2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Learning End-to-End Dexterous Arm-Hand VLA Policies with Shared Autonomy: DexGrasp AI Copilot for Efficient Teleoperation

Yu Cui[1,†], Yujian Zhang[1], Lina Tao[1], Yang Li[1], Xinyu Yi[1], Zhibin Li[1]

***Abstract*— Achieving human-like dexterous manipulation is essential for general-purpose robots but remains a challenge. Recent advances in Vision-Language-Action (VLA) models offer the potential to learn flexible skills from demonstration data. However, training effective VLAs requires a large amount of high-quality data, which is difficult to obtain: fully manual teleoperation cognitively overloads human operators, while automated planning produces unnatural motions and lacks data diversity. We present a *Shared Autonomy* framework: a human operator teleoperates the arm for global motion, while an autonomous *DexGrasp-VLA* policy, as an *AI Copilot*, generates force-adaptive actions for a five-finger hand with tactile feedback – drastically reducing human effort and enabling efficient collection of high-quality demonstrations. Using these data, we train an end-to-end VLA policy with a novel *Arm-Hand Feature Enhancement* module – shared representations are conjunct with separate arm and hand latent features, representing the distinct dynamics of macro and micro movements, leading to more robust and natural coordination of arm-hand motions. Our *Corrective Teleoperation* can further refine the policy with failure-recovery demonstrations via human intervention. Experiments show our approach efficiently generates high-quality data and learns policies with a high success rate and natural behaviors. The trained arm-hand VLA policy is effectively generalized to both seen and unseen objects, with a success rate of around 90% in more than 50 diverse objects.**

I. INTRODUCTION

The goal of general-purpose robots is to achieve physical intelligence comparable to humans for complex tasks through flexible and diverse manipulation. Humanoid robots with dexterous hands offer promising potential for this goal due to their anthropomorphic design, which naturally blends into human-centric environments. However, a critical bottleneck remains in unlocking their full potential: advanced dexterous manipulation [1], [2], [3]. This capability requires not only precise control of the robotic arm's spatial motion but also the execution of delicate hand actions and, crucially, seamless coordination between the two. Currently, learning such a level of arm-hand coordination remains an open question.

Recent advances in Vision-Language-Action (VLA) models [4], [5], [6], [7], [8] have shown great potential for dexterous manipulation. However, training such models requires large-scale, high-quality demonstration data [9], [10], [11]. Current data collection methods often rely on fully manual teleoperation [12], [13], which places a heavy cognitive load on operators by requiring simultaneous control of all Degrees-of-Freedom (DoFs) of the arm and hand. Alternative methods leverage motion planning [14], [15], [16], [17] or reinforcement learning [2], [18], [19] to automate data generation. However, these typically involve significant manual engineering and struggle to produce natural, human-like behaviors, especially in multi-finger dexterous tasks.

To address these limitations, we propose a novel *Shared Autonomy* framework. Our core insight is to share the load of the control tasks: a human operator teleoperates the robotic arm's end-effector via a VR interface—focusing on global motion and navigation through the workspace—while the autonomous *DexGrasp-VLA* controller acts as an AI copilot, managing fine-grained motor control of the dexterous hand. Specifically, we train *DexGrasp-VLA*, which integrates visual input from an in-hand camera, tactile feedback, and proprioceptive data to generate grasping actions. This division of labor significantly reduces the operator's cognitive load and enables more efficient collection of high-quality demonstration data for coordinated arm-hand motions.

Leveraging this data, we conduct Supervised Fine-Tuning (SFT) to train a coherent end-to-end VLA policy for arm-hand coordination. Central to this policy is our novel Arm-Hand Feature Enhancement module. It first processes shared visual-semantic representations through separate encoders for the arm and the hand, each of which is supervised by auxiliary losses. The resulting dedicated features are then fused with the shared representation to predict the actions. This design explicitly captures the distinct kinematic and control characteristics of arm macro-movements and hand micro-manipulation, while retaining the task context and coordination dynamics encoded in the shared features. Experiments demonstrate that this architecture learns complex coordination more efficiently and yields more natural and robust behavior compared to a shared, undifferentiated network for arm-hand action prediction.

Finally, we implemented a corrective human-in-the-loop teleoperation system for self-improvement: we deploy the trained arm-hand VLA policy for automated data collection, where successful trajectories are recorded, while naturally occurring failures are recovered through human-in-the-loop intervention. Both the successful and recovery data are aggregated into training datasets for iterative model refinement, enabling incremental improvements to the policy.

Our primary contributions are as follows:

- **Shared Autonomy for Data Collection:** human VR teleoperation combined with an autonomous AI Copilot for grasping, enabling efficient collection of high-quality demonstrations with low cognitive load.
- **Multimodal AI Copilot for Dexterous Grasping:** an AI Copilot, DexGrasp-VLA, that fuses visual, language, tac-

