# CofactVLA: Deconfounding Vision-Language-Action Models via Counterfactual Intervention

**Yan Zhang**
Department of Automation
Tsinghua University
yan_zhang@ieee.org

**Yinan Wu***
Department of Automation
Tsinghua University
yinanwu@ieee.org

**Haoran Duan**
Department of Automation
Tsinghua University
haoran.duan@ieee.org

**Jungong Han***†
Department of Automation
Tsinghua University
jghan@tsinghua.edu.cn

## Abstract

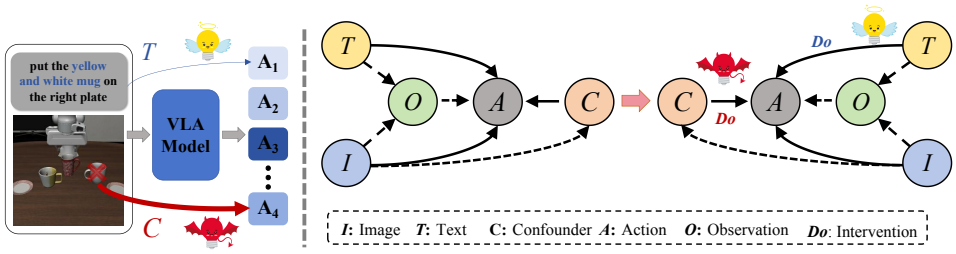
Vision-Language-Action (VLA) models have driven significant progress in robotic manipulation, yet they fundamentally struggle with the "vision-override" phenomenon. Driven by the severe modality imbalance between dense visual streams and sparse linguistic instructions, VLAs frequently fall prey to causal confusion. Instead of treating language as the primary causal driver, the policy entirely bypasses the original instruction by overfitting to spurious visual confounders, such as prominent objects or familiar layouts. To systematically alleviate this bias, we formalize the process of action generation as a Dual-path Deconfounding Graph (DDG) and propose **CofactVLA**, a novel causal intervention framework. By dynamically constructing a language-masked counterfactual branch within a single forward pass, CofactVLA estimates and attenuates observation-dominant components associated with visual shortcuts through two synergistic mechanisms. First, Action-Level Orthogonal Projection Guidance (OPG) geometrically projects the factual velocity field away from the counterfactual visual bias during continuous flow matching, emphasizing a language-aligned action component. Second, Feature-Level Counterfactual Covariance Reduction (CCR) mathematically deconfounds latent representations by penalizing the positive eigenspace of the covariance difference, explicitly suppressing dominant visual shortcuts while preserving the causal language intent. Extensive experiments demonstrate that CofactVLA establishes a new state-of-the-art across diverse simulation benchmarks. Beyond simulation, real-world robot experiments demonstrate the causal efficacy of our method in bridging the generalization gap, yielding a 52.3% absolute success rate gain under out-of-distribution scenarios.

## 1 Introduction

The integration of pre-trained Vision-Language Models (VLMs) into robotic control has driven remarkable progress in Vision-Language-Action (VLA) models[1, 2, 3, 4, 5, 6, 7, 8, 9]. However, the deployment of VLAs in open-world environments reveals a critical vulnerability known as the **Vision-Override Phenomenon**. Due to the severe modality imbalance in robotics datasets—where dense visual streams inevitably dominate sparse linguistic instructions—VLA models frequently

*Corresponding authors: Yinan Wu and Jungong Han.
†Jungong Han is also with Beijing National Research Center for Information Science and Technology, Tsinghua University.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).




**(a) Motivation: Vision-Override Phenomenon**

**(b) Dual-path Deconfounding Graph**

Figure 1: **(a) Motivation:** Visual modality inadvertently induces a spurious backdoor path ($I \dashrightarrow C \rightarrow A$). The visual confounder $C$ acts as a shortcut that dictates the action $A$, overriding the true semantic intent $T$. **(b) Our proposed Dual-path Deconfounding Graph:** The red $do$-operator signifies Feature-Level Counterfactual Covariance Reduction (CCR), which explicitly suppresses the spurious visual confounder $C$. The blue $do$-operator signifies Action-Level Orthogonal Projection Guidance (OPG), which geometrically extracts and enhances the pure language intent $T$.

suffer from **causal confusion**, as shown in Figure 1(a). Instead of treating the language instruction as the causal driver, the model overfits to spurious visual confounders, such as habitually grasping the most salient object while entirely ignoring the text prompt.

Recent efforts to address the vision-override phenomenon generally fall into two categories. Data-centric approaches (e.g., linguistic rephrasing [10] or data augmentation in specific domains [11, 12]) are notoriously difficult to scale across open-world tasks, thereby failing to systematically decouple entangled representations. At the action level, Counterfactual action guidance [13] attempt to suppress visual priors via dual-branch scalar subtraction. However, this post-hoc linear extrapolation often inflates unaligned noise, pushing continuous robotic trajectories out-of-distribution and causing unsafe executions. More importantly, these methods lack a principled mechanism to address structural modality collapse within the model's internal representations.

To this end, we approach the vision-override problem through the lens of **Counterfactual Intervention**. We first formalize the decision-making process as a **Dual-path Deconfounding Graph (DDG)**, as shown in Figure 1(b). Ideally, the robotic action $A$ should be properly conditioned on the language instruction $T$ and the image observation $I$. However, due to the severe modality imbalance in training data, the model inadvertently learns a spurious causal path $I \dashrightarrow C \rightarrow A$. Here, $C$ represents the visual confounder (e.g., prominent out-of-context objects or familiar scene layouts). This confounder acts as a dominant visual shortcut that heavily dictates the action $A$ while bypassing the text instruction $T$. To eliminate the spurious effect of $C$ and recover the true causal effect of the language instruction, we formulate a counterfactual question: *"What action would the VLA model inherently take in this exact visual scene if the language instruction were entirely absent?"* A language-masked counterfactual branch provides an observation-only reference that exposes observation-dominant behavior. This allows us to perform causal interventions to sever the spurious path (the red $do$-operator on $C \rightarrow A$) and increase the relative contribution of language-aligned action components (the blue $do$-operator on $T \rightarrow A$). To summarize, our main contributions are as follows:

- **A Unified Deconfounding Framework for VLAs**: We propose CofactVLA, which formulates the vision-override phenomenon as a causal confusion problem. By constructing the action generation process as a DDG, we introduce a language-masked counterfactual branch within a single forward pass, which aims to isolate spurious visual confounders.
- **Action-Level Orthogonal Projection Guidance (OPG)**: We introduce an action-level geometric decoupling strategy for continuous flow matching, which projects the factual velocity field orthogonally away from the counterfactual visual bias to extract the pure semantic intent, yielding strictly deconfounded action trajectories.
- **Feature-Level Counterfactual Covariance Reduction (CCR)**: To explicitly mitigate the interference of visual confounders at the feature level, we introduce CCR by penalizing the positive eigenspace of the covariance difference between the counterfactual and factual.
- **State-of-the-Art Performance and OOD Generalization**: CofactVLA achieves highly competitive, often state-of-the-art results across simulated and real-world physical deploy-

ments, while exhibiting robust zero-shot generalization in unseen Out-of-Distribution (OOD) scenarios.

## 2 Related Work

### 2.1 Vision-Language-Action Models

Vision-Language-Action (VLA) models have emerged as powerful end-to-end visuomotor policies that map raw visual observations and natural-language instructions directly to low-level control actions. Typically, these models adapt large-scale pre-trained Vision-Language Models (VLMs) via Supervised Fine-Tuning (SFT) or Reinforcement Learning (RL) [1, 2, 14, 15]. A predominant paradigm formulates continuous control as a sequence modeling task, discretizing action chunks and employing autoregressive token decoding to inherit the scalability and open-vocabulary generalization of foundation models [1, 2, 14, 16, 17]. To enhance deployment practicality and generalization, subsequent researches have explored optimized fine-tuning strategies [18], parameter-efficient adaptation at smaller scales [19, 20, 21, 22], improved perceptual grounding [23], and continual post-training from experience [24]. Concurrently, to alleviate the latency bottleneck of autoregressive generation, efficiency-oriented VLAs have investigated parallel decoding and system-level accelerations, such as token compression, speculative execution, and extreme quantization [18, 25, 26, 27].

Beyond autoregressive paradigms, diffusion and flow-matching frameworks have gained traction for modeling complex, multi-modal continuous action distributions, synthesizing action chunks via iterative denoising or ordinary differential equation (ODE) integration [28, 29, 30, 31, 32]. Recent advancements have successfully scaled these continuous formulations, utilizing distillation to accelerate sampling [33, 34], while flow-matching-based VLAs and hybrid diffusion-autoregressive architectures explicitly target high-frequency, generalist robotic control [35, 36, 17, 16].

### 2.2 Causality and Deconfounding in Robotic Manipulation

Causal inference has increasingly garnered attention for mitigating distribution shifts and enhancing the robustness [37, 38, 39, 13]. While early robotics efforts injected causal priors into imitation frameworks [40, 41, 29], recent language-conditioned VLAs primarily rely on counterfactual data augmentation to reduce spurious visual shortcuts [11, 12]. However, these data-centric methods inherently struggle to scale across open-world scenarios and fail to fundamentally disentangle collapsed internal representations. In contrast, our proposed CofactVLA explicitly formulates the action generation process as a **Dual-path Deconfounding Graph**. Guided by this structure, we execute a systematic dual-path intervention: operating simultaneously at the latent feature level to suppress visual confounders, and at the continuous action level to geometrically extract the pure semantic intent.

## 3 Methods

### 3.1 A Causal View of VLA

VLA models aim to learn a continuous policy $\pi(A \mid O, T)$ that maps multimodal observations $O$ and language instructions $T$ to actions $A$ via generative processes (e.g., flow matching). We formalize this decision-making process through structural causal model, as illustrated in Figure 1. Ideally, $A$ should be causally driven by the semantic intent $T$ and grounded by $O$. However, due to the severe modality imbalance in training data, the model inadvertently learns a spurious backdoor path $I \dashrightarrow C \rightarrow A$. Here, $C$ represents the latent visual confounder (e.g., prominent out-of-context objects or familiar layouts) that heavily dictates the action $A$ and overrides the causal effect of $T$.

To mitigate the spurious effect of $C$ and recover the true causal effect of $T$, we pose a counterfactual question: *"What action would the VLA inherently take if the language instruction were entirely absent?"* To operationalize this, we propose a language-masked dual-branch framework by combining a factual branch (conditioned on $O$ and $T$) with a counterfactual branch (conditioned on $O$). Thus, its generated representations and actions are driven exclusively by its inherent visual biases for the given scene. Building upon this design, CofactVLA deconfounds the model through two complementary interventions (the $do$-operators in Fig. 2), detailed in the following sections.

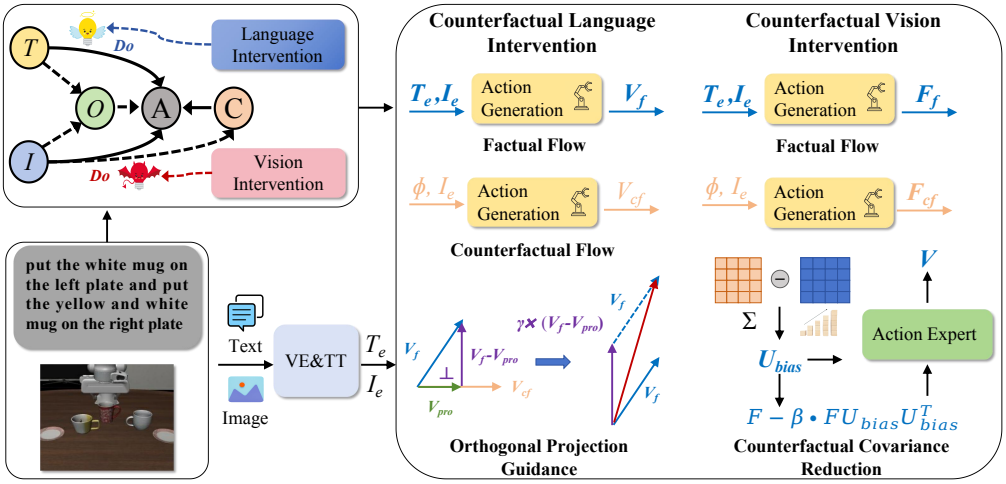



Figure 2: **The overall architecture of CofactVLA.** (Left) Driven by the proposed Dual-path Deconfounding Graph (DDG), we propose an intervention framework to address the vision-override phenomenon. (Right) The framework consists of two core components dynamically built upon factual and counterfactual flows. To execute Language Intervention, Orthogonal Projection Guidance (OPG) geometrically extracts the pure semantic intent by projecting the factual flow $V_f$ onto the counterfactual visual prior $V_{cf}$. To execute Vision Intervention, Counterfactual Covariance Reduction (CCR) extracts the dominant visual confounder bias $U_{bias}$ via the covariance difference of latent features to cleanse the features before feeding them to the action expert.

### 3.2 Action-Level Deconfounding: Orthogonal Projection Guidance (OPG)

Building upon the established counterfactual anchor, we first introduce our causal intervention at the action level. A fundamental challenge in robotic control lies in its inherently generative and multi-valued: a single semantic intent $T$ could be successfully executed via multiple valid trajectories.

