Tac-Bench: Benchmarking Diverse Tactile Manipulation and Multimodal World Models

Anonymous Author(s)
Affiliation
Address
email

Abstract: Contact-rich manipulation depends on touch: robots must modulate grip force, sense compliance, and guide insertions through occlusions that defeat vision alone. **weihan**: motivations not clear. why do you need visuotactile for robot learning? one sentence to start.We present **Tac-Bench**, a benchmark for visuotactile robot learning. Existing work on vision-touch fusion and visuotactile world models uses incompatible tasks, sensors, and metrics, making progress hard to measure. Tac-Bench targets two problems under controlled conditions. First, to study how vision and touch should be fused, we design contact-rich tasks that isolate occlusion, force sensitivity, and in-hand status, then fix the task and data while varying only the fusion architecture. Second, to evaluate physical realism in world models, we move beyond visual fidelity: we decode actions from generated rollouts with a pretrained inverse-dynamics model, and we test whether expert policies can reach goals inside the imagined dynamics. Tac-Bench reveals when tactile sensing helps and where current visuotactile models fail, establishing a rigorous, reproducible foundation for multimodal robot learning.**weihan**: Remove this, one sentence say the importance of this work

Keywords: Tactile Manipulation, Visuo-Tactile Policies, World Models, Benchmark

1 Introduction

Human dexterity is grounded in touch. We modulate grip force to prevent slip, infer object compliance on contact, and guide insertions through occlusions that vision cannot penetrate. Robotic manipulation increasingly aspires to replicate this tactile intelligence. High-resolution visuotactile sensors, such as GelSight [1] and DIGIT [2], have emerged to provide learned control policies with dense, rich contact signals. By fusing these tactile streams with vision and proprioception, recent policy learning frameworks have made significant strides in mastering complex, contact-rich tasks.

Learning visuotactile policies, however, poses challenges that vision-centric methods evade. Tactile sensing is inherently local and intermittent: a signal appears only on contact, yields information restricted to a minute patch of interaction, and often acts as the *only* informative cue precisely when the object of interest is occluded from the camera. A policy must therefore decide when to trust a sparse tactile stream over a rich but uninformative visual one, and how to fuse the two without letting the dominant visual modality wash out the tactile signal. **weihan**: are you trying to investigate why vision-tactile policy learning is difficult here? You need to have one topic sentence for this

The same properties that complicate policy learning also make touch difficult to model. While tactile sensing can intuitively prompt the world model to become deeply knowledgeable about real-world physics, capturing this modality presents a steep modeling hurdle. A visuotactile world model must predict a signal that is spatially local, sharply discontinuous at contact onset, and rich in fine-grained texture and force detail—structure that visually-oriented generative metrics neither capture

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

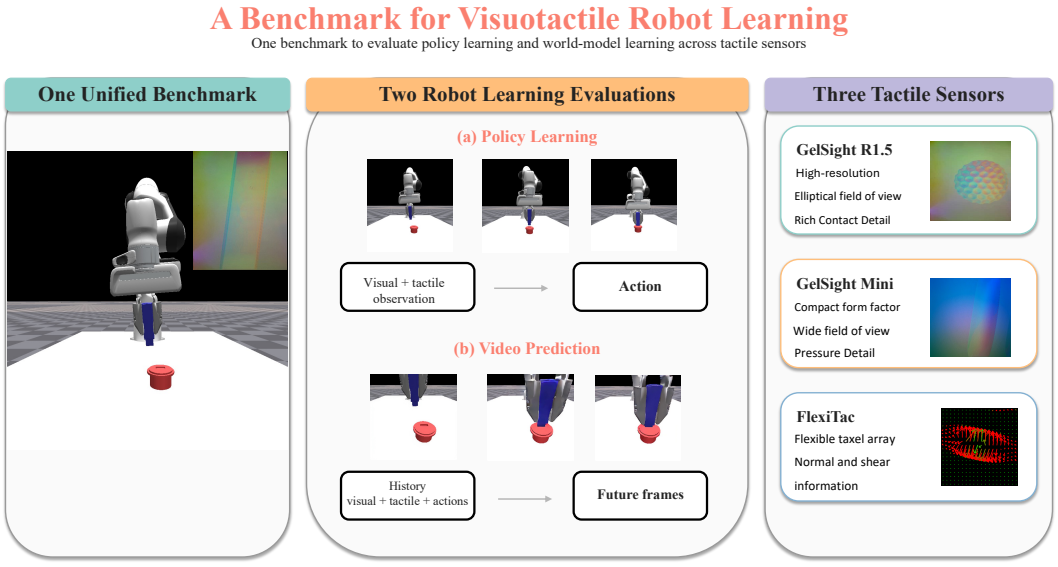

Figure 1: **Tac-Bench** unifies visuotactile policy learning and world-model evaluation on a single simulated platform. Fourteen contact-rich tasks span pick-and-place, insertion, assembly, classification, visual occlusion, in-hand orientation, and door manipulation, each instantiable with three tactile sensors (GelSight R1.5, GelSight Mini, FlexiTac) on a Franka platform.

