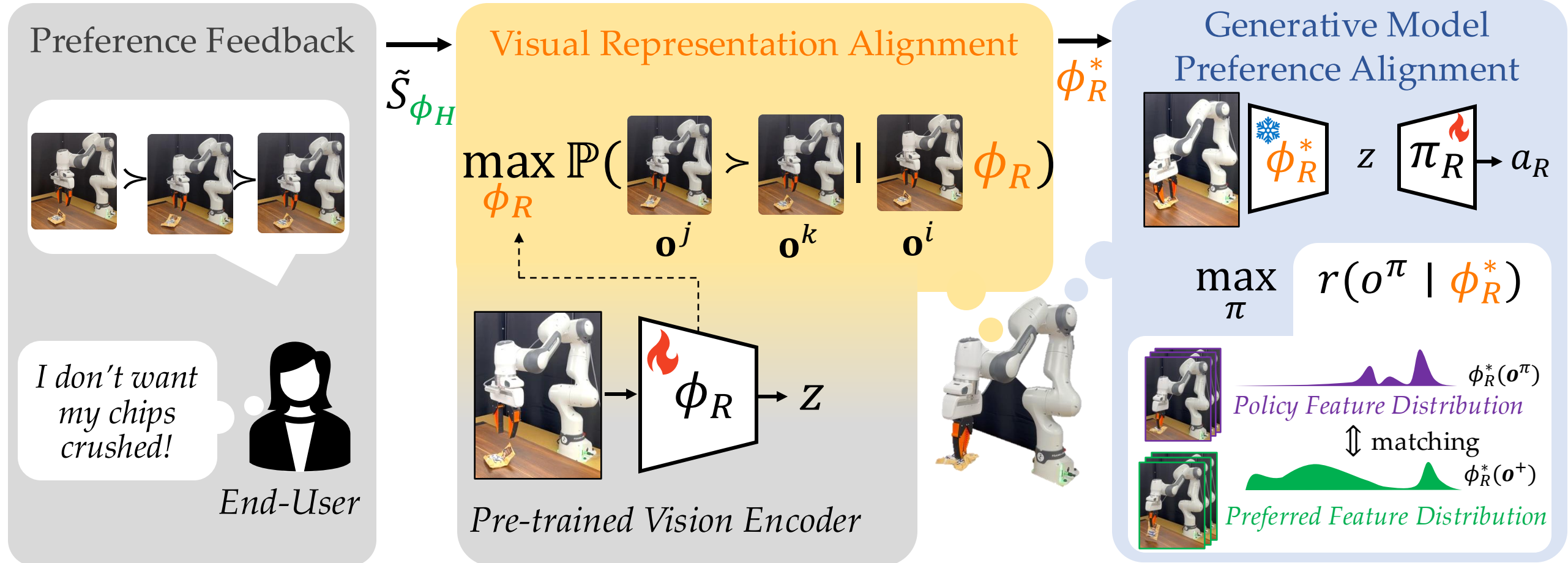


Representation-Aligned Preference-Based Learning (RAPL)

Action-free
& video only

Data efficient reward learning

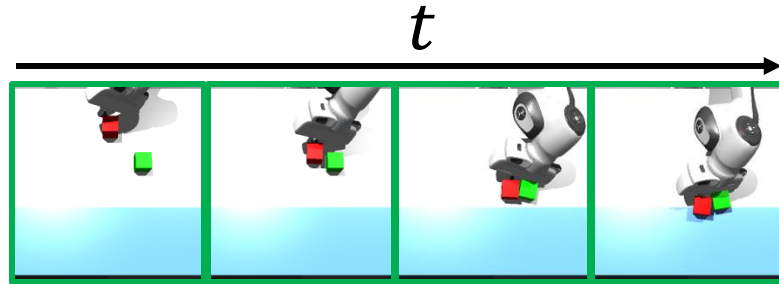
Effective
preference alignment



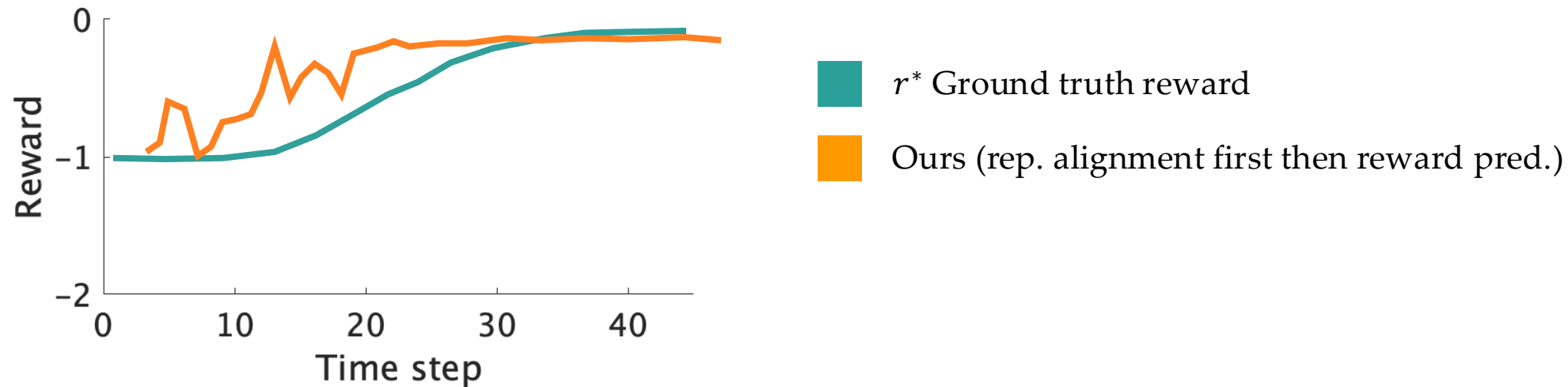
On the Value of Aligned Representation in Visual Reward Learning



$r^*(\phi_H(o))$: push objects move efficiently to goal region



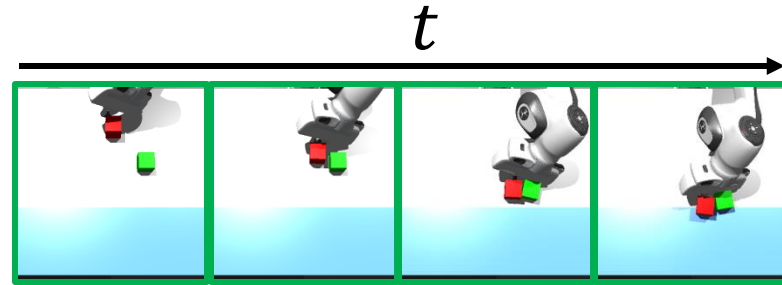
Reward of *Good* Behavior



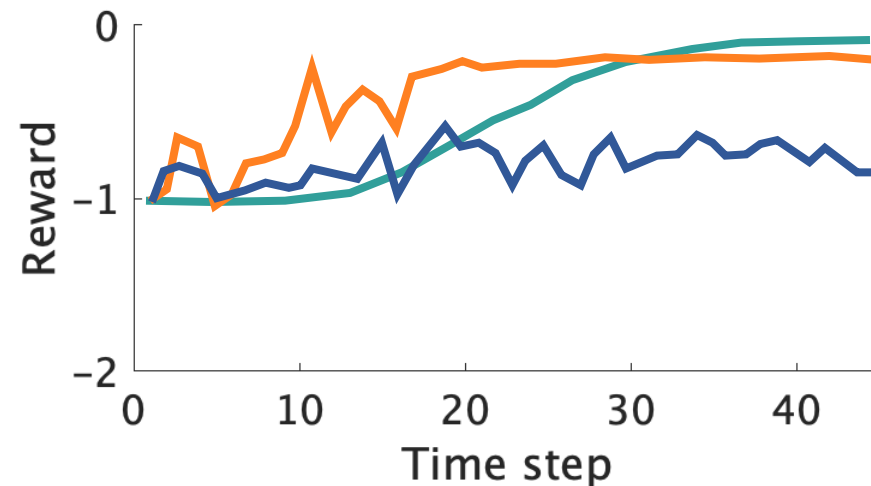
On the Value of Aligned Representation in Visual Reward Learning



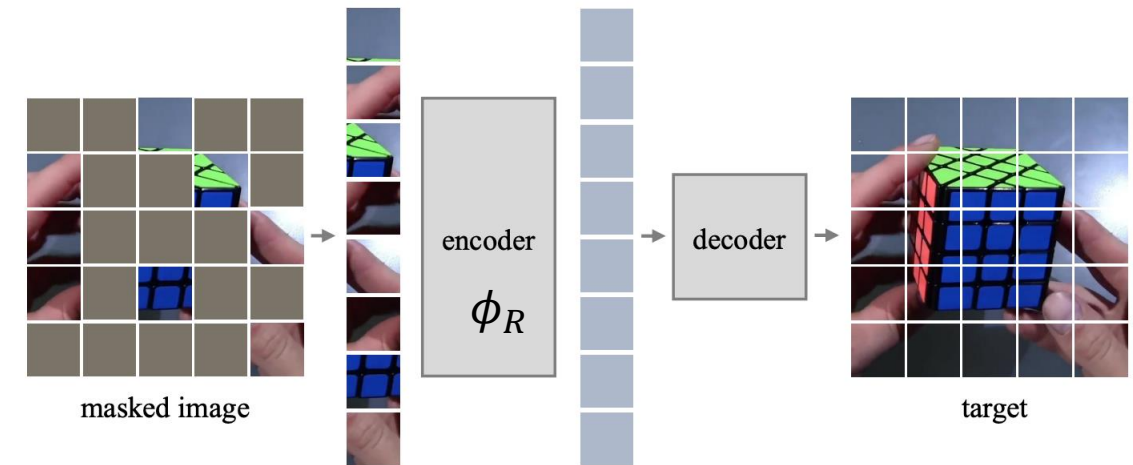
$r^*(\phi_H(\mathbf{o}))$: push objects move efficiently to goal region



Reward of *Good* Behavior



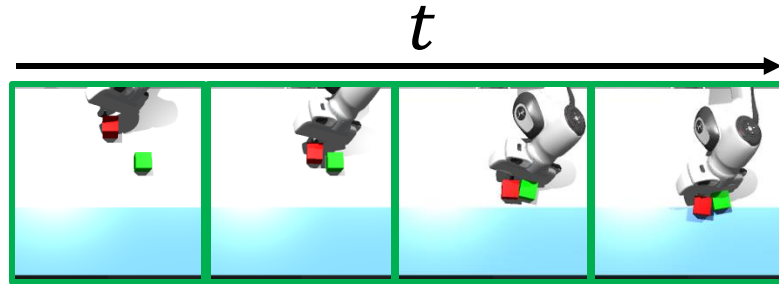
- r^* Ground truth reward
- Ours (rep. alignment first then reward pred.)
- MVP representation [1]



