ID2P: FUTURE-GUIDED POLICY LEARNING VIA IDM-TO-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Future observations reveal the consequences of executed actions, motivating existing methods to leverage them for robot policy learning. However, prior work primarily uses future observations for representation-level supervision, while direct action-level future supervision remains less explored. We propose IDM-to-Policy Distillation (ID2P), a training framework that combines action-level and representation-level future supervision for flow-based policies. ID2P first trains an inverse dynamics model (IDM) conditioned on current and future observations using the same architecture as the policy. The trained IDM then provides action supervision by mixing its predicted action velocity with the expert action velocity from expert trajectories, while its action expert hidden states provide complementary representation supervision for the policy. Under an ideal IDM, we show that the mixed action velocity objective preserves the expected gradient of expert-only supervision while achieving no greater target variance. ID2P requires neither future observations nor explicit future state prediction at inference time, preserving the policy architecture and inference cost. Extensive simulation and real-world experiments demonstrate that ID2P consistently improves task success rates across flow-based policies of different architectures and scales, including SmolVLA (0.45B) and $\pi_{0.5}$ (3B).

1 INTRODUCTION

Future observations capture the world states after a robot executes an action and can provide more reliable cues for action prediction. Motivated by this insight, recent work leverages future observations available during training to guide robot policy learning, thereby improving action prediction. Typically, existing methods first construct future-derived supervision signals from future observations, such as future images (Ye et al., 2026b; Li et al., 2026b; Zhang et al., 2026b), future visual features (Deng et al., 2026), 2D tracks (Guan et al., 2026), 3D tracks (Wang et al., 2026a), and latent actions (Liu et al., 2026a), and then use them to supervise the policy's intermediate outputs. Some methods explicitly predict these future states and condition action prediction on them, while others use them to supervise existing policy's intermediate outputs without predicting additional future states. In this paper, we refer to the policy's intermediate outputs as representations. Since explicitly predicted future states are themselves intermediate outputs, they also constitute policy representations. Thus, existing methods mainly perform representation-level future supervision, while actions are supervised by expert actions from expert trajectories.

Despite their success, direct action supervision from future observations remains insufficiently explored. Since representations are typically mapped to actions through complex transformations, representation-level future supervision alone does not guarantee that future observations can effectively improve final action prediction. A recent attempt, PFD (Fang et al., 2026), uses an IDM to correct policy action velocity, but its IDM is not explicitly trained for inverse dynamics prediction, limiting the reliability of its velocity supervision. To address this limitation, we propose **Inverse Dynamics Model-to-Policy Distillation (ID2P)**, a future-guided policy learning framework with both action-level and representation-level future supervision. During training, ID2P follows a two-stage paradigm. First, we train an Inverse Dynamics Model (IDM) on demonstration data using the same architecture as the flow-based policy. Second, we perform IDM-to-Policy Distillation. The frozen IDM's predicted action velocities are combined with expert action velocities to form a mixed action velocity target for supervising the policy's action velocity. Meanwhile, the IDM action expert's hid-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

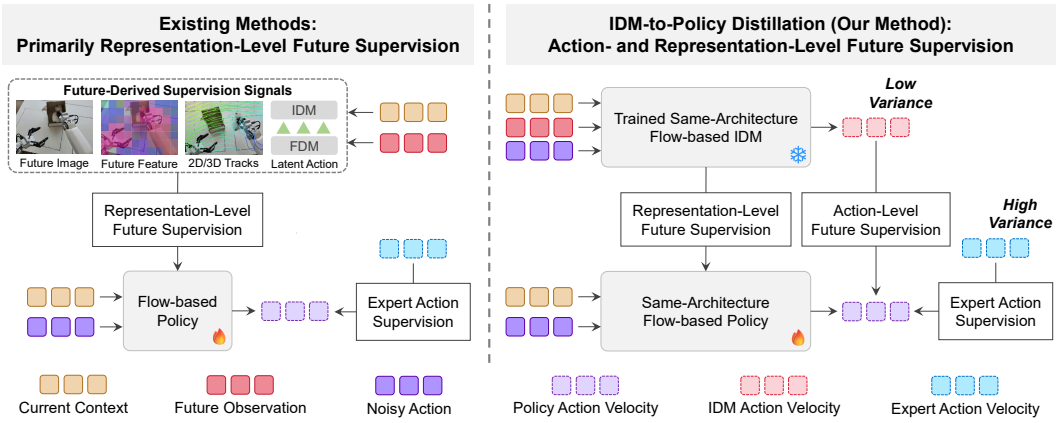

Figure 1: **Motivation of ID2P.** Existing methods primarily leverage future observations for representation-level supervision, while direct action-level future supervision remains insufficiently explored. ID2P trains a same-architecture flow-based IDM with future observations and distills both its action velocity and representations into the policy. By combining action-level and representation-level future supervision, ID2P more fully leverages future observations for policy learning.

den states supervise the corresponding policy hidden states. During inference, the policy directly predicts actions without predicting additional future states. ID2P is a general training framework for flow-based policies with no architectural changes. Figure 1 shows our motivation.

Intuitively, the IDM can predict actions more reliably by leveraging future images. ID2P enables an IDM with the same architecture as the policy to learn both the action velocities predicted with access to future images and the corresponding internal computations, and then transfers this knowledge to a policy that does not observe future images. In this way, ID2P aims to endow the policy with high-quality action prediction capabilities similar to those of the IDM.

For action-level future supervision, we show that, under an ideal IDM, mixed action velocity supervision preserves the expected gradient of expert-only supervision without increasing target variance. Although using the IDM action velocity target alone provides the greatest theoretical variance reduction, a learned IDM inevitably incurs estimation errors and is therefore not optimal. We thus incorporate the unbiased expert action velocity target to compensate for errors, and use the mixed action velocity target to supervise the policy's action velocity.

For representation-level future supervision, we align the hidden states of the policy's action expert with those of the frozen IDM. Since the IDM has access to future images, its hidden states encode action-relevant future information. Distilling these representations encourages the policy to internalize such information without requiring future images at inference time. Compared with conventional methods (Ye et al., 2026b; Li et al., 2026b), ID2P's representation-level future supervision does not require the policy to predict additional future states, eliminating extra inference overhead.

Our contributions are summarized as follows:

- We propose ID2P, a future-guided policy learning framework that distills IDM action velocities and representations into the policy. With action-level and representation-level future supervision, ID2P more fully leverages future observations to improve policy learning.
- We theoretically show that mixing ideal IDM action velocity with expert action velocity preserves the expected gradient of expert-only supervision without increasing target variance. We further characterize the effect of the mixing weight on variance reduction.
- Extensive simulation and real-world experiments demonstrate that ID2P consistently improves flow-based policies across different architectures and model scales, including SmolVLA (0.45B) and $\pi_{0.5}$ (3B), validating its effectiveness and generalizability.

2 Related Work

2.1 Vision-Language-Action Models

Vision-Language-Action (VLA) models control robots from visual observations and language instructions, showing promise toward generalist agents (Black et al., 2024; Kim et al., 2025; Zheng et al., 2026; Wu et al., 2026; Yuan et al., 2026a). As pioneering efforts, the RT series (Zitkovich et al., 2023; Belkhale et al., 2024) fine-tuned vision-language models on large-scale robot data (O'Neill et al., 2024), achieving strong task success and generalization. Subsequent work further improves VLAs through Chain-of-Thought reasoning (Zawalski et al., 2024; Zhao et al., 2025; Bai et al., 2026), 3D information (Qu et al., 2025; Sun et al., 2025; Li et al., 2026a; Liu et al., 2026b), and reinforcement learning (Guo et al., 2025; Lu et al., 2025), enhancing reasoning, spatial perception, and robustness. Despite these advances, existing VLAs still struggle to capture the relationship between actions and their resulting outcomes, limiting their ability to achieve desired outcomes. In contrast, our method learns these relationships with an IDM and transfers them to the policy, improving its understanding of action consequences and action prediction.

2.2 Future-Guided Policy Learning

To improve a policy's future awareness, recent methods leverage future observations to guide policy training (Bi et al., 2026; Ye et al., 2026a; Team et al., 2026; Kim et al., 2026; Zhang et al., 2026a; Guo et al., 2026). These methods can be broadly divided into two categories based on whether future states are explicitly modeled at inference time. The first explicitly predicts future images (Liang et al., 2025; Kim et al., 2026; Ma et al., 2026), future visual features (Deng et al., 2026), latent world dynamics (Tang et al., 2026), 2D tracks (Guan et al., 2026), or future-related latent reasoning (Luo et al., 2026), and then conditions action prediction on them. The second uses auxiliary future state prediction tasks (Lyu et al., 2026; Yuan et al., 2026b; Li et al., 2025; Zhu et al., 2025) or latent actions (Liu et al., 2026a) to supervise the policy's existing intermediate outputs, i.e., representations, without explicitly predicting future states at inference time. Overall, these methods mainly use future observations for representation supervision, while direct action supervision remains insufficiently studied. More recently, PFD (Fang et al., 2026) uses IDM action velocity to correct policy action velocity under the guidance of a video generation task. However, its IDM shares parameters with the policy during training and is not explicitly trained for inverse dynamics prediction, so its action velocity is not explicitly optimized to approximate the optimal flow matching velocity. Moreover, PFD lacks IDM-to-Policy representation distillation. In contrast, we first train an IDM to provide reliable supervision and then distill both its action velocity and representations into the policy, enabling more comprehensive knowledge transfer. We further establish the gradient-consistency and variance-reduction properties of IDM-to-Policy action distillation.

