# ACTIONS WITHOUT CONSEQUENCES: BINDING WORLD ACTION MODEL POLICIES TO WHAT THEY DECIDE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

World action models (WAMs) build on pretrained video models to jointly predict future observations and robot actions. Demonstration data, however, provides only one action for each observed state and therefore does not directly constrain how the predicted future should change under alternative actions. We study this gap with CF-BENCH, a counterfactual benchmark for measuring action sensitivity in WAM policies. ANCHORis introduced to compare the robot pose represented in the predicted future with the pose implied by the predicted action through forward kinematics and camera calibration. This connection also enables counterfactual training without ground-truth future observations: we perturb the action, enforce the corresponding kinematic change in the robot region, and constrain regions outside the arm's swept volume to remain unchanged. With counterfactual training, ANCHORimproves the command-normalised keypoint response to 0.79 for in-distribution perturbations and 0.71 for held-out perturbations. Under matched evaluation settings, it also improves task success over base-FT by 2.3–6.6 percentage across RoboCasa, RoboTwin 2.0, and LIBERO-Plus.

**Keywords:** World Action Models, Counterfactual Action Sensitivity, Kinematic Anchoring, Counterfactual Training

## 1 INTRODUCTION

Vision-language-action models typically predict robot actions from visual and language inputs (Kim et al., 2024; Black et al., 2025; Zheng et al., 2025; Ghosh et al., 2024; Chi et al., 2023). World action models (WAMs) adapt pretrained video models to predict both future observations and robot actions (Du et al., 2023; Wu et al., 2024; Ye et al., 2026b; Li et al., 2026; Kim et al., 2026). The video pretraining provides useful knowledge about how scenes change, which can support control (Reuss, 2026; Shen et al., 2026). Recent results suggest that this approach can produce competitive policies. DreamZero, for example, reported the top RoboArena position in February 2026 after training on DROID (Ye et al., 2026b; Atreya et al., 2025). Since RoboArena is a live leaderboard, we treat this as a dated result rather than a current state-of-the-art claim. For planning, verification, and test-time action selection, the predicted future needs to reflect the predicted action (Kim et al., 2026; Ruan et al., 2026; Aida, 2026). Different actions should therefore produce different future predictions when their outcomes differ. This issue has been studied in action-conditioned world models, where actions are provided as inputs (Chen et al., 2026b; Yang et al., 2026a; Co et al., 2026). WAM policies differ because the action and the future are generated by the same model.

Demonstrations provide only one observed action for each context. Alternative actions and their resulting futures are generally missing. A WAM may therefore match the training pairs without learning how future predictions should respond to action changes. We formalize this ambiguity in Section 3, where the demonstration objective leaves the counterfactual action–future relation unidentified. Related failures have been reported in existing WAMs. Generated futures may contain incorrect robot bodies (Reuss, 2026; Jiang et al., 2026), while failed rollouts may become static and still receive high consistency scores (Ruan et al., 2026). These findings motivate a direct evaluation of how predicted futures respond to action changes. The main challenge is that unexecuted actions have no observed outcomes. Demonstrations cannot tell us whether an alternative grasp would



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.

succeed or how an object would move. The commanded robot body is one exception. Given the current state and controller, an action defines a commanded robot configuration. Forward kinematics and camera calibration then map this configuration into the image.

For an unexecuted action, object motion and task outcome are unknown, but the commanded robot configuration can still be computed. We use this configuration as an action-dependent supervision signal. Kinematics specifies the commanded pose. The executed pose can differ because of contact, tracking error, and controller dynamics. We measure this mismatch and examine its relation to the gains from our method.

Several methods use robot geometry to represent actions. iMaC and EA-WM convert actions into visual or geometric conditions for future prediction (Wu et al., 2026; Yang et al., 2026b; Wang et al., 2025). Render-and-Diffuse, GENIMA, and Action Images encode actions with rendered robot geometry (Vosylius et al., 2024; Shridhar et al., 2024; Zhen et al., 2026). SelfWAM uses demonstrated actions to condition future prediction and adds robot self-mask prediction during training (Pan et al., 2026; Feingold et al., 2026). In all of these cases, the training signal comes from actions observed in the dataset.

We introduce ANCHOR to link the predicted action with the robot body in the imagined future, as illustrated in Figure 1. The policy's predicted action chunk is passed through forward kinematics and projected into the camera. A small keypoint head estimates the robot body in the imagined future. The anchor loss penalizes disagreement between these two quantities. The keypoint head is used as a measurement function. It is calibrated on demonstrated robot bodies, but its parameters do not receive gradients from the anchor loss. Anchor gradients instead update the action and video pathways. This prevents the model from satisfying the constraint simply by changing the detector used to measure its own prediction. The same relation also gives a training signal for alternative actions. For counterfactual training, we perturb the action and project the new commanded pose into the image. The robot region is trained to follow the resulting kinematic change. Outside an FK-derived swept region, we regularize the counterfactual prediction toward the factual one. Other approaches obtain counterfactual supervision by executing the perturbed action in simulation (Chen et al., 2026b) or by collecting failure trajectories (Peng et al., 2026). Our loss does not attempt to supervise the full counterfactual future. It supervises the part specified by robot kinematics, with the locality assumption outside the swept region. This avoids the need for a ground-truth counterfactual rollout, which requires a robot model and calibration. The anchor residual is used at test time. For a candidate action, we compare the body implied by the action detected in its imagined future. The resulting score is available before execution and requires only one generated future. It does not require a value head, consensus across multiple samples, or a separately trained verifier. The score measures geometric agreement between action and imagined robot motion. We include a coverage term to avoid assigning a favourable score. We introduce CF-BENCH, a policy-side adaptation of the off-expert protocol to evaluate the coupling Chen et al. (2026b). We perturb the action and measure how much the imagined future changes relative to the corresponding ground-truth change. Robot and environment responses are reported separately. We also evaluate held-out perturbation families. Finally, a second protocol is included in which the policy generates its own action.

The contributions are as follows. (i) We formalize the ambiguity left by demonstration training. (ii) We introduce kinematic anchoring to match the robot body in the imagined future with the body specified by the predicted action. (iii) We introduce CF-BENCH to measure action sensitivity in WAM policies. (iv) On RoboCasa, RoboTwin 2.0, and LIBERO-Plus, ANCHOR+CF reaches a command-normalised keypoint response of 0.79 in distribution and 0.71 on held-out perturbations.

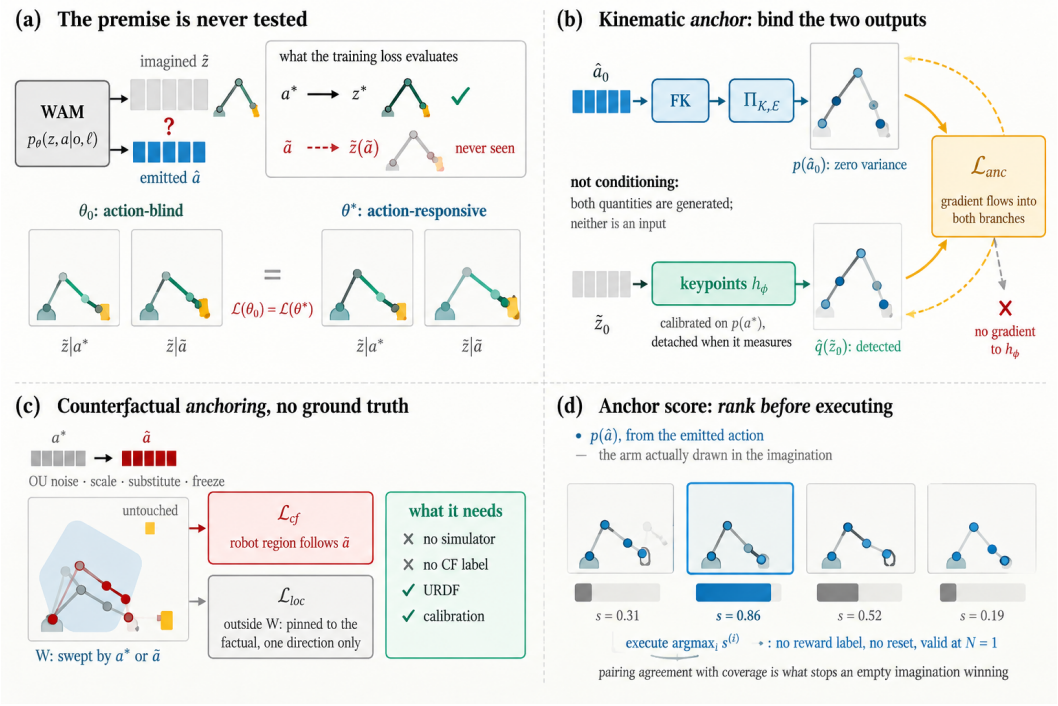


Figure 1: Overview of ANCHOR. (a) Action-blind and action-responsive mappings can agree on the observed demonstration pairs while differing on alternative actions. (b) Kinematic anchoring. The emitted action is mapped through forward kinematics and camera projection, while a keypoint head detects the robot body in the imagined future. (c) Counterfactual anchoring. We perturb the action to follow the corresponding kinematic change. (d) Candidate scoring. The same residual measures agreement between a candidate action and the robot body in its imagined future.