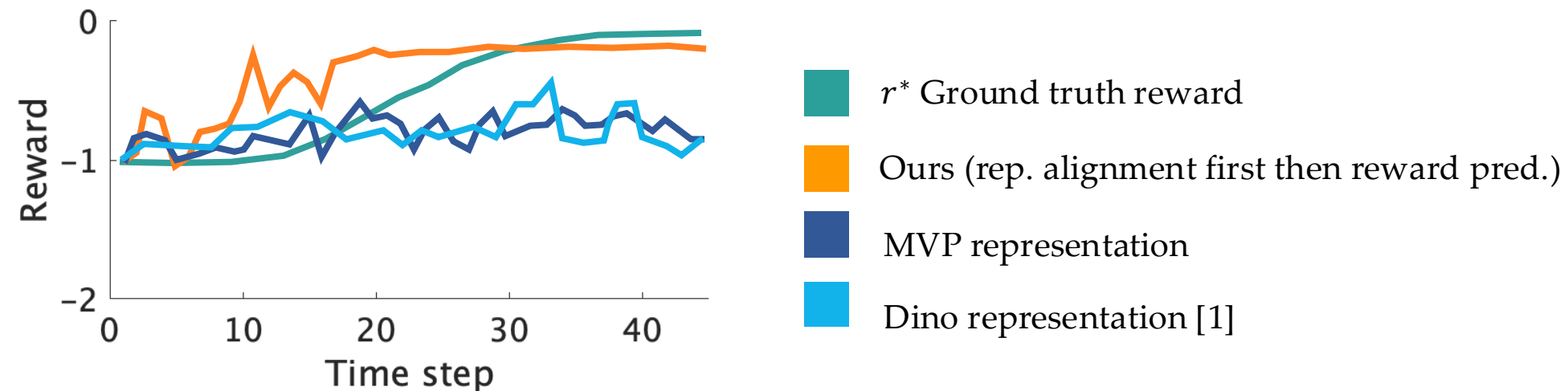
On the Value of Aligned Representation in Visual Reward Learning



$r^*(\phi_H(\mathbf{o}))$: push objects move efficiently to goal region



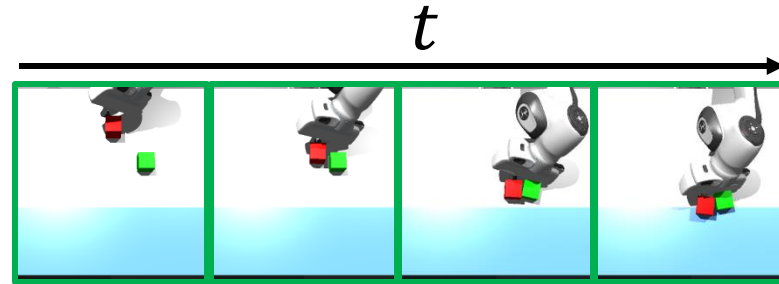
Reward of *Good* Behavior



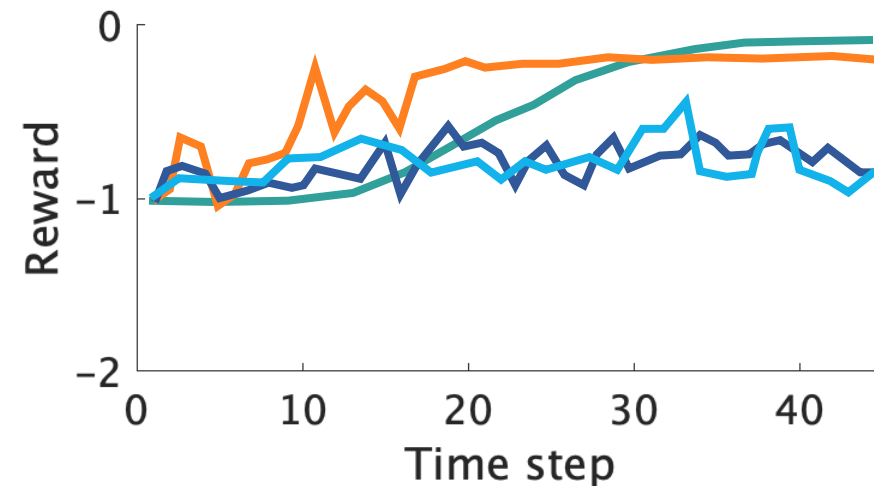
On the Value of Aligned Representation in Visual Reward Learning



$r^*(\phi_H(o))$: push objects move efficiently to goal region



Reward of *Good* Behavior



**DINO-WM: World Models on Pre-trained Visual Features
enable Zero-shot Planning**

Gaoyue Zhou¹ Hengkai Pan¹ Yann LeCun^{1,2} Lerrel Pinto¹

**Real-World Robot Learning with
Masked Visual Pre-training**

Ilija Radosavovic* Tete Xiao* Stephen James Pieter Abbeel Jitendra Malik[†] Trevor Darrell[†]

University of California, Berkeley

Abstract: In this work, we explore self-supervised visual pre-training on images from diverse, in-the-wild videos for real-world robotic tasks. Like prior work, our visual representations are pre-trained via a masked autoencoder (MAE), frozen, and then passed into a learnable control module. Unlike prior work, we show that the pre-trained representations are effective across a range of real-world robotic tasks and embodiments. We find that our encoder consistently outperforms CLIP (up to 75%), supervised ImageNet pre-training (up to 81%), and training from scratch (up to 81%). Finally, we train a 307M parameter vision transformer on a massive collection of 4.5M images from the Internet and egocentric videos, and demonstrate clearly the benefits of scaling visual pre-training for robot learning.

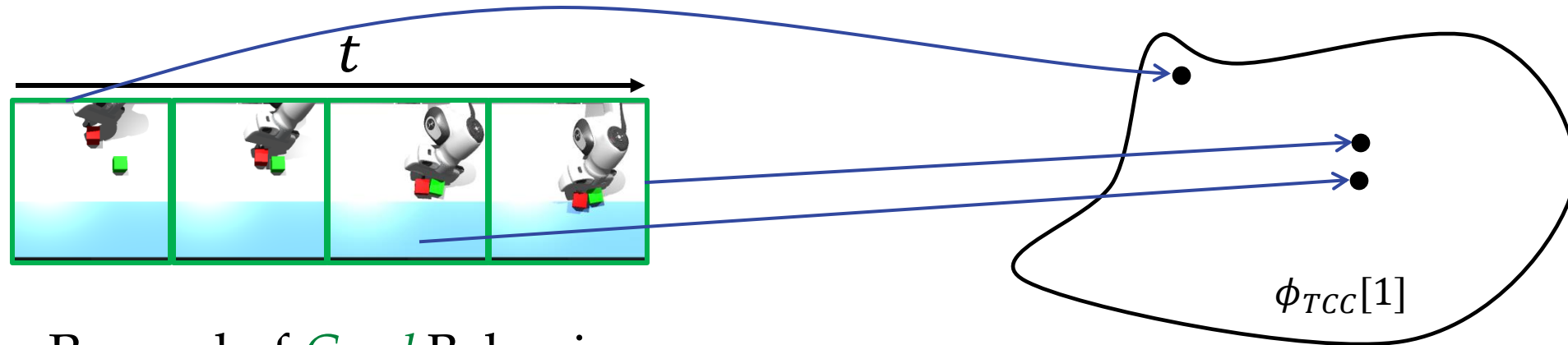
Keywords: Self-supervised Learning, Visual Representations, Robot Learning

Serve as **latent state**
representation for planning

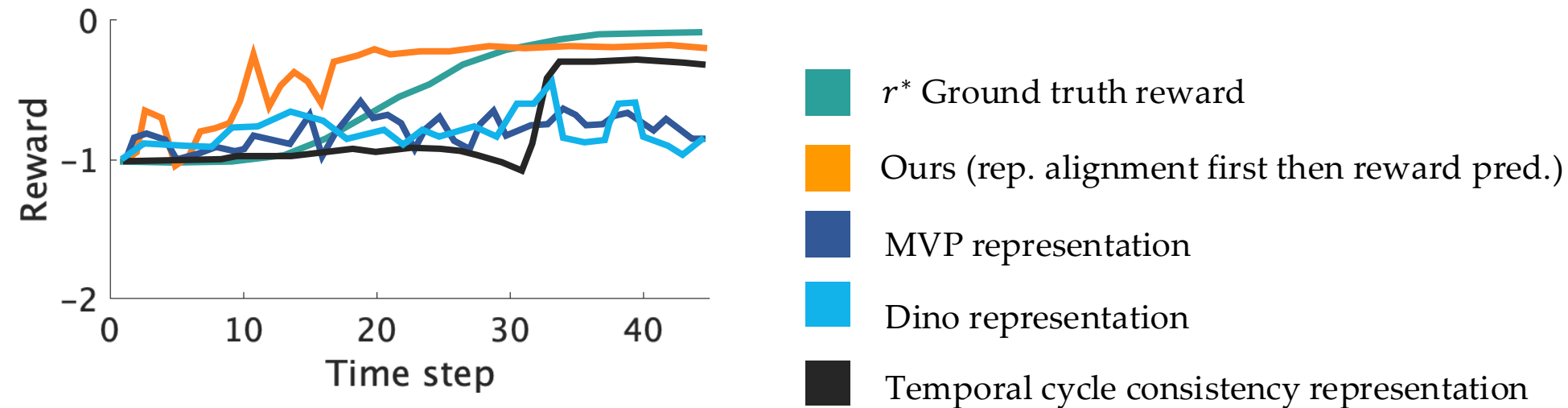
On the Value of Aligned Representation in Visual Reward Learning



$r^*(\phi_H(o))$: push objects move efficiently to goal region



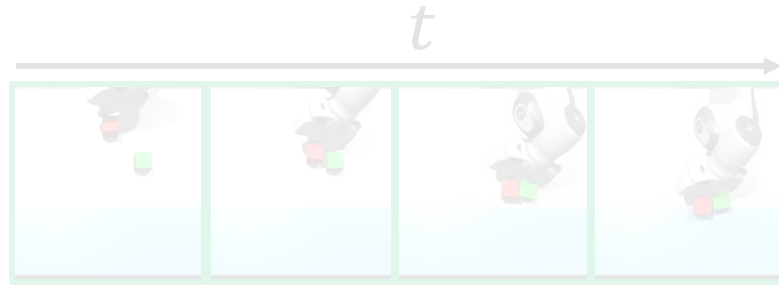
Reward of *Good* Behavior



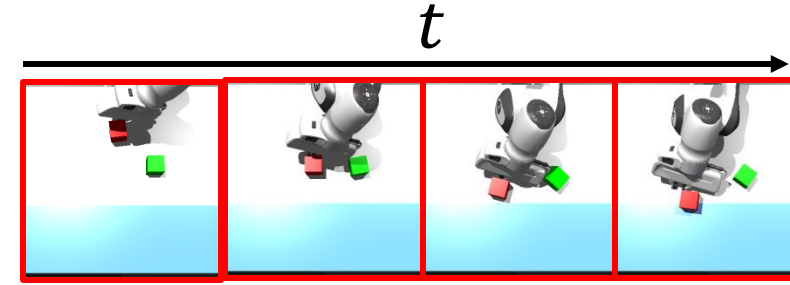
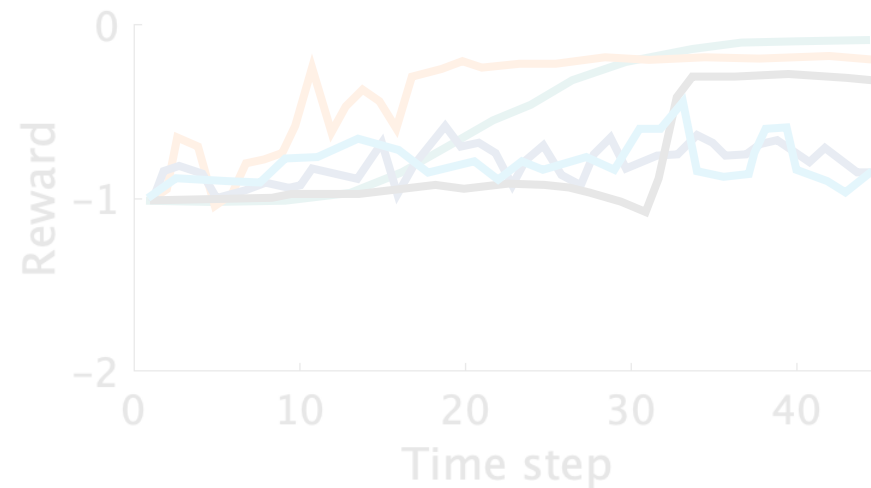
On the Value of Aligned Representation in Visual Reward Learning