3 Preliminaries

Problem Formulation. Given a set of expert trajectories $\mathcal{D} = \{\xi_i\}_{i=1}^{M}$ with $\xi_i \overset{\text{i.i.d.}}{\sim} p_E(\xi)$, where $p_E(\xi)$ denotes an unknown expert trajectory distribution, $\xi = \{(o_t, a_t)\}_{t=0}^{T-1}$ denotes a random expert trajectory, and T its length. $o_t = [I_t^1, \ldots, I_t^K, l_t, s_t]$ denotes the observation at time step t, I_t^k the image from the k-th camera, l_t the language instruction, s_t the robot proprioceptive state, and a_t the robot action. We aim to learn a policy $\pi_\theta(o_t)$ that outputs an action chunk $A_t = [a_t, a_{t+1}, \ldots, a_{t+N-1}]$, where N denotes the chunk size and θ the policy parameters.

Conditional Flow Matching Policies. Robot actions under the same observation are often multimodal (Chi et al., 2023). Since flow matching excels at modeling multimodal distributions, a common policy learning approach (Shukor et al., 2025; Intelligence et al., 2025) is to minimize the following Conditional Flow Matching (Lipman et al., 2022) behavior cloning loss:

$$\mathcal{L}_{\text{Policy}}(\theta) = \mathbb{E}_{\tau,\, p_E(o_t, A_t),\, p_\tau(A_t^\tau \mid A_t)} \left\| v_\theta(A_t^\tau, \tau, o_t) - v_\tau(A_t^\tau \mid A_t) \right\|^2, \tag{1}$$

where $\tau \sim \mathcal{U}(0,1)$ denotes flow time, $p_E(o_t, A_t)$ the joint expert distribution induced by $p_E(\xi)$ and the sampling of valid time steps, $p_\tau(A_t^\tau \mid A_t)$ the conditional probability path, $v_\theta(A_t^\tau, \tau, o_t)$ the neural velocity field parameterized by the policy, and $v_\tau(A_t^\tau \mid A_t)$ the conditional velocity field. Most work (Shukor et al., 2025; Intelligence et al., 2025) adopt a linear path $A_t^\tau = (1-\tau)\epsilon + \tau A_t$

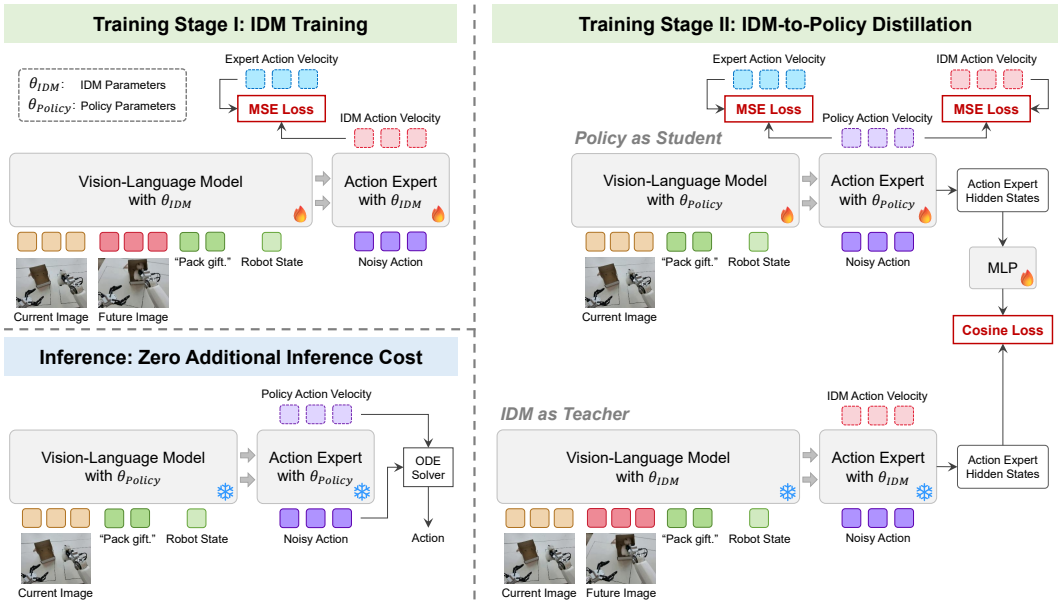

Figure 2: **Overview of ID2P.** We first train an IDM conditioned on current and future observations to learn action dynamics with access to future information. The frozen IDM then distills its future-conditioned velocity field and action expert representations into the policy, providing complementary action-level and representation-level future supervision without additional inference overhead.

with $\epsilon \sim \mathcal{N}(0, I)$, for which $v_\tau(A_t^\tau \mid A_t) = A_t - \epsilon$. In this paper, we refer to $v_\theta(A_t^\tau, \tau, o_t)$ as *poliy action velocity* and $v_\tau(A_t^\tau \mid A_t)$ as *expert action velocity*. Moreover, the minimizer of equation 1 is the true velocity field

$$v_\tau(A_t^\tau \mid o_t) = \mathbb{E}_{p_\tau(A_t \mid A_t^\tau, o_t)} \left[v_\tau(A_t^\tau \mid A_t) \right]. \tag{2}$$

Note that the expert action velocity $v_\tau(A_t^\tau \mid A_t)$ is only a single-sample Monte Carlo estimator of equation 2 and can have high variance (Hammersley, 2013). Following Xu et al. (2023) and Yang et al. (2026a), we define its variance at flow time τ as

$$\mathcal{V}_{\text{Policy}}(\tau) = \mathbb{E}_{p_\tau(A_t^\tau, o_t)} \operatorname{Tr}\left[\operatorname{Cov}_{p_\tau(A_t \mid A_t^\tau, o_t)} \left(v_\tau(A_t^\tau \mid A_t) \right) \right]. \tag{3}$$

4 IDM-to-Policy Distillation

To more fully leverage future observations to enhance action prediction, we propose ID2P, a training framework for flow-based policies. ID2P follows a two-stage training paradigm without modifying the policy model architecture or requiring additional future states prediction at inference time. Section 4.1 introduces the first stage of training, Section 4.2 and Section 4.3 present IDM-to-Policy action distillation and representation distillation in the second stage, and Section 4.4 discusses policy initialization. Figure 2 provides an overview of ID2P.

4.1 Stage I: Inverse Dynamics Model Training

In this stage, we train an IDM with the same architecture as the flow-based policy using the following Conditional Flow Matching objective:

$$\mathcal{L}_{\text{IDM}}(\phi) = \mathbb{E}_{\tau,\, p_E(o_t, F_t, A_t),\, p_\tau(A_t^\tau \mid A_t)} \left\| v_\phi(A_t^\tau, \tau, o_t, F_t) - v_\tau(A_t^\tau \mid A_t) \right\|^2, \tag{4}$$

where F_t denote the future observations associated with o_t, $p_E(o_t, F_t, A_t)$ the joint distribution induced by expert trajectories and the sampling of valid time steps, $v_\phi(A_t^\tau, \tau, o_t, F_t)$ the IDM neural velocity field, and ϕ the IDM parameters. In this paper, we refer to $v_\phi(A_t^\tau, \tau, o_t, F_t)$ as *IDM action velocity*. In practice, we approximate $p_E(o_t, F_t, A_t)$ by the empirical distribution induced by the robot dataset $\mathcal{D}$. The minimizer of equation 4 is the future-conditioned velocity field:

$$v_\tau(A_t^\tau \mid o_t, F_t) = \mathbb{E}_{p_\tau(A_t \mid A_t^\tau, o_t, F_t)} [v_\tau(A_t^\tau \mid A_t)]. \tag{5}$$

After Stage I, we freeze the learned IDM parameters $\widehat{\phi}$.

4.2 Stage II: IDM-to-Policy Action Distillation

Because the IDM observes future images that reveal the outcomes of executed actions, it has access to stronger cues for action prediction than the policy, making its action prediction more reliable. We therefore distill the IDM's action-prediction knowledge into the policy, using IDM action velocity as privileged supervision while requiring no future observations at inference.

To improve robustness, we retain the unbiased expert action velocity supervision from conventional behavior cloning and combine it with IDM supervision to form a mixed action velocity target

$$v^{\alpha}_{\tau,\widehat{\phi}} = (1-\alpha)\, v_\tau(A^\tau_t \mid A_t) + \alpha\ \mathrm{sg}\Big[v_{\widehat{\phi}}(A^\tau_t, \tau, o_t, F_t)\Big], \tag{6}$$

where $\alpha \in (0,1)$ and $\mathrm{sg}[\cdot]$ denotes stop-gradient. Then, we use this mixed action velocity target to supervise the policy action velocity through the following action distillation objective:

$$\mathcal{L}_{\mathrm{AD}}(\theta;\alpha) = \mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A^\tau_t|A_t)} \left\| v_\theta(A^\tau_t, \tau, o_t) - v^{\alpha}_{\tau,\widehat{\phi}} \right\|^2. \tag{7}$$

Gradient Equivalence. The following theorem shows that, under an ideal IDM, the policy Conditional Flow Matching objective (equation 1) and the action distillation objective (equation 7) have identical gradients w.r.t. θ. This property guarantees that, under an ideal IDM, mixed action velocity supervision and pure expert action velocity supervision share the same global minimizer, $v_\tau(A^\tau_t \mid o_t)$. Proofs are provided in Appendix D.1.