## 2 RELATED WORK

**World action models.** Imagining a future and acting on it is the oldest idea in model-based control (Hafner et al., 2025). Recovering actions from a generated future dates to UniPi (Du et al., 2023; Ko et al., 2024; Zhou et al., 2024), joint video–action prediction to GR-1 (Wu et al., 2024), and latent plans and actions to Play-LMP, Genie and LAPA (Lynch et al., 2020; Bruce et al., 2024; Ye et al., 2025; Zhen et al., 2025). Modern WAMs scale these in three formulations (Reuss, 2026; Shen et al., 2026): inverse dynamics on a predicted future (Li et al., 2026; Pai et al., 2025; Hu et al., 2025), joint denoising of video and action (Ye et al., 2026b; Kim et al., 2026; Zhu et al., 2025; Ye et al., 2026a), and latent foresight that skips pixels (Yuan et al., 2026; Luo et al., 2026). ANCHOR constrains what any video branch must agree with any action branch about; we evaluate one model of each of the first two kinds. For action-conditioned world models, where the action is an input, this has begun to be tested and found wanting: expert-trained simulators follow demonstrated actions and ignore or visually collapse under off-expert ones (Chen et al., 2026b; Yang et al., 2026a; Co et al., 2026). Wang et al. (2026b) name the related executability gap, which is visually coherent rollouts violate the robot's kinematics. For policies, where the action is generated, Ruan et al. (2026) show that action–state consistency separates success from failure and identify background collapse as its boundary condition. We ask the policy-side version of the question and supply a training-time answer, which Ruan et al. (2026) conjecture is what removing collapse requires.

**Kinematics.** That the commanded body is kinematically determined is known and used two ways. iMaC and EA-WM render the commanded body into the camera and condition a world model on it (Wu et al., 2026; Yang et al., 2026b), Wang et al. (2025) show a rendered skeleton beats text or state vectors as a control signal, and Render-and-Diffuse, GENIMA and Action Images represent the action itself as a rendered body (Vosylius et al., 2024; Shridhar et al., 2024; Zhen et al., 2026).

SelfWAM supervises a WAM's video branch with the self-mask of the demonstrated episode (Pan et al., 2026), MaskWAM predicts object masks (Yu et al., 2026), Mask2Real-WM bridges domains with embodiment masks (Feingold et al., 2026), and point tracks have long served as embodiment-agnostic targets (Bharadhwaj et al., 2024). Both leave Proposition 1 untouched, because the target exists only at $a^\star$. Self-modelling from proprioception and vision (Chen et al., 2022) shares the intuition that a robot's body is knowable, but learns it rather than reading it from a URDF.

**Off-manifold supervision and verification.** Intervention-effect alignment supplies counterfactual pairs by executing $\tilde{a}$ in a simulator (Chen et al., 2026b); FACT obtains coverage from collected failure rollouts (Peng et al., 2026). ANCHOR instead uses kinematic information from a URDF. For verification, candidates are scored by a learned value head (Kim et al., 2026), by consensus among samples (Ruan et al., 2026), by a forward–inverse cycle (Liu et al., 2026), by an inverse-dynamics drift signal (Tseng et al., 2026), or in latent space for progress and risk (Aida, 2026; Sun et al., 2026; Chen et al., 2026a); adaptive execution instead compares imagination with the realized observation (Wang et al., 2026a). A known physical function of the emitted action, defined for a single sample, needing no execution, labels or learned verifier. Our counterfactuals are grounded in a physical function and evaluated against ground truth.

## 3 PRELIMINARIES

Let $o$ be the observation, $\ell$ the instruction, $a \in \mathbb{R}^{K \times d_a}$ an action chunk and $z \in \mathbb{R}^{C \times T \times h \times w}$ the latent future video from the backbone's VAE (temporal stride $s$, $T \approx K/s$). A WAM parameterises $p_\theta(z, a \mid o, \ell)$ either jointly with one denoiser over $[z; a]$ (Ye et al., 2026b; Kim et al., 2026) or factored as $p_\theta(z \mid o, \ell)\, p_\theta(a \mid o, \ell, z)$ (Li et al., 2026), trained by flow matching (Lipman et al., 2023) on demonstrations $\mathcal{D} = \{(o_i, \ell_i, a_i^\star, z_i^\star)\}$; at flow time $\tau$ the model exposes $\hat{z}_0(\tau), \hat{a}_0(\tau)$.

The quantity of interest is the clean counterfactual relation $p_\theta(z \mid o, \ell, a)$ for an alternative action under the same context. Flow matching corrupts the demonstrated action tokens during training, but that corruption does not create an observed future for a different clean action. This distinction matters because common Gaussian flow-matching paths can have broad or full support in token space (Lipman et al., 2023); the missing supervision is therefore not a claim that noisy action inputs are confined to a bounded tube. It is the absence of a clean counterfactual action–future pair for the same demonstrated context.

**Proposition 1** (Clean counterfactual coupling is not identified by demonstrations)**.** *Let $F(o, \ell, a)$ denote the clean action-to-future response represented by a policy. The demonstration set observes only $F(o_i, \ell_i, a_i^\star) = z_i^\star$ for each context. Two response mappings $F$ and $F'$ that agree on all observed clean pairs but differ at some $F(o_i, \ell_i, \tilde{a})$, $\tilde{a} \neq a_i^\star$, induce the same set of observed clean demonstration pairs. Adding auxiliary noise or augmentations whose targets are functions of the observed tuple does not by itself provide the missing realized future $z_i(\tilde{a})$.*

An action-blind video predictor with an inverse-dynamics reader is therefore compatible with the observed clean action–future pairs whenever it fits the demonstrated futures. This is an observational identifiability statement, not a claim that every such parameterization attains exactly the same finite-sample flow-matching loss; architecture, noise schedules and optimization can introduce additional inductive pressure.

Near $\theta_0$ the body is rendered from the visual prior rather than the command, so pose and morphology drift (Reuss, 2026; Jiang et al., 2026); and if the policy stalls, the video branch predicts an unchanged scene because that is what context predicts, so the pair is consistent while failing (Ruan et al., 2026). Breaking the degeneracy requires evaluating at $\tilde{a} \neq a^\star$, for which $z(\tilde{a})$ is unavailable—but one functional of it is known. With forward kinematics FK from a URDF, camera $(\mathbf{K}, \mathbf{E})$ and $M$ body keypoints with known link offsets,

$$p(a) = \Pi_{\mathbf{K},\mathbf{E}}\big(\mathrm{FK}(q(a))\big) \in \mathbb{R}^{T \times M \times 2}, \tag{1}$$

where $q(a)$ maps actions to joint configurations (identity for joint-position control. Equation (1) is deterministic and differentiable in $a$, and it is worth being exact about what it is: $p(a)$ is the body configuration prescribed by the controller interface, not the one the arm will physically occupy. The two coincide in free space and diverge under contact.

