ACTRIGHT: COUNTERFACTUAL TRAINING FOR CAUSALLY-GROUNDED VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–Language–Action (VLA) models achieve strong performance in robotic manipulation, but existing approaches optimize task success without ensuring that actions are grounded in language semantics. As a result, models often rely on spurious visual correlations, leading to failures under distribution shifts and novel vision–language compositions. We propose ActRight, a training framework that enforces counterfactual causal constraints for cross-modal decision making, with three key technical contributions. First, we impose a structured V→L→A information flow by introducing an intermediate language representation as the sole input to the policy. Second, we introduce counterfactual objectives that constrain model behavior under interventions on vision and language, enforcing uncertainty under mismatched inputs and sensitivity to changes in language semantics. Third, we develop an adversarial mechanism that actively constructs challenging counterfactual scenarios to expose and penalize violations of causal consistency during training. Experiments on the LIBERO and CALVIN benchmarks show that ActRight consistently outperforms existing VLA methods in robustness and generalization, while improving sensitivity to language conditions.

1 INTRODUCTION

Vision–Language–Action (VLA) models have recently achieved strong performance in robotic manipulation and embodied AI. By learning policies that map multimodal inputs—including visual observations and language instructions—to actions, these models have demonstrated promising generalization across tasks and environments (Ghosh et al., 2024; Kim et al., 2025b). However, most existing approaches primarily optimize task success, without ensuring that actions are grounded in language semantics. As a result, models often rely on spurious visual correlations—such as background cues, object co-occurrence patterns, or dataset biases—which can lead to brittle behavior under distribution shifts and novel vision–language compositions (Guo et al., 2026; Glossop et al., 2025).

This limitation arises because current training paradigms constrain only the final outcome rather than the decision process itself. Although VLA models take both vision and language as input, their actions are not guaranteed to causally depend on these modalities. In practice, models may exploit shortcuts in the training data, producing correct actions for the wrong reasons. For example, a policy may predict actions based on scene layout or frequently observed patterns while largely ignoring language instructions, resulting in minimal changes in behavior when the instruction is altered. Such shortcut-driven policies undermine both robustness and generalization, and limit the reliability of VLA systems in real-world settings.

In this work, we address this limitation by introducing counterfactual causal constraints into VLA learning. We propose ActRight, a training framework that enforces cross-modal causal alignment by requiring that the influence of vision on action is mediated through language. Instead of solely optimizing task success, our method explicitly constrains how decisions are made.

Figure 1 illustrates the key distinction. Standard VLA models can exploit spurious visual shortcuts—producing correct actions for wrong reasons while language instructions are bypassed (Fig. 1a). ActRight enforces a structured **V→L→A** causal pathway through a language mediator f_ϕ, and con-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

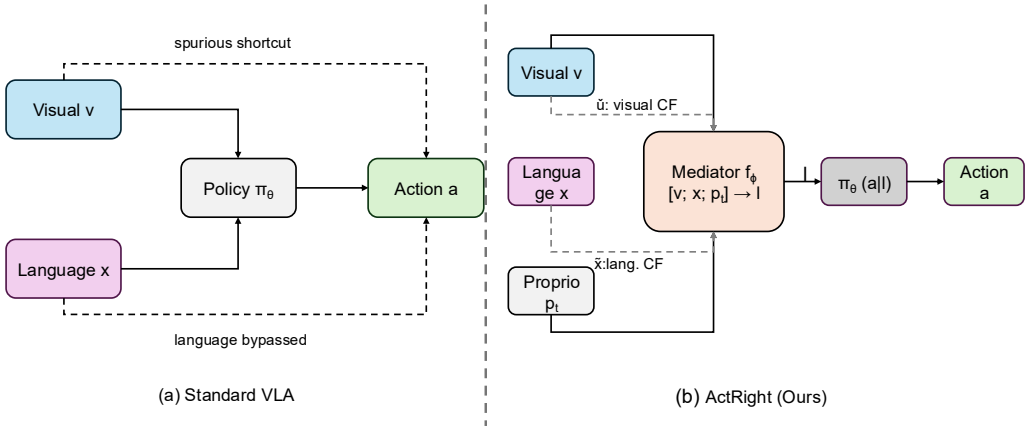


Figure 1: **Motivation. (a)** Standard VLA models can exploit spurious visual shortcuts, producing correct actions for the wrong reasons while bypassing language semantics. **(b) ActRight** enforces a structured **V→L→A** causal pathway via a language mediator f_ϕ, and constrains the policy under visual ($\tilde{v}$) and language ($\tilde{x}$) counterfactual interventions, ensuring that actions remain grounded in cross-modal causal dependencies.

strains the policy under counterfactual visual ($\tilde{v}$) and language ($\tilde{x}$) interventions so that actions remain grounded in cross-modal causal dependencies (Fig. 1b). Empirically, this yields **97.6%** average success rate on LIBERO and **99.5%** on CALVIN ABC→D, surpassing all prior VLA baselines.