[1] BtyeDance Seed
[†] Corresponding author, email: cuiyu.0627@bytedance.com

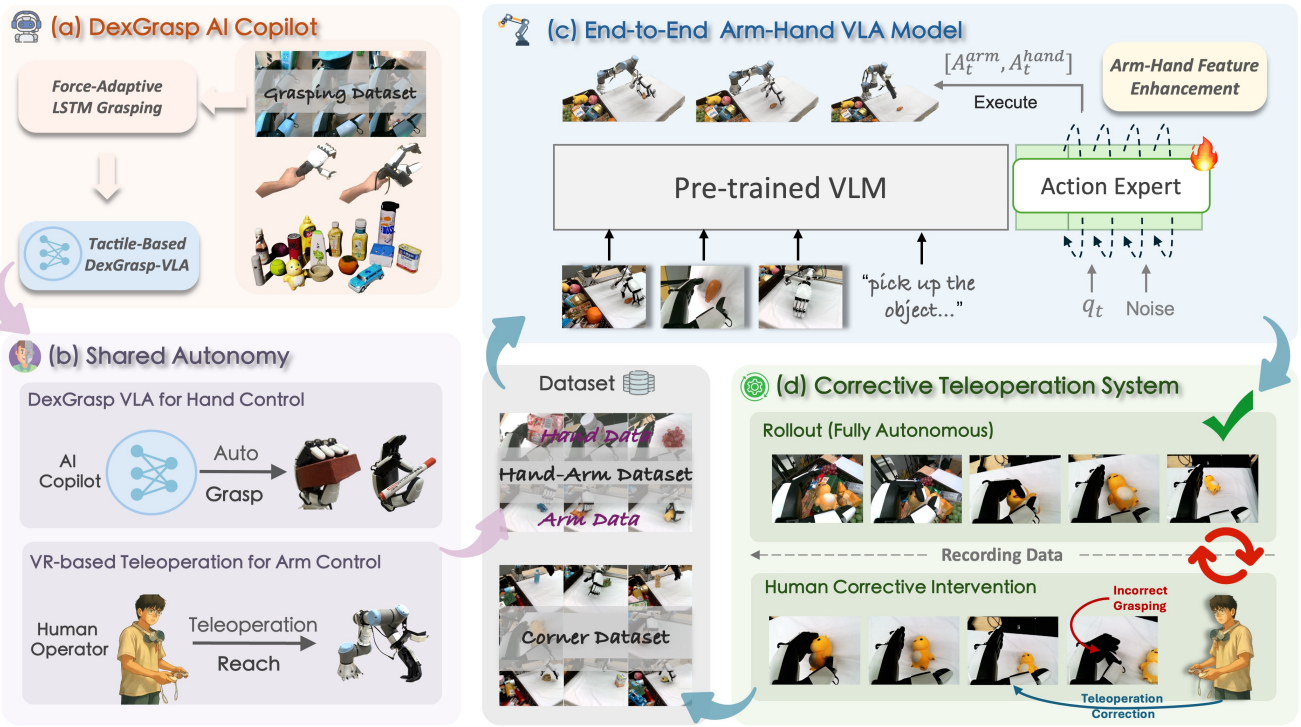

Fig. 1: Data collection and training pipeline for DexGrasp-VLA policy and arm-hand VLA policies. (a) Tactile-based DexGrasp-VLA policy for a five-finger dexterous hand, (b) Shared autonomy data collection, (c) End-to-end arm-hand policy learning with arm-hand feature enhancement, (d) Corrective human-in-the-loop teleoperation.

tile, and proprioceptive feedback to autonomously perform force-adaptive grasping for a five-finger hand.

- **End-to-End VLA with Arm-Hand Feature Enhancement:** an end-to-end VLA policy for holistic arm-hand control and coordination using a novel architecture with shared representations plus distinct arm and hand features, capturing both macro- and micro-motions for robust and natural control.
- **Corrective Human-in-the-Loop Teleoperation:** a continuous learning strategy that incorporates both successful trajectories plus human recovery data during deployment for iterative policy refinement.

Experimental results validated our approach: the framework efficiently generates high-quality demonstration data and learns a VLA policy that achieves a 90% success rate on a diverse set of over 50 objects. Ablation studies confirmed that the DexGrasp-VLA model, the feature enhancement module, and the corrective teleoperation system are all critical components, which significantly improve the success rate and robustness.

II. Related Works

A. Vision-Language-Action Models for Robot Control

Recent Vision-Language-Action models [4], [5], [6], [7], [8] have shown remarkable performance for the control of general-purpose robots. Most works [20], [21] are mainly limited to the application of two-finger grippers. While new research is addressing dexterous manipulation [22], [23], these methods typically train the arm and hand policies together as a single system, failing to distinguish between them. This approach overlooks different roles during reaching and grasping: the arm as the floating base is more for long-horizon reaching and moving the hand around, while the multiple fingers of the hand are more for fine-grained grasping. Based on this, our work approaches the dexterous grasping problem differently, by designing the data collection process via shared autonomy between the human and the robot, making the data collection easier and much more efficient, expediting the iteration of VLA models.

B. Data Collection Paradigm

High-quality robot demonstration datasets are fundamental for training policies in imitation learning. Conventional teleoperation methods—such as leader-follower [24], vision-based [25], or VR systems [26], [27]—can capture high-quality data shadowing the human dexterity, but are subject to the operator's skillfulness. In contrast, fully automated paradigms [2], [18], [19], [28], [29] using motion planners, e.g., CuRobo [30], can generate vast amounts of data efficiently. However, such trajectories often lack the fine-grained nuance and generalizability of human demonstrations. In this paper, we propose the AI-Copilot framework, which makes the large-scale, high-fidelity robot data collection much more efficient and easier to operate for ordinary operators.

C. Tactile Sensing for Robot Manipulation

Adding tactile signals provides information about physical interactions in contact-rich manipulation. Prior works typically apply reinforcement learning to fuse vision and touch, achieving success in assembly [31], [32] and dexterous in-hand control [33], [34]. More recently, imitation learning

approaches [35], [36], [37], [38] have emerged, emphasizing joint vision-tactile representation learning for fine-grained control. Some methods further extend this to VLA models [39], [40], [41], [42], improving generalization across tasks. However, such method of injecting tactile information directly into the vision-language model demands large-scale tactile data or separate tactile-language pretraining. In contrast, our method introduces tactile signals only *within* the expert model, avoiding full model re-training and retaining the benefits of visuo-tactile fusion.

III. Methodology

We present an integrated pipeline for learning dexterous arm-hand policies (Fig. 1). First, the DexGrasp-VLA controller performs autonomous grasping using multimodal feedback for the dexterous hand, serving as the core functionality in our shared autonomy framework to make human teleoperation easier. Within this framework, a human operator teleoperates the arm via VR while DexGrasp-VLA acts as an AI copilot, enabling force-adaptive grasping and easing data collection. Using the resulting synchronized hand-arm data, we train a holistic end-to-end policy augmented with an arm-hand feature enhancement module that extracts and fuses dedicated features to improve coordination. Finally, a corrective teleoperation system continuously improves the policy using successful and recovery demonstrations. Together, these components form a complete system for acquiring and refining dexterous manipulation skills.

