Dexora: Open-source VLA for High-DoF Bimanual Dexterity

***Abstract*— Vision-Language-Action (VLA) models have recently become a central direction in embodied AI, but current systems are restricted to either dual-gripper control or single-arm dexterous hand manipulation. While low-dimensional gripper control can often be handled with simpler methods, high-dimensional dexterous hand control benefits greatly from full end-to-end VLA learning. In this work, we introduce Dexora, the first open-source VLA system that natively targets dual-arm, dual-hand high-DoF manipulation. We design a hybrid teleoperation pipeline that decouples gross arm kinematics (captured with a custom exoskeleton backpack) from fine finger motion (markerless hand tracking via Apple Vision Pro), and that drives both a physical dual-arm dual-hand platform and an identical MuJoCo digital twin. Using that interface we assemble a large training corpus: an embodiment-matched synthetic corpus (100K simulated trajectories, 6.5M frames) and a real-world dataset of 10K teleoperated episodes (177.5 hours, 3.2M frames). To mitigate noisy teleoperation demonstrations, we propose a data-quality-aware training recipe: an offline discriminator provides clip-level weights for diffusion-transformer policy training, down-weighting low-quality demonstrations. Empirically, Dexora outperforms competitive VLA baselines on both basic and dexterous benchmarks (e.g., average dexterous success 66.7% vs. 51.7%), attains $\geq$90% success on basic tasks, and shows robust out-of-distribution and cross-embodiment generalization. Ablations confirm the importance of real data and the discriminator for dexterity. Demos, data, codes, and models can be found at https://dexoravla.github.io**

I. Introduction

Vision-Language-Action (VLA) models have emerged as a promising paradigm for embodied AI, yet existing systems remain fundamentally constrained: they are either designed for dual-arm, low-DoF grippers or single-arm dexterous hands, but not both [1]–[7]. As illustrated in Fig. 1 (top), such limitations prevent prior VLAs from handling tasks that intrinsically demand dual-arm coordination (e.g., piston insertion), or high-DoF dexterous fingers (e.g., bottle opening/complex book retrieval). *Dexora* is the first open-source VLA that addresses this gap by unifying dual-arm, dual-hand, and high-DoF dexterity into a single system (Fig. 2).

To enable such complex skill acquisition, *Dexora* introduces a hybrid teleoperation pipeline. Gross arm kinematics are captured with a lightweight exoskeleton backpack, while fine-grained finger articulation is driven by markerless hand tracking via Apple Vision Pro. This decoupling makes it feasible to control a physical dual-arm dual-hand platform with 36 DoF, while simultaneously mirroring demonstrations in a MuJoCo-based digital twin, thereby ensuring scalable and embodiment-matched data collection.

Using this interface, we construct a large-scale dataset for dual-arm, dual-hand dexterous manipulation (Fig. 1, §III-B). It consists of 100K simulated trajectories (361 hours, 6.5M frames) and 10K real teleoperated episodes (177.5 hours, 3.2M frames). The design follows the principle of sim-real complementarity: simulated data provide scale and task diversity, while real data provides fine-grained realism essential for high-DoF bimanual dexterity. Together, this dataset establishes a foundation for training VLA models under realistic dexterous settings.

A key challenge of teleoperated data is the presence of noisy or unstable demonstrations (Fig. 1, §III-C). To address this, *Dexora* employs discriminator-guided quality-aware training: an offline discriminator scores each demonstration, and the policy is trained with weighted diffusion-

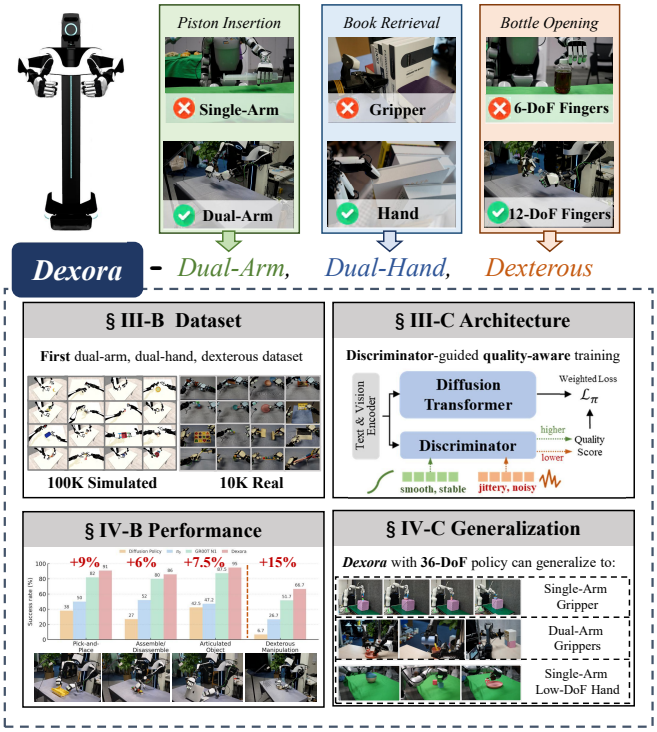

Fig. 1. ***Dexora* overview**. (a) **Motivation**: Three illustrative contrasts highlight the need for dual-arm, dual-hand dexterous VLA: piston insertion (requires two arms), book retrieval from a packed shelf (hands with fingers succeed where grippers fail), and bottle opening (12-DoF fingers with lateral swing outperform 6-DoF). (b) **Dataset** (§III-B): We pretrain on **100K** simulated bimanual-hand trajectories and post-train on **10K** real demonstrations, all collected with our dual-arm, dual-hand platform. (c) **Architecture** (§III-C): A trained discriminator scores dataset demonstration quality and guides training, driving the diffusion-transformer policy to prioritize high-quality trajectories while down-weighting low-quality ones. (d) **Performance** (§IV-B): *Dexora* achieves consistently higher average success rates on both basic (Pick-and-Place, Assemble/Disassemble, Articulated Object) and dexterous benchmarks compared to state-of-the-art VLA models. (e) **Embodiment generalization** (§IV-C): The same policy transfers across **single-arm gripper**, **dual-arm grippers**, and **single-arm low-DoF hand** without re-architecting the model.

Manuscript 209 submitted to 2026 IEEE International Conference on Robotics and Automation (ICRA). Received September 16, 2025.

