Tactile-WAM: Touch-Aware World Action Model with Tactile Asymmetric Attention

Siyu Wu[1,4,*], Linjing You[1,2,3,*], Junjie Zhu[1,†], Yaozu Liu[2], Huang Kaixiang[4], Chen Yonghang[4], Jituo Li[4], Changhao Zhang[1], Jian Liu[1], Zhu Hengshuo[1], Qi Li[2], Hengshuang Zhao[3]

[1]Ant Group
[2]Institute of Automation, Chinese Academy of Sciences
[3]The University of Hong Kong
[4]Zhejiang University

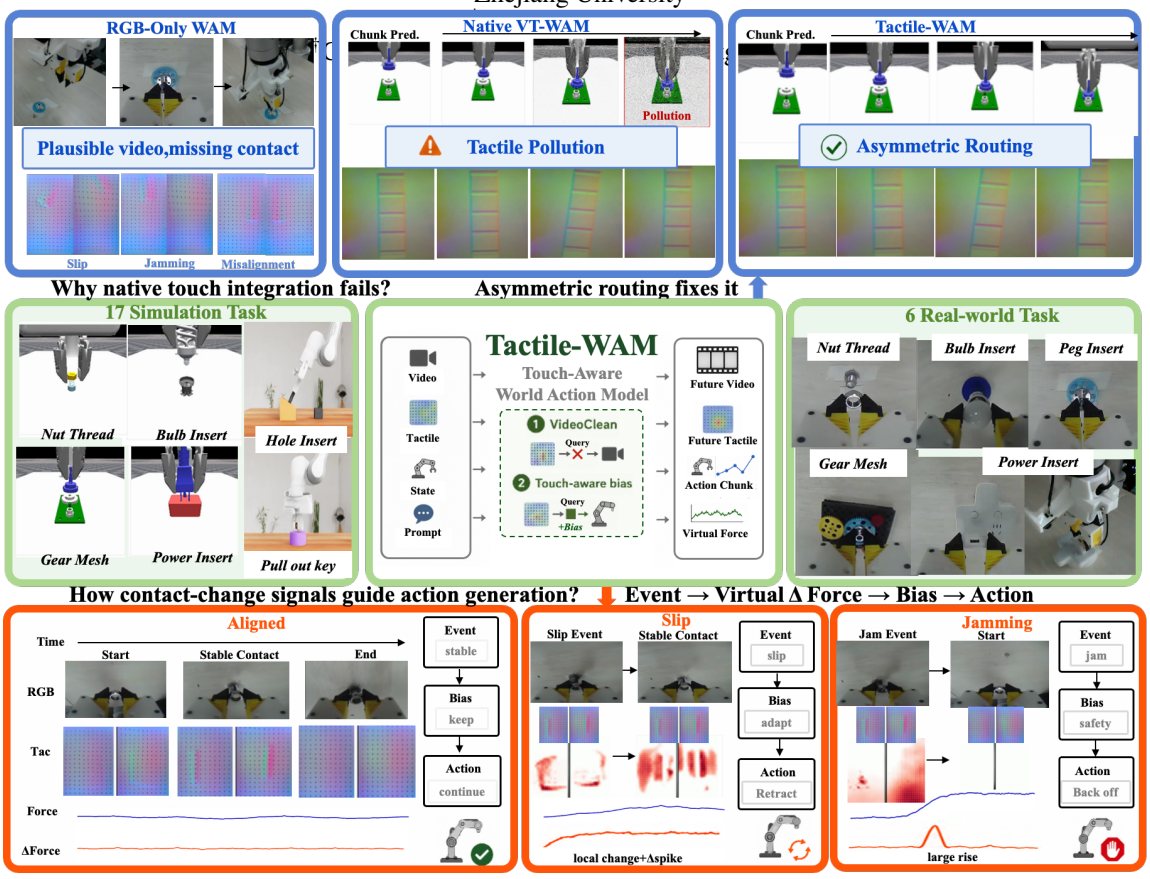

Figure 1: System overview. The top row contrasts RGB-only prediction, naive visuo-tactile modeling, and asymmetric routing; the middle summarizes the model inputs, outputs, and task suites; the bottom shows aligned, slip, and jamming contact traces.

Abstract

World Action Models (WAMs) jointly predict future visual observations and actions, but visual futures alone often miss slip, jamming, contact-direction changes, and subtle misalignment in contact-rich manipulation. Tactile signals reveal these hidden physical states, yet naive tactile-token injection can disrupt visual dynamics modeling due to the limited scale of tactile data, a phenomenon we term **tactile pollution**. We introduce **Tactile-WAM**, which uses asymmetric attention to block video queries from tactile keys while preserving tactile access for action queries. A contact-change-aware bias further strengthens action attention to touch. Because tactile pixel changes do not reliably reflect contact changes, we derive a six-dimensional touch-aware proxy from tactile optical flow. Observed proxy changes drive the attention bias, while future-proxy supervision preserves action-relevant contact dynamics in predicted tactile representations. On ManiFeel, visual-path isolation reduces deviation from the RGB-only trajectory by 21.8% in MSE at the step-matched 20K checkpoint without a statistically detectable change in ground-truth video quality. The full model improves average success from 15.6% to 32.7%, with VideoClean providing the largest gain. On five real-robot tasks, Tactile-WAM achieves 49.2% success.

1 Introduction

World Action Models (WAMs) jointly predict future states and robot actions, enabling policies to anticipate the consequences of an action sequence (Ye et al. 2026; Guo et al. 2024; Hu et al. 2024). Through large-scale video pretraining, they acquire strong visual appearance and motion priors. However, their predicted futures remain predominantly visual and cannot fully capture the physical states involved in contact-rich manipulation.

In tasks such as insertion, assembly, and object reorientation, success often depends on slip, jamming, contact-direction changes, and subtle misalignment. These events may be difficult to observe from RGB images but directly determine the next corrective action, as illustrated in Figure 1. Vision-based tactile sensors capture local deformation and contact transitions (Yuan, Dong, and Adelson 2017; Lambeta et al. 2020; Ward-Cherrier et al. 2018), and learned tactile representations have shown clear benefits in contact-rich manipulation (Xu et al. 2024; Higuera et al. 2024; Xue et al. 2025; Luu et al. 2025). A WAM for such tasks should therefore predict future tactile states and effectively use touch for action generation.

Directly incorporating touch, however, is not always beneficial. Tactile datasets are far smaller than video pretraining corpora, and tactile signals mainly describe sparse, local contact events. Unconstrained attention from video queries to tactile keys can therefore interfere with the pretrained visual representation and degrade both video and action prediction. We term this phenomenon *tactile pollution*. As shown in Figure 1, naive tactile fusion produces blur and distortion in predicted visual futures.

We further observe that pixel-level changes in tactile images do not reliably reflect actual contact changes. Small pixel differences may correspond to critical events such as slip, compression, or slight in-gripper rotation, while large pixel differences may result from visually prominent but physically minor variations (Chen, Long, and Shang 2026; Chen et al. 2026). Our analysis shows only weak correlation between tactile pixel similarity and deformation-derived contact similarity. Thus, effective tactile modeling must determine when touch is important and preserve contact-relevant dynamics in predicted tactile representations.

Motivated by these observations, we introduce Tactile-WAM, a Wan2.2-based tactile-aware world action model. Its core component, the Tactile Asymmetric Attention Mechanism (TAAM), prevents video queries from attending to tactile keys while preserving tactile access for action queries. We further derive a six-dimensional touch-aware proxy from optical flow between consecutive tactile images. Observed proxy changes generate a causal attention bias that strengthens action attention to touch, while future-proxy supervision encourages predicted tactile representations to preserve action-relevant contact dynamics.

Our main contributions are as follows:

- We identify and quantify *tactile pollution* in visually pretrained WAMs and introduce TAAM to protect video prediction while retaining tactile information for action generation;
- We reveal the mismatch between tactile pixel changes and contact changes, and propose an optical-flow-based touch-aware proxy for adaptive attention and future tactile supervision;
- We validate Tactile-WAM on nine simulation tasks and five real-robot tasks, demonstrating substantial improvements in contact-rich manipulation while preserving visual prediction quality.

2 Preliminaries and Notation

We briefly review World Action Models (WAMs) and vision-based tactile sensing to establish the notation used throughout this paper. Detailed related work is deferred to Appendix A.

World Action Models

WAMs jointly predict future visual observations and robot actions, conditioning action generation on environment evolution (Ye et al. 2026; Guo et al. 2024; Hu et al. 2024). At policy-call time t, a WAM receives RGB history $o^v_{\le t}$, proprioceptive states $s_{\le t}$, and language instruction ℓ, collectively denoted as

$$\mathcal{C}_t = \left(o^v_{\le t}, s_{\le t}, \ell\right). \tag{1}$$

Let E_v encode a future RGB sequence $o^v_{t+1:t+H}$ (horizon H) into latent $z^v_0 = E_v(o^v_{t+1:t+H})$, and let action chunk $a_0 = a_{t:t+K-1}$ (K: action horizon).

Following conditional flow matching (**?**), WAMs transform Gaussian noise into future latents and actions. For each target x^m_0 ($m \in \{v, a\}$), with noise $\epsilon^m \sim \mathcal{N}(0, I)$ and timestep $\rho \in [0, 1]$, the interpolant and velocity are

$$x^m_\rho = (1-\rho)x^m_0 + \rho\epsilon^m, \qquad u^m_\rho = \epsilon^m - x^m_0, \tag{2}$$

where $x^v_0 = z^v_0$ and $x^a_0 = a_0$. Conditioned on $\mathcal{C}_t$, the model predicts vector field v^m_θ via

$$\mathcal{L}^m_{\mathrm{FM}} = \mathbb{E}\left[\left\| v^m_\theta(x^m_\rho, \rho; \mathcal{C}_t) - u^m_\rho \right\|^2_2\right]. \tag{3}$$

Inference transforms sampled noise into future visual latents and executable actions.

Vision-Based Tactile Sensing

Vision-based tactile sensors capture contact-induced surface deformation via an internal camera (Yuan, Dong, and Adelson 2017; Lambeta et al. 2020; Ward-Cherrier et al. 2018). Let

$$o^{\tau,r}_t \in \mathbb{R}^{H_\tau \times W_\tau \times 3}, \qquad r \in \{L, R\}, \tag{4}$$

be the tactile image from the left/right sensor at time t, with history $o^{\tau,r}_{\le t}$. These images reflect local contact geometry, pressure, and temporal dynamics (e.g., shear, slip), but—unlike calibrated force measurements—are high-dimensional visual observations that do not directly yield physical force values.

3 Key Observations

Tactile Pollution

A straightforward approach to incorporate tactile information into WAMs is to jointly model tactile observations with

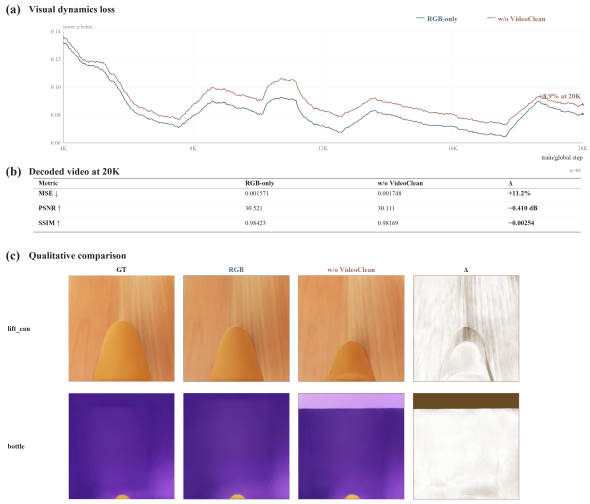

Figure 2: Tactile pollution on UniVTAC. (a) Training loss. (b) Prediction MSE vs. ground truth. (c) Qualitative comparison.

visual inputs and actions. In DiT-based architectures, tokens from all modalities are concatenated into a shared sequence, allowing visual queries to attend to tactile keys. For visual tokens, the attention operation is:

$$\mathrm{Attn}(Q^v, K, V) = \mathrm{Softmax}\left(\frac{Q^v K^\top}{\sqrt{d}}\right) V, \tag{5}$$

where Q^v accesses key-value representations from all modalities:

$$K, V = [K^v, K^\tau, K^a, K^s], \tag{6}$$

with K^v, K^τ, K^a, K^s denoting keys of visual, tactile, action, and state tokens, respectively. We refer to this direct fusion as *Naive VT-WAM*.

However, unrestricted tactile injection degrades visual prediction and action generation. On UniVTAC (800 training trajectories), Naive VT-WAM consistently exhibits higher training loss than DreamZero (Fig. 2(a)), yields 50% higher MSE in future frame prediction (Fig. 2(b)), and produces visible blur around manipulated objects (Fig. 2(c)). These results indicate that fine-tuning a visually pretrained WAM with limited tactile data interferes with the visual dynamics prior. We term this *Tactile Pollution*.

Pixel–Contact Misalignment

We further question whether pixel reconstruction—effective for RGB video generation—is suitable for tactile prediction. Unlike visual appearance, tactile sensing primarily captures local contact states and their transitions (slip, compression, object motion), making pixel-level changes an unreliable proxy for contact dynamics.

To verify this, we sample 5,000 tactile image pairs and analyze pixel variation against contact variation (Fig. 3). No clear linear correlation is observed. Samples in the blue region show large pixel changes despite limited contact variation, while the red region exhibits the opposite. Samples on

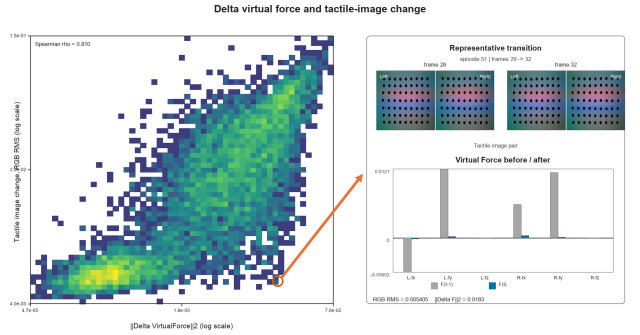

Figure 3: Pixel–contact misalignment. Each point denotes a pair of adjacent tactile images; x-axis: contact variation, y-axis: pixel variation.

the yellow dashed line share similar pixel variations but differ substantially in contact changes, and those on the green dashed line show the reverse.

These observations indicate that tactile representation space does not reliably reflect contact dynamics. Thus, pixel reconstruction objectives alone cannot ensure predicted tactile representations preserve contact changes critical for action generation.

4 Method

Architecture Overview

Figure 4 illustrates the overall framework of TACTILE-WAM. RGB and tactile history sequences are independently encoded into separate visual and tactile latent variables via the frozen Wan2.2 VAE, with corresponding modality embeddings added. The language-conditioned Wan DiT then jointly predicts future video frames, future tactile states, and action chunks within a unified framework.

The remainder of this section is organized as follows. We first describe the joint visual-tactile-action modeling based on Wan (Section 4.2). To address the tactile pollution and pixel–touch misalignment issues identified in Section 4.3, we introduce a tactile asymmetric attention mask (Section 4.4) and a touch-aware proxy (Section 4.5), respectively. Specifically, the mask module blocks visual queries from accessing tactile keys to mitigate tactile pollution, while simultaneously encouraging action queries to attend more to tactile information during contact changes via a touch-aware attention bias. To resolve pixel–touch misalignment, we employ a touch-aware proxy to derive the attention bias and additionally supervise the predicted tactile images with proxy signals during training. Further architectural details and training configurations are provided in Appendix B.

