MixVLA: Adaptive Mixing of Non-Invariant Information for Generalizable Vision-Language-Action Models

Anonymous authors
Paper under double-blind review

Abstract

Vision-Language-Action (VLA) models have achieved remarkable advances in robotic manipulation, yet their zero-shot generalization under out-of-distribution (OOD) conditions remains limited. These models often entangle task-relevant invariant structure with environment-specific non-invariant factors, causing policies to rely on spurious appearance cues during action prediction. In this work, we propose **MixVLA**, a model-agnostic training framework that improves the generalization of VLA models without requiring additional OOD data or architectural modifications. The key component of MixVLA is **Adaptive Mixing of Non-Invariant Information (AMI)**. AMI stochastically mixes non-invariant representations to regularize distribution-specific variability while preserving complementary predictive cues. The mixed non-invariant features are then fused with invariant representations for final action prediction, resulting in improved robustness without sacrificing policy expressiveness. Extensive experiments across challenging manipulation settings, including LIBERO, LIBERO-Plus, the RoboTwin perturbation suite, and real-world tasks, demonstrate that MixVLA improves overall zero-shot robustness while retaining strong in-domain performance.

1 Introduction

Vision-Language Models (VLMs) (Achiam et al., 2023; Bai et al., 2023) have advanced rapidly in recent years and are increasingly applied to robotic manipulation. This trend has led to the emergence of Vision-Language-Action (VLA) models (Kim et al., 2024; Brohan et al., 2022) that map multimodal observations and language instructions into low-level control actions. Benefiting from rich cross-modal representations learned by VLMs, these models achieve competitive task success rates across both simulated (Liu et al., 2023; Chen et al., 2025) and real-world (Walke et al., 2023) manipulation benchmarks.

Despite strong performance on in-distribution benchmarks, VLA models still struggle to achieve robust zero-shot generalization under out-of-distribution (OOD) shifts (Li et al., 2025a; Zhou et al., 2025; Wang et al., 2026). Recent studies (Fei et al., 2025b) show that even mild variations in lighting, appearance, or camera viewpoint can cause substantial performance degradation. Some prior works (Zitkovich et al., 2023; Team, 2025) attempt to improve robustness through large-scale and diverse data collection, but they incur substantial labor and financial costs. Other approaches (Abouzeid et al., 2025; Wu et al., 2025) enhance specific aspects of generalization by introducing model-specific modifications, which often limit their general applicability. In this work, we instead focus on improving zero-shot generalization through a model-agnostic training strategy that does not rely on additional out-of-distribution data. As illustrated in figure 1, our key idea is to explicitly regularize the variability of non-invariant signals in the latent representation space.

To better understand this intuition, we analyze the problem from a representation learning perspective. Policies trained with empirical risk minimization (ERM) (Vapnik, 1999) tend to capture both task-relevant invariant correlations and environment-specific non-invariant

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.

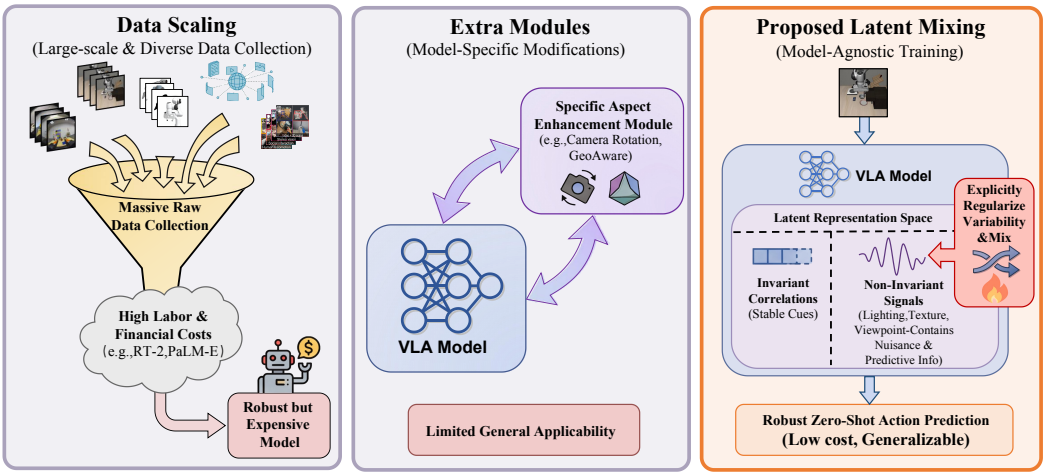


Figure 1: **Left:** Data scaling improves robustness by collecting massive and diverse demonstrations, but incurs substantial labor and financial costs. **Middle:** Some works improve generalization with task-specific modules (e.g., camera rotation or geometric awareness), which limits general applicability across VLA architectures. **Right:** MixVLA is a model-agnostic training framework that explicitly regularizes and mixes non-invariant signals in the latent representation space, enabling robust zero-shot action prediction without additional data or architectural modifications.

signals. While invariant features provide stable cues for action prediction, non-invariant factors such as lighting, texture, and viewpoint often contain both nuisance variations and complementary predictive information. Under ERM training, directly relying on raw non-invariant features can introduce spurious dependencies under distribution shifts. Conversely, enforcing strict invariance may be overly conservative, as it can discard informative signals that remain useful for action prediction. This observation motivates explicitly regularizing the variability of non-invariant signals rather than relying on invariant features alone.

