WT-UMI: Tactile-based Whole-Body Manipulation via Force-Supervised Contact-Aware Planning

Anonymous Author(s)
Affiliation
Address
email

Abstract: Whole-body humanoid manipulation of bulky, deformable, and shared-load objects requires distributed contact sensing and explicit force regulation, yet most imitation policies treat contact force only implicitly. On the other hand, different demonstration sources provide complementary modalities with inherent trade-offs: human demonstrations capture natural contact forces but not robot-executable actions, while teleoperation directly records robot actions but with less natural force regulation. This paper presents **WT-UMI**, a wearable whole-body tactile interface worn by human operators or mounted on humanoids, providing accurate observations of tactile images, contact forces, and end-effector poses across both human demonstration and humanoid teleoperation modes. We introduce a force-conditioned target-pose correction module that converts measured human poses into contact-aware robot targets by learning corrections from teleoperation data. To leverage the natural force interaction in human data, we propose a force-supervised planner that predicts end-effector pose chunks and contact-force trajectories. The predicted contact force serves as the reference for a tactile-based admittance controller. Across five contact-rich tasks spanning deformable objects, bulky rigid objects, and human–humanoid collaboration, WT-UMI improves success rate and reduces contact-position tracking error over four policy baselines. Our project page is available at https://wt-umi.github.io/WTUMI/.

Keywords: Humanoid Whole-body Manipulation, Tactile Sensing, Robot Learning, Force-aware Planning

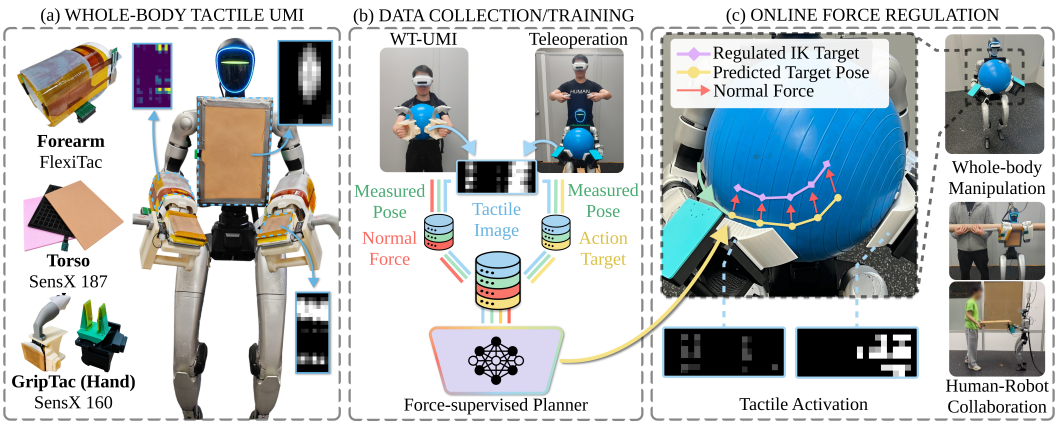

Figure 1: (a) **WT-UMI** is a shared interface between human demonstrators and humanoid robots for whole-body tactile data collection. (b) A human demonstrator wears WT-UMI, or a humanoid is controlled via teleoperation using the same hardware. (c) A force-supervised planner trained from WT-UMI data executes contact-rich tactile-aware tasks, spanning whole-body manipulation of deformable and large rigid objects and human–humanoid collaborative transport.

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

1 Introduction

Humanoid robots are increasingly expected to manipulate bulky, deformable, and shared-load objects in human environments. Tasks such as carrying a large box, reorienting a soft pillow, or transporting a beam with a human partner cannot rely on grasping alone; instead, they require coordinated contact across the torso, forearms, and hands to distribute interaction forces [1]. Small errors in contact location or force allocation, however, can lead to slips, collisions, or load loss. Reliable whole-body manipulation therefore requires distributed contact sensing and joint motion-force planning that actively regulates both contact location and interaction force.

Most learning-based manipulation systems rely on vision and proprioception [2]. However, vision is often occluded during contact and does not directly measure interaction forces, while proprioception cannot localize body-surface contact. Prior tactile sensing has focused primarily on fingertip arrays [3, 4, 5, 6, 7, 8], which do not capture the distributed contact central to whole-body interaction. Existing body-mounted tactile systems [9, 10] often rely on model-based controllers or per-task reward shaping, limiting their ability to learn diverse skills. Turning distributed whole-body tactile sensing into a contact force plan for humanoid whole-body manipulation remains an open problem.

We address this problem with **WT-UMI** (**W**hole-Body **T**actile **U**niversal **M**anipulation **I**nterface), a wearable tactile interface paired with a **force-aware** learning framework for humanoid whole-body manipulation (Fig. 1). WT-UMI enables scalable human demonstration collection while reducing the human–humanoid embodiment mismatch. The human demonstrations collected through WT-UMI capture whole-body, contact-rich interactions with calibrated contact forces, enabling explicit contact-force prediction during training and force regulation during deployment. However, human demonstrations often lack robot-executable action labels. To convert these force-aware human demonstrations into robot-executable actions, we introduce a **force-conditioned target-pose correction** module that learns target-pose corrections from robot-in-the-loop teleoperation. Human and teleoperation trajectories are paired using force-inferred contact modes, with teleoperation commands supervising the correction of measured human poses. The resulting corrected target poses serve as action label to supervise planner training. The action-labeled human data preserves accurate contact-force measurements, enabling a **force-supervised planner** to jointly predict contact forces and corrected actions (Fig. 2). During deployment, a tactile-based admittance controller then tracks the predicted force trajectory by modulating the predicted end-effector poses, achieving stable and force-regulated contact.

Overall, our framework offers four key contributions. (i) We introduce **WT-UMI**, a wearable whole-body tactile interface that collects tactile images and force-supervised demonstrations. The same sensing hardware supports both human demonstrations and humanoid teleoperation. (ii) We design a **force-conditioned target-pose correction** module that converts human trajectories into contact-aware robot actions by learning pose corrections from teleoperation data. (iii) We propose a **force-supervised planner** whose cross-attention head predicts the normal contact-force trajectory, with the predicted force serving as the reference for a tactile-based admittance controller. (iv) We validate the framework on five contact-rich whole-body tasks spanning deformable objects, bulky rigid objects, and human–humanoid collaboration, where it improves the success rate and reduces contact-position drift over four policy baselines.

2 Related Work

Whole-Body Tactile Sensing and Manipulation. Whole-body tactile sensing is an emerging modality for contact-rich manipulation, yet its adoption remains challenging [1]. Much prior work restricts contact to the hands, leaving the torso and forearms largely uncovered [11]. Compliant-body designs such as Punyo [10] extend coverage with pressure-sensitive skins on the arms and chest, while discrete-cell skins such as HEX-o-SKIN [9] cover larger areas but require per-cell kinematic calibration. In contrast, our thin-film piezoresistive arrays produce dense 2D contact images with large-area coverage that integrate naturally with vision encoders and imitation-learning pipelines.

Existing whole-body manipulation methods span model-based control [12, 13], planning [14], reinforcement learning [15, 16, 17, 18, 19, 20, 21], and imitation learning [2, 22]. However, these methods typically rely on accurate contact models, object-state estimation, and task-specific reward design, rather than learning an explicit contact-force plan for whole-body humanoid manipulation.

Joint Motion-Force Prediction. Model-based hybrid motion-force control requires explicit force sensing and careful contact modeling [12, 23, 24]. Learning-based approaches such as ForceMimic and UMI-FT learn motion-force representations from force-motion capture or wrist/finger force sensing, but their measurements remain local [25, 26]. Recent tactile- and force-aware policies further show the value of physical grounding for contact-rich manipulation, through tactile-conditioned force actions, tactile-force representation learning, force-feedback fusion, or force-token distillation [7, 27, 28, 29, 30]. However, these methods primarily focus on local fingertip, gripper, and wrist contacts for tabletop tasks. Humanoid Touch Dreaming [3] predicts future tactile signals in a learned latent space, but the predictions act as auxiliary regularization and contact is modeled only implicitly. Our method instead (i) predicts contact force explicitly from distributed tactile arrays spanning the robot body, and (ii) uses the predicted force as a reference for a tactile-based admittance controller, yielding explicit force-aware control during whole-body manipulation.