nor penalize when it is wrong. Consequently, a predicted rollout can appear visually plausible while hallucinating non-existent contacts or depicting forces that fundamentally violate physics. **weihan**: Same here, You need to have one topic sentence why visual-tactile world model is difficult. probably to combine with last paragraph. Two domiant learning method in this domain include xxx. why they are challenging

Progress in this area, however, is hard to measure. Methods are typically evaluated on inconsistent tasks, sensors, and baselines, which obscures genuine progress.**weihan**: these two sentences do not sound scientific Two issues compound the problem. First, visuotactile policies use different fusion strategies, but they are evaluated on different tasks, different sensors, and different assumptions about what the tactile stream provides. It is unclear which architectural choices matter and when, because the architecture is never the only thing that varies. Second, visuotactile world models are typically judged by visual fidelity. Metrics such as the Fréchet Video Distance reward visually smooth rollouts that may still violate contact dynamics: phantom contacts, inconsistent forces, impossible motion. Consequently, a model can achieve strong visual scores while producing physically invalid dynamics.

weihan: In order to solve, we present xx. need topic sentences To address these issues, we present **Tac-Bench**, a unified benchmark for visuotactile robot learning. The benchmark targets the two issues above with two controlled protocols. *For policy fusion,* **weihan**: topic sentences? we provide contact-rich simulation tasks that isolate three failure modes for unimodal vision: occlusion, force sensitivity, and in-hand status requirement. By holding task, data, and unimodal encoders fixed and varying only the fusion module, performance differences are attributable to the architecture itself rather than to incidental task or data choices. **weihan**: too detailed, you don't need to say this in introduction *For physical realism in world models,* **weihan**: topic sentences? we assess functional validity by using the world model to imagine future goal states, decoding the required actions via a pretrained inverse-dynamics model, and executing them in a closed-loop simulation to measure downstream task success. We also assess functional consistency by leveraging the world model in a predictive optimization framework, measuring the reachability gain it provides over an unassisted baseline policy. Together, these probes strictly quantify whether the model's hallucinated dynamics are physically consistent and functionally useful for active downstream control.

Tac-Bench makes three contributions:

(1) A simulation suite of 14 contact-rich tasks built on TacSL, covering insertion, assembly, classification, in-hand orientation, and door manipulation, with three tactile sensor types (GelSight R1.5, GelSight Mini, FlexiTac) mounted on a Franka platform.

(2) A controlled comparison of visuotactile policy architectures, isolating fusion strategy from confounding task and data choices, together with the imitation-learning baselines.

(3) A physical-realism protocol for visuotactile world models that directly measures the physical utility through closed-loop rollout success and optimization-based reachability gains.

Tac-Bench reveals where tactile sensing is important for contact-rich manipulation and where current visuotactile models fail. We open-source the entire platform, including tasks and baselines, to serve as a rigorous foundation for future research in multimodal robot learning.

2 Related Works

Manipulation Benchmarks. General-purpose manipulation benchmarks have driven much of the recent progress in robot learning, from early platforms such as OpenAI Gym [3], Meta-World [4], and Robosuite [5] to more recent suites such as ManiSkill [6], RLBench [7], BEHAVIOR [8], and FurnitureBench [9]. None of these expose contact signals to the policy: tactile feedback is either abstracted away or absent. A second line of work targets touch directly. Optical-tactile sensors such as GelSight [1] and DIGIT [2] provide dense contact images and underpin most real-world tactile datasets, including YCB-Slide [10], ManiFeel [11], Tactile MNIST [12], and Sparsh [13]. These efforts are limited by the throughput of physical data collection: hardware wears out, ground-truth contact state is rarely accessible, and the resulting datasets are small. Handheld interfaces such as ViTaMIn [14] ease the data-collection bottleneck but still cannot scale to RL training, and partial-simulation efforts like VTDexManip [15] cover only narrow task slices. Tac-Bench fills this gap with a fully simulated, contact-rich task suite that delivers ground truth at scale.

Tactile Simulation. Accurately simulating contact physics has been a recurring challenge. To avoid the cost of finite-element methods, the community has converged on optical and approximation-based renderers: Taxim [16] uses data-driven look-up tables, TACTO [17] renders deformation maps with OpenGL, and FOTS [18] simulates the underlying force tensor for Sim2Real transfer. TacSL [19] extends this line with a GPU-accelerated pipeline that parallelizes high-resolution tactile rendering across thousands of environments. Tac-Bench builds on TacSL and uses it as the substrate for a standardized task suite for evaluating closed-loop visuotactile policies and world models.