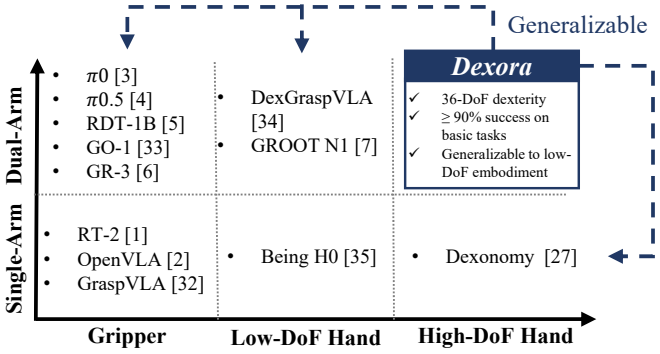

Fig. 2. **Comparison of embodiment coverage.** Prior works cover either single-arm or low-DoF dual-arm settings. *Dexora* is the first system positioned in the dual-arm, high-DoF dexterous region, while also generalizing across simpler embodiments without re-architecture.

transformer loss that down-weights low-quality clips. This design effectively stabilizes learning, ensuring that the policy benefits from large-scale data while mitigating the impact of teleoperation artifacts.

We evaluate *Dexora* across both basic manipulation and dexterous benchmarks (Fig. 1, §IV-B). Quantitatively, *Dexora* achieves over 90% success on basic pick-and-place and open articulated objects tasks, while improving dexterous success from 51.7% (baseline) to 66.7% (+15%). Qualitatively, the system demonstrates torsional manipulation and complex dual-arm coordination. These results highlight the critical role of both real-world data and quality-aware training in attaining high-DoF dexterity.

Finally, *Dexora* exhibits strong generalization beyond its native embodiment (Fig. 1, §IV-C). Despite being trained on a 36-DoF dual-arm dual-hand platform, the learned policy successfully transfers to single-arm gripper, dual-arm grippers, and single-arm low-DoF hand. This suggests that VLA policies trained under rich dexterous settings can serve as universal controllers, generalizing across embodiments. Fig. 2 situates this result in the broader landscape: prior VLAs mainly focus on single-/dual-arm grippers or low-DoF hand. *Dexora* is positioned in the dual-arm, high-DoF hands quadrant while remaining downward-compatible to the other regions of the grid. This suggests a practical route to universal controllers: train in the dexterous, high-DoF setting and deploy by projecting to simpler robots.

II. Related Work

A. Teleoperation System

Teleoperation enables us to acquire large-scale robot demonstrations by translating human motions into robot-executable control signals. Existing platforms can be categorized into five classes: (i) leader–follower systems with kinesthetic teaching rigs [8], [9]; (ii) VR/MR headset–based pose tracking (e.g., Vision Pro pipelines) [10], [11]; (iii) vision-only retargeting [12], [13]; (iv) exoskeleton interfaces for joint-level arm and finger tracking [14], [15]; and (v) joystick/button controllers [16], [17]. We adopt a hybrid teleoperation setup: exoskeletons provide precise arm-level kinematics, while the Vision Pro offers convenient, high-resolution capture of fine-finger motions. This combination produces high-DoF, dual-arm and dual-hand demonstrations that are both accurate and operator-friendly, and are natively compatible with Vision-Language-Action (VLA) model training [18], [19].

B. Dexterous Manipulation

Dexterous manipulation includes grasping, in-hand reconfiguration, tool use, and coordinated bi-manual skills [20], [21]. Prior research generally falls into two categories: *grasp synthesis* and *policy learning*. **On the synthesis side**, the field has undergone a paradigm shift from analytical sampling to generative modeling. Diffusion [22], [23], normalizing flows [24], and latent generative models such as VAE [25], [26], complemented by optimization-based pipelines [27], now enable scalable production of physically consistent grasps across diverse hands and objects. **On the policy side**, reinforcement learning [28] and imitation learning [29] have driven progress toward closed-loop robustness and sim-to-real transfer in high-DoF hands. Emerging *data engines* leverage automated imitation [30] and egocentric supervision [31] to expand coverage, accelerating policy learning at unprecedented scale. Despite this rapid progress, most pipelines remain hand-centric, reward-sensitive, and limited in multi-arm coordination. In contrast, we pursue a vision-language-action (VLA) model that operates in **dual-arm, dual-hand** high-dimensional action space.

C. Vision-Language-Action (VLA) Model

Vision-Language-Action (VLA) models have recently emerged as a promising paradigm yet most existing systems remain confined to low-DoF or single-arm embodiments. Representative efforts such as RT-2 [1], OpenVLA [2], and GraspVLA [32] output manipulation policies for single-arm grippers. More recent generalist policies extend to bimanual settings—e.g., π_0 [3], $\pi_{0.5}$ [4], RDT [5], GO-1 [33], GR-3 [6], GR00T [7], and DexGraspVLA [34]—but these typically simplify embodiment to parallel-jaw grippers, limiting dexterity. In parallel, large-scale data engines such as Being-H0 [35] and DreamGen [36] have enriched supervision, but they still fall short of enabling **high-DoF** dual-hand control.

Our work introduces a **dual-arm, dual-hand high-DoF VLA** that learns to output synchronized arm–hand trajectories end-to-end. The formulation admits natural downshifting to lower-DoF embodiments via finetuning, offering a unified pathway toward cross-embodiment generalization.

III. Dexora

In this section, we first introduce the hardware setup and teleoperation system (Sec. III-A), followed by the construction of our dataset, assembling an embodiment-aligned corpus of large-scale synthetic and real-world demonstrations (Sec. III-B). We then present the VLA framework with a learned data-quality discriminator that scores demonstrations and weights training (Sec. III-C). Finally, we specify the three-stage data-quality-aware training recipe (Sec. III-D).

Manuscript 209 submitted to 2026 IEEE International Conference on Robotics and Automation (ICRA). Received September 16, 2025.

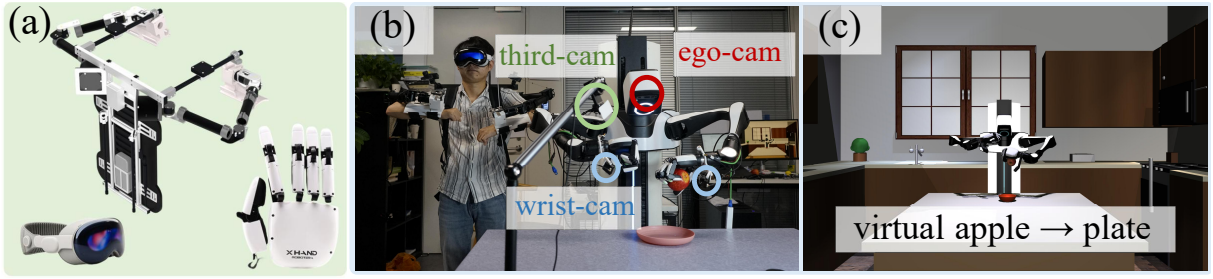

Fig. 3. **Hardware and teleoperation system**. (a) Hybrid teleoperation interface and 12-DoF XHAND. (b)-(c) The operator teleoperates the physical robot and its MujoCo digital twin, so *apple→plate* demonstrations are collected in real and simmulation under the same interface, thereby reducing the sim-to-real gap.

