HUMAN VIDEOS ARE NOT JUST FOR IMITATION

Robot Self-Improvement via Human-Video Dynamics Models

PREPRINT 2026

Hanzhi Chen[*,1,2,4], Anran Zhang[*,1,2,4], Simon Schaefer[1,2,4], Kejia Chen[2], Shi Chen[1], Daniel Cremers[2,4], Oier Mees[†,3,1], Stefan Leutenegger[†,1]

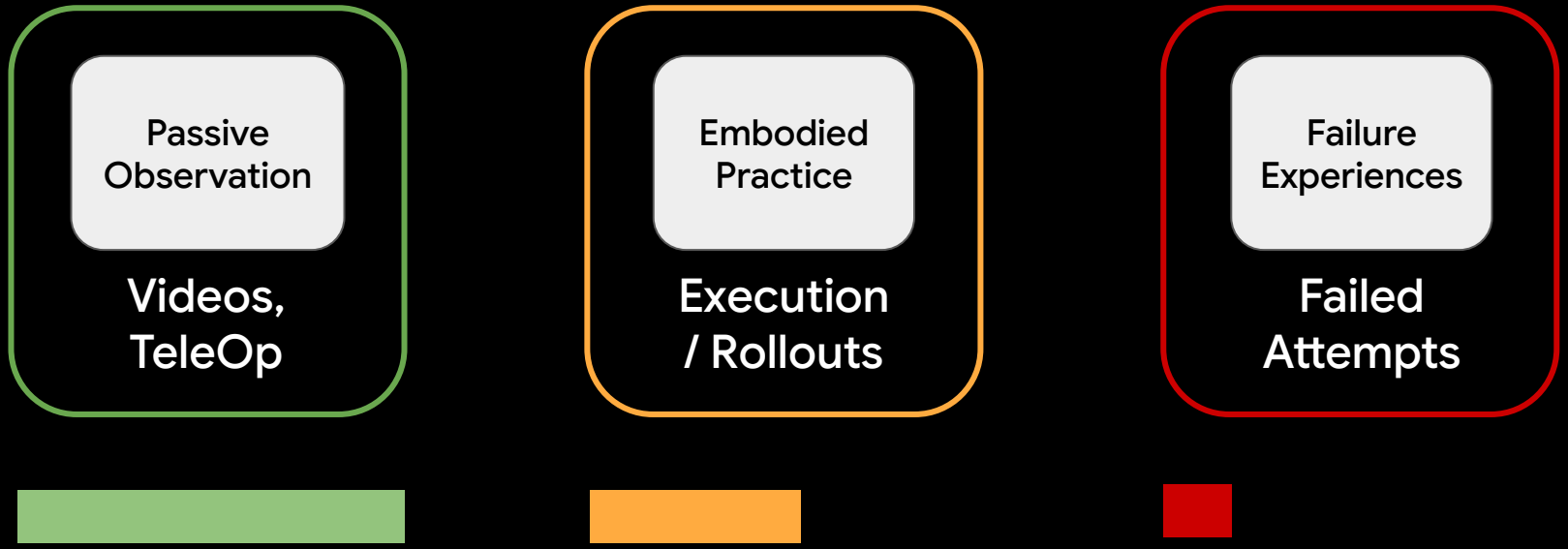
[*]Equal Contribution [†]Equal Advising

[1]ETH Zürich [2]Technical University of Munich [3]Microsoft [4]MCML

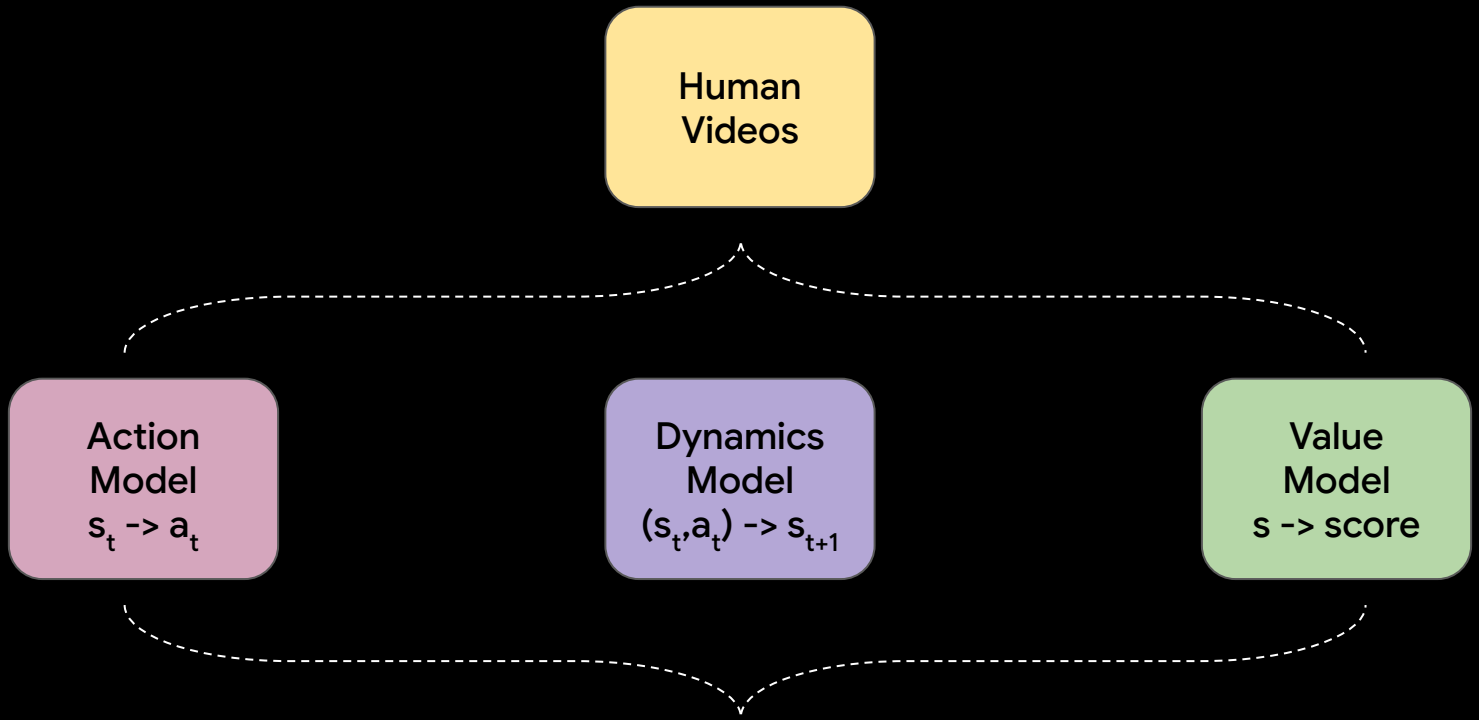
arXiv | Video | Code (Coming Soon)