From a probabilistic perspective, let $\mathcal{A}^H$ be the continuous action-chunk space with action horizon $H$, partitioned into discrete valid modes $Z = h(A)$ (e.g., different grasping styles or approach angles). Even if the model perfectly perceives the text instruction $T$, the multi-valued nature of the expert demonstrations allows the visual confounder $C$ to spuriously bias the mode selection. Formally, if the demonstrations exhibit nuisance-correlated mode selection $MI(Z;C \mid O) > 0$, data processing inequalities yield $MI(A;C \mid O) > 0$, where $MI(\cdot;\cdot|\cdot)$ represents the conditional mutual information. This induces a strict negative log-likelihood (NLL) gap under distribution shifts:

$$H(A \mid O) - H(A \mid O, C) = MI(A;C \mid O) > 0, \tag{1}$$

where $H(\cdot \mid \cdot)$ denotes conditional entropy. This identity suggests that nuisance-correlated cues may influence action-mode selection and impair generalization when these correlations change across environments. To eradicate this bias without destroying the valid action manifold, an explicit intervention is required at the generative output space to re-weight the mode selection probabilities.

We formulate this intervention within the continuous flow matching framework. Let $p^\tau_{cond} := p(A^\tau \mid O)$ and $p^\tau_{uncond} := p(A^\tau \mid T)$ denote the marginal densities of the factual and counterfactual branches at diffusion time $\tau$, respectively. To extract the pure semantic intent, we aim to perform a Product-of-Experts (PoE) composition in the probability density space. Crucially, under the Gaussian forward process $A^\tau = \tau A + (1-\tau)\omega$, the score function $\nabla_{A^\tau} \log p^\tau$ is linearly convertible from the predicted flow-matching velocity field $v_\theta(A^\tau, \tau \mid \cdot)$:

$$\nabla_{A^\tau} \log p^\tau(A^\tau \mid \cdot) = -\frac{A^\tau + \tau v_\theta(A^\tau, \tau \mid \cdot)}{1-\tau}. \tag{2}$$

Because of this exact linear equivalence, manipulating the generated velocity fields mathematically corresponds to composing the causal score functions. Counterfactual action guidance [13] attempts this via simple scalar extrapolation: $v_{CFG} = v_{uncond} + \gamma(v_{cond} - v_{uncond})$. However, in the high-dimensional non-linear action space, this blind scalar subtraction assumes orthogonal separability. When this assumption fails, it exponentially inflates unaligned visual noise, pushing the generated

flow out-of-distribution and causing unsafe physical executions. To robustly perform the causal composition, we propose OPG. Let $v_{cond}$ and $v_{uncond}$ be the predicted velocity fields from the factual and counterfactual branches. We geometrically project the factual field onto the counterfactual visual instinct:

$$v_{proj} = \frac{\langle v_{cond}, v_{uncond} \rangle}{\|v_{uncond}\|_2^2 + \epsilon} v_{uncond}. \tag{3}$$

By explicitly subtracting this collinear visual bias, we extract the strictly orthogonal velocity component $v_\perp$, which represents the pure, deconfounded language semantic intent:

$$v_\perp = v_{cond} - v_{proj}. \tag{4}$$

The final causally intervened velocity field is then constructed as:

$$v_{causal} = v_{cond} + \gamma \cdot v_\perp, \tag{5}$$

where $\gamma > 0$ is the guidance scale. By virtue of the score equivalence, this construction amplifies the component of the factual velocity field that is orthogonal to the observation-only reference, aiming to reduce observation-dominant bias while avoiding indiscriminate scalar extrapolation.

### 3.3 Feature-Level Deconfounding: Counterfactual Covariance Reduction (CCR)

While OPG robustly corrects the multi-modal action distribution at the output level, the root cause of the vision-override phenomenon originates within the latent representation space. Due to the severe modality imbalance during training, visual confounders within the VLA backbone often dominate the attention mechanisms, establishing spurious shortcuts in the Key-Value (KV) cache before the action generator even receives the features. To systematically eradicate this, we propose Counterfactual Covariance Reduction (CCR), a causal intervention applied directly to the latent attention features.

Let $F_f$ and $F_{cf} \in \mathbb{R}^{N \times d}$ denote the flattened valid attention features (e.g., Keys and Values) from the transformer layers of the factual and counterfactual branches, respectively. We compute their centered empirical covariance matrices: $\Sigma_f = \mathrm{Cov}(F_f)$ and $\Sigma_{cf} = \mathrm{Cov}(F_{cf})$, and formulate our intervention based on the covariance difference:

$$\Delta\Sigma = \Sigma_{cf} - \Sigma_f \tag{6}$$

To mathematically justify how this matrix isolates the visual bias, we introduce a gain–bias decomposition and a contrastive eigengap that separates nuisance- and observation-related directions.

**Assumption 1 (Gain–bias decomposition on the interface)** *The latent feature space $\mathbb{R}^{d_r}$ could be orthogonally decomposed into two structural subspaces: the causal observation-driven intent subspace $\mathcal{S}_O \subset \mathbb{R}^{d_r}$ and the spurious visual confounder subspace $\mathcal{S}_C \subset \mathbb{R}^{d_r}$, i.e., $\mathbb{R}^{d_r} = \mathcal{S}_O \oplus \mathcal{S}_C$. There exist measurable functions $h_O, h_C$ almost surely*

$$\widetilde{\Delta} = \widetilde{\Delta}_O + \widetilde{\Delta}_C, \widetilde{\Delta}_O = h_O(O) \in \mathcal{S}_O, \widetilde{\Delta}_C = h_C(C) \in \mathcal{S}_C. \tag{7}$$

**Assumption 2 (Contrastive eigengap)** *Let $M := \Sigma_0^{-1/2} \Sigma_\Delta \Sigma_0^{-1/2}$. The eigenvalues of $M$ restricted to $\Sigma_0^{1/2}\mathcal{S}_C$ are strictly larger than those restricted to $\Sigma_0^{1/2}\mathcal{S}_O$:*

$$\lambda_{\min}\big(M|_{\Sigma_0^{1/2}\mathcal{S}_C}\big) \;>\; \lambda_{\max}\big(M|_{\Sigma_0^{1/2}\mathcal{S}_O}\big). \tag{8}$$

Then, the visual confounder is geometrically identifiable. Let $\Delta\Sigma = U\Lambda U^\top$ be the eigendecomposition of the covariance difference. We isolate the positive eigenspace by selecting the top-$k$ eigenvectors corresponding to $\lambda > \epsilon > 0$, constructing the nuisance basis $U_{bias} \in \mathbb{R}^{d \times k}$. First, we project the latent features orthogonally away from $U_{bias}$ to eliminate the visual confounder while strictly preserving the causal language semantics located in the orthogonal negative eigenspace.

**Theorem 1 (Delete nuisance bias, keep content gain)** *By extracting the principal components of the positive spectrum of $\Delta\Sigma$, $U_{bias}$ spans exactly $\mathcal{S}_C$. The causal feature $F_{causal}$ is obtained by subtracting the collinear nuisance projection from the original features:*

$$F_{causal} = F + \Delta_O = F - \beta(FU_{bias})U_{bias}^\top, \tag{9}$$

*where $\beta \in (0, 1]$ controls the intervention strength.*

Please refer to the Appendix A.1 for methodological proofs. Since $U_{bias}$ is orthogonal to $\mathcal{S}_O$, the components in the assumed complementary subspace are approximately preserved ($F_{causal} \perp \mathcal{S}_C$). In practice, we intercept the computation of the Key and Value states in the final layers of the VLM.

# 4 Experiments

## 4.1 Experimental Setup

We evaluate CofactVLA in both **simulation** and **real-world robot environments**. For simulation, we build our model upon the $\pi_{0.5}$ architecture [17] and evaluate it on the standard LIBERO suite [42], as well as the LIBERO-Plus [10] benchmarks to systematically assess zero-shot generalization and robustness against environmental perturbations. For physical deployments, we employ a 6-DoF AgileX PiPer robotic arm equipped with overhead and wrist Intel RealSense cameras operating at 10 Hz. To strictly test the policy's causal grounding against visual distractors, we design four distinct tasks: *I. remove the cuboid from the blue plate, II. move the tennis ball from the yellow plate to the blue plate, III. put the apple on the yellow plate,* and *IV. pick the red cube into the yellow plate*. For each task, we employ about 100 expert trajectories for training and 100 independent trials for evaluation. Comprehensive details regarding benchmark metrics, training configurations, and hardware setups are provided in Appendixes A.2 and A.3. Note that the results of $\pi_{0.5}$ baseline uses the same initialization, optimization schedule, training steps, and evaluation protocol as CofactVLA.

## 4.2 Main Results on Simulation Benchmarks

As shown in Tables 1-2, we present the quantitative evaluation of CofactVLA against state-of-the-art baselines on the standard LIBERO, the highly challenging LIBERO-Plus benchmarks. Note that the best and second-best results are highlighted in **bold** and underlined, and the methods in Table 2 are trained exclusively on the standard LIBERO dataset to evaluate zero-shot generalization.

**Performance on Standard LIBERO.** As shown in Table 1, CofactVLA achieves a state-of-the-art average success rate of **98.5%**, reaching **100.0%** on the Object suite and **99.0%** on the Spatial suite. This confirms that our dual-path intervention effectively enhances fundamental manipulation capabilities, yielding higher execution precision than the base model $\pi_{0.5}$ (96.9%) and the SOTA models X-VLA (98.1%).

Table 1: **Performance comparison (%)** of different VLA models on the **LIBERO benchmark.**

| Model | Spatial | Object | Goal | Long | Average |
|---|---|---|---|---|---|
| OpenVLA-OFT[18] | 97.6 | 98.4 | 97.9 | 94.5 | 97.1 |
| $\pi_0$[35] | 96.8 | 98.8 | 95.8 | 85.2 | 94.2 |
| $\pi_{0.5}$[17] | 98.8 | 98.2 | **98.0** | 92.4 | 96.9 |
| DreamVLA[43] | 97.5 | 94.0 | 89.5 | 89.5 | 92.6 |
| X-VLA[44] | 98.2 | 98.6 | 97.8 | **97.6** | 98.1 |
| CofactVLA | **99.0** | **100.0** | **98.0** | 97.0 | **98.5** |

Table 2: **Zero-shot performance (%)** of different VLA models on the **LIBERO-Plus benchmark.**

| **Method** | **Camera** | **Robot** | **Language** | **Light** | **Background** | **Noise** | **Layout** | **Total** |
|---|---|---|---|---|---|---|---|---|
| OpenVLA [14] | 0.8 | 3.5 | 23.0 | 8.1 | 34.8 | 15.2 | 28.5 | 15.6 |
| NORA [45] | 2.2 | 37.0 | 65.1 | 45.7 | 58.6 | 12.8 | 62.1 | 39.0 |
| WorldVLA [46] | 0.1 | 27.9 | 41.6 | 43.7 | 17.1 | 10.9 | 38.0 | 25.0 |
| UniVLA [47] | 1.8 | 46.2 | 69.6 | 69.0 | 81.0 | 21.2 | 31.9 | 42.9 |
| $\pi_0$ [35] | 13.8 | 6.0 | 58.8 | 85.0 | 81.4 | 79.0 | 68.9 | 53.6 |
| $\pi_0$-Fast [48] | **65.1** | 21.6 | 61.0 | 73.2 | 73.2 | 74.4 | 68.8 | 61.6 |
| OpenVLA-OFT_w [18] | 10.4 | 38.7 | 70.5 | 76.8 | **93.6** | 49.9 | 69.9 | 55.8 |
| Openvla-OFT_m [18] | 55.6 | 21.7 | **81.0** | **92.7** | 91.0 | **78.6** | 68.7 | 67.9 |
| CofactVLA | 44.7 | **49.7** | 71.8 | 85.6 | 83.6 | 78.0 | **70.2** | **69.1** |

**OOD Generalization and Robustness.** We evaluate out-of-distribution capabilities on the LIBERO-Plus benchmark (Table 2), which introduces seven distinct visual and environmental perturbations. CofactVLA demonstrates exceptional robustness against these distribution shifts, achieving the highest total success rate of **69.1%** and significantly outperforming the foundational $\pi_0$ baseline (53.6%). CofactVLA performs particularly well under novel robotic embodiments (**Robot, 49.7%**) and unseen spatial configurations (**Layout, 70.2%**), scenarios where standard models often overfit to familiar visual cues and fail. Furthermore, it maintains highly competitive performance across lighting, background, and noise variations. Overall, these results indicate that explicitly isolating visual biases practically improves the model's ability to seamlessly generalize to novel environments and dynamic physical perturbations.

### 4.3 Main Results on Real-World Robot

As illustrated in Figure 3, we rigorously evaluate the policies in both standard and out-of-distribution (OOD) physical setups. In the standard environments (Figure 3a), CofactVLA achieves a superior average success rate of **90.8%**, outperforming the baseline $\pi_{0.5}$ (**71.0%**). The performance discrepancy across specific tasks strongly highlights the necessity of our deconfounding mechanisms against the "vision-override" phenomenon. For instance, in Task IV—which involves complex visual distractors—$\pi_{0.5}$ struggles significantly, scoring only **31%**. In contrast, CofactVLA maintains precise semantic grounding to achieve **88%**.

