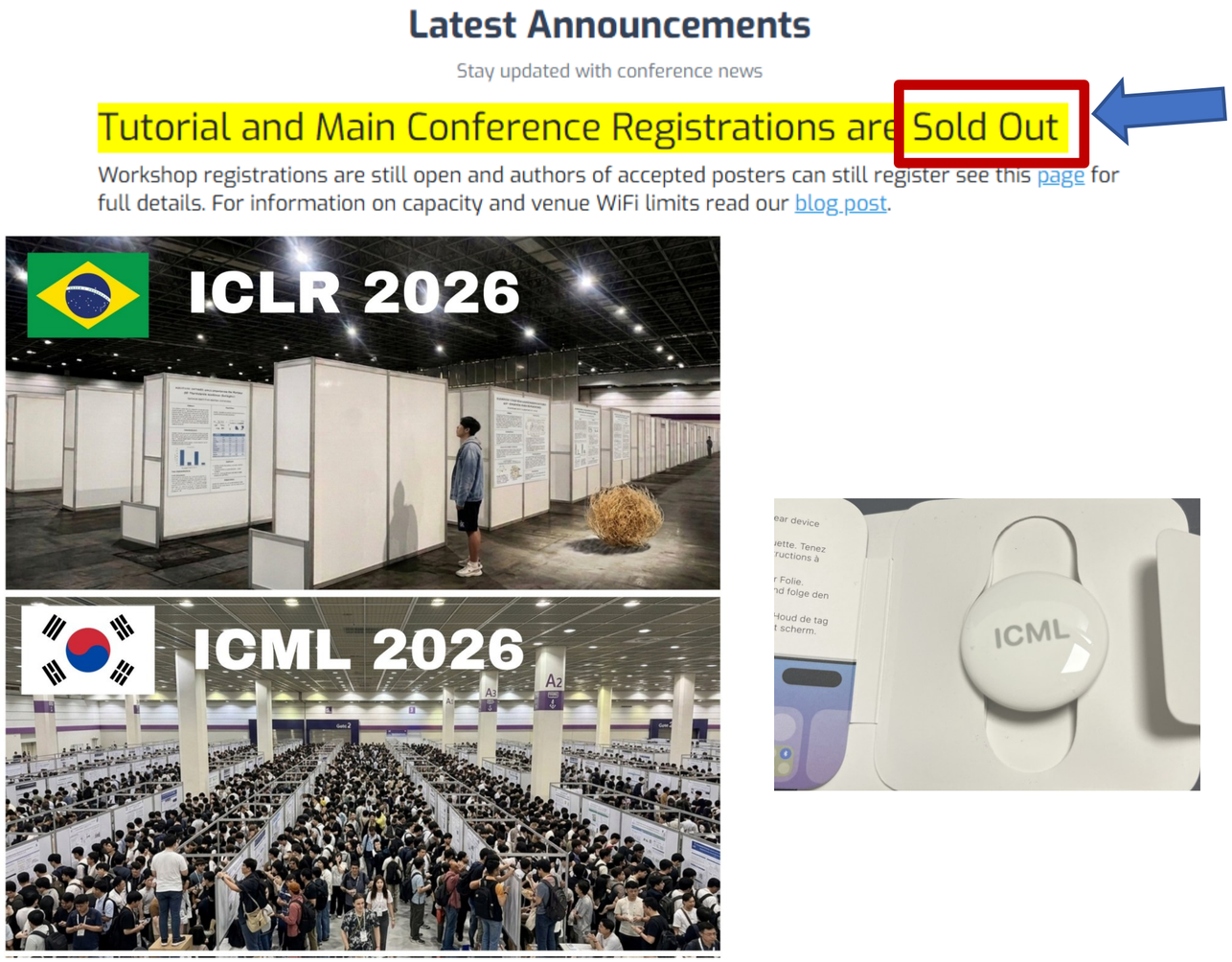
ICML Robotics Research Trends

Prof. Nuri Kim
Jeonbuk National University

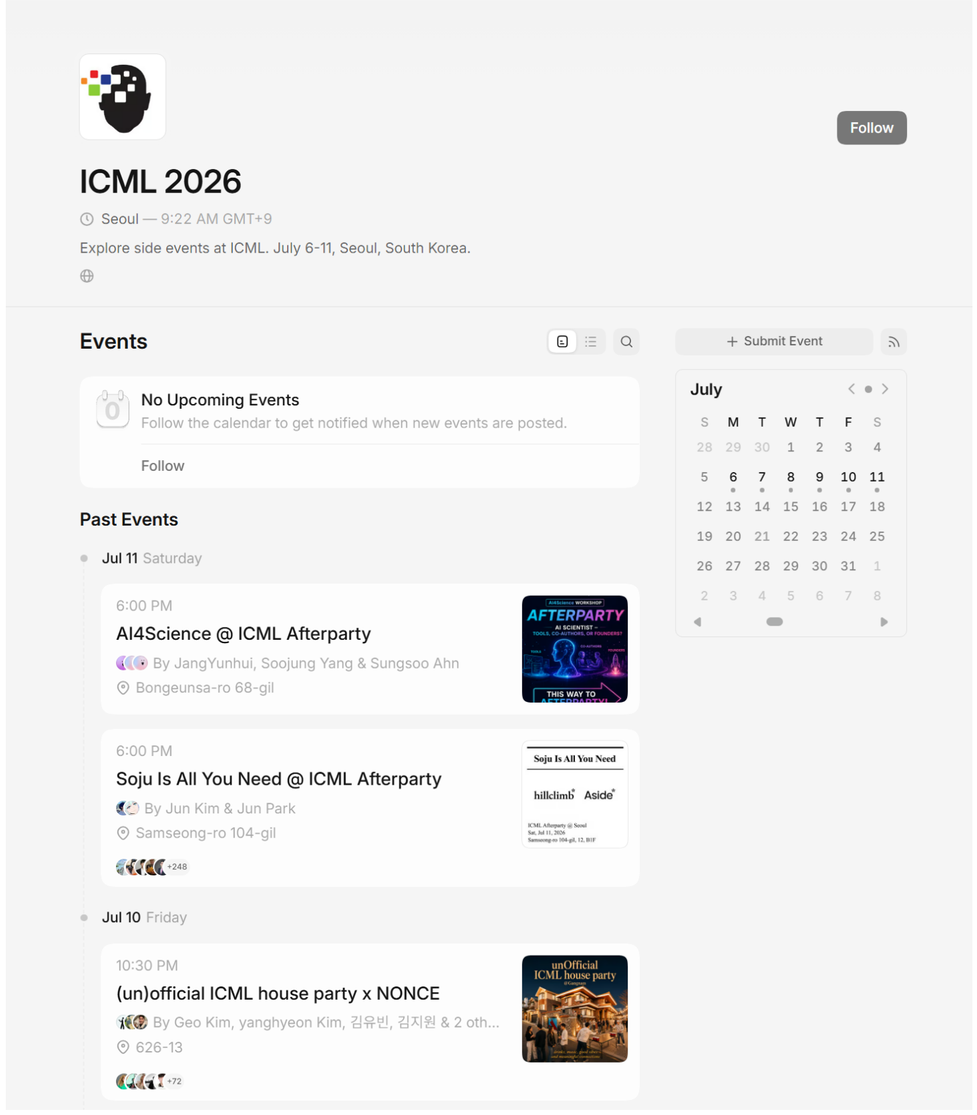
2026.07.20

ICML Overview



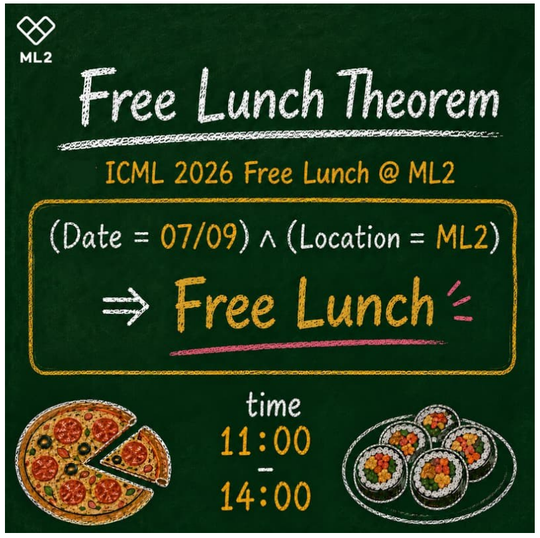
- A Massive Crowd
  - 20,456 attendees were reportedly on site
  - Koreans seem to enjoy attending conferences and studying
  - ICML rented every hall at COEX, making it much more comfortable than last year's CoRL
- ICML Has Deep Pockets
  - Quant firms must have plenty of money (Citadel, Jane Street, etc.)
  - Corporate booths served free ice cream and barista-made coffee (at nearly every booth—mostly quant firms)
  - Quiz contests that felt like IQ tests offered prizes (I won an AirTag)

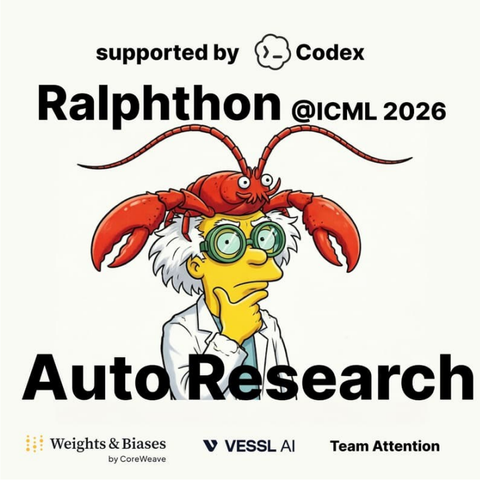
ICML Overview



- Many Luma networking events brought together researchers with similar interests
  - These days, almost every computer vision lab has at least one robot arm
  - Some professors use Claude Code/Codex for solo research and paper writing (Prof. Min-hwan Oh, Seoul National University)
  - More robotics researchers attended ICML than expected
  - Many VCs want to invest in AI/robotics

Networking Events





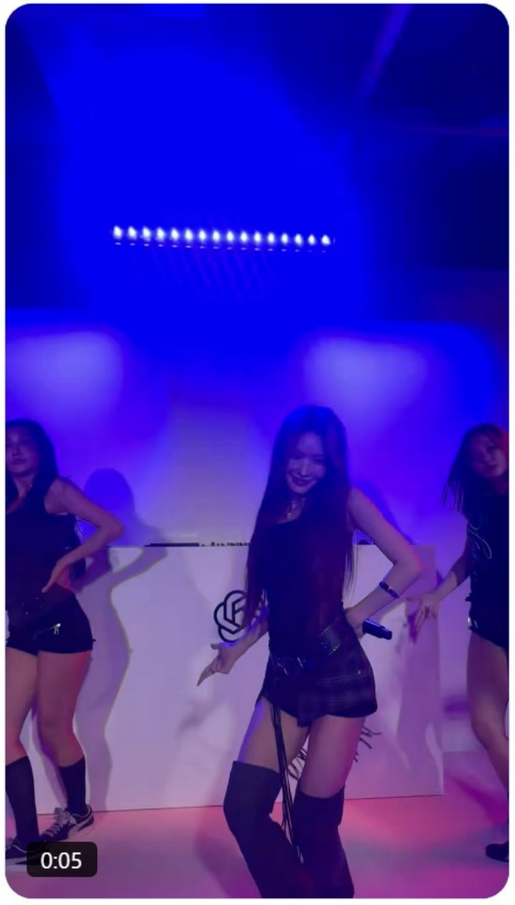
Networking Events

AI for Science Meetup

what happens when you put a kpop girl group in front of a room of ai researchers? #icml

OpenAI Meetup

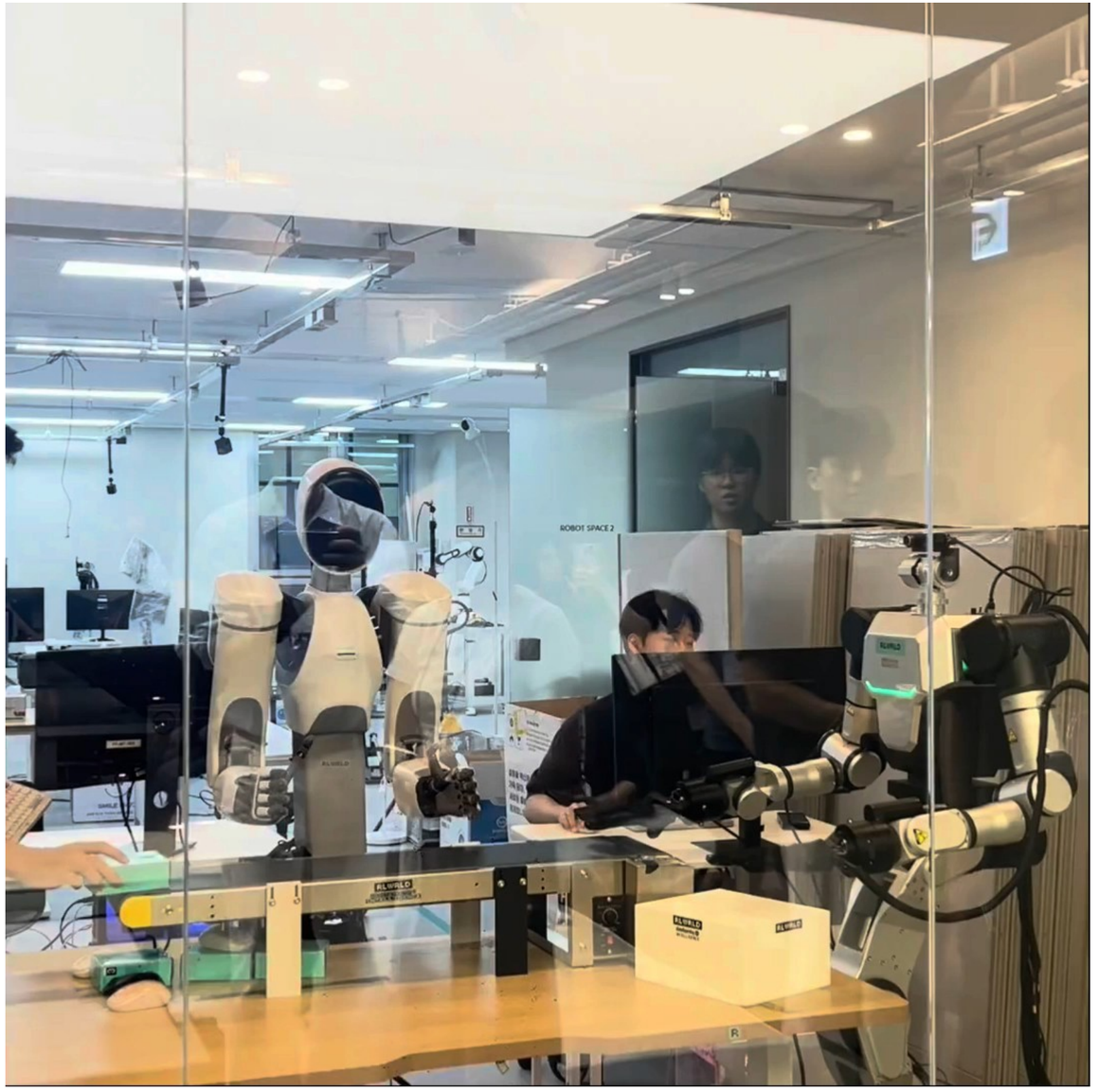
guys I just cancelled my Claude plan I don’t know what happened



Chungha at the OpenAI party was so awesome that some people canceled Claude

RLWORLD Networking

RLWORLD Networking



Academic Celebrities

- More robotics researchers attended ICML than expected!

