2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

CG-THWM: Curriculum-Guided Temporal Haptic World Modeling for Peg-in-Hole Tasks

Xinli Zhong[1,2,3,5], Feng Han[3], Manya Xu[4], Mu Li[3], Daqiang Zhang[6], and Jianwei Niu[*,3,5]

***Abstract*—Fine-tolerance peg-in-hole manipulation demands high precision under contact-rich, nonsmooth dynamics, where irregular geometries, inclinations, and tight-clearance interference often cause model-free reinforcement learning (RL) to fail. We propose the Curriculum-Guided Temporal Haptic World Model (CG-THWM), which couples a world model with temporal haptic information and trains it via a staged curriculum. The world model supports efficient long-horizon planning with value estimation, while temporal haptic signals expose critical contact events; the curriculum stabilizes training and improves generalization. To enable rigorous evaluation, we construct a dataset for complex insertions that covers irregular, inclined, and interference-rich settings. In simulation, CG-THWM attains a 100% success rate on standard baselines and a 70% mean success rate in scenarios where conventional RL fails. These results highlight CG-THWM's potential for industrial and service applications.**

I. Introduction

In modern manufacturing, many contact-rich tasks require high-precision operation and carry risk [1–3]: improper actions can fail the task and even damage components. Among these tasks, peg-in-hole is the most common [1, 4]. Peg-in-hole typically consists of multiple phases with distinct dynamics, nonsmooth contact transitions, and cumulative error [5]. Existing approaches either rely on high-fidelity simulation with long-horizon planning—costly and highly sensitive to modeling error—or approximate policies through extensive trial-and-error, which is sample-inefficient and unsafe [6–9]. Traditional model-based reinforcement learning (MBRL) often models task-irrelevant details, leading to compounding bias [10–12].

In real-world assembly, peg-in-hole rarely reduces to an idealized alignment-and-insertion step: **inclined, rotated, and polygonal holes introduce nonsmooth edge/corner contacts.** Haptic signals vary sharply with pose, and slight misalignment can cause jamming with transient spikes. In factory settings, compliance-based control combined with heuristic search increases cycle time under substantial initial misalignment, requires extensive threshold and state-machine parameter tuning, and exhibits pronounced sensitivity to station disturbances and part tolerances. **Temporal modeling is also a significant gap.** Static aggregation or frame-wise reconstruction fails to capture these transient [13] contact dynamics, thereby precluding timely initiation of recovery behaviors such as backing off, releasing contact force, and reattempting. Therefore, we adopt a **latent-space world model with temporal haptic information** that uses tactile signals to infer latent contact states. Within this latent space, we perform short-horizon planning and complement it with **long-horizon value** estimation, enabling the policy to generate appropriate actions conditioned on contact signatures and thereby reducing failure rates. **Complex peg-in-hole lacks a standardized dataset with systematic coverage of inclination, rotation, diverse hole shapes, and tight clearance.** This leads to inconsistent evaluation [14, 15] and poor reproducibility. Assembly is contact-rich: haptic information and geometry vary across inclination, rotation, diverse hole shape and tight clearances, so policies often fail to transfer between scenes. Most public studies emphasize a few cylindrical parts in constrained setups, limiting comparability. We propose the **Curriculum-Guided Temporal Haptic World Model (CG-THWM)**—a world-model planner that integrates temporal haptics, accompanied by a

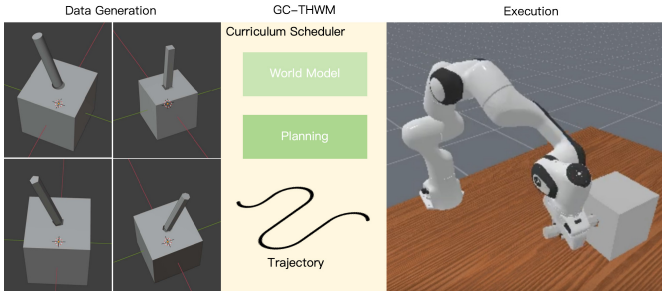

Fig. 1: **Overview of CG-THWM.** *Left:* In simulation, we instantiate the ComplexPeg-Hole benchmark by randomly sampling inclinations, relative rotations, diverse hole shapes, and clearances. *Middle:* We train a multimodal world model that aligns state and temporal haptic encoders via haptic attention; a curriculum scheduler gradually increases geometric complexity and initialization perturbations. *Right:* During execution, the agent applies the first action of the predicted trajectory and replans online, enabling fast and stable insertions.

*Corresponding author
[1]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China.
[2]Chongqing College, University of Chinese Academy of Sciences, Chongqing, China.
[3]Hangzhou Innovation Institute, Beihang University, Hangzhou, China.
[4]Institute of Software, Chinese Academy of Sciences, Beijing, China.
[5]Zhejiang Key Laboratory of Industrial Big Data and Robot Intelligent Systems, Hangzhou Innovation Institute of Beihang University, Hangzhou, China.
[6]Tongji University, Shanghai, China.
Emails:zhongxinli23@mails.ucas.ac.cn, dqzhang@tongji.edu.cn, niujianwei@buaa.edu.cn

new dataset and curriculum (Fig. 1). By combining temporal haptic world model with Contact-Geometry Curriculum Learning, CG-THWM guides the agent to generate correct trajectories in complex peg-in-hole scenarios. Planning performs trajectory planning in a latent space, and a longer-horizon return estimator provides terminal value targets. Joint training under value objectives allows haptics to dominate decision-making at critical moments. In addition, we design a **contact–geometry Curriculum** that schedules difficulty along inclination, rotation, clearance, and diverse hole shapes, while adaptively adjusting training using the success-rate metric. Our contributions are fourfold:

- To address the challenges of complex peg-in-hole scenarios, we propose the **Temporal Haptic World Model (THWM)**, which leverages temporal haptic signals to guide the agent, accelerating fine-manipulation tasks and achieving higher success rates.
- For complex peg-in-hole settings (e.g., holes with inclination, rotation, diverse hole shapes, and tight clearances), we release the **ComplexPeg-Hole** dataset comprising 100,000 objects, annotated along eight axes: hole roll, pitch, yaw; clearance; hole shape; hole-opening location; peg length; peg shape.
- Because complex peg-in-hole operations demand high precision and prolonged exploration, with difficulty varying widely across scenarios, we introduce a **Contact-geometry Curriculum** that couples geometric difficulty with disturbances, enabling the agent to learn faster and more stably and to attain higher success rates.
- We validate the dataset and the method with simulation experiments, including component-wise ablations.

