sharpa

WiTac: Wrist-Informed Tactile Representations for Contact-Rich Manipulation

Haotian He[1,2,*], **Yiyang Chen**[1,3,*], **Congsheng Xu**[1,3], **Wenchao Yue**[1,5]
Mingze Ye[1,3], **Shuai Chen**[1,4], **Yi Yang**[1,3], **Yaping Li**[1,5]
Haonan Wang[1], **Weinan E**[2], **Wenzhao Lian**[3], **Zhijie Deng**[3]

[1]Sharpa Robotics, [2]PKU, [3]SJTU, [4]ShanghaiTech, [5]CUHK

[*]Equal contribution

Abstract

Similar tactile contacts can call for different corrections depending on the configuration of the hand, object and target. Rather than supplying touch and wrist vision separately, fusing them before policy injection better grounds tactile features in visual context, yielding more accurate contact representations. We introduce Wrist-Informed Tactile Representations (WiTac), a pretrained encoder that learns such representations before action generation. Pretrained on nearly 300 hours of SharpaDex data, WiTac deeply fuses tactile signals with hand motion and wrist views into a comprehensive contact representation. Conditioning downstream VLA policies (GR00T N1.7 and $\pi_{0.5}$) on WiTac substantially improves task success over vision-only base policies across four contact-rich manipulation tasks (+28.8% and +15.0% on average), and further outperforms a wrist-free tactile encoder by 18.8% and 7.5%, respectively. Offline analyses further show that incorporating wrist vision yields more accurate contact descriptors, stronger joint task–contact discrimination, greater feature stability under tactile noise, and clearer action-direction readout from similar contacts. Project page: https://wrist-informed-tactile-representations.github.io/.

Correspondence: zhijied@sjtu.edu.cn

1 Introduction

Contact-rich manipulation requires actions that respond jointly to local touch and surrounding task geometry. When inserting a battery into a slot, similar finger loading can accompany different battery–slot alignments and call for distinct corrective motions. Local tactile deformation alone leaves this spatial configuration ambiguous, whereas close-range wrist cameras observe the hand, grasped object, and nearby target geometry, providing the spatial reference needed to interpret contact during manipulation.

Generalist vision-language-action (VLA) policies such as GR00T [17] and $\pi_{0.5}$ [18] already receive camera images in their vision-language streams, yet they lack pre-fusion with tactile signals to ground grasp- and contact-related states. Recent visuotactile designs incorporate touch into policies by supplying standalone tactile features alongside visual streams [9, 25], modulating visual features with tactile signals [15], or encoding discrete residuals from visually predicted touch [30]. However, separate tactile inputs leave local touch ungrounded in the surrounding visual geometry, while visual modulation and residual coding do not directly form continuous contact representations that preserve detailed touch within its wrist visual context.

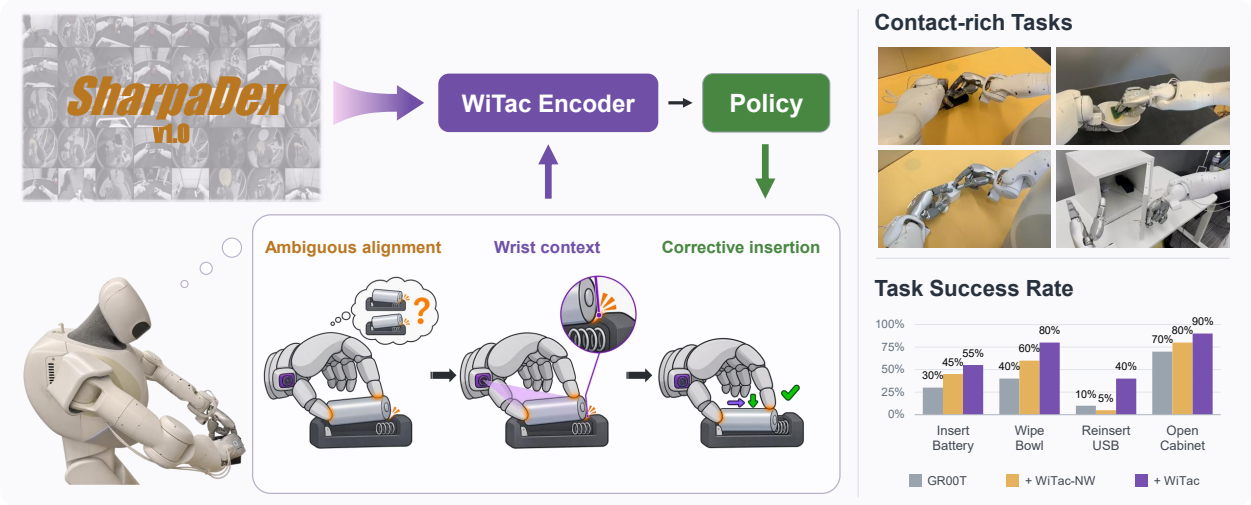


Figure 1 Wrist context complements fingertip contact. Left: WiTac is pretrained on SharpaDex v1.0 [19] and conditions a downstream policy. The battery-insertion cartoon illustrates how wrist context helps resolve alignment ambiguity during contact. Right: evaluation tasks and success rates.

To address this limitation, we introduce Wrist-Informed Tactile Representations (WiTac), a pretrained encoder that fuses touch with wrist vision before policy injection to produce grounded visuotactile contact tokens (figure 1). Pretrained on nearly 300 hours of synchronized dexterous manipulation data from SharpaDex v1.0 [19], WiTac summarizes each finger's tactile history with robot state and wrist-relative fingertip kinematics, then queries wrist patch features through cross-attention to form finger-level contact tokens. Pretraining jointly supervises contact-state and history reconstruction, masked wrist-feature recovery, and action-conditioned wrench-response prediction, encouraging the tokens to retain physical contact detail as they incorporate visual context. During policy learning, the frozen WiTac encoder conditions VLA policies through a lightweight gated cross-attention interface alongside their native visual, language and state inputs. Across four real-world contact-rich tasks, WiTac consistently boosts closed-loop success on both GR00T N1.7 and $\pi_{0.5}$ over vision-only base policies (+28.8% and +15.0% on average) and its wrist-free variant WiTac-NW (+18.8% and +7.5%), while offline analyses confirm improved contact reconstruction, joint task–contact discrimination, noise stability, and action-direction readout.

Our contributions are:

- We propose WiTac, a pretrained contact encoder that fuses tactile histories, hand kinematics, and wrist observations into finger-level contact tokens before action generation, through joint physical and visual supervision.
- We demonstrate the utility of frozen WiTac representations across two VLA backbones and four real-world contact-rich tasks, achieving higher success than vision-only policies and wrist-free tactile encoding on every task.
- Through representation analyses, we show that WiTac provides more accurate contact descriptors, stronger joint task–contact discrimination, lower feature drift under tactile noise, and more accurate action-direction readout than wrist-free encoding.

2 Related Work

Tactile and visuotactile representations. Pretrained tactile encoders provide reusable features for contact perception and manipulation. Sparsh [9] learns touch representations through self-supervised objectives on tactile images, while T3 [32] combines sensor-specific encoders with a shared Transformer across sensors and

tasks. AnyTouch [4] unifies static tactile images and dynamic tactile videos across multiple visuo-tactile sensors through masked modeling and cross-sensor alignment, and Sparsh-X [10] extends representation learning to image, audio, motion, and pressure signals from a multisensory tactile sensor. Physical supervision offers another route: UniVTAC [2] reconstructs tactile appearance and predicts deformation, marker locations, and relative object pose from simulated tactile observations. External vision and language can also shape tactile representations: early work uses self-supervision to learn compact visuotactile features for contact-rich policy learning [12], VITaL [6] aligns visual and tactile embeddings through temporal contrastive pretraining, OmniVTLA [3] aligns touch with visual and linguistic concepts in a dual-encoder architecture, and VT-MUSE [22] pretrains sequential visuotactile latents through cross-modal alignment and masked sensory reconstruction.

Visuotactile policies for parallel-jaw grippers. Vision-language-action (VLA) models such as GR00T [17] and $\pi_{0.5}$ [18] provide generalist visual, language, and motor capabilities that recent gripper-based systems extend with touch through policy conditioning or predictive world modeling. For reactive control, RDP [24] pairs a slow latent diffusion policy with high-frequency closed-loop tactile feedback. Within VLA policies, SimpleTouch [25] passes tokens from a frozen pretrained tactile encoder to a tactile expert trained with action supervision and future-latent prediction, TacFiLM [15] maps pretrained tactile features to feature-wise scale and shift parameters that modulate the visual encoder, and ResTacVLA [30] encodes the quantized residual between wrist-predicted and observed touch with uncertainty gating. To anticipate contact evolution, DreamTacVLA [26] aligns multi-scale views with touch and refines draft actions using future tactile states predicted by a world model, while OmniVTA [33] couples a two-stream visuo-tactile world model and contact-aware fusion policy with a 60 Hz reflexive controller. World-action models further unify dynamics prediction and control: VTAM [29] treats optical tactile images as an additional view of a pretrained video backbone with a flow-derived force proxy, Dream-Tac [14] jointly predicts future visual observations, tactile dynamics, and actions using a contact-aware attention bias, and FAWAM [7] and ForceWAM [23] predict future wrenches alongside actions.