Visuotactile Policies and World Models. Two open questions motivate Tac-Bench. The first is fusion: most visuotactile policies adopt one of three patterns—early concatenation, late fusion of modality-specific encoders, or cross-modal attention—and a complementary thread, such as Multi-Modal Policy Consensus [20], factorizes the policy itself by combining per-modality diffusion experts. Comparisons across these designs are difficult because prior work varies the task, sensor, and training pipeline together with the architecture. Tac-Bench fixes the task, data, and unimodal encoders, and varies only the fusion module. The second question is evaluation of visuotactile world models, which predict future observations conditioned on actions and enable model-based planning and offline RL [21, 22]. Standard generative metrics such as the Fréchet Video Distance [23] reward distributional similarity to real video and are blind to physical consistency: a rollout can score well on FVD while depicting phantom contacts or force-motion inconsistencies. Tac-Bench addresses this with two physics-aware probes: an inverse-dynamics action-consistency score that flags out-of-distribution decoded actions, and a policy-reachability test that measures whether expert policies retain control inside the imagined dynamics. Table 1 positions Tac-Bench against prior manipulation benchmarks along sensor coverage, task count, and skill breadth. **weihan**: why you put this sentence here?

Benchmark	Tasks	Sensor Type					Skill Coverage						
		GS	R15	Mini	FlexiTac	FC	PnP	I	A	C	VO	O	DM
RoboSuite	9	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
RLBench	106	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Meta-World	50	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
ManiSkill-ViTac	3	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
DiffHand	4	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
ManiFeel	6	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗
TacSL	3	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
VT-Refine	1	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗
Tac-Bench (Ours)	**14**	✗	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓

Sensor types: **GS**: GelSight 2017; **R15**: GelSight R1.5; **Mini**: GelSight Mini; **FlexiTac**: FlexiTac. **Skills:** **PnP**: Pick & Place; **I**: Insertion; **A**: Assembly; **C**: Classification; **VO**: Visual Occlusion; **O**: In-Hand Orientation; **DM**: Door Manipulation.

Table 1: Comparison of tactile sensor types and skill coverage across manipulation benchmarks.

3 Method

3.1 Simulation Platform and Data Collection

To enable systematic and reproducible research into visuo-tactile manipulation, we develop a high-fidelity simulation platform built upon a physics engine capable of modeling realistic contact dynamics. The simulator exposes three synchronized sensory streams at each timestep t: an RGB visual observation $\mathbf{v}_t \in \mathbb{R}^{H\times W\times 3}$ from one or more workspace cameras, a tactile observation $\boldsymbol{\tau}_t \in \mathbb{R}^{M\times M\times C}$ rendered from a simulated gel-based tactile sensor (e.g., a GelSight-style sensor) mounted on the robot's fingertip, and a proprioceptive state vector $\mathbf{p}_t \in \mathbb{R}^d$ comprising joint positions, joint velocities, and end-effector pose.

We construct a task suite of K contact-rich manipulation tasks spanning a spectrum of sensory demands. Tasks are selected to specifically stress-test conditions where unimodal policies are expected to fail:

(i) **Occlusion-dominant** scenarios, in which the grasped object is fully hidden from the camera during a critical phase,

(ii) **Force-critical** scenarios, such as peg insertion and surface wiping, requiring fine force modulation that is invisible to the camera, and

(iii) **In-hand-status-detection** scenarios, such as test tube arrangement, in which we need the tactile feedback to do the in-hand orientation to fulfill the task.

For each task, expert demonstrations are collected via a privileged controller with access to full ground-truth state, producing a dataset $\mathcal{D} = \{(\mathbf{v}_t, \boldsymbol{\tau}_t, \mathbf{p}_t, \mathbf{a}_t)\}$ used for imitation learning baselines and world model training.

3.2 Visuo-Tactile Policy Architectures

To isolate the effect of the multimodal fusion strategy, we define a shared backbone and systematically vary only the fusion module. All policies share identical unimodal encoders: a convolutional encoder $f_v(\cdot)$ for visual observations, a convolutional encoder $f_\tau(\cdot)$ for tactile observations, and a linear encoder $f_p(\cdot)$ for proprioception. Each encoder maps its respective input to a fixed-dimensional embedding $\mathbf{e}_v, \mathbf{e}_\tau, \mathbf{e}_p \in \mathbb{R}^d$, respectively.

We benchmark three fusion architectures shown in Figure 2, ranging from simple concatenation baselines to our proposed mixture-of-transformers design.