# 4 METHOD

## 4.1 KINEMATIC ANCHOR

A small head $h_\phi$ on the video branch maps $\hat{z}_0$ to $M$ heatmaps per latent frame and, via soft-argmax (Finn et al., 2016; Sun et al., 2018), to coordinates $\hat{q}(\hat{z}_0) \in \mathbb{R}^{T\times M\times 2}$ with confidences $c \in [0,1]^{T\times M}$. We treat $h_\phi$ as a measurement operator, calibrated on demonstrations, never asked to help satisfy the constraint it measures. Writing $\mathrm{sg}[\cdot]$ for the stop-gradient,

$$\mathcal{L}_{\mathrm{anc}} = \underbrace{\frac{1}{\sum v}\sum_{t,m} v_{t,m}\Big\|\tfrac{1}{\sigma}\big(\hat{q}_{t,m}(\hat{z}_0;\mathrm{sg}[\phi]) - p_{t,m}(\hat{a}_0)\big)\Big\|_2^2}_{\text{measures the model; updates } \hat{z}_0, \hat{a}_0} + \lambda_h \underbrace{\mathcal{L}_{\mathrm{hm}}\big(h_\phi(\mathrm{sg}[\hat{z}_0]),\, p(a^\star)\big)}_{\text{calibrates the instrument; updates } \phi}, \tag{2}$$

with visibility weights $v$ (rendered depth in simulation, detector confidence on real data; keypoints projecting outside the image are excluded) and scale $\sigma = 0.1\min(H, W)$, expressed relative to the image so that the loss does not silently change when a target robot's camera differs.

Both stop-gradients are load-bearing. Without $\mathrm{sg}[\phi]$ the detector drifts toward $p(\hat{a}_0)$ and satisfies the first term without the imagination changing, px of drift against 6.4 px when detached, with held-out sensitivity back at 0.33. Without $\mathrm{sg}[\hat{z}_0]$ the calibration term would supervise the imagination with the demonstrated body, detaching it means every gradient the imagination receives from the anchor comes from the action the model emitted. The gradient reaching $\hat{a}_0$ cannot drag the action far, since $\hat{a}_0$ is pinned by its own flow-matching loss. Early in denoising $\hat{a}_0$ is near noise, so we clamp $q(\hat{a}_0)$ to the joint limits and ramp $\lambda_1(\tau) = \lambda_1(1-\tau)^2$; a constant $\lambda_1$ costs 1.4 SR.

Three properties distinguish this formulation from the alternatives, the anchor is evaluated at the model's $\hat{a}_0$, not only at an action supplied as input (Wu et al., 2026; Wang et al., 2025) or at a demonstration-derived label (Chen et al., 2026b; Pan et al., 2026); gradients can reach both generated branches; and the commanded projection $p(\cdot)$ is deterministic given the action and calibration.

## 4.2 COUNTERFACTUAL ANCHORING AND THE CAUSAL LOCALITY PRIOR

With probability $\rho$ we draw $\tilde{a} \sim \mathcal{P}(a^\star)$ from a structured family. Ornstein–Uhlenbeck noise, amplitude scaling, substitution of another episode's chunk and freeze, which holds the arm still and targets background collapse directly. We condition the video branch on $\tilde{a}$—teacher-forcing clean action tokens in a joint model, or the action-to-video pathway in a mixture-of-transformers model to obtain $\tilde{z}_0$ under the same noise and flow time as $\hat{z}_0$. We have no ground truth for $\tilde{z}_0$.

$$\mathcal{L}_{\mathrm{cf}} = \frac{1}{\sum \tilde{v}}\sum_{t,m} \tilde{v}_{t,m}\Big\|\tfrac{1}{\sigma}\big(\hat{q}_{t,m}(\tilde{z}_0;\mathrm{sg}[\phi]) - p_{t,m}(\tilde{a})\big)\Big\|_2^2, \tag{3}$$

$$\mathcal{L}_{\mathrm{loc}} = \frac{\big\|(1-W)\odot\big(\tilde{z}_0 - \mathrm{sg}[\hat{z}_0]\big)\big\|_2^2}{\|1-W\|_1}, \quad W = \mathrm{dilate}\Big(\mathrm{cummax}_t\big[\mathrm{swept}(p(a^\star)) \cup \mathrm{swept}(p(\tilde{a}))\big]\Big), \tag{4}$$

with $\tilde{v} = \mathrm{visibility}(p(\tilde{a}))$: an action that swings the arm behind an object should not be penalised for failing to draw what the camera cannot see. Equation (3) makes the robot region follow the counterfactual action, Equation (4) pins everything outside the time-accumulated swept region to the factual prediction. The stop-gradient there runs opposite to intuition and is essential: without it the term is minimisable by dragging the factual prediction toward the counterfactual, so counterfactual training would corrupt the prediction the policy executes from, at a cost of 3.2 SR and 14% on the factual flow-matching loss. Equation (4) also presumes a camera the action cannot move.

$$\mathcal{L} = \mathcal{L}^z_{\mathrm{FM}} + \mathcal{L}^a_{\mathrm{FM}} + \lambda_1\mathcal{L}_{\mathrm{anc}} + \rho\big(\lambda_2\mathcal{L}_{\mathrm{cf}} + \lambda_3\mathcal{L}_{\mathrm{loc}}\big), \tag{5}$$

where the $\rho$ factor denotes the expectation over counterfactual sampling. In a Monte Carlo training step that first samples the branch with probability $\rho$, the sampled counterfactual loss should not be multiplied by $\rho$ a second time; otherwise its expected weight becomes proportional to $\rho^2$.

This objective recovers the base model exactly at $\lambda_{1,2,3} = 0$. Intervention-effect supervision (Chen et al., 2026b) instead aligns $\tilde{z}_0 - \hat{z}_0$ with a ground-truth $z^\star(\tilde{a}) - z^\star$ and so needs a simulator;

Equations (3) and (4) need only the URDF and calibration, and therefore apply to real-robot logs. Equation (3) does two separable things: it exposes the model to actions the demonstrator never took, and it says what the body should do there. We build a control, CF-Contrast, that keeps the first and drops the second—same $\mathcal{P}$, same $\rho$, same $\mathcal{L}_{\text{loc}}$, but Equation (3) replaced by a hinge demanding only that the robot region move.

$$\mathcal{L}_{\text{contrast}} = \max\left(0,\ \delta - \frac{\|\mathbf{1}_{\text{rob}} \odot (\tilde{z}_0 - \text{sg}[\hat{z}_0])\|_2}{\|\tilde{a} - a^\star\|_2}\right). \tag{6}$$

Because this control retains $\mathcal{L}_{\text{loc}}$ as written, its swept-region mask still uses the URDF and calibration; what it removes is the directional keypoint target in Equation (3). If an implementation instead removes all FK-derived information, then the locality prior must also be changed and the control no longer differs only in the target.

Nothing in Equations (3) and (4) says what happens to the mug, but a model that must reproduce its own arm from its own action cannot explain arm motion as generic scene change, and one whose untouched regions are pinned cannot smear arm uncertainty across the scene.

### 4.3 Anchor score: execution-free self-verification

For $N$ candidates $\{(\hat{z}^{(i)}, \hat{a}^{(i)})\}$,

$$\begin{aligned} s^{(i)} = &- \frac{\sum_{t,m} c_{t,m}^{(i)} \left\|\sigma^{-1}\big(\hat{q}_{t,m}(\hat{z}^{(i)}) - p_{t,m}(\hat{a}^{(i)})\big)\right\|_2^2}{\epsilon + \sum_{t,m} c_{t,m}^{(i)}} \\ &- \beta\left(1 - \frac{1}{TM}\sum_{t,m} c_{t,m}^{(i)}\right), \qquad i^\star = \arg\max_i s^{(i)}. \end{aligned} \tag{7}$$

and we execute $\hat{a}^{(i^\star)}$; winner-takes-all, since averaging actions mixes motion hypotheses and mishandles rotations (Ruan et al., 2026). We use $\epsilon{=}10^{-6}$ only for numerical stability. The coverage term prevents a degenerate solution: if a candidate contains no recognisable arm, all $c_{t,m}$ can approach zero, reducing the stabilised agreement penalty and artificially improving the score. Such a candidate is a collapsed future, exactly the failure this paper detects. With $\beta{=}0.5$, collapsed futures fall from 29.4% to 6.2% of selections and AUC rises from 0.68 to 0.81. The same degeneracy can arise in confidence-normalised agreement scores.