A. Dual-Arm Dual-hand System

As shown in Fig. 3 (a), *Dexora* integrates two 6-DoF AIRBOT arms with a pair of XHAND dexterous hands, each offering 12 fully actuated joints. All finger joints are independently driven, and the thumb and index additionally support lateral ab/adduction, enabling human-like in-hand reorientation and torsional manipulation (e.g., cap twisting).

To achieve scalable teleoperation, we decouple gross arm motion from fine finger control. A custom dual-arm exoskeleton backpack captures the operator's shoulder–elbow–wrist angles and maps them directly to robot joint space. This design yields drift-free, low-latency trajectories while avoiding the inverse-kinematics jitter and singularities that often degrade vision-only retargeting pipelines. Apple Vision Pro provides markerless 3D finger skeletons that we retarget to XHAND with a short calibration phase while enforcing joint limits and safety constraints. This hybrid interface combines the precision of joint-space control for the arms and the convenience of lightweight, glove-free finger input, making long data-collection sessions practical (Fig. 3 (a)).

Our interface drives both the physical robot and a MuJoCo digital twin of the same embodiment. All sensing streams share a time-aligned I/O system: four RGB views and full 36-DoF joint states are logged at 20 Hz. The twin mirrors the real robot's kinematics and controllers, and the same teleop drivers run in real and sim, yielding low latency and high fidelity; operators can switch seamlessly between hardware and simulation to collect demonstrations (Fig. 3 (b)-(c)).

B. Dataset Construction

Synthetic Data. We generate a large, embodiment-matched simulation corpus in MuJoCo. Using Qwen2.5-VL [37], we mine Objaverse [38] to select manipulable objects and automatically assign physical parameters (Fig. 4 (a)). On top of this, we build a set of 200 tasks covering three basic families in Fig. 4 (c). For each task, we collect 3–5 teleoperated seed demonstrations and follow the DexMimicGen [30] recipe to synthesize trajectories: we randomize initial states and retarget the seed actions to new scenes, yielding 500 trajectories per task. Scene layouts and success criteria are auto-generated by Qwen. All simulated episodes are logged with the same observation–action protocol as in the real system, which keeps the interface consistent and reduces the sim-to-real gap. In total, the synthetic set contains about 6.5M frames, 361h video.

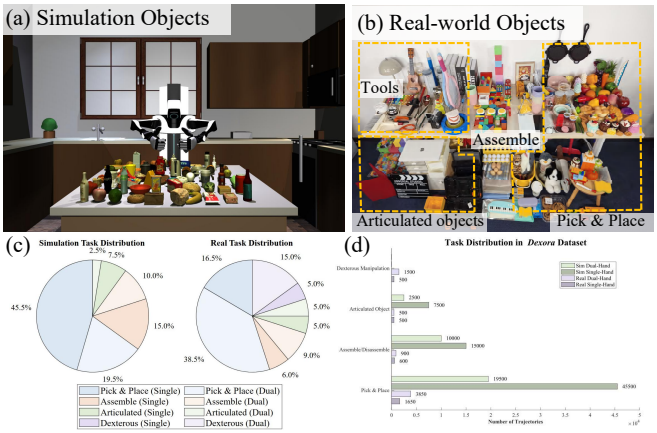

Fig. 4. **Dataset demonstration**. (a) Simulation objects subset: our simulator includes 297 objects across 30 categories. (b) Real-world objects (347 objects, 17 categories), covering both basic and dexterous use cases. (c) Per-family task distribution in simulation vs. real. The simulation data only includes basic tasks, while the real-world set shifts weight toward dexterity (20%). (d) Trajectory counts per family and embodiment (sim/real; single-/dual-hand).

Real World Data. We collect real-world data on the same embodiment used in the simulation. Beyond common objects and basic tasks, we add dexterous tool-use scenarios that are difficult to stage in simulation (Fig. 4 (b)) and the dexterous scenes in Fig. 4 (c)–(d). In total, we curate 200 tasks and acquire 50 teleoperated demonstrations per task via the hybrid teleoperation interface, yielding 10K episodes. The dataset amounts to 177.5 hours and 3,195,795 frames. All recordings are converted to the LIBERO-2.0 standard and open source. We use this to fine-tune the VLA to specialize basic competence into dexterous, bimanual skills.

C. Framework

Data Quality Criteria. Real-world teleoperation demonstrations exhibit substantial variability due to operator skill, sensing noise, inherent limitations (such as occlusion during hand keypoint tracking), and latency. Training on such heterogeneous data without constraints often degrades policy learning. We therefore establish **episode-level** quality criteria with two pillars: (i) kinematic smoothness and steadiness, proxied by low acceleration A_{ep} and jerk J_{ep}—for pre-screening; (ii) replay success as the decisive indicator of data

Manuscript 209 submitted to 2026 IEEE International Conference on Robotics and Automation (ICRA). Received September 16, 2025.