Joint Visual-Tactile-Action Modeling

To incorporate vision-based tactile sensing into the unified world-action-model framework, tactile observations are first encoded into a latent space shared with vision. Specifically, the left and right tactile images $o_t^{\tau,L}, o_t^{\tau,R}$ (Eq. (6)) are resized, horizontally mosaicked via $\mathcal{M}$, and encoded by the

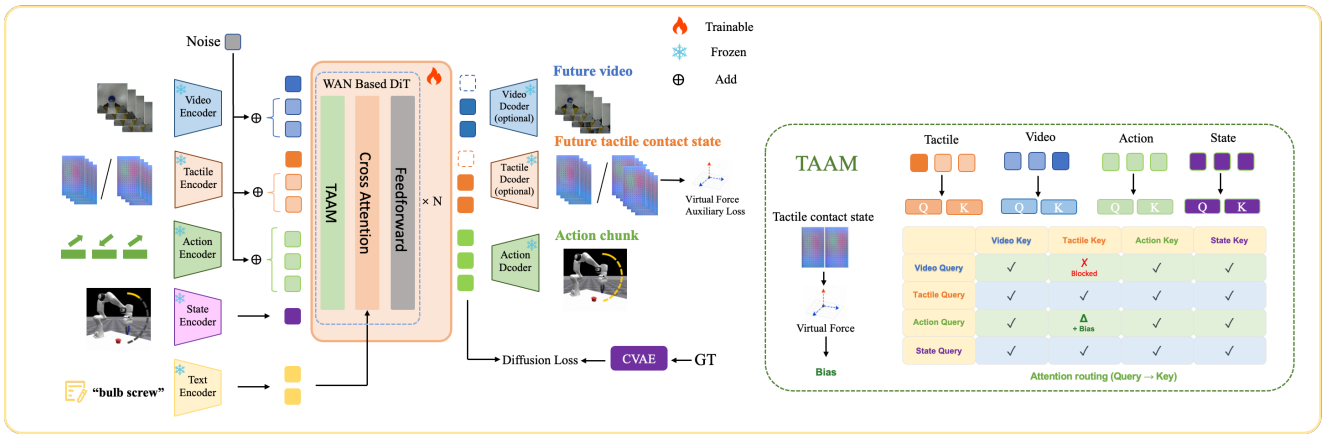

Figure 4: Overview of TACTILE-WAM. RGB and tactile histories are encoded by a shared frozen VAE, and a Wan DiT jointly predicts future visual latents, tactile latents, and action chunks. A VIDEOCLEAN mask prevents visual queries from accessing tactile keys, while a touch-aware bias routes action queries to tactile features within each causal block. A differentiable deformation proxy supervises tactile prediction and drives the bias computation.

causal VAE encoder E_{Wan} of the Wan2.2 backbone:

$$z^{\tau} = E_{\text{Wan}}\big(\mathcal{M}(o^{\tau,L}, o^{\tau,R})\big)\,, \tag{7}$$

where the future prediction target is denoted as z_0^{τ}, defined in parallel with the visual latent z_0^{v} in Eq. (2).

The conditional context is accordingly extended to explicitly incorporate the left and right tactile histories:

$$\mathcal{C}_t^{\tau} = \Big(o_{\le t}^{v},\ \{o_{\le t}^{\tau,r}\}_{r=L,R},\ s_{\le t},\ \ell\Big)\,, \tag{8}$$

where historical tactile latents, after learnable linear projection and modality embedding, are injected into the denoising network as auxiliary conditional signals.

Training and inference for the tactile modality directly follow the flow-matching paradigm defined in Section 2: substituting the conditional context in Eq. (4) with $\mathcal{C}_t^{\tau}$ yields the tactile flow-matching loss $\mathcal{L}_{\text{FM}}^{\tau}$; during inference, the clean latent $\widehat{z}_0^{\tau}$ is reconstructed from the sampled noise ϵ^{τ} following Eq. (5), and then decoded by D_{Wan} to produce the predicted tactile images. Each action chunk is temporally aligned with the corresponding tactile latent frames. Finally, the flow-matching losses for vision, action, and tactile modalities are jointly optimized.

Tactile Asymmetric Attention Mask

Naive symmetric routing allows visual queries to directly attend to sparse tactile keys. Since local tactile events do not necessarily predict global future appearance, this pathway can corrupt the pre-trained visual dynamics representation, leading to the tactile pollution failure mode. To address this, we propose a tactile asymmetric attention mask that protects visual prediction while retaining tactile information for action generation.

VIDEOCLEAN Mask. Let $G(q)$ and $G(k)$ denote the token groups to which query q and key k belong. We block only the access from visual queries to tactile keys:

$$M_{qk}^{\text{vc}} = \begin{cases} -\infty, & G(q) = V \ \wedge\ G(k) = T, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Action queries retain full access to tactile keys, and tactile queries preserve their multimodal context. Thus, this mask protects the visual prediction pathway in a unidirectional manner, rather than completely isolating modalities.

Touch-aware Attention Bias. To enable action prediction to perceive the current contact state, we design a touch-aware attention bias that guides action queries to attend to tactile representations within the same temporal block. The tactile sequence is partitioned into causal blocks according to the action-chunk boundaries, where each block relies solely on historical observations up to its end time to ensure causality.

For each tactile causal block c, we extract a touch-aware perception vector from the tactile frames at the previous and current policy-calling instants:

$$F_c^{\text{obs}} = \Phi(o_{u_c-8}^{\tau}, o_{u_c}^{\tau}), \tag{10}$$

where $\Phi(\cdot)$ denotes the differentiable deformation estimator detailed in Section 4, and u_c is the anchor timestamp of the block. To obtain a compact gating signal, we convert this high-dimensional vector into a contact intensity score:

$$s_c = \tanh\left(\frac{\text{ReLU}(\|F_c^{\text{obs}}\|_2 - \theta)}{T}\right), \tag{11}$$

which is normalized to $[0, 1)$ via threshold θ and temperature T, reflecting the presence and strength of contact. Each tactile patch key k within the block receives the following bias enhancement:

$$b_{c,k} = \text{clip}(\alpha s_c, 0, b_{\max})\,, \tag{12}$$

where α controls the amplification factor and $b_{\max}$ caps the upper bound. This bias acts as a block-level modality gate, applied between all action queries and tactile keys. Let I_{qk}^{c} indicate whether action query q and tactile key k belong to the same block (1 if yes, 0 otherwise). The final biased attention logit is:

$$A'_{qk} = A_{qk} + M_{qk}^{\text{vc}} + I_{qk}^{c} b_{c(k),k}, \tag{13}$$

Table 1: UniVTAC success counts after 100K training steps, with 20 trials per task. Bold marks the best result per task.

Method	Grasp cls.	Hole ins.	Tube ins.	HDMI ins.	Lift bottle	Lift can	Pull key	Bottle shelf	Overall
Tactile-WAM	11/20	4/20	**17/20**	0/20	1/20	2/20	4/20	9/20	48/160 (30.0%)
w/o VIDEOCLEAN	10/20	0/20	12/20	**4/20**	6/20	4/20	4/20	2/20	42/160 (26.3%)
DreamZero	16/20	4/20	4/20	2/20	2/20	2/20	2/20	2/20	34/160 (21.3%)
$\pi_{0.5}$	**19/20**	**8/20**	4/20	0/20	**11/20**	**8/20**	**9/20**	**10/20**	**69/160** (43.1%)

Table 2: ManiFeel success counts after 60K training steps, with 50 trials per task. Bold marks the best result per task.

Method	Peg ins.	USB ins.	Power ins.	Gear ins.	Bulb ins.	Bolt-nut	Object search	Peg reorient.	Ball sort	Overall
Tactile-WAM	**8/50**	1/50	9/50	**19/50**	**34/50**	**20/50**	2/50	17/50	37/50	147/450 (32.7%)
DreamZero	4/50	0/50	4/50	9/50	17/50	10/50	0/50	8/50	18/50	70/450 (15.6%)
$\pi_{0.5}$	5/50	**5/50**	**14/50**	11/50	18/50	15/50	**28/50**	**21/50**	**48/50**	**165/450** (36.7%)

where A is the original attention score and M^{vc} is the video–tactile masking matrix from Eq. (10). Note that the bias is computed solely from observed tactile forces, independent of future tactile targets or intermediate denoising predictions.

Touch-aware Proxy

Proxy Construction. We employ a differentiable deformation estimator Φ as a touch-aware proxy, mapping successive decoded tactile frames to a 3D deformation proxy signal for each sensor. Specifically, for sensor $s \in \{L, R\}$, let I_t^s denote the grayscale tactile image and $\Delta I_t^s = I_t^s - I_{t-1}^s$ its temporal difference (with zero-initialization for the first frame). Using the centered spatial gradients $(g_x^s, g_y^s) = \nabla I_t^s$, we construct a differentiable gradient-aligned deformation field:

$$u_{x,t}^s = \frac{\Delta I_t^s g_x^s}{\sqrt{(g_x^s)^2 + (g_y^s)^2 + \epsilon}}, \qquad u_{y,t}^s = \frac{\Delta I_t^s g_y^s}{\sqrt{(g_x^s)^2 + (g_y^s)^2 + \epsilon}}, \tag{14}$$

where $\epsilon = 10^{-6}$. The proxy signal for each sensor is then defined as $f_t^s = [\langle u_{x,t}^s \rangle, \langle u_{y,t}^s \rangle, \langle \partial_x u_{x,t}^s + \partial_y u_{y,t}^s \rangle]$, where $\langle \cdot \rangle$ denotes spatial averaging. The first two components summarize tangential deformation, while the third measures the average divergence. Each action chunk corresponds to 8 frames of tactile transition; applying Φ to the predicted tactile images yields the predicted proxy signal:

$$\widehat{F} = \Phi(\widehat{o}^{\tau,L}) \oplus \Phi(\widehat{o}^{\tau,R}) \in \mathbb{R}^{C \times 8 \times 6}, \tag{15}$$

where C is the number of action chunks.

Proxy Supervision. Applying the same estimator to the ground-truth tactile trajectories yields the reference signal F^*. We supervise the proxy signal with:

$$\mathcal{L}_F = 0.05 \ \text{SmoothL1}(\widehat{F}, F^*). \tag{16}$$

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{action}} + \lambda_\tau \mathcal{L}_{\text{FM}}^\tau + \mathcal{L}_F. \tag{17}$$

During training, gradients are propagated back to the predicted tactile latents through the frozen decoder and the differentiable estimator Φ; all Wan VAE parameters remain frozen.

5 Experiments

Experimental Setup

We compare with $\pi_{0.5}$, DreamZero, and Naive VT-WAM on eight UniVTAC tasks (Xu et al. 2024) (100K training steps and 20 trials per task), nine ManiFeel tasks (Luu et al. 2025) (60K steps and 50 trials per task, including ablations), and six real-robot conditions (20 trials each); $\pi_{0.5}$ is initialized from released weights, whereas all WAM variants are trained from scratch, and we report success counts and rates under fixed trial budgets. Additional experimental results and protocol details are provided in Appendix C.

Simulation Results

As shown in Table 1, pretrained $\pi_{0.5}$ achieves the highest overall rate (43.1%). Among models trained from scratch, Tactile-WAM reaches 30.0%, exceeding DreamZero (21.3%) and the variant without VIDEOCLEAN (26.3%), with the largest gain on tube insertion ($17/20$ versus $4/20$ for DreamZero and $\pi_{0.5}$). Failures on HDMI insertion and lifting indicate that the gains remain task dependent.

On ManiFeel, Tactile-WAM achieves $147/450$ (32.7%), improving over DreamZero by 17.1 percentage points and leading on peg insertion, gear insertion, bulb insertion, and bolt-nut assembly (Table 2). Although $\pi_{0.5}$ attains a higher overall rate (36.7%), Tactile-WAM performs best on these four contact-dominated tasks. Across both suites, pretraining provides a strong general task prior, while the proposed tactile pathway is most effective when contact determines the corrective action.

Real-Robot Results

Tactile-WAM succeeds in $59/120$ trials (49.2%), outperforming $\pi_{0.5}$ (24.2%), DreamZero (21.7%), and VT-WAM

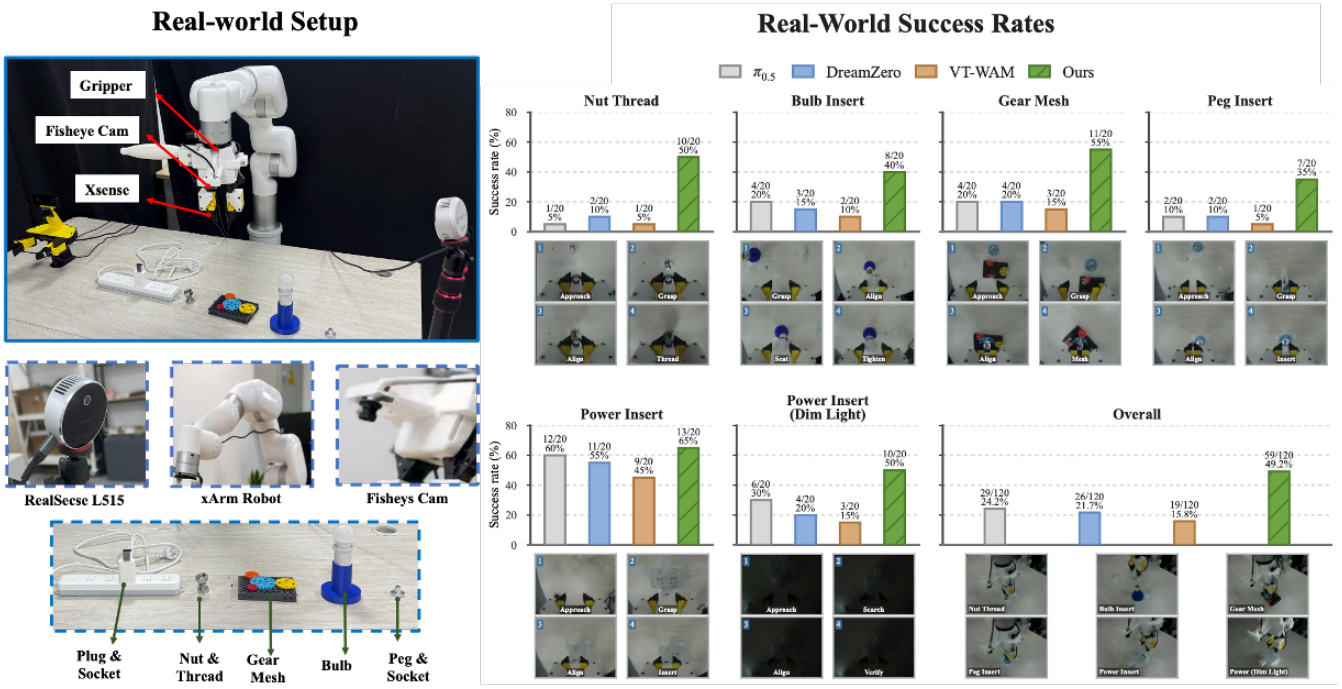

Figure 5: Real-robot success rates over six conditions (20 trials each). Tactile-WAM ranks first in every condition and retains 50% success on dim-light power insertion.

(15.8%) by 25.0–33.3 percentage points (Figure 5). It ranks first in all six conditions, including dim-light power insertion, where it retains 50% success while DreamZero drops from 55% to 20%. This robustness indicates that tactile state complements external vision during post-contact alignment.

Tactile Pollution and VideoClean

We test whether VideoClean preserves DreamZero's visual dynamics using 84 paired held-out UniVTAC samples at step-matched 20K checkpoints, separate from the 100K control evaluation in Table 1. Both tactile variants use identical samples, conditioning frames, and diffusion seeds, while the 20K DreamZero checkpoint serves as the reference. Their paired prediction-to-prediction distance measures how much tactile conditioning perturbs the learned visual trajectory.

VideoClean reduces prediction-to-DreamZero MSE and MAE by 21.8% and 17.8%, respectively, and is closer to DreamZero on 84.5% and 90.5% of samples under these metrics (Table 3). The paired confidence intervals support more reliable preservation of DreamZero's visual dynamics than unrestricted tactile fusion.

VideoClean is comparable to unrestricted fusion at H1 but becomes increasingly effective as autoregressive error accumulates, reducing MSE by 3.3%, 22.4%, and 38.5% at H2–H4 (Table 4; Figure 6). At H4, it is closer to DreamZero on 90.5% of samples, showing that its main effect is to suppress compounding tactile-induced visual drift.

All confidence intervals in Table 5 include zero, indicating no statistically detectable change in ground-truth prediction quality. Thus, VideoClean protects the visual prior and limits long-horizon drift, while virtual-force supervision and contact-aware routing determine how tactile representations guide action generation.