Chelsea Finn

Perhaps they were attending RSS?

Workshop Talks

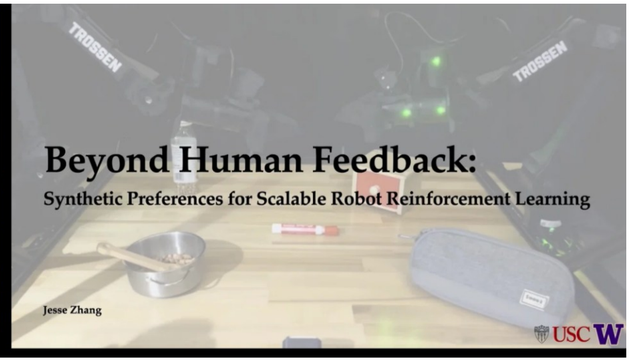
Data-Driven Control

Sergey Levine
UC Berkeley
Physical Intelligence

Emergent Physical Generalization

Chelsea Finn

Stanford Artificial Intelligence
IRIS LAB
physicalintelligence.company



Generative Modeling for Multimodal AI Agents: Reasoning, Action, and Simulation

ICML 2026 SCALE Workshop

Minhyuk Sung

KAIST

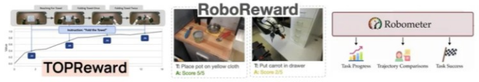
The problem with scalar rewards & preferences

People are **really bad** at specifying it.

a. Leads to major misalignment

b. Leads to overlooked dimensions of behavior.

Strong focus on basic task completion.

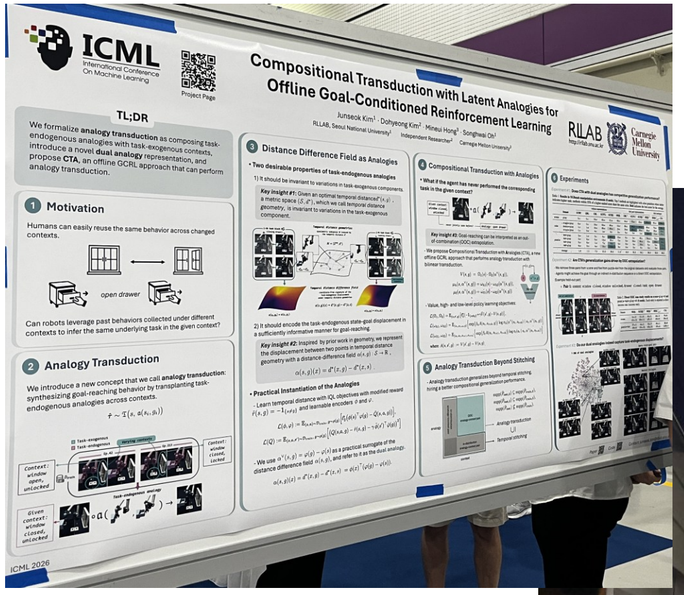


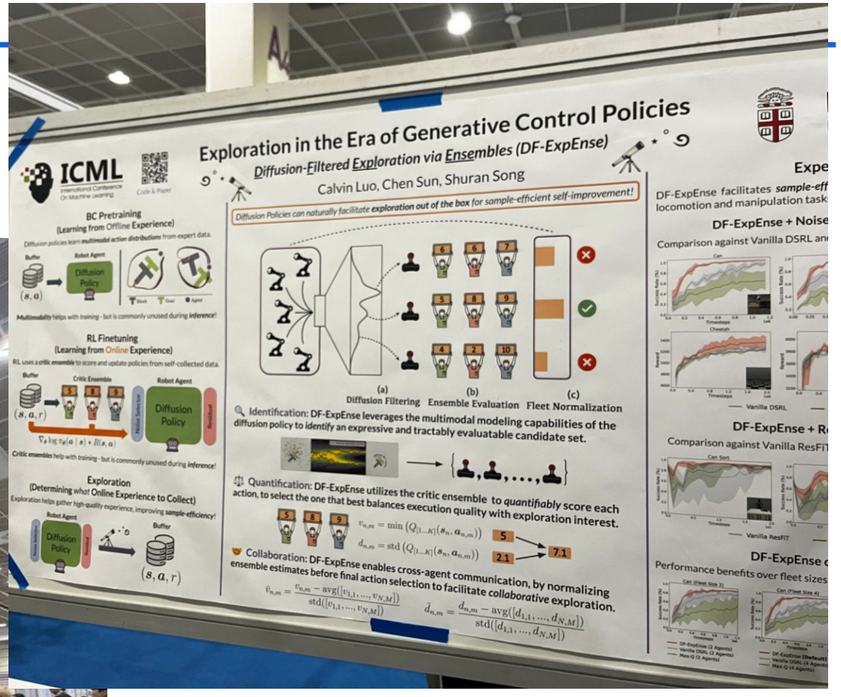
Generative Adversarial Harness

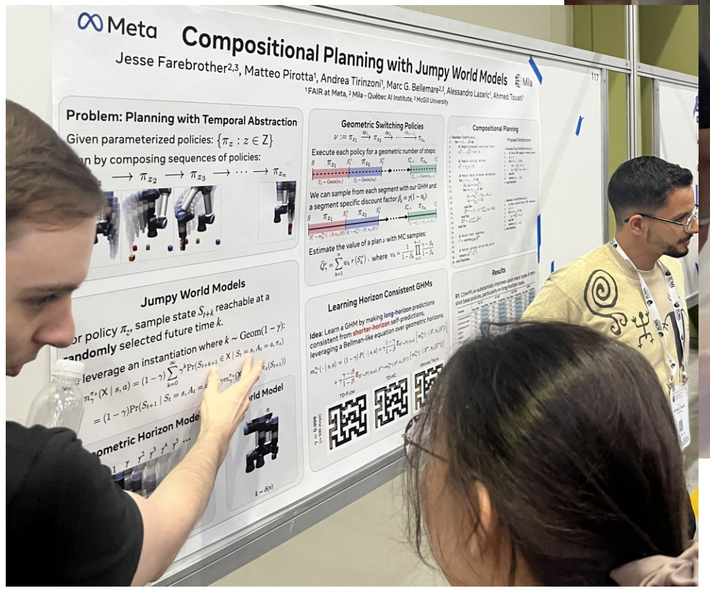
Kangwook Lee
KRAFTON / Ludo Robotics

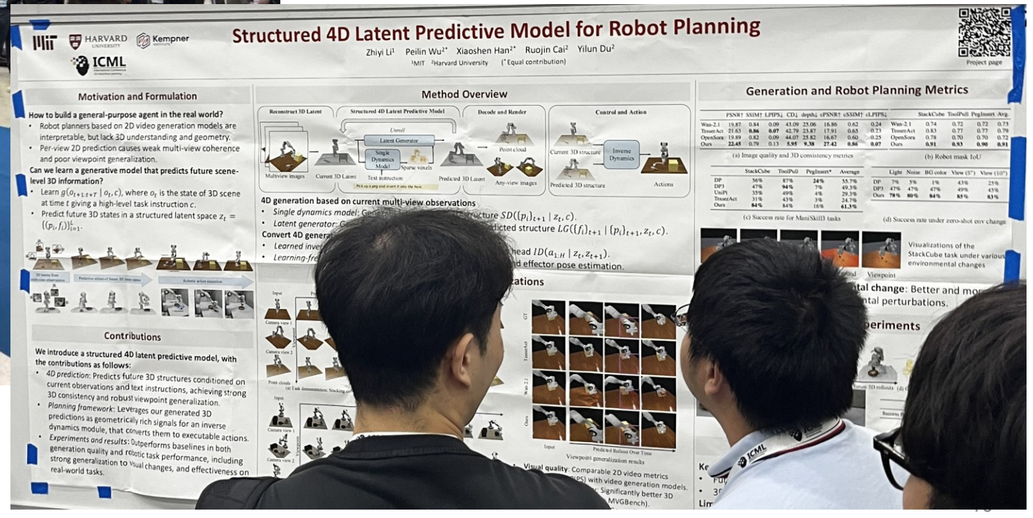
Joint work with Beomhan Baek, Junsoo Oh, Seonho Lee, and Junhyuck Kim.

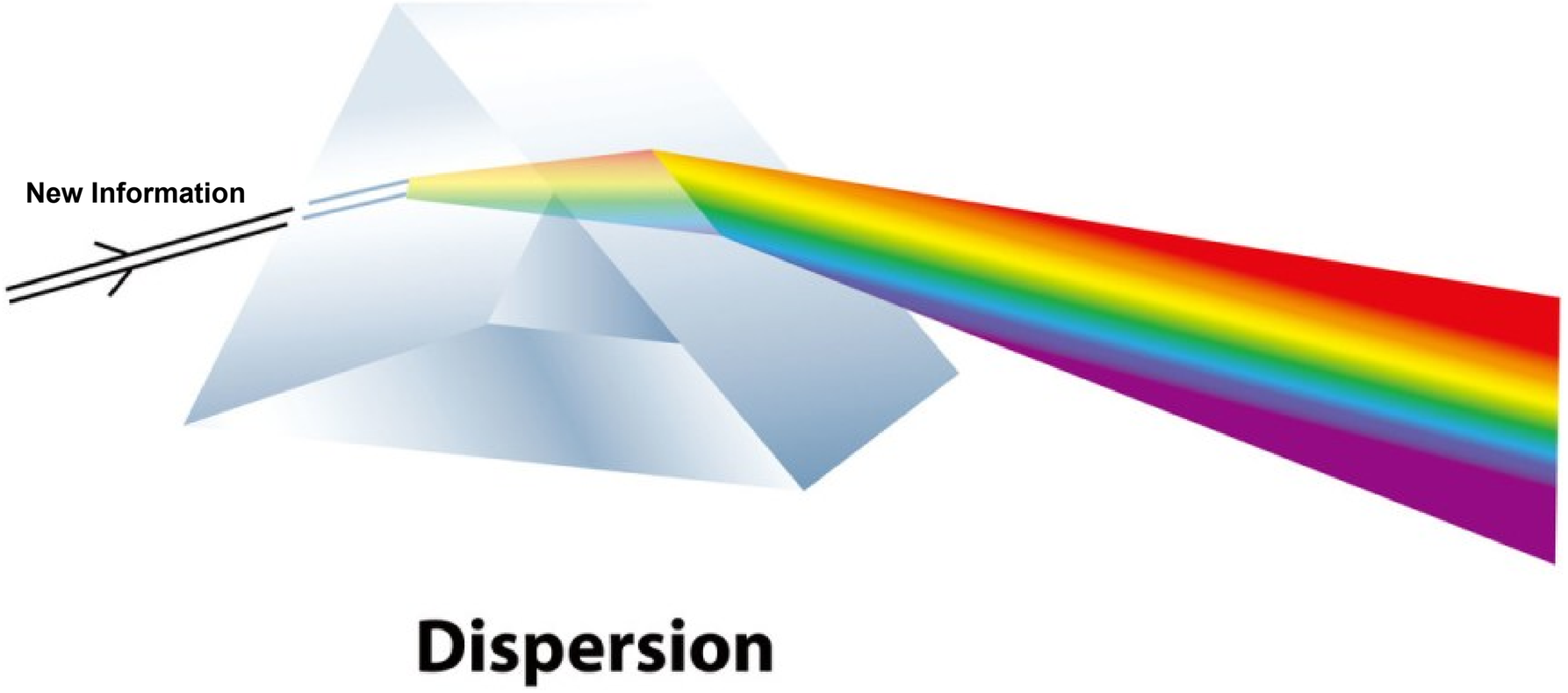
Poster Session










New Information
Research Ideas
.
.
Fresh Inspiration
.
.
Next Paper Topic

Dispersion

Key Issues in Robotics Research

- Data-Driven **vs**. Theory-Driven **Improvements: Which Is Better?**
- RL **vs.** SFT**: Which Is Better?**
- RL Limitations: How Can **We Address Them?**
  - Limitation 1: Reward Design Is Difficult
  - Limitation 2: Rewards Are Scalar
  - Limitation 3: Actor Distributions Tend to Collapse to a Single Mode
  - Limitation 4: RL Fine-Tuning After IL Leads to Poor Exploration
  - Limitation 5: Long-Horizon Policies Are Hard to Learn
  - Limitation 6: Poor Generalization

Key Issues in Robotics Research

- Data-Driven **vs**. Theory-Driven **Improvements: Which Is Better?**
- RL **vs.** SFT**: Which Is Better?**
- RL Limitations: How Can **We Address Them?**
  - Limitation 1: Reward Design Is Difficult
  - Limitation 2: Rewards Are Scalar
  - Limitation 3: Actor Distributions Tend to Collapse to a Single Mode
  - Limitation 4: RL Fine-Tuning After IL Leads to Poor Exploration
  - Limitation 5: Long-Horizon Policies Are Hard to Learn
  - Limitation 6: Poor Generalization

Data-Driven vs. Theory-Driven Improvements

- Data-Driven
  - Pure Data Scaling
  - As with LLMs and VLMs, scaling data may produce emergent abilities

- Theory-Driven Improvements
  - Structured Representations
  - Physics in the loop

My Opinion

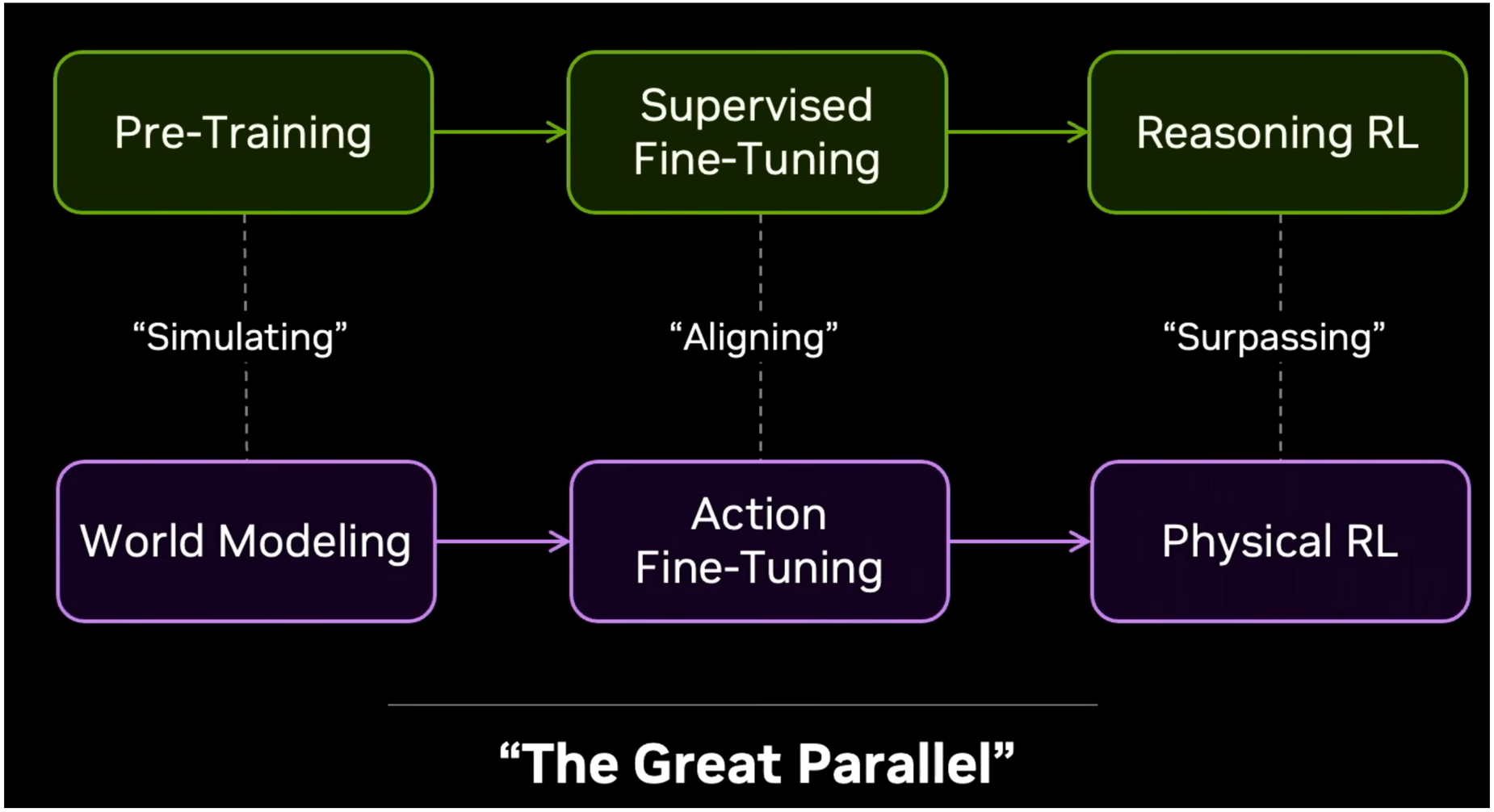
- After the discussion, I was left thinking that China may win through sheer scale in data collection, given its weaker privacy and human-rights protections

- China's large market is said to accelerate technological progress, as many Chinese professionals return after working at U.S. Big Tech firms

My Opinion

- After the discussion, I was left thinking that China may win through sheer scale in data collection, given its weaker privacy and human-rights protections
- China's large market is said to accelerate technological progress, as many Chinese professionals return after working at U.S. Big Tech firms
- But we lack data, large-scale GPUs for training, and top talent returning from U.S. Big Tech, so we need to focus on theoretical solutions
- In any case, both high-quality data and theoretical advances **are** essential to learning
  - **A central research question: where and how should structural knowledge be injected into data-driven** models?