Scaling tactile manipulation to dexterous hands. Moving from parallel-jaw grippers to multi-fingered dexterous hands introduces distributed contacts across fingers, higher-dimensional kinematics, and heavier visual self-occlusion. To bridge embodiments, FTP-1 [27] maps heterogeneous tactile signals into morphology-aware tokens processed by a shared tactile Transformer expert, and ME-Dex 1.0 [31] aligns gripper and dexterous-hand tactile inputs onto a canonical hand template within a three-expert world-action model. For multi-fingered hands specifically, SaTA [11] anchors tactile features to wrist-frame fingertip poses, ViTac-Former [8] fuses vision and touch via cross-attention with an autoregressive tactile-prediction head, and STAR [13] jointly predicts masked wrist-image and tactile patches during pretraining before passing separate tactile tokens to a VLA policy. At the policy and world-model level, T-Rex [16] combines a temporal tactile VQ-VAE with a variable-rate Mixture-of-Transformers to exploit high-frequency tactile feedback, DeCAL [5] pairs contact-gated fusion with latent visuo-tactile co-imagination, and DexTacWAM [28] compresses per-finger touch into a video diffusion world model to jointly predict visual and contact dynamics. Rather than training a generative video-tactile world model during policy learning or supplying deployment tactile tokens encoded without wrist views, WiTac pretrains a frozen wrist-informed contact encoder that fuses ten-finger tactile histories, kinematics, and wrist imagery before conditioning VLA policies through lightweight gated cross-attention.

3 Wrist-Conditioned Contact Representation

WiTac fuses fingertip touch, hand kinematics, and wrist views into finger-level contact tokens before action generation. We pretrain the encoder via self-supervision, then freeze it to condition a VLA policy through gated cross-attention. Figure 2(a–c) overviews contact encoding, representation pretraining, and policy injection.

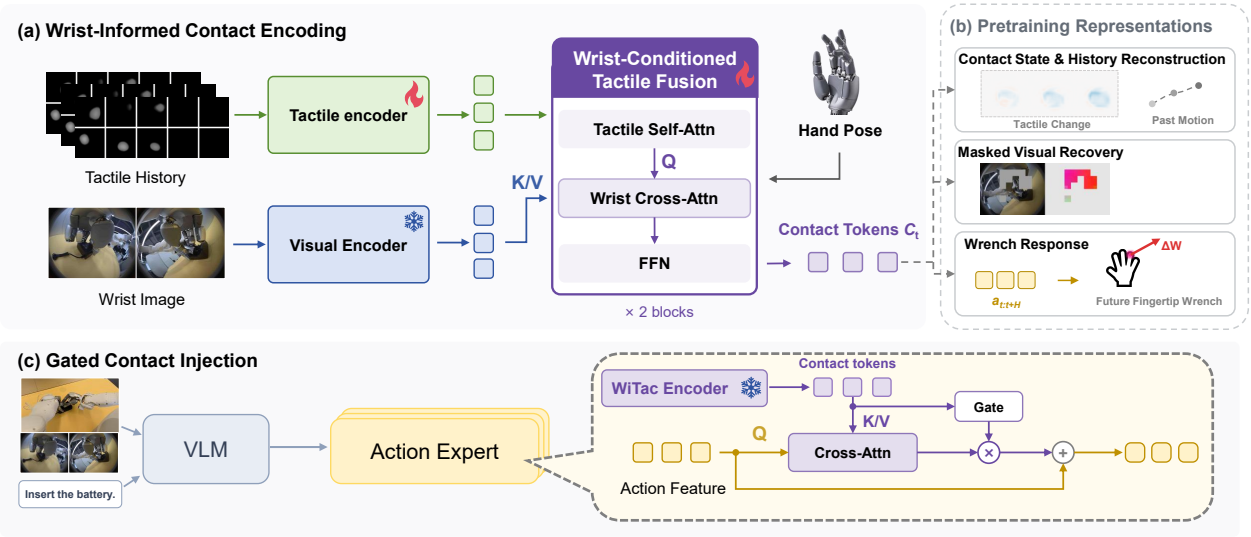


Figure 2 WiTac learns contact tokens before policy injection. (a) State-conditioned tactile histories query wrist features to form contact tokens C_t. (b) Pretraining-only heads supervise contact state, deformation and fingertip-motion histories, conditional visual recovery, and action-conditioned wrench responses. (c) Tokens from the frozen encoder condition the action expert through gated cross-attention.

3.1 Wrist-Informed Contact Encoding

Local fingertip deformation alone can be ambiguous without the surrounding hand–object configuration. Let T_t denote the 240×240 scalar deformation maps of all ten fingers at time t, and $\mathcal{T}_t$ their eight-frame history including the current frame. WiTac receives $\mathcal{T}_t$, current RGB images I_t from both wrist cameras, and the robot state s_t, while forward kinematics provides each finger's pose p_t (3D position x_t and orientation) in its respective wrist frame. From these inputs, the encoder produces one contact token per finger (figure 2(a)):

$$C_t = E_\theta(\mathcal{T}_t, I_t, s_t, p_t) \in \mathbb{R}^{10\times 768}. \tag{1}$$

To encode tactile dynamics, a per-frame spatial backbone initialized from SharpaWave Deform Encoder [20] maps each finger's deformation map to a 512-dimensional feature, and a causal temporal encoder compresses the eight-frame sequence into per-finger history features $H_t \in \mathbb{R}^{10\times 512}$. Each finger feature is projected to the 768-dimensional fusion width and spatially anchored on Fourier-encoded wrist-relative fingertip poses p_t via feature-wise linear modulation (FiLM), following SaTA [11]. Adding a learned projection of the current robot state s_t forms the state-conditioned tactile tokens entering cross-modal fusion.

In parallel, a frozen DINOv3 [21] ViT-B/16 backbone extracts patch features V_t concatenated across both wrist views, with camera-identity and 2D patch-coordinate embeddings specifying each patch's view and spatial location. Two Transformer blocks fuse the modalities: in each block, the tactile tokens first interact through finger self-attention, then query the wrist features V_t as keys and values through cross-attention, followed by a feed-forward network. Final layer normalization yields C_t, while validity masks handle unavailable observations. Each contact token thus grounds a finger's tactile history in close-range wrist visual context before policy conditioning.

3.2 Pretraining Contact Representations

We pretrain WiTac on synchronized tactile, wrist-image, robot-state, and fingertip force–torque recordings from SharpaDex v1.0 [19]. Five self-supervised objectives in three complementary groups (figure 2(b)) supervise C_t so that wrist fusion preserves local contact detail while incorporating visual context and action-dependent physical responses. All prediction heads read the shared token set C_t during pretraining and are

removed for policy inference. Finger-indexed outputs follow the ten-finger ordering of C_t, with hats denoting predictions.

Contact state and history reconstruction. Cross-attending to dense visual features risks washing out fine tactile signals. Following the sensory self-supervision rationale of Sparsh [9], three heads reconstruct current contact descriptors alongside recent deformation and fingertip-motion histories:

$$\begin{aligned} \widehat{d}_t &= D_{\text{desc}}(C_t), \\ \widehat{\Delta T}_{t,k} &= D_{\text{def}}(C_t)_k, \quad \Delta T_{t,k} = T_t - T_{t-k}, \\ \widehat{\Delta x}_{t,k} &= D_{\text{mot}}(C_t)_k, \quad \Delta x_{t,k} = x_t - x_{t-k}. \end{aligned} \tag{2}$$

Here d_t contains contact area, deformation intensity, and 2D center for each finger, while k indexes past-frame offsets from separate offset lists for deformation and motion. Because x_t and x_{t-k} are each expressed in their respective wrist frames at those timestamps, $\Delta x_{t,k}$ measures wrist-relative fingertip motion independent of global arm movement. Reconstructing these multi-step histories requires C_t to retain recent contact changes beyond the current snapshot.

Masked visual recovery. Conversely, tactile objectives alone could let the fusion blocks ignore wrist imagery. Following feature-space masked prediction [1], we extract visible wrist features V_t^{vis} from a separate DINOv3 pass with native patch masking and decode masked patch features through C_t:

$$\widehat{V}_t = D_{\text{vis}}(C_t, V_t^{\text{vis}}). \tag{3}$$

The loss compares $\widehat{V}_t$ with the detached full-image features V_t only at masked patch positions. Because complete wrist features V_t enter the fusion encoder, this objective trains conditional feature recovery rather than occluded-region inpainting. Routing information from 640 wrist patches through only ten finger tokens creates a bottleneck that compels C_t to carry wrist visual context.

