GEOACT: RECONCILING VIDEO AND ACTION WITH ROBOT GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

World action models generate a future video and an action chunk for a robot in one sampling process. The training loss does not tie the two outputs to the robot's kinematics. As a result, the video and the action can describe different arm motions. We propose GEOACT, a sampler that uses robot geometry to reconcile the two outputs during generation. At selected denoising steps, a frozen geometric backbone and a trained joint readout estimate the arm configurations shown in the clean video prediction. Forward kinematics places this estimate and the predicted action in the robot base frame. A bounded proximal step moves the clean action toward the visual estimate. GEOACT reinjects the corrected action at the current noise level and evaluates the world action model again. The remaining video generation can then respond to the correction. On a shared backbone, GEOACT raises task success from 94.5% to 97.5% on LIBERO and from 81.8% to 89.0% on RoboTwin 2.0. Post-hoc correction with the same readout and the same budget reaches 95.6% and 83.9%. Ablations show that the gain requires a fresh reading of the updated video and a bound on each update. Two interventions per action chunk cost $1.22\times$ the base latency.

1 INTRODUCTION

Video generation has become a basis for robot control (Du et al., 2023; Yang et al., 2023; Wu et al., 2023; Cheang et al., 2024). World action models generate a future video and an action chunk from the same sampling process (Li et al., 2025; Ye et al., 2026; Yuan et al., 2026). DreamZero (Ye et al., 2026) is one example: it predicts future images and joint targets from camera views, robot state, and a language instruction. The video shows what the robot expects to happen. The action tells the controller what to do. Both outputs should describe the same motion of the same robot. For example, a video of the gripper reaching a cup should come with joint targets that reach the same cup.

Joint training does not enforce this agreement. The training loss makes the video and the action statistically dependent. The training loss does not require the two outputs to describe the same robot configurations. Robot kinematics makes the difference measurable. The video shows a sequence of arm configurations. The action specifies a sequence of target configurations. Forward kinematics maps both sequences to end-effector positions in the robot base frame. We call the distance between the two position sequences the *video–action agreement gap*.

A simple fix corrects the action after generation. Post-hoc correction moves the final action toward the motion shown in the final video. The final video cannot respond to this edit. A changed approach path can change contact, object motion, and the following frames. Post-hoc correction never sees these consequences. Prior work adds geometry or feedback at other points. Geometry Forcing aligns video features with a geometric foundation model during training (Wu et al., 2026). The Geometric Action Model predicts future geometry and actions inside a geometric foundation model (Han et al., 2026). Feedback Flow Matching constrains sampling with observed states and committed actions (Li et al., 2026b). Diffusion guidance imposes external measurements or recognition signals during sampling (Chung et al., 2022; Bansal et al., 2024). None of these methods uses the disagreement between two predictions that are still being generated.

We propose GEOACT, a sampler that corrects the action inside joint generation, as Figure 1 shows. At selected denoising steps, GEOACT decodes the clean video prediction. A frozen VGGT backbone (Wang et al., 2025) and a trained joint readout estimate the arm configurations in this video.

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

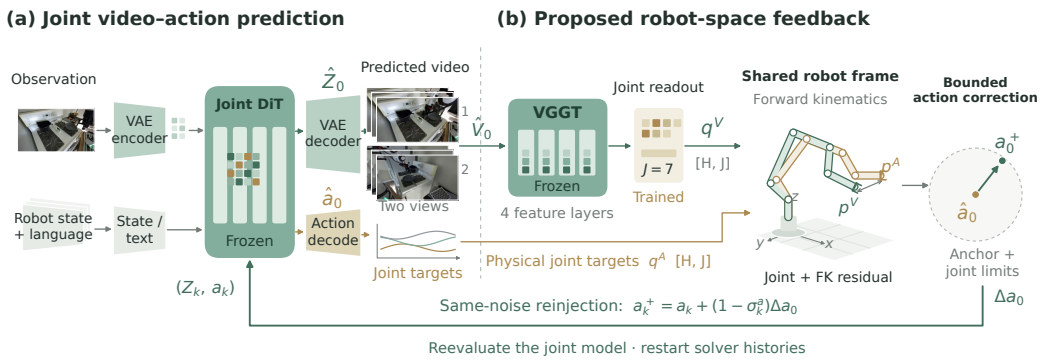


Figure 1: **The GEOACT feedback loop.** (a) A frozen world action model predicts video and action from the same sampling state. (b) The joint readout and forward kinematics place both motions in the robot base frame. The difference between the two motions drives a bounded action correction. Same-noise reinjection returns the corrected action to sampling, and the world action model runs again. The world action model and the geometric backbone stay frozen. Only the joint readout is trained, on real robot videos. Predicted frames come from an unedited DreamZero sample. Robot poses and action curves are schematic.