A. Problem Formulation

Our ultimate goal is to learn an end-to-end VLA model capable of controlling all the joints coherently for coordinated arm-hand grasping. Both DexGrasp-VLA policy for the hand, and the end-to-end arm-hand VLA policy for the hand and arm, are fine-tuned from π_0 [6] using the open-source framework LeRobot [43]. Specifically, the model predicts a sequence of future actions conditioned on the current observation. Let $\mathcal{A}_t = [a_t, a_{t+1}, \ldots, a_{t+H-1}]$ denote a horizon-H action sequence. The model aims to approximate the conditional distribution $\pi(\mathcal{A}_t \mid o_t)$, where o_t is the observation at time t. Each observation o_t includes multiple RGB views, a language command, and proprioceptive states: $o_t = [I_t^1, \ldots, I_t^m, l, q_t]$. Here, I_t^i represents the i-th camera image, l is the tokenized instruction, and q_t is the robot's joint state vector. All modalities are encoded through modality-specific encoders and projected into a shared embedding space for cross-modal reasoning and control.

B. DexGrasp-VLA: Autonomous Dexterous Grasping Policy

Our shared autonomy framework centers on DexGrasp-VLA, a high-performance controller that acts as an AI copilot for dexterous hand grasping. To ensure robustness and generalization, DexGrasp-VLA is developed through a two-stage training pipeline. It begins by learning a hand-only LSTM policy from a hybrid dataset consisting of parameterized force-control and teleoperated demonstrations. This compact, "blind" policy captures rich contact behaviors using tactile

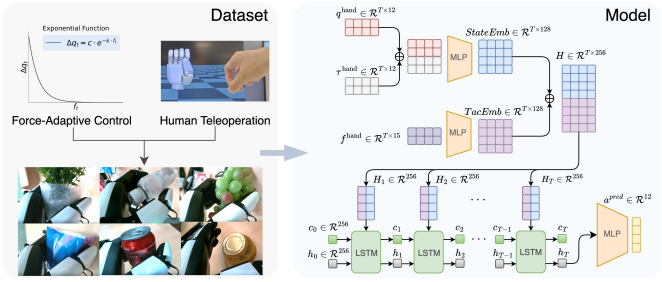

Fig. 2: Force-adaptive grasping policy learned via LSTM using datasets collected by parameterized control and teleoperated human strategies, ensuring data diversity.

sensing and can autonomously collect diverse grasping data. Building upon this, we train a hand-centric VLA policy that further integrates visual and tactile sensing, enabling perceptual grasping that is context-aware and reactive.

*1) **Force-Adaptive Grasping Policy Learned by LSTM**:* To bootstrap the learning of underlying contact dynamics without complex visual perception, a "blind" policy is trained first [44]. This initial stage focuses on learning the reactive closing and force-adaptive gripping of the hand using a hybrid dataset, combining grasping force control and human teleoperation. The controller updates joint commands q_t^{c} as:

$$\Delta q_t = c\, e^{-k f_t}, \quad q_t^{\mathrm{c}} = q_t^{\mathrm{m}} + \Delta q_t, \tag{1}$$

where q_t^{m} is the measured joint angle, f_t the fingertip normal force, c a position scale, and k a force gain. The increment Δq_t decreases with f_t, controlling the hand to close quickly and then gradually increase its grip force for a stable grasp. In addition, human demonstrations are collected via hand retargeting to enhance data diversity. The dataset comprises hand joint positions q_t^{hand}, torques τ_t^{hand}, and tactile forces f_t^{hand}. An LSTM policy (Fig. 2) is then trained via behavior cloning to unify diverse strategies—including parameterized force control and teleoperation—into a compact state-based policy that generalizes across various object shapes and materials. Once trained, this lightweight, fast-to-compute policy autonomously generates diverse and stable grasping trajectories, which are used to collect data quickly for training a hand-only VLA policy.

*2) **Tactile-based DexGrasp-VLA π_{hand}**:* While the LSTM policy provides robust low-level force adaptation, it lacks scene understanding. To incorporate visual context and enable task-aware grasping (e.g., timing to grasp the target object in clutter), we use the data autonomously collected by the LSTM policy to train a multimodal VLA policy that integrates tactile sensing for robust dexterous grasping.

*a) **Tactile Feature Extraction**:* The raw tactile data ($F_{\mathrm{raw}} \in \mathbb{R}^{10\times12\times3}$ for each fingertip) are high-dimensional and unsuitable for feeding into a VLA policy directly. To derive compact and meaningful representations, we extract two complementary tactile features from raw sensor data, as shown in Fig. 3. First, we introduce the resultant force feature $f_t^{\text{tac-f}} \in \mathbb{R}^{5\times3}$, defined as the vector sum of contact forces on each fingertip. This feature explicitly reflects the magnitude of the net forces, offering a direct measure of the

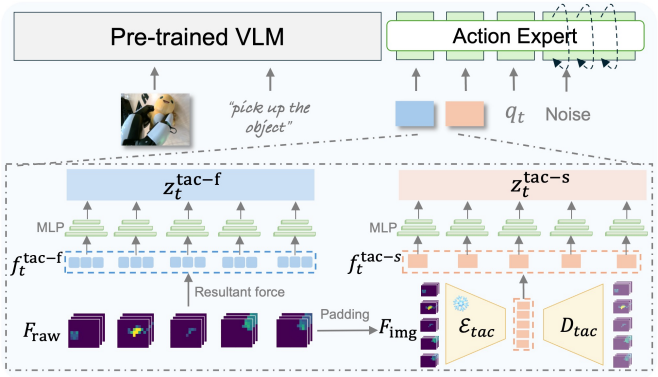

Fig. 3: Tactile-based DexGrasp-VLA for hand control. Two complementary tactile features are extracted: (i) resultant forces, representing the net contact force vector at each fingertip, and (ii) tactile image latents, capturing spatial contact patterns through a convolutional autoencoder.

interaction's strength. However, this single value does not capture the detailed spatial distribution of the contact points of the fingertip surface. To enrich the representation, we introduce a latent embedding $f_t^{\text{tac-s}} \in \mathbb{R}^{5\times128}$ derived from tactile images. Each fingertip signal F_{raw} is first padded to construct the tactile image $F_{\text{img}} \in \mathbb{R}^{16\times16\times3}$. The five tactile images are then encoded by a convolutional autoencoder into compact latent vectors that capture detailed contact patterns. By integrating both tactile features, the policy perceives both physically grounded force magnitude and spatially detailed contact patterns. This dual representation enhances robustness in contact-rich interactions and improves grasp stability across various objects.