Wrench response prediction. To capture how the current contact configuration physically responds to upcoming actions [12], we train an action-conditioned head to predict fingertip wrench changes over a future horizon h:

$$\widehat{\Delta w}_{t,h} = D_{\text{wr}}(C_t, A_{t,h}), \qquad \Delta w_{t,h} = w_{t+h} - w_t. \tag{4}$$

Here w_t collects each finger's six-dimensional force–torque vector, and $A_{t,h}$ is the future action prefix up to $t+h$. Future actions enter only this training-time decoder, whereas H_t and C_t depend solely on causal observations available at time t.

We minimize the weighted objective

$$\mathcal{L}_{\text{pre}} = \sum_{j=1}^{5} \lambda_j \mathcal{L}_j, \tag{5}$$

with validity masks for missing entries and balanced weighting between changing and nearly static regions in deformation reconstruction. Gradients through C_t jointly optimize the spatial and temporal tactile encoders together with the fusion blocks, while DINOv3 remains frozen. Temporal offsets, horizons, and objective weights are listed in Appendix A.1 (table S1), and pretraining uses no task-success labels.

3.3 Gated Contact Injection into Action Policies

For policy learning, we freeze WiTac and condition the action expert of a VLA policy through a shared gated cross-attention interface, while preserving the policy's native visual, language, and state inputs (figure 2(c)). A learned projection P maps C_t to the hidden dimension of the action expert, and selected action blocks query the projected contact tokens as keys and values:

$$X_l' = X_l + g_l(C_t)\ \text{Cross-Attn}_l(X_l, P(C_t)), \tag{6}$$

where X_l denotes the action features at block l. A shared gate predictor averages valid tokens in C_t, applies layer normalization and an MLP, and passes the outputs through a rescaled sigmoid function in $[0.05, 1]$ to produce one scalar gate $g_l(C_t)$ per injected block. This bounded gate keeps a small nonzero floor when valid touch is present while scaling up the contact residual during physical interaction. Because the interface modifies only selected action blocks, a pretrained WiTac encoder can condition different VLA backbones through policy-specific projections and insertion layers.

We jointly optimize action prediction and contact-gate supervision:

$$\begin{aligned} \mathcal{L}_{\text{policy}} &= \mathcal{L}_{\text{action}} + \lambda_{\text{gate}} \mathcal{L}_{\text{gate}}, \\ \mathcal{L}_{\text{gate}} &= \text{mean}_{t,l}\, \text{BCE}(g_l(C_t), y_t)\,, \end{aligned} \tag{7}$$

where $\mathcal{L}_{\text{action}}$ is the base policy's flow-matching loss and $\mathcal{L}_{\text{gate}}$ averages binary cross-entropy over injected blocks and labeled timesteps. The contact proxy $y_t \in \{0, 1\}$ is computed from the active-pixel fraction on the hands in T_t, encouraging stronger conditioning during physical contact. Throughout policy training, WiTac remains frozen while the contact interface and action expert are optimized jointly.

4 Experiments

Our experiments address three questions:

Q1. Does WiTac improve real-world contact-rich manipulation over base policies and tactile encoders without wrist fusion?

Q2. How does wrist conditioning affect contact reconstruction, context and action-direction readout, and sensitivity to tactile noise?

Q3. How do representation pretraining, visual supervision, and pre-injection fusion contribute to task success?

4.1 Experimental Setup

Robot and tasks. We evaluate on a Sharpa North robot with two Sharpa Wave dexterous hands. Its 65-dimensional joint-command action space comprises 7 joints per arm, 22 per hand, 2 head/neck joints, and 5 lower-body joints. We evaluate four bimanual contact-rich tasks: battery insertion (seating a battery into a slot), bowl wiping (scrubbing the inner bowl wall with a sponge), USB reinsertion (extracting and reinserting a connector on a power strip), and cabinet opening (turning a key and pulling open a door), with detailed task definitions in Appendix B.1. None of these four tasks are included in the SharpaDex v1.0 pretraining dataset; for each task, we collect 200 demonstrations through exoskeleton teleoperation for policy training.

Baselines. We integrate WiTac into GR00T N1.7 [17] (GR00T hereafter) and $\pi_{0.5}$ [18]. For each backbone, we compare against the base policy and its wrist-free variant WiTac-NW, which removes only wrist cross-attention and masked visual recovery while keeping everything else unchanged. On GR00T, we additionally pretrain the visuotactile encoder of VT-MUSE [22] on SharpaDex v1.0 and inject it into GR00T (VT-MUSE joint repr.). We also compare with two standalone tactile policies: FTP-1 [27], which unifies heterogeneous sensor inputs through morphology-aware tactile tokens and a shared Transformer expert, and T-Rex [16], which couples a temporal tactile encoder with a variable-rate Mixture-of-Transformers architecture to incorporate high-frequency touch into action generation.

Evaluation protocols and metrics. Each method is evaluated over 20 trials per task. We additionally measure contact-related behaviors and analyze frozen representations through contact-reconstruction error, joint task–contact macro-F1, feature/action stability under tactile noise, and action-direction cosine alignment. Each analysis uses its specified evaluation split, and learned readouts use matched head capacity across representations. Training settings are detailed in Appendix A.1 and Appendix A.2; full analysis protocols are provided in Appendix B.3.

Table 1 Closed-loop success rates (%). Average is the unweighted mean across all four tasks. WiTac-NW denotes WiTac without wrist conditioning.

Policy	Insert battery	Wipe bowl	Reinsert USB	Open cabinet	Average
FTP-1 [27]	40	55	5	60	40.0
T-Rex [16]	45	65	15	75	50.0
GR00T N1.7 [17]	30	40	10	70	37.5
+ WiTac-NW	45	60	5	80	47.5
+ VT-MUSE [22] joint repr.	40	70	15	75	50.0
+ WiTac	**55**	**80**	**40**	**90**	**66.3**
$\pi_{0.5}$ [18]	55	65	10	45	43.8
+ WiTac-NW	60	65	15	65	51.3
+ WiTac	**70**	**75**	**20**	**70**	**58.8**

4.2 Real-World Manipulation (Q1)

Manipulation performance across policies. WiTac improves closed-loop success over the base policy and WiTac-NW on all four tasks with both VLA backbones (table 1). With GR00T, WiTac reaches an average success rate of 66.3%, exceeding the vision-only base policy (37.5%) by 28.8%, WiTac-NW (47.5%) by 18.8%, and the standalone tactile policies FTP-1 (40.0%) and T-Rex (50.0%) on every task. Notably, on USB reinsertion, adding wrist-free touch (WiTac-NW) drops success from 10% to 5%, whereas WiTac raises it to 40% (+35%), indicating that ungrounded fingertip touch can mislead fine extraction and insertion without wrist spatial context. Under the same SharpaDex v1.0 pretraining and GR00T backbone, VT-MUSE joint repr. lifts average success to 50.0% over WiTac-NW (47.5%), yet WiTac surpasses it by 16.3% on average, demonstrating the effectiveness of our wrist-conditioned, finger-level contact encoder design. Reusing the frozen WiTac encoder with $\pi_{0.5}$ similarly lifts average success from 43.8% (base) and 51.3% (WiTac-NW) to 58.8%, with consistent 5–10% gains over WiTac-NW across all tasks. Qualitative rollouts for all four tasks are provided in Appendix B.2 (figure S1).

Contact-related behaviors. We examine two critical contact stages on GR00T: battery alignment and USB extraction. Figure 3 contrasts WiTac-NW and WiTac rollouts from matched initial conditions. WiTac-NW rotates the battery out of alignment and lifts its fingers without extracting the USB connector, whereas WiTac adjusts contact to seat the battery and extract the connector. These examples accompany the aggregate results in table 2: across 20 trials per task, WiTac reduces battery misalignment failures and improves USB extraction success.

Table 2 Contact-stage outcomes with GR00T (20 trials per task). Misalignment failures during battery insertion and partial-task extraction success on USB reinsertion.

Method	Misalignment failures ↓	Extraction success (%) ↑
GR00T	7	50
GR00T + WiTac-NW	4	60
GR00T + WiTac	**1**	**95**

4.3 Representation Analysis (Q2)

To understand how wrist conditioning shapes the learned tokens, we probe frozen WiTac and WiTac-NW across contact reconstruction, context readout, tactile-noise perturbations, and action-direction readout.

