PHI-WM 1.0 ACTEFFECT: FROM PREDICTIVE FORESIGHT TO CONSEQUENCE FEEDBACK IN ROBOT LEARNING

SunRisingAI Lab

ABSTRACT

Robot actions are interventions on the physical world, and their quality is ultimately determined by the state transitions and task outcomes they induce. Yet most robot policies are trained primarily to match demonstrated actions, while predictive world models are typically used as foresight for action generation rather than as feedback on the policy's own proposals. We present **ActEffect**, an action-first robot model that uses a controlled world model to evaluate the predicted consequences of proposed actions and guide progressive action refinement. ActEffect uses vision-language representations for task-conditioned action generation and a separate frozen DINO latent space for consequence modeling. Its policy exposes complete feedforward, coarse, and refined action proposals, which are evaluated by a shared controlled world model under the same current state. A model-based counterfactual consequence-ranking objective encourages successive proposals to produce predicted futures progressively closer to the demonstrated future, converting future prediction into direct training feedback. The controlled world model is used only during training and removed at deployment, requiring no future prediction, world-model rollout, or candidate consequence evaluation at test time. On three established and complementary benchmarks, ActEffect achieves average success rates of 98.8% on LIBERO, 80.3% on LIBERO-PLUS, and 67.5% on RoboCasa-GR1, attaining state-of-the-art performance.

1 INTRODUCTION

A robot action is not merely an output prediction, but an intervention that changes the physical world. Its quality is therefore ultimately determined by the state transition and task outcome it induces, rather than only by its numerical similarity to a demonstrated action. This distinction is particularly important in contact-rich manipulation and long-horizon tasks, where small differences in pose, velocity, contact timing, or action sequence can lead to substantially different outcomes. Although imitation learning has enabled increasingly capable and general robot policies Chi et al. (2025); Kim et al. (2025b); Bjorck et al. (2025); Li et al. (2026a), minimizing action prediction error alone does not necessarily guarantee successful physical execution. A central challenge is thus to make robot policies aware not only of what action should be generated, but also of how that action will change the future.

Recent robot learning systems have advanced along two complementary directions. Vision-Language-Action (VLA) policies Kim et al. (2025b); Intelligence et al. (2025); Bjorck et al. (2025); Zheng et al. (2026); Yuan et al. (2026a) leverage pretrained multimodal representations to connect task understanding with continuous action generation, while World-Action Models (WAMs) Kim et al. (2026); Bi et al. (2026); Ye et al. (2026b); Li et al. (2026a); Yuan et al. (2026b) augment policy learning with predictions of future visual frames or latent representations. Despite substantial progress, predictive representations still primarily serve as *foresight* for action generation, through either policy conditioning Yuan et al. (2026b); Ma et al. (2026) or coupling with the generation process Li et al. (2026a); Ye et al. (2026b); Bi et al. (2026). This leaves a question comparatively underexplored: can a model predict the consequence of the policy's *own proposed action* and use it as feedback to improve the proposal itself? This question follows from the temporal-causal structure of physical interaction, in which the current state is followed by an action intervention and then its resulting outcome. Such a perspective naturally calls for shifting future prediction from foresight to

feedback—from supplying predictive context for action generation to evaluating proposed actions through their predicted consequences and guiding their refinement. Unlike conventional model-based planning Hansen et al. (2024); Wang et al. (2026); Zhang et al. (2026a), which evaluates candidate actions through model rollouts at deployment, our goal is to use consequence evaluation during training, so that its benefit is incorporated into a direct policy requiring neither world model rollout nor candidate search at test time.

Using the predicted consequences of a policy's own action proposals as training-time feedback requires more than merely forecasting a plausible future from the current observation. First, the feedback world model should be *controllable*: when the current state is held fixed, its predicted consequences should vary meaningfully with the proposed action, rather than being confounded by task-semantic information from other modalities. Second, such a controlled world model provides a shared consequence space in which alternative action proposals, whose outcomes cannot be jointly observed from the same physical state, can be evaluated through their predicted consequences, enabling model-based counterfactual comparison. Third, the resulting consequence comparison should provide a learning signal that directly improves successive action proposals.

