Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

X-DiffVLA: X-Embodied Diffusion Action Heads for Vision-Language-Action Models

Boyu Li[1,2,3], **Chaoyi Xu**[4], **Haoqi Yuan**[4,5], **Xinrun Xu**[2],
Börje F. Karlsson[3], **Dongbin Zhao**[1,2,3], **Haoran Li**[1,2,3*], **Zongqing Lu**[4,5*]
[1]SKL-MAIS, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Beijing Academy of Artificial Intelligence, [4]BeingBeyond
[5]School of Computer Science, Peking University
* Correspondence: lihaoran2015@ia.ac.cn, zongqing.lu@pku.edu.cn

Abstract*—Learning universal policies from cross-embodied data remains a fundamental challenge in robotics. Although Vision-Language-Action (VLA) models are pre-trained on large and diverse datasets, they typically rely on embodiment-specific fine-tuning to achieve strong performance in downstream tasks. This requirement severely limits their generalization capability and restricts knowledge transfer across embodiments performing similar tasks. To overcome these limitations, we focus on cross-embodied settings with shared robotic bases and heterogeneous end-effectors, and propose X-DiffVLA, a diffusion-based VLA model featuring a unified cross-embodied action head. X-DiffVLA can leverage the generative strengths of diffusion models to capture both the diversity and latent correlations in cross-embodied datasets. Specifically, we introduce Embodiment Forcing, a classifier-free guidance technique to implicitly steer action generation toward embodiment-specific functional components, capturing fine-grained structural nuances without explicit supervision. In addition, a Morphological Tree Diffusion approach is designed to strengthen behavioral correlations across diverse end-effectors, maximizing the transferability of heterogeneous demonstrations. Experimental results across RoboCasa and Isaac Gym, covering different embodiments from grippers to dexterous hands, show that X-DiffVLA achieves state-of-the-art performance, with improvements of 15.3% and 12.5%, respectively. Real-world evaluations further validate the robustness of the proposed framework and its effectiveness in scalable cross-embodied policy learning. Videos are available on our project page https://boyuli-xh.github.io/X-DiffVLA/.

I. Introduction

In recent years, training Vision-Language-Action (VLA) models on large-scale, heterogeneous, multi-task, and multi-embodiment datasets has emerged as a prominent research frontier [13, 43]. Models such as OpenVLA [21], GR00T [2], and $\pi_{0.5}$ [19] have demonstrated that large-scale pre-training equips models with superior visual understanding [6] and long-horizon task planning capabilities [23], facilitating diverse robotic control tasks. However, existing architectures typically necessitate fine-tuning embodiment-specific action heads [2] during downstream evaluation to ensure performance. Consequently, any variation in robot morphology requires the finetuning of a new action head. This paradigm not only impairs training efficiency, particularly in scenarios involving multiple end-effectors, but also obstructs cross-embodiment data utilization for similar tasks, leading to a landscape where specialized models excel in isolated tasks. Therefore, designing a cross-embodied action head for VLA is a unified path toward general intelligence in the field of robotics.

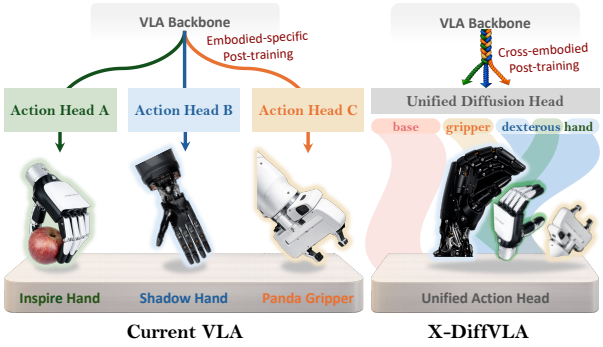

Fig. 1: **The motivation of X-DiffVLA.** While embodied-specific post-training restricts VLA models to isolated tasks and end-effectors, our architecture enables cross-embodied learning and knowledge transfer, leveraging diverse data to transition from specialized experts to a unified, general robotic controller.

A fundamental challenge in developing a cross-embodied action head lies in effectively characterizing the difference across heterogeneous embodiments to facilitate morphology-aware action generation. Unlike visual [28] or auditory modalities [34], robotic data exhibits substantial structural divergence [22]; for instance, the joint optimization across datasets including simple grippers and dexterous hands often precipitates distributional interference, leading to sub-optimal performance. Consequently, a control policy must possess morphological discrimination capabilities. Liu et al. [27] demonstrates that utilizing a unified action space and

diffusion transformer architectures is effective to leverage cross-gripper data on small-scale models. Extending to current VLA architecture with more end-effectors, achieving conditioned action generation for heterogeneous embodiments remains unexplored.

Effective knowledge transfer via cross-embodied action head is another challenge for VLA models. Cross-embodied data is valuable because there exist many tasks share underlying similarities across different embodiments [36]. A universal action head should learn shared representations from cross-embodiment data to enhance generalization. A representative scenario, frequently encountered in both simulation and real-world environments, involves a shared robotic platform equipped with diverse end-effectors [32, 37]. Since the primary agents remain consistent, tasks across these embodiments exhibit significant functional similarities, suggesting an implicit correlation between their control policies. Prior research has demonstrated that this correlation exists and can be transferred; for instance, DemoGrasp [39] utilize motion retargeting to map control strategies from human hand poses to various dexterous hands. Similarly, Bauer et al. [1] find that learning a semantically aligned latent action space for dexterous hands, human hands, and parallel jaw grippers, can enable the optimization of a joint policy by refining only one single embodiment. Therefore, how cross-embodied action heads can effectively mine these latent correlations is another challenge we want to solve for achieving robust cross-embodied control within Vision-Language-Action (VLA) models.

