MANIPULATION

Visual-tactile pretraining and online multitask learning for humanlike manipulation dexterity

Qi Ye[1]*†, Qingtao Liu[1]†, Siyun Wang[1], Jiaying Chen[1], Yu Cui[1], Ke Jin[1], Huajin Chen[1], Xuan Cai[1], Gaofeng Li[1], Jiming Chen[1,2]*

Achieving humanlike dexterity with anthropomorphic multifingered robotic hands requires precise finger coordination. However, dexterous manipulation remains highly challenging because of high-dimensional action-observation spaces, complex hand-object contact dynamics, and frequent occlusions. To address this, we drew inspiration from the human learning paradigm of observation and practice and propose a two-stage learning framework by learning visual-tactile integration representations via self-supervised learning from human demonstrations. We trained a unified multitask policy through reinforcement learning and online imitation learning. This decoupled learning enabled the robot to acquire generalizable manipulation skills using only monocular images and simple binary tactile signals. With the unified policy, we built a multifingered hand manipulation system that performs multiple complicated tasks with low-cost sensing. It achieved an 85% success rate across five complex tasks and 25 objects and further generalized to three unseen tasks that share similar hand-object coordination patterns with the training tasks.

INTRODUCTION

Human hands have remarkable dexterity, enabling a wide range of complex manipulation tasks—such as rotating a bottle cap, sliding a lever, or reorienting an object in the palm—with speed, precision, and fluid coordination. Despite substantial advances in robotic manipulation, dexterous hands still fall short in performing such tasks with comparable versatility and control.

A core challenge lies in the need to coordinate multiple fingers in a high-dimensional action space while managing dynamic, contact-rich interactions with objects (*1*–*6*). Tasks that feel effortless to humans—such as object rotation—require robotic hands to precisely control many actuated joints and to determine when and where to establish contact. Even minor misalignments between fingers can disrupt the manipulation process, causing the object to slip or deviate from the intended trajectory. These subtle dynamics make dexterous manipulation particularly difficult to learn and generalize in robotic systems.

A variety of control strategies have been explored to address this challenge. Among them, model-based control frameworks (*7*–*12*) attempt to generate contact-rich behaviors by relying on accurate dynamics models. However, this reliance on analytical modeling limits their applicability to real-world scenarios with complex dynamics and contact uncertainty. To overcome these limitations, deep model-free reinforcement learning (RL) has been explored (*13*–*18*), enabling policy learning through trial-and-error interactions. However, such learning demands large samples, and the high-dimensional action space of dexterous hands further exacerbates poor sample efficiency and training instability. To improve efficiency, many works have incorporated demonstrations via imitation learning (*19*–*22*), leveraged goal-conditioned objectives or affordance priors to guide exploration (*23*–*26*), or refined reference trajectories through residual learning (*27*–*29*). Although effective, these methods often rely on task-specific demonstrations, goals, or references, which require precise robotic data collection (especially the action of a robotic hand for each state). The difficulty of collecting diverse demonstrations for various and precise dexterous robotic hands challenges the learning of a general-purpose dexterous manipulation policy.

In addition to complex control, dexterous manipulation presents substantial perceptual challenges. The observation of the input state space for manipulation is high dimensional, involving diverse combinations of object shapes, textures, hand-object displacements, contact patterns, and lighting conditions. Although simulation environments provide privileged state information (e.g., object pose, velocity, contact points) to facilitate policy learning, such information is often unavailable in the real world because of sensing limitations and frequent occlusions from articulated fingers. To bridge this gap, prior works (*1*, *30*, *31*) have distilled a state-based policy trained in simulation into a vision-based or visual-tactile policy. Although effective, this inevitably introduces information loss when mapping partial and noisy observations to full states, limiting the performance of the student policy. Another line of work (*32*–*34*) used multicamera systems to enhance observability, but such setups increase cost and complexity, and vision alone remains insufficient for recovering occluded contact states. Humans naturally integrate visual and tactile feedback for precise manipulation, inspiring the use of tactile sensing as a complementary modality in robotic systems (*30*, *31*, *35*–*43*).

Despite advances in learning algorithms and sensing techniques, building a real-world dexterous manipulation system with a unified policy on multifingered hands remains challenging. A common solution is to use teleoperated demonstrations with full trajectories with observation-action pairs for imitation learning or RL (*44*–*50*). Although effective, such methods depend on sophisticated teleoperation platforms that are difficult to scale for contact-rich dexterous tasks. To reduce the difficulties of data collection, recent work has explored learning from human video demonstrations (*51*–*55*), pretrained visual encoders to acquire task-agnostic priors. Although

[1]College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China. [2]School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China.
†These authors contributed equally to this work.
*Corresponding author. Email: qi.ye@zju.edu.cn (Q.Y.); cjm@zju.edu.cn (J.C.)

scalable, these methods have mainly been validated on simple gripper-based manipulation and struggle with multifingered hands because of frequent occlusions and complex contact dynamics that vision alone cannot resolve.

Tactile sensing complements vision by providing fine-grained contact information. In gripper-based manipulation, combining vision with high-resolution tactile sensors enables accurate force control, especially when dealing with soft or deformable objects (*36*, *39*, *40*). However, these sensors are typically bulky and hard to integrate into multifingered hands, which compromise dexterity. To simplify system design and alleviate the sim-to-real gap, recent studies (*31*, *56*, *57*) represent tactile input as sparse contact events (touch/no-touch) and show that even this coarse signal, when concatenated with vision, benefits RL of manipulation skills. Still, mastering complex dexterous tasks like in-hand orientation with raw sensory input remains difficult; thus, prior work often adopts the framework of a state-based teacher and a visual-tactile student (*30*, *31*) despite the issue of information loss. Although these pipelines improve performance, they are usually limited to single-skill learning because of the training difficulty. Meanwhile, as aforementioned, pretraining on human video demonstrations (*51*–*55*) can provide strong perceptual representations transferable to diverse downstream tasks. Yet, little attention has been paid to how human visuotactile demonstrations—where visual observations are naturally paired with tactile cues experienced by human hands—can be exploited for learning multiple complex manipulation skills. Motivated by this, we explore whether pairing monocular RGB (red, green, and blue) inputs with sparse binary tactile cues from human video demonstrations can capture meaningful interaction patterns and whether these representations and interaction patterns can enable robots to acquire diverse dexterous skills using a lightweight sensing setup.

To achieve this goal, a key challenge lies in how to effectively integrate multisensory information and extract useful priors from human video demonstrations. Existing methods for visual-tactile manipulation typically process each modality independently and fuse them later during policy learning (*30*, *57*, *58*). Although this approach may suffice for simple manipulators, the high-dimensional action space of multifingered hands results in sparse rewards, making the joint learning of multimodal perception and action policies inefficient and unstable. Neuroscience studies suggest that in the human brain, separate areas of the cortex are developed for sensing and control, and there are neurons in a certain region of the human brain that integrate visual and tactile cues for our hand (*59*, *60*). In addition, "inferior parietal lobule (IPL) neurons provide representation of actions with multisensory information. These neurons fire during the observation of an act, before the beginning of the subsequent acts specifying the action" (*61*). Given the sparse supervision of learning via interactions and inspired by the findings of neuroscience, we propose to decouple the learning of the multisensory integration and action control policy: The integration module is implemented as a visual-tactile encoder trained via self-supervised learning on large-scale human demonstration videos, whereas the action control policy is learned with environment interactions.

In computer vision, a masked autoencoder (MAE) (*62*, *63*) is designed to encode the inherent patterns in images by forcing a neural network to recover masked image patches of the input. Building on this, we propose to align modality-specific tokens from vision and touch through cross-modal attention and masked input recovery and to integrate them via a dedicated integration token by observing a large corpus of human demonstration videos collected with tactile gloves (*64*). The integration token is conceptually analogous to neurons in the IPL of the human brain, which integrate multisensory information (Fig. 1A). Through cross-modal supervision, it encourages the emergence of integration units that encode contact-relevant cues (when and where contact occurs) despite occlusion or perceptual noise. Consequently, the network captures a compact, low-dimensional manifold of task-relevant features in the otherwise high-dimensional observation space (*65*–*67*), allowing for more efficient and generalizable policy learning. Although prior work (*64*) also adopted an MAE for the modality alignment, it did not learn to integrate the modalities because of the absence of a dedicated IPL token and encountered the issue of ineffective learning because of the challenge of modality integration via sparse rewards during RL.

Building on the learned perceptual representations that integrate RGB and tactile events, the next challenge was to train a unified control policy capable of handling multiple manipulation tasks. Multitask learning has been extensively explored in robotics (*68*–*72*), showing that parameter sharing across tasks can improve sample efficiency and generalization. In manipulation domains, recent work has scaled policy learning across tasks using imitation learning (*47*, *73*, *74*) and RL (*68*, *69*), typically with vision-based inputs and parallel grippers. However, these approaches face limitations when applied to dexterous hands. Imitation learning directly imitates the robotic hand actions given observations but is highly sensitive to small observation or action errors; in multifingered settings, such deviations compound quickly, leading to unstable contacts and task failure. RL, in contrast, requires task-specific reward engineering, which becomes increasingly difficult across heterogeneous objectives such as orientation, force, or position. The resulting misalignment in reward magnitudes and optimization landscapes often causes instability and low sample efficiency. To address these issues, we adopted an online imitation learning strategy inspired by Ross *et al.* (*75*). Instead of training on offline trajectories generated by expert policies, we iteratively collected states visited by the unified policy during learning and queried the corresponding expert policies for supervision. This online aggregation process aligns the observation distributions of the student and expert policies, reducing compounding errors and enabling stable multitask learning in a single control policy.

Using the proposed method, we trained a unified policy capable of performing multiple dexterous manipulation tasks that require fine-grained, coordinated finger movements, relying solely on monocular RGB images and binary tactile events. We deployed this policy on a Shadow Hand platform (*76*). In contrast with many existing systems that depend on expensive depth cameras and (or) high-precision optical tactile sensors—often costing thousands of dollars—our setup requires only a standard webcam and low-cost piezoresistive tactile sensors, with a total cost of ~$250. Despite the hardware simplicity, the trained policy successfully executed five complex manipulation tasks on 25 different objects in the real world, achieving an average success rate of ~85%. Furthermore, the policy generalized effectively to three unseen tasks that shared similar hand-object coordination patterns with the training set and to different tactile sensor types and challenging lighting conditions. Beyond task success, our experimental results show that policies pretrained with human demonstrations and tactile events produce contact behaviors that are more human-like than those trained with vision alone.