Overall, by combining an implicit task-variable representation with world-model reasoning over temporal haptics, we replace heuristic search and threshold tuning in late-stage assembly and improve success rate, stability, and safety in inclined, rotated, and polygonal scenarios, while providing a reproducible and comparable foundation for future research.

II. Related Work

Research on contact-rich manipulation spans classical control, model-based reinforcement learning, and curriculum strategies. To organize this literature, we group prior work into four areas: (1) complex contact and insertion tasks, (2) world models and MPC, and (3) curriculum learning for progressive training.

A. Complex Contact and Insertion Tasks

Peg-in-hole insertion has long been a benchmark in robotics [16, 17]. While previous analytical and model-based studies have successfully tackled insertion tasks with complex planar geometries and sub-millimeter tolerances [18, 19], traditional methods depend on accurate system models and cannot handle unknown deviations [17, 20]. Imitation learning also suffers from limited generalization due to small or biased datasets. Reinforcement learning (RL) faces sparse rewards and poor transferability. To mitigate these issues, some works divide the task into alignment and insertion [21, 22], while others apply domain randomization or data augmentation [23–25]. Survey papers further confirm insufficient validation of RL in real-world assembly, citing generalization bottlenecks as a major limitation.

B. World Models and MPC

World models simulate dynamics to improve efficiency and enable long-horizon reasoning. Dreamer achieves multi-task generalization without demonstrations [26, 27]. MPC has been integrated with learning, as in TD-MPC [28]. QT-TDM further replaces RNNs with Transformers for improved sequence modeling and efficiency [29, 30]. However, most world-model approaches still rely primarily on vision. Large multimodal models such as Palm-E and RT-2 emphasize semantic grounding but lack tactile integration and fine-manipulation validation [31–33]. Beyond vision-centric world models, safety filters based on CBF-constrained optimization and neural dynamics have shown promise for nonconvex control and real-time feasibility, offering complementary inner-loop safeguards to latent planning [19, 20].

C. Curriculum Learning For Progressive Training

Curriculum learning (CL) accelerates training and improves success in sparse-reward tasks [34]. Difficulties are gradually increased by reducing clearance or introducing misalignment. With domain randomization, sim-to-real transfer becomes feasible. While automatic curricula are emerging [25, 35], most fineassembly studies still rely on human-designed curricula due to stability concerns. Overall, CL has proven effective in contactrich manipulation, and our work applies it within CG-THWM.

III. Preliminaries

Problem Setup: We consider an infinite-horizon Markov Decision Process (MDP) defined by the sextuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma, p_0)$, where $\mathcal{S} \subset \mathbb{R}^n$ and $\mathcal{A} \subset \mathbb{R}^m$ are continuous state and action spaces, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and p_0 denotes the initial state distribution. To jointly capture perceptual and haptic information, we represent the state at time t as $s_t := (s_t^{\text{prop}}, f_t) \in \mathcal{S}$, where $s_t^{\text{prop}} \in \mathbb{R}^{n_s}$ is a low-dimensional proprioception, and $f_t \in \mathbb{R}^{n_f}$ is a temporal haptic vector that renders the process Markov; we define$f_t := [F_{t-k+1:t}, T_{t-k+1:t}]$ where $F_{t-k+1:t} = [F_{t-k+1}, \ldots, F_t]$, $T_{t-k+1:t} = [T_{t-k+1}, \ldots, T_t]$, $k \geq 2$. By stacking the most recent k force and torque measurements, the state is augmented to satisfy the Markov property. Our objective is to learn a parameterized policy $\Pi_\theta : \mathcal{S} \to \mathcal{A}$ that maximizes the discounted return $\mathbb{E}_{\Pi_\theta}[\sum_{t=0}^{\infty} \gamma^t r_t]$, $r_t \sim \mathcal{R}(\cdot \mid s_t, a_t)$ and at each decision step t we sample $a_t \sim \Pi_\theta(\cdot \mid s_t)$.

TD-MPC2: TD-MPC2 [36] is a model-based reinforcement learning algorithm that couples a latent-space world model with model predictive control (MPC) style local planning and a terminal value learned via temporal-difference (TD) methods [28]. Concretely, it learns a representation function $z = h_\psi(o)$ that maps high-dimensional observations