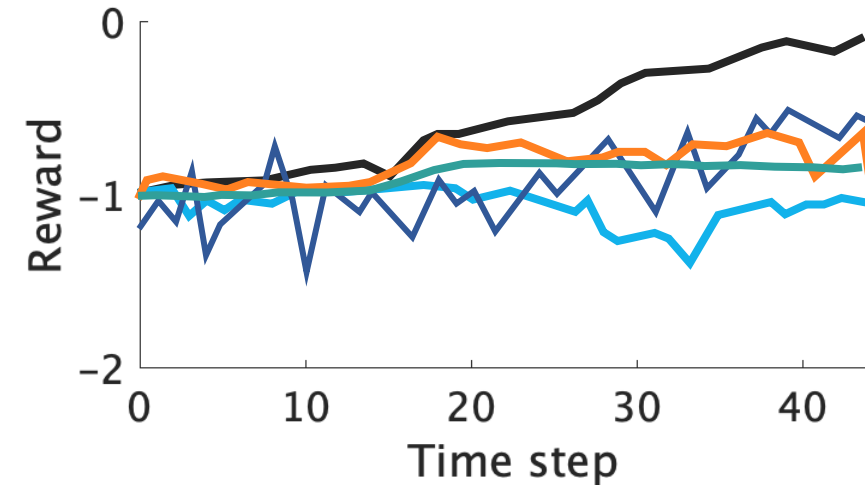
$r^*(\phi_H(\mathbf{o}))$: push objects move efficiently to goal region



Reward of *Good* Behavior



Reward of *Disliked* Behavior

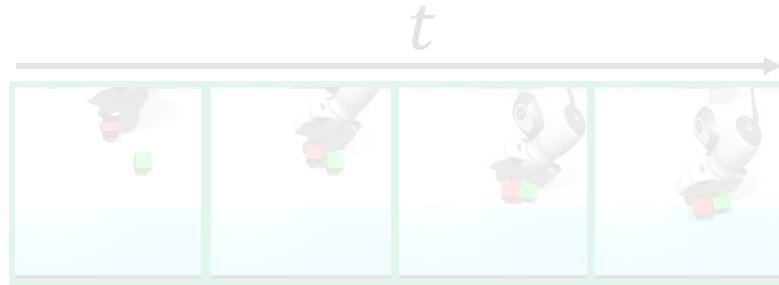


GT Ours MVP TCC Dino

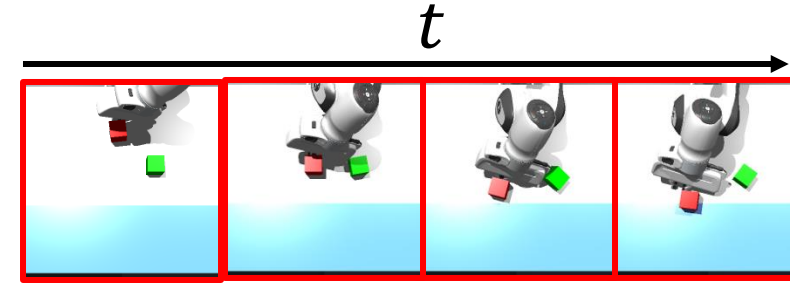
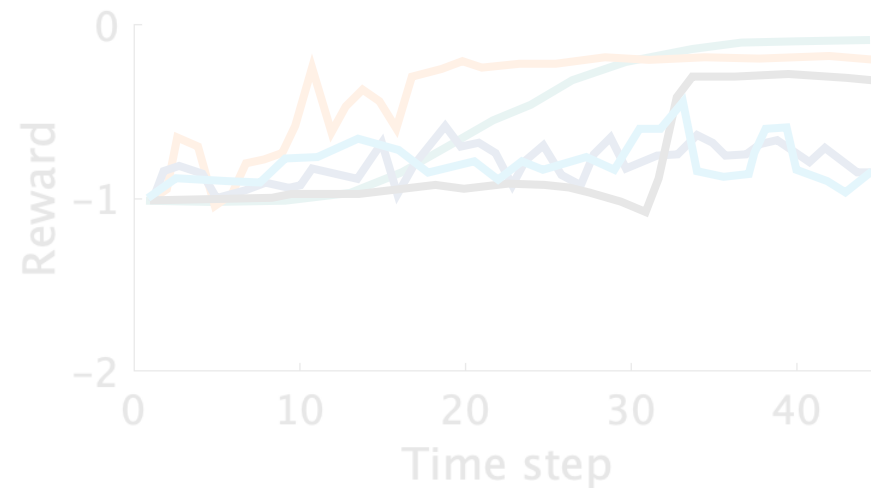
On the Value of Aligned Representation in Visual Reward Learning



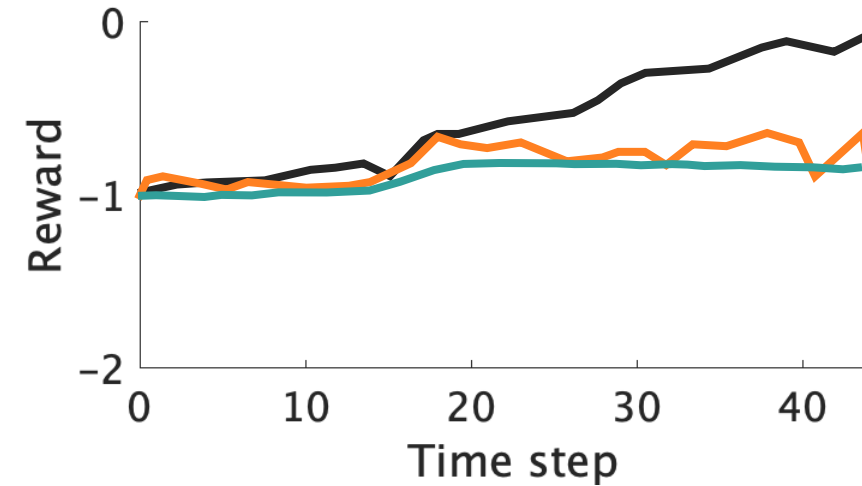
$r^*(\phi_H(\mathbf{o}))$: push objects move efficiently to goal region



Reward of *Good* Behavior

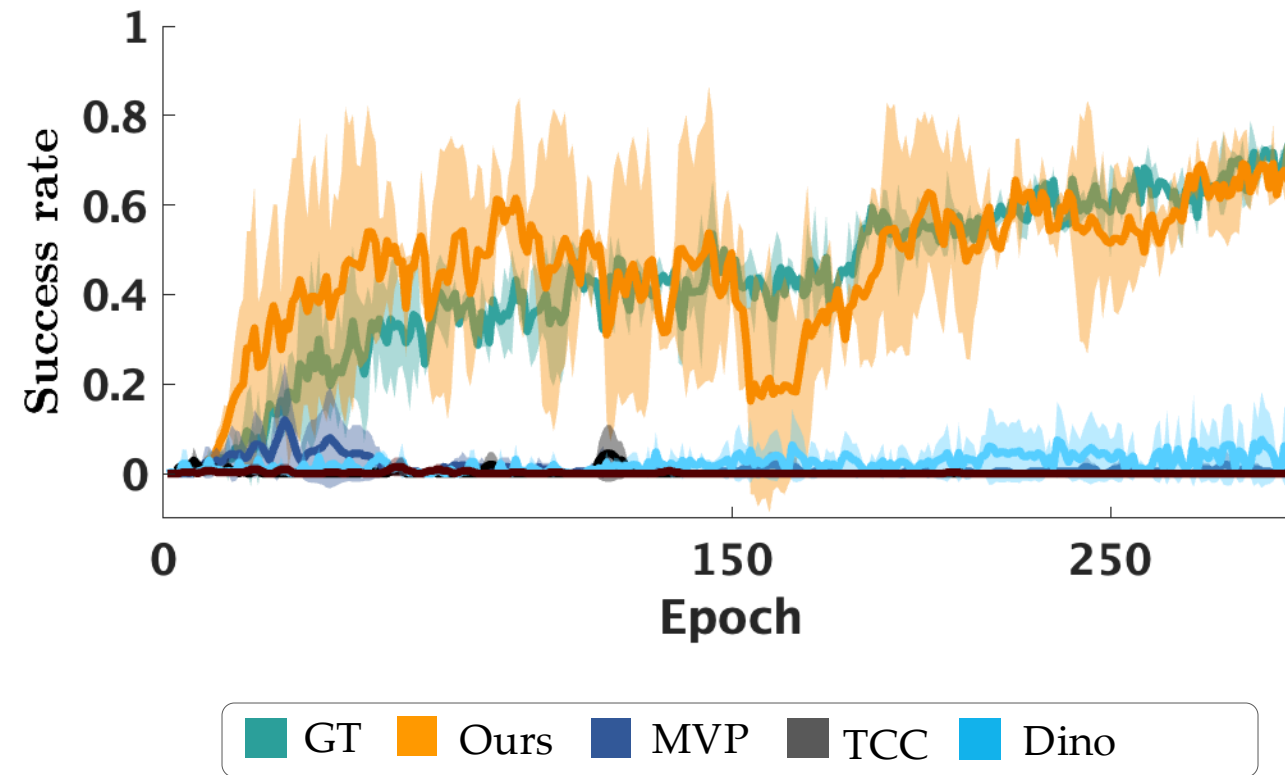


Reward of *Disliked* Behavior



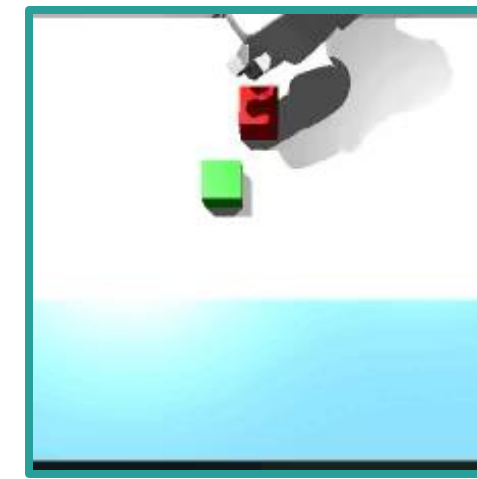
GT Ours MVP TCC Dino

On the Value of Aligned Representation in Visual Reward Learning

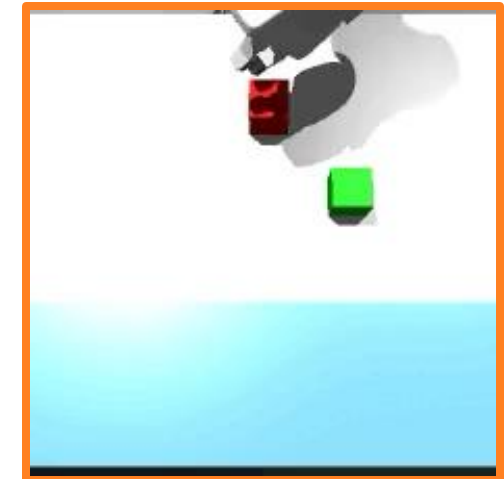


Insight

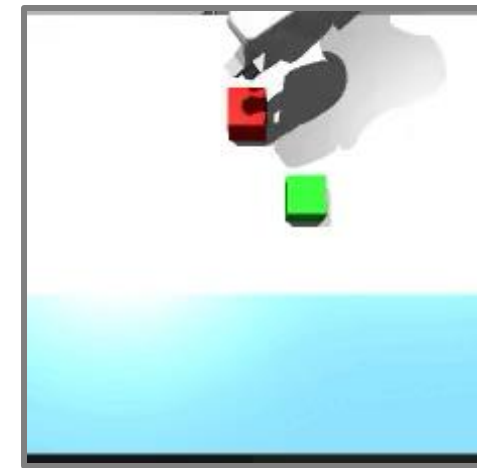
Pre-trained unaligned representations might miss important features that matter to the task



