G2P-WAM: From Geometry Distillation to Preference Alignment for World-Action Models

Anonymous authors
Paper under double-blind review

Abstract

World-action models (WAMs) jointly predict future observations and robot actions, yet visually plausible predictions do not guarantee successful execution. Geometric distillation improves representation grounding, but using geometry for subsequent policy improvement requires relating geometric scores to task outcomes. We propose G2P-WAM, a post-training framework that carries geometric supervision from representation learning into preference alignment. GeoSFT first aligns WAM features with static 3D representations and dynamic 4D trajectories. Reward diagnosis then evaluates candidate geometric scores against execution outcomes and compares their behavior on imagined and observed videos. GeoDPO uses outcomes to order preferences and reconstruction confidence to select trajectories for reference-anchored optimization. Experiments on simulation benchmarks and two physical-robot platforms show that preference alignment further improves the geometrically distilled policy. Joint static and dynamic supervision outperforms either branch alone, while confidence filtering improves on outcome-only preference training with matched pair counts and optimization settings. These results support using geometry for both representation learning and preference selection, without adding geometric teachers or reward scorers at deployment. Project page: https://anonymous.4open.science/w/G2P-WAM-F7BF/.

1 Introduction

World-action models (WAMs) jointly predict future video and robot actions, learning control together with its visual consequences. Manipulation depends on scene structure and object motion, yet plausible predictions alone do not guarantee correct contact. Recent methods use static or dynamic geometry during supervised post-training (Li et al., 2026a;c; Zhang et al., 2026b; Yang et al., 2026; Zhang et al., 2026c; Guo et al., 2026) to improve representations and predictions. The next question is how geometry can guide learning from the policy's own successful and failed executions.

Outcome labels order preferences but do not distinguish geometric properties within each outcome group. Geometry can guide this selection, provided that the score carries information about execution. In our calibration, re-projection consistency is near chance at separating outcomes, whereas reconstruction confidence is more informative. This motivates evaluating scores on the policy's own rollouts before using them to select outcome-ordered preferences.

We present G2P-WAM, a three-stage framework shown in Figure 1. **GeoSFT** combines static-3D features with dynamic-4D trajectories to supervise scene structure and temporal changes. **Reward diagnosis** examines outcome separation, imagined–observed agreement, and motion-related statistics. **GeoDPO** then refines GeoSFT using high-confidence successes and low-confidence failures, with a frozen reference and label smoothing. Geometry thus guides both representation learning and the selection of subsequent training trajectories.

Experiments across simulation benchmarks and real-world manipulation tasks demonstrate the effectiveness of G2P-WAM. Joint static and dynamic supervision outperforms either branch alone, while preference alignment further improves task success. Ablation studies also show gains from

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.

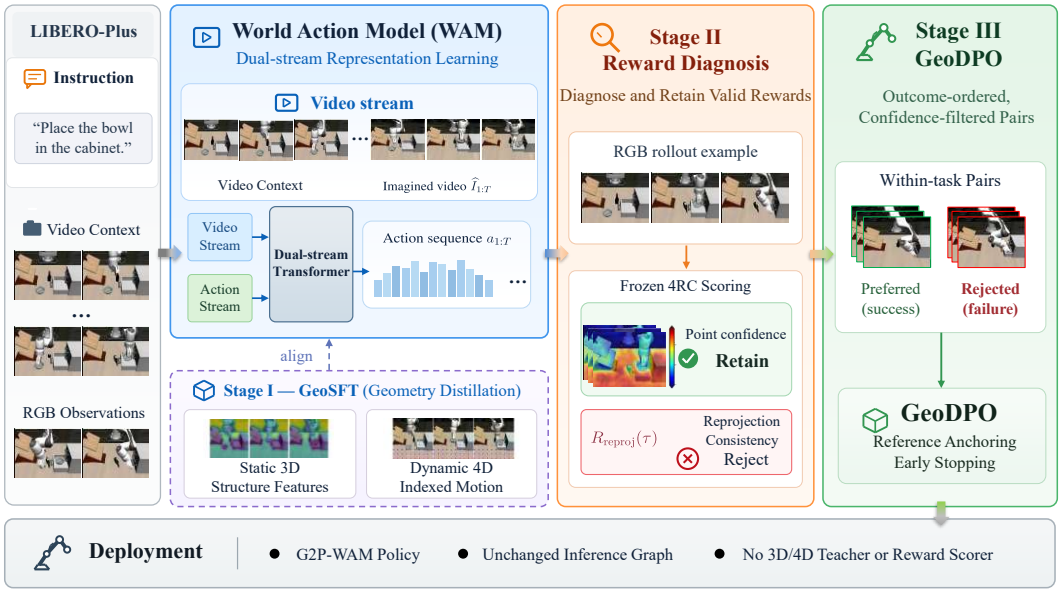


Figure 1: **G2P-WAM: geometry-to-preference post-training.** GeoSFT distills geometry; diagnosis selects a score; GeoDPO uses outcome-ordered, confidence-filtered pairs. LIBERO-Plus examples illustrate the pipeline: VGGT and SpatialTrackerV2 share an RGB clip; imagined video uses task-matched predictions. Preference pools are schematic. Only the WAM is deployed.

confidence-based pair selection. These results support using geometry for both representation learning and subsequent policy refinement.

Our contributions are threefold:

- We formulate **geometry-to-preference post-training**, connecting static-3D and dynamic-4D supervision to outcome-ordered, geometry-filtered policy alignment. GeoSFT supplies the initialization and frozen reference, with the original deployment graph preserved.
- We introduce **policy-relative reward diagnosis**, testing outcome separability and imagined–observed agreement alongside motion-related diagnostics. The analysis identifies point confidence as a useful calibration signal and characterizes how geometric scores vary across tasks and rollout conditions.
- We develop **reference-anchored GeoDPO** with outcome-ordered, confidence-filtered pairs. Simulation and physical-robot experiments show improved task success; the ablations in Section 4.6 establish gains from joint distillation and confidence-based pair selection under matched training settings.

2 Related Work

Geometric supervision for world-action models. WAMs combine future prediction with robot control, as in LingBot-VA (Li et al., 2026b) and Fast-WAM (Yuan et al., 2026). Geometric alignment also applies beyond WAMs: Spatial Forcing (Li et al., 2026a) aligns VLA representations, while Geometry Forcing (Wu et al., 2026) aligns video-diffusion features. For WAMs, WAM4D (Li et al., 2026c) and MECo-WAM (Zhang et al., 2026b) transfer geometric priors through auxiliary supervision. X-WAM (Guo et al., 2026) predicts RGB-D futures, while GeoSem-WAM (Ma et al., 2026) adds geometric and semantic prediction targets. For temporal structure, GEM-4D (Zhou et al., 2026) distills 4D correspondence into a video world model and derives actions through an inverse dynamics module; 4D-WAM (Yang et al., 2026) aligns trajectory-field representations within a WAM. These works use geometry as a representation or prediction target. We extend geometric supervision to subsequent preference training by evaluating geometric scores on the distilled policy's rollouts and using the selected score to filter training pairs.

Geometric rewards and policy preference optimization. DPO (Rafailov et al., 2023) and Diffusion-DPO (Wallace et al., 2024) provide objectives for learning from pairwise preferences. VideoGPA (Du et al., 2026) and epipolar preference alignment (Kupyn et al., 2025) apply geometric feedback to generated videos, while VGGRPO (An et al., 2026) and GeoFlow (Ackermann et al., 2026) pursue world consistency through geometric rewards. Stream4D (Ban et al., 2026) identifies a related failure mode: a rigid reconstruction critic can penalize genuine motion and favor static videos, motivating a dynamic 4D critic. These objectives improve visual futures; applying them to action policies also requires assessing their relation to execution outcomes. Policy post-training already extends beyond imitation: RIPT-VLA (Tan et al., 2025) learns from sparse execution rewards, NORA-1.5 (Hung et al., 2025) constructs action preferences using world-model rewards and demonstration-action deviations, and WMPO (Zhu et al., 2026) performs on-policy optimization through pixel-space imagined rollouts. G2P-WAM connects geometric distillation to preference alignment by diagnosing candidate scores on GeoSFT rollouts for outcome separability and imagined–observed agreement. Execution outcomes order preferences, while reconstruction confidence selects candidate trajectories.

3 METHOD

3.1 OVERVIEW AND PROBLEM FORMULATION

Given an instruction and observation context c, a WAM π_θ jointly predicts future video latents z^v and action latents z^a with a dual-stream flow-matching transformer (Li et al., 2026b; Lipman et al., 2023). For modality $q \in \{v, a\}$, let x_ξ^q be the interpolation between the clean latent x_0^q and noise ϵ^q, and let u_θ^q denote its velocity prediction. We use ξ for denoising time, t for video-frame indices, and τ for rollouts. The base supervised objective is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\big[\|u_\theta^v - (\epsilon^v - x_0^v)\|_2^2\big] + \lambda_a \mathbb{E}\big[\|u_\theta^a - (\epsilon^a - x_0^a)\|_2^2\big]\,. \tag{1}$$

Let $H_v^{(\ell)}$ and $H_a^{(\ell)}$ denote video- and action-stream hidden states at WAM layer ℓ. Geometry heads supervise video states through auxiliary readouts, preserving the original token sequences. G2P-WAM follows the three stages in Figure 2: two parameter-optimization stages, GeoSFT and GeoDPO, connected by policy-relative reward diagnosis. We optimize the GeoSFT and GeoDPO objectives sequentially.

3.2 STAGE I: GEOMETRY-GROUNDED SFT (GEOSFT)

GeoSFT adds static and dynamic supervision through training-only readout heads. The resulting checkpoint supplies calibration rollouts and initializes both the trainable policy and frozen reference in GeoDPO.

Static-3D representation alignment. A frozen VGGT teacher (Wang et al., 2025) is applied independently to each view and frame, yielding multi-level patch features $Y^{(j)}$. For selected student–teacher layer pairs $(i, j) \in \mathcal{P}$, a projector f_i maps WAM video states $H_v^{(i)}$ to the teacher dimension. Direction is the primary structural target and a lightweight scale head g_i regularizes feature magnitude, following the direction–scale decomposition of Geometry Forcing (Wu et al., 2026):

$$\mathcal{L}_{\mathrm{S3D}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \Big[\lambda_{\mathrm{ang}}\big(1 - \cos(f_i(H_v^{(i)}), \mathrm{sg}(\widetilde{Y}^{(j)}))\big) + \lambda_{\mathrm{scale}} \|g_i(\mathrm{norm}\, f_i(H_v^{(i)})) - \mathrm{sg}(\|\widetilde{Y}^{(j)}\|_2)\|_2^2\Big], \tag{2}$$

where $\widetilde{Y}^{(j)}$ is resized to the student patch layout and sg stops teacher gradients. The per-frame terms receive the noise weighting specified in Appendix A.1. Processing views independently avoids cross-view leakage while preserving the WAM's original multi-camera interface.

Dynamic-4D trajectory alignment. Dynamic supervision complements static features by capturing the motion of contacted objects and robot parts. Frozen SpatialTrackerV2 (Xiao et al., 2025)

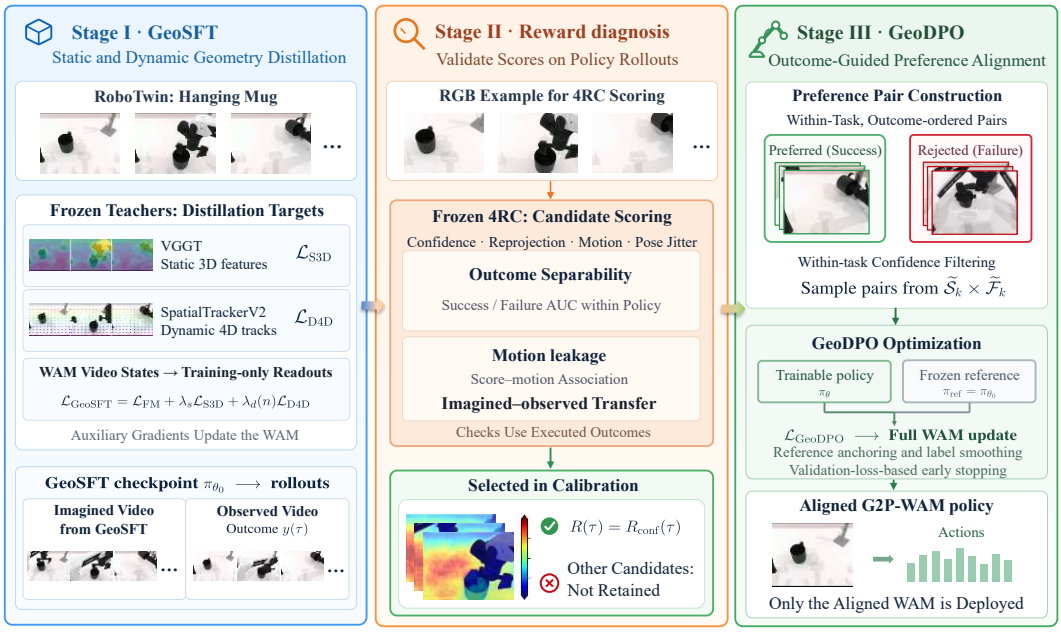


Figure 2: **G2P-WAM training framework with RoboTwin examples.** I: geometry distillation; II: frozen 4RC scoring and diagnostic criteria; III: outcome-ordered, confidence-filtered pairs and updates against a frozen GeoSFT reference.

extracts indexed trajectories $p_{k,t}^{c_t}$, camera poses, visibility $v_{k,t}$, and confidence $\alpha_{k,t}$ from a short RGB clip. We express every point in the first-camera coordinate system and remove ego motion:

$$x_{k,t}^{c_0} = \pi_3\left(T_{c_0 \leftarrow w} T_{w \leftarrow c_t} \bar{p}_{k,t}^{c_t}\right), \qquad \Delta_{k,t} = \frac{x_{k,t+1}^{c_0} - x_{k,t}^{c_0}}{s_{\text{clip}} + \varepsilon}, \tag{3}$$

where s_{clip} is the median valid displacement norm. A register indexed by the query location and transition time reads selected video layers and predicts $\widehat{\Delta}_{k,t}$. With $m_{k,t}$ masking invisible or low-confidence tracks,

$$\mathcal{L}_{\text{D4D}} = \frac{\sum_{k,t} m_{k,t}\, \rho_\delta(\widehat{\Delta}_{k,t} - \text{sg}(\Delta_{k,t}))}{\sum_{k,t} m_{k,t} + \varepsilon}, \tag{4}$$

where ρ_δ is coordinate-wise Smooth-ℓ_1. This objective supervises clip-normalized 4D motion in the first-camera coordinate system. Here $\bar{p} = (p_x, p_y, p_z, 1)^\top$ is homogeneous and π_3 drops its final coordinate. Figure 6 visualizes both teacher targets on a shared clip. VGGT features are displayed using PCA; training aligns the full feature vectors.