More accurate contact reconstruction. We assess three complementary aspects of contact: area describes its spatial extent, intensity measures the magnitude of tactile deformation, and center locates the contact on the sensor. We evaluate the pretrained reconstruction heads on 9,632 validation windows from 602 episodes in the SharpaDex v1.0 dataset. WiTac reduces area and intensity MAE by 7.57% and 10.00% over all windows, and contact-center error by 2.64%, with area and intensity errors also decreasing on contact-containing samples by 1.30% and 5.53%, respectively. These results support more accurate recovery of contact

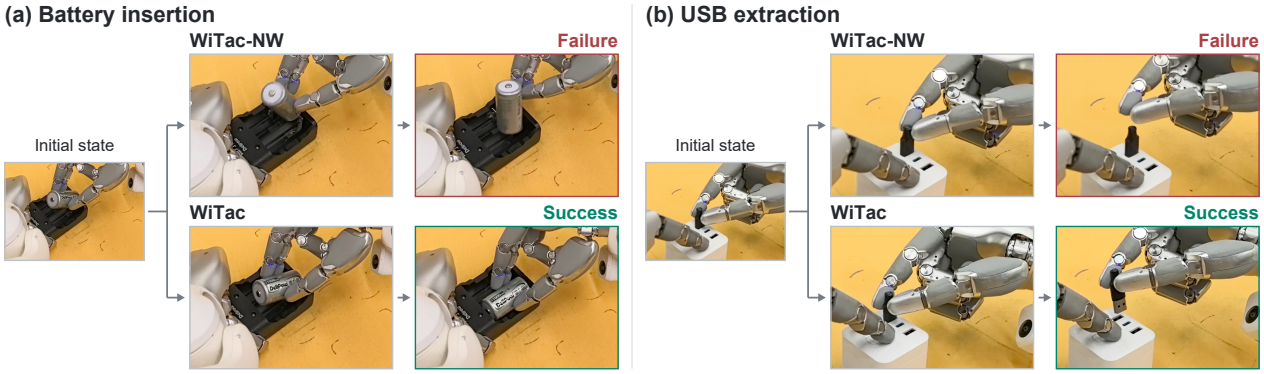


Figure 3 Contact-stage rollouts from matched initial conditions. Sequences progress left to right. In these examples, WiTac-NW rotates the battery out of alignment and leaves the USB connector inserted, whereas WiTac seats the battery and extracts the connector.

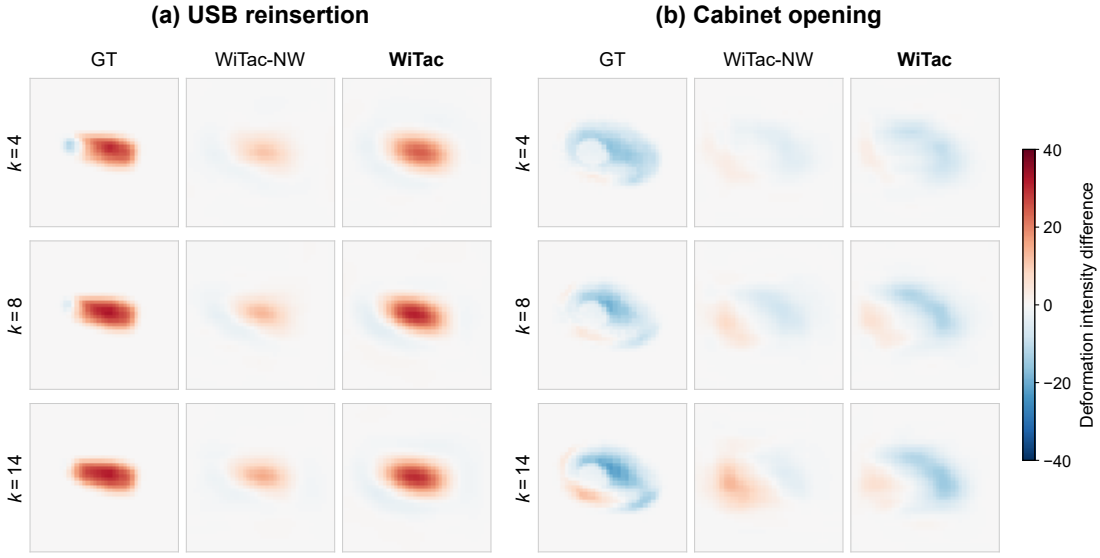


Figure 4 Qualitative examples of contact-change reconstruction. Right-index deformation differences $T_t - T_{t-k}$ at offsets $k \in \{4, 8, 14\}$ frames. Methods share a fixed crop within each case and a signed color scale.

descriptors (full results in table S4). The USB-reinsertion and cabinet-opening examples in figure 4 illustrate improved recovery of deformation changes.

Joint readout of task context and contact state. We next compare WiTac against WiTac-NW, wrist DINO features, and WiTac-NW+DINO concatenation on predicting both the task and its contact state. Each target is a task–state pair (e.g., battery insertion with stable contact), combining the four tasks with short-window tactile labels: the 12-class setting uses no, stable, and changing contact, while the 20-class setting adds contact onset and release. Each representation is reduced to 50 PCA dimensions and evaluated with an equal-capacity classifier using macro-F1 on the original test split and a held-out confirmation split.

WiTac improves macro-F1 over DINO by 4.39–10.14% and over WiTac-NW by 1.83–7.65% across all four evaluations (figure 5; full scores in table S5). These gains support the complementary roles of visual context and tactile information in distinguishing task–contact states. Compared with WiTac-NW+DINO concatenation, the absolute differences are at most 1.03% in three settings, with a larger 4.86% gain on the 20-class confirmation split. Thus, while joint encoding does not consistently surpass concatenation on this readout,

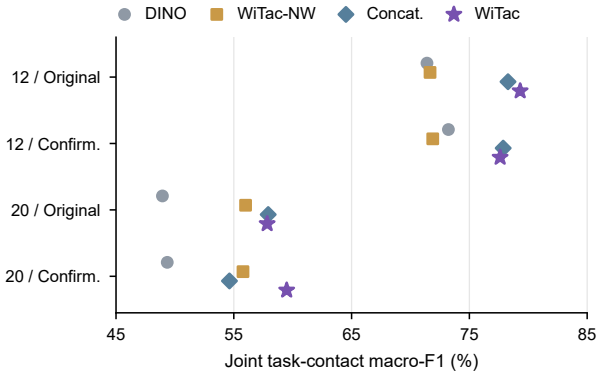


Figure 5 Joint task–contact readout. Macro-F1 (%) over four tasks with three or five contact states, giving 12 or 20 classes. Original/Confirm. identify the two test splits; Concat. denotes WiTac-NW+DINO.

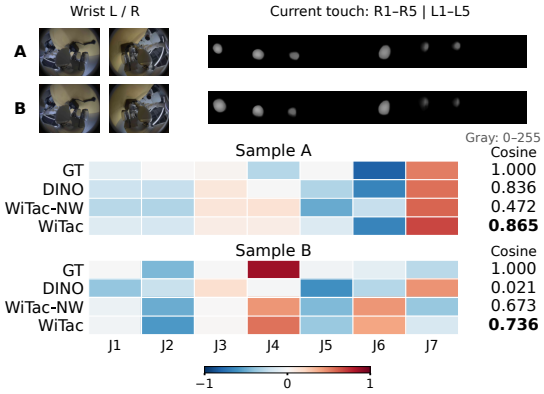


Figure 6 Action-direction readout on a battery test pair. Top: wrist views and current touch for observations A and B with similar contact patterns. Bottom: unit-normalized GT and linear-probe right-arm joint adjustments ($\Delta q_t/\|\Delta q_t\|_2$, J1–J7), with cosine similarity to GT on the right.

stronger fusion mechanisms for fine-grained contact-state discrimination remain a promising direction.

Feature stability and noisy-input action accuracy. Real-world tactile sensors often exhibit hardware instability and measurement noise during deployment. To evaluate stability under such perturbations, we inject Gaussian noise ($\sigma = 3$ in 8-bit intensity) into tactile inputs while keeping wrist images and robot state unchanged. We measure two metrics on identical perturbed windows: *feature drift* $D = 1-\cos(z_{\text{clean}}, z_{\text{noisy}})$ averaged over valid contact tokens z, and *noisy-input action MAE* (mean joint-angle error in degrees against demonstrated actions). Table 3 reports the percentage error reduction of WiTac relative to WiTac-NW (G_D and G_{MAE}, where positive values indicate lower drift or error than WiTac-NW). WiTac reduces feature drift by 35.6–44.1% across all four tasks. Downstream action MAE also drops by 7.35–22.76% across all four GR00T tasks, while $\pi_{0.5}$ improves on USB reinsertion and cabinet opening with minor regressions on battery insertion and bowl wiping. Overall, wrist conditioning consistently stabilizes contact features against tactile noise, whereas action-level gains depend on the downstream policy and task (full metrics in tables S6 and S7, Appendix B.3).

Table 3 Feature stability improves; action gains vary. G_D (shared across policies for the frozen encoder) and G_{MAE} are relative reductions (%) in feature drift and noisy-input action MAE versus WiTac-NW under Gaussian noise ($\sigma = 3$). Positive values indicate improvement.