To address these challenges, we present **X-DiffVLA**, a general VLA framework that enables cross-embodied post-training via a diffusion-based action head. By leveraging the capability of diffusion models to capture multi-peaked distributions in complex action spaces, X-DiffVLA can effectively accommodate diverse embodied data with a unified action space. We further propose **Embodied Forcing (EBF)**, a mechanism that integrates morphological information into the denoising process to guide action token generation. Additionally, a **Morphological Tree Diffusion (MPTD)** method is introduced to capture behavioral correlations across disparate end-effectors, facilitating cross-embodied knowledge transfer. Evaluated in both the task-complex RoboCasa and the high-dynamics Isaac Gym environments, spanning simple grippers to different dexterous hands, X-DiffVLA achieves state-of-the-art performance with improvements of 15.3% and 12.5%, respectively. Real-world experiments further underscore its potential as a scalable pathway toward universal robotic control.

In summary, our contributions encompass the following key advancements:

1) We propose **X-DiffVLA**, a general VLA framework featuring a diffusion-based action head designed for post-training across heterogeneous robotic embodiments.
2) We introduce **Embodied Forcing** and **Morphological Tree Diffusion**, two key techniques that enhance both discrimination and behavioral association across diverse embodiments during the action generation process.
3) We conduct experiments in RoboCasa and Isaac Gym, demonstrating that X-DiffVLA significantly improves cross-embodied data utilization, achieving performance gains of 15.3% and 12.5% over state-of-the-art VLA baselines.

II. Preliminaries

A. Action Heads in VLA Models

For current VLA models, continuous action space can be categorized into three types: the multilayer perceptron (MLP) head, the diffusion head, and the flow-matching head. The MLP head [25] is straightforward but tends to learn the mean of various action distributions, that may be not allowed by physical executions. The diffusion head possesses superior capabilities for fitting multi-peaked distributions and has proven effective for robotic action generation in multi-task scenarios [13, 26]. The flow-matching head, adopted by current popular embodied models [19], improves upon the diffusion head by directly predicting the denoising vector field, thereby achieving fewer sampling iterations and higher control frequencies.

The action head employed in X-DiffVLA is also improved from the diffusion head, leveraging its exceptional multi-modal fitting capabilities and conditional generation performance. We compare it with the flow-matching head in our experiments, demonstrating that the diffusion head exhibits a significant advantage under cross-embodied post-training conditions, as further substantiated in Section IV-C3.

B. Diffusion Generation for Multi-Peaked Distributions

Diffusion models have proven to be exceptional generative models for fitting multi-peaked data distributions [15], while demonstrating robust conditional generation capabilities. The core optimization revolves around a denoising process. It first considers a forward diffusion process that systematically injects Gaussian noise into a data point over a sequence of timesteps. Let $q(\mathbf{x})$ represent the data distribution, where we define $\mathbf{x} \sim q$. This process is formalized as a Markov chain, where the data $\mathbf{x}$ at each step k is incrementally added noise by:

$$q(\mathbf{x}^k|\mathbf{x}^{k-1}) = \mathcal{N}(\mathbf{x}^k; \sqrt{1-\beta_k}\mathbf{x}^{k-1}, \beta_k\mathbf{I}), \quad (1)$$

where $\mathcal{N}$ denotes the normal distribution and $\beta_k \in (0, 1)$ represents the noise variance at each step. This process

