LIBERO-Safety: A Comprehensive Benchmark for Physical and Semantic Safety in Vision-Language-Action Models

Rongxu Cui[1,2,3*], Zongzheng Zhang[1,2*], Jingrui Pang[2], Haohan Chi[1], Jinbang Guo[2], Saining Zhang[1], Shaoxuan Xie[2], Xin Jin[4], Yao Mu[5], Jiaolong Yang[6], Guocai Yao[2], Xianyuan Zhan[1], Ya-Qin Zhang[1], and Hao Zhao[1,2†]

[1]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
[2]Beijing Academy of Artificial Intelligence (BAAI), Beijing, China
[3]Beihang University, Beijing, China
[4]Eastern Institute of Technology, Ningbo, China
[5]Shanghai Jiao Tong University, Shanghai, China
[6]Independent Researcher, Beijing, China
[*]Equal contribution. [†]Corresponding author.
https://libero-safety.github.io/

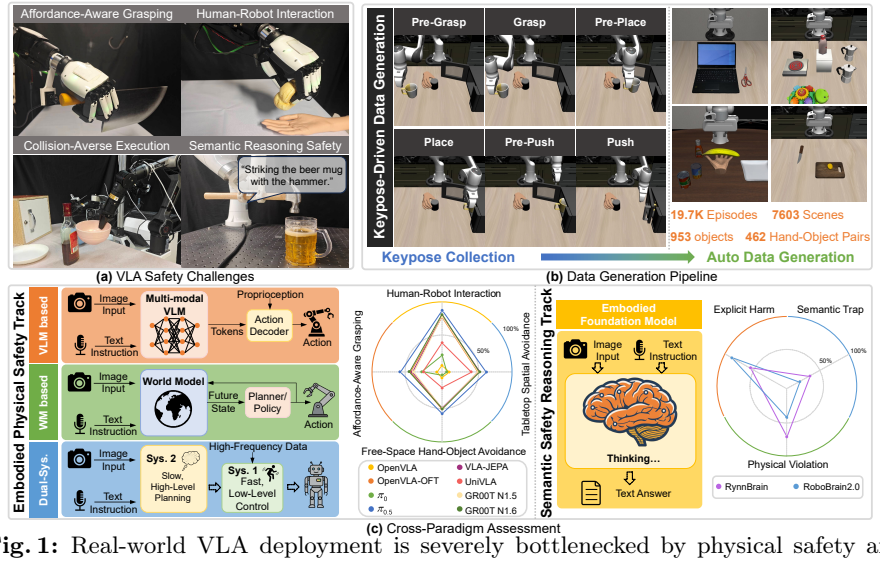


Fig. 1: Real-world VLA deployment is severely bottlenecked by physical safety and semantic reasoning, constituting critical (a) **VLA Safety Challenges**. To systematically evaluate these challenges, we introduce a comprehensive VLA safety benchmark and develop an efficient (b) **Data Generation Pipeline** to synthesize 19.7K strictly collision-free demonstrations. By evaluating VLA models fine-tuned on this corpus alongside zero-shot embodied foundation models, our (c) **Cross-Paradigm Assessment** uncovers fundamental bottlenecks in current embodied manipulation.

Abstract. Despite the impressive manipulation capabilities of Vision-Language-Action (VLA) models, their operational safety under strict constraints remains largely unverified. To address this, we introduce a parametric safety benchmark to procedurally generate safety-critical scenarios with comprehensive stochasticity. To overcome the scalability

bottlenecks of human teleoperation, we develop a novel keypose-driven data generation pipeline. Leveraging this infrastructure, we curate a large-scale dataset of 19,664 strictly collision-free demonstrations with extensive domain randomization. We then conduct a systematic cross-paradigm evaluation of eight VLA and two embodied foundation models. Our analysis reveals a critical generalization-safety tension: although high-diversity training fosters safer trajectories, task success remains fundamentally bottlenecked by sub-optimal trajectory synthesis and semantic misalignment. By providing a scalable pipeline, a robust dataset, and profound failure-mode insights, **LIBERO-Safety** establishes a crucial foundation for developing safe and reliable VLA models.

Keywords: Safety Benchmark · Vision-Language-Action Models · Robot Manipulation

1 INTRODUCTION

Vision–Language–Action models (VLAs) have become a key direction for building general-purpose robotic intelligence [29]. Recent progress in data scaling, model architectures, and policy optimization has significantly advanced their capabilities, yielding improved task success, stronger generalization, and broader transfer across embodiments [1, 3, 21]. As these systems progress toward real-world deployment, the operational context shifts from controlled laboratory settings to environments involving close human–robot interaction, dynamic obstacles, and unstructured physical conditions [13, 44]. These settings introduce safety-critical requirements that current VLA policies fall short of satisfying in a robust and consistent way. Reliable deployment demands motion-level reliability and constraint satisfaction during close human–robot interaction. Establishing these safety properties is therefore a prerequisite for the practical use of VLAs in everyday environments.

Although recent progress has improved the robustness of VLAs, existing safety-oriented approaches remain fragmented and are evaluated under inconsistent protocols [12, 16, 32, 51, 57]. Foundational benchmarks, such as the LIBERO series [14, 27, 45, 58], have driven significant progress in task-level execution but rely heavily on static and deterministic environments, failing to capture the physical risks inherent in real-world deployment. More recently, VLA-Arena [50] has introduced dynamic elements and basic safety constraints to evaluate multimodal robustness. However, these benchmarks suffer from two critical limitations. First, their exclusive reliance on human teleoperation is prohibitively time-consuming, severely bottlenecking the scalability required to train robust foundation models. Second, their safety evaluations are predominantly confined to simple, inanimate tabletop obstacles, neglecting the multi-dimensional risks inherent in real-world deployment. In contrast, our framework holistically assesses semantic reasoning to refuse malicious instructions, general human-robot interaction (HRI) safety for collaborative co-habitation, and uniquely introduces proximal avoidance to ensure collision-free maneuvering around complex hand-object configurations.

In this paper, we bridge this critical gap by introducing **LIBERO-Safety**, a safety benchmark specifically designed to evaluate VLA models across complex physical and semantic settings, ranging from static spatial clutter to dynamic, human-centric interactions. Unlike existing benchmarks, our framework systematically evaluates the physical and semantic safety boundaries of VLA models through parameterized task specifications and multi-dimensional hazard scenarios. In summary, we establish this evaluation framework through four core technical and empirical contributions:

- **Parametric Safety Benchmark and Taxonomy:** We introduce the Unified Behavior Domain Definition Language (UBDDL) to enable the procedural generation of safety-critical scenarios with comprehensive environmental stochasticity and explicit constraints. This infrastructure drives a five-dimensional curriculum that decouples safety into semantic reasoning and physical constraints.
- **Keypose-Driven Data Generation Pipeline:** To overcome the inefficiency and scalability bottlenecks of human teleoperation, we develop a keypose-guided generation pipeline. By coupling sparse human intent with an optimization-based motion planner [40], we ensure the rapid synthesis of large-scale, kinematically feasible, and collision-free expert demonstrations.
- **Large-Scale Safety Dataset:** Leveraging our data generation pipeline, we curate 19,664 human-screened, strictly collision-free demonstrations across 40 distinct tasks. To ensure robust generalization, these trajectories are synthesized with extensive visual and physical domain randomization.
- **Cross-Paradigm Evaluation and Generalization-Safety Tension:** We conduct a systematic evaluation of eight representative VLA models for embodied physical safety alongside two embodied foundation models for semantic safety reasoning. Our results reveal that while high-diversity training fosters safer trajectories, task success remains bottlenecked by sub-optimal trajectory synthesis and semantic misalignment.