The Success of LLMs



Data-Driven: Pure Data Scaling

- As with LLMs and VLMs, scaling data may produce emergent abilities

Data Scaling Laws

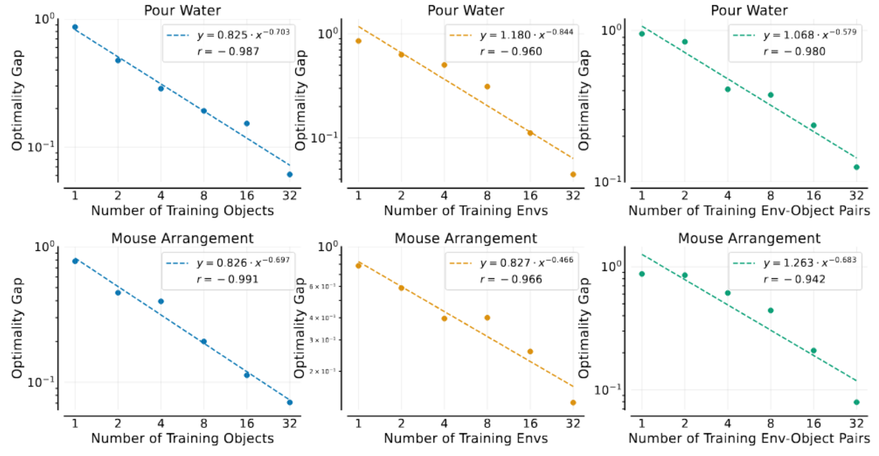
in Imitation Learning for Robotic Manipulation

ICLR 2025 Oral Presentation

Best Paper Award at Workshop on X-Embodiment Robot Learning, CoRL 2024

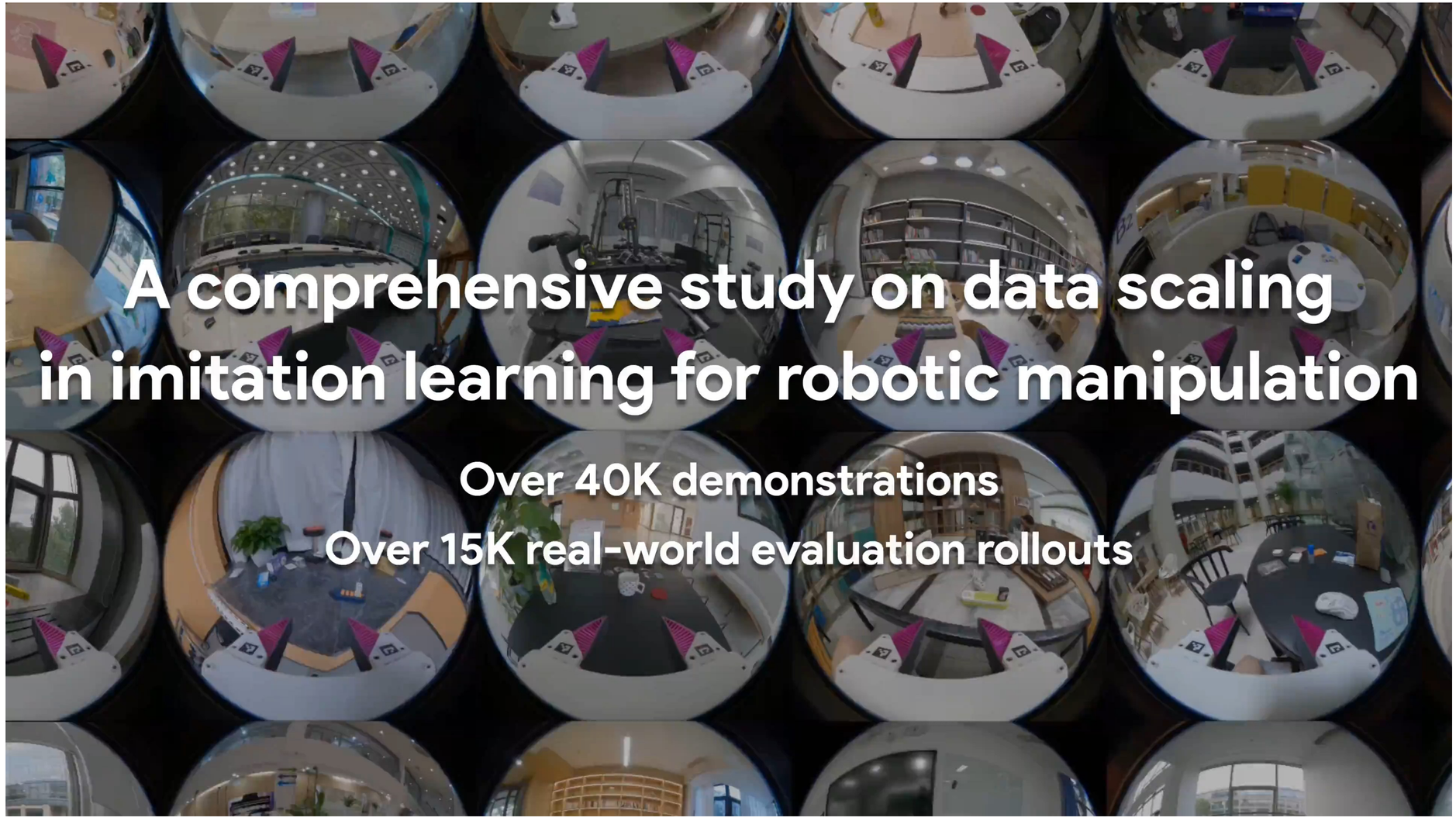
Fanqi Lin[1,2,3*], Yingdong Hu[1,2,3*], Pingyue Sheng[1], Chuan Wen[1,2,3], Jiacheng You[1], Yang Gao[1,2,3]

[1]Tsinghua University, [2]Shanghai Qi Zhi Institute, [3]Shanghai Artificial Intelligence Laboratory



Data-Driven: Pure Data Scaling

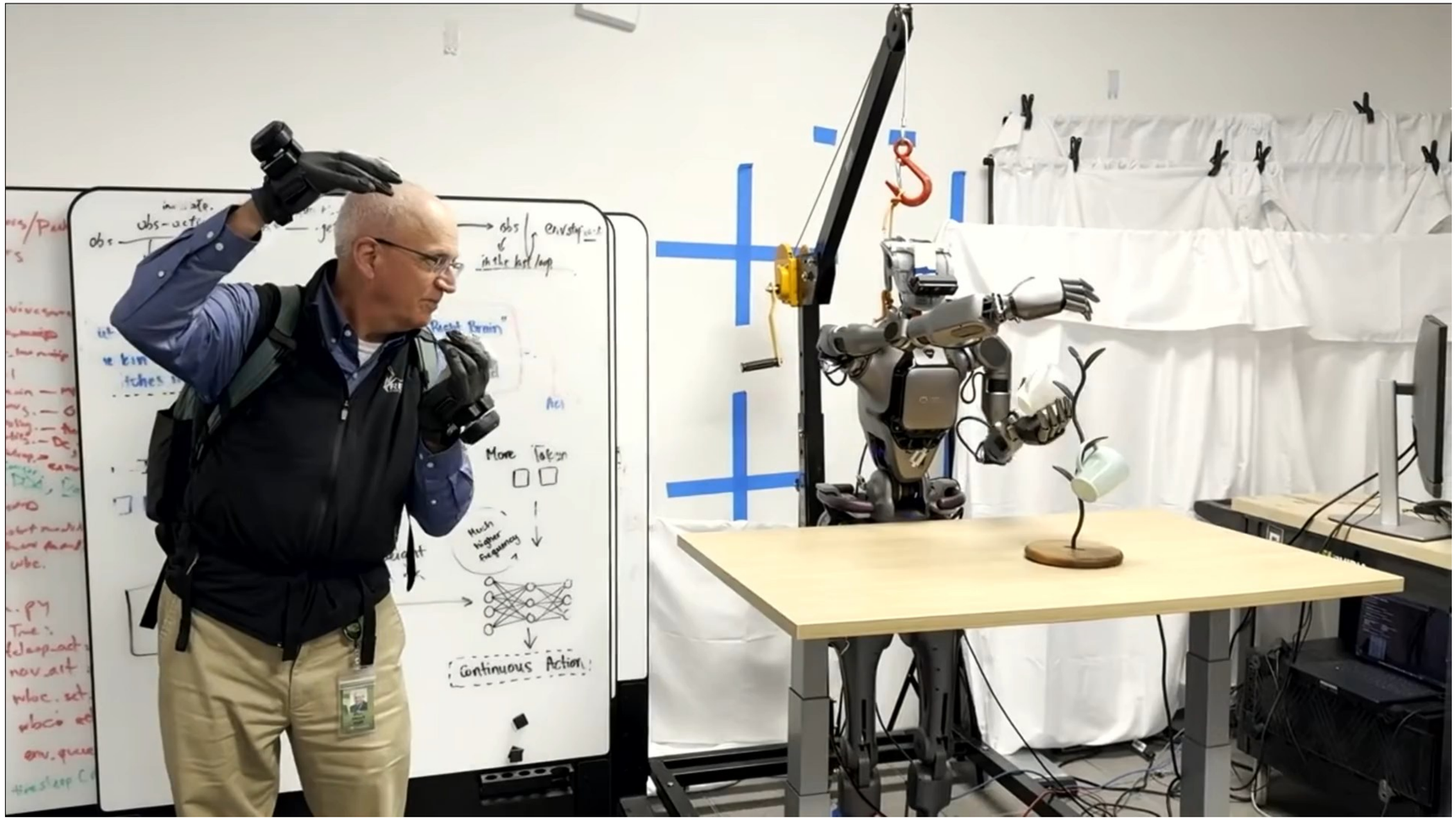
- As with LLMs and VLMs, scaling data may produce emergent abilities



The **Problem: Not Enough Robot** Data

How Can We Collect Robot Data?

- Teleoperation



How Can We Collect Robot Data?

- UMI

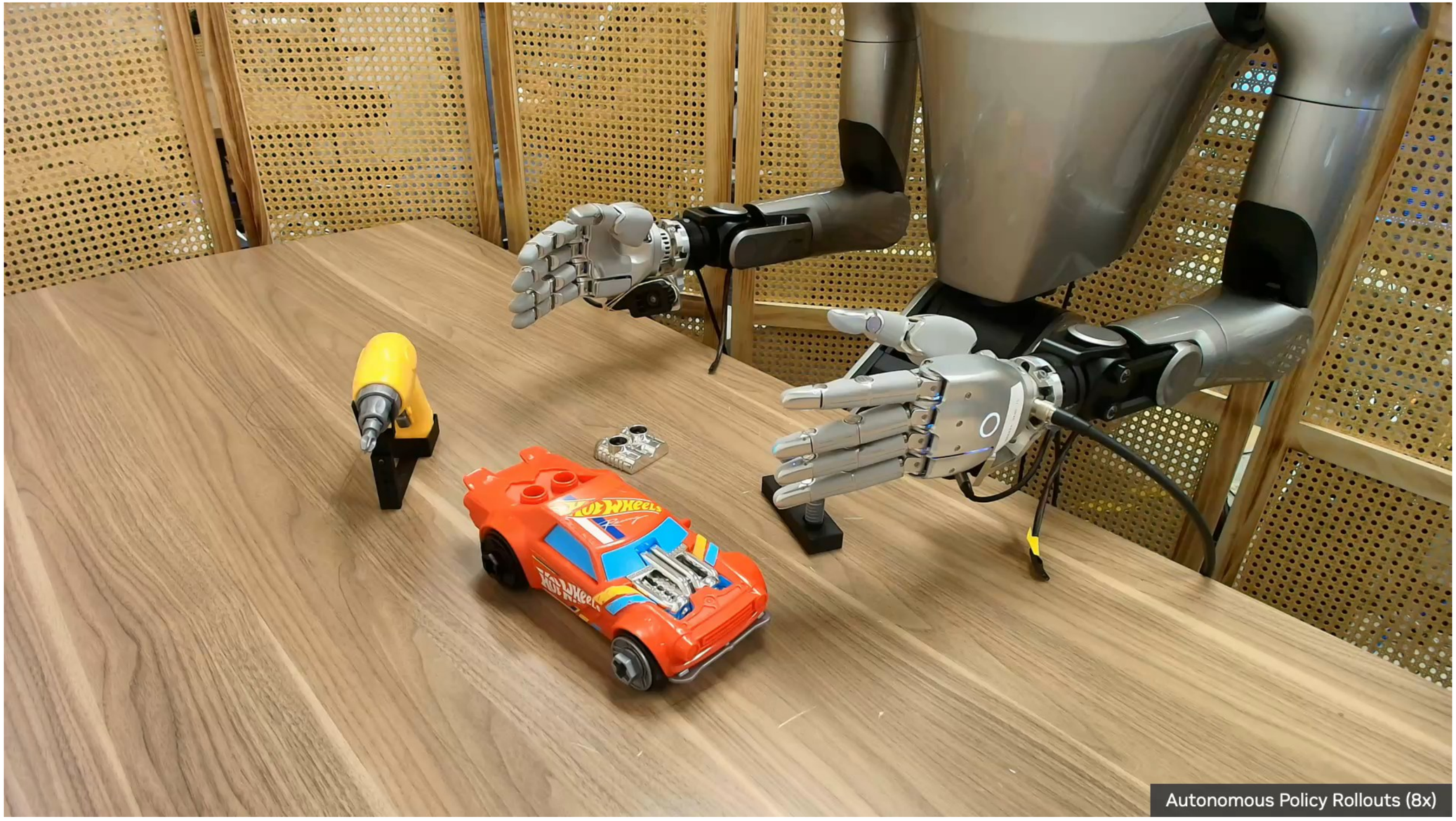
Source: Forbes article on Generalist. Apr. 2026

Source: Sunday Robotics, 2026

How Can We Collect Robot Data?