Our approach consists of three key technical contributions. First, we impose a structured V→L→A information flow by introducing an intermediate language representation as the sole input to the policy, making language a mediator between vision and action. Second, we design counterfactual objectives that constrain model behavior under interventions on vision and language, enforcing uncertainty under mismatched inputs and sensitivity to changes in language semantics. Third, we develop an adversarial intervention mechanism that actively constructs challenging counterfactual scenarios to expose and penalize violations of causal consistency during training.

We evaluate ActRight on standard VLA benchmarks, including LIBERO and CALVIN. Experimental results show that our method outperforms existing VLA approaches in robustness and generalization, particularly under distribution shifts and unseen vision–language compositions.

Our contributions are as follows. (1) We formulate cross-modal causal alignment in VLA as a counterfactual learning problem, highlighting the need to enforce language-mediated decision making. (2) We propose a training framework that enforces counterfactual causal constraints via counterfactual objectives, built on a structured V→L→A formulation and strengthened by an adversarial intervention mechanism. (3) We demonstrate that ActRight outperforms existing VLA methods on standard benchmarks, achieving improved robustness and generalization.

2 Related Work

Generalist VLA foundation models. Recent Vision–Language–Action (VLA) research has rapidly advanced toward large-scale pretraining and generalist robot policies. Representative works such as Octo (Ghosh et al., 2024) and OpenVLA (Kim et al., 2025b) demonstrate that large-scale robot datasets and vision-language backbones can be unified to produce generalist policies that transfer across tasks, embodiments, observation modalities, and action spaces. In parallel, cross-embodiment learning has been explored to improve policy generalization across manipulation and navigation settings (Yang et al., 2024). Subsequent works extend this paradigm with more efficient architectures, stronger generalization, and richer spatial representations. For example, RoboMamba introduces an

efficient Mamba-based VLA architecture (Liu et al., 2024); $\pi_{0.5}$ improves open-world generalization via large-scale training (Black et al., 2025); and SpatialVLA investigates spatial representations for VLA models (Qu et al., 2025). Overall, these works establish VLA models as a promising foundation for general-purpose robot control. However, their training and evaluation are still largely centered on task success and empirical transfer, leaving the underlying causal relationships among visual observations, language instructions, and actions insufficiently explored.

Structured, robust, and efficient VLA models. Beyond generalist pretraining, recent works have introduced structured reasoning and system-level improvements to enhance long-horizon and history-dependent behavior. These include hierarchical policies and memory mechanisms (Shi et al., 2025; Koo et al., 2026), latent planning and dual-system reasoning (Huang et al., 2025; Chen et al., 2025), and spatiotemporal or world-aware modeling (Zhang et al., 2025a;b). Other studies focus on improving robustness and post-training generalization, including knowledge-preserving adaptation (Driess et al., 2025), parameter merging (Yadav et al., 2025), reinforcement learning fine-tuning (Liu et al., 2025), robustness to multi-modal perturbations (Guo et al., 2026), and cross-task generalization analysis (Zhou et al., 2025). In parallel, interpretability and efficiency-oriented methods aim to analyze or optimize VLA internals via mechanistic analysis (Häon et al., 2025), routing and sparsification (Li et al., 2025), token caching (Xu et al., 2025), and training-free acceleration or compression (Yang et al., 2025). While these approaches improve reasoning ability, robustness, or efficiency, they primarily treat intermediate representations as tools for empirical performance gains. In contrast, we explicitly model VLA learning as a cross-modal causal alignment problem, constraining the $V \to L \to A$ pathway and suppressing spurious correlations through counterfactual objectives.

Counterfactual learning. Counterfactual learning has been widely used to diagnose and mitigate shortcut learning in vision-language reasoning and decision-making (Niu et al., 2021; Lai et al., 2024; Ding et al., 2023; Armengol Urpí et al., 2024).

It has also been explored in embodied decision-making such as autonomous driving. For instance, CF-VLA (Peng et al., 2025) introduces a self-reflective reasoning loop that simulates alternative action outcomes to refine planned trajectories (Peng et al., 2025).

However, these approaches focus on planning-time reflection or robustness evaluation, without incorporating counterfactual reasoning into policy learning or explicitly modeling cross-modal causal dependencies. In contrast, we integrate counterfactual intervention directly into training to enforce causal alignment along the $V \to L \to A$ pathway.

3 METHOD

3.1 OVERVIEW

Given a visual observation v and an instruction x, our goal is to predict an action a that completes the task. As shown in Fig. 2, we introduce an intermediate language representation l and require the model to make decisions through it, forming a structured $V \to L \to A$ pathway. Our framework consists of three components. (i) We introduce a language-mediated policy architecture, where actions are predicted from an intermediate language representation that connects visual observations and instructions. (ii) We use counterfactual training to ensure the policy ignores irrelevant variations while responding to meaningful changes. (iii) We actively generate challenging counterfactual examples to expose shortcut behaviors and enforce correct visual–language dependencies. Together, these components encourage actions to be grounded in stable visual-language causal relations.

3.2 LANGUAGE-MEDIATED POLICY ARCHITECTURE

As illustrated in Fig. 2, we model VLA decision making as a two-stage process: first mapping raw perception into an instruction-relevant semantic representation, then predicting actions from this representation alone.