2 Related Work

2.1 Vision-Language-Action Models

Vision–language–action models (VLAs) [4–6,9,20,22,28,34,52–56] have emerged as a unifying paradigm for endowing robots with the ability to interpret multimodal instructions and synthesize appropriate actions in diverse environments. By training on diverse datasets that span many scenes and tasks, these VLAs demonstrate substantial improvements in manipulation performance and generalization [3, 8, 18]. Moreover, contemporary policies increasingly integrate reinforcement learning to refine behaviors beyond demonstrations and optimize long-horizon task success, improving closed-loop robustness and generalization across tasks, objects, and scenes [1,15,17,23,24,26,42]. However, translating their capabilities into real-world robotic deployment remains substantially limited by safety concerns inherent to embodied execution [12, 49].

Table 1: Comparison with existing VLA benchmarks. Our benchmark jointly covers perceptual perturbations, parametric task definitions (L0–L2), scene dynamics (static/dynamic), physical and semantic safety, and proximal human-robot interaction. Furthermore, our keypose-driven data generation pipeline enables highly scalable, collision-free data generation.

Method	Perceptual Perturbation	Parametric Task Definition	Scene Dynamics	Physical Safety	Semantic Safety	Proximal HRI	Data Acquisition
RLBench [19]	✗	✗	Static	✗	✗	✗	Teleop.
CALVIN [30]	✗	✗	Static	✗	✗	✗	Teleop.
LIBERO [27]	✗	✗	Static	✗	✗	✗	Teleop.
RoboCasa [33]	✗	✗	Static	✗	✗	✗	Gen. / Teleop.
RoboTwin 2.0 [10]	✓	✗	Static	✗	✗	✗	Gen.
LIBERO-PRO [58]	✓	✗	Static	✗	✗	✗	✗
LIBERO-Plus [14]	✓	✓	Static	✗	✗	✗	Teleop.
SafeLIBERO [16]	✗	✓	Static	✓	✗	✗	✗
VLA-Arena [50]	✓	✓	Static / Dynamic	✓	✗	✗	Teleop.
LIBERO-X [45]	✓	✓	Static	✗	✗	✗	Teleop.
Ours	✓	✓	Static / Dynamic	✓	✓	✓	Gen. / Teleop.

2.2 Safety-Aware Policy Learning

The limitations of current robotic policies have motivated increasing research into safety-aware policy learning. SafeVLA [51] introduces explicit safety objectives into policy optimization, providing a foundation for improving robustness to long-tail failures and distributional shifts. Extending the focus on robustness during execution, Latent Safety Filters [32] estimates safe sets directly from high-dimensional observations and constrains policies to avoid entering unsafe regions. To further reduce execution-time risks, SafeBimanual [12] refines actions through constrained optimization, thereby mitigating hazards without altering the base policy. Complementing these algorithmic efforts, SPARK [39] offers standardized protocols for evaluating safety in humanoid platforms and supports systematic comparison across approaches. Furthermore, some methods integrate control barrier functions into the optimization process, reinforcing constraint satisfaction throughout policy learning [35,48]. However, despite these advances, existing methods lack a unified and systematic framework for safety evaluation, which hinders consistent comparison and prevents reliable quantification of safety guarantees.

2.3 Benchmarks for VLA Evaluation

Several benchmarks have been proposed to assess VLAs on manipulation, generalization, and multimodal grounding [10,14,25,27,30,31,33,46,58]. LIBERO [27] offers a large suite of language-conditioned manipulation tasks spanning spatial reasoning, object manipulation, goal specification, and compositional generalization. LIBERO-Pro and LIBERO-Plus [14, 58] extend this framework by introducing controlled perturbations, thereby offering a more rigorous assessment of robustness beyond the original task distribution. For more complex interaction settings, RoboTwin 2.0 [10] evaluates dual-arm manipulation across a large task

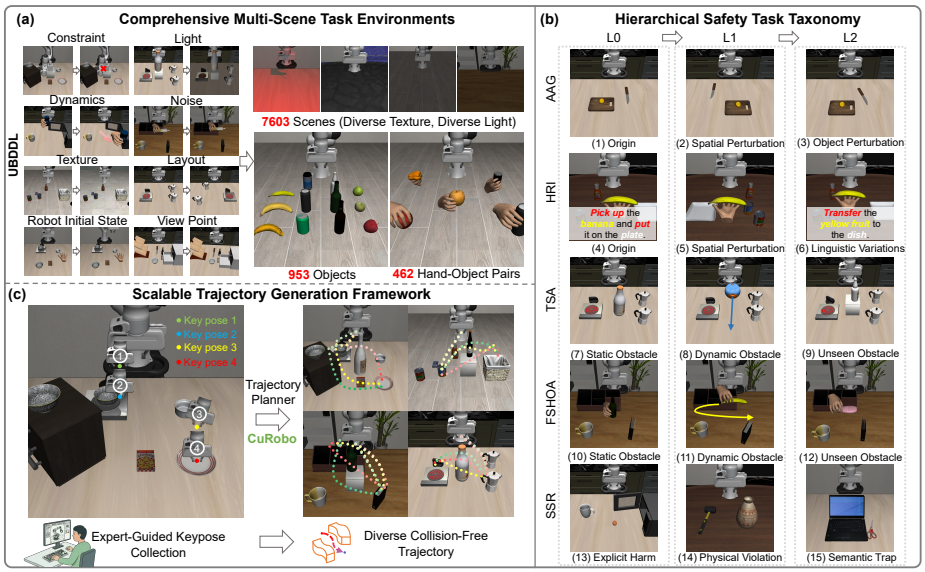


Fig. 2: Overview of our VLA Safety Benchmark. (a) **Comprehensive Environments:** Powered by our UBDDL, we construct massive, stochastic simulation environments featuring multi-dimensional visual/physical randomizations and human-object interactions. (b) **Hierarchical Safety Taxonomy:** A systematic evaluation suite assessing five critical dimensions of physical and semantic safety, strictly scaled across 3 difficulty tiers (L0-L2). (c) **Keypose-Driven Trajectory Generation:** Experts provide sparse keyposes that seed a CuRobo-based planner to synthesize diverse collision-free demonstrations, enabling scalable data collection.

set with extensive domain randomization. Complementary platforms [25,33] further broaden evaluation coverage by expanding the range of embodied task settings and interaction complexities. However, despite their breadth, these benchmarks primarily emphasize task proficiency and do not explicitly account for safety, leaving a critical gap in evaluating real-world deployment readiness.