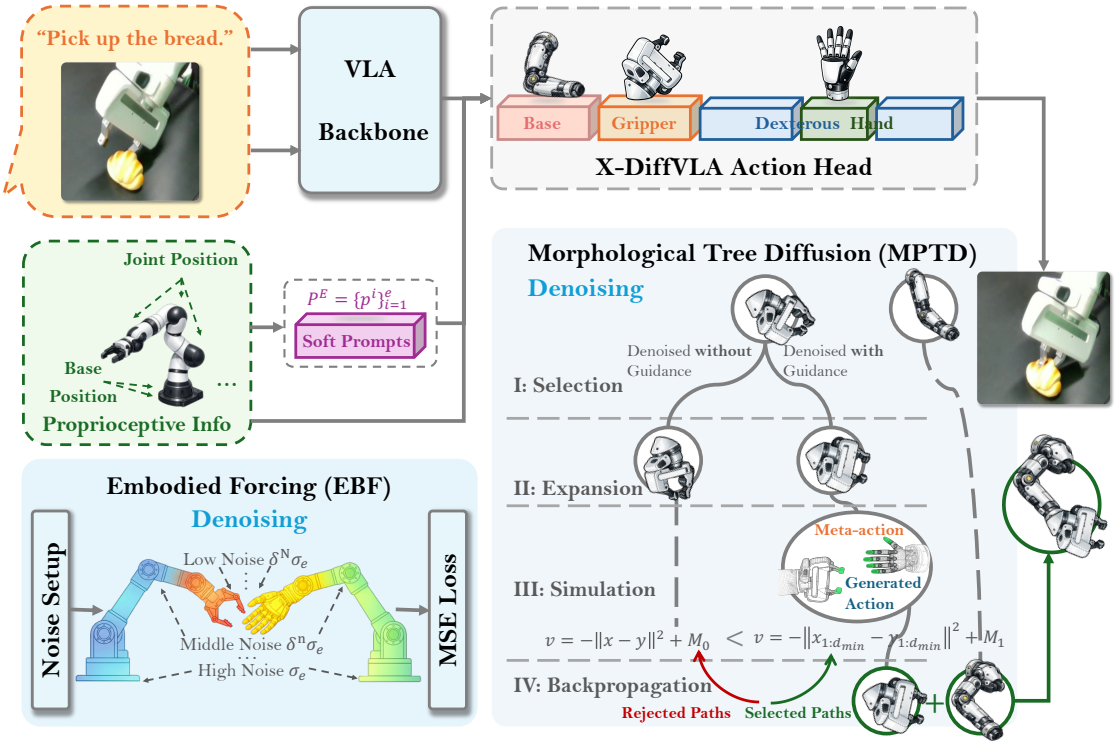

Fig. 2: **Architectural Overview of X-DiffVLA.** Our framework leverages a unified action space and a diffusion head to model multi-peaked action distributions for cross-embodied post-training. Key components include: (1) Embodied Forcing (EBF), which integrates morphological priors into the denoising process to enhance embodiment-specific discernment; and (2) Morphological Tree Diffusion (MPTD), designed to capture behavioral correlations across diverse end-effectors, thereby facilitating robust cross-embodied knowledge transfer.

continues until the original data is transformed into white noise at $\mathbf{x}^K$.

The reverse process is also a Markov chain, which aims to reconstruct the original data from noise using a parameterized model p_θ:

$$p_\theta(\mathbf{x}^{k-1}|\mathbf{x}^k) = \mathcal{N}(\mathbf{x}^{k-1}; \boldsymbol{\mu}(\mathbf{x}^k, k), \gamma_k \mathbf{I}), \tag{2}$$

where $\boldsymbol{\mu}$ is a neural network, and the covariance of it can be set to the identity scaled by a fixed constant γ_k. Furthermore, in the reverse transition $p_\theta(\mathbf{x}^{k-1}|\mathbf{x}^k)$, the model effectively directs the latent variables from a high-entropy Gaussian state toward the various local maximum modes of the underlying data manifold. Unlike GANs [14], which often suffer from mode collapse by focusing on a single point estimate, the stochastic feature of the reverse paths ensures that the diffusion model can cover multiple high-probability regions of $q(\mathbf{x})$. This allows for the synthesis of diverse samples that faithfully represent the complex, multi-peaked distribution of the cross-embodied data.

C. Monte Carlo Tree Diffusion

Monte Carlo Tree Search (MCTS) [9] is a planning algorithm widely used in Reinforcement Learning that combines tree search with stochastic simulations to effectively balance exploration and exploitation. Typically, MCTS consists of four stages: selection, expansion, simulation, and backpropagation. By iteratively exploring potential future states, MCTS allows an agent to estimate the value of long-term trajectories in complex decision-making environments.

Since the diffusion denoising process is inherently stochastic, identical initial conditions sometimes yield significantly different outcomes. To enhance generation stability, Monte Carlo Tree Diffusion (MCTD) [38] adapts the MCTS algorithm by employing the diffusion model as the core engine for state transitions and node expansion. During the search, the algorithm utilizes node-specific reward feedback to dynamically modulate the denoising guidance intensity. This mechanism deeply couples the generative capacity of diffusion models with the heuristic search of MCTS, ensuring thorough exploration of the search space while accelerating denoising convergence toward high-reward states. More related works can be seen in Appendix B.

III. Method

The architectural framework of **X-DiffVLA** is illustrated in Fig.2. To facilitate effective cross-embodied

post-training, we employ a unified action space and a diffusion head capable of modeling multi-peaked action distributions. We introduce Embodied Forcing to enhance the model's ability to distinguish between diverse robotic embodiments and functional segments, ensuring precise conditional generation. This mechanism integrates specific morphological information directly into the denoising process to guide action token generation. Furthermore, to capture behavioral correlations across varied end-effectors, we propose Morphological Tree Diffusion, which promotes robust cross-embodied knowledge transfer in similar tasks.

A. Embodied Forcing

First, X-DiffVLA establishes a unified action space by defining a standardized representation that encompasses the maximum dimensions of robot bases, grippers, and dexterous hands. Within each category, various robot platforms share a common action subspace; for robots with lower degrees of freedom (DoF), the action vectors are aligned via zero-padding in the trailing dimensions. This partitioning of the unified action space is strategically designed to distinguish between locomotion and interaction tasks, thereby facilitating conditional generation by the diffusion head.

To address the challenge of convergence in multi-peaked optimization and achievement of morphology-guided action generation for cross-embodied data, X-DiffVLA introduces **Embodied Forcing (EBF)** technique. In the field of video generation, Diffusion Forcing [5] is often employed to enhance the causality of diffusion models. While traditional diffusion models are trained using uniform noise levels across all tokens, Diffusion Forcing proposes training sequence diffusion models with independently varied noise levels per frame. This approach prioritizes the generation of significant inter-frame variations while minimizing fluctuations in static backgrounds. As a form of classifier-free guidance, it ultimately improves the ability of diffusion models to generate fine-grained local video details. Inspired by this, Embodied Forcing refines the initial noise distribution of the denoising process from the dual perspectives of global robotic structure and localized functional components.