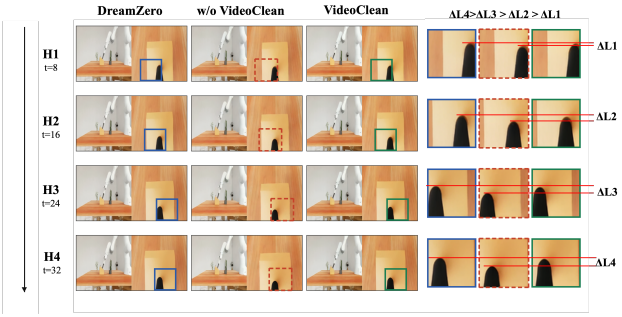

Figure 6: Step-matched 20K rollouts from H1 (top) to H4 (bottom). Blue, red dashed, and green boxes denote DreamZero, w/o VideoClean, and VideoClean; the right crops enlarge the same region. VideoClean better preserves the DreamZero trajectory at long horizons.

Effect of Contact-Aware Attention Bias

During bulb insertion, the bias redirects action attention toward tactile keys after initial contact, enabling correction before the error is visible in external RGB and weakening as contact stabilizes (Figure 7). Without the bias, residual contact develops into slip. This comparison supports the pro-

Table 3: Agreement with DreamZero at the step-matched 20K checkpoint. Positive paired improvements favor VideoClean; all 95% bootstrap confidence intervals exclude zero.

Metric	w/o VideoClean	VideoClean	VideoClean change	Paired improvement [95% CI]	VideoClean closer
MSE ↓	0.001568	**0.001227**	**−21.8%**	+0.000342 [0.000117, 0.000555]	84.5%
MAE ↓	0.020904	**0.017190**	**−17.8%**	+0.003714 [0.002877, 0.004591]	90.5%
PSNR ↑	31.016	**32.020**	**+1.005 dB**	+1.005 [0.646, 1.356]	81.0%
SSIM ↑	0.98440	**0.98683**	**+0.00243**	+0.00243 [0.00005, 0.00468]	70.2%

Table 4: Prediction-to-DreamZero MSE across four rollout chunks at 20K. The reduction from VideoClean reaches 38.5% at H4.

Chunk	w/o VideoClean	VideoClean	Error reduction	VideoClean closer
H1	**0.000695**	0.000706	−1.7%	13.1%
H2	0.001395	**0.001349**	+3.3%	70.2%
H3	0.001735	**0.001346**	**+22.4%**	88.1%
H4	0.002448	**0.001505**	**+38.5%**	90.5%

Table 5: Ground-truth prediction quality at the step-matched 20K checkpoint. Positive Δ favors VideoClean; all paired 95% confidence intervals include zero.

Metric	w/o VideoClean	VideoClean	Paired Δ [95% CI]
MSE ↓	**0.001748**	0.001777	−0.000029 [−0.000257, 0.000202]
MAE ↓	0.021034	**0.020976**	+0.000058 [−0.000765, 0.000915]
PSNR ↑	30.111	**30.305**	+0.194 [−0.078, 0.469]
SSIM ↑	0.98169	**0.98231**	+0.00062 [−0.00134, 0.00255]

posed mechanism but, without a bias-only real-robot variant, does not isolate its success-rate contribution.

Ablation Study

Naive future tactile prediction. Naive VT-WAM reaches 15.1%, slightly below DreamZero's 15.6%, suggesting that direct tactile injection may interfere with the shared representation.

Effect of VideoClean. Adding VideoClean raises success from 15.1% to 28.4% (+13.3 percentage points), the largest ablation gain, indicating that visual–tactile interaction is a greater bottleneck than tactile availability alone.

Virtual-force supervision and contact-aware bias. Virtual-force supervision and contact-aware bias jointly raise success from 28.4% to 32.7% (+4.3 percentage points). The two components target representation content and attention timing, respectively.

Overall gain and identification limit. Full Tactile-WAM improves over DreamZero from 15.6% to 32.7% (+17.1 percentage points; 2.10×). Because force supervision and attention bias are introduced together, the ablation measures only their joint contribution; their individual effects require force-only and bias-only variants.

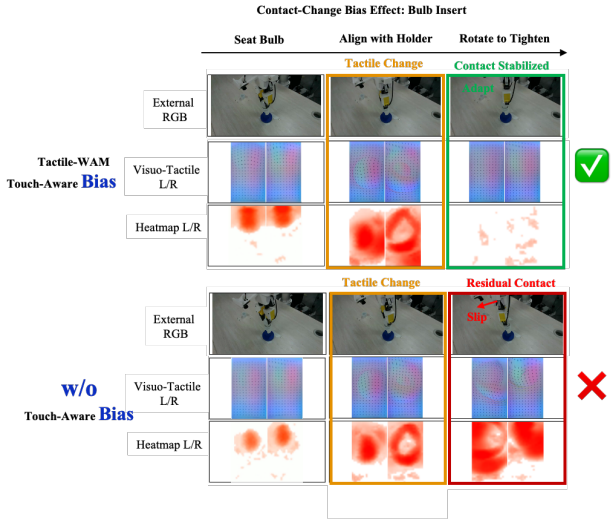

Figure 7: Contact-aware attention during bulb insertion. With the bias (top), tactile change redirects action attention and stabilizes contact; without it (bottom), residual contact leads to slip.

Table 6: Cumulative ablation on ManiFeel at 60K steps. Force denotes virtual-force auxiliary supervision; Bias denotes the contact-aware tactile attention bias.

Method	Future Touch	VideoClean	Force	Bias	Success
RGB-only WAM (DreamZero)	–	–	–	–	0.156
Naive VT-WAM	✓	–	–	–	0.151
+VideoClean	✓	✓	–	–	0.284
Full Tactile-WAM	✓	✓	✓	✓	**0.327**

6 Conclusion

We presented Tactile-WAM, a tactile-aware world action model that addresses tactile pollution via asymmetric attention and deformation-based proxy. It achieves state-of-the-art results on UniVTAC and ManiFeel, with 32.7% success and a 17.1-point improvement over DreamZero, while reducing visual trajectory MSE by 21.8%. Physical evaluation ranks first across all conditions. Future work will investigate individual component effects and generalizability.

References

Chen, K.; Long, Y.; and Shang, M. 2026. PIPHEN: Physical Interaction Prediction with Hamiltonian Energy Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18170–18179.

Chen, Z.; Gao, C.; Shao, L.; Shi, J.; Huo, J.; and Gao, Y. 2026. ManiLong-Shot: Interaction-Aware One-Shot Imitation Learning for Long-Horizon Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18189–18197.

Cheng, N.; Xu, J.; Chen, J.; Fang, B.; and Han, W. 2026. STOLA: Self-Adaptive Touch-Language Framework for Tactile Commonsense Reasoning in Open-Ended Scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18198–18206.

Guo, J.; Li, Q.; Li, P.; Chen, Z.; Sun, N.; Su, Y.; Wang, H.; Zhang, Y.; Li, X.; and Liu, H. 2026. Unified 4D World Action Modeling from Video Priors with Asynchronous Denoising. arXiv:2604.26694.

Guo, Y.; Hu, Y.; Zhang, J.; Wang, Y.-J.; Chen, X.; Lu, C.; and Chen, J. 2024. Prediction with Action: Visual Policy Learning via Joint Denoising Process. arXiv:2411.18179.

Higuera, C.; Arnaud, S.; Boots, B.; Mukadam, M.; Hogan, F. R.; and Meier, F. 2026. Visuo-Tactile World Models. arXiv:2602.06001.

Higuera, C.; Sharma, A.; Bodduluri, C. K.; Fan, T.; Lancaster, P.; Kalakrishnan, M.; Kaess, M.; Boots, B.; Lambeta, M.; Wu, T.; and Mukadam, M. 2024. Sparsh: Self-Supervised Touch Representations for Vision-Based Tactile Sensing. arXiv:2410.24090.

Hu, Y.; Guo, Y.; Wang, P.; Chen, X.; Wang, Y.-J.; Zhang, J.; Sreenath, K.; Lu, C.; and Chen, J. 2024. Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations. arXiv:2412.14803.

Jia, P.; Li, X.; Zhu, T.; Wu, R.; Lin, X.; and Sun, Y. 2025. Multi-Fingered Hand Grasps with Visuo-Tactile Fusion via Multi-Agent Deep Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14594–14601.

Lambeta, M.; Chou, P.-W.; Tian, S.; Yang, B.; Maloon, B.; Most, V. R.; Stroud, D.; Santos, R.; Byagowi, A.; Kammerer, G.; Jayaraman, D.; and Calandra, R. 2020. 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(3): 3838–3845.

Li, L.; Zhang, Q.; Luo, Y.; Yang, S.; Wang, R.; Han, F.; Yu, M.; Gao, Z.; Xue, N.; Zhu, X.; Shen, Y.; and Xu, Y. 2026. Causal World Modeling for Robot Control. arXiv:2601.21998.

Liu, L.; Wang, W.; Han, Y.; Xie, Z.; Yi, P.; Li, J.; and Lian, W. 2026. FoAM: Foresight-Augmented Multi-Task Imitation Policy for Robotic Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18460–18468.

Luu, Q.-K.; Zhou, P.; Xu, Z.; Zhang, Z.; Qiu, Q.; and She, Y. 2025. ManiFeel: Benchmarking and Understanding Visuotactile Manipulation Policy Learning. arXiv:2505.18472.

Lyu, K.; Xiao, L.; Zeng, J.; Dong, J.; Liu, X.; Zou, Z.; Yang, H.; Shu, L.; and Hao, J. 2026. TouchFormer: A Robust Transformer-Based Framework for Multimodal Material Perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18496–18504.

Pan, S.; Xu, Y.; Xu, R.; Zhou, Z.; Wu, S.; and Yu, Z. 2025. Self-Correcting Robot Manipulation via Gaussian-Splatted Foresight. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 26642–26650.

Sheng, J.; Wang, Z.; Li, P.; and Liu, M. 2026. MP1: MeanFlow Tames Policy Learning in 1-Step for Robotic Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18532–18539.

Song, W.; Zhou, Z.; Zhao, H.; Chen, J.; Ding, P.; Yan, H.; Huang, Y.; Tang, F.; Wang, D.; and Li, H. 2026. ReconVLA: Reconstructive Vision-Language-Action Model as Effective Robot Perceiver. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18549–18557.

Wan, C.; Wang, K.; Si, Y.; Zhang, P.; and Li, M. 2026. WorldAgen: Unified State-Action Prediction with Test-Time World Model Training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18584–18592.

Wan Team; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv:2503.20314.

Ward-Cherrier, B.; Pestell, N.; Cramphorn, L.; Winstone, B.; Giannaccini, M. E.; Rossiter, J.; and Lepora, N. F. 2018. The TacTip Family: Soft Optical Tactile Sensors with 3D-Printed Biomimetic Morphologies. *Soft Robotics*, 5(2): 216–227.

Xie, W.; Ding, Y.; He, Y.; Wang, L.; Bai, B.; Zhao, Z.; Wang, C.; and Yu, F. R. 2026. ForeDiffusion: Foresight-Conditioned Diffusion Policy via Future View Construction for Robot Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18665–18673.

Xu, Z.; Uppuluri, R.; Zhang, X.; Fitch, C.; Crandall, P. G.; Shou, W.; Wang, D.; and She, Y. 2024. UniT: Data Efficient Tactile Representation with Generalization to Unseen Objects. arXiv:2408.06481.

Xue, H.; Ren, J.; Chen, W.; Zhang, G.; Fang, Y.; Gu, G.; Xu, H.; and Lu, C. 2025. Reactive Diffusion Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation. arXiv:2503.02881.

Ye, S.; Ge, Y.; Zheng, K.; Gao, S.; Yu, S.; et al. 2026. World Action Models Are Zero-Shot Policies. arXiv:2602.15922.

Yuan, W.; Dong, S.; and Adelson, E. H. 2017. GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force. *Sensors*, 17(12): 2762.

Zhang, Q.; Li, L.; Zhang, L.; Yang, S.; Luo, Y.; Li, S.; Wang, R.; Wang, J.; Shao, J.; Xu, G.; et al. 2026. Native Video-Action Pretraining for Generalizable Robot Control. arXiv:2607.08639.

Zhou, J.; Wu, Q.; Li, J.; Chen, Z.; Xiong, X.; and Xu, R. 2026a. Gentle Manipulation Policy Learning via Demonstrations from VLM-Planned Atomic Skills. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18855–18863.

Zhou, Y.; Xu, M.; Shi, J.; Li, Q.; and Chen, J. 2026b. Collaborative Representation Learning for Alignment of Tactile, Language, and Vision Modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18864–18872.

Zhu, Y.; Shao, R.; Liu, Z.; He, J.; Liu, J.; Wang, J.; and Yu, Z. 2026. H-GAR: A Hierarchical Interaction Framework via Goal-Driven Observation-Action Refinement for Robotic Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18882–18890.

Supplementary Material

A Related Work

World action models. Joint observation-action denoising connects generative video priors to control (Guo et al. 2024; Hu et al. 2024; Ye et al. 2026), while action-conditioned foresight compares predicted and observed scenes for failure recovery (Pan et al. 2025). X-WAM adds multi-view RGB-D prediction and 4D reconstruction to Wan2.2 (Guo et al. 2026). WorldAgen uses separated heads and a mixed unidirectional mask for state-action prediction and test-time adaptation (Wan et al. 2026); LingBot-VA controls interaction through modality-specific transformer pathways (Li et al. 2026). H-GAR couples predicted observations with hierarchical action refinement (Zhu et al. 2026). Tactile-WAM differs by predicting touch as a physical state and constraining its interaction with video and action tokens.

Touch for manipulation and prediction. Optical tactile sensing exposes local contact geometry and deformation that may be weakly observable in RGB (Yuan, Dong, and Adelson 2017; Lambeta et al. 2020; Ward-Cherrier et al. 2018). UniT and Sparsh learn transferable representations (Xu et al. 2024; Higuera et al. 2024); recent work studies tactile reasoning, adaptive visual-tactile fusion, and cross-modal alignment (Cheng et al. 2026; Lyu et al. 2026; Zhou et al. 2026b). Reactive and multi-fingered policies use current touch for correction (Xue et al. 2025; Jia et al. 2025), while force-aware learning targets gentle manipulation (Zhou et al. 2026a). ManiFeel provides contact-rich tasks and matched evaluation (Luu et al. 2025), and visuo-tactile world models predict future touch for control (Higuera et al. 2026). Tactile-WAM instead encodes touch with the visual Wan VAE and supervises decoded tactile latents.

Cross-modal interference and tactile pollution. LingBot-VA prevents modality-specific representation interference with separately parameterized video and action transformer pathways (Li et al. 2026). Our setting instead concatenates video, tactile, action, and state tokens inside one Wan-based DiT. VideoClean therefore removes tactile-key access only for video queries, leaving action-to-tactile attention intact; *tactile pollution* names the failure in which sparse tactile events perturb the pretrained visual path.

Auxiliary supervision. FoAM predicts action consequences, ReconVLA reconstructs task-relevant visual regions, and ForeDiffusion aligns a policy with constructed future views (Liu et al. 2026; Song et al. 2026; Xie et al. 2026). LingBot-VA 2.0 extends this idea over longer visual horizons with training-only heads (Zhang et al. 2026), while MP1 improves generalization through a loss that does not slow inference (Sheng et al. 2026). Tactile-WAM applies the same training-time principle to virtual force derived from decoded tactile futures.

B Model Details

This section describes the architecture of Tactile-WAM, including its shared latent representation, multimodal tokenization, causal attention structure, touch-aware routing, prediction heads, and optimization configuration.

Backbone and Latent Representation

Tactile-WAM is built on the Wan2.2-TI2V-5B diffusion Transformer backbone (Wan Team et al. 2025). The backbone contains 30 Transformer blocks with a hidden dimension of 3072, 24 attention heads, and a feed-forward dimension of 14336. Each feed-forward block follows $\mathrm{Linear}(3072, 14336)$–GELU– $\mathrm{Linear}(14336, 3072)$.