To this end, we introduce **ActEffect**, a robot policy that uses a controlled world model to evaluate the predicted consequences of its own action proposals and convert their counterfactual comparison into training-time feedback for progressive action refinement. To support controllable consequence modeling, we decouple the world model from the vision-language task representation used for policy conditioning and predict action-induced future states in a separate frozen DINO latent space Siméoni et al. (2025), reducing interference from instruction semantics when modeling physical state transitions. We further introduce a counterfactual consequence ranking objective that evaluates multiple action proposals under the same current state in this shared consequence space, encouraging different action interventions to induce distinguishable predicted consequences and successive proposals to yield progressively better outcomes. We incorporate this consequence feedback as an auxiliary training objective to optimize an iterative action regression policy, instantiated with Minimal Iterative Policy (MIP) Pan et al. (2026), thereby directly guiding the progressive refinement of its action proposals. The world model is used only during training and removed at deployment, enabling ActEffect to directly output refined actions without world-model rollout or candidate evaluation at test time. We evaluate ActEffect on three complementary benchmarks, including LIBERO Liu et al. (2023), LIBERO-PLUS Fei et al. (2025), and RoboCasa-GR1 Nasiriany et al. (2024), covering multi-task manipulation, controlled distribution shifts, and humanoid bimanual manipulation. ActEffect achieves average success rates of 98.8%, 80.3%, and 67.5% on the three benchmarks, respectively, demonstrating state-of-the-art performance, while extensive ablations further validate the effectiveness of the proposed components.

Specifically, our contributions can be summarized as follows:

1. We introduce ActEffect, a robot policy learning framework that follows the temporal-casual structure of physical interaction, which shifts world modeling from foresight for action generation to training-time feedback for refining the policy's own action proposals.
2. To enable controllable consequence feedback, we decouple world modeling from vision-language representations in a separate DINO latent space to reduce task-semantic interference, and introduce a counterfactual consequence ranking objective that compares alternative action proposals under the same state to guide progressive action refinement.
3. Experiments across three established robot manipulation benchmarks demonstrate state-of-the-art performance and strong robustness, while ablation studies further validate the effectiveness of the proposed components.

2 Related Work

2.1 Vision Language Action Models

Vision-Language-Action models (VLAs) extend pretrained vision-language models to robotic control, leveraging their rich semantic representations to map visual observations and language instructions to executable robot actions Kim et al. (2025b); Black et al. (2024); Intelligence et al. (2025); Bjorck et al. (2025); Zheng et al. (2026); Yuan et al. (2026a); Yang et al. (2026). With large-scale

robot data and pretrained multimodal backbones, recent VLAs have demonstrated strong instruction following, generalization, and multi-task manipulation capabilities Intelligence et al. (2025); Yuan et al. (2026a). However, most VLAs are fundamentally trained to map the current task context directly to demonstrated actions, without modeling how the predicted actions affect the future state of the environment. Policy learning is thus driven almost entirely by imitation supervision, leaving the policy without feedback on the consequences its own actions induce—precisely the signal needed when action quality is determined by physical outcome rather than proximity to a demonstration.

2.2 World Action Models

World-Action Models (WAMs) address this gap by bringing predictive world modeling into generalist robot policies, jointly or sequentially modeling future scene evolution and robot actions Zhu et al. (2025); Bi et al. (2026); Kim et al. (2026); Ye et al. (2026b); Li et al. (2026a). Because explicitly generating future visual frames is costly at inference, recent approaches increasingly exploit predictive representations without decoding full future observations: DiT4DiT Ma et al. (2026) conditions action generation on intermediate features from video denoising, ImageWAM Zhang et al. (2026b) extracts predictive representations from an image-editing model without decoding the resulting image, and Fast-WAM Yuan et al. (2026b) shows that video modeling can benefit the policy even when explicit future prediction is removed at test time. Despite these different realizations, world modeling here plays a single role: it enriches the representation or context used to generate an action Bi et al. (2026); Li et al. (2026a); Yuan et al. (2026b), providing foresight before an action is committed rather than evaluating the consequences of the policy's own proposals.

More recent studies use future prediction not as a precondition but as a signal that shapes the policy's own outputs. A shared characteristic of this line is that the world modeling is aligned within the vision-language or policy representation space, coupled with action generation implicitly Zheng et al. (2025); Won et al. (2026); Ye et al. (2026a). FLARE Zheng et al. (2025) aligns the intermediate features of a diffusion policy with the latent embeddings of future observations. DUST Won et al. (2026) maintains separate action and vision streams in a dual-stream diffusion transformer, coupling them through shared attention with asynchronous denoising. GigaWorld-Policy Ye et al. (2026a) conditions future-video generation on the policy's predicted actions during training, entangling the predicted consequence with action generation in a shared representation. While effective, these approaches share two properties that limit their use as feedback. First, the future is aligned in a space entangled with task semantics, so the predicted consequence reflects task-conditioned scene evolution rather than the isolated effect of the action intervention. Second, the coupling is implicit—consequences are absorbed into the alignment or denoising process rather than explicitly evaluated—so the model is not required to distinguish how alternative actions, under an identical current state, lead to different outcomes. Our work departs on both points: we evaluate consequences in a separate, task-agnostic DINOv3 latent space, and we make the comparison explicit, ranking the predicted consequences of multiple action proposals under the same current state. This decoupled, explicit formulation provides a more direct feedback signal for action refinement than implicit alignment or denoising-based coupling, while requiring no additional inference at deployment.