The complete supervised objective is

$$\mathcal{L}_{\text{GeoSFT}} = \mathcal{L}_{\text{FM}} + \lambda_s \mathcal{L}_{\text{S3D}} + \lambda_d(n)\mathcal{L}_{\text{D4D}}, \tag{5}$$

with warm-up schedule $\lambda_d(n)$. Auxiliary gradients update the shared WAM trunk, but both branches only read video states. After training, teachers, projectors, and the register head are removed. The GeoSFT policy π_{θ_0} retains the original inference graph.

3.3 Stage II: Policy-Relative Geometric Reward Diagnosis

Candidate geometric scores. Let $\tau = (\widehat{I}_{1:T}, a_{1:T}) \sim \pi_{\theta_0}(\cdot \mid c)$ denote an imagined video and its paired action rollout from the GeoSFT checkpoint. A frozen 4D reconstructor Φ maps $\widehat{I}_{1:T}$ to point maps P_t, camera poses, depth, and point confidences $c_{t,p}$. Executing the paired actions gives real video and a binary outcome $y(\tau) \in \{0, 1\}$, with success encoded as one. The same Φ scores

imagined and real video for the transfer test. Its outputs induce candidate rewards:

$$R_{\text{reproj}}(\tau) = -\frac{1}{T}\sum_{t}\left[\,\text{MSE}(\widehat{I}_t, \widetilde{I}_t) + \lambda\,\text{LPIPS}(\widehat{I}_t, \widetilde{I}_t)\right], \tag{6}$$

$$R_{\text{conf}}(\tau) = \bar{c}(\tau) = \frac{1}{T}\sum_{t}\frac{1}{|P_t|}\sum_{p\in P_t} c_{t,p}. \tag{7}$$

We also consider motion and pose-jitter statistics. Here $\widetilde{I}_t$ is a point-cloud re-projection, and LPIPS measures perceptual distance (Zhang et al., 2018). For multi-camera WAM output, we score the teacher-aligned head-camera crop rather than the full mosaic.

We diagnose scores before using them for preference selection. On calibration rollouts sampled from the GeoSFT policy, we examine **outcome separability** using success/failure AUC and bootstrap uncertainty, and **imagined–observed agreement** by scoring both videos from the same episode. **Motion-related diagnostics** characterize how geometric scores relate to estimated motion and pose statistics. Calibration uses rollouts from the same policy and environment, aligning score selection with the rollout distribution of the policy to be refined.

As shown in Section 4.2, re-projection consistency provides near-chance outcome discrimination, whereas point confidence gives stronger separation and similar imagined/observed AUCs. We therefore select $R(\tau) = R_{\text{conf}}(\tau) = \bar{c}(\tau)$, averaging 4RC point confidence over at most eight uniformly sampled frames and their spatial locations. Reconstruction confidence provides a geometric signal for trajectory selection; execution outcomes supply the preference labels. The score selected in this calibration is then applied across tasks with task-specific quantile thresholds. Separate cross-task diagnostics examine how this association varies across tasks and evaluation conditions.

3.4 Stage III: Preference Alignment with GeoDPO

Outcome-stratified confidence filtering. For each task k, stochastic decoding from GeoSFT produces rollout records $(\tau, c, y(\tau), R(\tau))$, where execution supplies the binary outcome and $R(\tau)$ is computed from the WAM-predicted video. We first split rollouts by the executed outcome,

$$\mathcal{S}_k = \{\tau : y(\tau) = 1\}, \qquad \mathcal{F}_k = \{\tau : y(\tau) = 0\}. \tag{8}$$

Within each task, successful rollouts are sorted by confidence from high to low and the top 50% are retained, while failed rollouts are sorted from low to high and the bottom 50% are retained. Denoting these filtered pools by $\widetilde{\mathcal{S}}_k$ and $\widetilde{\mathcal{F}}_k$, GeoDPO randomly samples winner–loser pairs from $\widetilde{\mathcal{S}}_k \times \widetilde{\mathcal{F}}_k$; trajectory reuse across sampled pairs is allowed. Outcomes therefore establish the preference order, while confidence selects which trajectories enter the candidate pools. The score and filtering direction are fixed across tasks, while within-task, outcome-stratified quantiles adapt the selection thresholds. Cross-task AUCs characterize score behavior under this fixed direction.

Reference-anchored diffusion DPO. Following DPO and its diffusion-model formulation (Rafailov et al., 2023; Wallace et al., 2024), initialize the trainable policy from π_{θ_0} and freeze a copy $\pi_{\text{ref}} = \pi_{\theta_0}$. Let $v_\theta = u^a_\theta$ and v_{ref} be their action-stream velocity predictors. For clean action latent x, noised version x_ξ, and flow-matching target $d^a = \epsilon^a - x$, define the reference-relative score

$$s_\theta(x, \xi, c) = -w_\xi\Big(\|d^a - v_\theta(x_\xi, \xi, c)\|_2^2 - \|d^a - v_{\text{ref}}(x_\xi, \xi, c)\|_2^2\Big), \tag{9}$$

with denoising weight $w_\xi \geq 0$, using the parameterization of Eq. 1. For a sampled winner–loser pair with its own conditioning contexts c_w, c_l, let $z_\theta = \beta[s_\theta(x_w, \xi, c_w) - s_\theta(x_l, \xi, c_l)]$. With smoothing $\eta = 0.1$, GeoDPO minimizes

$$\mathcal{L}_{\text{GeoDPO}} = -\mathbb{E}\left[(1-\eta)\log\sigma(z_\theta) + \eta\log\sigma(-z_\theta)\right] + \lambda_{\text{ref}}\mathcal{L}_{\text{ref}}, \tag{10}$$

where σ is the logistic sigmoid and $\mathcal{L}_{\text{ref}} = \mathbb{E}\|v_\theta - v_{\text{ref}}\|_2^2$ anchors predictions on the same noised inputs. The denoising-error difference provides the diffusion-DPO surrogate score, and smoothing regularizes the preference margin. Each trajectory is evaluated under its own conditioning context, yielding a within-task, cross-context ranking surrogate when c_w and c_l differ.

Optimization updates the full WAM while freezing the reference. Anchoring and smoothing regularize the margin. RoboTwin variants B1/B2 share the maximum training budget and validation-loss stopping rule described in Section 4.1, with stopping steps determined separately for each run. Deployment uses only the aligned WAM, without the reference, reconstructor, or auxiliaries.

4 EXPERIMENTS

The experiments address three questions: whether geometric scores provide useful feedback, whether preference alignment improves control, and how each training component contributes to the final policy.

4.1 SETUP

Benchmarks. We evaluate closed-loop success on RoboTwin (Chen et al., 2025), long-horizon LIBERO (Liu et al., 2023), seven-factor LIBERO-Plus (Fei et al., 2025), and video quality on Physics-IQ (Motamed et al., 2026) and RBench (Deng et al., 2026). Physical-robot evaluation is reported in Section 4.4; simulation "real" video refers to post-execution simulator observations.

GeoSFT training. Experiments use 32 NVIDIA A100 GPUs. We initialize from LingBot-VA (Li et al., 2026b) and jointly train the full WAM and auxiliary heads with the flow-matching and geometry objectives in Section 3.2, keeping VGGT and SpatialTrackerV2 frozen. AdamW uses learning rates of 10^{-5} for the trunk and 10^{-4} for auxiliary heads, with an effective batch size of 64. The learning-rate warm-up lasts 100 updates; the dynamic loss weight increases linearly to 0.1 over 500 updates. RoboTwin training jointly covers all 50 tasks using mixed Clean/Randomized demonstrations. Appendix A.1 gives the training configuration; Figure 6 visualizes the teacher targets.

GeoDPO training and evaluation. GeoSFT initializes both the policy and its frozen reference. Frozen 4RC (Luo et al., 2026) supplies the reconstruction-confidence score selected by reward diagnosis. GeoDPO updates the full WAM with $\beta = 25$, label smoothing of 0.1, learning rate 10^{-4}, and reference weight $\lambda_{\text{ref}} = 0.2$. We collect preferences with SDE decoding and evaluate with deterministic decoding. B1 and B2 use at most 2,000 GeoDPO updates under the same validation-loss early-stopping rule. Main and ablation evaluations use 100 trials per task and setting, evaluating the same checkpoint on Clean and Randomized. Diagnostic studies use separate rollout sets. Appendix A.13 details the comparison protocol.

4.2 GEOMETRIC REWARD DIAGNOSIS ACROSS BENCHMARKS

We assess whether geometry provides a useful criterion for selecting preference trajectories. Cross-task diagnostics on RoboTwin and LIBERO-Plus examine how outcome discrimination varies across tasks and perturbations. Calibration on GeoSFT then compares candidate scores on imagined videos and their observed executions.

Cross-task and cross-benchmark behavior. Metric rankings and raw score directions vary across six eligible RoboTwin tasks and seven LIBERO-Plus perturbation families in the official-backbone diagnostics, detailed in Appendices A.9 and A.11. With official-policy successes fixed, replacing official-policy/randomized failures with weak-base/clean failures increases confidence AUC on five of six tasks, as shown in Figure 3b. This sensitivity to the joint policy and environment shift motivates diagnosis on rollouts from GeoSFT, the policy that supplies preference-training trajectories.

Calibration on GeoSFT rollouts. We compare candidate 4RC scores on paired imagined and observed videos from GeoSFT rollouts, as shown in Figure 3a and Table 13. Mean point confidence achieves similar outcome discrimination on imagined and observed videos, with AUCs of 0.744 and 0.752, respectively, whereas re-projection consistency remains near chance. These results support using confidence computed from WAM predictions to select training trajectories within outcome groups. Motion-related analysis also shows that confidence filtering changes the pose-motion distribution of selected rollouts, as detailed in Appendix A.13. The effect of this selection on policy performance is evaluated through the controlled ablations in Section 4.6.

4.3 SUCCESS RATES ACROSS THREE SIMULATION BENCHMARKS

Tables 2a–2b and Table 1 compare G2P-WAM with existing methods; baseline scores are taken from their original papers. On RoboTwin-50, GeoDPO raises mean success from 92.46% to 92.95%, with gains of 0.42 and 0.56 percentage points on Clean and Randomized, respectively. Improving

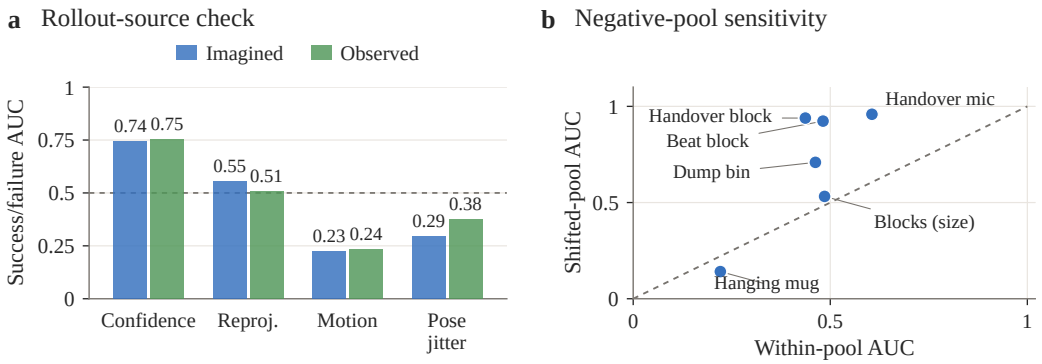


Figure 3: **Geometry-reward diagnosis. (a)** GeoSFT calibration on 122 Turn Switch episodes; the dashed line indicates chance. **(b)** Official-policy successes paired with within-pool or shifted-pool negatives; the diagonal indicates equal AUC. The shifted pool changes both model and setting and contains eight negatives per task. Raw score directions are retained.

in both settings shows that preference alignment remains beneficial after geometric distillation, including under environmental randomization. G2P-WAM also achieves 98.58% average success on LIBERO, reaching 99.0% on Long, and 79.5% overall on LIBERO-Plus. Performance on LIBERO-Long and LIBERO-Plus further supports the framework's applicability to extended action sequences and diverse evaluation conditions.

Table 1: **Success rates (%) on LIBERO-Plus.** Appendix A.6 provides the GeoSFT comparison. Bold marks the best listed score per axis. Overall uses each source's aggregation; our score equally weights the seven factors.