GEOACT compares these configurations with the decoded action in joint space and in end-effector space. A proximal objective moves the clean action toward the visual estimate and keeps the action close to its current prediction. GEOACT returns the corrected action to the current noise level and keeps the inferred noise endpoint fixed. The world action model then runs again on the corrected state. The next video prediction can respond to the corrected action. The next intervention reads this new video. The world action model and the geometric backbone stay frozen. Only the joint readout needs training.

We test GEOACT on LIBERO and RoboTwin 2.0. The main baseline is Post-hoc correction with the same readout, the same objective, and the same correction budget. GEOACT beats this baseline on every LIBERO suite and in both RoboTwin scene settings. The margin is largest on long-horizon tasks and in randomized scenes. Ablations trace the gain to two parts: the fresh reading of the updated video and the bound on each update. Two interventions per action chunk give most of the gain at a small latency cost.

This paper makes three contributions. First, it defines video–action agreement in robot coordinates. This definition makes the depicted motion and the commanded motion directly comparable. Second, it introduces a sampler that corrects the clean action, reinjects the correction with the same noise, and lets joint generation respond. Third, it shows in simulation that correction inside generation beats an equally strong correction after generation, in Section 4.

2 RELATED WORK

Generative robot policies. Diffusion Policy models visuomotor actions with a conditional denoising process and executes them with receding-horizon control (Chi et al., 2025). Diffuser plans whole trajectories with a diffusion model (Janner et al., 2022), and ACT predicts action chunks with a transformer (Zhao et al., 2023). Vision-language-action models map images and instructions to actions at scale (Brohan et al., 2023; Octo Model Team et al., 2024; Kim et al., 2024; Black et al., 2024). A second line of work predicts future frames and derives actions from them. UniPi generates a video plan and recovers actions with an inverse dynamics model (Du et al., 2023). GR-1, GR-2, Seer, and Video Prediction Policy pretrain on video prediction and decode actions from predictive features (Wu et al., 2023; Cheang et al., 2024; Tian et al., 2024; Hu et al., 2024). The Unified Video Action model generates video and action with one diffusion model (Li et al., 2025). DreamZero predicts video and action jointly on top of a pretrained video backbone (Ye et al., 2026). Motus, LingBot-VA, and Fast-WAM are also world action models (Bi et al., 2025; Li et al., 2026a; Yuan

et al., 2026). In these models, the future video is an extra output next to the action. Fast-WAM keeps video co-training but skips video generation at inference, and its success rate stays close to the joint variants (Yuan et al., 2026). This finding suggests that current world action models make little use of the generated video at test time. GEOACT gives the generated video a test-time role: the video checks the action before execution.

Geometry for visual prediction and control. DUSt3R and MASt3R regress point maps from image pairs (Wang et al., 2024a; Leroy et al., 2024). VGGT predicts camera parameters, depth, point maps, and tracks from images with one network (Wang et al., 2025). Geometry Forcing aligns the internal features of a video diffusion model with a geometric foundation model during training (Wu et al., 2026). The Geometric Action Model adds a causal future predictor to a geometric foundation model and decodes future geometry and actions (Han et al., 2026). These methods put geometry into training. DREAM estimates the camera-to-robot transform from a single image (Lee et al., 2020), and RoboPose also estimates the joint angles (Labbé et al., 2021). The GEOACT readout solves a related problem on generated frames with a frozen VGGT backbone. GEOACT uses geometry at sampling time and does not retrain the world action model.

Feedback during execution and sampling. Future Forward Dynamics Causal Attention compares predicted futures with observations to choose how long to execute an action chunk (Wang et al., 2026). Feedback Flow Matching constrains active sampling with observed states and with the overlap of committed actions (Li et al., 2026b). Real-time chunking keeps consecutive action chunks consistent through inpainting during sampling (Black et al., 2025). Bidirectional Decoding selects action chunks at test time by their consistency with earlier decisions (Liu et al., 2024). Inference-time policy steering guides a pretrained policy with human inputs (Wang et al., 2024b). These methods take feedback from the environment, from past actions, or from a user. GEOACT takes feedback from the model's own video prediction, before any action is executed.

Guidance for pretrained generators. Flow Matching and rectified flow learn a vector field along a chosen path between noise and data (Lipman et al., 2022; Liu et al., 2022). Classifier guidance and classifier-free guidance steer diffusion sampling toward a condition (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). Diffusion Posterior Sampling adds measurement consistency to sampling for noisy inverse problems (Chung et al., 2022). FreeDoM and Universal Guidance steer a diffusion model with pretrained energy or guidance functions (Yu et al., 2023; Bansal et al., 2024). GEOACT also edits the clean prediction during sampling. The measurement in GEOACT is a second output of the same generator, compared through robot kinematics.