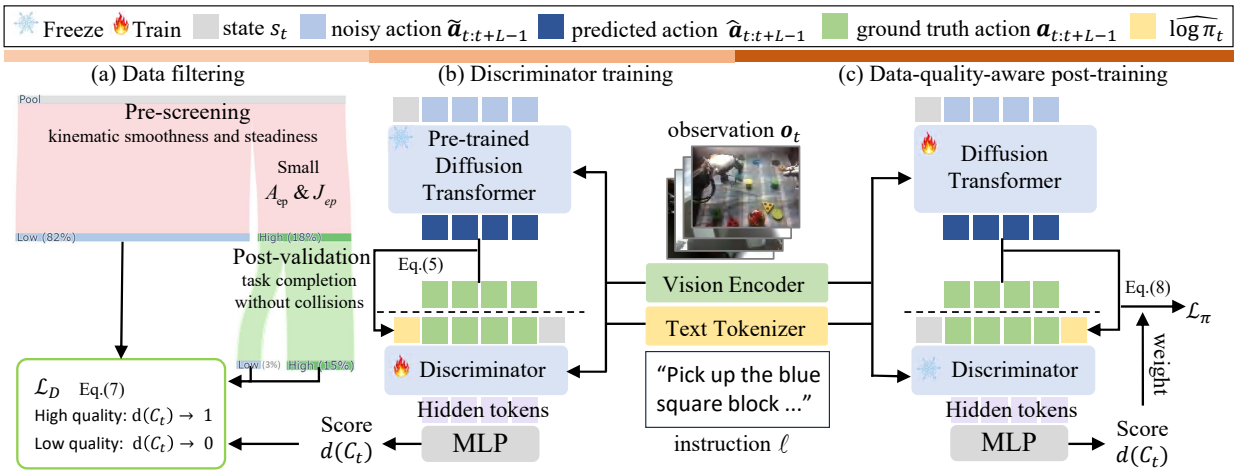

Fig. 5. ***Dexora* framework.** (a) **Data filtering**: From the real-world dataset we pre-screen demonstrations by kinematic smoothness (low acceleration and jerk), then replay them for post-validation and keep the clips that complete the task without collisions, forming a high-quality subset. (b) **Discriminator training**: With the pretrained diffusion–transformer policy frozen, we compute a log-π proxy for each clip and train a discriminator that, conditioned on observations and language, outputs a quality score $d(C_t) \in (0,1]$. (c) **Data-quality-aware post-training**: During post-training, the score $d(C_t)$ is converted to weights w_i and used in the diffusion loss $\mathscr{L}_\pi$. At inference time, only the policy is used.

reliability (task completion without collisions)—for post-validation. This two-stage design yields a clean positive set for training the discriminator (Fig. 5 (a)).

Let an episode be denoted by $\tau = \{s_t\}_{t=1}^{T}$, where $s_t \in \mathbb{R}^D$ is the proprioceptive state vector ($D = 36$). The sampling interval is Δt. Because state dimensions have heterogeneous numeric ranges, we first apply per-dimension min–max normalization. We compute velocity, acceleration, and jerk using centered finite differences ($t = 4, \ldots, T-3$):

$$v_t = \frac{s_{t+1} - s_{t-1}}{2\Delta t}, \quad a_t = \frac{v_{t+1} - v_{t-1}}{2\Delta t}, \quad j_t = \frac{a_{t+1} - a_{t-1}}{2\Delta t}, \tag{1}$$

For an episode τ, acceleration and jerk are defined via the root mean square (RMS) across both time and dimensions:

$$A_{\text{ep}}(\tau) = \sqrt{\frac{1}{(T-6)D} \sum_{t=4}^{T-3} \sum_{k=1}^{D} a_{t,k}^2}, \tag{2}$$

$$J_{\text{ep}}(\tau) = \sqrt{\frac{1}{(T-6)D} \sum_{t=4}^{T-3} \sum_{k=1}^{D} j_{t,k}^2}. \tag{3}$$

Lower values of A_{ep} and J_{ep} indicate smoother, steadier demonstrations. We rank episodes by A_{ep} and by J_{ep} separately, keep the lowest 20% in each list, and take their intersection: $\mathscr{S}_{\text{pre}} = \Big\{\tau : \ \tau \in \text{Low-20\%}(A_{\text{ep}}) \ \wedge \ \tau \in \text{Low-20\%}(J_{\text{ep}})\Big\}$, which retains about 18% of episodes in our data. From $\mathscr{S}_{\text{pre}}$, we designate positives by open-loop replay success—task completion without collisions: $\mathscr{S}_{\text{high}} = \{\tau : \tau \in \mathscr{S}_{\text{pre}} \wedge \text{Success}(\tau) = 1 \ \wedge \ \text{CollisionFree}(\tau) = 1\}$, yielding roughly 15% high-quality demonstrations. Note that we score quality at the episode not chunk-level: stationary chunks can trivially exhibit low acceleration/jerk yet be uninformative. Episode-level aggregation, paired with a movement-coverage guard, suppresses such false positives and better captures overall stability and task competence.

Discriminator Model. After selecting the top-quality subset, we use an offline discriminator to score every real episode. For each episode, we uniformly sample K sub-clips $\{C_k\}_{k=1}^{K}$, and construct a tokenized input per clip: $\xi_t = \big(s_t,\ \mathbf{o}_t,\ \ell,\ \mathbf{a}_{t:t+L-1},\ \widehat{\log \pi_t}\big)$, where $\mathbf{o}_t$ are multi-view RGB observations, ℓ is the language instruction, $\mathbf{a}_{t:t+L-1}$ is an action chunk of length L, and $\widehat{\log \pi_t}$ is a log-π chunk score (policy-compatibility proxy) computed from the pretrained diffusion policy over that clip.

Given a pretrained diffusion-transformer policy π_θ, we define a surrogate for $\log \pi(\mathbf{a}_{t:t+L-1} \mid \ell, \mathbf{o}_t)$ via the negative denoising residual energy:

$$E_t = \frac{1}{|\mathscr{S}|L} \sum_{s \in \mathscr{S}} \sum_{\tau=t}^{t+L-1} \|\varepsilon_\theta(\mathbf{o}_\tau, \ell, \mathbf{a}_{\tau:\tau+L-1}, s_\tau) - \varepsilon\|_2^2, \tag{4}$$

$$\widehat{\log \pi_t} = -\,\text{zscore}(E_t) = -\frac{E_t - \text{Mean}(E)}{\sqrt{\text{Var}(E) + \varepsilon}}, \tag{5}$$

where $\mathscr{S}$ is a small set of diffusion steps. Intuitively, larger $\widehat{\log \pi_t}$ indicates that the policy explains the chunk better.

Each clip is projected into a token sequence: $[\ s_t;\ \mathbf{a}_{t:t+L-1};\ \widehat{\log \pi_t}\ ]$, equipped with learned positional embeddings. Language and image tokens are concatenated as a condition stream. A shallow stack of Transformer blocks produces hidden tokens, which are globally averaged and passed through a small MLP head with sigmoid to output a clip score $d(C_k) \in (0,1]$ (Fig. 5 (b)).