Demonstration Interfaces for Behavior Cloning. Teleoperation provides direct robot action labels, with VR, retargeting, and recent portable systems improving ergonomics and coverage [31, 32, 33, 34, 35]. However, it still requires robot access, task setup, and skilled operators. In contrast, UMI-style interfaces bypass robot-in-the-loop collection by recording in-the-wild human demonstrations with handheld devices and transferring them to robot policies [36]. Existing interfaces mainly provide kinematic supervision and visual observations, without distributed whole-body contact force for explicit force planning. When touch is included, human-to-robot transfer becomes more challenging since human and robot embodiments, as well as their tactile sensors, may differ. TactAlign tackles this issue by aligning human and robot tactile observations in a shared latent space using rectified flow, without paired data or manual labels [37]. WT-UMI takes a complementary route: it uses shared wearable/robot-mounted tactile hardware to reduce sensing mismatch, while combining human force-rich demonstrations with robot teleoperation labels for whole-body humanoid behavior cloning.

3 Methods

Our system architecture is outlined in Fig. 2. In *data collection*, WT-UMI records distributed tactile readings and calibrated contact force from a human operator or via teleoperation on a humanoid robot. In *pre-trained target-pose correction*, a force-conditioned model learns to correct human target poses into robot-executable actions. In *planner training*, an action denoiser predicts corrected bimanual pose chunks, while a force head predicts contact-force trajectories. During *force-aware deployment*, the predicted forces drive a tactile-based admittance controller that regulates target poses. A pre-trained RL policy simultaneously controls lower-body locomotion.

3.1 Whole-Body Tactile Universal Manipulation Interface (WT-UMI) for Data Collection

We develop WT-UMI, an extensible wearable system for capturing operator motion and distributed tactile feedback. As shown in Fig. 3, it integrates tactile sensors on hardware modules shared between humans and humanoids: a chest plate, forearm covers, and handheld GripTacs, which are tactile-instrumented handheld interfaces with interchangeable end-effectors. By using the same sensing modules for human data collection, robot teleoperation, and robot execution, WT-UMI reduces the domain gap between demonstration and deployment (see Fig. 1).

WT-UMI supports demonstration collection in two modes (Fig. 2). In *teleoperation mode*, WT-UMI is mounted on the robot, and an operator streams bimanual pose commands through a PICO VR headset [38] and two handheld controllers via XRoboToolkit [39]. In *human mode*, a human operator

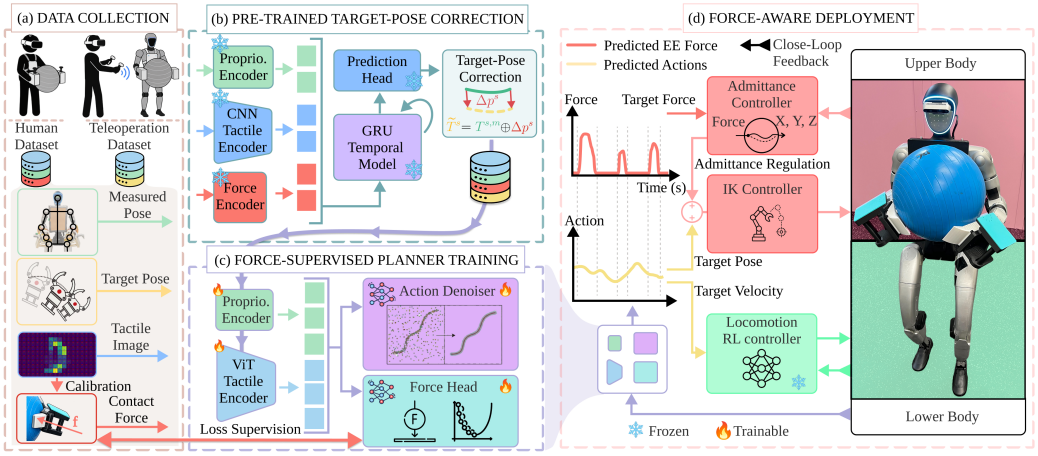

Figure 2: A force-conditioned target-pose correction module creates action labels for human data. A force-supervised planner produces a contact-force trajectory in addition to end-effector poses. The predicted forces are used for online force regulation via a tactile-based admittance controller. (gu: need more improvement)

wears the chest plate, forearm covers, and GripTacs, with a PICO controller on each GripTac to track the bimanual pose. Human mode requires no robot in the loop. Both modes stream hand poses, tactile images, and calibrated force measurements; teleoperation additionally records robot proprioception and VR commands. More details about sensor specifications and force calibration are provided in Appendix 8.1.

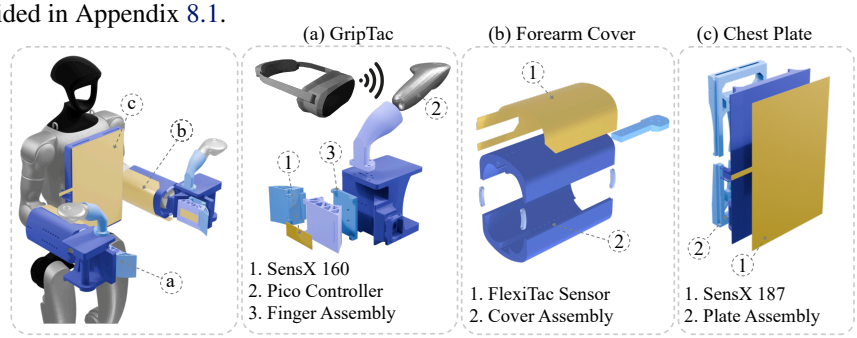

Figure 3: **WT-UMI** includes GripTac end-effectors, forearm covers, and a chest plate, each equipped with a thin-film tactile sensor.

3.2 Force-Conditioned Target-Pose Correction

We introduce a force-conditioned target-pose correction module that converts the measured human hand pose $T^{s,\mathrm{m}}$, contact force $f^{s,\mathrm{m}}$, and tactile observation $\mathbf{I}^{s,\mathrm{m}}$ into a robot target pose, where $s \in \{l, r\}$ indexes the left and right hands and m denotes measured quantities. This module is pre-trained first and then applied to human demonstrations to generate robot-executable action labels to supervise planner training. The correction module is pre-trained offline using paired robot teleoperation and human demonstration trajectories. For both data sources, we use contact forces to infer contact modes, such as right-contact, both-contact, and left-contact. We then match contact modes between teleoperation and human data using Dynamic Time Warping (DTW) over the end-effector motion. Given the aligned contact modes, the target-pose offset from the robot teleoperation data is used to supervise the corresponding offset in the human data, gated by inferred contact state: only hands that are in contact for that segment receive a nonzero target-pose offset. For each hand s, a lightweight CNN-GRU correction network g_θ predicts the hand-specific translation offset from the measured tactile image, pose, and force, $\Delta\mathbf{p}^s = g_\theta(\mathbf{I}^{s,\mathrm{m}}, T^{s,\mathrm{m}}, f^{s,\mathrm{m}}) \in \mathbb{R}^3$, over sampled aligned teleoperation–human trajectory pairs. After pre-training, the learned correction module is applied to human demonstration data to generate target-pose action labels. The corrected target pose in SE(3) is $\tilde{T}^s = T^{s,\mathrm{m}} \oplus \mathbf{1}[f^{s,\mathrm{m}} \geq f_{\text{threshold}}]\Delta\mathbf{p}^s$, where $\oplus$ shifts the translation component. The resulting corrected poses $\tilde{T}^s$ serve as robot-executable action labels for training an action denoiser.

The model is trained with a Smooth-L_1 offset loss and a temporal smoothness penalty; additional training details are provided in Appendix 8.3.

3.3 Force-Supervised Planner

Existing generative planners often do not explicitly model contact forces, leaving force regulation to position-based feedback alone. WT-UMI addresses this limitation by providing calibrated normal force from tactile measurements. We leverage these calibrated forces in the force-conditioned target-pose correction module and use them as supervision to augment the planner with a force head that predicts future contact-force trajectories. During deployment, the predicted forces serve as explicit references for a tactile-based admittance controller.