- Data From Human Video

Learning From Human Video



How Can We Use Robot **Data** Efficiently?

Data-Efficient Learning

- DexMachina: Functional Retargeting for Bimanual Dexterous Manipulation (Shuran Song, Dieter Fox, et al.)
  - Functional Retargeting enables a robot to interact with an object from a single demo

Data-Efficient Learning

- DexMachina: Functional Retargeting for Bimanual Dexterous Manipulation (Shuran Song, Dieter Fox, et al.)
  - Functional Retargeting enables a robot to interact with an object from a single demo

Many human demonstrations:

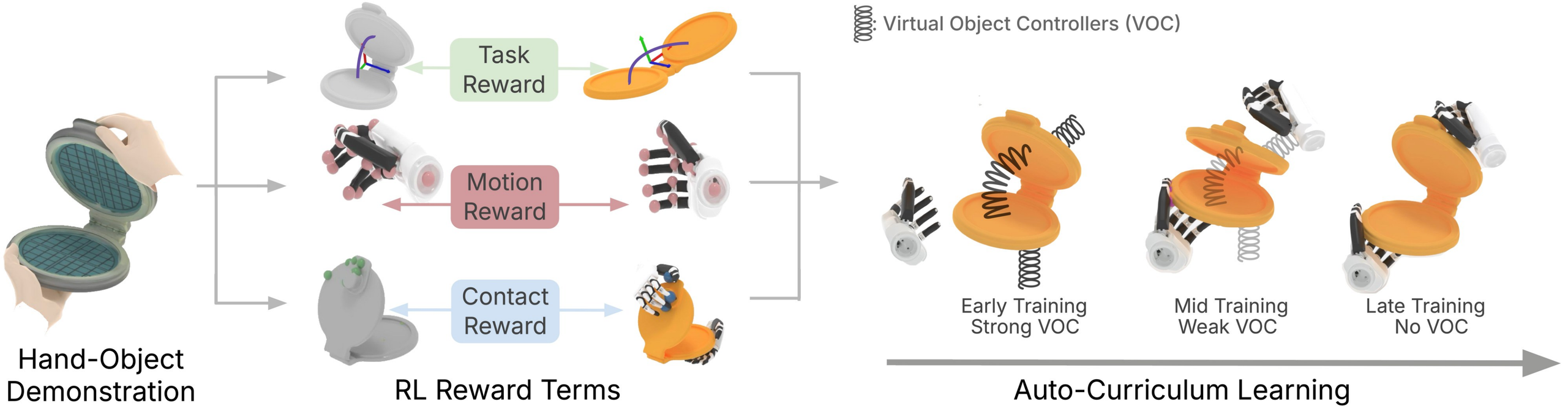
- Are naturally hong-horizon
- Contain mid-air object manipulation
- Require precise bimanual coordination

But a naive RL policy:

- Gets stuck at early failures
- Choose suboptimal actions that cannot achieve the full sequence

Data-Efficient Learning

- DexMachina: Functional Retargeting for Bimanual Dexterous Manipulation (Shuran Song, Dieter Fox, et al.)
  - Functional Retargeting enables a robot to interact with an object from a single demonstration



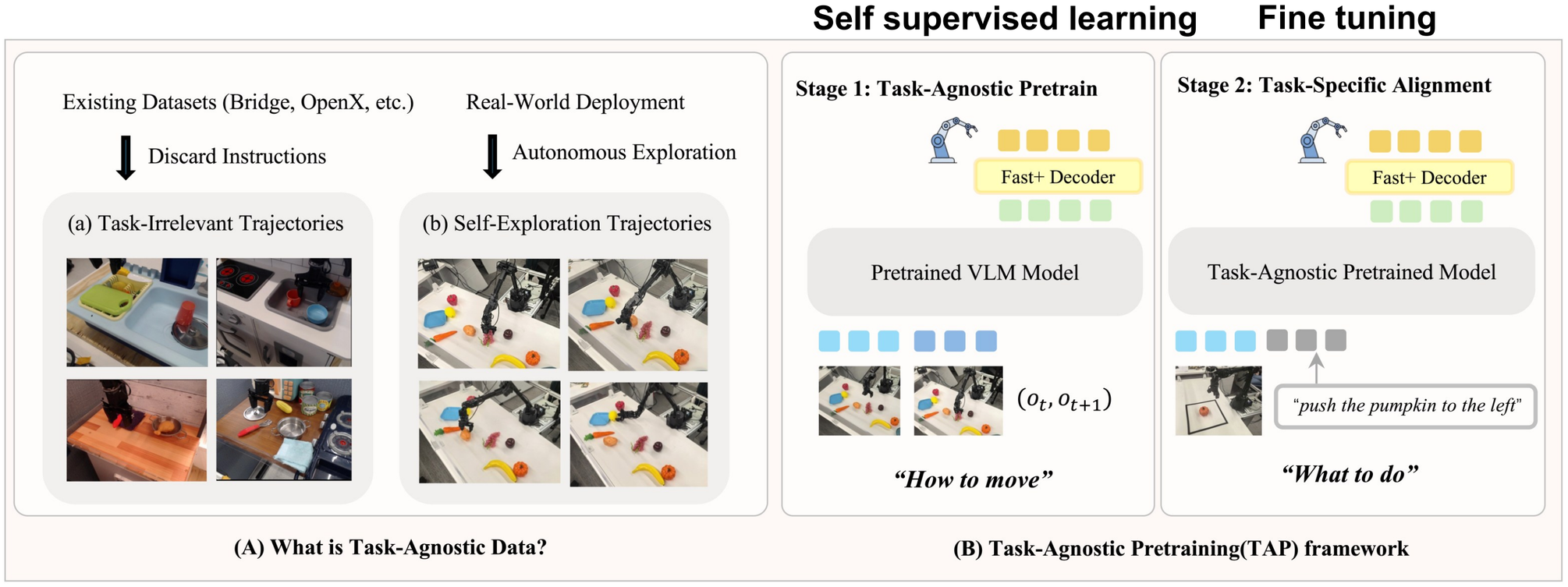
Data-Efficient Learning

- DexMachina: Functional Retargeting for Bimanual Dexterous Manipulation

Hand-Object
Demonstration

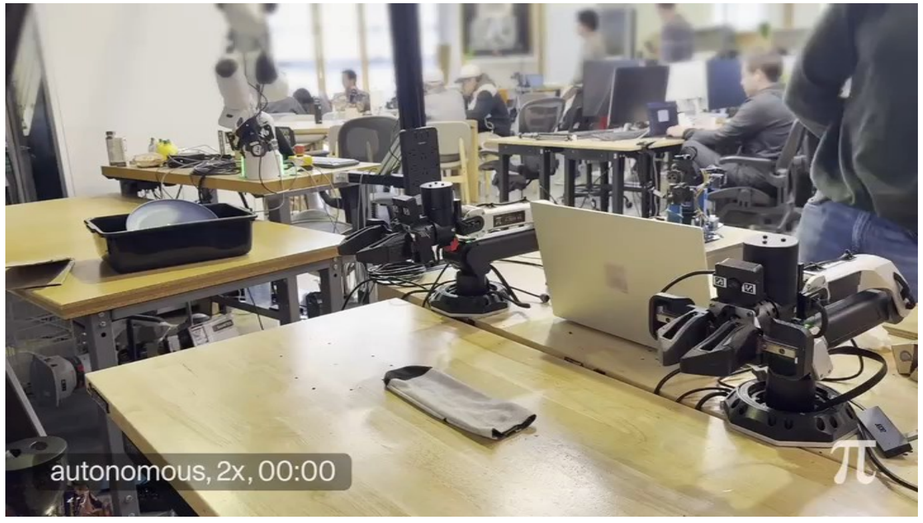
Data-Efficient Learning

- Learning to Move Before Learning to Do: Task-Agnostic Pretraining for VLAs
  - Use failed trials and random play—not just expensive expert demonstrations—as physical pretraining data



How to move: physical/kinematic competence
What to do: language–task alignment

Human Ingenuity from Data Collection

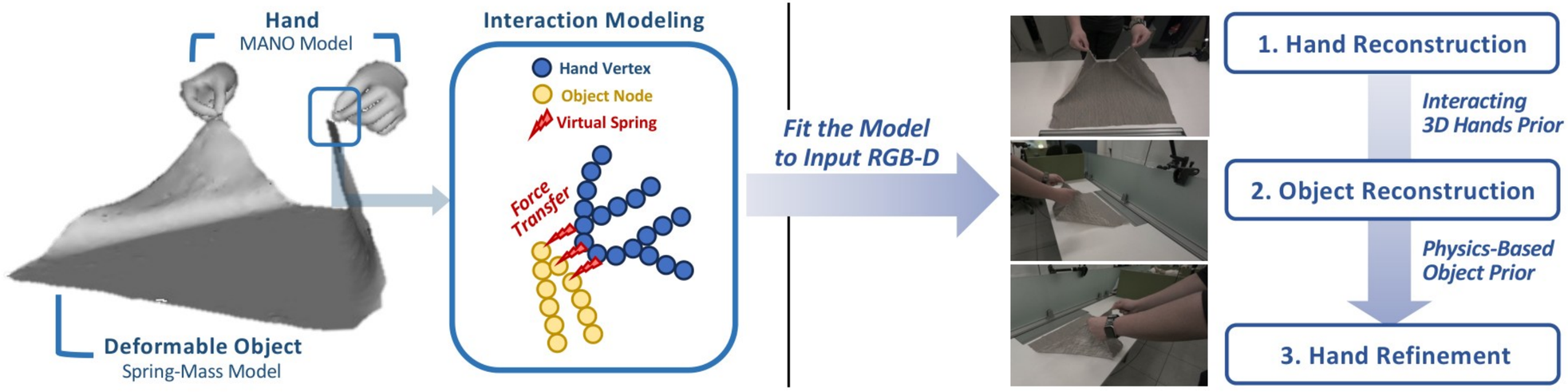


Wrapping sock on gripper to invert it

Mashing plastic bag on the table to open it

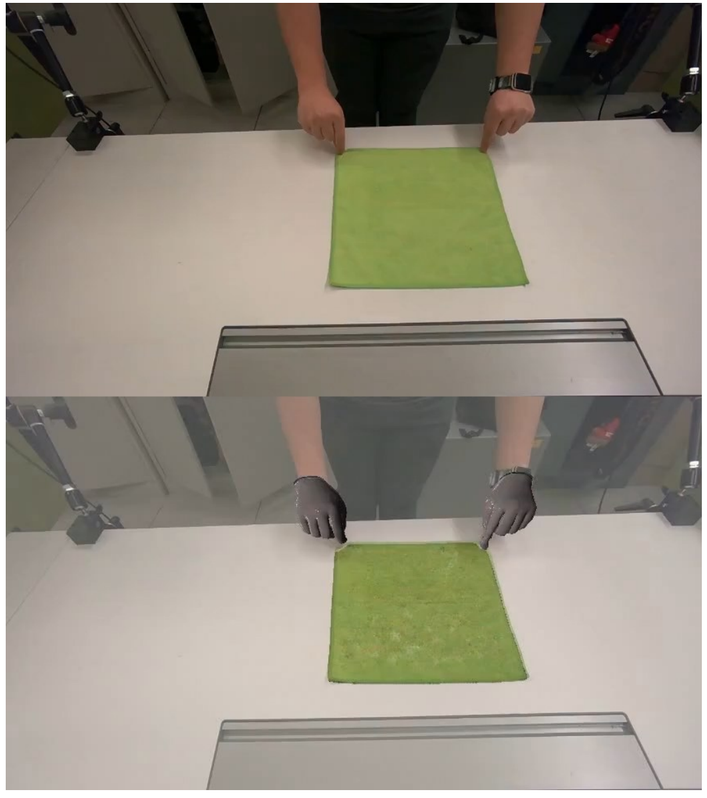
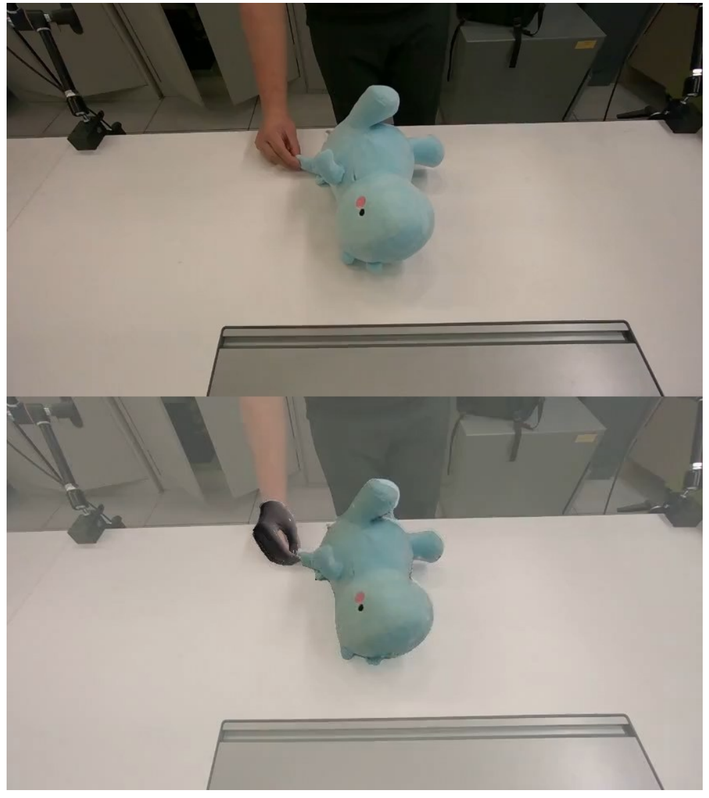
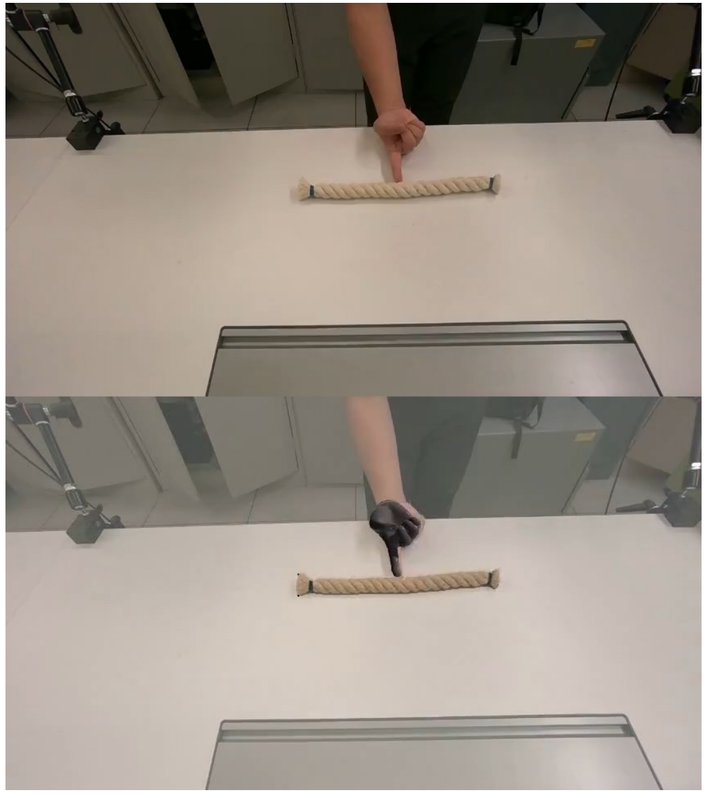
Physics in the loop

- **PhysHanDI**: Physics-Based Reconstruction of Hand-Deformable Object Interactions
  - Estimate 3D hand motion from video
  - Run deformable-object simulations using hand-applied forces
  - Generate physically plausible object deformations
  - Use inverse physics to refine the hand pose