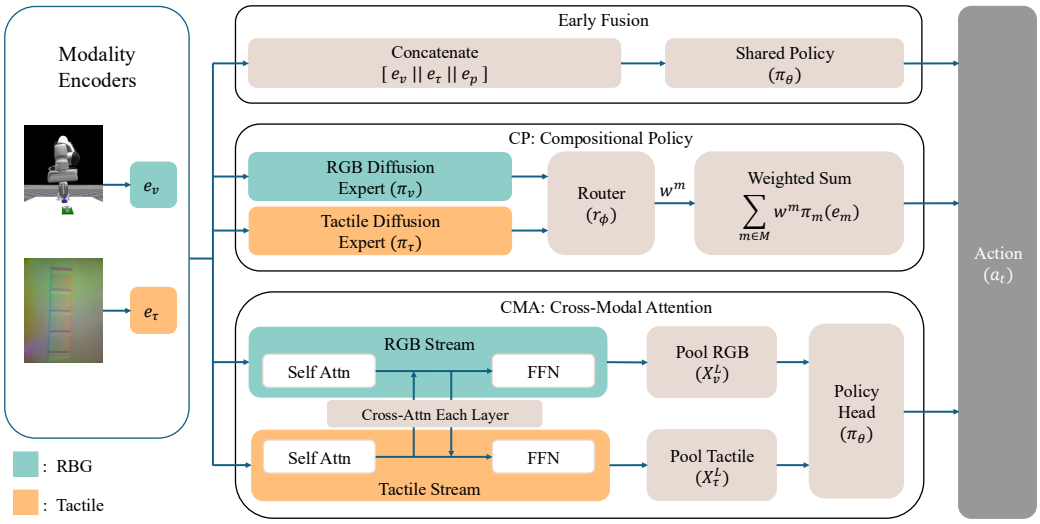

Figure 2: Visualization for three different policy architectures.

Early Fusion. All encoded modality features are concatenated immediately after the encoder stage and passed to a shared policy network π_θ:

$$\mathbf{z}_t = \big[\mathbf{e}_v \parallel \mathbf{e}_\tau \parallel \mathbf{e}_p\big], \quad \mathbf{a}_t = \pi_\theta(\mathbf{z}_t). \tag{1}$$

Despite its simplicity, this baseline is representative of the dominant approach in prior work. Its key limitation is that the fixed concatenation provides no mechanism for the model to up- or down-weight individual modalities at inference time, making it brittle when one stream is corrupted or occluded.

Multi-Modal Compositional Policy (CP) [20]. We include the method of Chen et al. [20] as a strong baseline. Rather than fusing features, CP factorizes the policy into a set of modality-specific diffusion models $\{\pi_m\}_{m \in \mathcal{M}}$, each specialized for a single representation (e.g., vision or touch). A learned router network r_ϕ produces consensus weights $\mathbf{w}_t \in \Delta^{|\mathcal{M}|-1}$ that adaptively combine the action score estimates from each diffusion model:

$$\mathbf{a}_t = \sum_{m \in \mathcal{M}} w_t^{(m)} \, \pi_m(\mathbf{a}_t \mid \mathbf{e}_m)\,, \tag{2}$$

where $\Delta^{|\mathcal{M}|-1}$ denotes the probability simplex over modalities. CP explicitly prevents dominant modalities such as vision from overwhelming sparse tactile signals, and supports incremental addition of new modalities without full retraining. It therefore serves as the most competitive late-fusion baseline in our study.

Cross-Modal Attention Fusion (CMA). Our proposed architecture is built on a **Mixture of Transformers (MoT)** design [24], in which each modality is assigned a dedicated Transformer stream rather than sharing a monolithic backbone. Formally, let $\mathbf{X}_m^{(\ell)} \in \mathbb{R}^{n_m \times d}$ denote the token sequence of modality $m \in \{v, \tau, p\}$ at layer ℓ. Each stream applies its own self-attention and feed-forward network:

$$\tilde{\mathbf{X}}_m^{(\ell)} = \mathrm{FFN}_m\Big(\mathrm{SelfAttn}_m\Big(\mathbf{X}_m^{(\ell)}\Big)\Big)\,. \tag{3}$$

At every S-th layer, cross-modal attention gates are inserted to allow controlled information exchange between streams. Concretely, visual tokens act as queries against tactile keys and values, capturing contact-relevant features that complement the global visual context:

$$\mathbf{X}_v^{(\ell+1)} = \tilde{\mathbf{X}}_v^{(\ell)} + \mathrm{CrossAttn}\Big(\mathbf{Q} = \tilde{\mathbf{X}}_v^{(\ell)},\ \mathbf{K} = \tilde{\mathbf{X}}_\tau^{(\ell)},\ \mathbf{V} = \tilde{\mathbf{X}}_\tau^{(\ell)}\Big)\,. \tag{4}$$

A symmetric gate is applied in the opposite direction so that tactile tokens can also attend to the visual context. After L layers, the final token representations from all streams are pooled and concatenated with the proprioceptive embedding before being passed to the policy head:

$$\mathbf{z}_t^{\mathrm{MoT}} = \left[\mathrm{Pool}(\mathbf{X}_v^{(L)}) \parallel \mathrm{Pool}(\mathbf{X}_\tau^{(L)}) \parallel \mathbf{e}_p\right], \quad \mathbf{a}_t = \pi_\theta\left(\mathbf{z}_t^{\mathrm{MoT}}\right). \tag{5}$$

By keeping modality-specific streams intact while injecting selective cross-modal signals only at periodic layers, CMA avoids the modality collapse observed in early fusion and retains the flexibility to up-weight tactile information in occlusion-dominant or force-critical scenarios.