### 4.4 Embodiment adaptation and implementation

FK, $\mathbf{K}, \mathbf{E}$ are inputs, not parameters: a new robot means a new URDF, calibration and keypoint set, after which $h_\phi$ and the adapters are fine-tuned on target data. This is adaptation assisted by a URDF, not transfer by swapping one—the action space itself changes across embodiments and zero-shot success is near chance for every method—but the anchor supplies $M$ geometric targets per latent frame in addition to the action-vector supervision. The low-data experiments in test whether this additional signal is associated with larger gains. We fine-tune with LoRA (Hu et al., 2022) and train $h_\phi$ (three convolutions on per-frame latents) from scratch, pretraining it against $p(a^\star)$ so that it is a competent instrument before the first term of Equation (2) is enabled. Under delta control we integrate from the commanded pose—never the realized one, which depends on contact and would make $p(\cdot)$ a random variable given the action—and anchor the tool centre and both fingertips.

### 4.5 CF-Bench

Proposition 1 says training cannot reveal action sensitivity, so evaluation must. Chen et al. (2026b) established the protocol for action-conditioned world models: query off the expert action and compare against ground truth. We instantiate it for policies, where the action is generated rather than given. From $S$ saved simulator states per task on held-out episodes, stratified into pre-, in- and post-contact, we sample $J$ perturbed chunks $\tilde{a}_j \sim \mathcal{P}(a^\star)$, execute them to obtain ground-truth counterfactual futures $z_j^\star$ and per-frame robot/environment segmentation $R \in \{\text{rob}, \text{env}\}$ at latent resolution, and query the model as $\tilde{z}_j = f(o, \ell, \tilde{a}_j)$. Per region, $\text{Sens}_R = \|\mathbf{1}_R \odot (\tilde{z}_j - \hat{z}_0)\|_2 / \|\tilde{a}_j - a^\star\|_2$ is how far the imagination moves per unit of action change and $\text{Sens}_R^\star$ the same for reality ($z_j^\star$ against $z^\star$); we report $\text{ratio}_R = \text{Sens}_R / \text{Sens}_R^\star$ and $\text{CFErr}_R = \|\mathbf{1}_R \odot (\tilde{z}_j - z_j^\star)\|_2$. A controllable policy

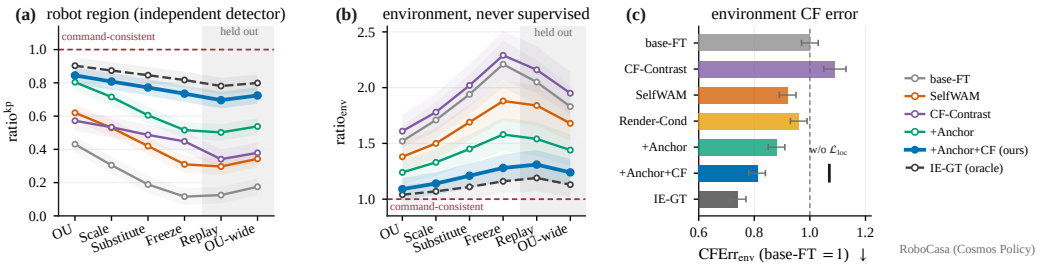


Figure 2: **CF-BENCH on RoboCasa (Cosmos Policy).** (a) Robot-region sensitivity by perturbation family, measured in keypoint space by an independent detector on decoded frames, so this is not the quantity Equation (3) optimises. (b) Environment-region ratio. values above the reference line mean the imagination changes the scene more than reality does. (c) Environment counterfactual error normalised to base-FT, with the $\mathcal{L}_{\text{loc}}$ ablation marked.

has $\text{ratio}_R \approx 1$, the degenerate $\theta_0$ has $\text{ratio}_{\text{rob}} \approx 0$, and CFErr penalises moving a lot but wrongly. An oracle returning $z_j^\star$ scores ratio=1, CFErr=0; a copy-factual model scores $\text{ratio}_{\text{rob}}$=0.

$\text{ratio}_{\text{rob}}$ is uncomfortably close to what Equation (3) optimises, so the robot region is additionally reported with a command-normalised keypoint ratio read by an instrument the method never touches: $r^{\text{kp}}_{\text{cmd}}$ replaces $\hat{z}$ by decoded pixels and $h_\phi$ by an independent detector trained on rendered frames, never on WAM outputs or on any projection of an emitted action. No robot-region measure can be fully independent of an objective that trains robot-region sensitivity, so we also report $\text{CFErr}_{\text{env}}$, $\text{ratio}_{\text{env}}$, success and transfer. The environment metrics use ground-truth counterfactual rollouts only for evaluation, training never observes a ground-truth counterfactual environment target, although $\mathcal{L}_{\text{loc}}$ regularises changes outside the swept volume. The same detector gives the hallucination rate. Teacher forcing is not how a policy runs: it samples $(\hat{z}, \hat{a})$ jointly, and improving the conditional need not improve the joint. We therefore add a protocol with no teacher forcing, drawing $N$ candidates from one observation and reporting $R^2$ between pairwise action distance and pairwise imagined-body distance, and a matching margin $\Delta_{\text{match}}$ that is $\leq 0$ when the candidates are interchangeable. Neither uses a simulator, so both can be computed on real-robot logs.

## 5 EXPERIMENTS

### 5.1 EXPERIMENTAL SETUP.

Base models: Cosmos Policy (Kim et al., 2026) (2B, joint prediction) on RoboCasa (Nasiriany et al., 2024) and LingBot-VA (Li et al., 2026) (5B, mixture-of-transformers inverse dynamics) on RoboTwin 2.0 (Chen et al., 2025). LIBERO-Plus (Fei et al., 2025) measures transfer of the RoboCasa models. Because a simulator can flatter a policy tuned inside it (Li et al., 2024), we use LIBERO-Plus transfer as the main out-of-training-distribution check in this manuscript. Variants: base-FT (identical LoRA budget, no anchor), +ANCHOR ($\rho$=0), +ANCHOR+CF. All variants share data, steps and batch size; $\lambda_2$=$\lambda_1$ and $\lambda_h$ is fixed, and $\lambda_1, \lambda_3, \rho, \beta$ were chosen once on a three-task development split excluded from every reported number, then frozen everywhere.

### 5.2 MAIN RESULTS

Figure 2 and Table 1 report CF-BENCH. Released checkpoints are largely action-blind off the demonstration manifold: base-FT's $r^{\text{kp}}_{\text{cmd}}$ is 0.26 in distribution and 0.15 on held-out families, with hallucination on 41.7% and 36.2% of imagined frames for the two backbones. Sensitivity declines monotonically from small OU noise to substitution and freeze, matching the empirical ordering. The environment inverts the pattern: $\text{ratio}_{\text{env}}$ sits above one while $\text{CFErr}_{\text{env}}$ is large, so the imagination is not frozen—it changes the wrong things, smearing what should have been arm motion across the scene. Anchoring raises $r^{\text{kp}}_{\text{cmd}}$ to 0.66 and counterfactual anchoring to 0.79, with 0.71 on families never trained on; the latent metric is more generous (0.87 / 0.78), and the gap measures how much it flatters the method. Environment error falls from 1.00 to 0.81; removing $\mathcal{L}_{\text{loc}}$ raises the separately