Feature extraction. Given a visual observation v, a vision encoder ϕ_v extracts patch-level features $\phi_v(v) \in \mathbb{R}^{N_p \times d_v}$, which are mean-pooled into $\bar{v} \in \mathbb{R}^{d_v}$. A language backbone ϕ_x embeds the

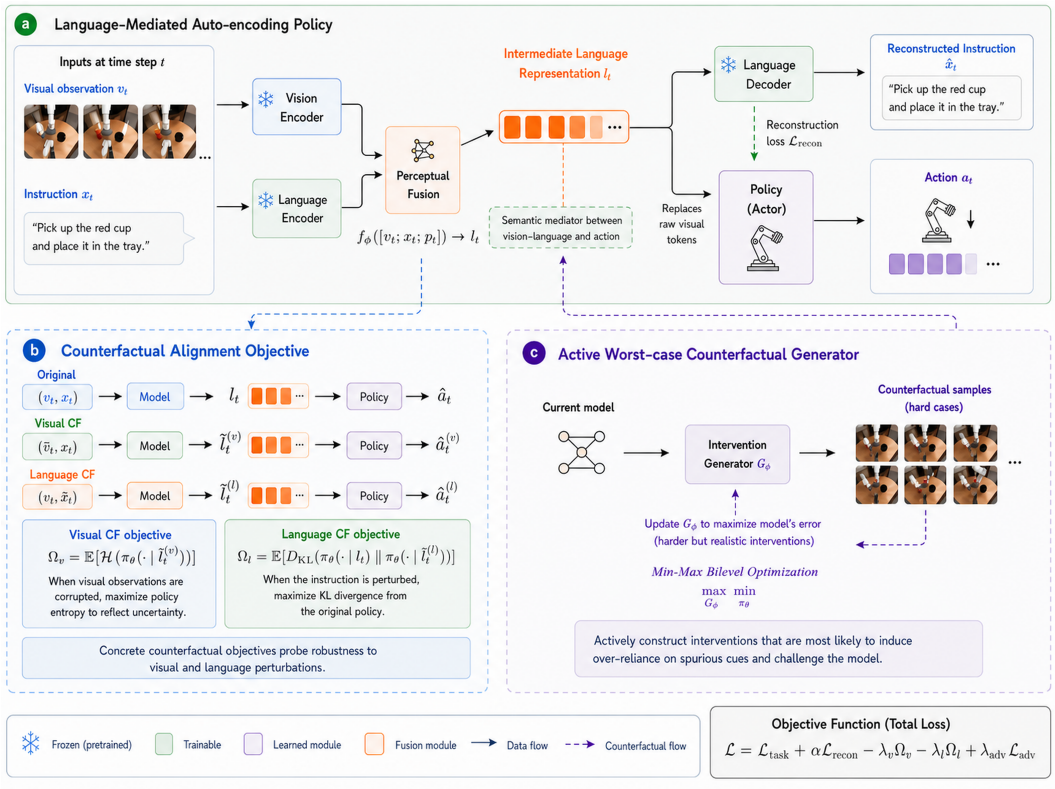


Figure 2: **Overview of the ActRight framework.** The framework comprises three tightly coupled components. **(i) Language-Mediated Architecture:** Visual observation v_t and language instruction x_t are encoded by frozen encoders and fused with proprioceptive state p_t by a learnable mediator f_ϕ, producing an intermediate language representation l_t. l_t is tiled across all visual patch slots of the LLM decoder, replacing raw patch features entirely, so the policy $\pi_\theta(a \mid l_t)$ can only access vision through the language-grounded mediator—enforcing the V↛L→A causal pathway by construction. **(ii) Counterfactual Alignment Objectives:** Two counterfactual branches test whether l_t captures causally relevant factors rather than spurious shortcuts. The visual branch substitutes a semantically inconsistent visual input $\tilde{v}_t$ (within-batch permutation) and maximizes policy entropy Ω_v, penalizing over-confident predictions under vision–language mismatch. The language branch substitutes a permuted instruction embedding $\tilde{x}_t$ and maximizes the KL divergence Ω_l between policies under original and counterfactual language, ensuring the policy remains sensitive to instruction semantics. **(iii) Active Worst-Case Generation:** An adversarial generator g_ψ targets the current policy via min-max bilevel optimization, adaptively synthesizing the hardest counterfactual pairs $(\tilde{v}, \tilde{l})$ at each training step to continuously expose the policy's most exploitable shortcut dependencies as they evolve. The total loss $\mathcal{L} = \mathcal{L}_{\text{task}} - \lambda_v\Omega_v - \lambda_l\Omega_l + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}$ jointly optimizes task performance and cross-modal causal consistency.

instruction tokens; the prompt embeddings (excluding supervised action-token positions) are pooled to give $\bar{x} \in \mathbb{R}^{d_l}$. A proprioceptive state vector $p_t \in \mathbb{R}^{d_p}$ (e.g., 7-DOF joint states) is appended directly.