3 Preliminaries

3.1 Problem Setup

Let o_t^v denote the current multi-view visual observation, o_t^p the proprioceptive observation, and l the language instruction at time t. The policy predicts an action chunk $a_{t:t+H-1}$, where H is the action horizon. We denote the demonstrated action chunk by $a^\star_{t:t+H-1}$, and the corresponding future visual observations by H by $o^v_{t+1:t+H}$.

The standard imitation policy, as commonly instantiated by VLA models, directly maps the current observation and language instruction to an action chunk as $\pi_\theta(a_{t:t+H-1} \mid o_t, l)$ Bjorck et al. (2025); Intelligence et al. (2025). WAMs extend this formulation by jointly modeling actions and future observations as $\pi_\theta(o_{t+1:t+H}, a_{t:t+H-1} \mid o_t, l)$ Lyu et al. (2026); Bi et al. (2026); Ye et al. (2026b). This joint modeling admits two complementary factorizations, corresponding to different roles of

future prediction in policy learning:

$$\begin{aligned} p(o_{t+1:t+H}, a_{t:t+H-1} \mid o_t, l) &= p(o_{t+1:t+H} \mid o_t, l)\pi_\theta(a_{t:t+H-1} \mid o_{t:t+H}, l), \quad \text{(Foresight)} \\ &= \pi_\theta(a_{t:t+H-1} \mid o_t, l)p(o_{t+1:t+H} \mid o_t, a_t, l). \quad \text{(Feedback)} \end{aligned} \tag{1}$$

The first factorization predicts future scene evolution before action generation and uses the predicted future as additional policy context, corresponding to the role of world modeling as *foresight* that underlies many existing approaches Yuan et al. (2026b); Ma et al. (2026); Zhang et al. (2026b). In this work, we focus on the second factorization, which follows the temporal-causal structure of physical interaction: an action is first proposed from the current task context, and its predicted consequence is subsequently modeled to provide feedback for refining the policy's action proposal.

3.2 Minimal Iterative Policy

Minimal Iterative Policy (MIP) Pan et al. (2026) is a regression policy that retains two key ingredients of diffusion and flow matching policies, namely noise-perturbed training and iterative action generation, while reducing the denoising process to only two supervised prediction stages. This lightweight refinement process achieves competitive control performance while exposing explicit intermediate action proposals during training.

Given an expert action $a^\star$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, MIP constructs an intermediate action: $I_\tau = \tau a^\star + (1-\tau)\epsilon$, and supervises two action predictions:

$$\mathcal{L}_{\text{MIP}} = \|\pi_\theta(c, 0, 0) - a^\star\|_2^2 + \|\pi_\theta(c, I_\tau, \tau) - a^\star\|_2^2, \tag{2}$$

where c denotes the policy condition. At inference time, stochastic perturbation is removed and the two stages are executed sequentially:

$$a^0 = \pi_\theta(c, 0, 0), \qquad a^1 = \pi_\theta(c, \tau a^0, \tau). \tag{3}$$

Here, a^0 and a^1 are both complete action proposals rather than intermediate noisy states. This property is particularly suitable for our setting, as it exposes successive policy proposals whose predicted consequences can be independently evaluated and compared by the controlled world model during training. In contrast, standard diffusion or flow-matching policies typically supervise individual noisy states at sampled timesteps and do not explicitly expose successive complete action proposals during training.

4 Method

4.1 Model Overview

As illustrated in Fig. 1, ActEffect consists of a vision-language policy for task-conditioned action generation and a controlled world model for training-time consequence evaluation. The policy produces successive complete action proposals through MIP, while the world model predicts their action-conditioned consequences in a separate frozen DINOv3 latent space Siméoni et al. (2025). We instantiate the vision-language backbone with Qwen3-VL 4B Bai et al. (2025) and use a DiT action head Bjorck et al. (2025) following the MIP formulation. The controlled world model is implemented as a compact 4-layer Transformer with a hidden dimension of 768, while the DINOv3 encoder remains frozen throughout training.

