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

LIBERO-X: Robustness Litmus for Vision-Language-Action Models

Guodong Wang*[1], Chenkai Zhang*[2], Qingjie Liu[2], Jinjin Zhang[2], Jiancheng Cai[1], Junjie Liu†[1], Xinmin Liu‡[1]
[1]Meituan, [2]Beihang University
https://github.com/meituan/LIBERO-X

***Abstract*—Reliable benchmarking is critical for advancing Vision–Language–Action (VLA) models, as it reveals their generalization, robustness, and alignment of perception with language-driven manipulation tasks. However, existing benchmarks often provide limited or misleading assessments due to insufficient evaluation protocols that inadequately capture complex distribution shifts. This work systematically rethinks VLA benchmarking from both evaluation and data perspectives, introducing LIBERO-X, a benchmark featuring: 1) A hierarchical evaluation protocol with progressive difficulty levels targeting three core capabilities: spatial generalization, object recognition, and task instruction understanding. This design enables fine-grained analysis of performance degradation under increasing environmental and task complexity; 2) A high-diversity training dataset collected via human teleoperation, where each scene supports multiple fine-grained manipulation objectives to bridge the train-evaluation distribution gap. Experiments with representative VLA models reveal significant performance drops under cumulative perturbations, exposing persistent limitations in scene comprehension and instruction grounding. By integrating hierarchical evaluation with diverse training data, LIBERO-X offers a more reliable foundation for assessing and advancing VLA development.**

I. Introduction

Vision–Language–Action (VLA) models [41, 14, 3, 32, 1, 9, 15, 18, 27] have recently made significant strides in robotic manipulation, emerging as a key paradigm for integrating perception, language understanding, and control. By processing visual observations and natural language instructions in an end-to-end manner, these models directly map multimodal inputs to low-level control signals and have achieved strong performance across a wide range of complex manipulation tasks. With rapid advances in VLA architectures and training algorithms, a central challenge for the field is now the systematic evaluation of their actual capabilities [31, 35, 20].

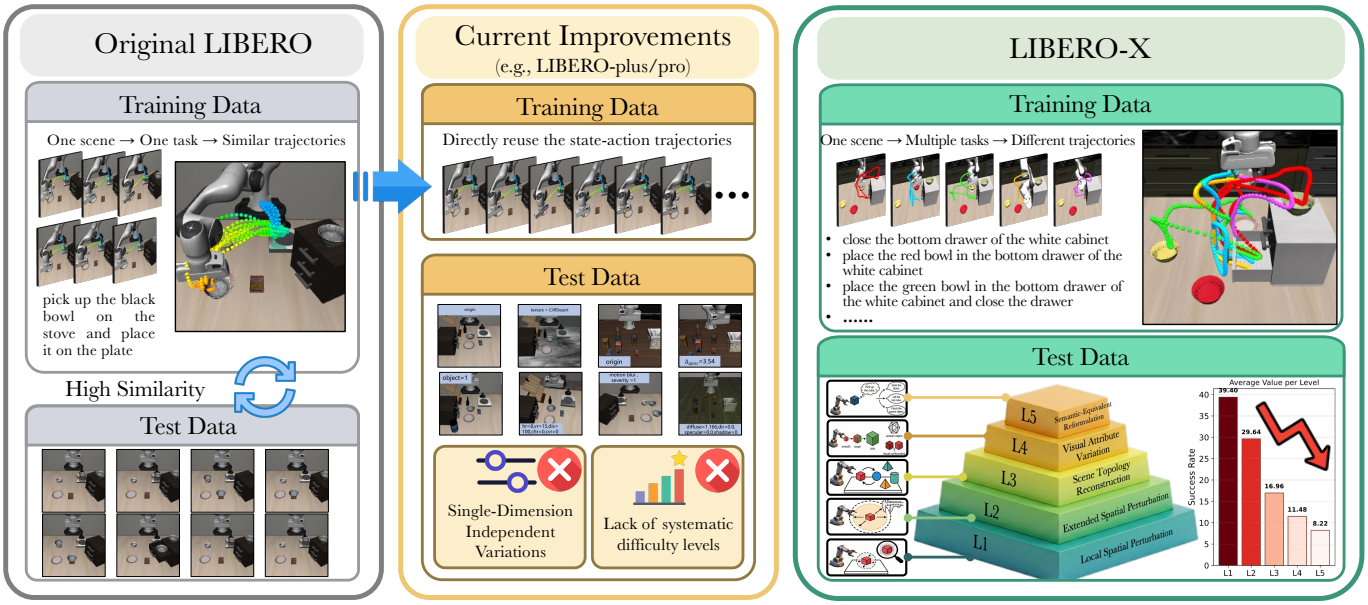
In response, the research community has introduced a series of standardized benchmarks, such as LIBERO [21] and SimpleEnv [20], which provide controlled, reproducible environments covering diverse task configurations and are widely used to assess multimodal understanding, long-horizon reasoning, and knowledge transfer. Despite impressive results on these existing benchmarks, current VLA evaluations are often confounded by limitations in both **evaluation protocols** and **data distribution**, leading to overly optimistic or even misleading conclusions about model capability.

First, existing benchmarks typically focus on individual perturbation types, primarily spatial ones. Extensions such as SafeLIBERO [12], LIBERO-PRO [40], and LIBERO-Plus [7] increase the diversity of test conditions by introducing additional perturbation factors (e.g., random obstacles for safety studies, or initial states, task instructions, and environment configurations). However, as shown in Figure 1 **(b)**, they still suffer from two major issues: (i) perturbations along different dimensions are mostly modeled independently, failing to capture the multi-source distribution shifts that arise in realistic deployment; and (ii) there is no systematic progression of difficulty from easy to hard, making it difficult to characterize performance degradation as environmental and task complexity increase. Consequently, these limitations hinder a comprehensive assessment of model robustness under realistic, compounded distribution shifts.