Language mediator. The intermediate language representation is produced by:

$$l = f_\phi(v, x, p_t) = \text{MLP}_\phi([\bar{v}; \bar{x}; p_t]), \tag{1}$$

where $[\cdot;\cdot]$ denotes concatenation and MLP_ϕ is a three-layer MLP with GELU activations, projecting from $(d_v + d_l + d_p)$ to d_l. The mediator token $l \in \mathbb{R}^{d_l}$ is *tiled* across all N_p visual patch slots in the input sequence, so the action decoder π_θ never accesses raw patch features and must rely solely on the language-grounded mediator:

$$a \sim \pi_\theta(\cdot \mid l). \tag{2}$$

This enforces the V↛L→A pathway *by construction*.

Reconstruction regularizer. To prevent l from collapsing into a task-specific latent that discards linguistic structure, we add a reconstruction objective:

$$\hat{x} = g_\psi(l), \qquad \mathcal{L}_{\text{rec}} = \left\|\bar{x} - g_\psi(l)\right\|_2^2, \tag{3}$$

where g_ψ is a two-layer MLP decoder. This encourages l to retain semantically meaningful language-level information.

Two-stage training. In the first stage, f_ϕ and g_ψ are jointly trained to minimize $\mathcal{L}_{\text{rec}}$, establishing an instruction-aligned representation before any action supervision. In the second stage, the policy π_θ is jointly fine-tuned with f_ϕ via the full ActRight objective (Eq. 8), with $\mathcal{L}_{\text{rec}}$ retained as an auxiliary regularizer. This staged initialization reduces the risk of l becoming a trivial visual shortcut for the policy.

3.3 Counterfactual Alignment Objectives

While the language-mediated architecture enforces a structured V→L→A pathway, it does not guarantee that l captures *causally* relevant factors. To address this, we introduce counterfactual objectives that explicitly test whether the policy relies on the correct visual–language relationships.

Constructing counterfactual inputs. For each training batch $\mathcal{B} = \{(v_i, x_i, p_i)\}_{i=1}^B$, we construct counterfactual pairs via *within-batch permutation*: the visual counterfactual $\tilde{v}_i$ is taken from another sample $v_{\sigma(i)}$ (σ is a cyclic shift), yielding a semantically inconsistent pair $(\tilde{v}_i, x_i)$. The language counterfactual embedding $\tilde{x}_i$ is taken analogously from $\bar{x}_{\sigma(i)}$. Both require no extra data and introduce negligible overhead. Harder counterfactuals can additionally be supplied by the adversarial generator (Section 3.4).

Visual counterfactual objective. Given a mismatched visual input $\tilde{v}$, the mediator produces:

$$\tilde{l}^{(v)} = f_\phi(\tilde{v},\, x,\, p_t). \tag{4}$$

Under visual–language mismatch, the instruction can no longer be reliably grounded in the observation. A causally correct policy should therefore *withhold confident predictions*. We enforce this by maximizing the entropy of the action distribution:

$$\Omega_v = \mathbb{E}\Big[\mathcal{H}\Big(\pi_\theta(\cdot \mid \tilde{l}^{(v)})\Big)\Big], \tag{5}$$

where $\mathcal{H}(\cdot)$ is the Shannon entropy over the discrete action-token vocabulary. Intuitively, Ω_v penalizes over-confident predictions under inconsistent visual context, discouraging the encoder from storing spurious visual cues in l.

Language counterfactual objective. To verify that l exerts genuine causal influence on actions, we substitute a counterfactual language embedding $\tilde{x}$ and compute:

$$\tilde{l}^{(l)} = f_\phi(v,\, \tilde{x},\, p_t). \tag{6}$$

If l is a true causal mediator, replacing it should yield meaningfully different action predictions. We enforce this via a KL-divergence objective:

$$\Omega_l = \mathbb{E}\Big[D_{\text{KL}}\Big(\pi_\theta(\cdot \mid l) \,\Big\|\, \pi_\theta(\cdot \mid \tilde{l}^{(l)})\Big)\Big], \tag{7}$$

Maximizing Ω_l ensures the policy remains sensitive to semantic changes in the instruction, preventing convergence to a language-agnostic action prior. In practice, both $\mathcal{H}$ and D_{KL} are computed via a lightweight projection onto the action-token sub-vocabulary of the LLM head, avoiding a full autoregressive rollout during training.

Overall objective.

$$\max_{\theta,\phi}\; J_{\text{task}}(\theta, \phi) + \lambda_v\, \Omega_v + \lambda_l\, \Omega_l. \tag{8}$$

This enforces two complementary conditions: (i) the policy becomes uncertain when visual inputs are inconsistent with the instruction, and (ii) the policy changes its behavior when instruction semantics are altered. Together, these constraints encourage the model to rely on stable, causally grounded visual–language dependencies.

3.4 Active Worst-Case Counterfactual Generation

The within-batch permutations in Section 3.3 cover a fixed distribution of failure modes. In practice, shortcut dependencies evolve during training: early shortcuts may be suppressed, but new ones can emerge in their place. To continuously expose these vulnerabilities, we introduce a *learnable adversarial generator* g_ψ that actively synthesizes the most challenging counterfactual examples tailored to the current policy.