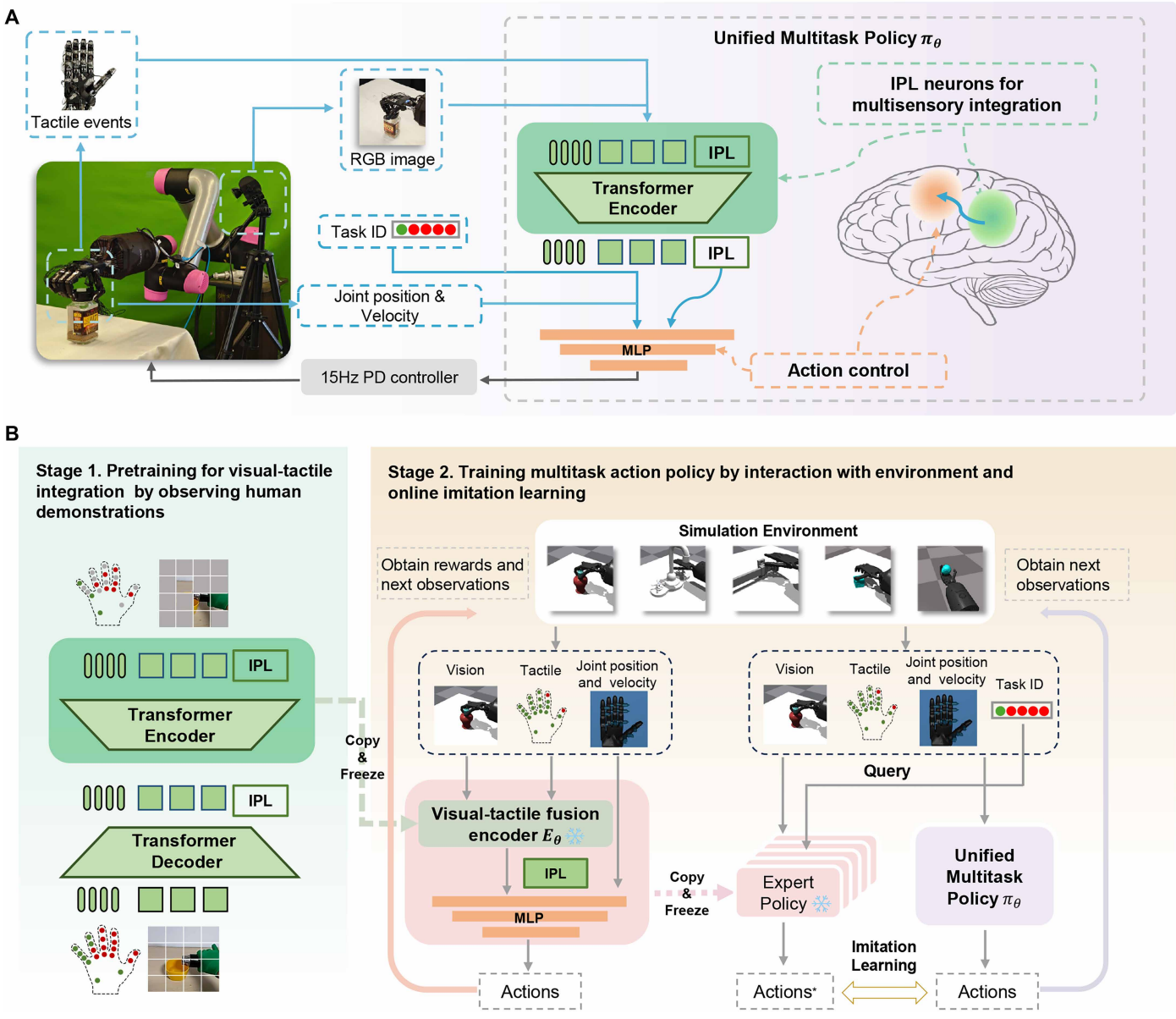

Fig. 1. Illustration of the real-world robot system and the full learning pipeline of the system. (**A**) Our multitask policy makes action decisions with two steps—multisensory integration and action control—resembling our human brain, which initially integrates multisensory information in IPL neurons and then produces signals to execute the actions in the motor cortex. In the first step, the policy integrates egocentric RGB images and tactile events into a fused representation. The representation, with the proprioception information and task identifier (ID), is then passed to an MLP for action prediction. (**B**) Multisensory integration (stage 1) is learned by observing human demonstrations. Given the integrated representation, the multitask action policy (stage 2) is learned by RL for expert policies for different tasks and distillation of the experts into a unified policy by online imitation learning.

RESULTS

Method overview and real-world system

This work addressed the challenge of acquiring multiple dexterous manipulation skills in a unified control policy using only simple sensing modalities—RGB vision and tactile feedback. An overview of the proposed framework is illustrated in Fig. 1B. The method consisted of two key stages: visual-tactile representation pretraining from human demonstrations, followed by online multitask policy learning through interaction with the environment combined with imitation learning. The unified policy was capable of switching between different tasks by conditioning on a task-specific input identifier.

As shown in Fig. 1B, the unified policy was completely trained in simulation, using visual, tactile, and proprioceptive cues, along with a task-specific one-hot encoded identifier to guide decision-making. For real-world deployment, we constructed a physical system comprising a five-fingered Shadow Hand (*76*), a monocular RGB web camera, and a custom tactile sensing system. The Shadow Hand was mounted on a robotic arm and equipped with 20 piezoresistive tactile sensors distributed across different hand surfaces, each with a

resolution of 1 pixel by 1 pixel. The camera was positioned to replicate the egocentric viewpoint used in the simulation. These hardware components and action policies communicated using Robot Operating System (ROS). During deployment, the unified policy received multimodal real-time inputs and produced control actions at 15 Hz. The entire system ran in real time on a standard laptop equipped with an Intel i9-12900K processor and an NVIDIA GeForce RTX 4070 GPU. We further built a visual-tactile manipulation platform based on an open-source four-fingered robotic hand to demonstrate the applicability of our method to low-cost robotic hands, which is described in the "LEAP Hand visual-tactile manipulation system and experiments" section in the Supplementary Materials.

Manipulation tasks

We conducted experiments on five tasks in simulation and eight tasks in the real world. Among the real-world experiments, five tasks—bottle cap turning, faucet screwing, lever sliding, tabletop reorientation, and in-hand reorientation—were included in the training set during simulation, whereas three tasks—pencil sharpening, screw unfastening, and snack sleeve sliding—were used to evaluate generalization. These unseen tasks shared similar hand-object coordination patterns with the seen ones but differed in key aspects. Detailed specifications of all tasks—including hand and object configurations, differences across tasks, and success criteria—are provided in the "Task specifications" section in the Supplementary Materials.

Figure 2 shows the intermediate frames capturing all tasks performed by the robot system with our unified policy. Movie S1 provides video recordings of these tasks and shows demonstrations using tactile sensors with different resolutions and principles as well as experiments under varying lighting conditions.

Experimental protocols and evaluation metrics

All policies were trained in simulation using 40 objects. For real-world evaluation on each seen task, we used five physical objects: three three-dimensional (3D) printed replicas of training objects (in-distribution, Fig. 3A) and two household items with novel shapes, materials, or textures (out-of-distribution, Fig. 3B). The 3D printed objects shared the same geometry as those in simulation but differed in physical properties such as color, weight, and friction, which were not explicitly modeled during training. The household objects further tested generalization to unfamiliar visual and material properties, including transparency and complex textures. Each object was tested in 10 trials with randomized initial positions and orientations. For unseen tasks, three household objects were selected to evaluate generalization beyond the distribution of training tasks, as shown in Fig. 3C.

For all experiments, we used success rate (%) as the primary evaluation metric. Following prior works (*70*, *77*–*79*), we set the maximum episode length to 600 steps for simulation training. A manipulation trial was considered successful if the predefined goal was achieved in a single episode. In the real world, a trial was considered successful if the task was completed within 40 s, which approximately corresponded to one simulation episode (the control frequency was 15 Hz in the real world and 60 Hz in simulation). Trials exceeding this time limit were considered failures. To ensure robustness and reproducibility, all simulation experiments were conducted with four different random seeds. Moreover, we used the success completion time (s) as a supplementary metric of system efficiency, computed as the average duration of successful trials. If no successful trials were observed, then the metric was set to 40 s. Given the article page limit, we refer readers to the Supplementary Materials ("Extended experimental protocols and additional analyses" section) for details on heterogeneous tactile sensor calibration, lighting variation setups, baseline implementations, and additional tactile pattern analyses.

Evaluation on real objects and new tasks

To demonstrate the versatility of our learned multitask control policy in the real world, experiments were conducted on both seen tasks (using 3D printed replicas and daily objects) and unseen tasks (using daily objects), as shown in Fig. 3 (A to C). The evaluation results are provided in Fig. 3 (D and E).

Deployment to real objects

Figure 3D(i) presents the success rate of our method in five manipulation tasks using in-distribution objects—3D printed shapes that were part of the simulation training set. In all tasks, which required intricate finger coordination, the system achieved an average success rate of ~87%. To evaluate generalization, we also tested our policy on out-of-distribution objects from daily life, as shown in Fig. 3D(ii). These objects varied considerably in shape, material (e.g., plastic bottles, metallic faucet handles, and soft fruits), and surface texture (e.g., reflective finishes, transparent materials, and complex visual patterns such as printed labels or jelly-like translucency). Despite these variations, our policy maintained robust performance, achieving an average success rate of 85%. This demonstrated strong generalization capabilities across both visual and tactile domains in real-world settings.

Deployment to unseen task settings

We further evaluated the policy in three task settings that were unseen during training. These tasks shared similar finger coordination with the five tasks used during training but differed in the coordination required for object stability control and subtle hand position adjustments. They also differed substantially in object composition and appearance.

The first two tasks, pencil sharpening and screw unfastening, shared key motion patterns with bottle cap turning, so the policy was conditioned on the bottle cap turning identifier during testing. As shown in Fig. 3D(iii), it achieved 9 of the 10 and 6 of the 10 successful trials for pencil sharpening and screw fastening, respectively, demonstrating generalization to new tasks, although with performance drops compared with bottle cap turning. This gap reflected differences in contact dynamics: Pencil sharpening required stable torque control to prevent slippage, given that the pencil provided only a narrow support point; screw unfastening required continuous finger height adjustment to maintain contact as the screw was gradually unfastened, in contrast with the minimal vertical displacement in bottle cap turning. For the third task, snack sleeve sliding, the policy was conditioned on the lever sliding identifier. Despite differences in geometry, material, and appearance, the policy successfully transferred the sliding behavior and achieved 8 of the 10 successful trials. These results highlighted the policy's ability to generalize to unseen tasks with related coordination patterns and showed that the degree of generalization depended on contact dynamics and hand-object displacement.

Applicability to different tactile sensors

To evaluate the applicability of our method across different tactile sensing modalities, we further tested on the Shadow Hand with three

Fig. 2. Snapshots of different tasks performed by our manipulation system. (**A**) to (**E**) show tasks used during training, including (A) bottle cap turning, (B) faucet screwing, (C) lever sliding, (D) tabletop reorientation, and (E) in-hand reorientation, whereas (**F**) to (**H**) present unseen tasks for evaluating generalization, including (F) pencil sharpening, (G) screw unfastening, and (H) snack sleeve sliding.

