VLA-Arena: Benchmarking Vision-Language-Action Models via Structured Task Design

Anonymous CVPR submission

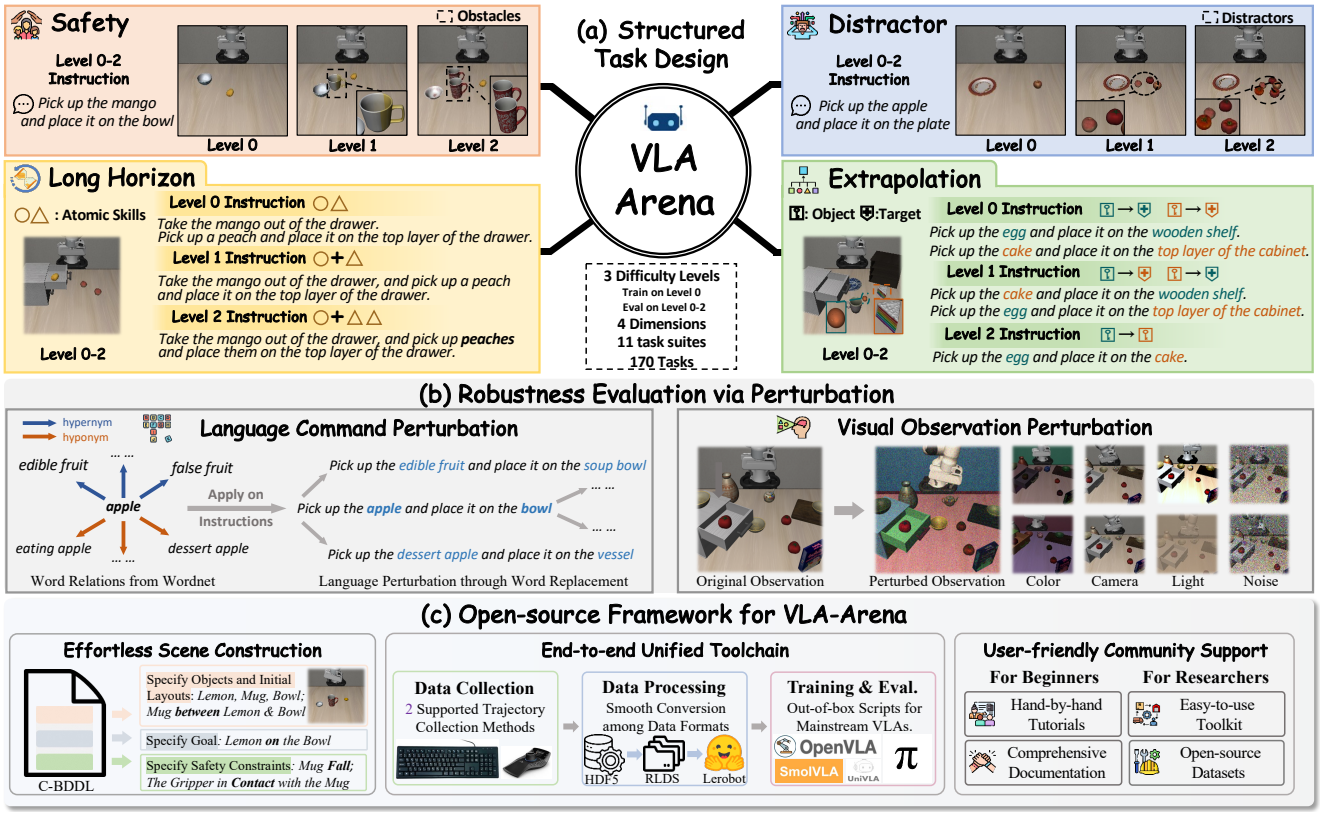

Figure 1. **Overview of the VLA-Arena Benchmark and Framework.** The VLA-Arena is an open-source framework for comprehensive evaluation of VLA models. **(a) Structured Task Design:** Span four key dimensions: Safety, Distractor, Extrapolation, and Long Horizon, covering 11 task suites with three difficulty levels (L0–L2), totaling 170 tasks. **(b) Robustness Evaluation via Perturbation:** Test robustness using systematic perturbations in both modalities: language command perturbations via semantically informed WordNet-based replacements, and visual observation perturbations through controlled environmental variations. **(c) Open-source Framework for VLA-Arena:** Build scenes declaratively and use the unified toolchain for data collection, processing, training, and evaluation of VLA models, supported by tutorials, advanced tools, rich documentation, and open-source datasets.

Abstract

*While Vision-Language-Action models (VLAs) are rapidly advancing towards generalist robot policies, it remains difficult to quantitatively understand their limits and failure modes. To address this, we introduce a comprehensive benchmark called **VLA-Arena**. We propose a novel structured task design framework to quantify difficulty across three orthogonal axes: **(1) Task Structure**, **(2) Language Command**, and **(3) Visual Observation**. This allows us to systematically design tasks with fine-grained difficulty levels, enabling a precise measurement of model capability*

frontiers. For Task Structure, VLA-Arena's 170 tasks are grouped into four dimensions: ***Safety***, ***Distractor***, ***Extrapolation***, *and* ***Long Horizon***. *Each task is designed with three difficulty levels (L0-L2), with fine-tuning performed exclusively on L0 to test for extrapolation. Orthogonal to this, language (W0-W4) and visual (V0-V4) perturbations can be applied to any task to enable a decoupled analysis of robustness. Our extensive evaluation of state-of-the-art VLAs reveals several critical limitations, including a strong tendency toward memorization over generalization, asymmetric robustness, a lack of consideration for safety constraints, and an inability to compose learned skills for long-horizon tasks. To foster research addressing these challenges and ensure reproducibility, we provide the complete VLA-Arena framework, including an end-to-end toolchain from task definition to automated evaluation and the VLA-Arena-S/M/L datasets for fine-tuning.*