Presented by: Jishnu Jaykumar P | 9/4/26
Reading Group | IRVL @ UTD

Acquire skills from the kinds of data that humans learn from



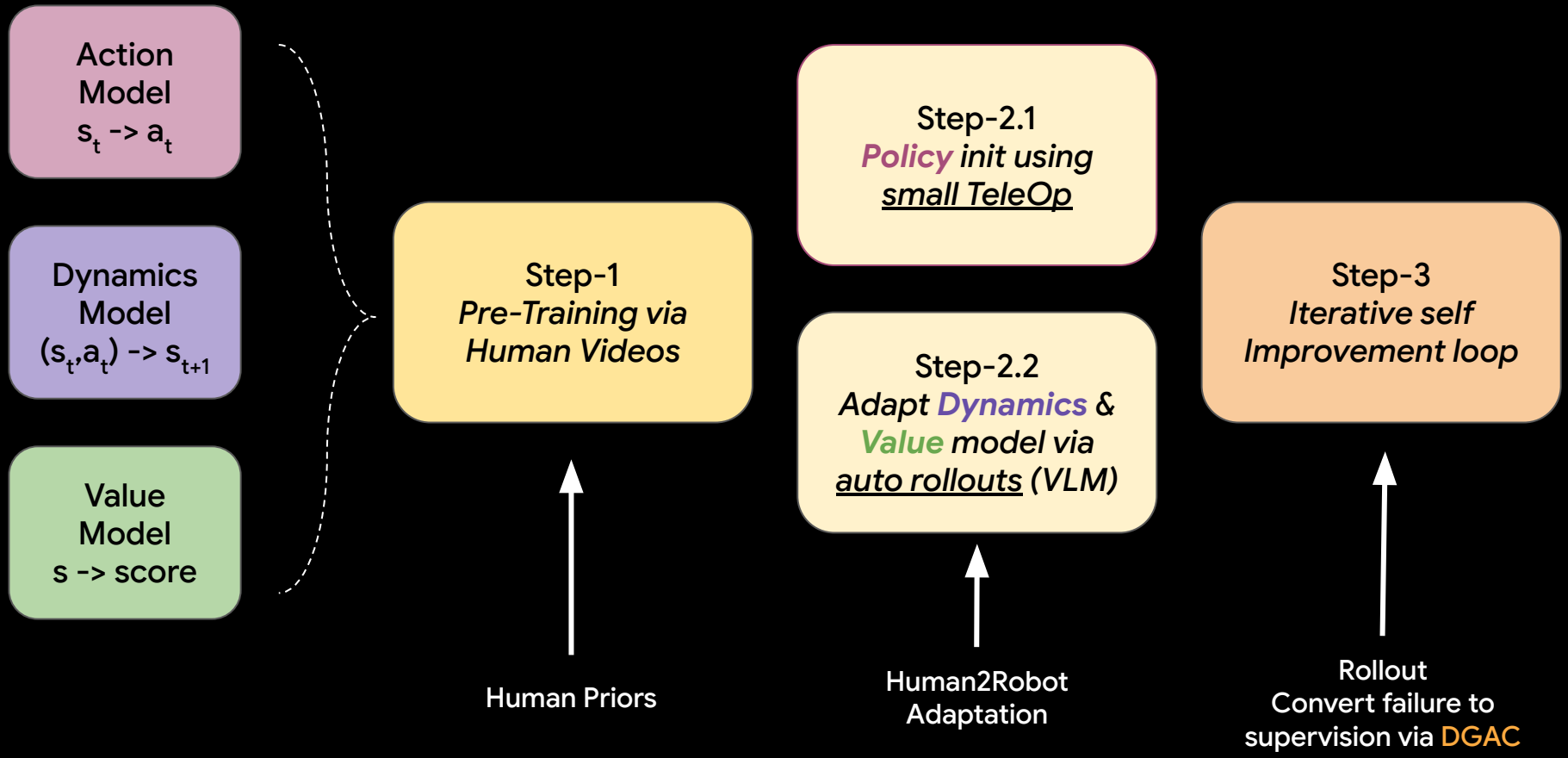
This work



Claim: Representation enables X-embodiment

Goal: Autonomously improve from own rollouts and failures

Pipeline


Action Model
s_t -> a_t
Dynamics Model
(s_t,a_t) -> s_{t+1}
Value Model
s -> score
Step-1
Pre-Training via Human Videos
Step-2.1
Policy init using small TeleOp
Step-2.2