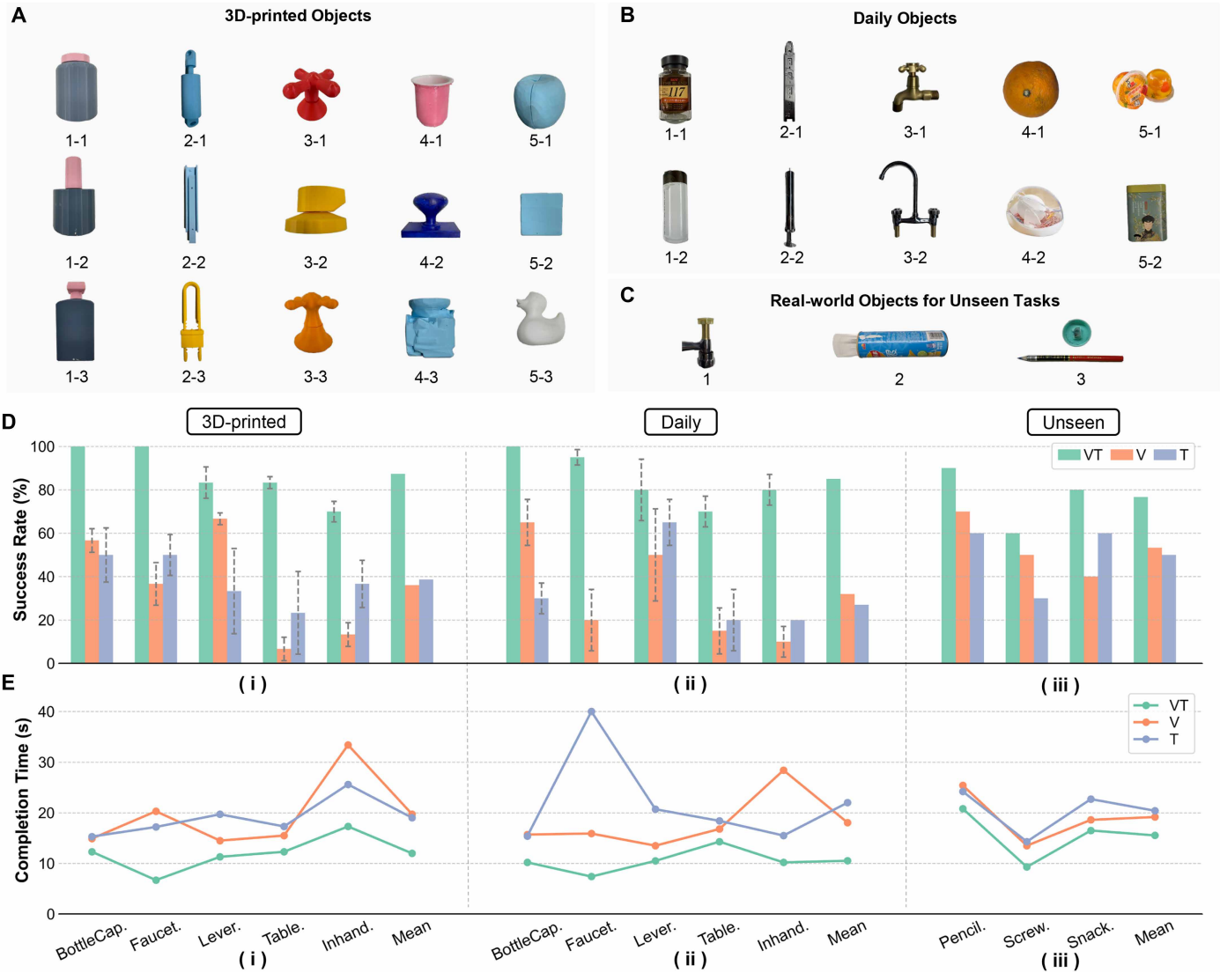

Fig. 3. Comparison of policies with different input modalities in real-world experiments. (**A** to **C**) Objects used for testing, including 3D-printed replicas of training objects, household objects with novel appearances, and three additional objects for unseen task settings. (**D**) Success rates of the unified policy with visual-tactile (VT), vision-only (V), and tactile-only (T) inputs on the three object sets in (A) to (C). Error bars denote mean ± SE across instances. (**E**) Completion time (s) of successful trials under the same settings.

alternative sensors (Fig. 4A): two piezoresistive arrays with different spatial resolutions (4 pixels by 1 pixel and 6 pixels by 4 pixels) and the built-in pressure and temperature sensors that measured fingertip air pressure. We evaluated different tactile sensor setups in the bottle cap turning task, and they all succeeded in 10 trials. Figure 4B also visualizes binary tactile signal sequences coupled with RGB images when the policy executed the actions with the RGB and tactile inputs. The success of our method in transferring across different tactile sensors was largely due to the use of binary tactile events, which simplified raw sensor signals into a shared contact/no-contact representation. Given that tactile sensors vary widely in sensing principles, resolution, and signal format, simulating them accurately is challenging. By focusing on binary events, which were easier to model and more consistent across sensors, we avoided these limitations. During training, we randomized the binarization threshold to further enhance robustness. This allowed our policy to generalize across diverse tactile sensors without requiring retraining or architectural modifications (see the "Details for heterogeneous tactile sensors" section in the Supplementary Materials for more details).

Robustness to challenging lighting conditions

To systematically evaluate robustness, we conducted simulation experiments where lighting direction and intensity were varied in a controlled manner: Lighting direction was modified by rotating the light source relative to its default orientation (i.e., varying the deviation angle), and lighting intensity was adjusted by scaling illumination relative to the default values. We evaluated policies with (VT) and without (V) tactile sensing on unseen objects. Representative conditions and the results in Fig. 4 (C and D) show that the VT policy maintained consistently high performance under all lighting variations, whereas the V policy degraded substantially, particularly under

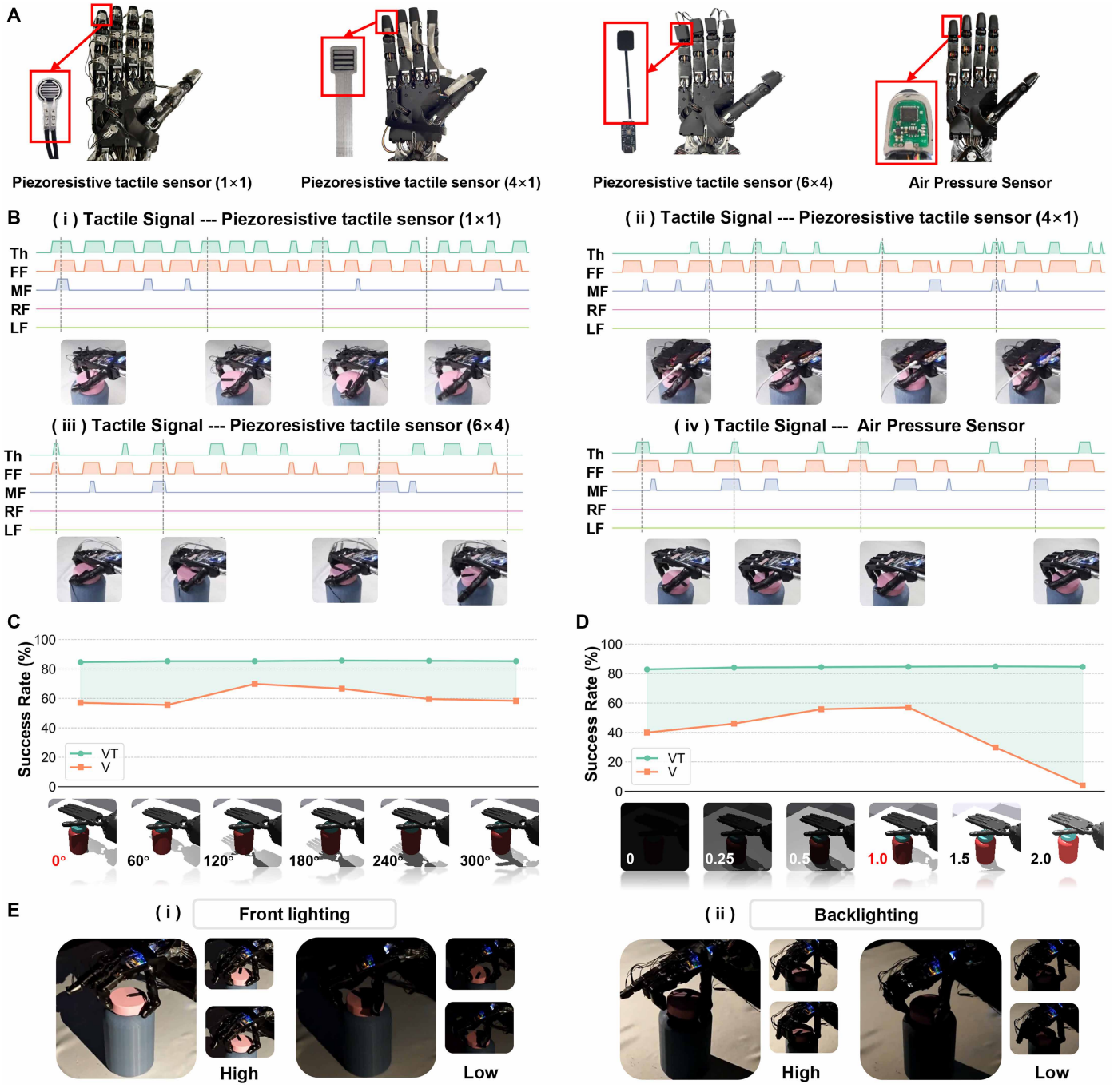

Fig. 4. Testing the unified policy under different tactile sensors and lighting conditions. (**A**) Four tactile sensors with varying resolutions and sensing principles. (**B**) Visualizations of fingertip-mounted signals for each finger (Th, thumb; MF, middle finger; RF, ring finger; LF, little finger) synchronized with egocentric camera images. The signals represent discrete contact events encoded as binary values (0 or 1). (**C** and **D**) Success rates on unseen objects across varying light directions and intensities. Red *x*-axis labels denote the default lighting configuration. (**E**) Representative snapshots of the bottle cap turning task under different illumination setups, including side lighting at high and low intensities.

extreme intensities. This highlighted the crucial role of tactile input in ensuring robustness when visual observations were unreliable. We further demonstrated real-world experiments under diverse natural illumination conditions (snapshots are shown in Fig. 4E and video examples in movie S1). More details are provided in the Supplementary Materials (the "Details for lighting variation experiments" section).

The effect of visual-tactile integration

To investigate the role of multisensory integration in multitask manipulation, we compared our proposed method with two ablation baselines that relied on a single sensory modality—either vision or tactile events. The two baselines were constructed by removing one modality from the input while keeping all other training procedures

identical to our full method. One baseline received only RGB images, and the other used only binarized tactile events as exteroceptive input.

Figure 5A shows the training curves of the methods. After convergence, our method achieved a success rate of greater than 80% across tasks on the training object set, whereas both single-modality baselines plateaued at less than 70%. Figure 5B further shows the results on identical unseen 3D-printed objects in both simulation and reality. For the unseen objects in simulation, the single modality baselines (V and T) achieved a success rate of about 60%, whereas in the real world, the success rate dropped to less than 40%. In contrast, as shown in Figs. 3D and 5 (A and B), the visual-tactile policy maintained a consistently high success rate of about 80% across seen and unseen task settings and objects, in both simulation and the real world.