The planner operates at 50 Hz with 400 ms observation and prediction horizons ($H_o = H_a = 20$). At timestep t, the model consumes an observation history $\mathbf{o}_{t-H_o+1:t}$, where each $\mathbf{o}_t = (\mathbf{I}_t^l,\ \mathbf{I}_t^r,\ \mathbf{I}_t^{ch},\ \mathbf{T}_t)$ consists of normalized tactile images from the left/right contact sites and the chest, alongside the measured bimanual end-effector pose $\mathbf{T}_t = [\mathrm{vec}(T_t^l);\ \mathrm{vec}(T_t^r)] \in \mathbb{R}^{18}$, where $\mathrm{vec}(\cdot)$ maps each SE(3) pose $T_t^s \in \mathbb{R}^{4\times4}$, $s \in \{l, r\}$, to its 3D translation and 6D continuous rotation representation [40] in $\mathbb{R}^9$. A vision transformer (ViT) [41] encodes the channel-stacked tactile images and a two-layer MLP encodes the hand poses; per timestep, their outputs are concatenated into one observation token of width $D = 320$. The hand pose is additionally linearly projected to a separate pose token of the same width, yielding a conditioning sequence $\mathbf{h} \in \mathbb{R}^{2H_o\times D}$. Conditioned on $\mathbf{h}$, a transformer-based action denoising policy (either flow matching [42] or diffusion [2]) generates a predicted action chunk $\mathbf{a} = [\mathbf{T}_{t+1}, \ldots, \mathbf{T}_{t+H_a}] \in \mathbb{R}^{H_a\times18}$. The action denoising policy is supervised by the target-pose-corrected labels $\mathbf{a}^\star = [\tilde{\mathbf{T}}_{t+1}, \ldots, \tilde{\mathbf{T}}_{t+H_a}] \in \mathbb{R}^{H_a\times18}$ from Sec. 3.2.

Unlike the action denoising policy, a force head performs direct regression without a diffusion process. It consists of a cross-attention decoder that takes H_a learnable positional queries $\mathbf{Q} \in \mathbb{R}^{H_a\times D}$, whose t-th row $\mathbf{q}_t \in \mathbb{R}^D$ is a positional query embedding for predicting the force at the t-th future action step, where $D = 320$ matches the encoder token width, and uses the shared observation embeddings $\mathbf{h}$ as keys and values. By performing self-attention across the queries followed by cross-attention to $\mathbf{h}$, the decoder models the temporal force dynamics before a linear projection maps the output to the predicted force chunk of both hands $\mathbf{F} \in \mathbb{R}^{H_a\times2}$. Supervision is provided by the calibrated ground-truth forces $\mathbf{F}^\star \in \mathbb{R}^{H_a\times2}$, with a softplus projection enforcing physical non-negativity. Because the force head directly outputs a force trajectory, its gradients flow back through $\mathbf{h}$, allowing the shared encoder to be jointly optimized by both the action and force objectives. To suppress gradient spikes from contact onset, the force head is supervised via an element-wise Huber loss ($\mathrm{SmoothL1}$), yielding the combined objective: $\mathcal{L} = \mathcal{L}_{\text{gen}}(\mathbf{a}, \mathbf{a}^\star) + \lambda_F\,\mathrm{SmoothL1}(\mathbf{F}, \mathbf{F}^\star)$, where $\mathcal{L}_{\text{gen}}$ is a flow-matching or diffusion denoising MSE and λ_F is the weight balancing the action denoiser and force head losses. More training details can be found in Appendix 8.4.

3.4 Tactile-based Admittance Controller

The low-level controller is split into a lower body (12 leg joints and 3 waist joints) and an upper body (14 arm joints). The lower body tracks pelvis-frame velocity $\mathbf{v} = [v_x, v_y, \omega_z]^\top$ with a pre-trained RL locomotion policy [43] that maintains balance under arbitrary upper-body motion. The upper body is regulated by a tactile-based admittance controller. At each control timestep t, we track the predicted end-effector pose target $\tilde{T}_t^s$ and a normal-force reference f_t^s for hand s. In addition to the learned target-pose correction, a proportional admittance controller regulates the motion and force simultaneously. This controller refines the planner-generated pose using both normal-force and contact-centroid feedback: $T_{\text{cmd},t}^s = \tilde{T}_t^s\,\Delta T_t^s$, where $T_{\text{cmd},t}^s$ is the commanded target pose. The admittance regulation $\Delta T_t^s \in \mathbb{R}^{4\times4}$ is a corrective SE(3) increment for palm s expressed in the palm frame, incorporating the corrective rotation $\Delta R_{xy,t}^s$ and translation $\Delta\mathbf{p}_{z,t}^s$ derived from proportional admittance law based on force-measurement feedback. The commanded poses are then passed to an optimization-based inverse kinematics solver [44], $q_{\text{cmd},t} = \mathrm{IK}(T_{\text{cmd},t}^l, T_{\text{cmd},t}^r, q_t, \dot{q}_t)$. Joint-level tracking uses a PD controller with gravity compensation, $\tau_t = K_p(q_{\text{cmd},t} - q_t) + K_d(-\dot{q}_t) +$

$G(q_t)$, where $q_t, \dot{q}_t$ are measured joint positions and velocities, $q_{\text{cmd},t}$ is the IK solution, τ_t is the commanded joint torque, K_p, K_d are joint-PD gains, and $G(q_t)$ is the gravity-compensation torque.

4 Experiment Setup

Tasks. We demonstrate five contact-rich tasks in three categories, with representative deployments shown in Fig. 4. For **deformable object manipulation**, *yoga ball manipulation* (**T1**) stabilizes and repositions an inflated ball, and *pillow reorientation* (**T2**) reorients a soft pillow. For **bulky rigid object manipulation**, *bucket manipulation* (**T3**) repositions a cone-shaped container with diverse loads. For **human–humanoid collaborative manipulation**, *beam transport* (**T4**) and *table transport* (**T5**) require the robot and a human partner to jointly carry a beam and a table, respectively, while inferring human physical intent from tactile feedback and following the partner's motion.

Policy Baselines. We train four widely used behavior-cloning policies as baselines: *ViT-FMT* [42], *ViT-DiT* [2], $\pi_{0.5}$ [45], and Ψ_0 [46]. *ViT-FMT* and *ViT-DiT* share the same vision transformer encoder but differ in the generative process: flow matching for *ViT-FMT* and denoising diffusion for *ViT-DiT*. For ViT-FMT and ViT-DiT, the ViT tactile encoder is trained end-to-end with both heads. $\pi_{0.5}$ and Ψ_0 are fine-tuned foundation policies: $\pi_{0.5}$ is a vision-language-action model conditioned on tactile images and a fixed per-task language instruction; we adopt the OpenPI-based implementation in FASTER [47]. Ψ_0 is a humanoid foundation policy conditioned on tactile images and proprioception. For both foundation policies, encoders are frozen during fine-tuning.

Hardware Setup. We use a Unitree G1 humanoid for teleoperation data collection and policy deployment. Raw tactile data streams at 100 Hz and proprioception streams at 500 Hz; both are resampled to a synchronized 50 Hz stream during training and deployment. The ViT-FMT and ViT-DiT policies run on an RTX 4500 and perform asynchronous inference with a 100 ms chunk latency. $\pi_{0.5}$ and Ψ_0 similarly run at 12 Hz with 83 ms per chunk on an RTX 6000. The low-level tactile-based admittance controller runs at 200 Hz. For ViT-DiT, we use DDIM [48] for faster inference. All policies adopt training-time real-time chunking (RTC) [49] to compensate for inference delay.

5 Results

We evaluate WT-UMI across five tasks introduced in Sec. 4. Representative deployments for T1–T3 are shown in Fig. 4 and used for quantitative evaluation throughout Sec. 5. For the collaborative tasks T4 and T5, we detail their setup and successful deployments in Appendix 8.5.

5.1 Force Head Evaluation

This section investigates whether the force head produces accurate, smooth, and temporally aligned force predictions. To address this, we evaluate our force head (in Sec. 3.3) within a *ViT-FMT* policy. We employ a 90/10 train-validation split across both teleoperation and human datasets derived from task T1. We evaluate the trained force-supervised planner on 10 held-out demonstrations and compare predicted forces with ground-truth forces.

Table 1 shows the force prediction accuracy using RMSE, smoothness using force rate $|\mathrm{d}F/\mathrm{d}t|$ RMS averaged across episodes, and temporal alignment using the cross-correlation peak offset between predicted and ground-truth forces. Overall, the force head tracks the ground-truth force trajectory with low prediction error and small temporal lag. Notably, the force head trained on human demonstrations achieves an RMSE of 1.05 N out of a 5.56 N force RMS and a 2.2× lower temporal

Source	Force RMSE (N)	Lag (ms)	Force Rate RMS (N/s)	
			Meas.	Pred.
Human	**1.05**	**68**	**5.86**	**3.74**
Teleop	2.07	151	30.62	19.80

Table 1: Force signal quality. Human demonstrations exhibit more accurate and smoother force profiles than teleoperation, improving the quality of supervision for learning.