Adapt Dynamics & Value model via auto rollouts (VLM)
Step-3
Iterative self Improvement loop
Human Priors
Human2Robot Adaptation
Rollout
Convert failure to supervision via DGAC

Step-1
Pre-Training using Human Videos

Step-1: Pre-Training using Human Videos

Action Model

Dyn Model

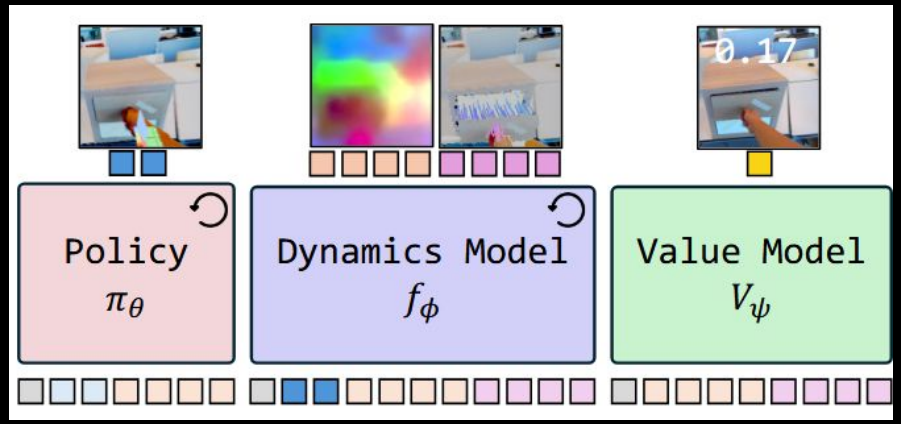
Value Model

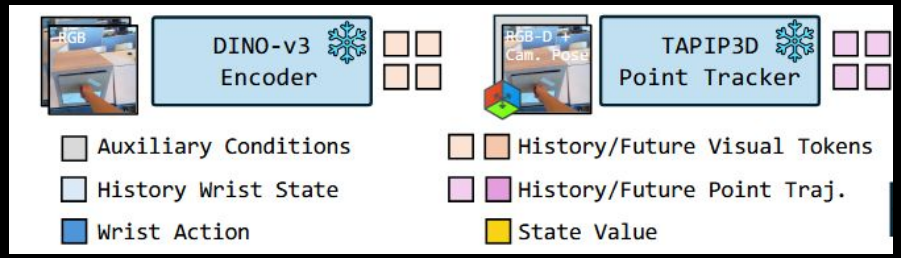
$a_t = [\xi_t, c_t]$

absolute 6-DoF wrist motion and $c_t \in [0, 1]$

Visual Tokens
Point Traj.