The vulnerability of the baseline is further exposed under environmental distribution shifts (Figure 3b). While $\pi_{0.5}$ perfectly solves straightforward tasks (e.g., Task II, **100%**) in familiar settings, it completely collapses (**0%**) when transferred to unseen OOD scenarios. Its overall average plummets to a mere **23.5%**, highlighting a severe reliance on memorized visual shortcuts. Conversely, CofactVLA effectively prevents overfitting to these spurious confounders, demonstrating remarkable zero-shot resilience. It achieves an average success rate of **75.8%** under OOD conditions, marking a massive **+52.3%** absolute improvement over $\pi_{0.5}$ (Please refer to Figures 6-7 and Appendix A.4 for qualitative details). These real-world experiments explicitly confirm that successfully decoupling visual biases inherently bridges the generalization gap, empowering the robot to execute safe, language-grounded manipulation in complex physical environments.

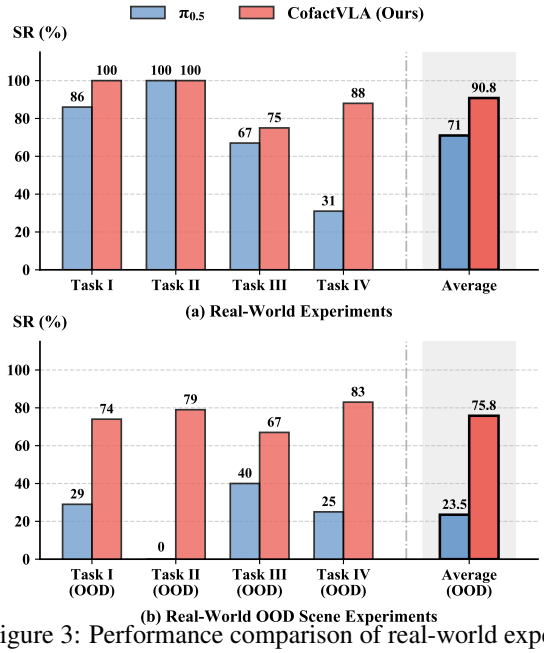


Figure 3: Performance comparison of real-world experiments and the evaluation of OOD scene.

### 4.4 Ablation Studies

We further conduct additional ablation studies on the four LIBERO suites to thoroughly validate the effectiveness of our CofactVLA model, encompassing both the model architecture and the different counterfactual designs.

Table 3: Ablation studies on **model architecture, counterfactual designs on the LIBERO benchmark.**

| Variants | Spatial | Object | Goal | Long | Avg |
|---|---|---|---|---|---|
| **Model Architecture:** | | | | | |
| Baseline | 97.0 | 99.0 | 96.0 | 96.0 | 97.0 |
| w/ CCR | 99.0 | 99.0 | 98.0 | 94.0 | 97.5 |
| w/ OPG | 100.0 | 99.0 | 98.0 | 95.0 | 98.0 |
| Full CofactVLA | 99.0 | 100.0 | 98.0 | 97.0 | 98.5 |
| **Counterfactual design:** | | | | | |
| Add | 98.0 | 98.0 | 92.0 | 88.0 | 94.0 |
| Sub | 97.0 | 99.0 | 98.0 | 90.0 | 96.0 |
| CAG[13] | 98.0 | 96.0 | 99.0 | 97.0 | 97.5 |
| OPG (Ours) | 99.0 | 100.0 | 98.0 | 97.0 | 98.5 |

**Model Architecture.** As shown in Table 3, applying either Feature-Level CCR or Action-Level OPG individually improves the baseline average from 97.0% to 97.5% and 98.0%, respectively, validating their standalone effectiveness. Crucially, combining them yields the highest overall performance (98.5%). This confirms our dual-path design is highly synergistic: CCR explicitly cleanses the latent representation to provide a structurally pure semantic intent, which in turn allows OPG to perform a much cleaner and precise geometric projection during action generation.

**Superiority of Orthogonal Projection.** We compare OPG against traditional scalar-based interventions: naive Addition (94.0%), Subtraction (96.0%), and CAG [13](97.5%). As theoretically argued, direct scalar subtraction (as in CAG) blindly amplifies non-orthogonal noise, risking out-of-distribution (OOD) action flows. In contrast, our OPG mathematically isolates and projects away only the collinear visual bias. This ensures the semantic guidance strictly enhances the causal intent without distorting the valid robotic action manifold, achieving the best performance (98.5%).

### 4.5 Sensitivity Analyses of Intervention Intensity

We investigate the model's sensitivity to the action-level causal scale $\gamma$ and the feature-level intervention strength $\beta$, as illustrated in Figure 4.

**Action-Level Causal Scale ($\gamma$):** Sweeping $\gamma \in \{0, 0.5, 1, 2, 3\}$ reveals that increasing the guidance scale yields substantial improvements on complex tasks, with Libero-Spatial and Libero-Long peaking at 99% and 97%, respectively. Meanwhile, the performance on Libero-Object and Libero-Goal remains consistently high (98%–100%). However, excessive guidance ($\gamma = 3$) causes minor performance drops. This aligns with continuous generative modeling theory[49, 50]: overly aggressive projection could distort the valid action manifold, causing unnatural trajectories. Thus, we set $\gamma = 2$ for maximizing causal deconfounding and ensuring physical execution safety.

**Feature-Level Intervention Strength ($\beta$):** Similarly, sweeping $\beta$ across $\{0, 0.05, 0.1, 0.15, 0.2\}$ demonstrates a clear and consistent performance peak at $\beta = 0.15$, achieving the highest average success rate of 98.5%. At this specific threshold, tasks requiring strict semantic grounding and long-horizon planning (i.e., Libero-Goal and Libero-Long) hit their maximum success rates of 98% and 97%, respectively. Conversely, applying an overly aggressive covariance penalty ($\beta = 0.2$) leads to noticeable performance degradation. This indicates that while moderate penalization successfully suppresses visual confounders, excessive feature-level intervention risks inadvertently washing out the essential, language-aligned visual features required for accurate manipulation.

**Impact of real-world causal scale.** The default $\gamma = 2$ is selected in simulation and transferred directly to all physical-robot tasks without task-specific or environment-specific tuning. To evaluate deployment sensitivity, we additionally sweep $\gamma$ on the four real-world tasks. Table 4 shows a broad effective range: $\gamma \in [0.5, 2]$ yields average success rates from 90.0% to 94.0%, consistently above the 82.5% result without OPG. The real-world optimum is $\gamma = 1$, rather than the simulation-selected default, which indicates that the reported default result is not obtained by post-hoc tuning on the physical platform.

Table 4: Real-world success rate (%) under different action-level causal scales. The reported default $\gamma = 2$ is selected in simulation and transferred unchanged.

| $\gamma$ | Task I | Task II | Task III | Task IV | Average |
|---|---|---|---|---|---|
| 0 (w/o OPG) | 86 | 78 | 83 | 83 | 82.5 |
| 0.5 | **100** | 86 | 86 | 88 | 90.0 |
| 1.0 | **100** | 88 | 88 | **100** | **94.0** |
| 2.0 (Default) | **100** | **100** | 75 | 88 | 90.8 |
| 3.0 | 88 | 86 | **89** | 88 | 87.8 |

### 4.6 Analyses of Computational Overhead

We compare the fine-tuned $\pi_{0.5}$ baseline and full CofactVLA under identical inference settings: batch size 1 on a single NVIDIA A800 GPU, identical observation inputs and numerical precision, the same number of flow-matching steps, and a 50-step action chunk. We report 50 timed runs after 5 warm-up runs. Latency is the wall-clock time required to generate one complete action chunk and excludes environment I/O and robot communication. Peak allocated memory is measured with `torch.cuda.max_memory_allocated()`.

The additional latency primarily arises from evaluating the language-masked counterfactual branch. The factual and counterfactual branches share parameters and are evaluated jointly, so the overhead remains below a naive twofold increase. The covariance estimate and eigendecomposition used by CCR are computed once during prefix-cache construction rather than at every flow-matching step or individual action. Thus, the reported overhead is incurred once per generated 50-step chunk.

Table 5: Inference cost for one 50-step action chunk.

| Method | Latency /chunk(ms)↓ | Peak GPU memory (GB)↓ |
|---|---|---|
| Fine-tuned $\pi_{0.5}$ | $234.94 \pm 3.79$ | 8.87 |
| CofactVLA | $323.90 \pm 3.51$ | 8.98 |
| **Overhead** | **+88.96 (+37.86%)** | **+0.11 (+1.2%)** |

### 4.7 Qualitative Results and Analyses

We present several qualitative visualizations in both simulation and real-world environments. Please refer to the Appendix A.6 for more comprehensive qualitative results and analyses.

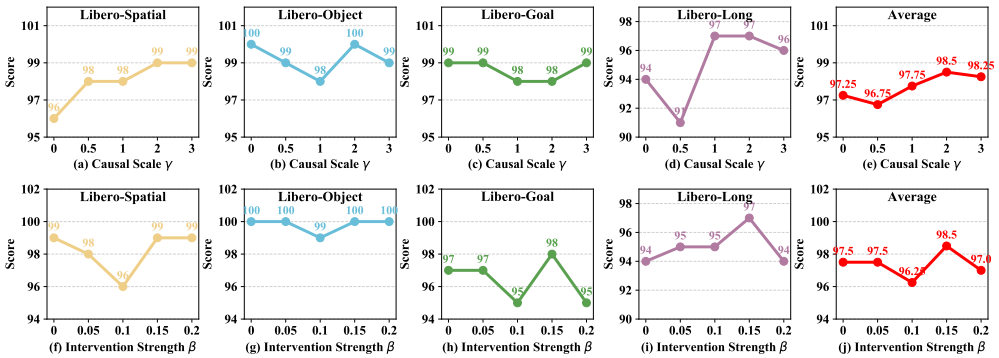



Figure 4: **Sensitivity analyses of intervention intensities.** Top row: action-level causal scale $\gamma$. Bottom row: feature-level intervention strength $\beta$.

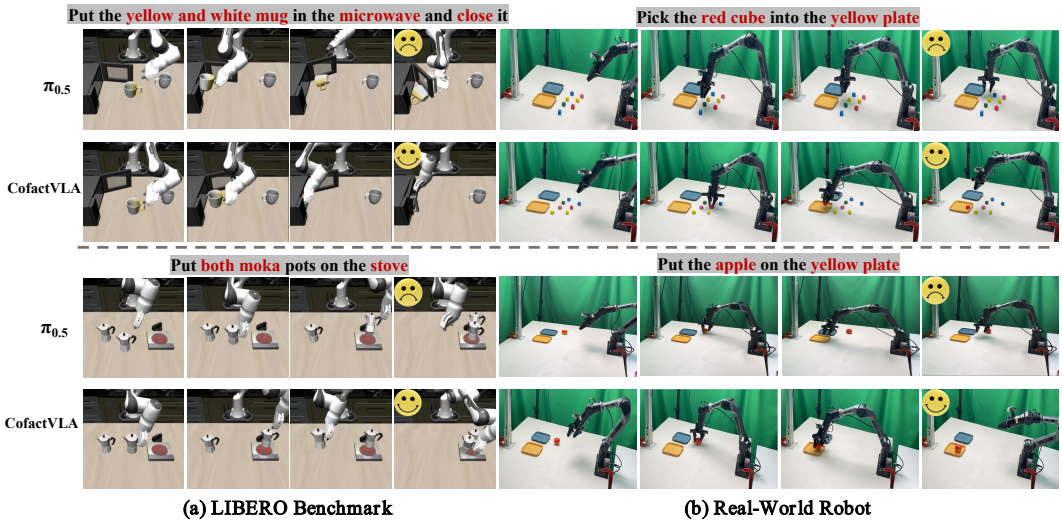



Figure 5: Qualitative comparison of mitigating the "vision-override" phenomenon on the **(a) LIBERO Benchmark and the (b) Real-World Robot.**

**Mitigating Causal Confusion.** As shown in Figure 5, the $\pi_{0.5}$ struggles with precise semantic grounding in visually dense scenes. For instance, $\pi_{0.5}$ fails to close the microwave door in the first example of LIBERO, indicating an incomplete understanding of long-horizon linguistic intents. Similarly, in the physical real-world task, $\pi_{0.5}$ fails to accurately locate the target among multiple colorful distractors. By suppressing dominant visual shortcuts, CofactVLA ensures that the generated trajectories are robust, complete, and driven by the causal language instructions.

**Generalization to Unseen OOD Environments.** Figure 6 highlights our method's OOD robustness. While both models succeed on standard backgrounds (left), the baseline $\pi_{0.5}$ collapses under an unseen checkered texture (right). Conversely, CofactVLA successfully neutralizes this spurious confounder, maintaining accurate spatial localization. Furthermore, as illustrated in Figure 7, our method extends this zero-shot generalization to object-level and semantic distribution shifts. Even when confronted with novel objects or instruction perturbations, CofactVLA robustly grounds the core semantic intent without overfitting to memorized visual attributes. These results confirm that isolating visual bias directly bridges the generalization gap in diverse and dynamic physical deployments.