Second, evaluation is frequently unfaithful due to shortcomings in both the training and testing datasets. Specifically, training data lack sufficient diversity, and there is limited divergence between training and testing scenarios and tasks. Benchmarks such as LIBERO typically adopt test configurations that closely resemble the training settings, introducing only minor perturbations, such as small changes in initial object positions, as shown in Figure 1 **(a)**. Under this narrow train–test gap, models often attain near-saturated success rates, illustrating that superficially varied test scenarios do not necessarily reveal the model's true generalization ability. On the training data side, existing datasets exhibit limited diversity in tasks, scenes, and trajectories, which fundamentally constrains the attainable capability of learned models. A single scene is usually associated with only one or a few tasks, and demonstrations for the same task follow highly homogeneous action sequences and intermediate states, effectively collapsing into a small set of template-like strategies. Overfitting is further exacerbated under such narrow data distributions, causing models to memorize a limited number of stereotyped behaviors and to mechanically replay training actions [7, 40].

In this work, we improve benchmarking by jointly addressing evaluation hierarchy and training data diversity, as shown in Figure 1 **(c)**. Regarding the evaluation, we focus on three core capabilities of VLA models for robotic manipulation: spatial generalization, object recognition, and instruction understanding. Centered on these capabilities, we introduce LIBERO-X, an evolved version of the original LIBERO benchmark, where "X" denotes *eXtreme* to reflect its goal of pushing evaluation limits. LIBERO-X first defines a new evaluation protocol that delivers a progressively chal-

∗ Equal Contribution. † Corresponding author. ‡ Project Leader.



(a) Original LIBERO **(b)** Current Improvements **(c)** LIBERO-X

Fig. 1: **Comparison with related simulation benchmarks. (a)**: LIBERO lacks diversity by coupling scenes with homogeneous trajectories, inducing action overfitting, while its test set closely mirrors training data. **(b)**: Recent extensions reuse original training trajectories, introducing perturbations but model them independently, failing to capture complex distribution shifts. **(c)**: LIBERO-X enhances training via multi-task scenes with diverse trajectories. Its multi-level evaluation protocol progressively escalates complexity, revealing significant performance degradation across VLA models as difficulty increases, thus enabling rigorous robustness assessment.

lenging, multi-level test suite for assessing the generalization of VLA models in complex environments. Specifically, we adopt a *reuse–adjust–reconstruct* strategy over training scenes to construct a five-level hierarchy of fine-grained manipulation tasks. The first two levels introduce increasing magnitudes of spatial perturbations, progressing from minor positional jitter to broader spatial randomization. The third level alters scene topology (e.g., swapping target and placement locations), breaking fixed spatial associations learned during training. The fourth level incorporates unseen and confounding objects, and varies fine-grained attributes such as color and texture. The fifth level further incorporates semantic variations in task instructions to evaluate vision-language–action alignment. Higher levels are formed by adding new perturbation dimensions on top of lower levels, thereby progressively escalating difficulty to enable precise analysis of performance degradation as environmental and task complexity increase. Complementing this hierarchy, we implement a multi-label evaluation strategy where each task is annotated with fine-grained attributes: interaction type, subtask count, spatial relation, and object attribute. This granular labeling facilitates in-depth diagnostics, allowing researchers to pinpoint specific failure modes across distinct manipulation dimensions.

To mitigate the overfitting risks associated with the relatively homogeneous scenes and trajectories in the original LIBERO training set, we further construct a high-diversity training dataset via human teleoperation. This dataset fundamentally increases scene and task diversity by incorporating variations in object attributes (e.g., size, color, texture) and spatial relations (e.g., left/right, near/far, front/behind), together with language instructions. Importantly, a single scene is associated with multiple manipulation tasks, avoiding template-like scene–task–trajectory mappings. This design provides a data foundation by exposing models to broader task–scene distributions during training. Consequently, performance on LIBERO-X more faithfully reflects models' transferability and robustness in complex, dynamic environments.

Overall, by combining richer training diversity and multi-dimensional hierarchical perturbations at test time, LIBERO-X represents a systematic rethinking of VLA model evaluation for robotic manipulation. We expect it to serve as a comprehensive benchmark that can more effectively guide VLA development toward reliable deployment in robotics.

The main contributions of this work are three-fold.

- We propose LIBERO-X, a comprehensive benchmark for robotic manipulation that introduces progressively challenging difficulty levels. It incorporates multi-label annotations and jointly perturbs spatial layouts, object properties, and instruction semantics to systematically evaluate model robustness against multi-dimensional distribution shifts.
- We construct a high-diversity training dataset comprising 2,520 demonstrations, 600 tasks, and 100 scenes collected via human teleoperation, substantially increasing scene diversity and task granularity.
- Extensive experiments on fine-tuned representative VLA

models reveal notable performance degradation as task and scene complexity increase, highlighting critical limitations in scene comprehension and instruction grounding and offering empirical insights for future model design.

II. Related Work

Generalist Robot Policies. Early imitation learning methods, such as ACT [36] and Diffusion Policy (DP) [5], addressed sequential decision-making via action chunking and conditional denoising, respectively. Building on these, Octo [9] and RDT-1B [23] scaled these architectures on large-scale robot dataset (e.g. the Open X-Embodiment dataset [26]), demonstrating that increasing model capacity significantly improves multi-task manipulation.

Inspired by the generalization and reasoning capabilities of Large Language Models (LLMs), a new paradigm of VLA models has emerged. These methods aim to leverage the rich semantic representations of pretrained VLM backbones for robotic control, diverging primarily in their action modeling strategies. Discrete approaches like RT-2 [41] and OpenVLA [14] discretize continuous actions into text tokens, treating manipulation as a next-token prediction task. Recent work like FAST [27] further optimizes this tokenization for higher efficiency. Conversely, continuous approaches retain the precision of the action space but vary in their formulation. Some methods, such as RoboFlamingo [19] and OpenVLA-OFT [15], adopt direct regression heads optimized via mean squared error (MSE) loss. Others integrate more expressive generative policies: CogACT [16] separates high-level understanding from low-level execution by attaching a diffusion head to the VLM, while π_0 [2], $\pi_{0.5}$ [2] and GR00T [1] achieve state-of-the-art performance by incorporating flow matching directly with VLM backbones. To enhance the grounding of VLA models, recent approaches also focus on explicitly or implicitly incorporating additional information, such as spatial structure and geometry [29, 37] and reasoning and planning signals [33, 2].

Despite these advancements, evaluating VLA models remains challenging due to the high cost and reproducibility of real-robot experiments. To address this, we develop a comprehensive simulation benchmark to enable the systematic and reproducible assessment of VLA methods.