These results highlighted the importance of visual-tactile integration for learning generalizable and robust manipulation policies. Monocular RGB input alone was vulnerable to occlusions, lighting variations, and texture ambiguity, whereas tactile signals could suffer from spatial sparsity and lack of global context. The pretraining stage, which leveraged human demonstrations, allowed the network to align and fuse complementary cues from both modalities. This integration enhanced the model's perception of contact-rich interactions and led to considerably better transfer across tasks, objects, and the real world compared with using a single modality (see the "Discussion on the sim-to-real degradation for single modalities" section in the Supplementary Materials for more discussion).

To evaluate system efficiency, Fig. 3E reports the average completion time of successful trials across real-world objects (Fig. 3, A to C). The results revealed three consistent observations. First, our VT

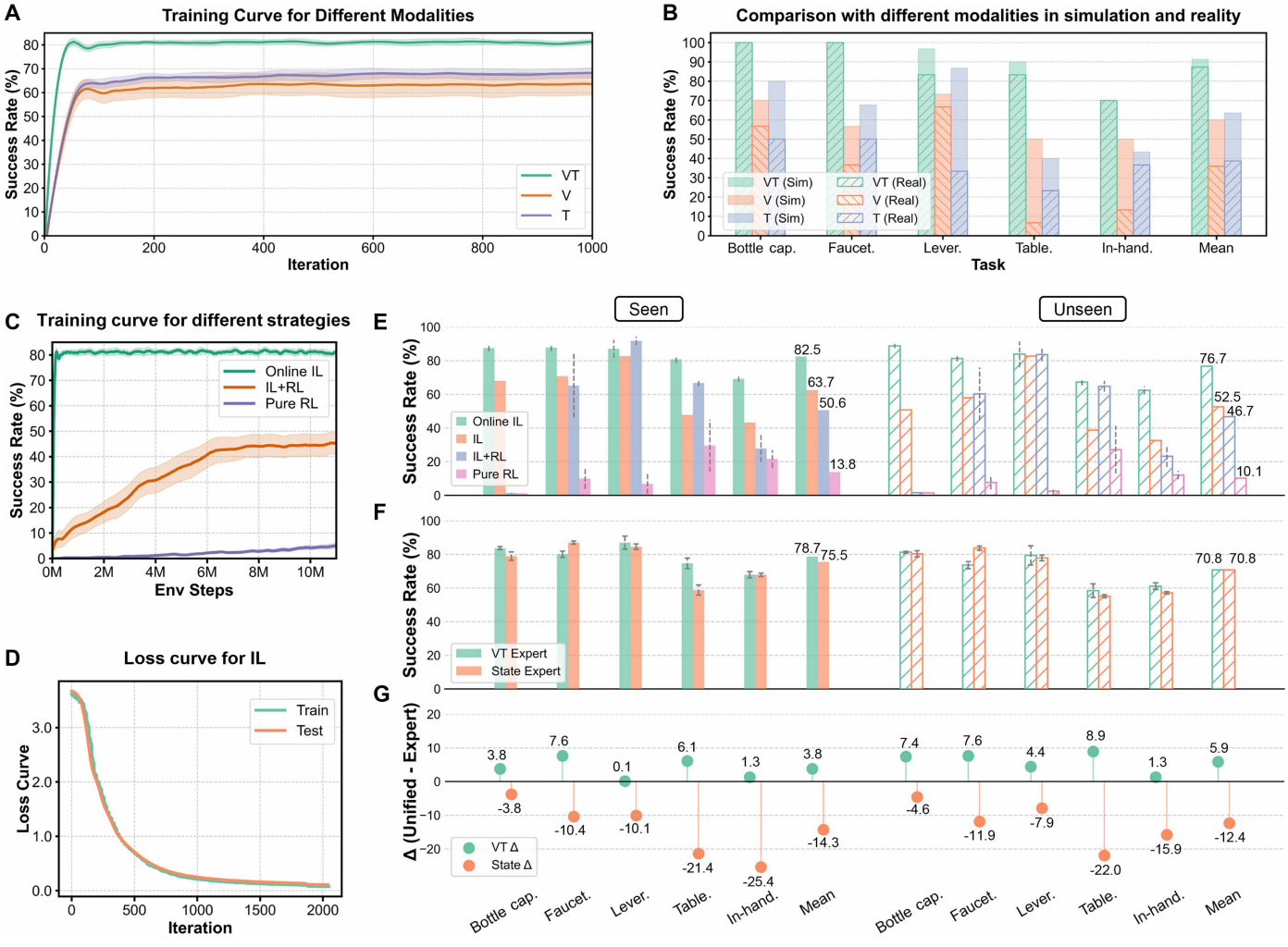

Fig. 5. Quantitative evaluation of sensor modalities and learning baselines. (**A**) and (**B**) compare our method with single modality baselines. (A) Training curves of the multitask policy with visual-tactile (VT), vision-only (V), and tactile-only (T) modalities, evaluated on the training object set in simulation. Shaded regions denote mean ± SE over four seeds. (B) Sim-to-real performance gap on 3D-printed objects, comparing VT with V and T modalities. (**C**) to (**G**) compare our method with existing multitask learning methods. (C) Training curve for unified policy training strategies: our online imitation learning (online IL) versus IL + RL and pure RL. (D) Training loss curves for the multitask imitation learning baseline (IL). (E) Success rates of online IL, IL + RL, pure RL, and IL, tested on seen and unseen simulation objects. (F) Comparison of our VT expert policy with state-based expert policies on seen and unseen simulation objects. (G) Distillation analysis for the two pipelines in (F). We plot the per-task difference in success rates (Δ = unified − expert). Positive values indicate that the unified policy improves upon the experts' success rates. The error bars in (E) and (F) represent the mean ± SE computed over different object instances for each task.

policy achieved the shortest and most stable completion times across all conditions, indicating both robustness and efficiency. Second, unimodal policies (vision-only or tactile-only) tended to be slower and less stable. In particular, the tactile-only policy failed completely on the faucet task, leading to the maximum capped time of 40 s, which highlighted the limitation of relying on a single modality. Third, in unseen tasks, the VT policy maintained both high success rates and low completion times, confirming that the learned visual-tactile representation enabled efficient generalization.

The effect of online multitask imitation learning

To validate the effectiveness of our online multitask imitation learning strategy, we compared it with three baseline methods adopted in state-of-the-art methods (*19*, *47*, *68*, *69*, *72*) for multitask learning: RL from scratch (pure RL), offline imitation learning from expert demonstrations (IL), and imitation learning followed by RL fine-tuning (IL + RL). Detailed implementation descriptions are provided in the Supplementary Materials ("Implementation of other baselines"). All baselines were trained under identical conditions as our approach, including the same input modalities, task identifiers, and network architectures, to ensure a fair comparison. The training curves of these baselines are illustrated in Fig. 5 (C and D).

As Fig. 5C shows, pure RL took millions of training steps to start improving. The success rates of the policy evaluated on seen and unseen objects for different tasks in simulation after ~16 million steps are depicted in Fig. 5C. The policy succeeded several times in tasks such as tabletop reorientation and in-hand reorientation, whereas it completely failed in the bottle cap–turning task. These results indicate that the difficulty of learning different tasks varied.

IL is supervised learning. In Fig. 5D, the multitask policy converged to an action loss close to zero, but the success rate of the policy was about 20% lower than our online IL. This effect arose from the accumulation of prediction errors over time in sequential decision-making, as discussed in the Introduction.

To mitigate divergence during execution, IL + RL added a RL stage after IL, as in (*19*, *21*, *22*). It achieved considerable improvements over pure RL but could not reach the success rates of expert policies. Although imitation learning for initialization was applied, the RL stage still faced the challenges of multitask RL mentioned above. In addition, action supervision during the imitation learning stage and rewards in the RL stage were defined by different aspects. Although rewards provided physical feedback, they potentially led to larger observation drift from expert policies.

Compared with these methods, Fig. 5E shows that our unified policy gained substantial improvements. In contrast with imitation learning that collected demonstrations from experts offline, we instead sampled the current multitask policy for observations during learning and queried the expert policies for actions for supervision. Given that the observation was acquired online from the current unified policy, it encountered less drift between observations rolled out by expert policies and the unified student policy. In all tasks, although objects and contact dynamics differed substantially, objects were placed under the robotic hand, and the overall pattern of finger movements was open-close-rotate. When training these tasks in one policy, the policy benefited from the greater variation of manipulation experiences introduced by similar but different tasks.

Comparison with the state-based expert pipeline

We further compared our proposed pipeline with the state-of-the-art skill distillation pipeline using state-based experts (*1*, *30*, *31*). We conducted an experiment by following this learning pipeline (state expert → VT unified): Per-task experts were trained in simulation using full state information (robot proprioception + object 6D pose/velocity) as in (*1*, *30*, *31*) and subsequently distilled into a unified policy with raw images and tactile events. In contrast, our pipeline was VT expert → VT unified: Per-task VT experts were trained in simulation and distilled into a unified policy, where both the experts and the unified policy used the pretrained visual-tactile encoder.

As shown in Fig. 5 (F and G), the state expert → VT unified approach obtained a success rate of 70.8% for the expert policies and 58.8% for the unified student policy on unseen objects; the unified policy suffered a marked drop of 12% relative to its experts. In contrast, our pipeline achieved an unexpected improvement of 6% over the experts. In the existing state-based distillation pipeline, the student policy had to simultaneously learn to map the sensory input to the state and control hand actions. The information loss from state to sensory input during learning resulted in degraded performance. In our method, when an expert and a student shared the same representation for the environment, distillation could be performed without degradation, even boosting the student's performance through shared experience.

Humanlike manipulation behaviors

To gain insight into the manipulation behaviors of learned policies, we performed a statistical analysis of tactile contact patterns across models pretrained with different modalities and compared them with human demonstrations. Specifically, we analyzed the average durations of contact segments from the three most frequently activated tactile sensors in each task, which reflected the temporal structure of finger-object interactions. The overall activation frequency of each tactile sensor across all tasks is provided in fig. S3.

As shown in Fig. 6A, the model pretrained with both visual and tactile human demonstrations (VT) exhibited contact patterns that were more similar to those of humans compared with models trained with a single modality. We further quantified this similarity using kernel density estimates and measured the mean squared error (MSE) between the human and robot contact duration distributions. This analysis highlighted the benefit of visual-tactile pretraining: By incorporating both modalities, the learned policies more closely reproduced the temporal structure of human hand-object interactions, contributing to contact patterns that were more similar to those observed in human demonstrations.

Multisensory correlation learned from human demonstrations

To understand the underlying reasons behind the improved human likeness, robustness under lighting variations, and sim-to-real transfer of our visual-tactile pretrained policy, we visualized the attention maps of the IPL token used in the pretraining network. Additional implementation details of the attention visualization are provided in the "Details for attention map visualization" section in the Supplementary Materials.