3 GEOMETRY-GUIDED JOINT SAMPLING

3.1 COMPARING DEPICTED AND COMMANDED MOTION

The comparison needs physical action units, aligned timestamps, and one robot coordinate frame. Let $V \in \mathbb{R}^{T\times M\times H_v\times W_v\times 3}$ denote a decoded future video with M views and T timestamps. The model state holds video latents Z and normalized action coordinates $a \in \mathbb{R}^{H_a\times d_a}$. The conditioning $y = (o, s, c)$ holds visual observations, the measured robot state, and a language instruction. The pretrained world action model defines a joint distribution $p_\theta(Z, a \mid y)$, and the video decoder gives $V = D(Z)$.

Forward kinematics needs physical joint targets. We write $Q_s(a) \in \mathbb{R}^{H\times J}$ for the map from normalized actions to joint targets at aligned timestamps. Here J is the number of arm joints. The map Q_s undoes the normalization and the relative-action convention of the checkpoint, with the observed state as reference. The robot model $\mathcal{R}$ defines forward kinematics $f_{\mathcal{R}}(q) \in \mathbb{R}^3$, the end-effector position in the robot base frame. The commanded motion at an aligned timestamp h is

$$q_h^A = [Q_s(a)]_h, \qquad p_h^A = f_{\mathcal{R}}(q_h^A). \tag{1}$$

These positions describe commanded targets. The executed motion also depends on the controller.

The visual readout gives a second configuration estimate at the same timestamps. A learned map G_ϕ predicts q_h^V from the video, and the same kinematic chain gives $p_h^V = f_{\mathcal{R}}(q_h^V)$. Both estimates share

the joint convention, the end-effector definition, the base frame, and the metric scale. The metric scale comes from the known robot geometry and the joint supervision. The comparison needs no metric reconstruction of the scene. Video frames and action targets are aligned by their physical timestamps and the command convention of the controller.

We report the agreement gap in centimeters:

$$d_{\mathrm{EE}}(V, a) = \frac{1}{H} \sum_{h=1}^{H} \left\| f_{\mathcal{R}}(q_h^V) - f_{\mathcal{R}}(q_h^A) \right\|_2 . \tag{2}$$

The correction energy adds a joint term. The joint term separates configurations that place the end effector at nearby positions:

$$E(a; q^V, w) = \frac{1}{2H} \sum_{h=1}^{H} w_h \left[\left\| q_h^A - q_h^V \right\|_{W_q}^2 + \frac{\beta}{\ell^2} \left\| p_h^A - p_h^V \right\|_2^2 \right] . \tag{3}$$

Here W_q normalizes squared joint differences, ℓ is a reference length in meters, and β balances the two dimensionless terms. The reliability weights $w_h \in [0, 1]$ stay fixed during one correction. GEOACT uses equal weights in all experiments.

3.2 READING JOINT CONFIGURATIONS FROM VIDEO

A frozen geometric backbone supplies visual features. Learned joint queries convert these features into robot configurations. The readout takes synchronized exterior camera views and runs the original VGGT backbone (Wang et al., 2025). For each timestamp, features from four backbone layers are concatenated along the channel dimension. The reader keeps spatial, camera, and register tokens. Each joint query attends across both views. Register tokens summarize each input image. Spatial tokens keep local evidence about the arm.

The head predicts joint angles directly. Forward kinematics then gives end-effector positions. A linear projection produces head features. Each following block applies cross-attention to visual tokens, self-attention among joint queries, and a feed-forward network. Each joint query outputs a scalar u_j, and the joint angle is

$$q_j^V = q_j^{\mathrm{mid}} + r_j \, b \tanh(u_j), \tag{4}$$

where q_j^{mid} and r_j are the midpoint and the half-range of the joint limits. The factor $b = 1.05$ leaves a small margin for regression. The sampler checks physical limits separately. The bounded form removes distant periodic solutions. Such solutions can give similar end-effector positions under forward kinematics.

Training uses measured configurations from real videos. For a measured configuration q, the loss combines normalized joint error and end-effector distance:

$$\mathcal{L}_{\mathrm{read}} = \frac{1}{J} \sum_{j=1}^{J} \left| \frac{q_j^V - q_j}{r_j} \right| + \alpha \left\| f_{\mathcal{R}}(q^V) - f_{\mathcal{R}}(q) \right\|_2 . \tag{5}$$

The coefficient α balances joint error against distance in meters. The end-effector term starts after a joint-only warmup. Gradients update only the readout head. The geometric backbone stays frozen in training and in inference. The DROID head uses seven joint queries, three attention blocks with eight heads each, and a 256-channel projection. Section 4.5 reports its data and validation.

