WHAT WILL THE FINGERS FEEL NEXT? A PROGRESSIVE MULTI-STEP TACTILE WORLD MODEL FOR PERSISTENT IN-HAND MANIPULATION

Anonymous authors
Paper under double-blind review

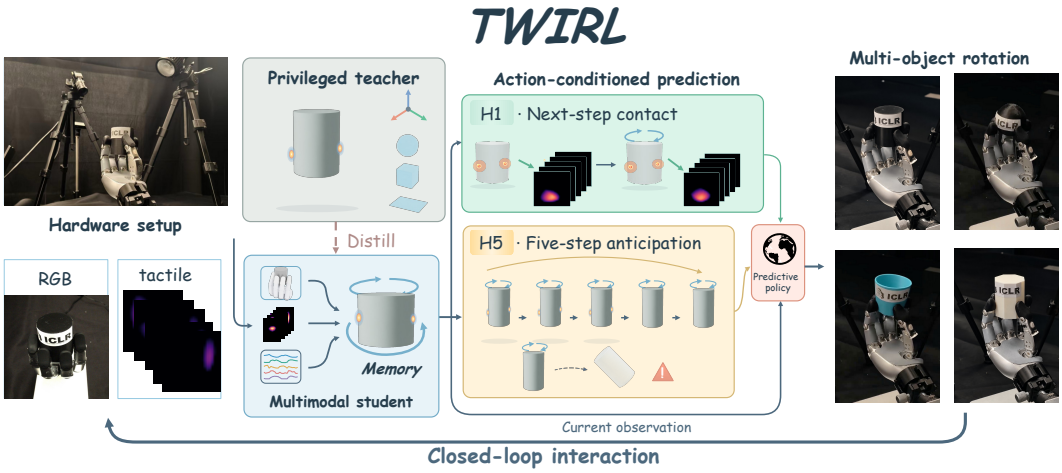

Figure 1: Persistent in-hand manipulation requires a policy to track changing contacts over time. TWIRL learns progressive object–tactile dynamics and fingertip-aligned visual fusion to make those contact changes controllable.

ABSTRACT

Dexterous in-hand manipulation requires a hand to repeatedly change contacts while sustaining object motion and avoiding drops. Vision and touch provide complementary observations, but simply combining their features leaves the consequences of actions on future contact implicit. We present TWIRL (Tactile World Model for In-hand Reinforcement Learning), a predictive visual–tactile policy built on a progressive multi-step tactile world model. TWIRL first learns immediate object and tactile dynamics, then extends this representation to anticipate contact evolution, rotation progress, and drop risk over a longer horizon. It predicts compact, action-conditioned object–tactile variables to inform control without future-image generation or online planning. A complementary fingertip-aligned fusion module uses calibrated kinematics to associate local visual features with tactile observations, grounding the policy in where contact occurs as well as how it may evolve. Persistent-rotation experiments across diverse objects in simulation show improved success over direct visual–tactile fusion, while real-world trials on multiple objects demonstrate effective transfer of the learned policy to a physical dexterous hand. Our results suggest that integrating progressive tactile dynamics prediction with fingertip-aligned visual feedback improves policy learning for sustained in-hand manipulation.

1 INTRODUCTION

Turning an object continuously inside a hand requires more than reaching a target orientation. Fingers must release and re-establish contacts while supporting the object against gravity. A small slip can change which fingers carry the load and whether the next action produces rotation or a drop. The policy must therefore anticipate contact transitions and coordinate force, geometry, and action timing

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.

across fingers throughout the rotation. Persistent in-hand rotation therefore provides a useful setting for studying how a dexterous policy represents changing contact.

Vision and touch offer complementary information. An external camera reveals object geometry and hand configuration, while fingertip sensing captures local contact that is often occluded. Recent policies combine these modalities to improve dexterous manipulation (Qi et al., 2023; Liu et al., 2025). However, combining current observations leaves two questions open: how to associate visual evidence with the finger that measures contact, and how to learn a representation of the contact changes induced by an action.

We present TWIRL, a visual–tactile policy that combines progressive object–tactile prediction with fingertip-aligned local fusion. A recurrent model first learns Immediate Tactile Dynamics (ITD), predicting the next object and tactile response. Anticipatory Contact Dynamics (ACD) then extends prediction to a five-step horizon. Both operate on compact physical targets. A complementary fusion module uses calibrated forward kinematics to query local image features at the fingertips. The resulting tokens enter the policy through residual interfaces, and actions are produced in a single forward pass without future-image generation or online action search.

Our central hypothesis is that learning contact dynamics and associating sensory evidence with physical contact locations can improve policy learning. This hypothesis has three distinct parts. Prediction losses may improve the current representation. The actor may use state-specific future information during execution. Spatial correspondence may also improve multimodal fusion. Predictive accuracy alone cannot distinguish these mechanisms. We therefore evaluate auxiliary supervision, online predictive conditioning, and fingertip alignment with separate training controls and inference interventions.

The experiments begin with cross-object simulation results and temporal rotation profiles, followed by matched supervision controls and analyses of predictive accuracy and policy reliance. The results show that training-time future supervision can shape the learned representation even when the actor does not receive a future token, while the value of direct future-token conditioning depends on the surrounding training protocol. Physical-hand tests also show transfer of the learned policy across multiple objects.

Our contributions are:

1) We propose a progressive tactile dynamics model that learns compact one- and five-step object–tactile targets and exposes them through residual policy interfaces.

2) We develop a fingertip-aligned fusion module that associates local visual evidence with force and tactile-depth observations using calibrated kinematics.

3) We evaluate TWIRL through simulation rotation and physical-hand retention trials. Matched supervision controls and policy-input interventions distinguish predictive supervision from online predictive conditioning.

2 Related Work

Dexterous in-hand manipulation. Large-scale reinforcement learning and privileged teacher–student training have enabled complex dexterous behaviors in simulation and on hardware (OpenAI et al., 2018; 2019). HORA studies continuous in-hand rotation through rapid motor adaptation from proprioceptive history (Qi et al., 2022). RotateIt adds vision and touch to generalize rotation across objects (Qi et al., 2023). Our task follows this continuous-rotation line of work. General reorientation across diverse geometries (Chen et al., 2021) and point-cloud policies for sim-to-real dexterous grasping (Qin et al., 2022) study broader object variation. The YCB benchmark formalizes object-diverse manipulation evaluation (Calli et al., 2015). EGAD provides another object-diversity benchmark for grasping analysis and reproducibility (Morrison et al., 2020). Our study asks whether explicit future-contact prediction and spatially aligned visual–tactile features improve a fixed persistent rotation protocol. We therefore compare against proprioceptive, tactile, visual, and direct multimodal policies under the same hand, horizon, and success definition. SimToolReal instead trains an object-centric policy to track goal-pose sequences across procedurally generated tools, enabling zero-shot tool use (Kedia et al., 2026). Our per-object rotation study focuses on the representation of contact during sustained motion. It does not test the unseen-tool generalization studied there.

Visual–tactile policy learning. Vision supplies global context while tactile sensors reveal local contact under occlusion. VTDexManip demonstrates visual–tactile pretraining and feedback for reinforcement-learned dexterous policies (Liu et al., 2025). Fusion spans global, attention, and per-contact designs, including Contact-Anchored Policies that condition utility models on explicit spatial contacts (Cui et al., 2026). Our calibrated force/depth and kinematic signals query local camera features to produce five finger-centered tokens, which we call fingertip-aligned local fusion. Earlier work studies vision–tactile regrasping and self-supervised contact representations (Calandra et al., 2018; Lee et al., 2019). Optical tactile sensing, tactile rendering, and differentiable contact simulation provide complementary tools (Lambeta et al., 2020; Wang et al., 2020; Si et al., 2024). PTLD uses privileged sensing for tactile-state distillation and studies online latent distillation in asymmetric actor–critic training (Chen et al., 2026), making auxiliary supervision a relevant comparison. Our current-label control tests whether future labels add value beyond it. The PTLD simulation baseline adapts its sensing choices rather than reproducing the original data pipeline.