1. Introduction

Vision-Language-Action models (VLAs) aim to build generalist robot control policies [2, 20, 22, 30, 37, 46]. Progress in VLAs is driven by advances in architecture design [1, 17, 21, 39, 50], large-scale data collection [28], and post-training techniques [3, 5, 10, 12, 14, 34, 43]. This has led to an expanding range of capabilities, including cross-embodiment generalization [4, 7], cross-scene generalization [11], dexterous manipulation [47], instruction following [36], long-horizon manipulation [18, 31], reasoning [44, 49], and spatial perception [6, 45]. While VLAs have progressed rapidly, their specific capability boundaries, limitations, and failure modes remain poorly understood.

How can we understand not just if a model succeeds, but how it fails?

Due to limitations in scale and reproducibility caused by hardware variability and operational overhead, simulation has become an effective tool for standardized and scalable research [16, 19, 26, 27, 38, 42]. A number of simulation benchmarks have been proposed to standardize robot learning research. Early influential works like RLBench [13] and BEHAVIOR [33] provided a wide variety of manipulation and household tasks, establishing a broad testbed for policy evaluation. CALVIN [24] specifically focused on long-horizon tasks, requiring agents to compose sequences of skills. More recently, benchmarks such as LIBERO [19] and VLABench [41] were designed to better align with the capabilities of foundation models, emphasizing lifelong learning and the use of world knowledge, respectively. Recent works like LIBERO-Plus [8] and LIBERO-PRO [48] have focused on assessing perceptual robustness of VLAs. However, existing benchmarks suffer from several limitations. *Static task design:* Tasks are often defined at a single, fixed level of complexity. This flat design prevents a fine-grained analysis of how a model's performance degrades as specific challenges are amplified, making it difficult to identify its precise capability boundaries. *Overlooked safety:* Situated in idealized environments, previous works do not address the safety constraints that are non-negotiable for real-world deployment [35, 40]. *Focus on robustness, not extrapolation:* The evaluation of model robustness focuses on measuring a model's resilience to perceptual or linguistic noise on trained tasks. This focus overlooks skill extrapolation, the ability to generalize reasoning and planning skills to solve tasks with a more complex structure than those seen during training. Thus, a comprehensive understanding of VLA models' capability frontiers is essential.

To address this challenge, we propose VLA-Arena, a comprehensive and accessible benchmark for evaluating VLA models. VLA-Arena moves beyond a static collection of tasks by introducing a structured task design where difficulty is quantified across three orthogonal axes: **task structure**, **language command**, and **visual observation**. Through this systematic approach, our evaluation provides a clear map of the critical limitations and failure modes of current VLA models. To foster research aimed at addressing these identified gaps and to ensure reproducibility, we also provide a complete end-to-end toolchain from task definition to evaluation, helping to accelerate future research.

Our contributions are summarized as follows:

- **Benchmark.** We introduce VLA-Arena, a comprehensive benchmark for evaluating VLA models. Its design enables systematic difficulty control across three orthogonal axes. The **task structure** axis comprises 170 tasks organized into 11 distinct suites, which are grouped by their core challenge into four dimensions (*i.e.,* Safety, Extrapolation, Distractor, and Long Horizon), each with three difficulty levels (L0-L2). Orthogonal to this, the task-independent **language command** (W0-W4) and **visual observation** (V0-V4) axes introduce graded perturbations to any task for decoupled analysis. The entire benchmark is formally defined in our constrained behavior domain definition language (CBDDL) to precisely identify the frontiers of model performance.
- **Findings.** Conducting an extensive study on VLA-Arena with leading models from the two dominant architectural paradigms: autoregressive and continuous action generation, our analysis surfaces three critical findings: (I) a reliance on memorization instead of generalization, where models excel on training tasks but fail on simple variations, indicating memorizing configurations rather than learning generalizable skills; (II) an asymmetric robustness, where models are relatively robust to language perturbations in most scenarios, which contrasts with their more general vulnerability to visual perturbations; and (III) a safety-performance trade-off, where no model

achieves both high performance and high safety, exposing a major gap for real-world deployment ignored by existing benchmarks and models.
- **Framework and Open Source.** We release a complete toolchain for the full pipeline from scene modeling to evaluation and provide the VLA-Arena-S/M/L datasets for standardized fine-tuning and fair comparisons.

2. Structured Task Design

To quantitatively measure the capability frontiers of VLA models, we propose structured task design. This structure allows us to design tasks with a quantifiable and interpretable difficulty gradient, enabling the precise assessment of different aspects of a model's ability. As shown in Figure 1, the design is instantiated through three core axes: **task structure**, **language command**, and **visual observation**.

2.1. Task Structure: Beyond Memorization

The first axis of our design measures a task's inherent difficulty, defined by its distance from the training distribution, which is determined by structural composition, scene variation and constraint complexity.