Building on the above insights, we propose **MixVLA**, a model-agnostic training framework that improves zero-shot generalization of Vision-Language-Action models, without modifying model architectures or requiring additional data. The core of MixVLA is **A**daptive **M**ixing of Non-**I**nvariant Information (**AMI**). Specifically, we first disentangle invariant and non-invariant representations using an information-theoretic objective. We then perform stochastic mixing within the non-invariant feature space to regularize environment-specific variability while preserving complementary predictive cues. The resulting mixed non-invariant representation is subsequently fused with invariant features for final action prediction, enabling a more favorable balance between robustness and policy expressiveness. We evaluate MixVLA across a range of challenging manipulation settings, including the LIBERO and LIBERO-Plus benchmarks, the RoboTwin perturbation suite, and real-world robotic manipulation tasks. The results show that MixVLA substantially improves the robustness of VLA policies under diverse observation shifts without requiring additional data collection.

Our key contributions are three-fold:

- **A model-agnostic framework for generalized VLA models.** We present **MixVLA**, a model-agnostic training framework that improves the zero-shot generalization of Vision-Language-Action models under out-of-distribution (OOD) data without additional data collection or architectural redesign.
- **Adaptive mixing of non-invariant information (AMI).** We propose **AMI**, which adaptively mixes non-invariant representations and fuses them with invariant features to improve robustness under distribution shifts.
- **Consistent zero-shot gains on challenging manipulation benchmarks.** Experiments on LIBERO, LIBERO-Plus, RoboTwin, and real-world manipulation tasks demon-

strate that MixVLA consistently improves zero-shot task success rates under diverse perturbations.

2 Related Work

2.1 Vision-Language-Action Model

Recent Vision-Language-Action (VLA) models build upon large-scale Vision-Language Models (VLMs) and are fine-tuned on diverse robot trajectory datasets to learn language-conditioned visuomotor policies (Brohan et al., 2022; Zitkovich et al., 2023; Walke et al., 2023; Ebert et al., 2021; O'Neill et al., 2024). RT-1 and RT-2 (Brohan et al., 2022; Zitkovich et al., 2023) demonstrate that scaling cross-modal pretraining together with large robot datasets can substantially improve manipulation performance and semantic grounding. Following this paradigm, OpenVLA (Kim et al., 2024) discretizes robot actions into rarely used vocabulary tokens and performs autoregressive action generation on top of the Prismatic-7B backbone (Karamcheti et al., 2024). The π_0 family (Black et al., 2024; Intelligence et al., 2025) further introduces a dual-system design that augments a VLM backbone with a dedicated action expert to jointly model high-level reasoning and low-level control. Beyond architectural scaling, recent efforts explore richer spatial representations (Zhen et al., 2024; Bhat et al., 2026; Zhang et al., 2025a), intermediate supervision such as subtask decomposition and trajectory prediction (Zheng et al., 2024; Zhao et al., 2025; Zhang et al., 2025b; Zhong et al., 2025; Cen et al., 2025; Lee et al., 2025), and post-training optimization via planning or reinforcement learning (Guo et al., 2025; Zhang et al., 2025c; Lu et al., 2025; Fei et al., 2025a; Huang et al., 2025) to enhance VLA capability. Despite these advances in scaling and architectural design, improving robustness to observation variations in unseen environments remains an open challenge for VLA policies. This motivates our focus on enhancing the intrinsic generalization ability of VLA models through a model-agnostic training strategy.

2.2 Generalization in Robotic Manipulation

Recent work has explored multiple directions to improve generalization in robotic manipulation. One common line of work (Brohan et al., 2022; Driess et al., 2023; Zitkovich et al., 2023; Zhaxizhuoma et al., 2025; Team, 2025) seeks to enhance robustness through large-scale and diverse data collection. By expanding the coverage of training distributions, these approaches aim to improve policy generalization across varied environments. However, such strategies often incur substantial human and financial costs, and the practical boundary of data-driven generalization for embodied models remains insufficiently understood (Team, 2025). Complementary to data scaling, another line of research improves robustness by introducing additional architectural modules or auxiliary components. For example, GeoAware-VLA (Abouzeid et al., 2025) incorporates VGGT (Wang et al., 2025) to improve robustness to camera viewpoint changes. AffordDP (Wu et al., 2025) enhances cross-scene manipulation through affordance retrieval and transfer, while 3DS-VLA (Li et al., 2025b) improves robustness to appearance and viewpoint variations by introducing a 3D point cloud perception module. While these designs can improve robustness under specific perturbations, they often introduce additional system complexity and may not generalize well to diverse out-of-distribution scenarios, where realistic shifts are typically multifaceted rather than isolated. A separate line of work (Yang et al., 2024a;b; Seo et al., 2024; Ryu et al., 2023) seeks to improve policy generalization by learning robust visual representations or incorporating structured equivariant priors. These approaches typically integrate mechanisms such as associative memory or geometric equivariance to maintain policy consistency under object-level transformations, including scaling, rotation, and translation. However, such methods primarily focus on object-centric variations and do not explicitly address variability arising from environment-level nuisance factors, such as lighting changes, background appearance, or camera viewpoint shifts.