Theorem 1. *Assume that the learned IDM realizes a global minimizer of $\mathcal{L}_{\mathrm{IDM}}$, such that $v_{\widehat{\phi}}(A^\tau_t, \tau, o_t, F_t) = v_\tau(A^\tau_t \mid o_t, F_t)$ a.s. For any $\alpha \in \mathbb{R}$, $\mathcal{L}_{\mathrm{AD}}(\theta;\alpha)$ and $\mathcal{L}_{\mathrm{Policy}}(\theta)$ are equal up to a constant independent of θ. Hence, $\nabla_\theta \mathcal{L}_{\mathrm{AD}}(\theta;\alpha) = \nabla_\theta \mathcal{L}_{\mathrm{Policy}}(\theta)$.*

Variance Reduction. Under the same ideal-IDM assumption, the action distillation target in equation 7 has no greater variance than the expert action velocity target used in the policy Conditional Flow Matching objective equation 1, while preserving gradient equivalence. To characterize this property, we define the action distillation target variance as

$$\mathcal{V}_{\mathrm{AD}}(\tau;\alpha) = \mathbb{E}_{p_\tau(A^\tau_t, o_t)}\, \mathrm{Tr}\Big[\mathrm{Cov}_{p_\tau(A_t, F_t | A^\tau_t, o_t)}\left(v^{\alpha}_{\tau,\widehat{\phi}}\right)\Big], \tag{8}$$

and the residual target variance conditioned on future observations F_t as

$$\mathcal{V}_{\mathrm{res}}(\tau) = \mathbb{E}_{p_\tau(A^\tau_t, o_t, F_t)}\, \mathrm{Tr}\big[\mathrm{Cov}_{p_\tau(A_t | A^\tau_t, o_t, F_t)}\left(v_\tau(A^\tau_t \mid A_t)\right)\big]. \tag{9}$$

We formalize the variance reduction property in the following theorem. Proofs are provided in Appendix D.2.

Theorem 2. *For a fixed $\tau \in (0,1)$, assume that the learned IDM satisfies $v_{\widehat{\phi}}(A^\tau_t, \tau, o_t, F_t) = v_\tau(A^\tau_t \mid o_t, F_t)$ a.s. Then, for any $\alpha \in \mathbb{R}$,*

$$\mathcal{V}_{\mathrm{Policy}}(\tau) - \mathcal{V}_{\mathrm{AD}}(\tau;\alpha) = \alpha(2-\alpha)\mathcal{V}_{\mathrm{res}}(\tau). \tag{10}$$

Hence, if $\mathcal{V}_{\mathrm{res}}(\tau) > 0$, then $\mathcal{V}_{\mathrm{AD}}(\tau;\alpha) < \mathcal{V}_{\mathrm{Policy}}(\tau)$ for every $\alpha \in (0,2)$.

Practical Choice of α. Theorem 2 shows that $\alpha = 1$ maximizes target variance reduction. In practice, however, a learned IDM may deviate from the true future conditioned velocity field. We therefore retain unbiased expert action velocity supervision and interpolate it with IDM action velocity, as in equation 6 , restricting $\alpha \in (0,1)$ so both signals contribute to policy learning.

4.3 Stage II: IDM-to-Policy Representation Distillation

Action distillation transfers only the IDM's action velocity, leaving its future-conditioned internal computation unexploited. Since the IDM action expert encodes future information directly relevant to action velocity prediction, we further distill its representations into the policy. Specifically, we define the following representation distillation objective:

$$\mathcal{L}_{\mathrm{RD}}(\theta,\psi) = \mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A^\tau_t|A_t)} \left[\frac{1}{LN} \sum_{\ell=1}^{L} \sum_{j=1}^{N} d_{\cos}\Big(f_\psi(h^{\mathrm{Policy}}_{t,\ell,j}), \mathrm{sg}[h^{\mathrm{IDM}}_{t,\ell,j}] \Big) \right], \tag{11}$$

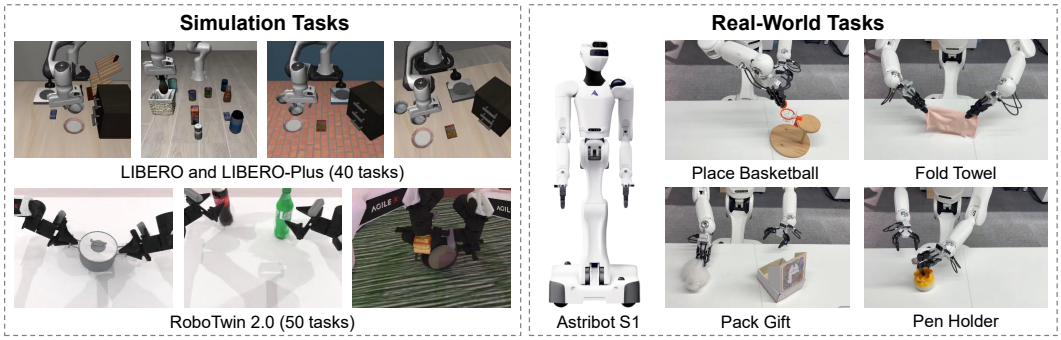

Figure 3: **Experimental setups.** Representative tasks from LIBERO (Liu et al., 2023), LIBERO-Plus (Fei et al., 2025), and RoboTwin (Chen et al., 2025), together with our real-world evaluation tasks conducted on the Astribot S1 platform.

where $h_{t,\ell,j}^{\text{Policy}}$ and $h_{t,\ell,j}^{\text{IDM}}$ denote the hidden states of the j-th action token at the ℓ-th action expert layer, and f_ψ is a learnable two-layer MLP projector with GELU activation function (Hendrycks & Gimpel, 2016). L denotes the number of action expert layers. The distance metric is defined as $d_{\cos}(x, y) = 1 - \text{CosSim}(x, y)$.

Finally, we optimize Stage II with the following objective:

$$\mathcal{L}_{\text{Stage II}}(\theta, \psi) = \mathcal{L}_{\text{AD}}(\theta; \alpha) + \lambda_{\text{RD}} \mathcal{L}_{\text{RD}}(\theta, \psi), \tag{12}$$

where λ_{RD} controls the weight of the representation distillation loss.

4.4 Policy Initialization

Since the IDM and policy share the same architecture, a natural design is to initialize the policy with the IDM parameters. However, we empirically find that this initialization makes the gains from action distillation and representation distillation unstable and degrades performance. We hypothesize that the IDM parameters become specialized to future-conditioned inputs and encode dependencies on future observations, causing a mismatch when transferred to a policy that only observes current information. Therefore, we do not initialize the policy with the IDM parameters by default.

5 Experiments

In this section, we conduct extensive experiments in simulated and real-world environments to evaluate the effectiveness of ID2P, as illustrated in Figure 3.

5.1 Implementation Details

To evaluate ID2P's generality, we apply it to SmolVLA (0.45B) (Shukor et al., 2025) and $\pi_{0.5}$ (3B) (Intelligence et al., 2025), both comprising a VLM and a flow-based action expert. In both training stages, SmolVLA initializes the VLM from pretrained SmolVLM2 (Marafioti et al., 2025) and the action expert randomly, while $\pi_{0.5}$ initializes both from a robot-pretrained checkpoint. We use a linear path $A_t^\tau = (1 - \tau)\epsilon + \tau A_t$ with $\epsilon \sim \mathcal{N}(0, I)$ for all objectives and set $\alpha = 0.5$, $\lambda_{\text{RD}} = 2$. The chunk size is set to $N = 16$, and we use the images in o_{t+16} corresponding to the current observation o_t as the future observations F_t. We train each stage for 30k steps using AdamW with a learning rate of 10^{-4}. Detailed hyperparameters are provided in Appendix B (Table 19).

5.2 Simulation Experiments

Experimental Setup. We evaluate ID2P on LIBERO (Liu et al., 2023), LIBERO-Plus (Fei et al., 2025), and RoboTwin (Chen et al., 2025) with SmolVLA (0.45B) and $\pi_{0.5}$ (3B) to assess task execution and robustness across model scales. LIBERO comprises 40 tasks across four suites: Spatial, Object, Goal, and Long, with 50 evaluation episodes per task, totaling 2,000 episodes. LIBERO-Plus

Table 1: Results on LIBERO and RoboTwin. P.T. denotes additional embodied policy pretraining. RoboTwin Easy and Hard denote Clean2Clean and Clean2Random, respectively.