RGB and optical tactile observations share the frozen Wan2.2 VAE. The VAE maps both modalities to a 48-channel latent space with spatial and temporal downsampling factors of 16 and 4, respectively. For the two tactile sensors, we form a horizontal mosaic

$$\mathcal{M}(o^{\tau,L}, o^{\tau,R}) = [\, o^{\tau,L} \mid o^{\tau,R} \,], \tag{18}$$

and resize the mosaic to the VAE input size. Sharing the VAE places visual and tactile observations in a common generative latent space without introducing a separate tactile tokenizer.

Both RGB sequences and complete tactile mosaics are resized bilinearly to 320×160 before VAE encoding. Integer-valued images are first divided by 255 and then normalized with channel-wise mean and standard deviation 0.5, giving the VAE input range $[-1, 1]$. Resizing is applied to the complete left-right mosaic rather than independently to the two sensor images.

Multimodal Tokenization

Video latents use Wan's native three-dimensional patchification with patch size $(1, 2, 2)$. Tactile latents are unfolded into 2×2 spatial patches. Each tactile patch therefore contains $48 \times 2 \times 2 = 192$ values and is projected to the 3072-dimensional DiT space by a learnable linear layer. A learnable tactile-modality embedding is added after projection.

An action $a \in \mathbb{R}^{D_a}$ is first mapped to 3072 dimensions and combined with a sinusoidal diffusion-time embedding through

$$\mathrm{Linear}(6144, 3072) \to \mathrm{SiLU} \to \mathrm{Linear}(3072, 3072). \tag{19}$$

Proprioceptive states are projected by a two-layer MLP. Language instructions are encoded by UMT5-XXL into 4096-dimensional token features, projected to 3072 dimensions, and injected into each DiT block through cross-attention.

Video tokens use the native three-dimensional rotary positional encoding (3-D RoPE), while tactile, action, and state tokens use one-dimensional RoPE. Within each modality, tokens are ordered by causal block. Their conceptual layout is

```
[condition and noisy video tokens]
[clean tactile-history tokens]
[noisy future-tactile tokens]
[action tokens]
[state tokens]
```

RoPE and the causal block mask preserve temporal correspondence across these modality groups.

Temporal Block Organization

The model receives 33 image frames, which the temporal VAE compresses to nine latent frames. The first latent frame

is retained as a clean condition; the remaining eight latent frames are divided into four non-overlapping causal blocks, with two latent frames per block. Each block is aligned with a 24-step action chunk and one proprioceptive state token. Visual, tactile, action, and state tokens are stored in modality-wise groups rather than interleaved frame by frame. A block identifier attached to every token provides the temporal alignment used by RoPE and the attention mask.

Actions and states use a shared width of eight dimensions. Lower-dimensional inputs are zero-padded and accompanied by validity masks. Each valid dimension x_d is normalized with its first and 99th percentiles:

$$\widetilde{x}_d = \mathrm{clip}\left(2\frac{x_d - q_{01,d}}{q_{99,d} - q_{01,d}} - 1, -1, 1\right). \tag{20}$$

Dimensions with zero percentile range retain their original value. Padding dimensions are excluded from the action loss.

Causal Multimodal Attention

Let M^{causal} denote the native block-causal attention mask. VideoClean adds a directional mask that prevents visual queries from reading tactile keys:

$$M^{\mathrm{vc}}_{qk} = \begin{cases} -\infty, & G(q) = V \ \wedge \ G(k) = T, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

where $G(\cdot)$ identifies the modality of a token. Action queries retain access to tactile keys, and tactile queries preserve their causally permitted multimodal context. The mask therefore isolates only the pathway that can inject tactile information into visual prediction.

The touch-aware bias selectively strengthens action-to-tactile attention. For causal block c, an observed tactile pair is mapped to a six-dimensional proxy F^{obs}_c. We convert its magnitude into a bounded block-level score

$$b_c = \mathrm{clip}\left[\alpha \tanh\left(\frac{\mathrm{ReLU}(\|F^{\mathrm{obs}}_c\|_2 - \theta)}{T}\right), 0, b_{\max}\right]. \tag{22}$$

The complete attention logit is

$$A'_{qk} = \frac{q^\top k}{\sqrt{d}} + M^{\mathrm{causal}}_{qk} + M^{\mathrm{vc}}_{qk} + B^{\tau}_{qk}, \tag{23}$$

where $B^{\tau}_{qk} = b_c$ only for an action query and a tactile key in the same causally accessible block. Because the bias modifies only connections already admitted by M^{causal}, it cannot reveal future tokens.

Touch-Aware Proxy

The fixed differentiable operator Φ summarizes deformation between two tactile frames. RGB inputs are mapped to grayscale using $0.299R + 0.587G + 0.114B$, after conversion to $[0, 1]$. Let $\Delta I = I_t - I_{t-1}$. Spatial derivatives use the centered one-dimensional kernels $k_x = \frac{1}{2}[-1, 0, 1]$ and $k_y = k_x^\top$, with one-pixel zero padding. Writing the resulting gradients as g_x and g_y, the gradient-aligned deformation field is

$$u_x = \frac{\Delta I\, g_x}{\sqrt{g_x^2 + g_y^2 + \epsilon}}, \qquad u_y = \frac{\Delta I\, g_y}{\sqrt{g_x^2 + g_y^2 + \epsilon}}, \tag{24}$$

where $\epsilon > 0$ is a numerical stabilizer. Each sensor produces

$$f^\tau = [\langle u_x \rangle, \langle u_y \rangle, \langle \partial_x u_x + \partial_y u_y \rangle], \tag{25}$$

where $\langle \cdot \rangle$ denotes spatial averaging. Concatenating the left and right outputs gives

$$F^\tau = [\langle u_x^L \rangle, \langle u_y^L \rangle, \langle \mathrm{div}^L \rangle, \langle u_x^R \rangle, \langle u_y^R \rangle, \langle \mathrm{div}^R \rangle] \in \mathbb{R}^6. \tag{26}$$

The proxy is used at two temporal resolutions. For the auxiliary target, each causal block contains eight consecutive frame-to-frame transitions:

$$F^*_c = \left[\Phi(I_{u_c}, I_{u_c+1}), \ldots, \Phi(I_{u_c+7}, I_{u_c+8})\right] \in \mathbb{R}^{8\times 6}. \tag{27}$$

Across C blocks, the target tensor therefore has shape $\mathbb{R}^{C\times 8\times 6}$, ordered as $(\text{block}, \text{transition}, \text{feature})$. In contrast, the observed cue used by the attention bias is a single non-adjacent difference,

$$F^{\mathrm{obs}}_c = \Phi(I_{u_c-8}, I_{u_c}), \tag{28}$$

and is not a sum of eight differences.

The operator Φ has no learnable parameters. For tactile prediction, the proxy loss is computed from decoded tactile latents and remains differentiable with respect to the predicted latents. The VAE parameters stay frozen. For attention routing, the observed proxy is detached before the bias is constructed.

Prediction Heads and Objective

The tactile output head applies $\mathrm{Linear}(3072, 192)$ to each tactile token and unpatchifies the result into a 48-channel latent flow field. The action head is a two-layer MLP,

$$\mathrm{Linear}(3072, 1024) \to \mathrm{ReLU} \to \mathrm{Linear}(1024, D_a). \tag{29}$$

The shared implementation width is $D_a = 8$, with invalid padded dimensions masked as described above. The model jointly predicts visual latents, tactile latents, and action tokens. Its objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{video}} + \mathcal{L}_{\mathrm{action}} + \lambda_\tau \mathcal{L}^\tau_{\mathrm{FM}} + \lambda_{\mathrm{proxy}} \mathcal{L}_{\mathrm{proxy}}, \tag{30}$$

where $\mathcal{L}_{\mathrm{proxy}} = \mathrm{SmoothL1}(\widehat{F}, F^*)$. Each modality-specific loss is normalized over its own feature dimensions before the terms are combined.

Training Configuration

Optimization. We fully fine-tune the DiT backbone together with the multimodal projections, modality embeddings, modified attention layers, and prediction heads. The pretrained VAE and language encoder remain frozen and are kept in inference mode. All trainable parameters are optimized jointly with AdamW using a peak learning rate of 1×10^{-5}, $\beta_1 = 0.95$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Weight decay is 1×10^{-5} for eligible weight matrices and zero for bias and normalization parameters. The learning rate is linearly warmed up during the first 5% of optimization and then decayed with a cosine schedule.

Flow-matching construction. The scheduler contains 1000 discrete training noise levels. For a clean latent or action target x_0, we sample $\epsilon \sim \mathcal{N}(0, I)$ and construct

$$x_t = (1 - \sigma_t)x_0 + \sigma_t\epsilon, \qquad v_t^* = \epsilon - x_0. \tag{31}$$

One noise level is sampled uniformly for each causal block. Frames within the same block share this level, and the temporally aligned visual, tactile, and action targets use the same block-level timestep. The first conditioning frame remains clean. The video, tactile, and action branches minimize timestep-weighted mean-squared errors on v_t^*, with the scheduler weights normalized to unit mean.

Loss balancing and numerical stability. The modality coefficients in Eq. (30) are kept fixed throughout optimization. Action padding and unavailable targets are masked before reduction. Each branch is averaged over its valid tokens and feature dimensions before the weighted losses are added. We use BF16 mixed precision, enable TF32 matrix multiplications and activation recomputation, and clip the global gradient norm to 1.0. The random seed is set to 42.

Training scale and hardware. Training uses eight NVIDIA A100 GPUs with a per-device micro-batch size of 1 and one gradient-accumulation step, giving an effective global batch size of 8. Full-parameter optimization is distributed with DeepSpeed ZeRO-2; optimizer states and gradients are sharded across devices, without CPU optimizer offloading.

C Experiments Details

Experimental Setup

Simulation protocols. We evaluate on UniVTAC (Xu et al. 2024) and ManiFeel (Luu et al. 2025). UniVTAC contains eight manipulation tasks; each method is trained for 100K steps and evaluated for 20 trials per task, giving 160 trials per method. ManiFeel contains nine tasks; each model is trained for 60K steps and evaluated for 50 trials per task, giving 450 trials per method. The ablation study is conducted only on ManiFeel with the same 60K-step budget. For ManiFeel, observations are sampled at 10 fps, the simulator runs at 60 Hz with control decimation 4, and the policy predicts 24 actions and executes 8 actions before replanning. Six-dimensional task actions are zero-padded to the model's 7-D interface and unpadded before execution.

Real-robot protocol. The physical evaluation covers nut threading, bulb insertion, gear meshing, peg insertion, and power insertion. Power insertion is additionally evaluated under dim lighting, yielding six conditions. Each method performs 20 trials per condition and 120 trials in total. Hardware, sensor installation, data collection, control frequency, task initialization, success criteria, and reset procedures are provided in the supplementary material.

Baselines and training fairness. We compare with $\pi_{0.5}$, DreamZero, and Naive VT-WAM. $\pi_{0.5}$ is fine-tuned from its released base weights and therefore benefits from pretrained initialization, whereas DreamZero, Naive VT-WAM, and our variants are trained from scratch in each environment. We report all absolute results, but comparisons among the from-scratch WAM variants provide the most controlled measure of the contribution of tactile modeling. Success rate is the primary control metric; every table reports the number of successful executions over the fixed trial count.

Qualitative Execution Trajectories

This section presents successful task-level execution trajectories for all simulation and real-robot tasks evaluated in the paper. Each visualization preserves the temporal ordering, task-specific stages, synchronized camera views, and tactile observations used during execution.

Qualitative Results on Simulation Tasks Figures S1–S17 present successful execution trajectories from all simulation tasks. Figures S1–S9 cover the nine ManiFeel tasks, with front, side, wrist, and bilateral visuo-tactile observations. Figures S10– S17 cover the eight UniVTac tasks, with external, wrist, and bilateral visuo-tactile observations.

Qualitative Results on Real-Robot Experiments Figures S18–S24 illustrate successful execution sequences from all seven real-world robotic experiments. Each visualization synchronizes the external RGB view, wrist RGB view, bilateral visuo-tactile images, and tactile heatmaps.

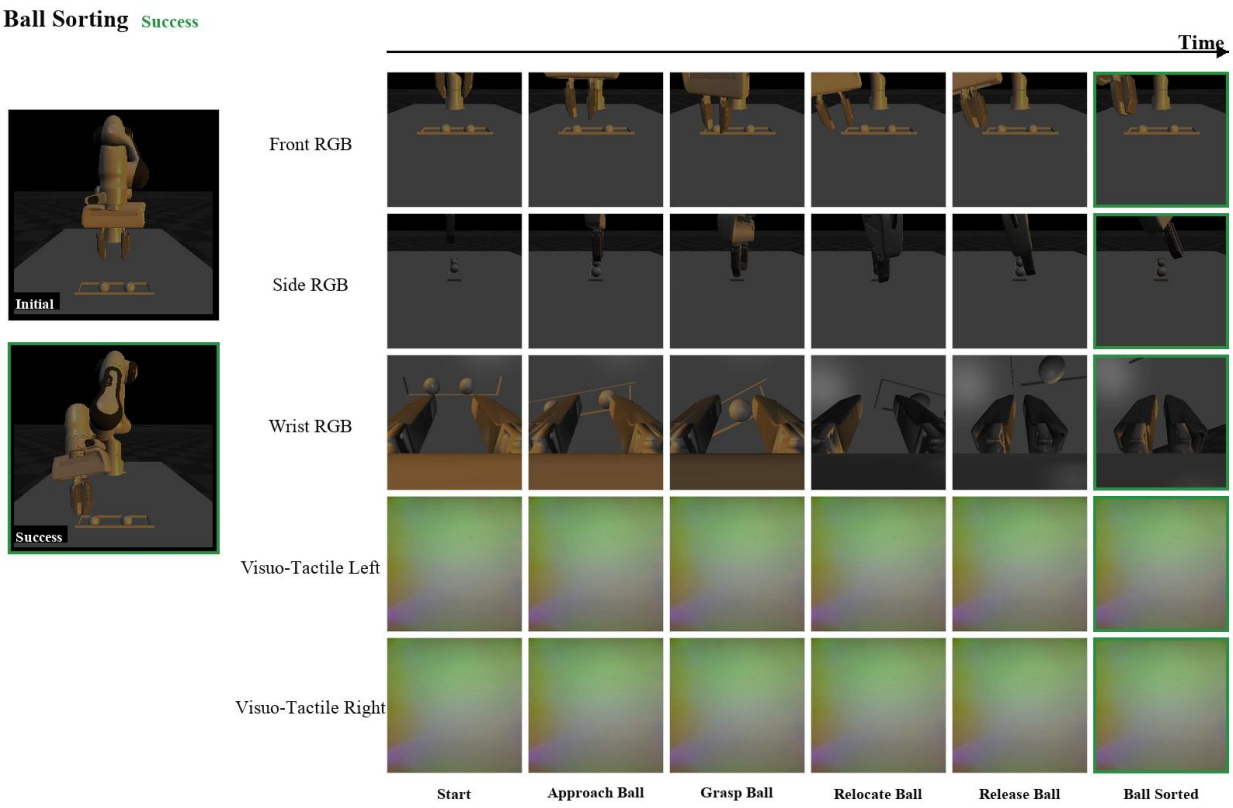

Figure S1: Successful ManiFeel execution for ball sorting. The robot approaches and grasps the ball, relocates it toward the target region, and releases it at the desired location.

