

Unified Visual-Tactile-Action Modeling from Human Demonstrations for Dexterous Manipulation

Wenqiao Li^{*1}, Qianyou Zhao^{*3}, Jiawen Hao⁴, Xuezhou Zhu³, Tengyu Liu²,
Kaifeng Zhang³, Chuan Wen^{✉1}, Siyuan Huang^{✉2}

¹ Shanghai Jiao Tong University

² State Key Laboratory of General Artificial Intelligence, BIGAI

³ Sharpa Robotics

⁴ Beijing Institute of Technology

^{*} Denotes equal contributions [✉] Corresponding authors

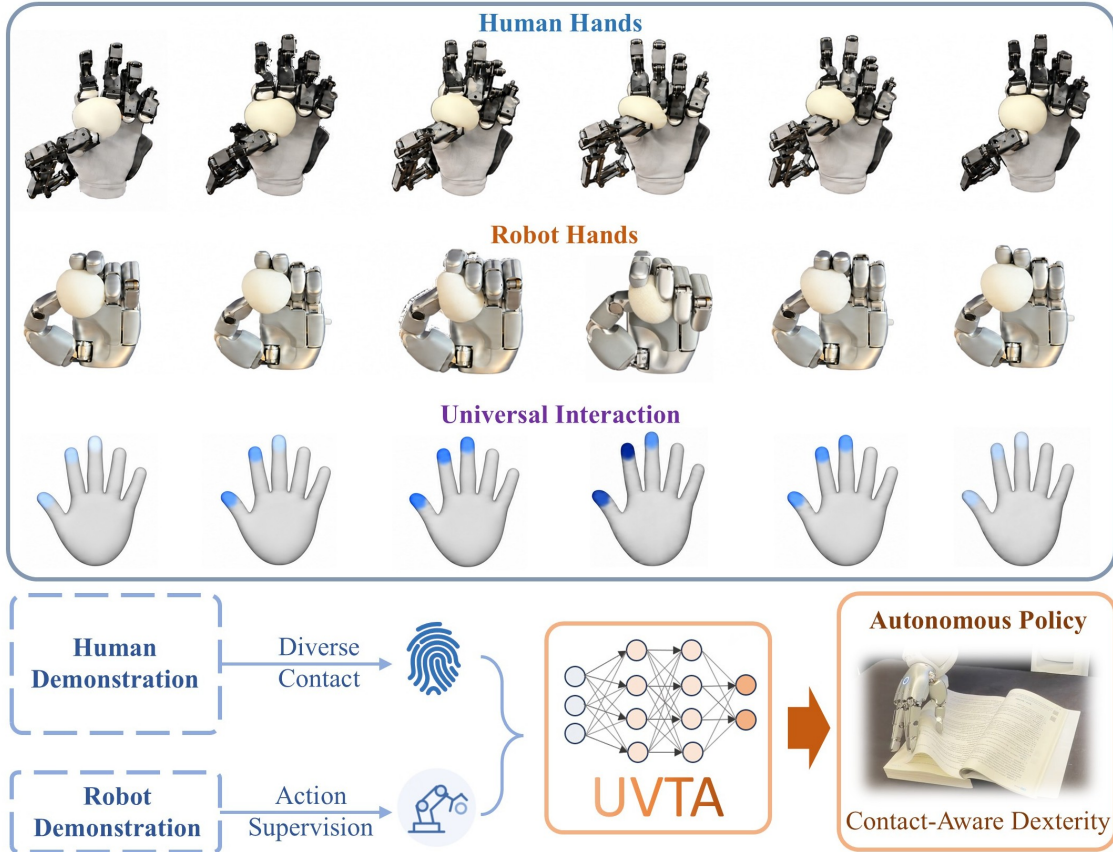


Fig. 1. We propose Unified Visual Tactile Action (UVTA) modeling, a framework that enables contact-aware dexterity learning from human tactile data. Under the UVTA framework, **Human hand** and **robot hand** share a **universal interaction representation**.

Abstract—Dexterous manipulation requires tactile feedback. However, robot tactile demonstrations are difficult to scale because dexterous-hand teleoperation provides limited tactile feedback to the operator. In contrast, human demonstrations offer a substantially more scalable source of diverse tactile interactions. Motivated by a simple premise—*hands can change, but the underlying physics of interaction does not*—we leverage human tactile data to improve dexterous manipulation policies. Specifically, we first build a tactile motion-capture system that synchronously records images, tactile signals, and hand motions. Using this system, we construct the UVTA dataset spanning five contact-rich tasks, with 1,000 human demonstrations covering diverse interaction patterns and 150 robot demonstrations per task. To transfer the underlying physics of human interaction to robot control, we propose

a Unified Visual-Tactile-Action Model that maps both embodiments into aligned tactile and action representations and jointly predicts future action and tactile trajectories. The joint objective enables human demonstrations to supervise contact-aware representation learning, while only robot actions are executed during deployment. In real-robot evaluations across five tasks, our method achieves an average success rate of 70%, outperforming the strongest visual-tactile baseline, which achieves 29%, and an architecture ablation, which achieves 42%. Performance improves consistently with additional human demonstrations and exhibits no saturation at 1,000 demonstrations per task, validating the effectiveness of scalable human tactile data for dexterous manipulation. Project page is available at <https://uni-vta.github.io/>.

I. INTRODUCTION

Tactile feedback is indispensable for dexterous manipulation. Contact-rich tasks depend on contact onset, pressure, deformation, and slip—physical states that are often ambiguous or occluded in visual observations. Access to these states has enabled increasingly precise and reactive manipulation [1, 2, 3, 4, 5]. Nevertheless, learning tactile policies typically requires demonstrations collected through dexterous robot-hand teleoperation. Such data are difficult to scale: teleoperating a high-dimensional hand is demanding, and conventional interfaces provide the operator with limited tactile feedback. Consequently, collecting synchronized visual, tactile, and action trajectories remains costly, constraining the diversity of objects, environments, and contact patterns represented in robot datasets.

Human demonstrations provide a substantially more scalable source of diverse physical interactions. Large-scale human video has supported visual representation learning and long-horizon robot control [6, 7, 8, 9, 10], while motion reconstruction, retargeting, and human-robot co-training transfer human behavior more directly [11, 12, 13, 14, 15, 16]. However, human and robot hands differ in morphology, kinematics, sensing layouts, and action spaces, making direct motion correspondence unreliable. Moreover, video alone does not fully capture the local contact dynamics that determine manipulation success. We therefore seek to transfer human interaction experience through a quantity that is more fundamental than embodiment-specific motion. Our central premise is that *hands can change, but the underlying physics of interaction does not*: although embodiments differ, contact formation, deformation, slip, release, and object response retain shared physical structure.