All policies are trained with behavioral cloning on $\mathcal{D}$ using an ℓ_2 regression loss on continuous actions, and evaluated using task success rate over N rollouts. A structured ablation, removing individual modality streams and cross-attention gates, further quantifies the marginal contribution of each design choice.

3.3 Multimodal World Model Evaluation Protocol

Existing world models for robotic manipulation are predominantly evaluated with generative metrics such as the Fréchet Video Distance (FVD), which measure the distributional similarity of generated video sequences to real ones. While useful for visual fidelity, FVD is blind to physical consistency: a rollout can score well on FVD while depicting physically impossible contact events or violating force-motion relationships. We introduce an evaluation protocol that directly probes the **physical utility** of a world model across two axes.

3.3.1 Functional Validity via Closed-Loop Rollout Execution

To evaluate whether the generated sequences translate to effective, goal-directed behavior, we introduce a benchmark assessing **Closed-Loop Rollout Success (CLRS)**. Given a current visuo-tactile observation pair $(\mathbf{v}_t, \boldsymbol{\tau}_t)$, we leverage the pre-trained world model $\mathcal{W}$ to imagine a future goal state $(\hat{\mathbf{v}}_g, \hat{\boldsymbol{\tau}}_g)$. The pre-trained inverse dynamics model $\mathcal{I}_\psi$ is then utilized as a goal-conditioned policy to decode the immediate action $\hat{\mathbf{a}}_t$ required to reach that generated goal:

$$\hat{\mathbf{a}}_t = \mathcal{I}_\psi(\mathbf{v}_t, \boldsymbol{\tau}_t, \hat{\mathbf{v}}_g, \hat{\boldsymbol{\tau}}_g). \tag{6}$$

This decoded action is executed directly within the simulation environment to transition the true state. By repeating this process in a closed-loop fashion over a horizon T, we measure the task's final **Success Rate**. CLRS strictly quantifies the functional utility of the world model's generations when deployed for active downstream control.

3.3.2 Functional Consistency via Optimization-Based Reachability Gain

To assess whether the world model functions effectively as a planning substrate, we evaluate the performance gain achieved through world-model-guided trajectory optimization. We first establish a baseline by rolling out a pre-trained policy directly within the true environment and recording its unassisted success rate, denoted as S_{base}.

We then introduce a model-predictive optimization framework leveraging the predictive capabilities of the world model. At each execution step, the pre-trained policy is treated as a proposal distribution from which we sample $M = 16$ candidate action sequences, $\{\mathbf{a}_{t:t+H}^{(i)}\}_{i=1}^{M}$, over a planning horizon H. Each candidate sequence is rolled out through the action-conditioned world model $\mathcal{W}^{\text{cond}}$ to predict a corresponding final observation $\hat{\mathbf{v}}_{t+H}^{(i)}$. Given a pre-defined goal image $\mathbf{v}^*$, we evaluate the candidate trajectories by calculating the Mean Squared Error (MSE) between their predicted final states and the goal state. The optimal action sequence is chosen by selecting the sample that minimizes this visual loss:

$$i^* = \arg\min_i \left\| \hat{\mathbf{v}}_{t+H}^{(i)} - \mathbf{v}^* \right\|_2^2. \tag{7}$$

The first action of the selected sequence, $\mathbf{a}_t^{(i^*)}$, is executed in the environment. This optimization process is repeated in a closed-loop manner, yielding an optimized success rate S_{opt}.

We define the **Reachability Gain** ($\mathcal{R}$) as the absolute improvement in task success rate brought by the optimization framework:

$$\mathcal{R} = S_{\text{opt}} - S_{\text{base}}. \tag{8}$$

A world model with physically inconsistent or drifting dynamics will provide inaccurate MSE gradients during planning, leading to sub-optimal action selection and a low or negative reachability gain. Conversely, a high $\mathcal{R}$ score signifies that the world model's imagined dynamics are accurate enough to consistently refine and enhance policy execution.

Unified evaluation. Together, the components above form a unified evaluation framework for both policy architectures and world models. For policy evaluation, we report task success rate broken down by task category (occlusion, force-critical, in-hand status) and modality ablation. For world model evaluation, we report the standard FVD alongside our proposed CLRS and $\mathcal{R}(\mathcal{W}^{\text{cond}}, \pi^*)$ to provide a complete picture of both visual and physical quality. This multi-axis protocol ensures that progress in the field can be measured not only by how realistic generated rollouts appear, but by whether they are grounded in physical reality and useful for downstream robotic control.