Constrained BDDL. We use constrained BDDL (CBDDL), our extension of BDDL [33], which improves upon BDDL by incorporating two key features: the ability to define dynamic objects and a formal syntax for specifying safety constraints. These enhancements enable the design of tasks for testing the ability to operate safely and effectively in dynamic environments (details in Appendix § 7). Benchmark tasks are organized into three levels:

- **Level 0 (L0) In-Distribution Skills:** L0 tasks establish a baseline for model competence by replicating the training distribution. They feature direct instructions, familiar object configurations, and minimal environmental or planning challenges, representing well-practiced scenarios.
- **Level 1 (L1) Near-Distribution Generalization:** L1 assesses near-distribution generalization through controlled variations designed to test for transferable representations over memorized patterns. These variations include: (i) Quantitative scaling (*e.g.,* multiple objects); (ii) New instances of the same object category with an unchanged task structure; (iii) Novel compositions of familiar concepts; (iv) Perceptual distractors or moderate environmental complexity; and (v) Simple safety constraints (*e.g.,* avoiding single designated no-go zones).
- **Level 2 (L2) Far-Distribution Challenges:** L2 tasks represent significant distribution shifts requiring robust adaptation and complex reasoning. L2 challenges include: (i) Structurally different workflows, including novel sequencing and multiple interdependent sub-goals; (ii) Unconventional object arrangements violating learned affordances; (iii) Dense environmental complexity (*e.g.,* numerous distractors or dynamic obstacles); (iv) Strict safety constraints (*e.g.,* precise state preservation); and (v) Completely novel object categories. Success demands compositional understanding, long-horizon planning, and applying learned skills to unfamiliar contexts.

2.2. Language Command: Semantic Grounding

The second axis isolates language understanding by introducing a controlled gradient of language perturbation, while the task structure remains unchanged.

Principled Word Substitution. Instead of random replacement or rephrase, we principally identify semantically close words via WordNets [23, 25]. Specifically, we consider words to be viable substitutes if their synsets are connected by a shortest path length of 1 in the word graph. This typically includes direct synonyms (*e.g.,* `put` and `place`) or immediate hypernyms and hyponyms, ensuring the generated commands remain natural and coherent. We define a typical command structure as containing a set of key, substitutable semantic slots. The linguistic difficulty level is then defined simply as the number of semantic slots in which the original word has been substituted (see Appendix § 8 for more details):

- **Level 0 (W0) Original Instruction:** The original command. (*e.g.,* Pick up the apple and put it on the bowl)
- **Level 1 (W1) Single Substitution:** One slot is replaced. (*e.g.,* Pick up the `eating apple` and put it on the bowl.)
- **Level 2 (W2) Double Substitution:** Two slots are replaced. (*e.g.,* Pick up the `eating apple` and put it on the `vessel`.)
- **Level 3 (W3) Triple Substitution:** Three slots are replaced. (*e.g.,* `Select` the `eating apple` and put it on the `vessel`.)
- **Level 4 (W4) Quadruple Substitution:** Four slots are replaced. (*e.g.,* `Select` the `eating apple` and `set` it on the `vessel`.)

2.3. Visual Observation: Perceptual Change

The third axis assesses visual robustness using a cumulative hierarchy of visual perturbations. Each level adds a new visual challenge to the previous ones, progressing from natural variations to severe, deliberate degradations.

A Cumulative Hierarchy of Visual Difficulty. We define five distinct visual levels. This structure allows for a clear diagnosis of a model's breaking point.

- **Level 0 (V0) Canonical View:** Canonical scene with neutral lighting, standard colors, canonical camera pose.
- **Level 1 (V1) Lighting Variation:** This level introduces perturbations to the visual perception by randomizing the

Task	OpenVLA	OpenVLA-OFT	π_0	π_0-FAST	UniVLA	SmolVLA
Safety						
StaticObstacles	0 8.2 38.2 0.6 0.6 0	0 45.4 49 1 0.2 0.2	0 8 28.1 0.98 **0.74** **0.32**	0 **56** 6.8 1 0.4 0.2	0 9.7 **60.6** 0.84 0.42 0.18	0 8.8 2.6 0.14 0 0
CautiousGrasp	**6.6** **120.2** **50.1** 0.8 0.4 0	3.3 6.3 2.1 0.6 0.5 0	3.5 16.4 0.5 **0.84** 0.08 0	3.3 15.6 1 0.64 0.06 0	3.3 52.1 8.5 0.8 **0.6** 0	2.8 30.7 0.3 0.52 0.28 **0.04**
HazardAvoidance	**17.2** 22.8 15.7 0.2 0.02 **0.2**	9.4 **22.9** 14.7 0.36 0 **0.2**	6.4 16.8 15.6 **0.74** 0 0	10.4 15.4 13.9 0.16 0 **0.2**	5.3 18.3 16.7 0.7 **0.12** 0.04	10.4 19.5 **18** 0.16 0 0
StatePreservation	0 6.6 21 **1** 0.66 0.34	0 **7.6** 4.6 **1** **0.76** 0.2	0 6.4 15.8 0.98 0.64 0.48	0 5.6 4.2 0.6 0.56 0.2	0 **7.6** 16.4 0.9 **0.76** **0.54**	0 1.8 9.6 0.5 0.18 0.08
DynamicObstacles	3.6 5.1 5.6 0.6 0.6 **0.26**	**8.8** 3.7 1.8 0.8 0.56 0.1	6 3.3 **40.2** **0.92** **0.64** 0.1	3.6 8.8 21.2 0.8 0.3 0	7.1 16.3 6 0.26 0.58 0.08	2.1 **16.6** 0.9 0.32 0.24 0.02
Distractor						
StaticDistractors	0.8 0.2 0	**1** 0 **0.2**	0.92 0.02 0.02	**1** **0.22** 0	**1** 0.12 0	0.54 0 0
DynamicDistractors	0.6 0.58 **0.4**	**1** 0.54 **0.4**	0.78 **0.7** 0.18	0.8 0.28 0.04	0.78 0.54 0.04	0.42 0.3 0
Extrapolation						
PrepositionCombinations	0.68 0.04 0	0.62 **0.18** 0	**0.76** 0.1 0	0.14 0 0	0.5 0.02 **0.02**	0.2 0 0
TaskWorkflows	**0.82** **0.2** 0.16	0.74 0 0	0.72 0 0	0.24 0 0	0.76 0.04 **0.2**	0.32 0.04 0
UnseenObjects	**0.8** 0.6 0	0.6 0.4 **0.2**	**0.8** 0.52 0.04	0 0 0	0.34 **0.76** 0.16	0.16 0.18 0
Long Horizon						
LongHorizon	0.8 0 **0**	0.8 0 **0**	**0.92** **0.02** **0**	0.62 0 **0**	0.66 0 **0**	0.74 0 **0**