3.3 CORRECTING THE CLEAN ACTION ENDPOINT

GEOACT edits the clean prediction extracted from the current noisy state. For a normalized flow state $x_\sigma = (1 - \sigma)x_0 + \sigma\epsilon$, we use the velocity convention $v = \epsilon - x_0$. The velocity prediction of the model then gives the clean endpoint and the noise endpoint:

$$\hat{x}_0 = x_\sigma - \sigma v_\theta, \qquad \hat{\epsilon} = x_\sigma + (1 - \sigma) v_\theta. \tag{6}$$

Video and action can have different noise levels. Each endpoint uses the noise level of its own modality. At intervention step k, GEOACT decodes $\hat{V}_0 = D(\hat{Z}_0)$ and computes the visual target $q^V = G_\phi(\hat{V}_0)$. The visual target stays fixed during the correction.

The correction solves a proximal problem:

$$a_0^+ \in \arg\min_{a \in \mathcal{C}(\hat{a}_0)} \underbrace{E(a; q^V, w) + \frac{1}{2\eta}\|a - \hat{a}_0\|_2^2}_{\mathcal{J}_k(a)}, \tag{7}$$

Here $\eta > 0$ sets the correction strength. The set $\mathcal{C}$ enforces joint limits and a maximum displacement $\|a - \hat{a}_0\|_\infty \leq \delta$ in normalized action units. Coordinates outside the arm joints stay fixed. In our setting, the gripper coordinate stays fixed. We solve the problem approximately with projected gradient steps. GEOACT accepts a candidate only if the candidate is feasible and lowers $\mathcal{J}_k$. The gradient passes through the action decoder and forward kinematics. The visual target is detached.

The anchored objective gives a descent guarantee for the fixed target. Assume that $\hat{a}_0$ is feasible and that the accepted candidate satisfies $\mathcal{J}_k(a_0^+) \leq \mathcal{J}_k(\hat{a}_0)$. Then

$$E(a_0^+; q^V, w) + \frac{1}{2\eta}\|a_0^+ - \hat{a}_0\|_2^2 \leq E(\hat{a}_0; q^V, w). \tag{8}$$

The guarantee holds for one intervention with one fixed target. An infeasible anchor or an inadmissible readout makes GEOACT skip the intervention.

3.4 RETURNING THE CORRECTION TO JOINT GENERATION

GEOACT keeps the inferred noise endpoint and replaces the clean endpoint. With the action noise level σ_k^a and the endpoint estimate of Equation equation 6, the corrected action state is

$$\begin{aligned} a_k^+ &= (1 - \sigma_k^a)a_0^+ + \sigma_k^a \hat{\epsilon}_a \\ &= a_k + (1 - \sigma_k^a)(a_0^+ - \hat{a}_0). \end{aligned} \tag{9}$$

Substituting $\hat{\epsilon}_a = a_k + (1 - \sigma_k^a)v_a$ from Equation equation 6 gives the second line. The second line shows that the edit is a deterministic shift of the current state. A zero correction leaves the state unchanged. The shift scales with $1 - \sigma_k^a$. A clean edit therefore moves the state little at high noise, and GEOACT intervenes at low-noise steps. The video latent keeps its current value.

GEOACT then evaluates the joint field again at (Z_k, a_k^+) with the same conditioning and the same noise levels. The joint model attends to the corrected action. The new video velocity can therefore differ from the old one. Continued denoising gives a new video estimate. The next intervention reads this new video. This loop is the difference between GEOACT and post-hoc correction.

The solver must handle the jump in state. Cached predictions from before the intervention no longer match the state. GEOACT discards these predictions and restarts the multistep solver history of both modalities. Conditioning caches stay valid. Algorithm 1 includes the reevaluation and the restart.

3.5 RELIABILITY AND COMPUTATION

A wrong visual estimate is the main risk of the correction. The bound δ and the proximal anchor limit the effect of a wrong estimate. GEOACT rejects nonfinite estimates and configurations outside the physical joint limits. GEOACT does not gate corrections on a confidence score. Section 4.5 shows that a small geometric fitting residual does not certify readout accuracy.

GEOACT intervenes only at low-noise steps that leave at least one later denoising step. Each intervention decodes a clean video estimate and runs the geometric backbone and the readout. An accepted intervention also needs one more joint field evaluation and a solver restart. With m attempted and m_{acc} accepted interventions, the extra compute is approximately

$$\Delta C \approx m(C_D + C_G + C_{\text{prox}}) + m_{\text{acc}} C_{\text{WAM}}. \tag{10}$$

The four terms are video decoding, geometry readout, proximal optimization, and joint field evaluation. Section 4.4 measures the realized latency, including the solver restart.

4 EXPERIMENTS

The experiments answer three questions. Does geometric feedback inside generation raise task success over the same correction after generation? Which parts of the feedback loop produce the gain? How much latency does the feedback add?