GT



RAPL

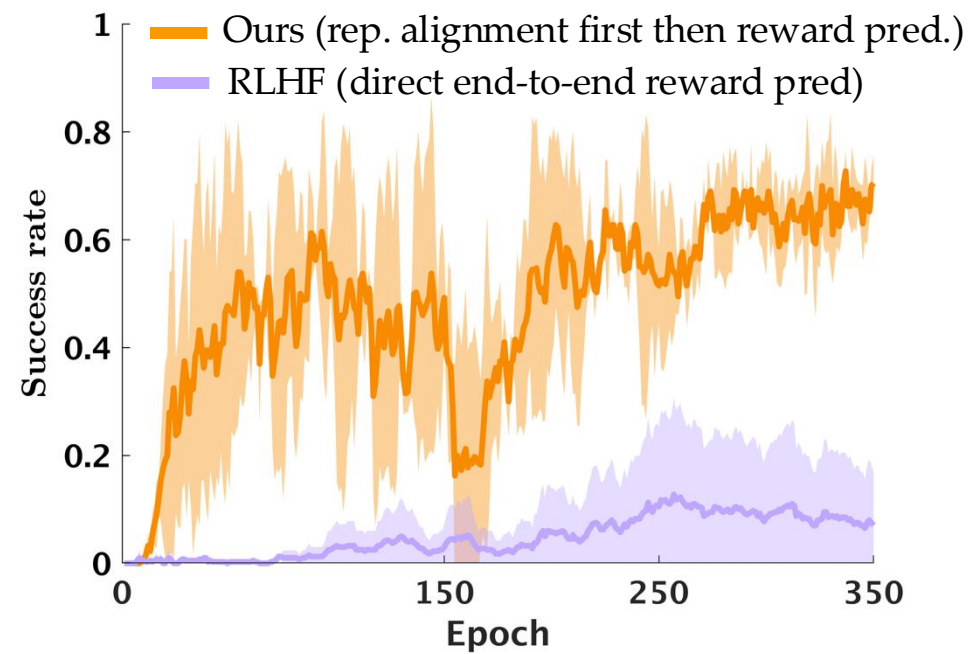


TCC

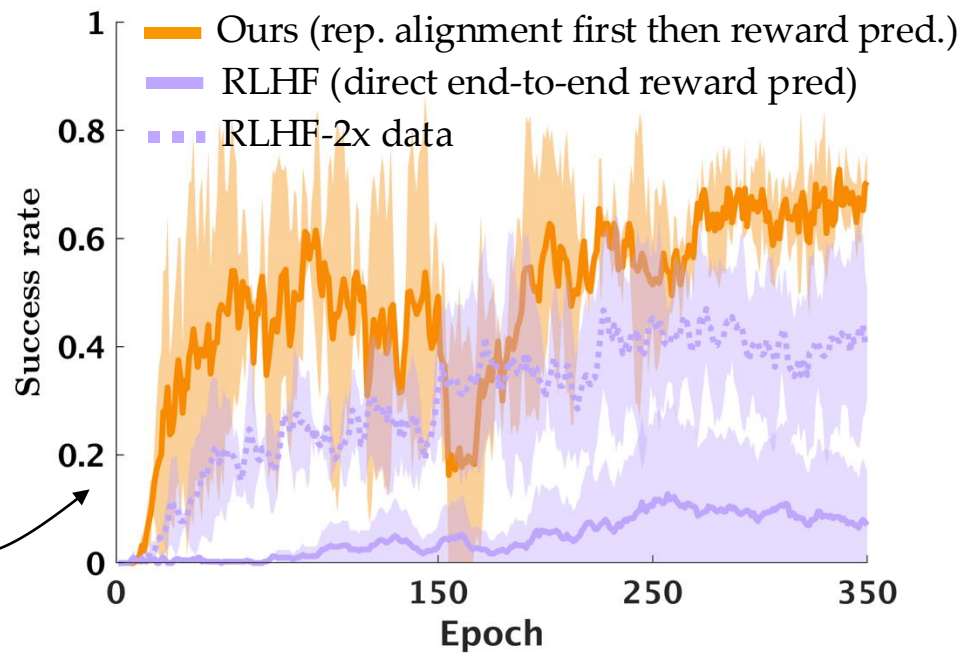


Dino

On the Value of Aligned Representation – *Sample Efficiency*

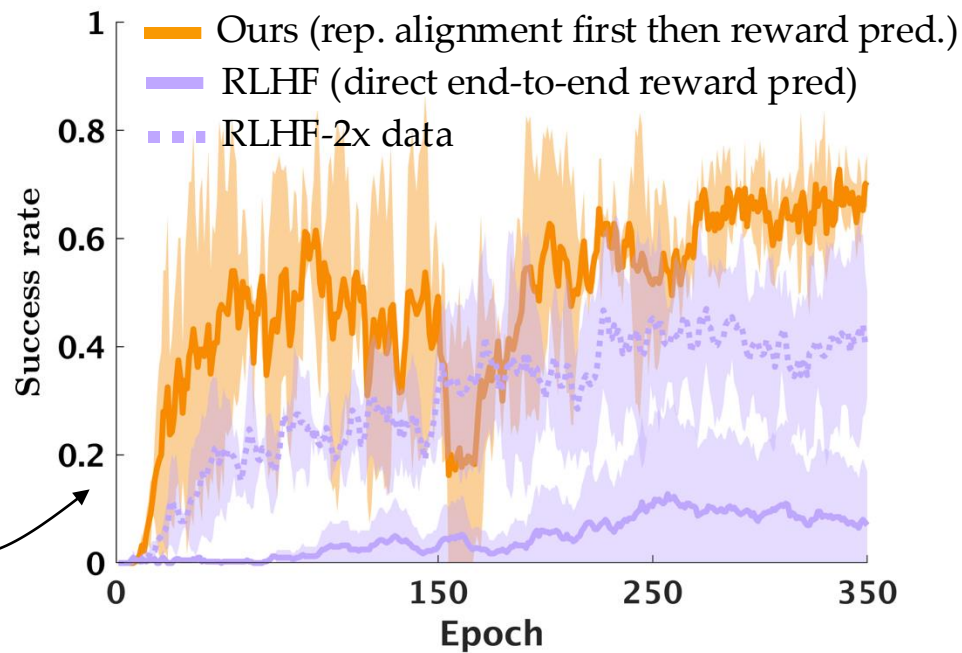


On the Value of Aligned Representation – *Sample Efficiency*

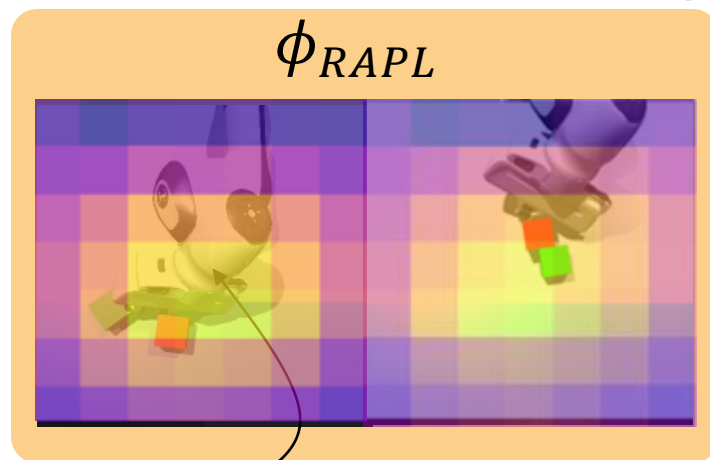


***RAPL** outperforms
RLHF by 75% with 50%
less preference data*

On the Value of Aligned Representation – *Sample Efficiency*



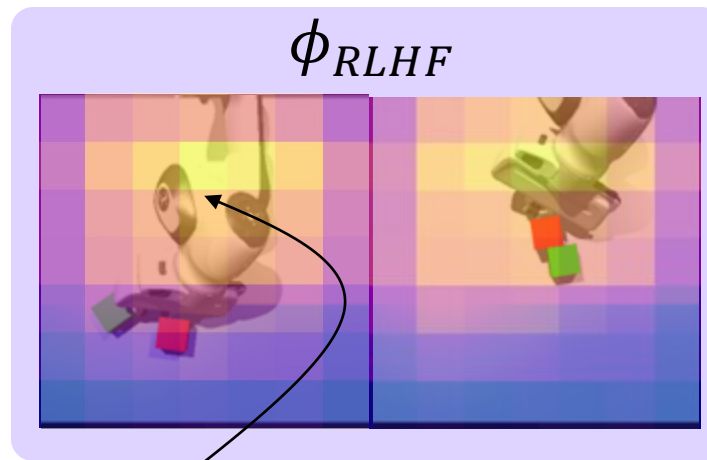
RAPL outperforms
RLHF by 75% with 50%
less preference data



Contribution to final embed.

high low

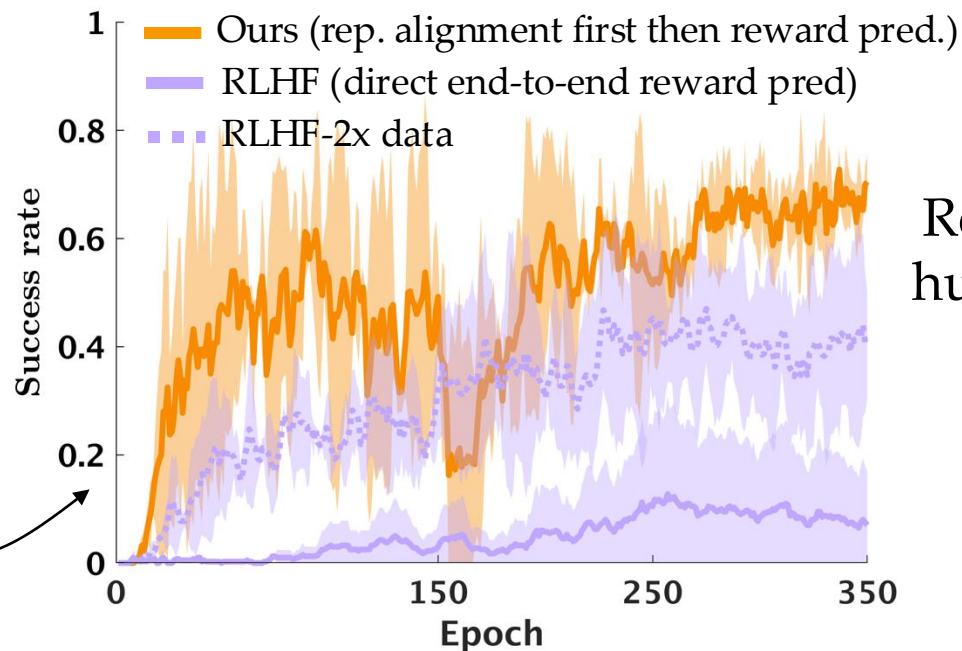
Attending to *task-relevant*
goal region & objects



Attending to *embodiment*

On the Value of Aligned Representation – *Sample Efficiency*

RAPL outperforms
RLHF by 75% with 50%
less preference data



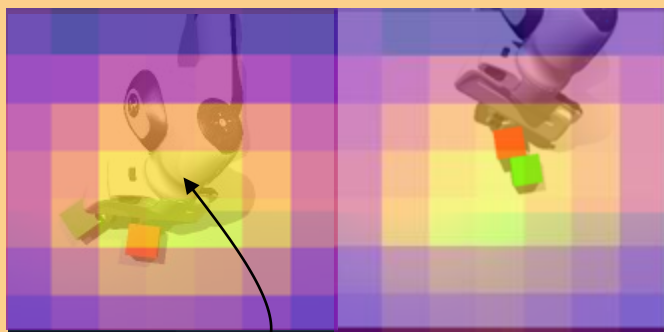
Insight

Representation alignment **reduces** the reliance on human preference feedback for achieving effective preference alignment.