We append H_a learnable action-query tokens $\mathcal{Q}$ to the language instruction and process them together with the current visual observation using the vision-language backbone. Let q_t denote the resulting task representation and q_t^{act} the hidden states corresponding to the action-query tokens. In parallel, the current visual observation is encoded by the frozen DINOv3 encoder into a visual latent representation z_t, while the proprioceptive observation is mapped by a lightweight state encoder:

$$(q_t, q_t^{\text{act}}) = E_{\text{VLM}}(o_t^v, [l; \mathcal{Q}]), \quad z_t = \text{sg}(E_{\text{DINO}}(o_t^v)), \quad e_t^p = E_{\text{State}}(o_t^p). \tag{4}$$

The DINO tokens are projected to the policy hidden dimension and combined with the vision-language and proprioceptive representations to form the condition for MIP to generate action:

$$c_t = [q_t; z_t; e_t^p]. \tag{5}$$

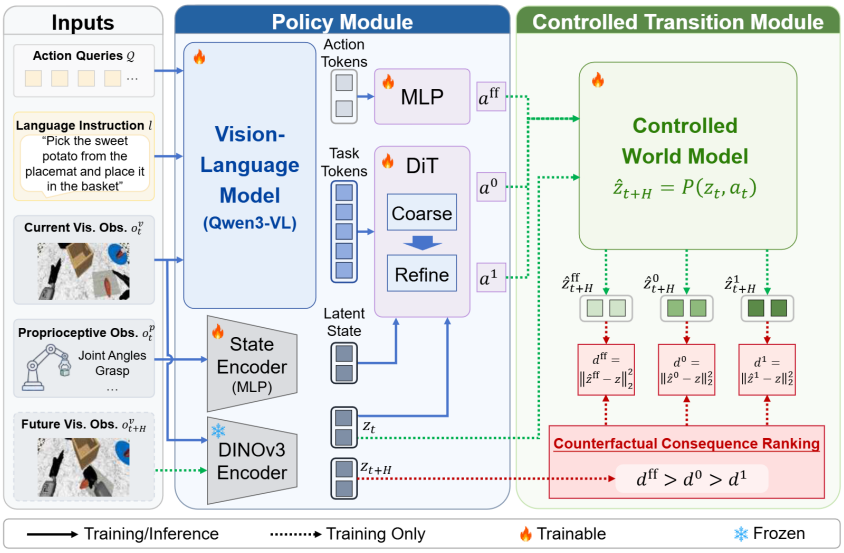

Figure 1: **Model Architecture of ActEffect.** The policy generates feedforward, coarse, and refined action proposals, while a controlled world model predicts their consequences in a frozen DINOv3 latent space and provides counterfactual consequence-ranking feedback during training; the world model is removed at inference.

The action queries additionally provide a direct feedforward pathway from the vision-language representation to the continuous action space, producing a complete feedforward proposal:

$$a_t^{\text{ff}} = \text{MLP}(q_t^{\text{act}}). \tag{6}$$

This pathway encourages the vision-language representation to retain action-relevant information and provides an initial coarse action hypothesis for subsequent consequence evaluation Kim et al. (2025a); Intelligence et al. (2025); Zhong et al. (2026). In parallel, the DiT action head follows the two-stage MIP refinement process introduced in Sec. 3.2, producing a coarse proposal a_t^0 followed by a refined proposal a_t^1. During training, the refinement stage is conditioned on the noise-perturbed expert action prescribed by MIP, whereas at inference it is conditioned on the coarse prediction a_t^0. ActEffect therefore exposes three complete action proposals,

$$\mathcal{A}_t = \{a_t^{\text{ff}}, a_t^0, a_t^1\},$$

corresponding to the feedforward, coarse, and refined stages. These proposals share the same current state and task context but represent different action interventions, providing the alternatives required for counterfactual consequence evaluation by the controlled world model.

4.2 Controlled Consequence Modeling

The controlled world model is designed to evaluate the consequence induced by a proposed action, rather than to use future prediction as additional context for generating that action. We model the controlled transition as $p(z_{t+1:t+H} \mid z_t, a_t)$, where a_t serves as the control variable governing the transition from the current state to its future consequence.

This formulation naturally separates task-semantic reasoning from consequence modeling. The language instruction determines which action should be selected to accomplish the task. Once the current state and action are fixed, however, the resulting transition is determined by their interaction and should no longer depend on the language instruction. Accordingly, the controlled transition satisfies

$$p(z_{t+1:t+H} \mid z_t, a_t, l) = p(z_{t+1:t+H} \mid z_t, a_t). \tag{7}$$

This property motivates modeling action consequences in a representation space separated from the task-semantic space used for action generation. Language-conditioned reasoning remains within

the vision-language policy, while the controlled world model receives only the current visual latent state and the proposed action. Such decoupling prevents task semantics from directly entering consequence evaluation and allows the model to focus on action-induced state transitions.