Task	G_D	G_{MAE}	
		GR00T	$\pi_{0.5}$
Insert battery	+37.4	+22.76	−0.56
Wipe bowl	+35.6	+7.35	−1.90
Reinsert USB	+37.0	+9.18	+1.91
Open cabinet	+44.1	+21.59	+2.19

Action-direction readout. Similar contacts can call for different corrective motions depending on hand–object alignment. Using frozen features and equal-capacity linear probes (PCA16) fitted on separate training episodes, we predict upcoming joint-command changes $\widehat{\Delta q}_t$ and measure cosine similarity $\cos(\widehat{\Delta q}_t, \Delta q_t)$ against demonstrated short-horizon adjustments Δq_t across 25 contact-matched test pairs per task, where paired observations share similar touch but precede distinct Δq_t (table 4; protocol in Appendix B.3). Averaging cosine similarity over both observations in each pair, WiTac achieves the highest directional alignment across all four tasks. Figure 6 illustrates a battery test pair on the 7-DoF right arm

Table 4 Action-direction readout on matched test pairs. Mean cosine similarity $\cos(\widehat{\Delta q}_t, \Delta q_t) \uparrow$ across 25 test pairs per task.

Task	DINO	WiTac-NW	WiTac
Insert battery	0.6814	0.6794	**0.7325**
Reinsert USB	0.3281	0.3621	**0.4295**
Wipe bowl	0.7246	0.7696	**0.7881**
Open cabinet	0.6993	0.6669	**0.7386**

(J1–J7), where WiTac recovers both adjustment directions (0.865 and 0.736) while DINO and WiTac-NW fail on one.

4.4 Ablation Studies (Q3)

Using GR00T as the policy backbone, we ablate representation pretraining, masked visual supervision, and pre-injection fusion on downstream task success. Variant definitions are provided in Appendix B.4.

Contribution of representation pretraining. In the variant without representation pretraining, the SharpaWave Deform Encoder remains frozen while the remaining WiTac modules are trained jointly with the policy during task fine-tuning. Pretrained WiTac achieves 5–25% higher success across all four tasks, improving the average success rate from 53.8% to 66.3% (table 5), with the largest gain on USB reinsertion (15% to 40%). These gains demonstrate the effectiveness of our representation pretraining prior to policy learning.

Table 5 Design ablations on GR00T. Closed-loop success rates (%); full WiTac combines representation pretraining, masked visual recovery, and pre-injection joint encoding.

Variant	Insert battery	Wipe bowl	Reinsert USB	Open cabinet	Average
WiTac w/o pretraining	45	75	15	80	53.8
WiTac w/o visual recovery	45	**80**	25	85	58.8
Separate injection	40	65	15	80	50.0
Joint (WiTac)	**55**	**80**	**40**	**90**	**66.3**

Effect of masked visual recovery. To test whether cross-attending to wrist features suffices without visual supervision, we pretrain *WiTac w/o visual recovery* on SharpaDex v1.0 by keeping wrist inputs and all contact, motion, and wrench losses while removing masked visual recovery (D_{vis}). Although wrist views without D_{vis} raise average success to 58.8% over WiTac-NW (47.5%) and unpretrained WiTac (53.8%), omitting D_{vis} reduces average success by 7.5% versus full WiTac (66.3%, table 5), with the sharpest drops on fine-alignment tasks such as USB reinsertion (25% vs. 40%) and battery insertion (45% vs. 55%). Supervising masked wrist recovery thus keeps the ten-token bottleneck grounded in visual context.

Joint encoding also exceeds separate injection. Both joint encoding and separate injection supply tactile and wrist visual features to the policy, differing in whether the two modalities enter the policy through independent attention branches or are fused beforehand into unified contact tokens. Fusing them before policy injection yields 10–25% higher success across all four tasks, raising the four-task average from 50.0% to 66.3% (table 5). This confirms that grounding touch in wrist visual context prior to policy injection is substantially more effective than leaving cross-modal fusion to the action policy.

5 Conclusion

We presented WiTac, a pretrained encoder that integrates tactile histories, hand kinematics, and wrist vision into finger-level contact tokens before action generation. Across four contact-rich tasks, WiTac improves average success over its wrist-free variant WiTac-NW by 18.8% and 7.5% with GR00T N1.7 and $\pi_{0.5}$, respectively. Offline analyses further show that WiTac reconstructs contact descriptors more accurately, better distinguishes joint task–contact states, exhibits lower feature drift under synthetic tactile noise, and better aligns with demonstrated adjustment directions than WiTac-NW. Together, these findings support incorporating wrist context into tactile representations for dexterous manipulation.

Limitations. WiTac uses feature-level cross-attention without depth-based alignment of vision and touch in a shared 3D space, despite incorporating fingertip kinematics. It combines tactile histories with only current wrist images, without explicitly modeling their joint temporal evolution. Our evaluation is also limited to four tasks on one robotic platform; transfer across hand morphologies, tactile sensors, and camera configurations remains to be validated.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

[2] Baijun Chen, Weijie Wan, Tianxing Chen, Xianda Guo, Congsheng Xu, Yuanyang Qi, Haojie Zhang, Longyan Wu, Tianling Xu, Zixuan Li, Yizhe Wu, Rui Li, Xiaokang Yang, Ping Luo, Wei Sui, and Yao Mu. UniVTAC: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking. arXiv preprint arXiv:2602.10093, 2026. URL https://arxiv.org/abs/2602.10093.

[3] Zhengxue Cheng, Yiqian Zhang, Anni Tang, Keyu Wang, Wenkang Zhang, Haoyu Li, Hengdi Zhang, and Li Song. OmniVTLA: Vision-tactile-language-action models with semantic-aligned tactile sensing. arXiv preprint arXiv:2508.08706, 2025. URL https://arxiv.org/abs/2508.08706.

[4] Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. AnyTouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. arXiv preprint arXiv:2502.12191, 2025. URL https://arxiv.org/abs/2502.12191.

[5] Yankai Fu, Ning Chen, Junkai Zhao, Heng Zhang, Guocai Yao, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. DeCAL: Towards physically-grounded dexterous vision-language-action models via contact-aware latent co-imagination. arXiv preprint arXiv:2609.09119, 2026. URL https://arxiv.org/abs/2609.09119.

[6] Abraham George, Selam Gano, Pranav Katragadda, and Amir Barati Farimani. VITaL pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. arXiv preprint arXiv:2403.11898, 2024. URL https://arxiv.org/abs/2403.11898.

[7] Haotian He, Zeyu Yan, Qipeng Liu, Ning Guo, and Wenzhao Lian. FAWAM: Force-aware world action models for closed-loop contact-rich manipulation. arXiv preprint arXiv:2606.08555, 2026. URL https://arxiv.org/abs/2606.08555.

[8] Liang Heng, Haoran Geng, Kaifeng Zhang, Pieter Abbeel, and Jitendra Malik. ViTacFormer: Learning cross-modal representation for visuo-tactile dexterous manipulation. arXiv preprint arXiv:2506.15953, 2025. URL https://arxiv.org/abs/2506.15953.

[9] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. arXiv preprint arXiv:2410.24090, 2024. URL https://arxiv.org/abs/2410.24090.

[10] Carolina Higuera, Akash Sharma, Taosha Fan, Chaithanya Krishna Bodduluri, Byron Boots, Michael Kaess, Mike Lambeta, Tingfan Wu, Zixi Liu, Francois Robert Hogan, and Mustafa Mukadam. Tactile beyond pixels: Multisensory touch representations for robot manipulation. arXiv preprint arXiv:2506.14754, 2025. URL https://arxiv.org/abs/2506.14754.

[11] Jialei Huang, Yang Ye, Yuanqing Gong, Xuezhou Zhu, Yang Gao, and Kaifeng Zhang. Spatially-anchored tactile awareness for robust dexterous manipulation. arXiv preprint arXiv:2510.14647, 2025. URL https://arxiv.org/abs/2510.14647.

[12] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. arXiv preprint arXiv:1810.10191, 2018. URL https://arxiv.org/abs/1810.10191.

[13] Xiangcheng Liu, Tianhao Wu, Le Zheng, Yidong Wang, Bowen Jiang, Mingjie Pan, Xinlin Ren, Yi Liu, and Jianlan Luo. STAR: Sparse tactile representation learning in vision-tactile-language-action models for dexterous manipulation. arXiv preprint arXiv:2609.12549, 2026. URL https://arxiv.org/abs/2609.12549.

[14] Yunfan Lou, Yifan Ye, Yankai Fu, Jun Cen, Xiaowei Chi, Yaoxu Lyu, Peidong Jia, Sirui Han, Zhihe Lu, and Shanghang Zhang. Dream-Tac: A unified tactile world action model for contact-rich robot manipulation. arXiv preprint arXiv:2606.08737, 2026. URL https://arxiv.org/abs/2606.08737.

[15] Charlotte Morissette, Amin Abyaneh, Wei-Di Chang, Anas Houssaini, David Meger, Hsiu-Chin Lin, Jonathan Tremblay, and Gregory Dudek. Tactile modality fusion for vision-language-action models. arXiv preprint arXiv:2603.14604, 2026. URL https://arxiv.org/abs/2603.14604.

[16] Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, Ryan Punamiya, Mengda Xu, Yuqi Xie, Yunfan Jiang, Letian Fu, Konstantinos Kallidromitis, Matteo Gioia, Junyi Zhang, Jiaxin Ge, Haiwen Feng, Fabio Galasso, Wei Zhan, David M.