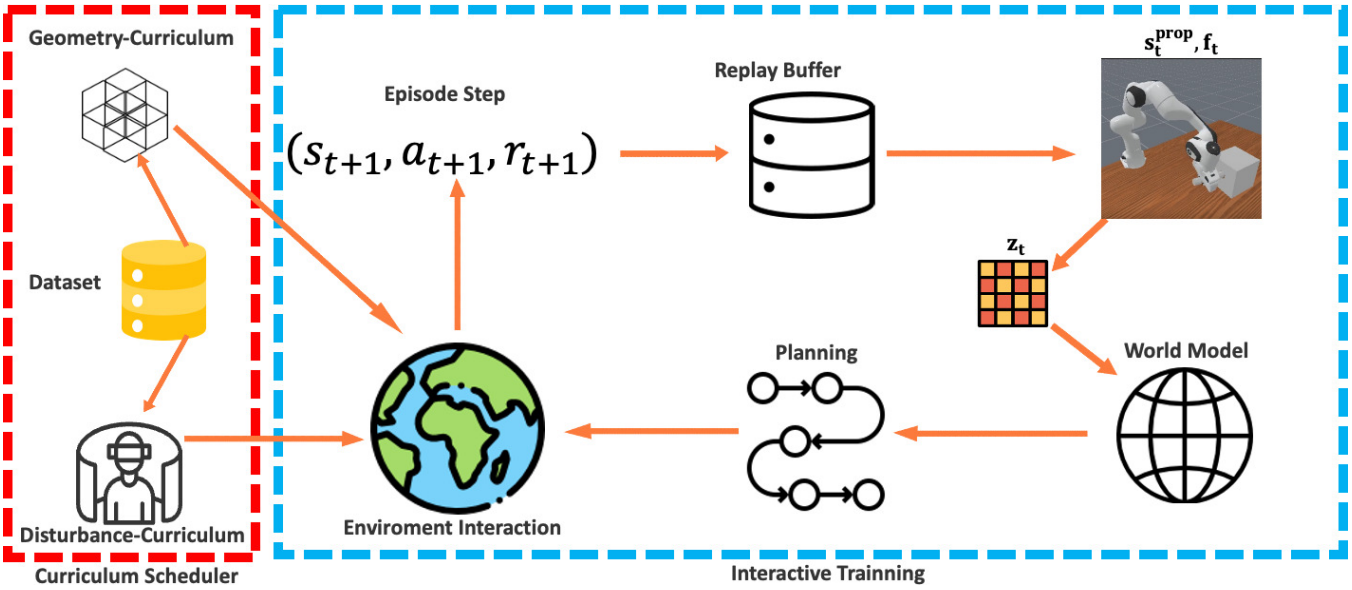

Fig. 2: **Framework overview.** A curriculum-driven interactive MBRL loop couples a **Contact-Geometry Curriculum Scheduler** (left, red) with **Interactive Training** (right, blue). The scheduler maintains Geometry-Curriculum and Disturbance-Curriculum, generating task distributions and adapting their difficulty using statistics from the Dataset. During interactive training, an agent plans with a learned World Model to interact with the environment, and then the produced transitions (s_t, a_t, r_t) are stored in a Replay Buffer,closing the collect–learn–schedule loop. All collected data update the world model and planner, which in turn enable better planning in the next round.

o to a compact latent z and latent dynamics $z' = d_\psi(z, a)$. In addition, it trains prediction heads R_ψ, Q_ψ, π_ψ for (i) instantaneous reward $r = R_\psi(z, a)$, (ii) state–action value $Q_\psi(z, a)$, and (iii) a policy prior $a \sim \pi_\psi(z)$ that biases sampling-based planning toward high-return trajectories. During interaction, the agent performs receding-horizon planning in the learned world model and executes the resulting action.

IV. Curriculum-Guided Temporal Haptic World Model

Building on the analysis above and targeting three challenges in complex peg-in-hole tasks—**contact complexity with a temporal-modeling gap, dataset and evaluation gaps with wide difficulty ranges, and slow, brittle heuristic engineering**—we propose the Curriculum-Guided Temporal Haptic World Model (CG-THWM). The overall framework is shown in Fig. 2. The core idea is to align system states with temporal haptic signals in a unified latent space and to focus on key contact events via haptic-aware attention; then, leveraging world-model planning, we perform long-horizon decision making in that latent space; finally, a contact–geometry curriculum stabilizes training and improves generalization across geometries and poses, thereby reducing reliance on exhaustive search and ad hoc thresholds. Our approach comprises three modules: the ComplexPeg-Hole dataset, the temporal haptic world model, and the contact–geometry curriculum learning module.

A. ComplexPeg-Hole Dataset

Overall, existing datasets for peg-in-hole manipulation have substantially advanced benchmarking and reproducibility [37, 38]; however, they remain somewhat limited in the diversity of geometric families and hole poses, as well as in the systematic coverage of tolerances and contact details. For example, ManiSkill's *PegInsertionSide-v1* mainly performs planar randomization on a tabletop, and fixes the hole diameter as "peg radius/half-width plus tolerance," thereby emphasizing ease of use rather than comprehensive pose variation and tight-fit challenges.

To enable robots to reliably complete assembly in complex, contact-rich scenes featuring inclination, rotation, and diverse hole shapes [39], we construct the ComplexPeg-Hole Dataset. The dataset leverages Blender Geometry Nodes to build a fully parameterized generator that can automatically produce virtually unbounded pairs of peg-in-hole modules. Shapes, sizes, and poses are all controllable via parameters, ensuring sufficiently diverse training scenes to improve policy generalization.

Our domain randomization spans three layers: geometry, physics, and pose. To ensure physical consistency, hole geometries are constructed by a Boolean subtraction with tolerance-aware dilation:

$$B = C \setminus (P \oplus \tau), \tag{1}$$

where C is a base cuboid, P is the peg geometry, and $\oplus$ denotes a Minkowski dilation that adds a clearance τ. This construction guarantees a well-defined peg-in-hole match. With large-scale parallel simulation, the framework

efficiently generates stratified samples with near-uniform coverage within each difficulty tier, yielding a rich and balanced training distribution for reinforcement learning.

B. Temporal Haptic World Model (THWM)

To address the long-horizon and multi-stage planning challenges of peg-in-hole manipulation, it is necessary to adopt a framework that achieves long-horizon planning under complex environment, uses a terminal value to account for beyondhorizon returns, and fuses haptic information to robustly handle contact-rich tasks. Based on these considerations, we take TD-MPC2 as the backbone implementation of such a framework. Although TD-MPC2 is an excellent and general world model planning approach, its original formulation lacks explicit modeling and utilization of haptic modalities, making it insufficient to fully capture the contact dynamics uncertainty in peg-in-hole settings. At the same time, temporal limitations remain: TD-MPC2's reward and planning objective are effectively weighted by a near-uniform temporal discount, lacking attention or gating mechanisms anchored to haptic features. Building on these observations, we propose the Temporal Haptic World Model (THWM) atop the TD-MPC2 backbone, which introduces temporal haptic attention to support long-horizon planning under contactrich, nonsmooth dynamics. THWM is a control-centric and haptics-enhanced implicit world model; compared with TDMPC2, we additionally introduce a temporal haptic encoder and a multimodal fusion module. The model is trained by joint representation learning, reward prediction, and temporaldifference (TD) learning, and is subsequently embedded in an MPC framework for local trajectory optimization. This design avoids long-horizon pixel/state reconstruction, reduces mismatch between reconstruction and control objectives, lowers multi-step compounding error, and enables efficient use of large-scale interaction data with moderate model capacity. The world model supports long-horizon imagination planning for MPC to roll out high-quality action candidates; the resulting interaction data are then fed back to continuously train the model, forming a closed loop.