4 Tactile Benchmark

Tac-Bench includes 14 contact-rich tasks grouped into seven skill classes: pick-and-place, insertion, assembly, classification, visual occlusion, in-hand orientation, and door manipulation. Each task is designed so that at least one phase requires tactile feedback to solve reliably. Figure 3 shows representative tasks from the suite.

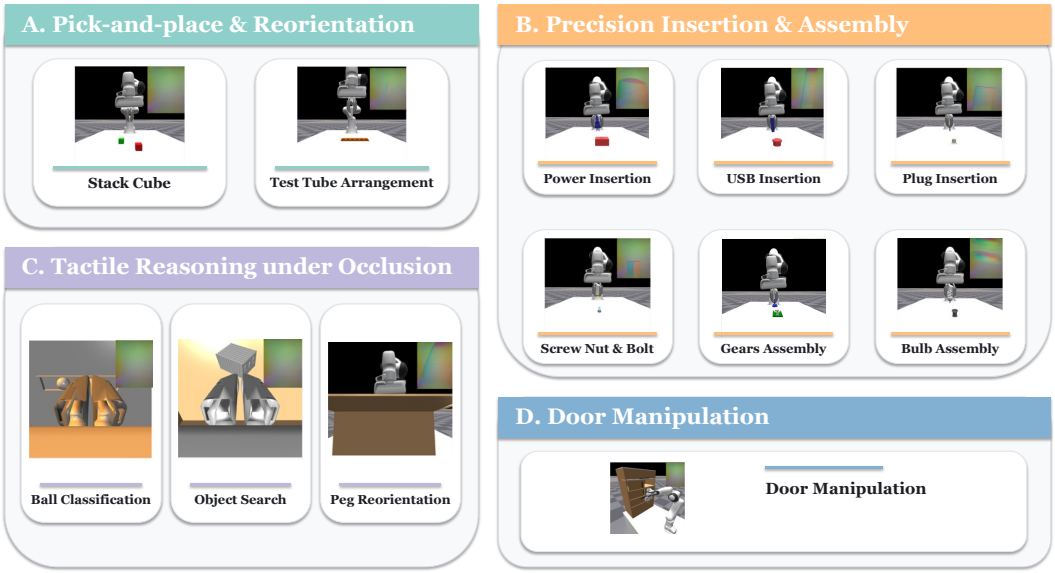

Figure 3: **Tac-Bench** includes 14 contact-rich tasks grouped into seven skill classes: pick-and-place, insertion, assembly, classification, visual occlusion, in-hand orientation, and door manipulation.

4.1 Task Suite

Pick-and-place. *Stack Cube* requires lifting and stacking a cube, with grip-force modulation under partial visual occlusion.

Insertion. *Peg*, *USB*, and *Plug Insertion* require sub-millimeter alignment of a connector into a fixed socket. The connector occludes the contact site once it enters the hole.

Assembly. *Screw Nut & Bolt* requires threading a nut onto a bolt; *Gears Assembly* requires meshing two gears with correct tooth alignment; *Bulb Assembly* requires placing the bulb into the socket.

Classification. *Ball Classification* requires inferring a hidden object property from tactile interaction alone.

Visual Occlusion. *Object Search* requires locating an object using tactile feedback alone while vision is occluded. *Blind Insertion* requires searching for the target socket and inserting an object into it with no informative visual input.

In-hand orientation. *Test Tube Arrangement* requires reorienting a tube in the gripper before placing it.

Door manipulation. *Sliding Door Open*, *Pulling Door Open*, and *Hinge Door Open* require maintaining contact while pulling or sliding, with the hand-handle interface largely occluded.

For each task we provide: a reset distribution, a dense reward used for RL, a sparse success criterion, and a privileged expert that generates demonstration data for imitation learning and world-model training. A subset of tasks have tuned RL reward shaping; the remaining tasks are evaluated only with imitation learning. Each task can be instantiated with any of the three tactile sensors from Section A.

5 Benchmarking Results

We report results across two axes: imitation-learning baselines with a controlled comparison of visuotactile fusion architectures, and the visuotactile world-model evaluation protocol. All numbers below are the mean success rate (or metric value) over N evaluation episodes; standard deviations are computed over three seeds.

5.1 Imitation Learning Baselines and Fusion Ablation

For each task, we collect 200 expert demonstrations from a privileged controller to establish imitation learning baselines and perform a direct ablation of visuotactile fusion architectures. To ensure a fair central comparison, we hold the dataset and unimodal encoders fixed, varying only the policy architecture and fusion module. We evaluate our *CMA* mixture-of-transformers design against a standard Diffusion Transformer policy, a Unet Diffusion Policy utilizing an *Early Fusion* encoder, and the Multi-Modal Policy Consensus (*CP*) method of Chen et al. [20].