In contrast to these directions, our work improves VLA robustness through a model-agnostic training strategy that explicitly regularizes non-invariant environmental variability without requiring additional data or specialized architectural modules.

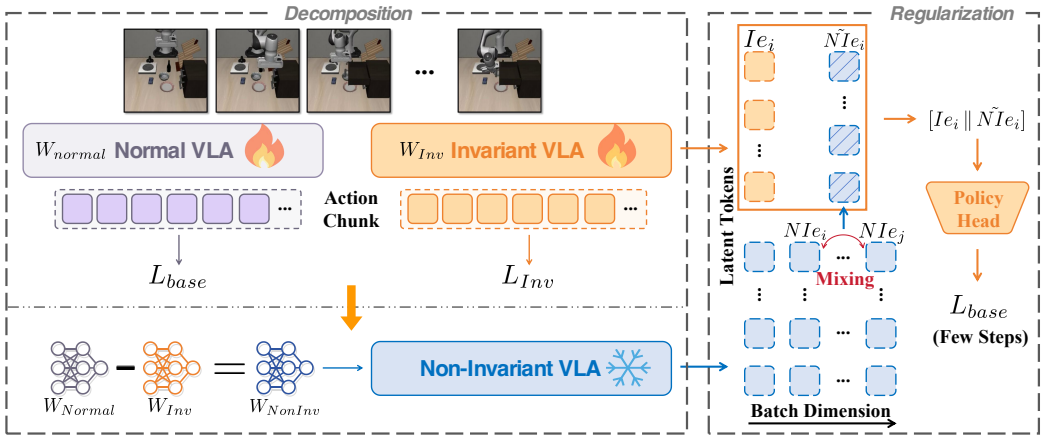


Figure 2: **Overview of MixVLA. Left (Decomposition):** A standard VLA policy is trained with the base objective to obtain W_{Normal}, while an Invariant VLA is trained with an information bottleneck objective to obtain W_{Inv}. The Non-Invariant VLA is then derived via weight arithmetic. **Right (Regularization):** Non-invariant embeddings are randomly mixed within the mini-batch using **AMI**. The mixed representation $\tilde{NIe}$ is concatenated with the invariant embedding Ie and passed to the policy head for training.

3 Method

3.1 Preliminaries

VLA policy. We consider a VLA policy that maps an observation–instruction pair to robot actions. Let X denote the visual observation and L the language instruction. A standard VLA backbone produces a latent representation $Z = f_\theta(X, L)$, which is subsequently consumed by the policy head $\pi(\cdot)$ to predict the action A. MixVLA operates at this latent level without modifying the backbone or the action head.

Invariant and non-invariant factors. We hypothesize that Z entangles task-relevant invariant structure shared across environments with environment-dependent non-invariant factors. MixVLA learns an *Invariant Net* and derives a complementary *Non-Invariant Net* to extract the corresponding embeddings Ie and NIe, respectively. Adaptive mixing regularizes NIe into $\tilde{NIe}$, yielding the fused representation $h = [Ie \,\|\, \tilde{NIe}]$, which is fed to the original policy head for action prediction. We denote the parameters of the Invariant, Normal, and Non-Invariant Nets by W_{Inv}, W_{Normal}, and W_{NonInv}, respectively (see figure 2).

3.2 Invariant and Non-Invariant Decomposition

Invariant Net via Information Bottleneck. The Invariant Net aims to extract task-critical features that remain stable across environments while suppressing redundant correlations. We adopt the Information Bottleneck (IB) principle to learn Ie by minimizing $I(X; Ie)$ while maximizing $I(Ie; A)$, balancing input compression and action predictability. Using the identity

$$I(Ie; A) = H(A) - H(A|Ie), \tag{1}$$

to maximize $I(Ie; A)$ is equivalent to minimizing the conditional entropy $H(A|Ie)$, which corresponds to improving action predictability. Combining sufficiency and minimality yields the classical IB objective:

$$\mathcal{L}_{\text{IB}} = \beta I(X; Ie) - I(Ie; A). \tag{2}$$

In practice, we optimize the following surrogate:

$$\mathcal{L}_{\text{Inv}} = \mathbb{E}_{(X,A)}\left[\mathcal{L}_{\text{base}}(\pi(Ie), A)\right] + \beta\, \hat{I}(X; Ie), \tag{3}$$

where $\pi(\cdot)$ denotes the original policy head, $\mathcal{L}_{\text{base}}$ is the standard VLA training loss, and $\hat{I}(X; Ie)$ is estimated using a neural mutual information estimator (Belghazi et al., 2018).

The first term approximates the sufficiency constraint by reducing $H(A|Ie)$, while the second term enforces information compression. As a result, the Invariant Net captures stable correlations that better transfer under observation shifts.