State Value





Policy Model: $\hat{\boldsymbol{a}}_{t:t+H-1} = \pi_\theta(\boldsymbol{s}_{t-H'+1:t}, \boldsymbol{z}_{t-H'+1:t}; \boldsymbol{m}_t^\pi)$,

$$\mathcal{L}_\pi = \left\| \hat{\boldsymbol{a}}^\tau_{t:t+H-1} - \boldsymbol{a}_{t:t+H-1} \right\|_2^2$$

Dynamics Model: $\hat{\boldsymbol{o}}_{t+H-1} = f_\phi(\boldsymbol{o}_{t-H'+1:t}, \boldsymbol{a}_{t:t+H-1}; \boldsymbol{m}_t^f)$,

$$\mathcal{L}_f = \left\| \hat{\boldsymbol{o}}^\tau_{t+H-1} - \boldsymbol{o}_{t+H-1} \right\|_2^2$$

Value Model: $\hat{V}_t = V_\psi(\boldsymbol{o}_t; \boldsymbol{m}_t^V)$.

$$\mathcal{L}_V = \left\| \hat{V}_t - V_t \right\|_2^2$$

$\sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} = \gamma^{T-t} r_T$, where γ is the discount factor, T is the terminal timestep, all intermediate rewards are 0, and the terminal reward $r_T \in \{1, -1\}$ denotes success or failure.

$\boldsymbol{m}_t^\pi$, $\boldsymbol{m}_t^f$, and $\boldsymbol{m}_t^V$

Aux conditions.
E.g. language
Human Datasets:
HOI4D, Arti4D, EgoDex

Step-2
Human 2 Robot Adaptation

2.1 Policy
2.2 Dynamics and Value

Step-2.1: Policy init

Action Model Π_{ref}

From Human Vids

01 / Data Collection

- Collect a small amount of teleoperation data D_{tele}
- Anchor the policy to the target environment
- Support multiple interaction modes

02 / Policy Init

- Start from human-pretrained policy
- Initialize robot policy with same architecture
- Train on D_{tele} using action loss **L**

03 / Conditioning

- Condition model on gripper-view observations m_t
- Set ***improvement indicator*** $I_t = 1$ for all initialization samples

$$I_t \in \{0, 1\}$$

$$\hat{a}_{t:t+H-1} = \Pi(s_{t-H'+1:t'}\ z_{t-H'+1:t}\ ;\ m_{t'}\ I_t)$$

$$\hat{A}(o_t, a_{t:t+H-1}) = \sum_{t'=t}^{t+H-1} r_{t'} + V_\psi(o_{t+H}) - V_\psi(o_t)$$

$$\mathcal{L}_\pi = \left\| \hat{a}^{\tau}_{t:t+H-1} - a_{t:t+H-1} \right\|_2^2$$

$$\hat{\pi}(a_{t:t+H-1} \mid o_t) \propto \pi_{\text{ref}}(a_{t:t+H-1} \mid o_t)\, p(I_t \mid \hat{A}(o_t, a_{t:t+H-1}))^\beta$$

$\mathbf{1}(\hat{A}(o_t, a_{t:t+H-1}, \ell) > \epsilon)$ with a task-specific threshold ϵ

Step-2.2: Adapt Dynamics & Value

Dynamics Model

Value Model

01 / Why Adapt?

- Models don't transfer perfectly because robot bodies are different
- Mistakes and biases only show up when testing on the real robot

02 / How to Adapt

- Use real robot practice data to ground the AI models
- Tune predictions to fit the specific robot's unique actions

03 / Autonomous Scale

- Skip the high cost of humans guiding the robot manually
- Let AI suggest tasks while the robot practice-runs them on its own
- GPT-4 proposes plausible atomic interaction instructions (e.g., close the drawer), which the frozen human-pretrained policy then executes on the robot.
- These rollouts expose the models to robot states and natural failures