## 5 Conclusion and Limitations

**Conclusion.** In this paper, we address the "vision-override" phenomenon in VLA models with CofactVLA, a unified causal deconfounding framework based on a Dual-path Deconfounding Graph (DDG). By simultaneously applying Counterfactual Covariance Reduction (CCR) to latent features and Orthogonal Projection Guidance (OPG) to continuous actions, CofactVLA ensures that the

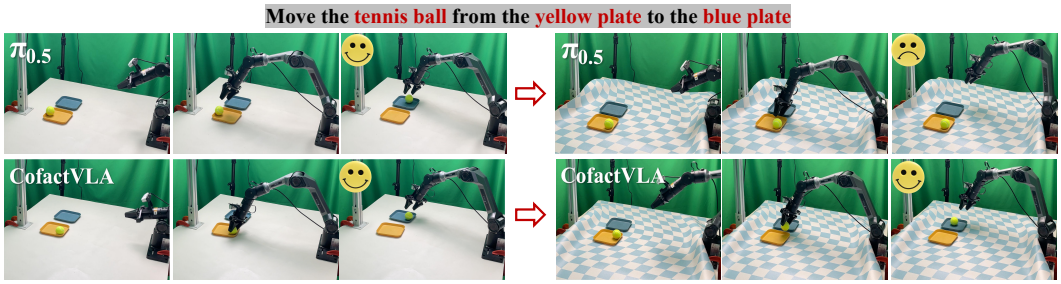



Figure 6: **Qualitative comparison of OOD generalization.** While $\pi_{0.5}$ succeeds on the standard background (left), it fails completely on the unseen checkered texture (right). In contrast, CofactVLA robustly completes the task by decoupling visual confounders.

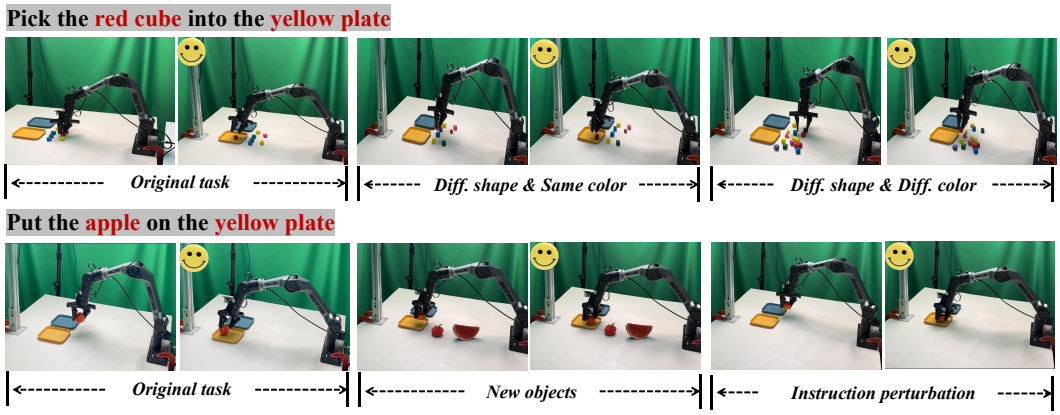



Figure 7: **OOD generalization of CofactVLA:** Visualizations from real-world experiments.

generated manipulation trajectories are strictly and consistently grounded in the causal semantic intents of the language instructions. Extensive simulated and real-world experiments demonstrate state-of-the-art performance, highlighting its exceptional robustness and generalization against dynamic visual distractors in open-world out-of-distribution (OOD) scenarios.

**Limitations.** CofactVLA is inherently bottlenecked by the base VLM's zero-shot grounding, as causal interventions cannot synthesize unlearned concepts. Additionally, it remains vulnerable to severe physical occlusions (e.g., the arm obstructing the camera). Future work will explore dynamic multi-view fusion to enhance occlusion-aware robustness.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants 92570204, 62441235, 62602354, and 62601484) and the Beijing Natural Science Foundation (Grant L257005).

**Competing Interests.** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] 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] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R Sanketi, Grecia Salazar, Michael S Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J Joshi, Alex Irpan, brian ichter,

Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *7th Annual Conference on Robot Learning*, 2023.

[3] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[6] Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19752–19763, 2025.

[7] 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. In *The Thirteenth International Conference on Learning Representations*, 2025.

[8] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, Zhizheng Zhang, and He Wang. GraspVLA: a grasping foundation model pre-trained on billion-scale synthetic action data. In *Proceedings of The 9th Conference on Robot Learning*, pages 1004–1029. PMLR, 2025.

[9] Lucy Xiaoyang Shi, brian ichter, Michael Robert Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, Adrian Li-Bell, Danny Driess, Lachy Groom, Sergey Levine, and Chelsea Finn. Hi Robot: Open-ended instruction following with hierarchical vision-language-action models. In *Forty-second International Conference on Machine Learning*, 2025.

[10] 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.

[11] Catherine Glossop, William Chen, Arjun Bhorkar, Dhruv Shah, and Sergey Levine. CAST: Counterfactual labels improve instruction following in vision-language-action models. *IEEE Robotics and Automation Letters*, 2026.

[12] Zhenghao Peng, Wenhao Ding, Yurong You, Yuxiao Chen, Wenjie Luo, Thomas Tian, Yulong Cao, Apoorva Sharma, Danfei Xu, Boris Ivanovic, Boyi Li, Yan Wang, and Marco Pavone. Counterfactual VLA: Self-reflective vision-language-action model with adaptive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4022–4031, June 2026.

[13] Yu Fang, Yuchun Feng, Dong Jing, Jiaqi Liu, Yue Yang, Zhenyu Wei, Daniel Szafir, and Mingyu Ding. When vision overrides language: Evaluating and mitigating counterfactual failures in vlas. *arXiv preprint arXiv:2602.17659*, 2026.

[14] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[15] Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Yang Zhaohui, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, Dehui Wang, Dingxiang Luo, Yuchen Fan, Youbang Sun, Jia Zeng, Jiangmiao Pang, Shanghang Zhang, Yu Wang, Yao Mu, Bowen Zhou, and Ning Ding. SimpleVLA-RL: Scaling VLA training via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[16] 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.

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

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

[19] Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, et al. VLA-Adapter: An effective paradigm for tiny-scale vision-language-action model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18638–18646, 2026.

[20] Yan Zhang, Zhong Ji, Yanwei Pang, Jungong Han, and Xuelong Li. Modality-experts coordinated adaptation for large multimodal models. *Science China Information Sciences*, 67(12):220107, 2024.

[21] Xudong Wang, Zebin Han, Zhiyu Liu, Gan Li, Jiahua Dong, Baichen Liu, Lianqing Liu, and Zhi Han. Lifelong language-conditioned robotic manipulation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18629–18637, 2026.

[22] Yufei Guo, Yinan Wu, Haoran Duan, Guiguang Ding, and Jungong Han. PhaseLoRA: Control-regime-conditioned low-rank adaptation for continuous-action vision-language-action policies. *arXiv preprint arXiv:2608.15285*, 2026.

[23] Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. ReconVLA: Reconstructive vision-language-action model as effective robot perceiver. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18549–18557, 2026.

[24] Danny Driess, Jost Tobias Springenberg, brian ichter, LILI YU, Adrian Li-Bell, Karl Pertsch, Allen Z. Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, and Sergey Levine. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[25] Rongyu Zhang, Menghang Dong, Yuan Zhang, Liang Heng, Xiaowei Chi, Gaole Dai, Li Du, Dan Wang, Yuan Du, and Shanghang Zhang. MoLe-VLA: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18764–18772, 2026.

[26] Songsheng Wang, Rucheng Yu, Zhihang Yuan, Chao Yu, Feng Gao, Yu Wang, and Derek F Wong. Spec-VLA: speculative decoding for vision-language-action models with relaxed acceptance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26916–26928, 2025.

[27] Wei Li, Renshan Zhang, Rui Shao, Jie He, and Liqiang Nie. CogVLA: Cognition-aligned vision-language-action models via instruction-driven routing & sparsification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020.

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

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[31] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[32] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

[33] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: a diffusion foundation model for bimanual manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[34] Zhendong Wang, Max Li, Ajay Mandlekar, Zhenjia Xu, Jiaojiao Fan, Yashraj Narang, Linxi Fan, Yuke Zhu, Yogesh Balaji, Mingyuan Zhou, Ming-Yu Liu, and Yu Zeng. One-step diffusion policy: Fast visuomotor policies via diffusion distillation. In *Forty-second International Conference on Machine Learning*, 2025.

[35] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi_0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

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

[37] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[38] Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024.

[39] Yan Zhang, Zhijin Qin, Feng Xu, Guiguang Ding, and Jungong Han. CofactGVR: Counterfactual intervention for grounded visual reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

[40] Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wuthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. In *International Conference on Learning Representations*, 2021.

[41] Jiahua Ma, Yiran Qin, Yixiong Li, Xuanqi Liao, Yulan Guo, and Ruimao Zhang. CDP: Towards robust autoregressive visuomotor policy learning via causal diffusion. In *Proceedings of the 9th Conference on Robot Learning (CoRL)*, pages 4190–4205. PMLR, 2025.

[42] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, pages 44776–44791, 2023.

[43] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, XinQiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[44] Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Tai Wang, Ya-Qin Zhang, Jingjing Liu, and Xianyuan Zhan. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026.

[45] Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U Tan, Navonil Majumder, Soujanya Poria, et al. NORA: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025.

[46] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. WorldVLA: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

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

[48] 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.

[49] Jinhao Liang, Yixuan Sun, Anirban Samaddar, Sandeep Madireddy, and Ferdinando Fioretto. Chance-constrained flow matching for high-fidelity constraint-aware generation. *arXiv preprint arXiv:2509.25157*, 2025.

[50] Jean-Baptiste Bouvier, Kanghyun Ryu, Kartik Nagpal, Qiayuan Liao, Koushil Sreenath, and Negar Mehr. DDAT: Diffusion policies enforcing dynamically admissible robot trajectories. *arXiv preprint arXiv:2502.15043*, 2025.

[51] Remi Cadene, Simon Alibert, Francesco Capuano, Michel Aractingi, Adil Zouitine, Pepijn Kooijmans, Jade Choghari, Martino Russi, Caroline Pascal, Steven Palma, Dana Aubakirova, Mustafa Shukor, Jess Moss, Alexander Soare, Quentin Lhoest, Quentin Gallouédec, and Thomas Wolf. LeRobot: An open-source library for end-to-end robot learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[52] 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.

# A Technical appendices and supplementary material

In this appendix, we further provide the following additional details, which are omitted in the main paper owing to the limited space.

The structure of this Appendix is as follows:

- Section A.1 contains **additional methodological proofs**.
- Section A.2 contains **introduction of our real-world robot platform**.
- Section A.3 contains **training configuration and implementation details**.
- Section A.4 contains **details of perturbation experiment design**.
- Section A.5 contains **additional real-world robot experiments**.
- Section A.6 contains **comprehensive qualitative results and analyses**.
- Section A.7 contains **limitations and outlooks**.
- Section A.8 contains **broader impacts**.

## A.1 Additional Methodological Proofs

### A.1.1 Preliminaries: Whitening and Generalized Eigenvectors

Throughout, assume $\Sigma_0 \succ 0$. Define the $\Sigma_0$-whitened matrix

$$M \;:=\; \Sigma_0^{-1/2}\,\Sigma_\Delta\,\Sigma_0^{-1/2}, \tag{10}$$

which is symmetric positive semidefinite. A pair $(\lambda, b)$ solves the generalized eigenvalue problem

$$\Sigma_\Delta b \;=\; \lambda \Sigma_0 b. \tag{11}$$

If and only if $(\lambda, u)$ solves the standard eigenproblem

$$Mu \;=\; \lambda u, \qquad u := \Sigma_0^{1/2} b. \tag{12}$$

Indeed, multiplying (11) on the left by $\Sigma_0^{-1/2}$ and setting $u = \Sigma_0^{1/2} b$ gives (12); conversely, $b = \Sigma_0^{-1/2} u$ maps eigenvectors of $M$ to generalized eigenvectors of $(\Sigma_\Delta, \Sigma_0)$.