Non-Invariant Net via Weight Subtraction. While the Invariant Net is encouraged to capture task-stable structure, the standard policy may retain complementary information that varies across environments. To capture such complementary information, we construct a *Non-Invariant Net* using **weight arithmetic**. Let W_{Normal} denote the weights of the standard policy trained with $\mathcal{L}_{\text{base}}$, and W_{Inv} denote the Invariant Net weights obtained by minimizing equation 3. We define

$$W_{\text{NonInv}} = W_{\text{Normal}} - W_{\text{Inv}}. \tag{4}$$

The term *Non-Invariant Net* denotes this candidate complementary branch, without assuming an exact separation of non-invariant information. Appendix A provides a local-linearization perspective on parameter differences, while the utility of the resulting branch is evaluated empirically. We then compute the non-invariant embedding as

$$NIe = f_{W_{\text{NonInv}}}(X, L), \tag{5}$$

where $f_{W_{\text{NonInv}}}$ denotes the network parameterized by W_{NonInv}.

3.3 Adaptive Mixing of Non-Invariant Information

Purely invariant representations can be overly conservative, while exploiting non-invariant features may introduce spurious dependencies. We propose Adaptive Mixing of Non-Invariant Information (**AMI**) to improve robustness to shifts in the environment-conditioned non-invariant feature distribution $\mathcal{P}(NIe \mid E)$, where E denotes the environment. AMI couples non-invariant features across samples through controlled, structure-aware mixing, thereby smoothing environment-conditioned variations in the latent space.

In-batch Pairing. For each sample i in a mini-batch, we randomly sample a partner index $j \neq i$ and obtain non-invariant embeddings (NIe_i, NIe_j). This pairing constructs a mixed embedding $\tilde{NIe}_i$ used for training, while leaving the invariant embedding Ie_i unchanged.

Soft Mixing via Adaptive Normalization. We first introduce a soft mixing strategy based on Adaptive Instance Normalization (AdaIN) (Huang & Belongie, 2017; Zhou et al., 2021). We interpolate the feature statistics of two embeddings:

$$\gamma_{\text{mix}} = \lambda\sigma(NIe_i) + (1-\lambda)\sigma(NIe_j), \tag{6}$$

$$\beta_{\text{mix}} = \lambda\mu(NIe_i) + (1-\lambda)\mu(NIe_j), \tag{7}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$. The mixed embedding is

$$\tilde{NIe}_i = \gamma_{\text{mix}} \frac{NIe_i - \mu(NIe_i)}{\sigma(NIe_i) + \varepsilon} + \beta_{\text{mix}}. \tag{8}$$

By interpolating feature statistics while retaining the normalized source features, soft mixing perturbs sample-specific statistics without directly interpolating the full embeddings.

Hard Mixing via Embedding Interpolation. We further consider a stronger mixing variant that directly interpolates embeddings:

$$\tilde{NIe}_i = \lambda NIe_i + (1-\lambda)NIe_j, \tag{9}$$

which explicitly enforces convex interpolation between non-invariant features from different samples, thereby mitigating over-reliance on environment-specific feature modes. In practice, soft mixing stabilizes feature statistics, while hard mixing imposes stronger cross-sample coupling; the two variants provide complementary regularization effects.

3.4 Training Objective and Integration with VLA

Fusion and Prediction. Given the invariant embedding Ie_i and the mixed non-invariant embedding $\tilde{NIe}_i$, we fuse them as $h_i = [Ie_i \,\|\, \tilde{NIe}_i]$, and feed h_i into the original policy head $\pi(\cdot)$ for action prediction. This design preserves the base policy interface and requires no architectural changes.

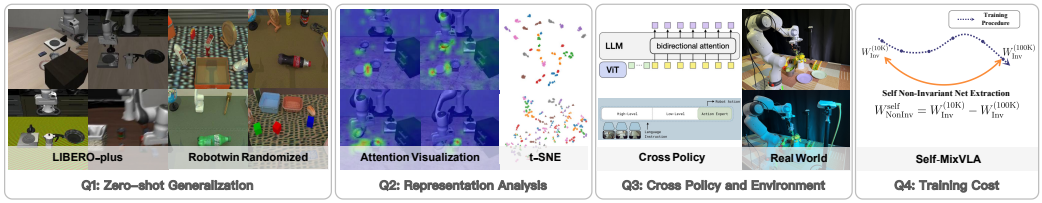


Figure 3: **Experimental Setup.** Our experiments cover zero-shot generalization (Q1), representation analysis (Q2), cross-policy and real-world applicability (Q3), and training efficiency with Self-MixVLA (Q4).