Physics in the loop

- **PhysHanDI**: Physics-Based Reconstruction of Hand-Deformable Object Interactions

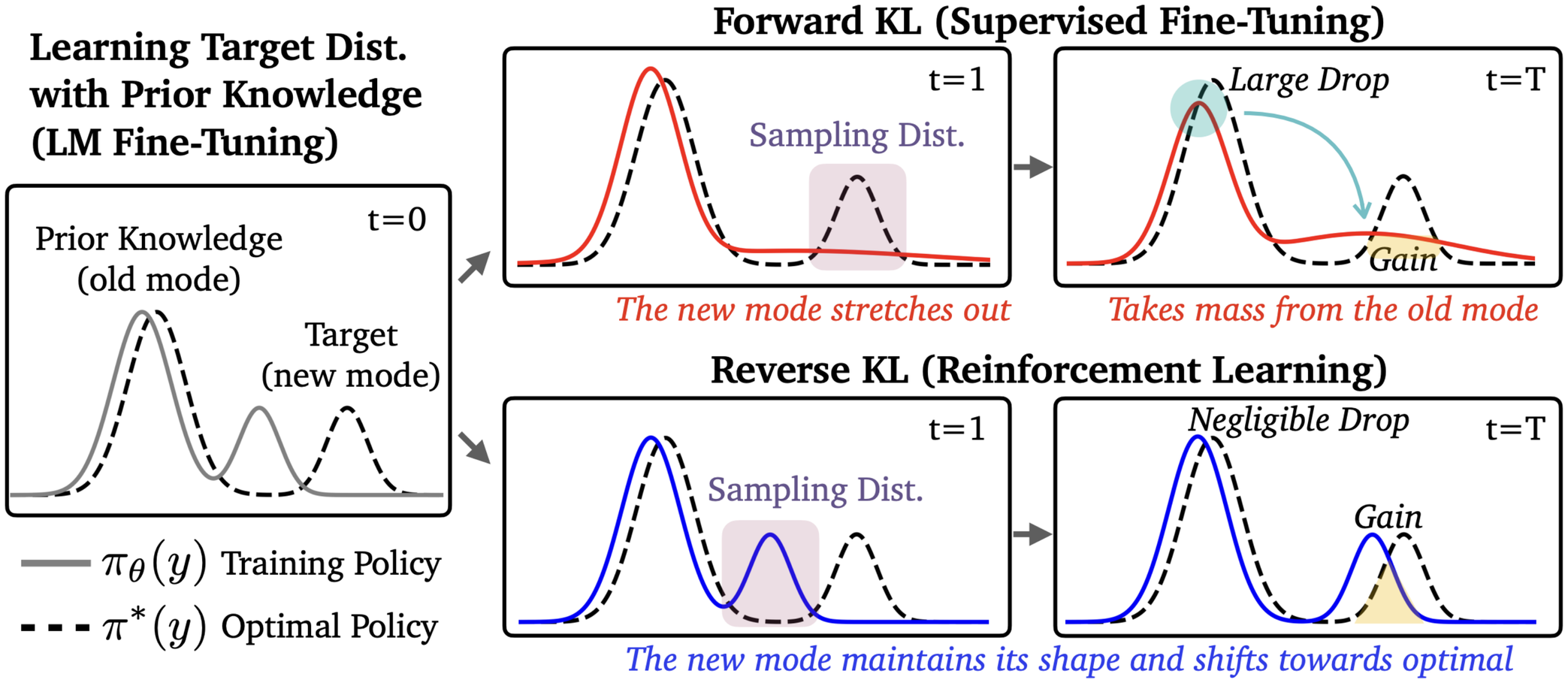




Key Issues in Robotics Research

- Data-Driven **vs**. Theory-Driven **Improvements: Which Is Better?**
- RL **vs.** SFT**: Which Is Better?**
- RL Limitations: How Can **We Address Them?**
  - Limitation 1: Reward Design Is Difficult
  - Limitation 2: Rewards Are Scalar
  - Limitation 3: Actor Distributions Tend to Collapse to a Single Mode
  - Limitation 4: RL Fine-Tuning After IL Leads to Poor Exploration
  - Limitation 5: Long-Horizon Policies Are Hard to Learn
  - Limitation 6: Poor Generalization

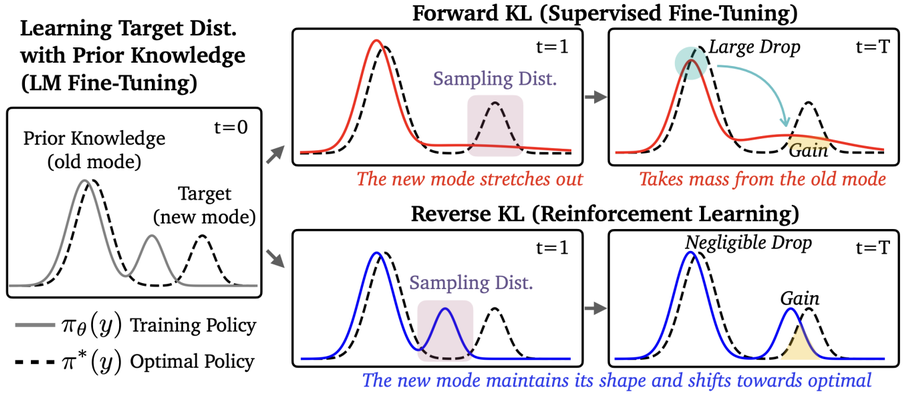
RL vs. SFT: Which Is Better?

- Retaining by Doing: The Role of On-Policy Data in Mitigating Forgetting



RL vs. SFT: Which Is Better?

- Retaining by Doing: The Role of On-Policy Data in Mitigating Forgetting
  - When adapting to the same new task, compare whether **RL or SFT better preserves prior** capabilities
  - RL tends to forget less than SFT while achieving similar or higher target-task performance
  - This is **due to differences in data distribution**
    - SFT: off-policy learning that imitates fixed teacher responses
    - RL: on-policy learning using responses generated by the current model
  - **Whether learning is on-policy is the key factor explaining differences in forgetting**



RL Limitations: How Can **We Address Them?**

Key Issues in Robotics Research

- Data-Driven **vs**. Theory-Driven **Improvements: Which Is Better?**
- RL **vs.** SFT**: Which Is Better?**
- RL Limitations: How Can **We Address Them?**
  - Limitation 1: Reward Design Is Difficult
  - Limitation 2: Rewards Are Scalar
  - Limitation 3: Actor Distributions Tend to Collapse to a Single Mode
  - Limitation 4: RL Fine-Tuning After IL Leads to Poor Exploration
  - Limitation 5: Long-Horizon Policies Are Hard to Learn
  - Limitation 6: Poor Generalization

RL Limitation 1: Reward Design Is Difficult

- Complex human objectives are difficult to specify with hand-crafted rewards
  - Inverse Reinforcement Learning
  - RL from Preference

RL Limitation 1: Reward Design Is Difficult

- Complex human objectives are difficult to specify with hand-crafted rewards
- Relative preferences can address this to some extent

RL Limitation 1: Reward Design Is Difficult

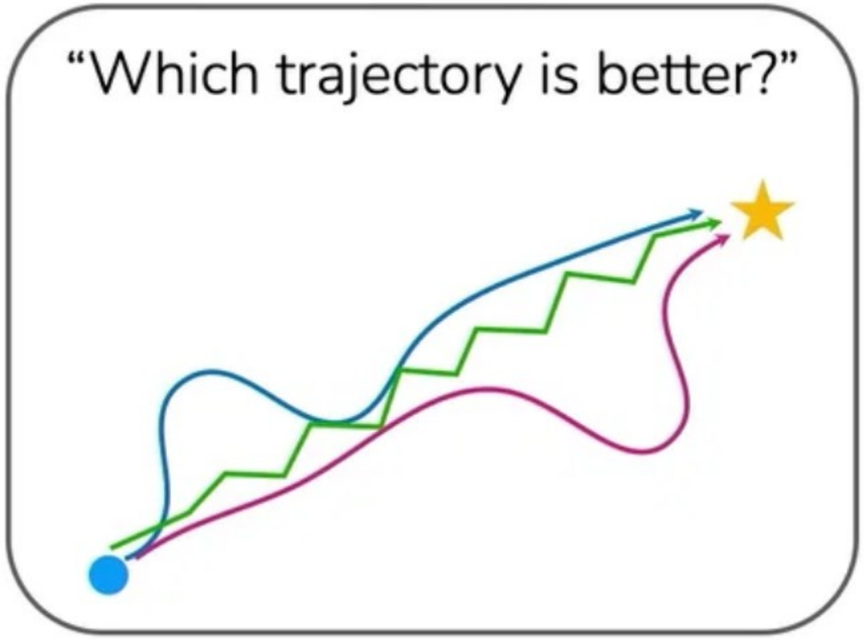
- **Learn rewards indirectly from human preferences! (PbRL)**
  - From Reward-Free Representations to Preferences: Rethinking Offline Preference-Based Reinforcement Learning
  - Video-Based Optimal Transport for Feedback-Efficient Offline Preference-Based Reinforcement Learning
  - PAWS: Preference Learning with Advantage-Weighted Segments
  - Rethinking the Hardness of PbRL: A Provable General Regret Bound

RL Limitation 1: Reward Design Is Difficult

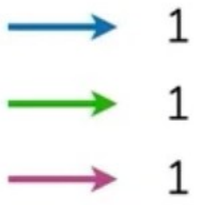
- **Learn the policy directly from human preferences! (No Explicit Reward Model)**
  - TUR-DPO: Topology- and Uncertainty-Aware Direct Preference Optimization
  - Task-Aware Preference Calibration for Direct Preference Optimization
  - Layer-wise Gradient Disentanglement: Decoupling Semantics and Preferences in Direct Preference Optimization
  - Distributed Direct Preference Optimization
  - Spurious Correlation Learning in Preference Optimization: Mechanisms, Consequences, and Mitigation via Tie Training
  - Causal Direct Preference Optimization for Distributionally Robust Generative Recommendation

RL Limitation 2: Rewards Are Scalar

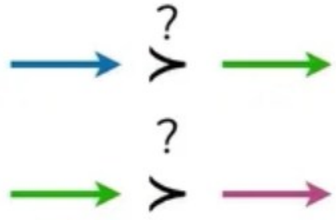
- In standard RL, rewards are sparse and represented as scalars



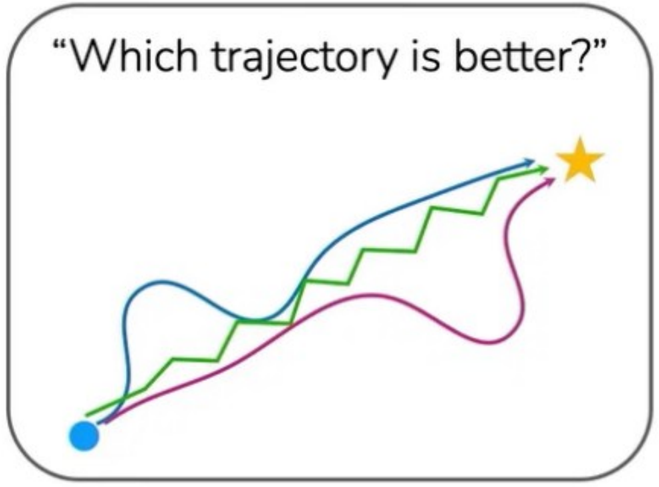
Standard sparse reward



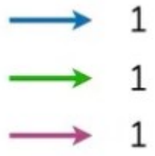
Standard preferences



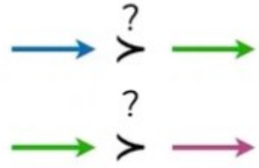
RL Limitation 2: Rewards Are Scalar



Standard sparse reward



Standard preferences



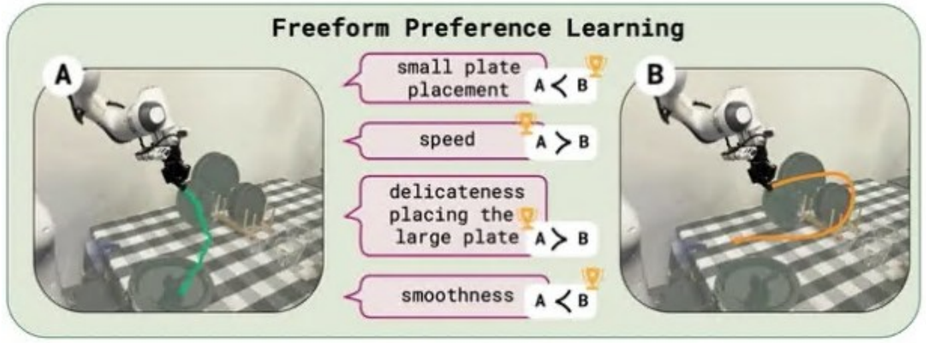
Freeform preferences

smoother

more direct

RL Limitation 2: Rewards Are Scalar

- **Enrich Reward Information with Natural-Language Feedback**



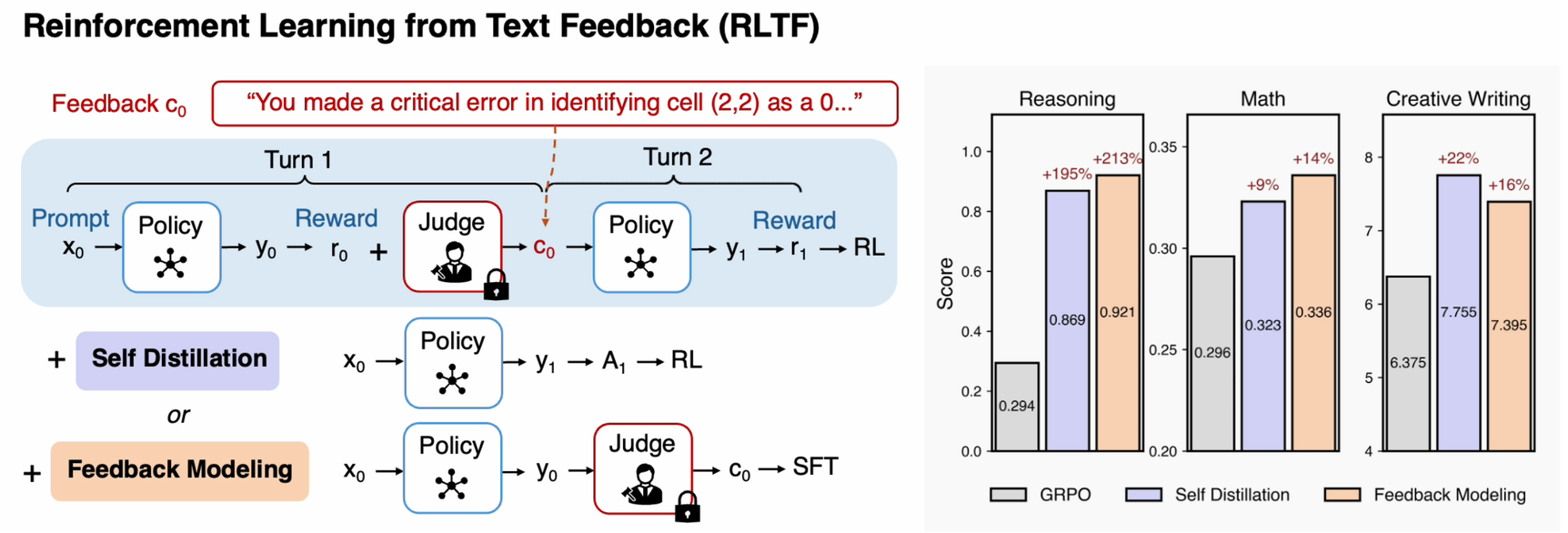
1. Describe axes of feedback (up front or on-the-fly!)
2. Label preferred trajectory along each axis.