Method	Camera	Robot	Lang.	Light	BG	Noise	Layout	Overall
π_0 (Black et al., 2025)	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
π_0-FAST (Pertsch et al., 2025)	65.1	21.6	61.0	73.2	73.2	74.4	68.8	61.6
UniVLA (Bu et al., 2025)	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
RIPT-VLA (OFT) (Tan et al., 2025)	55.2	31.2	77.6	88.4	91.6	73.5	74.2	68.4
OpenVLA-OFT (Kim et al., 2025)	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
X-VLA (Zheng et al., 2026)	23.4	**89.7**	75.7	88.2	**96.0**	62.7	71.8	71.4
Fast-WAM (Yuan et al., 2026)	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
Fast-WAM-Joint (Yuan et al., 2026)	34.0	55.1	88.9	90.0	44.9	33.3	73.4	59.0
LingBot-VA (Li et al., 2026b)	52.0	87.5	68.3	90.0	84.1	76.0	71.2	75.6
4D-WAM (Yang et al., 2026)	45.15	64.26	90.63	**94.29**	57.71	69.08	79.21	71.01
GaussianWAM (Cosmos) (Zhang et al., 2026c)	**79.11**	56.52	**92.18**	89.84	66.17	86.26	70.55	77.30
G2P-WAM (ours)	33.1	82.8	92.0	90.0	76.1	**86.4**	**95.8**	79.5

Table 2: **Full-benchmark success (%) on LIBERO and RoboTwin. (a)** Standard LIBERO suites. **(b)** RoboTwin Clean/Randomized; Emb. PT. denotes embodied-policy pretraining. Bold marks the best comparable score.

(a) LIBERO

Method	Spatial	Object	Goal	Long	Avg.
π_0 (Black et al., 2025)	96.8	98.8	95.8	85.2	94.15
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	**98.8**	98.2	98.0	92.4	96.85
OpenVLA-OFT (Kim et al., 2025)	97.6	98.4	97.9	94.5	97.10
Motus (Bi et al., 2026)	96.8	99.8	96.6	97.6	97.70
LingBot-VA (Li et al., 2026b)	98.5	99.6	97.2	98.5	98.5
Fast-WAM (Yuan et al., 2026)	98.2	**100.0**	97.0	95.2	97.60
MECo-WAM (Zhang et al., 2026b)	**98.8**	**100.0**	98.2	95.8	98.20
4D-WAM (Yang et al., 2026)	**98.8**	99.5	**98.6**	97.4	98.6
G2P-WAM (ours)	98.5	99.6	97.2	**99.0**	98.58

(b) RoboTwin

Method	Emb. PT.	Clean	Random	Avg.
π_0 (Black et al., 2025)	Yes	65.92	58.40	62.16
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	Yes	82.74	76.76	79.75
Motus (Bi et al., 2026)	Yes	88.66	87.02	87.84
LingBot-VA (Li et al., 2026b)	Yes	92.90	91.50	92.20
Fast-WAM (Yuan et al., 2026)	No	91.88	91.78	91.83
WAM4D (Li et al., 2026c)	–	**93.82**	89.86	91.84
MECo-WAM (Zhang et al., 2026b)	No	93.26	91.98	92.62
4D-WAM (Yang et al., 2026)	No	91.92	90.76	91.34
G2P-WAM (ours)	Yes	93.40	**92.50**	**92.95**

Task-level changes. Relative to GeoSFT, G2P-WAM improves on 24 Clean tasks, ties on 13, and regresses on 13. The corresponding counts for Randomized are 26, 10, and 14. A paired bootstrap over the 50 task identities gives 95% intervals for the GeoSFT-to-G2P-WAM gains of

$[0.02, 0.82]$ percentage points on Clean, $[0.16, 0.96]$ on Randomized, and $[0.18, 0.81]$ for the equal-setting mean. More tasks improve than regress in both settings, indicating that the aggregate gain is accompanied by improvements across the task suite. Section 4.6 distinguishes this full-stage gain from the contribution of confidence-based selection.

Qualitative rollouts. Figure 4 compares selected GeoSFT and G2P-WAM rollouts on three benchmarks. Appendix A.5 provides extended sequences and task descriptions.

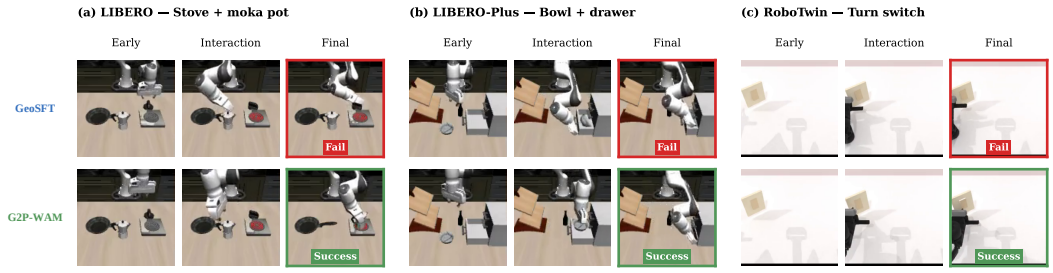


Figure 4: **Closed-loop rollout examples.** GeoSFT appears above G2P-WAM on three benchmarks. Appendix A.5 provides extended sequences and task descriptions.

4.4 Real-world robot experiments

We evaluate an AgileX PiPER X dual-arm robot on three tasks: storing fruit in a drawer, placing a plate on the cabinet, and placing a banana on the plate. We compare $\pi_{0.5}$ (Physical Intelligence et al., 2025), LingBot-VA (Li et al., 2026b), GeoSFT, and G2P-WAM under the same observation, instruction, action, reset, maximum-horizon, and final-state scoring protocol, with 30 trials per task and method. Figure 5 shows executions and success rates. G2P-WAM reaches 80.0%, 86.7%, and 90.0%, respectively. GeoDPO raises GeoSFT's mean success from 77.8% to 85.6%, a gain of 7.8 percentage points. LingBot-VA and $\pi_{0.5}$ achieve 73.3% and 60.0%. Plate placement shows the largest gain over GeoSFT, while improvements on all three tasks extend the benefit of preference training to physical manipulation. Appendix A.4 gives the success criteria and complete trial counts.

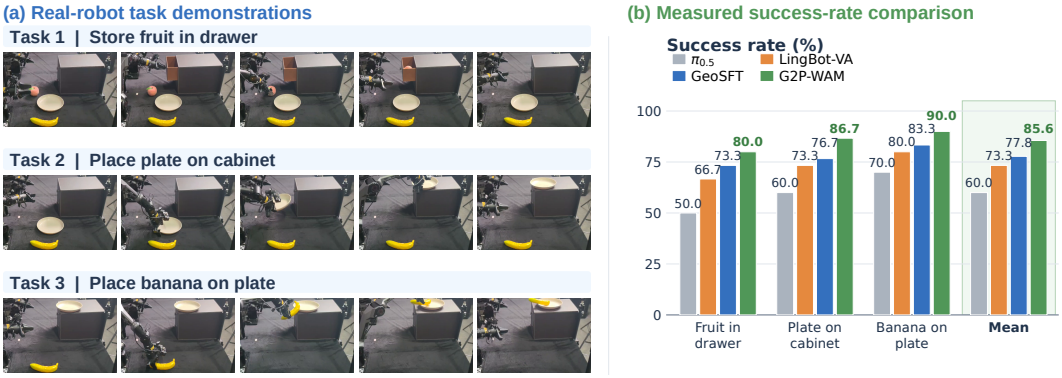


Figure 5: **Physical-robot evaluation on the AgileX PiPER X dual-arm platform.** **(a)** Representative keyframes for three tasks. **(b)** Four-method comparison over 30 trials per task; GeoSFT is the checkpoint before GeoDPO. Mean equally weights the tasks.

4.5 Reward controls and world-model evaluation

Reward controls. Exploratory runs in Table 14 motivate the combination of outcome-based diagnosis and reference anchoring in GeoDPO. Re-projection and unregularized cycle-consistency rewards reduce task success; D4RT feature rewards (Zhang et al., 2026a), evaluated with the available community implementation, show a similar decline. Reference-regularized cycle consistency reduces

predicted-rollout error by 3.3% while preserving success, and point confidence supports a success-rate gain in its reported run. These observations guide the choice of outcome-ordered preferences and confidence-based selection.

World-model quality and transfer. Compared with GeoSFT, G2P-WAM improves PSNR and SSIM and reduces LPIPS on held-out RoboTwin continuations, as shown in Table 11. Its official Physics-IQ (Motamed et al., 2026) score also rises from 28.5 to 29.6. On RBench (Deng et al., 2026), independent Qwen2.5-VL (Bai et al., 2025) and Q-Align (Wu et al., 2024) evaluation favors G2P-WAM on all four morphology composites and all five task-oriented tracks. Instruction-adherence scores remain close, with small decreases across the four morphologies. The gains in reconstruction and perceptual quality indicate that preference alignment can benefit video prediction as well as control. Appendix A.7 reports the full evaluation.

4.6 Ablation study on RoboTwin 2.0

Table 3: **RoboTwin 2.0 ablations (%).** S/D: static/dynamic distillation; Avg. averages Clean/Randomized. A0: official checkpoint; †: scores reused from the main evaluation.

ID	Configuration	S	D	Pair selection	Clean	Random.	Avg.
A. Official-checkpoint reference and geometry distillation (A1–A3)							
A0	LingBot-VA (official weights)	–	–	None	92.90	91.50	92.20
A1	Static-only SFT	✓	–	None	92.86	91.70	92.28
A2	Dynamic-only SFT	–	✓	None	92.96	91.84	92.40
A3	**GeoSFT**†	✓	✓	None	92.98	91.94	92.46
B. Preference alignment from GeoSFT (B1/B2: matched pair count and training rule)							
B0	+ Outcome-only DPO	✓	✓	Full outcome pool	93.08	92.26	92.67
B1	+ Matched outcome-only DPO	✓	✓	Random, count-matched	93.12	92.16	92.64
B2	+ **GeoDPO (G2P-WAM)**†	✓	✓	Confidence-filtered	93.40	92.50	92.95

Static and dynamic geometry. A1–A3 each train a joint 50-task policy on the same mixed Clean/Randomized demonstrations, providing a matched comparison of the geometric objectives. A0 provides the official-checkpoint reference before additional SFT. Joint GeoSFT achieves 92.46% average success, exceeding static-only and dynamic-only supervision by 0.18 and 0.06 percentage points, respectively. The advantage of joint supervision supports combining scene-structure targets with temporal motion supervision.

Confidence-based pair selection. B1 and B2 share the GeoSFT initialization and frozen reference, preference-pair count, maximum training budget, and validation-loss early-stopping rule. B1 samples unfiltered, outcome-ordered pairs; B2 applies the confidence filter in Section 3.4. GeoDPO exceeds B1 by 0.28 percentage points on Clean and 0.34 on Randomized, yielding 0.31 points overall; it exceeds B0 by 0.28 points overall. With pair counts matched, the B2–B1 gain supports the contribution of confidence-based trajectory selection beyond random subsampling. The 0.49-point gain over GeoSFT measures the benefit of the complete preference-training stage. Appendix A.13 details the training controls, pair construction, and evaluation protocol.

5 Conclusion

In this work, we proposed G2P-WAM to extend geometric supervision from WAM representation learning to preference alignment. GeoSFT combines static 3D features with dynamic 4D trajectories, while reward diagnosis identifies confidence as a candidate-selection signal. GeoDPO then refines the distilled policy through outcome-ordered, confidence-filtered preferences. Simulation and physical-robot experiments show that this additional stage improves task success. Ablations support both geometric targets and confidence-based selection beyond outcome-only training. Together, these results show that geometric knowledge can guide both what a WAM learns from demonstrations and how it improves from its own rollouts. By unifying geometric grounding and policy refinement without adding inference-time modules, G2P-WAM offers a practical approach to translating geometric priors into more effective embodied control.

Ethics Statement

This work studies world-action models through simulated manipulation, physical-robot experiments, and generated-video evaluation. The experiments focus on robotic object manipulation and do not involve human-subject studies. Deployment beyond the evaluated settings requires task-specific assessment of execution reliability and interaction safety.

Reproducibility Statement

Appendix A.1 and A.12 specify distillation and preference construction. RoboTwin uses a joint 50-task policy, mixed Clean/Randomized SFT data, and full-WAM updates. B0–B2 share a frozen GeoSFT reference; B1/B2 use the same maximum optimization budget and validation-loss early-stopping rule. Confidence is computed by uniformly sampling at most eight frames from each WAM-predicted rollout and averaging 4RC point confidence. Appendix A.13 separates the official-weight A0 reference from trained variants and documents the task-level confidence filter. Checkpoint identifiers, teacher/scoring scripts, configurations, evaluation harnesses, and trial manifests are part of the planned release. Project page: https://anonymous.4open.science/w/G2P-WAM-F7BF/.

AI Use Statement

Generative AI tools assisted with language polishing and LaTeX formatting to improve the manuscript's clarity and presentation. The authors reviewed all AI-assisted revisions and take responsibility for the final manuscript.

References

Jan Ackermann, Shengqu Cai, Boyang Deng, Zhengfei Kuang, Songyou Peng, and Gordon Wetzstein. GeoFlow: Enforcing Implicit Geometric Consistency in Video Generation. *arXiv preprint arXiv:2605.18365*, 2026.

Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. Vggrpo: Towards world-consistent video generation with 4d latent reward. In *European Conference on Computer Vision*, pp. 305–322. Springer, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, Junyang Lin, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

Yuanhao Ban, Jiaqi Feng, Hengguang Zhou, Xiaohuan Pei, Justin Cui, and Cho-Jui Hsieh. Stream4D: 4D-Consistency for Streaming Autoregressive Diffusion Video Models. *arXiv preprint arXiv:2608.19556*, 2026.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A Unified Latent Action World Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Robotics: Science and Systems*, 2025.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to Act Anywhere with Task-centric Latent Actions. In *Robotics: Science and Systems*, 2025.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, 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.

Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruoqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking Video Generation Model for the Embodied World. *Forty-third International Conference on Machine Learning*, 2026.

Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. VideoGPA: Distilling Geometry Priors for 3D-Consistent Video Generation. In *International Conference on Machine Learning*, 2026.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth Robustness Analysis of Vision-Language-Action Models. *arXiv preprint arXiv:2510.13626*, 2025.

Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4D World Action Modeling from Video Priors with Asynchronous Denoising. *arXiv preprint arXiv:2604.26694*, 2026.

Chia-Yu Hung, Navonil Majumder, Haoyuan Deng, Renhang Liu, Yankang Ang, Amir Zadeh, Chuan Li, Dorien Herremans, Ziwei Wang, and Soujanya Poria. NORA-1.5: A Vision-Language-Action Model Trained Using World Model- and Action-Based Preference Rewards. *arXiv preprint arXiv:2511.14659*, 2025.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. In *Robotics: Science and Systems*, 2025.

Orest Kupyn, Théo Uscidda, Marta Tintore Gazulla, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar Geometry Improves Video Generation Models. *arXiv preprint arXiv:2510.21615*, 2025.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial Forcing: Implicit Spatial Representation Alignment for Vision-Language-Action Model. In *International Conference on Learning Representations*, volume 2026, pp. 132324–132345, 2026a.

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. *Robotics: Science and Systems*, 2026b.

Ying Li, Xiaobao Wei, Jiajun Cao, Hao Wang, Xiaowei Chi, Chengyu Bai, Qianpu Sun, Jiajun Li, Xiaojie Zhang, Peidong Jia, Jian Tang, Sirui Han, and Shanghang Zhang. WAM4D: Fast 4D World Action Model via Spatial Register Tokens. *arXiv preprint arXiv:2606.14048*, 2026c.

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*, 2023.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791, 2023.

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4RC: 4D Reconstruction via Conditional Querying Anytime and Anywhere. *arXiv preprint arXiv:2602.10094*, 2026.

Fulong Ma, Daojie Peng, Wenjun Yue, Jiahang Cao, Bintao Wang, Qiang Zhang, and Jun Ma. GeoSem-WAM: Geometry- and Semantic-Aware World Action Models. *arXiv preprint arXiv:2606.03188*, 2026.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient Action Tokenization for Vision-Language-Action Models. In *Robotics: Science and Systems*, 2025.

Physical Intelligence, 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. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive Post-Training for Vision-Language-Action Models. *arXiv preprint arXiv:2505.17016*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion Model Alignment Using Direct Preference Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 54015–54029. PMLR, 2024.

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.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: Advancing 3D Point Tracking with Explicit Camera Motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6726–6737, 2025.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, Qiao Sun, Pengwei Wang, Lingqiao Liu, Yan Wang, Yuxiang Gao, Feras Dayoub, and Haoang Li. 4D-WAM: Infusing Spatiotemporal Awareness into World Action Models through Trajectory Fields. *arXiv preprint arXiv:2608.08023*, 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.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently Reconstructing Dynamic Scenes One D4RT at a Time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7382–7392, 2026a.

Jianjun Zhang, Jian Zhu, Taiyi Su, Chong Ma, Zitai Huang, Yi Xu, and Hanli Wang. Learning 4D Geometric Priors for Inference-Efficient World Action Models. *arXiv preprint arXiv:2607.05468*, 2026b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zijian Zhang, Yuqing Jiang, Weitao Zhou, Minglei Li, Jinhao Zhang, Yao Mu, Xiaofan Li, Hao Zhao, and Haibao Yu. GaussianWAM: Distilling Geometry and Semantics from 3D Gaussian Fields into World-Action Models. *arXiv preprint arXiv:2608.24714*, 2026c.

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 Liu, Tai Wang, and Xianyuan Zhan. X-VLA: Soft-Prompted Transformer as Scalable Cross-Embodiment Vision-Language-Action Model. In *International Conference on Learning Representations*, volume 2026, pp. 60580–60606, 2026.

Kaichen Zhou, Yuzhen Chen, Fangneng Zhan, Hang Hua, Grace Chen, Xinhai Chang, Ao Qu, Yilun Du, Zhuang Liu, Paul Pu Liang, and Mengyu Wang. GEM-4D: Geometry-Enhanced Video World Models for Robot Manipulation. *arXiv preprint arXiv:2605.22882*, 2026.

Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. WMPO: World Model-Based Policy Optimization for Vision-Language-Action Models. In *International Conference on Learning Representations*, volume 2026, pp. 62486–62502, 2026.

A Additional Experiments and Implementation Details

A.1 GeoSFT: static-3D and dynamic-4D distillation

GeoSFT grounds WAM representations in static scene structure and dynamic correspondence. This section details the complementary distillation objectives and the training configuration in Table 4. The resulting geometry-grounded checkpoint supports policy-relative reward diagnosis and initializes both the trainable policy and frozen reference in GeoDPO. This design connects representation learning with subsequent preference alignment while preserving the WAM's deployment interface.

Table 4: **GeoSFT training configuration.** Negative teacher indices count backward from the final layer; Smooth-ℓ_1 δ is distinct from the GeoDPO temperature β.

Component	Setting
WAM trunk	30-layer dual-stream Wan-DiT; original video/action interface
Static teacher	Frozen VGGT; clean RGB, independently per view and frame
Static read points	Student $\{10, 20, 29\}$ paired with teacher $\{-9, -5, -1\}$
Static projector and loss	Layer-specific 2048-D MLP; $\lambda_{\text{ang}} = 0.5$, $\lambda_{\text{scale}} = 0.05$, $\lambda_s = 1$
Static noise weighting	$\omega(\xi) = \max(1 - \sigma_{\text{noise}}(\xi), 0.05)$ per video frame
Dynamic teacher and input	Frozen SpatialTrackerV2; $T = 8$ clean head-camera frames
Dynamic queries	8×8 frame-0 grid; $N = 64$ indexed trajectories
Dynamic read points	Video layers $\{12, 16, 18\}$; no action-token readout
Register head	768-D; 12 heads; four self-/cross-attention blocks; zero-initialized output projection
Motion target and loss	Frame-0-camera displacement, clip-scale normalized; masked Smooth-ℓ_1, $\delta = 0.01$
Dynamic loss schedule	$\lambda_d(n) = 0.1 \min(1, n/500)$, with optimizer update n
Optimization	AdamW; full WAM with trainable auxiliary heads; trunk/auxiliary learning rates $10^{-5}/10^{-4}$; main warmup 100 updates
AdamW coefficients	$\beta_1 = 0.9$, $\beta_2 = 0.95$; weight decay 0.1
Effective batch size	64 training examples per optimizer update
Deployment	Remove both teachers, static projectors, and dynamic register head

Static targets for per-frame structure. VGGT (Wang et al., 2025) extracts multi-level patch features from the clean demonstration frame. Each camera view is processed separately; its feature grid is resized to the corresponding student footprint before restoring the WAM layout. For RoboTwin's head-plus-two-wrist input, this reconstructs the original T-shaped mosaic. Processing views and frames independently keeps this branch focused on per-frame structure, complementing the temporal supervision supplied by the dynamic branch. Equation 2 follows the direction–scale decomposition of Geometry Forcing (Wu et al., 2026) and averages the angular and magnitude terms over matched patches and layer pairs. In our GeoSFT configuration, both per-frame terms also receive $\omega(\xi)$, where $\sigma_{\text{noise}}(\xi)$ is the frame's video-noise level; very noisy student frames are downweighted but not discarded.

Dynamic targets: one reconstruction frame, one scale per clip. SpatialTrackerV2 (Xiao et al., 2025) supplies trajectories, camera poses, visibility, and confidence from the same RGB-only reconstruction. Its points and poses share a consistent within-clip coordinate system for the motion compensation in Eq. 3. This RGB-only teacher path supplies geometric supervision without requiring simulator geometry or calibrated robot-camera poses. Writing $d_{k,t} = x^{c_0}_{k,t+1} - x^{c_0}_{k,t}$, we define

$$a_{k,t} = \min(\alpha_{k,t}, \alpha_{k,t+1}), \qquad \widetilde{a} = \text{median}_{(k,t):v_{k,t}v_{k,t+1}=1}\, a_{k,t}, \tag{11}$$

$$m_{k,t} = \mathbb{1}[v_{k,t} = 1 \wedge v_{k,t+1} = 1]\,\mathbb{1}[a_{k,t} \geq \widetilde{a}], \tag{12}$$

$$s_{\text{clip}} = \text{median}_{(k,t):m_{k,t}=1} \|d_{k,t}\|_2, \qquad \Delta_{k,t} = d_{k,t}/(s_{\text{clip}} + \varepsilon). \tag{13}$$

Thus both endpoints must be visible and the weaker endpoint must pass the clipwise confidence threshold. Using the same scale for every transition preserves relative motion magnitudes within the clip. Clips with an empty valid set contribute zero dynamic loss. The positive floor ε stabilizes normalization for near-static clips. This normalization applies only to the auxiliary displacement

targets; the original action flow-matching targets are unchanged. The resulting pseudo-labels encode scale-normalized within-clip motion.

Query-indexed readout and gradient paths. A register for query k and transition t is initialized as $r_{k,t} = r_0 + W_{uv}\gamma_8(u_k) + e_t$, combining a learned base token, an eight-band Fourier embedding of normalized frame-0 location u_k, and a learned time embedding. It attends to selected video states and predicts an array of shape $B \times N \times (T-1) \times 3$. In Eq. 4, Smooth-$\ell_1$ is averaged over the three coordinates and then over valid query-transition pairs. Teacher outputs are detached, but the student video states are not: auxiliary gradients update the WAM through the readout heads. Registers and teacher features never become keys or values for the WAM's own video/action tokens. This read-only forward interface preserves the base attention constraints and permits deletion of all auxiliary branches after GeoSFT.

Complementary roles of distillation and reward diagnosis. The Stage-I tracker confidence $\alpha_{k,t}$ selects reliable demonstration pseudo-labels for dynamic supervision. Stage II applies a separate frozen 4D reconstructor, 4RC (Luo et al., 2026), to policy-generated futures and evaluates its candidate scores against executed outcomes. This policy-relative diagnosis connects geometric feedback to the behavior of the distilled policy and guides trajectory selection for GeoDPO. Static structure, dynamic correspondence, and outcome-informed selection thus contribute at different stages of the pipeline (Table 5).

Table 5: **Geometry roles across G2P-WAM.** Distillation targets, diagnostic signals, and preference rewards serve different functions; all geometry modules are removed at deployment.

Role	Input and target	Reliability test	Effect on WAM
Static-3D distillation	Demonstration frames; multi-level structure	Patch/view alignment; teacher frozen	Supervise video representations
Dynamic-4D distillation	Demonstration clips; indexed displacement	Endpoint visibility; track confidence; shared scale	Supervise temporal correspondence
Reward diagnosis and GeoDPO	Imagined futures plus executed outcomes; preferences	Within-policy AUC; motion diagnostics; imagined–observed agreement	Select pairs, then update the action policy

From complementary supervision to preference alignment. GeoSFT integrates the structural features of VGGT (Wang et al., 2025), the dynamic trajectories of SpatialTrackerV2 (Xiao et al., 2025), and the direction–scale alignment of Geometry Forcing (Wu et al., 2026) into a joint WAM training objective. The framework couples this representation-grounding stage with policy-relative score diagnosis and confidence-guided preference selection. All auxiliary geometry branches are removed after training, so the learned policy retains its original inference interface. Section 4.6 evaluates static-only SFT (A1), dynamic-only SFT (A2), and joint GeoSFT (A3) under shared demonstration data and joint training across all 50 tasks. These comparisons show the benefit of combining static and dynamic supervision, while the preference controls evaluate the subsequent contribution of confidence-based selection. The official-weight LingBot-VA reference (A0) and detailed training protocols are reported in Appendix A.13.

Geometry-aware baselines and evaluation conventions. WAM4D (Li et al., 2026c) is represented by its full 50-task RoboTwin results; the mean of Clean/Random 93.82/89.86 is 91.84. MECo-WAM (Zhang et al., 2026b) contributes its LIBERO and RoboTwin main results, and the 4D-WAM (Yang et al., 2026) row uses the FastWAM-Joint protocol. GaussianWAM (Zhang et al., 2026c) uses Cosmos Policy and reports an official LIBERO-Plus aggregate, whereas G2P-WAM reports the seven-axis mean. The benchmark tables retain each source's evaluation and aggregation conventions and indicate unavailable results explicitly.

Teacher targets and cross-stream attention. Figure 6 shows the Stage-I supervision sources at six temporal samples. VGGT feature responses and indexed point tracks illustrate static and dynamic targets; the PCA response is for visualization, while Equation 2 aligns full feature vectors. The bottom row shows WAM action-to-video attention over six latent stages on a separate rollout. The 4RC reward readout appears in Figure 12.