*b) **Grasping VLA Policy Learning**:* To adapt the extracted tactile features to the VLA input space, the per-fingertip features $f_t^{\text{tac-f}}$ and $f_t^{\text{tac-s}}$ are processed through MLPs into tactile embeddings $z_t^{\text{tac-f}}$ and $z_t^{\text{tac-s}}$, which are then fused with other pre-embedded modalities and fed into the action expert model, enabling the model to learn multimodal representations for manipulation. The observation space of the hand-level VLA policy is defined as: $o_t^{\text{hand}} = [I_t^{\text{hand}}, l_t^{\text{hand}}, q_t^{\text{hand}}, z_t^{\text{tac-f}}, z_t^{\text{tac-s}}]$, where I_t^{hand} denotes the eye-in-hand camera image, l_t is the language command, q_t^{hand} is the hand joint state, and $z_t^{\text{tac-f}}, z_t^{\text{tac-s}}$ are the embeddings derived from our dual tactile features. This approach enables $\pi_{\text{hand}}(\mathcal{A}_t^{\text{hand}} \mid o_t^{\text{hand}})$ to produce *firm* grasps across various objects, benefiting from both the expert's sense of touch and rich multimodal context.

C. Shared Autonomy for Data Collection

Building upon DexGrasp-VLA, we introduce a shared autonomy framework for efficient data collection, combining the autonomous grasping capability of DexGrasp-VLA with global guidance from the human. The framework divides control strategically: a human operator teleoperates the robotic arm's end-effector via a VR interface for navigation and positioning, while the pre-trained DexGrasp-VLA policy autonomously controls the dexterous hand for fine grasping. This approach dramatically reduces the operator's cognitive load by eliminating the need to simultaneously coordinate both the high-DoF arm and hand, while preserving the naturalness and quality of the demonstrations.

*1) **Arm Teleoperation System Based on VR Headsets**:* The human operator focuses solely on controlling the 6-DoF pose of the arm's end-effector. A VR-based teleoperation system has been developed on the foundation of the XRoboToolkit [27], to achieve intuitive and seamless manipulator control. When the clutch button is pressed, the initial poses of the controller and end-effector, $T_{\text{c},0}$ and $T_{\text{e},0}$, are recorded. The controller's current pose $T_{\text{c},t}$ defines the target end-effector pose as:

$$T_{\text{e},t} = T_{\text{e},0} \cdot (T_{\text{c},0}^{-1} \cdot T_{\text{c},t}), \tag{2}$$

which is resolved via inverse kinematics (IK) to ensure smooth and intent-aligned motion.

*2) **Coordinated Arm-Hand Data Collection**:* The data collection process employs a dual-threaded architecture that seamlessly integrates human teleoperation with autonomous policy execution. While the operator controls the arm's motion through the VR interface, the DexGrasp-VLA policy runs concurrently to generate appropriate grasping actions based on real-time visual and tactile feedback. This parallel execution enables natural and efficient collection of coordinated arm-hand demonstrations. The resulting dataset incorporates temporally synchronized observations and actions from both control sources:

$$\begin{aligned} \mathcal{D}_{\text{uni}} &= \{(o_t^{\text{uni}}, a_t^{\text{arm}}, a_t^{\text{hand}})\}_{t=1}^{T}, \\ a_t^{\text{arm}} &\sim p_{\text{teleop}}, \quad a_t^{\text{hand}} \sim \pi_{\text{hand}}(\cdot \mid o_t^{\text{hand}}), \end{aligned} \tag{3}$$

where the combined observation vector is formally defined as $o_t^{\text{uni}} = [I_t, l_t, q_t^{\text{arm}}, q_t^{\text{hand}}]$, incorporating multi-view RGB images I_t, language instruction l_t, and the joint states of the arm q_t^{arm} and hand q_t^{hand}. This comprehensive dataset provides the foundation for training end-to-end arm-hand manipulation policies that learn effective coordination strategies from human-guided arm motions and autonomous hand actions.

D. Learning End-to-End Arm-Hand VLA Policy π_{uni}

Building upon the arm-hand demonstration data collected through our shared autonomy framework, we perform SFT of π_0 to learn an arm-hand coordinated dexterous grasping policy $\pi_{\text{uni}}(\mathcal{A}t^{\text{uni}} \mid o_t^{\text{uni}})$. Unlike the hand-only policy π_{hand}, tactile information is not included in o_t^{uni} to avoid disrupting stable arm-centric coordination during reaching. A key challenge lies in effectively handling the distinct characteristics of arm motion and hand movement. To address this, we introduce an Arm-Hand Feature Enhancement module that extends the base architecture with explicit mechanisms to capture both shared task context and individualized features for arm and hand dynamics.

*1) **Arm-Hand Feature Enhancement**:* As illustrated in Fig. 4, we extend the base π_0 model, which encodes multimodal observations and language instructions into a shared task representation $z_t^{\text{share}} \in \mathbb{R}^{d_s}$ using `PaliGemma` and `Gemma Expert` and relies on z_t^{share} to predict actions. To better handle the dual challenges of high-DoF dexterous hand

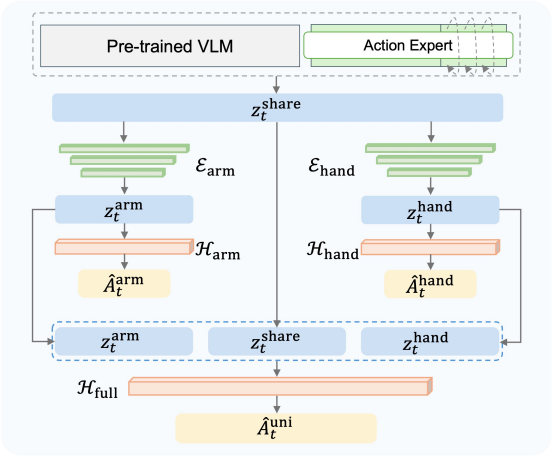

Fig. 4: Arm-hand feature enhancement for the end-to-end VLA policy. z_t^{share} is extracted and distilled into distinctively enhanced z_t^{arm} and z_t^{hand} via specific encoders and auxiliary predictors, and then the main action head fuses all representations to produce coherent and precise joint actions.