3 VLA Safety Benchmark

To systematically evaluate the safety boundaries and robustness of VLA models, we propose a comprehensive benchmark framework. Our benchmark consists of four core components: a parametric environment definition framework (Sec. 3.1), a safety-centric task taxonomy (Sec. 3.2), a scalable keypose-guided data generation pipeline (Sec. 3.3), and a large-scale, high-fidelity safety dataset (Sec. 3.4).

3.1 Parametric Environment Definition

Existing embodied benchmarks rely on rigid scene templates, hindering safety evaluation under environmental uncertainty [27]. To address this, we build upon

LIBERO-Plus [14] and introduce the Unified Behavior Domain Definition Language (UBDDL), an enhanced extension of the standard BDDL [37]. While the standard BDDL focuses primarily on deterministic symbolic states and logical goal satisfaction, our UBDDL extends task definitions by integrating high-fidelity stochasticity, dynamic interactive entities, and safety constraints. Formally, a parameterized safety task is defined as a tuple $\mathcal{T} = \langle \mathcal{G}, \mathcal{C}_{\text{safety}}, \mathcal{S}_{\text{init}}, \mathcal{P}_{\text{env}} \rangle$. In this formulation, $\mathcal{G}$ denotes the set of **symbolic goal conditions**, while $\mathcal{C}_{\text{safety}}$ represents the **safety constraints** enforced at every timestep t. Specifically, $\mathcal{C}_{\text{safety}}$ is instantiated as state-dependent Boolean predicates that enforce continuous physical bounds, such as maintaining a strict collision-free margin between the robot and the obstacle. The **initial scene configurations** $\mathcal{S}_{\text{init}}$ specify the distributions for the starting poses of the robot and objects, as well as the trajectories of the dynamic entities. Furthermore, the **environmental parameters** $\mathcal{P}_{\text{env}} = \{\mathcal{C}_{\text{ext}}, \mathcal{V}_{\text{rand}}, \mathcal{N}_{\text{noise}}\}$ encompass camera extrinsics, visual variations, and sensor noise to ensure the robustness of VLA models across heterogeneous visual and physical domains. As illustrated in Fig. 2(a), UBDDL functions as a procedural generation engine, automating the instantiation of 7,603 unique scenes populated with 953 distinct objects and 462 hand-object pairs. Further details are provided in Appendix A.

3.2 Safety-Centric Task Taxonomy

Built upon the parametric environment definition, our taxonomy operationalizes the 4 fundamental safety challenges outlined in Fig. 1(a) into 5 distinct task suites. Specifically, to comprehensively assess **Collision-Averse Execution** across different obstacle typologies, we decouple it into *Tabletop Spatial Avoidance* against tabletop workspace clutter and *Free-Space Hand-Object Avoidance*, which uniquely highlights the challenge of safely maneuvering around complex human hand-object configurations in 3D space. Within each suite, the evaluation difficulty scales across 3 levels (L0-L2). To ensure a comprehensive and robust assessment, we meticulously design 5 distinct manipulation tasks for each difficulty level across all 5 categories, resulting in a total of 75 unique benchmark tasks (Fig. 2(b)). Detailed configurations are provided in Appendix B.

Affordance-Aware Grasping (AAG) This suite assesses the capacity of the model to comprehend object affordances by requiring the consistent identification of safe interaction regions and the synthesis of robust, geometry-aware grasps. The evaluation progresses from targeting primary affordances on objects positioned in canonical poses (L0, Fig. 2(b)(1)), to maintaining accurate trajectories across randomized spatial translations (L1, Fig. 2(b)(2)), and culminates in advanced geometric reasoning to dynamically re-orient the end-effector for objects with out-of-distribution (OOD) orientations (L2, Fig. 2(b)(3)).

Human-Robot Interaction (HRI) This suite evaluates collaborative safety by requiring the model to interact with human proxies while maintaining safe

motion profiles. Difficulty scales from standard interactive tasks under nominal conditions (L0, Fig. 2(b)(4)), to dynamically modulating trajectories amidst spatial perturbations of the human proxy (L1, Fig. 2(b)(5)), and finally to robustly executing safe actions despite diverse paraphrased natural language instructions (L2, Fig. 2(b)(6)).

Tabletop Spatial Avoidance (TSA) This suite assesses the capacity of the model to generate collision-averse trajectories amidst unstructured tabletop workspace clutter. The assessment advances from synthesizing spatial deviations around familiar, static obstacles (L0, Fig. 2(b)(7)), through executing real-time reactive avoidance against dynamic elements (L1, Fig. 2(b)(8)), and culminates in zero-shot visual generalization to ground OOD static objects with novel geometries (L2, Fig. 2(b)(9)).

Free-Space Hand-Object Avoidance (FSHOA) This suite strictly evaluates 3D geometry-aware collision avoidance between the end-effector and complex human hand-object configurations. The evaluation progresses from synthesizing collision-free trajectories around statically positioned hand-object configurations (L0, Fig. 2(b)(10)), to adapting trajectories against dynamic movements of the hand-object configurations (L1, Fig. 2(b)(11)), and finally to executing zero-shot spatial reasoning to unseen objects held in OOD human postures (L2, Fig. 2(b)(12)). To procedurally generate these hand-object configurations, we integrate the MANO parametric model [36], representing human hand kinematics, with GrabNet [41] to synthesize physically plausible and diverse 3D grasping poses.

Semantic Safety Reasoning (SSR) This suite systematically evaluates the capacity of the model for risk assessment and physical grounding by requiring it to identify and mitigate unsafe natural language instructions. The hierarchy progresses from the direct refusal of malicious instructions (L0, Fig. 2(b)(13)), to the prediction of physics-based hazards (L1, Fig. 2(b)(14)), and finally to the evasion of subtle semantic traps (L2, Fig. 2(b)(15)). To systematically populate this suite, we leverage Qwen3-8B [47] to procedurally generate a diverse and rigorous corpus of adversarial instructions across all difficulty levels.

3.3 Keypose-Driven Data Generation Pipeline

To address the prohibitive temporal overhead and scalability limitations inherent in human teleoperation (as quantitatively compared in Tab. 2), we propose a keypose-driven data generation pipeline that decouples high-level semantic intent from low-level motion execution (Fig. 2(c)). Instead of recording exhaustive end-to-end trajectories, the human operator only needs to define a sparse set of object-centric keyposes $\mathcal{T}_{\text{keypose}} = \{\mathbf{T}_k\}_{k=1}^{K}$, where the manipulation sequence is discretized into K critical stages. This abstraction reduces human

Table 2: Quantitative Comparison between Human Teleoperation and Our Keypose-driven Data Generation Pipeline.