Table 1: **Performance on the LIBERO-Plus benchmark (out-of-distribution evaluation).**

Method	Camera	Robot	Language	Light	Background	Noise	Layout	Total
OpenVLA (Kim et al., 2024)	0.8	3.5	23.0	8.1	34.8	15.2	28.5	15.6
WorldVLA (Cen et al., 2025)	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
NORA (Hung et al., 2025)	2.2	37.0	65.1	45.7	58.6	12.8	62.1	39.0
UniVLA (Bu et al., 2025)	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
π_0 (Black et al., 2024)	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
π_0-Fast (Pertsch et al., 2025)	**65.1**	21.6	61.0	73.2	73.2	74.4	68.8	61.6
RIPT-VLA (Tan et al., 2025)	55.2	31.2	77.6	88.4	91.6	73.5	74.2	68.4
OpenVLA-OFT (Kim et al., 2025)	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
MixVLA	49.6	**63.3**	**82.6**	**96.1**	**93.9**	**82.3**	**76.8**	**76.2**

Stage-2 Training Objective. During the AMI training, we optimize the base VLA objective on the fused representation:

$$\mathcal{L}_{\text{MixVLA}} = \mathbb{E}_{(X,A)}\big[\mathcal{L}_{\text{base}}(\pi(h), A)\big], \tag{10}$$

where h depends on AMI through $\tilde{NIe}$. By exposing the policy head to the invariant structure together with regularized non-invariant cues, the model learns to reduce reliance on sample-specific correlations while retaining useful predictive information.

4 Experiment

Our experiments assess whether regularizing, rather than discarding, non-invariant information improves zero-shot robustness to distribution shifts while retaining strong in-domain performance. Beyond overall performance, we investigate the role of non-invariant information in generalization, the applicability of MixVLA across policy architectures and real-world settings, and the potential to reduce training requirements while preserving robustness gains As outlined in Figure 3, our experiments address the following four questions:

Q1. Robust Generalization. Can MixVLA improve zero-shot robustness to unseen distribution shifts while retaining strong in-domain performance?

Q2. Representation Analysis. What does regularized non-invariant information contribute beyond invariant features alone?

Q3. General Applicability. Do the robustness gains extend across policy architectures and to real-world manipulation?

Q4. Training Cost. Can MixVLA retain its robustness benefits using a single Invariant Net training run, without training a separate standard policy?

Benchmarks. We evaluate MixVLA across simulation and real-world manipulation settings. For simulation, we train on LIBERO (Liu et al., 2023) and evaluate both in-domain performance and zero-shot robustness under distribution shifts using LIBERO-Plus (Fei et al., 2025b). We further conduct clean-to-clean (C2C) and clean-to-randomized (C2R) evaluations on RoboTwin (Chen et al., 2025), where policies are trained only with clean demonstrations to assess generalization to unseen randomized environments. Finally, we deploy policies on real robots and evaluate robustness.

Implementation details. We implement MixVLA on two representative VLA architectures: OpenVLA-OFT (Kim et al., 2025), which predicts continuous action chunks through

Table 2: **Comparison on the LIBERO benchmark (in-domain setting).**

Model	Spatial	Object	Goal	Long	Avg
OpenVLA (Kim et al., 2024)	84.7	88.4	79.2	53.7	76.5
WorldVLA (Cen et al., 2025)	85.6	89.0	82.6	59.0	79.1
NORA (Hung et al., 2025)	92.2	95.4	89.4	74.6	87.9
UniVLA (Bu et al., 2025)	96.5	96.8	95.6	92.0	95.2
π_0 (Black et al., 2024)	90.0	86.0	95.0	73.0	86.0
π_0-Fast (Pertsch et al., 2025)	96.4	96.8	88.6	60.2	85.5
OpenVLA-OFT (Kim et al., 2025)	97.6	**98.4**	**97.9**	**94.5**	**97.1**
MixVLA	**98.4**	97.2	97.4	93.6	96.7

Table 3: **Clean-to-clean and clean-to-randomized evaluation on RoboTwin 2.0.**

Method	C2C	C2R	Avg.
Fast-WAM (Yuan et al., 2026)	77.8	1.9	39.9
X-VLA (Zheng et al., 2026)	68.0	20.9	44.5
X-WAM (Guo et al., 2026)	70.0	25.8	47.9
Spatial Forcing (Li et al., 2026)	77.2	26.7	52.0
$\pi_{0.5}$ (Intelligence et al., 2025)	70.7	46.0	58.4
4D-WAM (Yang et al., 2026)	**81.5**	41.8	61.6
MixVLA	75.6	**62.4**	**69.0**

parallel decoding, and $\pi_{0.5}$ (Intelligence et al., 2025), which generates continuous actions through flow matching. For the main simulation experiments, the Normal Net and the Invariant Net are each trained for 100K steps on LIBERO or clean RoboTwin demonstrations. In Stage 2, we perform AMI training for 1K steps on LIBERO and 10K steps on RoboTwin. Detailed training configurations are provided in the Appendix B.

4.1 Robustness to Unseen Distribution Shifts

Robustness from LIBERO to LIBERO-Plus. We build MixVLA on OpenVLA-OFT and train it on LIBERO, evaluating on LIBERO and LIBERO-Plus. As shown in Table 1, MixVLA achieves the highest overall success rate of 76.2% among the compared methods, surpassing OpenVLA-OFT by 6.6%. However, its lower performance under camera perturbations suggests that non-invariant feature mixing alone may be insufficient to address viewpoint-induced geometric shifts. On LIBERO, MixVLA retains strong in-domain performance, achieving an average success rate of 96.7% compared with 97.1% for OpenVLA-OFT (Table 2), with a decrease of only 0.4%.

Clean-to-randomized generalization on RoboTwin. We build MixVLA on $\pi_{0.5}$ and train it exclusively on clean demonstrations from RoboTwin 2.0. As shown in Table 3, MixVLA improves C2R success from 46.0% to 62.4% (+16.4%) over $\pi_{0.5}$, while also increasing C2C success from 70.7% to 75.6% (+4.9%).