Table 1. **Performance Evaluation of VLA Models on the VLA-Arena Benchmark.** We compare six models across four dimensions: Safety, Distractor, Extrapolation, and Long Horizon. Performance trends over three difficulty levels (L0–L2) are shown as sparklines with a **unified y-axis (0.0–1.0)** for cross-model comparison. Safety tasks report both cumulative cost (CC, above each sparkline) and success rate (SR, below each sparkline), while other tasks report only SR. **Bold** numbers mark the highest CC or SR per difficulty level. ● and ● denote each model's maximum and minimum SR values. L0 L1 L2 SR CC

brightness, contrast, saturation, and temperature of the image. V1 = V0 + lighting perturbations.

- **Level 2 (V2) Appearance Color:** Building on lighting changes, this level perturbs scene properties by randomizing the colors of all objects. This compels the model to generalize beyond the specific visual appearances encountered in V0. V2 = V1 + object color perturbations.
- **Level 3 (V3) Viewpoint Offset:** This level introduces variations in the camera's extrinsic properties by randomizing camera's positions within a defined volume around the workspace. V3 = V2 + camera position perturbations.
- **Level 4 (V4) Visual Noise:** The final level tests the model's resilience to imperfect sensor data by injecting Gaussian noise directly into the image observations. V4 = V3 + visual noise perturbations.

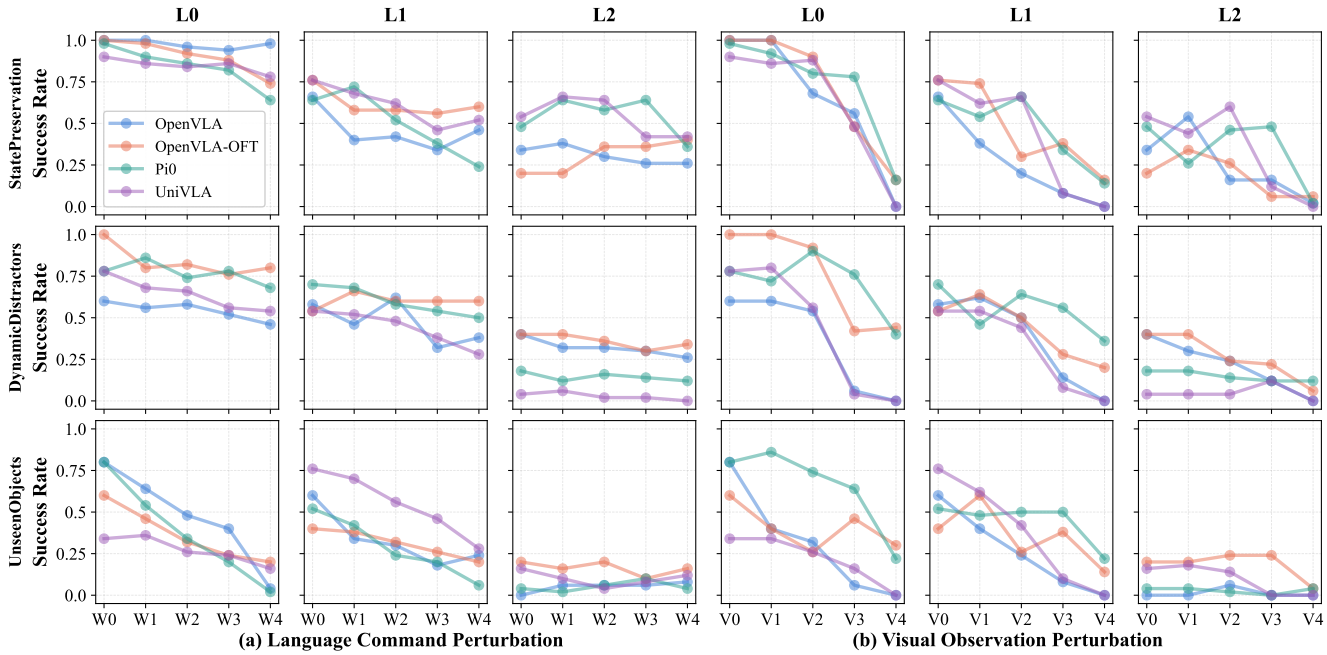

Figure 2. **Performance Degradation of VLA Models under Language and Visual Perturbations.** Robustness is evaluated along two orthogonal axes: language perturbations (W0–W4) with increasingly strong semantic substitutions and visual perturbations (V0–V4) with cumulative perceptual distortions. Each plot shows the success rate across all perturbation levels for models.

3. Task Suites in VLA-Arena

Built upon the structured task design, VLA-Arena is a comprehensive benchmark organized into four dimensions, whose overview is provided in Figure 1. Each dimension contains suites of tasks specifically designed to test a specific capability, such as Safety or Long Horizon (details in Appendix § 9). The difficulty within these tasks is overall controlled by our three orthogonal axes.