To capture this structure at scale, we first develop a wearable tactile motion-capture system that synchronously records wrist-centric images, fingertip tactile measurements, and hand motion. We then use the system to construct the UVTA dataset over five contact-rich tasks. Each task contains 1,000 human demonstrations that cover diverse interaction patterns and 150 robot demonstrations. This composition combines the diversity and scalability of human interaction with the embodiment-specific supervision required for robot execution.

Building on this dataset, we propose a *Unified Visual-Tactile-Action Model* that maps human and robot trajectories into aligned tactile and action representations. The model jointly predicts future action and tactile trajectories, using human demonstrations to supervise contact-aware representation learning while executing only robot actions at deployment. Predicting tactile outcomes connects motion to its physical consequences, providing a shared learning signal across embodiments [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. Across five real-robot tasks, our method achieves a 70% average success rate, compared with 29% for the strongest visual-tactile baseline and 42% for the ablated architecture. Performance continues to improve as the number of human demonstrations increases, with no

saturation at 1,000 demonstrations per task. These results demonstrate that scalable human tactile data can substantially improve contact-rich dexterous manipulation.

Our main contributions are threefold:

- We develop a wearable tactile motion-capture system that synchronously captures visual observations, fingertip tactile measurements, and hand motion without requiring robot teleoperation.
- We introduce the UVTA dataset, comprising five contact-rich tasks with 1,000 diverse human demonstrations and 150 robot demonstrations per task.
- We propose a unified visual-tactile-action model that transfers contact-aware physical representations across embodiments. Real-robot experiments and scaling studies validate both its performance and its ability to benefit from increasing amounts of human tactile data.

II. RELATED WORK

Robot Learning from Human Data. Large-scale egocentric datasets capture diverse human interactions across tasks, objects, and environments [6, 7, 30, 31]. This experience has been used to learn reusable visual representations and latent plans for downstream robot control [8, 9, 10, 32, 33], while a complementary line transfers human motor behavior through pose reconstruction, action retargeting, portable motion capture, or human-robot co-training [11, 12, 13, 14, 15, 16, 34]. Although these methods substantially reduce the demand for robot demonstrations, direct policy transfer still relies on recoverable human motion or an aligned action representation. Predictive approaches instead use embodiment-invariant interaction outcomes as supervision [25]. Our work jointly predicts action and future tactile trajectories from both human and robot demonstrations. Human actions contribute training supervision but are not treated as robot commands; deployment selects only robot-action predictions, while paired action-touch dynamics provide the shared cross-embodiment representation.

Tactile Learning for Dexterous Manipulation. Touch reveals contact, force, slip, and deformation that may be ambiguous or occluded in vision, enabling closed-loop dexterity in contact-rich tasks [1, 2, 3, 35, 36]. Beyond task-specific control, tactile representations have been learned from robotic play, masked multimodal objectives, large unlabeled tactile corpora, and heterogeneous sensor data [37, 38, 39, 40, 41]. Policy-level systems further combine vision and touch for bimanual or multifingered manipulation, high-frequency reactive control, and physically grounded vision-language-action learning [28, 29, 42, 43, 44]. Recent scaling efforts introduce tactile mid-training and human-centric tactile pre-training [4, 5]. Nevertheless, existing systems predominantly learn from robot touch or transfer human touch through staged training with aligned human-robot action spaces. We instead co-train human tactile motion-capture and robot teleoperation trajectories with the same action and future-tactile objectives. Their coupling lets both embodiments shape the visual representation, while deployment retains robot-specific execution.

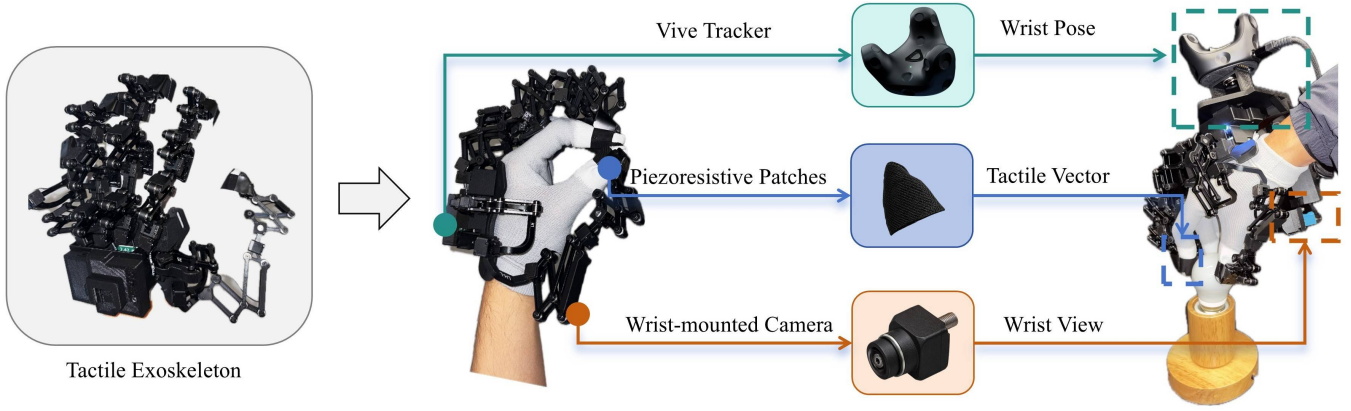


Fig. 2. **Overview of the Tactile Motion-Capture System.** A passive tactile exoskeleton captures hand motion and tactile signals. The integrated system further includes a VIVE tracker and a wrist-mounted RGB camera for capturing the wrist pose and wrist-centric images, respectively. The camera and VIVE tracker are calibrated to the Sharpa Wave wrist frame.

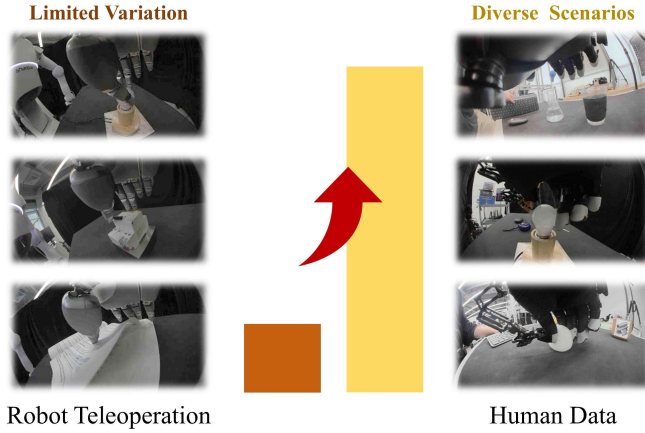


Fig. 3. **Dataset Overview.** Each task contains 150 robot demonstrations collected under limited variation, and 1,000 human demonstrations with diverse scenarios and contact patterns.