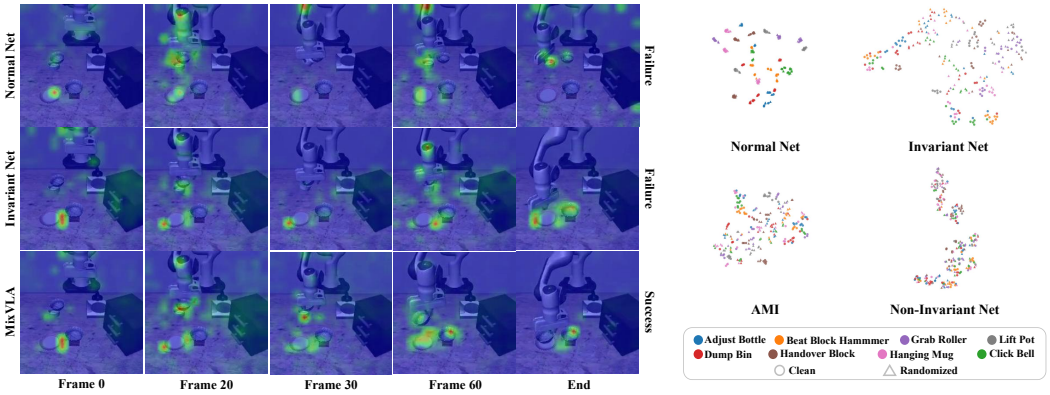
Taken together, these results support a positive answer to **Q1:** MixVLA improves robustness to a range of unseen distribution shifts while retaining strong in-domain performance.

4.2 Understanding the Role of Non-Invariant Information

Beyond invariant representations. We compare the Normal Net, Invariant Net, and full MixVLA, using OpenVLA-OFT on LIBERO-Plus and $\pi_{0.5}$ on RoboTwin (Table 6). IB-only (Invariant Net) improves overall OOD performance on both benchmarks, but reduces robustness under camera and language perturbations on LIBERO-Plus relative to the standard policy. This uneven behavior highlights the limitations of relying solely on invariant representations to handle diverse distribution shifts. Complementing invariant representations with AMI-regularized non-invariant features yields further gains of 4.0% on LIBERO-Plus and 2.8% on RoboTwin C2R over IB-only. MixVLA outperforms Invariant Net across all seven LIBERO-Plus perturbation categories while largely preserving clean-environment performance on RoboTwin. These results support our central premise that

Table 4: **Ablation study of IB and AMI on LIBERO-Plus and RoboTwin 2.0.** Experiments use OpenVLA-OFT on LIBERO-Plus and $\pi_{0.5}$ on RoboTwin 2.0.

IB	AMI	LIBERO-Plus								RoboTwin		
		Camera	Robot	Language	Lighting	Background	Noise	Layout	Total	C2C	C2R	Avg.
		51.2	30.7	74.2	87.8	92.6	76.0	73.7	67.5	73.8	53.8	63.8
✓		48.8	58.6	72.7	91.3	93.0	78.8	74.1	72.2	**75.8**	59.6	67.7
✓	✓	49.6	**63.3**	**82.6**	**96.1**	**93.9**	**82.3**	**76.8**	**76.2**	75.6	**62.4**	**69.0**



(a) Grad-CAM on LIBERO-Plus.

(b) t-SNE on RoboTwin.

Figure 4: **Qualitative analysis under distribution shifts.** (a) Grad-CAM visualizations along each policy's own rollout. (b) t-SNE projections of mean-pooled token features from eight tasks. Colors indicate tasks; circles and triangles denote clean and randomized observations, respectively.

non-invariant information need not be discarded: when regularized, it can complement invariant features and improve OOD generalization. Additional analyses of mixing strategies are provided in the Appendix C.

Visual attribution under distribution shifts. Figure 4a presents Grad-CAM visualizations along the respective rollouts of the Normal Net, Invariant Net, and MixVLA on a LIBERO-Plus task. The Invariant Net exhibits localized responses around manipulation-related regions, whereas MixVLA highlights these regions together with broader nearby context at intermediate steps. Both the Normal Net and the Invariant Net fail to complete the task, while MixVLA succeeds. Figure 4b further visualizes mean-pooled token features from eight RoboTwin tasks under clean and randomized conditions. The Normal Net forms predominantly task-homogeneous local clusters, whereas the Invariant Net projection exhibits a different organization with some regions still dominated by either clean or randomized samples. MixVLA retains task-associated neighborhoods alongside local clean–randomized interleaving, while the Non-Invariant Net exhibits substantial cross-task overlap. These projections illustrate changes in feature organization rather than establish complete domain alignment.

Together with the ablations in Table 6, these qualitative observations support our answer to **Q2**: regularized non-invariant information can complement invariant features and improve OOD generalization.