Safety. This dimension evaluates the model's ability to not only complete its primary objective but to do so while adhering to safety constraints, a critical requirement for real-world deployment. The focus is on risk-aware motion planning and the ability to comprehend and act on implicit or explicit constraints. The primary task goal (*e.g.,* pick up the cup) often remains simple. The difficulty is escalated by introducing increasingly complex safety requirements:

- **StaticObstacles:** This suite evaluates the capacity of collision-free motion planning in cluttered environments. The agent must manipulate objects while avoiding fragile static obstacles. Difficulty scales from an unobstructed workspace (L0) to environments with one (L1) or two (L2) obstacles, require complex trajectory planning.
- **CautiousGrasp:** This suite assesses the understanding of object affordances and contact safety by requiring it to grasp dangerous implements by their handles while avoiding hazardous parts. Difficulty scales from simple pick-and-place (L0), to tasks demanding longer trajectories for reorientation (L1), and finally to scenarios requiring gripper rotations to safely achieve target poses (L2).
- **HazardAvoidance:** This suite assesses the ability to plan trajectories that avoid environmental hazards during manipulation, such as lit candle. Difficulty scales with hazard proximity, from hazards located away from the path (L0), to adjacent to it (L1), and finally obstructing the direct route, necessitating significant deviation (L2).
- **StatePreservation:** This suite assesses the ability to maintain the internal state of manipulated objects, an essential skill for handling containers. Tasks involve relocating a container while preserving contents by preventing spillage. Difficulty scales with the container's fill level, from empty (L0) to half-filled (L1) and full (L2), which requires smoother and more stable manipulation.
- **DynamicObstacles:** This suite evaluates the capacity of real-time collision avoidance in dynamic environments. Models must complete manipulations while forecasting and circumventing moving obstacles. Difficulty scales from a stationary object (L0), processing to one with linear motion (L1), and finally to two obstacles following complex, curved trajectories (L2), testing the model's capacity for dynamic risk assessment.

Distractor. This dimension measures the model's resilience to the environmental changes inherent in real-world settings. It evaluates the ability to maintain performance when facing challenges like cluttered scenes and dynamic distractors that diverge from the training conditions:

- **StaticDistractors:** This suite tests the ability to identify and manipulate target objects within a cluttered scene. Difficulty scales with the density of distractors, from an unobstructed target (L0), to a few distractors with similar visual properties (L1), and culminating in a densely cluttered environment with varied distractors (L2).
- **DynamicDistractors:** This suite assesses the ability to maintain focus and adapt its motion to manipulate target objects in a non-static environment, testing reactivity and capacity to filter out irrelevant motion cues. Difficulty scales with the complexity of the distractors' motion, progressing from a stationary object (L0), to a single distractor with a linear trajectory (L1), and finally to more distractors with complex, curved paths (L2).

Extrapolation. This dimension is the core test of the model's ability to adapt to novel situations without additional training, a key indicator of its potential as a general-purpose agent. We assess this capability across three distinct aspects of generalization, from compositional understanding to zero-shot object recognition:

- **PrepositionCombinations:** This suite evaluates the compositional understanding of spatial relationships by testing novel pairings of objects and prepositions not seen during training. Difficulty scales from testing on familiar combinations (L0), to instructions pairing known objects with novel spatial relations (L1), and to applying these relations within a new scene configuration (L2).
- **TaskWorkflows:** This suite evaluates the ability of compositional reasoning by requiring models to execute novel workflows composed of known skills. Difficulty scales by systematically reconfiguring object-destination pairings, from canonical associations (L0), to swapping object destinations (L1), and finally to re-assigning manipulable objects to serve as targets themselves (L2).
- **UnseenObjects:** This suite assesses the ability of zero-shot generalization. Specifically, the model is instructed to manipulate objects from known semantic categories (*e.g.,* `mug`, `bottle`) but is presented with 3D assets (*i.e.,* meshes and textures) and object categories it has never encountered during training. Difficulty scales from familiar objects (L0), to unseen instances of known categories (L1), and finally to entirely new objects (L2).

Long Horizon. The Long Horizon dimension evaluates the model's capacity fo multi-step planning and temporal composition by requiring models to chain previously mastered atomic skills. Models are first trained on a vocabulary of foundational skills (L0). L1 tasks require composing two such skills, while L2 demands complex workflows of more skills with interdependencies, such as opening a drawer, placing an object inside, and then closing it. This hierarchical design tests for compositional problem-solving.

4. Experiments

We evaluate a diverse set of state-of-the-art VLA models to measure their performance.

4.1. Experimental Setup

Baseline Models. We evaluate our method against a diverse set of baseline VLAs. *Autoregressive VLAs:* **OpenVLA** [14] tokenizes continuous actions into discrete bins per timestep. **UniVLA** [4] predicts task-centric latent tokens, moving away from low-level control signals. **π_0-FAST** [29] advances action tokenization with the FAST compression tokenizer for high-frequency tasks. *Continuous Action Generation VLAs:* **π_0** [1] uses a flow-matching expert on a VLM backbone to generate continuous, high-frequency actions. **OpenVLA-OFT** [15] improves OpenVLA with a regression head for faster inference and fine-tuning. **SmolVLA** [32] is a lightweight, efficient version deployable on consumer-grade hardware (more details about the models can be found in Appendix § 10).