Contribution to final embed.

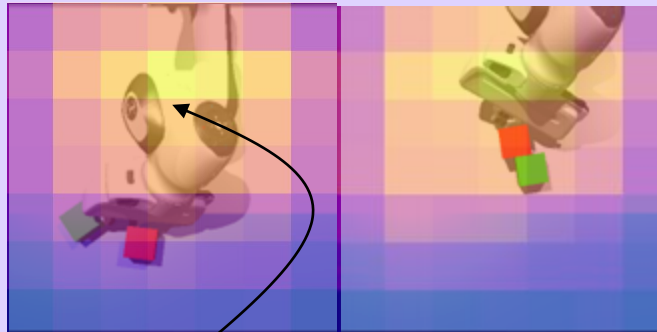
high low

ϕ_{RAPL}



Attending to *task-relevant*
goal region & objects

ϕ_{RLHF}



Attending to *embodiment*

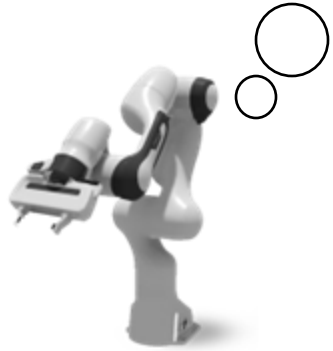
$\phi_{RLHF-2x \text{ data}}$



Preference Alignment of Real-World Robot Visuomotor Policies

Behavior Cloning

$$\max \mathbb{P}(\mathbf{a}_t^{\text{demo}} | \mathbf{o}_t; \text{task})$$



Pick up chips



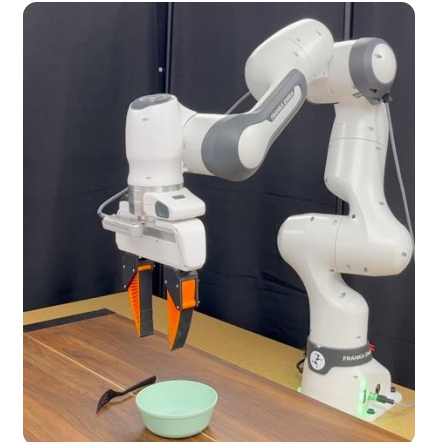
Crush the chips

Pick up cup



Contaminate water

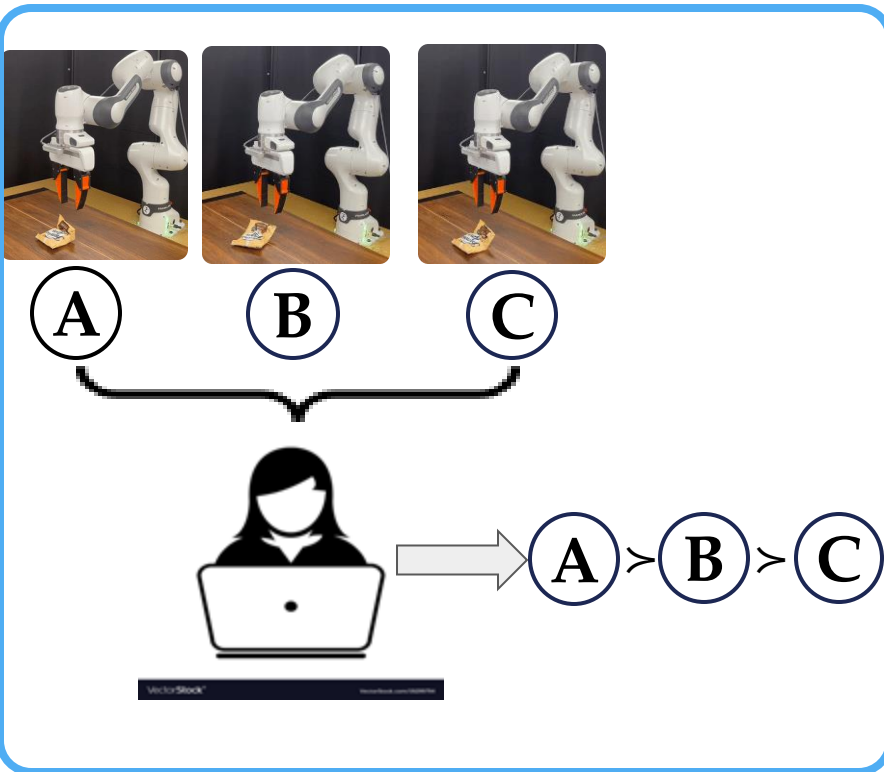
Pick up and place fork



Drop the fork

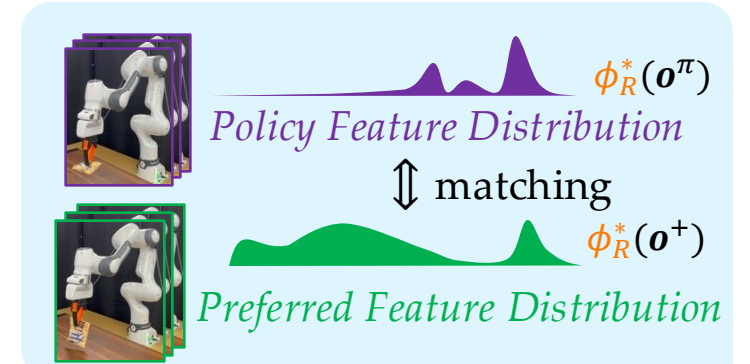
Preference Alignment of Real-World Robot Visuomotor Policies

No access to high-fidelity
simulators in deployment setting.



Visual Representation
Alignment

Reward via Feature Matching

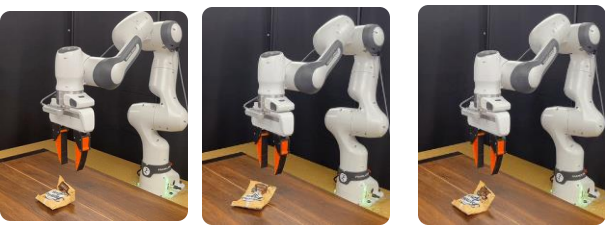


$$\max r_R(o^{\pi_{ref}}; \phi_R^*)$$

Motion Generation
Model

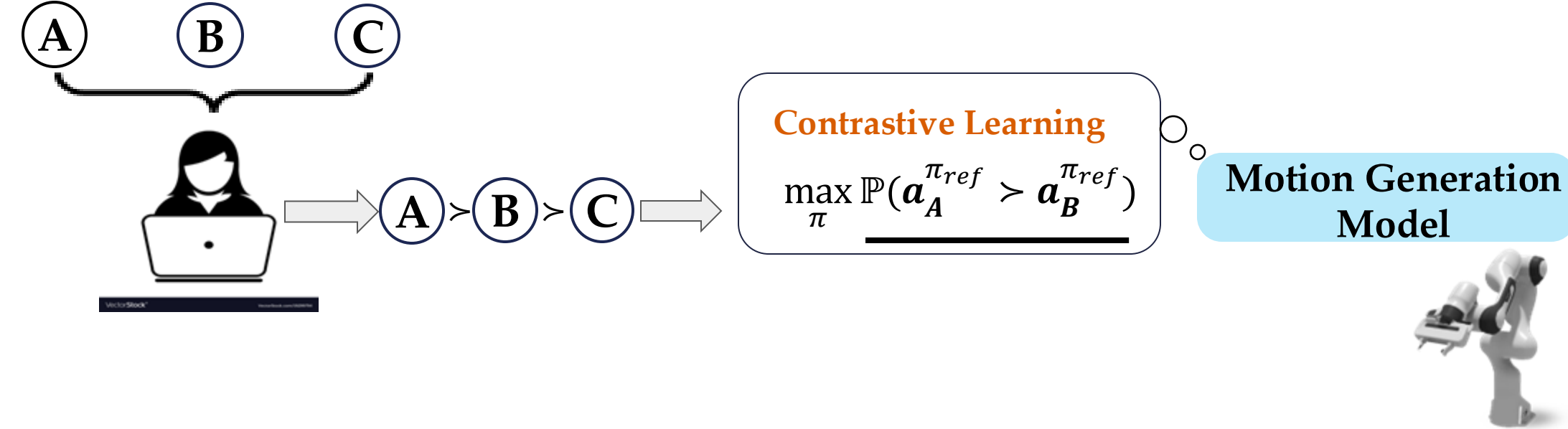


Preference Alignment of Real-World Robot Visuomotor Policies



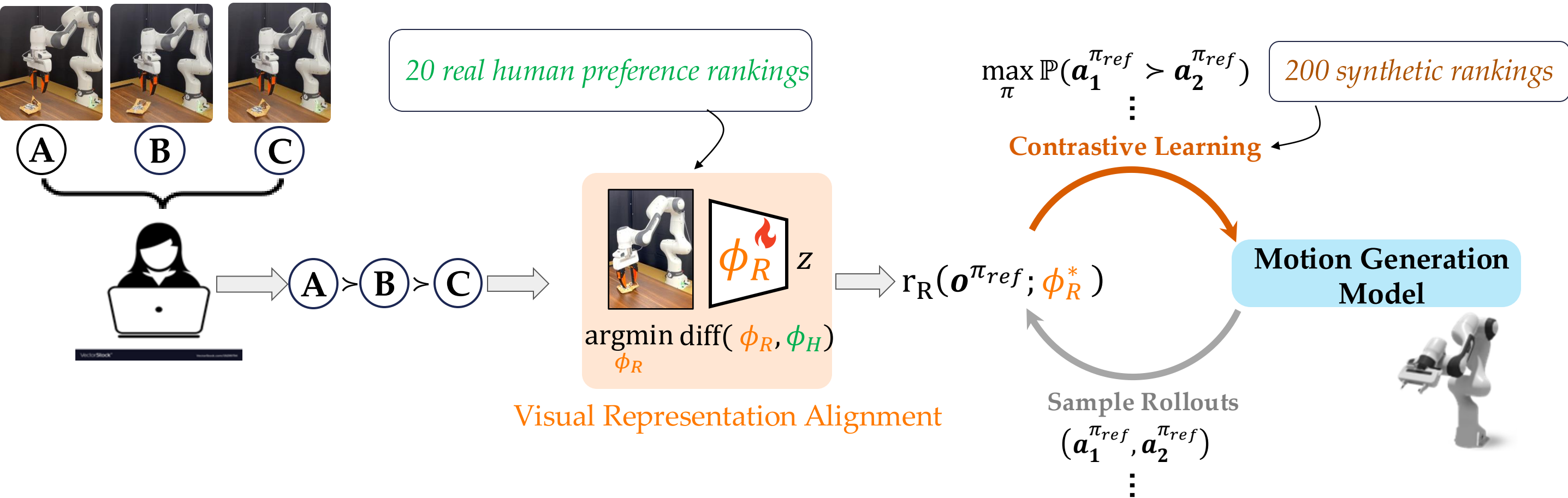
Need a huge number of rankings!