World Models for Robot Learning. World models support robot learning by predicting future observations or latent states and coupling these predictions to planning or action generation [17, 18, 19, 45, 46, 47, 48]. Recent systems integrate future-video or future-feature prediction directly into generalist policies and jointly model visual dynamics and actions [20, 21, 22]. Scaling this paradigm with heterogeneous robot and human videos has produced world action models that transfer physical knowledge across tasks and embodiments [23, 24, 25]. However, most such models predict visual futures, in which contact onset, force, and local deformation are only weakly observable. Tactile prediction offers a complementary world-modeling target, from anticipating local touch observations to generating tactile subgoals and multimodal interaction dynamics [5, 26, 27, 29, 37]. Our model jointly predicts future tactile and action trajectories from both human and robot interactions, coupling physical consequences with the motions that produce them. At deployment, only the robot-action predictions are executed.

III. DATA COLLECTION SYSTEM AND DATASET

We collect human demonstrations rich in contact using the wearable tactile exoskeleton glove system shown in Fig. 2. The system records wrist fisheye images, fingertip tactile signals, and 22-DoF joint angles. After processing and temporal synchronization, the raw signals captured by this system are aligned with Sharpa Wave Hand in both kinematic and contact space.

A. Tactile Motion-Capture System

Tactile sensing. We employ conventional force-sensing resistors (FSR) to capture human tactile signals. Specifically, each fingertip of the tactile-sensing exoskeleton glove is equipped with an FSR array containing four sensing regions, each with an active area of $15 \times 15\text{mm}^2$. When external pressure is applied to the fingertip, the resistance of the corresponding sensing point changes. This change is converted into an electrical signal and recorded as the tactile measurement at that location.

Hand Motion. Wrist and finger motions are captured using a pair of VIVE Tracker 3.0 devices and a passive linkage exoskeleton, respectively. One tracker is fixed to the table to establish a reference frame, whereas the other is rigidly attached to the wrist via a 3D-printed adapter. At each timestep, the pose of the wrist tracker relative to the reference tracker is composed with the calibrated tracker-to-wrist transformation to recover the 6-DoF wrist pose. Finger motion is measured by a passive 22-DoF linkage exoskeleton mounted on the dorsum of the hand. The exoskeleton combines rigid carbon-fiber links with a compliant dorsal support plate. Forward kinematics converts the encoder measurements into the positions and orientations of the five fingertips, which are then transformed into the Sharpa Wave palm frame using a fixed calibration. A constrained inverse-kinematics optimization recovers the corresponding 22-DoF hand configuration by matching fingertip poses while preserving the relative geometry between the thumb and the other fingertips. Joint limits are enforced as hard constraints,

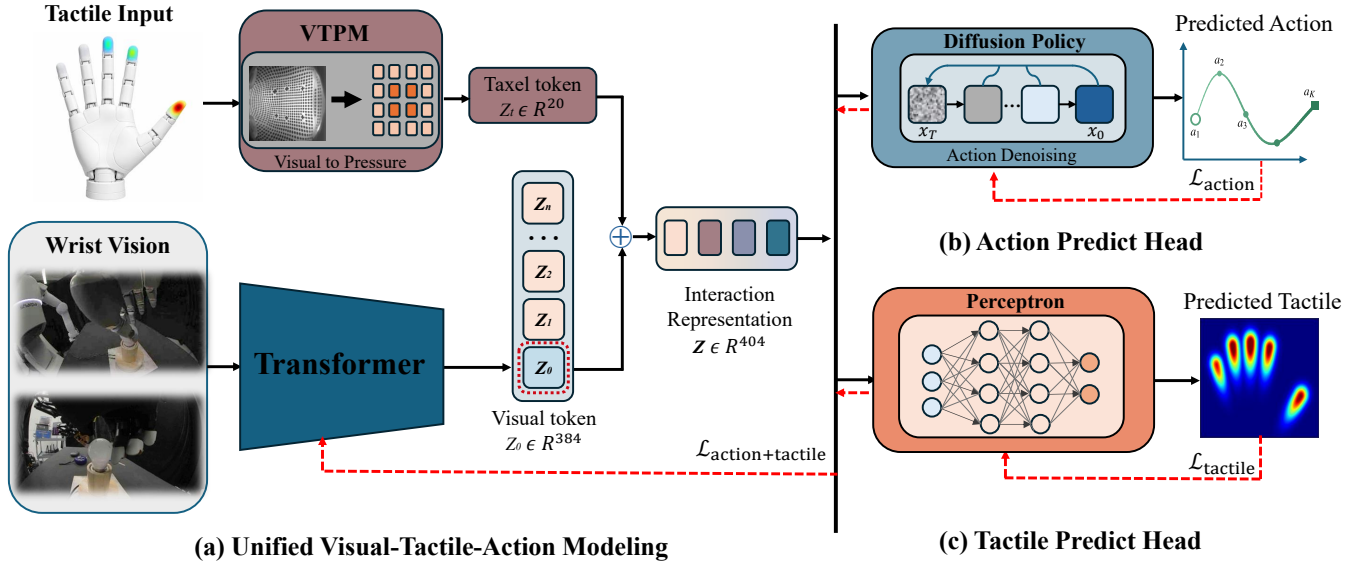


Fig. 4. **Overview of Our Unified Visual-Tactile-Action Model.** VTPM converts the visual-tactile input into a 20-D piezoresistive taxel token, while a Transformer extracts a 384-D visual token. The two tokens are concatenated into a unified 404-D interaction representation and passed to a diffusion action head and a tactile regression head. The action and tactile losses supervise their respective prediction heads, while their joint objective optimizes the shared Transformer to learn a unified tactile-aware visual representation.

whereas the thumb carpometacarpal (CMC) and four-finger metacarpophalangeal (MCP) couplings are modeled as soft linear constraints.

Image Stream. A monocular fisheye RGB camera is rigidly mounted near the wrist using a fixed adapter. Its camera-to-wrist extrinsic transformation is calibrated to match the camera-to-end-effector transformation of the Sharpa North robot, thereby maintaining consistent camera–hand geometry across embodiments. The wrist image stream, tactile signals, wrist poses, and joint motion are timestamped using a shared host clock and temporally resampled at 30 Hz.

B. Dataset

Our dataset covers five contact-rich manipulation tasks, each with 150 robot and 1,000 human demonstrations. Robot demonstrations are collected through teleoperation on the Sharpa North platform, with the scene layout, initial robot configuration, and table height fixed for each task. Human demonstrations are collected using our wearable tactile motion-capture system without a robot in the loop. We vary object instances and scene backgrounds, and use black and white gloves to vary hand appearance. Fig. 3 contrasts the fixed robot collection settings with the more varied human demonstrations.

Robot demonstrations pair visual-tactile observations with executable wrist and finger commands, providing direct supervision for control on the target platform. Human demonstrations contribute more varied motion and contact sequences for learning contact-aware representations across objects and scenes. Both sources supervise action and future-tactile prediction, training the shared representation to cap-

ture how contact changes with motion. Human motions serve as training targets rather than directly executable robot commands.