Method	P.T.	LIBERO					RoboTwin		
		Spatial	Object	Goal	Long	Avg.	Easy	Hard	Avg.
π_0 (Black et al., 2024)	✓	96.8	98.8	95.8	85.2	94.2	46.4	16.3	31.4
Fast-WAM (Yuan et al., 2026b)	×	98.2	100.0	97.0	95.2	97.6	77.8	1.9	39.9
ImageWAM (Zhang et al., 2026b)	×	97.2	99.2	98.8	98.4	98.4	84.4	18.3	51.4
ABot-M0 (Yang et al., 2026b)	✓	98.8	99.8	99.0	96.6	98.6	57.4	30.4	43.9
SmolVLA (Shukor et al., 2025)	×	**96.8**	98.2	94.4	91.0	95.1	31.6	2.8	17.2
SmolVLA + ID2P (Ours)	×	**96.8**	**99.4**	**96.8**	**93.4**	**96.6**	**32.7**	**3.0**	**17.9**
ID2P Relative Improvement		0.0%	+1.2%	+2.5%	+2.6%	+1.6%	+3.5%	+7.1%	+4.1%
$\pi_{0.5}$ (Intelligence et al., 2025)	✓	97.8	**100.0**	97.6	**98.0**	98.4	80.6	37.7	59.2
$\pi_{0.5}$ + ID2P (Ours)	✓	**99.6**	**100.0**	**98.0**	**98.0**	**98.9**	**81.1**	**40.5**	**60.8**
ID2P Relative Improvement		+1.8%	0.0%	+0.4%	0.0%	+0.5%	+0.6%	+7.4%	+2.7%

Table 2: Results on LIBERO-Plus. P.T. denotes additional embodied policy pretraining.

Method	P.T.	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
WorldVLA (Cen et al., 2025)	✓	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
π_0 (Black et al., 2024)	✓	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
OpenWAM-α (Wang et al., 2026b)	✓	33.8	76.1	88.0	97.0	87.1	39.8	77.5	69.2
StarVLA (Community, 2026)	×	52.5	49.8	88.5	95.7	95.7	73.0	76.9	74.1
Being-H0.7 (Luo et al., 2026)	✓	82.0	59.0	82.8	97.8	90.0	93.5	88.5	82.1
ImageWAM (Zhang et al., 2026b)	×	80.8	50.3	91.4	98.1	85.5	93.8	80.5	83.1
SmolVLA (Shukor et al., 2025)	×	10.0	10.8	46.7	76.1	**63.8**	11.7	**58.0**	36.6
SmolVLA + ID2P (Ours)	×	**19.2**	**25.7**	**64.0**	**76.6**	62.6	**33.4**	54.4	**45.9**
ID2P Relative Improvement		+92.0%	+138.0%	+37.0%	+0.7%	−1.9%	+185.5%	−6.2%	+25.4%
$\pi_{0.5}$ (Intelligence et al., 2025)	✓	**71.5**	87.6	88.0	**99.4**	97.6	82.2	88.7	86.8
$\pi_{0.5}$ + ID2P (Ours)	✓	70.7	**92.4**	**89.7**	**99.4**	**99.0**	**89.0**	**89.2**	**89.0**
ID2P Relative Improvement		−1.1%	+5.5%	+1.9%	0.0%	+1.4%	+8.3%	+0.6%	+2.5%

contains 10,030 perturbed task instances spanning camera, robot, language, lighting, background, noise, and layout variations. We evaluate one episode per task using models trained on LIBERO training data. RoboTwin contains 50 tasks. We train on clean demonstrations and evaluate 100 episodes per task in each of the Easy (Clean2Clean) and Hard (Clean2Random) settings, totaling 5,000 episodes per setting. For a fair comparison, we reproduce both baselines using the same evaluation protocols as their ID2P counterparts and match all training hyperparameters not specific to ID2P. We further compare with representative VLA and world action models in Tables 1 and 2.

Results. Tables 1 and 2 show that ID2P improves both backbones across all three benchmarks. On LIBERO, SmolVLA gains most on Goal and Long, while $\pi_{0.5}$ further improves on Spatial and Goal. On LIBERO-Plus, SmolVLA substantially outperforms its baseline, and both backbones show notable gains under robot and noise perturbations; SmolVLA also improves under camera and language variations. On RoboTwin, ID2P improves both Easy and Hard, with larger relative gains on Hard, indicating stronger transfer to randomized environments. ID2P also consistently improves over the baseline with or without robot pretrained initialization. Overall, $\pi_{0.5}$ + ID2P achieves the highest average performance among the compared methods across all three benchmarks, validating the effectiveness and generality of ID2P.

5.3 Real-World Experiments

Experimental Setup. We further evaluate $\pi_{0.5}$ (Intelligence et al., 2025) and $\pi_{0.5}$ + ID2P on the Astribot S1 robot across four tasks: placing a basketball into a hoop, folding a towel, packing a gift, and placing pens into a holder. The policy receives three RGB streams from one head camera and two wrist cameras and operates at 30Hz with 34-dimensional actions. We train a separate policy for each task for 30k steps. Each method is evaluated over 30 trials per task, for a total of 120 trials. Table 3 reports per-task and average success rates.

Table 3: Results on real-world tasks. P.T. denotes additional embodied policy pretraining.

Method	P.T.	Basketball	Fold Towel	Pack Gift	Pen Holder	Avg.
ImageWAM	×	60.0	50.0	46.7	50.0	51.7
$\pi_{0.5}$ (Intelligence et al., 2025)	✓	60.0	53.3	**80.0**	26.7	55.0
$\pi_{0.5}$ + ID2P (Ours)	✓	**66.7**	**80.0**	73.3	**36.7**	**64.2**
ID2P Relative Improvement		+11.2%	+50.1%	−8.4%	+37.5%	+16.7%

Table 4: Future supervision ablations with shared policy and trained IDM teachers on SmolVLA. A checkmark under $\mathcal{L}_{\text{AD}}$ sets $\alpha = 0.5$ (otherwise 0), while a checkmark under $\mathcal{L}_{\text{RD}}$ sets $\lambda_{\text{RD}} = 2$ (otherwise 0).

IDM Type	$\mathcal{L}_{\text{AD}}$	$\mathcal{L}_{\text{RD}}$	LIBERO	LIBERO-Plus	Avg.
–			95.1	36.6	65.9
Shared Policy	✓		57.8	15.4	36.6
		✓	93.5	39.3	66.4
	✓	✓	37.1	10.6	23.9
Trained IDM (Ours)	✓		96.0	39.0	67.5
		✓	96.4	43.3	69.9
	✓	✓	**96.6**	**45.9**	**71.3**

Table 6: Ablation of α on SmolVLA.

α	LIBERO	LIBERO-Plus	Avg.
0	95.1	36.6	65.9
1	95.6	38.4	67.0
0.5 (Ours)	**96.0**	**39.0**	**67.5**

Table 7: Ablation of Representation Target on SmolVLA.

Representation Target	LIBERO	LIBERO-Plus	Avg.
VLM	96.4	38.1	67.3
VLM + Action Expert	**96.8**	42.2	69.5
Action Expert (Ours)	96.4	**43.3**	**69.9**

Results. Table 3 shows that ID2P improves over $\pi_{0.5}$ on most tasks, extending the gains observed in simulation to real-world manipulation. The largest gains occur in towel folding and pen placement, highlighting benefits for both deformable object manipulation and precise placement. Deformable objects involve complex contact dynamics, making action consequences difficult to infer from current observations alone. By distilling future-conditioned knowledge from the IDM, ID2P better captures such action consequence relationships and improves policy learning. Gains in basketball further demonstrate the effectiveness of ID2P across diverse real-world tasks. Real-world rollouts are visualized in Appendix C (Figure 4).

5.4 Ablation Study

We conduct all ablations on SmolVLA using the LIBERO and LIBERO-Plus benchmark. The LIBERO and LIBERO-Plus columns report average success rates, while Avg. denotes their arithmetic mean. Appendix A provides the complete suite-level and perturbation-level results.

Q1: Are action-level and representation-level future supervision effective? To answer this question, we train multiple policies using the same trained IDM while varying the loss terms. As shown in the bottom three rows of Table 4, using either IDM action supervision or representation supervision alone outperforms the baseline without either supervision on both LIBERO and LIBERO-Plus, while combining both achieves the best performance. These results demonstrate that action-level and representation-level future supervision each independently improve policy performance, while their combination provides complementary benefits and yields further gains.

Q2: Does dedicated IDM training improve distillation effectiveness? PFD (Fang et al., 2026) adopts a shared-parameter design in which the privileged branch is not trained as a separate IDM. Motivated by this distinction, we compare our independently trained IDM with a shared-policy teacher within the ID2P framework. As shown in the top four rows of Table 4, supervision from the shared-policy teacher is less effective in our setting, especially when action and representation supervision are combined. In contrast, the separately trained IDM consistently yields stronger performance, suggesting that explicitly optimizing IDM is beneficial for ID2P.

Q3: How does IDM parameter initialization affect performance? Table 5 shows that independent policy initialization performs best on both benchmarks. Initializing the policy from a trained IDM reduces performance on LIBERO and LIBERO-Plus, suggesting that direct parameter transfer introduces IDM-specific bias into policy learning. We therefore initialize the policy independently and transfer action-related future knowledge solely through action and representation distillation.

Table 5: Ablation of policy parameter initialization from the IDM on SmolVLA.

Method	LIBERO	LIBERO-Plus	Avg.
SmolVLA (Shukor et al., 2025)	95.1	36.6	65.9
SmolVLA + ID2P w/ IDM param. init.	93.9	44.4	69.2
SmolVLA + ID2P w/o IDM param. init. (Ours)	**96.6**	**45.9**	**71.3**

Table 8: Ablation of training budget on SmolVLA. Epochs and GPU Hours aggregate all training stages; ID2P allocates 14.0 epochs to each stage. GPU Hours reports total cost on 8 GPUs.