DPO: update generative model without RL!



Preference Alignment of Real-World Robot Visuomotor Policies

Key idea: leverage the reward built from the aligned representation to scalably generate synthetic rankings.



Preference Alignment of Real-World Robot Visuomotor Policies

Imitation Learning

$$\max \mathbb{P}(\mathbf{a}_t^{\text{demo}} | \mathbf{o}_t; \text{task})$$



Before alignment

Pick up chips



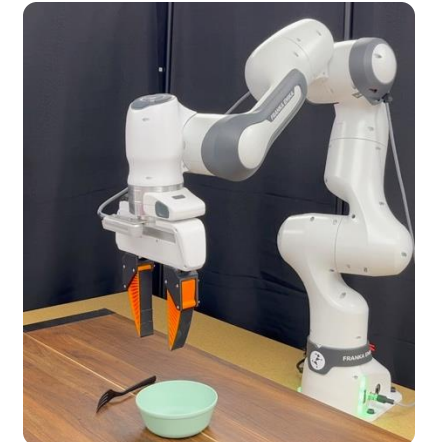
Crush the chips

Pick up cup



Contaminate water

Pick up and place fork



Drop the fork

$$\max \mathbb{P}(\mathbf{a}_+^{\pi_{\text{ref}}} > \mathbf{a}_-^{\pi_{\text{ref}}}; \phi_R^*, \text{task})$$

After alignment

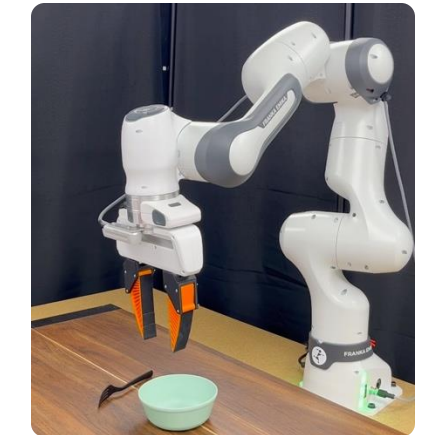
Preference Alignment



Hold the packaging
by its edges



Pick the cup by the
handle

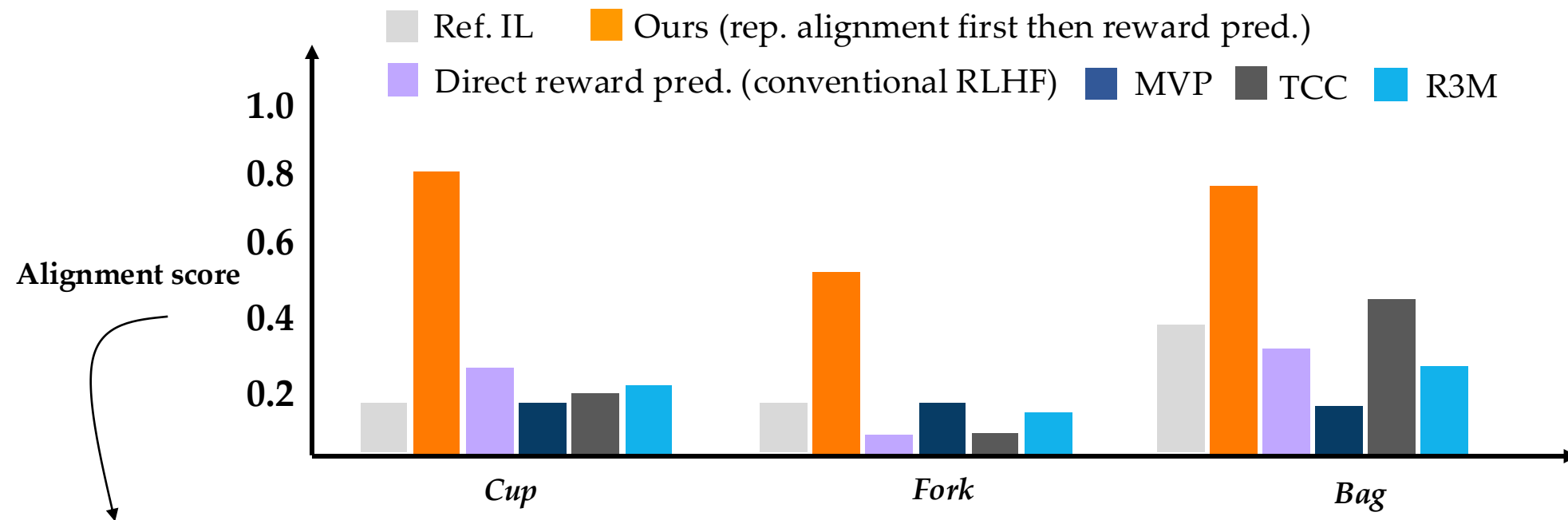


Gently place the fork

Preference Alignment of Real-World Robot Visuomotor Policies

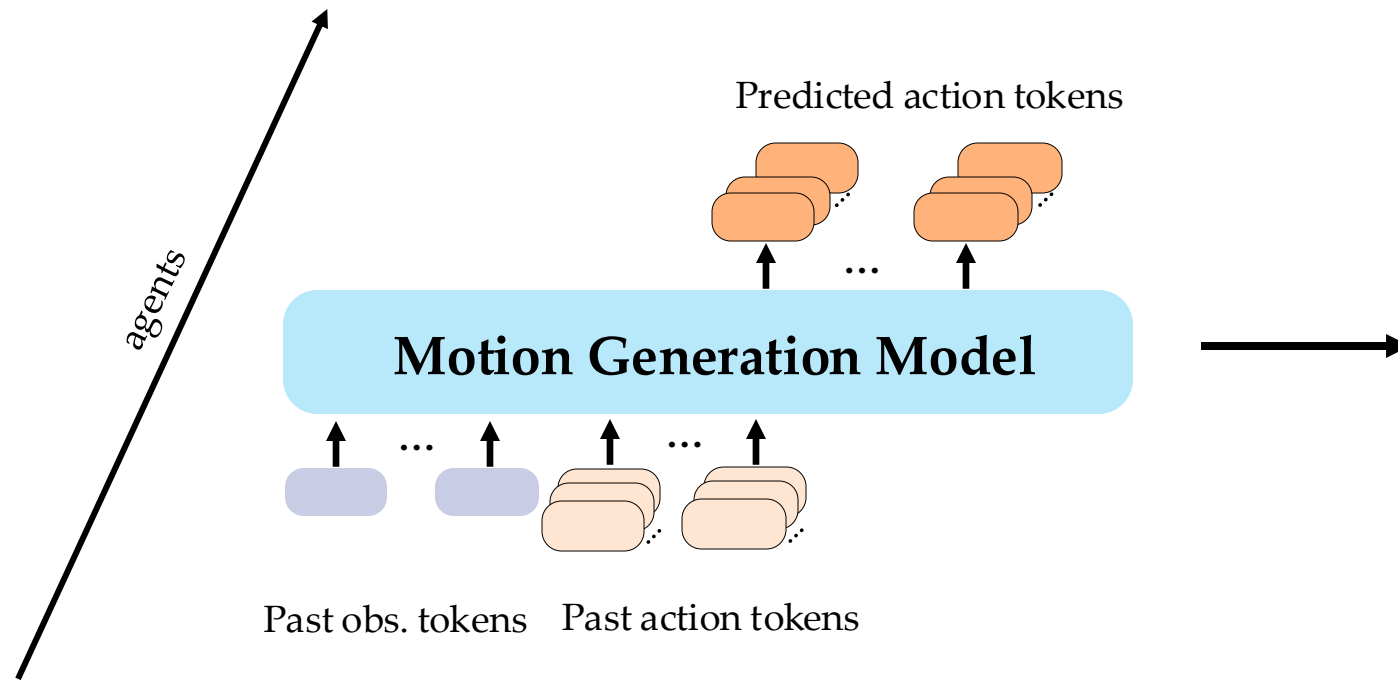
Result – alignment performance

Ours achieves **2x** (in average) better alignment scores compared to baselines under same amount of human budget

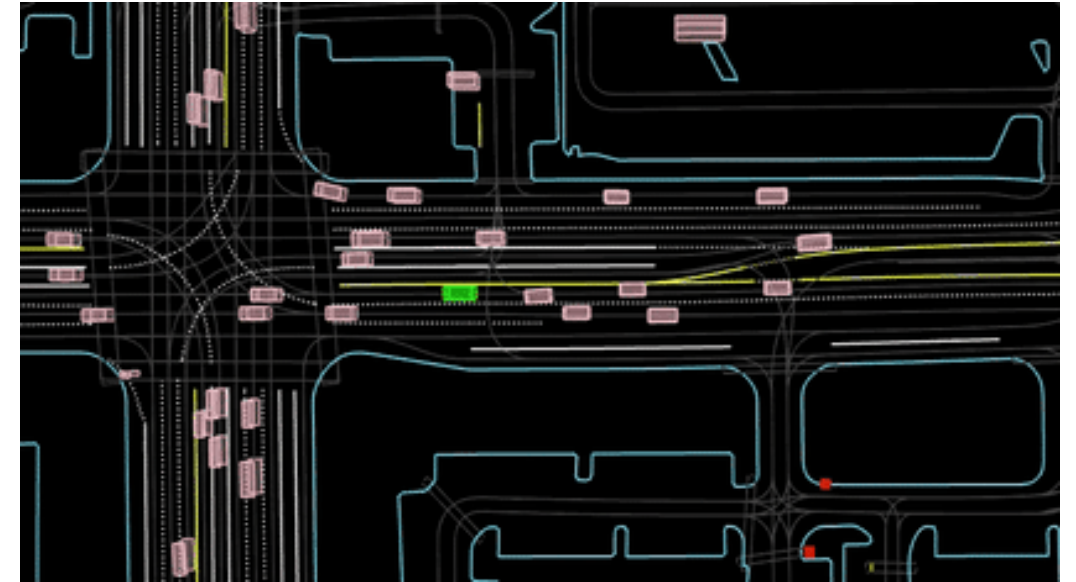


$$\frac{\# \text{ most-likely mode is preferred (graded by human)}}{\# \text{ test task configurations}}$$

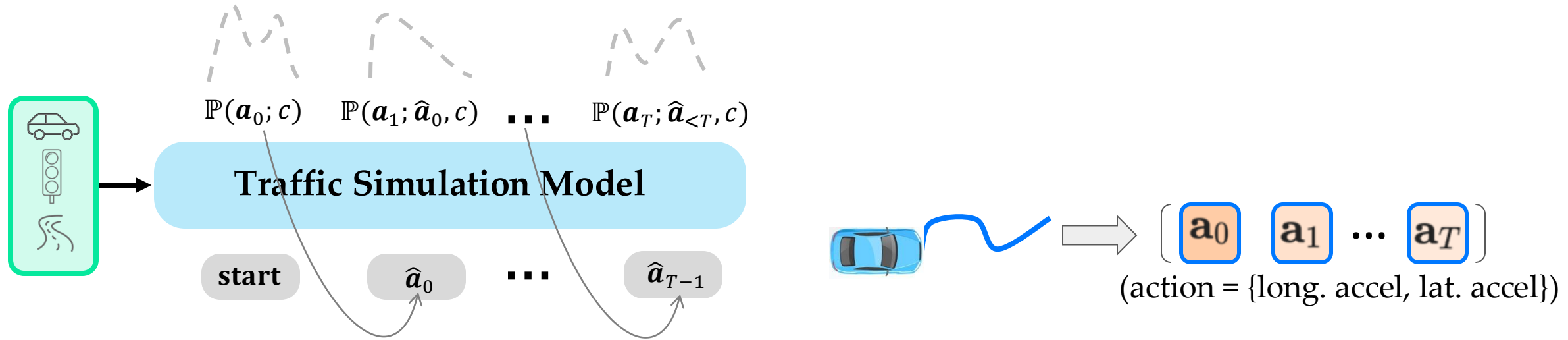
Let's Scale it up to Multi-agent!