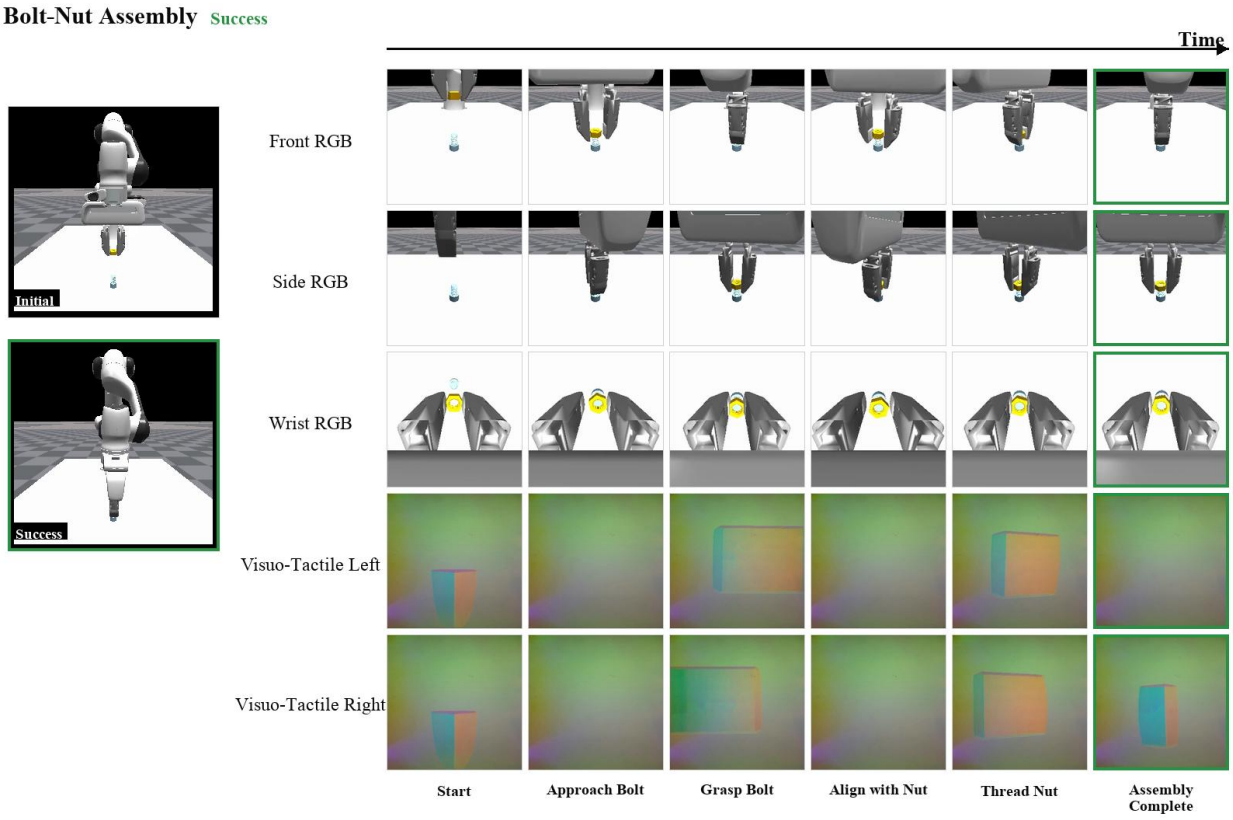

Figure S2: Successful ManiFeel execution for bolt–nut assembly. The sequence shows approach, grasp, contact establishment, axis alignment, and threading.

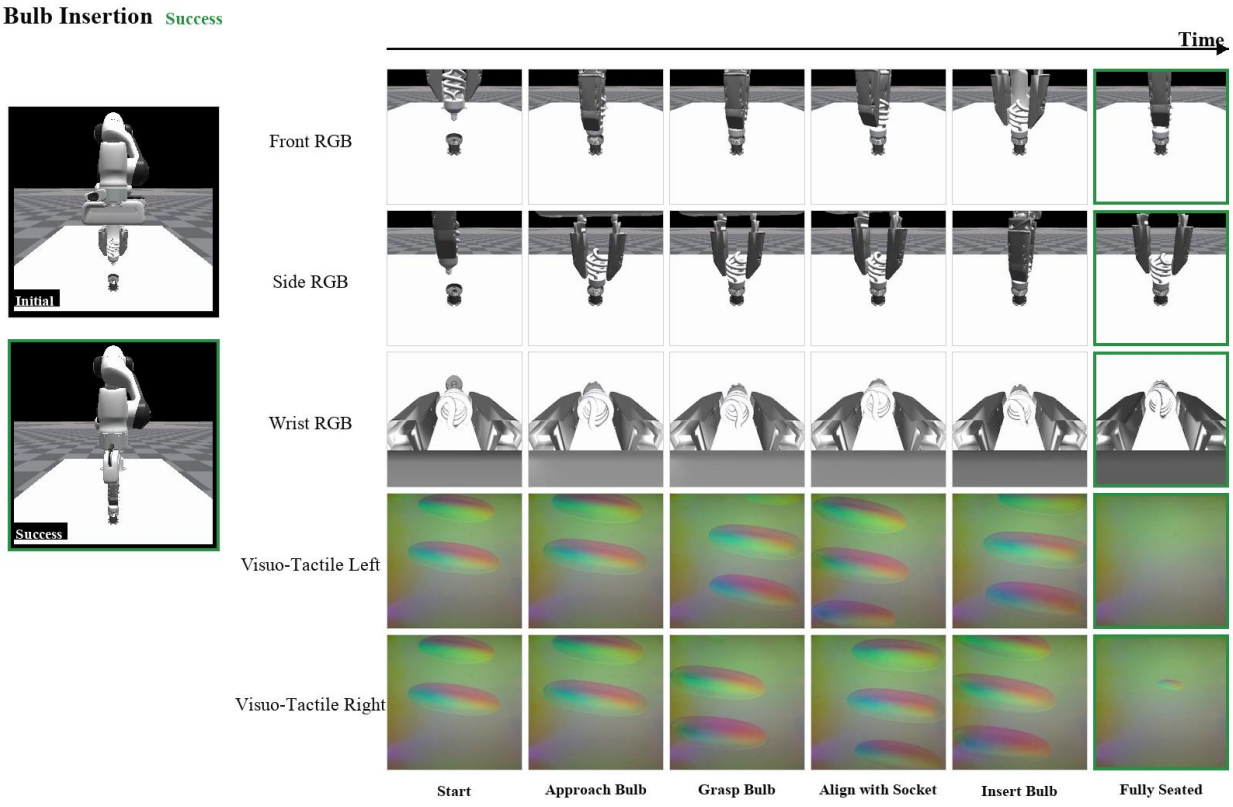

Figure S3: Successful ManiFeel execution for bulb insertion. The bulb is grasped, aligned with the socket, inserted, and seated while tactile observations capture the evolving contact state.

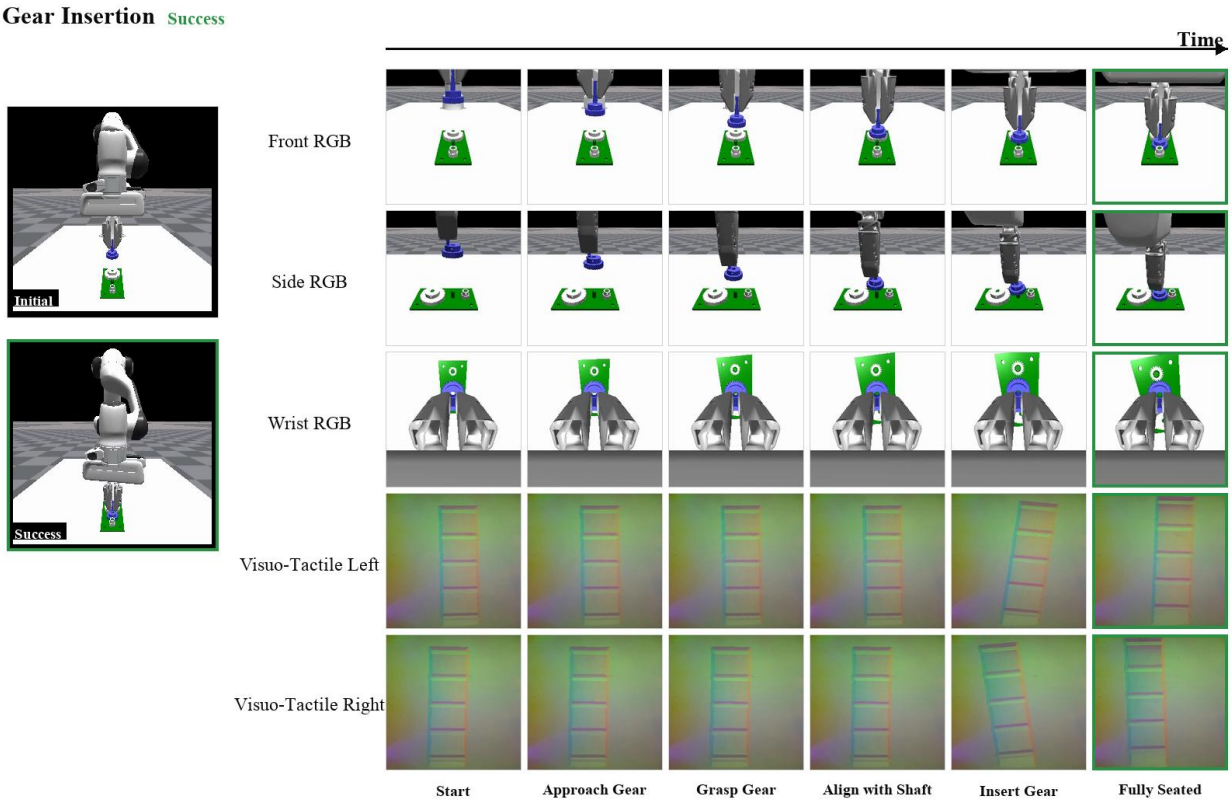

Figure S4: Successful ManiFeel execution for gear insertion, including approach, grasp, alignment with the shaft, insertion, and final seating.

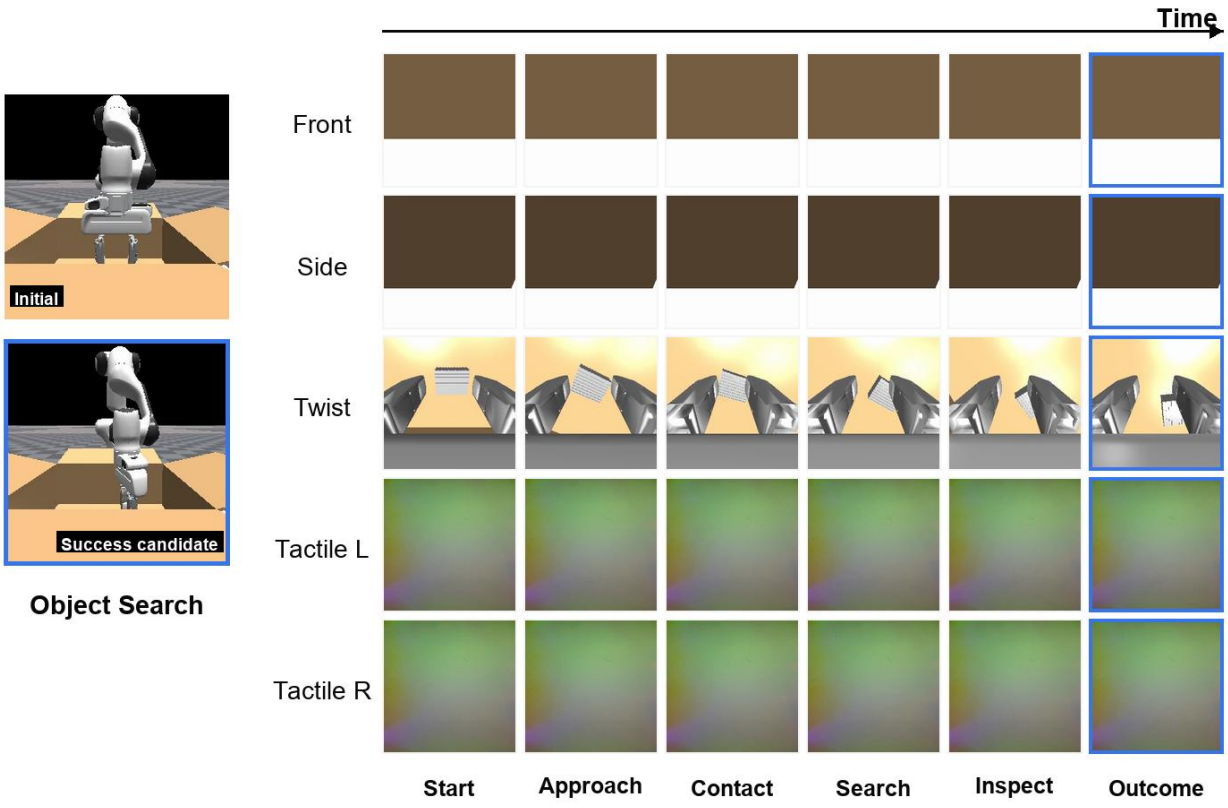

Figure S5: Successful ManiFeel execution for object search. The robot explores the workspace, establishes tactile contact, localizes the target, and completes the search.

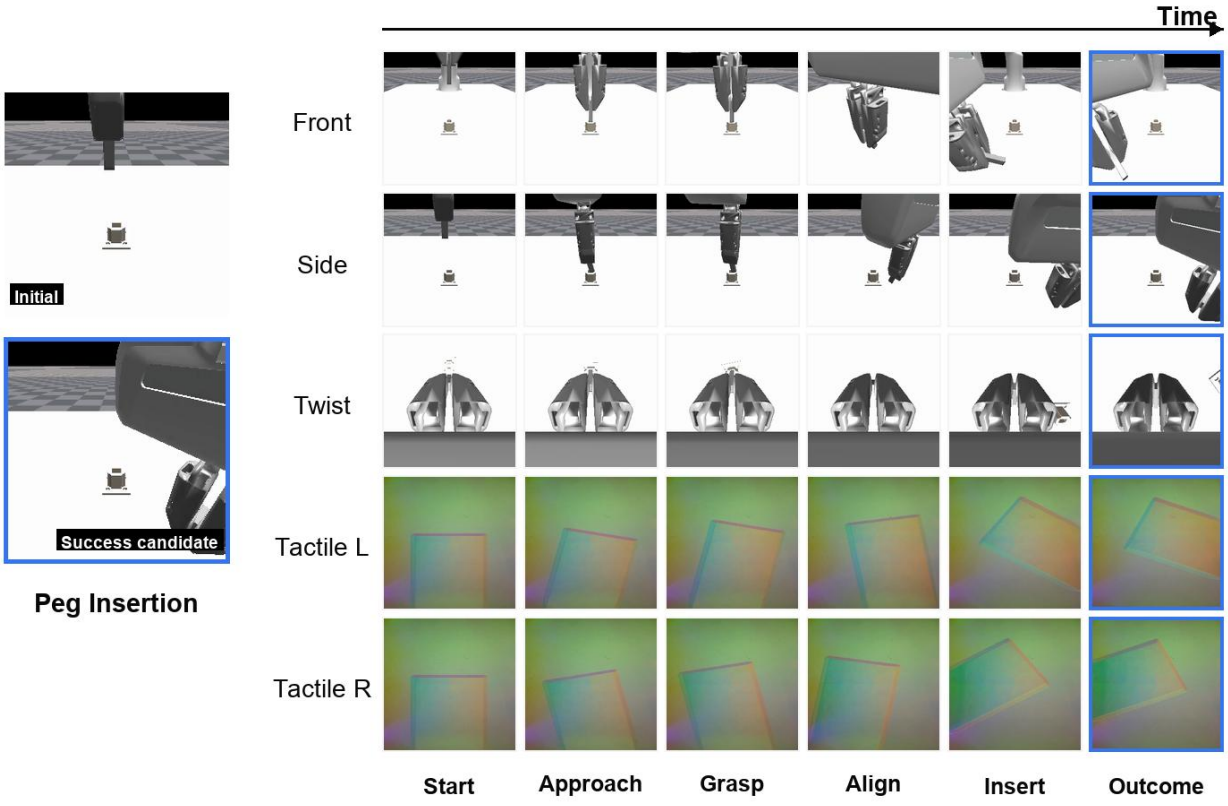

Figure S6: Successful ManiFeel execution for peg insertion. The robot grasps the peg, aligns it with the opening through contact, and completes insertion.