Global Structure-Aware Noise Initialization. Different embodied datas represent different distributions, therefore the diffusion head first need to realize the generation of robot structure-guided action. As we introduce in Section II-B, diffusion head generate actions through denosing process. Specifically, at each timestep t, the diffusion head predicts a H steps action chunk τ_t conditioned on an embodiment feature e, which specifies the robot's morphology and dynamics:

$$\tau_t = [a_t, a_{t+1}, \dots, a_{t+H}]. \tag{3}$$

We train the diffusion head, $\hat{\tau}_t = x_{0,\theta}(\tau_t^k, p_t, n_t, k, e)$, to reconstruct a clean trajectory, where τ_t^k is the original trajectory τ added with Gaussian noise at level k in embodiment type e. p_t represents the proprioceptive information, and n_t represents the output tokens by VLM backbone. The forward process for a specific embodiment e is defined as:

$$\tau_t^k = \sqrt{1-\beta_k}\tau_t^{k-1} + \sqrt{\beta_k} \cdot \sigma_e \epsilon, \epsilon \sim \mathcal{N}(0, I), \tag{4}$$

where α_k is the hyper-parameter in traditional diffusion models and σ_e represents the unique noise magnitude scale for embodiment e. This allows the diffusion head to characterize the difference across heterogeneous embodiments from different Gaussian noise regions.

Local Function-Aware Noise Initialization. Beyond global structural differences across embodiments, the functional heterogeneity of robot components also leads to multi-peaked distributions in cross-embodied data. For example, end-effector action distribution demands higher precision than that of robot base. Therefore, we add function-guided noise injection to assist the diffusion head in generating more accurate, function-aligned actions.

We consider that the spatial causality of robot actions mirrors the temporal causality observed in videos. In video diffusion models, pyramid noise [5] is typically applied to historical frames according to their temporal distance to diminish the influence of distant frames. Leveraging this intuition, we propose a spatial pyramid noise injection strategy, where noise scales are assigned based on a component's proximity to the end-effector. The forward process is redefined as:

$$\tau_t^k = \sqrt{1-\beta_k}\tau_t^{k-1} + \sqrt{\beta_k} \cdot \mathbf{\Sigma}_{e,f} \epsilon, \epsilon \sim \mathcal{N}(0, I), \tag{5}$$

$$\mathbf{\Sigma}_{e,f} = \text{diag}(\sigma_e, \delta\sigma_e, \delta^2\sigma_e, \dots), \tag{6}$$

where $\mathbf{\Sigma}_{e,f}$ represents a diagonal covariance matrix and σ serves as the noise attenuation coefficient. In experiments, we find that a window-based pyramid outperforms a step-wise approach in conditional generation. This suggests that the underlying distribution of robotic action data is organized around functional blocks (e.g., arm, gripper, fingers) rather than simple spatial positions. Detailed experimental results are provided in Section IV-C2.

Finally, the network is optimized using the Mean Squared Error (MSE) loss against the ground-truth clean trajectory:

$$\mathcal{L} = \text{MSE}(x_{0,\theta}(\tau_t^k, p_t, n_t, k, e, f), \tau_t). \tag{7}$$

B. Morphological Tree Diffusion

After solving the multi-peaked optimization problem of cross-embodied action heads, we further consider how

to capture the correlation between different embodiments on similar tasks. This serves as the primary motivation for cross-embodied post-training to effectively mine transferable knowledge from heterogeneous robot data, thereby providing a robust solution to the distributional challenges in contemporary robotics datasets.

We propose **Morphological Tree Diffusion (MPTD)**, which extends the principles of MCTD to the processing of cross-embodied data. MCTD is a conditional diffusion generation framework designed for long-horizon planning tasks; it decomposes complex planning tasks into hierarchical sub-tasks and utilizes critical meta-points to guide the trajectory of the denoising path. By integrating MCTS and fast jumpy denoising techniques, MCTD efficiently gets the path values of sub-planning nodes, thereby steering the global denoising process toward higher-value plans. This approach effectively mitigates the inherent instability of diffusion denoising paths and provides a clue for mining action correlations across heterogeneous embodiments.

Under the influence of windowed pyramid noise via Embodied Forcing, the action space is functionally partitioned into several segments, which we define as the sub-tasks required for the diffusion-based generation process. To provide effective guidance, we construct meta-action nodes derived from the nearest-neighbor points of heterogeneous embodiments performing the same task. The value of these sub-tasks during the denoising process is evaluated based on the distance between joint positions. Given a sub-task generated action $\mathbf{x}$ and a reference meta-action $\mathbf{y}$, the node value v is defined as follows:

$$v = \begin{cases} -\|\mathbf{x} - \mathbf{y}\|^2 + M_0, & \text{if } \dim(\mathbf{x}) = \dim(\mathbf{y}) \\ -\|\mathbf{x}_{1:d_{\min}} - \mathbf{y}_{1:d_{\min}}\|^2 + M_1, & \text{if } \dim(\mathbf{x}) \neq \dim(\mathbf{y}) \end{cases} \tag{8}$$

where M_0 and M_1 represent the base values of a node, with M_1 slightly exceeding M_0 to incentivize denoising conditioned on heterogeneous embodied data. $d_{\min}$ denotes the minimum functional dimensionality. This formulation is particularly critical for end-effector segments where zero-padding is utilized to account for morphological discrepancies.