Generator architecture. g_ψ consists of two lightweight MLPs. Given (v, x, l, ξ) where $\xi \sim \mathcal{N}(0, I)$ injects diversity, the generator produces bounded perturbations:

$$\tilde{v} = \bar{v} + \delta_{\max} \cdot \tanh\big(\mathrm{MLP}^v_\psi([\bar{v}; \bar{x} + \xi])\big), \tag{9}$$

$$\tilde{l} = \bar{x} + \delta_{\max} \cdot \tanh\big(\mathrm{MLP}^l_\psi(\bar{x} + \xi)\big), \tag{10}$$

where $\delta_{\max}$ caps the perturbation magnitude to maintain semantic plausibility.

Min-max formulation. We formulate the joint optimization as:

$$\min_{\theta,\phi} \max_{\psi} \; J_{\text{task}}(\theta, \phi) + \lambda\, \mathcal{V}(\theta, \phi, \psi) + \eta\, \mathcal{C}(\psi), \tag{11}$$

where $\mathcal{V}$ measures violations of causal consistency and $\mathcal{C}$ constrains generated counterfactuals to remain near the originals. Specifically:

$$\mathcal{V}(\theta, \phi, \psi) = \mathbb{E}\Big[-\mathcal{H}\Big(\pi_\theta(\cdot \mid \tilde{l}^{(v)})\Big) - \beta\, D_{\mathrm{KL}}\Big(\pi_\theta(\cdot \mid l) \,\Big\|\, \pi_\theta(\cdot \mid \tilde{l}^{(l)})\Big)\Big], \tag{12}$$

$$\mathcal{C}(\psi) = \big\|\tilde{v} - \bar{v}\big\|_2^2 + \big\|\tilde{l} - \bar{x}\big\|_2^2, \tag{13}$$

where $\tilde{l}^{(v)} = f_\phi(\tilde{v}, x, p_t)$ and $\tilde{l}^{(l)}$ uses $\tilde{l}$ in place of $\bar{x}$. The generator g_ψ *minimizes* $\mathcal{V}$ (i.e., maximizes causal violations) subject to $\mathcal{C}$, while the policy (θ, ϕ) minimizes the resulting adversarial loss.

Efficient optimization via gradient reversal. To realize the min-max update within a single forward pass—required for FSDP distributed training compatibility—we attach a *gradient reversal layer* between generator outputs and the policy computation. During the forward pass it acts as an identity; during the backward pass it negates gradients flowing into g_ψ, converting the policy's loss-minimization gradient into a violation-maximization update for the generator. The adversarial contribution to the total loss is:

$$\mathcal{L}_{\text{adv}} = \mathcal{V}(\theta, \phi, \psi) + \eta\, \mathcal{C}(\psi). \tag{14}$$

The full training procedure is summarized in Algorithm 1.

Compared to standard counterfactual augmentation with fixed perturbations, g^{adv}_ψ adaptively discovers the hardest failure cases throughout training, ensuring the policy is continuously challenged by its current weaknesses rather than by a static failure distribution.

4 Experiment

4.1 Experiment Setup

Dataset Description We evaluate our method on two embodied manipulation benchmarks, **LIBERO** and **CALVIN ABC→D**, covering structured generalization and out-of-distribution task compositions. **LIBERO** contains **130 language-conditioned tasks** across four suites: **Spatial**, **Object**, **Goal**, and **Long**. Each task provides human demonstration trajectories with RGB observations and continuous actions, testing spatial, object, goal, and long-horizon generalization. **CALVIN** consists of **30+ atomic skills** in a household environment. In the **ABC→D** setting, models are trained on task subsets (A/B/C) and evaluated on unseen compositions (D), requiring long-horizon skill chaining under distribution shifts.

Algorithm 1 ActRight Training

Require: Dataset $\mathcal{D}$; policy π_θ; mediator f_ϕ; decoder g_ψ; adversarial generator g_ψ^{adv}; weights $\lambda_v, \lambda_l, \lambda_{\text{adv}}, \eta$; CF ratio ρ.

Stage 1: Pre-train mediator
for each batch $\{(v_i, x_i, p_i)\}$ from $\mathcal{D}$ **do**
  $l_i \leftarrow f_\phi(v_i, x_i, p_i)$; Update f_ϕ, g_ψ to minimize $\mathcal{L}_{\text{rec}}$ (Eq. 3)
end for
Stage 2: Joint counterfactual training
for each batch $\{(v_i, x_i, p_i, a_i)\}$ from $\mathcal{D}$ **do**
  $l_i \leftarrow f_\phi(v_i, x_i, p_i)$; compute $\mathcal{L}_{\text{task}}$ from $\pi_\theta(\cdot|l_i)$
  if with probability ρ **then**
    Sample permutation σ; set $\tilde{v}_i \leftarrow v_{\sigma(i)}$, $\tilde{x}_i \leftarrow \bar{x}_{\sigma(i)}$
    Compute Ω_v (Eq. 5) and Ω_l (Eq. 7)
    $(\tilde{v}^{\text{adv}}, \tilde{l}^{\text{adv}}) \leftarrow g_\psi^{\text{adv}}(v_i, x_i, l_i, \xi)$; compute $\mathcal{L}_{\text{adv}}$ via gradient reversal
    Update g_ψ^{adv} to maximize violations $\mathcal{V}$ (inner max)
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{task}} - \lambda_v \Omega_v - \lambda_l \Omega_l + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}$
  else
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{task}}$
  end if
  Update (θ, ϕ) to minimize $\mathcal{L}$ (outer min)
end for