Method	Steps	Epochs	GPU Hours	LIBERO	LIBERO-Plus	Avg.
SmolVLA (Shukor et al., 2025)	30k Policy	14.0	15.6	95.1	36.6	65.9
	60k Policy	28.0	30.3	96.1	38.7	67.4
	80k Policy	37.3	41.4	95.7	39.1	67.4
SmolVLA + ID2P (Ours)	30k IDM + 30k Policy	28.0	40.1	**96.6**	**45.9**	**71.3**

Q4: Does mixing IDM and expert action velocity supervision improve performance? In Section 4.2, we show that $\alpha = 1$ achieves the maximum variance reduction under an ideal IDM. In practice, however, a learned IDM is imperfect and its velocity can deviate from the true future conditioned velocity field. Choosing $\alpha \in (0, 1)$ introduces the unbiased expert action velocity to mitigate such IDM errors. To isolate the effect of α, Table 6 evaluates IDM-to-Policy action distillation only, without representation distillation. As shown, $\alpha = 0.5$ outperforms both endpoints, $\alpha = 0$ and $\alpha = 1$, demonstrating the benefit of mixed supervision. Notably, $\alpha = 1$ still outperforms $\alpha = 0$, showing that IDM action velocity alone provides stronger supervision than pure expert action velocity despite estimation errors, further supporting action-level future supervision.

Q5: Which IDM representations should be distilled? In VLA-based IDMs, the action expert accesses future information, while the VLM often uses bidirectional attention (Shukor et al., 2025; Intelligence et al., 2025), allowing non-future token hidden states to encode future information as well. Thus, both can serve as representation distillation targets. To isolate representation distillation, we disable IDM action distillation by setting $\alpha = 0$. As shown in Table 7, jointly distilling the VLM and action expert performs similarly to distilling the action expert alone, with the latter slightly better. Since action expert representations directly support action velocity prediction, we distill only the action expert for more targeted action relevant supervision.

Q6: Does ID2P remain effective under matched training budgets? ID2P requires training an IDM before policy training. Table 8 compares ID2P with the baseline under matched training budgets. With the same total 28 epochs, ID2P improves the average success rate from 67.4% to 71.3%. Under comparable GPU hours, ID2P also outperforms the 80k-step baseline, with a particularly large gain on LIBERO-Plus. These results show that ID2P's gains cannot be explained by additional training alone, but stem from the future supervision provided by IDM-to-Policy Distillation.

6 Conclusion

Limitations. A limitation of ID2P is the additional IDM training cost compared with standard behavior cloning. Its effectiveness also depends on IDM quality, as imperfect IDMs may introduce biased supervision. Our current evaluation focuses on flow-based policies and selected manipulation tasks, leaving broader policy families and larger scale real-world settings for future work.

Conclusion. We presented ID2P, a training framework that transfers action-related future knowledge from a separately trained IDM to a flow-based policy through action and representation distillation. ID2P provides both action-level and representation-level future supervision without modifying the policy architecture or adding inference overhead. We theoretically established the gradient equivalence and variance reduction properties of IDM action distillation, and extensive simulation and real-world experiments demonstrated consistent gains across model scales and tasks. Overall, ID2P offers a simple and general way to exploit future observations more fully for policy learning.

REFERENCES

Shuanghao Bai, Jing Lyu, Wanqi Zhou, Zhe Li, Dakai Wang, Lei Xing, Xiaoguang Zhao, Pengwei Wang, Zhongyuan Wang, Cheng Chi, et al. Latent reasoning vla: Latent thinking and prediction for vision-language-action models. *arXiv preprint arXiv:2602.01166*, 2026.

Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.026.

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

Yueci Deng, Guiliang Liu, and Kui Jia. Dexworldmodel: Causal latent world modeling towards automated learning of embodied tasks. *arXiv preprint arXiv:2604.16484*, 2026.

Pengcheng Fang, Hongli Chen, and Xiaohao Cai. Privileged foresight distillation: Zero-cost future correction for world action models. *arXiv preprint arXiv:2604.25859*, 2026.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Jiarui Guan, Wenshuai Zhao, Yue Pei, Ziliang Chen, Arno Solin, and Juho Kannala. Point tracking improves world action models. *arXiv preprint arXiv:2605.23856*, 2026.

Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026.

Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. Improving vision-language-action model with online reinforcement learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15665–15672, 2025. doi: 10.1109/ICRA55743.2025.11127299.

John Hammersley. *Monte carlo methods*. Springer Science & Business Media, 2013.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Chengmeng Li, Junjie Wen, Yaxin Peng, Yan Peng, and Yichen Zhu. Pointvla: Injecting the 3d world into vision-language-action models. *IEEE Robotics and Automation Letters*, 11(3):2506–2513, 2026a. doi: 10.1109/LRA.2026.3653303.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026b.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified Video Action Model. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.074.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Mengya Liu, Baoxiong Jia, Jiangyong Huang, Jingze Zhang, and Siyuan Huang. Lara: Latent action representation alignment for vision-language-action models. *arXiv preprint arXiv:2606.07100*, 2026a.

Zhenyang Liu, Yongchong Gu, Yikai Wang, Xiangyang Xue, and Yanwei Fu. Activevla: Injecting active perception into vision-language-action models for precise 3d robotic manipulation. *arXiv preprint arXiv:2601.08325*, 2026b.

Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719*, 2025.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-h0. 7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026.

Jiangran Lyu, Kai Liu, Xuheng Zhang, Haoran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, et al. Lda-1b: Scaling latent dynamics action model via universal embodied data ingestion. *arXiv preprint arXiv:2602.12215*, 2026.

Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv preprint arXiv:2603.10448*, 2026.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi Jim Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick Tree Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al.

Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. Geovla: Empowering 3d representations in vision-language-action models. *arXiv preprint arXiv:2508.09071*, 2025.

Peijun Tang, Shangjin Xie, Baifu Huang, Binyan Sun, Haotian Yang, Kuncheng Luo, Weiqi Jin, Shilin Fang, and Jianan Wang. Towards predictive, aligned, and scalable robot learning. *arXiv preprint arXiv:2607.11270*, 2026.

MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, et al. Motubrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792*, 2026.

Chenyi Wang, Xinkai Wang, Bokai Lin, Jialin Tian, Fucheng Zhang, Cewu Lu, and Lixin Yang. Track4action: Distilling world-centric 3d tracker into vision-language-action policies. *arXiv preprint arXiv:2608.03727*, 2026a.

Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, et al. Openwam: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026b.

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, et al. A pragmatic vla foundation model. *arXiv preprint arXiv:2601.18692*, 2026.

Yilun Xu, Shangyuan Tong, and Tommi Jaakkola. Stable target field for reduced variance score estimation in diffusion models. *arXiv preprint arXiv:2302.00670*, 2023.

Donglin Yang, Yongxing Zhang, Xin Yu, Liang Hou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Renjie Liao. Stable velocity: A variance perspective on flow matching. *arXiv preprint arXiv:2602.05435*, 2026a.

Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, et al. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026b.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026b.

Haoqi Yuan, Zhixuan Liang, Anzhe Chen, Ye Wang, Haoyang Li, Pei Lin, Yiyang Huang, Zixing Lei, Tong Zhang, Jiazhao Zhang, et al. Qwen-robotmanip technical report: Alignment unlocks scale for robotic manipulation foundation models. *arXiv preprint arXiv:2606.17846*, 2026a.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026b.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems*, 38: 24195–24228, 2026a.

Yuyang Zhang, Wenyao Zhang, Zekun Qi, He Zhang, Haitao Lin, Jingbo Zhang, Yao Mu, Xiaokang Yang, Wenjun Zeng, and Xin Jin. Imagewam: Do world action models really need video generation, or just image editing? *arXiv preprint arXiv:2606.19531*, 2026b.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713, 2025. doi: 10.1109/CVPR52734.2025.00166.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 60580–60606, 2026.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 06–09 Nov 2023. URL `https://proceedings.mlr.press/v229/zitkovich23a.html`.

APPENDIX

AI USE STATEMENT

Generative AI tools were used for English translation and language editing. The authors reviewed the AI-assisted text and take responsibility for the final content, claims, and presentation.

A DETAILED ABLATION RESULTS

This section expands the ablations in Tables 4–8. All ablations use SmolVLA and the evaluation protocols from the main experiments.

A.1 FUTURE SUPERVISION

Table 9 compares shared policy and independently trained IDM teachers, with policy parameter initialization from the IDM disabled. Under every distillation setting, the trained IDM improves all four LIBERO suites over its shared policy counterpart. Combining both objectives yields the strongest suite average.

Table 9: Future supervision is ablated on SmolVLA using LIBERO.

IDM Type	$\mathcal{L}_{\text{AD}}$	$\mathcal{L}_{\text{RD}}$	Spatial	Object	Goal	Long	Avg.
–			96.8	98.2	94.4	91.0	95.1
Shared Policy	✓		73.8	63.2	53.0	41.2	57.8
		✓	94.4	97.4	93.4	88.8	93.5
	✓	✓	62.8	32.8	40.0	12.8	37.1
Trained IDM (Ours)	✓		**97.4**	98.4	96.0	92.0	96.0
		✓	96.4	98.6	96.4	**94.0**	96.4
	✓	✓	96.8	**99.4**	**96.8**	93.4	**96.6**

The LIBERO-Plus breakdown in Table 10 further distinguishes the two supervision signals. With the trained IDM, joint distillation yields the strongest Camera, Robot, Language, and Noise results, while representation-only distillation excels under Light and Background changes. The strongest overall robustness comes from combining action and representation distillation.