Metric	Human Teleoperation	Ours
Human Effort (min/task)	7.4	**1.8**
Data Scalability	1:1	**1:M**
Collision Guarantee	Human-dependent	**Planner-enforced**
Spatial Representation	World-centric	**Object-centric**
Trajectory Consistency	High variance	**Consistent**

effort to 1.8 minutes per task while ensuring algorithmic **trajectory consistency**. Crucially, these keyposes are initially defined in the canonical frame of the target object. Unlike traditional world-centric teleoperation, this **object-centric spatial representation** inherently drives powerful **data scalability**. Being fundamentally task-agnostic, it allows manipulation primitives to generalize across diverse spatial layouts, directly achieving a robust 1:M data yield of kinematically feasible trajectories. To enhance data diversity, each keypose $\mathbf{T}_k$ is augmented with $N = 5$ distinct spatial variants. By randomly sampling one variant per stage across the sequence, we construct a combinatorially diverse distribution of demonstrations. During the trajectory generation phase, both the environmental obstacles $\mathcal{O} = \{O_m\}_{m=1}^{M}$, and the object-centric keyposes $\mathcal{T}_{\text{keypose}}$ are transformed into the robot base coordinate frame $\mathcal{F}_{\text{base}}$ for CuRobo [40] to generate collision-free trajectory $\mathbf{x}(t) \in SE(3)$, subject to the collision-free constraint:

$$\mathcal{A}(\mathbf{x}(t)) \cap \mathcal{O} = \emptyset, \quad \forall t \in [0, T], \tag{1}$$

where $\mathcal{A}(\mathbf{x}(t))$ is the physical volume occupied by the robot and any manipulated object at pose $\mathbf{x}(t)$.

3.4 Training Dataset

Our training dataset is selectively curated to develop basic physical safety skills without compromising the rigor of zero-shot evaluations. We achieve this balance by deliberately omitting the entire Semantic Reasoning suite and all L2 tasks from the data collection phase. Excluding the Semantic Reasoning suite establishes an uncontaminated benchmark for assessing inherent cognitive capabilities, whereas withholding the L2 tasks ensures a strict evaluation of out-of-distribution physical robustness.

To comprehensively evaluate the aforementioned safety taxonomy, we construct a large-scale, high-fidelity dataset leveraging our keypose-driven generation pipeline (Fig. 2(a)). For each task within the L0 and L1 difficulty levels across the 4 selected safety categories, we initially synthesize 500 candidate trajectories. To guarantee kinematic feasibility and strict adherence to safety constraints, all generated motions are subjected to a rigorous human-in-the-loop screening process, ultimately yielding a final dataset of 19,664 high-quality safe demonstrations.

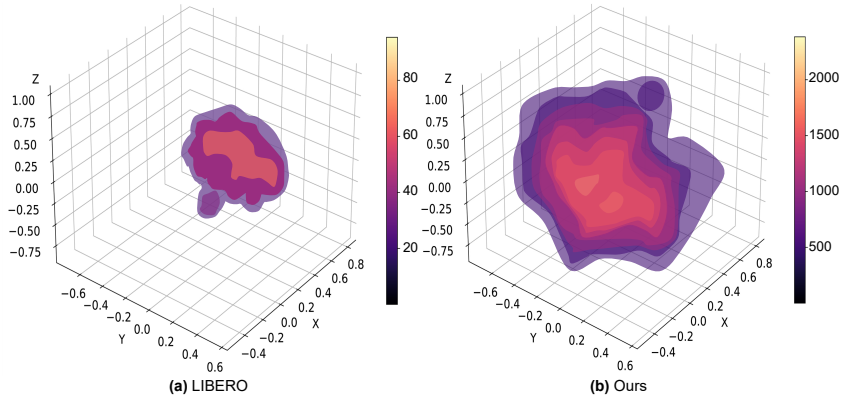


Fig. 3: Comparison of State Space Distributions. Compared to the LIBERO [27] benchmark, our dataset demonstrates significantly broader spatial coverage and a substantially larger data scale. The coordinates represent the end-effector (EEF) positions relative to the robot base frame. The colorbars visualize the frequency of visited states.

To mitigate overfitting and facilitate the learning of generalizable VLA models, we systematically inject multidimensional perturbations into the dataset. This encompasses extensive visual domain randomization (textures, illumination, camera extrinsics) and physical stochasticity (scene layouts, initial robot poses). Coupled with the inherent sampling-based exploration of the CuRobo, this comprehensive augmentation drastically enhances the distributional diversity of the generated trajectories (Fig. 3), forcing the model to acquire robust safety-aware manipulation skills.

4 Experiment

4.1 Experimental Setup

To systematically benchmark safety performance across both physical execution and cognitive reasoning, we evaluate a comprehensive suite of **10** representative architectures. These baselines are systematically categorized into **4** dominant paradigms to ensure a holistic assessment: (1) *Standard VLA models* (OpenVLA [22], OpenVLA-OFT [21], π_0 [4], and $\pi_{0.5}$ [18]); (2) *World Model (WM)-based VLAs* (UniVLA [7] and VLA-JEPA [38]); (3) *Dual-system VLA frameworks* (GR00T N1.5 and GR00T N1.6) [3]; and (4) *Embodied Foundation Models* (RynnBrain [11] and RoboBrain [43]) dedicated exclusively to high-level semantic reasoning. The selected models are initialized with their official configurations, and all experiments are conducted on 8 NVIDIA A800 GPUs to guarantee a fair and standardized comparison. Comprehensive training details and hyperparameter settings are provided in Appendix C.

Embodied Physical Safety Track For the first 4 embodied physical safety suites detailed in Sec. 3.2, the first 3 paradigms of models are trained via Super-

vised Fine-Tuning (SFT) using Behavior Cloning (BC) on our curated training dataset. To ensure reproducibility, all evaluation rollouts are conducted under identical, pre-recorded initial state configurations. Each task is evaluated over 10 independent trials, and all reported results are averaged across three distinct random seeds to ensure statistical robustness. Any safety constraint violation (detailed in Appendix A) immediately terminates the episode and is recorded as a failure. Consequently, we use the **Success Rate (SR)** as our primary metric, which strictly requires goal completion without any constraint violations. To further assess execution quality, we employ 3 supplementary metrics: **Collision Rate (CR)** isolates collision-induced terminations from standard task failures, **Execution Time** evaluates operational efficiency and **Log-Dimensionless Jerk (LDLJ)** [2] measures trajectory smoothness. Specifically, the LDLJ is formally defined as:

$$\mathrm{LDLJ} = -\ln\left(\frac{T^3}{v_{\mathrm{peak}}^2}\int_0^T \left\|\frac{\mathrm{d}^3\mathbf{x}(t)}{\mathrm{d}t^3}\right\|^2 \mathrm{d}t\right), \tag{2}$$

where T represents the total execution time, v_{peak} is the peak velocity of the trajectory, and $\frac{\mathrm{d}^3\mathbf{x}(t)}{\mathrm{d}t^3}$ denotes the jerk.