Robotic Manipulation Benchmarks. Standardized benchmarks are crucial for evaluating the generalization and robustness of robotic manipulation policies. While numerous benchmarks have been developed for real-world systems, this section primarily focuses on simulation-based benchmarks due to their experimental convenience and reproducibility. Early efforts such as RLBench [13] and LIBERO [21] have significantly advanced the field by establishing unified evaluation protocols. Recent benchmarks [7, 28, 40, 39, 17] have explored specific facets of generalization. For example, AGNOSTOS [39] and CALVIN [24] investigate transfer across tasks, while RoboCerebra [11] and VLABench [35] focus on long-horizon manipulation and complex scene-task understanding. Meanwhile, works such as LIBERO-PRO [40] and LIBERO-Plus [7] expand the evaluation by introducing controlled perturbations across dimensions like target placement, object attributes, and robot pose. These studies consistently find that high nominal success rates often decline sharply under minor perturbations, suggesting a reliance on memorization of training data rather than generalized skill acquisition.

Moreover, such evaluations usually test isolated distribution shifts, like visual appearance or object position, rather than the coupled, multi-factor shifts encountered in real environments. Other works organize test tasks into progressively harder levels to probe model limits. For example, GemBench [8] uses a multi-level protocol from novel placements to long-horizon tasks, observing performance degradation across levels. Yet its limited training scale and task coverage restrict its use for large-scale VLA assessment. Additionally, LIBERO-Plus [7] stratifies difficulty via multi-model averaging, which can introduce bias and lacks interpretability.

Another limitation of many benchmarks [13, 21] is that training diversity is often secondary to test design, with assessments frequently conducted using off-the-shelf checkpoints [7, 40]. However, the fine-tuning datasets for these VLA models generally lack sufficient variation in scenes, tasks, and trajectories, thereby fundamentally constraining the generalization capacity of the resulting policies [35]. While LIBERO-Plus [7] attempts to increase variation by re-rendering scenes using state-action pairs from the original LIBERO [21], it still inherits the limited diversity of trajectories.

By contrast, our benchmark integrates a highly diverse, teleoperation-collected training set with a multi-level evaluation protocol that systematically increases difficulty through spatial perturbations, object substitutions, and instruction rewrites. This approach enables a more faithful and systematic assessment of VLA models under complex, multi-dimensional distribution shifts.

III. LIBERO-X benchmark

We introduce LIBERO-X, a comprehensive benchmark designed to advance the evaluation and development of VLA models for robotic manipulation. LIBERO-X provides not only a multi-dimensional evaluation framework but also a highly diverse training dataset tailored to complex environmental conditions. Specifically, we first describe the training dataset collected via human teleoperation, which covers extensive task and scene variations. We then present the multi-level evaluation framework with fine-grained multi-label metrics, progressively incorporating spatial, object, and semantic perturbations. These dimensions capture the core factors governing perception, reasoning, and action coupling in VLA-based robotic manipulation, enabling a systematic assessment of model generalization and robustness across diverse environmental and task conditions.

A. Evaluation Framework

The LIBERO-X test set employs a multi-level evaluation framework combined with a fine-grained multi-label system, forming a two-dimensional assessment design (Figure 2).

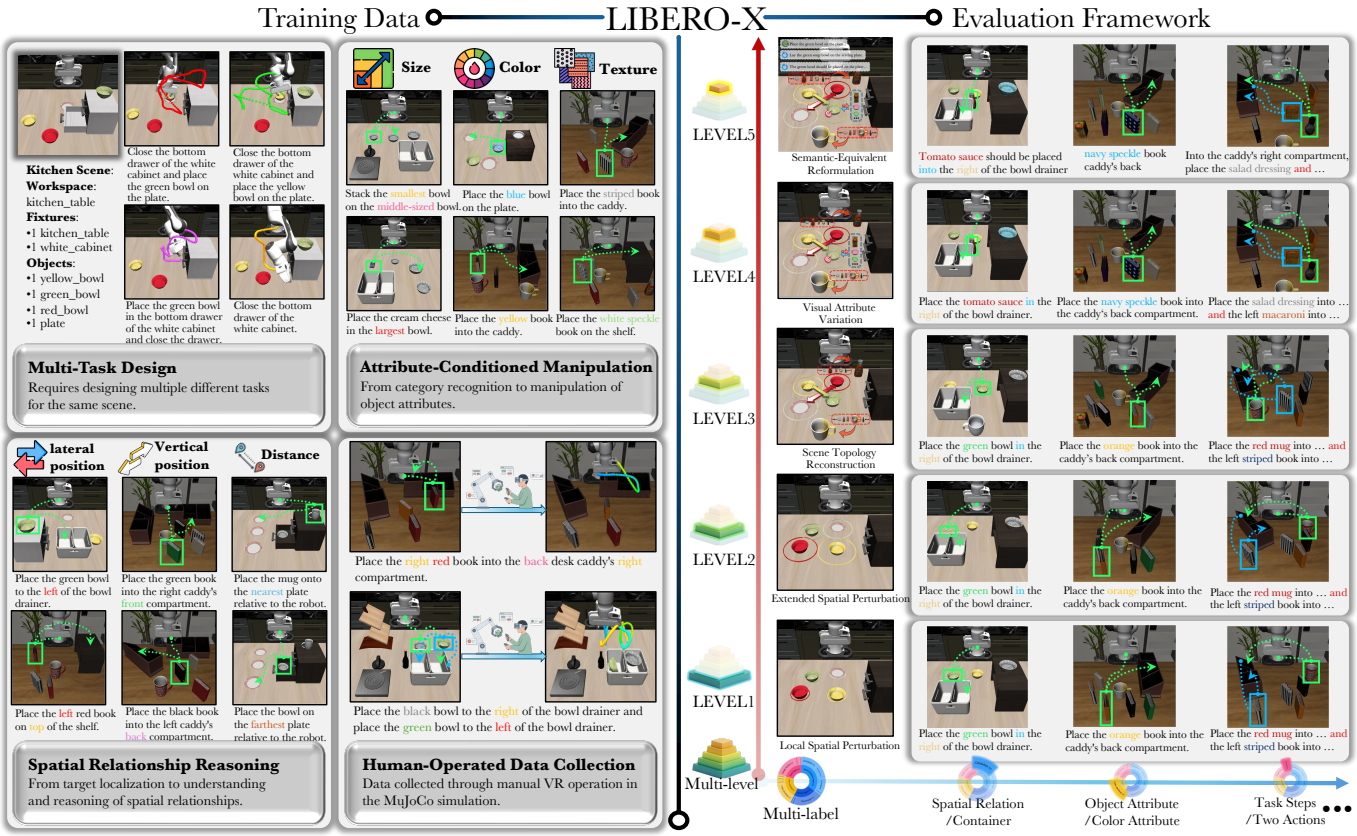


Fig. 2: **Overview of LIBERO-X.** LIBERO-X provides a high-diversity training dataset constructed through human teleoperation, along with multi-level and multi-label evaluation data.