Components: The world model THWM comprises seven components: a visual–proprioceptive state encoder h_θ, a temporal haptic encoder g_θ, a multimodal fusion module F_θ, latent dynamics d_θ, a reward head R_θ, a terminal-value head Q_θ, and a policy prior π_θ. Fig. 3 illustrates the architecture of the THWM model and its constituent components:

$$
\begin{aligned}
\text{State encoder:} \quad & x_t = h_\theta(s_t^{\text{prop}}) \\
\text{Temporal haptic encoder:} \quad & u_t = g_\theta(f_t) \\
\text{Multimodal fusion:} \quad & z_t = \phi_\theta(x_t, u_t) \\
\text{Latent dynamics:} \quad & z_{t+1} = d_\theta(z_t, a_t) \\
\text{Reward head:} \quad & \hat{r}_t = R_\theta(z_t, a_t) \\
\text{Terminal-value head:} \quad & \hat{q}_t = Q_\theta(z_t, a_t) \\
\text{Policy prior:} \quad & \hat{a}_t \sim \pi_\theta(z_t)
\end{aligned}
\tag{2}
$$

Here s_t^{prop} and f_t denote the state and temporal haptic information, a_t is the action, z_t is the fused latent, and r_t the immediate reward. Crucially, the Markovized temporal

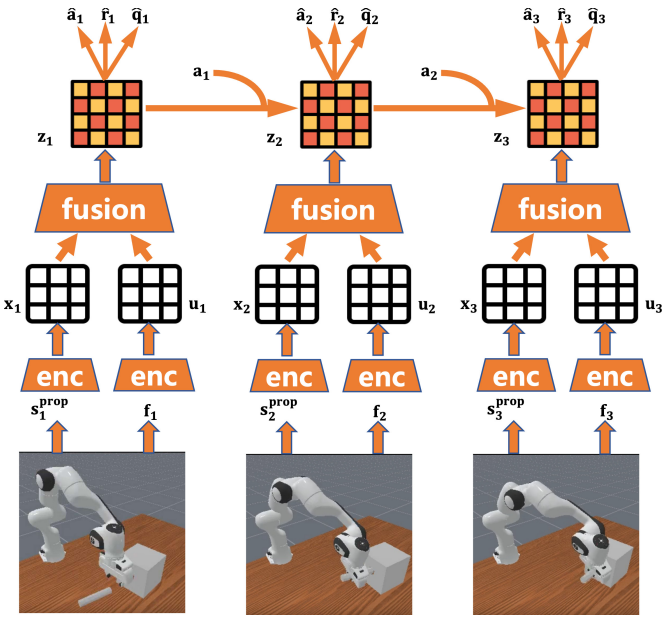

Fig. 3: **Latent World Model.** At each step i, the proprioceptive state s_i^{prop} together with haptic features f_i are encoded by two encoders into x_i and u_i, respectively. A fusion module combines them into a latent state z_i. From z_i, the model jointly predicts actions, rewards, and terminal values $(\hat{a}_i, \hat{r}_i, \hat{q}_i)$, and a latent dynamics rolls forward to z_{i+1} conditioned on the executed action a_i. Planning and rollouts happen entirely in latent space, without decoding future observations.

haptic signal is integrated into both representation learning and the control loop: via the haptic encoder and multimodal fusion, contact cues directly influence the latent state, value estimation, policy prior, and planning at every step. Given the proprioception s_t^{prop} and temporal haptic feedback f_t, the encoders produce x_t and u_t, which are fused into z_t. Conditioned on a_t, the model then (i) predicts the next latent z_{t+1} via d_θ, (ii) predicts the one-step reward $\hat{r}_t$ via R_θ, (iii) estimates the state–action value $\hat{q}_t$ via Q_θ, and (iv) samples $\hat{a}_t$ from the policy prior π_θ to guide planning and execution.

Temporal Unrolling and Gradient Propagation: To curb compounding error, we roll out from predicted future latents for K steps in latent space, accumulate losses on $\{\hat{r}, \hat{q}\}$ and related heads, and backpropagate through time (BPTT) across the unrolled computation graph to update θ. This multi-step, latent-space training aligns the model with control objectives while preserving stability under contact-rich, nonsmooth dynamics. MPPI is an MPC algorithm that iteratively updates the parameters of a family of action-sequence distributions via importance-weighted averaging over the top (expected-return) samples. We execute MPPI on the fused latent $z_t = \phi_\theta(h_\theta(s_t^{\text{prop}}), g_\theta(f_t))$. For each time step $h = 0, \ldots, H-1$, we maintain a diagonal-Gaussian family $\mathcal{N}(\mu_h, \text{diag}(\sigma_h^2))$ and maximize the receding-horizon return plus a terminal

value:

$$(\mu^*, \sigma^*) = \arg\max_{\mu,\sigma} \mathbb{E}_{a_{t:t+H}\sim\mathcal{N}(\mu,\sigma^2)} \left[\mathcal{J}_{\text{term}} + \mathcal{J}_{\text{stage}}\right] \quad (3)$$

where $\mathcal{J}_{\text{term}} = \gamma^H Q(z_{t+H}, a_{t+H})$, $\mathcal{J}_{\text{stage}} = \sum_{h=0}^{H-1} \gamma^h R(z_{t+h}, a_{t+h})$, $z_t = F_\theta(h_\theta(s_t), g_\theta(f_t))$. Here $\mu, \sigma \in \mathbb{R}^{H\times m}$ with m the action dimension. In practice, (3) is solved by repeatedly sampling action sequences from $\mathcal{N}(\mu, \sigma^2)$, evaluating their expected return, and updating (μ, σ) by importance-weighted averaging. Notably, (3) bootstraps beyond the horizon H using the learned terminal value, thereby approximating the full RL objective. After a fixed number of planning iterations, we execute the first action $a_t \sim \mathcal{N}(\mu_t^*, (\sigma_t^*)^2)$ in the environment. To accelerate convergence, a subset of candidate sequences is sampled from the policy prior p, and (μ, σ) is warm-started by time-shifting the previous solution by one step.