IV. METHOD

We aim to learn generalizable interaction representations from human tactile demonstrations for dexterous manipulation. To bridge the visual, tactile, and action embodiment gaps between human and robot data, we align both into a unified cross-embodiment space (Sec. IV-B). As shown in Fig. 4, our Unified Visual-Tactile-Action Model conditions action generation and future-tactile prediction on the same wrist-centric visual-tactile observation (Sec. IV-C). It is trained jointly on human motion-capture and robot teleoperation samples using a balanced co-training recipe; during inference, only robot-normalized action predictions are converted into wrist and finger commands.

A. Problem Formulation

We consider contact-rich dexterous manipulation with two data sources: a human dataset $\mathcal{D}^H$ collected by our tactile motion-capture system and a robot dataset $\mathcal{D}^R$ collected by teleoperation. A sample from either embodiment contains, at each timestep, a synchronized wrist RGB image $\mathbf{I}_t$, a tactile token $Z_t \in \mathbb{R}^{20}$, and an action $\mathbf{a}_t \in \mathbb{R}^{31}$. Given the current visual-tactile observation, the model predicts a future action chunk and a future tactile trajectory over horizon H :

$$\mathbf{A}_t = \mathbf{a}_{t:t+H-1}, \quad \mathbf{Y}_t = Z_{t+1:t+H}, \quad (1)$$

where $H = 16$ in our implementation. Robot actions are executable teleoperation commands, whereas human actions

are derived from consecutive motion-capture states. We formulate policy learning as the joint prediction of future human and robot interaction trajectories:

$$(\hat{\mathbf{A}}_t, \hat{\mathbf{Y}}_t) \sim \pi_{\Theta}(\cdot \mid \mathbf{I}_t, Z_t), \quad (2)$$

where $\Theta = \{\theta, \phi, \psi\}$ denotes the parameters of the visual encoder, action head, and tactile head, respectively. At test time, only the robot action prediction $\hat{\mathbf{A}}_t$ is sent to the controller.

B. Unified Action and Tactile Space

We aim to learn general physical representations from human tactile demonstrations that are directly useful for dexterous manipulation. However, human tactile motion-capture data differs from robot data in both tactile representation and kinematics, while inevitably containing tactile and motion noise. To align robot and human tactile observations, we introduce Visual-Tactile-to-Piezoresistive Taxel Mapping (VTPM), which converts visual-tactile deformation maps captured by the Wave sensors into pressure-like piezoresistive taxel signals. For each finger, valid pixels are assigned to four glove-aligned regions in a canonical fingertip frame, and the deformation intensity within each region is aggregated into one scalar value. Concatenating all five fingers produces a unified 20-dimensional tactile representation shared by human and robot demonstrations. For the action space, human and robot motions are represented using the same 31-dimensional layout. Wrist motion is expressed relative to the current wrist frame using 3-D translation and a continuous 6-D rotation representation, while finger motion is represented by 22 absolute joint angles. This relative-wrist and absolute-finger formulation reduces dependence on the global workspace while preserving precise hand configurations for dexterous manipulation.

C. Model Architecture

As shown in Fig. 4, the model has a shared visual-tactile condition and two output branches. We omit the embodiment superscript on intermediate features for clarity. A trainable transformer encoder f_{θ} maps the current 224×224 wrist image to a 384-D visual token,

$$Z_0 = f_{\theta}(\mathbf{I}_t) \in \mathbb{R}^{384}. \quad (3)$$

The encoder uses a ViT-S/8 architecture and is optimized from scratch in our implementation. In parallel, VTPM provides the 20-D tactile token Z_t , after normalization with statistics specific to embodiment e . The model performs early feature-level fusion by direct concatenation,

$$Z = [Z_0; Z_t] \in \mathbb{R}^{404}. \quad (4)$$

The action branch models a 16-step, 31-D action chunk with conditional diffusion. Let $\mathbf{x}_0 = \mathcal{N}_e(\mathbf{A}_t) \in \mathbb{R}^{16 \times 31}$, where $e \in \{\text{H}, \text{R}\}$ selects the normalization statistics of the sample’s embodiment, and each step contains a 9-D relative wrist target (3-D translation and 6-D rotation) and

22 absolute finger-joint targets. At diffusion step k , Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is added as

$$\mathbf{x}_k = \sqrt{\bar{\alpha}_k} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_k} \epsilon. \quad (5)$$

A FiLM-conditioned 1-D U-Net $\epsilon_{\phi}(\mathbf{x}_k, k, Z)$ predicts the injected noise. Its channel widths are 256, 512, and 1024. In parallel, the tactile world head is an MLP comprising two linear layers: the first maps the 404-D interaction representation to a 512-D hidden feature and applies a Mish activation, while the second produces a 320-D output. It directly regresses the complete future tactile trajectory from the noise-free condition:

$$\hat{\mathbf{Y}}_t = g_{\psi}(Z) \in \mathbb{R}^{16 \times 20}. \quad (6)$$

The tactile target is shifted one step ahead and aligned with the action horizon. Keeping it outside the diffusion trajectory prevents tactile denoising errors from entering the iterative action-generation process, while its regression loss still updates the shared visual encoder through Z .

D. Training Recipe

We train each task-specific model in a single human–robot co-training run, rather than using sequential pretraining, mid-training, and post-training stages. The training set is the union $\mathcal{D}^{\text{H}} \cup \mathcal{D}^{\text{R}}$. A weighted random sampler assigns equal probability mass to the human and robot datasets and samples with replacement, preventing their different numbers of frames from determining the mixture ratio.

We apply min–max normalization separately to each embodiment so that both human and robot channels occupy the same numerical range; the saved robot statistics are used to decode predictions at deployment. To deal with tactile sensor native noise problem, tactile signals are baseline-corrected per recording before normalization. For human data, the baseline is the mean of the first five frames; for robot data, it is the first frame.

For each sample, we draw a diffusion step k uniformly from 50 training noise levels. The denoising objective averages the wrist and hand errors separately and weights the two groups equally, so the 22-D hand block does not dominate the 9-D wrist block:

$$\mathcal{L}_{\text{action}} = \frac{1}{2} \text{MSE}(\epsilon_{\text{wrist}}, \hat{\epsilon}_{\text{wrist}}) + \frac{1}{2} \text{MSE}(\epsilon_{\text{hand}}, \hat{\epsilon}_{\text{hand}}), \quad (7)$$

where each MSE is itself averaged over batch, time, and dimensions within that group. The future-tactile loss is

$$\mathcal{L}_{\text{tactile}} = \text{MSE}(g_{\psi}(Z), \mathcal{N}_e(Y_t)), \quad (8)$$

and the optimized objective is

$$\mathcal{L}_{\text{action+tactile}} = \mathcal{L}_{\text{action}} + \lambda \mathcal{L}_{\text{tactile}}. \quad (9)$$

The tactile-loss weight λ is set to 0.2 in all experiments. We optimize all modules jointly for 300 epochs with batches of 250 using AdamW.