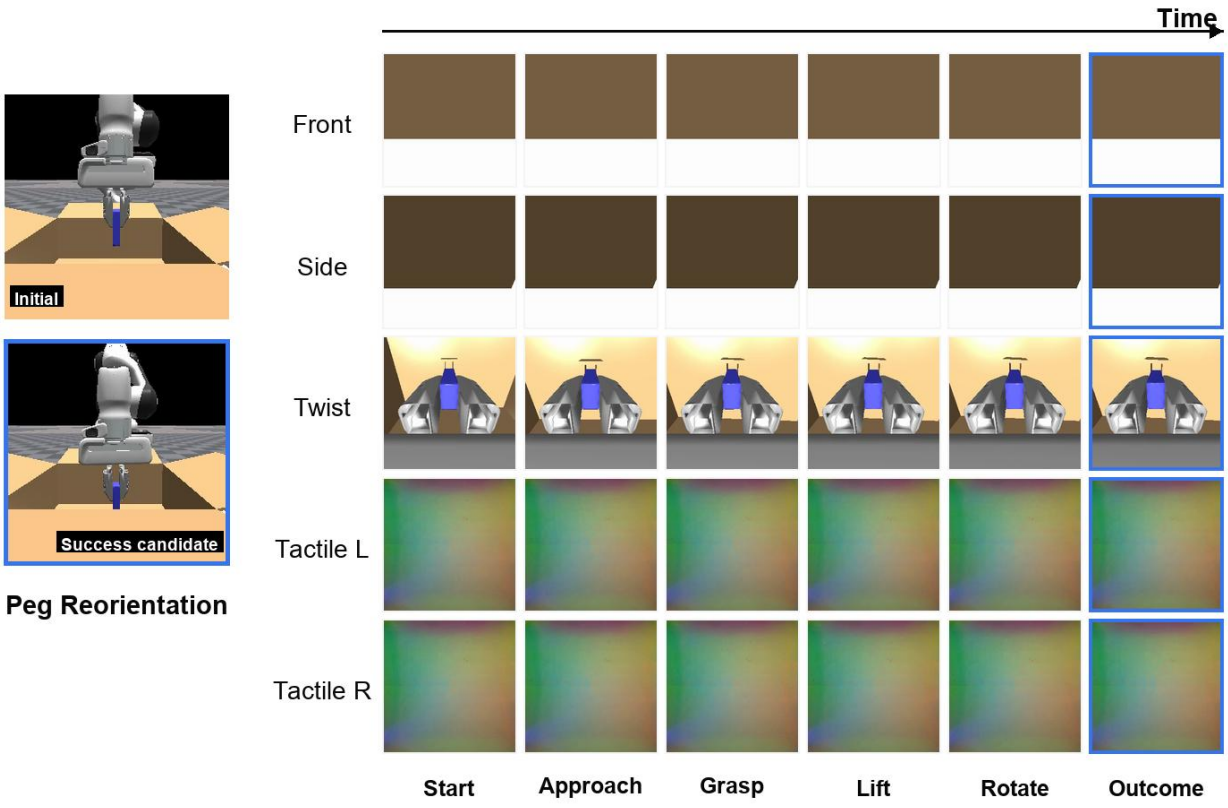

Figure S7: Successful ManiFeel execution for peg reorientation. The peg is grasped, rotated under contact, and placed in the required final orientation.

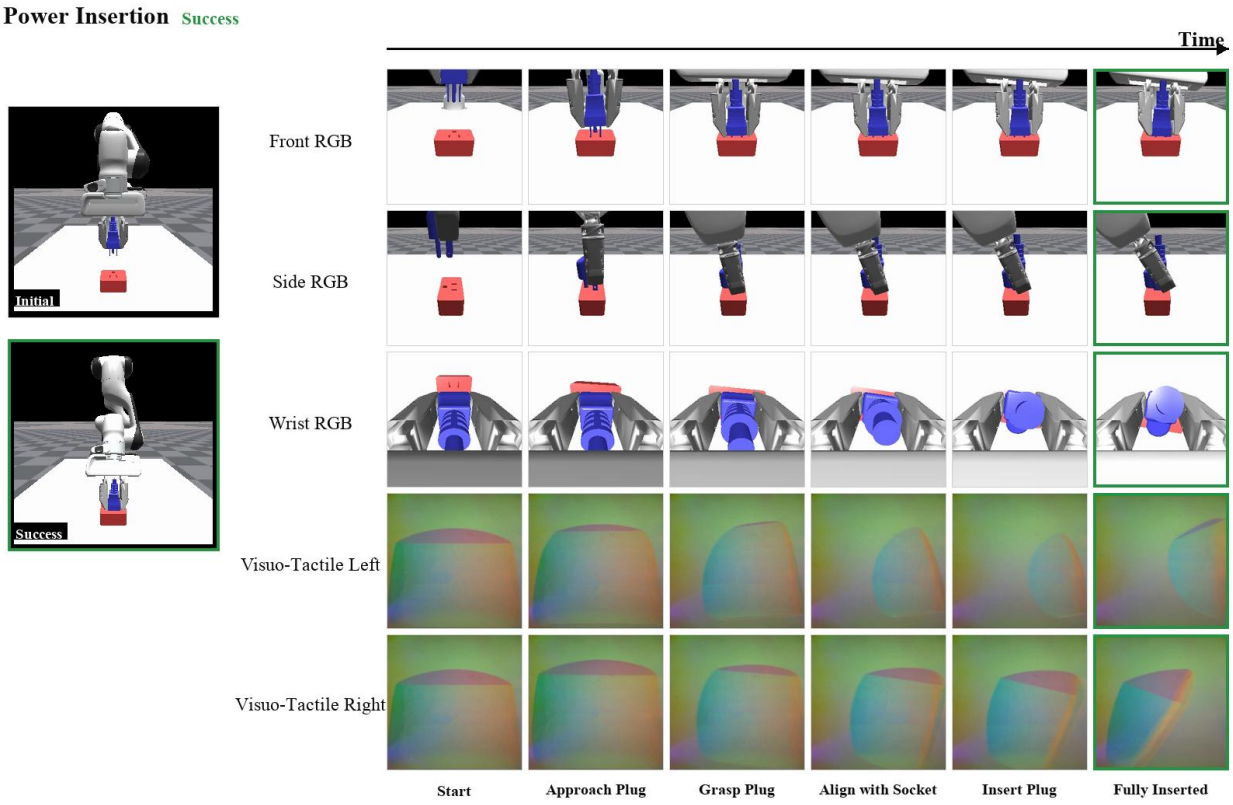

Figure S8: Successful ManiFeel execution for power-plug insertion. The robot approaches the plug, grasps it, aligns it with the socket, and completes insertion.

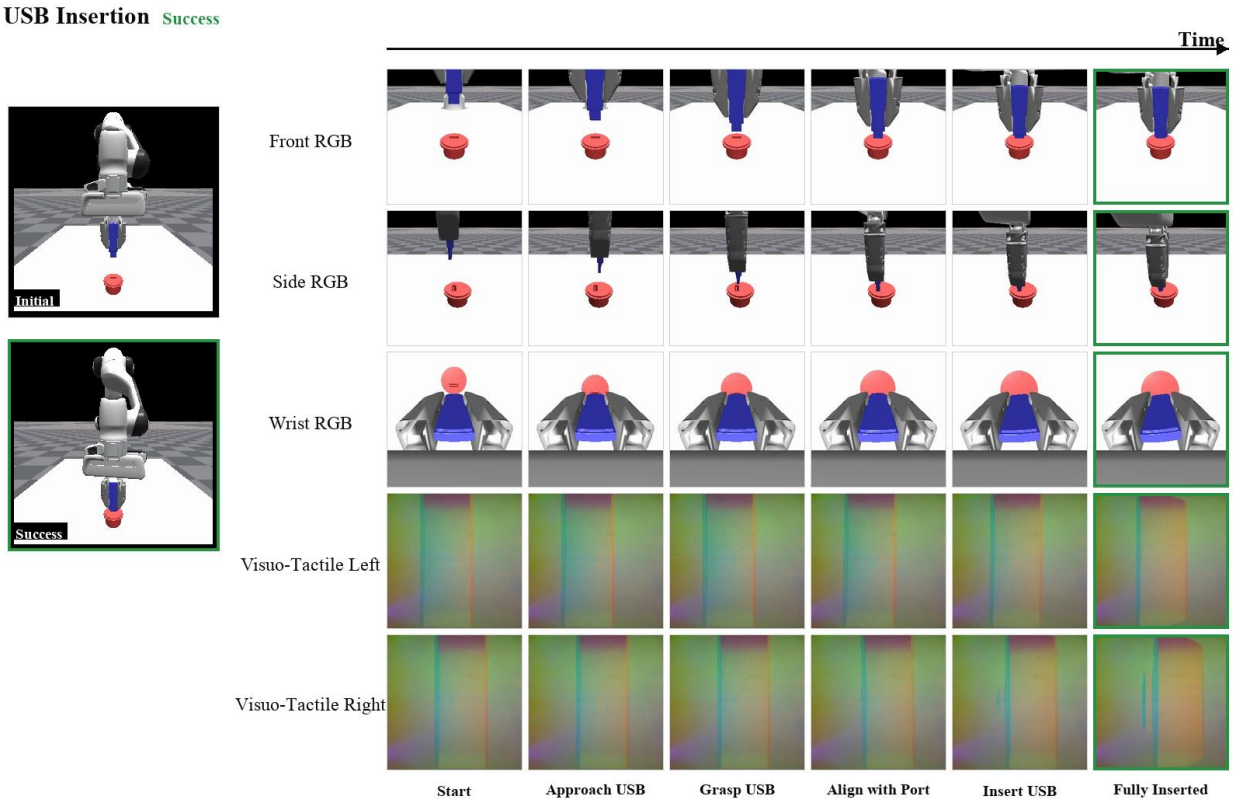

Figure S9: Successful ManiFeel execution for USB insertion, showing approach, grasp, port alignment, insertion, and full seating.

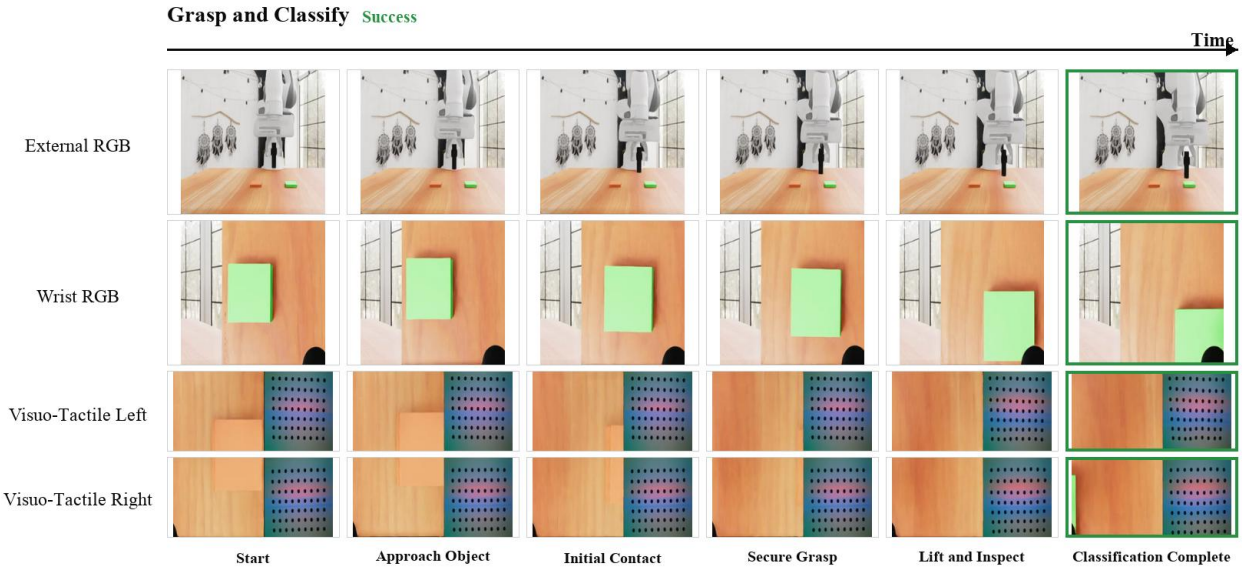

Figure S10: Successful UniVTac execution for grasp and classify. The robot establishes a secure grasp, lifts and inspects the object, and maintains stable contact through classification.

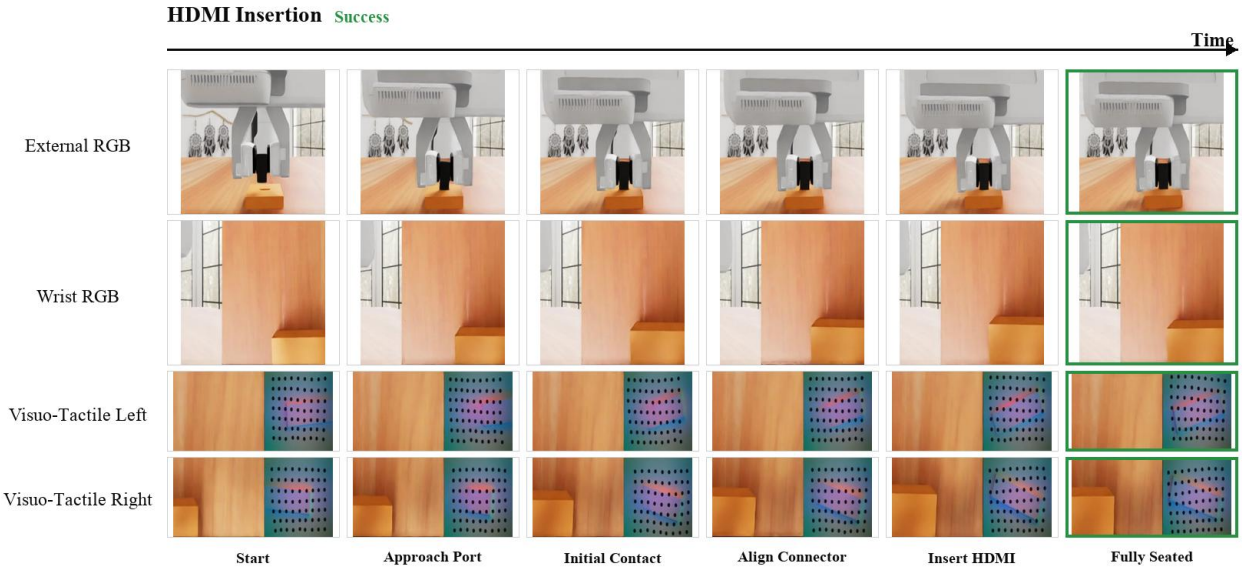

Figure S11: Successful UniVTac execution for HDMI insertion. The connector is grasped, aligned with the port, inserted, and fully seated.

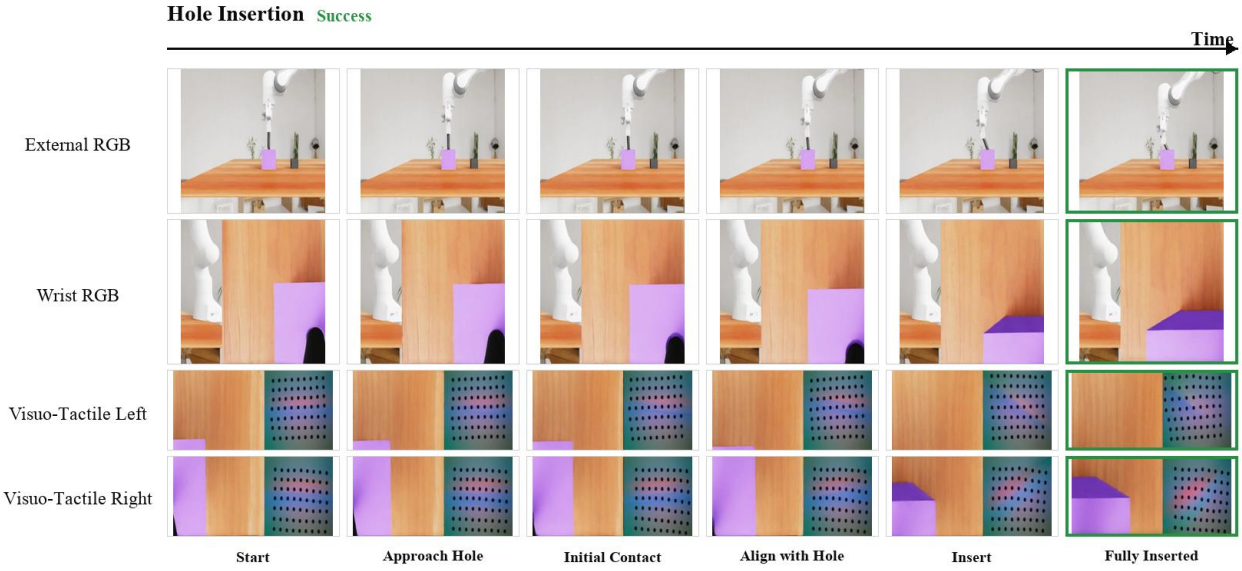

Figure S12: Successful UniVTac execution for hole insertion. Contact-guided alignment is followed by insertion to the target depth.