The denoising process is structured as an iterative tree search consisting of four distinct phases:

- **Selection**: Starting from the white noise within the morphological tree, the diffusion head can either conduct self-denoising or leverage external embodiments as conditioning signals for the denoising process.
- **Expansion**: From the selected node, the diffusion model generates candidate action segments, conditioning on Embodied Forcing noise initialization.
- **Simulation**: To evaluate the expanded paths, we follow fast jumpy denoising to perform a rapid rollout evaluation. The potential value is caculated by Equation 8.
- **Backpropagation**: The value derived from the simulation step is propagated back through the tree hierarchy.

MPTD ensures that the denoising gradient is strategically biased toward action distributions that maintain functional consistency across different robotic morphologies, effectively mining action correlations within cross-embodied datasets.

C. Soft Prompts for Embodied Feature

To enhance the discriminative capacity of X-DiffVLA across heterogeneous robots, we utilize a soft-prompting mechanism inspired by X-VLA framework [43]. Recognizing that diffusion-based models are particularly sensitive to the latent prompt space, we incorporate domain-specific learnable parameters $P^E = \{p^i\}_{i=1}^e$ to effectively absorb cross-embodiment heterogeneity.

IV. Experiments

A. Experiments Setup

To evaluate the generalizability of our method across diverse robotic platforms, we conduct experiments in Robocasa [32], Isaac Gym [31], and real-world environments. While Robocasa serves as a high-fidelity benchmark with rich task scenarios and object rendering, its native support is limited to two parallel grippers and one dexterous hand. To increase the comprehensiveness of the experimental results, we integrate Isaac Gym to expand our evaluation of X-DiffVLA to a broader range of embodiments, including one gripper and two distinct types of dexterous hands across ten unique objects. These results demonstrate that X-DiffVLA exhibits robust universality when facing both cross-embodiment diversity and task complexity, ensuring a comprehensive evaluation of the model. Finally, we deploy X-DiffVLA on both parallel-gripper and dexterous-hand robots in real-world environments. This deployment serves to validate the effectiveness of our model's transition from simulation to physical hardware.

We utilize Being-H0 2B [29] as the backbone model to validate the performance of our action head. Within the Robocasa benchmark, we employ the Rethink Mount and GR1 as robotic bases, paired with Panda gripper, Robotiq-85 gripper, and Inspire hands as end-effectors. For Isaac Gym, we follow the experimental configuration of DexGraspNet [40], utilizing Panda gripper, Inspire hands, and Shadow hands to perform grasping tasks across ten distinct objects. Finally, for real-world validation, we deploy the Panda gripper and Inspire hands mounted on a 7-DoF Franka Research 3 (FR3) arm.

TABLE I: Results of X-DiffVLA in RoboCasa environment. We examine the impact of diffusion head under cross-embodied post-training by analyzing the reduction of two failure categories.

Method	Success Rate			Avg.	Failure Categories	
	Panda	Robotiq-85	Inspire		#1	#2
UniVLA	39.5%	37.0%	35.1%	37.2%	889	241
XR-1	45.2%	43.8%	34.9%	41.3%	863	193
X-VLA	50.1%	46.9%	45.8%	47.6%	816	127
GR00T-N1	39.6%	40.4%	38.5%	39.5%	854	235
π_0 +FAST	44.3%	45.8%	39.2%	43.1%	820	204
π_0	45.1%	45.2%	37.8%	42.7%	834	197
$\pi_{0.5}$	52.3%	51.1%	44.2%	49.2%	784	130
X-DiffVLA(Ours)	**67.2%**	**68.0%**	**58.3%**	**64.5%**	550	89

B. Baselines

To rigorously evaluate the performance of **X-DiffVLA**, we compare it with several baselines that represent different paradigms in VLA models:

- **GR00T-N1 [2], π_0 [3], π_0+FAST, $\pi_{0.5}$ [19]** represent the state-of-the-art (SOTA) methods in this field, employing flow-matching action heads.
- **UniVLA [4], X-VLA [43], XR-1 [12]** constitute the mainstream research in cross-embodied VLA. These frameworks facilitate cross-embodied knowledge transfer in pre-train stage through latent action representations, soft-prompt-based hardware descriptions, and unified video-motion encoding, respectively.

To maintain evaluation fairness, all models are post-trained using a unified action space across heterogeneous embodiments. Detailed configurations are available in the Appendix A.

C. Experimental Results

*1) **Is Diffusion Head Effective to Handle Multi-peaked Distributions in Different Embodiments**:* As a generative model capable of fitting multimodal distributions, diffusion models provide the theoretical foundation for generating action distributions tailored to diverse robotic embodiments. Within the Robocasa environment, we establish a benchmark based on the GR00T dataset, encompassing 30 task categories for each end-effector, as detailed in Appendix A. For each task, we perform post-training using a mixture of 50 trajectories. Evaluation is conducted through 20 test trials per task, with the results illustrated in the TABLE I.