Task	Transformer	Early Fusion (Diffusion)	CP [20]	CMA (Ours)
Peg Insertion	0.12	0.06	0.14	0.10
Usb Insertion	0.22	0.06	0.08	0.08
Plug Insertion	0.18	0.10	0.14	0.10
Ball Classification	0.66	0.74	0.82	0.68
Assemble Gears	0.10	0.02	0.06	0.04
Assemble Bulb	0.04	0.04	0.02	0.06
Screw Nut & Bolt	0.22	0.16	0.28	0.06

Table 2: Imitation learning success rates (↑) and fusion architecture ablation. All policies share identical encoders and training data (200 demonstrations per task), with changes restricted to the fusion module and base policy architecture. Evaluated over 50 rollouts $\times$ 3 seeds.

5.2 Visuotactile World Model Evaluation

We evaluate the visuo-tactile world model along four primary metrics: Fréchet Video Distance (FVD, lower is better) to evaluate video quality, the Closed-Loop Rollout Success (CLRS, higher is better) to quantify downstream functional validity and measure physical plausibility, and the policy reachability rate $\mathcal{R}$ (higher is better) to assess imagination-based planning.

6 Conclusion

World Model	FVD ↓	CLRS (%) ↑	$\mathcal{R}$ ↑
Visual-only baseline	—	—	—
Visuotactile (concat)	—	—	—
Visuotactile (cross-attn)	—	—	—

Table 3: World-model evaluation across video fidelity, physical consistency, and functional utility. FVD measures distributional visual quality; CLRS measures the closed-loop success rate when executing world-model-guided trajectories in the actual simulation; $\mathcal{R}$ measures whether an expert policy can navigate inside the imagined dynamics.

World Model (CLRS ↑)	Peg Insertion	USB Insertion	Plug Insertion	Gear Assembly	Bulb Assembly
Visual-only baseline	0.49	0.22	0.06	0.67	0.00
Visuotactile (concat)	0.58	0.27	0.04	0.77	0.00

Table 4: Task-specific downstream rollout evaluation. This task set measures the actual closed-loop execution success rate (CLRS) inside the simulation environment.

World Model	Sorting	Peg Insertion	Bulb Assembly	USB Insertion
Visual-only baseline	14	18	6	—
Visuotactile (concat)	16	14	12	—

Table 5: Policy reachability evaluation ($\mathcal{R}$ ↑) evaluated across a distinct set of validation benchmarks to quantify the navigational capabilities of an expert policy within the imagined latent dynamics.

References

[1] W. Yuan, S. Dong, and E. H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors (Basel, Switzerland)*, 17, 2017.

[2] M. Lambeta, P. wei Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5:3838–3845, 2020.

[3] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym, 2016. URL https://arxiv.org/abs/1606.01540.

[4] R. McLean, E. Chatzaroulas, L. McCutcheon, F. Röder, T. Yu, Z. He, K. Zentner, R. Julian, J. K. Terry, I. Woungang, N. Farsad, and P. S. Castro. Meta-world+: An improved, standardized, RL benchmark. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=1de3azE606.

[5] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, K. Lin, A. Maddukuri, S. Nasiriany, and Y. Zhu. robosuite: A modular simulation framework and benchmark for robot learning, 2025. URL https://arxiv.org/abs/2009.12293.

[6] T. Mu, Z. Ling, F. Xiang, D. C. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=zQIvkXHS_U5.

[7] S. James, Z. Ma, D. Rovick Arrojo, and A. J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.

[8] S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, K. Liu, S. Savarese, H. Gweon, J. Wu, and L. Fei-Fei. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 477–490. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/srivastava22a.html.

[9] M. Heo, Y. Lee, D. Lee, and J. J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In *Robotics: Science and Systems*, 2023.

[10] S. Suresh, Z. Si, S. Anderson, M. Kaess, and M. Mukadam. MidasTouch: Monte-Carlo inference over distributions across sliding touch. In *Proc. Conf. on Robot Learning, CoRL*, Auckland, NZ, Dec. 2022.

[11] Q. K. Luu, P. Zhou, Z. Xu, Z. Zhang, Q. Qiu, and Y. She. Manifeel: Benchmarking and understanding visuotactile manipulation policy learning. *arXiv preprint arXiv:2505.18472*, 2025.

[12] T. Schneider, G. Duret, C. de Farias, R. Calandra, L. Chen, and J. Peters. Tactile MNIST: Benchmarking active tactile perception, 2026. URL https://openreview.net/forum?id=QXq5BoaNPI.

[13] C. Higuera, A. Sharma, C. K. Bodduluri, T. Fan, P. Lancaster, M. Kalakrishnan, M. Kaess, B. Boots, M. Lambeta, T. Wu, and M. Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. 2024. URL https://openreview.net/forum?id=xYJn2e1uu8.

[14] F. Liu, C. Li, Y. Qin, J. Xu, P. Abbeel, and R. Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface, 2025. URL https://arxiv.org/abs/2504.06156.