We perform consequence prediction in the frozen DINOv3 latent space. The demonstrated future observations are encoded as the target representation $z_{t+1:t+H} = \mathrm{sg}\left(\mathrm{DINO}(o_{t+1:t+H})\right)$. For each proposal $a_t^i \in \mathcal{A}_t$, where $i \in \{\mathrm{ff}, 0, 1\}$, the controlled world model predicts

$$\hat{z}^i_{t+1:t+H} = P(z_t, a_t^i). \tag{8}$$

The same controlled world model F_ϕ is shared across all proposals, such that different actions are evaluated under the same transition model and in the same consequence space. The complete action chunk is projected into a control embedding and fused with the current DINO tokens before being processed by the controlled world model.

For each proposal, we measure the discrepancy between its predicted consequence and the demonstrated future as

$$d_t^i = \frac{1}{HKD} \left\| \hat{z}^i_{t+1:t+H} - z_{t+1:t+H} \right\|_2^2, \tag{9}$$

where K and D denote the number and dimensionality of DINO tokens, respectively. A smaller d_t^i indicates that the consequence induced by a_t^i is predicted to be closer to the demonstrated future.

4.3 Counterfactual Consequence Ranking

Only one action can be executed from a given physical state, since executing an action changes the subsequent state. The controlled world model instead enables multiple candidate interventions to be evaluated while holding the current latent state z_t fixed. We therefore evaluate a_t^{ff}, a_t^0, and a_t^1 independently under the same current state, yielding a model-based counterfactual comparison among alternative action proposals.

Because these proposals correspond to progressive stages of policy refinement, their predicted consequences are expected to improve accordingly:

$$d_t^1 < d_t^0 < d_t^{\mathrm{ff}}. \tag{10}$$

We enforce this ordering through pairwise consequence ranking. Let $\mathcal{P} = \{(1, 0), (0, \mathrm{ff})\}$, where (i, j) indicates that proposal i should yield a better consequence than proposal j. The ranking objective is

$$\mathcal{L}_{\mathrm{rank}} = \sum_{(i,j)\in\mathcal{P}} \beta_{ij} \,\mathrm{softplus}\left(\frac{d_t^i - \mathrm{sg}(d_t^j) + m_{ij}}{T_{ij}} \right), \tag{11}$$

where m_{ij} and T_{ij} denote the ranking margin and temperature, respectively, and β_{ij} controls the contribution of each pairwise comparison.

We stop gradients through the reference discrepancy d_t^j. This prevents the ranking objective from being satisfied by simply increasing the error of the weaker proposal and instead encourages the preferred proposal to move closer to the demonstrated future. In this way, the predicted consequences provide explicit counterfactual feedback for progressive action refinement.

4.4 Training and Inference

We jointly optimize action prediction, controlled consequence modeling, and counterfactual consequence ranking. The feedforward proposal is directly supervised by the demonstrated action chunk through $\mathcal{L}_{\mathrm{ff}} = \|a_t^{\mathrm{ff}} - a_t^\star\|_1$, while the coarse and refined proposals are optimized using the supervised iterative objective $\mathcal{L}_{\mathrm{MIP}}$ introduced in Sec. 3.2.

The controlled world model is trained over all three action proposals using

$$\mathcal{L}_{\mathrm{pred}} = \sum_{i\in\{\mathrm{ff},0,1\}} \alpha_i d_t^i, \tag{12}$$

where α_i controls the prediction weight for each proposal. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{MIP}} + \lambda_{\mathrm{ff}}\mathcal{L}_{\mathrm{ff}} + \lambda_{\mathrm{pred}}\mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}}. \tag{13}$$

Table 1: **Success rate (%) on LIBERO.** Each suite contains 10 tasks evaluated over 50 trials per task; Avg. is the macro-average across suites. Model Size denotes the deployed policy parameter count.