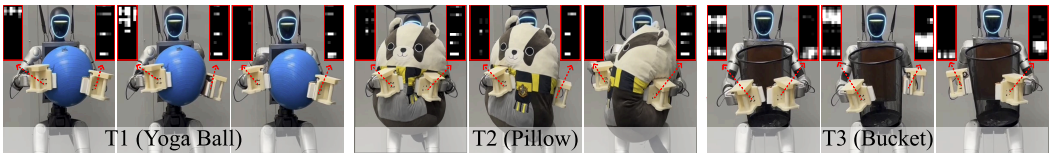

Figure 4: Deployment of our framework on three whole-body manipulation tasks.

lag. This improved temporal consistency is further reflected in the lower force-rate RMS, which decreases to 3.74 N/s compared to the 19.80 N/s from the force head trained on teleoperation data.

5.2 Effect of Training Data Sources and Force-Conditioned Target-Pose Correction

This section evaluates the complementary roles of teleoperated robot data and human demonstrations collected with WT-UMI. We evaluate ViT-FMT because it is the best-performing policy among all baselines, as demonstrated in Sec. 5.4. Among the representative tasks **T1**–**T3**, T2 involves a soft pillow whose compliance keeps contact-force readings small in magnitude, and T3's bucket requires precise contact on its cone-shaped surfaces, making both tasks challenging. We adopt two data-source settings: pure teleoperation data (**Tel.**) or a combined dataset (**Comb.**) that augments teleoperation with target-pose-corrected human demonstrations. Across T1–T3, **Tel.** contains 2.2, 2.2, and 1.4 min, respectively, while **Comb.** contains 13.2, 15.7, and 8.8 min. Thus, teleoperation accounts for only 16.7%, 14.0%, and 15.9% of the corresponding **Comb.** demonstrations. Each setting is evaluated with $N = 25$ trials per task. In each trial, the object is rotated once, starting and ending in the same hugging pose. We report the following metrics: (i) success rate; (ii) contact off-center drift, defined as the distance between the measured contact centroid and the geometric center of the tactile sensor, which indicates how well the contact position remains centered; (iii) mean contact force, which measures contact firmness during required contact phases; and (iv) translational and rotational accelerations of end-effector poses, which assess motion smoothness.

Table 2: Effect of data source over tasks **T1**-**T3** (Sec. 4). Best values are highlighted in bold.

Task	**Succ. (%)**		**Cont. Drift (mm)**		**Cont. Force (N)**		**Smooth.-Trans. (m/s^2)**		**Smooth.-Rot. (rad/s^2)**	
Data $\rightarrow$	Tel.	Comb.	Tel.	Comb.	Tel.	Comb.	Tel.	Comb.	Tel.	Comb.
T1	**100.0**	**100.0**	20.41	**18.12**	3.27	**4.77**	4.08	**3.14**	22.93	**20.29**
T2	**100.0**	**100.0**	24.61	**21.04**	0.17	**0.52**	6.03	**1.87**	34.54	**12.63**
T3	60.0	**80.0**	26.01	**25.00**	0.66	**0.96**	3.49	**1.85**	27.82	**14.05**

Table 2 presents the quantitative results. The **Tel.** policy alone achieves a strong 86.7% success rate on average, because teleoperation provides robot-feasible action labels. However, teleoperation lacks accurate force feedback during data collection, reflected in its occasional missed contacts or over-pressing behaviors, which result in larger contact-region drift of about 11% on average compared to **Comb.**. Conversely, human demonstrations offer valuable contact information because the demonstrator directly perceives interaction forces and naturally regulates contact. However, pure human data is not directly robot-executable due to missing action labels and the human–humanoid kinematics gap, which leads to poor policy performance when used alone. Our proposed human-data correction addresses this limitation by converting human data into robot-feasible pose targets. In the **Comb.** policy, the corrected human data supplements teleoperation data by reducing contact center drift and increasing mean contact force by approximately 2 times on average. Motion smoothness also improves consistently across all tasks with the **Comb.** policy: translational and rotational accelerations are reduced by 46.3% and 41.5% on average, respectively. The **Comb.** policy also increases the success rate by 20% for T3, where the most common failure mode is loss of contact and motion freeze due to out-of-distribution observations. Notably, the target-pose correction module requires only a small amount of teleoperation data to convert a much larger set of human demonstrations into robot-executable data, demonstrating the teleoperation-data efficiency of the proposed module.

Table 3: Admittance ablation across policy backbones over tasks **T1**, **T2**, and **T3**. Each metric is evaluated without (w/o) and with (w/) our admittance control. Best or tied values are bolded.

Policy	**Task**	**Cont. Drift (mm)**		**Cont. Force (N)**		**Smooth.-Trans. (m/s^2)**		**Smooth.-Rot. (rad/s^2)**	
Admi. (Ours) $\rightarrow$		w/o	w/	w/o	w/	w/o	w/	w/o	w/
ViT-FMT	T1	18.12	**15.67**	4.77	**5.50**	3.14	**2.98**	20.29	**18.56**
	T2	21.04	**19.44**	**0.52**	0.13	**1.87**	1.95	**12.63**	13.61
	T3	25.00	**22.08**	0.96	**1.61**	1.85	**1.29**	14.05	**10.38**

5.3 Effect of Force-Conditioned Admittance Control

This section evaluates whether tactile-based admittance control improves contact quality and motion stability. We evaluate ViT-FMT across tasks **T1**–**T3**, all trained on **Combined** human and teleoperation datasets. The result is shown in Table 3. Across tasks **T1**–**T3**, enabling admittance control consistently improves motion smoothness, reducing translational acceleration by 10.4% and rotational acceleration by 9.0% on average. Contact quality also improves: contact off-center drift decreases by 10.9% and mean contact force increases by 2.7% on average, indicating more stable and better-centered contact when contact is desired. The overall result confirms that our admittance controller substantially improves motion smoothness while maintaining stable, firm contact.

5.4 Effect of Policy Backbone on Whole-body Manipulation Tasks

In this section, we compare the baseline policies $\pi_{0.5}$, Ψ_0, ViT-DiT, and ViT-FMT. All policies are trained on the same combined dataset with force prediction and deployed with the admittance controller (deployment videos in the supplementary material). Overall, ViT-FMT achieves the best motion smoothness, as reflected by the lowest translational and rotational accelerations (2.07 m/s^2 and 14.18 rad/s^2, respectively), and most closely reproduces the motion patterns in the dataset. $\pi_{0.5}$ (4.62 m/s^2, 28.45 rad/s^2) shows more jittery motion, often with rapid end-effector swings. The ViT-DiT (2.90 m/s^2, 18.35 rad/s^2) and Ψ_0 (3.38 m/s^2, 22.83 rad/s^2) tend to stuck in static poses or fail to continue the motion, leading to frequent start-stop behavior and larger accelerations.

6 Conclusion

In this study, we presented WT-UMI, a whole-body humanoid manipulation system built together with a force-conditioned target-pose correction module and a force-supervised planner that leverage force-rich human demonstrations. The correction module converts human hand poses into contact-aware robot target poses by applying learned corrections from teleoperated robot data, yielding action labels for planner training. The force-supervised planner uses a cross-attention force head to predict a contact-force trajectory via direct regression. At deployment, a tactile-based admittance controller consumes the predicted force as the normal-force reference, maintaining stable contact. Together, the target-pose correction, force-supervised planner, human-robot co-training data, and admittance controller improve contact-rich whole-body manipulation across four policy backbones, spanning deformable, large rigid, and human-collaborative tasks.

7 Limitations and Future Work

Our system has three main limitations. First, tactile coverage is constrained by available sensor configurations, so only the palms, forearms, and chest are instrumented. Extending coverage to dexterous hands, legs, and back would broaden the set of tactile-driven tasks WT-UMI can support. Second, our policy does not yet consume RGB vision input, because third-person views introduce a human–humanoid embodiment gap and finding a camera angle that consistently avoids object occlusion is challenging, especially for transport tasks. Fusing tactile sensing with vision is an important next step toward enabling the policy to anticipate future contact and re-establish lost contact. Third, the force head predicts only a scalar normal force; extending it to multi-axis contact wrenches and distributed force maps would support finer force regulation in more complex contact configurations.