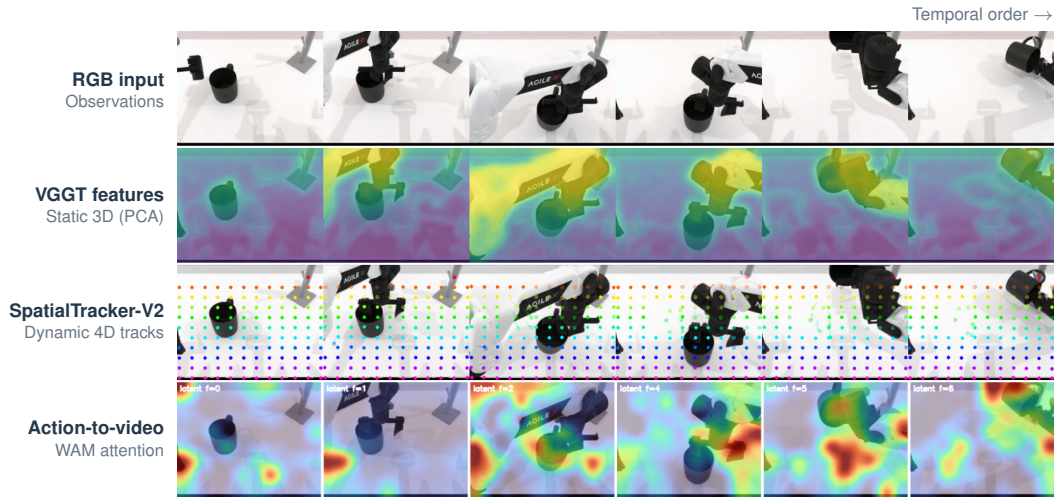


Figure 6: **Geometry supervision and WAM attention.** Rows show RGB inputs, VGGT features (PCA visualization), SpatialTracker-V2 point tracks, and WAM action-to-video attention, from top to bottom. The first three rows share six temporal samples; track colors identify points. The attention row shows six ordered latent stages from a separate rollout.

A.2 Visual guide to the control benchmarks

Figure 7 groups the control-benchmark observations and LIBERO-Plus perturbation examples in one place, illustrating the scene diversity and robustness settings used in the experiments. Geometry supervision is shown in Figure 6.

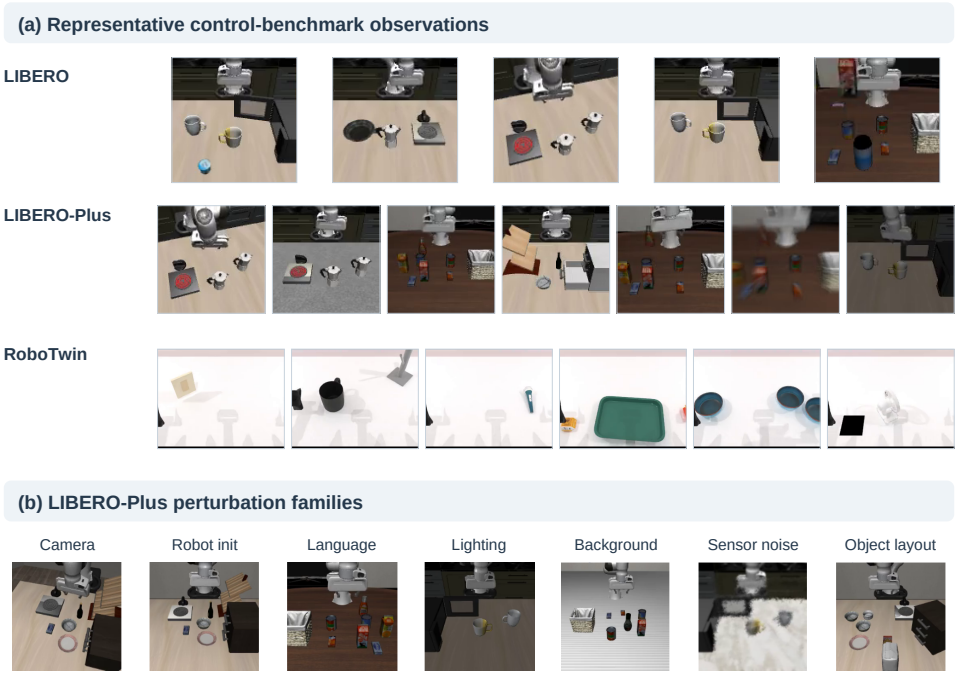


Figure 7: **Benchmark observations and robustness settings.** **(a)** Representative LIBERO, LIBERO-Plus, and RoboTwin observations. **(b)** The seven LIBERO-Plus perturbation families.

A.3 Full per-task results on RoboTwin 2.0

Table 6 lists all 50 tasks for WAM4D, LingBot-VA, and Fast-WAM, complementing the aggregate comparison in Table 2. Separate GeoSFT and G2P-WAM columns distinguish geometry distillation from subsequent GeoDPO alignment under the same evaluation protocol.

Table 6: **Per-task RoboTwin 2.0 success rates (%).** Easy/Hard denote Clean/Randomized; each GeoSFT and G2P-WAM cell contains 100 trials.

		WAM4D		LingBot-VA		Fast-WAM		**GeoSFT**		**G2P-WAM**	
Simulation Task	Horizon	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Adjust Bottle	1	100	99	90	94	100	100	92	95	93	95
Beat Block Hammer	1	99	97	96	98	99	97	97	96	95	96
Blocks Ranking RGB	3	100	97	99	98	100	100	98	97	97	96
Blocks Ranking Size	3	98	84	94	96	94	98	93	96	93	95
Click Alarmclock	1	100	100	99	100	100	100	99	99	100	99
Click Bell	1	100	100	100	100	100	100	100	100	100	98
Dump Bin Bigbin	1	92	94	89	96	97	96	92	95	95	96
Grab Roller	1	100	100	100	100	100	100	100	99	100	98
Handover Block	2	96	93	99	78	95	81	93	80	95	81
Handover Mic	2	94	94	94	96	99	100	95	96	97	98
Hanging Mug	2	64	57	40	28	58	62	47	44	48	45
Lift Pot	1	100	88	100	99	100	100	99	99	100	99
Move Can Pot	1	96	97	94	97	90	88	95	98	95	97
Move Pillbottle Pad	1	100	99	99	99	100	99	98	99	99	99
Move Playingcard Away	1	100	100	100	99	100	100	98	99	99	98
Move Stapler Pad	1	79	76	91	79	77	64	86	75	87	79
Open Laptop	1	100	66	92	94	98	100	95	96	95	97
Open Microwave	1	61	65	82	86	62	45	84	87	86	88
Pick Diverse Bottles	2	98	92	89	82	80	85	87	84	85	86
Pick Dual Bottles	2	100	99	100	99	100	96	100	98	100	99
Place A2B Left	1	97	86	97	93	95	93	98	92	98	93
Place A2B Right	1	94	79	97	95	93	99	94	95	95	98
Place Bread Basket	1	93	97	97	95	91	93	96	95	95	97
Place Bread Skillet	2	93	87	95	90	90	93	96	93	96	95
Place Burger Fries	2	96	97	97	95	96	99	98	99	97	98
Place Can Basket	2	87	83	81	84	71	69	86	88	89	85
Place Cans Plasticbox	2	100	100	100	99	99	96	100	98	99	97
Place Container Plate	1	98	91	99	97	96	100	98	96	97	98
Place Dual Shoes	2	96	97	94	89	94	88	95	86	93	85
Place Empty Cup	1	100	100	100	100	100	100	98	97	99	96
Place Fan	1	100	95	99	93	96	96	99	94	99	95
Place Mouse Pad	1	96	97	93	96	83	89	94	95	94	96
Place Object Basket	2	88	90	91	88	89	88	89	87	90	86
Place Object Scale	1	99	97	96	95	90	97	94	93	95	93
Place Object Stand	1	98	100	99	96	90	94	98	95	97	95
Place Phone Stand	1	94	81	97	97	97	99	96	97	95	96
Place Shoe	1	100	99	98	98	96	99	97	96	98	96
Press Stapler	1	97	94	85	82	90	97	89	86	90	87
Put Bottles Dustbin	3	87	85	87	91	95	90	86	89	88	91
Put Object Cabinet	2	81	53	85	87	94	89	88	89	89	90
Rotate QRcode	1	95	95	96	91	93	89	95	93	95	96
Scan Object	2	92	68	96	91	89	92	91	89	93	91
Shake Bottle Horizontally	1	100	99	100	99	100	100	100	99	99	99
Shake Bottle	1	100	99	100	97	100	100	100	96	99	97
Stack Blocks Three	3	98	97	99	98	95	97	97	96	99	98
Stack Blocks Two	2	100	100	100	98	100	100	98	100	98	99
Stack Bowls Three	3	83	84	86	83	80	81	85	83	87	86
Stack Bowls Two	2	96	95	94	98	92	98	95	96	95	97
Stamp Seal	1	96	87	96	97	90	94	95	93	92	93
Turn Switch	1	60	64	44	45	61	59	56	60	61	63
Average (50 tasks)	–	93.82	89.86	92.90	91.50	91.88	91.78	92.98	91.94	**93.40**	**92.50**

GeoSFT and G2P-WAM each contain 5,000 trials per setting. Because every task has the same trial count, task-macro and pooled success rates coincide; Horizon is not a weight.

Evaluation protocol and baseline results. Easy/Hard denote Clean/Randomized settings. Baseline scores are taken from the corresponding original papers (Li et al., 2026c;b; Yuan et al., 2026). For both of our methods, every cell contains $N{=}100$ trials, giving 5,000 trials per setting. GeoSFT records $4{,}649/5{,}000$ Clean and $4{,}597/5{,}000$ Randomized successes ($92.98/91.94$); G2P-WAM records $4{,}670/5{,}000$ and $4{,}625/5{,}000$ ($93.40/92.50$). Thus GeoDPO changes the two settings by $+0.42/+0.56$ points and the equal-setting mean from 92.46 to 92.95.

Confidence intervals for the mean improvement. We repeatedly draw 50 tasks with replacement and average the success-rate improvement of G2P-WAM over GeoSFT on the selected tasks. Using 200,000 bootstrap resamples gives 95% confidence intervals of $[0.02, 0.82]$ percentage points on Clean, $[0.16, 0.96]$ on Randomized, and $[0.18, 0.81]$ for the equal-setting mean. The intervals measure sensitivity to task composition for the same evaluated models.

A.4 Real-robot evaluation protocol and detailed results

Tasks and final-state criteria. On AgileX PiPER X, we evaluate three goal-level physical manipulation tasks. **Store fruit in drawer** is successful only if the drawer is opened, the fruit is placed fully inside, and the drawer is closed at the end of the episode. **Place plate on cabinet** requires grasping the plate, transferring it onto the cabinet, and releasing it in a stable resting state. **Place banana on plate** requires grasping the banana and releasing it on the plate without dropping it outside the target region. The keyframes in Figure 5(a) come from a separate continuous demonstration of the three tasks. The hardware platform is an AgileX PiPER X dual-arm robot system with parallel-jaw grippers.

Comparison protocol. We compare $\pi_{0.5}$ (Physical Intelligence et al., 2025), LingBot-VA (Li et al., 2026b), GeoSFT, and G2P-WAM on the same physical setup. GeoSFT is the geometry-distilled checkpoint from which GeoDPO produces G2P-WAM. All methods use the same observation and instruction interface, action representation, task horizon, reset protocol, and final-state scoring rule. Each method is evaluated on 30 trials for each of the three tasks. Every trial starts from a reset following the same specification across methods. We report raw success counts together with percentages; the overall Mean is the unweighted average of the three per-task rates.

GeoSFT succeeds in $22/30$, $23/30$, and $25/30$ trials on the three tasks, giving a mean of 77.8%. G2P-WAM succeeds in $24/30$, $26/30$, and $27/30$ trials, giving 85.6% and an improvement of 7.8 percentage points.

Table 7: **AgileX PiPER X success rates (%).** Thirty trials per task and method; cells show rates and success counts. Mean equally weights the three tasks. GeoSFT is the distilled checkpoint before GeoDPO.

Method	Fruit in drawer	Plate on cabinet	Banana on plate	Mean
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	50.0 (15/30)	60.0 (18/30)	70.0 (21/30)	60.0
LingBot-VA (Li et al., 2026b)	66.7 (20/30)	73.3 (22/30)	80.0 (24/30)	73.3
GeoSFT	73.3 (22/30)	76.7 (23/30)	83.3 (25/30)	77.8
G2P-WAM (ours)	**80.0 (24/30)**	**86.7 (26/30)**	**90.0 (27/30)**	**85.6**

Confidence intervals for task success. Table 8 gives two-sided 95% Wilson score intervals for each task and method on both robot platforms, using $z = 1.96$ and independent binomial trials within each cell. The intervals quantify trial-sampling uncertainty for each evaluated checkpoint.

Table 8: **Per-task 95% confidence intervals for real-robot success rates (%).** Wilson intervals use the success counts in Tables 7 and 9, with 30 trials per task and method.

AgileX PiPER X			
Method	Fruit in drawer	Plate on cabinet	Banana on plate
$\pi_{0.5}$	[33.2, 66.8]	[42.3, 75.4]	[52.1, 83.3]
LingBot-VA	[48.8, 80.8]	[55.6, 85.8]	[62.7, 90.5]
GeoSFT	[55.6, 85.8]	[59.1, 88.2]	[66.4, 92.7]
G2P-WAM	[62.7, 90.5]	[70.3, 94.7]	[74.4, 96.5]
UNT David A1C			
Method	Object handover	Object stacking	Drawer opening
$\pi_{0.5}$	[70.3, 94.7]	[59.1, 88.2]	[52.1, 83.3]
LingBot-VA	[74.4, 96.5]	[66.4, 92.7]	[55.6, 85.8]
GeoSFT	[78.7, 98.2]	[74.4, 96.5]	[62.7, 90.5]
G2P-WAM	[83.3, 99.4]	[78.7, 98.2]	[70.3, 94.7]

Additional experiments on UNT David A1C. We additionally evaluate three tasks on the **UNT David A1C** robot platform: bimanual object handover, object stacking, and drawer opening. Figure 8 shows representative executions. The handover transfers a red object between the hands and back before release onto the table; stacking places it on a blue base; drawer opening ends with the drawer open and the hand withdrawn. Each row is a separate recording. These tasks complement the AgileX PiPER X evaluation in Table 7.