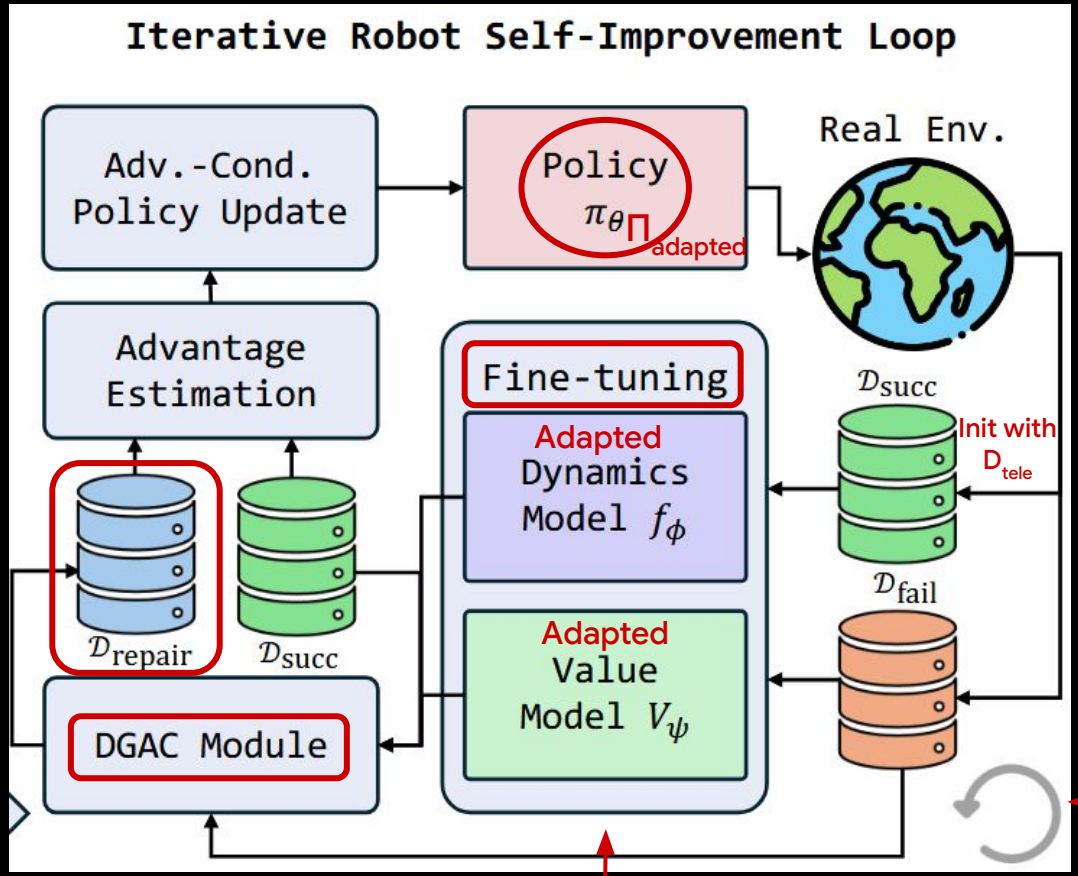
For each scene, we first use the pretrained affordance model from VidBot [8] to detect plausible contact regions. We then select a target contact point and plan a pre-contact trajectory that moves the gripper to the corresponding region while avoiding collisions. Once the gripper reaches the contact region, we query the human-pretrained policy to generate the post-contact interaction trajectory. The robot follows the predicted end-effector poses using its low-level controller, while the gripper is smoothly adjusted along the trajectory to improve execution stability. This strategy separates coarse contact acquisition from post-contact interaction, allowing the robot to explore semantically meaningful behaviors even when the human-pretrained policy does not provide precise contact poses. The collected rollouts contain both successful and failed interactions caused by robot-specific contacts, dynamics, and scene configurations. We use these autonomous rollouts to adapt the dynamics and value models before task-level self-improvement.

Step-3
Iterative self improvement

$$\hat{\pi}(\boldsymbol{a}_{t:t+H-1} \mid \boldsymbol{o}_t)$$

$$\pi_{\text{ref}}(\boldsymbol{a}_{t:t+H-1} \mid \boldsymbol{o}_t)\, p(I_t \mid \hat{A}(\boldsymbol{o}_t, \boldsymbol{a}_{t:t+H-1}))^{\beta}$$

$$\hat{A}(\boldsymbol{o}_t, \boldsymbol{a}_{t:t+H-1}) = \sum_{t'=t}^{t+H-1} r_{t'} + V_\psi(\boldsymbol{o}_{t+H}) - V_\psi(\boldsymbol{o}_t)$$

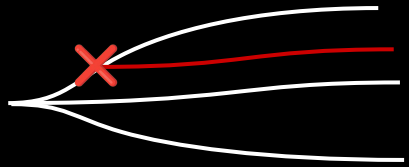


Only place where human is needed for determining the outcome Succ / Fail

Keep on doing this. Here authors have used 2 rounds

This column models are from step-2: adaptation

DGAC: Dynamics Guided Action Correction



Naive copy or repair ✅

Algorithm 1: Dynamics-Guided Action Correction (DGAC)

Require: Policy π_θ, dynamics model f_ϕ, value model V_ψ, thresholds $\delta_\mathrm{v}, \delta_\mathrm{s}$

Input: Failed state $\boldsymbol{o}_t$, successful dataset $\mathcal{D}_\mathrm{succ}$

Output: Corrected action $\boldsymbol{a}^\star_{t:t+H-1}$ or $\varnothing$

Find $\mathcal{O}_t^{(k)} \subset \mathcal{D}_\mathrm{succ}$ and select $\tilde{\boldsymbol{o}}_t$ by Eq. 4

if $|V_\psi(\tilde{\boldsymbol{o}}_t) - V_\psi(\boldsymbol{o}_t)| > \delta_\mathrm{v}$ **or** $\mathrm{sim}(\boldsymbol{o}_t, \tilde{\boldsymbol{o}}_t) < \delta_\mathrm{s}$ **then**

  return $\varnothing$

Sample action candidates $\{\boldsymbol{a}^{(n)}_{t:t+H-1}\}_{n=1}^N$ via Eq. 5

for $n = 1$ **to** N **do**

  $\hat{\boldsymbol{o}}^{(n)}_{t+H-1} \leftarrow f_\phi(\boldsymbol{o}_{t-H'+1:t}, \boldsymbol{a}^{(n)}_{t:t+H-1})$