References

[1] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu, A. Kheddar, X. B. Peng, Y. Zhu, G. Shi, Q. Nguyen, G. Cheng, H. Gao, and Y. Zhao. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *IEEE/ASME Transactions on Mechatronics*, 31(2):2300–2330, 2026.

[2] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

[3] Y. Niu, Z. Fang, B. Chen, S. Zhou, R. Senthilkumaran, H. Zhang, B. Chen, C. Qiu, H. E. Tseng, J. Francis, and D. Zhao. Learning versatile humanoid manipulation with touch dreaming, 2026.

[4] Y. Hou, Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, S. Feng, B. Burchfiel, and S. Song. Adaptive compliance policy: Learning approximate compliance for diffusion guided control. In *IEEE International Conference on Robotics and Automation*, pages 4829–4836, 2025.

[5] Y. Han, K. Yu, R. Batra, N. Boyd, C. Mehta, T. Zhao, Y. She, S. Hutchinson, and Y. Zhao. Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer. *IEEE/ASME Transactions on Mechatronics*, 30(1):554–566, 2024.

[6] E. Helmut, L. Dziarski, N. Funk, B. Belousov, and J. Peters. Learning force distribution estimation for the gelsight mini optical tactile sensor based on finite element analysis. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 8553–8560, 2025.

[7] E. Helmut, N. Funk, T. Schneider, C. de Farias, and J. Peters. Tactile-conditioned diffusion policy for force-aware robotic manipulation. In *IEEE International Conference on Robotics and Automation*, 2026.

[8] K. Zhang, H. Zhang, Z. Xu, Z. Zhang, M. R. I. Prince, X. Li, X. Han, Y. Zhou, A. Ajoudani, and Y. She. Tacvla: Contact-aware tactile fusion for robust vision-language-action manipulation, 2026.

[9] P. Mittendorfer, E. Yoshida, T. Moulard, and G. Cheng. A general tactile approach for grasping unknown objects with a humanoid robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4747–4752, 2013.

[10] J. A. Barreiros, A. Özgün Önol, M. Zhang, S. Creasey, A. Goncalves, A. Beaulieu, A. Bhat, K. M. Tsui, and A. Alspach. Learning contact-rich whole-body manipulation with example-guided reinforcement learning. *Science Robotics*, 10(105):eads6790, 2025.

[11] T. Cheng, K. Chen, L. Chen, L. Zhang, Y. Zhang, Y. Ling, M. Hamad, Z. Bing, F. Wu, K. Sharma, and A. Knoll. TacUMI: A multi-modal universal manipulation interface for contact-rich tasks. In *Extended Abstracts of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 342–343, 2026.

[12] S. Armleder, F. Bergner, J. R. Guadarrama-Olvera, J. Nakanishi, and G. Cheng. Real-time control of a humanoid robot for whole-body tactile interaction. *Advanced Intelligent Systems*, 7(12):e202500149, 2025.

[13] M. Murooka, K. Fukumitsu, M. Hamze, M. Morisawa, H. Kaminaga, F. Kanehiro, and E. Yoshida. Whole-body multi-contact motion control for humanoid robots based on distributed tactile sensors. *IEEE Robotics and Automation Letters*, 9(11):10620–10627, 2024.

[14] R. Subburaman and O. Stasse. A whole-body multi contact large object manipulation and estimation framework for humanoids using skin patches. In *IEEE-RAS International Conference on Humanoid Robots*, pages 1–8, 2025.

[15] C. Zheng, K. Chen, Z. Bi, Y. Li, L. Pan, J. Zhou, H. Li, and J. Ma. Embracing bulky objects with humanoid robots: Whole-body manipulation with reinforcement learning. In *IEEE International Conference on Robotics and Automation*, page 16930, 2026.

[16] F. Liu, Z. Gu, Y. Cai, Z. Zhou, H. Jung, J. Jang, S. Zhao, S. Ha, Y. Chen, D. Xu, and Y. Zhao. Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation. *IEEE Robotics and Automation Letters*, 10(11):12261–12268, 2025.

[17] S. Chen, Z.-a. Cao, Z. Luo, F. Castañeda, C. Li, T. Wang, Y. Yuan, L. Fan, C. K. Liu, and Y. Zhu. Chip: Learning adaptive compliance for humanoid control through hindsight perturbation. 2025.

[18] L. Wei, X. Peng, R.-Z. Qiu, T. Huang, X. Cheng, and X. Wang. Hmc: Learning heterogeneous meta-control for contact-rich loco-manipulation. In *IEEE International Conference on Robotics and Automation*, 2026.

[19] G. B. Margolis, M. Wang, N. Fey, and P. Agrawal. SoftMimic: Learning compliant whole-body control from examples. 2025.

[20] F. Wu, X. Nal, J. Jang, W. Zhu, Z. Gu, A. Wu, and Y. Zhao. Learn to teach: Sample-efficient privileged learning for humanoid locomotion over real-world uneven terrain. In *IEEE Robotics and Automation Letters*, pages 9048–9055, 2025.

[21] Z. Gu, Y. Chen, Z. Chai, A. Cueva, T. Nguyen, Y. Wu, H. Xue, M. Kim, I. Legene, F. Liu, M. Kim, A. Barula, Y. Chen, and Y. Zhao. Refine-dp: Diffusion policy fine-tuning for humanoid loco-manipulation via reinforcement learning, 2026.

[22] M. Murooka, T. Hoshi, K. Fukumitsu, S. Masuda, M. Hamze, T. Sasaki, M. Morisawa, and E. Yoshida. Tact: Humanoid whole-body contact manipulation through deep imitation learning with tactile modality. *IEEE Robotics and Automation Letters*, 10(8):7819–7826, 2025.

[23] O. Khatib, M. Jorda, J. Park, L. Sentis, and S.-Y. Chung. Constraint-consistent task-oriented whole-body robot formulation: Task, posture, constraints, multiple contacts, and balance. *The International Journal of Robotics Research*, 41(13-14):1079–1098, 2022.

[24] L. Wijayarathne, Z. Zhou, Y. Zhao, and F. L. Hammond. Real-time deformable-contact-aware model predictive control for force-modulated manipulation. *IEEE Transactions on Robotics*, 39(5):3549–3566, 2023.

[25] W. Liu, J. Wang, Y. Wang, W. Wang, and C. Lu. Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation. In *IEEE International Conference on Robotics and Automation*, pages 1105–1112, 2025.

[26] H. Choi, Y. Hou, C. Pan, S. Hong, A. Patel, X. Xu, M. R. Cutkosky, and S. Song. In-the-wild compliant manipulation with umi-ft. In *IEEE International Conference on Robotics and Automation*, 2026.

[27] Y. Huang, P. Lin, W. Li, D. Li, J. Li, J. Jiang, C. Xiao, and Z. Jiao. Taf-vla: Tactile-force alignment in vision-language-action models for force-aware manipulation, 2026.

[28] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai, W. Zhang, and C. Lu. Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation. In *Advances in Neural Information Processing Systems*, volume 38, pages 93409–93439, 2025.

[29] Y. Li, Zhaxizhuoma, H. Jiang, J. Xia, H. Zhang, J. Du, Y. Zhou, J. Zeng, C. Hao, J. Ren, Q. Yu, C. Lu, Y. Qiao, and J. Pang. Forcevla2: Unleashing hybrid force-position control with force awareness for contact-rich manipulation, 2026.

[30] R. Zhao, W. Wang, Y. Ma, X. Li, F. E. H. Tay, M. H. Ang, Jr., and H. Zhu. Fd-vla: Force-distilled vision-language-action model for contact-rich manipulation, 2026.

[31] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *IEEE International Conference on Robotics and Automation*, pages 5628–5635, 2018.

[32] Y. Qin, W. Yang, B. Huang, K. Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox. AnyTeleop: A General Vision-Based Dexterous Robot Arm-Hand Teleoperation System. In *Robotics: Science and Systems*, 2023. ISBN 978-0-9923747-9-2.

[33] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pages 4066–4083, 2025.

[34] J. Aldaco, T. Armstrong, R. Baruch, J. Bingham, S. Chan, K. Draper, D. Dwibedi, C. Finn, P. Florence, S. Goodrich, et al. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation. 2024.