Table 10: Future supervision is ablated on SmolVLA using LIBERO-Plus.

IDM Type	$\mathcal{L}_{\text{AD}}$	$\mathcal{L}_{\text{RD}}$	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
–			10.0	10.8	46.7	76.1	63.8	11.7	58.0	36.6
Shared Policy	✓		4.3	4.1	25.1	33.0	19.9	7.8	20.2	15.4
		✓	13.8	24.0	53.9	62.7	54.5	30.9	47.2	39.3
	✓	✓	2.9	1.4	21.7	21.9	12.3	5.5	12.7	10.6
Trained IDM (Ours)	✓		9.8	13.9	54.7	77.1	63.5	14.7	**59.3**	39.0
		✓	11.8	21.3	63.7	**77.6**	**68.3**	22.0	56.9	43.3
	✓	✓	**19.2**	**25.7**	**64.0**	76.6	62.6	**33.4**	54.4	**45.9**

A.2 POLICY PARAMETER INITIALIZATION FROM THE IDM

Table 11 evaluates whether to initialize policy parameters from the IDM before distillation. Independent policy initialization matches or improves every LIBERO suite relative to both the baseline and the IDM-initialized variant.

On LIBERO-Plus, independent initialization achieves the best benchmark average and the strongest Camera, Light, and Noise results (Table 12). Its gains on Camera and Noise demonstrate that future knowledge can be transferred effectively through the distillation objectives without parameter transfer.

Table 11: IDM-based policy initialization is ablated on SmolVLA using LIBERO.

Method	Spatial	Object	Goal	Long	Avg.
SmolVLA (Shukor et al., 2025)	**96.8**	98.2	94.4	91.0	95.1
SmolVLA + ID2P w/ IDM param. init.	96.4	96.6	90.6	91.8	93.9
SmolVLA + ID2P w/o IDM param. init. (Ours)	**96.8**	**99.4**	**96.8**	**93.4**	**96.6**

Table 12: IDM-based policy initialization is ablated on SmolVLA using LIBERO-Plus.

Method	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
SmolVLA (Shukor et al., 2025)	10.0	10.8	46.7	76.1	**63.8**	11.7	58.0	36.6
SmolVLA + ID2P w/ IDM param. init.	12.4	**28.6**	**65.6**	74.6	54.3	26.4	**62.0**	44.4
SmolVLA + ID2P w/o IDM param. init. (Ours)	**19.2**	25.7	64.0	**76.6**	62.6	**33.4**	54.4	**45.9**

A.3 Action-Supervision Mixing Weight

We vary α while disabling representation distillation to isolate the effect of action supervision. In Table 13, the mixed target achieves the strongest LIBERO average.

Table 13: The mixing weight α is ablated on SmolVLA using LIBERO.

α	Spatial	Object	Goal	Long	Avg.
0	96.8	98.2	94.4	91.0	95.1
1	96.8	**98.6**	94.6	**92.2**	95.6
0.5 (Ours)	**97.4**	98.4	**96.0**	92.0	**96.0**

On LIBERO-Plus, equal mixing also maximizes the benchmark average and performs best under Light and Layout perturbations (Table 14). These results support combining expert action velocities with future-conditioned IDM predictions to obtain effective supervision across task execution and robustness evaluations.

Table 14: The mixing weight α is ablated on SmolVLA using LIBERO-Plus.

α	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
0	10.0	10.8	46.7	76.1	**63.8**	11.7	58.0	36.6
1	**12.8**	**20.4**	**55.1**	62.6	57.3	**25.4**	49.3	38.4
0.5 (Ours)	9.8	13.9	54.7	**77.1**	63.5	14.7	**59.3**	**39.0**

A.4 Representation Target

To isolate the representation target, we disable action distillation and compare supervision of the VLM, the action expert, or both. Table 15 shows that joint supervision performs best on Goal and Long, while action expert supervision gives the strongest Object result.

Table 15: Representation targets are ablated on SmolVLA using LIBERO.

Representation Target	Spatial	Object	Goal	Long	Avg.
VLM	**97.6**	98.2	96.6	93.2	96.4
VLM + Action Expert	97.2	97.2	**97.0**	**95.8**	**96.8**
Action Expert (Ours)	96.4	**98.6**	96.4	94.0	96.4

Table 16 shows that the action expert achieves the best overall score and leads on Language, Light, Background, and Noise. This robustness advantage supports distilling action-related future knowledge directly into the module responsible for action generation.

Table 16: Representation targets are ablated on SmolVLA using LIBERO-Plus.

Representation Target	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
VLM	10.3	18.7	59.9	70.1	52.5	14.2	56.1	38.1
VLM + Action Expert	**11.9**	**25.6**	62.5	68.2	63.6	21.8	**57.5**	42.2
Action Expert (Ours)	11.8	21.3	**63.7**	**77.6**	**68.3**	**22.0**	56.9	**43.3**

A.5 Training Budget

Table 17 compares standard policy training at three budgets with two-stage ID2P. ID2P has the same aggregate data exposure as the 60k-step baseline and uses slightly fewer GPU hours than the 80k-step baseline. It achieves the highest LIBERO average and the strongest Goal performance.

Table 17: Training budgets are ablated on SmolVLA using LIBERO.

Method	Steps	Epochs	GPU Hours	Spatial	Object	Goal	Long	Avg.
SmolVLA (Shukor et al., 2025)	30k Policy	14.0	15.6	96.8	98.2	94.4	91.0	95.1
	60k Policy	28.0	30.3	96.4	99.2	94.8	**93.8**	96.1
	80k Policy	37.3	41.4	**98.4**	**99.4**	94.4	90.6	95.7
SmolVLA + ID2P (Ours)	30k IDM + 30k Policy	28.0	40.1	96.8	**99.4**	**96.8**	93.4	**96.6**

The LIBERO-Plus results in Table 18 reinforce this advantage under distribution shifts. ID2P outperforms both extended-training baselines in six of seven perturbation categories and achieves the highest benchmark average. The gains span Camera, Robot, Language, Light, Background, and Noise, supporting the value of future supervision under matched data and comparable compute budgets.

Table 18: Training budgets are ablated on SmolVLA using LIBERO-Plus.

Method	Steps	Epochs	GPU Hours	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
SmolVLA (Shukor et al., 2025)	30k Policy	14.0	15.6	10.0	10.8	46.7	76.1	**63.8**	11.7	**58.0**	36.6
	60k Policy	28.0	30.3	13.8	16.4	48.4	66.5	60.0	24.0	57.0	38.7
	80k Policy	37.3	41.4	18.5	19.7	47.6	73.6	55.2	21.9	52.7	39.1
SmolVLA + ID2P (Ours)	30k IDM + 30k Policy	28.0	40.1	**19.2**	**25.7**	**64.0**	**76.6**	62.6	**33.4**	54.4	**45.9**

B Training Hyperparameters

Table 19 summarizes the key ID2P settings for the simulation experiments. Optimization settings apply to both training stages. LIBERO-Plus uses the models trained on LIBERO.

Table 19: Key training hyperparameters. Paired values follow the order LIBERO / RoboTwin.

Hyperparameter	SmolVLA	$\pi_{0.5}$
Optimizer	AdamW	AdamW
Learning rate	10^{-4}	10^{-4}
Final learning rate	2.5×10^{-6}	2.5×10^{-6}
Learning rate schedule	Cosine	Cosine
Warmup steps	1k	1k
Global batch size	128 / 256	128 / 256
Steps per stage	30k	30k
Action chunk length	16	16
Future observation offset	16	16
Image augmentation	None	None

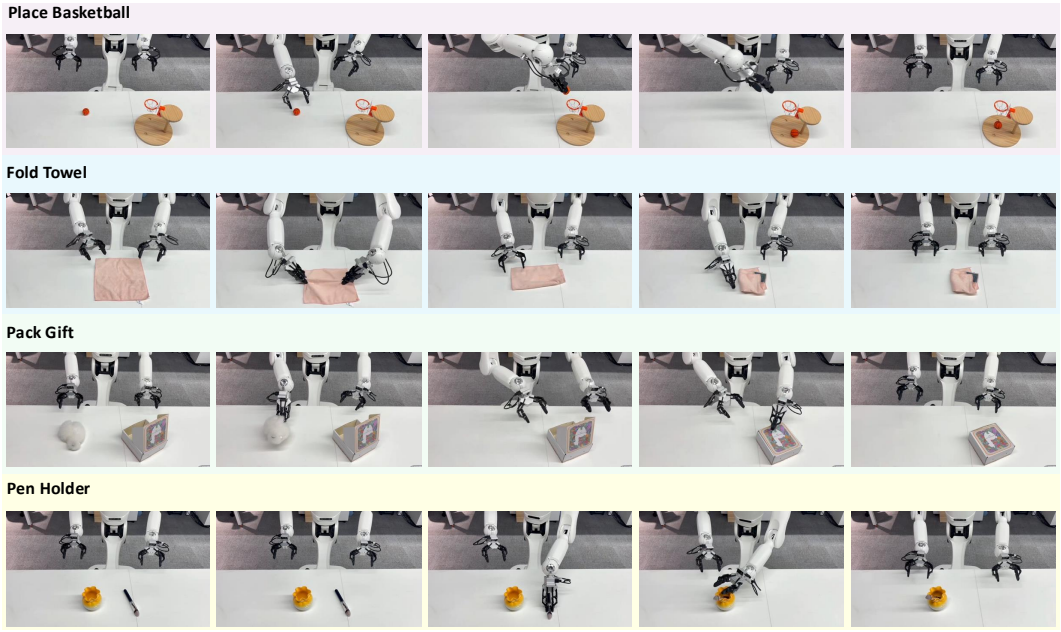

Figure 4: **Real-world task visualizations on the Astribot S1 robot.** Each row shows five frames sampled in temporal order from a representative execution. From top to bottom, the tasks are Place Basketball, Fold Towel, Pack Gift, and Pen Holder.