Chan, Yutong Bai, Roei Herzig, Jiahui Lei, Fei-Fei Li, Ken Goldberg, Jitendra Malik, Pieter Abbeel, Yuke Zhu, Danfei Xu, Linxi Fan, and Trevor Darrell. T-Rex: Tactile-reactive dexterous manipulation. arXiv preprint arXiv:2606.17055, 2026. URL https://arxiv.org/abs/2606.17055.

[17] NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734, 2025. URL https://arxiv.org/abs/2503.14734.

[18] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025. URL https://arxiv.org/abs/2504.16054.

[19] Sharpa. SharpaDex v1.0: Large-scale bimanual dexterous manipulation demonstrations. https://huggingface.co/datasets/Sharpa-Robotics/SharpaDex-v1.0, 2026. CC BY-NC 4.0, https://creativecommons.org/licenses/by-nc/4.0/.

[20] Sharpa Group. SharpaWave Deform Encoder. https://huggingface.co/Sharpa-Robotics/sharpawave-deform-encoder, 2026.

[21] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. arXiv preprint arXiv:2508.10104, 2025. URL https://arxiv.org/abs/2508.10104.

[22] Congsheng Xu, Qiaochu Yang, Fangyuan Shi, Yifan Han, Baijun Chen, Yiming Wang, Haonan Zhao, Zhe Liu, Yao Mu, Daolin Ma, Xiaokang Yang, and Hesheng Wang. VT-MUSE: Multimodal unified sequential visuotactile representation learning for manipulation. arXiv preprint arXiv:2608.21290, 2026. URL https://arxiv.org/abs/2608.21290.

[23] Zijun Xu, Wendi Chen, Xinkai Wang, Jiachen Zou, Hongjin Chen, Hairuo Liu, Bokai Lin, Shihao Ma, Ke Ma, Zhiwei Zhang, Zhuohao Yan, Ke Wu, Haoming Song, Xueqi Li, Chuan Wen, Jieru Zhao, Wenchao Ding, Cewu Lu, and Lixin Yang. ForceWAM: Decoupling future and response learning from demo and play data for contact-rich manipulation. arXiv preprint arXiv:2610.18852, 2026. URL https://arxiv.org/abs/2610.18852.

[24] Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. arXiv preprint arXiv:2503.02881, 2025. URL https://arxiv.org/abs/2503.02881.

[25] Chen Yang, Linzhe Shi, Changjie Wu, Hang Zhang, Ronghan Chen, Lingjun Zhang, Xu Hu, Mu Xu, Jiansheng Fan, and Chen Wang. SimpleTouch: Can vision-language-action models master contact-rich manipulation without tactile policy pretraining? arXiv preprint arXiv:2610.02784, 2026. URL https://arxiv.org/abs/2610.02784.

[26] Guo Ye, Zexi Zhang, Xu Zhao, Shang Wu, Haoran Lu, Shihan Lu, and Han Liu. Learning to feel the future: DreamTacVLA for contact-rich manipulation. arXiv preprint arXiv:2512.23864, 2025. URL https://arxiv.org/abs/2512.23864.

[27] Chengbo Yuan, Zicheng Zhang, Mingjie Zhou, Wendi Chen, Yi Wang, Zhuoyang Liu, Dantong Niu, Shuo Wang, Hui Zhang, Wenkang Zhang, Yingdong Hu, Yuanqing Gong, Wanli Xing, Chuan Wen, Cewu Lu, Kaifeng Zhang, and Yang Gao. FTP-1: A generalist foundation tactile policy across tactile sensors for contact-rich manipulation. arXiv preprint arXiv:2606.13102, 2026. URL https://arxiv.org/abs/2606.13102.

[28] Haoran Yuan, Zekai Wang, Boning Shao, Haoran Lu, Trevor Darrell, Ismini Lourentzou, and Wei Zhan. DexTacWAM: A visuo-tactile world-action model for dexterous manipulation. arXiv preprint arXiv:2609.24976, 2026. URL https://arxiv.org/abs/2609.24976.

[29] Haoran Yuan, Weigang Yi, Zhenyu Zhang, Wendi Chen, Yuchen Mo, Jiashi Yin, Xinzhuo Li, Yue Luo, Jialin Li, Xiangyu Zeng, Chuan Wen, Cewu Lu, Katherine Driggs-Campbell, and Ismini Lourentzou. VTAM: Efficient visuo-tactile world-action models for contact-rich manipulation without tactile pretraining. arXiv preprint arXiv:2603.23481, 2026. URL https://arxiv.org/abs/2603.23481.

[30] Pengwei Zhang, Bin Xie, Xinpan Meng, Xinyu Guo, Ce Hao, Fang Deng, Long Cheng, and Tiancai Wang. Feeling

the unexpected: ResTacVLA for contact-rich manipulation via residual tactile representation. arXiv preprint arXiv:2607.03387, 2026. URL https://arxiv.org/abs/2607.03387.

[31] Xuancheng Zhang, Xuetao Liu, Qianying Tang, Jizhe Wang, Runsheng Wang, Zhijing Cheng, Bochen Lin, Haoran Wen, Ming Li, Kun Zhan, and Yu Liu. ME-Dex 1.0: Bringing heterogeneous tactile sensing into world action modeling. arXiv preprint arXiv:2609.21449, 2026. URL https://arxiv.org/abs/2609.21449.

[32] Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 3766–3779. PMLR, 2025. URL https://proceedings.mlr.press/v270/zhao25c.html.

[33] Yuhang Zheng, Songen Gu, Yupeng Zheng, Weize Li, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, Haoran Li, Yilun Chen, Shuicheng Yan, and Wenchao Ding. OmniVTA: Visuo-tactile world modeling for contact-rich robotic manipulation. arXiv preprint arXiv:2603.19201, 2026. URL https://arxiv.org/abs/2603.19201.

Appendix

A Method and Implementation Details

This section supplements Section 3 of the main paper with the temporal offsets, objective weights, and training hyperparameters for WiTac and downstream policy fine-tuning.

A.1 Contact Encoding and Pretraining (Main Sec. 3.1 & 3.2)

At each decision step t (30 Hz), the encoder takes an eight-frame tactile history $\mathcal{T}_t$ at frame offsets $(-14, -12, -10, -8, -6, -4, -2, 0)$ across all ten fingers, together with the 65D robot joint state s_t, wrist-relative fingertip poses p_t from forward kinematics, and dual-wrist fisheye RGB images I_t resized to 256×320 and encoded by a frozen DINOv3 ViT-B/16 [21]. The spatial tactile encoder is initialized from the SharpaWave Deform Encoder [20], followed by causal temporal convolutions, FiLM pose modulation, and two Transformer fusion blocks (8 heads, width 768) that output ten contact tokens $C_t \in \mathbb{R}^{10\times768}$. The wrist-free variant WiTac-NW uses the same tactile, state, and kinematic encoding architecture, omitting only the wrist cross-attention branch and masked visual recovery objective.

Table S1 lists the temporal offsets, prediction horizons, and loss weights λ_j for the five self-supervised pretraining objectives in Eq. (5) of the main paper. In deformation-change reconstruction (D_{def}), changing and static pixels are weighted equally in a balanced L1 loss so that static background regions do not dominate training. In masked visual recovery (D_{vis}), 50% of patch blocks within the hand–object wrist regions are masked before ViT encoding to form V_t^{vis}, and the decoder reconstructs the LayerNorm-normalized full-image DINOv3 features at masked positions. Pretraining hyperparameters on SharpaDex v1.0 [19] are summarized in table S2.

Table S1 Pretraining objectives, temporal offsets, and loss weights (Main Sec. 3.2). Frame offsets and horizons are defined at 30 Hz. All prediction heads decode the shared contact tokens C_t during pretraining and are discarded at deployment.

Objective	Target	Offsets / horizons	Loss & weight (λ_j)
Contact descriptor (D_{desc})	Contact area, intensity, 2D center	Current frame (t)	Smooth L1 (0.05)
Deformation change (D_{def})	$\Delta T_{t,k} = T_t - T_{t-k}$ (64×64)	$k \in \{4, 8, 14\}$ past frames	Balanced L1 (0.10)
Fingertip motion (D_{mot})	Wrist-frame $\Delta x_{t,k} = x_t - x_{t-k}$	$k \in \{2, 6, 10, 14\}$ past frames	Smooth L1 (0.05)
Visual recovery (D_{vis})	Masked full-image DINOv3 features	Current frame (t)	LayerNorm MSE (0.10)
Wrench response (D_{wr})	6D wrench change $\Delta w_{t,h} = w_{t+h} - w_t$	$h \in \{4, 8, 16\}$ future frames	Smooth L1 (0.10)

A.2 Policy Injection and Fine-Tuning (Main Sec. 3.3)

During downstream policy learning, the pretrained contact encoder (WiTac or WiTac-NW) is frozen and injected into alternating action-expert blocks of GR00T N1.7 [17] and $\pi_{0.5}$ [18] via gated cross-attention ($\lambda_{\text{gate}} = 0.02$). Both policies receive head stereo RGB (512×640), dual-wrist fisheye RGB (384×480 for the native policy visual stream; the same dual-wrist fisheye feeds are resized to 256×320 in native orientation for WiTac), language instructions, and 65D robot joint states, and predict relative joint-angle action chunks. Binary contact-gate labels $y_t \in \{0, 1\}$ are derived from the active-pixel fraction on the primary manipulation fingers (thumb, index, and middle of both hands). Table S3 summarizes the downstream fine-tuning hyperparameters; on $\pi_{0.5}$, the pretrained policy backbone is fine-tuned with learning rate 2.5×10^{-5} while the newly introduced contact cross-attention and gate modules use 10^{-4}.