Method	Model Size	Long ↑	Goal ↑	Object ↑	Spatial ↑	Avg. ↑
VLA						
OpenVLA-OFT Kim et al. (2025a)	7B	94.5	97.9	98.4	97.6	97.1
π_0 Black et al. (2024)	3.3B	88.4	94.4	96.8	98.0	94.4
$\pi_{0.5}$ Intelligence et al. (2025)	2.3B	92.4	98.0	98.2	98.8	96.9
GR00T-N1.6 Bjorck et al. (2025)	3B	94.4	97.5	98.5	97.7	97.0
VLA-JEPA Sun et al. (2026)	3B	95.8	97.2	99.6	96.2	97.2
WAM						
FLARE Zheng et al. (2025)	3B	92.2	95.6	100.0	96.8	96.2
DUST Won et al. (2026)	3B	92.6	96.0	99.8	96.2	96.2
Motus Bi et al. (2026)	8B	97.6	96.6	99.8	96.8	97.7
Fast-WAM Yuan et al. (2026b)	6B	95.2	97.0	100.0	98.2	97.6
Cosmos-Policy Kim et al. (2026)	2.1B	97.6	98.2	100.0	98.1	98.5
LingBot-VA Li et al. (2026a)	5.5B	98.5	97.2	99.6	98.5	98.5
DIT4DIT Ma et al. (2026)	2.2B	97.6	98.6	99.6	98.4	<u>98.6</u>
LaWAM Chen et al. (2026)	2.3B	97.0	98.4	99.6	99.4	<u>98.6</u>
PHI-WM						
ActEffect (Ours)	4.6B	97.0	98.4	100.0	99.8	**98.8**

The DINOv3 encoder remains frozen throughout training, and the demonstrated future representation is treated as a fixed target. The feedforward proposal a_t^{ff} is detached before being passed to the controlled world model, such that the direct feedforward branch is optimized only by $\mathcal{L}_{\text{ff}}$. Within each ranking pair, the reference discrepancy is also detached, while consequence feedback from the preferred MIP proposals is allowed to propagate to the iterative policy.

At deployment, the controlled world model and the future-observation branch are removed. The current-observation DINOv3 pathway remains part of the policy, and the MIP action head directly produces the coarse proposal a_t^0 followed by the refined action a_t^1. No future prediction, world-model rollout, or candidate consequence evaluation is required at test time.

5 EXPERIMENTS

We evaluate ActEffect to answer three primary questions: **(Q1)** How does ActEffect compare with state-of-the-art imitation-learning-based VLA and WAM policies? **(Q2)** How much does each component of ActEffect contribute to its overall performance? **(Q3)** Does the controlled world model provide meaningful feedback for action refinement? We investigate these questions across a diverse range of manipulation tasks.

5.1 EXPERIMENTAL SETUP

Benchmarks We comprehensively evaluate ActEffect on three simulation benchmarks.

1. **LIBERO** Liu et al. (2023) focuses on tabletop manipulation tasks with a Franka Emika Panda robot arm. We follow the standard protocol and train on four standard suites: Spatial, Object, Goal, and Long. Each suite contains 10 tasks with 50 demonstrations per task, resulting in a total of 2,000 training demonstrations. For evaluation, each task is evaluated over 50 rollouts, and we report the average success rate.
2. **LIBERO-Plus** Fei et al. (2025) extends LIBERO to systematically evaluate policy robustness under controlled distribution shifts. It introduces perturbations along seven dimensions, including camera viewpoint, robot initial state, language rephrasing, lighting, background texture, sensor noise, and object layout, while retaining the original LIBERO manipulation tasks and training demonstrations. We report the success rate under each perturbation dimension.
3. **RoboCasa-GR1** Nasiriany et al. (2024) evaluates generalist manipulation on a substantially more complex embodiment. Built upon the RoboCasa simulation framework, the

Table 2: **Success rate (%) on LIBERO-PLUS.** Columns correspond to the seven controlled perturbation dimensions; Avg. is their macro-average. Model Size denotes the deployed policy parameter count.

Method	Model Size	Camera	Robot	Language	Light	Bg.	Noise	Layout	Avg. ↑
				VLA					
UniVLA Bu et al. (2025)	7B	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
OpenVLA-OFT Kim et al. (2025a)	7B	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
VLA-JEPA Sun et al. (2026)	3B	64.2	67.7	88.1	91.8	93.4	65.8	83.9	79.5
π_0 Black et al. (2024)	3.3B	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
$\pi_{0.5}$ Intelligence et al. (2025)	2.3B	75.4	77.5	85.6	96.9	94.6	89.7	85.7	**85.7**
				WAM					
Fast-WAM Yuan et al. (2026b)	6B	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
SG-WAM He et al. (2026)	4.7B	54.8	78.3	81.7	96.5	88.9	90.7	85.3	81.3
Cosmos-Policy Kim et al. (2026)	2.1B	75.8	63.3	81.7	96.5	88.9	92.7	82.2	82.2
ImageWAM Zhang et al. (2026b)	4.6B	80.8	50.3	91.4	98.1	85.5	93.8	80.5	<u>83.1</u>
				PHI-WM					
ActEffect (Ours)	4.6B	66.4	55.2	87.1	96.7	95.8	82.6	78.8	80.3