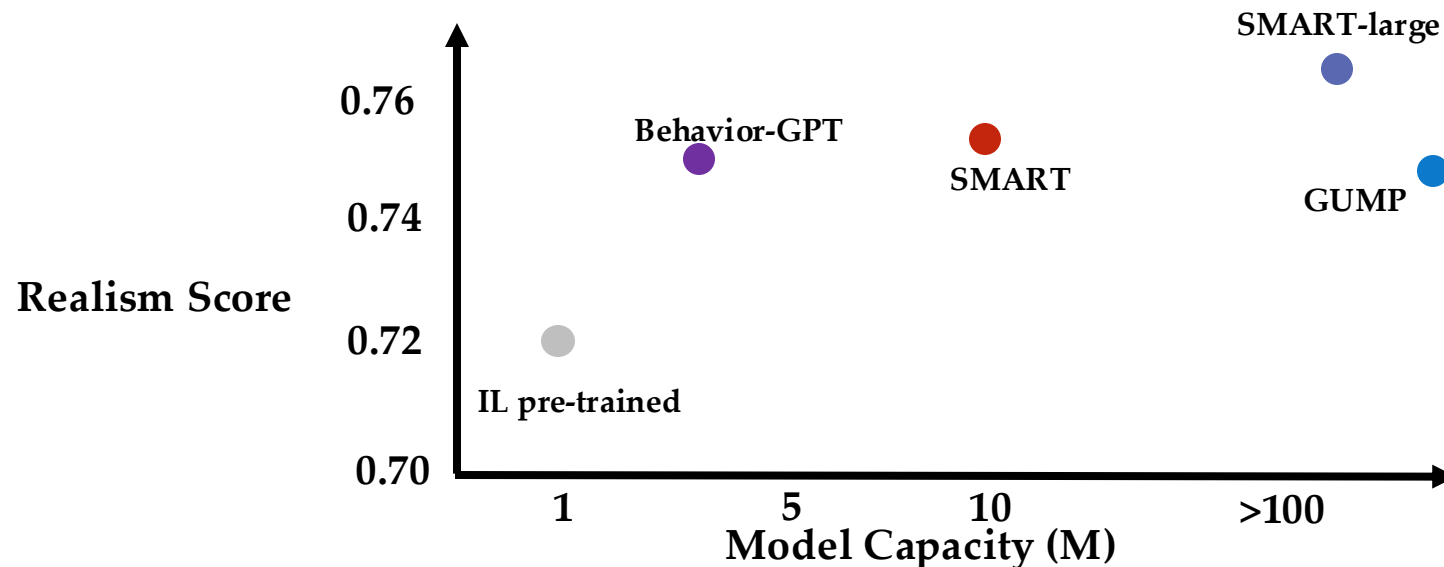
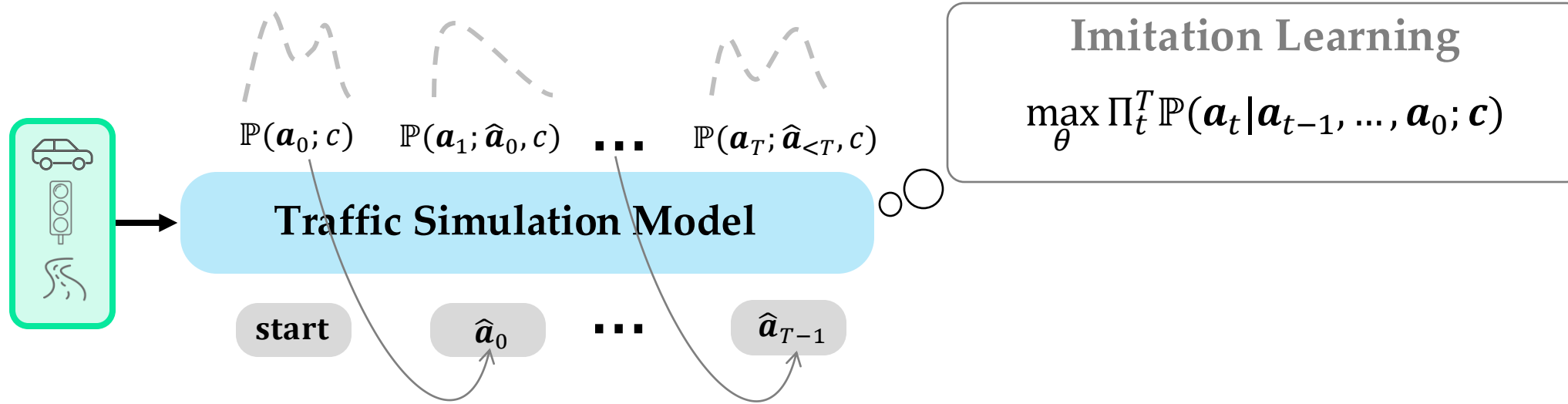
Large scale traffic simulation



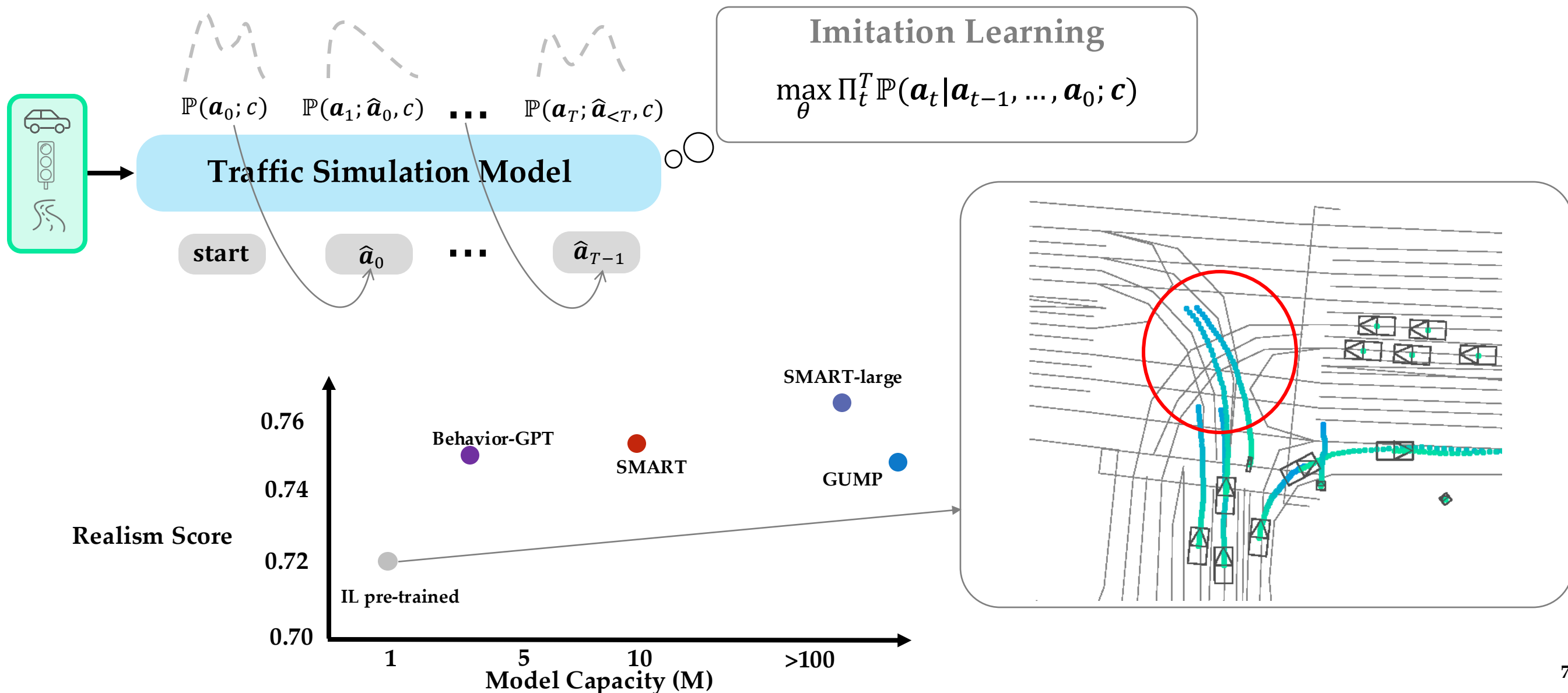
Imitation-learning based Traffic Simulation Model



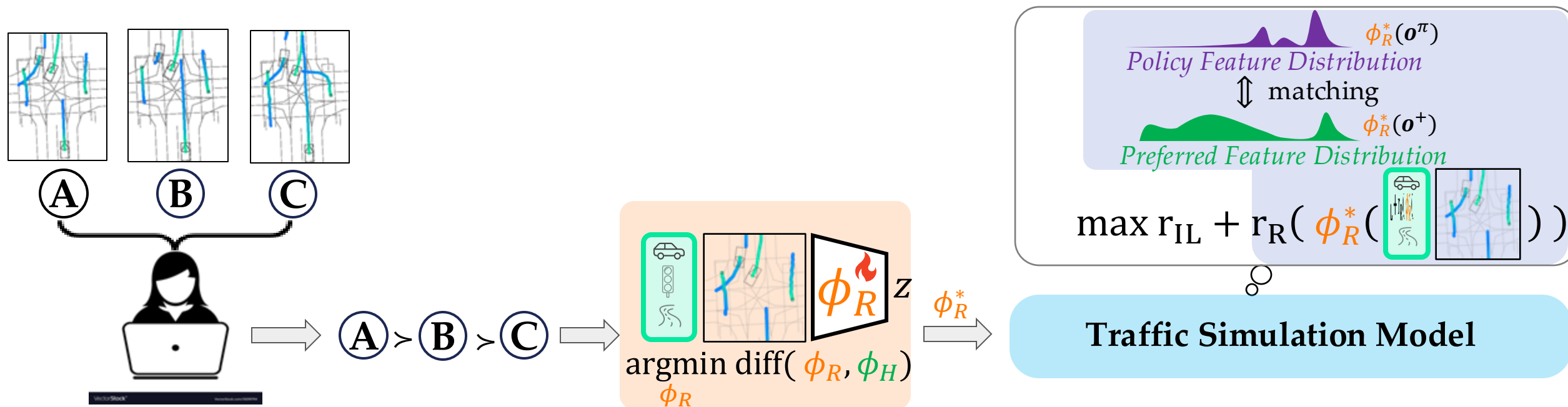
Imitation-learning based Traffic Simulation Model