C Real-World Experimental Setup and Visualizations

We evaluate ID2P on the Astribot S1, a bimanual humanoid robotic platform equipped with one head-mounted RGB camera and two wrist-mounted RGB cameras. The head camera provides a global view of the workspace, while the wrist cameras capture close-range visual observations during object interaction. The policy takes the three RGB streams together with the robot proprioceptive state as input and predicts 34-dimensional actions at 30 Hz. For each task, we train a separate policy for 30k steps using an action chunk length of 16. Both $\pi_{0.5}$ and $\pi_{0.5}$ + ID2P are evaluated over 30 trials per task.

Our real-world benchmark consists of four tabletop manipulation tasks with diverse interaction patterns: *Place Basketball*, where the robot grasps a basketball and places it into a miniature hoop; *Fold Towel*, where the robot manipulates and folds a towel on the table; *Pack Gift*, where the robot places a gift item into a box and closes the package; and *Pen Holder*, where the robot picks up a pen and inserts it into a holder. Together, these tasks cover rigid-object manipulation, deformable-object handling, coordinated bimanual control, and precise object placement, providing a complementary evaluation of the policy under different manipulation requirements.

Figure 4 presents representative real-world rollouts for all four tasks. Each row contains five temporally ordered frames from a single execution, illustrating the progression of the manipulation from the initial scene, through object approach and interaction, to successful task completion. These visualizations provide a qualitative view of the policy's closed-loop behavior and its ability to execute coherent multi-stage manipulation in real-world environments.

D Proofs for IDM-to-Policy Action Distillation

D.1 Proof of Theorem 1

Theorem 1. *Assume that the learned IDM realizes a global minimizer of $\mathcal{L}_{\mathrm{IDM}}$, such that $v_{\widehat{\phi}}(A_t^\tau, \tau, o_t, F_t) = v_\tau(A_t^\tau \mid o_t, F_t)$ a.s. For any $\alpha \in \mathbb{R}$, $\mathcal{L}_{\mathrm{AD}}(\theta; \alpha)$ and $\mathcal{L}_{\mathrm{Policy}}(\theta)$ are equal up to a constant independent of θ. Hence, $\nabla_\theta \mathcal{L}_{\mathrm{AD}}(\theta; \alpha) = \nabla_\theta \mathcal{L}_{\mathrm{Policy}}(\theta)$.*

Proof. To ensure the existence of all expectations and to allow changing the order of integration by Fubini's theorem, we assume that all integrands below are absolutely integrable and that the regularity conditions required to interchange differentiation and expectation are satisfied.

Under the assumption of the theorem, $v_{\widehat{\phi}}(A_t^\tau, \tau, o_t, F_t) = v_\tau(A_t^\tau \mid o_t, F_t)$ a.s. Since stop-gradient does not alter target values, $v^\alpha_{\tau,\widehat{\phi}} = v^\alpha_\tau$ a.s.

Expanding the squared norm,

$$\begin{aligned}\mathcal{L}_{\text{Policy}}(\theta) = \mathbb{E}_{\tau,\, p_E(o_t,A_t),\, p_\tau(A_t^\tau|A_t)} \|v_\theta(A_t^\tau, \tau, o_t)\|^2 \\ - 2\,\mathbb{E}_{\tau,\, p_E(o_t,A_t),\, p_\tau(A_t^\tau|A_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v_\tau(A_t^\tau \mid A_t)\rangle \\ + \mathbb{E}_{\tau,\, p_E(o_t,A_t),\, p_\tau(A_t^\tau|A_t)} \|v_\tau(A_t^\tau \mid A_t)\|^2,\end{aligned} \tag{13}$$

and

$$\begin{aligned}\mathcal{L}_{\text{AD}}(\theta;\alpha) = \mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \|v_\theta(A_t^\tau, \tau, o_t)\|^2 \\ - 2\,\mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v^\alpha_\tau\rangle \\ + \mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \|v^\alpha_\tau\|^2.\end{aligned} \tag{14}$$

Since $v_\theta(A_t^\tau, \tau, o_t)$ does not depend on F_t,

$$\begin{aligned}\mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \|v_\theta(A_t^\tau, \tau, o_t)\|^2 \\ = \mathbb{E}_{\tau,\, p_E(o_t,A_t),\, p_\tau(A_t^\tau|A_t)} \|v_\theta(A_t^\tau, \tau, o_t)\|^2.\end{aligned} \tag{15}$$

By definition, $v^\alpha_\tau = (1-\alpha)v_\tau(A_t^\tau \mid A_t) + \alpha v_\tau(A_t^\tau \mid o_t, F_t)$. Therefore,

$$\begin{aligned}&\mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v^\alpha_\tau\rangle \\ &= (1-\alpha)\mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v_\tau(A_t^\tau \mid A_t)\rangle \\ &\quad + \alpha\mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v_\tau(A_t^\tau \mid o_t, F_t)\rangle\,.\end{aligned} \tag{16}$$

For the second term,

$$\begin{aligned}&\mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v_\tau(A_t^\tau \mid o_t, F_t)\rangle \\ &= \mathbb{E}_{\tau,\, p_\tau(A_t^\tau, o_t, F_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v_\tau(A_t^\tau \mid o_t, F_t)\rangle \\ &= \mathbb{E}_{\tau,\, p_\tau(A_t^\tau, o_t, F_t)} \left\langle v_\theta(A_t^\tau, \tau, o_t), \mathbb{E}_{p_\tau(A_t|A_t^\tau, o_t, F_t)} v_\tau(A_t^\tau \mid A_t)\right\rangle \\ &= \mathbb{E}_{\tau,\, p_\tau(A_t^\tau, o_t, F_t)} \mathbb{E}_{p_\tau(A_t|A_t^\tau, o_t, F_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v_\tau(A_t^\tau \mid A_t)\rangle \\ &= \mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v_\tau(A_t^\tau \mid A_t)\rangle\,.\end{aligned} \tag{17}$$

Substituting equation 17 into equation 16,

$$\begin{aligned}&\mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v^\alpha_\tau\rangle \\ &= \mathbb{E}_{\tau,\, p_E(o_t,A_t),\, p_\tau(A_t^\tau|A_t)} \langle v_\theta(A_t^\tau, \tau, o_t), v_\tau(A_t^\tau \mid A_t)\rangle\,.\end{aligned} \tag{18}$$

Hence, for all $\alpha \in \mathbb{R}$,

$$\begin{aligned}\mathcal{L}_{\text{Policy}}(\theta) - \mathcal{L}_{\text{AD}}(\theta;\alpha) = \mathbb{E}_{\tau,\, p_E(o_t,A_t),\, p_\tau(A_t^\tau|A_t)} \|v_\tau(A_t^\tau \mid A_t)\|^2 \\ - \mathbb{E}_{\tau,\, p_E(o_t,F_t,A_t),\, p_\tau(A_t^\tau|A_t)} \|v^\alpha_\tau\|^2.\end{aligned} \tag{19}$$

The right-hand side is independent of θ. Therefore, $\mathcal{L}_{\text{AD}}(\theta;\alpha)$ and $\mathcal{L}_{\text{Policy}}(\theta)$ are equal up to a constant independent of θ. Hence,

$$\nabla_\theta \mathcal{L}_{\text{AD}}(\theta;\alpha) = \nabla_\theta \mathcal{L}_{\text{Policy}}(\theta),$$

which proves the theorem. □

D.2 Proof of Theorem 2

Theorem 2. *For a fixed $\tau \in (0,1)$, assume that the learned IDM satisfies $v_{\widehat{\phi}}(A_t^\tau, \tau, o_t, F_t) = v_\tau(A_t^\tau \mid o_t, F_t)$ a.s. Then, for any $\alpha \in \mathbb{R}$,*

$$\mathcal{V}_{\mathrm{Policy}}(\tau) - \mathcal{V}_{\mathrm{AD}}(\tau;\alpha) = \alpha(2-\alpha)\mathcal{V}_{\mathrm{res}}(\tau). \tag{10}$$

Hence, if $\mathcal{V}_{\mathrm{res}}(\tau) > 0$, then $\mathcal{V}_{\mathrm{AD}}(\tau;\alpha) < \mathcal{V}_{\mathrm{Policy}}(\tau)$ for every $\alpha \in (0,2)$.

Proof. Under the assumption of the theorem, $v_{\widehat{\phi}}(A_t^\tau, \tau, o_t, F_t) = v_\tau(A_t^\tau \mid o_t, F_t)$ a.s. Since stop-gradient does not alter target values, $v^\alpha_{\tau,\widehat{\phi}} = v^\alpha_\tau$ a.s.