As shown in Fig. 6B, the difference in the attention patterns for visual-tactile and vision-only models was distinctive: Visual-tactile attention maps focused on hands and objects, whereas vision

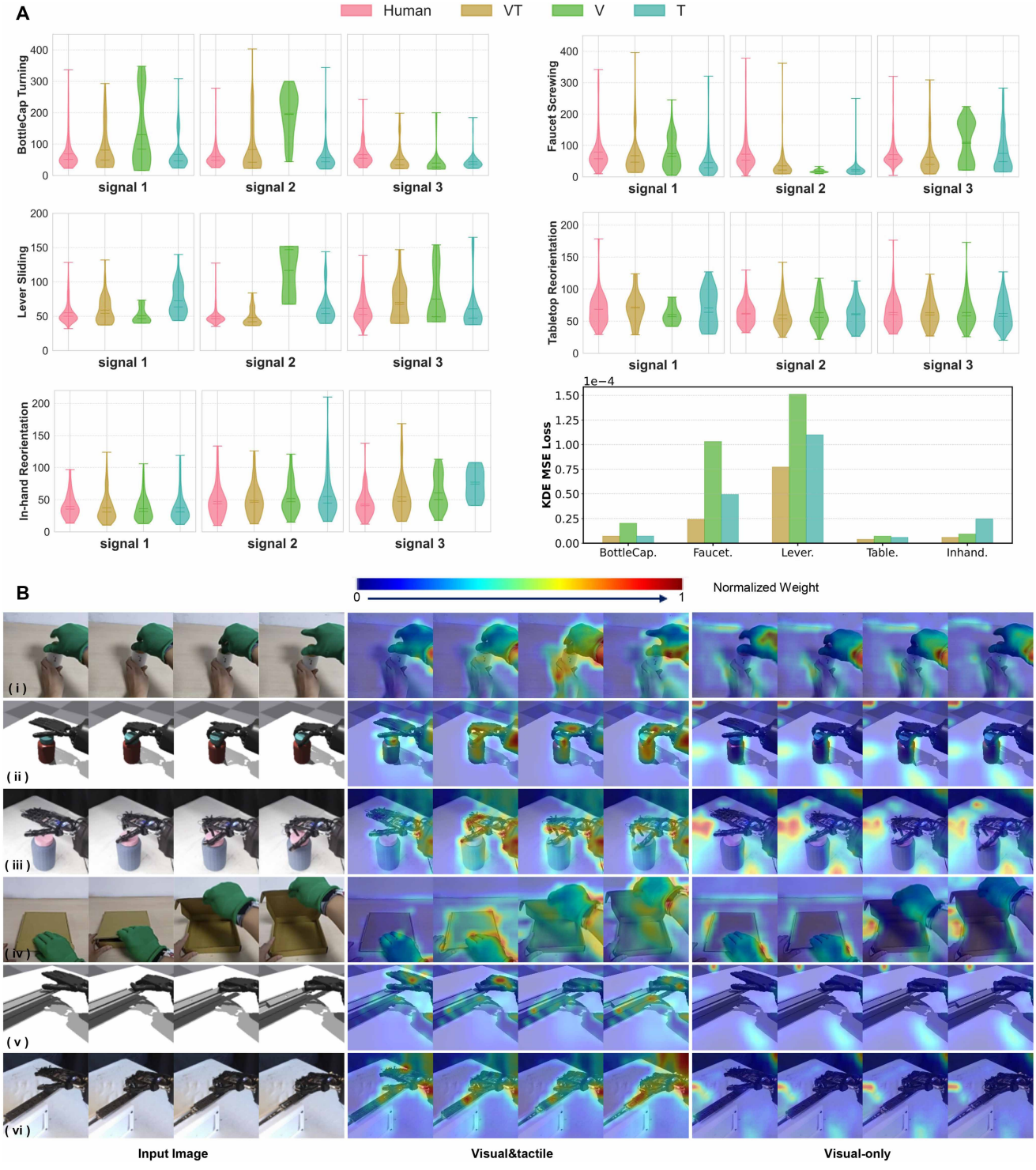

Fig. 6. Analysis of humanlike tactile patterns and attention for the integration token IPL. (**A**) Violin plots of contact-segment durations for three representative tactile channels per task, comparing our method with unimodal baselines (V and T) and human demonstrations. The KDE MSE quantifies similarity to human contact dynamics via kernel density estimation (lower is more humanlike). (**B**) Attention maps to IPL: The visual-tactile model consistently attends to hands and manipulated objects with action-dependent shifts, whereas the vision-only model shows less stable, task-relevant attention.

attention maps did not show this focus. The attention maps revealed the reasons for the superior success rate, the small sim-to-real gap, and the lighting insensitivity of the visual-tactile policy compared with the vision-only policy. With the tactile modality, the pretraining network encoded more relevant information for learning. Although the input images from human hand demonstrations, the robotic hand in simulation, and the robotic hand in the real world for the same task differed in appearance [an example is shown in Fig. 6B (i to iii)], the attention maps focused more on the areas related to interaction and reduced the interference from other areas in the vision modality. Integrating this highly related visual information and tactile events helped to compensate for the differences between simulation and reality.

A more unexpected finding of the visualization was that the attention on object areas varied with the dynamics of hands and objects, although no temporal information or status of hands and objects was provided during pretraining and pretraining was conducted on individual images rather than video sequences. Figure 6B (iv) shows a box, which is an articulated object consisting of two major parts. When the hand was not in contact with objects, the areas of attention were on the hand and the weights were moderate; when the hand touched the box, the attention weights of the box and the hand immediately became very high, particularly near the box edges and fingertips, and after the box was wide open, the attention distributed more evenly to all inner areas of the box. Similar attention changes with object status were observed in the downstream lever sliding task in both simulation and real-world deployment [shown in Fig. 6B (v and vi)]: The shaft gained more attention after being slid out from the slot. In comparison, attention from pretraining with vision showed little variation or variation patterns that did not show a clear correlation with object motions.

These attention dynamics explained the generalization of the pretraining models to unseen downstream manipulation tasks. The visual-tactile pretraining models learned to attend to object areas relating to hand-object dynamics, indicating that the integration token learned the representations of hand dynamics combined with visual-tactile sensory information. Although the pretraining network could not provide information on how to open the box without explicit action commands during training, it established a correlation between the hand status and the object areas that might change with the hand status. This correlation demonstrated some degree of "intent understanding" by gathering information more related to action decision-making; when learning a new task, this correlation helped action policy learning by providing more relevant visual information.

DISCUSSION

In our results, visual-tactile pretraining from human demonstrations consistently led to superior performance: enhanced learning efficiency; a reduced sim-to-real transfer gap; more humanlike manipulation behaviors; and improved generalization to novel objects, varying lighting conditions, and unseen tasks. We attribute these advantages to the critical role of tactile events in enriching sensory representations of action. Specifically, tactile events provide precise temporal cues about when contact occurs—information that is often ambiguous from vision alone. When combined with visual input during representation pretraining, tactile signals guide the model to focus on task-relevant regions, implicitly answering both when and where to attend. As shown in Fig. 6B, attention maps from the pretrained model consistently highlight hands and manipulated objects, particularly in regions associated with contact onset and dynamic interactions. This attention pattern suggests that the learned integration token captures abstract representations of sensorimotor interactions. Interestingly, this resembles the function of IPL neurons in the human brain, which are believed to encode actions through multisensory integration by "observing the acts done by others" (*61*). Our results suggest that visual-tactile pretraining serves a similar role, enabling the model to internalize task-relevant perception-action priors through observation. Why, then, can a model learn where to attend using only binary tactile events—i.e., whether contact occurred—without any explicit spatial annotations such as contact locations or segmentation masks?

This can be understood by analogy to standard image classification. Consider training a neural network (e.g., a convolutional neural network or transformer) on an image dataset labeled only with category tags (e.g., cat or dog), without any bounding boxes or pixelwise labels. Despite the lack of spatial supervision, such networks often learn to assign higher attention weights or feature activations to regions containing the object of interest. This emerges because during training, the network adjusts its internal weights to minimize classification errors. Tokens or features associated with background regions are inconsistent across samples and contribute little to the correct label, whereas those consistently aligned with the target concept (e.g., cat faces or bodies) converge to represent that concept. Similarly, in our case, although we did not provide explicit annotations of hand position, contact points, or object motion, the tactile events implicitly indicated the presence of contact involving specific parts of the hand. Across large-scale visual-tactile datasets, images associated with touch events tended to share consistent visual patterns—such as proximity between the hand and an object or particular grasping poses—whereas background regions varied randomly. Through pretraining, the model learned to associate these consistent patterns with the binary concept of "contact occurring" and, in doing so, implicitly learned where in the image to focus for contact-related information. This mechanism explains why the attention maps of the IPL token consistently highlight hand-object interaction regions: These are the visually informative areas that correlate most strongly with the tactile contact signal, even in the absence of explicit supervision.

MATERIALS AND METHODS

Our method was structured into three sequential stages: visual-tactile pretraining, manipulation skill learning, and real-world policy deployment. First, we leveraged human demonstration data to pretrain a model that learned a fused representation of perceived visual and tactile events. Then, this pretrained representation was used to enable the agent to efficiently acquire a unified dexterous manipulation policy for multiple tasks through RL in a simulated environment and online imitation learning. Last, we used domain randomization to bridge the gap between simulation and reality, enabling the deployment of the learned policy on real-world robotic systems.

Visual-tactile integration by pretraining with human demonstrations

The visual-tactile pretraining stage was designed to learn a fused representation of visual and tactile information from human demonstration data. This stage was crucial to enable the agent to

effectively perceive and interpret multimodal sensory input during dexterous manipulation tasks. The human demonstration dataset used in this work was collected with the proposed hardware system (*64*) (see the "Manipulation tasks in human demonstrations" section in the Supplementary Materials for more details).

Visual-tactile pretraining

To leverage the visual and tactile data from human manipulation tasks, we proposed a self-supervised pretraining framework inspired by an MAE (*62*), which consisted of a fusion encoder with a learnable integration token (IPL token) and a decoder to recover masked input.