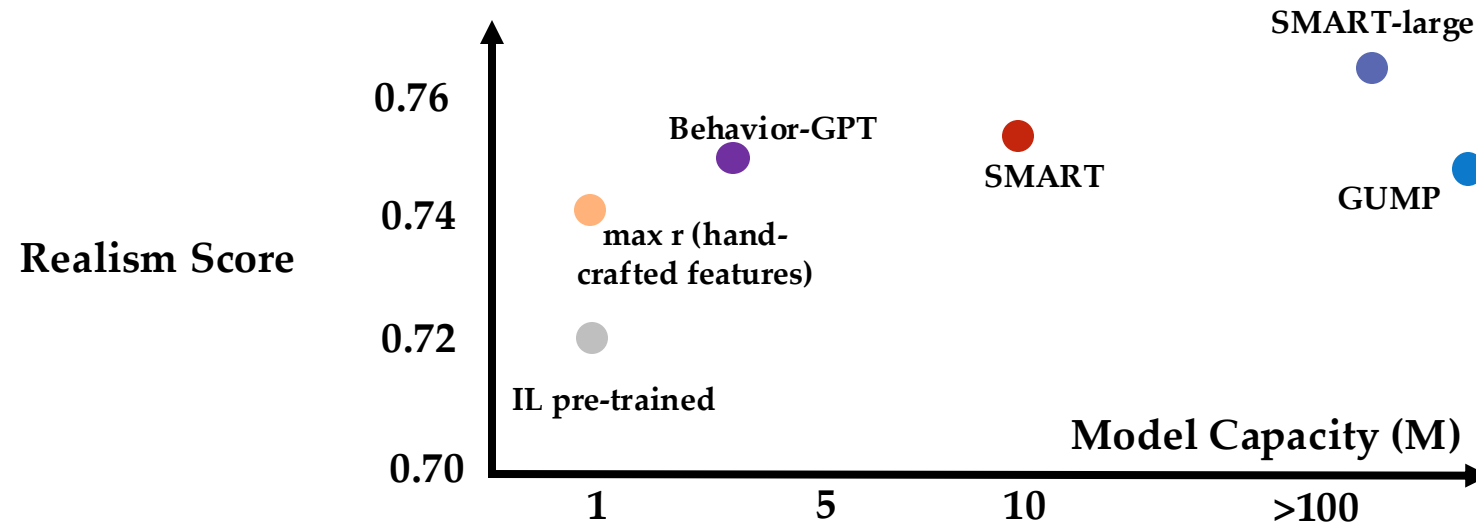
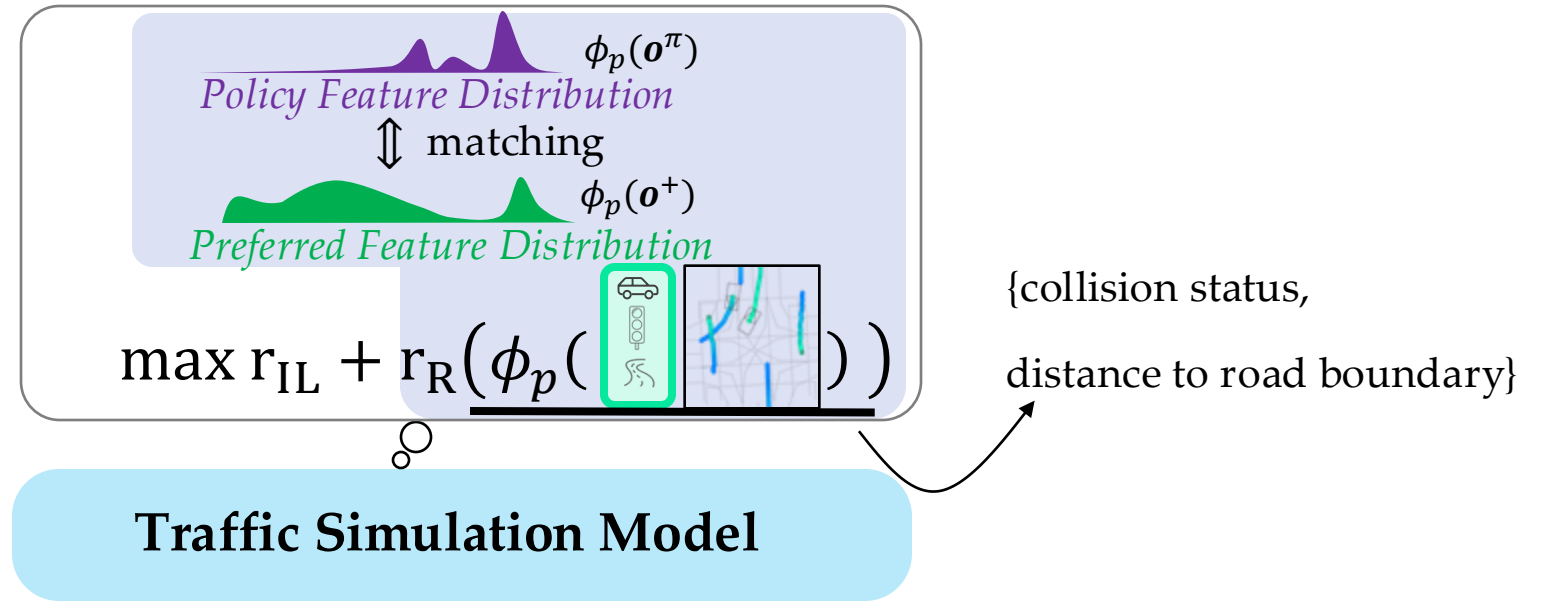
Imitation-learning based Traffic Simulation Model



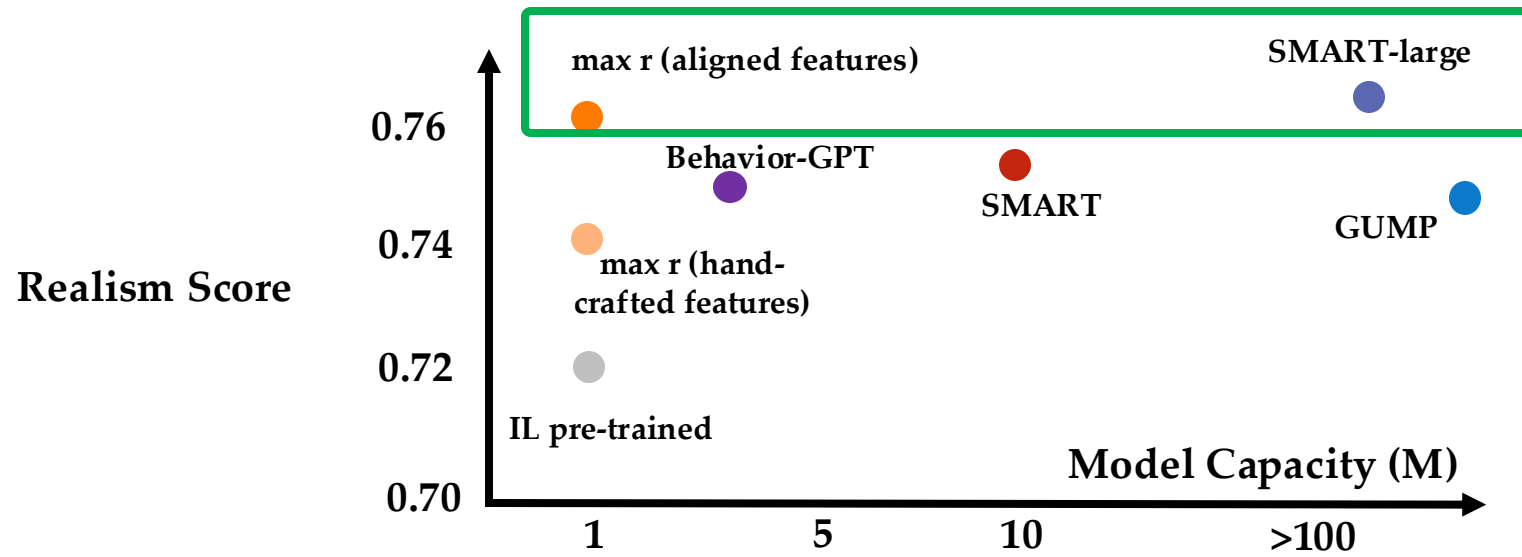
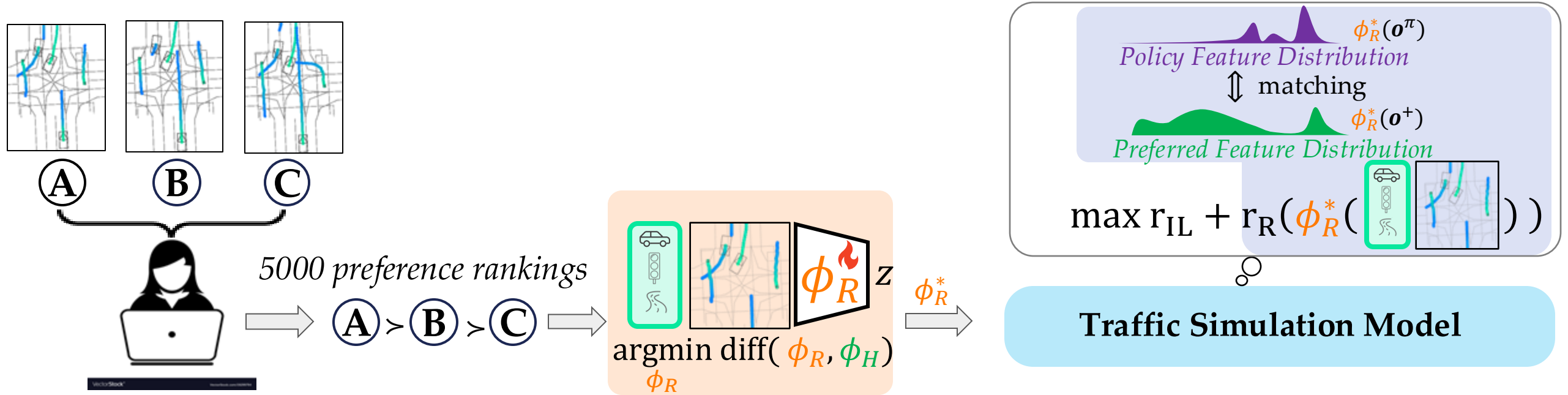
Preference-Alignment of Traffic Simulation Model



Preference-Alignment with *Hand-designed* Features



Preference-Alignment with *Aligned* Features

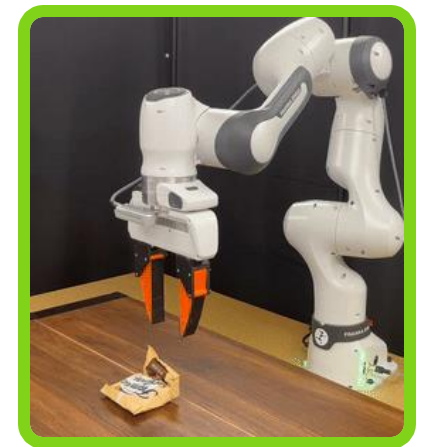


Ours is comparable to **100x larger** SOTA model

Takeaways

Model-based principles can help improve frontier robotics foundation models, but we need to “generalize” them to make them compatible with generative AI.

- ① Robotics foundation model’s training objective is only a proxy, we need post-training preference alignment
- ② We need robot representations to understand *what “matters” to us*
 - Require **10x less** human budget to achieve high preference alignment in robotics manipulation
 - Make a **1M** traffic model comparable to **100x larger** SOTA model after alignment with only a fraction of human data



VS.