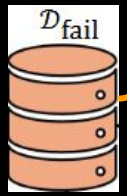
  $V^{(n)}_{t+H-1} \leftarrow V_\psi(\hat{\boldsymbol{o}}^{(n)}_{t+H-1})$

$\boldsymbol{a}^\star_{t:t+H-1} \leftarrow \mathrm{argmax}_n\, V^{(n)}_{t+H-1}$

return $\boldsymbol{a}^\star_{t:t+H-1}$

retrieve the steering reference from D_{succ} in two stages

Failed state $\boldsymbol{o}_t$



$$|V_\psi(\tilde{\boldsymbol{o}}_t) - V_\psi(\boldsymbol{o}_t)| \le \delta_\mathrm{v}$$ Nearest task progress candidates

$$\mathrm{sim}(\boldsymbol{o}_t, \tilde{\boldsymbol{o}}_t) \ge \delta_\mathrm{s}$$ Nearest similar candidate

$$\mathcal{O}_t^{(k)} = \mathrm{TopK}_{\boldsymbol{o}' \in \mathcal{D}_\mathrm{succ}}(-|V_\psi(\boldsymbol{o}') - V_\psi(\boldsymbol{o}_t)|)$$

$$\tilde{\boldsymbol{o}}_t = \arg\max_{\boldsymbol{o}' \in \mathcal{O}_t^{(k)}} \mathrm{sim}(\boldsymbol{o}_t, \boldsymbol{o}')$$

We know the timestep. Query the velocity field over noisy action chunks

$$\boldsymbol{v}_\theta(\boldsymbol{a}^\tau_{t:t+H-1}, \tau; \cdot)$$

Compose o_t and retrieved o_t Velocity fields

$$w_n \sim \mathcal{U}(0.1, 1)$$

$$\boldsymbol{v}^{(n)}_\theta = \boldsymbol{v}_\theta(\boldsymbol{a}^\tau_{t:t+H-1}, \tau; \boldsymbol{h}_t) + w_n \Big[\boldsymbol{v}_\theta(\boldsymbol{a}^\tau_{t:t+H-1}, \tau; \tilde{\boldsymbol{h}}_t) - \boldsymbol{v}_\theta(\boldsymbol{a}^\tau_{t:t+H-1}, \tau; \boldsymbol{h}_t)\Big]. \quad (5)$$

Dynamics Guided Action Correction: Failure becomes supervision signal now for self improvement

DGAC: Dynamics Guided Action Correction

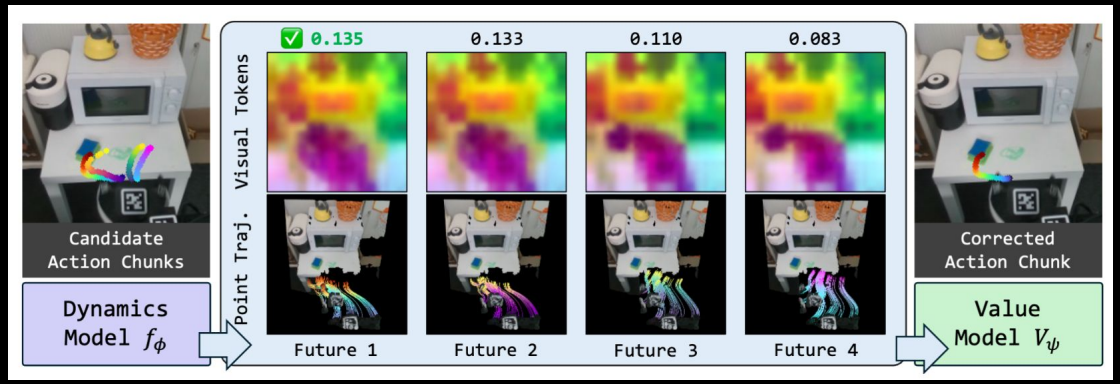
Sample action candidates $\{\boldsymbol{a}_{t:t+H-1}^{(n)}\}_{n=1}^{N}$ via Eq. 5

for $n = 1$ **to** N **do**

$\hat{\boldsymbol{o}}_{t+H-1}^{(n)} \leftarrow f_\phi(\boldsymbol{o}_{t-H'+1:t}, \boldsymbol{a}_{t:t+H-1}^{(n)})$