Diffusion Transformer. We employ a decoder-only Transformer as the diffusion model for the policy. Its architecture resembles the discriminator, but the input consists of the current observation $\mathbf{o}_t$, and the instruction ℓ, forming a vision–language conditioned policy:

$$\pi_\theta(s_t,\ \mathbf{o}_t,\ \ell) = \hat{\mathbf{a}}_{t:t+L-1}. \tag{6}$$

The current joint angle state information state s_t, and noisy actions $\tilde{\mathbf{a}}_{t:t+L-1}$ are projected into the latent space and

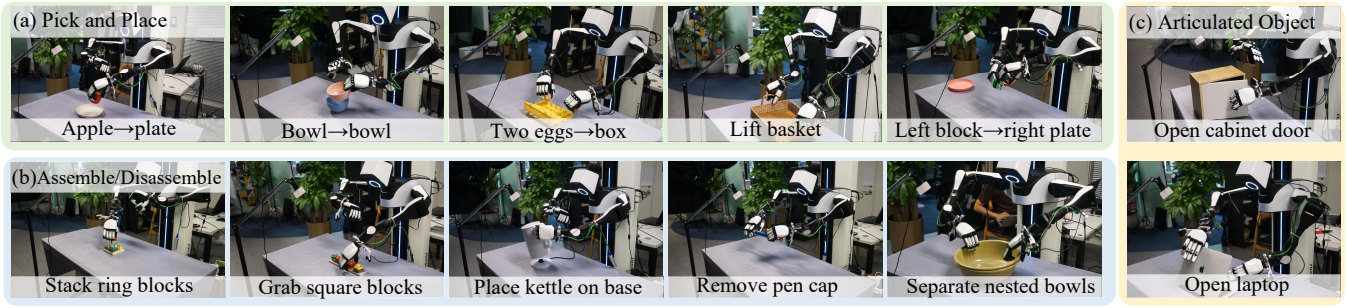

Fig. 6. **Basic tasks suite.** (a) Pick and Place (5 tasks). (b) Assemble/Disassemble (5 tasks). (c) Articulated Objects (2 tasks).

concatenated with the diffusion timestep t to form the input tokens for the transformer. Natural language and multi-view image inputs are encoded into conditional tokens via the T5 [39] and SigLip [40] encoders, respectively, and alternately injected into the transformer blocks. The model predicts the action noise $\widehat{\theta}$, thereby yielding the predicted action sequence $\widehat{\mathbf{a}}_{t:t+L-1}$ (Fig. 5 (c)). We use the standard DDPM for sampling during training and employ DPMSolver++ for acceleration during action generation.

D. Data-quality-aware Training Recipe

We first **pretrain** the diffusion-transformer policy π_θ on simulation data to endow the VLA with basic competence (pick & place, assemble, etc.). This policy is then used to compute the *log-π proxy* for training the discriminator model.

Let the positive set be the replay-validated high-quality subset $\mathscr{S}_{\text{high}}$ (about 15%) and the unlabeled pool be $\mathscr{U} = \mathscr{D}_{\text{real}} \setminus \mathscr{S}_{\text{high}}$. We optimize a positive–unlabeled objective:

$$\mathscr{L}_D = \eta \underbrace{\mathbb{E}_{\tau\in\mathscr{S}_{\text{high}}}\big[-\log d(\tau)\big]}_{\text{positive BCE}\to 1} + \underbrace{\mathbb{E}_{\tau\in\mathscr{U}}\big[-\log(1-d(\tau))\big]}_{\text{unlabeled as negative}\to 0}, \tag{7}$$

where $\eta = 0.5$. We apply clip scores to $d \in [0.1, 0.9]$ for stability (Fig. 5 (b)). Following the DWBC mapping from [41], we convert calibrated scores to weights w_i.

Finally, we **post-train** π_θ on the real dataset to upgrade this base competence into dexterous skills, using the precomputed weights. For diffusion training,

$$\mathscr{L}_\pi = \sum_{i=1}^{L} w_i \left\|\varepsilon_\theta(\cdot) - \varepsilon\right\|_2^2, \tag{8}$$

with a short weight warm-up (Fig. 5 (c)).

IV. EXPERIMENT

We evaluate *Dexora* across three axes: (1) **Performance**: higher success on basic and dexterous tasks, especially on bimanual skills (Sec. IV-B). (2) **Generalization**: Robust to OOD shifts and transfers across embodiments (Sec. IV-C). (3) **Ablations**: contributions of training data composition and the learned data-quality discriminator (Sec. IV-D).

A. Experimental Setup and Baselines

Setup. Our policy model has 28 layers, a hidden size of 1024, and 16 attention heads, totaling 400M parameters. The discriminator is smaller, with 12 layers, a hidden size of 512, and 8 attention heads, for 30M parameters. We pretrain the policy model for 100K gradient steps and train the discriminator model for 10K steps, using distributed data parallelism across 8 × NVIDIA L20 GPUs with a total batch size of 64. Both models are optimized using AdamW.

Baselines. We compare against three representative baselines: **Diffusion Policy (DP)** [42]—a conditional denoising policy for visuomotor imitation; π_0 [3]—a VLA with a flow-matching action generator; and **GR00T N1** [7]—an open VLA (VLM + DiT) designed for humanoid control.

Action-space Adaptation. DP natively regresses continuous actions, so we train it directly on our 36-D vector commands. For π_0 and GR00T N1, we append a 2-layer MLP projector that maps each model's native action output to our 36-D joint command. The projector is factorized by physical groups (L/R arm, L/R hand), and learns the expansion from lower-DoF end-effector outputs (e.g., GR00T N1's 6-DoF hand) to our 12-DoF hands via learned synergies.

Protocol. All other settings are identical across methods: control frequency, action chunk length $L = 32$, camera intrinsics/extrinsics, and the number of views. For each task, we collect 100 demonstrations to train/fine-tune the baselines for 50K steps. Fine-tuning runs on 4 × NVIDIA L20 GPUs with LoRA; inference is performed on a single RTX 4090. We report the success rate over 20 rollouts per task.

B. Evaluation Results in Real World

Basic Tasks Evaluation. We group basic tasks into three types—**Pick-and-Place** (5 tasks), **Assemble/Disassemble** (5 tasks), and **Articulated Objects** (2 tasks). Each type mixes single-hand and bimanual problems. Representative bimanual examples include placing a distant block into a tray via a two-hand handover with temporal ordering, and separating two stacked bowls that require simultaneous two-hand prying (Fig. 6). *Dexora* is evaluated zero-shot. Results in Tab. I show that *Dexora* attains the highest overall success, reaching $\geq$90% on 7/12 tasks and consistently leading the bimanual tasks. GR00T N1 [7] is competitive on simpler, mostly single-hand tasks. π_0 [3] degrades most after mapping a gripper-centric action space to high-DoF hands, confirming that the low→ high DoF mapping is ill-posed without embodiment-matched data. Benefiting from many dual-arm episodes in training, *Dexora* shows clear gains on bimanual coordination while maintaining strong performance. Over-

Manuscript 209 submitted to 2026 IEEE International Conference on Robotics and Automation (ICRA). Received September 16, 2025.