World models and tactile prediction. World models learn action-conditioned dynamics and have been used for latent imagination, planning, and policy learning (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2023). TD-MPC uses task-oriented latent dynamics for short-horizon optimization, whereas our forecasts enter a policy without online search (Hansen et al., 2022). Auxiliary objectives can improve shared representations without becoming online plans (Jaderberg et al., 2016). Recent tactile world models improve imagined contact rollouts for planning (Higuera et al., 2026), generate aligned trajectories for augmentation and evaluation (Huang et al., 2026), predict tactile latents for real-time control (Zang et al., 2026), or jointly predict visual and tactile outcomes (Mandil & Ghalamzan-E, 2023). Thus the novelty is not the phrase "tactile world model", but our control design and evidence for persistent rotation: progressive ITD→ACD object–tactile targets, direct residual conditioning, and interventions that test use of predicted futures.

3 METHOD

Overview. Figure 2 summarizes the policy. TWIRL starts from the direct-fusion B4 policy, adds a progressively trained object–tactile world model, and then aligns local RGB evidence with each fingertip. B4 encodes RGB, five-finger force and kinematics, and five tactile depth maps into a fused observation $x_t \in \mathbb{R}^{128}$ and actor input u_t. All variants start from the same locked checkpoint. The B4 encoders remain frozen while new recurrent and fusion modules are optimized, so the deployable sensing interface stays fixed. The privileged teacher only initializes B4. Its state and latent are never inputs to the deployed policy.

3.1 IMMEDIATE TACTILE DYNAMICS (ITD)

ITD models the shortest useful consequence of an action: how the object and fingertip signals change at the next step. It converts the multimodal history into a compact belief that can receive privileged supervision during training without exposing privileged state at deployment. This short horizon supplies an immediate contact-transition target while the recurrent state retains information needed for later multi-step prediction. The separation keeps the deployable actor causal and lets the auxiliary losses shape its representation without adding privileged inputs at execution time.

A 128-dimensional posterior summarizes the current observation, a learned proprioceptive feature, the previous action, and its recurrent history. In this equation, x_t is the fused deployable observation, p_t is the proprioceptive history, ϕ_p is its learned embedding, a_{t-1} is the previous action, and h_{t-1} is the previous posterior state. The operator $\mathrm{GRU}_{\mathrm{post}}$ is the recurrent posterior encoder.

$$h_t = \mathrm{GRU}_{\mathrm{post}}\left([x_t, \phi_p(p_t), a_{t-1}], h_{t-1}\right). \tag{1}$$

The completed B4 actor produces a detached proposal $\bar{a}_t$, meaning that the proposal is used as a conditioning action without sending gradients into the completed actor. The action-conditioned transition f_1 then predicts the next latent state $\hat{h}_{t+1}$ from the current state h_t and proposal:

$$\hat{h}_{t+1} = f_1(h_t, \bar{a}_t). \tag{2}$$

Prediction heads decode a 9D object-motion target, 15D fingertip force, 15D depth summaries, and five contact logits. Object state and contact labels are privileged supervision only. The posterior must

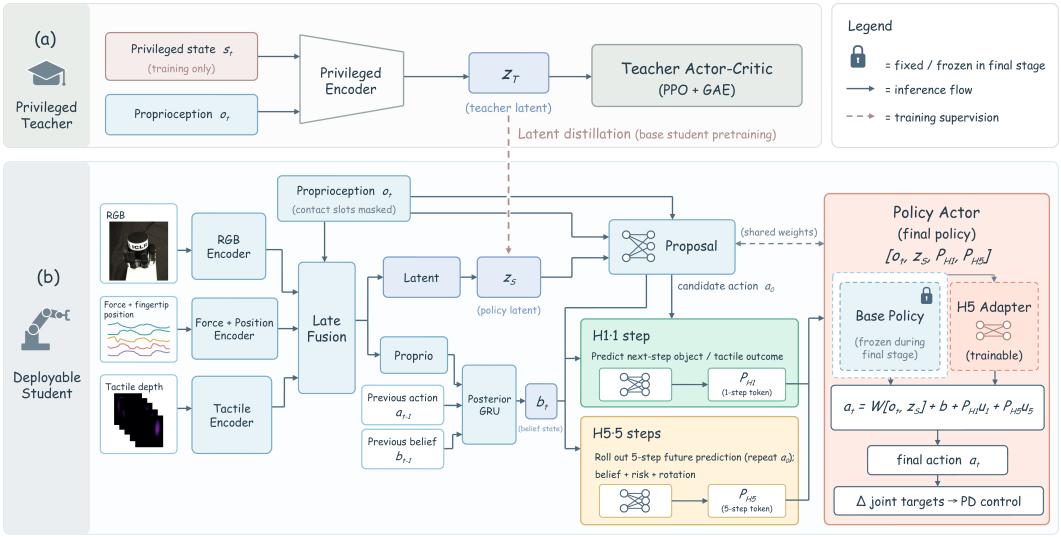

Figure 2: Teacher–student architecture and training order. The privileged teacher consumes simulator state and proprioception, and its latent is distilled to initialize the deployable B4 student. The teacher state and latent are removed at deployment, while the subsequent ITD/H1 and ACD/H5 stages use deployable sensing with training-only future supervision. Their tokens enter the frozen actor through zero-initialized residuals, and fingertip fusion queries calibrated local RGB features.

infer them from deployable observations at test time. Robust normalized regression losses supervise the continuous targets and a class-balanced binary loss supervises contacts:

$$\mathcal{L}_{\text{ITD}} = \lambda_o \ell_o + \lambda_f \ell_f + \lambda_d \ell_d + \lambda_c \ell_c. \tag{3}$$

Here ℓ_o, ℓ_f, ℓ_d, and ℓ_c are the object, force, depth, and contact losses, while the λ coefficients set their relative weights. The decoded future is embedded as a 32D anticipation token c_t^1. A zero-initialized projection sends this token to the first actor layer. Future RGB is neither reconstructed nor predicted.

3.2 Anticipatory Contact Dynamics (ACD)

Immediate prediction is insufficient for sustained rotation because contact changes accumulate over several actions. ACD therefore extends the learned transition to a causal five-step rollout and predicts task progress, contact evolution, and drop risk for the actor interface.

ACD extends the one-step representation with a separate transition for medium-horizon contact dynamics. Its transition f_5 is initialized from f_1 and recurrently unrolled for five action inputs. Here $\hat{h}^5_{t+k}$ is the predicted latent at rollout step k, a_{t+k-1} is the corresponding proposal action, and k indexes the five predicted steps:

$$\hat{h}^5_{t+k} = f_5(\hat{h}^5_{t+k-1}, a_{t+k-1}), \qquad k = 1, \ldots, 5. \tag{4}$$

Starting from the current recurrent belief, the rollout summarizes how contact may evolve under the action proposal. Recomputing this summary at every control step allows subsequent decisions to incorporate new sensory evidence as the fingers release and re-establish contact. ITD and the B4 backbone remain frozen. ACD heads predict a compact task state, finger contact evolution, medium-horizon rotation progress, and drop hazard. The five-step outputs form another 32D token c_t^5. The two predicted tokens condition the actor directly:

$$z_t^{(1)} = \sigma\left(W_u u_t + W_1 c_t^1 + W_5 c_t^5 + b\right), \qquad W_1, W_5 \text{ zero initialized.} \tag{5}$$

In this expression, $z_t^{(1)}$ is the actor's first-layer preactivation, σ is its nonlinearity, u_t is the B4 actor input, W_u is the base input projection, W_1 and W_5 project the one- and five-step tokens, and b is the bias.