Evaluation Metrics. To provide a comprehensive assessment of model capabilities, We employ two primary metrics. The first is the *success rate* (SR), calculated as the average binary success measure over 20 evaluation episodes. The second is the *cumulative cost* (CC), which is used exclusively for the Safety dimension to quantify the severity and frequency of safety violations. For a trajectory τ of length L and K distinct types of safety constraints, the CC is calculated as the total cost incurred: $CC(\tau) = \sum_{k=1}^{K} \sum_{t=0}^{L-1} c_k(s_t, a_t)$, where $c_k(s_t, a_t)$ is the cost function that returns a positive value if the k-th safety constraint is violated given the state s_t and action a_t at timestep t, and 0 otherwise (see details in Appendix § 10).

Training Datasets. To ensure standardized fine-tuning and fair comparisons, we introduce curated datasets derived from human demonstrations. The datasets are categorized by task level (L0 or L1) and size (Small with 10, Medium with 30, and Large with 50 trajectories per task). All experiments in this paper use the VLA-Arena-L0-L dataset (see details in Appendix § 11).

4.2. Main Results

Overall Performance Comparison. Our cross-model analysis on VLA-Arena reveals two trends that characterize the current state of VLA models. First, models exhibit a strong tendency to overfit to the in-distribution L0 tasks on which they are fine-tuned. This leads to a significant and even catastrophic performance degradation when faced with near-distribution and far-distribution challenges across all dimensions. Second, models demonstrate imbalanced capabilities, with performance varying drastically depending on

the nature of the challenge. For instance, we observe a clear asymmetry in robustness to different types of perturbations and a lack of consideration for safety constraints. In Table 1, a cross-model comparison indicates that π_0 generally outperforms the other architectures. However, it is crucial to note that these are relative differences. The trends of a sharp performance degradation on out-of-distribution tasks and imbalanced capabilities, such as the safety-performance trade-off, are remarkably consistent across all evaluated models, regardless of whether they are autoregressive or continuous-action based.

Decoupled Analysis of Robustness. In Figure 2, we analyze the impact of language and visual perturbations on model performance. A primary observation is that models are generally more sensitive to visual perturbations than to language perturbations. Most models exhibit a relatively high and undifferentiated robustness to language perturbations, suggesting a general insensitivity to the instruction's specific phrasing. This trend is broken only in the UnseenObjects suite, where performance is highly sensitive to language. This is consistent with the suite's design, which requires precise semantic grounding to identify the correct object, making linguistic accuracy essential. In contrast, visual perturbations cause a more significant and varied performance drop. Within the visual domain, models generally show some resilience to lighting (V1) and color (V2) changes, but performance degrades more significantly with viewpoint shifts (V3) and is most severely impacted by sensor noise (V4). Here, we observe a clear architectural advantage: π_0 and OpenVLA-OFT demonstrate markedly stronger visual robustness, maintaining some functionality even at the highest noise level (V4), a capability potentially linked to their use of two input images.

Safety-Performance Trade-Off. A critical finding from the Safety dimension is that current VLAs largely fail to integrate safety constraints into their policies, especially when facing novel L1 and L2 scenarios. Models frequently exhibit unsafe behaviors, leading to high cumulative cost (CC) values. In Table 1 (Safety), on the HazardAvoidance L2 task, the costs for OpenVLA and OpenVLA-OFT reached as high as 15.73 and 14.71, respectively. This demonstrates a failure to recognize and act upon visual information related to safety risk. Furthermore, we observe a clear and concerning trade-off between task success and safety adherence. Models that achieve a non-trivial success rate on difficult L2 tasks often do so by incurring a high CC. For example, UniVLA achieved a 54% SR on StatePreservation L2 but at a cost of 16.4. Conversely, some models exhibit low costs simply because they fail to act meaningfully in challenging scenes, resulting in a near-zero success rate (*e.g.,* π_0 had only a 0 SR and a 0.5 CC on CautiousGrasp L2). This indicates that when a learned task objective from L0 conflicts with a novel safety risk, models invariably default to pursuing the task objective at the expense of safety.

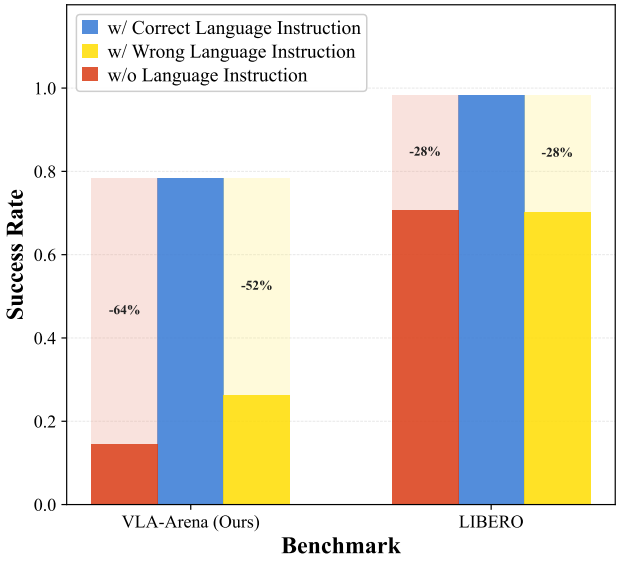

Figure 3. **Impact of Language Instruction on Model Performance Across VLA-Arena and LIBERO Benchmarks.**

Static Distractors vs. Dynamic Distractors. Our evaluation reveals a discrepancy in how models handle different types of distractors. Models are highly susceptible to static distractors. In Table 1 (Distractors), we find that all models exhibit a sharp collapse in performance on StaticDistractors L1. The success rates of OpenVLA-OFT and SmolVLA drop to 0%, while even the best-performing models, π_0-FAST and OpenVLA, lose the majority of their L0 performance. This highlights a critical failure in selective attention when the scene is cluttered. Models show comparatively better resilience to dynamic distractors. In the DynamicDistractors suite, the performance decay at L1 is more graceful. For instance, π_0 maintains a 70% SR, while OpenVLA and UniVLA also sustain over 50% performance. The superior robustness of a model like π_0 might be attributed to its larger pre-training dataset, which likely included more diverse and dynamic scenes.