First,

$$\begin{aligned}
&\mathrm{Cov}_{p_\tau(A_t \mid A_t^\tau, o_t)}\left(v_\tau(A_t^\tau \mid A_t)\right) \\
&= \mathbb{E}_{p_\tau(A_t \mid A_t^\tau, o_t)}\Big[\big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t)\big)\big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t)\big)^\top\Big] \\
&= \mathbb{E}_{p_\tau(A_t, F_t \mid A_t^\tau, o_t)}\Big[\big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t, F_t) + v_\tau(A_t^\tau \mid o_t, F_t) - v_\tau(A_t^\tau \mid o_t)\big) \\
&\qquad\qquad \cdot \big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t, F_t) + v_\tau(A_t^\tau \mid o_t, F_t) - v_\tau(A_t^\tau \mid o_t)\big)^\top\Big].
\end{aligned} \tag{20}$$

The first term is

$$\begin{aligned}
&\mathbb{E}_{p_\tau(A_t, F_t \mid A_t^\tau, o_t)}\Big[\big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t, F_t)\big)\big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t, F_t)\big)^\top\Big] \\
&= \mathbb{E}_{p_\tau(F_t \mid A_t^\tau, o_t)}\left[\mathrm{Cov}_{p_\tau(A_t \mid A_t^\tau, o_t, F_t)}\left(v_\tau(A_t^\tau \mid A_t)\right)\right].
\end{aligned} \tag{21}$$

For the second term,

$$\begin{aligned}
&\mathbb{E}_{p_\tau(F_t \mid A_t^\tau, o_t)} v_\tau(A_t^\tau \mid o_t, F_t) \\
&= \mathbb{E}_{p_\tau(F_t \mid A_t^\tau, o_t)} \mathbb{E}_{p_\tau(A_t \mid A_t^\tau, o_t, F_t)} v_\tau(A_t^\tau \mid A_t) \\
&= \mathbb{E}_{p_\tau(A_t, F_t \mid A_t^\tau, o_t)} v_\tau(A_t^\tau \mid A_t) \\
&= \mathbb{E}_{p_\tau(A_t \mid A_t^\tau, o_t)} v_\tau(A_t^\tau \mid A_t) = v_\tau(A_t^\tau \mid o_t).
\end{aligned} \tag{22}$$

Hence,

$$\begin{aligned}
&\mathbb{E}_{p_\tau(A_t, F_t \mid A_t^\tau, o_t)}\Big[\big(v_\tau(A_t^\tau \mid o_t, F_t) - v_\tau(A_t^\tau \mid o_t)\big)\big(v_\tau(A_t^\tau \mid o_t, F_t) - v_\tau(A_t^\tau \mid o_t)\big)^\top\Big] \\
&= \mathrm{Cov}_{p_\tau(F_t \mid A_t^\tau, o_t)}\left(v_\tau(A_t^\tau \mid o_t, F_t)\right).
\end{aligned} \tag{23}$$

The third term is

$$\begin{aligned}
&\mathbb{E}_{p_\tau(A_t, F_t \mid A_t^\tau, o_t)}\Big[\big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t, F_t)\big)\big(v_\tau(A_t^\tau \mid o_t, F_t) - v_\tau(A_t^\tau \mid o_t)\big)^\top\Big] \\
&= \mathbb{E}_{p_\tau(F_t \mid A_t^\tau, o_t)}\Bigg[\mathbb{E}_{p_\tau(A_t \mid A_t^\tau, o_t, F_t)}\big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t, F_t)\big) \\
&\qquad\qquad \cdot \big(v_\tau(A_t^\tau \mid o_t, F_t) - v_\tau(A_t^\tau \mid o_t)\big)^\top\Bigg] = 0.
\end{aligned} \tag{24}$$

The fourth term is

$$\begin{aligned}
&\mathbb{E}_{p_\tau(A_t, F_t \mid A_t^\tau, o_t)}\Big[\big(v_\tau(A_t^\tau \mid o_t, F_t) - v_\tau(A_t^\tau \mid o_t)\big)\big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t, F_t)\big)^\top\Big] \\
&= \mathbb{E}_{p_\tau(F_t \mid A_t^\tau, o_t)}\Bigg[\big(v_\tau(A_t^\tau \mid o_t, F_t) - v_\tau(A_t^\tau \mid o_t)\big) \\
&\qquad\qquad \cdot \mathbb{E}_{p_\tau(A_t \mid A_t^\tau, o_t, F_t)}\big(v_\tau(A_t^\tau \mid A_t) - v_\tau(A_t^\tau \mid o_t, F_t)\big)^\top\Bigg] = 0.
\end{aligned} \tag{25}$$

Substituting equation 21– equation 25 into equation 20,

$$\begin{aligned}&\mathrm{Cov}_{p_\tau(A_t|A_t^\tau,o_t)}\left(v_\tau(A_t^\tau \mid A_t)\right)\\&=\mathrm{Cov}_{p_\tau(F_t|A_t^\tau,o_t)}\left(v_\tau(A_t^\tau \mid o_t,F_t)\right)\\&\quad+\mathbb{E}_{p_\tau(F_t|A_t^\tau,o_t)}\left[\mathrm{Cov}_{p_\tau(A_t|A_t^\tau,o_t,F_t)}\left(v_\tau(A_t^\tau \mid A_t)\right)\right].\end{aligned}\tag{26}$$

Next,

$$\begin{aligned}&\mathbb{E}_{p_\tau(A_t,F_t|A_t^\tau,o_t)}v_\tau^\alpha\\&=(1-\alpha)\mathbb{E}_{p_\tau(A_t|A_t^\tau,o_t)}v_\tau(A_t^\tau \mid A_t)+\alpha\mathbb{E}_{p_\tau(F_t|A_t^\tau,o_t)}v_\tau(A_t^\tau \mid o_t,F_t)\\&=v_\tau(A_t^\tau \mid o_t).\end{aligned}\tag{27}$$

Therefore,

$$\begin{aligned}&\mathrm{Cov}_{p_\tau(A_t,F_t|A_t^\tau,o_t)}\left(v_{\tau,\widehat{\phi}}^\alpha\right)=\mathrm{Cov}_{p_\tau(A_t,F_t|A_t^\tau,o_t)}\left(v_\tau^\alpha\right)\\&=\mathbb{E}_{p_\tau(A_t,F_t|A_t^\tau,o_t)}\left[\left(v_\tau^\alpha-v_\tau(A_t^\tau \mid o_t)\right)\left(v_\tau^\alpha-v_\tau(A_t^\tau \mid o_t)\right)^\top\right]\\&=\mathbb{E}_{p_\tau(A_t,F_t|A_t^\tau,o_t)}\Bigg[\Big((1-\alpha)\big(v_\tau(A_t^\tau \mid A_t)-v_\tau(A_t^\tau \mid o_t,F_t)\big)+v_\tau(A_t^\tau \mid o_t,F_t)-v_\tau(A_t^\tau \mid o_t)\Big)\\&\qquad\cdot\Big((1-\alpha)\big(v_\tau(A_t^\tau \mid A_t)-v_\tau(A_t^\tau \mid o_t,F_t)\big)+v_\tau(A_t^\tau \mid o_t,F_t)-v_\tau(A_t^\tau \mid o_t)\Big)^\top\Bigg]\\&=(1-\alpha)^2\mathbb{E}_{p_\tau(F_t|A_t^\tau,o_t)}\left[\mathrm{Cov}_{p_\tau(A_t|A_t^\tau,o_t,F_t)}\left(v_\tau(A_t^\tau \mid A_t)\right)\right]\\&\quad+\mathrm{Cov}_{p_\tau(F_t|A_t^\tau,o_t)}\left(v_\tau(A_t^\tau \mid o_t,F_t)\right),\end{aligned}\tag{28}$$

where the two cross terms are given by equation 24 and equation 25.

Subtracting equation 28 from equation 26,

$$\begin{aligned}&\mathrm{Cov}_{p_\tau(A_t|A_t^\tau,o_t)}\left(v_\tau(A_t^\tau \mid A_t)\right)-\mathrm{Cov}_{p_\tau(A_t,F_t|A_t^\tau,o_t)}\left(v_{\tau,\widehat{\phi}}^\alpha\right)\\&=\alpha(2-\alpha)\mathbb{E}_{p_\tau(F_t|A_t^\tau,o_t)}\left[\mathrm{Cov}_{p_\tau(A_t|A_t^\tau,o_t,F_t)}\left(v_\tau(A_t^\tau \mid A_t)\right)\right].\end{aligned}\tag{29}$$

Taking the trace and expectation over $p_\tau(A_t^\tau, o_t)$,

$$\begin{aligned}\mathcal{V}_{\mathrm{Policy}}(\tau)-\mathcal{V}_{\mathrm{AD}}(\tau;\alpha)&=\alpha(2-\alpha)\mathbb{E}_{p_\tau(A_t^\tau,o_t,F_t)}\,\mathrm{Tr}\left[\mathrm{Cov}_{p_\tau(A_t|A_t^\tau,o_t,F_t)}\left(v_\tau(A_t^\tau \mid A_t)\right)\right]\\&=\alpha(2-\alpha)\mathcal{V}_{\mathrm{res}}(\tau),\end{aligned}\tag{30}$$

which proves equation 10.

If $\mathcal{V}_{\mathrm{res}}(\tau) > 0$ and $\alpha \in (0, 2)$, then $\alpha(2-\alpha) > 0$, and hence $\mathcal{V}_{\mathrm{AD}}(\tau;\alpha) < \mathcal{V}_{\mathrm{Policy}}(\tau)$. □