[35] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. In *IEEE International Conference on Robotics and Automation*, 2026.

[36] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems*, page p045, 2024.

[37] Y. Wi, J. Yin, E. Xiang, A. Sharma, J. Malik, M. Mukadam, N. Fazeli, and T. Hellebrekers. Tactalign: Human-to-robot policy transfer via tactile alignment, 2026.

[38] PICO Immersive Pte. Ltd. PICO 4 Ultra: An All-New Mixed Reality Experience. https://www.picoxr.com/global/products/pico4-ultra, 2023.

[39] Z. Zhao, L. Yu, K. Jing, and N. Yang. XRoboToolkit: A cross-platform framework for robot teleoperation. In *IEEE/SICE International Symposium on System Integration*, pages 15–20, 2026.

[40] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[42] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

[43] NVIDIA, J. Bjorck, N. C. Fernando Castañeda, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. Llontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu. GR00T N1: An open foundation model for generalist humanoid robots. 2025.

[44] S. Caron, Y. De Mont-Marin, R. Budhiraja, S. H. Bang, I. Domrachev, S. Nedelchev, P. Du, A. Escande, J. Vaillant, B. Wingo, S. Patapati, and D. San José Pro. Pink: Python inverse kinematics based on Pinocchio, 2026.

[45] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305, pages 17–40, 2025.

[46] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, J. Liu, X. Liu, K. Kang, S. Zang, W. Yuan, M. Pavone, D. Huang, and Y. Wang. ψ_0: An open foundation model towards universal humanoid loco-manipulation. In *Proceedings of Robotics: Science and Systems*, 2026.

[47] Y. Lu, Z. Liu, X. Fan, Z. Yang, J. Hou, J. Li, K. Ding, and H. Zhao. Faster: Rethinking real-time flow vlas, 2026.

[48] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[49] K. Black, A. Z. Ren, M. Equi, and S. Levine. Training-time action conditioning for efficient real-time chunking, 2025.

[50] TouchTronix Robotics Inc. SensX thin-film tactile sensors. https://www.touchtronix.io/.

[51] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li. 3D-ViTac: Learning fine-grained manipulation with visuo-tactile sensing. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pages 2557–2578, 2025.

8 Supplementary

8.1 Sensor Specification and Force Calibration

As shown in Fig. 1, WT-UMI uses a TouchTronix SensX 187 sensor on the chest, SensX 160 sensors on the palm end-effectors [50], and custom 16×26 FlexiTac-style sensors on the forearms [51].

We calibrate the palm TouchTronix SensX 160 sensors under controlled quasi-static normal loading from 0 to 25 N. Two contact configurations are evaluated: direct loading on the bare sensor surface and loading through a compliant gel pad with a thickness of $5\,\mathrm{mm}$ and a Shore-A-20 hardness. The calibrated tactile response is used as a proxy measurement of contact normal force. In this work, only the palm sensors are quantitatively force-calibrated, while the forearm and chest sensors are primarily used for contact localization and interaction-state inference.

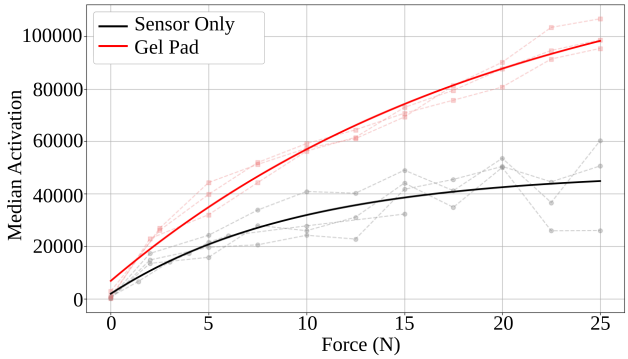

Figure 5: Force calibration of the SensX 160 palm sensor with and without a gel pad. The bare sensor response rises quickly and saturates near 20 N. The gel-pad configuration distributes contact pressure more evenly, yielding a smoother and less rapidly saturating response while extending the usable calibration range.

As shown in Fig. 5, the bare sensor produces a steep activation increase under low loads before saturating near $20\,\mathrm{N}$. In contrast, the gel-pad configuration spreads the contact load across the sensing surface, yielding a smoother and less rapidly saturating response. This improved response facilitates more stable force calibration and extends the force range before saturation. We use this gel-pad configuration for all manipulation experiments in our study. In addition to producing a more stable and consistent calibration response, the compliant gel surface increases contact friction, which improves grasp stability during contact-rich interactions while preserving a consistent force scale between WT-UMI demonstrations and robot deployment.

The contact normal force is calibrated based on the tactile activation map. For each palm sensor, each tactile frame is subtracted by the reading at zero load to be converted into a calibrated activation map. Let $X(t) \in \mathbb{R}^{H \times W}$ denote the calibrated tactile activation map at time t over the taxel grid $\Omega = \{1, \ldots, H\} \times \{1, \ldots, W\}$, with $X_{ij}(t)$ representing the activation at taxel row i and column j. Active taxels above a fixed threshold τ_a form the index set $\mathcal{I}(t)$, whose values are summed into an aggregate activation score $A(t)$. The threshold τ_a is selected empirically from no-contact recordings to suppress background sensor noise and inactive taxels, and is kept fixed across all calibration and deployment data:

$$\mathcal{I}(t) = \{(i,j) \in \Omega \mid X_{ij}(t) > \tau_a\}, \qquad A(t) = \sum_{\substack{(i,j)\in\Omega \\ X_{ij}(t)>\tau_a}} X_{ij}(t). \tag{1}$$

The calibration fits an inverse-exponential response model between the activation score $A(t)$ and applied normal force $f(t)$. This form captures the rapidly saturating response commonly observed in piezoresistive tactile sensors under increasing normal load:

$$A(t) = c_1 \left(1 - \exp(-c_2 f(t))\right) + c_3, \tag{2}$$

where $c_1, c_2, c_3 \in \mathbb{R}$ are scalar parameters fitted from the calibration data. Inverting this model gives an intermediate force estimate

$$\tilde{f}(t) = -\frac{1}{c_2} \log \left(1 - \text{clip} \left(\frac{A(t) - c_3}{c_1}, 0, 1 - \epsilon \right) \right), \tag{3}$$

where $\epsilon > 0$ is a small numerical margin that keeps the argument of $\log(\cdot)$ strictly positive. The final normal-force estimate f_t^m is obtained by applying a linear post-calibration scaling and offset correction to compensate for sensor-dependent gain and baseline variation. Specifically, a scale factor α and offset β are fitted from the calibration measurements and applied to the intermediate force estimate, followed by clipping to the calibrated operating range $[0, f_{\max}]$,

$$f_t^m = \text{clip} \left(\alpha \, \tilde{f}(t) + \beta, \, 0, \, f_{\max} \right). \tag{4}$$

8.2 Tactile-based Calibration Procedure for Human Data Collection

Collecting human demonstrations with WT-UMI requires an online calibration procedure. The calibration process records transformations among the PICO headset, the handheld GripTacs, and the chest plate, ensuring that the recorded human motion can be consistently mapped to the corresponding robot poses.

During calibration, the operator wears the headset and chest plate while holding the GripTacs. Each GripTac's tip is pressed against the chest tactile sensor, activating exactly one cell on the tactile sensor. The resulting tactile image allows the sensor to precisely localize the contact location of the GripTac on the chest plate. Combined with the known GripTac and chest plate geometry, this measurement estimates the transformation from the chest-plate center to the handheld GripTac, denoted as $T^s_{\text{Chest}\to\text{Hand}}$, with $s \in \{l, r\}$ representing the left or right hand. After calibration, the GripTac positions are expressed in the chest-plate frame.

To convert human poses into robot configurations, we use the known transformation from the robot base to the mounted chest-plate center, $T_{\text{Base}\to\text{Chest}}$. The corresponding robot end-effector pose is then computed as $T^s_{\text{Base}\to\text{Hand}} = T_{\text{Base}\to\text{Chest}} \cdot T^s_{\text{Chest}\to\text{Hand}}$. During data inspection, the resulting bimanual targets are passed to an inverse-kinematics solver, and poses that exceed the feasibility tolerance or risk self-collision are rejected. The retained demonstrations are then verified in simulation or on robot hardware. This calibration process ensures that WT-UMI produces consistent demonstrations.