Table 9 reports 30 trials per task and method on this platform. G2P-WAM achieves a mean success rate of 92.2%, compared with 87.8% for GeoSFT, 82.2% for LingBot-VA, and 77.8% for $\pi_{0.5}$. The improvement over GeoSFT is 4.4 percentage points, with gains on all three tasks. These results are reported separately from the AgileX PiPER X averages in the main text.

Table 9: **UNT David A1C success rates (%).** Thirty trials per task and method; cells show rates and success counts. Mean equally weights the three tasks.

Method	Object handover	Object stacking	Drawer opening	Mean
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	86.7 (26/30)	76.7 (23/30)	70.0 (21/30)	77.8
LingBot-VA (Li et al., 2026b)	90.0 (27/30)	83.3 (25/30)	73.3 (22/30)	82.2
GeoSFT	93.3 (28/30)	90.0 (27/30)	80.0 (24/30)	87.8
G2P-WAM (ours)	**96.7 (29/30)**	**93.3 (28/30)**	**86.7 (26/30)**	**92.2**

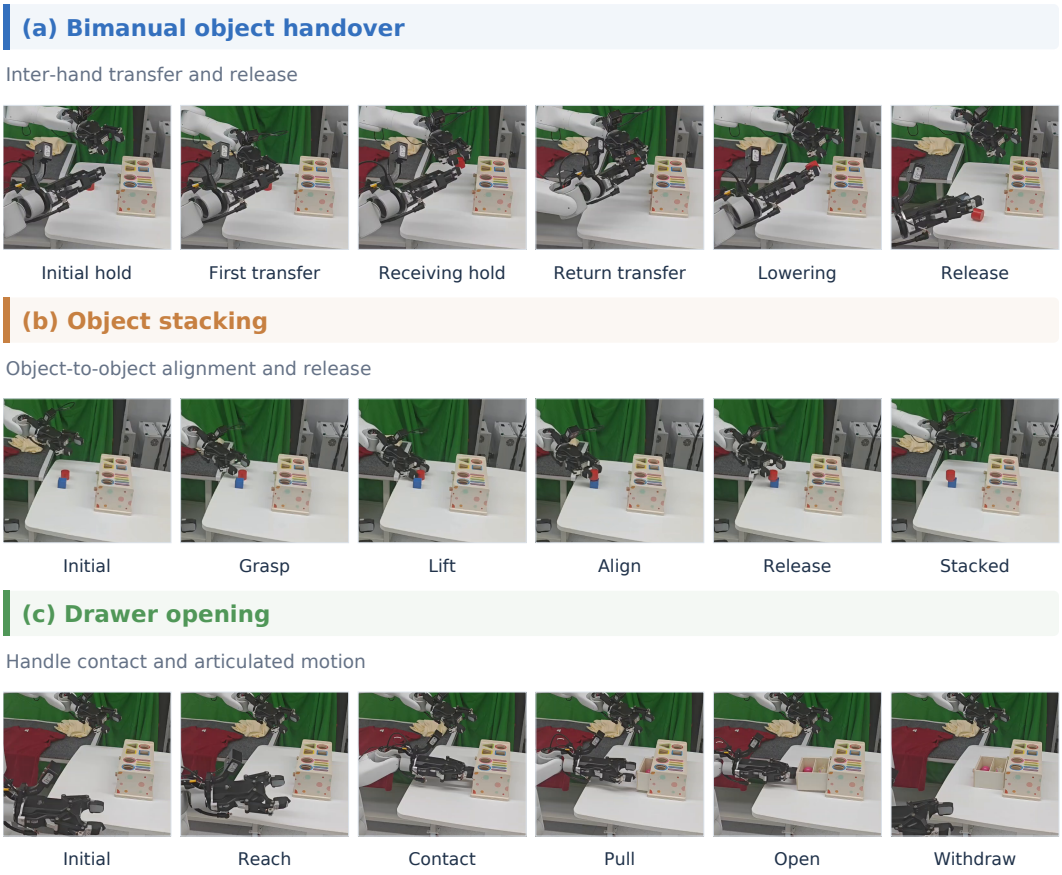


Figure 8: **Additional manipulation sequences on UNT David A1C. (a)** Bimanual object handover with return transfer and final release. **(b)** Stacking a red object on a blue base. **(c)** Opening a drawer and withdrawing the hand. Each row contains six recorded keyframes ordered from left to right, using a fixed crop of the interaction region.

A.5 Closed-loop rollout examples

Figure 4 shows three stages from each selected episode, while Figure 9 provides six ordered snapshots and both camera views. The compact figure uses the global camera for LIBERO and LIBERO-Plus, and a fixed square crop of the high camera for RoboTwin. Within each benchmark, the same cropping rule is applied to both policies.

GeoSFT and G2P-WAM rollouts. The examples cover stove and moka-pot manipulation in LIBERO, bowl and drawer manipulation under object-layout perturbations in LIBERO-Plus, and

switch manipulation in RoboTwin. Upper rows show the geometry-distilled GeoSFT policy; lower rows show its GeoDPO continuation, G2P-WAM. The selected GeoSFT failures and G2P-WAM successes illustrate the interaction sequence and final task outcome, complementing the quantitative evaluation in Section 4.3. The LIBERO-Plus examples use paired initializations. The RoboTwin episodes follow different switch instructions and illustrate separate task executions.

Successful GeoSFT executions. Figure 10 shows successful GeoSFT episodes on Hanging Mug and Turn Switch in RoboTwin. The sequences illustrate object placement and switch interaction supported by the distilled policy. Together with the failure cases above, they provide a view of GeoSFT behavior before preference alignment.

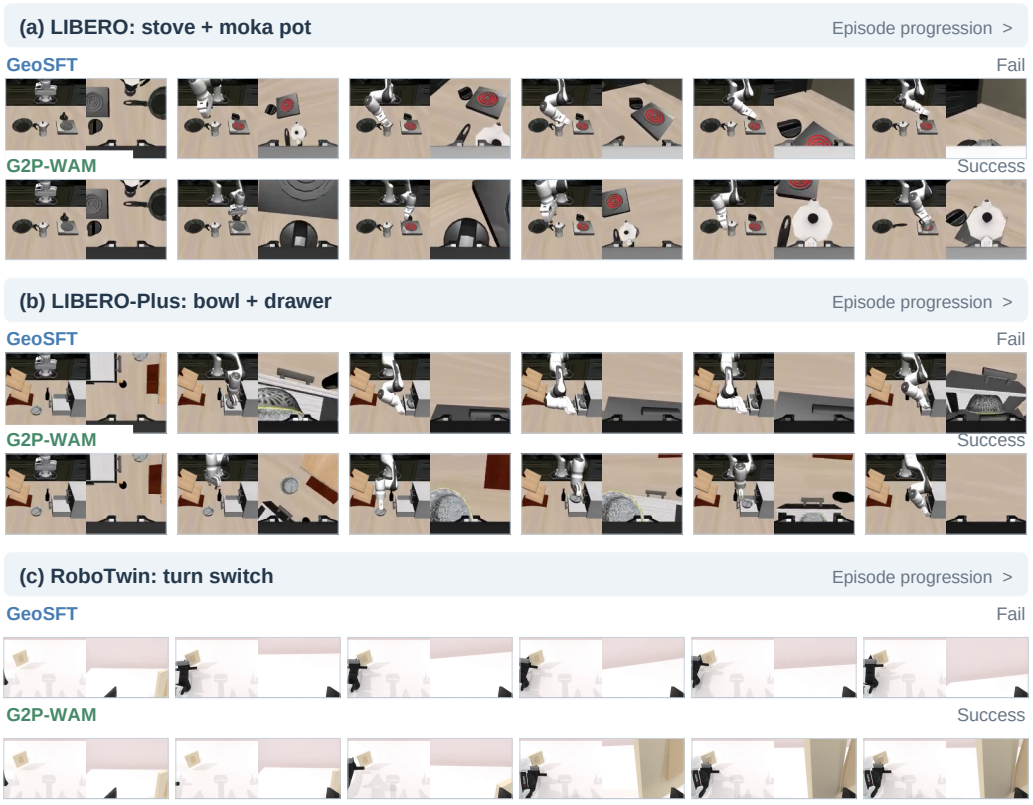


Figure 9: **Selected closed-loop rollouts.** GeoSFT failures appear above G2P-WAM successes on LIBERO, LIBERO-Plus, and RoboTwin. Each row shows six snapshots with both camera views.

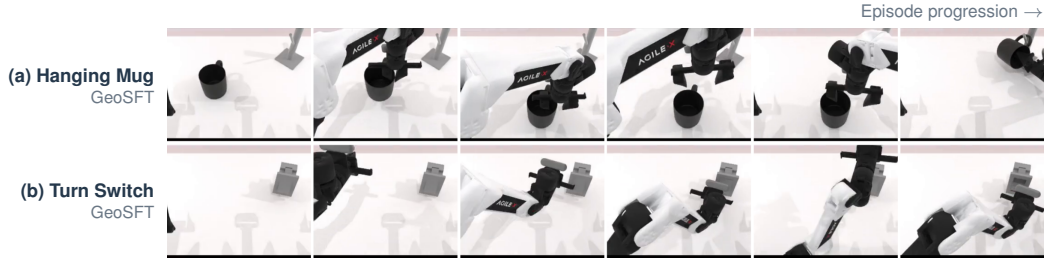


Figure 10: **Successful GeoSFT rollouts.** RoboTwin episodes on **(a)** Hanging Mug and **(b)** Turn Switch. Each row shows six ordered snapshots from initialization to task completion.

A.6 Per-factor results on LIBERO-Plus

Table 10 compares GeoSFT and G2P-WAM across the seven perturbation factors of the full LIBERO-Plus benchmark. The G2P-WAM scores are the same as those reported in Table 1; this

table additionally reports the GeoSFT baseline and the improvement for each factor. Viewpoint corresponds to Camera in the main table. The mean success rate equally weights the seven factors.

Effect of preference alignment. Compared with GeoSFT, G2P-WAM improves success on six factors and maintains the same performance on Language. The mean success rate rises from 76.7% to 79.5%, with an average gain of 2.7 percentage points. The largest gains occur for Background, Layout, and Viewpoint, indicating improved robustness to changes in visual appearance, object arrangement, and camera perspective. Performance also improves under Lighting, Robot initialization, and Sensor noise perturbations.

Training dynamics. With checkpoints saved every 25 steps, peak success occurs before preference saturation, and later checkpoints regress. This observation motivates reference anchoring, early stopping, and checkpoint selection on data disjoint from the final test.

Table 10: **LIBERO-Plus per-factor success rates.** Results are reported as percentages on the full LIBERO-Plus benchmark. The G2P-WAM scores match Table 1. Mean equally weights the seven factors, and Δ denotes the improvement over GeoSFT. The mean improvement is calculated before rounding.

Method	Layout	Backgr.	Light	Robot init	Viewpoint	Language	Noise	Mean
GeoSFT	91.7	71.6	87.5	81.3	29.0	92.0	84.1	76.7
G2P-WAM	**95.8**	**76.1**	**90.0**	**82.8**	**33.1**	92.0	**86.4**	**79.5**
Δ	+4.1	+4.5	+2.5	+1.5	+4.1	0.0	+2.3	+2.7

A.7 World-model quality and external video-benchmark evaluation

Table 11 compares GeoSFT and its GeoDPO continuation, G2P-WAM, on in-domain video prediction and the official Physics-IQ I2V benchmark. On seven held-out continuations of the Move Stapler Pad task in RoboTwin, G2P-WAM increases PSNR by 0.44 dB and SSIM (Wang et al., 2004) by 0.032, while reducing LPIPS (Zhang et al., 2018) by 0.060. These changes indicate improved reconstruction fidelity and perceptual similarity on the evaluated videos. The local PhysIQ proxy also rises from 0.788 to 0.792. On the 198 official Physics-IQ I2V cases (Motamed et al., 2026), the score increases from 28.5 to 29.6, a gain of 1.1 points. Together, these results show improvements in in-domain prediction quality and external video-benchmark performance after preference alignment.

In-domain predicted-future quality

Method	PhysIQ↑	PSNR↑	SSIM↑	LPIPS↓
GeoSFT	0.788	15.03	0.463	0.602
G2P-WAM	**0.792**	**15.47**	**0.495**	**0.542**

Official Physics-IQ

Method	Score↑
GeoSFT	28.5
G2P-WAM	**29.6**

Table 11: **World-model quality of GeoSFT and G2P-WAM.** In-domain prediction quality is evaluated on seven held-out continuations of the Move Stapler Pad task in RoboTwin. PhysIQ is a local proxy, and PSNR is reported in dB. The official Physics-IQ I2V evaluation includes 198 cases.

RBench external evaluation. We compare GeoSFT and G2P-WAM on RBench (Deng et al., 2026) with the same local Qwen2.5-VL-7B judge (Bai et al., 2025). The embodiment track uses 100 matched clips per morphology, scoring subject stability, physical plausibility, task adherence, and Q-Align quality (Wu et al., 2024); the task-oriented tracks use 50 matched clips each. Table 12 shows gains for G2P-WAM on all four morphology composites and all five task-oriented tracks. The mean task-track score rises from 4.03 to 4.24, with an average gain of 0.21 points. Long-horizon planning and visual reasoning show the largest gains, followed by spatial relationships. These results support improvements across robot morphologies and task types, although some component metrics decline, including instruction adherence across all four morphologies.

Qualitative model-imagined rollouts. Figure 11 compares GeoSFT and G2P-WAM on three RBench examples, retaining seven sampled frames per generated video. Each pair shares the same reference image; these examples complement the aggregate evaluation in Table 12. Temporal confidence readouts are grouped in Appendix A.8.