Along the vertical axis, the multi-level framework evaluates models on tasks of progressively increasing difficulty, creating a curriculum-style, incremental assessment. Along the horizontal axis, the multi-label system assigns detailed labels to tasks, enabling fine-grained quantitative analysis. Together, these two components allow systematic measurement of performance under various perturbations and across diverse tasks and environments. Table I compares LIBERO-X with other simulation benchmarks, highlighting its multi-level framework and fine-grained labeling scheme.

1) Multi-level Evaluation: The vertical axis of the framework implements a multi-level evaluation following a progressive scene design that incrementally increases task and environmental complexity across spatial reasoning, object recognition, and instruction understanding. Each level builds upon the preceding one by superimposing additional perturbations, ensuring a cumulative assessment of the model's robustness:

Level 1 - Local Spatial Perturbation: Test scenes closely resemble the training distribution, with minor variations in object positions, designed to measure model stability under small-scale spatial changes.

Level 2 - Extended Spatial Perturbation: Building upon Level 1, spatial variations are amplified by dynamically adjusting the sampling radius according to object proximity, with a safety margin to expand layout space. This level assesses the model's ability to cope with large-scale positional changes.

Level 3 - Scene Topology Reconstruction: The relative positions of objects are altered, including swapping targets with distractors and adding irrelevant objects, to increase environmental complexity. This stage evaluates task execution under restructured object relationships and modified scene configurations.

Level 4 - Visual Attribute Variation: Objects vary in texture, color, and size; target objects may belong to unseen categories, and visually similar confounders are introduced. This level tests the model's discriminative ability, robustness to attribute changes, and generalization to unseen objects.

Level 5 - Semantic-Equivalent Reformulation: Instructions are rephrased while preserving task semantics, including synonym replacement (L5-1), word compression (L5-2), word order adjustment (L5-3), voice conversion (L5-4), and redundant descriptions (L5-5). This level measures the model's ability to generalize across diverse linguistic formulations of the same task.

2) Multi-label Evaluation: The horizontal axis of the framework incorporates a fine-grained multi-label evaluation system. This system enables a detailed, multi-dimensional analysis of model performance, particularly regarding its ability to manage fine-grained aspects of task execution. The multi-label evaluation is organized hierarchically, with each top-level category further divided into sub-categories (detailed definitions are provided in the appendix).

TABLE I: **Comparison with other simulation benchmarks. Name:** Name of each benchmark. **Simulator:** ManiSkill2 [10], MuJoCo [30], RLBench [13] and NVIDIA Isaac Sim [25]. **Date:** Release date. **Training Data:** Availability of training set. If provided, 🤖 indicates automation, 🧑‍🔧 indicates manual annotation. **Testing Data:** Availability of independent testing set. ✓ represents full compliance, ✗ represents non-compliance, and ✓✗ indicates partial compliance. **Multi-level Evaluation:** Adoption of multi-level framework. L1: Local Spatial Perturbation, L2: Extended Spatial Perturbation, L3: Scene Topology Reconstruction, L4: Visual Attribute Variation, and L5: Semantic-Equivalent Reformulation. **Multi-label Evaluation:** Task label across multiple dimensions. **IT:** Interaction Type. **SC:** Subtask Count. **SR:** Spatial Relation. **OA:** Object Attribute.

Name	Simulator	Date	Training Data	Testing Data	Multi-level Evaluation					Multi-label Evaluation			
					L1	L2	L3	L4	L5	IT	SC	SR	OA
RoboMIND [34]	Isaac Sim	2024.12	🧑‍🔧	✓	✗	✗	✗	✓✗	✗	✗	✗	✗	✗
COLOSSEUM [28]	RLBench	2024.05	✗	✓	✗	✗	✓✗	✓✗	✗	✗	✗	✗	✓
AGNOSTOS [39]	RLBench	2025.10	✗	✓	✗	✗	✗	✗	✗	✗	✓✗	✗	✗
GemBench [8]	RLBench	2025.03	🤖	✓	✓✗	✗	✗	✓✗	✗	✗	✓✗	✗	✓✗
VLATest [31]	Maniskill2	2025.05	✗	✓	✓	✗	✓✗	✓✗	✓✗	✓✗	✓✗	✓✗	✗
INT-ACT [6]	Maniskill2	2025.06	✗	✓	✓✗	✗	✗	✓	✓✗	✗	✗	✗	✗
RL4VLA [22]	Maniskill2	2026.01	🤖	✓	✓	✓	✓	✗	✓✗	✓✗	✓✗	✓✗	✓✗
LIBERO [21]	MuJoCo	2023.06	🧑‍🔧	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
LIBERO-OOD [17]	MuJoCo	2025.05	✗	✓	✗	✗	✓✗	✓✗	✗	✗	✗	✗	✗
RoboCerebra [11]	MuJoCo	2025.06	🧑‍🔧	✓	✗	✓✗	✓✗	✗	✗	✗	✗	✗	✗
LIBERO-PRO [40]	MuJoCo	2025.10	✗	✓	✗	✗	✓✗	✓✗	✓✗	✓✗	✓	✓✗	✓✗
LIBERO-Plus [7]	MuJoCo	2025.10	🤖	✓	✗	✓	✗	✗	✓✗	✓✗	✓✗	✗	✗
SafeLIBERO [12]	MuJoCo	2025.12	✗	✓	✗	✓✗	✗	✓✗	✗	✗	✗	✗	✗
LIBERO-X (Ours)	MuJoCo	2026.01	🧑‍🔧	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Interaction Type: This dimension categorizes object interactions into three types: (i) Pick-and-Place, where objects are grasped and moved to targets; (ii) Switch-based, such as opening/closing drawers or turning knobs, requiring precise gripper control; and (iii) Combined, integrating both types and requiring flexible transitions. It evaluates the model's ability to execute diverse manipulation strategies.