The actor is causal in both training and deployment. Before choosing a_t, its ACD token is rolled out by repeating the current clamped proposal. During training, the auxiliary ACD loss teacher-forces the recorded five-action sequence, while deployment uses only the causal repeated-proposal rollout. Future executed actions therefore do not enter the action that generated them, and privileged future targets supervise the predictive representation without becoming actor inputs. The selected schedule first optimizes prediction and then enables adapter-only PPO. It uses ITD and ACD loss weights $1.0/0.5$, drop-loss weight 0.02 with positive-class weight 3, and rotation-progress weight 1.0. We evaluated a 20-step extension as a negative ablation and exclude it from the final model.

3.3 Fingertip-aligned local fusion

The predictive tokens describe what may happen, while the visual–tactile fusion module identifies where contact is occurring on the hand. It uses the calibrated fingertip geometry to retrieve local RGB evidence and complements the global predictive state under occlusion. This design preserves global context while using geometry to focus visual evidence on likely contact regions when fingers occlude the object.

Let $q_{i,t}$ be fingertip i projected by calibrated forward kinematics into the spatial RGB feature map V_t. A finger query attends to a 3×3 neighborhood $\mathcal{N}(q_{i,t})$ around that location. For feature-map index v, $\alpha_{i,v,t}$ is the attention weight, and Q and K are learned query and key projections:

$$r_{i,t} = \sum_{v \in \mathcal{N}(q_{i,t})} \alpha_{i,v,t} V_t[v], \tag{6}$$

$$\alpha_{i,v,t} = \mathrm{softmax}_v \left((Q e_{i,t})^\top K V_t[v] \right), \tag{7}$$

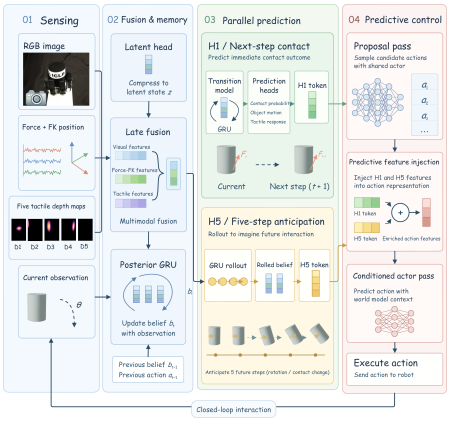

Figure 3: Action-conditioned visual–tactile world model: one-step prediction and five-step rollout condition the policy.

where $e_{i,t}$ is the finger descriptor containing force, depth, finger identity, and projected coordinates. A learned reliability score is multiplied by a hard in-frame mask. The local RGB feature, tactile feature, and reliability score form a finger-centered token. Thus $r_{i,t}$ is the pooled local RGB feature for finger i. Pooling the five tokens produces a compact fusion latent, and a zero-initialized residual sends this latent to the actor.

3.4 Training and deployment

The training schedule follows the same dependency order as the architecture: each completed parent is frozen before the next residual interface is added. Deployment then uses only causal sensing and the current action proposal, so the privileged targets used for supervision cannot leak into control.

Training begins with the direct visual–tactile backbone. ITD then learns the one-step targets, followed by the ACD stage initialized from ITD. Prediction warm-up precedes PPO optimization (Schulman et al., 2017) of the new policy interfaces. The combined model initializes from the completed predictive and local-fusion modules. Freezing the completed backbone limits each stage's trainable scope. The exact budgets, learning rates, and module-freezing schedule are given in Appendix A. The matched supervision controls in Section 4.2 use a separately specified common recurrent architecture to isolate auxiliary targets and future actor inputs.

At deployment, the policy encodes the current sensors, updates its recurrent state, generates a proposal, and computes the one- and five-step tokens. The five-step rollout repeats the proposal because future executed actions are unavailable. The actor combines these tokens with the aligned sensory features to produce the next joint-position increment. Object state, contact labels, and physical parameters are used as supervision during training and as diagnostic references during evaluation. They are not direct actor inputs.

4 EXPERIMENTS

We organize the experiments around four questions.

Q1. How does TWIRL compare with sensing baselines across objects, and does rotation persist over longer horizons?

Q2. What do auxiliary supervision and future actor inputs contribute under matched training conditions?

Q3. How accurately does the model predict contact dynamics, and does the policy use state-specific future information?

Q4. Can the learned policy be deployed on a real dexterous hand for sustained in-hand rotation?

Common setup. We use the Revo3 right hand in Isaac Lab under full gravity. An episode is successful only when the object survives the fixed horizon and completes a signed world-z turn. The same deployable inputs, reward, termination rule, and evaluation protocol are used across the matched rows. Object comparisons and matched supervision controls use two training seeds, each evaluated with three evaluation seeds. Reported spreads are reproduced from the supplied summaries. Historical fixed-checkpoint diagnostic tables instead report mean $\pm$ sample standard deviation across three evaluation batches of 1,000 episodes per checkpoint. Those spreads describe evaluation variability for a fixed policy. We also measure windowed rotation progress and late-episode angular velocity. All canonical stages fit on a single NVIDIA RTX 5090 with 32 GB of memory.

Baselines and comparison scope. HORA uses proprioceptive history, PTLD uses force, fingertip kinematics, and tactile depth, and B4 concatenates RGB with these tactile inputs. Because the Revo3 force and tactile sensors are not identical to the hardware used in the original PTLD and RotateIt studies, PTLD and RotateIt are matched adaptations to our available sensor suite rather than exact reproductions. All rows use the same Revo3 observations, task protocol, and evaluation seeds. The adaptation details are in the appendix. Historical reference checkpoints and the new matched training controls are reported separately. The latter fix initialization, additional transitions, optimizer schedule, and terminal-checkpoint selection.

4.1 SIMULATION PERFORMANCE AND SUSTAINED ROTATION

Table 1: Simulation rotation success (%) across four objects. The final column is the arithmetic mean across objects. Each geometry uses a separately trained policy.

Method	Cylinder	Ball	Flowerpot	Octagonal	Avg. success
HORA	12.73 ± 0.15	37.27 ± 1.40	7.07 ± 1.23	19.57 ± 0.87	19.16
PTLD	29.80 ± 2.25	47.07 ± 1.15	19.63 ± 1.54	30.01 ± 1.27	31.63
RotateIt	29.33 ± 1.75	46.59 ± 1.31	16.53 ± 1.07	27.83 ± 1.45	30.07
Direct fusion (B4)	31.43 ± 1.72	50.06 ± 1.68	22.56 ± 1.81	29.54 ± 2.03	33.40
TWIRL	$\mathbf{43.57 \pm 3.66}$	$\mathbf{60.60 \pm 2.20}$	$\mathbf{30.67 \pm 2.67}$	$\mathbf{42.54 \pm 3.17}$	**44.35**

Setup. Table 1 reports Cylinder, Ball, Flowerpot, and Octagonal rotation success. Each geometry uses a separate grasp and rotation policy and is evaluated on new episodes of that geometry to assess applicability across objects. Each reported mean averages two training seeds, each evaluated with three evaluation seeds. The final column is the arithmetic mean of the four object means, not an additional evaluation, and the reported $\pm$ values are reproduced from the supplied summaries. Additional object and sensing comparisons appear in Appendices L and M. This aggregation captures variation from both optimization and evaluation seeds, while the final object average summarizes the four separately trained geometry policies. The resulting table compares the same sensing interface and success definition across every listed method.