manipulation and large-workspace arm motion, we introduce Arm-Hand Feature Enhancement, as shown in Fig. 4. Specifically, the shared context z_t^{share} is processed by two MLPs, $\mathcal{E}_{\text{arm}}$ and $\mathcal{E}_{\text{hand}}$, producing task-specific latent vectors $z_t^{\text{arm}} \in \mathbb{R}^{d_a}$ and $z_t^{\text{hand}} \in \mathbb{R}^{d_h}$. Each latent vector is guided by an auxiliary prediction head, $\mathcal{H}_{\text{arm}}$ and $\mathcal{H}_{\text{hand}}$, which supervise the corresponding sub-actions $\hat{\mathcal{A}}_t^{\text{arm}}$ and $\hat{\mathcal{A}}_t^{\text{hand}}$ to enforce arm-specific and hand-specific latent feature learning. For coordinated execution, the main action head $\mathcal{H}_{\text{main}}$ consumes the fused representation $z_t^{\text{fused}} = [z_t^{\text{share}}, z_t^{\text{arm}}, z_t^{\text{hand}}]$, and predicts the final joint action $\hat{\mathcal{A}}_t^{\text{uni}} = [\hat{\mathcal{A}}_t^{\text{arm}}, \hat{\mathcal{A}}_t^{\text{hand}}]$. The direct connection from z_t^{share} ensures that the global semantic context is preserved. This design allows $\mathcal{H}_{\text{main}}$ to adaptively balance global strategy with local precision, enabling coordinated actions that are both coherent and fine-grained.

*2) **Learning Objective**:* The model is trained using a composite loss that integrates a primary coordinated action generation loss with two auxiliary expert-specific losses.

*a) **Main Loss**:* The primary objective, $\mathcal{L}_{\text{main}}$, follows the conditional flow matching formulation (as in π_0) and is applied to the concatenated action chunk vector $\mathcal{A}_t^{\text{uni}} = (\mathcal{A}_t^{\text{arm}}, \mathcal{A}_t^{\text{hand}})$:

$$\mathcal{L}_{\text{main}}^{\tau}(\theta) = \mathbb{E}\left\| \mathcal{H}_{\text{main}}(z_t^{\text{fused}}) - u(\mathcal{A}_t^{\tau,\text{uni}} \mid \mathcal{A}_t^{\text{uni}}) \right\|^2, \quad (4)$$

where $\mathcal{A}_t^{\tau} = \tau \mathcal{A}_t + (1-\tau)\epsilon$ is the noisy action chunk, and $u(\mathcal{A}_t^{\tau} \mid \mathcal{A}_t) = \epsilon - \mathcal{A}_t$ denotes the target vector field.

*b) **Auxiliary Expert Losses**:* To supervise and enhance the individual, distinct features of the arm and the hand, two auxiliary objectives are introduced:

$$\mathcal{L}_{\text{hand}}^{\tau}(\theta) = \mathbb{E}\left\| \mathcal{H}_{\text{hand}}(z_t^{\text{hand}}) - u_{\text{hand}}(\mathcal{A}_t^{\tau,\text{hand}} \mid \mathcal{A}_t^{\text{hand}}) \right\|^2. \quad (5)$$

$$\mathcal{L}_{\text{arm}}^{\tau}(\theta) = \mathbb{E}\left\| \mathcal{H}_{\text{arm}}(z_t^{\text{arm}}) - u_{\text{arm}}(\mathcal{A}_t^{\tau,\text{arm}} \mid \mathcal{A}_t^{\text{arm}}) \right\|^2. \quad (6)$$

*c) **Total Loss**:* The final training objective combines the main and auxiliary losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda\Big(\mathcal{L}_{\text{hand}} + \mathcal{L}_{\text{arm}}\Big), \quad (7)$$

where λ is a weighting factor balancing global coordination with individual motions of the arm and the hand.

E. Corrective Human-in-the-Loop Teleoperation System

To enable robust deployment of robotic policies in unstructured real-world environments, where distribution shifts and long-tail scenarios often occur, we implement a corrective teleoperation system. The core of it is an incremental SFT framework that enables the policy to continuously learn and adapt from real-world data. This framework uses a human-in-the-loop corrective intervention paradigm for data collection and policy refinement. Specifically, during real-world deployment, successful trajectories executed by the unified arm-hand policy π_{uni} are recorded as positive demonstration data. When the policy fails or exhibits suboptimal behavior, the system triggers a shared-autonomy paradigm that enables human-in-the-loop corrective intervention. A human operator then takes over via teleoperation to recover from the failure. After successful recovery, the system seamlessly transitions back to its autonomous inference mode. The data from both the initial failure and the successful recovery is captured as a valuable corrective demonstration. This process forms an improvement cycle that leverages both successful and failed trials, allowing the policy to continually enhance its robustness and task performance.

IV. Experimental Validation

A. Experimental Settings

*1) **Robot System**:* Our grasping platform comprises a UR3e robotic arm paired with a five-fingered dexterous hand `Xhand` [45], which features 12 DoFs and is equipped with 120 triaxial force sensors on each fingertip. The vision module integrates three RGB-D cameras: two Intel RealSense D435i cameras in third-person view for global scene perception, and one eye-in-hand Intel RealSense D405 for close-range observations.

*2) **Datasets**:* We construct different datasets to train different stages within our framework. For the LSTM-based grasping policy, 218 trajectories are collected, including 150 human teleoperation demonstrations and 68 autonomous trajectories generated by a force-adaptive controller. The hand-only DexGrasp-VLA dataset comprises 180 cluttered-scene grasping trajectories over 60 objects, collected using the trained LSTM grasping policy. For arm-hand VLA fine-tuning, 100 single-object demonstrations across 20 everyday objects are collected through shared autonomy, forming $\mathcal{D}_{\text{uni}}$. Each trajectory contains synchronized RGB observations, joint states, and action sequences. To support iterative human-in-the-loop refinement, two additional corrective datasets are constructed: $\mathcal{D}$orient (50 trajectories for orientation failures) and $\mathcal{D}$corner (50 trajectories for other challenging corner cases).