### A.1.2 Closed-form Solution of the Generalized Eigenproblem

**Proposition 1** *Consider the optimization problem:* $\max_{B \in \mathbb{R}^{d_r \times d_v}} tr(B^\top \Sigma_\Delta B) \quad$ *subject to* $\quad B^\top \Sigma_0 B = I_{d_v}$. *Any optimizer $B^\star$ for this problem spans the top $d_v$ generalized eigenspace of $(\Sigma_\Delta, \Sigma_0)$:*

$$\Sigma_\Delta b_i \;=\; \lambda_i \Sigma_0 b_i, \qquad \lambda_1 \geq \cdots \geq \lambda_{d_r} \geq 0, \qquad B^\star = [b_1, \ldots, b_{d_v}]. \tag{13}$$

**Proof 1** *Let $C := \Sigma_0^{1/2} B \in \mathbb{R}^{d_r \times d_v}$. Then the constraint becomes*

$$B^\top \Sigma_0 B = I_{d_v} \quad \Longleftrightarrow \quad (\Sigma_0^{1/2} B)^\top (\Sigma_0^{1/2} B) = C^\top C = I_{d_v}. \tag{14}$$

*The objective becomes*

$$\mathrm{tr}(B^\top \Sigma_\Delta B) = \mathrm{tr}\big(B^\top \Sigma_0^{1/2} \underbrace{\Sigma_0^{-1/2} \Sigma_\Delta \Sigma_0^{-1/2}}_{M} \Sigma_0^{1/2} B\big) = \mathrm{tr}(C^\top M C). \tag{15}$$

*Let $M = Q \Lambda Q^\top$ be an eigendecomposition with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d_r} \geq 0$ and orthonormal eigenvectors $Q = [q_1, \ldots, q_{d_r}]$. By the Ky Fan maximum principle (a trace form of Rayleigh–Ritz), the maximum attained by*

$$C^\star = [q_1, \ldots, q_{d_v}]\, U, \tag{16}$$

*for any orthogonal $U \in \mathbb{R}^{d_v \times d_v}$ (i.e., any orthonormal basis of the top-$d_v$ eigenspace). Mapping back yields*

$$B^\star = \Sigma_0^{-1/2} C^\star. \tag{17}$$

*Each column $b_i = \Sigma_0^{-1/2} q_i$ satisfies the generalized eigen-equation:*

$$\Sigma_\Delta b_i = \Sigma_\Delta \Sigma_0^{-1/2} q_i = \Sigma_0^{1/2} (\Sigma_0^{-1/2} \Sigma_\Delta \Sigma_0^{-1/2}) q_i = \Sigma_0^{1/2} M q_i = \lambda_i\, \Sigma_0^{1/2} q_i = \lambda_i\, \Sigma_0 b_i. \tag{18}$$

*Hence any optimizer spans the top $d_v$ generalized eigenspace of $(\Sigma_\Delta, \Sigma_0)$, proving* (13)*.*

### A.1.3 $\Pi_V$ is the $\Sigma_0$-orthogonal Projector

**Lemma 1** *Let $B \in \mathbb{R}^{d_r \times d_v}$ have full column rank and define*

$$\Pi_V := B(B^\top \Sigma_0 B)^{-1} B^\top \Sigma_0. \tag{19}$$

*Then (i) $\Pi_V^2 = \Pi_V$ (idempotent), (ii) $\mathrm{range}(\Pi_V) = \mathrm{span}(B)$, and (iii) $\Pi_V$ is self-adjoint w.r.t. the $\Sigma_0$-inner product:*

$$\langle u, \Pi_V v\rangle_{\Sigma_0} = \langle \Pi_V u, v\rangle_{\Sigma_0}, \qquad \langle a, b\rangle_{\Sigma_0} := a^\top \Sigma_0 b. \tag{20}$$

*Consequently, $\Pi_V$ is the $\Sigma_0$-orthogonal projector onto $\mathrm{span}(B)$ and $\Pi_\perp := I - \Pi_V$ projects onto the $\Sigma_0$-orthogonal complement.*

**Proof 2** *Let $G := B^\top \Sigma_0 B \in \mathbb{R}^{d_v \times d_v}$; $G$ is invertible since $\Sigma_0 \succ 0$ and $B$ has full column rank. Then*

$$\Pi_V^2 = BG^{-1}B^\top \Sigma_0\, BG^{-1}B^\top \Sigma_0 = BG^{-1}\underbrace{(B^\top \Sigma_0 B)}_{G} G^{-1} B^\top \Sigma_0 = BG^{-1}B^\top \Sigma_0 = \Pi_V. \tag{21}$$

*Thus $\Pi_V$ is idempotent. Also, for any vector $z$, $\Pi_V z \in \mathrm{span}(B)$ by construction, so $\mathrm{range}(\Pi_V) \subseteq \mathrm{span}(B)$. Conversely, for any $y = Bc$,*

$$\Pi_V(Bc) = BG^{-1}B^\top \Sigma_0 Bc = BG^{-1}Gc = Bc = y, \tag{22}$$

*so $\mathrm{span}(B) \subseteq \mathrm{range}(\Pi_V)$. Hence $\mathrm{range}(\Pi_V) = \mathrm{span}(B)$.*

*For self-adjointness, note $\Pi_V^\top \Sigma_0 = \Sigma_0 \Pi_V$:*

$$\Pi_V^\top \Sigma_0 = (\Sigma_0^\top BG^{-1}B^\top)\Sigma_0 = \Sigma_0 BG^{-1}B^\top \Sigma_0 = \Sigma_0 \Pi_V, \tag{23}$$

*where we used $\Sigma_0^\top = \Sigma_0$. Therefore for any $u, v$,*

$$\langle u, \Pi_V v\rangle_{\Sigma_0} = u^\top \Sigma_0 \Pi_V v = u^\top \Pi_V^\top \Sigma_0 v = (\Pi_V u)^\top \Sigma_0 v = \langle \Pi_V u, v\rangle_{\Sigma_0}. \tag{24}$$

*This is exactly $\Sigma_0$-orthogonal projection onto $\mathrm{span}(B)$.*

### A.1.4 CCR: Delete Nuisance Bias, Keep Content Gain

We proceed in three steps.

**Step 1: Top generalized eigenspace equals top eigenspace of $M$.** By the whitening equivalence (12), generalized eigenvectors of $(\Sigma_\Delta, \Sigma_0)$ correspond to eigenvectors of $M = \Sigma_0^{-1/2}\Sigma_\Delta \Sigma_0^{-1/2}$ under the invertible map $u = \Sigma_0^{1/2} b$.

**Step 2: Subspace separation implies the top-$d_v$ eigenspace is $\Sigma_0^{1/2}\mathcal{S}_O$.** Let

$$\mathcal{T}_O := \Sigma_0^{1/2}\mathcal{S}_O, \qquad \mathcal{T}_C := \Sigma_0^{1/2}\mathcal{S}_C. \tag{25}$$

Since $\Sigma_0^{1/2}$ is invertible and $\mathbb{R}^{d_r} = \mathcal{S}_C \oplus \mathcal{S}_O$, we also have $\mathbb{R}^{d_r} = \mathcal{T}_O \oplus \mathcal{T}_C$ and $\dim(\mathcal{T}_O) = d_o$.

Assumption 2 states that eigenvalues of $M$ on $\mathcal{T}_C$ are strictly larger than those on $\mathcal{T}_O$:

$$\lambda_{\min}\big(M|_{\mathcal{T}_C}\big) \;>\; \lambda_{\max}\big(M|_{\mathcal{T}_O}\big). \tag{26}$$

This implies a spectral separation: the top $d_c$ (let $dim(\mathcal{T}_C) = d_c$) eigenvalues of $M$ come from the restriction to $\mathcal{T}_c$, and therefore the top-$d_c$ eigenspace of $M$ equals $\mathcal{T}_C$. Concretely, if $U_{bias}$ is any orthonormal basis of $\mathcal{T}_C$, then $U_{bias}$ spans the maximizer and thus spans the top-$d_o$ eigenspace of $M$ (uniquely up to rotations within that eigenspace).

Mapping back via $b = \Sigma_0^{-1/2}u$, the top generalized eigenspace equals $\mathcal{S}_C$. Hence the matrix $B$ formed by the top generalized eigenvectors satisfies $\mathrm{span}(B) = \mathcal{S}_C$.

**Step 3: $\Pi_\perp$ removes exactly the nuisance component of $\widetilde{\Delta}$.** By Lemma 1, let $\Pi_C$ be the $\Sigma_0$-orthogonal projector onto $span(B) = \mathcal{S}_C$, and $\Pi_\perp = I - \Pi_C$ projects onto the $\Sigma_0$-orthogonal complement, which is $\mathcal{S}_O$. Therefore, $\Pi_\perp \tilde{\Delta} = (I - \Pi_C)(\tilde{\Delta}_C + \tilde{\Delta}_O) = \tilde{\Delta}_O$. This isolated $\tilde{\Delta}_O$ allows us to extract the causal semantic features.

Finally, with $\Delta = \Delta_C + \Delta_O$ (the non-token version of the same decomposition),

$$F_{causal} = F + \Delta_O = F - \beta(FU_{bias})U_{bias}^\top, \tag{27}$$

which is exactly (9).

### A.2 Introduction of our Real-world Robot Platform

**Hardware Platform.** As illustrated in Figure 8, our physical experiments are conducted using a 6-DoF AgileX PiPer robotic arm, which provides precise and dexterous manipulation capabilities. The end-effector is a standard 1-DoF parallel gripper. To perceive the environment and provide visual feedback for the continuous flow matching policy, we deploy two Intel RealSense D435 cameras. One camera is mounted overhead to capture a global scene view (*Third Person View*), while the other is attached to the robot's wrist to capture fine-grained, localized spatial details (*Wrist View*) during manipulation. The diverse set of objects used in our experiments, including various fruits, vegetables, blocks, and plates, is shown in Figure 8(b).



**(a) 6-DoF AgileX PiPer Robotic Arm**



**(b) Experiment Objects**




**(c) Hardware Platform**

Figure 8: Our hardware platform for the real-world robot experiments.

### A.3 Training Configuration and Implementation Details

We build CofactVLA upon the Hugging Face `lerobot` [51]framework, where the core architecture is initialized from the pre-trained $\pi_{0.5}$ [17] checkpoint. As summarized in Table 6, all model fine-tuning is conducted on a computing node equipped with 4×NVIDIA H100 (96GB) GPUs. Detailed hyperparameters are summarized in Table 6.

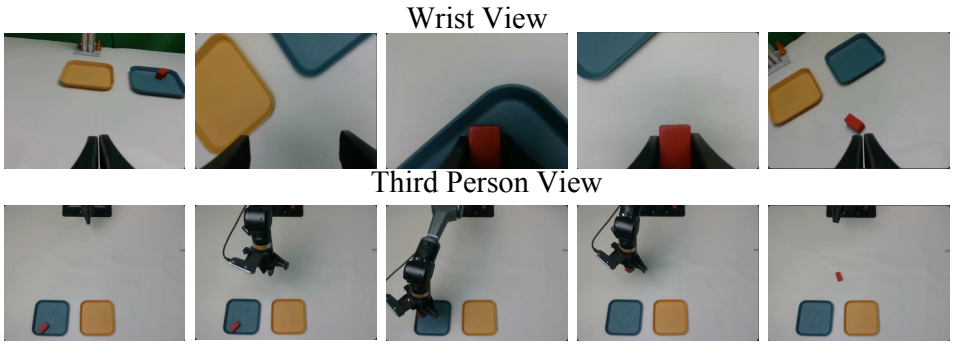
**Datasets and Evaluation Protocols.** For **simulation**, our model is trained exclusively on the standard LIBERO dataset [42], which encompasses four benchmark suites: **LIBERO-Spatial**, **LIBERO-Object**, **LIBERO-Goal**, and **LIBERO-Long** (or LIBERO-10). These suites focus on various spatial

relationships, object categories, goal objectives, and extended sequential challenges, respectively. We evaluate these standard suites over 10 episodes to ensure statistical reliability. To assess zero-shot robustness, we directly deploy the trained policy on the LIBERO-Plus [10] benchmarks. Due to the extensive scale of LIBERO-Plus, its tasks are evaluated for a single episode. For **real-world deployments**, we collect a custom dataset comprising approximately 400 expert trajectories (about 100 per task). Real-world evaluations report the average success rate over 100 independent trials per task. We provide several trajectories in Figures 9-10.

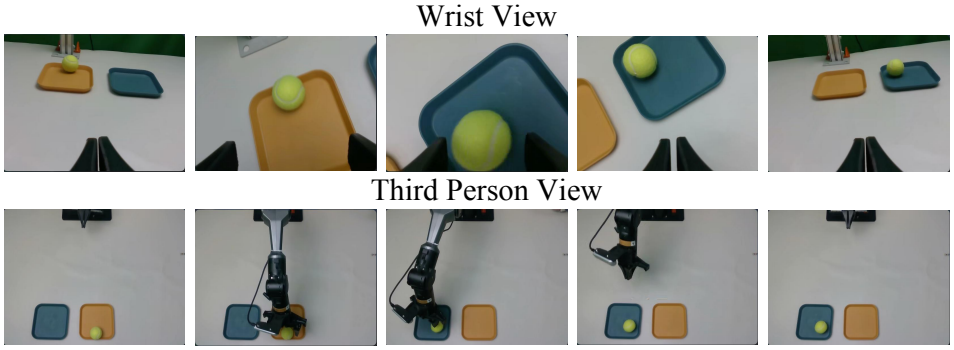
**Asset Licenses.** Our implementation relies on several open-source assets. The LIBERO benchmark suites are licensed under the MIT License. The `lerobot`[51] framework and the pre-trained $\pi_{0.5}$ checkpoints are distributed under the Apache License 2.0. Our usage of these datasets and models strictly adheres to their respective terms of use.