Algorithm 1 Proposed GEOACT sampler

Require: Frozen field v_θ; conditioning y; readout G_ϕ; action decoder Q_s; robot model $\mathcal{R}$
Require: Noise schedules $\{\sigma_k^v, \sigma_k^a\}$; intervention indices $\mathcal{I}$; correction settings $(\eta, \mathcal{C})$
Initialize video and action noise (Z_1, a_1) and the native solver state
for each solver index k **do**
  $(v_Z, v_a) \leftarrow v_\theta(Z_k, a_k, \sigma_k^v, \sigma_k^a; y)$
  if $k \in \mathcal{I}$ **then**
    $\hat{Z}_0 \leftarrow Z_k - \sigma_k^v v_Z$; $\hat{a}_0 \leftarrow a_k - \sigma_k^a v_a$
    $q^V \leftarrow G_\phi(D(\hat{Z}_0))$; obtain fixed weights w
    if the readout is admissible and $\hat{a}_0 \in \mathcal{C}(\hat{a}_0)$ **then**
      Find a feasible a_0^+ decreasing $\mathcal{J}_k$ in Equation equation 7
      if a nonzero correction is accepted **then**
        $a_k \leftarrow a_k + (1 - \sigma_k^a)(a_0^+ - \hat{a}_0)$
        Invalidate affected predictions and restart the coupled solver history
        $(v_Z, v_a) \leftarrow v_\theta(Z_k, a_k, \sigma_k^v, \sigma_k^a; y)$
      end if
    end if
  end if
  $(Z_{k+1}, a_{k+1}) \leftarrow \mathrm{SolverStep}(Z_k, a_k, v_Z, v_a)$
end for
return Decoded video and physical actions at the final sampling state

4.1 EXPERIMENTAL SETUP

Benchmarks. LIBERO (Liu et al., 2023) has four suites: Spatial, Object, Goal, and Long. Each suite has 10 tasks, and each task has 50 evaluation episodes. RoboTwin 2.0 (Chen et al., 2025) tests coordinated bimanual manipulation in clean and randomized scenes. Each RoboTwin task has 100 trials in each scene setting. The benchmark choice follows Fast-WAM (Yuan et al., 2026). The comparison includes OpenVLA (Kim et al., 2024), π_0 (Black et al., 2024), $\pi_{0.5}$ (Physical Intelligence et al., 2025), Motus (Bi et al., 2025), LingBot-VA (Li et al., 2026a), and Fast-WAM. Published baseline scores come from the evaluation tables of Yuan et al. (2026). The PT column marks embodied pretraining as reported there. A dash marks a score that the source does not report.

Model and action interfaces. Within each benchmark, Base, Post-hoc, and Interleaved use one adapted joint video–action checkpoint. The three samplers also share camera inputs, action horizon, and training data. The world action model and the geometric backbone stay frozen. Only the readout for each embodiment is trained with joint supervision. The readout keeps the architecture of Section 3.2. For RoboTwin, each arm has its own joint queries and its own kinematic chain. LIBERO uses end-effector actions. For LIBERO, a differentiable map anchored at the measured state converts end-effector actions to joint targets. The inverse map returns executable end-effector commands.

Matched comparisons. Base uses the native sampler. Post-hoc corrects the final action against the final video. Interleaved runs Algorithm 1 before sampling completes. The three samplers share initial scenes, instructions, random seeds, initial noise, and execution horizons. Post-hoc and Interleaved share the readout, the geometric objective, the admissibility rules, and the total number of correction steps. Post-hoc and Interleaved therefore differ only in when the correction happens. Readout selection and correction settings use development episodes. The development episodes are disjoint from the test episodes.

Metrics. Task success is the primary metric. Suite averages weight tasks equally. End-effector error, written EE error, compares simulator ground-truth motion with task-defined target poses at annotated task events. EE error does not use the visual reader. The correction objective therefore does not optimize EE error directly. Failed trials are evaluated at the common timeout. The agreement gap follows Equation equation 2 and measures internal consistency between video and action. Latency covers video decoding, geometry readout, correction, field reevaluation, and solver restart. All timings use the same hardware, precision, and cache settings.

Table 1: **LIBERO task success (%, ↑).** Baseline scores and embodied-pretraining (PT) status follow Yuan et al. (2026). Average weights the four suites equally. The last three rows share one checkpoint and one evaluation protocol.

Method	PT	Spatial	Object	Goal	Long	Avg.
OpenVLA	Yes	84.7	88.4	79.2	53.7	76.5
π_0	Yes	96.8	98.8	95.8	85.2	94.1
$\pi_{0.5}$	Yes	98.8	98.2	98.0	92.4	96.9
Motus	Yes	96.8	99.8	96.6	97.6	97.7
LingBot-VA	Yes	98.5	99.6	97.2	98.5	98.5
Fast-WAM	No	98.2	100.0	97.0	95.2	97.6
Shared-backbone sampler comparison						
Base	Yes	96.0	97.2	95.6	89.2	94.5
Post-hoc	Yes	96.8	97.6	96.4	91.6	95.6
Interleaved (GEOACT)	Yes	**98.0**	**98.4**	**98.0**	**95.6**	**97.5**