Table S2 Representation pretraining hyperparameters (Main Sec. 3.2).

Setting	Value
Dataset	SharpaDex v1.0 [19] (59 tasks)
Train / val split	98% / 2% episodes
Training steps	30,000
Global batch size	1,024 (`bfloat16`)
Optimizer	AdamW (weight decay 10^{-3}, grad clip 1.0)
Learning rate	10^{-5} (spatial backbone), 10^{-4} (other modules)
LR schedule	2,000-step warmup, cosine decay to 0.1×

Table S3 Policy fine-tuning settings (Main Sec. 3.3 & 4.1). $\pi_{0.5}$ LR: policy backbone / contact interface.

Setting	GR00T N1.7 [17]	$\pi_{0.5}$ [18]
Action horizon	40 steps	25 steps
Demonstrations	200 per task	200 per task
Training schedule	30 epochs	30 epochs
Optimizer	AdamW	AdamW
Learning rate	10^{-4}	2.5×10^{-5} / 10^{-4}
Warmup & decay	2,000 steps, cosine	1,000 steps, cosine

B Additional Experimental Details and Results

This section supplements Section 4 of the main paper with task and baseline descriptions (Sec. 4.1), qualitative real-world rollouts (Sec. 4.2), complete numerical tables for the representation analyses (Sec. 4.3), and ablation setups (Sec. 4.4).

B.1 Tasks and Baselines (Main Sec. 4.1)

Task definitions. We evaluate four bimanual contact-rich manipulation tasks (none of which are included in the 59-task SharpaDex v1.0 pretraining dataset) over 20 real-world trials per method and task, with all policies fine-tuned for 30 epochs on 200 teleoperated demonstrations per task (recorded at 30 Hz):

- **Insert battery:** One hand stabilizes the battery holder while the other grasps a cylindrical battery, aligns it with the compartment slot under contact, and presses it down until fully seated.
- **Wipe bowl:** One hand stabilizes a bowl while the other grasps a sponge and maintains continuous interior contact to wipe the inner wall.
- **Reinsert USB:** One hand stabilizes a power strip while the other grasps and extracts a plugged connector, re-aligns the plug with a target socket, and presses it in until seated.
- **Open cabinet:** The robot grasps and turns a key in the cabinet lock to unlatch it, then pulls the articulated door open along its hinge arc.

External baseline adaptations. All external baselines are fine-tuned on the same 200 demonstrations per task for 30 epochs:

- **VT-MUSE joint repr. [22]:** We adapt the sequential visuotactile encoder of VT-MUSE to eight-frame histories of the ten tactile sensors and two wrist cameras, pretrain it on the same SharpaDex v1.0 split using its two-stage alignment and masked reconstruction objectives, and inject the frozen encoder into GR00T N1.7 via the same gated cross-attention interface as WiTac.
- **FTP-1 [27] and T-Rex [16]:** Fine-tuned with global batch size 256 from the pretrained checkpoint for FTP-1 and the mid-trained checkpoint for T-Rex, adapting their respective visual, tactile, and proprioceptive input pipelines to the Sharpa North observation and relative joint-action space.

B.2 Qualitative Real-World Rollouts (Main Sec. 4.2)

Figure S1 compares successful WiTac rollouts with failure rollouts from WiTac-NW on GR00T N1.7 across all four real-world tasks, complementing the closed-loop success rates and contact-stage behaviors reported in Section 4.2 of the main paper.

B.3 Complete Representation Analysis Results (Main Sec. 4.3)

Contact reconstruction. Table S4 reports the complete absolute errors and relative reductions on the 9,632 SharpaDex v1.0 validation windows corresponding to the contact-reconstruction analysis in Section 4.3, alongside the qualitative deformation-change reconstructions illustrated in Figure 4 of the main paper.

Joint task–contact readout. Table S5 provides the complete macro-F1 scores and sample counts corresponding to Figure 5 of the main paper. Short-window contact states are categorized from the eight-frame tactile history into *no contact*, *onset*, *release*, *stable contact*, and *changing contact*. The 12-class setting pairs the four tasks with {no, stable, changing} contact, while the 20-class setting uses all five contact states. Frozen representations are reduced to 50 PCA dimensions and evaluated with a class-balanced multinomial logistic regression probe.

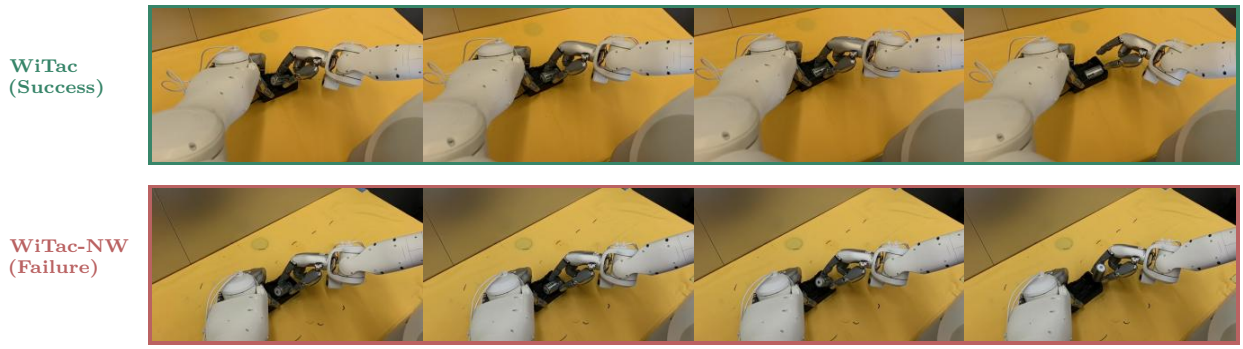
Table S4 WiTac improves contact-descriptor reconstruction. Rel. denotes relative error reduction.

Error ↓ (units)	WiTac-NW	WiTac	Rel. (%)
Area, all (%)	0.3948	**0.3649**	7.57
Area, contact (%)	0.7723	**0.7623**	1.30
Intensity, all (0–255)	0.3706	**0.3335**	10.00
Intensity, contact (0–255)	0.7406	**0.6997**	5.53
Center (pixels)	6.5204	**6.3481**	2.64

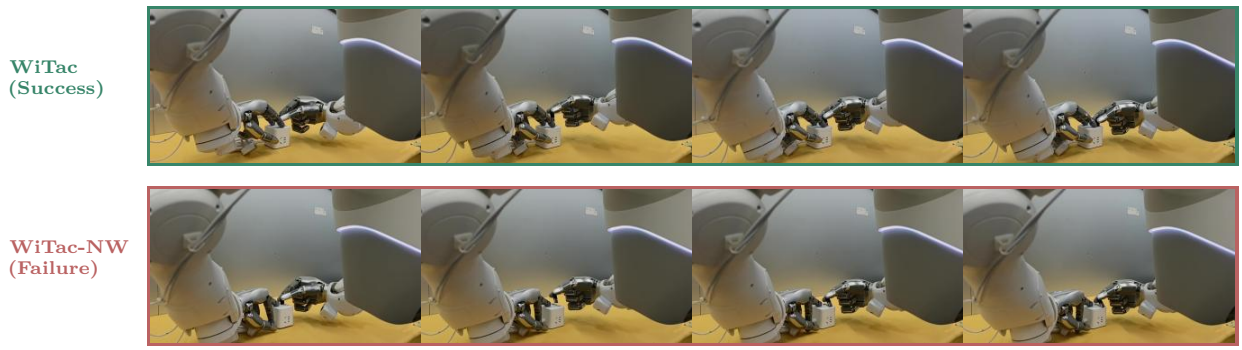
Table S5 Joint task–contact macro-F1 ↑. O/C: original/confirmation split; Concat.: WiTac-NW+DINO; N: windows.