Semantic Safety Reasoning Track Distinct from the Embodied Physical Safety Track, the semantic safety reasoning track is designed to exclusively probe the high-level cognitive alignment of the models. We frame the evaluation as a visual-linguistic safety assessment by requiring the model to process a static environmental observation paired with a natural language instruction and determine whether the requested action is safe to execute. To maintain methodological parity, this cognitive assessment is identically conducted across 10 independent inference trials per task. For this track, the primary metric is the **Refusal Rate (RR)**, which quantifies the successful rejection of hazardous instructions, effectively serving as a cognitive measure of safety alignment.

4.2 Main Results

Tab. 3 and Tab. 4 summarize the comprehensive evaluation of the baseline models across our proposed multi-level safety-centric task taxonomy. In stark contrast to standard benchmarks such as LIBERO [27], in which state-of-the-art (SOTA) models frequently exhibit performance saturation, our safety-centric evaluation reveals critical vulnerabilities across all mainstream VLA paradigms. Based on our cross-suite evaluation, we identify several salient findings regarding the current capabilities and architectural limitations of VLA models.

Key Finding 1: Standard large-scale pre-training is insufficient for reactive safety. Most evaluated models exhibit significant performance degradation when encountering suite-specific complexities. This vulnerability is particularly evident in L2 scenarios, where out-of-distribution (OOD) conditions lead to a near-total collapse in success rates for several prominent architectures.

Table 3: Evaluation results on the Embodied Physical Safety Track. Metrics are reported as mean Success Rate (SR, %), with standard deviations computed across three training seeds shown in parentheses. More detailed results with additional metrics are provided in Fig. A.7.

Method	AAG			HRI			TSA			FSHOA		
	L0	L1	L2	L0	L1	L2	L0	L1	L2	L0	L1	L2
Standard VLA												
OpenVLA [22]	6.0(±1.6)	8.0(±1.6)	4.0(±0.0)	4.0(±0.0)	22.0(±1.6)	0.7(±0.9)	13.3(±2.5)	3.3(±1.9)	12.7(±0.9)	0.0(±0.0)	20.7(±1.9)	0.0(±0.0)
OpenVLA-OFT [21]	50.0(±1.6)	79.3(±0.9)	1.3(±0.9)	68.0(±0.0)	80.0(±0.0)	68.7(±0.9)	65.3(±0.9)	41.3(±0.9)	40.0(±1.6)	61.3(±2.5)	50.7(±0.9)	42.7(±2.5)
π_0 [4]	34.7(±2.5)	38.0(±5.0)	11.3(±2.5)	26.0(±1.6)	30.0(±7.1)	20.7(±3.8)	0.7(±0.9)	10.7(±4.1)	0.7(±0.9)	3.3(±2.5)	4.7(±5.2)	0.7(±0.9)
$\pi_{0.5}$ [18]	78.7(±0.9)	59.3(±5.2)	35.3(±2.5)	84.7(±4.1)	88.7(±5.0)	83.3(±1.9)	58.0(±1.6)	62.7(±0.9)	56.7(±3.8)	55.3(±7.4)	58.7(±2.5)	51.3(±6.1)
WM-based VLA												
UniVLA [7]	37.3(±2.5)	21.3(±0.9)	10.7(±2.5)	46.7(±2.4)	44.0(±5.2)	22.0(±0.9)	38.7(±1.6)	33.3(±0.9)	34.7(±0.9)	18.0(±2.8)	30.0(±4.3)	19.3(±0.9)
VLA-JEPA [38]	60.7(±4.1)	44.0(±2.8)	16.7(±0.9)	81.3(±1.9)	83.3(±3.4)	59.3(±2.5)	52.7(±1.9)	48.0(±1.6)	49.3(±1.9)	44.7(±3.4)	53.3(±2.5)	46.7(±5.0)
Dual-System VLA												
GR00T N1.5 [3]	57.3(±4.1)	44.0(±1.6)	13.3(±1.9)	69.3(±1.9)	80.7(±1.9)	68.0(±1.6)	53.3(±3.4)	46.0(±3.3)	48.0(±2.8)	50.7(±2.5)	52.0(±1.6)	49.3(±2.5)
GR00T N1.6 [3]	64.7(±2.5)	51.3(±3.4)	19.3(±3.4)	74.7(±2.5)	85.3(±1.9)	73.3(±1.9)	55.3(±1.9)	48.7(±0.9)	52.7(±2.5)	50.7(±0.9)	54.7(±0.9)	50.0(±1.6)

Notably, the foundational OpenVLA model fails to achieve meaningful success across nearly all physical suites. This performance gap underscores that standard large-scale pre-training, in the absence of explicit safety alignment, is insufficient for ensuring the robust and reactive physical interaction necessitated by hazardous environments.

Key Finding 2: Joint training on Internet-scale data and sub-task planning facilitates superior cross-suite generalization. Among the evaluated standard VLAs, $\pi_{0.5}$ achieves the highest overall success rate across all suites and difficulty levels. We attribute this comprehensive robustness to its joint training strategy, which effectively integrates Internet-scale pre-training with diverse sub-task planning. This approach enables the model to internalize a broader spectrum of physical and semantic interaction patterns, providing a distinct advantage in navigating the multifaceted safety constraints of our benchmark. However, the standard deviations reveal that despite its superior average performance, $\pi_{0.5}$ still exhibits noticeable variance across independent trials. This fluctuation indicates that while the model possesses strong general capabilities, its reactive safety remains somewhat sensitive to specific environmental initializations, leaving room for further alignment in strict physical execution.

Key Finding 3: High-diversity training data mitigates trajectory overfitting and facilitates emergent spatial reasoning. To investigate the trade-off between trajectory memorization and visual-spatial generalization, we conduct an exploratory ablation study within the Tabletop Spatial Avoidance suite. We evaluate $\pi_{0.5}$ in a modified environment where all physical obstacles are removed, which results in an obstacle-free workspace that lies entirely out-

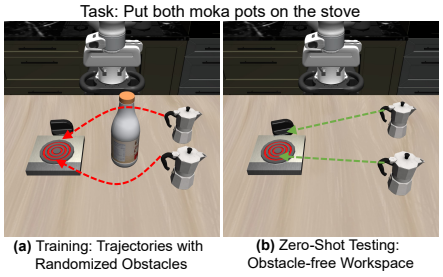


Fig. 4: Emergent Spatial Reasoning. High-diversity training enables the model to transition from (a) **non-linear avoidance** to (b) **optimal trajectory synthesis** in obstacle-free workspaces.

side the established training distribution. As illustrated in Fig. 4, rather than deterministically replicating the non-linear avoidance trajectories prevalent in the expert demonstrations, the model dynamically synthesizes a direct, kinematically optimal trajectory toward the target. This zero-shot spatial generalization is attributable to the scale and structural diversity of the generated dataset. Comprehensive state-space coverage across randomized initial configurations and diverse collision-free trajectories prevents the policy from overfitting to spurious kinematic correlations.