Table 6: CofactVLA Training Configuration.

| **Hyperparameters** | **Value** | **Hyperparameters** | **Value** |
|---|---|---|---|
| Compute resources | 4×H100-96GB | Causal scale $\gamma$ | 2.0 |
| Backbone | $\pi_{0.5}$[17] | Intervention strength $\beta$ | 0.15 |
| Batch size | 32 / GPU | Intervention layers | [15, 16] |
| Learning rate (LR) | 2.5e-5 | Freeze vision encoder | False |
| Training steps | 6K | Freeze action expert | False |
| Weight decay | 0.01 | Gradient checkpointing | True |
| Optimizer | AdamW | Action chunk size | 50 |
| Betas | [0.9, 0.95] | Warm-up steps | 1K |




**(a) Remove the cuboid from the blue plate**

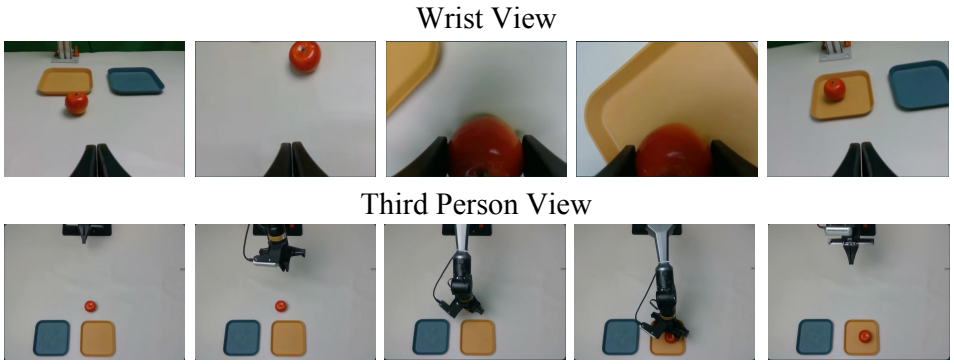



**(b) Move the tennis ball from the yellow plate to the blue plate**

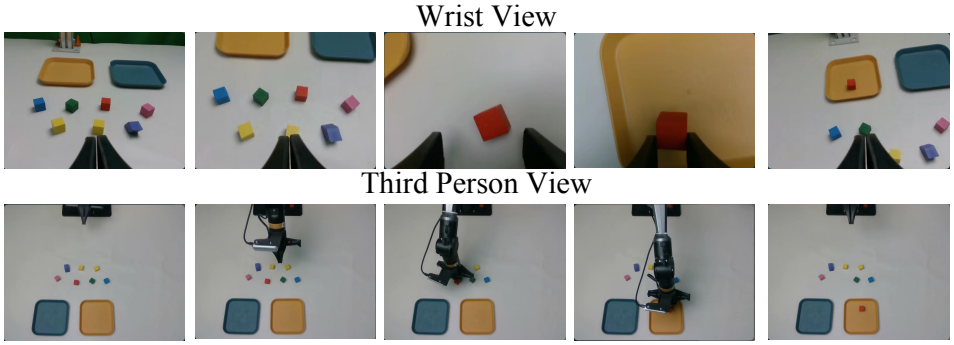
Figure 9: Example of dual-view sample collection for Task I and Task II.

### A.4 Perturbation Experiment Design

Inspired by the task design of LIBERO-Pro [52], we delicately design several comprehensive perturbation experiments to assess anti-perturbation ability of the VLA models beyond the visual layouts and




**(a) Put the apple on the yellow plate**




**(b) Pick the red cube into the yellow plate**

Figure 10: Example of dual-view sample collection for Task III and Task IV.

object configurations. In this section, we introduce perturbations along three dimensions, including *object perturbation, environment perturbations, and instruction perturbations*, to reveal whether the VLA models rely on genuine perception and language grounding or merely memorize task-specific trajectories. Figure 11 provides an intuitive demonstration, and the specific design of perturbation for different tasks are as follows:

- For *Task I: remove the cuboid from the blue plate*, we introduce both *object perturbation* and *environment perturbation*. For object-level perturbation, several objects with different shapes and colors are placed around the target red cuboid, making the target less visually salient and increasing the difficulty of object detection and grounding for the VLA models. For environment-level perturbation, the original white tabletop is covered with a green grass-like surface, introducing a significant background texture shift. This setting evaluates whether the model could still accurately localize the target under foreground distractors or background appearance changes.
- For *Task II: move the tennis ball from the yellow plate to the blue plate*, we further introduce more challenging *environment perturbations* to test the robustness of the learned policy. Specifically, the original plates are replaced with lighter-colored yellow plates, and the tabletop is wholly covered with a blue checkered tablecloth. These modifications change the surroundings of both the source and the target workspace. This perturbation is designed to examine whether the model could maintain reliable manipulation when the environment distribution differs from the training data.
- For *Task III: put the apple on the yellow plate*, we apply both *object perturbation* and *instruction perturbation*. For object-level perturbation, we place visually similar distractors near the target apple, including a red tomato and a piece of watermelon. Since these objects share similar category-level or color-level cues with the red apple, they inevitably interfere with fine-grained object recognition of the VLA models. In addition, we introduce instruction-level perturbation by paraphrasing the original prompt, such as replacing the

expression “put ... on” with “pick ... into”. This setting tests whether the model could jointly handle visual distractors and linguistic variation.

- For *Task IV: pick the red cube into the yellow plate*, we design a more cluttered and fine-grained object perturbation scenario. The original set of seven small wooden blocks is expanded to eight objects, and multiple blocks are randomly scattered across the workspace. For example, we place a red sphere whose appearance is highly similar to the target red cube in color and size but differs in shape. This setting aims to evaluate the ability of correctly grounding the compositional description “red cube” by jointly considering both color and shape, rather than relying only on the most salient color cue.

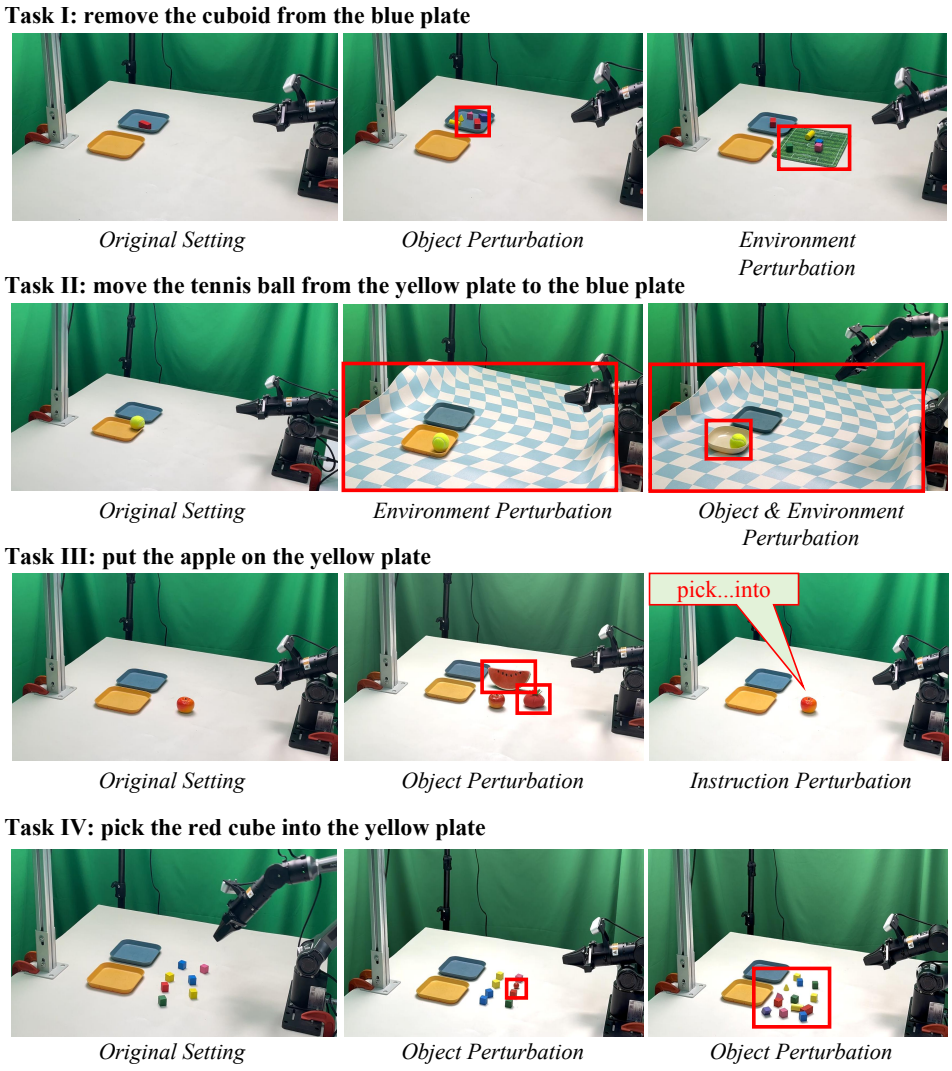



Figure 11: Visualizations of the perturbation design for different tasks.

### A.5 Additional Real-world Robot Experiments

In this section, we provide additional real-world robot experiments to validate the robustness and generalization of our method. As it is shown in Table 7, we evaluate real-robot manipulation on these tasks and compare against OpenVLA[14], $\pi_0$[35], and $\pi_{0.5}$[17]. The tasks span a range of manipulation primitives, including: (i) object placement into receptacles (apple→plate, tennis ball→bowl, eggplant→bowl, tissue→plate), (ii) stacking/placing object onto another surface (bag→plate), (iii) push and pull (opening and closing a drawer), (iv) longer-horizon sequential tasks (open the drawer → pick up the apple, insert a pen into a cup).

Table 7: Task IDs and their instructions.

| Task ID | Instructions |
|---|---|
| A-1 | Pick the apple into plate |
| A-2 | Pick the tennis into bowl |
| A-3 | Place the eggplant into the bowl |
| A-4 | Place the tissue into the plate |
| A-5 | Stack the bag on the plate |
| A-6 | Open the drawer |
| A-7 | Close the drawer |
| A-8 | Open the drawer and pick up the apple |
| A-9 | Pick the pen and insert into cup |

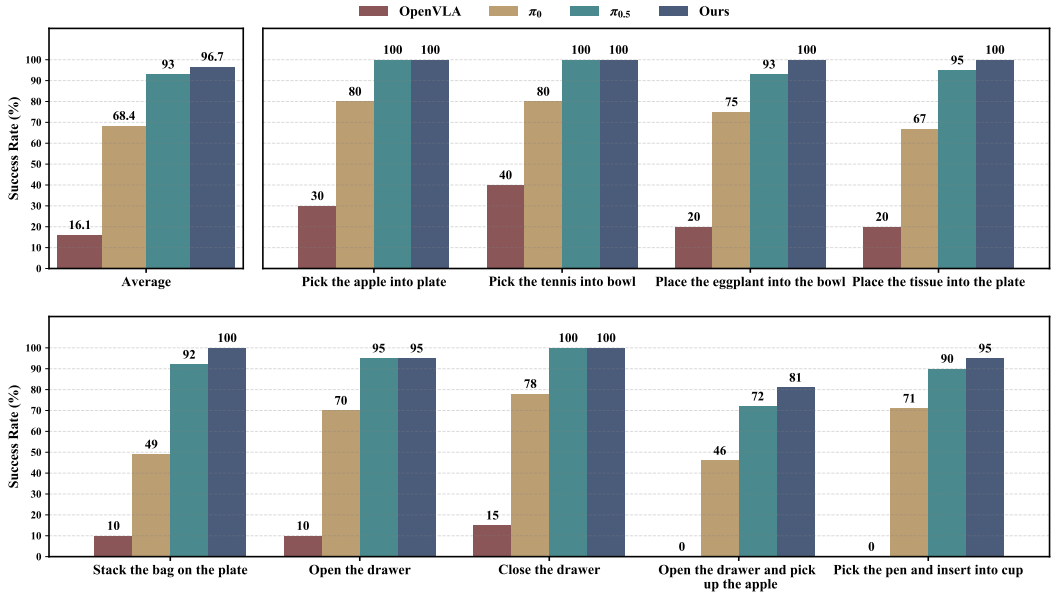



Figure 12: We report success rate (100 trials) for 9 skills (x-axis) and the average over all tasks (left inset), comparing OpenVLA, $\pi_0$, $\pi_{0.5}$, and Ours. Our method achieves the best overall performance (96.7% average) and reaches better performance on most tasks.

All methods are evaluated under the same real-robot setup and identical execution protocol. Figure 12 summarizes success rates, and our method achieves the best average performance (**96.7%**), improving over the strongest baseline $\pi_{0.5}$ (93.0%) by **+3.7** points and clearly outperforming $\pi_0$ (68.4%) and OpenVLA (16.1%). While both our method and $\pi_{0.5}$ saturate on simple pick-and-place tasks (100%), our gains concentrate on different object manipulation (e.g., tissue placement and bag stacking, up to **+8** over $\pi_{0.5}$), as well as long-horizon compositions (e.g., Open drawer + pick and Pen insertion).

### A.6 Comprehensive Qualitative Results and Analyses

To provide a deeper understanding of our proposed CofactVLA framework, we present extensive qualitative visualizations across both simulated benchmarks and real-world physical deployments. These examples explicitly demonstrate our method's superiority in semantic grounding, robustness against distribution shifts, and current operational boundaries.

**Comparisons against Baselines on LIBERO.** Figure 13 provides a detailed step-by-step comparison between the baseline $\pi_{0.5}$[17] and our CofactVLA in the standard LIBERO simulation environment. As illustrated, while $\pi_{0.5}$ frequently suffers from causal confusion by prematurely interacting with visually prominent but task-irrelevant objects, CofactVLA successfully maintains strict alignment with the language instruction. This visual evidence further corroborates that our dual-path deconfounding mechanism effectively suppresses spurious visual shortcuts.