Results. TWIRL has the highest reported mean on all four objects. Its margins over the strongest available baseline are 12.14, 10.54, 8.11, and 12.53 percentage points on Cylinder, Ball, Flowerpot, and Octagonal, respectively. These are descriptive differences in the supplied aggregates. Appendix M

reports the extended Cylinder sensing comparison. The ranking is reported for the four objects with complete matched aggregates. The consistent object-wise ordering indicates that the improvement is not driven by a single favorable geometry, while the final average preserves this pattern.

Sustained motion and failures. A one-turn threshold does not guarantee continued rotation. The fresh long-horizon audit in Table 6 therefore reports both unconditional and survivor-only trajectories. It shows a retention advantage for the historical A checkpoint, while the comparison has unequal budgets and does not evaluate the new supervision controls. Stall diagnostics appear in Appendix C.

The full horizon-by-horizon table is reported in Appendix E. Together, these complementary views distinguish completing one signed turn from maintaining useful angular motion after the first successful event. They also provide a temporal context for interpreting the cross-object averages in Table 1. This temporal breakdown further reveals whether successful turns remain stable as contacts are repeatedly exchanged and small disturbances accumulate. It links the aggregate success metric to the underlying contact sequence and highlights when rotation slows, stalls, or recovers after a disturbance.

4.2 Supervision and Policy-Input Ablations

Architecture contributions. The historical component ladder in Appendix N motivates the design, but it changes training schedule and budget across rows. We therefore do not present it as a causal component ablation. The same-capacity four-cell prediction–alignment factorial was not completed and is excluded from the quantitative claims. The mechanistic conclusion is restricted to the matched supervision screen below and to the explicit inference interventions in Section 4.3.

Matched supervision setup. The completed R0–R4 experiment on Cylinder in simulation asks whether gains come from auxiliary targets or from exposing their predictions to the actor (Table 2). Every row starts from the same B4 checkpoint and uses the same additional training budget, optimizer schedule, and terminal-checkpoint rule. All retain aligned fusion and the current recurrent route. Each row averages six evaluations (two training seeds $\times$ three evaluation seeds). Speed is mean angular velocity over 10–20 seconds, with zero assigned after termination. R0 disables auxiliary supervision. R1 uses current labels. R2 uses future labels without a future actor token. R3 adds a learned constant token. R4 adds a dynamic future token. Labels cover task, contact, and physics targets. The modules and optimizer membership match, although their gradient-active capacity differs. Protocol details are in Appendix O, and R4 is the dynamic-token condition within this common protocol rather than the final TWIRL training protocol used for Table 1.

Table 2: Matched controls on simulated Cylinder. Mean $\pm$ reported spread.

Row	Success (%)	Speed (rad/s)
R0	54.83 ± 4.16	0.419 ± 0.033
R1	67.33 ± 2.41	0.518 ± 0.024
R2	68.60 ± 2.69	0.552 ± 0.023
R3	61.13 ± 3.05	0.453 ± 0.015
R4	53.20 ± 1.92	0.392 ± 0.010

Matched supervision results. Table 2 shows that auxiliary supervision improves the matched screen. Future labels without a future actor token (R2) reach 68.60%, only 1.27 points above current-label supervision (R1), whereas the dynamic future-token condition (R4) reaches 53.20%. Thus future supervision helps in this screen through a training-time representation signal: R2 receives privileged future labels during training, but its actor never receives a future token. Direct conditioning is not uniformly useful in the matched screen. Under the canonical recipe, the bridge in Appendix O finds that zeroing H5 lowers success from 43.57% to 38.53%, so we retain H5 in the final method. This indicates that the causal predicted token can provide an additional control interface under the final recipe, while the main matched-screen gain is attributable to training-time auxiliary supervision rather than ground-truth future information at action time.

One-step prediction and training schedule. A separate historical terminal comparison and a progressive-versus-joint schedule screen are reported in the appendix, where freezing, update scope, and learning-rate changes remain coupled. The screen therefore does not establish a general advantage for progressive training.

4.3 Predictive Accuracy and Policy Reliance

Prediction protocol. We evaluate one-step object–tactile prediction and five-step task/contact forecasts on held-out simulation trajectories. Probe fitting, calibration, and testing use separate evaluation splits of 300 episodes each. One-step accuracy is measured against a persistence baseline that copies the current target into the next step. We report MSE skill as $100(1 - \mathrm{MSE}_{\mathrm{model}}/\mathrm{MSE}_{\mathrm{persistence}})$, so positive values indicate improvement over persistence. The anchored prediction protocol isolates the predicted change from current-state reconstruction. The simulator-state anchor is diagnostic only and is not an actor input (Appendix D).

Table 3: Prediction accuracy metrics.

Target	Value
One-step MSE skill (%)	
Force	$+29.04$
Object	$+1.57$
One-step contact prediction	
Future contact AP / F1	$0.986/0.949$
Contact-change AP	0.339
Event prevalence	0.052
Contact-change F1	0.378
Five-step R^2 (deployable)	
Task	0.074
Contact	0.167

Forecast quality. Table 3 shows a 29.04% reduction in one-step force MSE relative to persistence and a smaller improvement for object prediction. The full target-wise regression audit is in Appendix D. Future-contact macro AP/F1 is $0.986/0.949$. Since contact states often persist, we also evaluate contact changes: AP is 0.339 at event prevalence 0.052, with F1 0.378. This provides evidence of contact-transition signal, while leaving substantial room for improvement in event detection. For five-step prediction, we repeat the current action proposal, matching the information available when the policy chooses an action. Task and contact forecasts have positive but modest R^2, indicating uneven predictive quality across targets. Replaying the realized future actions improves these scores, but those actions are unavailable at decision time. Replay and drop-risk diagnostics are therefore reported separately in Appendix D. All displayed MSE skill, AP/F1, and deployable R^2 values are positive and higher values are better. Event prevalence is the empirical fraction of contact changes rather than a predictive score. AP is average precision and the contact-state scores use macro averaging. The full regression audit is in Appendix D.

Policy reliance. To distinguish predictive accuracy from its use in control, we hold the checkpoint and evaluator fixed and replace the five-step actor token with its calibration mean, a shuffled token, or zero (Appendix P.1). Mean and shuffled tokens change success little, while zeroing has a larger effect because it removes both the token's mean contribution and its state-dependent variation. This sensitivity to the actor interface, together with the matched supervision results, does not establish a success benefit from state-specific future information.

4.4 Real-World Deployment

Hardware and control loop. We deploy the learned policy on a Revo3 dexterous hand with 21 degrees of freedom. An external Hikrobot RGB camera provides visual observations, which the policy combines with proprioception and fingertip tactile inputs. The policy operates at 20 Hz, updating the joint-position command every 50 ms. Each update encodes the observations, updates the recurrent state, and computes the action-conditioned tactile predictions before producing the control command. Prediction conditions the policy directly, without online trajectory optimization or future-image generation.

Table 4: Real-world retention.

Object	HORA	TWIRL
Cylinder	3/10 = 30%	8/10 = 80%
Ball	4/10 = 40%	10/10 = 100%
Flowerpot	2/10 = 20%	7/10 = 70%
Octagonal	3/10 = 30%	9/10 = 90%

Camera calibration and sensory alignment. Deployment requires camera intrinsics and the extrinsic transformation between the camera and hand coordinate frames. This calibration maps fingertip positions obtained from forward kinematics into the RGB image, allowing each fingertip's tactile observation to query the corresponding local visual features. The hand–camera registration connects the spatial fusion mechanism used in simulation to the physical sensor setup. It is distinct from learning an object-specific visual correspondence online.