benchmark features the Fourier GR-1 humanoid equipped with two 7-DoF arms, two 6-DoF dexterous hands, and a 3-DoF waist, resulting in a 29-dim action space. The benchmark consists of 24 household tabletop manipulation tasks spanning diverse pick-and-place and articulated-object interactions, with 1,000 human-collected demonstrations provided for each task. We evaluate all 24 tasks and report the average task success rate.

Baselines. We compare against two groups of methods. **VLA policies** include representative state-of-the-art methods on each benchmark, such as $\pi_{0.5}$ Intelligence et al. (2025) and GR00T-N1.6 Bjorck et al. (2025). **WAM policies** include methods that use world models for foresight, such as DiT4DiT Ma et al. (2026) and Fast-WAM Yuan et al. (2026b), as well as methods that leverage future modeling as feedback for policy learning, such as FLARE and DUST.

5.2 BENCHMARK VALIDATIONS

LIBERO. As shown in Table 1, ActEffect achieves the best overall success rate of 98.8%. Despite the near-saturated performance of recent WAMs on this benchmark, ActEffect still improves over foresight-based methods such as DiT4DiT. This suggests that using world model to evaluate proposed actions provides a more effective supervision signal than using predicted futures only as additional context for action generation. Compared with feedback-based WAMs such as FLARE and DUST, the further improvement may stem from explicitly modeling action consequences in a decoupled space and directly ranking alternative proposals during refinement.

LIBERO-PLUS. Table 2 evaluates robustness under seven distribution shifts. ActEffect achieves an average success rate of 80.3%, substantially outperforming the foresight-based Fast-WAM (51.5%). This large gap suggests that providing explicit training-time feedback on the predicted consequences of policy actions can improve robustness beyond relying solely on predictive foresight.

RoboCasa-GR1. ActEffect reaches 67.5% average success, exceeding Fast-WAM and DiT4DiT by 10.8 and 16.7 percentage points, respectively. The large margin over these foresight-based WAMs indicates that explicit consequence feedback remains beneficial even under a higher-dimensional action space and more complex manipulation dynamics. Meanwhile, ActEffect also surpasses FLARE and DUST by a considerable margin, suggesting that explicitly evaluating and ranking the consequences of complete action proposals can provide more effective feedback than implicitly coupling future representations with action generation.

Table 3: **Average success rate (%) on the 24 RoboCasa-GR1 tabletop tasks.** Model Size denotes the deployed policy parameter count.

Method	Model Size	Avg. ↑
VLA		
$\pi_{0.5}$ Intelligence et al. (2025)	2.3B	37.0
GR00T N1.6 Bjorck et al. (2025)	3B	47.6
FoMoVLA Li et al. (2026b)	4.6B	56.9
ABot-M0 Yang et al. (2026)	4.2B	<u>58.3</u>
WAM		
UWM Zhu et al. (2025)	1B	19.3
FLARE Zheng et al. (2025)	3B	36.3
DUST Won et al. (2026)	3B	42.0
DIT4DIT Ma et al. (2026)	2.2B	50.8
LDA-1B Lyu et al. (2026)	1B	55.4
Fast-WAM Yuan et al. (2026b)	6B	56.7
PHI-WM		
ActEffect (Ours)	4.6B	**67.5**

Table 4: **Ablation study on LIBERO.** We report the average success rate (%) across the four LIBERO suites.

Method	Feedback	Consequence Space	FF Anchor	Feedback Proposals	Ranking	Avg. ↑
w/o Feedback	×	–	✓	–	×	97.0
w/o Decoupled Space	✓	VLM	✓	$a^{\mathrm{ff}}, a^0, a^1$	✓	97.3
w/o Feedforward Anchor	✓	DINO	×	a^0, a^1	✓	98.2
w/o Ranking	✓	DINO	✓	a^1	×	98.1
Full	✓	DINO	✓	$a^{\mathrm{ff}}, a^0, a^1$	✓	**98.8**

5.3 ABLATION STUDY

We conduct ablation studies on LIBERO to examine the contribution of each major component in ActEffect. As shown in Table 4, removing consequence feedback reduces the average success rate from 98.8% to 97.0%, confirming that explicitly supervising the policy through predicted action consequences provides complementary information beyond standard action imitation.