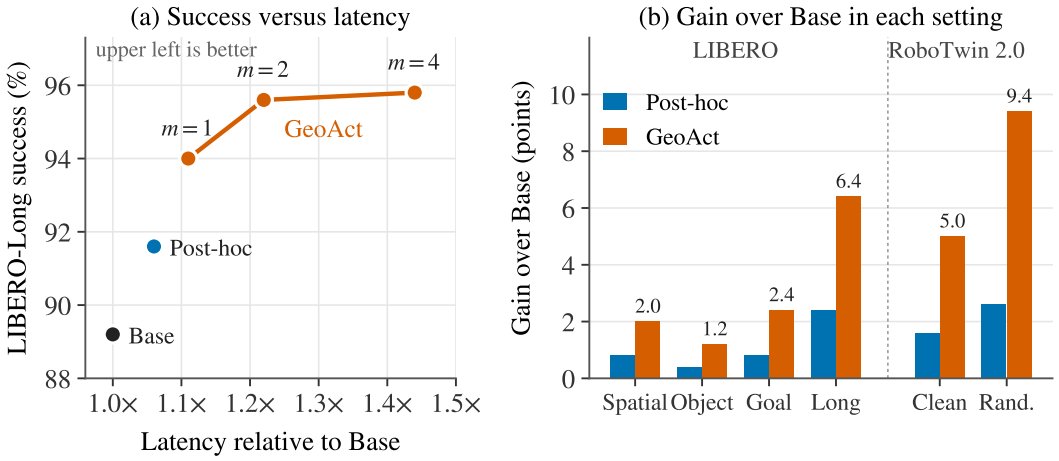


Figure 2: **Feedback inside generation beats the same correction after generation at a small cost.** (a) LIBERO-Long success against latency relative to Base on the same device. The value m counts interventions per action chunk. GEOACT uses $m = 2$ in all tables. (b) Gain over Base in each LIBERO suite and each RoboTwin 2.0 scene setting. Post-hoc and GEOACT share the readout, the objective, and the correction budget.

4.2 SIMULATION PERFORMANCE

Feedback inside generation beats the same correction after generation. Table 1 reports LIBERO. Interleaved raises the LIBERO average from 94.5% for Base to 97.5%. Post-hoc reaches 95.6% with the same readout and the same correction budget. Interleaved leads Post-hoc by 1.9 points on average and leads Post-hoc on every suite. The largest margin is on LIBERO-Long: Base reaches 89.2%, Post-hoc 91.6%, and Interleaved 95.6%.

The margin doubles in randomized scenes. Table 2 reports RoboTwin 2.0. Interleaved raises the average from 81.8% for Base to 89.0%. Post-hoc reaches 83.9%. Interleaved leads Post-hoc by 3.4 points in clean scenes and by 6.8 points in randomized scenes. Scene randomization costs Base 6.8 points. Scene randomization costs Interleaved 2.4 points. Figure 2b shows the gain over Base in each setting: GEOACT gains more than Post-hoc everywhere, and the gap is widest on LIBERO-Long and in randomized scenes.

A sampler change closes most of the gap to the strongest published models. The shared-backbone rows isolate the effect of the sampler. The published rows also differ in backbone, adaptation data, and pretraining. GEOACT does not retrain the world action model. On LIBERO,

Table 2: **RoboTwin 2.0 task success (%, ↑).** Baseline scores and embodied-pretraining (PT) status follow Yuan et al. (2026). Average uses both scene settings. Rows marked from Wan2.2 start from the Wan video model (Wan Team et al., 2025). Fast-WAM does not report the randomized score of LingBot-VA from Wan2.2.

Method	PT	Clean	Randomized	Avg.
π_0	Yes	65.92	58.40	62.2
$\pi_{0.5}$	Yes	82.74	76.76	79.8
Motus	Yes	88.66	87.02	87.8
LingBot-VA	Yes	92.90	91.50	92.2
Motus from Wan2.2	No	77.56	77.00	77.3
LingBot-VA from Wan2.2	No	80.60	–	–
Fast-WAM	No	91.88	91.78	91.8
Shared-backbone sampler comparison				
Base	Yes	85.20	78.40	81.8
Post-hoc	Yes	86.80	81.00	83.9
Interleaved (GEOACT)	Yes	**90.20**	**87.80**	**89.0**

Table 3: **Mechanism ablations on LIBERO-Long.** Each row uses the same 500 evaluation episodes. Gap is the internal video–action agreement of Equation equation 2. EE error uses task-defined targets and does not use the visual reader.

Variant	Success (%) ↑	EE (cm) ↓	Gap (cm) ↓
Base	89.2	5.8	10.6
Base, matched field-evaluation count	90.0	5.6	10.2
Post-hoc correction	91.6	4.9	5.1
Interleaved, no readout refresh	92.4	4.6	4.8
Interleaved, joint-only objective	93.2	4.2	4.6
Interleaved, no update bound	90.8	5.2	**2.7**
Interleaved (GEOACT)	**95.6**	**3.5**	3.8

Base trails LingBot-VA, the best published average, by 4.0 points. GEOACT closes 3.0 of these points and moves the shared backbone above $\pi_{0.5}$. On RoboTwin 2.0, Base trails LingBot-VA by 10.4 points. GEOACT closes 7.2 of these points and moves the shared backbone above Motus.