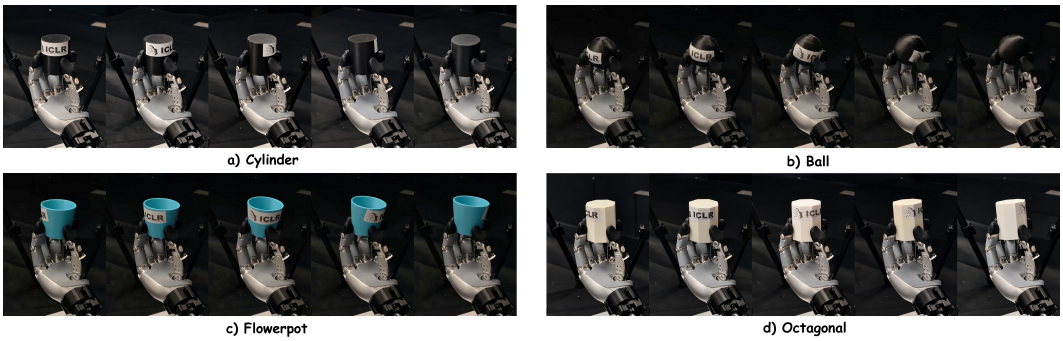

Figure 4: Qualitative real-hand sequences for Cylinder, Ball, Flowerpot, and Octagonal. The quantitative hardware study reports 20-second retention for all four objects in Table 4.

Evaluation protocol. We evaluate HORA and TWIRL using Cylinder, Ball, Flowerpot, and Octagonal, with ten trials per method and object. The policy executes the in-hand rotation task, while the reported endpoint measures retention: the object must remain in the hand without dropping for the full 20 seconds. Unlike simulation success, this criterion does not require a completed turn. Table 4 reports retained-trial counts and percentages, so each entry directly records the outcome of ten physical trials. The table uses a no-drop retention criterion over 20 seconds and imposes no rotation threshold. This endpoint isolates physical retention from the stricter simulation criterion and makes the counts directly comparable across methods and objects.

Transfer results. TWIRL retains Cylinder, Ball, Flowerpot, and Octagonal in 8/10, 10/10, 7/10, and 9/10 trials, respectively, compared with HORA's 3/10, 4/10, 2/10, and 3/10. The observed retention advantage is consistent across these objects, with gains on both compact and elongated geometries, indicating useful contact coordination across varied object shapes. The four objects span different silhouettes and contact patterns, so consistent retention suggests that the learned coordination transfers beyond a single favorable shape. The results validate transfer of the learned policy to the physical hand. A qualitative sequence is shown in Fig. 4.

5 Discussion and Conclusion

TWIRL combines progressive tactile dynamics prediction with fingertip-aligned visual feedback and achieves the strongest reported mean on Cylinder, Ball, Flowerpot, and Octagonal. Simulation demonstrates sustained in-hand rotation, and real-hand trials validate transfer of the learned policy across objects. Matched controls further show how training-time auxiliary supervision shapes the predictive representation.

Limitations and future work. The two training seeds provide limited evidence about training-population variability. More object geometries are needed to test whether the observed gains persist beyond the four evaluated objects, and a single policy trained to cover several objects would provide a stronger generalization test. Such studies would clarify whether the learned representation remains stable as geometry, contact layout, and initialization vary across broader object sets. The tactile hardware and calibration pipeline also merit higher-resolution sensing and improved contact fidelity. Reproducibility details are recorded in Appendix A.

Reproducibility Statement

Section 3 specifies the model and interfaces, and Section 4 defines the task, metrics, and seed aggregation. Appendix A records the uniform 100M-step stages and the exact-rerun checklist. The prediction and matched-control protocols are in Appendices D and O. Reported aggregates use two training seeds and three evaluation seeds, with fixed-checkpoint diagnostics kept separate. The supplementary archive contains the world-model source, installation metadata, smoke test, and qualitative real-hand video.

REFERENCES

Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H. Adelson, and Sergey Levine. More Than a Feeling: Learning to Grasp and Regrasp using Vision and Touch. *arXiv preprint arXiv:1805.11085*, 2018. URL https://arxiv.org/abs/1805.11085.

Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols. *arXiv preprint arXiv:1502.03143*, 2015. URL https://arxiv.org/abs/1502.03143.

Rosy Chen, Mustafa Mukadam, Michael Kaess, Tingfan Wu, Francois R. Hogan, Jitendra Malik, and Akash Sharma. PTLD: Sim-to-real privileged tactile latent distillation for dexterous manipulation. *arXiv preprint arXiv:2603.04531*, 2026. URL https://arxiv.org/abs/2603.04531.

Tao Chen, Jie Xu, and Pulkit Agrawal. A System for General In-Hand Object Re-Orientation. *arXiv preprint arXiv:2111.03043*, 2021. URL https://arxiv.org/abs/2111.03043.

Zichen Jeff Cui, Omar Rayyan, Haritheja Etukuru, Bowen Tan, Zavier Andrianarivo, Zicheng Teng, Yihang Zhou, Krish Mehta, Nicholas Wojno, Kevin Yuanbo Wu, et al. Contact-anchored policies: Contact conditioning creates strong robot utility models. *arXiv preprint arXiv:2602.09017*, 2026. URL https://arxiv.org/abs/2602.09017.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019. URL https://arxiv.org/abs/1912.01603.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Domains through World Models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal Difference Learning for Model Predictive Control. *arXiv preprint arXiv:2203.04955*, 2022. URL https://arxiv.org/abs/2203.04955.

Carolina Higuera, Sergio Arnaud, Byron Boots, Mustafa Mukadam, Francois Robert Hogan, and Franziska Meier. Visuo-tactile world models. *arXiv preprint arXiv:2602.06001*, 2026. URL https://arxiv.org/abs/2602.06001.

Yunao Huang, Shiyu Sang, Haotao Lu, Suting Ni, Shijie Wu, Ziyang Guo, Ye Shi, and Jingya Wang. ViTacWorld: Scaling visuo-tactile world models for contact-rich robot manipulation. *arXiv preprint arXiv:2607.22530*, 2026. URL https://arxiv.org/abs/2607.22530.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement Learning with Unsupervised Auxiliary Tasks. *arXiv preprint arXiv:1611.05397*, 2016. URL https://arxiv.org/abs/1611.05397.

Kushal Kedia, Tyler Ga Wei Lum, Jeannette Bohg, and C. Karen Liu. SimToolReal: An object-centric policy for zero-shot dexterous tool manipulation. *arXiv preprint arXiv:2602.16863*, 2026. URL https://arxiv.org/abs/2602.16863.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. DIGIT: A Novel Design for a Low-Cost Compact High-Resolution Tactile Sensor with Application to In-Hand Manipulation. *arXiv preprint arXiv:2005.14679*, 2020. URL https://arxiv.org/abs/2005.14679.

Michelle A. Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks. *arXiv preprint arXiv:1907.13098*, 2019. URL https://arxiv.org/abs/1907.13098.

Qingtao Liu, Yu Cui, Zhengnan Sun, Gaofeng Li, Jiming Chen, and Qi Ye. VTDexManip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e19b6f65791e350347bcff8a3955cb5b-Abstract-Conference.html.

Willow Mandil and Amir Ghalamzan-E. Multi-modal world model for physical robot interactions: Simultaneous visual and tactile predictions for enhanced accuracy. *arXiv preprint arXiv:2304.11193*, 2023. URL https://arxiv.org/abs/2304.11193.

Douglas Morrison, Peter Corke, and Jürgen Leitner. EGAD! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *arXiv preprint arXiv:2003.01314*, 2020. URL https://arxiv.org/abs/2003.01314.

OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafał Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *arXiv preprint arXiv:1808.00177*, 2018. URL https://arxiv.org/abs/1808.00177.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019. URL https://arxiv.org/abs/1910.07113.

Haozhi Qi, Ashish Kumar, Roberto Calandra, Yi Ma, and Jitendra Malik. In-hand object rotation via rapid motor adaptation. *arXiv preprint arXiv:2210.04887*, 2022. URL https://arxiv.org/abs/2210.04887.

Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. In *Conference on Robot Learning*, 2023. URL https://arxiv.org/abs/2309.09979.

Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Hao Su, and Xiaolong Wang. DexPoint: Generalizable Point Cloud Reinforcement Learning for Sim-to-Real Dexterous Manipulation. *arXiv preprint arXiv:2211.09423*, 2022. URL https://arxiv.org/abs/2211.09423.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zilin Si, Gu Zhang, Qingwei Ben, Branden Romero, Zhou Xian, Chao Liu, and Chuang Gan. DIFFTACTILE: A physics-based differentiable tactile simulator for contact-rich robotic manipulation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1cc70be9fb6a83bc46cf4ac21a91e0b0-Abstract-Conference.html.

Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. TACTO: A Fast, Flexible, and Open-source Simulator for High-Resolution Vision-based Tactile Sensors. *arXiv preprint arXiv:2012.08456*, 2020. URL https://arxiv.org/abs/2012.08456.

Yujie Zang, Yuhang Zheng, Xian Nie, Yupeng Zheng, Shuai Tian, Songen Gu, Chen Gao, Zining Wang, Shuicheng Yan, and Wenchao Ding. TacForeSight: Force-guided tactile world model for contact-rich manipulation. *arXiv preprint arXiv:2606.11184*, 2026. URL https://arxiv.org/abs/2606.11184.

AI USE STATEMENT

OpenAI Codex was used for research ideation and execution support, including discussing experimental framing, organizing supplementary artifacts, and checking supporting code. It was also used to polish manuscript text. The authors verified all scientific claims, equations, citations, figures, quantitative results, and implementation details, and take responsibility for the final submission.

A Implementation Details

The recovered recipe below describes the diagnostic checkpoints. Object, sensing, and matched-supervision comparisons average two training seeds and three evaluation seeds. Fixed-checkpoint diagnostic protocols and checkpoint-specific settings are reported separately.

The deployed observation contains proprioception, one RGB frame, five force histories, and five tactile depth maps. For the revised training accounting, each canonical stage uses a uniform 100M environment-step budget: B4 pretraining, ITD continuation, ACD continuation, independent contact-fusion pretraining, and the combined stage. Posterior and transition width are 128 and each anticipation token has width 32. ITD and ACD use 512 parallel environments. The actor adapter is enabled only after prediction warm-up. The separate R0–R4 matched screen is a short common continuation from B4 rather than an additional canonical stage. The object and sensing aggregates use two training seeds. This recovered configuration describes the shared training recipe rather than a single aggregate entry. The selected A setting uses world-model learning rate 10^{-4}, critic/interface rates $10^{-5}/5 \times 10^{-6}$, rollout length 16, two PPO epochs, ratio clip 0.1, and KL threshold 0.005. These specialized multiscale settings override generic PPO fields in the saved YAML. Action distillation is disabled. The optional five-dimensional physics-consistency head predicts randomized friction, mass, and object-frame center of mass against privileged simulator targets during training. It is not an actor input at deployment. The recovered A config and completion metadata are archived with this revision. The exact-rerun checklist covers reward coefficients, target transforms, normalization fitting/freezing, randomization ranges, PPO schedules, optimizer groups, contact-label thresholds, evaluator resets, checkpoint hashes, and seed mapping so that the released configuration can be checked against the runtime overrides.

B Evaluation Details

For each policy checkpoint, the evaluator runs 1000 episodes for each of three evaluation seeds. The matched R0–R4 comparison applies this evaluator to two independently trained checkpoints per condition. Its aggregate mean therefore covers six evaluation runs. The success indicator is

$$\mathbf{1}[T_{\text{episode}} = 400]\,\mathbf{1}[\Delta\theta_z \geq 2\pi]. \tag{8}$$

For every method we also save termination cause, episode duration, time-to-drop, signed and absolute rotation, mean angular velocity, direction consistency, and per-step contact traces. Independent policy-training seeds, when complete, are aggregated above the episode level rather than being mixed with evaluation-environment seeds.

C Temporal Motion Reanalysis

We reanalyse archived held-out diagnostic traces for A and D, each containing 300 episodes, using signed per-step rotation increments. These are descriptive post-hoc results on existing traces, not new evaluations of the headline checkpoints under the main evaluation settings. The traces store float16 increments. We report unconditional five-second progress with zero additional rotation after a drop, and separately report survivors and historically successful episodes.

For the 161 successful A episodes, mean turns in windows 0–5, 5–10, 10–15, and 15–20 seconds are $0.466/0.487/0.489/0.491$. For D's 153 successful episodes they are $0.440/0.449/0.454/0.446$. Mean second-half angular velocities are 0.616 and 0.565 rad/s, respectively. Neither set contains a successful episode with nonpositive final-five-second progress or second-half mean velocity below 0.1 rad/s. Thus these traces do not exhibit an aggregate early-burst-then-hold pattern, despite the metric permitting it.

For stall diagnostics we smooth signed velocity with a trailing one-second mean, discard the initial incomplete smoothing window, and measure contiguous near-stationary intervals ($|\bar{\omega}_z| < \epsilon$), separately from non-forward intervals ($\bar{\omega}_z < \epsilon$). We examine $\epsilon \in \{0.05, 0.1, 0.2\}$ rad/s without changing historical success. At $\epsilon = 0.1$, the mean longest near-stationary interval among successful episodes is 0.068 s for A and 0.081 s for D. Duration counts eligible smoothed endpoints and excludes time after a drop. These observations do not establish improvement over B4 or constant-token policies, which

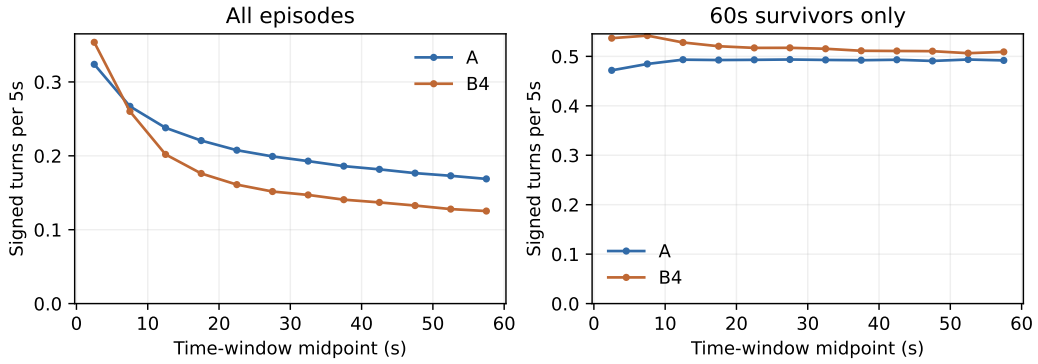

Figure 5: Signed rotation progress in fresh 60-second rollouts of historical A and B4, 3000 episodes each. Left: all episodes, with zero progress after termination. Right: episodes surviving the full horizon. The survivor sets differ between policies. Both retain positive windowed motion. A has greater unconditional progress despite slower survivor-only rotation.

Table 5: Physics-consistency ablation for the WM-only ACD-V1.1 control.

Variant	Success (%)	Δ (pp)
Physics loss on	39.90 ± 3.90	–
Physics loss off	38.33 ± 3.85	-1.57

lack matched traces here, or persistence beyond 20 seconds. Future confirmatory evaluation must collect all conditions under one temporal protocol.

D Predictive Diagnostic Protocol

Prediction diagnostics use separate evaluation seeds for linear-probe fitting, threshold selection and drop-risk calibration, and untouched testing. Each split contains 300 episodes. Reported ITD intervals use 2,000 episode-cluster bootstrap replicates. The anchored ITD forecast $x_t + (\hat{x}_{t+1} - \hat{x}_t)$ isolates transition fidelity from reconstruction of the current state. Simulator object state is used only for this diagnostic anchor and is never exposed to the actor.