Effect of the consequence space. Replacing the decoupled DINO consequence space with the VLM representation space leads to a substantial drop from 98.8% to 97.3%. This suggests that separating task-semantic representations used for action generation from the visual space used for consequence evaluation is important for learning a reliable action-conditioned transition signal.

Effect of the feedforward proposal. Removing the feedforward action proposal decreases the performance to 98.2%. The feedforward branch provides an additional task-conditioned action hypothesis, enriching the set of interventions considered during consequence evaluation and providing a broader reference for progressive refinement.

Effect of counterfactual consequence ranking. When consequence feedback is applied only to the refined proposal without counterfactual ranking, the success rate decreases to 98.1%. In contrast, the full model jointly evaluates the feedforward, coarse, and refined proposals and explicitly encourages their predicted consequences to improve progressively, achieving the best performance of 98.8%. These results support the use of structured consequence ranking across multiple action proposals rather than relying on consequence prediction alone.

6 CONCLUSION

We presented ActEffect, an action-first robot model that uses a controlled world model to evaluate the predicted consequences of complete action proposals and turn future prediction into training-time feedback for action refinement. ActEffect uses vision-language representations for task-

conditioned action generation and a separate frozen DINOv3 latent space for consequence modeling. By evaluating feedforward, coarse, and refined action proposals under the same current state and enforcing model-based counterfactual consequence ranking, ActEffect encourages successive proposals to produce predicted futures progressively closer to the demonstrated future. The controlled world model is used only during training and removed at deployment, requiring no future prediction, world-model rollout, or candidate consequence evaluation at inference. Across LIBERO, LIBERO-PLUS, and RoboCasa-GR1, ActEffect achieves average success rates of 98.8%, 80.3%, and 67.5%, respectively. These results demonstrate that explicitly modeling counterfactual action consequences and using them as a training signal for ranking alternative futures can effectively shape policy learning, highlighting action-consequence feedback as a promising direction for robot policy learning.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

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.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, et al. Lawam: Latent world action models for efficient dynamics-aware robot policies. *arXiv preprint arXiv:2606.15768*, 2026.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

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.

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024.

Junjie He, Junfeng Li, Zhide Zhong, Haodong Yan, Ruixin Li, Yangyang Zheng, Jiaguan Zhu, Tianran Zhang, Yuqiao Du, Wen Chen, et al. Sg-wam: Text-grounded and spatial-aware semantic guidance for world-action models. *arXiv preprint arXiv:2608.08839*, 2026.

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, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025a.

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 2025b.

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.

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*, 2026a.

Wei Li, Peijin Jia, Yuan Ma, Xuefeng Jiang, Titong Jiang, Sheng Sun, Yujian Li, Xin Wen, Han Hong, Zhikang Liu, et al. Fomovla: Bridging visual foresight and motion guidance for vision-language-action models. *arXiv preprint arXiv:2607.14739*, 2026b.

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.

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.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

Chaoyi Pan, Giri Anantharaman, Nai-Chieh Huang, Claire Jin, Daniel Pfrommer, Chenyang Yuan, Frank Permenter, Guannan Qu, Nicholas Matthew Boffi, Guanya Shi, and Max Simchowitz. Much ado about noising: Dispelling the myths of generative robotic control. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=LzWKuxTKuW`.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. Vla-jepa: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

Likun Wang, Xiangteng Zhang, Yinuo Wang, Guojian Zhan, Wenxuan Wang, Haoyu Gao, Jingliang Duan, and Shengbo Eben Li. Off-policy reinforcement learning with model-based exploration augmentation. *Advances in Neural Information Processing Systems*, 38:92121–92151, 2026.

John Won, Kyungmin Lee, Huiwon Jang, Dongyoung Kim, and Jinwoo Shin. Dual-stream diffusion for world-model augmented vision-language-action model. In *Forty-third International Conference on Machine Learning*, 2026.

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*, 2026.

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.

Tianyi Zhang, Likun Wang, Guojian Zhan, Feihong Zhang, Yang Guan, Yao Lyu, and Shengbo Eben Li. Langevin rollout optimization for modelic reinforcement learning. In *Forty-third International Conference on Machine Learning*, 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.

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.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot learning with implicit world modeling. In *9th Annual Conference on Robot Learning*, 2025.

Linqing Zhong, Yi Liu, Yifei Wei, Ziyu Xiong, Si Liu, and Guanghui Ren. Acot-vla: Action chain-of-thought for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8152–8162, 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. *arXiv preprint arXiv:2504.02792*, 2025.