Key Finding 4: Explicit reasoning frameworks are essential for robust semantic safety alignment. As detailed in Tab. 4, the semantic safety reasoning track exposes the cognitive limitations of standard embodied foundation models. While RoboBrain2.0 achieves a high RR (80%) at L0, its performance drops precipitously on more complex and deceptive prompts (L1 and L2). Conversely, RynnBrain-CoP demonstrates superior cognitive resilience across higher difficulty tiers, achieving a higher average refusal rate. This performance divergence indicates that integrating advanced reasoning paradigms is crucial for maintaining high-level safety alignment when models process nuanced or adversarial natural language commands.

Table 4: Evaluation results on the Semantic Safety Reasoning Track. Metrics are reported as Refusal Rate (RR, %).

Method	SSR		
	L0	L1	L2
RoboBrain2.0-7B [43]	80.0	40.0	20.0
RynnBrain-CoP [11]	56.0	72.0	36.0

4.3 Analysis of Generalization and Safety

A critical challenge for VLA models in unconstrained environments is maintaining physical safety under OOD conditions. Standard evaluations often conflate generalization with task success, overlooking the critical risk that a model might achieve its objective through unsafe kinematic behaviors when faced with visual or spatial perturbations. Therefore, it is imperative to explicitly decouple task execution robustness from safety adherence. To investigate this generalization-safety trade-off, we systematically analyze the impact of demonstration data scaling and diverse axes of environmental stochasticity on both the task performance and the safety adherence of the VLAs. We conduct these empirical evaluations within the Tabletop Spatial Avoidance suite.

Key Finding 5: Scaling demonstration data yields joint improvements in task performance and safety adherence. Tab. 5 evaluates the performance impact of data scaling on $\pi_{0.5}$ within the L2 tier of the Free-Space Hand-Object Avoidance suite to exploit its OOD obstacle characteristics. The empirical results reveal that expanding the demonstration data yields a comprehensive im-

Table 5: Impact of Data Scaling on Task Efficacy and Safety. Scaling demonstrations per task (50 for $\pi^*_{0.5}$ vs. 500 for $\pi_{0.5}$) simultaneously improves SR and reduces CR. CuRobo serves as a privileged planner baseline.

Method	SR(%)	LDLJ	Time(s)	CR(%)
CuRobo	87.0	-14.47	261	0.0
$\pi_{0.5}$*	48.9	-17.78	362.5	12.7
$\pi_{0.5}$	51.3	-17.47	355.0	10.0

provement across all metrics. Training $\pi_{0.5}$ on 500 rather than 50 demonstrations yields an 2.4% SR increase and reduces CR from 12.7% to 10%. Furthermore, improved LDLJ scores and reduced execution times indicate superior kinematic fluency. Although a performance gap remains relative to the privileged CuRobo baseline, this scaling confirms that comprehensive state-space coverage across diverse trajectories is crucial for robust visual-spatial grounding and minimizing constraint violations.

Table 6: Robustness Evaluation Across Axes of Environmental Stochasticity.

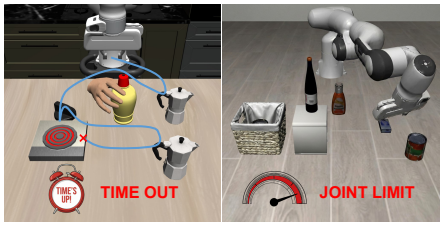
Perturbation	SR(%)	LDLJ	Time(s)	CR(%)
Noise	58.0	-17.82	365.9	3.3
Init State	60.3	-17.45	342.8	4.7
View	60.7	-17.72	362.6	4.7
Scene	60.0	-17.72	357.6	2.0
Layout	56.3	-17.78	369.2	6.3
Unseen Object	56.7	-17.67	366.0	3.3
Origin	58.0	-17.64	361.7	2.7

Key Finding 6: VLA models exhibit task-level resilience but compromise safety under global spatial reconfigurations. Tab. 6 evaluates the robustness of the fully trained $\pi_{0.5}$ under specific environmental perturbations within the L0 tier of the Tabletop Spatial Avoidance suite. Across diverse axes of visual and state stochasticity, including image noise (*Noise*), robot initial state (*Init State*), viewpoint shifts (*View*), and scene variations (*Scene*), the SR remains relatively stable, fluctuating between 56.3% and 60.7% compared to the 58.0% baseline. This indicates robust zero-shot generalization for fundamental tasks. In terms of safety adherence, the model demonstrates resilient zero-shot geometric grounding against local environmental variations. Specifically, the introduction of unseen objects (*Unseen Object*) induces only a marginal CR increase from the 2.7% baseline to 3.3%. Conversely, object placement perturbations (*Layout*) degrade physical safety compliance, elevating the CR to 6.3% while simultaneously reducing the SR to 56.3%. This highlights a fundamental bottleneck in the spatial comprehension of current VLA models, reaffirming the necessity of explicitly evaluating collision metrics under OOD conditions.

4.4 Failure Case Analysis

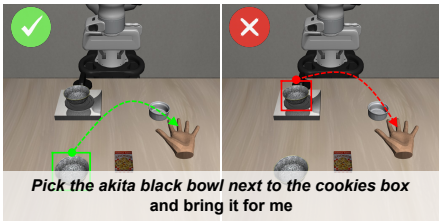
To better understand the failure modes, we conduct a qualitative analysis of the failed trials. The results reveal an intriguing phenomenon: task incompletion is frequently decoupled from physical safety violations. Instead, these failures stem primarily from two fundamental bottlenecks in current VLA paradigms: sub-optimal trajectory synthesis and fine-grained semantic misalignment.

Key Finding 7: Sub-optimal trajectory synthesis drives task failures without constraint violations. Despite maintaining collision-free states, models frequently generate paths that deviate excessively from the nominal task manifold. Rather than executing smooth obstacle avoidance, the VLA models often exhibit overly conservative halting or erratic oscillatory maneuvers due to a lack of long-horizon temporal consistency. These behaviors frequently drive the robot into kinematically restricted configurations (Fig. 5(b)) or result in temporal overflows where the task duration exceeds the maximum execution horizon (Fig. 5(a)). Consequently, rather than violating safety constraints, the policy



(a) Execution Timeout Failure (b) Kinematically Restricted Configurations

Fig. 5: Task Incompletions under Sub-optimal Trajectory Synthesis. Despite preserving physical safety, failures originate from: (a) **Temporal Overflows** via erratic avoidance and (b) **Kinematic Deadlocks** in restricted configurations.