[15] Q. Liu, Y. Cui, Z. Sun, G. Li, J. Chen, and Q. Ye. VTDexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jf7C7EGw21.

[16] Z. Si and W. Yuan. Taxim: An example-based simulation model for gelsight tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):2361–2368, 2022. doi:10.1109/LRA.2022.3142412.

[17] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra. TACTO: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters (RA-L)*, 7(2):3930–3937, 2022. ISSN 2377-3766. doi:10.1109/LRA.2022.3146945. URL https://arxiv.org/abs/2012.08456.

[18] Y. Zhao, K. Qian, B. Duan, and S. Luo. Fots: A fast optical tactile simulator for sim2real learning of tactile-motor robot manipulation skills. *IEEE Robotics and Automation Letters*, pages 5647 – 5654, May 2024. ISSN 2377-3766. doi:10.1109/LRA.2024.3396665.

[19] I. Akinola, J. Xu, J. Carius, D. Fox, and Y. Narang. Tacsl: A library for visuotactile sensor simulation and learning. *IEEE Transactions on Robotics*, 41:2645–2661, 2025. doi:10.1109/TRO.2025.3547267.

[20] H. Chen, J. Xu, H. Chen, K. Hong, B. Huang, C. Liu, J. Mao, Y. Li, Y. Du, and K. Driggs-Campbell. Multi-modal manipulation via multi-modal policy consensus. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[21] H. Chen, J. Xu, L. Sheng, T. Ji, S. Liu, Y. Li, and K. Driggs-Campbell. Learning coordinated bimanual manipulation policies using state diffusion and inverse dynamics models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[22] K. Hong, H. Chen, J. Xu, R. Wang, K. Wang, M. Zhang, S. Liu, Y. Zhu, Y. Li, and K. Driggs-Campbell. Gotta Scoop 'Em All: Sim-and-Real Co-Training of Graph-Based Neural Dynamics for Long-Horizon Scooping. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026. To appear.

[23] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv*, abs/1812.01717, 2018. URL https://api.semanticscholar.org/CorpusID:54458806.

[24] W. Liang, L. Yu, L. Luo, S. Iyer, N. Dong, C. Zhou, G. Ghosh, M. Lewis, W. tau Yih, L. Zettlemoyer, and X. V. Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models, 2025. URL https://arxiv.org/abs/2411.04996.

[25] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. URL https://arxiv.org/abs/2304.13705.

A Simulated Robots and Tactile Sensors

A.1 Robot Platforms

All tasks in the current suite run on the **Franka Panda**, a 7-DoF arm with a parallel-jaw gripper. We control the arm with an operational-space impedance controller exposing $(\Delta x, \Delta y, \Delta z, \Delta r_x, \Delta r_y, \Delta r_z)$ Cartesian deltas, and the gripper through a binary open/close command. The platform layer also supports the **ALOHA** bimanual setup [25]—two 6-DoF arms with

parallel-jaw grippers under joint-space PD control—as infrastructure for future bimanual extensions, though all tasks reported in this paper are single-arm.

A.2 Tactile Sensors

Three tactile sensors are simulated, covering both optical and force-based families.

- **GelSight R1.5** (R15). High-resolution optical-tactile sensor with a thick elastomer gel. We render a $320 \times 240 \times 3$ RGB image of the deformed gel surface per finger.
- **GelSight Mini**. Compact optical-tactile sensor with a smaller contact patch. Rendered at $320 \times 240 \times 3$.
- **FlexiTac**. Resilience-based piezoresistive array. Each finger exposes a $C \times C$ pressure map sampled from the contact patch.

All three sensors share a common interface so that the same task can be run with any sensor by changing a single configuration field.

A.3 Observation and Action Spaces

At each timestep t, the simulator returns

$$\begin{aligned} \mathbf{v}_t &\in \mathbb{R}^{H \times W \times 3} && \text{(workspace RGB camera)} \\ \boldsymbol{\tau}_t &\in \mathbb{R}^{M \times M \times C} && \text{(tactile observation, per finger)} \\ \mathbf{p}_t &\in \mathbb{R}^{d} && \text{(joint positions, velocities, end-effector pose)} \end{aligned}$$

where $(H, W) = (224, 224)$ for the workspace camera and (M, C) depends on the sensor. The action space matches the chosen controller above. Episode length and success criteria are task-specific (Section 4).

A.4 Tactile Rendering Pipeline

The tactile simulation pipeline generates the tactile signal in two sequential stages. First, a signed-distance-function (SDF) query between the indenter mesh and the gel surface yields per-pixel penetration depths and surface normals. Second, a Taxim-style data-driven renderer [16] maps these geometric quantities to the appearance of each sensor, including illumination and elastomer-specific marker fields for the GelSight variants and a calibrated pressure response for FlexiTac. The pipeline runs on the GPU in parallel across environments, following TacSL [19], which lets the benchmark scale to thousands of parallel rollouts for RL.