B. Main Results

*1) **Grasping Performance of π_{hand}**:* We evaluate the hand-only DexGrasp-VLA policy on cluttered tabletop scenarios, where the task is to clear all objects from the

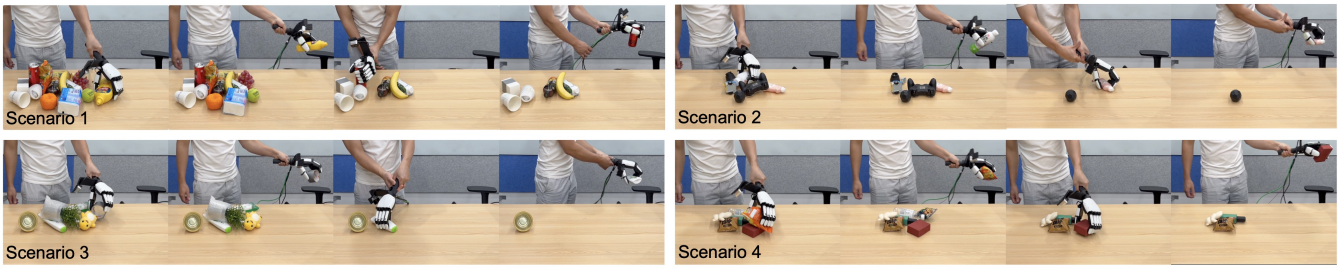

Fig. 5: Table bussing in clutter via Dex-Grasp VLA policies for a dexterous robotic hand–validated capability as an AI copilot with local autonomous tactile-based robust grasping.

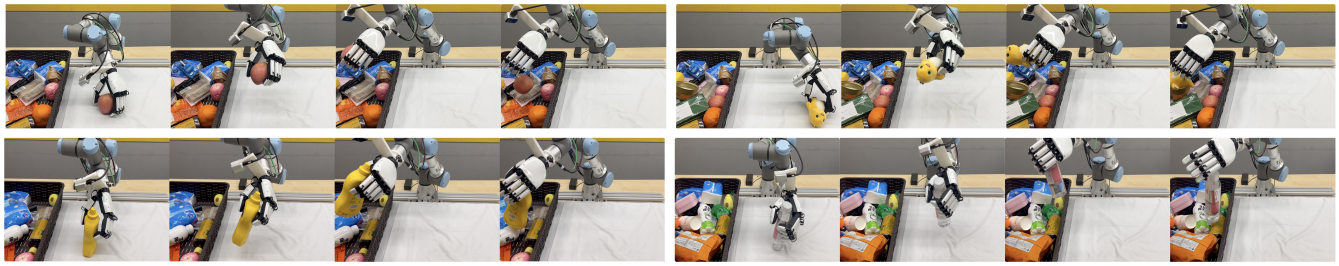

Fig. 6: Grasping and placing objects via end-to-end (arm-hand) VLA policies with a dexterous hand.

TABLE I: Success rates(%) of the final end-to-end arm-hand VLA policy π_{uni}-final across 50 objects.

Methods	Seen objects	Unseen objects	Average
π_{uni}-final	91.7	85.6	88.7

table sequentially. The evaluation includes five randomly arranged scenes, comprising a total of over 50 objects with varying sizes, colors, and materials. The policy achieves an overall success rate of 95.5%, successfully clearing most objects in each scene. Fig. 5 shows some representative examples. These results demonstrate that the hand VLA can robustly handle complex, densely cluttered environments and generalize effectively across diverse object properties.

*2) **Grasping Performance of** π_{uni}:* We evaluate the arm-hand VLA policy, π_{uni}, on a pick-and-place task involving 20 seen and more than 30 unseen objects of varying shapes. Fig. 6 illustrates the grasping results. Each object is tested in 3 trials with randomized positions and orientations within a 40cm × 40cm workspace. In each trial, the robot attempts to grasp the object and place it in a target basket. A trial is deemed successful if the object is securely grasped and accurately placed without slipping or being dropped. Table I summarizes the success rates of our policy, which achieves an average of about 90%. The model demonstrates consistently high performance on familiar objects, reflecting effective hand-arm coordination and stable grasping capabilities. Furthermore, π_{uni} generalizes robustly to novel objects and challenging orientations, enabling reliable pick-and-place across diverse geometries and configurations.

*3) **Data Collection Efficiency of Shared Autonomy**:* Our shared autonomy framework improves data collection efficiency. A single operator collects 110 trajectories per hour for the main dataset, compared to 90 under full teleoperation (+22.2%), and 100 trajectories per hour for corrective data, versus 80 with teleoperation (+25%). These improvements accelerate policy iteration and underscore the practical advantages of the shared autonomy paradigm.

C. Ablation study

*1) **Effectiveness of Tactile Sensing in** π_{hand}:* To evaluate how tactile information improves grasp robustness in DexGrasp-VLA, we designed a two-stage test. After grasping an object with the trained policy, the robot must first hold it steadily for 3 seconds, then continue to hold it for another 10 seconds after the camera input is occluded. A grasp is counted as successful only if the object remains secure throughout both stages. This criterion highlights the role of tactile sensing in ensuring stability when vision is unreliable. As shown in Table II, tactile feedback substantially improves performance. Without tactile input, the policy achieves only a 21% average success rate, while adding force-based feedback raises the success rate to 70%. With both tactile features combined, the success rate reaches 90%, confirming that spatially detailed and magnitude-aware tactile signals are crucial for maintaining stable grasps under visual occlusion. Fig. 7 demonstrates this effect: while vision is occluded, without tactile feedback, objects slip within seconds, whereas with dual tactile features, the dexterous hand can hold firmly for the full duration—even under perturbations such as intense shaking or knocking against the table. Contact distributions of tactile sensors during grasping are shown in Fig. 8. These tests confirm that tactile sensing not only compensates for temporary visual loss but also provides essential information for keeping firm gripping forces.