During rollout, DDIM uses 16 reverse steps to sample a 16-step robot action chunk. We invert the robot normalization, convert the 6-D rotations back to SO(3), and anchor

the relative predictions of the wrist to the live pose as $\hat{\mathbf{T}}_{t+k} = \mathbf{T}_t \Delta \hat{\mathbf{T}}_{t,k}$. The controller executes the first eight wrist and hand waypoints before replacing the newly visual-tactile feedback. The tactile-head output is not used for control.

V. EXPERIMENTS

In this section, we aim to answer the following questions through our experiments:

RQ1: *What is the value of human tactile data for contact-rich dexterous manipulation?*

RQ2: *Does scaling human tactile data improve real-robot performance?*

RQ3: *How important are the architecture and training design of the world model?*

A. Experiment Setup

Tasks: To evaluate policy performance, we design five contact-rich manipulation tasks. Each task is provided with 150 robot demonstrations and 1,000 human demonstrations. **(Task I) Flip Page.** The robot approaches the book, slows down near the page, and coordinates its fingers to rub and separate a single page before turning it to the left.

(Task II) Screw Light Bulb. The robot approaches a light bulb loosely placed in its socket and coordinates its wrist and fingers to screw the bulb into the socket until it lights up.

(Task III) Toggle Switch. The robot approaches a circuit breaker on the table, stabilizes it with the index and middle fingers, and uses the thumb to firmly push the switch until it is fully turned on.

(Task IV) Ball Classification. The robot approaches visually similar balls made of different materials, infers their material properties through tactile interaction, and sorts them into the corresponding cups.

(Task V) Liquid Transfer with a Dropper. The robot grasps a dropper, inserts it into a flask to draw water, transfers it to a target cup, and dispenses the water without spilling.

Robot Platform and Action Space: All real-world experiments use a fixed-base North robot with two 22-DoF Sharpa Wave dexterous hands. The policy receives the RGB images from the right-arm monocular wrist camera and calibrated tactile readings from the fingertips. Actions use relative end-effector delta control for the right arm and absolute joint control for the right-hand fingers.

Baselines: We compare our method with three representative visuo-tactile policies: (1) **ViTacFormer** [28], an ACT-style dexterous imitation policy that learns cross-modal visual-tactile representations through cross-attention and auxiliary future tactile prediction; (2) **Reactive Diffusion Policy (RDP)** [43], a slow-fast diffusion policy that combines low-frequency visual action-chunk generation with high-frequency tactile-reactive refinement for contact-rich manipulation; and (3) **T-Rex** [4], a pretrained tactile-reactive dexterous foundation model that uses a variable-rate Mixture-of-Transformers to couple low-frequency visuo-motor planning with high-frequency tactile action refinement.

For a fair comparison, all methods receive the same wrist RGB and per-finger tactile observations, predict the same right-arm and right-hand action space, and use identical task-specific human and robot demonstration splits and evaluation protocols. Each baseline retains its original tactile fusion and policy architecture and is trained or post-trained on the same number of robot demonstrations as our method.

Evaluation Metric: We evaluate each method and ablation over 10 real-robot trials per task, with randomized object positions and heights. Rather than scoring each rollout as a binary outcome, we report a stage-wise success rate that awards partial credit for reaching intermediate stages. For example, in the flip page task, a rollout receives a score of 0.3 for successfully separating the pages, 0.7 for turning multiple pages, and 1.0 for successfully turning exactly one page.

B. Main Results and Ablations

Table 1 jointly reports the baseline comparison and controlled ablations. Scores are averaged over 10 real-robot rollouts and reported as percentages. The final column is the unweighted mean over the five tasks.

Human Tactile Data is Key to Strong Contact-Rich Manipulation Policy. Our method achieves a 70% average success rate and performs best on all five tasks, exceeding the strongest baseline, T-Rex, by 41 percentage points. The gains are particularly pronounced on *Toggle Switch* (88% versus 24%) and *Screw Light Bulb* (80% versus 26%), where successful execution depends on contact events that are difficult to resolve from wrist vision alone. In contrast, training with human data alone yields 0% because human trajectories are not directly executable by the robot. Together with the 0% result obtained with no human trajectories in Fig. 5, these results show that neither source is sufficient in isolation under the available data budget. The improvement therefore comes from combining diverse human interaction experience with robot-specific action supervision through unified co-training.

Performance scales with human tactile data and has not saturated. Fig. 5 varies the number of human trajectories while keeping the 150 robot demonstrations per task fixed. Averaged over *Flip Page*, *Screw Light Bulb*, and *Toggle Switch*, success increases monotonically from 0% without human data to 8%, 29%, 51%, and 84% with 250, 500, 750, and 1,000 human trajectories, respectively. All three tasks exhibit the same positive trend, indicating that the benefit is not confined to a single manipulation behavior. Moreover, the largest average gain, 33 percentage points, occurs between 750 and 1,000 trajectories. The curve thus remains steep at the largest dataset size evaluated, suggesting that performance has not yet saturated and may continue to benefit from additional human tactile demonstrations.

Future Tactile Prediction is Critical to the Model. The ablations in Table 1 isolate the contributions of tactile conditioning and future-tactile prediction. Removing the tactile prediction head while retaining tactile input reduces the average success rate from 70% to 42%, demonstrating

Table 1: **Comparison of Our Method with Baseline and Ablation Methods Across Five Tasks.** Success rates (%) are computed over 10 rollouts per task, then averaged across tasks. The last row is our full model; the blocks above compare it against representative baselines and against ablations of human data only, future-tactile prediction, tactile input, and proprioceptive conditioning.

Configuration	Flip Page	Light Bulb	Toggle Switch	Ball Classification	Liquid Transfer	Average
<i>Baseline Methods</i>						
ViTacFormer [28]	0	4	0	0	0	1
RDP [43]	10	20	20	18	0	14
T-Rex [4]	43	26	24	36	15	29
<i>Ablation Methods</i>						
Human Data only	0	0	0	0	0	0
w/o Tactile Prediction	58	38	50	44	21	42
Vision-only	47	23	26	20	14	26
w/ Proprioception	22	20	14	20	15	18
Ours	83	80	88	56	44	70

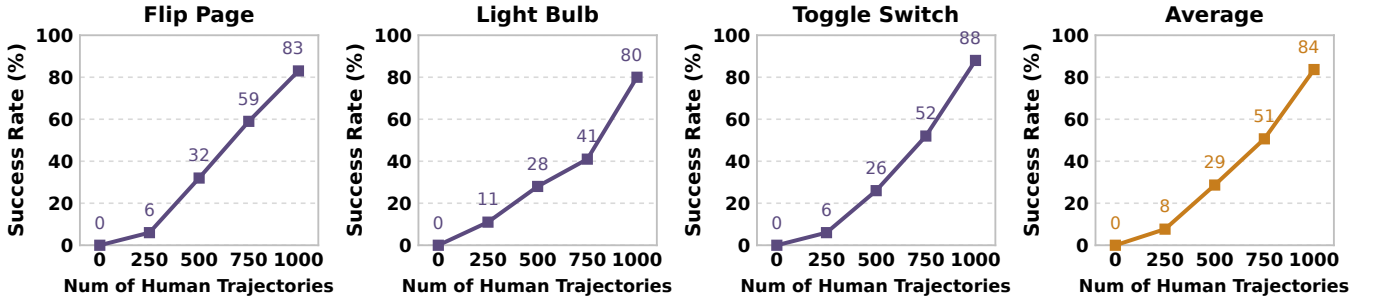


Fig. 5. **Human Tactile Data Scaling Curve.** Success rates on three contact-rich tasks and their average as we add human demonstrations to a fixed budget of robot demonstrations. Performance keeps improving with more human data and has not saturated at 1,000 trajectories.