(a) Robot-embodiment track

Morphology	Method	Stability↑	Plaus.↑	Adher.↑	Quality↑	Comp.↑
Single-arm	GeoSFT	11.98	4.27	**2.08**	**0.550**	18.33
	G2P-WAM	**12.17**	**4.31**	1.98	0.535	**18.46**
Dual-arm	GeoSFT	**11.65**	4.47	**2.15**	0.451	18.27
	G2P-WAM	11.64	**4.64**	2.06	**0.473**	**18.34**
Humanoid	GeoSFT	11.67	4.19	**2.52**	0.461	18.38
	G2P-WAM	**11.70**	**4.42**	2.49	**0.526**	**18.61**
Quadruped	GeoSFT	10.23	4.77	**2.93**	0.488	17.93
	G2P-WAM	**10.46**	**4.78**	2.85	**0.490**	**18.09**

(b) Task-oriented tracks

Task	GeoSFT	G2P-WAM	Δ	Task	GeoSFT	G2P-WAM	Δ
Common manip.	3.98	**4.16**	+0.18	Long-horizon	4.15	**4.44**	+0.29
Spatial rel.	4.62	**4.85**	+0.23	Multi-entity	3.59	**3.68**	+0.09
Visual reason.	3.83	**4.07**	+0.24	**Mean**	4.03	**4.24**	**+0.21**

Table 12: **RBench external evaluation** (Deng et al., 2026). GeoSFT and G2P-WAM on **(a)** the robot-embodiment track and **(b)** five task-oriented tracks. Higher is better; Mean equally weights the five task tracks, and Δ denotes the change from GeoSFT to G2P-WAM, computed before rounding.

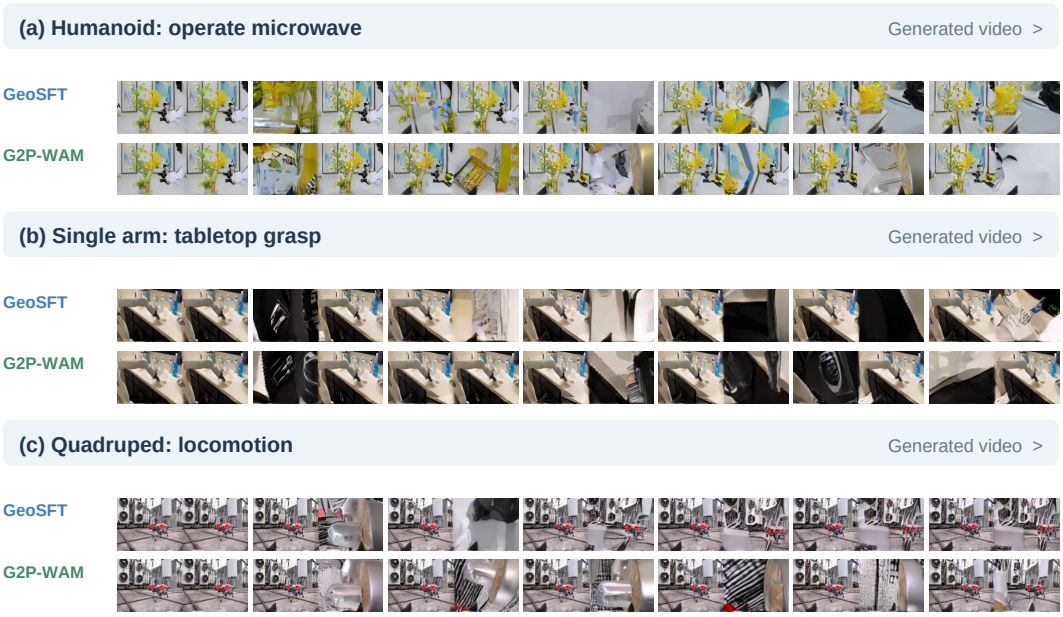


Figure 11: **Generated-video examples on RBench.** GeoSFT (top) and G2P-WAM (bottom) for humanoid, single-arm, and quadruped prompts; each pair shares the same reference frame.

A.8 Temporal confidence readouts across control benchmarks

Figure 12 shows the 4RC point-confidence signal used by GeoDPO on RoboTwin, LIBERO-Long, and LIBERO-Plus. Each RGB sequence is aligned with its confidence maps in temporal order, illustrating variation across tasks and frames. These visualizations complement the outcome-based reward analysis in Appendix A.9.

A.9 Cross-task reward validity on RoboTwin

We complement the GeoSFT calibration in Table 13 with exploratory score–outcome analyses of official-backbone policies. These diagnostics examine how geometric signals vary across tasks and rollout conditions, motivating calibration on the policy used for preference training. The diagnostic pools are separate from the uniform RoboTwin-50 evaluation. The official-policy pool contains 2,400 records across 14 tasks, and a weaker-base-policy pool contains 48 records.

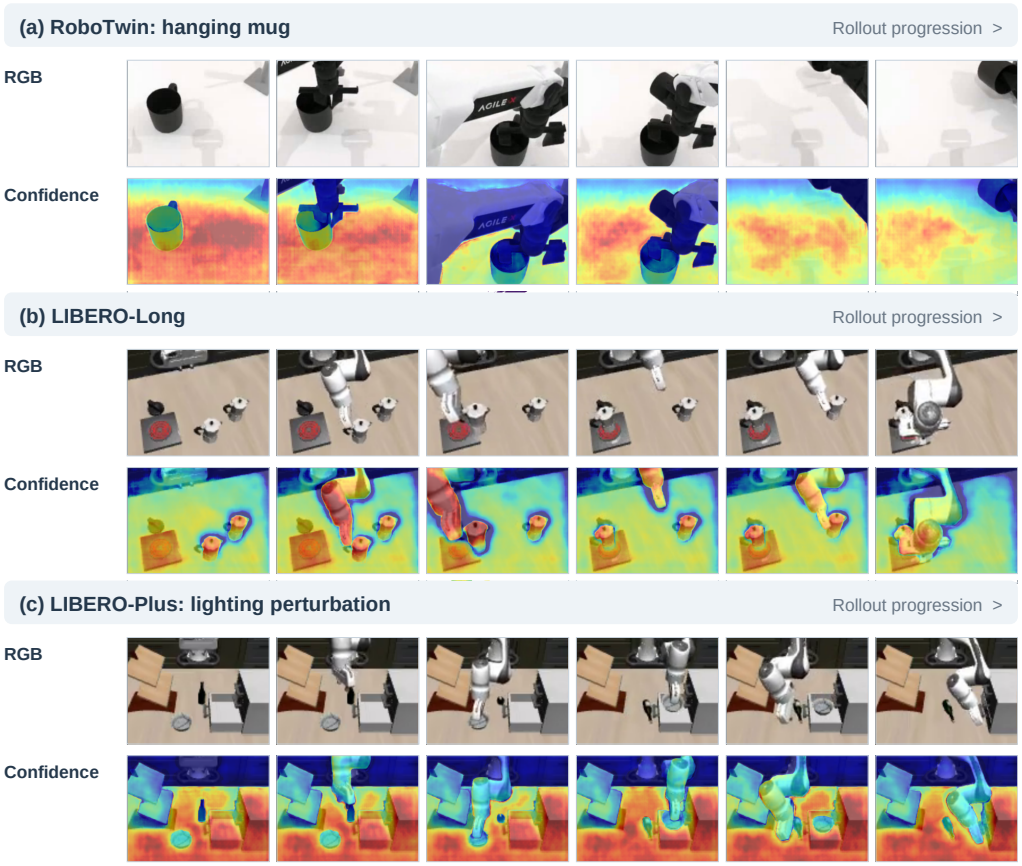


Figure 12: **Temporal confidence readouts across control benchmarks.** The panels show **(a)** Hanging Mug in RoboTwin, **(b)** LIBERO-Long, and **(c)** LIBERO-Plus with lighting perturbations. Each panel pairs RGB frames with corresponding 4RC point-confidence maps in temporal order from left to right. Mean point confidence provides the selected GeoDPO score.

Table 13: **Reward validity on Turn Switch.** AUCs measure how well 4RC scores distinguish successful from failed executions across 122 episodes, using either imagined or observed videos. An AUC of 0.5 indicates chance-level discrimination.

4RC quantity	Imagined AUC	Observed AUC	Interpretation
Reprojection consistency	0.553	0.510	Near-chance discrimination
Point confidence	**0.744**	**0.752**	Similar AUCs across video sources
Estimated motion	0.226	0.235	Lower scores associated with success
Pose jitter	0.293	0.377	Correlated with motion, $r = 0.89$
Depth variance and outliers	∼0.4–0.5	∼0.4–0.5	Limited outcome discrimination

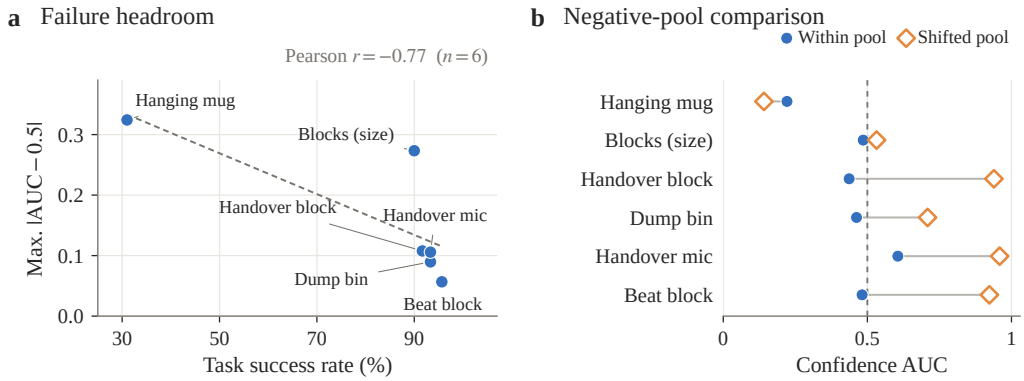
Task success and geometric discriminability. Six official-policy tasks have at least ten successes and ten failures; four contain 300 records each, and two contain 100 each. For each eligible task, we compute

$$D_{\text{task}} = \max_{m \in \mathcal{M}} |\text{AUC}_m - 0.5|,$$

where $\mathcal{M}$ contains the eight metrics in Figure 13(c). The horizontal axis shows task success rate, $100 n_{\text{succ}}/n$. Across the six tasks, the Pearson correlation between success rate and D_{task} is $r = -0.772$, with equal weight assigned to each task. The plotted line is a descriptive ordinary least-squares fit to the six task-level points. Each task's AUC uses all available scoring records. In this pool, tasks with greater failure headroom tend to exhibit stronger separation for at least one geometric metric. This observation motivates examining candidate scores on rollouts that include both successful and failed executions.

Task-dependent score direction. All AUCs retain the original score direction. Point-confidence AUC is 0.606 for Handover Mic and 0.221 for Hanging Mug, and the rankings of other metrics also vary by task. This variation motivates policy-relative calibration and the separation of preference ordering from geometric selection. GeoDPO uses execution outcomes to determine the preferred trajectory and applies a fixed confidence-filtering rule within each task and outcome group. Table 3 evaluates the policy benefit of this selection under matched training settings.

Sensitivity to the negative pool. Both comparisons use the same official-policy success pool for each task. Within-pool negatives are failures of the official policy in the Randomized setting. Shifted-pool negatives are failures of a weaker base policy in the Clean setting. Replacing the negative pool increases confidence AUC on five of the six tasks, with four shifted-pool values exceeding 0.70. Each shifted pool contains eight negative records and jointly changes the policy and environment. The resulting AUC changes illustrate the sensitivity of geometric discrimination to rollout conditions, supporting GeoSFT-based calibration before preference selection. Exact AUCs and sample counts accompany the source release.



c Task-dependent metric spectrum

	Conf.	Reproj.	Motion	Low-conf.	Depth std.	Depth outlier	Pose jitter	Point drift
Hanging mug (31/69)	0.22	0.67	0.45	0.82	0.75	0.25	0.51	0.50
Blocks (size) (90/10)	0.49	0.66	0.40	0.61	0.77	0.38	0.43	0.76
Handover block (275/25)	0.44	0.57	0.56	0.51	0.60	0.47	0.61	0.58
Dump bin (280/20)	0.46	0.53	0.52	0.41	0.50	0.56	0.54	0.49
Handover mic (280/20)	0.61	0.46	0.42	0.47	0.50	0.52	0.44	0.59
Beat block (287/13)	0.48	0.48	0.52	0.45	0.49	0.49	0.55	0.44

AUC: 1, 0.5, 0

Figure 13: **RoboTwin reward-validity diagnostics.** **(a)** Task success rate versus the maximum $|\mathrm{AUC} - 0.5|$ across eight metrics on six eligible tasks. **(b)** AUC comparison with a fixed success pool. Hollow diamonds mark shifted negative pools containing eight records per task, with both policy and environment changed. **(c)** Raw metric AUCs and success and failure counts. Fits summarize descriptive associations in the diagnostic pools.

A.10 Exploratory reward-source controls

Table 14 summarizes exploratory runs that informed reward selection and optimization in GeoDPO. Point confidence supports a success-rate gain in its reported run, while reference-regularized cycle consistency preserves task success as prediction error decreases.

These observations motivate GeoDPO's combination of outcome-ordered preferences, confidence-based selection, and reference anchoring. The matched outcome-only comparison in Section 4.6 evaluates the additional policy benefit of confidence selection under the final training protocol.

Table 14: **Exploratory comparisons of reward sources.** Section 4.6 reports the matched confidence-filtering comparison.