Policy Objective: The policy prior π_θ is a stochastic maximum-entropy policy trained to maximize

$$\mathcal{L}_{\pi_\theta}(\theta) = \mathbb{E}_{(s,a)_{0:H}\sim\mathcal{B}}\Bigg[\sum_{t=0}^{H} \lambda^t \Big(\alpha Q(z_t, \pi_\theta(z_t)) - \beta\mathcal{H}(\pi_\theta(\cdot \mid z_t))\Big)\Bigg], \quad (4)$$

where $z_{t+1} = d(z_t, a_t)$, $\mathcal{H}$ denotes the entropy of π_θ. Because the scales of $Q(z_t, \pi_\theta(z_t))$ and $\mathcal{H}(\pi_\theta(\cdot \mid z_t))$ can vary substantially across datasets and training phases, one must balance them to avoid premature entropy collapse. A common practice is to fix either α or β and tune the other via an entropy target or via running statistics; both α and β can also be adjusted adaptively using these schemes.

World-Model Objective: The components h, d, R, Q are trained jointly by minimizing

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')_{0:H}\sim\mathcal{B}}\Bigg[\sum_{t=0}^{H} \lambda^t \Bigg(\underbrace{\|z'_t - \hat{z}'_t\|_2^2}_{\text{joint-embedding prediction}} + \underbrace{\text{CE}(\hat{r}_t, r_t)}_{\text{reward prediction}} + \underbrace{\text{CE}(\hat{q}_t, q_t)}_{\text{value prediction}}\Bigg)\Bigg] \quad (5)$$

where $\hat{z}'_t = \text{sg}\big(\phi_\theta(h_\theta(s_t^{\text{prop}'}), g_\theta(f'_t))\big)$, $\text{sg}(\cdot)$ denotes the stop- gradient operator, $(z'_t, \hat{r}_t, \hat{q}_t)$ are defined in (2), and $q_t \triangleq r_t + \gamma \bar{Q}(z'_t, \pi_\theta(z'_t))$ is the TD target at step t with $\bar{Q}$ an EMA target for Q. Due to large cross-task variations in reward scale, we express reward and value prediction as discrete regression in log space and minimize cross-entropy against soft targets for r_t and q_t.

C. Contact-Geometry Curriculum Learning Module

Training directly on the full difficulty distribution in com- plex peg-in-hole tasks is inadequate—even with a world model: tiny errors can cause jams; contact dynamics are highly nonlinear; and large sample-to-sample variation. We therefore propose the Contact-Geometry Curriculum: a haptic-aware mechanism that preserves task difficulty while enabling stable progression. A layered seed-scheduling algorithm serves as the core scheduler for progressive difficulty control. It rests on a dual-axis curriculum that independently manages (i) geometry complexity and (ii) environment-initialization perturbations, while a unified controller advances both axes stage by stage.

Three-Stage Progressive Strategy: Training proceeds through three stages from simple to challenging. Along the geometry axis we keep three buckets: easy, medium, and hard. Along the perturbation axis we control translational offsets and rotational ranges of the peg and socket, forming three non- overlapping, increasing uncertainty intervals aligned with the geometry buckets. Stage advancement is triggered by success rate computed over a sliding evaluation window to avoid oscillations caused by short-term variance: let the window length be K and the stage threshold be η_s. A stage transition is permitted only when the arithmetic mean of the success rates within the window is no less than the threshold.

Mitigating Catastrophic Forgetting via Mixed Replay: We mitigate catastrophic forgetting with a mixture replay scheme. Separate replay buffers are maintained per stage, and higher- stage training mixes samples from the current, preceding, and early stages with preset proportions:

$$x_{\text{mixed}} = \alpha_{\text{current}}\, x_{\text{current}} + \alpha_{\text{prev}}\, x_{\text{prev}} + \alpha_{\text{early}}\, x_{\text{early}}, \quad (6)$$

where $\alpha_{\text{current}} + \alpha_{\text{prev}} + \alpha_{\text{early}} = 1$. Concretely, in Stage 2 we use $(\alpha_{\text{current}}, \alpha_{\text{prev}}, \alpha_{\text{early}}) = (0.8, 0.2, 0.0)$; in the hardest stage we adopt a three-way mix $(0.7, 0.2, 0.1)$.

Hard-Case Replay Pool: A hard-example replay pool fo- cuses learning on persistently difficult configurations. If the observed success rate for a configuration stays below threshold across multiple evaluations, that configuration is added to the pool and subsequently resampled with a low but non-negligible probability to improve generalization.

Weight Smoothing Between Stages: To ensure smooth transitions, sampling probabilities are slowly reweighted across difficulty buckets. During promotion, weights are linearly interpolated over a transition period of length T:

$$w(t) = (1 - \lambda)\, w_{\text{from}} + \lambda\, w_{\text{to}}, \quad (7)$$

where w_{from} and w_{to} are the pre- and post-promotion weight vectors, $\lambda = \min(1, t/T)$, and t counts steps within the transition. Together, progressive staging, success-rate triggers with dwell and M-of-K safeguards, mixed replay across stages, hard-case resurfacing, and weight smoothing yield stable learning dynamics and effective transfer from simple to complex tasks.

V. Experiments

We run all experiments in the *PegInsertionSide-v1* from ManiSkill3 environment, which provides the robot, table, and general scene setup. The task-specific peg–hole assets are drawn from our ComplexPeg-Hole dataset, which serves as the unified base for training and evaluation and spans key axes including hole inclination, relative rotation, diverse hole shapes, clearance. All experiments are conducted in

TABLE I: **Experimental setup.** We evaluate on *PegInsertionSide-v1* from ManiSkill3 using the ComplexPeg-Hole dataset. All curves are averaged over 5 random seeds. The interaction budget is 5M steps. DEMO^3 receives the sparse environment reward, while all other methods share the same dense reward. Demonstrations are used only by DEMO^3 and MoDem (10 demos each).