that future tactile supervision is more than an auxiliary output: it encourages the shared representation to encode the physical consequences of actions. Removing tactile input as well further reduces performance to 26%, confirming that current contact observations provide information unavailable from vision alone. Finally, adding proprioceptive conditioning produces the largest degradation, reaching only 18%. We observe that the proprioception-conditioned policy relies heavily on its current state when predicting actions and consequently produces behaviors resembling the replay of a nearly fixed demonstration trajectory. This reliance limits its ability to adapt actions in response to changing visual and tactile observations, supporting our choice to condition both prediction heads on the unified visual-tactile representation.

VI. CONCLUSION

We presented UVTA, a framework for leveraging scalable human tactile demonstrations in contact-rich dexterous manipulation. Motivated by the premise that hand embodiments may differ while the underlying physics of interaction remains shared, we developed a wearable tactile motion-capture system that synchronously records wrist-centric vision, fingertip tactile measurements, and hand motion. Using this system, we collected a dataset spanning five tasks, with 1,000 diverse human demonstrations and 150 robot demonstrations per task. Our Unified Visual-Tactile-Action Model aligns the two embodiments and jointly predicts future action

and tactile trajectories, allowing human data to supervise contact-aware representation learning while retaining robot-specific execution. Across five real-robot tasks, the resulting policy achieves a 70% average success rate, compared with 29% for the strongest visual-tactile baseline. Ablations further show that removing future-tactile prediction reduces performance to 42%, confirming the importance of modeling the physical consequences of actions. Finally, performance improves consistently as human data scales and remains unsaturated at 1,000 demonstrations per task. These results establish human tactile interaction as a practical and scalable source of physical supervision for dexterous robot learning.

REFERENCES

- [1] F. R. Hogan, J. Ballester, S. Dong, and A. Rodriguez, "Tactile dexterity: Manipulation primitives with tactile feedback," *arXiv preprint arXiv:2002.03236*, 2020.
- [2] I. Guzey *et al.*, "See to touch: Learning tactile dexterity through visual incentives," *arXiv preprint arXiv:2309.12300*, 2023.
- [3] X. Zhang *et al.*, "DexTac: Learning contact-aware visuotactile policies via hand-by-hand teaching," *arXiv preprint arXiv:2601.21474*, 2026.
- [4] D. Niu *et al.*, "T-Rex: Tactile-reactive dexterous manipulation," *arXiv preprint arXiv:2606.17055*, 2026.
- [5] C. Zhang *et al.*, "Human-centric transferable tactile pre-training for dexterous robotic manipulation," *arXiv preprint arXiv:2607.01067*, 2026.
- [6] K. Grauman *et al.*, "Ego4D: Around the world in 3,000 hours of egocentric video," *arXiv preprint arXiv:2110.07058*, 2021.
- [7] Y. Liu *et al.*, "HOI4D: A 4d egocentric dataset for category-level human-object interaction," *arXiv preprint arXiv:2203.01577*, 2022.

- [8] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv preprint arXiv:2203.06173*, 2022.
- [9] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.
- [10] C. Wang *et al.*, "MimicPlay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.
- [11] Y. Qin *et al.*, "DexMV: Imitation learning for dexterous manipulation from human videos," *arXiv preprint arXiv:2108.05877*, 2021.
- [12] C. Wang *et al.*, "DexCap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.
- [13] S. Kareer *et al.*, "EgoMimic: Scaling imitation learning via egocentric video," *arXiv preprint arXiv:2410.24221*, 2024.
- [14] J. Mu *et al.*, "DexImit: Learning bimanual dexterous manipulation from monocular human videos," *arXiv preprint arXiv:2602.10105*, 2026.
- [15] R. Yang *et al.*, "EgoVLA: Learning vision-language-action models from egocentric human videos," *arXiv preprint arXiv:2507.12440*, 2025.
- [16] R. Zheng *et al.*, "EgoScale: Scaling dexterous manipulation with diverse egocentric human data," *arXiv preprint arXiv:2602.16710*, 2026.
- [17] Y. Du *et al.*, "Learning universal policies via text-guided video generation," *arXiv preprint arXiv:2302.00111*, 2023.
- [18] H. Wu *et al.*, "Unleashing large-scale video generative pre-training for visual robot manipulation," *arXiv preprint arXiv:2312.13139*, 2023.
- [19] S. Zhou *et al.*, "RoboDreamer: Learning compositional world models for robot imagination," *arXiv preprint arXiv:2404.12377*, 2024.
- [20] Y. Hu *et al.*, "Video prediction policy: A generalist robot policy with predictive visual representations," *arXiv preprint arXiv:2412.14803*, 2024.
- [21] S. Li, Y. Gao, D. Sadigh, and S. Song, "Unified video action model," *arXiv preprint arXiv:2503.00200*, 2025.
- [22] R. Zheng *et al.*, "FLARE: Robot learning with implicit world modeling," *arXiv preprint arXiv:2505.15659*, 2025.
- [23] S. Gao *et al.*, "DreamDojo: A generalist robot world model from large-scale human videos," *arXiv preprint arXiv:2602.06949*, 2026.
- [24] S. Ye *et al.*, "World action models are zero-shot policies," *arXiv preprint arXiv:2602.15922*, 2026.
- [25] B. Li, X. Yin, M. Lin, Y. Zhang, and D. Xu, "EgoWAM: World action models beyond pixels with in-the-wild egocentric human data," *arXiv preprint arXiv:2607.08436*, 2026.
- [26] A. Ayad, A. Röfer, N. Heppert, and A. Valada, "Imagine2touch: Predictive tactile sensing for robotic manipulation using efficient low-dimensional signals," *arXiv preprint arXiv:2405.01192*, 2024.
- [27] J. Zhou *et al.*, "TouchWorld: A predictive and reactive tactile foundation model for dexterous manipulation," *arXiv preprint arXiv:2607.07287*, 2026.
- [28] L. Heng, H. Geng, K. Zhang, P. Abbeel, and J. Malik, "ViTacFormer: Learning cross-modal representation for visuo-tactile dexterous manipulation," *arXiv preprint arXiv:2506.15953*, 2025.
- [29] H. Yuan *et al.*, "VTAM: Video-tactile-action models for complex physical interaction beyond VLAs," *arXiv preprint arXiv:2603.23481*, 2026.
- [30] D. Damen *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100," *International Journal of Computer Vision*, vol. 130, pp. 33–55, 2022.
- [31] K. Grauman *et al.*, "Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [32] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "VIP: Towards universal visual reward and representation via value-implicit pre-training," in *International Conference on Learning Representations*, 2023.
- [33] A. Majumdar *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?," in *Advances in Neural Information Processing Systems*, 2023.
- [34] Z. Wang, B. He, K. Yu, S. Lee, R. Gao, F. Huang, and Y. Aloimonos, "HumanEgo: Zero-shot robot learning from minutes of human egocentric videos," *arXiv preprint arXiv:2605.24934*, 2026.
- [35] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?," in *Conference on Robot Learning*, 2017.
- [36] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *IEEE International Conference on Robotics and Automation*, 2019.
- [37] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens, "Touch and go: Learning from human-collected vision and touch," in *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [38] I. Guzey, B. Evans, S. Chintala, and L. Pinto, "Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play," *arXiv preprint arXiv:2303.12076*, 2023.
- [39] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel, "The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning," *arXiv preprint arXiv:2311.00924*, 2023.
- [40] C. Higuera *et al.*, "Sparsh: Self-supervised touch representations for vision-based tactile sensing," *arXiv preprint arXiv:2410.24090*, 2024.
- [41] R. Feng *et al.*, "AnyTouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors," *arXiv preprint arXiv:2502.12191*, 2025.
- [42] T. Lin *et al.*, "Learning visuotactile skills with two multifingered hands," *arXiv preprint arXiv:2404.16823*, 2024.
- [43] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu, "Reactive Diffusion Policy: Slow-Fast visual-tactile policy learning for contact-rich manipulation," in *Proceedings of Robotics: Science and Systems*, 2025.
- [44] J. Huang *et al.*, "Tactile-VLA: Unlocking vision-language-action model's physical knowledge for tactile generalization," *arXiv preprint arXiv:2507.09160*, 2025.
- [45] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.
- [46] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International Conference on Machine Learning*, 2019.
- [47] P. Wu, A. Escontrela, D. Hafner, K. Goldberg, and P. Abbeel, "Day-Dreamer: World models for physical robot learning," *arXiv preprint arXiv:2206.14176*, 2022.
- [48] N. Hansen, H. Su, and X. Wang, "TD-MPC2: Scalable, robust world models for continuous control," in *International Conference on Learning Representations*, 2024.