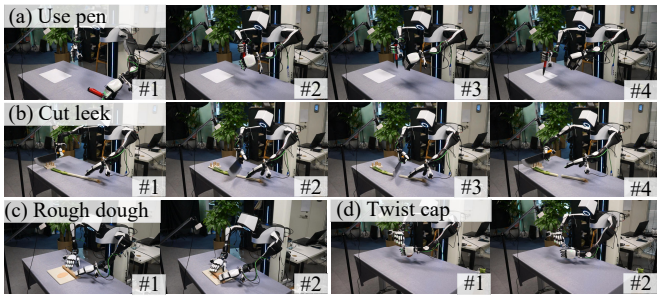

Fig. 7. **Dexterous manipulation sequences**. (a) **Use Pen**: The left hand picks up the pen (#1), hands it to the right hand (#2); the right thumb depresses the tip (#3) and writes on paper (#4). (b) **Cut Leek**: The right hand grasps the knife (#1), the left hand stabilizes the leek (#2); the right hand slices (#3) and returns the knife to the table (#4). (c) **Rough Dough**: Both hands press the rolling pin simultaneously (#1) and push forward to flatten the dough (#2). (d) **Twist Cap**: The left hand holds the bottle while the right thumb–index grip twists the cap (#1) and removes it (#2).

all, these trends support our design choice: embodiment-matched, high-DoF data are essential for performance.

Dexterous Manipulation Tasks Evaluation. Pure pick-and-place does not exploit high-DoF hands; grippers can also do that. The value of hands emerges on dexterous skills that require in-hand tool use and coordinated bimanual manipulation (Fig. 1(a)). We therefore benchmark 6 tasks (Fig. 7). All models are trained/fine-tuned on 100 demonstrations. Tab. II shows that *Dexora* gains the best average performance (66.7% vs. 51.7% for GR00T N1, 26.7% for π_0, and 6.7% for DP). GR00T N1 is the strongest baseline but uses a 6-DoF hand; it struggles on in-hand skills such as *Use pen* and fails on *Twist cap*, which require thumb-index synergies and lateral finger swing to generate a stable torsional wrench. *Dexora*'s gains arise from its 12-DoF hands and bimanual training corpus, enabling reliable in-hand and dual-arm coordination. We find that cap twisting exhibits the lowest success rate. The task requires generating a stable torsional wrench to overcome cap breakaway torque while preventing slip, which couples precise normal-force regulation, fingertip friction, and fine in-hand alignment. In our current setup, the absence of tactile feedback and relatively low-friction rigid fingertip pads leads to slip.

C. Generalization

Out-of-Distribution Generalization. We test OOD robustness on the "Pick apple to the plate" task across six conditions: unseen background, unseen lighting, unseen object, occlusion, clutter, and height change, and we report the success rate (Fig. 8). *Dexora* maintains high performance across all variants, showing excellent OOD generalization.

Cross-Embodiment Generalization. Our premise is that a dual-arm, dual-hand high-DoF policy contains lower-DoF embodiments as subspaces: projecting a 36-D joint action down to simpler robots is dimension reduction, not synthesis—far easier than "lifting" a gripper policy to dexterous hands. We therefore test three representative embodiment configurations: **EC-1: single-arm gripper** - Franka Emika Panda (6-DoF + 1-DoF gripper); **EC-2: dual-arm grippers** - Cobot Magic ALOHA (2 × (6-DoF arm + 1-DoF gripper));

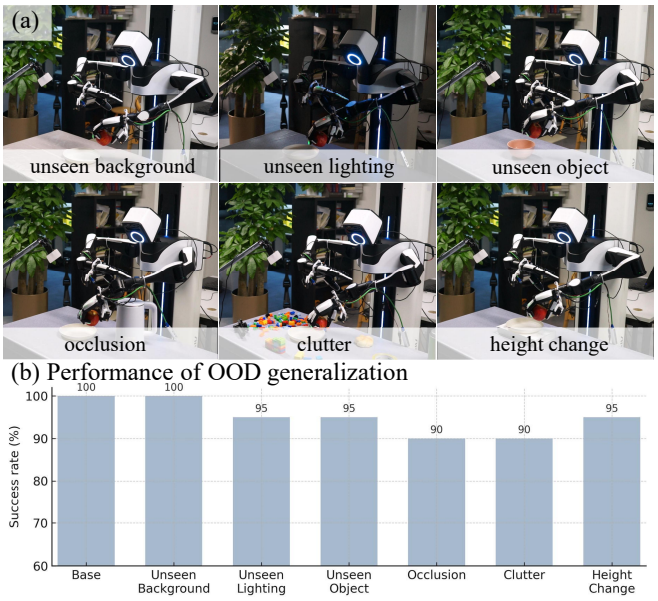

Fig. 8. Generalization of six Out-of-Distribution (OOD) conditions. We report success rate (%) over 20 rollouts.

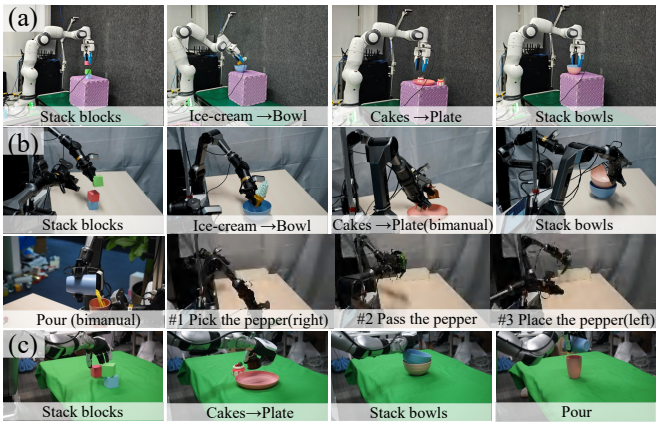

Fig. 9. **Cross-embodiment generalization**. The *Dexora* policy transfers to (a) single-arm gripper, (b) dual-arm grippers, and (c) single-arm single-hand, completing representative tasks like a three-step pepper hanover.