Experimental results demonstrate that X-DiffVLA achieves superior performance, validating the effectiveness of diffusion models in modeling cross-embodiment distributions. To further analyze the diffusion head, we conduct an error analysis on failed task trajectories, categorizing them into two failure types: Mobility Failure (Failure #1) and Interaction Failure (Failure #2). Specifically, Failure #1 occurs when the end-effector fails to reach an approachable range of the object, whereas Failure #2 involves operational errors despite the end-effector being within the interaction workspace. The results indicate that X-DiffVLA significantly reduces the probability of both failure types, particularly Mobility Failures. This underscores its ability to achieve functional partitioning of action space and facilitates the global optimization of multi-peaked distributions. What's more, this provides empirical evidence that X-DiffVLA facilitates positive knowledge transfer across embodiments, whereby post-training on heterogeneous data significantly augments the model's proficiency in executing task-specific behaviors for a given robot morphology.

*2) **What enables Diffusion Models to handle diverse embodiment distributions**:* In experiments, we find that EBF is critical for X-DiffVLA, and it exhibits extreme sensitivity to noise parameters. We find that the noise technique acts as a latent guide, enabling the diffusion head to fit cross-embodiment distributions and facilitate conditional generation. We conduct a series of sensitivity tests on the pyramid noise hyperparameters, the results are summarized in the TABLE II. It shows that while excessively large decay coefficients reduce noise discriminability and hinder functional partitioning, overly small coefficients weaken causal coherence in the action space, resulting in unexecutable action generation. Furthermore, the impact of window sizes suggests that the diffusion head favors a functional decomposition of action generation, providing a robust empirical basis for task partitioning in MPTD. A window size of $W_n = 3$ partitions the action space into [base, gripper, hand], whereas $W_n = 9$ denotes a full decomposition of base by degrees of freedom; notably, neither achieves the best results.

*3) **Are EBF Useful for Flow-Matching Action Head**:* Given that flow-matching heads are diffusion-based and increasingly adopted in VLA architectures, we investigate whether noise-guided generation methods can yield similar performance gains. To this end, we conduct

TABLE II: Performance comparison under different hyperparameter settings (δ and W_n) in RoboCasa environment. When evaluating one variable, the other is held at its optimal value to ensure the reliability of the experimental results.

Hyperparameter	Success Rate			Avg.
	Panda	Robotiq-85	Inspire	
$\delta = 0.9$	59.5%	61.7%	45.9%	55.7%
$\delta = 0.5$	**67.2%**	**68.0%**	**58.3%**	**64.5%**
$\delta = 0.1$	41.3%	39.5%	31.7%	37.5%
$W_n = 9$	43.6%	46.7%	41.1%	43.8%
$W_n = 4$	**67.2%**	**68.0%**	**58.3%**	**64.5%**
$W_n = 3$	46.3%	45.5%	37.2%	43.0%

parallel experiments within the RoboCasa environment. Specifically, EBF is applied during the noise initialization phase of the flow-matching head, with results summarized in Table III. Our findings indicate that EBF fails to effectively enhance cross-embodiment generalization for flow-matching architecture. A plausible explanation is that the denoising process in diffusion models may need a more granular sampling during the high-noise initial stages, which are critical for capturing robot-specific structural constraints and generating corresponding actions. Conversely, although flow-matching also uses the EBF method, its primary advantage lies in efficient few-step sampling. In cross-embodied tasks, the simplified probability paths of flow-matching may struggle to capture the structural and functional variances across different robotic morphologies.

TABLE III: Study of EBF under diffusion head (DH) and flow-matching head (FH). For all configurations, a unified action space is employed, and the MPTD approach is omitted.

Method	Success Rate			Avg.
	Panda	Robotiq-85	Inspire	
FH	43.3%	43.1%	38.1%	41.5%
FH + EBF	42.5%	44.7%	40.6%	42.6%
DH	41.2%	39.4%	41.8%	40.8%
DH + EBF	**56.9%**	**58.3%**	**47.1%**	**54.1%**

*4) **What's the Relationship Between Different Embodiments Data**:* To further enhance the versatility of the diffusion head across more end-effectors, we conduct additional evaluations within the Isaac Gym environment using a diverse set of robotic hands, including the Panda gripper, Inspire hand, and Shadow hand. Following the benchmarking and data sourcing protocols established in DemoGrasp [39], we select 10 distinct objects, with 10 trajectories per object utilized for model post-training. More details can be seen in Appendix A. These experiments are executed to investigate whether X-DiffVLA can identify latent correlations between the actions of low-dimensional grippers and high-dimensional dexterous hands, thereby improving both model performance and generalization. The results are illustrated in the TABLE IV.

TABLE IV: Results of X-DiffVLA in Isaac Gym environment. Additional dexterous hand data is incorporated to evaluate the model's capability in capturing correlations across heterogeneous embodiments.