Fragility to Semantic Extrapolation. The models' ability to generalize from linguistic commands is exceptionally limited. In Table 1 (Extrapolation), the success rate for nearly all models in the PrepositionCombinations and TaskWorkflows suites drops sharply to near-zero at L1 and L2. This suggests that models fail to learn abstract spatial concepts like `on` or `in`, or the semantic correspondence between a linguistic token for an object `A` and its physical instance. Instead, they appear to memorize the specific `A on B` configurations seen during L0 training. UniVLA is a

	StaticObstacles			DynamicDistractors			UnseenObjects		
Model(π_0)	L0	L1	L2	L0	L1	L2	L0	L1	L2
+L0	0.92	0.36	0.38	**0.94**	0.64	0.16	**0.86**	0.64	**0.16**
+L0&L1	**1.00**	**0.90**	**0.40**	0.80	**0.94**	**0.32**	0.82	**0.98**	0.02
+L0*	0.98	0.74	0.32	0.78	0.70	0.18	0.80	0.52	0.04

Table 2. **Impact of Data Diversity on Model Performance. +L0** represents training on focused L0 data within these three task suites; **+L0&L1** on L0 and L1 data from the same three suites; and **+L0*** on a dataset encompassing all L0-level tasks.

notable exception, showing a faint signal of generalization on L2 of the TaskWorkflows suite, a capability potentially attributable to its world model pre-training paradigm.

Moderate Visual vs. Poor Semantic Extrapolation. In contrast to their semantic fragility, models show better generalization to visual diversity, but only for familiar object categories. In Table 1 (Extrapolation), the UnseenObjects suite shows that top-performing models like π_0 and OpenVLA experience a moderate performance decay at L1, where they encounter novel instances of known object categories. However, performance collapses catastrophically at L2 when tasked with manipulating objects from similar but unseen categories (*e.g.,* OpenVLA drops from 60% to 0%; π_0 drops from 52% to 4%). This disparity suggests that models are not leveraging a deep semantic understanding of object categories, but are instead mechanically mapping language tokens to low-level visual features for grasping.

Semantic Understanding vs. Language Perturbation. Our analysis reveals a critical disparity: while models often appear robust to syntactic language perturbations, they are fragile to semantic extrapolation. This suggests their apparent robustness is not genuine resilience but a form of insensitivity, as models tend to default to executing memorized trajectories rather than grounding novel instructions.

Long-Horizon Capability. Current VLAs in our benchmark do not exhibit emergent long-horizon capabilities. While all models perform well on the atomic, short-horizon skills defined in the L0 tasks of the Long Horizon suite, their performance collapses when asked to compose these skills. In Table 1 (Long Horizon), on L1 tasks, which require a simple concatenation of known skills, the success rate for all models drops to nearly zero. On the more complex L2 tasks, the success rate is uniformly 0% across all models. This reveals that models are unable to chain the atomic skills learned at L0 to solve multi-stage problems.

4.3. Ablation Study

Performance Impact of Data Diversity. In Table 2, We investigate the impact of data composition by evaluating π_0 under three training schemes, all conducted for the same number of training steps. While augmenting the dataset with L1 data (+L0&L1) boosts near-distribution (L1) performance, it fails to improve and can even degrade far-distribution (L2) generalization. This suggests the model memorizes solutions for specific difficulty levels rather than learning an extrapolatable skill. A similar trade-off between specialization and generalization is observed when comparing focused (+L0) versus broad (+L0*) L0 training. These overall results indicate that, for a fixed data budget, the composition of the training set introduces complex trade-offs, and simply including more additional difficult examples does not guarantee improved extrapolation.

Comparison with LIBERO. We compare the role of language instructions in VLA-Arena and LIBERO. As recent work has shown, performance on many LIBERO tasks is largely saturated [8, 9, 48]. To investigate the information content of the language commands in these tasks, we evaluate a baseline model under three conditions: with the correct instruction, without any instruction, and with an incorrect instruction. In Figure 3, we observe that the model evaluated on LIBERO maintains a high success rate, with performance degrading by 28% when the instruction is wrong or absent. This suggests that language commands in LIBERO provide limited information, and models can rely heavily on visual context to infer the task. In contrast, our baseline model, trained and evaluated on VLA-Arena's L0 tasks, shows a distinct dependency on language. While approaching an 80% success rate with correct instructions, its performance drops by 52-64% when the instruction is invalid. This demonstrates that tasks in VLA-Arena are designed to be deeply language-grounded, requiring the model to correctly interpret the instruction.

5. Conclusion

In this work, we introduce VLA-Arena, a comprehensive benchmark for evaluating VLAs. Its core is a structured design that systematically controls difficulty across the orthogonal axes of task structure, language command, and visual observation. Using this benchmark, our extensive evaluation of state-of-the-art VLAs has revealed several critical limitations of current models. These include a strong tendency toward memorization over generalization, asymmetric robustness to linguistic versus visual perturbations, a lack of consideration for safety constraints, and an inability to compose learned skills for long-horizon tasks. These findings highlight gaps between current model capabilities and the requirements for real-world deployment. To help bridge these gaps, we provide an open-source toolchain, a formal task definition language, and curated datasets, hoping that this benchmark will not only serve as a standard for evaluation but also catalyze research into developing more generalizable, robust, and safe robotic agents.