Table 1: **Reliability and controllability of imagined futures.**

| | | Reliability | | | CF-BENCH controllability | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | robot region | | environment (no GT CF target) | |
| Base | Variant | Consist. ↑ | Collapse ↓ | Halluc. ↓ | $r^{\mathrm{kp}}_{\mathrm{cmd}}$ ↑ | latent | $\mathrm{ratio}_{\mathrm{env}} \to 1$ | $\mathrm{CFErr}_{\mathrm{env}}$ ↓ |
| Cosmos Policy RoboCasa | base-FT | 0.71 | 34.2% | 41.7% | 0.26 / 0.15 | 0.29 / 0.18 | 1.94 | 1.00 |
| | Aux-Track | 0.72 | 33.1% | 39.8% | 0.28 / 0.17 | 0.31 / 0.20 | 1.88 | 0.99 |
| | Render-Cond | 0.76 | 29.5% | 24.3% | 0.64 / 0.31 | 0.72 / 0.38 | 1.79 | 0.96 |
| | Traj-Head | 0.77 | 27.8% | 22.6% | 0.42 / 0.29 | 0.46 / 0.33 | 1.66 | 0.93 |
| | SelfWAM† | 0.78 | 25.4% | 19.8% | 0.47 / 0.32 | 0.51 / 0.35 | 1.61 | 0.92 |
| | CF-Contrast | 0.79 | 19.7% | 22.1% | 0.51 / 0.36 | 0.66 / 0.47 | 2.02 | 1.09 |
| | +ANCHOR | 0.83 | 21.6% | 12.4% | 0.66 / 0.52 | 0.72 / 0.57 | 1.42 | 0.88 |
| | +ANCHOR+CF | 0.86 | 12.7% | 7.9% | **0.79 / 0.71** | 0.87 / 0.78 | 1.19 | 0.81 |
| | IE-GT (oracle) | 0.88 | 11.3% | 6.8% | 0.86 / 0.79 | 0.91 / 0.84 | 1.12 | 0.74 |
| LingBot-VA RoboTwin 2.0 | base-FT | 0.74 | 29.8% | 36.2% | 0.31 / 0.17 | 0.36 / 0.20 | 1.81 | 1.00 |
| | +ANCHOR | 0.85 | 18.9% | 11.1% | 0.69 / 0.55 | 0.75 / 0.60 | 1.38 | 0.87 |
| | +ANCHOR+CF | 0.88 | 10.4% | 6.5% | **0.81 / 0.73** | 0.89 / 0.80 | 1.16 | 0.79 |

Table 2: **Success rate (%).**

| | RoboCasa | | RoboTwin 2.0 | | LIBERO-Plus (transfer) | |
|---|---|---|---|---|---|---|
| Method | SR | Δ [95% CI] | SR | Δ [95% CI] | SR | Δ [95% CI] |
| $\pi_0$ (reported by Kim et al. (2026)) | 62.5 | – | – | – | – | – |
| $\pi_{0.5}$ (reported by Li et al. (2026)) | – | – | 82.7 | – | – | – |
| X-VLA (Zheng et al., 2025) | – | – | 72.9 | – | – | – |
| Cosmos Policy‡ (Ruan et al., 2026) | 66.6 | – | – | – | – | – |
| LingBot-VA‡ (Ruan et al., 2026) | – | – | 90.2 | – | – | – |
| base-FT | 66.9±0.8 | – | 90.1±0.6 | – | 49.7±1.3 | – |
| Aux-Track | 67.4±0.9 | +0.5 [-0.6, 1.7] | – | – | 50.4±1.4 | +0.7 [-1.0, 2.4] |
| Render-Cond | 68.1±0.7 | +1.2 [0.1, 2.4] | – | – | 51.2±1.2 | +1.5 [-0.2, 3.2] |
| Traj-Head | 68.5±0.9 | +1.6 [0.4, 2.8] | – | – | 52.0±1.3 | +2.3 [0.5, 4.1] |
| SelfWAM† | 68.9±0.8 | +2.0 [0.8, 3.2] | – | – | 52.6±1.2 | +2.9 [1.0, 4.8] |
| CF-Contrast | 68.3±1.0 | +1.4 [0.0, 2.8] | – | – | 51.8±1.4 | +2.1 [0.1, 4.1] |
| +ANCHOR | 69.4±0.6 | +2.5 [1.3, 3.7] | 91.5±0.5 | +1.4 [0.3, 2.5] | 54.1±1.1 | +4.4 [2.4, 6.4] |
| **+ANCHOR+CF** | 70.3±0.7 | +3.4 [2.1, 4.7] | 92.4±0.6 | +2.3 [1.0, 3.6] | 56.3±1.2 | +6.6 [4.3, 8.9] |

reported error by 32%, which is consistent with the intended role of the locality regularizer. Hallucination falls to 7.9% and collapse among failures from 34.2% to 12.7%. A single-seed freeze-only ablation also gives lower collapse than the base model, but the row-level seed mismatch does not support a precise fraction-of-reduction claim.

CF-Contrast uses the same off-manifold actions and locality prior but removes the directional keypoint target. It reaches $r^{\mathrm{kp}}_{\mathrm{cmd}}$ 0.51 / 0.36; $\mathrm{CFErr}_{\mathrm{env}}$ rises to 1.09, above base-FT, hallucination is 22.1%, and the RoboCasa success difference is +1.4 points with a 95% CI of [0.0, 2.8]. These measurements show that exposure alone does not reproduce the full +ANCHOR+CF result. Because the reported robot sensitivity is a magnitude ratio, they do not directly measure whether the imagined motion has the correct direction.

Table 2 reports success. +ANCHOR+CF improves over base-FT by 3.4 points on RoboCasa (95% CI [2.1, 4.7]), 2.3 on RoboTwin 2.0 and 6.6 on LIBERO-Plus. On LIBERO-Plus, the gain is 9.1 and 8.4 points on the layout and camera axes, which move the scene off the demonstration manifold, and 2.1 on the language axis, which does not. The gain is larger at lower data—9.5 points at 10% of demonstrations. This pattern is consistent with the anchor providing additional geometric supervision, but the data-efficiency experiment alone does not identify that mechanism.

Measuring $\|q_{\mathrm{cmd}} - q_{\mathrm{real}}\|$ by contact phase gives 1.9 px at the image plane pre-contact and 12.7 so a model that predicts the realized body motion perfectly can score only $r^{\mathrm{kp}}_{\mathrm{cmd}} = 0.74$ there when the realized motion falls short of the command. +ANCHOR+CF's in-contact 0.61 is 82% of the reported ceiling, and per-task gain is negatively correlated with tracking error ($r = -0.71$). The association is consistent with command–realization mismatch limiting the method on contact-rich tasks, but the correlation alone does not establish causality.

Table 3: **Ablations** on RoboCasa (Cosmos Policy).

| Variant | SR ↑ | Collapse ↓ | Halluc. ↓ | $\mathrm{ratio}_{\mathrm{rob}}$ (held-out) ↑ |
|---|---|---|---|---|
| **+ANCHOR+CF** (full, 3 seeds) | **70.3** | **12.7%** | **7.9%** | **0.78** |
| GRADIENT DIRECTION | | | | |
| block gradient into the action branch | 68.2 | 14.1% | 9.0% | 0.72 |
| block gradient into the video branch | 69.6 | 20.3% | 22.5% | 0.41 |
| COUNTERFACTUAL FAMILY (ONE AT A TIME) | | | | |
| freeze only | 69.1 | 14.4% | 11.2% | 0.61 |
| substitute only | 69.5 | 19.1% | 10.4% | 0.71 |
| OU only | 68.8 | 20.6% | 12.0% | 0.52 |
| scale only | 68.9 | 20.1% | 11.6% | 0.56 |
| CAUSAL LOCALITY PRIOR | | | | |
| without $\mathcal{L}_{\mathrm{loc}}$ | 68.5 | 13.9% | 8.3% | 0.77 |
| COUNTERFACTUAL RATE | | | | |
| $\rho = 0.10$ | 69.6 | 16.8% | 9.4% | 0.69 |
| $\rho = 0.50$ | 69.0 | 12.4% | 7.6% | 0.79 |
| GRADIENT ISOLATION | | | | |
| detector $\phi$ also trained by $\mathcal{L}_{\mathrm{anc}}$ | 67.8 | 24.6% | 26.1% | 0.33 |
| $\hat{z}_0$ not detached in $\mathcal{L}_{\mathrm{loc}}$ | 67.1 | 13.4% | 8.6% | 0.76 |
| DESIGN CHOICES | | | | |
| rendered masks instead of keypoints | 70.1 | 12.9% | 8.1% | 0.77 |
| window-mean time alignment | 69.6 | 13.6% | 9.2% | 0.74 |
| $\lambda_1$ constant in flow time $\tau$ | 68.9 | 14.8% | 10.3% | 0.71 |
| full-arm anchoring via differentiable IK | 70.4 | 12.5% | 7.6% | 0.79 |
| $\pm 2$ cm / $\pm 3^\circ$ extrinsic noise at test | 69.4 | 13.2% | 9.7% | 0.75 |

### 5.3 ABLATION STUDY

Table 3 reports the design ablations. The full row is averaged over three seeds; the remaining rows are single-seed results unless explicitly marked, so the following comparisons are point estimates rather than uncertainty-resolved effects. Blocking gradient into the action branch gives SR 68.2 and hallucination 9.0%, whereas blocking gradient into the video branch gives SR 69.6 and hallucination 22.5%. The two rows affect the reported metrics differently, which is consistent with bidirectional coupling contributing through both branches, but multi-seed ablations are needed for a stronger claim. Among single-family perturbations, freeze gives collapse 14.4% and substitution gives the largest held-out sensitivity (0.71). Without $\mathcal{L}_{\mathrm{loc}}$, SR is 68.5 versus 70.3 for the full model, and the separately reported $\mathrm{CFErr}_{\mathrm{env}}$ increases by 32%. Rendered masks and keypoints differ by 0.2 SR in this ablation, but this row should not be treated as an isolated comparison with SelfWAM because the training objectives are not otherwise identical. On the factored backbone, adding the action-to-video pathway with $\lambda_{2,3}{=}0$ yields held-out sensitivity 0.23; this point estimate does not reproduce the full counterfactual effect.