Classes/split	N	DINO	WiTac-NW	Concat.	WiTac
12 / O	3,487	.7140	.7166	.7828	**.7931**
12 / C	3,233	.7322	.7189	**.7787**	.7761
20 / O	3,555	.4894	.5600	**.5792**	.5783
20 / C	3,288	.4935	.5578	.5463	**.5949**

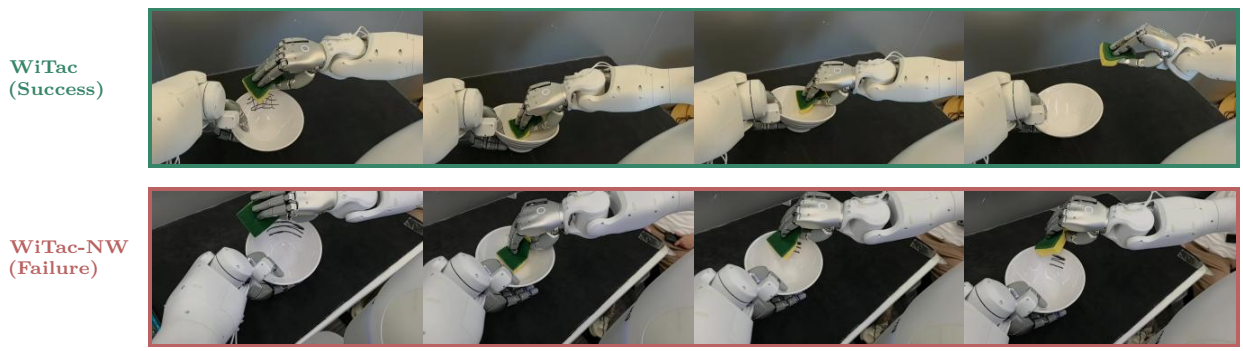
(a) Insert battery



(b) Reinsert USB



(c) Wipe bowl



(d) Open cabinet

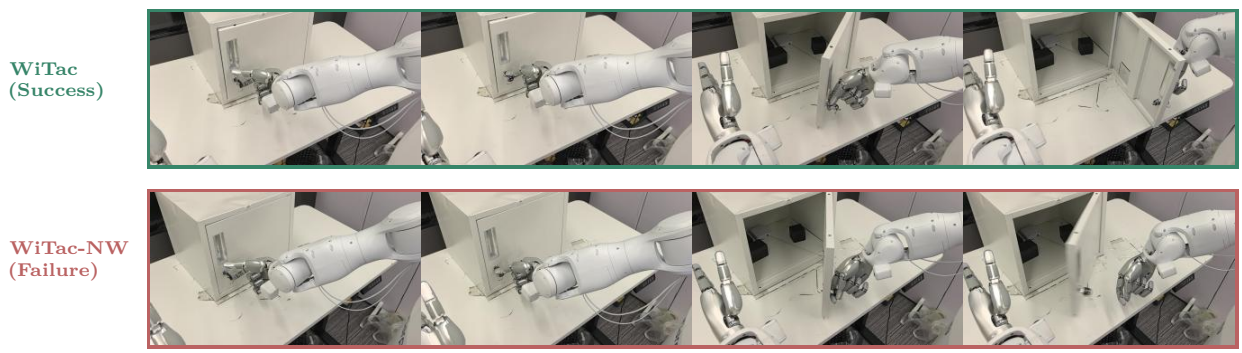


Figure S1 Qualitative real-world rollouts comparing WiTac and WiTac-NW across all four tasks. (a) Battery insertion, (b) USB reinsertion, (c) bowl wiping, and (d) cabinet opening. For each task, the top row shows a successful rollout with WiTac (green) and the bottom row shows a failure rollout with WiTac-NW (red) on GR00T N1.7, with four frames ordered temporally from left to right.

Synthetic tactile-noise perturbations. Tables S6 and S7 provide the full numerical metrics supplementing Table 3 of the main paper under Gaussian tactile noise ($\sigma = 3$ in 0–255 intensity units, with wrist images and robot state held fixed). Alongside feature drift $1 - \cos(z_{\text{clean}}, z_{\text{noisy}})$ and action MAE (averaged over the first 25 command steps), table S6 reports cross-attention Jensen–Shannon (JS) divergence and the relative ℓ_2 drift of the pre-gate contact-attention output r, defined as $\|r_{\text{noisy}} - r_{\text{clean}}\|_2 / (\|r_{\text{clean}}\|_2 + 10^{-8})$. WiTac reduces feature drift by 35.6–44.1% across all four tasks, as well as attention JS by 26.8–79.7% and contact-output drift by 15.7–28.8% across all eight policy–task pairs.

Table S6 Synthetic-noise stability at Gaussian $\sigma = 3$. Feature drift depends only on the frozen encoder and task; JS denotes cross-attention Jensen–Shannon divergence, and contact-output drift is relative ℓ_2 change before gating.

Policy / task	Feature drift ↓		Attention JS ↓		Contact-output drift ↓	
	WiTac-NW	WiTac	WiTac-NW	WiTac	WiTac-NW	WiTac
GR00T / Insert battery	0.5078	**0.3177**	0.1626	**0.0658**	1.4321	**1.1506**
GR00T / Wipe bowl	0.5413	**0.3484**	0.1437	**0.0878**	1.1974	**0.8520**
GR00T / Reinsert USB	0.4966	**0.3130**	0.1711	**0.0348**	1.6406	**1.2741**
GR00T / Open cabinet	0.5415	**0.3028**	0.2133	**0.1560**	1.4558	**1.1163**
$\pi_{0.5}$ / Insert battery	0.5078	**0.3177**	0.1023	**0.0353**	1.0884	**0.8215**
$\pi_{0.5}$ / Wipe bowl	0.5413	**0.3484**	0.0584	**0.0289**	1.3652	**1.1510**
$\pi_{0.5}$ / Reinsert USB	0.4966	**0.3130**	0.1059	**0.0384**	1.1275	**0.8561**
$\pi_{0.5}$ / Open cabinet	0.5415	**0.3028**	0.0789	**0.0248**	1.6630	**1.3620**

Table S7 Action error and noise-induced degradation at Gaussian $\sigma = 3$. ΔMAE is noisy minus clean MAE; positive reduction favors WiTac.

Policy / task	Clean MAE (°)		Noisy MAE (°)		ΔMAE (°)		Reduction in ΔMAE 95% interval (°)
	WiTac-NW	WiTac	WiTac-NW	WiTac	WiTac-NW	WiTac	
GR00T / Insert battery	**0.8785**	0.8931	1.4027	**1.0835**	0.5242	**0.1904**	[+0.2877, +0.3805]
GR00T / Wipe bowl	**0.7757**	0.7801	0.9709	**0.8996**	0.1952	**0.1195**	[+0.0393, +0.1198]
GR00T / Reinsert USB	0.7787	**0.7592**	0.9950	**0.9037**	0.2163	**0.1445**	[+0.0348, +0.1067]
GR00T / Open cabinet	**0.6370**	0.6425	1.1156	**0.8748**	0.4786	**0.2322**	[+0.1232, +0.3845]
$\pi_{0.5}$ / Insert battery	0.8263	**0.8250**	**0.8734**	0.8783	**0.0470**	0.0533	[-0.0157, +0.0035]
$\pi_{0.5}$ / Wipe bowl	**0.7212**	0.7277	**0.7359**	0.7499	**0.0146**	0.0222	[-0.0127, -0.0024]
$\pi_{0.5}$ / Reinsert USB	**0.6094**	0.6164	0.6735	**0.6606**	0.0641	**0.0442**	[-0.00005, +0.04273]
$\pi_{0.5}$ / Open cabinet	0.5428	**0.5352**	0.5500	**0.5380**	0.0072	**0.0027**	[-0.0020, +0.0104]

Action-direction readout. In Table 4 and Figure 6 of the main paper, demonstrated adjustment targets Δq_t are defined over a 16-step future command window (the difference between the mean of the last four and first four joint commands). Frozen DINO, WiTac-NW, and WiTac features are standardized on training episodes, projected to 16 whitened PCA components, and mapped to $\widehat{\Delta q}_t$ via ridge-regression linear probes with regularization selected by three-fold cross-validation over training episodes. Directional alignment is evaluated by $\cos(\widehat{\Delta q}_t, \Delta q_t) = \widehat{\Delta q}_t^{\top} \Delta q_t / (\|\widehat{\Delta q}_t\|_2 \|\Delta q_t\|_2)$, averaged over the two contact-matched observations in each held-out test pair and across the 25 test pairs per task.

B.4 Ablation Setup (Main Sec. 4.4)

Section 4.4 (Table 5 of the main paper) compares full pretrained WiTac against three design variants on GR00T N1.7 under the same 30-epoch training schedule and 200 demonstrations per task:

- **WiTac w/o pretraining:** The SharpaWave Deform Encoder remains frozen while the remaining WiTac modules are trained end-to-end with the GR00T policy and gated contact interface during task fine-tuning, without self-supervised pretraining on SharpaDex v1.0.
- **WiTac w/o visual recovery:** Pretrained on the same SharpaDex v1.0 split with dual-wrist cross-attention and all tactile, motion, and wrench objectives ($D_{\text{desc}}, D_{\text{def}}, D_{\text{mot}}, D_{\text{wr}}$), omitting only the masked visual recovery loss (D_{vis}) before freezing the encoder for GR00T fine-tuning.
- **Separate injection:** The frozen WiTac-NW encoder and frozen DINOv3 wrist patch tokens are injected into the GR00T action expert through two parallel cross-attention branches instead of fusing wrist vision into contact tokens prior to policy injection.