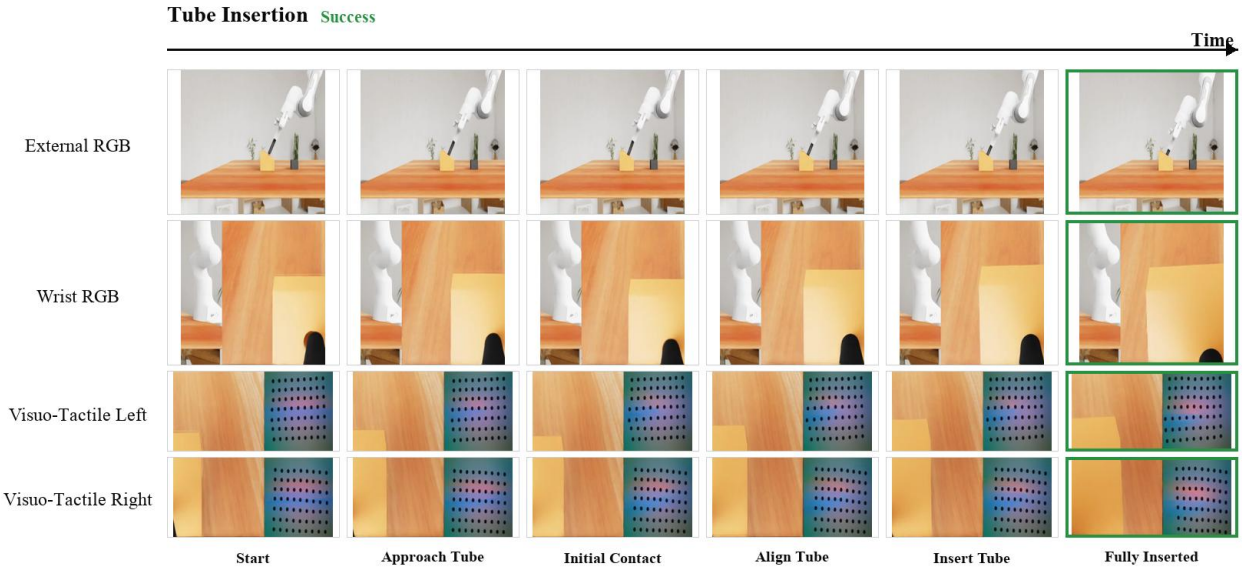

Figure S13: Successful UniVTac execution for tube insertion, including approach, grasp, coaxial alignment, insertion, and final seating.

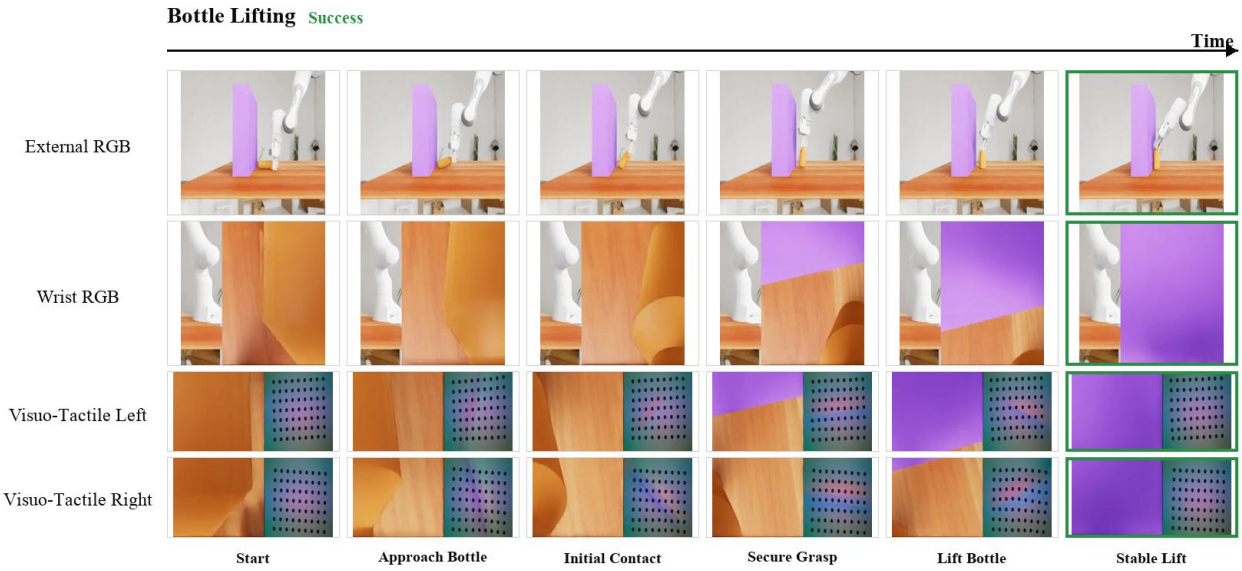

Figure S14: Successful UniVTac execution for bottle lifting. The robot forms a stable grasp and maintains contact throughout the lift.

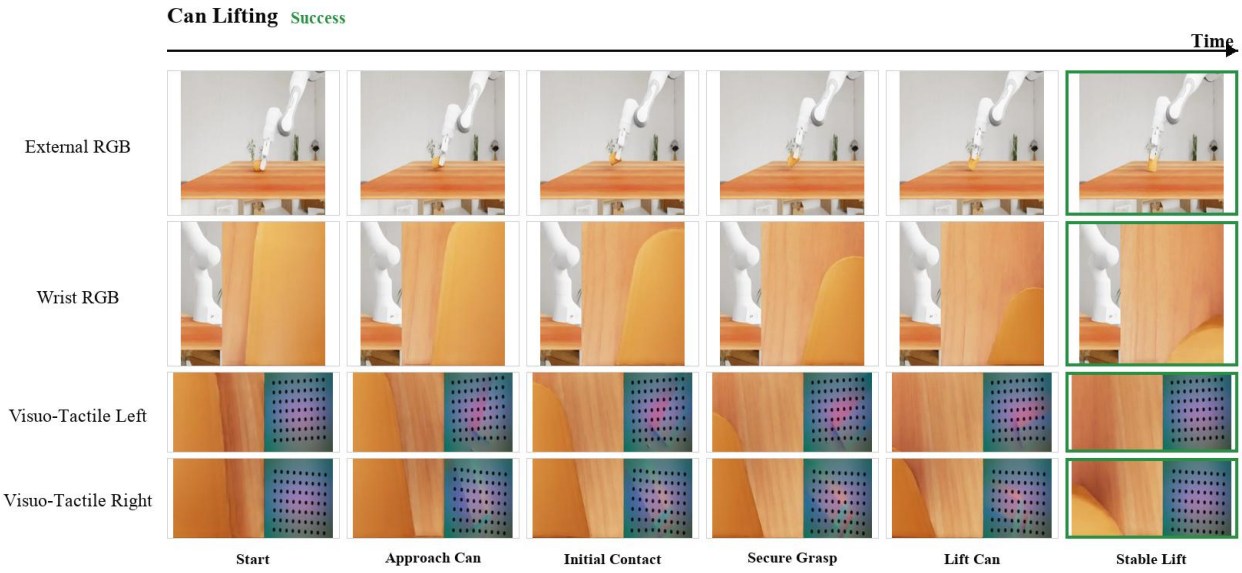

Figure S15: Successful UniVTac execution for can lifting. The can is securely grasped and lifted without slip.

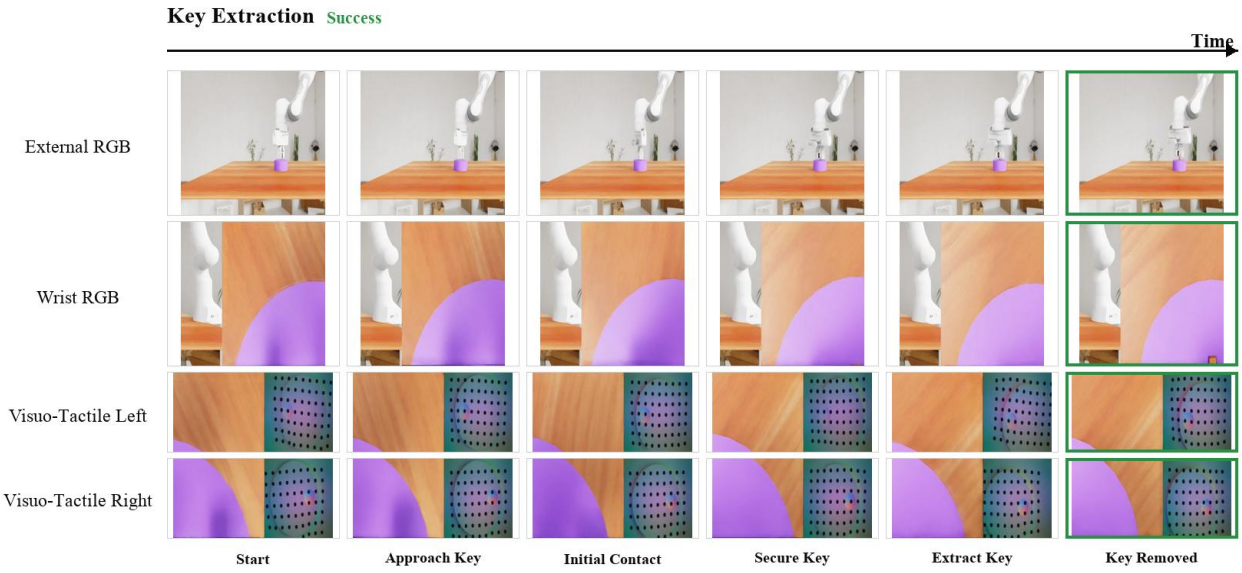

Figure S16: Successful UniVTac execution for key extraction. The robot grasps the key, maintains alignment under contact, and removes it completely.

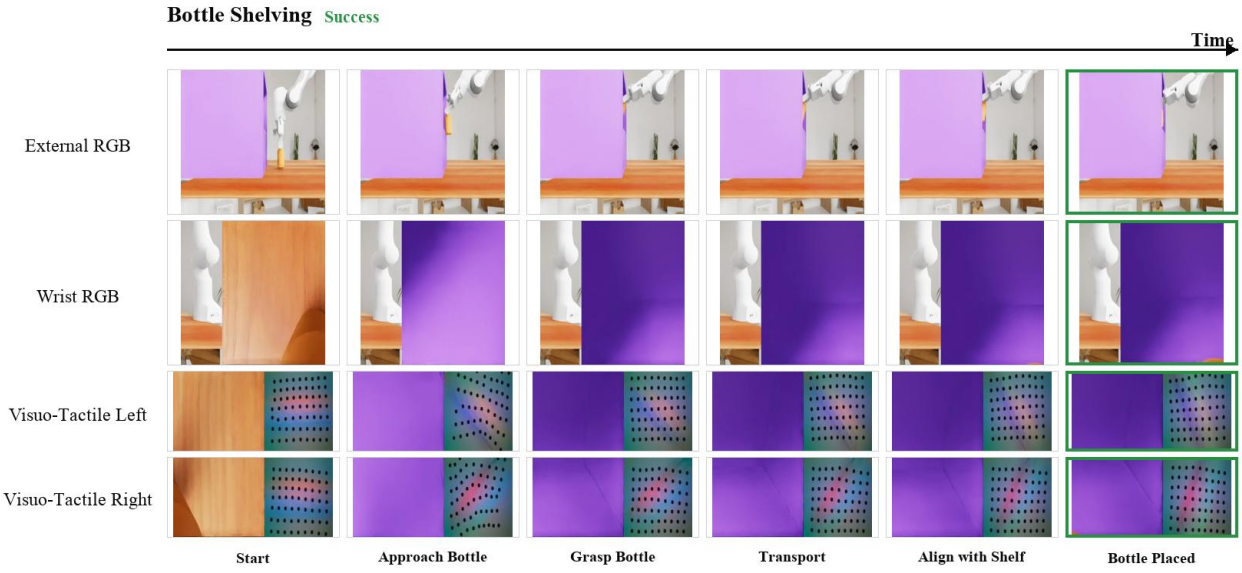

Figure S17: Successful UniVTac execution for bottle shelving. The robot grasps and transports the bottle, aligns it with the shelf, places it, and releases it.

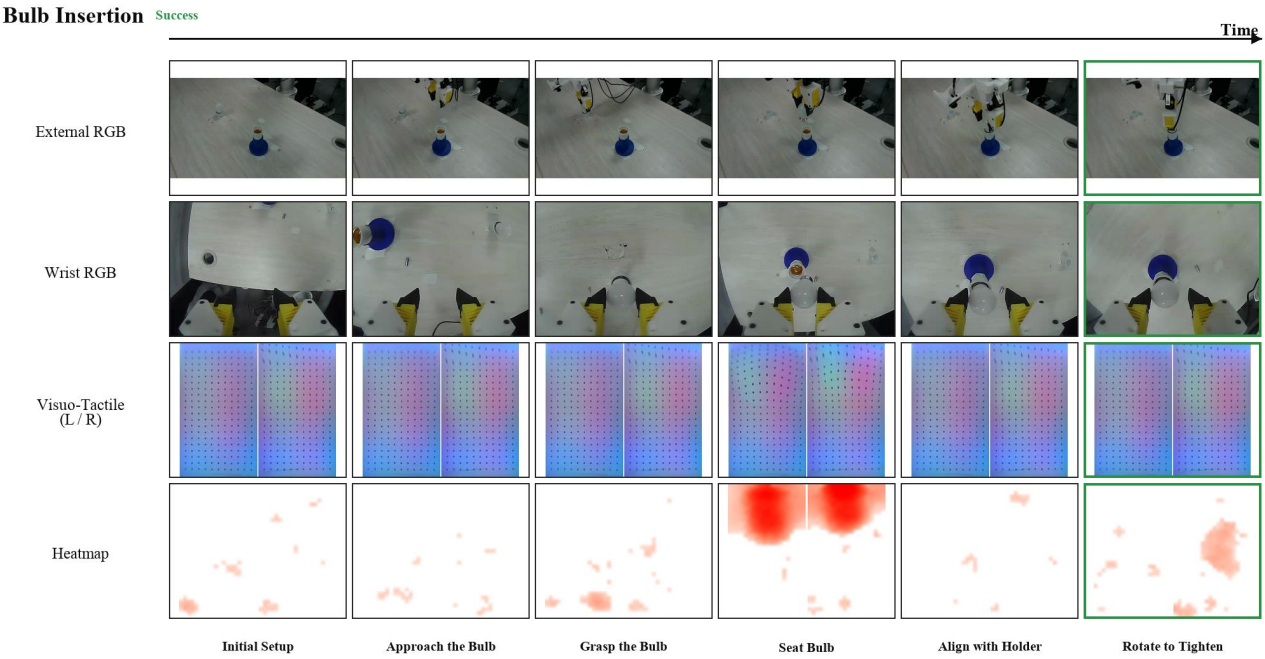

Figure S18: Successful real-robot bulb insertion. The final stages seat the bulb, align it while maintaining the grasp, and rotate it to tighten.