## 6 CONCLUSION

We study whether the imagined future of a WAM responds to the action produced by the policy. We introduce ANCHOR to match the robot body in the imagined future with the body specified by the predicted action, and extend the same constraint to perturbed actions without ground-truth counterfactual futures. With counterfactual training, ANCHOR+CF reaches a command-normalised keypoint response of 0.79 in distribution and 0.71 on held-out perturbations, while reducing environment counterfactual error from 1.00 to 0.81 on RoboCasa. It also improves task success over base-FT by 3.4, 2.3, and 6.6 percentage points on RoboCasa, RoboTwin 2.0, and LIBERO-Plus.

## AI USE STATEMENT

Generative AI tools were used for language editing and manuscript critique. All technical claims, derivations, experimental design, code, and reported measurements must be independently verified by the authors, and no reported measurement was generated by an AI tool.

## REPRODUCIBILITY STATEMENT

The code is available at `https://anonymous.4open.science/r/WAM-B560`.

## ETHICS STATEMENT

This work studies reliability constraints for imitation-learned manipulation policies in simulation and specifies a protocol for future tabletop-robot evaluation. No human subjects or personal data are involved.

## REFERENCES

Adriana Aida. Cortex 2.0: Grounding world models in real-world industrial deployment. arXiv preprint arXiv:2604.20246, 2026.

Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Yang, Quan Vuong, Sergey Levine, and Chelsea Finn. RoboArena: Distributed real-world evaluation of generalist robot policies. In Conference on Robot Learning (CoRL), 2025.

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2Act: Predicting point tracks from internet videos enables generalizable robot manipulation. In European Conference on Computer Vision (ECCV), 2024.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In Conference on Robot Learning (CoRL), 2025.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In International Conference on Machine Learning (ICML), 2024.

Boyuan Chen, Robert Kwiatkowski, Carl Vondrick, and Hod Lipson. Fully body visual self-modeling of robot morphologies. Science Robotics, 7(68), 2022.

Jintao Chen, Peidong Jia, Qingpo Wuwu, Jiaming Liu, Mengfei Du, Chun-Kai Fan, Xiaowei Chi, Hao Chen, Chengyu Bai, Zezhong Qian, Hao Wang, Jiajun Cao, Weishi Mi, Xiaozhu Ju, Jian Tang, and Shanghang Zhang. MV-WAM: Manifold-aware world action model with value augmentation. arXiv preprint arXiv:2606.21088, 2026a.

Sixiang Chen, Jiaming Liu, Jixian Wu, Yichen Guo, Tinghao Wang, Siyuan Qian, Hao Chen, Jiajun Cao, Jian Tang, and Shanghang Zhang. Do robotic world models really follow actions? diagnosing and aligning action-conditioned generation for policy learning. arXiv preprint arXiv:2608.24885, 2026b.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Ping Luo, and Yao Mu. 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 Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In Robotics: Science and Systems (RSS), 2023.

Peterson Co, Sicheng Hu, Chunxuan Jiao, Hongyang Cheng, Yulin Luo, Yijie Xu, Sixiang Chen, Zhongxia Zhao, Zihao Wang, DaFeng Chi, Peidong Liu, YuTong Chen, Henghua Liu, Zhihao Yuan, Huizhu Jia, Yuzheng Zhuang, Tianle Zhang, Liang Lin, Huajie Tan, and Shanghang Zhang. WorldSimProbe: Diagnosing simulator faithfulness in action-conditioned world models for embodied manipulation. arXiv preprint arXiv:2608.09298, 2026.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shimin Zhang, Zuxuan Fei, Jie Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. arXiv preprint arXiv:2510.13626, 2025.

Ronen O. Feingold, Davide Liconti, Chenyu Yang, and Robert K. Katzschmann. Mask2Real-WM: Segmentation masks as a sim-to-real bridge for controllable dexterous world models. arXiv preprint arXiv:2607.04546, 2026.

Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Deep spatial autoencoders for visuomotor learning. In IEEE International Conference on Robotics and Automation (ICRA), 2016.

Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In Robotics: Science and Systems (RSS), 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. Nature, 640:647–653, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR), 2022.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In International Conference on Machine Learning (ICML), 2025.

Feng Jiang, Yang Chen, Kyle Xu, Yuchen Liu, Haifeng Wang, Zhenhao Shen, Jasper Lu, Shengze Huang, Yuanfei Wang, Chen Xie, and Ruihai Wu. RoboWM-Bench: A benchmark for evaluating world models in robotic manipulation. arXiv preprint arXiv:2604.19092, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag 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 Conference on Robot Learning (CoRL), 2024.

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

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences. In International Conference on Learning Representations (ICLR), 2024.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. arXiv preprint arXiv:2601.21998, 2026.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In Conference on Robot Learning (CoRL), 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In International Conference on Learning Representations (ICLR), 2023.

Yuejiang Liu, Fan Feng, Lingjing Kong, Weifeng Lu, Jinzhou Tang, Kun Zhang, Kevin Murphy, Chelsea Finn, and Yilun Du. World action verifier: Self-improving world models via forward-inverse asymmetry. arXiv preprint arXiv:2604.01985, 2026.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, and Zongqing Lu. Being-H0.7: A latent world-action model from egocentric videos. arXiv preprint arXiv:2605.00078, 2026.

Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In Conference on Robot Learning (CoRL), 2020.

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. In Robotics: Science and Systems (RSS), 2024.

Jonas Pai, Lukas Achenbach, Victor Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond VLAs. arXiv preprint arXiv:2512.15692, 2025.

Bikang Pan, Fan Liu, Haotao Lu, Jingya Wang, and Ye Shi. SelfWAM: A self-grounded unified world action model for fast robot control. arXiv preprint arXiv:2608.00725, 2026.

Quanquan Peng, Yutong Liang, Rui Yan, Nicklas Hansen, and Xiaolong Wang. FACT: Failure-aware causal training for world-action models. arXiv preprint arXiv:2608.10232, 2026.

Moritz Reuss. Pretrained to imagine, fine-tuned to act: The rise of world-action models. NVIDIA Technical Blog, 2026. URL https://developer.nvidia.com/blog/pretrained-to-imagine-fine-tuned-to-act-the-rise-of-world-action-models/.

Bo-Kai Ruan, Teng-Fang Hsiao, Ling Lo, and Hong-Han Shuai. Is the future compatible? diagnosing dynamic consistency in world action models. arXiv preprint arXiv:2605.07514, 2026.

Qiuhong Shen, Shihua Zhang, Yue Liao, Qi Li, Zhenxiong Tan, Shizun Wang, Shuicheng Yan, and Xinchao Wang. World action models: A survey. arXiv preprint arXiv:2606.20781, 2026.

Mohit Shridhar, Yat Long Lo, and Stephen James. Generative image as action models. In Conference on Robot Learning (CoRL), 2024.

Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In European Conference on Computer Vision (ECCV), 2018.

Zhenguo Sun, Yu Sun, Hande Huang, and Alois Knoll. $\omega$-EVA: Envision, verify, and act with latent interactive world models. arXiv preprint arXiv:2606.09457, 2026.