8.3 Target-Pose Correction Training Details

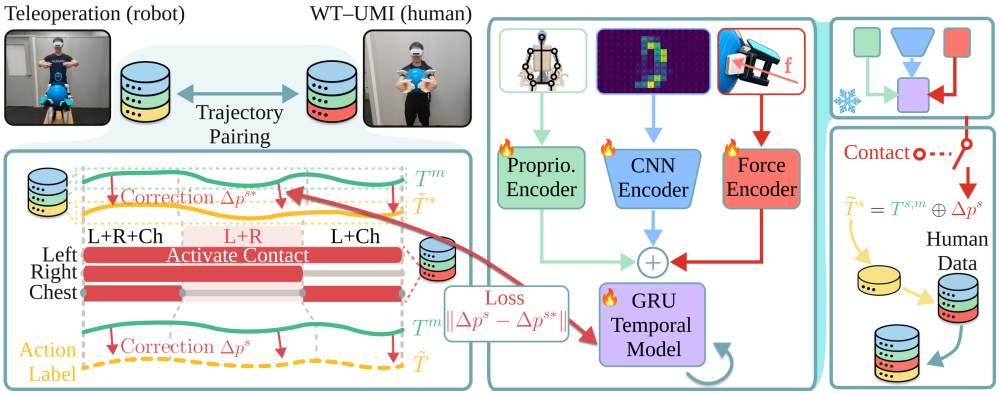

Figure 6: Target-pose correction training. Contact-mode-aligned teleoperation and human segments supervise contact-gated offsets that convert human poses into robot target poses.

This subsection provides details of the data alignment and training procedure for the target-pose correction module in Fig. 2(b), which produces robot-executable action labels for human data by learning contact-dependent hand-pose offsets from teleoperation data. The module is trained separately

for each task on paired teleoperation and human trajectories, as illustrated in Fig. 6. Trajectories are paired by matching their contact-mode sequences (e.g., left hand + right hand + chest → left hand + right hand → left hand + chest → ⋯). Each teleoperation trajectory supplies a residual target-pose offset $\Delta\mathbf{p}^\star$ in the local end-effector frame, that supervises the predicted correction $\Delta\mathbf{p}$ applied on top of the measured human hand pose. A deployment-time force gating ensures the correction is applied only to force-active hands.

The training data are resampled to 50 Hz. The model observes a 0.2 s history of hand pose, tactile images, contact force, and binary contact state for both hands, and predicts corrections over a 0.1 s future target-pose chunk. Because the correction is applied offline to recorded demonstrations, the human hand poses in the upcoming prediction chunk are also available and provided as input. The module then predicts a per-step correction that is applied to each upcoming hand pose to obtain the robot-executable target pose.

The network combines a per-block CNN tactile encoder with a GRU temporal model: a two-layer GRU over the history and a two-layer bidirectional GRU over the future window, both with hidden dimension 256 and dropout 0.3. An auxiliary head predicts the contact-segment class and per-hand contact direction, while the output head regresses the translational offset $\Delta\mathbf{p} = [\Delta\mathbf{p}^l; \Delta\mathbf{p}^r] \in \mathbb{R}^6$, where each $\Delta\mathbf{p}^s \in \mathbb{R}^3$ is predicted in the local end-effector frame of hand $s \in \{l, r\}$. Training minimizes a contact-masked Smooth-L_1 offset loss, a Smooth-L_1 temporal-smoothness loss on consecutive offset differences, and the auxiliary contact losses. We optimize with AdamW using cosine learning-rate decay, warmup, and gradient clipping, with early stopping on validation offset error and auxiliary contact accuracy.

8.4 Force Head Architecture and Hyperparameters

The force head uses a TransformerDecoder with two layers and four attention heads using a hidden dimension $D = 320$, with a total of 0.5M parameters. Crucially, the action denoiser and the force head are split into independent decoders that share only the encoder. The action denoiser takes the noisy action chunk a_k and the denoising step k as input, while the force head reads only the clean observation embedding $\mathbf{h}$. Keeping the decoders independent allows the force head's input to remain free of the noise schedule while still letting force supervision shape the shared encoder.

During training, both flow-matching and transformer backbones undergo the same number of gradient steps using AdamW, cosine learning rate decay, and an exponential moving average over weights. Image augmentations applied during training include additive noise, channel dropout, and random patch masking. The sensor cap $f_{\max}$=20 N is fitted to the calibrated palm range, and the calibration-aware force term is enabled throughout training with $\lambda_F = 1.0$. At inference, the force-conditioned target-pose correction (Sec. 3.2) is gated by the anticipated per-hand force using $f_{\text{threshold}} = 0.5\,\text{N}$.

8.5 Human–Humanoid Collaborative Manipulation

We further evaluate WT-UMI and our planning framework on two human–humanoid collaborative manipulation tasks: beam transport (T4) and table transport (T5), shown in Fig. 7. Unlike the single-agent tasks in Sec. 4, these tasks require the humanoid to maintain distributed whole-body contact while reading its partner's intent from tactile feedback: changes in force distribution and contact activation across the sensing surfaces signal whether the human is pushing or pulling, and the humanoid adapts its whole-body motion to accomplish coordinated transport.

Task Setup. Both tasks require continuous contact regulation and shared-load coordination. In the beam transport task, the humanoid stabilizes a cardboard beam (length 1.33 m, diameter 0.09 m, weight 8.72 N) using distributed contact across the forearms and chest, while a human partner applies pushing or pulling forces to guide the motion. The robot's forearms support most of the load, while its chest provides additional stabilization and prevents slip. The tactile signals in these regions contain visible patterns for intent inference (Fig. 7): during forward motion, left forearm activation

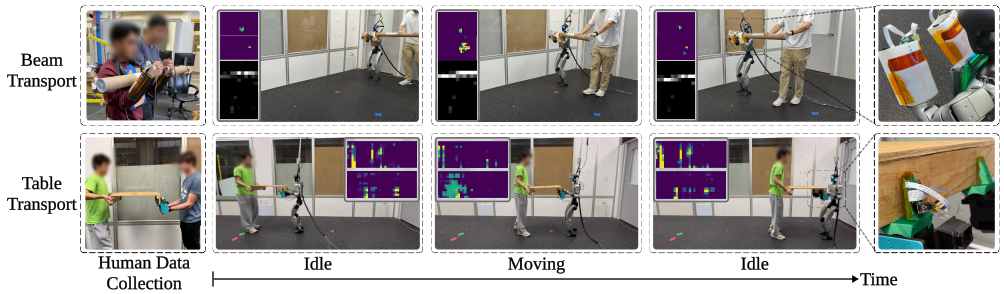

Figure 7: Human–humanoid collaborative manipulation tasks. Top images show the beam transport task, and bottom images show the table transport task. For each task, a depiction of the human-human data collection is included.

typically increases while right forearm activation decreases, and the chest activation distribution shifts with the interaction-force direction.

In the table-transport task, the humanoid supports a plywood table with both palms while adapting to translational motions initiated by a human partner. The table measures $609\,\mathrm{mm} \times 609\,\mathrm{mm} \times 125\,\mathrm{mm}$, with a weight of $39\,\mathrm{N}$. For this task, we use FlexiTac sensors mounted on a custom-designed gripper, as shown in Fig. 1.

Data Collection. For both tasks, demonstrations are collected through human-human collaborative transport, where one participant wears WT-UMI and physically mimics the robot role, while the other acts as the human partner. These demonstrations capture natural shared-load coordination behaviors and tactile contact patterns without requiring robot teleoperation. For the beam transport task, we further collect an additional robot-in-the-loop dataset to cover the embodiment mismatch between human and robot. In this data collection, WT-UMI is mounted directly on the humanoid robot. Overall, we collected 56 beam transport trajectories and 45 table transport trajectories containing pushing, pulling, and idle interaction behaviors, with each trajectory lasting approximately 15 seconds on average.

Deployment. During deployment, the humanoid supports both forward and backward locomotion in the beam and table task, and returns to an idle state when the contact distribution relaxes to the nominal stationary-holding pattern. The transition delay is approximately 0.7 s from idle to active transport and 0.8 s from active transport back to idle.

8.6 Comparison of Four Policy Backbones

To complement the aggregated results in Sec. 5.3 and Sec. 5.4, we compare four policy backbones ($\pi_{0.5}$, Ψ_0, ViT-DiT, and ViT-FMT) on tasks **T1**–**T3**, each evaluated with and without our tactile admittance controller. Each setting is evaluated with $N = 25$ trials per task. Table 4 reports per-task success rate together with contact-quality and motion-smoothness metrics.