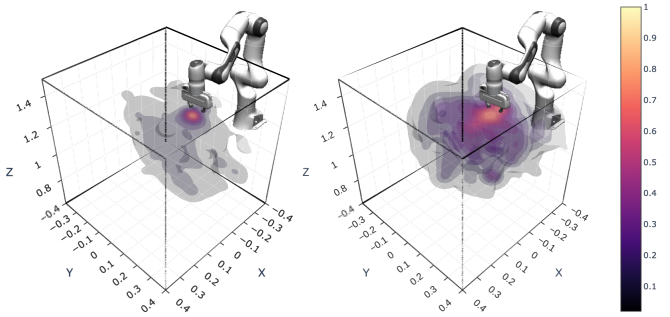
Subtask Count: This dimension captures the number of subtasks required to complete a task, ranging from single-step to three-step sequences. It assesses the model's planning, sequential reasoning, and ability to maintain task coherence over long-horizon tasks.

Spatial Relation: This dimension characterizes the relative positions and spatial arrangements of objects within a task. It evaluates the model's spatial reasoning and environmental perception by measuring how effectively it interprets spatial relationships and their impact on task execution.

Object Attribute: This dimension describes the visual object properties, including color, size, texture, and other fine-grained features. It evaluates the model's ability to recognize and leverage object attributes to accomplish tasks, emphasizing the role of precise visual perception in task success.

B. Training Dataset

To mitigate the overfitting risk arising from the limited scene diversity in the original LIBERO, where each scene is typically associated with a single task and the collected trajectories exhibit high similarity, LIBERO-X constructs a training dataset that better reflects complex environmental variability. Specifically, the dataset consists of 2,520 demonstration trajectories spanning 600 tasks and 100 scenes, as shown on the left side of Figure 2. Unlike the original LIBERO [21], which emphasizes manipulating targets by category thus encourages learning

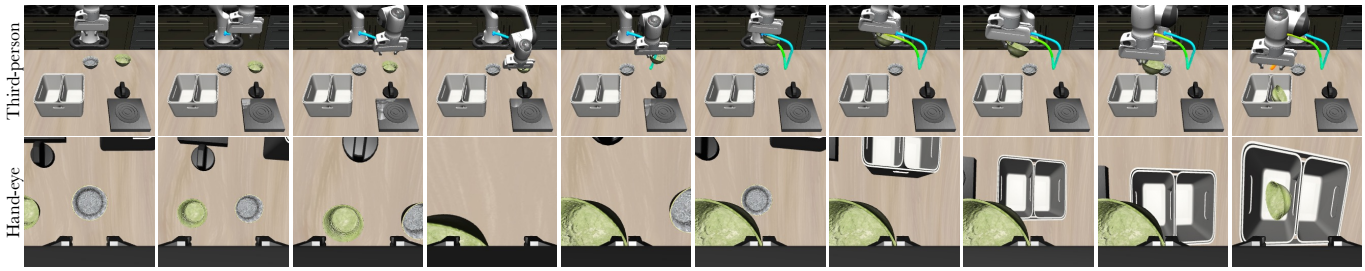


(a) LIBERO **(b)** LIBERO-X

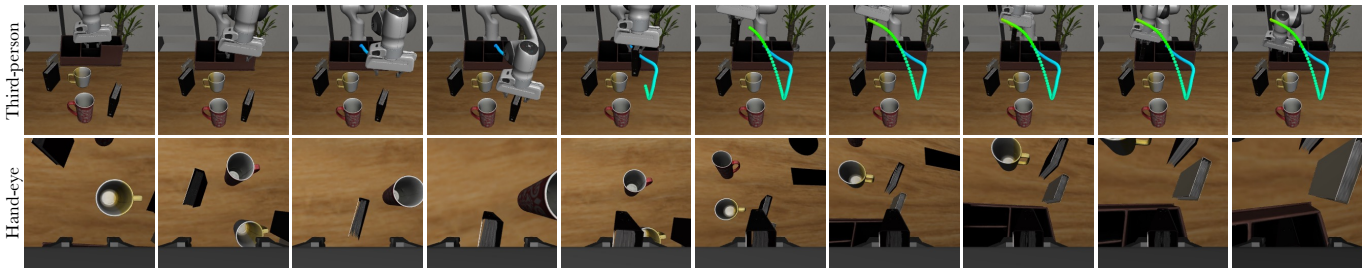
Fig. 3: **Distribution visualization of training trajectories.**

category-specific action patterns, LIBERO-X introduces finer-grained task-level extensions to expose models to diverse task formulations and workspace configurations:

- **Multi-Task Scene Design**: In contrast to LIBERO [21], which is predominantly composed of single-task scenes, resulting in an average of 2.6 tasks per scene, our dataset significantly increases task density to an average of 6 tasks per scene. This design entails multi-object interactions and versatile manipulation skills, requiring models to adapt their manipulation strategies under identical visual and physical conditions.
- **Attribute-Conditioned Manipulation**: Actions are explicitly conditioned on fine-grained object properties (e.g., size, color, texture) rather than broad categories. To rigorously test perceptual grounding, scenes include distractor objects with similar attributes, requiring the model to distinguish the target amidst visual ambiguity.
- **Spatial Relationship Reasoning**: Tasks extend beyond



(a) **Kitchen scene.** Task instruction: *Place the green bowl to the right of the bowl drainer.*



(b) **Study scene.** Task instruction: *Place the right black book into the desk caddy's left compartment.*

Fig. 4: **Practical examples from the LIBERO-X training dataset.**

target localization to require understanding and reasoning about spatial relationships among objects, including left/right, front/back, and near/far.

With this training setup, the model is encouraged to learn the composable semantic correspondence between target objects and their language descriptions, rather than relying on template instructions or fixed scene structures to make "memorized" decisions. By increasing the diversity of tasks and scenes at the data source, LIBERO-X raises the upper bound of model performance, allowing models to truly acquire generalizable manipulation strategies. As a result, performance variations observed across subsequent LIBERO-X test levels more faithfully reflect the model's transferability and robustness in dynamic environments.