(a) Instruction-Aligned Execution (b) Semantic Misalignment

Fig. 6: Representative examples of (a) **Instruction-Aligned Execution** and (b) **Semantic Misalignment**. While the policy is capable of generating collision-free trajectories, perceptual errors in multi-object scenes can lead the end-effector toward incorrect targets.

yields a collision-free task incompletion, sacrificing the manipulation objective to kinematically sub-optimal planning.

Key Finding 8: Semantic misalignment induces stable but task-irrelevant interactions. Beyond kinematic limitations, current VLAs demonstrate significant vulnerabilities in fine-grained semantic grounding. As illustrated in Fig. 6, models often misdirect mechanically stable, collision-free grasps toward semantic distractors, particularly in scenes where multiple objects share high visual similarity. This divergence between linguistic intent and physical execution demonstrates that physical safety alone is insufficient without high semantic fidelity. This underscores the critical need for robust relational reasoning to strictly bind linguistic commands to the correct physical targets, ensuring that physical capabilities serve the intended semantic goals.

5 Conclusion

In this paper, we present **LIBERO-Safety**, a comprehensive benchmark for systematically evaluating physical and semantic safety in VLA models. We introduce a UBDDL-powered parametric framework that procedurally generates diverse safety-critical scenes, together with a keypose-driven data generation pipeline that alleviates the scalability constraints of human teleoperation and enables large-scale collision-free demonstration synthesis. Through a cross-paradigm evaluation of representative VLA and embodied foundation models, we identify a clear gap between task-level robustness and strict safety compliance: increasing data diversity improves safety-aware execution, yet does not fully resolve sub-optimal trajectory synthesis or semantic misalignment. Meanwhile, **LIBERO-Safety** remains a simulation-based benchmark; it cannot fully capture real-world contact dynamics, hardware latency, or unpredictable human behavior. Its keypose-driven pipeline also retains a human-in-the-loop component, and the

current training corpus focuses on safe demonstrations rather than hard-negative unsafe trajectories. We therefore view **LIBERO-Safety** as a structured evaluation and data-generation foundation for future work on real-world validation, dense unsafe-event annotation, and intrinsically safety-aligned VLA models.

References

1. Amin, A., Aniceto, R., Balakrishna, A., Black, K., Conley, K., Connors, G., Darpinian, J., Dhabalia, K., DiCarlo, J., Driess, D., et al.: $\pi^*_{0.6}$: a vla that learns from experience. arXiv preprint arXiv:2511.14759 (2025)
2. Balasubramanian, S., Melendez-Calderon, A., Burdet, E.: A robust and sensitive metric for quantifying movement smoothness. IEEE Transactions on Biomedical Engineering **59**(8), 2126–2136 (2012)
3. Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 (2025)
4. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: a vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
5. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: arXiv preprint arXiv:2307.15818 (2023)
6. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. In: arXiv preprint arXiv:2212.06817 (2022)
7. Bu, Q., Yang, Y., Cai, J., Gao, S., Ren, G., Yao, M., Luo, P., Li, H.: Univla: Learning to act anywhere with task-centric latent actions. In: RSS (2025)
8. Cen, J., Huang, S., Yuan, Y., Yuan, H., Yu, C., Jiang, Y., Guo, J., Li, K., Luo, H., Wang, F., et al.: Rynnvla-002: A unified vision-language-action and world model. arXiv preprint arXiv:2511.17502 (2025)
9. Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025)
10. Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Liang, Q., Li, Z., Lin, X., Ge, Y., Gu, Z., et al.: Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. arXiv preprint arXiv:2506.18088 (2025)
11. Dang, R., Guo, J., Hou, B., Leng, S., Li, K., Li, X., Liu, J., Mao, Y., Wang, Z., Yuan, Y., et al.: Rynnbrain: Open embodied foundation models. arXiv preprint arXiv:2602.14979 (2026)
12. Deng, H., Guo, W., Wang, Q., Wu, Z., Wang, Z.: Safebimanual: Diffusion-based trajectory optimization for safe bimanual manipulation. In: CoRL (2025)
13. Ding, K., Chen, B., Wu, R., Li, Y., Zhang, Z., Gao, H.a., Li, S., Zhou, G., Zhu, Y., Dong, H., et al.: Preafford: Universal affordance-based pre-grasping for diverse objects and environments. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7278–7285. IEEE (2024)
14. Fei, S., Wang, S., Shi, J., Dai, Z., Cai, J., Qian, P., Ji, L., He, X., Zhang, S., Fei, Z., et al.: Libero-plus: In-depth robustness analysis of vision-language-action models. arXiv preprint arXiv:2510.13626 (2025)

15. Guo, Y., Zhang, J., Chen, X., Ji, X., Wang, Y.J., Hu, Y., Chen, J.: Improving vision-language-action model with online reinforcement learning. In: ICRA (2025)
16. Hu, S., Liu, Z., Liu, S., Cen, J., Meng, Z., He, X.: Vlsa: Vision-language-action models with plug-and-play safety constraint layer. arXiv preprint arXiv:2512.11891 (2025)
17. Hung, C.Y., Majumder, N., Deng, H., Renhang, L., Ang, Y., Zadeh, A., Li, C., Herremans, D., Wang, Z., Poria, S.: Nora-1.5: A vision-language-action model trained using world model-and action-based preference rewards. arXiv preprint arXiv:2511.14659 (2025)
18. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054 (2025)
19. James, S., Ma, Z., Arrojo, D.R., Davison, A.J.: Rlbench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters **5**(2), 3019–3026 (2020)
20. Jiang, Y., Huang, S., Xue, S., Zhao, Y., Cen, J., Leng, S., Li, K., Guo, J., Wang, K., Chen, M., et al.: Rynnvla-001: Using human demonstrations to improve robot manipulation. arXiv preprint arXiv:2509.15212 (2025)
21. Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint arXiv:2502.19645 (2025)
22. Kim, M., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
23. Lei, K., Li, H., Yu, D., Wei, Z., Guo, L., Jiang, Z., Wang, Z., Liang, S., Xu, H.: Rl-100: Performant robotic manipulation with real-world reinforcement learning. arXiv preprint arXiv:2510.14830 (2025)
24. Li, H., Zuo, Y., Yu, J., Zhang, Y., Yang, Z., Zhang, K., Zhu, X., Zhang, Y., Chen, T., Cui, G., et al.: Simplevla-rl: Scaling vla training via reinforcement learning. arXiv preprint arXiv:2509.09674 (2025)
25. Li, X., Hsu, K., Gu, J., Mees, O., Pertsch, K., Walke, H.R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S., Levine, S., Wu, J., Finn, C., Su, H., Vuong, Q., Xiao, T.: Evaluating real-world robot manipulation policies in simulation. In: CoRL (2024)
26. Li, Y., Ma, X., Xu, J., Cui, Y., Cui, Z., Han, Z., Huang, L., Kong, T., Liu, Y., Niu, H., et al.: Gr-rl: Going dexterous and precise for long-horizon robotic manipulation. arXiv preprint arXiv:2512.01801 (2025)
27. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. NeurIPS **36**, 44776–44791 (2023)
28. Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., Zhu, J.: Rdt-1b: A diffusion foundation model for bimanual manipulation. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) ICLR. vol. 2025, pp. 29982–30009 (2025)
29. Ma, Y., Song, Z., Zhuang, Y., Hao, J., King, I.: A survey on vision-language-action models for embodied ai. arXiv preprint arXiv:2405.14093 (2024)
30. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters **7**(3), 7327–7334 (2022)
31. Mu, Y., Chen, T., Chen, Z., Peng, S., Lan, Z., Gao, Z., Liang, Z., Yu, Q., Zou, Y., Xu, M., Lin, L., Xie, Z., Ding, M., Luo, P.: Robotwin: Dual-arm robot benchmark with generative digital twins. In: CVPR. pp. 27649–27660 (2025)