Ranked by success rate averaged across tasks **T1**–**T3** and both admittance settings (w/o and w/), ViT-FMT performs best (98.7% on average), followed by $\pi_{0.5}$, while ViT-DiT and the foundation policy Ψ_0 trail behind. ViT-DiT often stucks in the initial hugging pose and fails to continue the rotation on the yoga-ball and bucket tasks, whereas Ψ_0 produces indecisive motion and fails the bucket task entirely (0% on **T3**, leaving its quality metrics undefined).

Across all four backbones, the tactile-based admittance controller regulates the contact centroid and reduces contact drift by approximately 17% on average (Cont. Drift columns). Applying admittance control also adjusts the mean contact force as it tracks the predicted force reference (Cont. Force columns). Motion quality is on-average improved by the admittance controller: across backbones, translational and rotational accelerations decrease by roughly 14% and 13% (Smooth.-Trans.

Table 4: Admittance ablation across all policy backbones over tasks **T1**, **T2**, and **T3**. Each metric is evaluated without (w/o) and with (w/) our admittance control. The better or tied value is bolded for success rate, contact drift, and motion smoothness.
Note: these appendix evaluations are obtained from separate runs from those in Sec. 5.4; due to variations in experiments, the values may differ.

Policy	**Task**	**Succ. Rate (%)**		**Cont. Drift (mm)**		**Cont. Force (N)**		**Smooth.-Trans.** (m/s^2)		**Smooth.-Rot.** (rad/s^2)	
Admi. (Ours) $\rightarrow$		w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
ViT-FMT	T1	**100**	**100**	18.12	**15.67**	4.77	**5.50**	3.14	**2.98**	20.29	**18.56**
	T2	**100**	**100**	21.04	**19.44**	**0.52**	0.13	**1.87**	1.95	**12.63**	13.61
	T3	80	**92**	25.00	**22.08**	0.96	**1.61**	1.85	**1.29**	14.05	**10.38**
ViT-DiT	T1	**52**	**52**	21.79	**18.00**	2.21	**2.57**	2.92	**2.54**	18.73	**14.97**
	T2	**100**	**100**	21.22	**19.61**	0.19	**0.76**	4.67	**2.41**	25.93	**15.17**
	T3	**52**	**52**	26.22	**22.20**	**1.74**	0.93	2.33	**2.05**	16.47	**14.79**
$\pi_{0.5}$	T1	88	**92**	21.51	**11.78**	2.53	**2.80**	**4.69**	4.82	**29.33**	29.79
	T2	68	**76**	19.78	**15.97**	**0.50**	**0.50**	5.06	**5.03**	31.86	**31.22**
	T3	**84**	76	20.18	**19.40**	2.57	**3.38**	4.11	**3.53**	27.26	**23.56**
Ψ_0	T1	88	**92**	15.80	**13.56**	**3.14**	3.03	4.20	**3.27**	27.06	**23.01**
	T2	89	**96**	20.48	**18.69**	**2.50**	**2.50**	**4.94**	5.44	**32.12**	33.04
	T3	0	0	-	-	-	-	-	-	-	-

and Smooth.-Rot. columns). Success rates are mostly unchanged, indicating that the added force feedback improves contact centering without compromising task completion.

The two foundation policies, $\pi_{0.5}$ and Ψ_0, are less smooth, as reflected by their higher translational and rotational accelerations. Their jerkier motion can be attributed to the slower inference and more conservative real-time chunking (RTC) settings.

Overall, ViT-FMT is the strongest of the four baseline backbones, so we adopt it as the default backbone in the ablation studies in Sec. 5.3 and Sec. 5.4. Notably, the backbone itself is not a contribution of this work; our contributions, the force-conditioned target-pose correction, the force-supervised planner, and the tactile-based admittance controller, are backbone-agnostic and improve contact quality across all four baselines.

8.7 Ablation Study of Target-Pose Correction and Admittance Controller on Recorded Data

In addition to the policy evaluations in Sec. 5.2 and Sec. 5.3, we isolate the effects of our target-pose correction and admittance control directly on the recorded data, removing the confounding influence of policy backbones and training setups. We replay the collected yoga-ball trajectories on the robot hardware under four configurations: raw human data, raw teleoperation data, target-pose-corrected human data, and corrected human data with admittance control enabled. This comparison evaluates trajectory feasibility across both data sources. It also tests whether target-pose correction improves the feasibility of human data and whether admittance control improves contact quality. We report the same metrics as in Sec. 5.2.

Table 5: Force-modulation module ablation on data replay across four configurations. "Failed" indicates the configuration cannot complete the task. Best value in bold for success rate, contact drift, and motion smoothness.

Configuration	**Success Rate (%)**	**Contact Center Drift (mm)**	**Mean Contact Force (N)**	**Smooth.-Trans.** (m/s^2)	**Smooth.-Rot.** (rad/s^2)
Raw Human	Failed	–	–	–	–
Raw Teleoperation	85.35	15.59	**3.30**	3.93	26.42
Correction	89.29	17.88	3.23	**3.57**	**25.51**
Correction + Admittance	**96.15**	**11.46**	3.02	4.06	27.15

As shown in Table 5, **Raw Human** data fails to complete the task due to loss of contact, confirming that human motion is not directly robot-executable for lack of action labels. In contrast, **Raw Teleoperation** is executable and completes the task at an 85.35% success rate. Our proposed target-pose **Correction** improves the feasibility of the human trajectories and raises the success rate to 89.29%. Adding admittance control (**Correction + Admittance**) further raises the success rate to 96.15% and reduces contact center drift by 35.9%. This gain comes with a moderate loss of motion smoothness, which is expected because admittance control introduces reactive adjustments based on contact feedback. This ablation confirms that target-pose correction is the key component enabling the feasibility of human data, while admittance control improves contact quality.

8.8 Admittance Controller Details

The corrective SE(3) increment ΔT_t^s introduced in Sec. 3.4 contains a rotation $\Delta R_{xy,t}^s$ and a translation $\Delta \mathbf{p}_{z,t}^s$ derived from force feedback. Both follow proportional control laws,

$$\Delta R_{xy,t}^s = \mathrm{Exp}\big([(K_R(\mathbf{c}_t^s - \mathbf{c}_t^{s,\mathrm{m}})); 0]_\times\big), \qquad \Delta \mathbf{p}_{z,t}^s = K_F(f_t^s - f_t^{s,\mathrm{m}})\,[0,0,1]^\top, \tag{5}$$

where $\mathbf{c}_t^s, \mathbf{c}_t^{s,\mathrm{m}} \in \mathbb{R}^2$ are the desired and measured contact centroids in the sensor frame, f_t^s and $f_t^{s,\mathrm{m}}$ are the reference (planner-predicted) and measured normal forces, $K_R \in \mathbb{R}^{2\times 2}$ is the contact-centering gain matrix, and K_F is a scalar normal-force gain. The rotation term reorients the palm to re-center the contact centroid, while the translation term drives the local sensor-normal motion to track the reference normal force.

Force regulation evaluation. We evaluate the closed-loop force regulation of the tactile admittance controller from Sec. 3.4. The robot holds a yoga ball between its chest and palms under a fixed pose target, with both palms initially contacting the ball at 1 N. It then tracks step changes in the desired contact force f_t^s from 1 N to 4 N.

Fig. 8 shows the force-tracking response: both palms reach each setpoint with steady-state errors within $3.5\,\%$, confirming that the proportional admittance law tracks the planner-predicted force reference accurately.

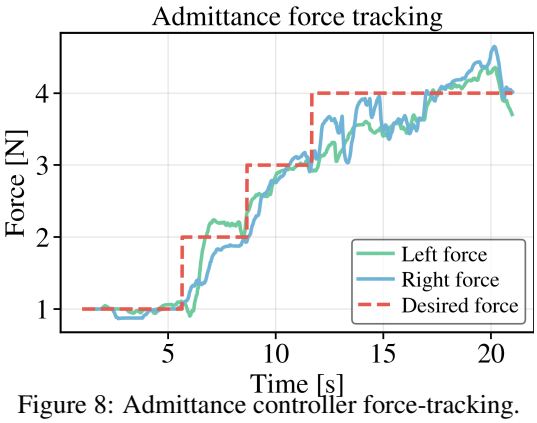

Figure 8: Admittance controller force-tracking.