EC-3: single-arm single-hand - Unitree G1 7-DoF arm + Inspire Hand 6-DoF. For adaptation, we pad unused action dimensions to keep tensor shapes fixed; for observations, we mask the absent camera. Each task is fine-tuned with 100 demonstrations, and all other settings are identical. On the evaluated tasks including single- and dual-arm setups (Fig. 9), grasping tasks transfer readily across embodiments, whereas dexterity-demanding tasks show the largest gaps (Tab. II). This supports our hypothesis that high→low mapping is better posed than the inverse; compressing a 12-DoF hand policy to a 1-DoF gripper is simpler than lifting a gripper policy to dexterous hands.

D. Ablation Study

Effectiveness of Training Data Composition. We compare three post-training regimes: Sim Only, Sim + 50% Real (100 tasks), and Sim + All Real (200 tasks). Four

TABLE I

BASIC TASKS EVALUATION. RESULTS ARE SUCCESS RATES (%) OVER 20 TRIALS. GRAY COLUMNS INDICATE BIMANUAL TASKS.

Method	Pick and Place					Assemble / Disassemble					Articulated Object		Avg.
	Apple → plate	Bowl → bowl	Two eggs → box	Lift basket	Left block → right plate	Stack ring blocks	Grab square blocks	Place kettle on base	Remove pen cap	Separate nested bowls	Open cabinet door	Open laptop	
DP	60	65	30	10	25	35	15	45	30	10	65	20	34.2
π_0	75	70	45	30	30	60	60	65	55	20	60	35	50.4
GR00T N1	95	**100**	75	60	80	**90**	**80**	90	80	60	95	80	82.1
Dexora	**100**	**100**	**85**	**80**	**90**	85	**80**	**95**	**90**	**80**	**100**	**90**	**89.6**

TABLE II

DEXTEROUS MANIPULATION TASKS EVALUATION.

Method	Use pen	Fetch book	Cut leek	Place plates	Rough dough	Twist cap
DP	5	10	10	0	15	0
π_0	20	45	60	20	15	0
GR00T N1	45	60	**85**	60	60	0
Dexora	**65**	**80**	80	**70**	**80**	**25**

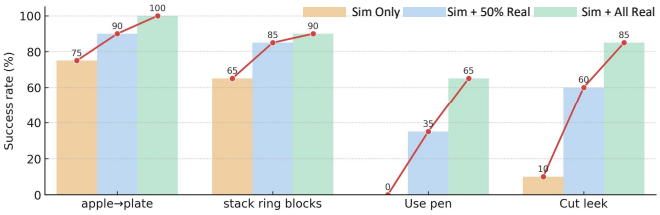

Fig. 10. **Effect of training data composition.** Success rate for four tasks under three training regimes: Sim Only, Sim + 50% Real, Sim + All Real.

tasks are evaluated, two basic (Apple→plate, Stack ring blocks) and two dexterous (Use pen, Cut leek). Success rises steadily with more real data; dexterous tasks improve from 0→35→65 and 10→60→85 (Fig. 10). These results show that simulation is effective for bootstrapping basic skills, while real, more complex data plays a crucial role in developing dexterous capabilities.

Effectiveness of Discriminator model. We compare vanilla post-training of the Diffusion Transformer with quality-aware post-training that uses a learned discriminator to score and weight demonstrations. Tab. III quantifies the gains: the discriminator improves success rate and reduces acceleration and jerk at inference. In both a single-hand and a bimanual task, the quality-aware model executes smoother, more coherent motions. The time-series traces show lower variance and fewer reversals (Fig. 11). Overall, the discriminator helps the policy learn from mixed-quality demonstrations by emphasizing high-quality segments and down-weighting suboptimal ones, enabling better strategies from imperfect data.

V. CONCLUSIONS

We present *Dexora*, the first open-source VLA system that natively controls dual-arm, dual-hand, 36-DoF robots. A hybrid teleoperation pipeline drives both hardware and a MuJoCo twin to build an embodiment-matched corpus, and a data-quality discriminator guides post-training so the policy learns most from high-quality demonstrations. *Dexora* outperforms strong baselines on basic and dexterous tasks,

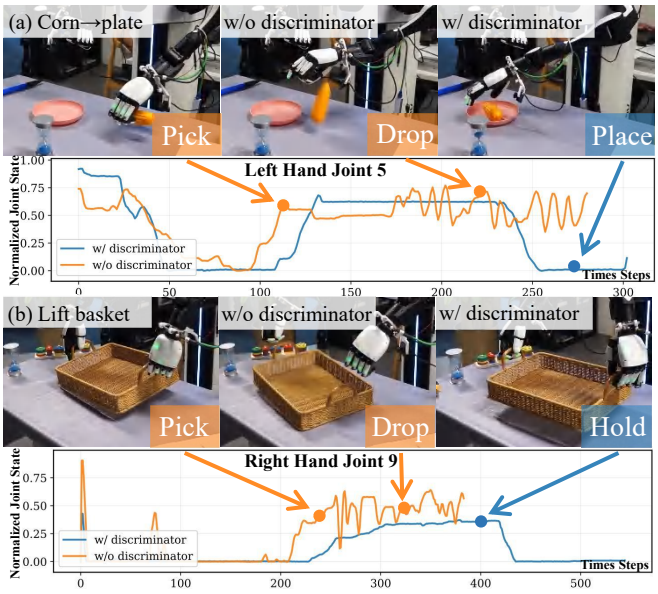

Fig. 11. **Effect of the data-quality discriminator**. (a) **Corn → plate**: with the discriminator, joint trajectories are smooth and the placement succeeds; without it, high-frequency oscillations in left-hand joint 5 cause the corn to drop. (b) **Lift basket (bimanual)**: with the discriminator, the basket is lifted; without it, jitter in right-hand joint 9 tilts the basket and it slips.

TABLE III

EFFECT OF THE DISCRIMINATOR MODEL. WE REPORT **S.R.** (SUCCESS RATE %) AND SMOOTHNESS METRICS—MEAN NORMALIZED JOINT **ACCELERATION** AND **JERK**, AVERAGED OVER 20 EPISODES.

Method	Corn → plate			Lift basket		
	S.R.	Acc. ↓	Jerk ↓	S.R.	Acc. ↓	Jerk ↓
w/o discriminator	85	0.034	0.043	55	0.041	0.052
w/ discriminator	95	0.020	0.032	80	0.023	0.036

is robust to OOD shifts, and transfers across embodiments with lightweight action projectors—evidence that training in a rich, high-DoF action space provides a well-posed path to lower-DoF controllers. Ablations show that simulation bootstraps basic competence, while real data and the discriminator are key for dexterity and smooth control.