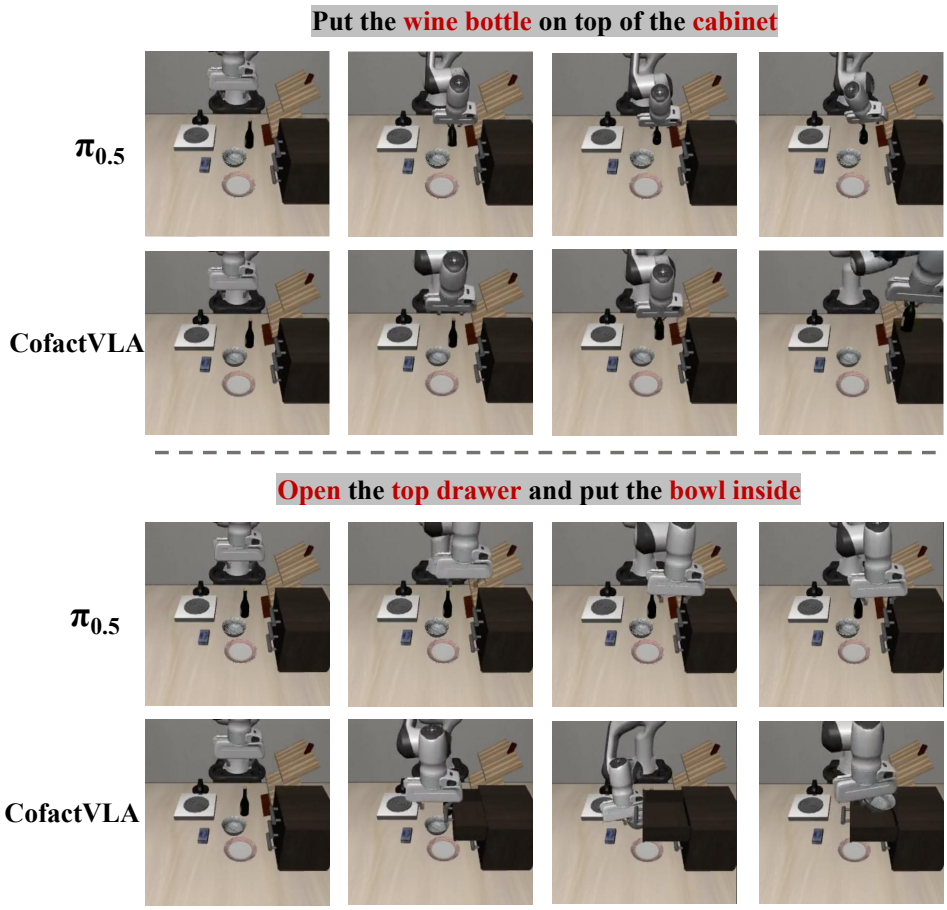



Figure 13: Qualitative comparison between our CofactVLA model and $\pi_{0.5}$ model on the LIBERO Benchmark.

**Robustness Across the LIBERO-Plus Suites.** To validate the out-of-distribution (OOD) robustness of our approach, Figures 14 to 17 showcase successful execution trajectories on the highly challenging LIBERO-Plus dataset. We categorize these visualizations into four distinct suites: *Spatial*, *Object*, *Goal*, and *Long*. Despite the introduction of complex visual perturbations (e.g., altered lighting, novel camera viewpoints, and unseen textures), CofactVLA exhibits remarkable zero-shot generalization. Notably, in the *Long* suite (Figure 17), our policy robustly completes extended sequential tasks without suffering from the compounding errors typically induced by visual confounders.

**Real-World Deployment and Generalization.** Figure 18 extends our qualitative analysis to physical environments, including the standard execution of the original tasks and additional generalization scenarios as introduced in Appendix A.4. For instance, when presented with entirely unseen table textures or dynamically added physical distractors, CofactVLA dynamically neutralizes these visual biases. The robotic arm smoothly reaches the target receptacles without hesitation, bridging the notoriously difficult sim-to-real and environment-to-environment generalization gaps.

**Failure Mode Analysis.** Finally, to provide a transparent and rigorous evaluation of our system, we present representative failure cases from both LIBERO-Plus and real-world experiments in Figure 19 and Figure 20, respectively. In the simulation environment, failures primarily occur when multiple perturbations are overlaid simultaneously (e.g., extreme lighting changes (the 4th instance in Figure 19) combined with drastic camera angle shifts), occasionally causing the vision encoder to lose spatial tracking. In the real-world deployments, execution failures are predominantly observed under severe physical occlusions—such as when the robot's end-effector entirely blocks the camera's field of view (FoV) during precision grasping.

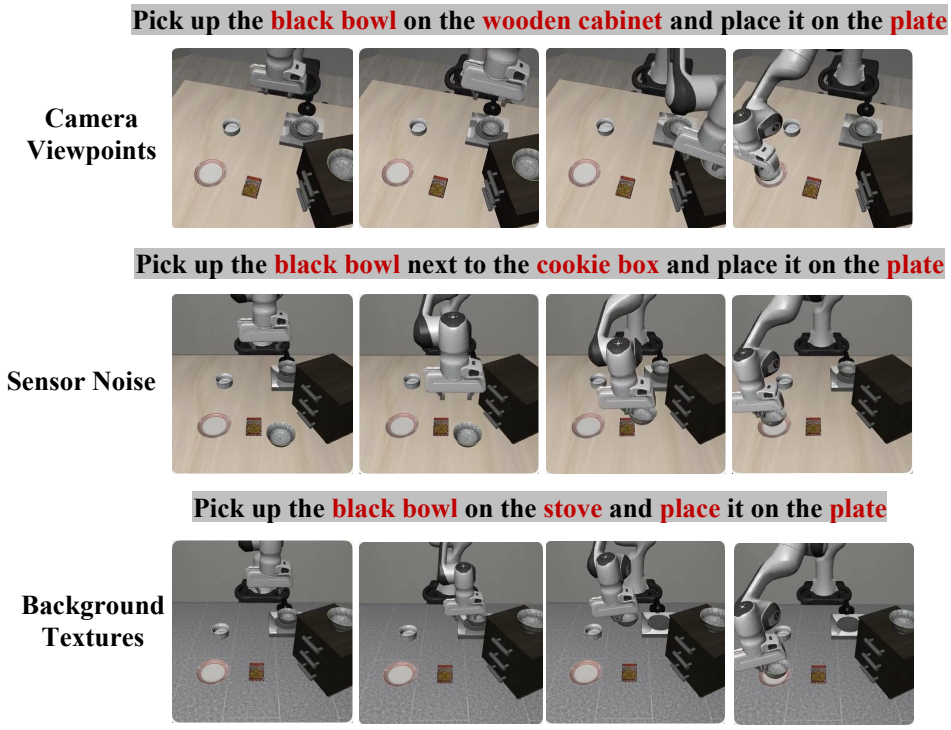



Figure 14: Qualitative results of our CofactVLA model on the **Spatial suite of LIBERO-Plus Benchmark.**

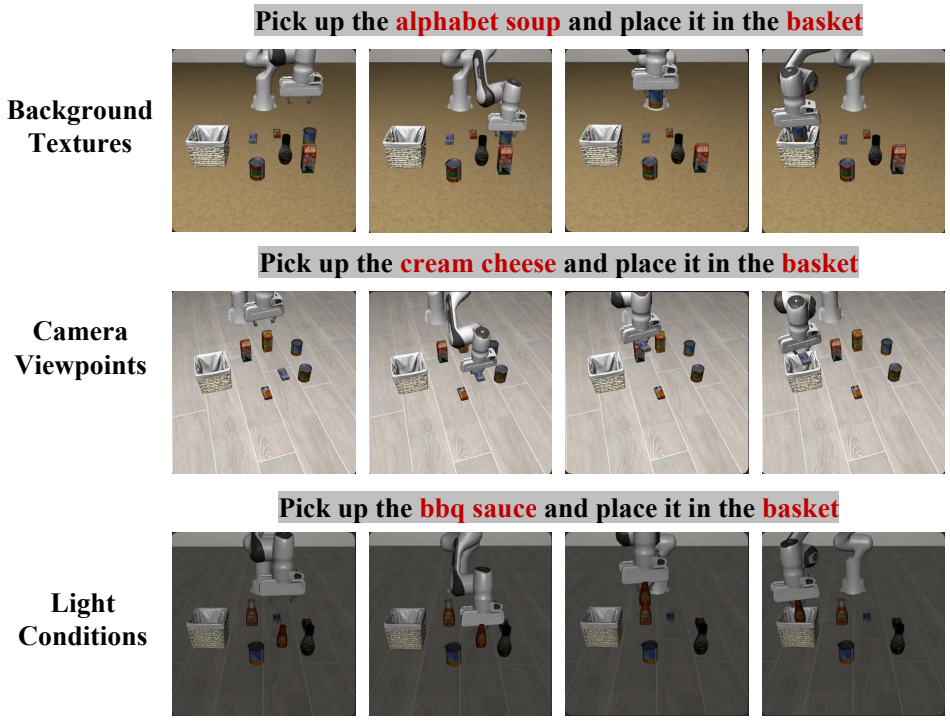



Figure 15: Qualitative results of our CofactVLA model on the **Object suite of LIBERO-Plus Benchmark.**

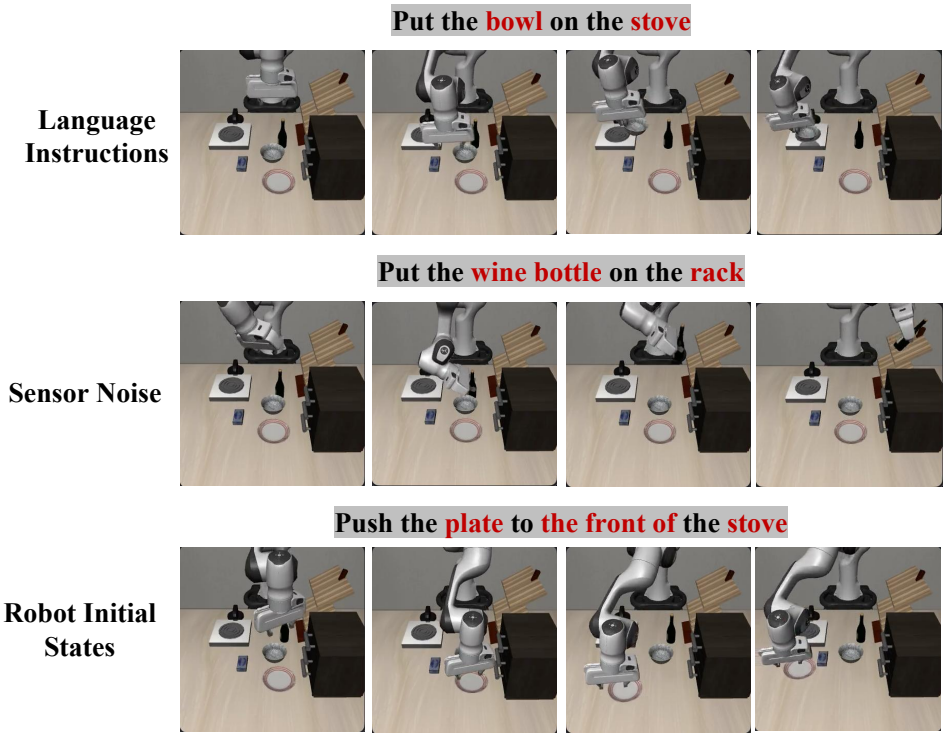



Figure 16: Qualitative results of our CofactVLA model on the **Goal suite of LIBERO-Plus Benchmark.**

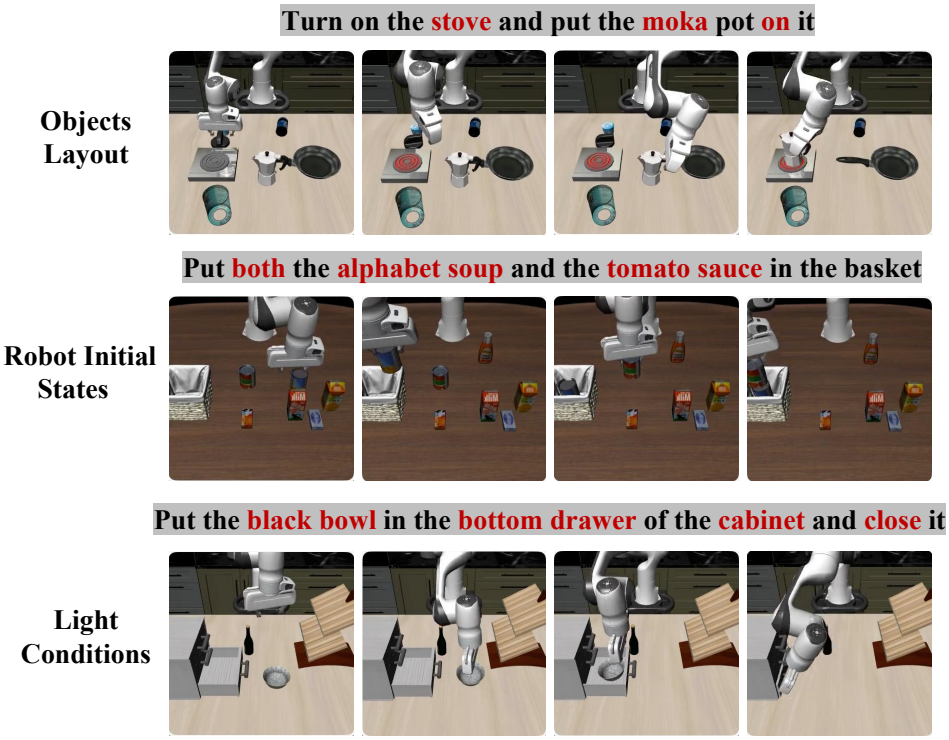



Figure 17: Qualitative results of our CofactVLA model on the **Long suite of LIBERO-Plus Benchmark.**