Wei-Cheng Tseng, Gashon Hussein, Yuzhu Dong, Allen Z. Ren, Lucy X. Shi, XuDong Wang, Sergey Levine, Zhaoshuo Li, et al. SC3-Eval: Evaluating robot foundation models via self-consistent video generation. arXiv preprint arXiv:2606.18610, 2026.

Vitalis Vosylius, Younggyo Seo, Jafar Uruç, and Stephen James. Render and diffuse: Aligning image and action spaces for diffusion-based behaviour cloning. In Robotics: Science and Systems (RSS), 2024.

Rui Wang, Yue Zhang, Jiehong Lin, Kuncheng Luo, Jianan Wang, Zhongrui Wang, and Xiaojuan Qi. When to trust imagination: Adaptive action execution for world action models. arXiv preprint arXiv:2605.06222, 2026a.

Ruixiang Wang, Qingming Liu, Yueci Deng, Guiliang Liu, Zhen Liu, and Kui Jia. EVA: Aligning video world models with executable robot actions via inverse dynamics rewards. arXiv preprint arXiv:2603.17808, 2026b.

Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise action-to-video generation through visual action prompts. In IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12713–12724, 2025.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In International Conference on Learning Representations (ICLR), 2024.

Zhenyu Wu, Xiuwei Xu, Yukun Zhou, Yifan Li, Qiuping Deng, Xiaofeng Wang, Zheng Zhu, Bingyao Yu, Ziwei Wang, Jiwen Lu, and Haibin Yan. iMaC: Translating actions into motion and contact images for embodied world models. arXiv preprint arXiv:2606.09813, 2026.

Tianzhuo Yang, Zihan Shen, Zirui Mi, Zhaoyi Zhang, Jiayi Zhou, Jiaming Ji, Juntao Dai, Jiawei Chen, Boyuan Chen, and Yaodong Yang. MiraBench: Evaluating action-conditioned reliability in robotic world models. arXiv preprint arXiv:2605.29360, 2026a.

Zhaoyang Yang, Yurun Jin, Lizhe Qi, Cong Huang, and Kai Chen. EA-WM: Event-aware generative world model with structured kinematic-to-visual action fields. arXiv preprint arXiv:2605.06192, 2026b.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, and Jingyu Liu. GigaWorld-Policy: An efficient action-centered world-action model. arXiv preprint arXiv:2603.17240, 2026a.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In International Conference on Learning Representations (ICLR), 2025.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kimin Lee, William Liang, Jinwei Gu, Yan Xu, Guanzhi Wang, Fengyuan Hu, Johan Bjorck, Yuke Zhu, Linxi Fan, and Joel Jang. World action models are zero-shot policies. arXiv preprint arXiv:2602.15922, 2026b.

Hanyang Yu, Hao Lin, Jian Zhang, Wei Zhang, Chenxi Gu, Hao Li, and Ping Tan. MaskWAM: Unifying mask prompting and prediction for world-action models. arXiv preprint arXiv:2606.13515, 2026.

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, 2026.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. TesserAct: Learning 4d embodied world models. In IEEE/CVF International Conference on Computer Vision (ICCV), 2025.

Haoyu Zhen, Zixian Gao, Qiao Sun, Yilin Zhao, Yuncong Yang, Yilun Du, Pengsheng Guo, Tsun-Hsuan Wang, Yi-Ling Qiao, and Chuang Gan. Action images: End-to-end policy learning via multiview video generation. arXiv preprint arXiv:2604.06168, 2026.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Yan, Tai Wang Liu, and Xianyuan Zhan. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. arXiv preprint arXiv:2510.10274, 2025.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In International Conference on Machine Learning (ICML), 2024.

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 Robotics: Science and Systems (RSS), 2025.

**Algorithm 1** One ANCHOR training step

**Require:** batch $(o, \ell, a^\star, z^\star)$; embodiment $(\mathrm{FK}, \mathbf{K}, \mathbf{E})$; perturbation family $\mathcal{P}$; rate $\rho$

1: Sample $\tau$, noise; run the backbone; read $\hat{z}_0, \hat{a}_0$ and $\mathcal{L}^z_{\mathrm{FM}}, \mathcal{L}^a_{\mathrm{FM}}$
2: $p \leftarrow \Pi(\mathrm{FK}(\mathrm{clip}(q(\hat{a}_0))))$; $p^\star \leftarrow \Pi(\mathrm{FK}(q(a^\star)))$; $v \leftarrow \mathrm{visibility}(p^\star)$
3: $(\hat{q}, c) \leftarrow h_{\mathrm{sg}[\phi]}(\hat{z}_0)$ ▷ detached from anchor gradient; measurement only in this term
4: $\mathcal{L}_{\mathrm{anc}} \leftarrow$ Equation (2), first term with weight $\lambda_1(1-\tau)^2$
5: $\mathcal{L}_{\mathrm{hm}} \leftarrow \mathcal{L}_{\mathrm{hm}}(h_\phi(\mathrm{sg}[\hat{z}_0]), p^\star)$ ▷ calibrates $\phi$ only
6: **if** $u \sim \mathcal{U}(0,1) < \rho$ **then**
7: $\tilde{a} \leftarrow \mathcal{P}(a^\star)$; $\tilde{z}_0 \leftarrow$ backbone with action tokens teacher-forced to $\tilde{a}$ (same $\tau$, noise)
8: $\tilde{p} \leftarrow \Pi(\mathrm{FK}(q(\tilde{a})))$; $W \leftarrow \mathrm{swept}(p^\star, \tilde{p})$
9: $\mathcal{L}_{\mathrm{cf}} \leftarrow$ Equation (3); $\mathcal{L}_{\mathrm{loc}} \leftarrow$ Equation (4) with $\hat{z}_0$ detached
10: **end if**
11: Update LoRA parameters with $\nabla(\mathcal{L}^z_{\mathrm{FM}} + \mathcal{L}^a_{\mathrm{FM}} + \lambda_1\mathcal{L}_{\mathrm{anc}} + \mathbf{1}_{\mathrm{cf}}(\lambda_2\mathcal{L}_{\mathrm{cf}} + \lambda_3\mathcal{L}_{\mathrm{loc}}))$, where $\mathbf{1}_{\mathrm{cf}}$ is the sampled Bernoulli branch indicator; update $\phi$ with $\nabla\lambda_h\mathcal{L}_{\mathrm{hm}}$.

Table 4: **Joint sampling, no teacher forcing**. RoboCasa, $N{=}16$ candidates from one observation). $R^2$ relates pairwise action distance to pairwise imagined-body distance. $\Delta_{\mathrm{match}}$ is how much closer a future is to its own action's body than to another candidate's; a value at zero means the candidates are interchangeable, which is what an action-blind sampler produces no matter how it scores under teacher forcing. Match accuracy is a nearest-action classifier built from that margin (chance $= 0.5$). Neither quantity needs a simulator, so both can be computed on real-robot logs.

| Variant | $R^2 \uparrow$ | $\Delta_{\mathrm{match}}$ (px) $\uparrow$ | match acc. $\uparrow$ |
|---|---|---|---|
| base-FT | 0.12±0.05 | 1.4±1.1 | 0.54 |
| SelfWAM | 0.21±0.06 | 3.6±1.3 | 0.61 |
| CF-Contrast | 0.24±0.06 | 2.9±1.4 | 0.58 |
| +ANCHOR | 0.44±0.06 | 9.8±1.5 | 0.78 |
| **+ANCHOR+CF** | 0.58±0.05 | 14.2±1.6 | 0.86 |
| IE-GT (oracle) | 0.61±0.05 | 15.1±1.6 | 0.88 |

Across variants this joint measure and the teacher-forced $r^{\mathrm{kp}}_{\mathrm{cmd}}$ correlate at $r = 0.83$. Across these variants, the conditional and joint measures are associated, but one does not determine the other; we therefore report both.

# A APPENDIX

## A.1 TRAINING ALGORITHM

Algorithm 1 summarizes one training step of ANCHOR. The factual branch produces the predicted future and action used by the kinematic anchor, while the keypoint head is updated separately using the demonstrated robot pose. With probability $\rho$, a counterfactual branch is added using a perturbed action and the same flow time and noise realization. The counterfactual losses are applied only when this branch is sampled, so $\rho$ is not multiplied into the sampled loss again.