*2) **Effectiveness of Arm-Hand Feature Enhancement in** π_{uni}:* We evaluate the enhanced policy π_{uni}-enhance (with Arm–Hand Feature Enhancement) against the baseline π_{uni}-origin (the original π_0 policy)—both fine-tuned

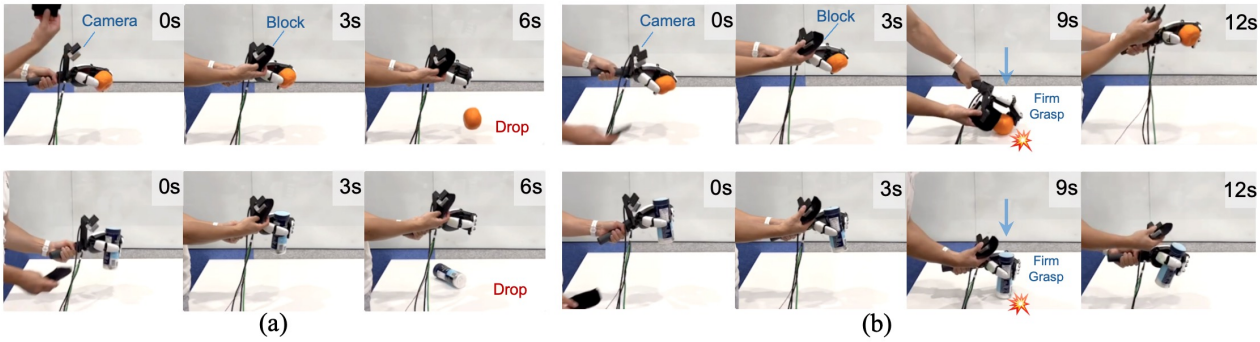

Fig. 7: Effectiveness of tactile-based DexGrasp-VLA π_{hand} and robust grasping performance with tactile sensing under camera occlusion: (a) Without the sense of touch (tactile), the object slips while camera is occluded; (b) With dual tactile features, the policy sustains firm and stable grasp throughout the test without any slippage, even under perturbations such as bumping against the table (9s) and intense shaking (9s-12s).

TABLE II: Success rates of dexterous grasping policy π_{hand} across 10 everyday objects under different tactile configurations.

Methods	Salt Can	Bottle 1	Bottle 2	Apple	Orange	Banana	Bowl	Mug	Ball	Gamepad	Average
π_{hand}-origin(π_0 [6])	1/10	2/10	2/10	0/10	5/10	0/10	1/10	5/10	3/10	2/10	21%
π_{hand}-tacf	8/10	8/10	7/10	6/10	9/10	4/10	7/10	9/10	6/10	6/10	70%
π_{hand}-tacf-tacs	8/10	10/10	10/10	8/10	10/10	7/10	9/10	10/10	10/10	8/10	90%

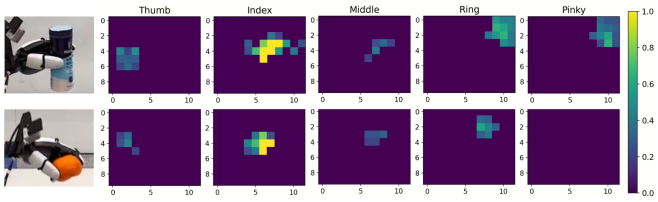

Fig. 8: Representative cases showing the contact distributions of tactile sensors.

on $\mathcal{D}_{\text{uni}}$—under varied conditions. As shown in Tab. III, which reports results from three experimental groups each testing 10 seen and unseen objects over 10 trials per object within a 20cm × 20cm region, π_{uni}-enhance consistently outperforms π_{uni}-origin on both `Xhand` and `RY-H2` [46] dexterous hand, with gains of 7% and 10% respectively. To evaluate robustness under degraded perception, we block the right-view camera. This occlusion reduces visibility in the grasping area, impairing shape cues near contact regions. In this setting, the success rate of π_{uni}-origin drastically drops to 19% , while π_{uni}-enhance still retains 58% performance. Such notable improvement confirms that our method enhances robustness in perception-limited scenarios. The occlusion experiment is particularly well-suited to validate the necessity of enhancing each individual latent feature, as it selectively disrupts visual perception, thereby compelling the policy to rely on the quality and completeness of its internal kinematic representations. The enhanced model's ability to compensate for missing views suggests that its arm and hand branches have learned distinct, complementary roles—such as robust reaching and contact strategies. These evidences suggest that our Arm-Hand Feature Enhancement enforces coordinated actions, compared to a standard architecture *without* explicit individual features extracted for the arm and hand.

TABLE III: Grasp success rates(%) on 10 objects of the end-to-end arm-hand VLA policy π_{uni}.

Methods	**Xhand**	**RY-H2**	**Xhand-Occlude**
π_{uni}-origin(π_0 [6])	88	71	19
π_{uni}-enhance	95	81	58

TABLE IV: Grasp success rates(%) across five everyday objects under corrective human-in-the-loop teleoperation.

Methods	Bottle	Apple	Nailoong	Chips	Bowl	Average
π_{uni}-enhance	4/10	4/10	5/10	4/10	3/10	40%
π_{uni}-orient	6/10	6/10	6/10	4/10	4/10	52%
π_{uni}-final	8/10	9/10	9/10	9/10	9/10	88%

*3) **Effectiveness of Corrective Teleoperation**:* To evaluate the efficacy of our corrective human-in-the-loop teleoperation system, we designed a challenging experimental setup featuring a larger workspace (40 cm × 40 cm). This configuration evaluates policy robustness through varied object placements (arranged in a 3×3 grid), orientations, and states (upright and inverted). As detailed in Table IV, the π_{uni}-enhance model, which is trained exclusively on the initial dataset $\mathcal{D}_{\text{uni}}$, failed to handle specific orientations and all corner-case scenarios. After one iteration, π_{uni}-orient was developed by fine-tuning the base model with 50 corrective trajectories $\mathcal{D}_{\text{orient}}$ for orientation failures. This model demonstrated improved generalization to varied orientations and exhibited emergent error recovery behaviors, yet it was still unable to address the corner cases. Finally, the π_{uni}-final model was further fine-tuned with an additional 50 trajectories $\mathcal{D}_{\text{corner}}$ targeting corner-case scenarios, which successfully generalized across all tested scenarios. These results clearly demonstrate that our system can precisely identify and inject data for corner

cases, thereby significantly enhancing the policy's robustness in complex scenarios through continuous iterations.