1) Training Trajectory Collection: The trajectories were collected via VR teleoperation using a Meta Quest 3 in MuJoCo [30], facilitated by LeRobot [4]. During data collection, strict protocols were followed: state-action sequences were recorded at 20 Hz, where the state included the 6D Cartesian pose of the end effector and the gripper open/close status, and actions consisted of Cartesian increments and gripper commands. Multi-view images (third-person and hand-eye), were rendered by rolling out these recorded actions. To ensure quality, trajectories with noticeable pauses or non-optimal paths were discarded, retaining only smooth and successful demonstrations. As shown in Figure 3, heatmaps visualize the trajectory distributions, where brighter regions correspond to areas with higher trajectory density. Comparing with LIBERO, LIBERO-X exhibits a broader spread and higher trajectory density, demonstrating its greater diversity.

2) Example Training Demonstrations: The training dataset is built upon **Kitchen** and **Study** base scenes. By integrating diverse object categories, we created 100 scenarios encompassing 600 distinct tasks. These tasks cover object types, ranging from rigid to articulated objects, and include multi-object and attribute-conditioned interactions. Representative examples are illustrated in Figure 4.

IV. Experiments

A. Experimental setup

We selected five representative models for evaluation: OpenVLA-OFT [15], X-VLA [38], GR00T-N1.5 [1], π_0 [3] and $\pi_{0.5}$ [2]. Architecturally, these methods share a unified paradigm where a VLM backbone is utilized to extract latent representations from visual observations and language instructions, followed by an action head to generate the control sequences. All models were evaluated using their official configurations. Specifically, we employed a behavior cloning approach by performing supervised fine-tuning (SFT) on our self-collected training dataset. Detailed configurations for each model are listed in the appendix. To ensure reproducibility, we recorded the initial state configurations, including the placement and status of all objects, to guarantee consistent starting conditions across runs. Each task was evaluated over 10 trials with slight spatial perturbations applied to object positions during rollouts, following the LIBERO benchmark protocol [21]. Furthermore, the task time limit was dynamically set to 1.1 times the average human completion time, scaling with task difficulty. This temporal buffer accounts for minor deviations in execution speed, such as a model approaching a target more slowly than a human demonstrator, thereby preventing the unfair penalization of near-complete trajectories while maintaining realistic evaluation standards.

B. Multi-level Evaluation

Table II summarizes the evaluation results of the models under the multi-level framework. Notably, in contrast to the near-saturated performance (often $\sim$90%) [7, 40] in LIBERO [21],

TABLE II: **Success Rates under Multi-level Evaluation.** * denotes the model fine-tuned with LIBERO-Plus [7].

Model	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
OpenVLA-OFT*	0.9	0.6	0.2	0.1	0.0
OpenVLA-OFT	29.0	17.6	8.8	6.4	4.2
π_0	29.4	21.9	11.0	7.6	5.1
X-VLA	30.1	22.6	10.3	6.0	4.1
GR00T1.5	43.3	32.9	18.7	13.3	9.7
$\pi_{0.5}$	65.2	53.2	36.0	24.1	18.0

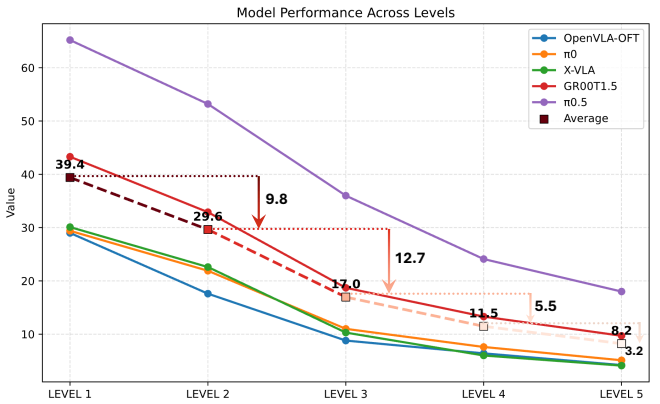


Fig. 5: **Success Rate Decline across Multi-level Evaluation.**

our benchmark prevents metric saturation through richer training diversity and multi-level test shifts, yielding a broader performance spread (4%∼65% success rates) that more clearly differentiates model capabilities. At Level 1, which involves only minor spatial perturbations relative to the training set, the average success rate across the five models is merely 39.4%, with even the top-performing $\pi_{0.5}$ capped at 65.2%. This sharp contrast not only underscores the challenging nature of our benchmark but also exposes the limitations of existing evaluation systems, where performance saturation often masks true model differences.

This limitation is further exemplified by the failure of data scaling alone to improve generalization. Specifically, fine-tuning OpenVLA-OFT on LIBERO-Plus, despite its much larger trajectory volume (20,000+ vs. 2,520) and shared simulation environment, yields near-zero performance on LIBERO-X test sets. The root cause lies in trajectory diversity. LIBERO-Plus reuses LIBERO trajectories and thus inherits their narrow distribution (Figure 3), whereas LIBERO-X leverages human teleoperation to generate significantly more diverse trajectories, scenes, and tasks. Consequently, models trained on LIBERO-X achieve substantially higher performance, demonstrating that trajectory diversity not merely data quantity is the key driver of robust generalization.

Despite the overall difficulty, $\pi_{0.5}$ maintains the most robust generalization across all levels. We attribute this robustness to its joint training strategy, which effectively incorporates pre-training on the Internet-scale data and sub-task planning. However, for all models, a significant downward trend is observed as task complexity increases. With the introduction of spatial, object-level, and semantic variations from Level 1 to Level 5, the models exhibit an average performance decline of 31.2%. These results highlight the increasing difficulty of the tasks and provide insights into the models' sensitivity to high-level perturbations. Based on these quantitative evaluations, we summarize our key findings as follows:

Finding 1: VLA models struggle with spatial extrapolation beyond training distributions. While models maintain reasonable performance under minor jitters (Level 1), the transition to Level 2, characterized by larger spatial shifts, triggers a notable 9.8% drop in average success rate (Figure 5). This decline suggests that current VLM-based policies tend to overfit to the specific spatial configurations in the training data, relying on memorized pixel-level patterns rather than acquiring a robust spatial understanding. Consequently, performance degrades significantly when targets are extrapolated to out-of-distribution positions.

Finding 2: Structural invariance is lacking when facing topological perturbations. Beyond mere coordinate shifts, Level 3 challenges the models' understanding of scene topology by altering the relative arrangement of objects (e.g., swapping positions). Crucially, as shown in Figure 5, this level witnesses the most substantial performance degradation of 12.7%, indicating that models lack structural invariance. They struggle to adapt to new scene layouts where the semantic relationships between objects remain valid but their topological organization has been reconstructed, highlighting a failure in reasoning about relative spatial relations.