Fusion encoder with IPL token. The fusion encoder E_θ took pairs of visual-tactile input $(\mathbf{V}, \mathbf{C}) \in \mathcal{D}_{\text{human}}$ and produced an integration representation for visual-tactile sensory information. It consisted of three steps: token extraction for the visual-tactile modality, random masking of the tokens, and integration of the multimodal tokens. First, the image and tactile input were tokenized. Specifically, the RGB image $\mathbf{V} \in \mathbb{R}^{H\times W\times 3}$ was divided into N_v patches and flattened to $\mathbf{v} \in \mathbb{R}^{N_v\times d_p}$, where H and W denote the height and width of the image, respectively; P represents the patch size; $N_v = (H\times W)/(P\times P)$ is the number of patches; and $d_p = P\times P\times 3$ is the dimension of each patch. Each patch was first linearly projected onto the dimension d_{en} via $\phi_\theta(\cdot)$ and then added with 2D sinusoidal positional embeddings $\mathbf{v}^{\text{pos}}$ that denote its position in the image, resulting in image patch embeddings $\bar{\mathbf{v}} \in \mathbb{R}^{N_v\times d_{\text{en}}}$. The tactile input $\mathbf{C} \in \{0,1\}^{20}$ for 20 tactile sensors was similarly split into patches and processed by multilayer perceptrons (MLPs) $\varphi_\theta(\cdot)$ with 1D positional embeddings $\mathbf{c}^{\text{pos}}$ to produce tactile patch embeddings $\bar{\mathbf{c}} \in \mathbb{R}^{N_c\times d_{\text{en}}}$, where $N_c = 20$ is the number of tactile patches. The process could be described as $\bar{\mathbf{v}} = \phi_\theta(\mathbf{v}) + \mathbf{v}^{\text{pos}}, \bar{\mathbf{c}} = \varphi_\theta(\mathbf{c}) + \mathbf{c}^{\text{pos}}$.

For the masking, a modality-specific masking function $M(\cdot, \gamma)$ was applied to randomly mask input patches at a specified ratio γ, generating visible patch embeddings $\bar{\mathbf{v}}_{\text{vis}} \in \mathbb{R}^{(1-\gamma_v)N_v\times d_{\text{en}}}$ and $\bar{\mathbf{c}}_{\text{vis}} \in \mathbb{R}^{(1-\gamma_c)N_c\times d_{\text{en}}}$ by $\bar{\mathbf{v}}_{\text{vis}} = M(\bar{\mathbf{v}}, \gamma_v), \bar{\mathbf{c}}_{\text{vis}} = M(\bar{\mathbf{c}}, \gamma_c)$. In (*64*), $(\bar{\mathbf{v}}_{\text{vis}}, \bar{\mathbf{c}}_{\text{vis}})$ was fed into the Transformer encoder $\mathbf{TransE}(\cdot)$, which consisted of stacked self-attention layers and feed-forward modules and enabled interaction and encoding between visual and tactile modalities. However, this attention mechanism only accomplished the goal of feature extraction and augmentation from the other modality. The updated features for $(\bar{\mathbf{v}}_{\text{vis}}, \bar{\mathbf{c}}_{\text{vis}})$ after attention were fed into the action policy directly. There was no process that resembled the multisensory integration in IPL neurons in the human brain before action execution in the motor cortex.

To integrate the visual-tactile modalities, an additional learnable integration token $\mathbf{IPL} \in \mathbb{R}^{1\times d_{\text{en}}}$ along with $(\bar{\mathbf{v}}_{\text{vis}}, \bar{\mathbf{c}}_{\text{vis}})$ was fed into the Transformer encoder $\mathbf{TransE}(\cdot)$. The integration token aggregated information from all visible visual-tactile tokens, serving as a critical perceptual embedding for downstream policy learning. After being processed by the encoder, the input tokens were updated with fused information from other tokens of the same modality or other modalities, which resulted in

$$\mathbf{h}_{\text{IPL}}, \mathbf{h}_{\text{v}}, \mathbf{h}_{\text{c}} = \mathbf{TransE}(\mathbf{IPL}, \bar{\mathbf{v}}_{\text{vis}}, \bar{\mathbf{c}}_{\text{vis}})$$

The updated tokens $\mathbf{h}_{\text{IPL}}, \mathbf{h}_{\text{v}}, \mathbf{h}_{\text{c}}$ encoded the integration, visual, and tactile features, respectively, maintaining the same dimensionality as the input tokens.

Reconstruction decoder. A reconstruction decoder was introduced to reconstruct the masked patches using the fused representation $\mathbf{h}_{\text{v}}, \mathbf{h}_{\text{c}}$ and mask tokens $\mathbf{m} \in \mathbb{R}^{(\gamma_v N_v + \gamma_c N_c)\times d_{\text{en}}}$. First, the Transformer-based decoder $\mathbf{TransD}(\cdot)$ inferred the restored vision and tactile embeddings $\hat{\mathbf{v}} \in \mathbb{R}^{N_v\times d_{\text{de}}}$ and (d_{de} is the output dimension of the decoder) $\hat{\mathbf{c}} \in \mathbb{R}^{N_c\times d_{\text{de}}}$. Then, MLPs were used to map these embeddings back to their original domains: the image domain $\hat{\mathbf{V}} \in \mathbb{R}^{H\times W\times 3}$ and the tactile domain $\hat{\mathbf{C}} \in \{0,1\}^{20}$.

Loss function. A weighted MSE loss evaluated the quality of the reconstruction of both modalities $\boldsymbol{L}(\boldsymbol{\theta}) = \lambda_{\text{v}} \cdot \mathbf{MSE}(\mathbf{V}, \hat{\mathbf{V}}) + \lambda_{\text{c}} \cdot \mathbf{MSE}(\mathbf{C}, \hat{\mathbf{C}})$, where λ_{v} and λ_{c} controlled the contributions of the vision and tactile modalities. This reconstruction objective encouraged the model to infer the contact state and appearance of the masked regions on the basis of the observed visual patches and tactile signals. In this way, the model could learn the complementary relationship between visual and tactile information in a self-supervised manner, producing an integration representation that associates both modalities.

Manipulation skill learning

The learning of dexterous manipulation skills was achieved by interaction with a simulated environment. The whole framework consisted of two main stages: task-specific expert policy learning and online multitask learning. We first trained task-specific policies for each dexterous manipulation task incorporating the pretrained visual-tactile representation. Given the learned task-specific policies, we further adopted an online learning strategy to obtain a unified policy that could generalize across all tasks given the learned task-specific expert policies.

Task-specific expert policy learning

We modeled dexterous manipulation tasks as a Markov decision process defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively. The policy π_θ: $\mathcal{S} \to \mathcal{A}$ maps states to actions, whereas $\mathcal{T}$: $\mathcal{S}\times\mathcal{A} \to \mathcal{S}$ represents the transition dynamics. The reward function $\mathcal{R}$: $\mathcal{S}\times\mathcal{A} \to \mathbb{R}$ guides the policy, with $\gamma \in (0,1)$ as the discount factor. Our goal was to optimize the expected discounted reward

$$\boldsymbol{J}(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}(\mathbf{s}_t, \mathbf{a}_t)\right]$$

We used proximal policy optimization (*80*) to train policies for dexterous manipulation skills.

State space. For all tasks, the state was defined as $\mathbf{s} = [\mathbf{h}_{\text{IPL}}; \mathbf{P}]$. h_{IPL} was the integration representation for RGB images $\mathbf{V}_{\text{sim}}$ and tactile signals $\mathbf{C}_{\text{sim}}$, as introduced in the previous section. $\mathbf{V}_{\text{sim}}$ was captured using an egocentric camera, whereas $\mathbf{C}_{\text{sim}}$ was obtained from tactile sensors with a threshold of 0.01 N. Proprioceptive information $\mathbf{P}$ includes joint positions and velocities of the dexterous hand.

Action space. In all tasks, we used the Shadow Hand, which has 24 degrees of freedom, including four tendon-driven joints. To simplify the learning process, we immobilized the arm, restricting actions to finger movements. Thus, the action $\mathbf{a} = \pi_\theta(\mathbf{s}) \in \mathbb{R}^{20}$.

Rewards. The criteria for a successful manipulation varied, and we shaped rewards for different tasks. The definition of the reward function can be found in the "Reward function for each task" section in the Supplementary Materials.

Online multitask learning

Given the expert policies learned for each task, in this section, our aim was to learn a unified policy for all tasks. The unified policy π_θ was trained to approximate the behavior of the task-specific expert policies $\{\pi_1^*, \pi_2^*, \ldots, \pi_N^*\}$. Each expert policy π_i^* was independently trained for its corresponding task T_i using RL in the previous section. Directly rolling out the manipulation demonstrations from the expert policies and learning the unified policy by imitation learning had the issue of observation drift with increasing steps. To mitigate the observation drift issue, we adopted an online learning similar to the learning strategies proposed (*1, 37*). In contrast with rolling out the expert demonstrations offline to learn the multitask policy, we sampled the observation states visited by the multitask policy online during learning and queried the expert policies for action supervision with the visited observation states. Specifically, online multitask learning aggregated the dataset to train the unified policy π_θ and learn the policy iteratively.

Unified policy architecture. The unified policy π_θ was parameterized as a MLP designed to process multitask inputs. To distinguish between tasks, each task in the set of tasks $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$ was assigned a unique hot-encoded task identifier $\mathbf{z}^i \in \mathbb{R}^N$, where N is the total number of tasks (five in this study). This ID of the task was concatenated with the state vector $\mathbf{s} \in \mathcal{S}$ to form an augmented state $\tilde{\mathbf{s}}^i = [\mathbf{s}^i; \mathbf{z}^i]$, where $[\cdot;\cdot]$ represents the concatenation of the vector. The task ID allowed the policy to differentiate tasks. The input layer of the MLP policy network took the augmented state $\tilde{\mathbf{s}}^i$ as input. The hidden layers consisted of three fully connected layers with dimensions (1024, 1024, 512), activated by exponential linear unit (ELU) functions. The output layer mapped the hidden representation to the action variable $\mathbf{a} \in \mathbb{R}^{20}$, which represented the expected joint positions of the fingers and the wrist of the hand.

Iterative dataset aggregation and policy training. At a training iteration, it was assumed that a dataset $\mathcal{D} = \cup_{i=1}^N \mathcal{D}^{\mathrm{n}}$ was given, consisting of state and action pairs $(\tilde{\mathbf{s}}^i, \mathbf{a}^i)$ for task i. The dataset aggregation consisted of three steps. For the policy interaction step, the unified policy π_θ in the iteration interacted with the environment to generate a set of state and action pairs $\tau = \{(\tilde{\mathbf{s}}^i, \mathbf{a}^i)\}$. For the expert query step, for each sampled state $\tilde{\mathbf{s}}^i$ in τ, we queried the expert policies for action supervision $\mathbf{a}^{i*}$. For the dataset update step, the dataset for task T_i was updated by adding the collected state-action pairs $\mathcal{D}_i \leftarrow \mathcal{D}_i \cup \{(\tilde{\mathbf{s}}^i, \mathbf{a}^{i*})\}$. The dataset $\mathcal{D}$ to train the unified policy was aggregated by updating all the tasks dataset with these states visited by the unified policy and the action supervision queried from the expert policies. Given the aggregated dataset, the policy was trained with imitation loss, which was defined as the MSE between the actions predicted by the unified policy π_θ and the expert actions $\mathbf{a}^*$

$$L_{\mathrm{imitate}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{(\tilde{\mathbf{s}}^i, \mathbf{a}^{i*}) \sim \mathcal{D}_i}\left[\left\|\pi_\theta(\tilde{\mathbf{s}}^i) - \mathbf{a}^{i*}\right\|^2\right]$$

Transferring to reality

Directly applying the learned policy in the real world often resulted in substantial performance degradation due to the domain gap between simulation and physical environments. To mitigate this sim-to-real gap, we adopted domain randomization (*24*, *81*, *82*). We randomized multiple sensory modalities. Specifically, for proprioception, we added Gaussian noise to joint positions and velocities, whereas for vision, we randomized object colors, textures, and lighting conditions. Last, for tactile sensing, we applied random perturbations to the binarization thresholds to account for sensor variability. These randomizations were sampled from predefined distributions and applied both at the start of each episode or during rollouts. A full list of randomization parameters and their noise configurations is provided in the Supplementary Materials (see the "Sim-to-real transfer via domain randomization" section).