Looking forward, we see two promising directions: (i) contact-aware control via tactile sensing to close the loop on tasks like cap twisting; (ii) long-horizon reasoning and hierarchical VLA planning that combines memory, subgoal decomposition, and language-guided tool use. We hope the released models, data, and code catalyze research toward broadly capable, dexterous robot assistants.

Manuscript 209 submitted to 2026 IEEE International Conference on Robotics and Automation (ICRA). Received September 16, 2025.

REFERENCES

[1] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *CoRL*, PMLR, 2023.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[4] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[5] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," *arXiv preprint arXiv:2410.07864*, 2024.

[6] C. Cheang, S. Chen, Z. Cui, Y. Hu, L. Huang, T. Kong, H. Li, Y. Li, Y. Liu, X. Ma, *et al.*, "Gr-3 technical report," *arXiv preprint arXiv:2507.15493*, 2025.

[7] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[8] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *RSS*, 2023.

[9] A. . Team, "Aloha 2: An enhanced low-cost hardware for bimanual teleoperation," *arXiv preprint arXiv:2405.02292*, 2024.

[10] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "Open teach: A versatile teleoperation system for robotic manipulation," *CoRL*, 2024.

[11] R. Ding, Y. Qin, J. Zhu, C. Jia, S. Yang, R. Yang, X. Qi, and X. Wang, "Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning," *arXiv preprint arXiv:2407.03162*, 2024.

[12] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," *RSS*, 2023.

[13] S. Li, X. Ma, H. Liang, M. Görner, P. Ruppel, B. Fang, F. Sun, and J. Zhang, "Vision-based teleoperation of shadow dexterous hand using end-to-end deep neural network," *ICRA*, 2019.

[14] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild," *ICRA*, 2024.

[15] M. Xu, H. Zhang, Y. Hou, Z. Xu, L. Fan, M. Veloso, and S. Song, "Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation," *CoRL*, 2025.

[16] A. Imdieke and K. Desingh, "Spark-remote: A cost-effective system for remote bimanual robot teleoperation," *arXiv preprint arXiv:2504.05488*, 2025.

[17] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," *IROS*, 2024.

[18] H. Li, Y. Cui, and D. Sadigh, "How to train your robots? the impact of demonstration modality on imitation learning," *arXiv preprint arXiv:2503.07017*, 2025.

[19] C. Pan, K. Junge, and J. Hughes, "Vision-language-action model and diffusion policy switching enables dexterous control of an anthropomorphic hand," *arXiv preprint arXiv:2410.14022*, 2024.

[20] J. Zhang, H. Liu, D. Li, X. Yu, H. Geng, Y. Ding, J. Chen, and H. Wang, "Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes," in *8th CoRL*, 2024.

[21] J. Ye, K. Wang, C. Yuan, R. Yang, Y. Li, J. Zhu, Y. Qin, X. Zou, and X. Wang, "Dex1b: Learning with 1b demonstrations for dexterous manipulation," *arXiv preprint arXiv:2506.17198*, 2025.

[22] Y. Ye, A. Gupta, K. Kitani, and S. Tulsiani, "G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis," in *CVPR*, pp. 1911–1920, 2024.

[23] Y. Zhong, Q. Jiang, J. Yu, and Y. Ma, "Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness," in *CVPR*, pp. 22584–22594, 2025.

[24] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen, *et al.*, "Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," in *CVPR*, pp. 4737–4746, 2023.

[25] K. Li, J. Wang, L. Yang, C. Lu, and B. Dai, "Semgrasp: Semantic grasp generation via language aligned discretization," in *ECCV*, 2024.

[26] Y. Liu, Y. Yang, Y. Wang, X. Wu, J. Wang, Y. Yao, S. Schwertfeger, S. Yang, W. Wang, J. Yu, *et al.*, "Realdex: Towards human-like grasping for robotic dexterous hand," *arXiv:2402.13853*, 2024.

[27] J. Chen, Y. Ke, L. Peng, and H. Wang, "Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy," *arXiv preprint arXiv:2504.18829*, 2025.

[28] H. Zhang, Z. Wu, L. Huang, S. Christen, and J. Song, "Robustdexgrasp: Robust dexterous grasping of general objects," *arXiv preprint arXiv:2504.05287*, 2025.

[29] K. Li, P. Li, T. Liu, Y. Li, and S. Huang, "Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning," in *CVPR*, pp. 6991–7003, 2025.

[30] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu, "Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning," *arXiv preprint arXiv:2410.24185*, 2024.

[31] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, H. Yin, S. Liu, *et al.*, "Egovla: Learning vision-language-action models from egocentric human videos," *arXiv arXiv:2507.12440*, 2025.

[32] S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, H. Cui, *et al.*, "Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data," *arXiv preprint arXiv:2505.03233*, 2025.

[33] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang, *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv preprint arXiv:2503.06669*, 2025.

[34] Y. Zhong, X. Huang, R. Li, C. Zhang, Y. Liang, Y. Yang, and Y. Chen, "Dexgraspvla: A vision-language-action framework towards general dexterous grasping," *arXiv preprint arXiv:2502.20900*, 2025.

[35] H. Luo, Y. Feng, W. Zhang, S. Zheng, Y. Wang, H. Yuan, J. Liu, C. Xu, Q. Jin, and Z. Lu, "Being-h0: Vision-language-action pretraining from large-scale human videos," *arXiv preprint arXiv:2507.15597*, 2025.

[36] J. Jang, S. Ye, Z. Lin, J. Xiang, J. Bjorck, Y. Fang, F. Hu, S. Huang, K. Kundalia, Y.-C. Lin, *et al.*, "Dreamgen: Unlocking generalization in robot learning through neural trajectories," pp. arXiv–2505, 2025.

[37] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[38] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, *et al.*, "Objaverse-xl: A universe of 10m+ 3d objects," *NeurIPS*, vol. 36, 2023.

[39] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, vol. 21, pp. 1–67, 2020.

[40] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

[41] H. Xu, X. Zhan, H. Yin, and H. Qin, "Discriminator-weighted offline imitation learning from suboptimal demonstrations," in *International Conference on Machine Learning*, pp. 24725–24742, PMLR, 2022.

[42] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *IJRR*, 2023.