V. Conclusion

This work presents a shared autonomy framework that combines human teleoperation for global arm motion with an autonomous DexGrasp-VLA hand control using tactile feedback for adaptive grasping, reducing operator cognitive load and thus enabling efficient collection of high-quality demonstrations. Our arm-hand feature enhancement design and corrective human-in-the-loop mechanism further support robust policy learning and continuous improvement. Experiments show a grasping success rate of about 90% across over 50 seen and unseen objects, validating both efficiency and generalization. While this study focuses on grasping tasks, the proposed shared autonomy paradigm—along with its submodules—can be applied to other manipulation skills and long-horizon tasks through additional specialized AI copilots. Future work will explore its generalization to bimanual and tool-use scenarios, robustness under real-world distribution shifts, and improved tactile integration for arm-hand coordination. Overall, this work provides a practical solution for collaborative data collection and outlines promising directions for dexterous manipulation research.

References

[1] A. Rajeswaran, *et al.*, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *arXiv:1709.10087*, 2017.

[2] Y. Chen, *et al.*, "Bi-dexhands: Towards human-level bimanual dexterous manipulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2804–2818, 2023.

[3] J. Ye, *et al.*, "Dex1b: Learning with 1b demonstrations for dexterous manipulation," *arXiv:2506.17198*, 2025.

[4] B. Zitkovich, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[5] M. J. Kim, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv:2406.09246*, 2024.

[6] K. Black, *et al.*, "π0: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550," *arXiv.2410.24164*.

[7] P. Intelligence, *et al.*, "π0. 5: a vision-language-action model with open-world generalization, 2025," *URL https://arxiv.org/abs/2504.16054*, vol. 1, no. 2, p. 3.

[8] J. Bjorck, *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv:2503.14734*, 2025.

[9] A. O'Neill, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation*. IEEE, 2024, pp. 6892–6903.

[10] A. Khazatsky, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv:2403.12945*, 2024.

[11] Q. Bu, *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv:2503.06669*, 2025.

[12] T. Z. Zhao, *et al.*, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv:2304.13705*, 2023.

[13] R. Ding, *et al.*, "Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning," *arXiv:2407.03162*, 2024.

[14] M. Pan, *et al.*, "Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 359–17 369.

[15] A. Curtis, *et al.*, "Trust the proc3s: Solving long-horizon robotics problems with llms and constraint satisfaction," in *Conference on Robot Learning*. PMLR, 2025, pp. 1362–1383.

[16] J. Duan, *et al.*, "Manipulate-anything: Automating real-world robots using vision-language models," *arXiv:2406.18915*, 2024.

[17] W. Huang, *et al.*, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," *arXiv:2409.01652*, 2024.

[18] T. Lin, *et al.*, "Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids," *arXiv:2502.20396*, 2025.

[19] S. Patel, *et al.*, "A real-to-sim-to-real approach to robotic manipulation with vlm-generated iterative keypoint rewards," *arXiv:2502.08643*, 2025.

[20] J. Liu, *et al.*, "Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model," *arXiv:2503.10631*, 2025.

[21] S. Deng, *et al.*, "Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data," *arXiv:2505.03233*, 2025.

[22] J. Wen, *et al.*, "Dexvla: Vision-language model with plug-in diffusion expert for general robot control," *arXiv:2502.05855*, 2025.

[23] Y. Zhong, *et al.*, "Dexgraspvla: A vision-language-action framework towards general dexterous grasping," *arXiv:2502.20900*, 2025.

[24] T. Z. Zhao, *et al.*, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv:2304.13705*, 2023.

[25] Y. Qin, *et al.*, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," *arXiv:2307.04577*, 2023.

[26] X. Cheng, *et al.*, "Open-television: Teleoperation with immersive active visual feedback," *arXiv:2407.01512*, 2024.

[27] Z. Zhao, *et al.*, "Xrobotoolkit: A cross-platform framework for robot teleoperation," *arXiv:2508.00097*, 2025.

[28] W. Thomason, Z. Kingston, and L. E. Kavraki, "Motions in microseconds via vectorized sampling-based planning," in *IEEE International Conference on Robotics and Automation*, 2024, pp. 8749–8756.

[29] N. Kumar, *et al.*, "Open-world task and motion planning via vision-language model inferred constraints," *arXiv:2411.08253*, 2024.

[30] B. Sundaralingam, *et al.*, "Curobo: Parallelized collision-free robot motion generation," in *IEEE International Conference on Robotics and Automation*, 2023, pp. 8112–8119.

[31] J. Hansen, *et al.*, "Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning," in *IEEE International Conference on Robotics and Automation*, 2022, pp. 8298–8304.

[32] M. A. Lee, *et al.*, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582–596, 2020.

[33] Q. Liu, *et al.*, "Vtdexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning," in *The Thirteenth International Conference on Learning Representations*.

[34] W. Hu, *et al.*, "Dexterous in-hand manipulation of slender cylindrical objects through deep reinforcement learning with tactile sensing," *Robotics and Autonomous Systems*, vol. 186, p. 104904, 2025.

[35] H. Xue, *et al.*, "Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation," *arXiv:2503.02881*, 2025.

[36] F. Liu, *et al.*, "Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface," *arXiv:2504.06156*, 2025.

[37] B. Huang, *et al.*, "3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing," *arXiv:2410.24091*, 2024.

[38] T. Lin, *et al.*, "Learning visuotactile skills with two multifingered hands," *arXiv:2404.16823*, 2024.

[39] J. Bi, *et al.*, "Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback," *arXiv:2507.17294*, 2025.

[40] J. Huang, *et al.*, "Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization," *arXiv:2507.09160*, 2025.

[41] Z. Cheng, *et al.*, "Omnivtla: Vision-tactile-language-action model with semantic-aligned tactile sensing," *arXiv:2508.08706*, 2025.

[42] C. Zhang, *et al.*, "Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation," *arXiv:2505.09577*, 2025.

[43] R. Cadene, *et al.*, "Lerobot: State-of-the-art machine learning for real-world robotics in pytorch," https://github.com/huggingface/lerobot, 2024.

[44] S. Wang, *et al.*, "Learning adaptive grasping from human demonstrations," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 5, pp. 3865–3873, 2022.

[45] Xhand1. [Online]. Available: https://www.robotera.com/en/goods1/4.html

[46] Ry-h2. [Online]. Available: http://www.ruiyanrobot.com/product/hand/35