4.3 WHICH PARTS OF GEOMETRIC FEEDBACK MATTER?

Table 3 removes one part of the feedback loop at a time on LIBERO-Long. LIBERO-Long shows the largest margin in Table 1. The matched field-evaluation baseline gives Base the same number of model evaluations as GEOACT. The no-refresh variant keeps the first admissible readout for all later interventions. The joint-only variant removes the end-effector term from Equation equation 3. The unbounded variant removes the proximal anchor and the displacement cap but keeps physical joint limits. All correction variants share the total correction budget.

Extra model evaluations do not explain the gain. Base with matched field evaluations reaches 90.0%. The extra evaluations add 0.8 points over Base. GEOACT adds 6.4 points over Base with the same number of evaluations.

Reading the updated video matters. The no-refresh variant reaches 92.4%, and full GEOACT reaches 95.6%. The no-refresh variant still denoises and reevaluates the model after each correction. The only missing part is a new geometric reading of the video after the video responds. This new reading is worth 3.2 points.

The end-effector term matters. The joint-only objective reaches 93.2% success and 4.2 cm EE error. Full GEOACT reaches 95.6% and 3.5 cm. The end-effector term weights joint errors by their effect on the gripper position.

The bound turns agreement into better motion. The unbounded variant reaches the smallest agreement gap, 2.7 cm. The unbounded variant also has the lowest success of all correction variants, 90.8%, and an EE error of 5.2 cm. The likely cause: the unbounded update copies readout errors into the action. The proximal anchor and the displacement cap keep the action near the model prediction. With the bound, GEOACT reaches 3.8 cm gap and the best EE error, 3.5 cm. Agreement gap is therefore a means of correction, not a target for evaluation.

4.4 FEEDBACK FREQUENCY AND COMPUTATIONAL COST

Figure 2a varies the number of late interventions m on LIBERO-Long. The total correction budget stays fixed and is split evenly across attempts. Every schedule ends before the final joint denoising step. Latency and memory are ratios to Base on the same device.

Two interventions give most of the gain. One intervention reaches 94.0% at $1.11\times$ Base latency. Two interventions reach 95.6% at $1.22\times$. Four interventions reach 95.8% at $1.44\times$. Going from two to four interventions adds 0.2 points and 0.22 of Base latency. We therefore use two interventions by default. With one intervention, Interleaved costs $1.11\times$ Base latency and Post-hoc costs $1.06\times$. At this similar cost, Interleaved leads Post-hoc by 2.4 points. Peak memory is $1.12\times$ Base for every correction variant.

4.5 READOUT AND FEEDBACK PATH ON REAL VIDEO

Three diagnostics on DROID and DreamZero test the parts of the loop in isolation.

The readout recovers arm configurations from real video. The reader is trained on 48,588 samples from 915 DROID (Khazatsky et al., 2024) episodes. Validation uses 850 samples from 30 disjoint episodes. The two-view, four-layer head reaches 10.55 cm median end-effector error at update 13,600. The training-mean joint predictor reaches 25.90 cm on the same samples. The final update 16,000 reaches 10.56 cm, so the result does not depend on the selected update. The validation set also serves for model selection. Two design choices matter at the same budget of 16,000 updates. A single view with four feature layers reaches 15.30 cm. Two views with only the final layer reach 14.25 cm. Two views with four layers reach 10.56 cm, and the 90th-percentile error falls from 30.82 cm to 20.16 cm. GEOACT therefore reads two views and four layers.

Action edits reach the generated video. One DreamZero example is sampled with a fixed seed and zero joint-state conditioning. Two opposite edits change the third normalized action coordinate during sampling. The edits change the final video latents by 6.53% and 5.91%. An unedited repeat changes the video latents by 0.00%. On the edited joint, the readout response has cosine similarity 0.841 with the action edit. The action-to-video path that GEOACT relies on therefore exists in a pretrained world action model.

A fitting residual is not a confidence score. A separate kinematic analyzer fits tracked image points and reports a fitting-residual proxy. At a 3 cm threshold, the proxy accepts 8 of 30 prospective episodes. Six accepted episodes meet the 3 cm level, and the worst accepted error is 5.09 cm. A small residual therefore does not certify a small error. This result motivates the design of Section 3.5: bounded updates with equal weights instead of a confidence gate.