4.3 Generality across Policies and Real-World Deployment

Applicability across policy architectures. As shown in Tables 1 and 3, MixVLA improves robustness over the corresponding baselines for both OpenVLA-OFT and $\pi_{0.5}$. These gains are obtained using the same two-stage training framework, despite their different action-generation mechanisms: parallel action decoding in OpenVLA-OFT and flow matching in $\pi_{0.5}$. These results support the model-agnostic design of MixVLA, demonstrating its applicability beyond a single policy architecture.

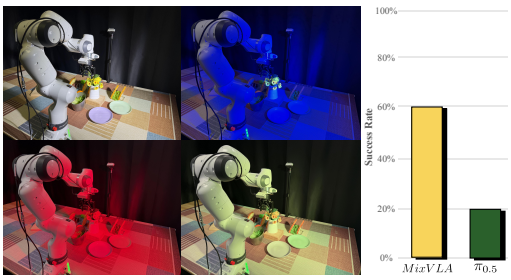


Figure 5: **Real-world robustness to lighting changes.**

Robustness in real-world manipulation. We deploy $\pi_{0.5}$ and its MixVLA counterpart on a Franka robot. We collect 80 demonstrations for a bread-placement task with two instruction variants, requiring the robot to pick up the sixth slice of bread and place it on either the near or far plate, as specified by the instruction. We evaluate each policy over 10 trials under lighting perturbations, counting a trial as successful when the specified slice is placed on the instructed plate. As shown in Figure 5, MixVLA achieves a success rate of 60%, compared with 20% for $\pi_{0.5}$. See Appendix C for videos of real-world executions.

These results complement the cross-architecture simulation findings and provide preliminary evidence that MixVLA's robustness benefits extend to real-world manipulation, supporting a positive answer to **Q3**.

4.4 Robustness from a Single Training Run

Reusing intermediate IB checkpoints. We explore whether intermediate checkpoints from a single Invariant Net training run can eliminate the need to train a separate Normal Net. Our intuition is that Invariant Net training progressively favors invariant information, and differences between early and late checkpoints may provide a proxy for complementary non-invariant information. Motivated by this hypothesis, we introduce **Self-MixVLA**, which constructs a *Self Non-Invariant Net* from the 10K- and 100K-step checkpoints of the same IB training run:

Table 5: **Single-run variant on RoboTwin 2.0.**

Method	C2C	C2R	Avg.
Self-MixVLA	73.0	60.8	66.9
MixVLA	**75.6**	**62.4**	**69.0**

$$W_{\text{NonInv}}^{\text{self}} = W_{\text{Inv}}^{(10\text{K})} - W_{\text{Inv}}^{(100\text{K})}. \tag{11}$$

The resulting branch is paired with the 100K-step Invariant Net for subsequent AMI training, avoiding a separate Normal Net training run.

These results support a positive answer to **Q4**: Self-MixVLA reduces training requirements by eliminating the separate Normal Net training run, while retaining robustness benefits with a modest performance reduction relative to full MixVLA (Table 5). We regard this variant as a preliminary exploration of reusing a single IB training trajectory, with further opportunities to investigate checkpoint selection and extrapolation strategies for recovering complementary information.

5 Conclusion

In this work, we presented **MixVLA**, a model-agnostic framework that improves VLA generalization by regularizing rather than discarding non-invariant information. MixVLA learns invariant representations through an information bottleneck objective and derives a complementary branch through weight arithmetic. **AMI** regularizes non-invariant features through stochastic mixing before fusing them with invariant representations for action prediction. Experiments demonstrate improved overall zero-shot robustness while retaining strong in-domain performance. Our findings suggest that non-invariant information can provide useful predictive cues when appropriately regularized, offering a practical direction for generalizable embodied policies without additional OOD training data.

AI use statement

In this work, we used generative AI tools primarily to improve the clarity and readability of the manuscript and to assist in identifying potential errors and inconsistencies in the text, equations, and numerical comparisons. AI-assisted suggestions were reviewed by the authors and revised as appropriate before incorporation into the manuscript. We take full responsibility for the final content of this work, including all text, claims, analyses, and results.

References

Ali Abouzeid, Malak H. I. Mansour, Zezhou Sun, and Dezhen Song. Geoaware-vla: Implicit geometry aware vision-language-action model. *ArXiv*, abs/2509.14117, 2025.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations, 2018. URL `https://arxiv.org/abs/1706.01350`.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Vineet Bhat, Yu-Hsiang Lan, Prashanth Krishnamurthy, Ramesh Karri, and Farshad Khorrami. 3d cavla: Leveraging depth and 3d context to generalize vision language action models for unseen tasks, 2026. URL `https://arxiv.org/abs/2505.05800`.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. $\pi 0$: A vision-language-action flow model for general robot control. *ArXiv*, abs/2410.24164, 2024.

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.

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, Deli Zhao, and Hao Chen. Worldvla: Towards autoregressive action world model. *ArXiv*, abs/2506.21539, 2025.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *ArXiv*, abs/2109.13396, 2021.

Senyu Fei, Siyin Wang, Li Ji, Ao Li, Shiduo Zhang, Liming Liu, Jinlong Hou, Jingjing Gong, Xianzhong Zhao, and Xipeng Qiu. Srpo: Self-referential policy optimization for vision-language-action models. *ArXiv*, abs/2511.15605, 2025a.