TABLE III: **Level 4 Accuracy - Overall, Unseen Objects (UO), and Confounding Objects (CO).**

Category	OpenVLA-OFT	π_0	X-VLA	GR00T1.5	$\pi_{0.5}$
Overall	6.4	7.6	6.0	13.3	24.1
UO	0.8	1.1	0.8	2.1	7.2
CO	8.6	9.6	7.5	16.9	28.9

Finding 3: Data diversity enables emergent generalization, though semantic grounding for novel objects remains a bottleneck. While models trained on standard datasets typically fail on out-of-distribution tasks [21], our diversified LIBERO-X dataset enables the model to maintain non-zero success rates, suggesting that diversity helps mitigate overfitting to specific training scenes. However, a closer inspection of Table III reveals a significant performance gap: success rates on confounding objects are notably higher than those on unseen objects. This disparity indicates that while the model gains robustness against visual distractions, it still struggles to ground instructions to novel visual concepts. The lower performance on unseen objects suggests that despite the VLM's capabilities, the alignment between pre-trained representations and the control policy remains imperfect for handling unfamiliar visual features.

Finding 4: Language instruction variations can influence model performance. Table IV reveals a modest 3.26% drop in success rate from Level 4 to Level 5, indicating that phrasing changes affect execution. Crucially, voice conversion proved to be the least disruptive variation, achieving the highest success rate among all rephrasing methods. This implies that the model

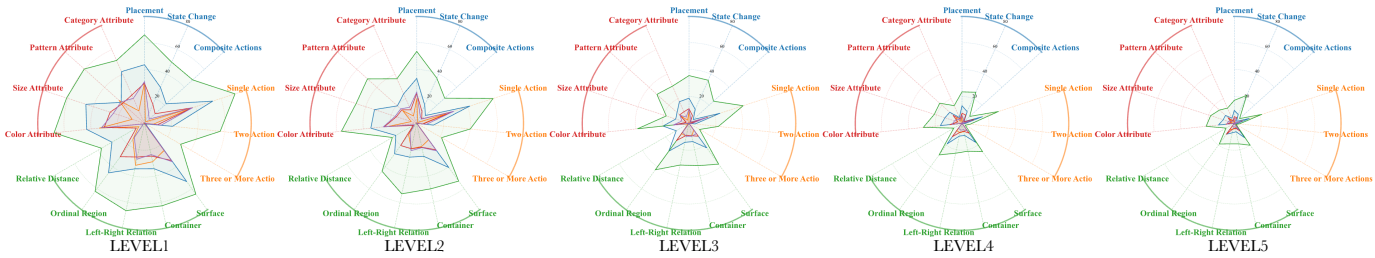


Fig. 6: **Multi-label Evaluation.** ■ denotes $\pi_{0.5}$, ■ GR00T1.5, ■ π_0, ■ X-VLA, and ■ OpenVLA-OFT.

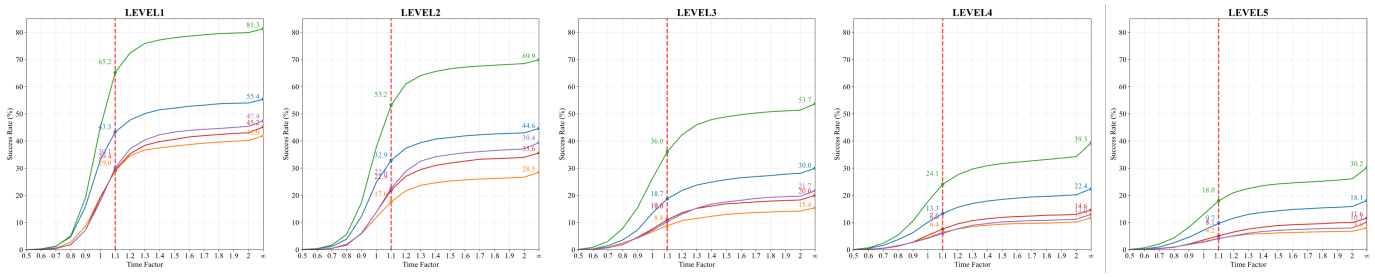


Fig. 7: **Impact of Time Factor Thresholds on Success Rate.** The x-axis displays time factor multipliers relative to the manual standard, where ∞ denotes an unbounded threshold.

TABLE IV: **Level 5 Accuracy - Overall and by Semantic Equivalence Reconstruction Method.**

Category	OpenVLA-OFT	π_0	X-VLA	GR00T1.5	$\pi_{0.5}$
Overall	4.2	5.1	4.1	9.7	18.0
5-1 (Synonym)	5.1	3.9	4.3	8.7	18.0
5-2 (Concise)	5.0	5.0	4.2	8.3	14.6
5-3 (Reorder)	4.4	5.1	3.0	8.4	18.0
5-4 (Voice)	5.2	7.5	5.1	10.8	19.6
5-5 (Verbose)	1.2	4.2	3.9	12.5	19.6

maintains better alignment with instructions under syntactic voice shifts compared to other forms of linguistic conversion.

C. Multi-label Evaluation

Figure 6 visualizes success rates for task labels across five levels. Macroscopically, the progressive shrinking of the enclosed area from Level 1 to 5 illustrates general performance degradation due to escalating complexity. Crucially, we observe a consistent alignment in performance patterns. Within the same level, methods exhibit synchronized variations; dimensions challenging for one model tend to be universally difficult, reflecting inherent task difficulty. Furthermore, across levels, the relative performance hierarchy remains stable. The degradation manifests as a proportional contraction, where models maintain comparative strengths and weaknesses even as absolute success rates decline.

Finding 5: Task horizon length critically limits performance. As illustrated by the "Subtask Count" category (orange axes ■), models perform best on single-step tasks but degrade sharply as sequence length increases, dropping to near zero for tasks with three or more steps. Notably, for tasks requiring three or more steps, success rates for nearly all models drop sharply toward zero. This non-linear performance drop-off suggests that current VLA models struggle with the compounding errors inherent in sequential manipulation. It reveals a fundamental deficiency in long-horizon reasoning and temporal consistency, where models fail to maintain precise execution over extended periods. Consequently, future research must prioritize error mitigation and long-term planning to enable robust multi-step manipulation.