APPENDIX I ADDITIONAL SYSTEM AND DATA DETAILS

A. Cross-Embodiment Calibration

The human and robot streams are aligned at both the observation and action levels. For vision, the wrist-camera adapter on the motion-capture system is calibrated so that its camera-to-wrist transformation matches the camera-to-end-effector transformation used on the robot. For motion, one VIVE tracker defines the table reference frame and a second tracker measures the human wrist pose. The calibrated tracker-to-wrist transformation converts these measurements into wrist poses, while constrained inverse kinematics maps the passive exoskeleton measurements to the 22 joint angles of the Sharpa Wave hand. Wrist targets are then expressed relative to the current wrist frame using 3-D translation and a continuous 6-D rotation representation. This construction avoids requiring the human and robot to share a global workspace.

B. Tactile Processing and Synchronization

Each human fingertip contains four FSR sensing regions. The five fingertips therefore produce the same 20-D tactile layout used by VTPM for robot observations. All image, tactile, wrist-pose, and hand-motion streams are timestamped with a shared host clock and resampled to 30 Hz. Before normalization, tactile readings are baseline corrected independently for each recording. We subtract the mean of the first five frames for human recordings and the first frame for robot recordings, reflecting the different resting-noise characteristics of the two sensors. Human and robot tactile channels are subsequently normalized using embodiment-specific statistics.

C. Dataset Composition

The complete dataset contains 5,750 demonstrations across five tasks: 5,000 human demonstrations and 750 robot demonstrations. Robot data provide executable commands on the target platform, whereas human data expand the variation in object instances, backgrounds, hand appearance, motion, and contact patterns. Every training sample contains a synchronized wrist image, a 20-D tactile token, and a 31-D motion target. Human targets supervise representation learning in the unified action space but are never issued directly as robot commands.

APPENDIX II IMPLEMENTATION DETAILS

Table 2 summarizes the model and optimization settings used in all experiments. We train one policy per task. The weighted sampler assigns equal probability mass to the human and robot datasets and samples with replacement, preventing the larger human dataset from determining the training mixture. Min-max statistics are computed separately for the two embodiments; at deployment, only the robot statistics are used to decode executable actions. Exponential moving average weights are used for inference.

The action loss first averages errors within the wrist and hand groups and then weights the two group means equally. This prevents the 22-D finger target from dominating the 9-D wrist target solely because it contains more coordinates. The tactile head regresses the complete 16×20 future trajectory from the noise-free interaction representation. Its loss updates both the head and the shared visual encoder, but its prediction is not passed into the diffusion trajectory or the controller.

APPENDIX III ROLLOUT AND EVALUATION DETAILS

A. Closed-Loop Execution

At every policy update, UVTA consumes the current wrist image and tactile token, samples a 16-step robot action chunk with 16 DDIM reverse steps, and denormalizes the chunk with the robot statistics. Relative wrist predictions are anchored to the

Table 2: **Implementation Details.** Settings shared by all task-specific UVTA policies.

Component	Setting
Visual observation	224×224 wrist RGB
Visual encoder	ViT-S/8, trained from scratch
Visual / tactile token	384-D / 20-D
Fused representation	404-D
Prediction horizon	16 steps
Action per step	9-D wrist + 22-D hand
Diffusion U-Net widths	256, 512, 1024
Tactile head	$404 \rightarrow 512 \rightarrow 320$, Mish
Training diffusion steps	50
DDIM reverse steps	16
Executed waypoints	First 8 of each 16-step chunk
Human-robot sampling	Equal probability mass
Tactile-loss weight λ	0.2
Optimizer	AdamW
Batch size / epochs	250 / 300

live wrist pose, while finger predictions are interpreted as absolute joint targets. The controller executes the first eight wrist and hand waypoints, then acquires a new visual-tactile observation and replans. This receding-horizon procedure preserves feedback during contact while avoiding a full diffusion sample at every low-level control step.

B. Stage-Wise Success Rate

Each method and ablation is evaluated in $N = 10$ real-robot rollouts per task, with randomized object positions and heights. Table 3 lists the complete scoring criteria, and Fig. 6 illustrates representative successful rollouts. The reported success rate is a stage-wise score that gives partial credit for task progress, rather than the fraction of fully successful trials.