$V_{t+H-1}^{(n)} \leftarrow V_\psi(\hat{\boldsymbol{o}}_{t+H-1}^{(n)})$

$\boldsymbol{a}_{t:t+H-1}^{\star} \leftarrow \text{argmax}_n \, V_{t+H-1}^{(n)}$



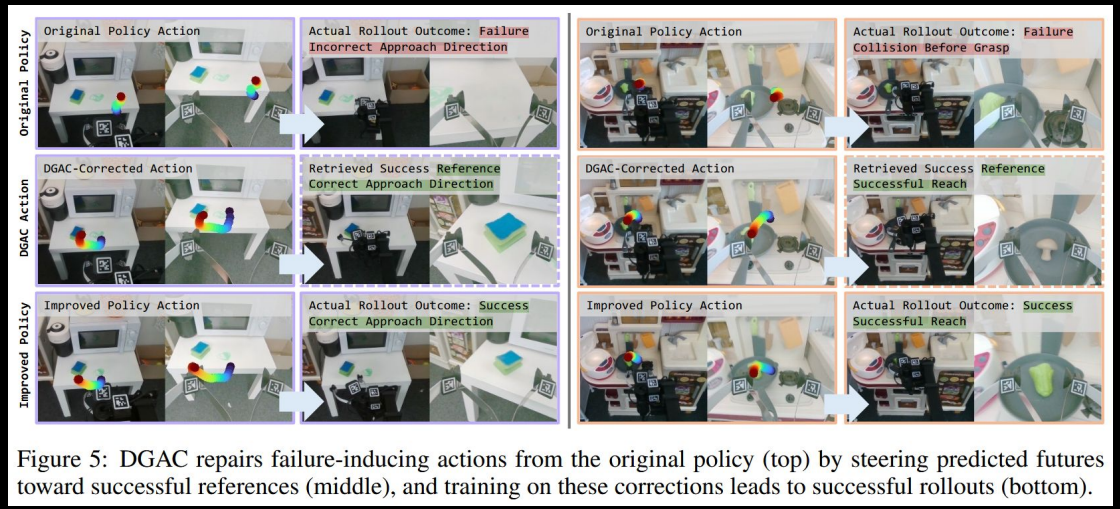
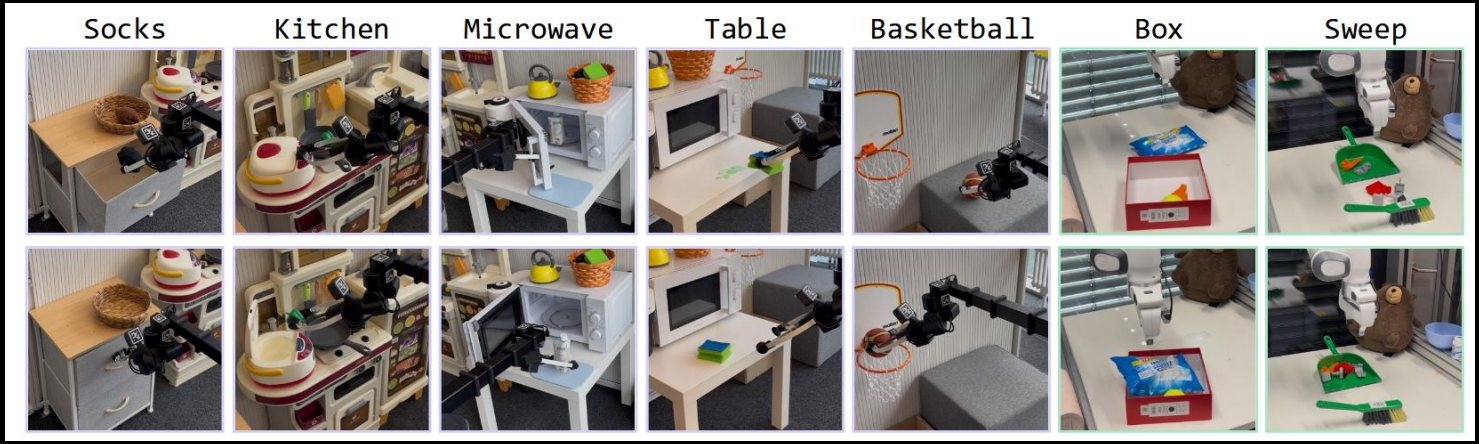


Figure 5: DGAC repairs failure-inducing actions from the original policy (top) by steering predicted futures toward successful references (middle), and training on these corrections leads to successful rollouts (bottom).