D. Time Limit Threshold Differences Evaluation

Figure 7 illustrates the impact of time limits set relative to the manual standard, which is defined as the average duration of human demonstration. For instance, by applying multipliers such as 0.8 (stringent) and 1.5 (relaxed) to this baseline, we evaluate the model's adaptability to varying temporal pressures ranging from 0.8 to 1.5 times the human average.

Finding 6: Relaxed time limits buffer execution imperfections, though benefits diminish beyond a saturation point. As shown in Figure 7, stringent constraints ($0.8\times$) severely penalize performance. We attribute this to inevitable inference or control errors that result in deviations from the optimal trajectory, rendering most task completion impossible under tight deadlines. Conversely, extending limits yields substantial gains, particularly between $1.0\times$ and $1.1\times$, justifying our default selection of $1.1\times$. We avoid excessive limits (e.g., $1.5\times$ or beyond) to prevent success via brute-force trial-and-error. Nevertheless, performance plateaus beyond $1.3\times$, suggesting that remaining failures arise from fundamental capability deficits like perceptual hallucinations or logic errors rather than temporal insufficiency, thus further time extension cannot compensate for these intrinsic limitations.

V. Conclusion

This work introduces LIBERO-X, a comprehensive benchmark for evaluating vision–language–action models under realistic, multi-dimensional distribution shifts. By combining a hierarchically structured evaluation protocol with a high-diversity teleoperation-based training dataset, LIBERO-X enables systematic analysis of model generalization across spatial layouts, object attributes, and task semantics. Experimental results reveal significant performance degradation as complexity increases, exposing limitations in scene understanding and instruction grounding. Overall, LIBERO-X provides a more faithful and rigorous framework for assessing VLA models and guiding future research toward robust robotic deployment.

Acknowledgements. We express our sincere gratitude to Prof. Lei Zhang, chief scientist at the International Digital Economy Academy (IDEA), for his valuable feedback and insightful suggestions on this paper. We are also deeply thankful to all the robot operators for their tireless efforts in collecting high-quality robot manipulation data.

References

[1] 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.

[2] Kevin Black, Noah Brown, James Darpinian, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of Conference on Robot Learning (CoRL)*, 2025.

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi 0$: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[4] Remi Cadene, Simon Alibert, Alexander Soare, et al. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024.

[5] 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. *International Journal of Robotics Research (IJRR)*, 44(10-11):1684–1704, 2025.

[6] Irving Fang, Juexiao Zhang, Shengbang Tong, and Chen Feng. From intention to execution: Probing the generalization boundaries of vision-language-action models. *arXiv preprint arXiv:2506.09930*, 2025.

[7] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. Liberoplus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

[8] Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[9] Dibya Ghosh, Homer Rich Walke, Karl Pertsch, et al. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[10] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[11] Songhao Han, Boxiang Qiu, Yue Liao, Siyuan Huang, Chen Gao, Shuicheng YAN, and Si Liu. Robocerebra: A large-scale benchmark for long-horizon robotic manipulation evaluation. In *Proceedings of the Conference on Neural Information Processing System Datasets and Benchmarks Track (NeurIPS)*, 2025.

[12] Songqiao Hu, Zeyi Liu, Shuang Liu, Jun Cen, Zihan Meng, and Xiao He. Vlsa: Vision-language-action models with plug-and-play safety constraint layer. *arXiv preprint arXiv:2512.11891*, 2025.

[13] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):3019–3026, 2020.

[14] 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 *Proceedings of Conference on Robot Learning (CoRL)*, 2024.

[15] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[16] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

[17] Quanyi Li. Task reconstruction and extrapolation for π_0 using text latent. *arXiv preprint arXiv:2505.03500*, 2025.

[18] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[19] Xinghang Li, Minghuan Liu, Hanbo Zhang, et al. Vision-language foundation models as effective robot imitators. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[20] Xuanlin Li, Kyle Hsu, Jiayuan Gu, et al. Evaluating real-world robot manipulation policies in simulation. In *Proceedings of Conference on Robot Learning (CoRL)*, 2024.

[21] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In

Proceedings of the Conference on Neural Information Processing System (NeurIPS), 2023.

[22] Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What can RL bring to VLA generalization? an empirical study. In *Proceedings of the Conference on Neural Information Processing System (NeurIPS)*, 2026.

[23] 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[24] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022.

[25] NVIDIA. NVIDIA Isaac Sim: Robotics simulation and synthetic data. https://developer.nvidia.com/isaac-sim, 2023.

[26] Abby O'Neill, Abdul Rehman, Maddukur, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration0. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[27] Karl Pertsch, Kyle Stachowicz, Brian Ichter, et al. Fast: Efficient action tokenization for vision-language-action models. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[28] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The COLOSSEUM: A benchmark for evaluating generalization for robotic manipulation. In *Proceedings of Robotics: Science and Systems (RSS) Workshop: Data Generation for Robotics*, 2024.

[29] Delin Qu, Haoming Song, Qizhi Chen, et al. Spatialvla: Exploring spatial representations for visual-language-action model. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[30] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[31] Zhijie Wang, Zhehua Zhou, Jiayang Song, Yuheng Huang, Zhan Shu, and Lei Ma. Vlatest: Testing and evaluating vision-language-action models for robotic manipulation. *Proceedings of the ACM on Software Engineering*, 2(FSE):1615–1638, 2025.

[32] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. Tinyvla: Toward fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters (RA-L)*, 10(4): 3988–3995, 2025.

[33] Junjie Wen, Yichen Zhu, Minjie Zhu, et al. DiffusionVLA: Scaling robot foundation models via unified diffusion and autoregression. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[34] Kun Wu, Chengkai Hou, Jiaming Liu, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[35] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[36] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[37] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, et al. 3d-vla: a 3d vision-language-action generative world model. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[38] Jinliang Zheng, Jianxiong Li, Zhihao Wang, et al. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

[39] Jiaming Zhou, Ke Ye, Jiayi Liu, Teli Ma, Zifan Wang, Ronghe Qiu, Kun-Yu Lin, Zhilin Zhao, and Junwei Liang. Exploring the limits of vision-language-action manipulation in cross-task generalization. In *Proceedings of the Conference on Neural Information Processing System (NeurIPS)*, 2025.

[40] Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. *arXiv preprint arXiv:2510.03827*, 2025.

[41] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of Conference on Robot Learning (CoRL)*, 2023.