Implementation Details Our implementation framework strictly follows the model architecture and training configuration of Open-VLA, including the pretrained visual encoder, language encoding module, and transformer-based action policy head, while adopting identical optimizer settings, learning rate schedules, and training steps. We set $\lambda_v = 0.1$, $\lambda_l = 0.1$, $\lambda_{adv} = 0.05$, $\delta_{max} = 0.1$, CF sampling ratio $\rho = 0.3$. Stage 1 pre-trains the mediator for 500 steps with lr = 1e-4 (cosine decay). Stage 2 uses LoRA (rank 16) on attention layers with lr = 2e-5. Training runs on 8× A100 (80GB) GPUs; Stage 1 takes 2 hours, Stage 2 takes 18 hours on LIBERO.

Evaluation Metrics The evaluation metrics are configured identically to OpenVLA to ensure valid cross-comparisons.We employ task success rate, compositional generalization performance, language sensitivity analysis, and visual robustness analysis as evaluation metrics.

We evaluate using four complementary metrics: (1) **task success rate** on both in- and out-of-distribution environments; (2) **compositional generalization** on unseen vision–language combinations; (3) **language sensitivity**, measuring action distribution shifts under instruction changes with fixed visual inputs; and (4) **visual robustness**, assessing over-reliance on spurious correlations under visual perturbations. Together, these metrics provide a comprehensive evaluation of performance, generalization, and cross-modal causal grounding.

4.2 Main Results

4.2.1 Overall comparison on simulation benchmark

LIBERO Table 1 presents a comparison of the success rate (SR) between the proposed method and representative existing Vision-Language-Action (VLA) approaches across four task suites: **LIBERO-Spatial**,**LIBERO-Object**, **LIBERO-Goal**, and **LIBERO-Long**. To comprehensively evaluate model capabilities, we report results under two settings: **"one policy for all 4 suites"** and **"one policy per suite"**. In the first setting, a single policy is trained to handle all four task suites simultaneously, emphasizing unified modeling and cross-task generalization. In the second setting, separate policies are trained for each suite, focusing on the performance upper bound within individual suites.

Under the **one policy for all 4 suites** setting, ActRight achieves 97.2% average success rate, outperforming OpenVLA-OFT (96.8%) and UnifiedVLA (95.5%), with the largest gains on Goal and Long tasks—demonstrating strong cross-task generalization from a single unified policy.

Table 1: Performance on LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long task suites

Method	Spatial SR (%)	Object SR (%)	Goal SR (%)	Long SR (%)	Average SR (%)
One policy for all 4 suites					
TraceVLA (Zheng et al., 2024)	84.6	85.2	75.1	54.1	74.8
WorldVLA (Cen et al., 2025)	87.6	96.2	83.4	60	81.8
π_0 (Black et al., 2024)	96.8	98.8	95.8	85.2	94.2
π_0-FAST (Pertsch et al., 2025)	96.4	96.8	88.6	60.2	85.5
UnifiedVLA (Wang et al., 2025)	95.4	96.8	93.6	94	95.5
OpenVLA-OFT (Kim et al., 2025a)	**97.7**	98	96.1	95.3	96.8
Ours	97.4	**99.2**	**97.7**	**97.5**	**97.2**
One policy per suite					
OpenVLA (Kim et al., 2025b)	84.7	88.4	79.2	53.7	76.5
SpatialVLA(Qu et al., 2025)	88.2	89.9	78.6	55.5	78.1
CoT-VLA (Zhao et al., 2025)	87.5	91.6	87.6	69	83.9
NORA (Hung et al., 2025)	92.2	95.4	89.4	74.6	87.9
PD-VLA (Song et al., 2025)	95.5	96.7	94.9	91.7	94.7
UniVLA (Bu et al., 2025)	96.5	96.8	95.6	92	95.2
OpenVLA-OFT (Kim et al., 2025a)	97.6	98.4	97.9	94.5	97.1
FLOWER (Reuss et al., 2025)	97.5	99.1	96.1	94.9	96.9
VLA-Adapter(Wang et al., 2026)	97.8	99.2	97.2	**95**	97.3
RIPT-VLA(Tan et al., 2025)	99	98.6	**98.6**	93.8	97.5
Ours	**99.1**	**99.4**	96.8	94.9	**97.6**

Under the **one policy per suite** setting, ActRight attains state-of-the-art results on Spatial (99.1%) and Object (99.4%) suites and the highest overall average of 97.6%, surpassing RIPT-VLA (97.5%), VLA-Adapter (97.3%), and OpenVLA-OFT (97.1%). Unlike methods that excel on a subset of tasks, ActRight maintains consistently high performance across all four suites, reflecting a more balanced and robust policy.

In summary, the causal alignment mechanism improves performance under both settings: it enhances cross-task generalization in the unified-policy setting and achieves the best overall average in the per-suite setting, confirming stronger robustness and adaptability across diverse task distributions.