Reward source	Training condition	Reported outcome
VGGT reprojection	WAM-DPO consistency objective	Collapse: 0 successes in 5 trials
Cycle consistency	Full data without reference anchoring	Success falls from 100% to 0%
Cycle consistency	Reference anchoring and early stopping	Error decreases by 3.3% with success unchanged
D4RT features (Zhang et al., 2026a)	Feature score from a community implementation	Success falls from 55% to 10–15%
Point confidence	Reconstruction-confidence score	Success rises from 80% to 83.3% with stable training

A.11 Reward-validity diagnostics on LIBERO-Plus

We analyze 2,624 scored records across seven LIBERO-Plus perturbation families (Fei et al., 2025), comprising 1,997 successes and 627 failures. All four panels in Figure 14 use this same diagnostic pool. The recorded scoring configuration specifies the official LIBERO-Long checkpoint and eight 4RC quantities. These exploratory analyses characterize geometric signals under different perturbations and complement the policy-level evaluation.

Outcome coverage and aggregation. For each family, AUC and success rate are computed directly from the same pooled records across perturbation variants. Pooling retains records from individual variants with sparse outcome coverage. Family sample sizes range from 128 to 704, and variant indices identify LIBERO-Plus perturbation settings.

Associations across perturbation families. Pearson correlations assign equal weight to each of the seven families. The correlations between raw AUC and success rate are -0.636 for the low-confidence fraction and -0.482 for mean confidence. Point drift and pose jitter yield $+0.665$ and $+0.745$, respectively. Both the direction and magnitude of these associations vary across metrics. This diversity shows that geometric diagnostics capture different aspects of the evaluated trajectories. It motivates comparing multiple candidate signals and calibrating their use on the policy that supplies preference-training data.

A.12 Implementation details and reproducibility

Training and inference configuration. SpatialTrackerV2 (Xiao et al., 2025) supplies the Stage-I trajectory targets. GeoDPO constructs within-task success–failure pairs after task-level confidence filtering, with each trajectory retaining its own initial observation and conditioning context. Evaluation uses the same attention kernel as training to maintain consistent numerical behavior. The Turn Switch calibration contains 122 rollouts, with 79 successes and 43 failures. Imagined and observed videos use matched high-camera crops. Table 13 reports the calibration AUCs.

Calibration and diagnostic datasets. The GeoSFT calibration in Figure 3a is distinct from the official-backbone pools analyzed in Appendices A.9 and A.11. Each calibration episode provides imagined and observed videos with the same executed outcome. For mean point confidence on imagined videos, the reported rollout-bootstrap 95% AUC interval is $[0.64, 0.84]$. Estimated motion has an AUC of 0.23 in its original direction and 0.77 when the score is reversed. These comparisons characterize the score directions and support selecting point confidence for the subsequent outcome-stratified filter.

The recorded diagnostic configuration uses a frozen 4RC scorer with at most eight uniformly sampled frames. The RoboTwin source pools correspond to an official policy evaluated in Randomized environments and a weaker base policy evaluated in Clean environments. These pools support the cross-task diagnostic comparisons, while GeoSFT supplies the rollouts used for current-policy calibration and preference training.

Numerical exports and score conventions. The diagnostic export contains 5,392 scoring records and 43,136 scalar values, with figure-level summaries for the reported analyses. AUCs in Figures 13 and 14 retain the original metric directions to make task-dependent behavior visible. GeoDPO uses the fixed confidence-selection direction specified in Section 3.4, with quantile thresholds computed within each task and outcome group.

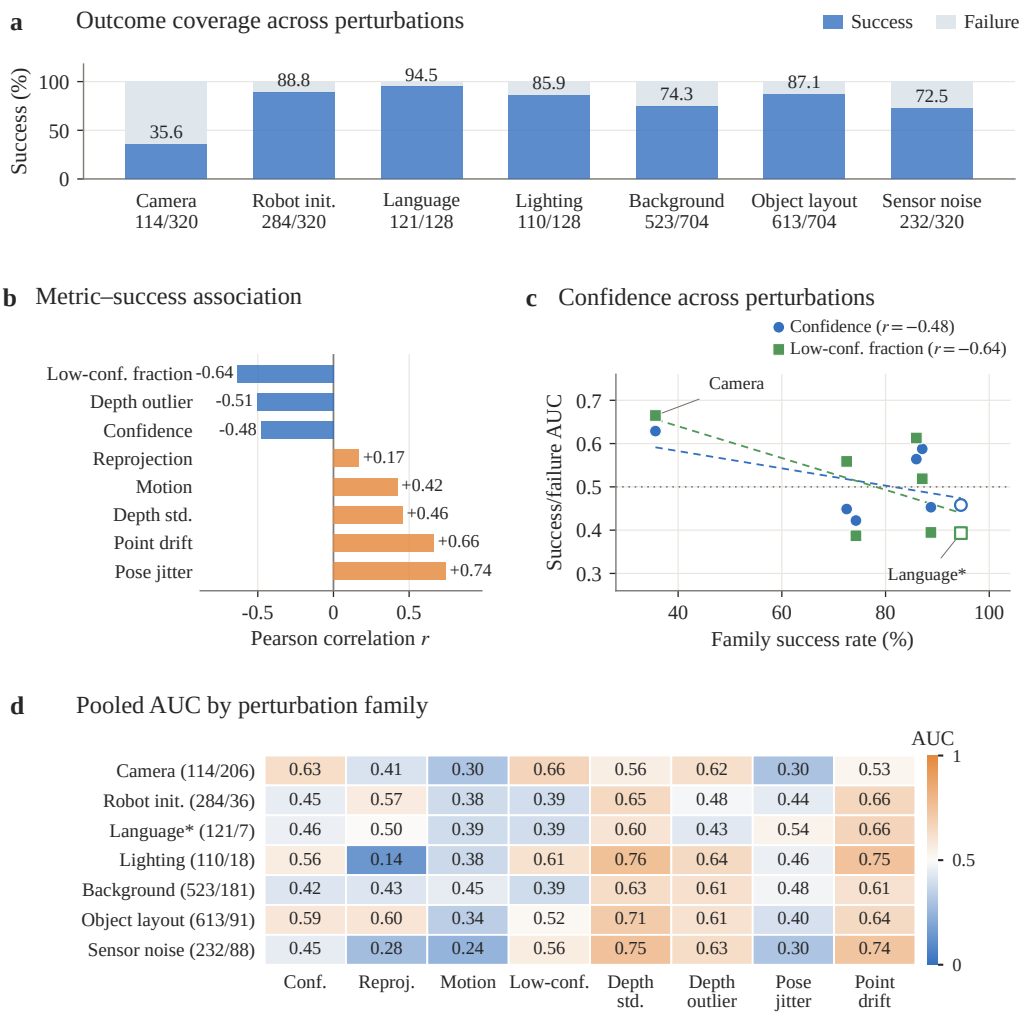


Figure 14: **LIBERO-Plus diagnostics from** 2,624 **scored records. (a)** Outcome coverage, labeled with successes and total records. **(b)** Family-level Pearson correlations between AUC and success rate. **(c)** Confidence AUCs with descriptive fits and a chance-level reference line. **(d)** Family-pooled AUCs, labeled with success and failure counts. Hollow markers and the asterisk identify the Language family. All panels use the same seven-family diagnostic pool.

Compute and release. Experiments used 32 NVIDIA A100 GPUs. The source includes figure-level numerical exports. Teacher and checkpoint configurations, scoring scripts, evaluation harnesses, and trial manifests are part of the planned release.

A.13 RoboTwin component-comparison protocol

Table 3 distinguishes the official-checkpoint reference, geometry-distilled variants, and preference-aligned policies. This section describes their training and evaluation protocols and connects the controlled comparisons to the contributions of representation grounding and preference alignment.

Official reference and representation grounding. A0 directly evaluates the official LingBot-VA checkpoint without additional SFT. A1–A3 use the same mixture of Clean and Randomized demonstrations, and each trains one policy jointly across all 50 tasks. Each resulting model is evaluated in both settings. All three variants update the full WAM rather than only the action branch. A1 uses $\mathcal{L}_{\mathrm{S3D}}$, A2 uses $\mathcal{L}_{\mathrm{D4D}}$, and A3 uses both, in addition to the flow-matching objective. Static loss weights are shared by A1 and A3. Dynamic loss weights and warm-up settings are shared by A2 and A3. A1–A3 therefore provide the matched comparisons of geometric supervision. A0 serves as

the official-checkpoint reference; its contrast with the trained variants captures the combined effect of further training and geometric supervision.

Episode-level scoring and task-level filtering. For each candidate rollout, GeoDPO uniformly samples at most eight frames from the WAM-predicted video and evaluates them with frozen 4RC. Point confidence is averaged over sampled frames and spatial locations. The score quantifies reconstruction confidence, and the executed rollout supplies the binary success outcome. These complementary signals support trajectory selection and preference ordering, respectively.

For task k, let $\mathcal{S}_k = \{\tau : y(\tau) = 1\}$ and $\mathcal{F}_k = \{\tau : y(\tau) = 0\}$ denote successful and failed rollout pools. GeoDPO ranks successful rollouts by descending confidence and retains the top 50%. Failed rollouts are ranked by ascending confidence, and the lowest-scoring 50% are retained. We denote the retained pools by $\widetilde{\mathcal{S}}_k$ and $\widetilde{\mathcal{F}}_k$. Preference pairs are sampled randomly from $\widetilde{\mathcal{S}}_k \times \widetilde{\mathcal{F}}_k$, with trajectory reuse allowed. Outcomes determine the preference order, while confidence selects the candidate trajectories. Filtering and pairing operate within each task across separately collected rollouts, including trajectories with different resets, initial observations, or robot states. The score and filtering direction remain fixed across tasks, with outcome-stratified quantile thresholds computed separately for each task.

Confidence and estimated pose motion. Across the 122 imagined-video calibration records, the Spearman correlations between confidence and the pose-motion proxy are -0.591 for successes and -0.360 for failures. The proxy adds mean adjacent camera-translation magnitude to 0.1 times the mean rotation angle. The proxy describes estimated camera-pose variation in reconstruction coordinates, whose translation scale is specific to the reconstruction; it is not a measure of physical robot displacement. Outcome-stratified filtering retains 39 of 79 successes and 21 of 43 failures. The corresponding changes in mean proxy are -27.2% and $+26.5\%$, respectively. These statistics characterize how confidence selection changes the motion composition of the retained trajectories. The matched B1–B2 comparison in Table 3 evaluates the resulting effect on policy performance.

Preference-alignment controls. B0 uses outcome-ordered pairs from the unfiltered rollout pool. B1 omits confidence filtering and randomly samples the same number of training pairs as B2. B2 retains half of each outcome pool according to confidence before random pair sampling. B1 and B2 share the GeoSFT initialization and frozen reference, reference anchoring, label smoothing, and all other training settings. Both use at most $2{,}000$ updates and the same validation-loss early-stopping criterion. The dagger in Table 3 identifies A3 and B2 as the GeoSFT and G2P-WAM results reused from the main evaluation.

Evaluation and aggregation. The ablation follows the main RoboTwin evaluation across 50 tasks, with 100 trials per task in each of the Clean and Randomized settings. The same model is evaluated in both settings. Let $y_{m,t,s,i} \in \{0, 1\}$ denote the outcome for method m, task t, setting s, and trial i. The task-macro success rate and equal-setting average are

$$\mathrm{SR}_{m,s} = \frac{1}{50} \sum_{t=1}^{50} \frac{100}{N_{t,s}} \sum_{i=1}^{N_{t,s}} y_{m,t,s,i}, \qquad \mathrm{Avg}_m = \tfrac{1}{2}(\mathrm{SR}_{m,\mathrm{Clean}} + \mathrm{SR}_{m,\mathrm{Randomized}}), \tag{14}$$

with $N_{t,s} = 100$, giving $5{,}000$ evaluation trials per setting. Table 3 reports the measured success rates and their arithmetic mean as point estimates. Appendix A.3 reports task-bootstrap confidence intervals for the GeoSFT–G2P-WAM comparison, quantifying sensitivity to task composition for the evaluated checkpoints. With separately collected within-task trajectories, Equation 10 serves as a cross-context ranking surrogate that compares each trajectory under its own conditioning context.

Interpretation of the reported contrasts. Joint GeoSFT exceeds static-only and dynamic-only supervision by 0.18 and 0.06 percentage points, respectively. The shared data and training settings among these variants support the benefit of combining scene structure with temporal correspondence. GeoDPO exceeds the count-matched B1 control by 0.28 percentage points on Clean and 0.34 on Randomized, with an overall gain of 0.31 points. The B2–B0 gain is 0.28 points and also reflects a change in pool coverage. The B2–A3 gain of 0.49 points measures the full GeoDPO stage. Together, these comparisons show consistent gains from joint geometric grounding and subsequent confidence-guided preference selection under the evaluated protocol.

A.14 LIMITATIONS AND FUTURE WORK

The current framework uses a fixed set of geometric teachers, and the usefulness of reconstruction confidence may vary across models and scene conditions. Future work will investigate stronger 4D reconstruction models and confidence calibration across tasks and environments. Beyond reconstruction accuracy, these comparisons will examine how well geometric scores identify trajectories that contribute to downstream policy improvement, providing a basis for selecting and combining geometric teachers.

Another direction is to enrich trajectory evaluation beyond the current combination of binary execution outcomes and reconstruction confidence. Incorporating task progress, object motion, and contact geometry could provide finer distinctions between trajectories with the same final outcome. Scoring specific interaction regions and task stages may further focus preference selection on behavior relevant to task completion. These extensions could make geometric feedback more adaptive to different manipulation tasks.

Finally, extending GeoDPO from offline preference training to online policy refinement is a promising direction. Iteratively collecting rollouts, updating geometric diagnostics, and refining preferences could allow the learning process to adapt as the policy improves. We also plan to investigate this approach on longer manipulation sequences, additional robot embodiments, and changing real-world environments, examining how geometric feedback supports continued policy improvement under broader operating conditions.