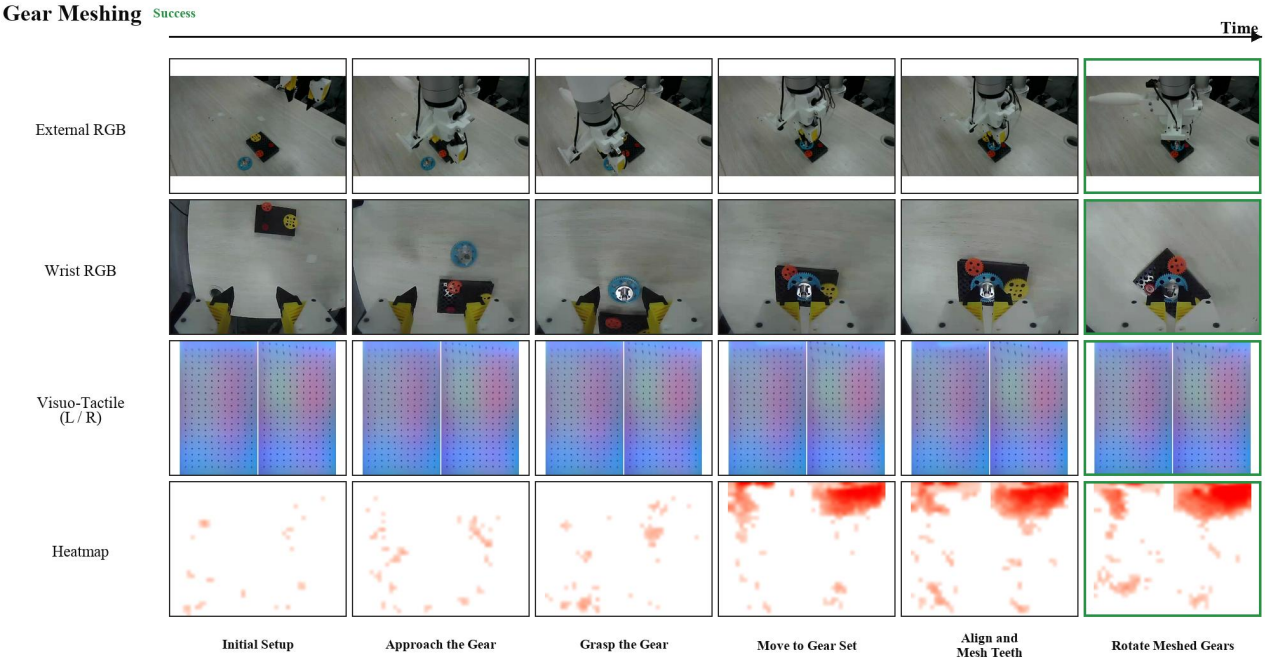

Figure S19: Successful real-robot gear meshing. After alignment and engagement, the robot rotates the meshed gear to verify functional contact.

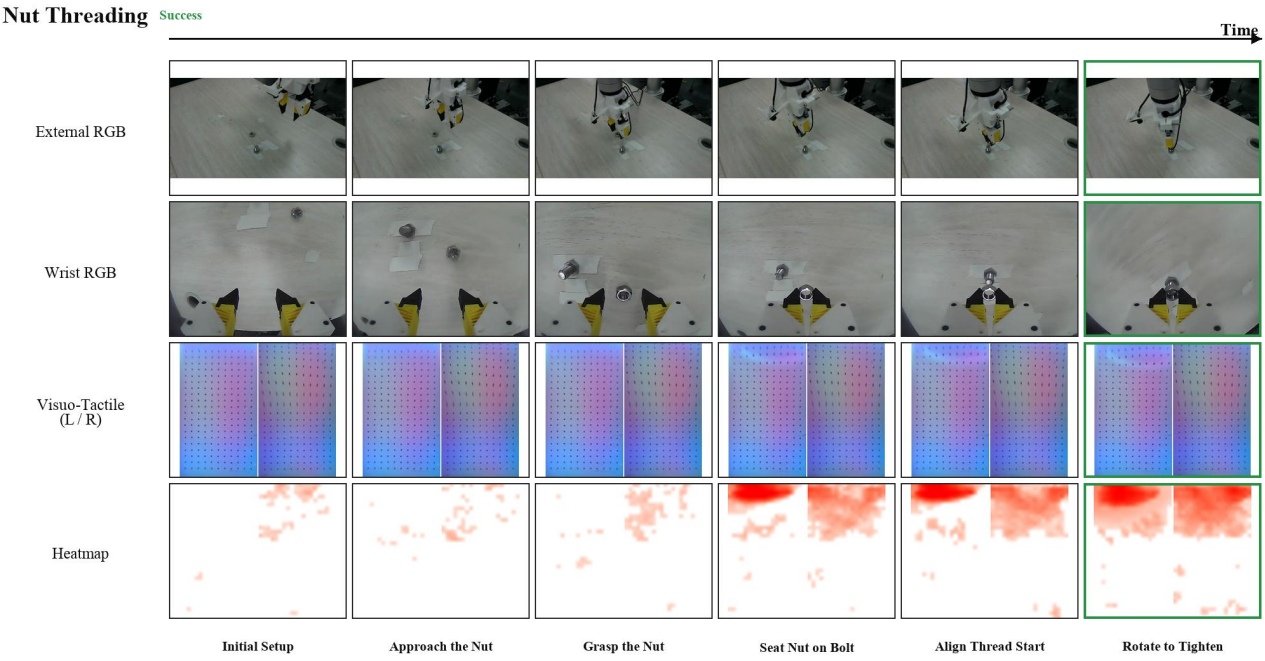

Figure S20: Successful real-robot nut threading, including approach, grasp, seating, thread alignment, and tightening.

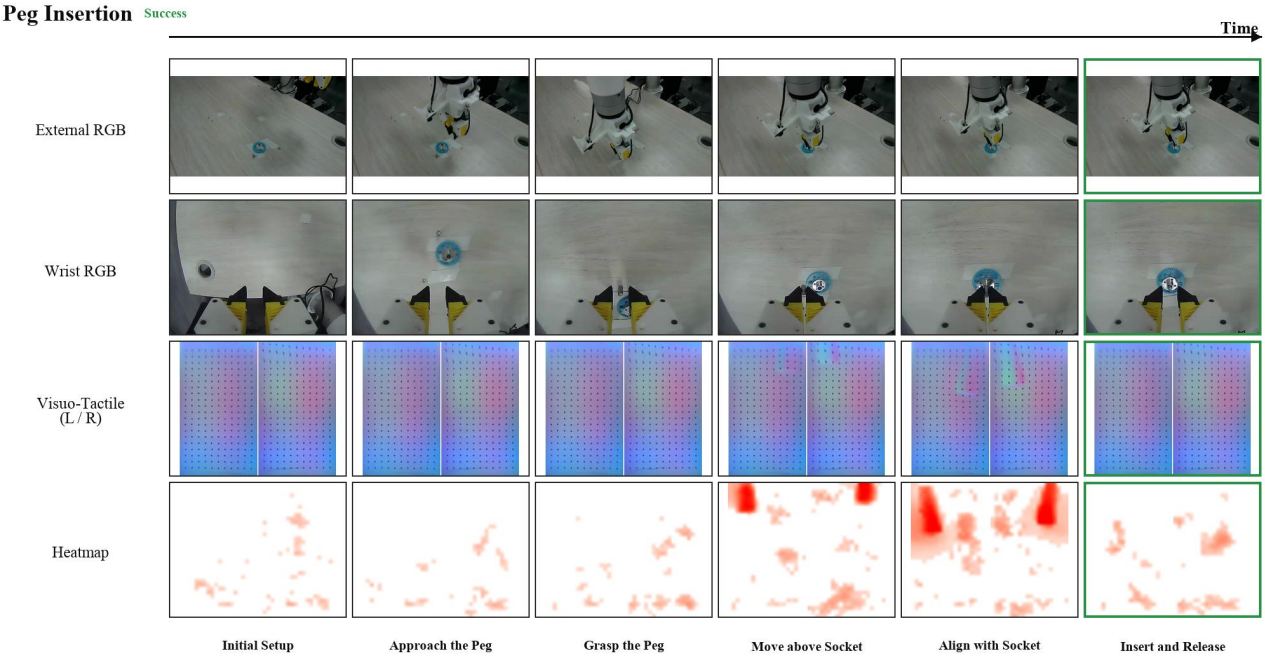

Figure S21: Successful real-robot peg insertion, from approach and grasp to socket alignment and final insertion.

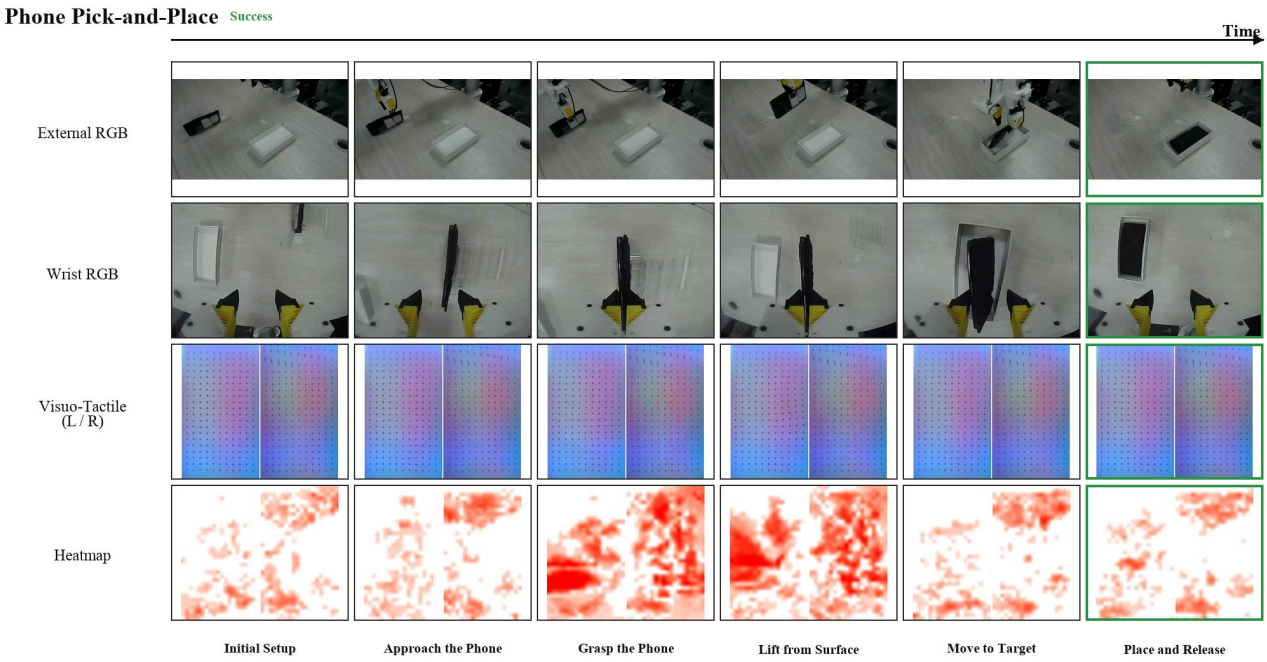

Figure S22: Successful real-robot phone pick-and-place execution, including grasp, lift, transport, placement, and release.

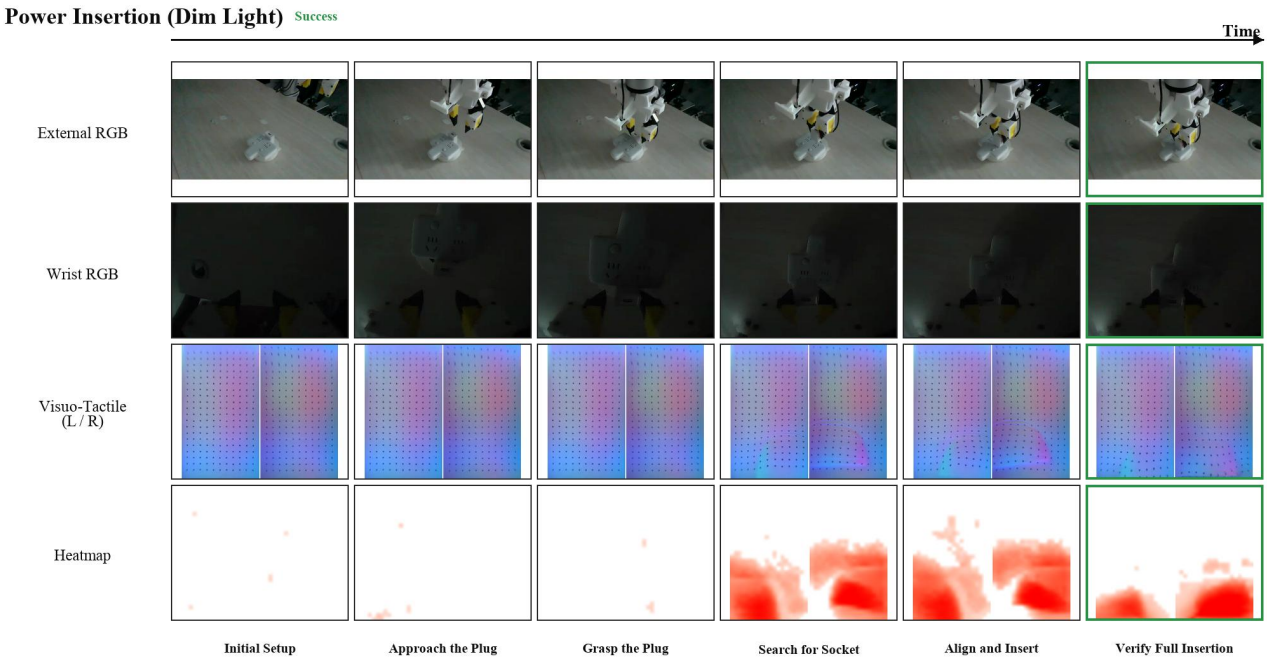

Figure S23: Successful real-robot power-plug insertion under dim lighting. The sequence shows socket search, contact-guided alignment, insertion, and verification.

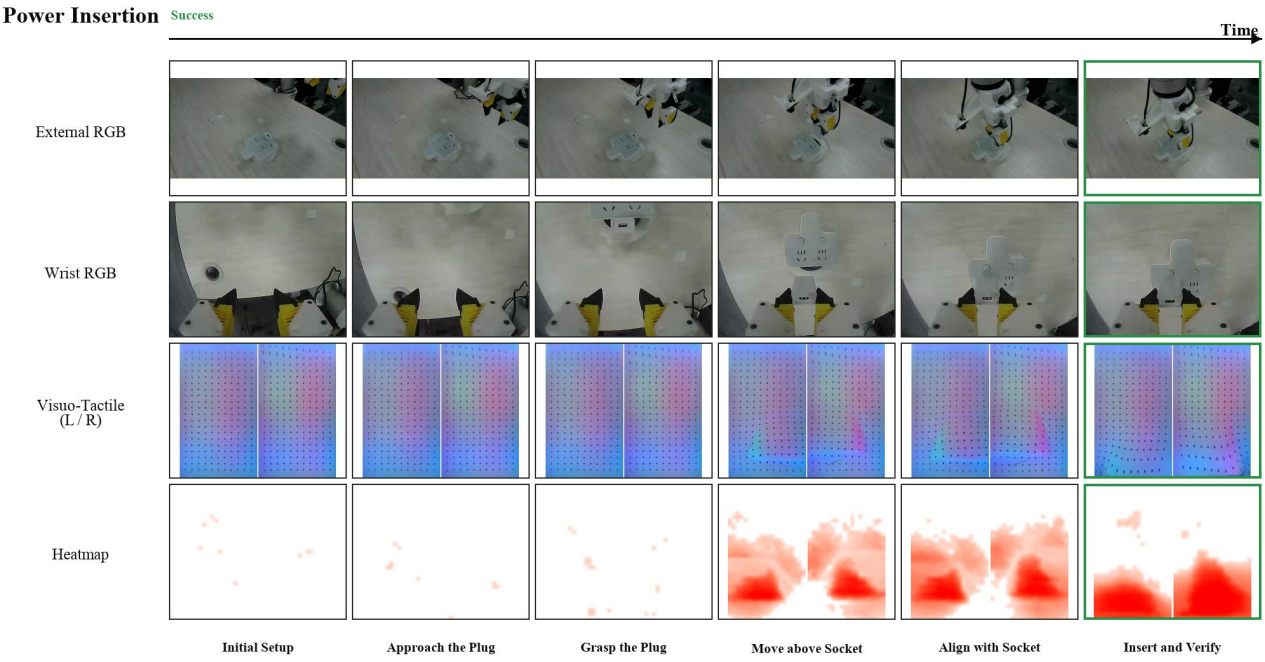

Figure S24: Successful real-robot power-plug insertion under standard lighting, including socket alignment, insertion, and final verification.