Supplementary Materials

The PDF file includes:

Methods
Figs. S1 to S12
Tables S1 to S9
Algorithm S1
References (*83–99*)

Other Supplementary Material for this manuscript includes the following:

Movies S1 to S3

REFERENCES AND NOTES

1. T. Chen, M. Tippur, S. Wu, V. Kumar, E. Adelson, P. Agrawal, Visual dexterity: In-hand reorientation of novel and complex object shapes. *Sci. Robot.* **8**, eadc9244 (2023).
2. G. Solak, L. Jamone, "Learning by demonstration and robust control of dexterous in-hand robotic manipulation skills," in *Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2019), pp. 8246–8251.
3. A. S. Morgan, K. Hang, B. Wen, K. Bekris, A. M. Dollar, Complex in-hand manipulation via compliance-enabled finger gaiting and multi-modal planning. *IEEE Robot. Autom. Lett.* **7**, 4821–4828 (2022).
4. G. Solak, L. Jamone, Haptic exploration of unknown objects for robust in-hand manipulation. *IEEE Trans. Haptics* **16**, 400–411 (2023).
5. F. Khadivar, A. Billard, Adaptive fingers coordination for robust grasp and in-hand manipulation under disturbances and unknown dynamics. *IEEE Trans. Robot.* **39**, 3350–3367 (2023).
6. X. Gao, K. Yao, F. Khadivar, A. Billard, Enhancing dexterity in confined spaces: Real-time motion planning for multifingered in-hand manipulation. *IEEE Robot. Autom. Mag.* **31**, 100–112 (2024).
7. I. Mordatch, Z. Popović, E. Todorov, "Contact-invariant optimization for hand manipulation," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (ACM, 2012), pp. 137–144.
8. V. Kumar, Y. Tassa, T. Erez, E. Todorov, "Real-time behaviour synthesis for dynamic hand-manipulation," in *Proceedings of 2014 IEEE International Conference on Robotics and Automation* (IEEE, 2014), pp. 6808–6815.
9. G. J. Pollayil, G. Grioli, M. Bonilla, A. Bicchi, "Planning robotic manipulation with tight environment constraints," in *Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2021), pp. 9385–9392.
10. A. Nagabandi, K. Konolige, S. Levine, V. Kumar, "Deep dynamics models for learning dexterous manipulation," in *Proceedings of 2020 Conference on Robot Learning* (PMLR, 2020), pp. 1101–1112.
11. X. Zhu, J. H. Ke, Z. Xu, Z. Sun, B. Bai, J. Lv, Q. Liu, Y. Zeng, Q. Ye, C. Lu, M. Tomizuka, L. Shao, "Diff-lfd: Contact-aware model-based learning from visual demonstration for robotic manipulation via differentiable physics-based simulation and rendering," in *Proceedings of 2023 Conference on Robot Learning* (PMLR, 2023), pp. 499–512.
12. Y. Jiang, M. Yu, X. Zhu, M. Tomizuka, X. Li, "Contact-implicit model predictive control for dexterous in-hand manipulation: A long-horizon and robust approach," in *Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2024), pp. 5260–5266.
13. H. Yin, A. Varava, D. Kragic, Modeling, learning, perception, and control methods for deformable object manipulation. *Sci. Robot.* **6**, eabd8803 (2021).
14. J. Ichnowski, Y. Avigal, V. Satish, K. Goldberg, Deep learning can accelerate grasp-optimized motion planning. *Sci. Robot.* **5**, eabd7710 (2020).
15. W. Yuan, J. A. Stork, D. Kragic, M. Y. Wang, K. Hang, "Rearrangement with nonprehensile manipulation using deep reinforcement learning," in *Proceedings of 2018 IEEE International Conference on Robotics and Automation* (IEEE, 2018), pp. 270–277.
16. M. Ishige, T. Taniguchi, Y. Kawahara, Dream to posture: Visual posturing of a tendon-driven hand using world model and muscle synergies. *Adv. Robot.* **37**, 1237–1252 (2023).

17. A. Handa, A. Allshire, V. Makoviychuk, A. Petrenko, R. Singh, J. Liu, D. Makoviichuk, K. Van Wyk, A. Zhurkevich, B. Sundaralingam, Y. Narang, J.-F. Lafleche, D. Fox, G. State, "Dextreme: Transfer of agile in-hand manipulation from simulation to reality," in *Proceedings of 2023 IEEE International Conference on Robotics and Automation* (IEEE, 2023), pp. 5977–5984.
18. W. Hu, B. Huang, W. W. Lee, S. Yang, Y. Zheng, Z. Li, Dexterous in-hand manipulation of slender cylindrical objects through deep reinforcement learning with tactile sensing. *Robot. Auton. Syst.* **186**, 104904 (2025).
19. A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Proceedings of Robotics: Science and Systems XIV* (RSS, 2018); 10.15607/RSS.2018.XIV.049.
20. Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, X. Wang, "Dexmv: Imitation learning for dexterous manipulation from human videos," in *Proceedings of the European Conference on Computer Vision* (Springer, 2022), pp. 570–587.
21. Y.-H. Wu, J. Wang, X. Wang, "Learning generalizable dexterous manipulation from human grasp affordance," in *Proceedings of 2023 Conference on Robot Learning* (PMLR, 2023), pp. 618–629.
22. Q. Liu, Y. Cui, Q. Ye, Z. Sun, H. Li, G. Li, L. Shao, J. Chen, "Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations," in *Proceedings of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2023), pp. 3153–3160.
23. P. Mandikal, K. Grauman, "Learning dexterous grasping with object-centric visual affordances," in *Proceedings of 2021 IEEE International Conference on Robotics and Automation* (IEEE, 2021), pp. 6169–6176.
24. P. Mandikal, K. Grauman, "Dexvip: Learning dexterous grasping with human hand pose priors from video," in *Proceedings of 2022 Conference on Robot Learning* (PMLR, 2022), pp. 651–661.
25. Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen, T. Liu, L. Yi, H. Wang, "Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), pp. 4737–4746.
26. J. Borja-Diaz, O. Mees, G. Kalweit, L. Hermann, J. Boedecker, W. Burgard, "Affordance learning from play for sample-efficient policy learning," in *Proceedings of 2022 International Conference on Robotics and Automation* (ACM, 2022), pp. 6372–6378.
27. T. Silver, K. Allen, J. Tenenbaum, L. Kaelbling, Residual policy learning. arXiv:1812.06298 [cs.RO] (2018).
28. K. Li, P. Li, T. Liu, Y. Li, S. Huang, "ManipTrans: Efficient dexterous bimanual manipulation transfer via residual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2025), pp. 6991–7003.
29. J. Zhang, Y. Zhang, L. An, M. Li, H. Zhang, Z. Hu, Y. Liu, ManiDext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* 10.1109/TPAMI.2025.3588302 (2025).
30. H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, J. Malik, "General in-hand object rotation with vision and touch," in *Proceedings of 2023 Conference on Robot Learning* (PMLR, 2023), pp. 2549–2564.
31. Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, X. Wang, "Robot synesthesia: In-hand manipulation with visuotactile sensing," in *Proceedings of 2024 IEEE International Conference on Robotics and Automation* (IEEE, 2024), pp. 6558–6565.
32. T.-W. Ke, N. Gkanatsios, K. Fragkiadaki, "3D diffuser actor: Policy diffusion with 3D scene representations," in *Proceedings of 2024 Conference on Robot Learning* (PMLR, 2024), pp. 1949–1974.
33. T. Zhang, Y. Hu, H. Cui, H. Zhao, Y. Gao, "A universal semantic-geometric representation for robotic manipulation," in *Proceedings of 2023 Conference on Robot Learning* (PMLR, 2023), pp. 3342–3363.
34. J. Duan, W. Yuan, W. Pumacay, Y. R. Wang, K. Ehsani, D. Fox, R. Krishna, "Manipulate-anything: Automating real-world robots using vision-language models," in *Proceedings of 2024 Conference on Robot Learning* (PMLR, 2024), pp. 5326–5350.
35. S. Li, H. Yu, W. Ding, H. Liu, L. Ye, C. Xia, Visual–tactile fusion for transparent object grasping in complex backgrounds. *IEEE Trans. Robot.* **39**, 3838–3856 (2023).
36. F. Zhang, Y. Demiris, Visual-tactile learning of garment unfolding for robot-assisted dressing. *IEEE Robot. Autom. Lett.* **8**, 5512–5519 (2023).
37. I. Guzey, B. Evans, S. Chintala, L. Pinto, "Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play," in *Proceedings of 2023 Conference on Robot Learning* (PMLR, 2023), pp. 3142–3166.
38. Y. Han, K. Yu, R. Batra, N. Boyd, C. Mehta, T. Zhao, Y. She, S. Hutchinson, Y. Zhao, Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer. *IEEE/ASME Trans. Mechatron.* **30**, 554–566 (2025).
39. R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?," in *Proceedings of 2017 Conference on Robot Learning* (PMLR, 2017), pp. 314–323.
40. C. Sferrazza, Y. Seo, H. Liu, Y. Lee, P. Abbeel, "The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning," in *Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2024), pp. 9698–9705.
41. R. Liu, X. Liu, "Mu-mae: Multimodal masked autoencoders-based one-shot learning," in *Proceedings of 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval* (IEEE, 2024), pp. 253–259.
42. T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, J. Malik, "Learning visuotactile skills with two multifingered hands," in *Proceedings of 2025 IEEE International Conference on Robotics and Automation* (IEEE, 2024), pp. 5637–5643.
43. M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, J. Bohg, Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Trans. Robot.* **36**, 582–596 (2020).
44. A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," in *Proceedings of Robotics: Science and Systems* (RSS, 2023).
45. S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, C. Finn, Robonet: "Large-scale multi-robot learning," in *Proceedings of 2019 Conference on Robot Learning* (PMLR, 2019), pp. 885–897.
46. H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, V. Kumar, "Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," in *Proceedings of 2024 IEEE International Conference on Robotics and Automation* (IEEE, 2024), pp. 4788–4795.
47. E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Proceedings of 2022 Conference on Robot Learning* (PMLR, 2022), pp. 991–1002.
48. A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, V. Guizilini, D. A. Herrera, M. Heo, K. Hsu, J. Hu, M. Z. Irshad, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, C. Finn, "Droid: A large-scale in-the-wild robot manipulation dataset," in *Proceedings of Robotics: Science and Systems* (RSS, 2024).
49. H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, A. Lee, K. Fang, C. Finn, S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Proceedings of 2023 Conference on Robot Learning* (PMLR, 2023), pp. 1723–1736.
50. Y. Liu, Y. Yang, Y. Wang, X. Wu, J. Wang, Y. Yao, S. Schwertfeger, S. Yang, W. Wang, J. Yu, X. He, Y. Ma, "RealDex: Towards human-like grasping for robotic dexterous hand," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (ACM, 2024), pp. 6859–6867.
51. S. Nair, A. Rajeswaran, V. Kumar, C. Finn, A. Gupta, "R3M: A universal visual representation for robot manipulation," in *Proceedings of 2022 Conference on Robot Learning* (PMLR, 2022), pp. 892–909.
52. S. Karamcheti, S. Nair, A. Chen, T. Kollar, "Language-driven representation learning for robotics," in *Proceedings of Robotics: Science and Systems* (RSS, 2023).
53. R. Tian, C. Xu, M. Tomizuka, J. Malik, A. Bajcsy, "VIP: Towards universal visual reward and representation via value-implicit pre-training," in *Proceedings of the Eleventh International Conference on Learning Representations* (ICLR, 2023).
54. I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, T. Darrell, "Real-world robot learning with masked visual pre-training," in *Proceedings of 2023 Conference on Robot Learning* (PMLR, 2023), pp. 416–426.
55. Y. J. Ma, W. Liang, V. Som, V. Kumar, A. Zhang, O. Bastani, D. Jayaraman, "LIV: Language-image representations and rewards for robotic control," in *Proceedings of International Conference on Machine Learning* (PMLR, 2023), pp. 23301–23320.
56. F. Ceola, E. Maiettini, L. Rosasco, L. Natale, "A grasp pose is all you need: Learning multi-fingered grasping with deep reinforcement learning from vision and touch," in *Proceedings of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2023), pp. 2985–2992.
57. M. Bauza, A. Bronars, Y. Hou, I. Taylor, N. Chavan-Dafle, A. Rodriguez, SimPLE, a visuotactile method learned in simulation to precisely pick, localize, regrasp, and place objects. *Sci. Robot.* **9**, eadi8808 (2024).