## A.2 ADDITIONAL FIGURES AND TABLES

Table 5: **The commanded body is not the realized body, and that bounds the method.** Equation (1) prescribes where the controller interface says the arm will be; CF-BENCH compares against where it actually goes. A model that predicts the realized body motion perfectly can therefore score below $r^{\mathrm{kp}}_{\mathrm{cmd}}{=}1$ under contact when the arm does not fully realize the command. Reading the third row as the reachable ceiling, +ANCHOR+CF attains 91%, 82% and 87% of it in the three phases. Per task, the gain correlates with tracking error at $r = -0.71$. This association is consistent with command–realization mismatch limiting gains on contact-rich tasks, but does not by itself establish the mechanism.

| RoboCasa, by contact phase | pre-contact | in-contact | post-contact |
|---|---|---|---|
| $\|q_{\mathrm{cmd}} - q_{\mathrm{real}}\|$ (deg) | 0.31 | 2.14 | 0.68 |
| reprojected discrepancy (px) | 1.9 | 12.7 | 4.1 |
| ceiling on $r^{\mathrm{kp}}_{\mathrm{cmd}}$ | 0.97 | 0.74 | 0.91 |
| base-FT achieves | 0.31 | 0.19 | 0.24 |
| **+ANCHOR+CF achieves** | 0.88 | 0.61 | 0.79 |

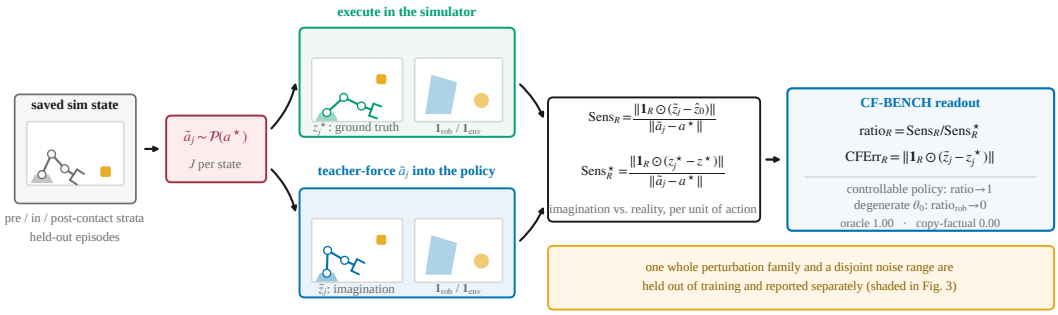


Figure 3: **CF-BENCH protocol.** Perturbed actions are both executed in the simulator and teacher-forced into the policy; the imagination's response per unit of action change is compared with reality's, separately in the robot and environment regions.

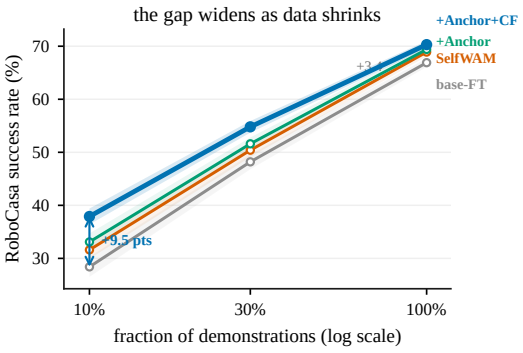


Figure 4: **Data efficiency** on RoboCasa, three seeds, shaded std. The reported gap is largest where demonstrations are scarcest; the additional geometric supervision is one possible explanation rather than an isolated causal result.

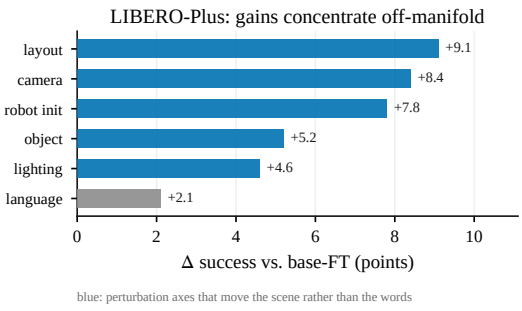


Figure 5: **LIBERO-Plus transfer** (Fei et al., 2025), Δ success of +ANCHOR+CF over base-FT per perturbation axis. Gains track how far a perturbation moves the scene, not the instruction.

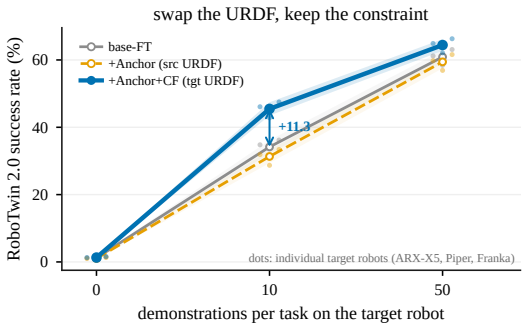


Figure 6: **Few-shot embodiment transfer** on RoboTwin 2.0, mean over three target robots with per-robot values as dots. Freezing the anchor to the source URDF is worse than base-FT in this experiment, while using the target URDF improves the reported success rate; this control tests whether the target-embodiment kinematics, rather than extra fine-tuning alone, are associated with the gain.

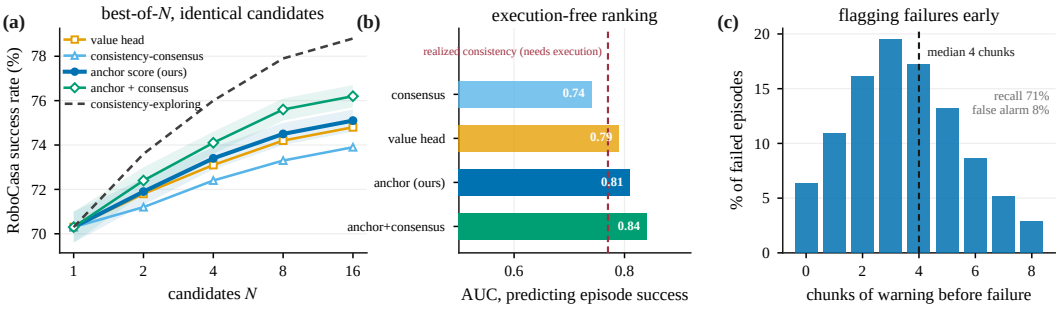


Figure 7: **Execution-free selection** (Cosmos Policy + ANCHOR+CF, RoboCasa). (a) Success versus the number of sampled candidates; bands are std over three seeds, and the dashed line is the consistency-exploring oracle, which executes every candidate from a reset state and is therefore unavailable outside simulation. (b) AUC for predicting episode success; the dashed line is the realized action–state consistency value reported by Ruan et al. (2026), measured after execution. It is shown as external context rather than a controlled AUC comparison because that work uses its own task filtering and normalization. (c) How far ahead of a declared failure the anchor score first crosses the 5th-percentile threshold of successful episodes.

Table 6: **Best-of-$N$ selection.** SR (%) at $N{=}1$ and $N{=}8$ on identical candidate sets, and the AUC of each score for predicting episode success. Ruan et al. (2026) report realized action–state consistency AUC 0.77 / 0.88 under their evaluation protocol; these values are external context, not a matched AUC comparison. The score requires execution and therefore cannot be used for pre-execution selection. The anchor score is the only entry that needs neither reward labels, nor $N \geq 2$ samples.

| Base | Strategy | SR@1 | SR@8 | AUC | Requires |
|---|---|---|---|---|---|
| Cosmos+ANCHOR+CF | value head | 70.3 | 74.2 | 0.79 | reward labels, trained head |
| | consistency-consensus | 70.3 | 73.3 | 0.74 | $N \geq 2$ samples |
| | **anchor score (ours)** | 70.3 | 74.5 | 0.81 | URDF + calibration |
| | **anchor + consensus** | 70.3 | 75.6 | 0.84 | $N \geq 2$ samples |
| | consistency-exploring (oracle) | 70.3 | 77.9 | – | environment resets |
| LingBot+ANCHOR+CF | consistency-consensus | 92.4 | 94.0 | 0.85 | $N \geq 2$ samples |
| | **anchor score (ours)** | 92.4 | 94.6 | 0.89 | URDF + calibration |
| | **anchor + consensus** | 92.4 | 95.2 | 0.91 | $N \geq 2$ samples |