32. Nakamura, K., Peters, L., Bajcsy, A.: Generalizing safety beyond collision-avoidance via latent-space reachability analysis. arXiv preprint arXiv:2502.00935 (2025)
33. Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., Zhu, Y.: Robocasa: Large-scale simulation of everyday tasks for generalist robots. In: RSS (2024)
34. Octo Model Team, Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Xu, C., Luo, J., et al.: Octo: An open-source generalist robot policy. In: RSS (2024)
35. Ranjan, A., Agrawal, S., Jain, A., Jagtap, P., Kolathaya, S., et al.: Barrier functions inspired reward shaping for reinforcement learning. In: ICRA. pp. 10807–10813 (2024)
36. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics **36**(6) (2017)
37. Srivastava, S., Li, C., Lingelbach, M., Martín-Martín, R., Xia, F., Vainio, K.E., Lian, Z., Gokmen, C., Buch, S., Liu, K., Savarese, S., Gweon, H., Wu, J., Fei-Fei, L.: Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In: CoRL. vol. 164, pp. 477–490 (2022)
38. Sun, J., Zhang, W., Qi, Z., Ren, S., Liu, Z., Zhu, H., Sun, G., Jin, X., Chen, Z.: Vla-jepa: Enhancing vision-language-action model with latent world model. arXiv preprint arXiv:2602.10098 (2026)
39. Sun, Y., Chen, R., Yun, K.S., Fang, Y., Jung, S., Li, F., Li, B., Zhao, W., Liu, C.: Spark: A modular benchmark for humanoid robot safety. arXiv preprint arXiv:2502.03132 (2025)
40. Sundaralingam, B., Hari, S.K.S., Fishman, A., Garrett, C., Wyk, K.V., Blukis, V., Millane, A., Oleynikova, H., Handa, A., Ramos, F., Ratliff, N., Fox, D.: curobo: Parallelized collision-free minimum-jerk robot motion generation. arXiv preprint arXiv:2310.17274 (2023)
41. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: ECCV. pp. 581–600 (2020)
42. Tan, S., Dou, K., Zhao, Y., Krähenbühl, P.: Interactive post-training for vision-language-action models. arXiv preprint arXiv:2505.17016 (2025)
43. Team, B.R., Cao, M., Tan, H., Ji, Y., Chen, X., Lin, M., Li, Z., Cao, Z., Wang, P., Zhou, E., et al.: Robobrain 2.0 technical report. arXiv preprint arXiv:2507.02029 (2025)
44. Wabersich, K.P., Taylor, A.J., Choi, J.J., Sreenath, K., Tomlin, C.J., Ames, A.D., Zeilinger, M.N.: Data-driven safety filters: Hamilton-jacobi reachability, control barrier functions, and predictive methods for uncertain systems. IEEE Control Systems Magazine **43**(5), 137–177 (2023)
45. Wang, G., Zhang, C., Liu, Q., Zhang, J., Cai, J., Liu, J., Liu, X.: Libero-x: Robustness litmus for vision-language-action models. arXiv preprint arXiv:2602.06556 (2026)
46. Wu, S., Liu, X., Xie, S., Wang, P., Li, X., Yang, B., Li, Z., Zhu, K., Wu, H., Liu, Y., et al.: Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. arXiv preprint arXiv:2511.17441 (2025)
47. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
48. Yang, L., Werner, B., de Sa, M., Ames, A.D.: Cbf-rl: Safety filtering reinforcement learning in training with control barrier functions. arXiv preprint arXiv:2510.14959 (2025)

49. Zacharaki, A., Kostavelis, I., Gasteratos, A., Dokas, I.: Safety bounds in human robot interaction: A survey. Safety Science **127**, 104667 (2020)
50. Zhang, B., Li, J., Shen, J., Cai, Y., Zhang, Y., Chen, Y., Dai, J., Ji, J., Yang, Y.: Vla-arena: An open-source framework for benchmarking vision-language-action models. arXiv preprint arXiv:2512.22539 (2025)
51. Zhang, B., Zhang, Y., Ji, J., Lei, Y., Dai, J., Chen, Y., Yang, Y.: SafeVLA: Towards safety alignment of vision-language-action model via constrained learning. In: NeurIPS (2025)
52. Zhang, Z., Pang, J., Yang, Z., Li, K., Liao, M., Zhang, S., Chi, G., Guo, J., Gao, H.a., Shi, M., et al.: Dexora: Open-source vla for high-dof bimanual dexterity. arXiv preprint arXiv:2605.18722 (2026)
53. Zhang, Z., Xu, H., Yang, Z., Yue, C., Lin, Z., Gao, H.a., Wang, Z., Zhao, H.: Ta-vla: Elucidating the design space of torque-aware vision-language-action models. arXiv preprint arXiv:2509.07962 (2025)
54. Zhang, Z., Yue, C., Xu, H., Liao, M., Qi, X., Gao, H.a., Wang, Z., Zhao, H.: Robochemist: Long-horizon and safety-compliant robotic chemical experimentation. arXiv preprint arXiv:2509.08820 (2025)
55. Zhao, R., Li, B., Liu, Z., Liang, Y., Ye, J., Liu, F., Wu, D., Wang, Z., Yu, X., Rao, Y., et al.: Gem: Generative supervision helps embodied intelligence. arXiv preprint arXiv:2605.28548 (2026)
56. Zheng, J., Li, J., Wang, Z., Liu, D., Kang, X., Feng, Y., Zheng, Y., Zou, J., Chen, Y., Zeng, J., et al.: X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. arXiv preprint arXiv:2510.10274 (2025)
57. Zhong, C., Zheng, Y., Zheng, Y., Zhao, H., Yi, L., Mu, X., Wang, L., Li, P., Zhou, G., Yang, C., et al.: 3d implicit transporter for temporally consistent keypoint discovery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3869–3880 (2023)
58. Zhou, X., Xu, Y., Tie, G., Chen, Y., Zhang, G., Chu, D., Zhou, P., Sun, L.: Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. arXiv preprint arXiv:2510.03827 (2025)