Method	Class	Reward	Demos	Interactions	Seeds
Ours	MBRL	dense	0	5M	5
TD-MPC2	MBRL	dense	0	5M	5
DEMO^3	MBRL	sparse	10	5M	5
MoDem	MBRL	dense	10	5M	5
TD-MPC	MBRL	dense	0	5M	5
SAC	MFRL	dense	0	5M	5
PPO	MFRL	dense	0	5M	5

ManiSkill3 using documentation build 3.0.0b21. Physics is powered by PhysX through the SAPIEN integration. The robot is a 7-DoF Franka Emika Panda driven with a PD end-effector position-increment controller. In simulation, we estimate the end-effector six-axis external wrench from joint torques using a dynamics model and an observer, thereby providing a haptic signal consistent with the real robot; the sampling frequency is matched to the Franka Control Interface (FCI). All settings and control interfaces follow the official ManiSkill3 and SAPIEN implementations and controller definitions.

All learning curves are reported on 5 random seeds. The interaction budget is 5M environment steps. The episode length is 100 environment steps. Except for DEMO³, which uses the environment sparse reward plus a discriminator-produced dense reward, all other methods, including CG-THWM, use the same dense reward function for fairness. Among the compared methods, only DEMO³ and MoDem use demonstrations, and both receive 10 demonstrations. The complete setup is provided in Table I. Under the unified setup in Table 1, we evaluate on PegInsertionSide-v1 from ManiSkill3 with the ComplexPeg-Hole dataset to answer three questions:

1) Under the same interaction budgets, how does CG-THWM compare to model-free RL methods and other model-based RL methods on peg-in-hole task using ComplexPeg-Hole dataset?

2) How do method ranking, sample efficiency, and training stability change when using ComplexPeg-Hole or not?

3) What are the relative contributions of our components? All methods share the same observation and action interfaces and evaluation method. Interaction and demonstration budgets are set per task.

We compare against two families of methods: Model-based RL algorithm: TD-MPC [28], TD-MPC2 [36], MoDem [40], DEMO³ [41]. Model-free RL algorithm: SAC [42], PPO [43]. We use official implementations or community-verified reproductions when available, tune hyperparameters under a shared validation budget, and keep observation/action/reset interfaces identical across methods.

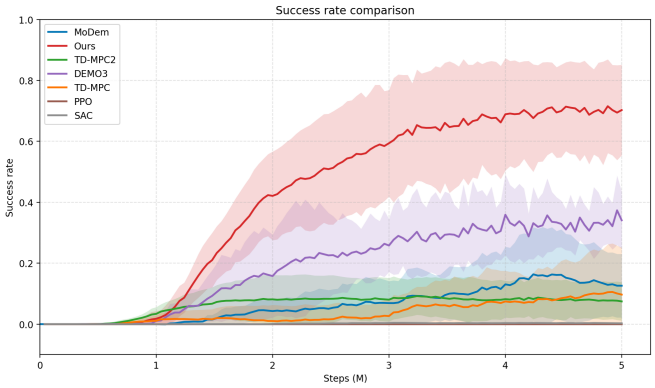

Fig. 4: **Comparative Experiments.** Success rate of our method and baselines averaged across 5 random seeds. Solid lines denote the mean over multiple random seeds, and shaded regions indicate ±1 standard deviation.

Comparative Experiments: We first compare against both model-based and model-free RL methods on complex pegin-hole tasks using the ComplexPeg-Hole dataset; aggregated results are shown in Fig. 4. Under identical settings, our CG-THWM achieves the best performance. This indicates that, in contact-rich dynamics, haptic-augmented latent-space planning paired with a geometry–dynamics curriculum improves final outcomes. We also observe that world-model approaches significantly outperform model-free RL, suggesting that shorthorizon planning in latent space, complemented by longhorizon value estimation, leads to better planning and higher task success rates. We further benchmark all methods on the communitystandard PegInsertionSide-v1 from ManiSkill3 as a reproducible sanity check (Fig. 5). All methods reach 100% success within a short training horizon, except PPO, learning curves and final performance exhibit no meaningful differences, which needs much more interactions. This suggests that, under current settings, the task is saturated: its geometry and contact conditions are sufficiently benign that it cannot differentiate algorithmic capability under non-smooth contact, strong disturbances, and tight tolerances. Consequently, we use PegInsertionSide-v1 primarily to verify consistency and implementation correctness rather than to support our main claims. On our ComplexPeg-Hole benchmark, by contrast, method differences are systematically amplified, enabling a faithful assessment of model capability on truly complex pegin-hole scenarios.

Ablation Experiments: We ablate the components of CG-THWM by (i) removing haptics, (ii) removing the contact-geometry curriculum, and (iii) varying the temporal window size. Results are shown in Fig. 6. Across settings, each proposed component contributes materially to performance. Haptics yields the largest gain: temporal haptic signals help the model disambiguate contact states and choose appropriate actions. Removing the contact–geometry curriculum produces markedly noisier learning curves and lowers overall success, indicating that the curriculum stabilizes training while improving the average success rate. Using a single

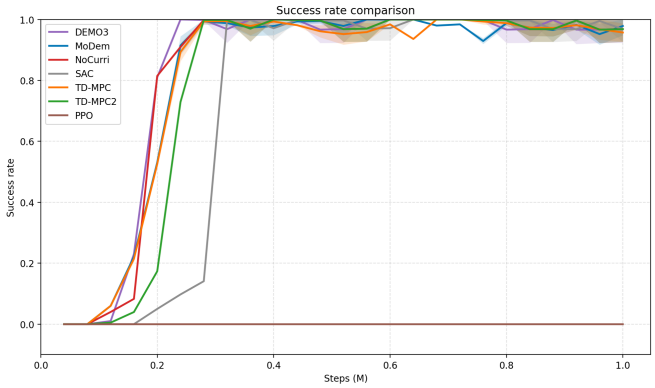

Fig. 5: **Comparative Experiments on Original PegInsertionSide-v1 from ManiSkill3 task.** Success rate of our method and baselines averaged across 5 random seeds over 1M steps.