Natural language text feedback

- can capture and describe many dimensions
- easier to specify *correctly* than single reward or preference
- can naturally express both coarse and detailed feedback

RL Limitation 2: Rewards Are Scalar

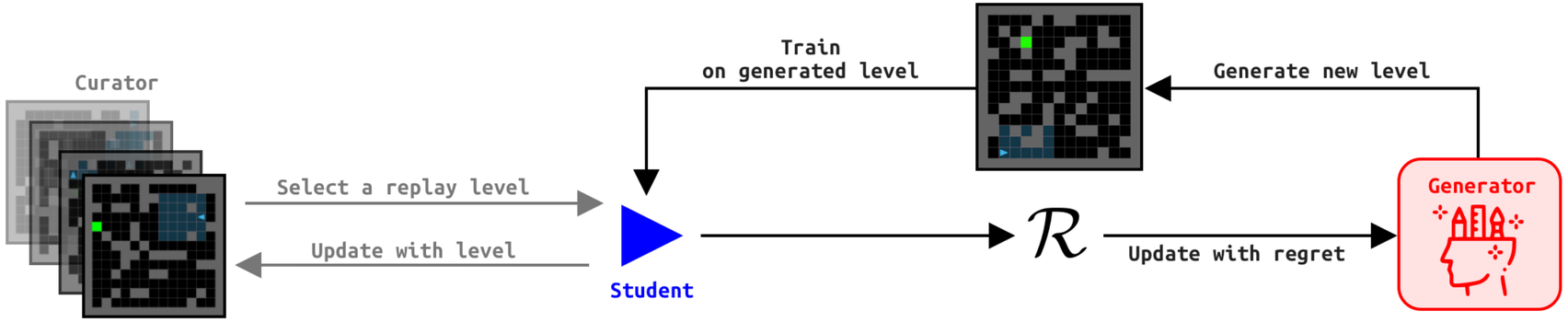
- **Enrich Reward Information with Natural-Language Feedback**
  - Expanding the Capabilities of Reinforcement Learning via Text Feedback
    - Standard RLVR receives only one bit of correctness feedback—1 or 0—for an entire rollout
    - The paper introduces text feedback as an intermediate form between scalar rewards and full demonstrations
    - For example, instead of simply giving reward = 0, provide feedback *such as: "The approach is correct, but independence was incorrectly assumed* in the third step."

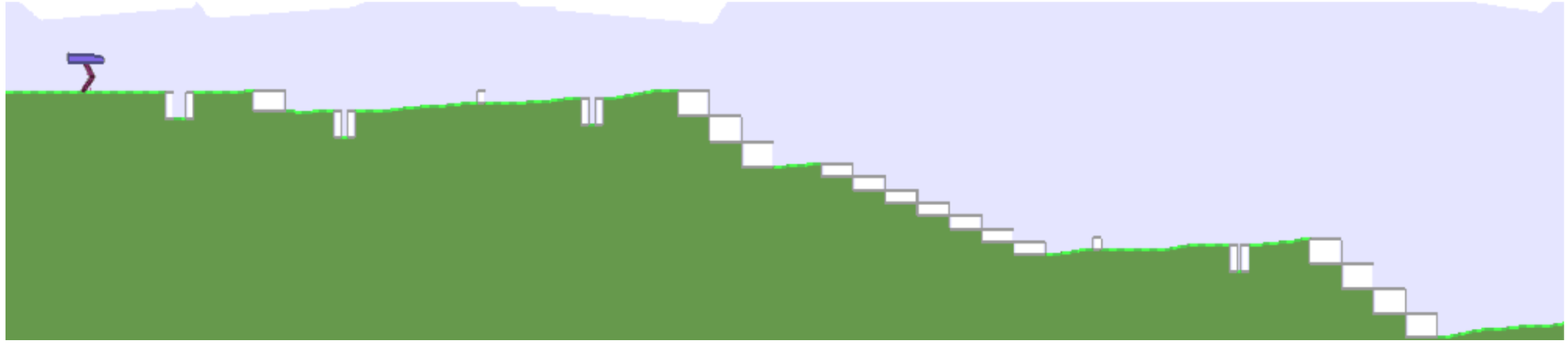


RL Limitation 3: Actor Mode Collapse

- Actors trained with recent RL algorithms often have highly unimodal distributions
  - E.g., Evolving Curricula with Regret-Based Environment Design (ICML 22)

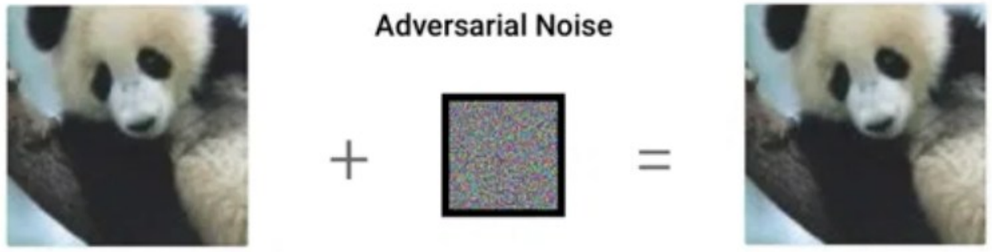
https://accelagent.github.io/





RL Limitation 3: Actor Mode Collapse

$$\text{learn } \hat{f}_\theta(\mathbf{x}) \leftarrow \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} \|\hat{f}_\theta(\mathbf{x}_i) - y_i\|^2 \qquad \mathbf{x}^\star = \arg\max_{\mathbf{x}} \hat{f}_\theta(\mathbf{x})$$

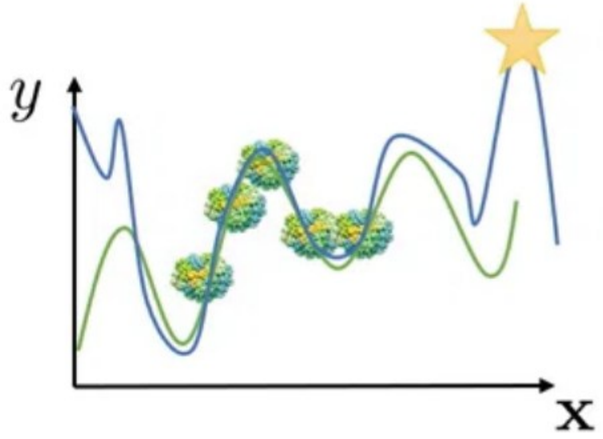


optimizing over the inputs to a neural network can yield almost any output

more formally: $\hat{f}_\theta(\mathbf{x})$ is correct for $\mathbf{x} \sim p(\mathbf{x})$

$\hat{f}_\theta(\mathbf{x})$ has no reason to be correct for *out of distribution* $\mathbf{x}$

a version of this challenge shows up in *every* data-driven decision-making problem



RL Limitation 3: Actor Mode Collapse

- Distribution shift also occurs in RL…

Q-learning:

$$Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \max_{\mathbf{a}'} Q(\mathbf{s}', \mathbf{a}')$$

actor-critic:

$$Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + E_{\mathbf{a}' \sim \pi}[Q(\mathbf{s}', \mathbf{a}')]$$

$$\pi = \arg\max_{\pi} E_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s}, \mathbf{a})]$$

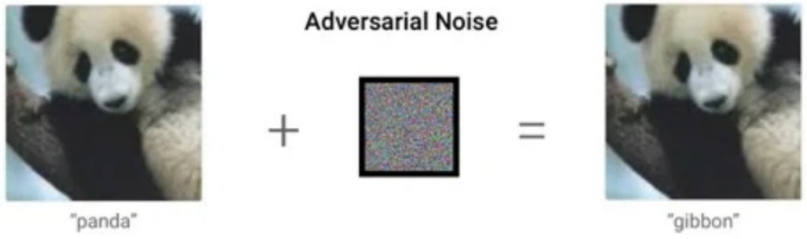
we have *distributional shift* when $\pi_\beta(\mathbf{a}|\mathbf{s}) \neq \pi(\mathbf{a}|\mathbf{s})$

what is the objective?

$$\min_Q E_{(\mathbf{s},\mathbf{a}) \sim \pi_\beta(\mathbf{s},\mathbf{a})} \left[(Q(\mathbf{s}, \mathbf{a}) - y(\mathbf{s}, \mathbf{a}))^2 \right]$$

behavior policy

target value



This is **just like** what we saw before in the simple case, just at **every time step** when learning the Q-function!

Kumar, Fu, Tucker, Levine. **Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction.** NeurIPS '19

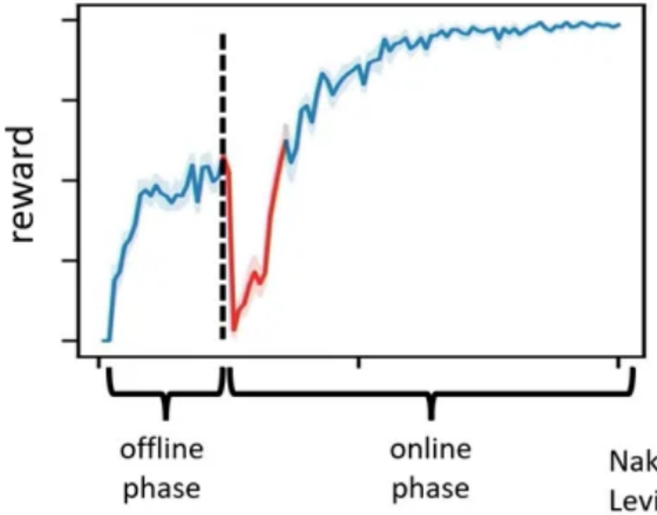
RL Limitation 3: Actor Mode Collapse

- Towards Online RL

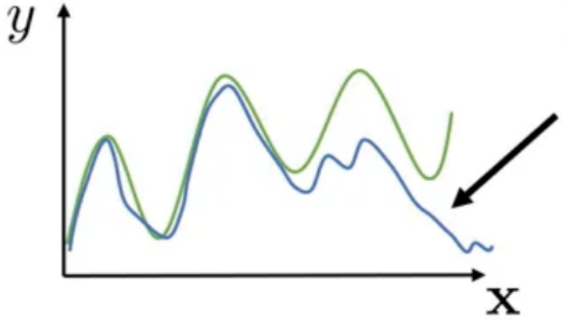
given MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma\}$
and $\mathcal{D} = \{(s_i, a_i, s'_i, r_i)\}_{i=1}^n$
and interactions with $\mathcal{M}$

offline-to-online RL ← sequential decision

given MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma$
and $\mathcal{D} = \{(s_i, a_i, s'_i, r_i)\}_{i=1}^n$
find policy $\pi : s \mapsto a$
with largest total reward



why does this happen?



It's very hard to figure out a method that trains value functions **both offline and online**

conservative value functions have high OOD error

when we start collecting online data, these points might no longer be OOD!

Nakamoto, Zhai, Singh, Mark, Ma, Finn, Kumar, Levine. **Cal-QL: Calibrated Offline RL Pre-Training for Efficient Online Fine-Tuning.** 2023

RL Limitation 3: Actor Mode Collapse

- Model the actor distribution with diffusion/flow matching

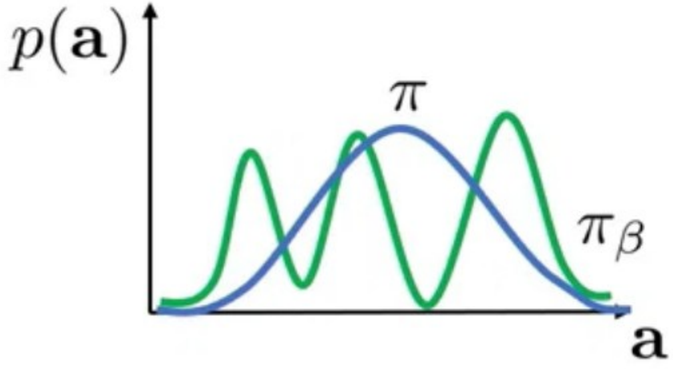
$$Q(\mathbf{s},\mathbf{a}) \leftarrow r(\mathbf{s},\mathbf{a}) + E_{\mathbf{a}'\sim\pi}[Q(\mathbf{s}',\mathbf{a}')]$$

$$\pi = \arg\max_\pi E_{\mathbf{a}\sim\pi(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})] \;\text{ s.t. }\; D(\pi,\pi_\beta) \le \epsilon$$

$$\min_Q E_{(\mathbf{s},\mathbf{a})\sim\pi_\beta(\mathbf{s},\mathbf{a})}\left[(Q(\mathbf{s},\mathbf{a}) - y(\mathbf{s},\mathbf{a}))^2\right]$$

"policy constraint method"

Problem: it can be very hard to make this work while avoiding "adversarial" actions



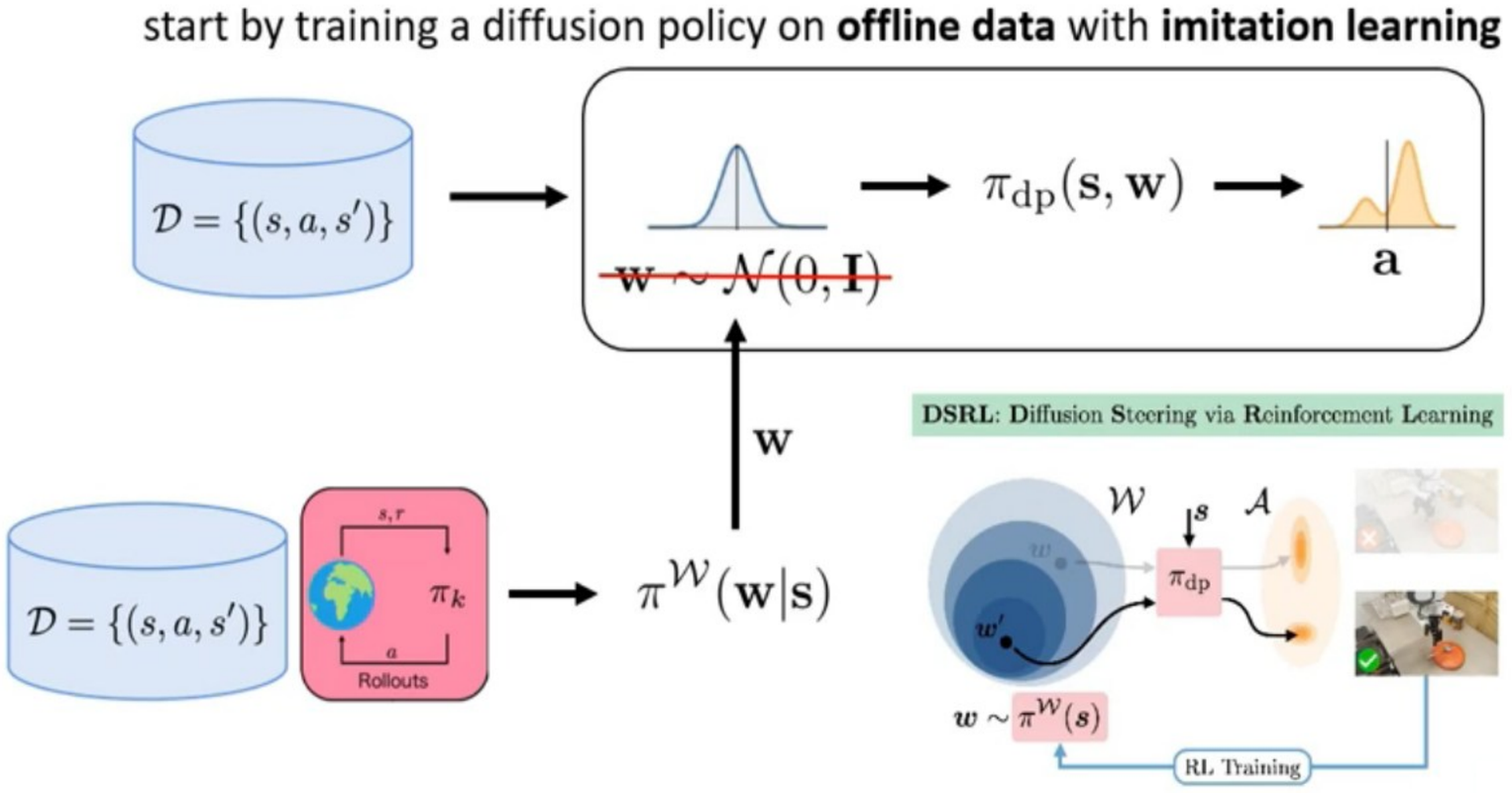
(surprising) key idea: use a very powerful policy class (diffusion, flow matching)

problem: RL with diffusion model actors is difficult
because it's hard to get gradients w.r.t. model likelihoods

if we can train very expressive actors (e.g., diffusion), policy constraints can work well **offline** and **online**