The complete one-step regression audit gives object, force, and tactile-depth MSE skills of $+1.57\%$, $+29.04\%$, and -4.78%, respectively, relative to persistence. Their conditional 95% intervals are $[-0.04, 3.08]$, $[27.61, 30.55]$, and $[-7.32, -2.20]$ percentage points. Thus regression quality is target-dependent. The force improvement does not extend to every prediction head.

For one-step future-contact prediction, the held-out macro AP/F1 is $0.986/0.949$. Contact states can persist without a transition, so we report contact-change detection separately: AP 0.339, event prevalence 0.052, and F1 0.378. Contact onset and offset AP are 0.502 and 0.134, respectively, showing asymmetric event difficulty. Event prevalence provides a class-balance reference, not the performance of a trained competing model. These scores do not establish how the actor uses the predictions.

For ACD we replay the learned transition with the realized five-action sequence after the episode has been collected. This action-aligned replay cannot enter the action that generated it. It measures whether the learned transition is conditionally predictive. The deployed policy instead repeats its current proposal five times. On primary model A, action alignment changes task/contact R^2 from $0.074/0.167$ to $0.269/0.252$. Drop ranking has AUROC 0.742 and AUPRC 0.0386. The drop head reaches calibrated ECE 0.00069, but detects only 43.4% of drop episodes with mean lead time 74.6 ms and approximately 9 false-alarm events per minute.

The optional belief-injection model B illustrates why control and prediction must be reported separately. Its held-out object/force/depth anchored RMSE is $0.03980/0.04299/0.04897$, versus $0.02449/0.03671/0.02718$ for A. We therefore select A even though B has a slightly higher development control point estimate.

Table 6: Fresh-rollout robustness of frozen historical A and B4.

Horizon	Policy	Success (%)	Positive every 5s (%)	Late speed
20s	A	42.37 ± 4.05	42.87 ± 3.97	0.287 ± 0.021
20s	B4	29.20 ± 1.25	30.57 ± 1.56	0.237 ± 0.008
40s	A	37.50 ± 4.98	37.43 ± 5.01	0.248 ± 0.026
40s	B4	24.60 ± 1.05	23.90 ± 0.95	0.186 ± 0.008
60s	A	34.03 ± 5.17	33.73 ± 5.27	0.226 ± 0.029
60s	B4	22.90 ± 1.74	21.93 ± 1.79	0.170 ± 0.013

Table 7: Legacy architecture diagnostics. These runs motivated progressive warm-starting and are not mixed into the confirmatory proof matrix.

Variant	Horizon	Mean success (%)
Structured fusion only	none	29.30
Structured fusion + dynamics	1	27.57
Structured residual only	none	25.30
Structured residual + WM	1	27.97
Joint 1/5-step predictor (final)	1/5	29.77
Joint 1/5/20-step predictor (final)	1/5/20	29.83
Joint 1/5/20-step @200M†	1/5/20	35.40

The physics table reports mean $\pm$ sample SD across three evaluation batches of 1,000 episodes for each fixed checkpoint. Both variants use the same 100M-stage recipe and terminal selection. Only the physics-loss weight differs. The point-estimate decrease alone does not establish a reproducible benefit.

E Fresh Rollout Robustness

These values are mean $\pm$ sample SD across three evaluation batches of 1,000 episodes for each fixed historical checkpoint. Success requires survival and one signed turn. Positive-every-5s requires survival and positive progress in every window. Late speed is measured over the second half of the horizon and is zero after termination. These rows are separate from R0–R4 and from the continuation diagnostics.

F Historical World-Model Controls

The historical matched suite compares B4 continuation, a recurrent no-prediction control, ITD-only, and progressive or joint predictive training. It is distinct from the completed R0–R4 supervision experiment, which fixes the current recurrent route and varies auxiliary labels and future actor input. Legacy runs that failed preflight or used incompatible evaluators are not entered as numerical results.

The completed bug-fixed WM-only terminal diagnostics reached 23.50% aggregate success for progressive ITD$\rightarrow$ACD and 20.27% for joint ITD+ACD across three evaluation seeds. Both used an aggressive full-actor PPO schedule. These negative terminal results limit the tested training recipe. Attributing the collapse to the schedule is a hypothesis, not a reason to exclude the results or a proven explanation.

G Additional Architecture Diagnostics

The 200M dagger row was selected on a single-seed exploratory screen and is not an unbiased test result.

The legacy runs in Table 7 motivated progressive warm starts. In particular, the strong intermediate joint 1/5/20-step checkpoint followed by a weak final policy indicates an interaction between prediction training and PPO, rather than evidence that a 20-step horizon is uniformly beneficial. These runs are not mixed with the matched proof matrix.

Table 8: Local convergence diagnosis on the cylinder task. Every entry uses the canonical 100M-stage budget, one policy-training seed, and the same 300-episode held-out evaluator. "Best" and "latest" refer to points within that diagnostic screen. These entries do not define the paper checkpoint rule.

Method	Best diagnostic	Latest diagnostic
Aligned Contact fusion	30.67	27.00
Capacity-matched nonaligned fusion	27.00	25.00
ITD-only	29.33	15.67
GRU, no prediction	12.33	6.67
RotateIt	23.67	29.33
Matched Full, final stage	13.67	11.67

H Development Checkpoint Screens

Each screen uses the canonical 100M-stage budget, one policy-training seed, and a 300-episode held-out evaluator. Success requires 400-step survival and at least 2π signed world-z rotation. These screens diagnose convergence and are excluded from confirmatory means and statistical claims.

Table 8 records the local convergence screen used to decide whether long DLC runs should be extended. It also exposes why training reward is not used for policy selection: ITD and GRU controls are strongly non-monotonic in closed-loop success, and the latest screened checkpoints can be much worse than earlier ones. All confirmatory rows nevertheless retain the predeclared terminal-checkpoint rule.

I RotateIt Adaptation Details

The published RotateIt policy fuses a history of foreground object depth, discretized fingertip contact locations, joint positions, and previous actions with a 32-dimensional, two-layer, two-head Transformer. Our matched adaptation retains this global temporal-fusion topology, but maps the available Revo3 modalities to RGB, five-finger force/center histories, and fingertip depth. It predicts the fixed-cylinder teacher's 8-dimensional privileged latent rather than RotateIt's 40-dimensional shape–physics latent. Joint, force/center, and reconstructed previous-action streams are temporal. Because the current environment exports only one RGB and fingertip-depth frame, those two features are broadcast across the three temporal tokens. We therefore label the row "RotateIt matched adaptation" and do not claim reproduction of the official sensors, training set, or reported RotateIt numbers. The PTLD row is likewise rewritten for the Revo3 force, fingertip-kinematic, and tactile-depth suite. Neither row is an exact sensor-level reproduction of the original system.

J Qualitative Rollouts

Qualitative synchronized RGB, tactile-depth, force, contact, and rotation traces are not used as quantitative evidence in this submission. The policy camera and any presentation camera must be kept distinct in future release artifacts. Projected fingertip locations should be overlaid on the policy image when such examples are added. This omission does not affect the numerical results or claims in the paper.

K Object and Leakage Audit

The main-paper object table reports Cylinder, Ball, Flowerpot, and Octagonal summaries. Each geometry uses an object-specific Stage-1 grasp policy and a separate rotation policy, so the held-out episodes are new episodes for a trained geometry rather than unseen-object tests. Uncompleted scanned-object assets are not included in the table or in any generalization claim. The first canonical Sapota Stage-1 migration failed its 300-episode qualification screen, so we reject that checkpoint and do not propagate it into B4 or TWIRL.

Table 9: Additional object comparisons: mean success (%).