58. J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, G. Dudek, "Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning," in *Proceedings of 2022 IEEE International Conference on Robotics and Automation* (IEEE, 2022), pp. 8298–8304.
59. D. M. Lloyd, D. I. Shore, C. Spence, G. A. Calvert, Multisensory representation of limb position in human premotor cortex. *Nat. Neurosci.* **6**, 17–18 (2003).
60. H. H. Ehrsson, C. Spence, R. E. Passingham, That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science* **305**, 875–877 (2004).
61. L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, G. Rizzolatti, Parietal lobe: From action organization to intention understanding. *Science* **308**, 662–667 (2005).
62. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), pp. 16000–16009.
63. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (ACL, 2019), pp. 4171–4186.
64. Q. Liu, Q. Ye, Z. Sun, Y. Cui, G. Li, J. Chen, "Masked visual-tactile pre-training for robot manipulation," in *Proceedings of 2024 IEEE International Conference on Robotics and Automation* (IEEE, 2024), pp. 13859–13875.
65. M. Li, H. Yin, K. Tahara, A. Billard, "Learning object-level impedance control for robust grasping and dexterous manipulation," in *Proceedings of 2014 IEEE International Conference on Robotics and Automation* (IEEE, 2014), pp. 6784–6791.
66. M. Li, Y. Bekiroglu, D. Kragic, A. Billard, "Learning of grasp adaptation through experience and tactile sensing," in *Proceedings of 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2014), pp. 3339–3346.
67. A. Bernardino, M. Henriques, N. Hendrich, J. Zhang, "Precision grasp synergies for dexterous robotic hands," in *Proceedings of 2013 IEEE International Conference on Robotics and Biomimetics* (IEEE, 2013), pp. 62–67.
68. T. Yu, D. Quillen, Z. He, R. Julian, A. Narayan, H. Shively, A. Bellathur, K. Hausman, C. Finn, S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Proceedings of 2020 Conference on Robot Learning* (PMLR, 2020), pp. 1094–1100.
69. D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, K. Hausman, "Mt-opt: Continuous multi-task robotic reinforcement learning at scale," in *Proceedings of 2021 Conference on Robot Learning* (PMLR, 2021), pp. 557–575.
70. Y. Chen, Y. Geng, F. Zhong, J. Ji, J. Jiang, Z. Lu, Bi-dexhands: Towards human-level bimanual dexterous manipulation. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 2804–2818 (2023).
71. C. Yang, K. Yuan, Q. Zhu, W. Yu, Z. Li, Multi-expert learning of adaptive legged locomotion. *Sci. Robot.* **5**, eabb2174 (2020).
72. L. Han, Q. Zhu, J. Sheng, C. Zhang, T. Li, Y. Zhang, H. Zhang, Y. Liu, C. Zhou, R. Zhao, J. Li, Y. Zhang, R. Wang, W. Chi, X. Li, Y. Zhu, L. Xiang, X. Teng, Z. Zhang, Lifelike agility and play in quadrupedal robots using reinforcement learning and generative pre-trained models. *Nat. Mach. Intell.* **6**, 787–798 (2024).
73. R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, S. Levine, "Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration," in *Proceedings of 2018 IEEE International Conference on Robotics and Automation* (IEEE, 2018), pp. 3758–3765.
74. S. Haldar, Z. Peng, L. Pinto, "BAKU: An efficient transformer for multi-task policy learning," in *Proceedings of Annual Conference on Neural Information Processing Systems* (NeurIPS, 2024), pp. 141208–141239.
75. S. Ross, G. Gordon, D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (PMLR, 2011), pp. 627–635.
76. D. Sharma, K. Tokas, A. Puri, K. Sharda, Shadow Hand. *J. Adv. Res. Appl. Sci.* **1**, 4–7 (2014).
77. Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, A. Anandkumar, "Eureka: Human-level reward design via coding large language models," in *Proceedings of the Twelfth International Conference on Learning Representations* (ICLR, 2024), pp. 26516–26560.
78. J. Wong, V. Makoviychuk, A. Anandkumar, Y. Zhu, "OSCAR: Data-driven operational space control for adaptive and robust robot manipulation," in *Proceedings of 2022 IEEE International Conference on Robotics and Automation* (IEEE, 2022), pp. 10519–10526.
79. J. Liu, C. Li, D. Delehelle, Z. Li, F. Chen, Skylark0924/Rofunc: v0.0.2.5 More examples (Zenodo, 2023); https://doi.org/10.5281/zenodo.10016946.
80. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms. arXiv:1707.06347 [cs.LG] (2017).
81. J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2017), pp. 23–30.
82. Open AI, M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. M. Grew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, W. Zaremba, Learning dexterous in-hand manipulation. *Int. J. Robot. Res.* **39**, 3–20 (2020).
83. V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, G. State, "Isaac Gym: High performance GPU-based physics simulation for robot learning," in *Proceedings of Annual Conference on Neural Information Processing Systems Track on Datasets and Benchmarks* (NeurIPS, 2021).
84. A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, F. Yu, Shapenet: An information-rich 3D model repository. arXiv:1512.03012 [cs.GR] (2015).
85. R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, H. Wang, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," in *Proceedings of 2023 IEEE International Conference on Robotics and Automation* (IEEE, 2023), pp. 11359–11366.
86. F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, H. Su, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), pp. 11097–11107.
87. B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *Proceedings of 2015 International Conference on Advanced Robotics* (IEEE, 2015), pp. 510–517.
88. D.-A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (ELUs). arXiv:1511.07289 [cs.LG] (2015).
89. K. Shaw, A. Agarwal, D. Pathak, "LEAP hand: Low-cost, efficient, and anthropomorphic hand for robot learning," in *Proceedings of Robotics: Science and Systems* (RSS, 2023).
90. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.
91. C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, S. Song, Diffusion policy: Visuomotor policy learning via action diffusion. *Int. J. Robot. Res.* **44**, 1684–1704 (2024).
92. Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, H. Xu, "3D diffusion policy: Generalizable visuomotor policy learning via simple 3D representations," in *Proceedings of Robotics: Science and Systems* (RSS, 2024).
93. W. Yuan, S. Dong, E. H. Adelson, Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors* **17**, 2762 (2017).
94. M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, R. Calandra, DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robot. Autom. Lett.* **5**, 3838–3845 (2020).
95. F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, A. Owens, "Touch and go: Learning from human-collected vision and touch," in *Proceedings of Annual Conference on Neural Information Processing Systems* (NeurIPS, 2022), pp. 8081–8103.
96. J. Kerr, H. Huang, A. Wilcox, R. Hoque, Jeffrey Ichnowski, R. Calandra, K. Goldber, "Self-supervised visuo-tactile pretraining to locate and follow garment features," in *Proceedings of Robotics: Science and Systems* (RSS, 2023).
97. Y. Dou, F. Yang, Y. Liu, A. Loquercio, A. Owens, "Tactile-augmented radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), pp. 26529–26539.
98. S. Yu, K. Lin, A. Xiao, J. Duan, H. Soh, "Octopi: Object property reasoning with large tactile-language models," in *Proceedings of Robotics: Science and Systems* (RSS, 2024).
99. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of International Conference on Machine Learning* (PMLR, 2021), pp. 8748–8763.

Acknowledgments: This system paper was made possible by the efforts of many people that contributed at different stages of the project. We acknowledge the help of T. Chen, Z. Sun, P. Xu, and S. Fan in making this paper a reality. **Funding:** The paper was supported by the National Natural Science Foundation of China (NSFC) under grant nos. 62088101, 62233013, and 62293511. **Author contributions:** Q.Y. organized the project, contributed all theoretical ideas including the multisensory representation and policy learning pipeline, and analyzed the results. Q.L. collected the dataset; proposed and implemented the online multitask learning strategy, pretraining network, and sim-to-real solutions; established the physical control pipeline; and implemented the baselines. S.W. and Jiaying Chen conducted the real-world experiments, optimized the data collection pipeline, and constructed the LeapHand platform. Q.L. and Y.C. implemented the expert policies and the RL baseline. H.C. constructed the tactile sensing systems. G.L. and Jiming Chen built the robot system and implemented low-level control. Jiming Chen led the project, providing overall supervision and financial support. Q.Y., Q.L., K.J., and X.C. wrote the manuscript and designed the figures. **Competing interests:** The authors declare that they have no competing interests. **Data, code, and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials. The data and code for this study have been deposited at https://doi.org/10.5281/zenodo.17986310. No new materials have been generated in this study.

Submitted 18 April 2025
Accepted 24 December 2025
Published 28 January 2026
10.1126/scirobotics.ady2869

Science Robotics

Visual-tactile pretraining and online multitask learning for humanlike manipulation dexterity

Qi Ye, Qingtao Liu, Siyun Wang, Jiaying Chen, Yu Cui, Ke Jin, Huajin Chen, Xuan Cai, Gaofeng Li, and Jiming Chen

Sci. Robot. **11** (110), eady2869. DOI: 10.1126/scirobotics.ady2869

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.