Table 2: Performance comparison on CALVIN ABC→D benchmark

CALVIN ABC→D	**Task completed in a row ↑**	**Avg. len ↑**
OpenVLA (Kim et al., 2025b)	91.3	77.8
UniVLA (Bu et al., 2025)	95.5	85.8
UnifiedVLA (Wang et al., 2025)	98.9	94.8
OpenVLA-OFT (Kim et al., 2025a)	96.9	92
FLOWER (Reuss et al., 2025)	99.4	95.8
VLA-Adapter (Wang et al., 2026)	99.1	94.6
Seer (Yue et al., 2024)	96.3	91.6
Ours	**99.5**	**96.9**

CALVIN ABC→D Table 2 presents the comparison results of various methods on the CALVIN ABC→D benchmark, which evaluates long-horizon sequential decision-making in robotic manipulation. This benchmark requires the agent to execute a sequence of language-conditioned sub-tasks, making it particularly suitable for assessing both short-term execution accuracy and long-term policy consistency. We report two metrics: *Task completed in a row* (consecutive sub-tasks completed without failure) and *Avg. len* (average trajectory length of successful executions), jointly measuring execution accuracy and long-horizon stability.

ActRight achieves state-of-the-art performance on both metrics, attaining 99.5% and 96.9 Avg. len, outperforming all baselines including FLOWER, VLA-Adapter,and UnifiedVLA. The advantage is most pronounced in challenging multi-step settings, where ActRight more effectively mitigates error accumulation and maintains stable behavior over extended sequences, demonstrating strong robustness and generalization in complex, long-horizon tasks.

Table 3: Ablation Study on LIBERO Task Suites. w/o VLA C. indicates w/o V→L→A Constraint;w/o C.R. indicates w/o Counterfactual Rewards;w/o V.CF R.indicates w/o Visual CF Reward;w/o L.CF R. indicates w/o Language CF Reward;R.L indicates Randomized L

Model	Spatial SR (%)	Obj. SR (%)	Goal SR (%)	Long SR (%)	Avg. SR (%)
Our Model	**99.1**	**99.4**	**96.8**	**94.9**	**97.6**
w/o VLA C.	97.8	98.6	94.1	91.2	95.4
w/o C.R.	97.2	97.9	93.6	89.8	94.6
w/o V.CF R.	98.9	98.7	95.1	92	96.2
w/o L.CF R.	98.3	98.1	94.4	90.9	95.4
R.L	95.6	96.2	90.7	86.5	92.3

4.2.2 Ablation Study

Table 3 summarizes the ablation study results on the four LIBERO task suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long. All experiments are conducted under datasets, training pipelines, and evaluation protocols strictly consistent with OpenVLA, with only targeted modifications to model architecture or training objectives. Therefore, the observed performance differences can be directly attributed to the proposed causal alignment mechanism.

The full model achieves 97.6% average success rate across all four suites, confirming that jointly enforcing the V→L→A constraint with counterfactual rewards yields the most robust and generalizable policy.

Effect of architectural constraint. Removing the V→L→A constraint drops performance to 95.4%, with the largest degradation on Goal and Long tasks, demonstrating that counterfactual rewards alone cannot prevent the model from bypassing the language mediator and exploiting visual shortcuts.

Effect of counterfactual rewards. Retaining the structural constraint but removing all counterfactual supervision further reduces performance to 94.6%, showing that architectural design alone is insufficient: without counterfactual objectives, the language representation collapses into a superficial bottleneck rather than a causally meaningful variable.

Component-wise analysis. Removing visual or language counterfactual rewards independently yields 96.2% and 95.4% respectively, indicating that the two objectives provide complementary constraints—suppressing spurious visual reliance and maintaining genuine language sensitivity—neither of which can substitute for the other.

Causal mediation validation. Randomizing the intermediate language representation causes the sharpest drop to 92.3%, most severely on Goal and Long tasks, confirming that l functions as an indispensable causal mediator rather than a redundant bottleneck.

5 Conclusion

We presented ActRight, a causal VLA training framework that enforces a structured $V \to L \to A$ pathway via a language-mediated representation, grounding actions in instruction-relevant semantics rather than spurious visual correlations. Augmented with counterfactual alignment objectives and an active worst-case generator, the framework yields consistent improvements in robustness and generalization across distribution shifts and novel compositions—highlighting that constraining *how* decisions are made, not just their outcomes, is key to reliable VLA deployment.

Limitations. Evaluation is limited to simulation (LIBERO, CALVIN); real-robot transfer remains unvalidated. Within-batch counterfactuals are semantically inconsistent but not guaranteed to be physically plausible, and the two-stage training adds ∼20–30% overhead. Mean-pooling before the mediator also discards spatial detail that may matter for precise low-level manipulation.

AI Use Statement

We used generative AI tools only to polish the language of this manuscript (grammar, wording, and readability). We did not use generative AI tools for research ideation, method design, implementation,

experiments, or analysis of results. All authors reviewed the edited text and take full responsibility for the content of this paper.