For task q and rollout j , let $w_{q,m}$ be the score assigned to milestone m , and let $C_{q,j,m}$ indicate whether that milestone is achieved. We assign a single rollout score

$$s_{q,j} = \max(\{0\} \cup \{w_{q,m} \mid C_{q,j,m} = 1\}). \quad (10)$$

Thus, a rollout receives 0 if it achieves no listed milestone and 1 only if it fully completes the task. Milestone scores are absolute values, not additive rewards. For example, turning multiple pages receives 0.7, while turning exactly one page receives 1.0.

The task-level percentage and the five-task average are

$$\text{SR}_q = \frac{100}{N} \sum_{j=1}^N s_{q,j}, \quad \overline{\text{SR}} = \frac{1}{5} \sum_{q=1}^5 \text{SR}_q. \quad (11)$$

The final column of Table 1 reports this unweighted five-task mean, rounded to the nearest integer percentage. These scores should not be interpreted as binary completion rates.

C. Human-Data Scaling Protocol

For the human-data scaling study, the robot dataset is fixed at 150 demonstrations per task while the human subset is increased from 0 to 250, 500, 750, and 1,000 demonstrations. The study covers Flip Page, Screw Light Bulb, and Toggle Switch, and uses the same policy architecture, training objective, and evaluation procedure at every data scale. Its average is computed over these three tasks using the same stage-wise scoring rule.

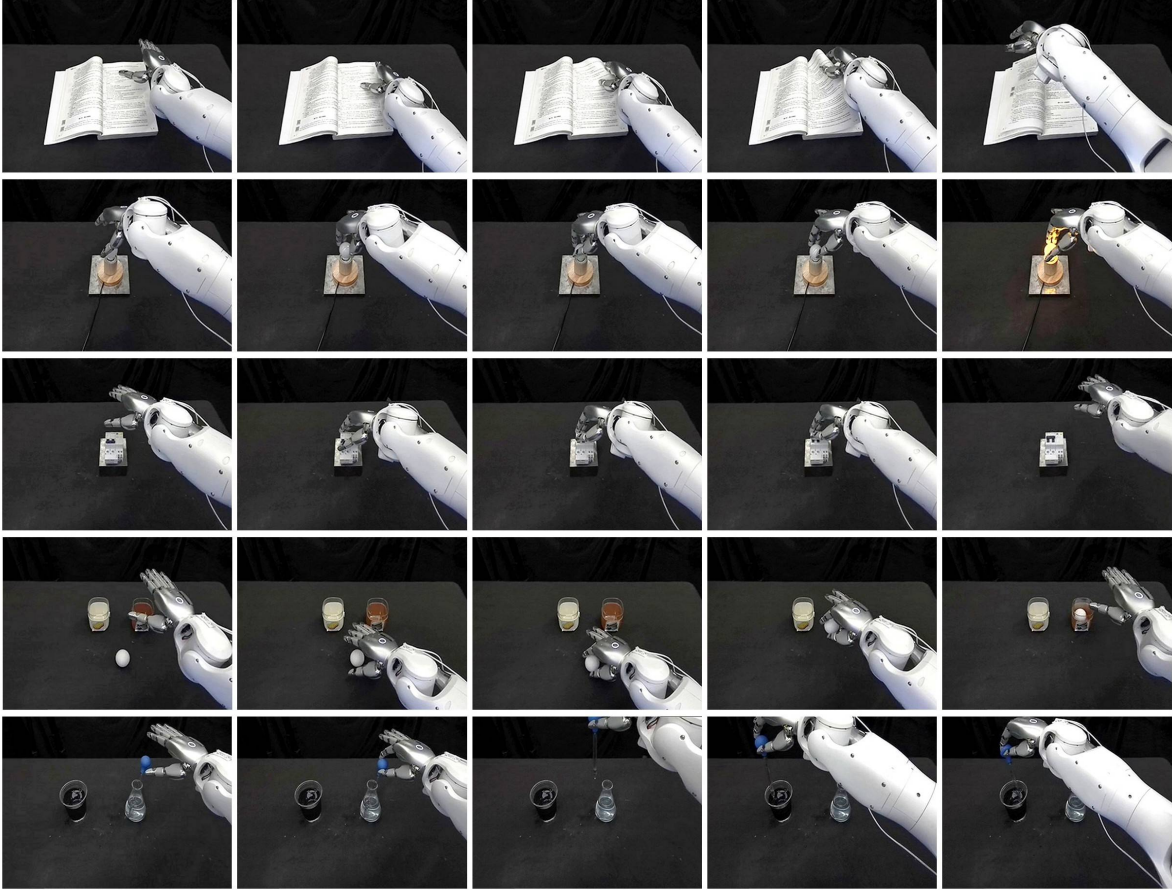


Fig. 6. **Real-Robot Rollout Trajectories.** Representative successful executions of UVTA on five contact-rich tasks. From top to bottom: Flip Page, Screw Light Bulb, Toggle Switch, Ball Classification, and Liquid Transfer with a Dropper. Each row shows an execution sequence, with time progressing from left to right.

Table 3: **Task-Specific Evaluation Criteria.** A rollout receives the highest score associated with a milestone it achieves; scores are not summed across milestones. A rollout that achieves none of the listed milestones receives 0. A score of 1.0 denotes full task success.

Task	Milestones and scores
Flip Page	0.3: separate and lift the page edge; 0.7: turn multiple pages together; 1.0: turn exactly one page.
Screw Light Bulb	0.2: establish stable contact with the bulb; 0.5: rotate the bulb through one full turn; 1.0: complete multiple turns until the bulb remains lit.
Toggle Switch	0.2: establish stable contact; 0.6: partially actuate the switch; 1.0: push the switch fully to its on position.
Ball Classification	0.2: securely grasp the ball; 0.6: place it in a container; 1.0: place it in the correct container for its material.
Liquid Transfer	0.2: establish stable contact with the dropper; 0.5: draw liquid into the dropper; 0.7: lift and transport the filled dropper; 1.0: dispense the liquid into the target container.

APPENDIX IV LIMITATIONS AND FUTURE WORK

Fingertip-only sensing. The current tactile motion-capture system measures contact only at the fingertips. This design is motivated by an unresolved sensing limitation: whole-hand tactile gloves available for our data-collection setting cannot yet reliably suppress motion-induced noise during large-range hand movements. Such artifacts make it difficult to distinguish object-contact signals from signals caused by the motion of the glove itself. We therefore restrict sensing to the fingertips to obtain more reliable contact supervision. This choice leaves palm and dorsal-hand contacts unobserved and limits the interaction patterns represented in the dataset. Extending coverage will require improved sensor integration and compensation

for motion-induced artifacts.

Data and model scale. The experiments use five laboratory tasks, a single robot platform, task-specific policies, at most 1,000 human demonstrations per task, and one model scale. Consequently, the present results do not establish how UVTA behaves with substantially larger and more diverse human datasets, larger models, or new robot embodiments. Future work will scale both data and model capacity and investigate multi-task and cross-platform policies alongside more extensive tactile coverage.