Method	Success Rate			Avg.
	Panda	Inspire	Shadow	
X-VLA	56.5%	57.0%	58.0%	57.2%
π_0	55.5%	52.5%	54.0%	54.0%
$\pi_{0.5}$	59.5%	58.5%	57.5%	58.5%
X-DiffVLA (Ours)	**73.5%**	**69.5%**	**70.0%**	**71.0%**

To provide a clearer analysis of the action generation results across different embodiments, we employ T-SNE [30] to visualize the final action distributions, as illustrated in Fig.3. To ensure consistency with the gripper's positions, we select the joint positions of the two most distant fingertips as the primary features for dexterous hands. Each task is evaluated over 50 trials, with the final joint positions being recorded for analysis, and we only choose one of them present in the paper. We can observe that despite the significant structural disparities between parallel grippers and dexterous hands, their action spaces exhibit a high degree of correlation. This demonstrates that X-DiffVLA successfully facilitates cross-embodiment knowledge transfer from similar tasks with the effective guidance of MPTD.

D. Real Robots

To evaluate the effectiveness of the X-DiffVLA action head, we conduct real-world validation using both Panda grippers and Inspire hands. For our hardware platform, we employ a unified FR3 robotic arm as the primary base. Real-robot datasets are collected via teleoperation, utilizing a GELLO device for arm control and Manus gloves for hand manipulation. More details can be seen in Appendix A. For post-training, we collect 10 trajectories for each of the 5 distinct objects in the pick-up tasks. And each task is rigorously evaluated over 20 independent trials. The experimental results are presented in the TABLE V and Fig.4.

From the Fig.4, we can observe that the Inspire hand prioritizes stabilizing objects using its two most distant fingers (the yellow part in the third image) before engaging the remaining digits for a full grasp. This behavior closely mirrors the grasping strategy characteristic of parallel grippers, providing compelling evidence that X-DiffVLA effectively leverages cross-embodiment data. Such strategic alignment further demonstrates the

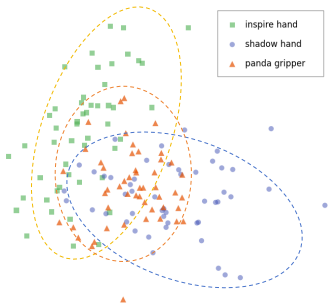

Fig. 3: **T-SNE visualization results of three embodiments in Isaac Gym.** The visualization compares the final joint positions of different end-effectors over 50 trials.

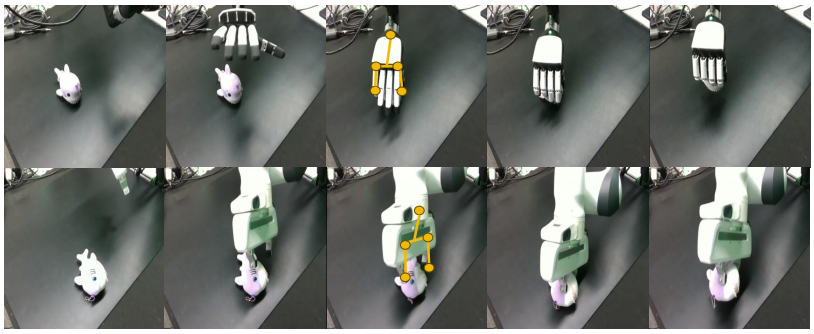

Fig. 4: **Visualization results of real world experiments.** The Inspire hand demonstrates a policy learned from parallel gripper data (highlighted in yellow). Such behavior underscores the effectiveness of X-DiffVLA in leveraging cross-embodied datasets to facilitate robust knowledge transfer.

TABLE V: Results of X-DiffVLA in real world experiments.

Method	Success Rate		Avg.
	Panda	Inspire	
X-VLA	51.0%	47.0%	49.0%
π_0	55.0%	47.0%	51.0%
$\pi_{0.5}$	59.0%	55.0%	57.0%
X-DiffVLA (Ours)	**67.0%**	**60.0%**	**63.5%**

TABLE VI: Ablation study of X-DiffVLA components in the RoboCasa environment. The experimental setup is consistent with the main experiments.

Method	Success Rate			Avg.
	Panda	Robotiq-85	Inspire	
X-DiffVLA	**67.2%**	**68.0%**	**58.3%**	**64.5%**
w/o SP	66.1%	65.4%	55.4%	62.3%
w/o MPTD	56.9%	58.3%	47.1%	54.1%
w/o EBF	46.7%	48.3%	35.5%	43.5%
w/o ALL	41.2%	39.4%	41.8%	40.8%

model's capacity for robust knowledge transfer between different robotic morphologies.

E. Ablation Study

We conduct ablation studies in the RoboCasa environment to investigate the individual contributions of each proposed module to X-DiffVLA. Specifically, w/o SP denotes the removal of the Soft Prompt mechanism; w/o MPTD and w/o EBF represent the models excluding Morphological Tree Diffusion and Embodied Forcing, respectively; and w/o ALL serves as the baseline utilizing only the initial diffusion head.

As presented in Table VI, the results demonstrate that both EBF and MPTD are indispensable, as they effectively capture the inherent disparities and underlying correlations across different embodiments, respectively. Furthermore, the soft prompt provides effective guidance for conditional generation, enabling the model to generate more precise actions tailored to specific robotic morphologies.