Ethics Statement

This work studies training objectives for robot manipulation policies and is evaluated entirely in simulation on public benchmarks (LIBERO and CALVIN). It involves no human subjects, personal data, or crowdsourcing. Any deployment on physical robots should include appropriate safety measures.

Reproducibility Statement

Section 3 and Algorithm 1 specify the architecture, objectives, and training procedure. Section 4.1 lists the hyperparameters, compute resources, and training schedules. Anonymized source code will be provided as supplementary material and released upon acceptance.

References

Núria Armengol Urpí et al. Causal action influence aware counterfactual data augmentation. In *ICML*, 2024.

Kevin Black, Noah Brown, Danny Driess, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Kevin Black, Noah Brown, Danny Driess, Chelsea Finn, Sergey Levine, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Conference on Robot Learning (CoRL)*, 2025.

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.

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.

Hao Chen, Jiaming Liu, Chenyang Gu, Zhuoyang Liu, Renrui Zhang, Xiaoqi Li, Xiao He, Yandong Guo, Chi-Wing Fu, Shanghang Zhang, and Pheng-Ann Heng. Fast-in-slow: A dual-system foundation model unifying fast manipulation within slow reasoning. *arXiv preprint arXiv:2506.01953*, 2025.

Wenhao Ding, Laixi Shi, et al. Seeing is not believing: Robust reinforcement learning against spurious correlation. In *NeurIPS*, 2023.

Danny Driess, Jost Tobias Springenberg, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. In *NeurIPS*, 2025.

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

Catherine Glossop et al. Cast: Counterfactual labels improve instruction following in vision-language-action models. *arXiv preprint arXiv:2508.13446*, 2025.

Jianing Guo, Zhenhong Wu, et al. On robustness of vision-language-action models against multimodal perturbations. In *ICLR*, 2026.

Bear Häon, Kaylene Stocking, et al. Mechanistic interpretability for steering vision-language-action models. In *CoRL*, 2025.

Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *arXiv preprint arXiv:2507.16815*, 2025.

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.

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

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713, 2025b.

Myungkyu Koo, Daewon Choi, et al. Hamlet: Switch your vision-language-action model into a history-aware policy. In *ICLR*, 2026.

Chengen Lai, Shengli Song, et al. Improving vision and language concepts understanding with multimodal counterfactual samples. In *ECCV*, 2024.

Wei Li, Renshan Zhang, Rui Shao, Jie He, and Liqiang Nie. Cogvla: Cognition-aligned vision-language-action models via instruction-driven routing and sparsification. In *NeurIPS*, 2025.

Jiaming Liu, Mengzhen Liu, Zhenyu Wang, et al. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. In *NeurIPS*, 2024.

Jijia Liu, Feng Gao, et al. What can rl bring to vla generalization? an empirical study. In *NeurIPS*, 2025.

Yulei Niu, Kaihua Tang, et al. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, 2021.

Zhenghao Mark Peng et al. Counterfactual vla: Self-reflective vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2512.24426*, 2025.

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.

Delin Qu, Haoming Song, et al. Spatialvla: Exploring spatial representations for vision-language-action models. In *Robotics: Science and Systems (RSS)*, 2025.

Moritz Reuss, Hongyi Zhou, Marcel Rühle, Ömer Erdinç Yağmurlu, Fabian Otto, and Rudolf Lioutikov. Flower: Democratizing generalist robot policies with efficient vision-language-action flow policies. *arXiv preprint arXiv:2509.04996*, 2025.

Lucy Xiaoyang Shi, Brian Ichter, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. In *ICML*, 2025.

Wenxuan Song, Jiayi Chen, Pengxiang Ding, Han Zhao, Wei Zhao, Zhide Zhong, Zongyuan Ge, Jun Ma, and Haoang Li. Accelerating vision-language-action model integrated with action chunking via parallel decoding. *arXiv preprint arXiv:2503.02310*, 2025.

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

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*, volume 40, pp. 18638–18646, 2026.

Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. *arXiv preprint arXiv:2506.19850*, 2025.

Siyu Xu, Yunke Wang, Chenghao Xia, Dihao Zhu, Tao Huang, and Chang Xu. Vla-cache: Efficient vision-language-action manipulation via adaptive token caching. In *NeurIPS*, 2025.

Yajat Yadav, Zhiyuan Zhou, et al. Robust finetuning of vision-language-action robot policies via parameter merging. *arXiv preprint arXiv:2512.08333*, 2025.

Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, et al. Pushing the limits of cross-embodiment learning for manipulation and navigation. In *Robotics: Science and Systems (RSS)*, 2024.

Yantai Yang, Yuhao Wang, et al. Efficientvla: Training-free acceleration and compression for vision-language-action models. *arXiv preprint arXiv:2506.10100*, 2025.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *Advances in Neural Information Processing Systems*, 37:56619–56643, 2024.

Jiahui Zhang, Yurui Chen, et al. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration. In *NeurIPS*, 2025a.

Wenyao Zhang, Hongsi Liu, et al. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. In *NeurIPS*, 2025b.

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

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

Jiaming Zhou, Ke Ye, et al. Exploring the limits of vision-language-action manipulations in cross-task generalization. *arXiv preprint arXiv:2505.15660*, 2025.