RL Limitation 3: Actor Mode Collapse

- Replace Gaussian actor distributions with conditional diffusion
  - Steering Your Diffusion Policy with Latent Space Reinforcement Learning (CoRL 25)

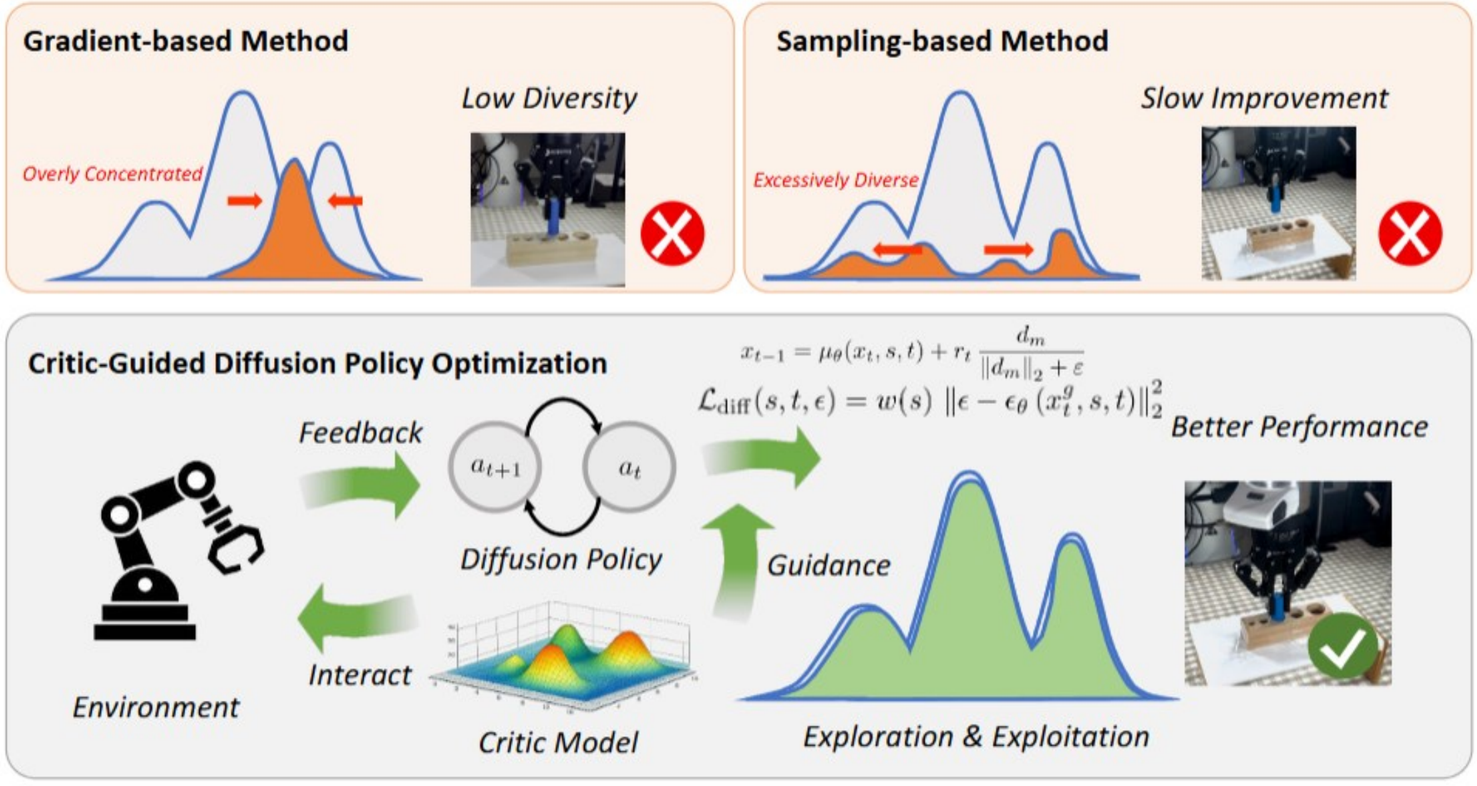


RL Limitation 3: Actor Mode Collapse

- Does conditional diffusion alone completely solve mode collapse? No.
- Even if a conditional diffusion actor initially represents multiple modes, maximizing expected return with RL can concentrate probability on the highest-reward mode
- Because Gaussian actions cannot easily represent multimodal distributions

RL Limitation 3: Actor Mode Collapse

- Representing Multimodal Distributions
  - Sample-Efficient Diffusion-based Reinforcement Learning with Critic Guidance



RL Limitation 4: Poor Exploration After IL

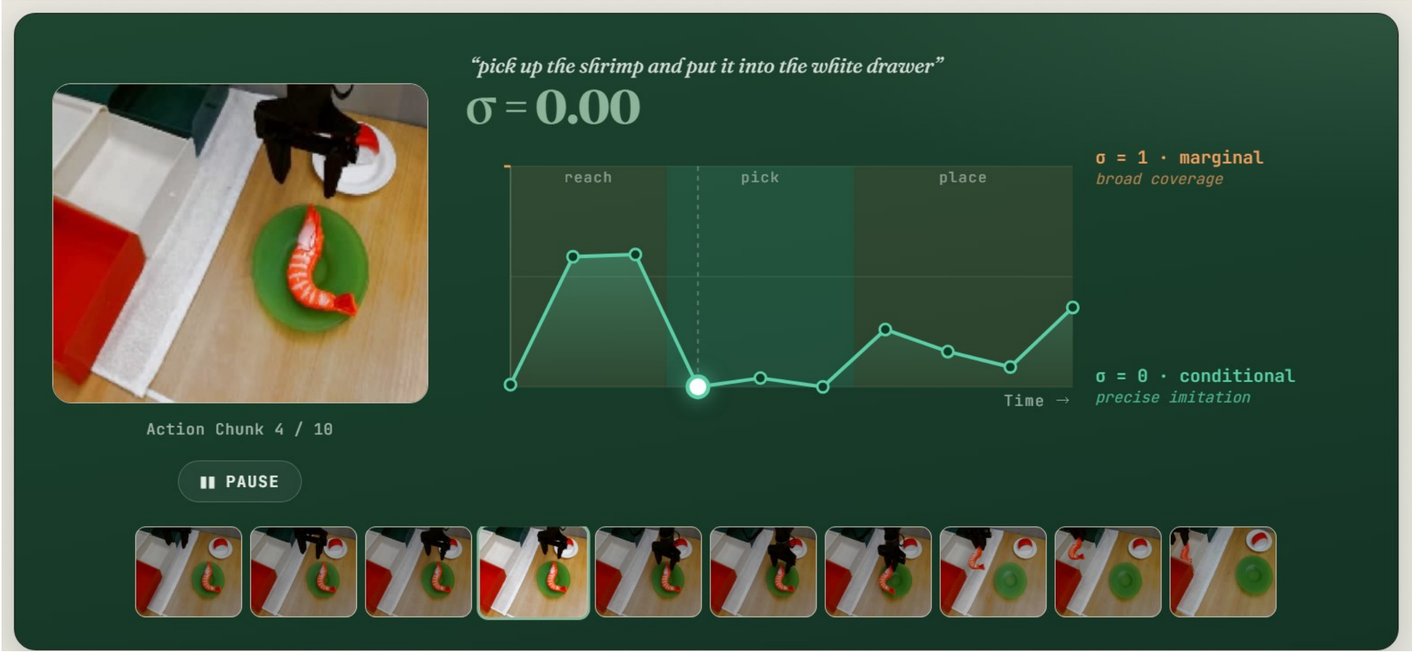
- A behavior-cloned policy has an action distribution concentrated around demonstrations, so subsequent RL fine-tuning cannot explore sufficiently novel actions

RL Limitation 4: Poor Exploration After IL

- TMRL: Diffusion Timestep-Modulated Pretraining Enables Exploration for Efficient Policy Finetuning (RSS 2026)

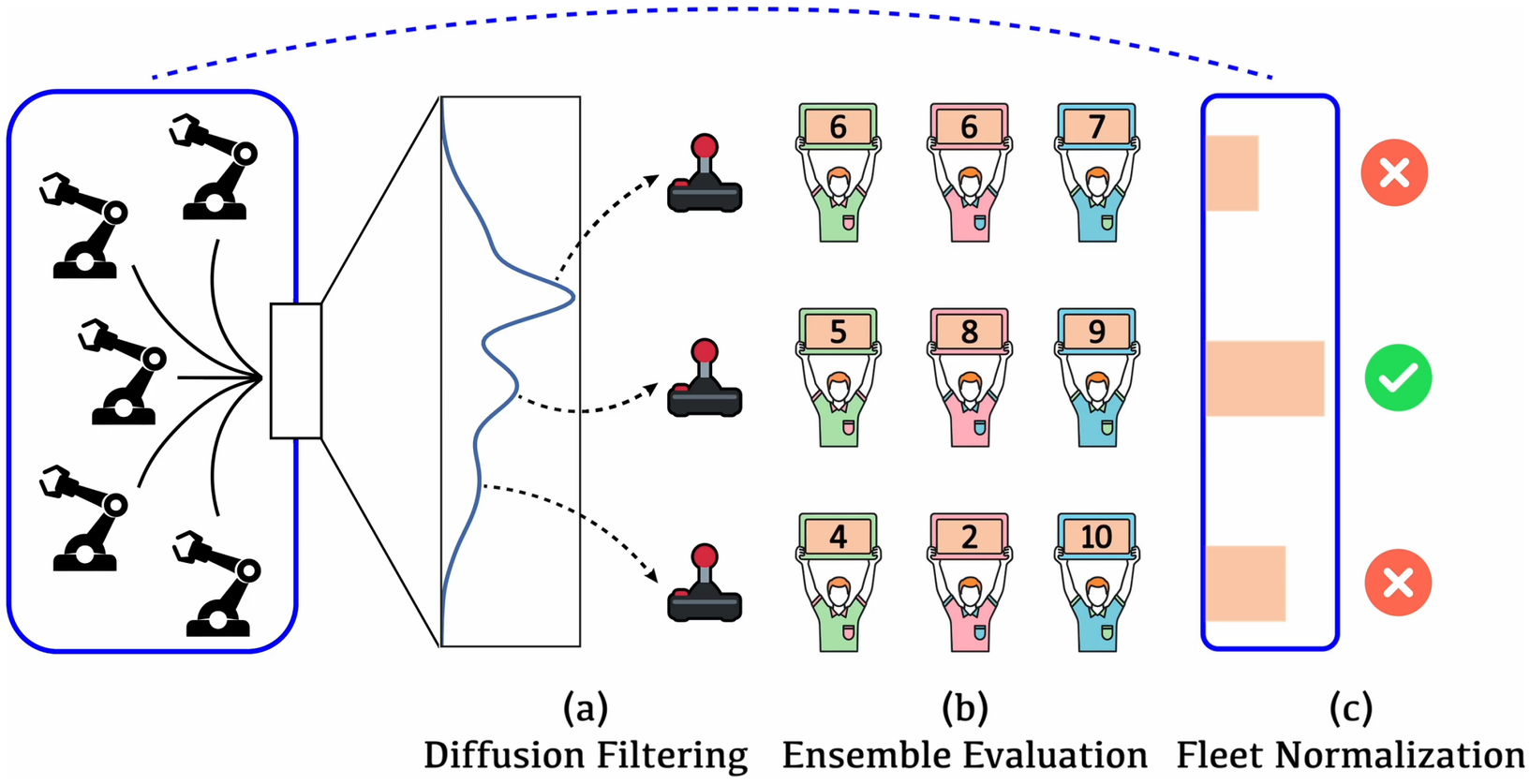
RL Limitation 4: Poor Exploration After IL

- TMRL: Diffusion Timestep-Modulated Pretraining Enables Exploration for Efficient Policy Finetuning (RSS 2026)
  - Two Proposed Methods
    - CSP (Context-Smoothed Pretraining): add diffusion noise to observations/contexts during pretraining so the policy learns distributions ranging from narrow BC behavior to broad action distributions
    - TMRL: directly control the diffusion timestep during RL to learn when to explore broadly and when to act precisely



RL Limitation 4: Poor Exploration After IL

- DF-ExpEnse: Diffusion Filtered Exploration for Sample Efficient Finetuning
  (shuran song)

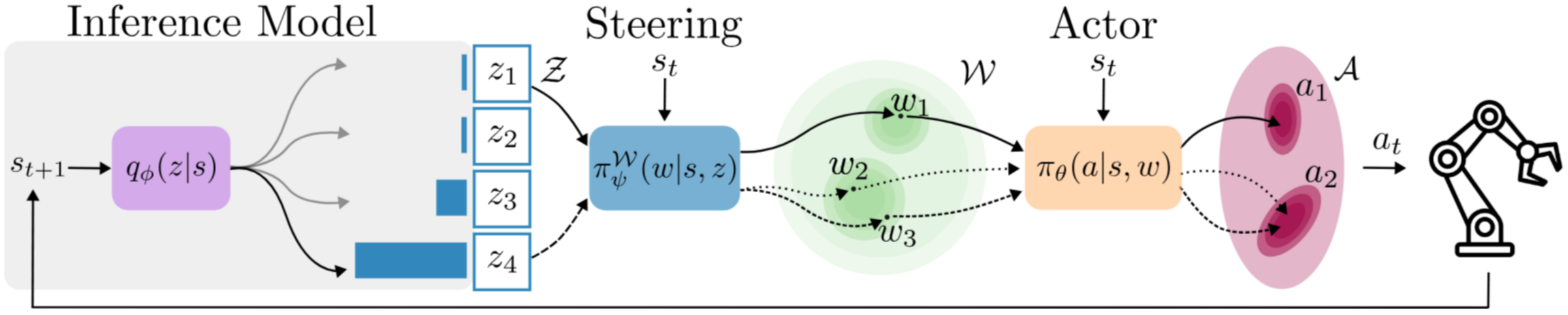


RL Limitation 4: Poor Exploration After IL

- DF-ExpEnse: Diffusion Filtered Exploration for Sample Efficient Finetuning (shuran song)
  - Starting point: a diffusion policy behavior-cloned from offline demonstrations
  - $\pi_{BC}(a|s)$: Fine-tuning the policy with online RL creates two problems
    - First, the BC policy repeatedly samples only near actions seen in demonstrations
    - Second, random exploration in continuous action spaces is too risky or inefficient
  - In short, the method addresses being trapped near the IL policy while avoiding useless data from completely random departures
  - Sample multiple action candidates from the diffusion policy, then use a critic ensemble to estimate each action's performance (Q-value) and uncertainty
  - Build a UCB objective with mean value + uncertainty bonus, then execute the highest-scoring action