5 CONCLUSION

GEOACT uses the video of a world action model as feedback for the action of the same model, before sampling ends. Robot kinematics places the depicted motion and the commanded motion in one frame. A bounded proximal step corrects the clean action. Same-noise reinjection and a new model evaluation let the video respond. On LIBERO and RoboTwin 2.0, this feedback inside generation beats an equally strong correction after generation. The gain is largest on long-horizon tasks and in randomized scenes. The ablations tie the gain to a fresh geometric reading of the updated video and to a bound on each update. Two interventions per action chunk cost $1.22\times$ the base latency.

AI USE STATEMENT

Generative AI tools were used only to assist with writing. Their use covered polishing the language of author-written text, improving readability, checking the manuscript for consistency and formatting errors, and formatting references. The authors reviewed and verified every suggested edit. Generative AI tools were not used for research ideation, formulating hypotheses, designing methods or experiments, implementing code, generating or processing data, running evaluations, analyzing or interpreting results, or creating figures. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work does not involve human subjects, crowdsourcing, or the collection of personal data, and we do not release new datasets. The joint readout is trained on the publicly released DROID dataset (Khazatsky et al., 2024), and all policy evaluations use the public LIBERO (Liu et al., 2023) and RoboTwin 2.0 (Chen et al., 2025) benchmarks. The pretrained world action model and the VGGT backbone (Wang et al., 2025) are used under their released licenses and stay frozen. For DROID, we use only robot camera frames and the recorded joint configurations as readout supervision. We make no attempt to identify people or locations that may appear in the scenes. All closed-loop results in this paper come from simulation, and no corrected action was executed on a physical robot.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify the full sampler, and Section 3.2 gives the readout architecture; the world action model and VGGT stay frozen, so only the readout is trained. Sections 4.1 and 4.5 describe the evaluation protocol and the DROID readout data, and all samplers share checkpoints, seeds, initial noise, and correction budget.

REFERENCES

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 51304–51323, 2024.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

Kevin Black, Noah Brown, Danny Driess, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Kevin Black, Manuel Y. Galliker, and Sergey Levine. Real-Time Execution of Action Chunking Flow Policies. *arXiv preprint arXiv:2506.07339*, 2025.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Conference on Robot Learning*, 2023.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, et al. GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, et al. RoboTwin 2.0: A Scalable Data Generator and Benchmark with Strong Domain Randomization for Robust Bimanual Robotic Manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Cheng Chi, 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.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, et al. Learning Universal Policies via Text-Guided Video Generation. In *Advances in Neural Information Processing Systems*, 2023.

Jisang Han, Seonghu Jeon, Jaewoo Jung, René Zurbrügg, Honggyu An, Tifanny Portela, Marco Hutter, Marc Pollefeys, Seungryong Kim, and Sunghwan Hong. Geometric action model for robot policy learning. *arXiv preprint arXiv:2606.17046*, 2026.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, et al. Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations. *arXiv preprint arXiv:2412.14803*, 2024.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with Diffusion for Flexible Behavior Synthesis. In *International Conference on Machine Learning*, 2022.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Single-view robot pose and joint angle estimation via render & compare. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Timothy E. Lee, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Oliver Kroemer, et al. Camera-to-Robot Pose Estimation from a Single Image. In *IEEE International Conference on Robotics and Automation*, 2020.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R. *arXiv preprint arXiv:2406.09756*, 2024.

Lin Li, Qihang Zhang, Yiming Luo, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a.

Peize Li, Ruimeng Zhang, Ru Zhang, Cong Huang, Kai Chen, and Shanghang Zhang. Fbfm: A training-free asynchronous feedback mechanism for flow-matching in world-action models execution. *arXiv preprint arXiv:2607.29235*, 2026b.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified Video Action Model. *arXiv preprint arXiv:2503.00200*, 2025.

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

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, et al. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *Advances in Neural Information Processing Systems*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2022.

Yuejiang Liu, Jubayer Ibn Hamid, Annie Xie, Yoonho Lee, Maximilian Du, and Chelsea Finn. Bidirectional Decoding: Improving Action Chunking via Guided Test-Time Sampling. *arXiv preprint arXiv:2408.17355*, 2024.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, et al. Octo: An Open-Source Generalist Robot Policy. In *Robotics: Science and Systems*, 2024.

Physical Intelligence et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, et al. Predictive Inverse Dynamics Models are Scalable Learners for Robotic Manipulation. *arXiv preprint arXiv:2412.15109*, 2024.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

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*, 2026.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Yanwei Wang, Lirui Wang, Yilun Du, Balakumar Sundaralingam, Xuning Yang, Yu-Wei Chao, et al. Inference-Time Policy Steering through Human Interactions. *arXiv preprint arXiv:2411.16627*, 2024b.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, et al. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In *International Conference on Learning Representations*, 2023.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, et al. Learning Interactive Real-World Simulators. In *International Conference on Learning Representations*, 2023.

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*, 2026.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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. URL `https://arxiv.org/abs/2603.16666`.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Robotics: Science and Systems*, 2023.