Object	HORA	PTLD	TWIRL
Cylinder	12.73	29.80	43.57
Ball	37.27	47.07	60.60
Flowerpot	7.07	19.63	30.67
Octagonal	19.57	30.01	42.54

Table 10: Cylinder sensing comparison.

Method	Success (%)
HORA	12.73 ± 0.15
RotateIt	29.33 ± 1.75
Force + FK	26.43 ± 1.67
Force + FK + depth	29.80 ± 2.25
RGB only	23.37 ± 2.27
Direct fusion (B4)	31.43 ± 1.72
TWIRL	43.57 ± 3.66

L Additional Object Comparisons

Table 9 summarizes the object comparisons using the same two-training-seed, three-evaluation-seed aggregation as Table 1. Cylinder, Ball, Flowerpot, and Octagonal repeat the corresponding main-table means for reference. The tactile baseline is PTLD, including force, fingertip kinematics, and tactile depth. Each mean averages two training seeds times three evaluation seeds, and each geometry uses a separately trained policy.

Strawberry and Duck both show promising results in the available screening runs. The Strawberry teacher screen has mean 45.57% across three evaluation seeds, while Duck is retained as a separate auxiliary screen. These runs use only one training seed, so they do not meet the canonical two-training-seed protocol and are omitted from the main text. Archived teacher and HORA/PTLD/TWIRL screens remain separate diagnostics and do not define the canonical 100M-stage recipe.

M Reference Sensing Results

Table 10 expands the Cylinder comparison with force/FK and RGB-only sensing variants. Each mean averages two training seeds, each with three evaluation seeds. Shared methods report the same results as Table 1. Force/FK/depth corresponds to PTLD. This sensing comparison does not isolate the effects of individual TWIRL components under an equal-budget architecture factorial. The reported spreads are reproduced as supplied. This is a sensing comparison, not an equal-budget component ablation, and it repeats shared entries from Table 1. Force/FK/depth denotes PTLD.

N Development Component Comparisons

The dagger row is an exploratory checkpoint screen rather than a formal 1,000-episode comparison. Based on the displayed rounded means, B exceeds A by only 0.16 points. A is selected because it preserves the predictive belief. The matched B4/fusion/world-model factorial remains future work.

The staged ladder changes training schedule and budget as well as modules. It motivates the four-cell architecture experiment but cannot substitute for it. The belief-injection variant has a slightly higher control point estimate and poorer held-out anchored prediction than the actor-only variant.

O Matched Supervision Protocol and Continuation

R0–R4 are evaluated on Cylinder in simulation, with 6,000 episodes per condition: two training seeds, each evaluated with three evaluation seeds and 1,000 episodes per evaluation. They share the B4 initialization, aligned fusion, a trainable current recurrent route, 30,007,296 additional transitions,

Table 11: Component ablation on the cylinder task. Fingertip-aligned visual–tactile fusion is a proposed component. Rows come from the indicated development packages and are not yet an equal-budget factorial comparison. "Screen" uses 200 rather than 1000 episodes per seed.

Variant	ITD	ACD	Aligned V–T	Belief route	Protocol	Success
Paired B4 for ITD	–	–	–	–	ITD eval.	30.60
Fusion only	–	–	✓	–	separate eval.	32.00
ITD (1-step)	✓	–	–	–	ITD eval.	34.13
ITD→ACD†	✓	✓	–	–	screen	39.83
Combined A (primary)	✓	✓	✓	–	A/B/C eval.	**43.57**
Combined B	✓	✓	✓	✓	A/B/C eval.	43.73

Table 12: Canonical-recipe bridge for the H5 actor token on Cylinder. The only intervention is to zero the H5 token during both training and deployment.

H5 actor token	Success (%)
Enabled	43.57 ± 3.66
Zeroed in training and deployment	38.53

117,216 Adam updates, and terminal selection. R1 uses current labels through the same forecast-head graph and future-valid mask as R2. The shared protocol uses finite-horizon terminal values and one-step recurrent gradients, and is distinct from historical A. Auxiliary targets cover task/contact/physics. Modules and optimizer membership match, while gradient-active capacity differs.

The reported conditional 95% intervals use 10,000 independent reset-wave bootstrap draws stratified by evaluation seed. They condition on the two evaluated training checkpoints per row and do not estimate broader training-population variability. No multiple-comparison correction is applied. R1–R0 success is $+12.50$ points $[11.03, 13.97]$. R2–R0 is $+13.77$ $[12.27, 15.27]$. R2–R1 is $+1.27$ $[-0.27, 2.77]$, whereas its late speed difference is $+0.0335$ rad/s $[0.0224, 0.0449]$ and total-turn difference is $+0.0974$ $[0.0674, 0.1271]$. R3–R2 success is -7.47 $[-8.93, -6.03]$, R4–R2 is -15.40 $[-17.07, -13.67]$, and R4–R3 is -7.93 $[-9.50, -6.27]$.

The canonical-recipe bridge uses the same final training configuration as the main TWIRL row. With H5 enabled it reaches 43.57%. Zeroing H5 during both training and deployment reaches 38.53%. Both means average two training seeds and three evaluation seeds. This bridge is separate from R0–R4 because the matched screen uses a common short continuation from B4, so R2 is treated as a mechanism probe rather than as a replacement for the final architecture.

Both bridge means average two training seeds, each evaluated with three evaluation seeds. The spread for the token-off bridge was not supplied. This canonical recipe is distinct from the matched R0–R4 screen.

P Additional Policy-Reliance Diagnostics

P.1 Primary intervention cohort

The calibration mean uses disjoint data. The primary normal/mean/empirical and supplemental shuffle/zero evaluations share the validated protocol. Zeroing changes both the offset and variation. The mean-minus-normal precision audit touches the strict -2-point equivalence boundary, so robust equivalence is not established.

The WM-only conditions use separate simulator launches but the same checkpoint, episode budget, and evaluator. Zeroing changes both the mean and variation at the actor interface and does not isolate predictive information. Mean replacement preserves performance, leaving a constant-bias explanation viable. These are descriptive differences, not an equivalence test.

The WM-only cohort in Table 14 is separate from the fresh historical-A cohort in the main text. In an earlier combined-A diagnostic, normal and shuffled ACD yield 43.03% and 42.50%, respectively. Neither is the 43.57% development reference. In the new primary cohort, normal/mean/empirical conditions yield $42.2333/41.8000/41.1000\%$. Mean-minus-normal is -0.4333 points. Two 200,000-draw precision audits give $[-2.00, 1.1333]$ and touch the strict equivalence boundary. Supplemental

Table 13: Policy reliance in the fresh historical-A cohort . Each condition uses 3000 episodes across three evaluation seeds. Interventions change the five-step actor token at inference.

Future actor input	Success (%)	Difference from normal (pp)
Normal prediction	42.23	–
Global calibration mean	41.80	-0.43
Independent empirical token	41.10	-1.13
Cross-environment shuffle	40.93	-1.30
Zero	35.20	-7.03

Table 14: Matched inference interventions on the WM-only progressive ITD→ACD diagnostic checkpoint. Each condition contains 1000 episodes for each of three evaluation seeds. Rotation is signed progress over all episodes, in revolutions.

ACD token	Success (%)	Δ (pp)	Timeout (%)	Rotation
Normal	39.07	–	39.80	0.958
Global calibration mean	39.93	$+0.87$	40.57	0.963
Cross-environment shuffle	37.50	-1.57	38.47	0.946
Zero	33.27	-5.80	34.20	0.803

lag 5/8/15 conditions yield $40.1667/41.8000/41.9333\%$, shuffle 40.9333%, and zero 35.2000%. Delay effects are nonmonotonic. Zeroing removes the mean contribution as well as state-dependent variation, so its larger effect cannot alone demonstrate use of future semantics.