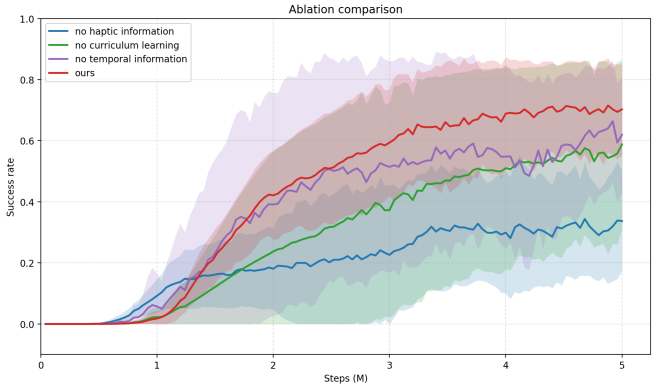

Fig. 6: **Ablation Experiments.** Success rate as a function of interaction steps for variations of our method. Averaged across 5 random seeds. Ablation components include haptic information,Contact-Geometry Curriculum Learning and temporal information.

haptic frame leads to a clear drop in success, which suggests that employing a temporal window aligned with events helps capture contact signatures and stabilizes optimization. We further study temporal horizon lengths $\{1, 2, 4, 6, 10\}$ and observe that larger horizons generally perform better up to six(Fig. 7) the difference between six and ten is small, and a horizon of ten can even degrade performance. We hypothesize that an overly long window dilutes informative haptic cues, reducing sensitivity to contact events and thereby lowering the success rate.

VI. Conclusion

We study peg-in-hole assembly under tight tolerances and contact-rich, nonsmooth dynamics, and propose CG-THWM. Within a unified latent space, the method introduces temporal haptic attention to highlight key contact events; it performs short-horizon planning on a learned world model combined with terminal value estimation, thereby supporting long-horizon decision making; in parallel, a contact–geometry curriculum stabilizes training and improves generalization.

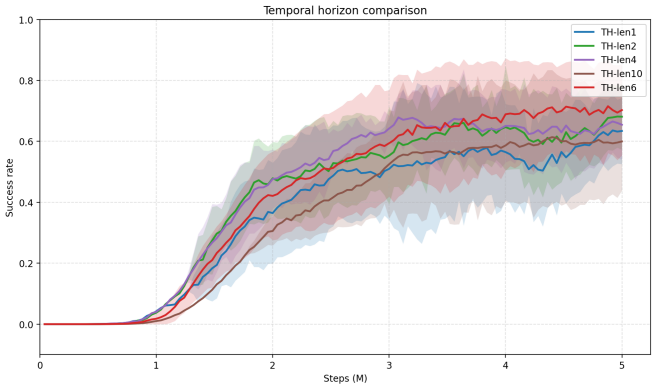

Fig. 7: **Temporal Horizon Ablation Experiments.** We compare success-rate trajectories versus training steps across temporal horizon lengths 1, 2, 4, 6, 10. Longer horizons generally achieve higher final success.

To enable rigorous evaluation, we construct the ComplexPeg-Hole dataset and a unified evaluation protocol covering inclination, relative rotation, diverse hole shapes, and tolerance. Under our insertion setup, CG-THWM attains 100% success on standard benchmark cases and a 70% mean success rate on harder scenarios where conventional RL fails, validating the effectiveness of a world-model approach with temporal haptic fusion coupled with a geometry-aware curriculum. Ablations show that removing temporal haptics markedly reduces success, while removing the curriculum induces training oscillations and lowers average performance.

Current validation is primarily in simulation; real-world friction, clearance, compliance, and sensor latency may introduce a sim-to-real gap. Planning overhead and high-frequency haptic encoding also increase runtime cost. To validate our approach on real-world assembly, we are conducting real robot experiments. Future work includes sim-to-real studies on hardware for contact-rich tasks, extending the approach to other tasks, augmenting ComplexPeg-Hole with real data, and releasing the implementation to promote reproducibility. Overall, CG-THWM goes beyond heuristic thresholds and exhaustive search, offering a practical path to high success rates, robustness, and safety for industrial and service robots in complex contact settings.

References

[1] Andrew S Morgan et al. “Vision-driven compliant manipulation for reliable, high-precision assembly tasks”. In: arXiv preprint arXiv:2106.14070 (2021).

[2] Florian Wirnshofer et al. “Controlling Contact-Rich Manipulation Under Partial Observability.” In: Robotics: Science and Systems. 2020.

[3] Tony Z Zhao et al. “Learning fine-grained bimanual manipulation with low-cost hardware”. In: arXiv preprint arXiv:2304.13705 (2023).

[4] Wenzhao Lian et al. “Benchmarking Off-The-Shelf Solutions to Robotic Assembly Tasks”. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE Press, 2021, pp. 1046–1053. DOI: 10.1109/IROS51168.2021.9636586. URL: https://doi.org/10.1109/IROS51168.2021.9636586.

[5] Yiting Chen et al. “Robust Peg-in-Hole Assembly under Uncertainties via Compliant and Interactive Contact-Rich Manipulation”. In: arXiv preprint arXiv:2506.22766 (2025).

[6] Kurtland Chua et al. “Deep reinforcement learning in a handful of trials using probabilistic dynamics models”. In: Advances in neural information processing systems 31 (2018).

[7] Ramanan Sekar et al. “Planning to explore via self-supervised world models”. In: International conference on machine learning. PMLR. 2020, pp. 8583–8592.

[8] Philipp Wu et al. “Daydreamer: World models for physical robot learning”. In: Conference on robot learning. PMLR. 2023, pp. 2226–2240.

[9] Shun Zuo et al. “Fast Robot Hierarchical Exploration Based on Deep Reinforcement Learning”. In: 2023 International Wireless Communications and Mobile Computing (IWCMC). 2023, pp. 138–143. DOI: 10.1109/IWCMC58020.2023.10183136.

[10] Danijar Hafner et al. “Dream to control: Learning behaviors by latent imagination”. In: arXiv preprint arXiv:1912.01603 (2019).

[11] Danijar Hafner et al. “Learning latent dynamics for planning from pixels”. In: International conference on machine learning. PMLR. 2019, pp. 2555–2565.