Results



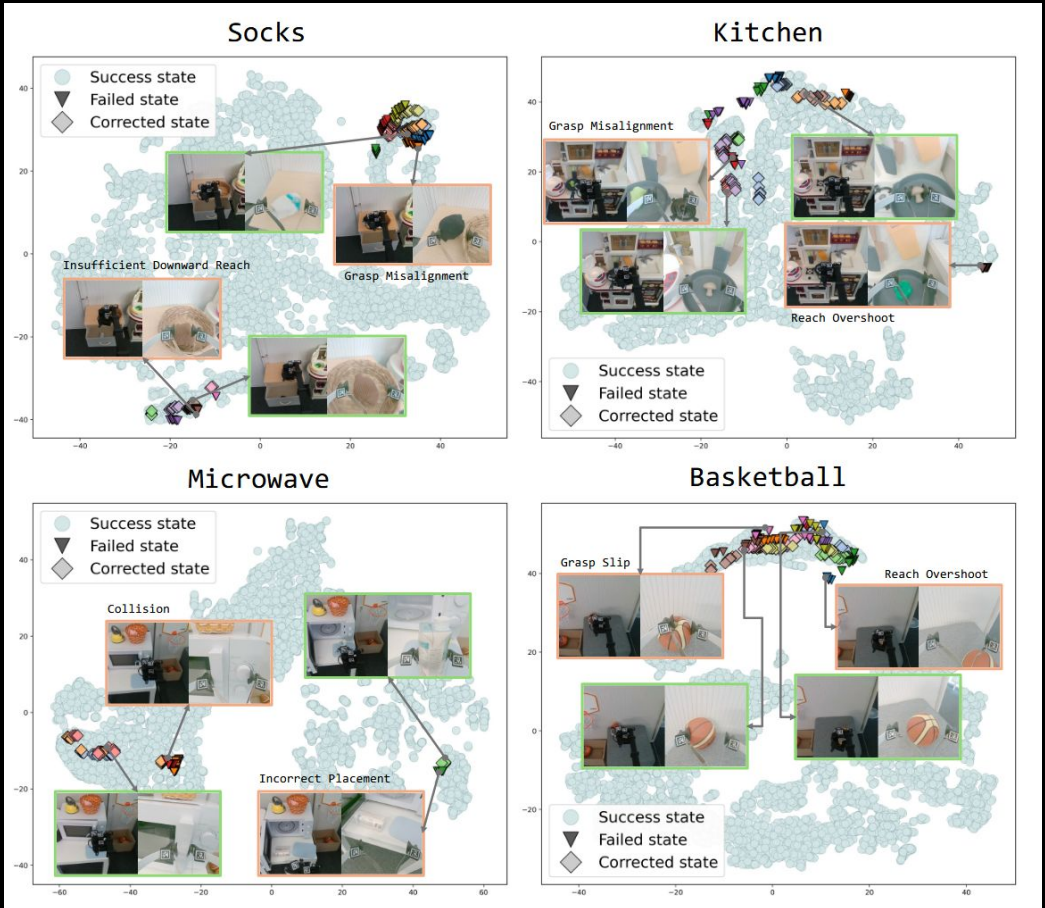
Method	T1	T2	T3	T4	T5	Avg.
Expert BC	33.3	46.7	46.7	40.0	40.0	41.3
SWIM [62]	53.3	40.0	46.7	46.7	60.0	49.3
LPB [63]	53.3	40.0	66.7	80.0	60.0	60.0
AWR [61]	53.3	53.3	**80.0**	60.0	53.3	60.0
RECAP† [2]	60.0	53.3	73.3	53.3	66.7	61.3
RISE† [3]	**86.7**	60.0	73.3	80.0	80.0	76.0
Ours	**86.7**	**80.0**	**80.0**	**93.3**	**86.7**	**85.3**

Table 1: Benchmark evaluated on success rate (%). T1: Socks, T2: Kitchen, T3: Microwave, T4: Table, T5: Basketball. †: no human-intervened corrections.

Method	T1	T2	T3	T4	T5	Avg.
$\pi_{0.5}$ + SFT	60.0	66.7	80.0	40.0	66.7	62.7
$\pi_{0.5}$ + RECAP†	66.7	66.7	73.3	53.3	80.0	68.0
$\pi_{0.5}$ + DGAC	**93.3**	**86.7**	**86.7**	86.7	**86.7**	**88.0**
Ours	86.7	80.0	80.0	**93.3**	**86.7**	85.3

Table 3: Success rates (%) of the baseline policy $\pi_{0.5}$ under different policy improvement schemes. †: no human-intervened corrections.

t-SNE Dynamics Viz



Ablations

For module specific ablations.
Please refer to the paper.

Limitations & Future Work

01

Human Supervision

- Relies on human annotations for rollout outcomes
- Needs foundation models for automated feedback

02

Horizon Limits

- Operates locally over short-horizon action chunks
- Requires extending to longer-horizon dynamics

03

Pretraining & Scale

- Scale pretraining to larger, diverse datasets
- Improves generalization across environments

Fin.

Q/A