V. Conclusion

In this paper, we introduce X-DiffVLA, a novel and general VLA framework designed to support cross-embodied post-training through a unified diffusion-based action head. By integrating the Embodied Forcing (EBF) mechanism and Morphological Tree Diffusion (MPTD), our approach successfully captures both the unique morphological discrimination and the underlying behavioral correlations to generate more precise actions. Extensive evaluations in RoboCasa and Isaac Gym demonstrate that X-DiffVLA significantly outperforms existing baselines, achieving an average improvement of 15.3% and 12.5%, respectively. Furthermore, our real-world experiments validate the model's ability to facilitate effective knowledge transfer between grippers and dexterous hands. These results suggest that X-DiffVLA provides a scalable and robust pathway toward universal robotic control, laying the groundwork for more versatile and generalized embodied AI systems.

VI. Acknowledgement

This work was supported by NSFC in part under Grant 62136008, 62450001 and 62476008. This work was also supported by Beijing Major Science and Technology Project under Contract no.Z251100008125023 and Beijing Academy of Artificial Intelligence(BAAI).

References

[1] Erik Bauer, Elvis Nava, and Robert K Katzschmann. Latent action diffusion for

cross-embodiment manipulation. *arXiv preprint arXiv:2506.14608*, 2025.

[2] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[4] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

[5] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

[6] Guangyan Chen, Meiling Wang, Qi Shao, Zichen Zhou, Weixin Mao, Te Cui, Minzhao Zhu, Yinan Deng, Luojie Yang, Zhanqi Zhang, et al. See once, then act: Vision-language-action model with task learning from one-shot video demonstrations. *arXiv preprint arXiv:2512.07582*, 2025.

[7] Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. Conrft: A reinforced fine-tuning method for vla models via consistency policy. In *Proceedings of Robotics: Science and Systems*, 2025.

[8] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

[9] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer, 2006.

[10] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023.

[11] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.

[12] Shichao Fan, Kun Wu, Zhengping Che, Xinhua Wang, Di Wu, Fei Liao, Ning Liu, Yixue Zhang, Zhen Zhao, Zhiyuan Xu, et al. Xr-1: Towards versatile vision-language-action models via learning unified vision-motion representations. *arXiv preprint arXiv:2511.02776*, 2025.

[13] Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems*, 2024.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63 (11):139–144, 2020.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

[17] R Huang, MWY Lam, J Wang, D Su, D Yu, Y Ren, and Z Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 4157–4163. IJCAI: International Joint Conferences on Artificial Intelligence Organization, 2022.

[18] Xiaoyu Huang, Takara Truong, Yunbo Zhang, Fangzhou Yu, Jean Pierre Sleiman, Jessica Hodgins, Koushil Sreenath, and Farbod Farshidian. Diffuse-cloc: Guided diffusion for physics-based character look-ahead control. *ACM Transactions on Graphics (TOG)*, 44(4):1–12, 2025.

[19] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[20] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pages 9902–9915. PMLR, 2022.

[21] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan

Vuong, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025.

[22] Boyu Li, Haoran Li, Yuanheng Zhu, and Dongbin Zhao. Mat: Morphological adaptive transformer for universal morphology policy learning. *IEEE Transactions on Cognitive and Developmental Systems*, 16(4):1611–1621, 2024.

[23] Boyu Li, Siyuan He, Hang Xu, Haoqi Yuan, Yu Zang, Liwei Hu, Junpeng Yue, Zhenxiong Jiang, Pengbo Hu, Börje F Karlsson, et al. Dualthor: A dual-arm humanoid simulation platform for contingency-aware planning. *arXiv preprint arXiv:2506.16012*, 2025.

[24] Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Tian Nian, Liuao Pei, Shunbo Zhou, Xiaokang Yang, Jiangmiao Pang, et al. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv:2508.20072*, 2025.

[25] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37:40085–40110, 2024.

[26] Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

[27] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[28] Xin Liu, Haoran Li, and Dongbin Zhao. Videos are sample-efficient supervisions: Behavior cloning from videos via latent representations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[29] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[31] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[32] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

[33] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

[34] Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *ICLR*, 2024.

[35] Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2508.19236*, 2025.

[36] Zhenyu Wei, Zhixuan Xu, Jingxiang Guo, Yiwen Hou, Chongkai Gao, Zhehao Cai, Jiayu Luo, and Lin Shao. D (r, o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. *arXiv e-prints*, pages arXiv–2410, 2024.

[37] Adina Yakefu, Bin Xie, Chongyang Xu, Enwen Zhang, Erjin Zhou, Fan Jia, Haitao Yang, Haoqiang Fan, Haowei Zhang, Hongyang Peng, et al. Robochallenge: Large-scale real-robot evaluation of embodied policies. *arXiv preprint arXiv:2510.17950*, 2025.

[38] Jaesik Yoon, Hyeonseo Cho, Doojin Baek, Yoshua Bengio, and Sungjin Ahn. Monte carlo tree diffusion for system 2 planning. In *Forty-second International Conference on Machine Learning*.

[39] Haoqi Yuan, Ziye Huang, Ye Wang, Chuan Mao, Chaoyi Xu, and Zongqing Lu. Demograsp: Universal dexterous grasping from a single demonstration. *arXiv preprint arXiv:2509.22149*, 2025.

[40] Jialiang Zhang, Haoran Liu, Danshi Li, XinQiang Yu, Haoran Geng, Yufei Ding, Jiayi Chen, and He Wang. Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In *8th Annual Conference on Robot Learning*, 2024.

[41] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, XinQiang Yu, Jiazhao Zhang, Run-

pei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[42] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025.

[43] Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*, 2025.

[44] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.