[12] Sangwoon Kim and Alberto Rodriguez. “Active Extrinsic Contact Sensing: Application to General Peg-in-Hole Insertion”. In: 2022 International Conference on Robotics and Automation (ICRA). 2022, pp. 10241–10247. DOI: 10.1109/ICRA46639.2022.9812017.

[13] Tobias Johannink et al. “Residual reinforcement learning for robot control”. In: 2019 international conference on robotics and automation (ICRA). IEEE. 2019, pp. 6023–6029.

[14] Wenzhao Lian et al. “Benchmarking off-the-shelf solutions to robotic assembly tasks”. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. 2021, pp. 1046–1053.

[15] Lujie Yang et al. “Physics-driven data generation for contact-rich manipulation via trajectory optimization”. In: arXiv preprint arXiv:2502.20382 (2025).

[16] Sangwoon Kim and Alberto Rodriguez. “Active extrinsic contact sensing: Application to general peg-in-hole insertion”. In: 2022 International Conference on Robotics and Automation (ICRA). IEEE. 2022, pp. 10241–10247.

[17] Karin Nottensteiner et al. “Robust, Locally Guided Peg-in-Hole Insertion Using Impedance-Controlled Robots”. In: 2020 International Conference on Robotics and Automation (ICRA). IEEE. 2020.

[18] A. Stemmer, A. Albu-Schaffer, and G. Hirzinger. “An Analytical Method for the Planning of Robust Assembly Tasks of Complex Shaped Planar Parts”. In: Proceedings 2007 IEEE International Conference on Robotics and Automation. 2007, pp. 317–323. DOI: 10.1109/ROBOT.2007.363806.

[19] Lars Johannsmeier, Malkin Gerchow, and Sami Haddadin. “A framework for robot manipulation: Skill formalism, meta learning and adaptive control”. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE. 2019, pp. 5844–5850.

[20] Jianlan Luo et al. “Reinforcement Learning on Variable Impedance Controller for High-Precision Robotic Assembly”. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE. 2019.

[21] Letian Fu et al. “Safe Self-Supervised Learning in Real of Visuo-Tactile Feedback Policies for Industrial Insertion”. In: 2023 International Conference on Robotics and Automation (ICRA). IEEE. 2023.

[22] Michael Haugaard et al. “Fast and Robust Peg-in-Hole Insertion by Controlling Visual In-Hand Pose”. In: Proceedings of the 5th Conference on Robot Learning (CoRL). PMLR. 2021.

[23] Josh Tobin et al. “Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World”. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. 2017.

[24] Xue Bin Peng et al. “Sim-to-Real Transfer of Robotic Control with Dynamics Randomization”. In: 2018 International Conference on Robotics and Automation (ICRA). IEEE. 2018.

[25] Denis Yarats et al. “Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels”. In: International Conference on Learning Representations (ICLR). 2021.

[26] Danijar Hafner et al. “Dream to Control: Learning Behaviors by Latent Imagination”. In: International Conference on Learning Representations (ICLR). 2020.

[27] Yifeng Wu et al. “DayDreamer: World Models for Continuous Control in the Real World”. In: Proceedings of the 6th Conference on Robot Learning (CoRL). PMLR. 2022.

[28] Nicklas Hansen, Hao Su, and Xiaolong Wang. “Temporal Difference Learning for Model Predictive Control”. In: Proceedings of the 39th International Conference on Machine Learning (ICML). PMLR. 2022, pp. 8512–8533.

[29] Michael Janner et al. “Trajectory Transformer: Offline Reinforcement Learning via Sequence Modeling”. In: Advances in Neural Information Processing Systems (NeurIPS). 2021.

[30] Lili Chen et al. “Decision Transformer: Reinforcement Learning via Sequence Modeling”. In: Advances in Neural Information Processing Systems (NeurIPS). 2021.

[31] Michael Ahn et al. “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances (SayCan)”. In: Proceedings of the 7th Conference on Robot Learning (CoRL). PMLR. 2023.

[32] Anthony Brohan et al. “RT-1: Robotics Transformer for Real-World Control at Scale”. In: Robotics: Science and Systems (RSS). 2023.

[33] Yunfan Jiang et al. “VIMA: General Robot Manipulation with Multimodal Prompts”. In: International Conference on Machine Learning (ICML). 2023.

[34] Marcin Andrychowicz et al. “Hindsight Experience Replay”. In: Advances in Neural Information Processing Systems (NeurIPS). 2017.

[35] Dmitry Kalashnikov et al. “QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation”. In: Proceedings of the 2nd Conference on Robot Learning (CoRL). PMLR. 2018.

[36] Nicklas Hansen, Hao Su, and Xiaolong Wang. “TD-MPC2: Scalable, Robust World Models for Continuous Control”. In: International Conference on Learning Representations (ICLR). 2024.

[37] Daniel Sliwowski et al. “Reassemble: A multimodal dataset for contact-rich robotic assembly and disassembly”. In: arXiv preprint arXiv:2502.05086 (2025).

[38] Stone Tao et al. “Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai”. In: arXiv preprint arXiv:2410.00425 (2024).

[39] Andrej Orsula et al. “Leveraging Procedural Generation for Learning Autonomous Peg-in-Hole Assembly in Space”. In: 2024 International Conference on Space Robotics (iSpaRo) (2024), pp. 357–364. URL: https://api.semanticscholar.org/CorpusID:269502315.

[40] Nicklas Hansen et al. MoDem: Accelerating Visual Model-Based Reinforcement Learning with Demonstrations. 2022. arXiv: 2212.05698. URL: https://arxiv.org/abs/2212.05698.

[41] Adrià López Escoriza et al. Multi-Stage Manipulation with Demonstration-Augmented Reward, Policy, and World Model Learning. 2025. arXiv: 2503.01837. URL: https://arxiv.org/abs/2503.01837.

[42] Tuomas Haarnoja et al. “Soft actor-critic algorithms and applications”. In: arXiv preprint arXiv:1812.05905 (2018).

[43] John Schulman et al. Proximal Policy Optimization Algorithms. 2017. arXiv: 1707.06347. URL: https://arxiv.org/abs/1707.06347.