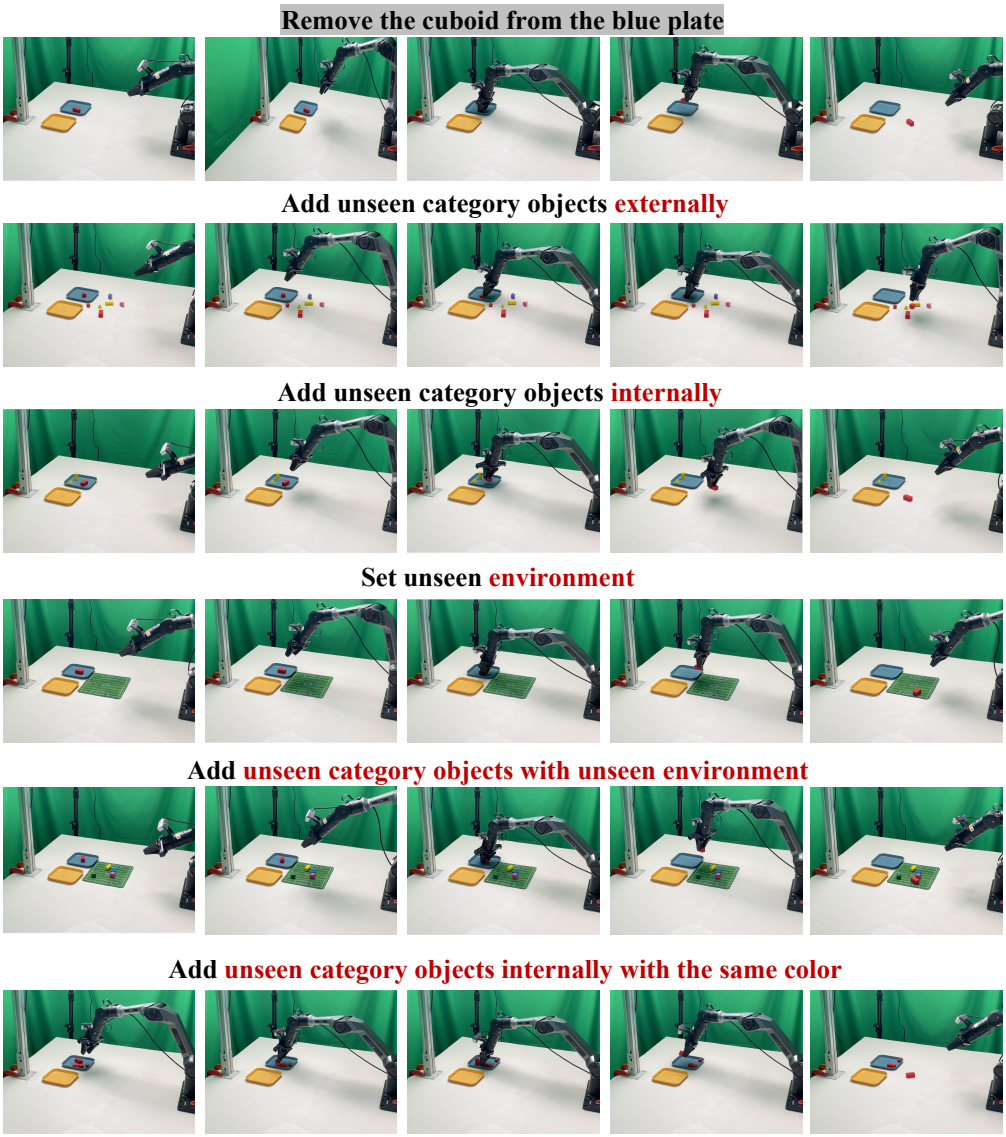



Figure 18: Qualitative results of our CofactVLA model on the **real-world robot experiments.**

### A.7 Limitations and Outlooks

**Dependence on Base VLM Priors.** First, the overall efficacy of our deconfounding framework is fundamentally bottlenecked by the inherent zero-shot semantic grounding capabilities of the underlying pre-trained Vision-Language Model (VLM). While our causal intervention effectively isolates and neutralizes spurious visual correlations, it cannot artificially synthesize semantic knowledge that the base model fundamentally lacks. Consequently, if the foundational model lacks the latent representations for highly specific visual attributes (e.g., rare object textures or novel geometries), our method could not compensate for this intrinsic knowledge deficit. Therefore, scaling up foundational vision-language pre-training remains a prerequisite to further pushing the boundaries of open-world generalization.

**Vulnerability to Severe Visual Occlusions.** Second, empirical evaluations in real-world physical deployments highlight a susceptibility to severe visual occlusions. During close-proximity manipulation, the robotic arm or the end-effector may temporarily obstruct the camera's field of view. This transient loss of critical visual feedback will disrupt the continuous flow-matching generation, occasionally leading to execution failures. Importantly, this stems from a hardware-induced perceptual limitation

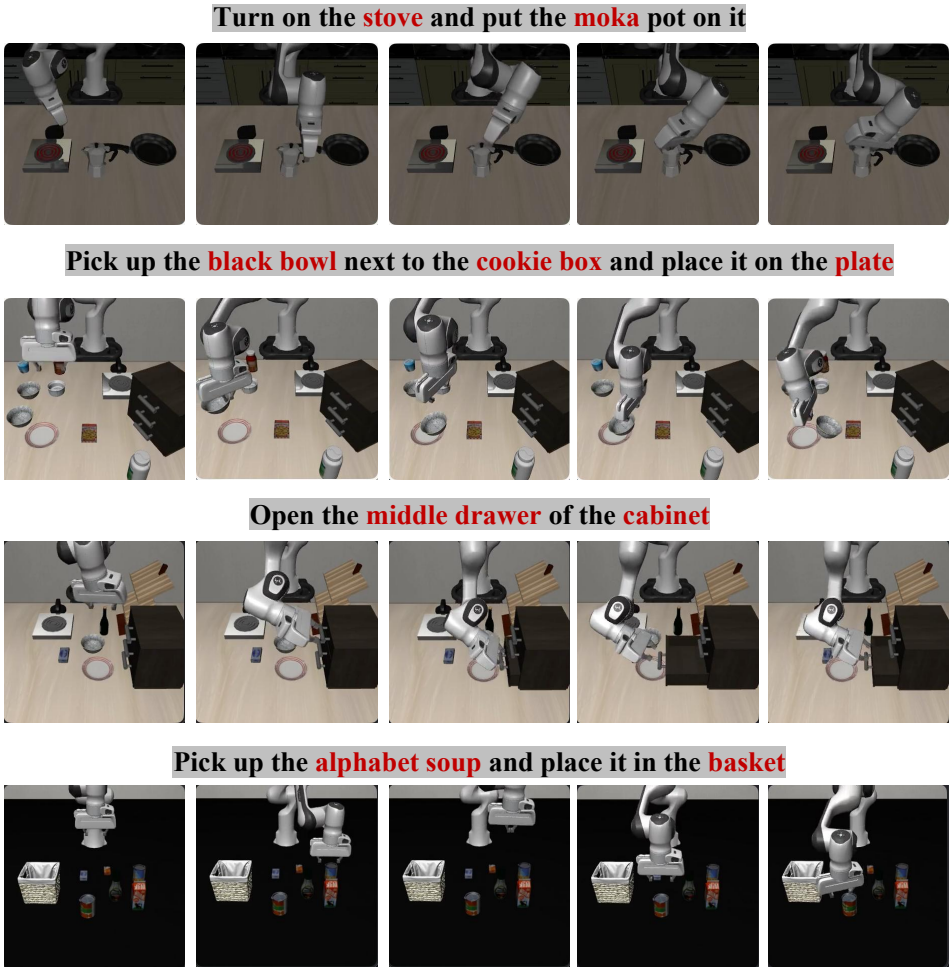



Figure 19: Failed results of our CofactVLA model on the **LIBERO-Plus benchmark.**

rather than an inherent algorithmic deficiency in our causal policy. To address this, future research could explore the integration of dynamic multi-view representation learning or real-time 3D scene understanding. Constructing a robust, occlusion-aware perceptual latent space would significantly unlock the full potential of our deconfounded VLA framework in highly cluttered and dynamic physical environments.

### A.8 Broader Impacts

The deployment of robust, language-grounded VLA models has significant positive societal potential, particularly in automating household chores, assisting the elderly, and operating in hazardous environments. By mitigating the vision-override phenomenon, CofactVLA may improve instruction following and reliability in the evaluated settings, but it does not provide formal safety guarantees. However, potential negative impacts include the acceleration of labor displacement in manual sectors. Furthermore, highly capable open-world manipulation models could theoretically be repurposed for harmful physical tasks if deployed without proper guardrails. We encourage future research to establish rigorous physical safety evaluations and alignment protocols before widespread commercial deployment.

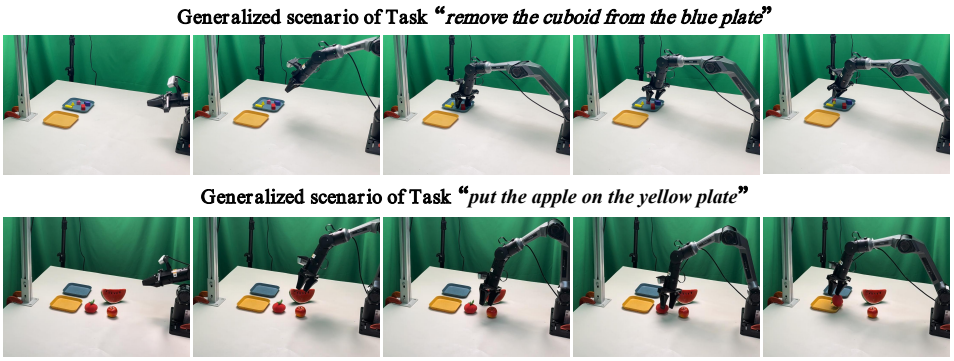



Figure 20: Failed results of our CofactVLA model on the **generalization experiment of real-world deployments.**

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction explicitly outline our core contributions (Our Dual-path Deconfounding Graph (DDG), Orthogonal Projection Guidance (OPG), and Counterfactual Covariance Reduction (CCR)), which are directly supported by the methodological details in Section 3 and the empirical evaluations in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the limitations of our framework in Section 5, specifically addressing the inherent upper-bound dictated by the base VLM's zero-shot grounding capabilities and the system's vulnerability to physical camera occlusions during real-world execution. Besides, Appendix A.6 and Appendix A.7 include qualitative results of failed cases and more detailed explanations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The formal theoretical assumptions (Assumptions 1 and 2) and the main theorem (Theorem 1) underlying Counterfactual Covariance Reduction are explicitly stated in Section 3.3. The complete, rigorous mathematical proofs, along with supporting propositions and lemmas, are fully detailed in Appendix A.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide comprehensive details to ensure reproducibility. Section 4.1 outlines the experimental setup for both simulation and real-world tasks. Furthermore, Appendix A.3 explicitly lists all training configurations and hyperparameters (e.g., base model, batch size, learning rate), and Appendix A.2 details the real-world hardware platform and execution protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The codebase for the proposed CofactVLA model has been cleaned and released at `https://github.com/zhangy0822/CofactVLA`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 4.1 introduces the experimental setups, evaluation protocols, and data configurations for both simulation benchmarks and real-world evaluations. Furthermore, comprehensive training details, including the optimizer (AdamW), learning rate, batch size, and other key hyperparameters, are explicitly listed in Appendix A.3 and Table 6.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to the high computational cost of training large-scale Vision-Language-Action models, reporting error bars across multiple random training seeds is prohibitive. However, to ensure statistical reliability, we report the average success rates computed over a rigorously large number of independent evaluation rollouts (10 episodes per task in simulation and 100 trials per task in real-world deployments).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Comprehensive details regarding the computational resources are provided in Appendix A.3. Specifically, Table 6 explicitly states the hardware configuration and the required training steps.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research focuses on improving the algorithmic robustness of robotic manipulation models. It does not involve human subjects, personally identifiable information, or any inherently harmful applications, and it strictly adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have included a dedicated discussion on the broader societal impacts in Appendix A.8, covering both the positive potentials (e.g., safer domestic/industrial assistance) and the negative risks (e.g., labor displacement, potential malicious physical deployments).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper introduces a framework for robotic manipulation policies evaluated in simulated and controlled physical environments. It does not involve the release of general-purpose large language models, unconstrained image generators, or scraped internet datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets, including the LIBERO benchmark suites, the pre-trained $\pi_{0.5}$ base model, and the `lerobot` framework, are appropriately credited via citations in the main text and Appendix A.3. Our usage strictly complies with their respective open-source licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper evaluates the proposed framework on existing public benchmarks and does not release new standalone public datasets.

    Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing, human subjects, or human-in-the-loop evaluations. The experiments are strictly algorithmic and evaluated on standard robotic simulation benchmarks and a physical robotic arm.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: As this research focuses exclusively on algorithmic interventions for robotic manipulation and does not involve human subjects or crowdsourcing, Institutional Review Board (IRB) approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

Justification: The core methodology of this research does not involve the use of LLMs as an external tool for data generation, evaluation, or as a non-standard pipeline component. The study focuses entirely on architectural and mathematical interventions within a Vision-Language-Action (VLA) model. Any usage of LLMs was strictly limited to standard writing and formatting assistance.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.