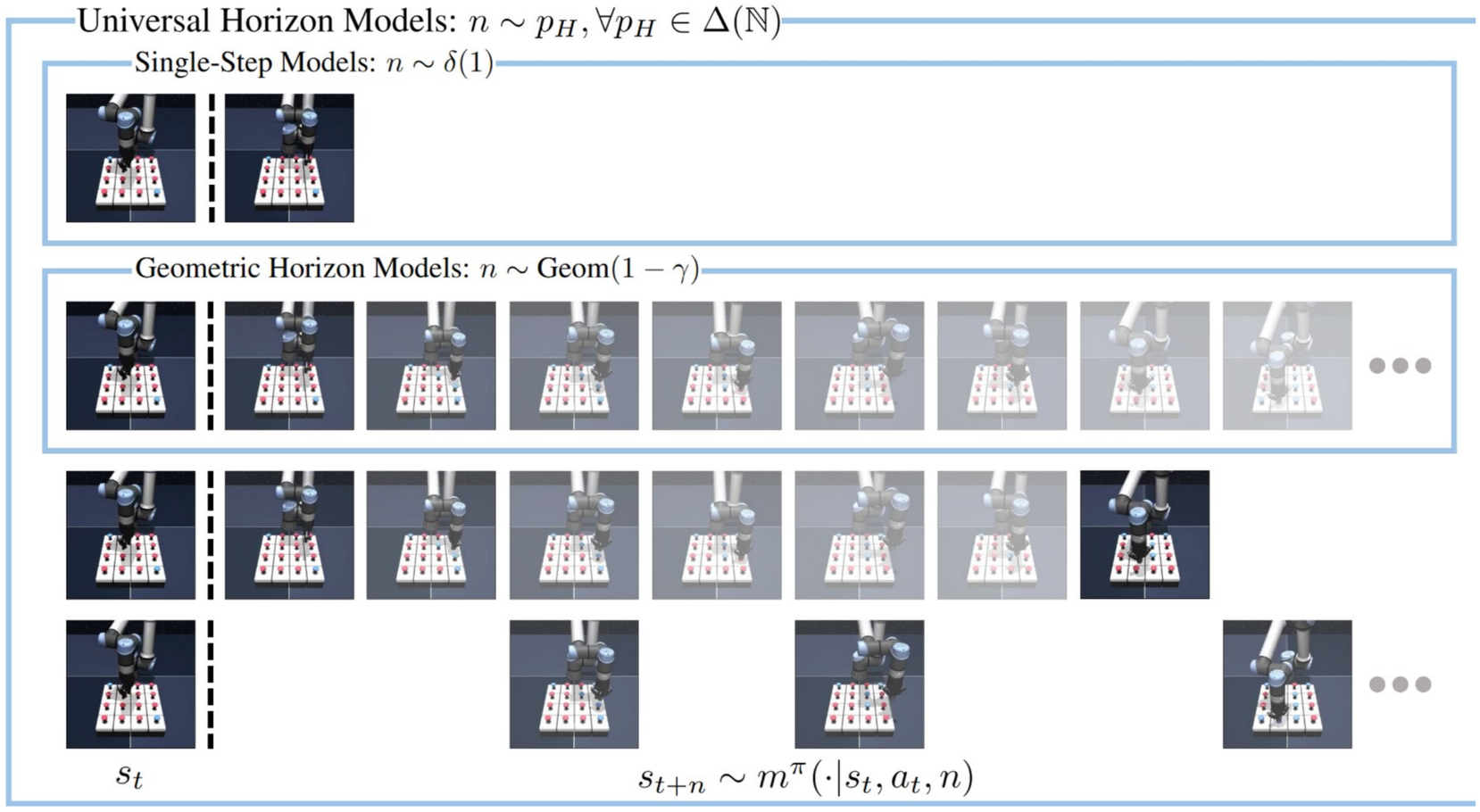
RL Limitation 4: Poor Exploration After IL

- Behavioral Mode Discovery for Fine-tuning Multimodal Generative Policies
  - Fine-tuning a multimodal generative policy such as a diffusion policy with RL improves success rates but can collapse behavioral diversity
  - First discover multiple behavioral modes already latent in the pretrained policy, then add an intrinsic reward during RL fine-tuning to keep the modes distinct



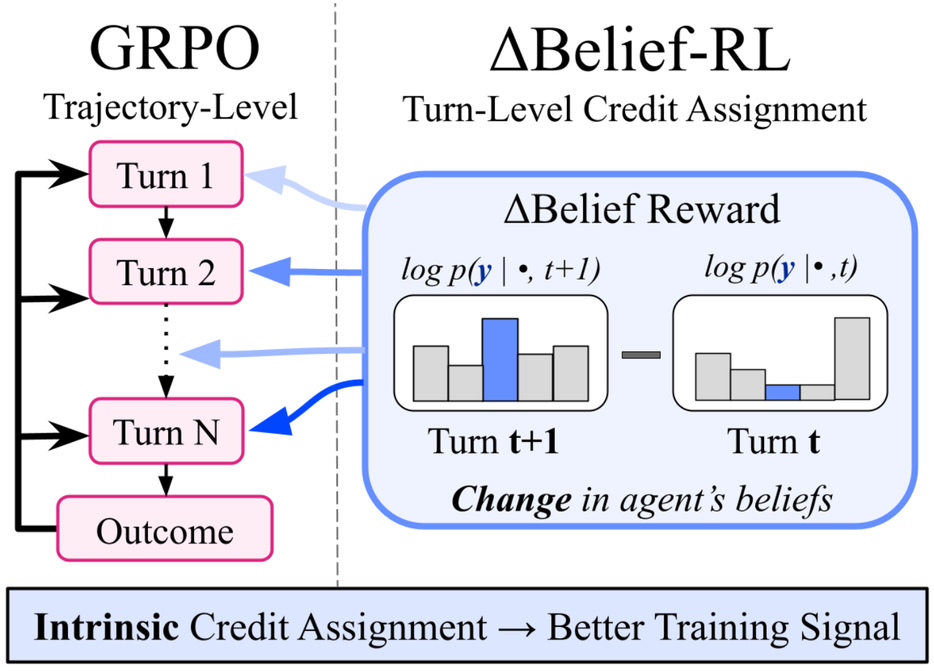
RL Limitation 5: Long-Horizon Learning

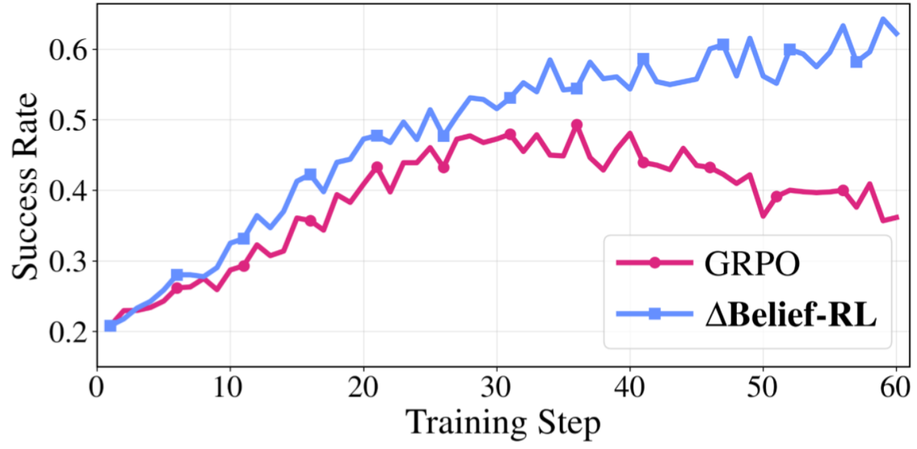
- Directly Improve Long Rollouts in a World Model
  - Offline Reinforcement Learning with Universal Horizon Models



RL Limitation 5: Long-Horizon Learning

- Tackle Temporal Credit Assignment in Long-Horizon Tasks
  - Use Intermediate Progress as Reward
  - Intrinsic Credit Assignment for Long Horizon Interaction
    - Instead of giving only a final-answer reward after a long interaction, reward the step-by-step increase in the agent's belief in the correct answer



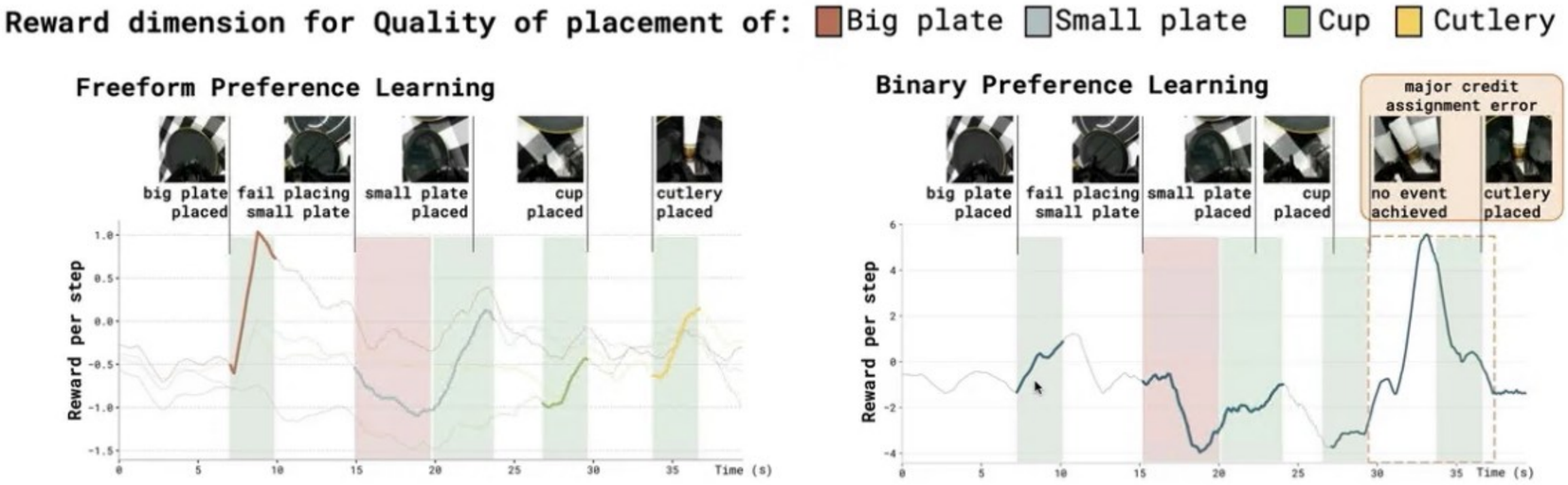


RL Limitation 5: Long-Horizon Learning

- Tackle Temporal Credit Assignment in Long-Horizon Tasks
  - Use Intermediate Progress as Reward

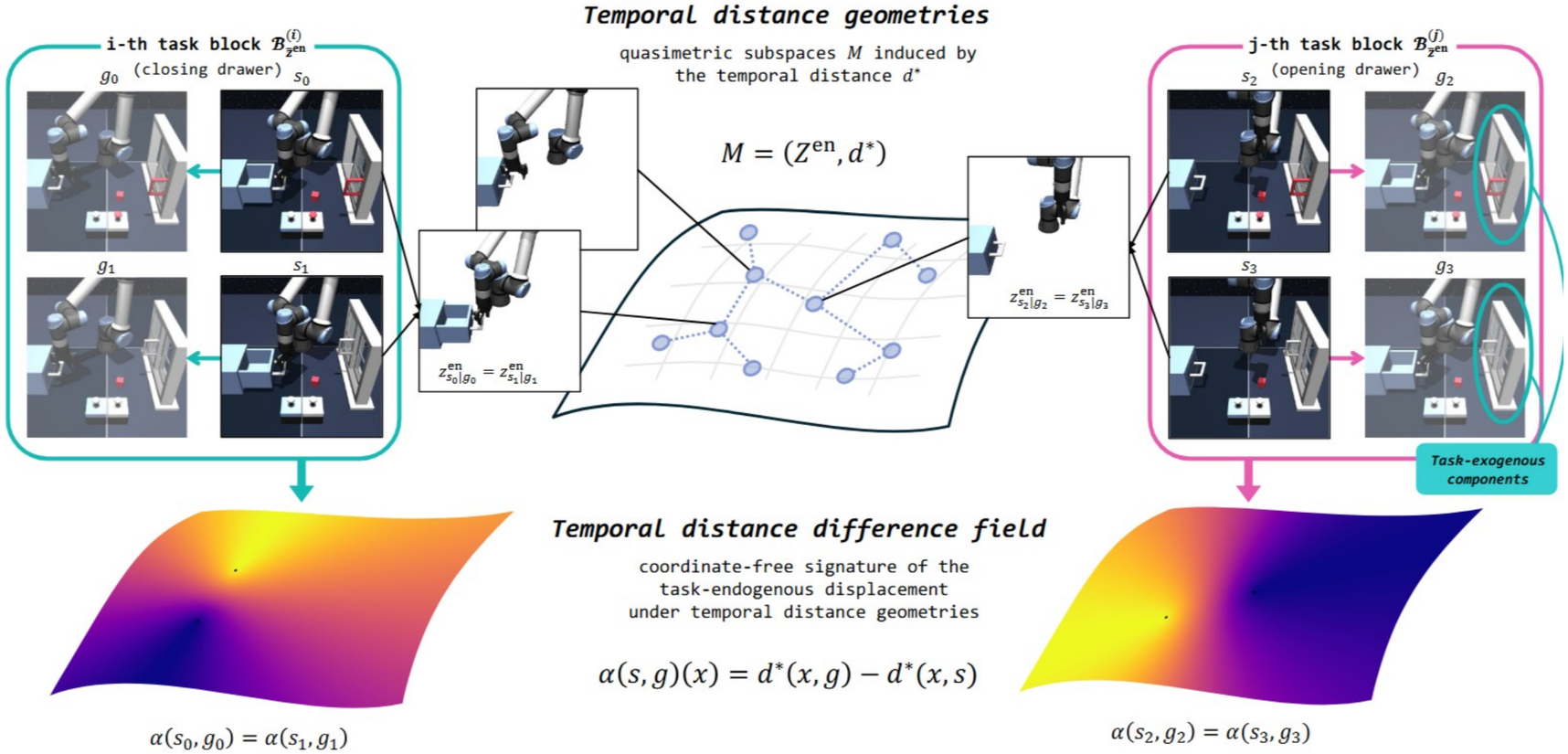
Can freeform preferences enable temporal credit assignment?

- freeform preference axes correspond to subtasks
- preferences provided ***at trajectory level***



RL Limitation 6: Poor Generalization

- Compositional Generalization After Learning Subtasks/Skills
  - Compositional Transduction with Latent Analogies for Offline Goal-Conditioned RL
    - Extract reusable behavioral changes/skills from the training data, then combine them in a new context to perform unseen tasks



Summary

- Data-Driven **vs**. Theory-Driven **Improvements: Which Is Better?**

A: Both matter, but we should prioritize theoretical advances

- RL **vs.** SFT**: Which Is Better?**

A: Results suggest RL is somewhat better (on-policy learning)

- RL Limitations: How Can **We Address Them?**
  - Limitation 1: Reward Design Is Difficult
  - Limitation 2: Rewards Are Scalar
  - Limitation 3: Actor Distributions Tend to Collapse to a Single Mode
  - Limitation 4: RL Fine-Tuning After IL Leads to Poor Exploration
  - Limitation 5: Long-Horizon Policies Are Hard to Learn
  - Limitation 6: Poor Generalization

A: RL is young, leaving ample room for impactful research