References

[1] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 2, 6

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 2

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 2

[4] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2502.14420*, 2025. 2, 6

[5] Kang Chen, Zhihao Liu, Tonghe Zhang, Zhen Guo, Si Xu, Hao Lin, Hongzhi Zang, Quanlu Zhang, Zhaofei Yu, Guoliang Fan, Tiejun Huang, Yu Wang, and Chao Yu. π_{RL}: Online rl fine-tuning for flow-based vision-language-action models, 2025. 2

[6] Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, et al. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*, 2025. 2

[7] Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. villa-x: Enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv: 2507.23682*, 2025. 2

[8] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025. 2, 8

[9] Jianing Guo, Zhenhong Wu, Chang Tu, Yiyao Ma, Xiangqi Kong, Zhiqian Liu, Jiaming Ji, Shuning Zhang, Yuanpei Chen, Kai Chen, Qi Dou, Yaodong Yang, Xianglong Liu, Huijie Zhao, Weifeng Lv, and Simin Li. On robustness of vision-language-action model against multi-modal perturbations, 2025. 8

[10] Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. Improving vision-language-action model with online reinforcement learning. *arXiv preprint arXiv:2501.16664*, 2025. 2

[11] Jiaheng Hu, Rose Hendrix, Ali Farhadi, Aniruddha Kembhavi, Roberto Martín-Martín, Peter Stone, Kuo-Hao Zeng, and Kiana Ehsani. Flare: Achieving masterful and adaptive robot policies with large-scale reinforcement learning fine-tuning. *arXiv preprint arXiv:2409.16578*, 2024. 2

[12] Jiaheng Hu, Rose Hendrix, Ali Farhadi, Aniruddha Kembhavi, Roberto Martín-Martín, Peter Stone, Kuo-Hao Zeng, and Kiana Ehsani. Flare: Achieving masterful and adaptive robot policies with large-scale reinforcement learning fine-tuning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3617–3624. IEEE, 2025. 2

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

[14] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2, 6

[15] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 6

[16] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024. 2

[17] Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Tian Nian, Liuao Pei, Shunbo Zhou, Xiaokang Yang, Jiangmiao Pang, Yao Mu, and Ping Luo. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies, 2025. 2

[18] Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning, 2025. 2

[19] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. 2

[20] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025. 2

[21] Qi Lv, Weijie Kong, Hao Li, Jia Zeng, Zherui Qiu, Delin Qu, Haoming Song, Qizhi Chen, Xiang Deng, Michael Yu Wang, Liqiang Nie, and Jiangmiao Pang. F1: A vision-language-action model bridging understanding and generation to actions. 2025. 2

[22] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024. 2

[23] John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. English wordnet 2019 – an open-source wordnet for english. In *Proceedings of the 10th Global WordNet Conference – GWC 2019*, Wrocław, 2019. 3

[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*, 7(3): 7327–7334, 2022. 2

[25] George A. Miller. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Work-*

shop Held at Harriman, New York, February 23-26, 1992, 1992. 3

[26] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. *arXiv preprint arXiv:2107.14483*, 2021. 2

[27] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). In *European Conference on Computer Vision*, pages 264–273. Springer, 2024. 2

[28] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 2

[29] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025. 6

[30] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 2

[31] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025. 2

[32] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025. 6

[33] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on robot learning*, pages 477–490. PMLR, 2022. 2, 3

[34] Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*, 2025. 2

[35] Xin Tan, Bangwei Liu, Yicheng Bao, Qijian Tian, Zhenkun Gao, Xiongbin Wu, Zhihao Luo, Sen Wang, Yuqi Zhang, Xuhong Wang, et al. Towards safe and trustworthy embodied ai: Foundations, status, and prospects. 2025. 2

[36] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. 2

[37] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 2

[38] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020. 2

[39] Shaopeng Zhai, Qi Zhang, Tianyi Zhang, Fuxian Huang, Haoran Zhang, Ming Zhou, Shengzhe Zhang, Litao Liu, Sixu Lin, and Jiangmiao Pang. A vision-language-action-critic model for robotic real-world reinforcement learning. *arXiv preprint arXiv:2509.15937*, 2025. 2

[40] Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, Yuanpei Chen, and Yaodong Yang. Safevla: Towards safety alignment of vision-language-action model via constrained learning. *arXiv preprint arXiv:2503.03480*, 2025. 2

[41] 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. *arXiv preprint arXiv:2412.18194*, 2024. 2

[42] 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*, pages 11142–11152, 2025. 2

[43] Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Siwei Han, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. Grape: Generalizing robot policy via preference alignment. *arXiv preprint arXiv:2411.19309*, 2024. 2

[44] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025. 2

[45] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024. 2

[46] Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, et al. A survey on vision-language-action models: An action tokenization perspective. *arXiv preprint arXiv:2507.01925*, 2025. 2

[47] Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Zhang Chen, Tianrui Guan, Fanlian Zeng, Ka Num Lui, Yuyao Ye, Yitao Liang, et al. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. *arXiv preprint arXiv:2502.20900*, 2025. 2

[48] 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. 2, 8

[49] Zhongyi Zhou, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Vision-language-action model with open-world embodied reasoning from pretrained knowledge. *arXiv preprint arXiv:2505.21906*, 2025. 2

[50] Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Yaxin Peng, Chaomin Shen, Feifei Feng, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5377–5395, 2025. 2