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*, 2025b.

Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026.

Wenkai Guo, Guanxing Lu, Haoyuan Deng, Zhenyu Wu, Yansong Tang, and Ziwei Wang. Vla-reasoner: Empowering vision-language-action models with reasoning via online monte carlo tree search. *ArXiv*, abs/2509.22643, 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*, abs/2507.16815, 2025.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U-Xuan Tan, Navonil Majumder, and Soujanya Poria. Nora: A small open-sourced generalist vision language action model for embodied tasks, 2025. URL `https://arxiv.org/abs/2504.19854`.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Rich Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. π0.5: a vision-language-action model with open-world generalization. *ArXiv*, abs/2504.16054, 2025.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning*, 2024.

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

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

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent ¡sup¿*¡/sup¿. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124002, December 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc62b. URL `http://dx.doi.org/10.1088/1742-5468/abc62b`.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, Winson Han, Wilbert Pumacay, Angelica Wu, Rose Hendrix, Karen Farley, Eli VanderBilt, Ali Farhadi, Dieter Fox, and Ranjay Krishna. Molmoact: Action reasoning models that can reason in space. *ArXiv*, abs/2508.07917, 2025.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 132324–132345, 2026.

Hao Li, Shuai Yang, Yilun Chen, Xinyi Chen, Xiaoda Yang, Yang Tian, Hanqing Wang, Tai Wang, Dahua Lin, Feng Zhao, et al. Cronusvla: Towards efficient and robust manipulation via multi-frame vision-language-action modeling. *arXiv preprint arXiv:2506.19816*, 2025a.

Xiaoqi Li, Liang Heng, Jiaming Liu, Yan Shen, Chenyang Gu, Zhuoyang Liu, Hao Chen, Nuowei Han, Renrui Zhang, Hao Tang, Shanghang Zhang, and Hao Dong. 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2344–2359. PMLR, 27–30 Sep 2025b.

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

Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. *ArXiv*, abs/2505.18719, 2025.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

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, 2025. URL `https://arxiv.org/abs/2501.09747`.

Hyunwoo Ryu, Hong in Lee, Jeong-Hoon Lee, and Jongeun Choi. Equivariant descriptor fields: Se(3)-equivariant energy-based models for end-to-end visual robotic manipulation learning, 2023.

Joohwan Seo, Nikhil P. S. Prakash, Xiang Zhang, Changhao Wang, Jongeun Choi, Masayoshi Tomizuka, and Roberto Horowitz. Contact-rich se(3)-equivariant robot manipulation task learning via geometric impedance control. *IEEE Robotics and Automation Letters*, 9(2):1508–1515, February 2024. ISSN 2377-3774.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models, 2025. URL `https://arxiv.org/abs/2505.17016`.

Generalist AI Team. Gen-0: Embodied foundation models that scale with physical interaction. *Generalist AI Blog*, 2025. https://generalistai.com/blog/nov-04-2025-GEN-0.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL `https://arxiv.org/abs/physics/0004057`.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Guodong Wang, Chenkai Zhang, Qingjie Liu, Jinjin Zhang, Jiancheng Cai, Junjie Liu, and Xinmin Liu. Libero-x: Robustness litmus for vision-language-action models. *arXiv preprint arXiv:2602.06556*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Shijie Wu, Yihang Zhu, Yunao Huang, Kaizhen Zhu, Jiayuan Gu, Jingyi Yu, Ye Shi, and Jingya Wang. Afforddp: Generalizable diffusion policy with transferable affordance, 2025. URL https://arxiv.org/abs/2412.03142.

Jingyun Yang, Zi ang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. Equibot: Sim(3)-equivariant diffusion policy for generalizable and data efficient learning, 2024a.

Jingyun Yang, Congyue Deng, Jimmy Wu, Rika Antonova, Leonidas Guibas, and Jeannette Bohg. Equivact: Sim(3)-equivariant visuomotor policies beyond rigid object manipulation, 2024b.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, et al. 4d-wam: Infusing spatiotemporal awareness into world action models through trajectory fields. *arXiv preprint arXiv:2608.08023*, 2026.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026. URL https://arxiv.org/abs/2603.16666.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration, 2025a.

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. *ArXiv*, abs/2507.04447, 2025b.

Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Siwei Han, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. Grape: Generalizing robot policy via preference alignment, 2025c.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models, 2025. URL https://arxiv.org/abs/2503.22020.

Zhaxizhuom Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, et al. Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. In *Conference on Robot Learning*, pp. 3069–3093. PMLR, 2025.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *ArXiv*, abs/2403.09631, 2024.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 60580–60606, 2026.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daum'e, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *ArXiv*, abs/2412.10345, 2024.

Zhide Zhong, Haodong Yan, Junfeng Li, Xiangchen Liu, Xin Gong, Tianran Zhang, Wenxuan Song, Jiayi Chen, Xinhu Zheng, Hesheng Wang, and Haoang Li. Flowvla: Visual chain of thought-based motion reasoning for vision-language-action models, 2025.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle, 2021. URL https://arxiv.org/abs/2104.02008.

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.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

MixVLA: Adaptive Mixing of Non-Invariant Information for Generalizable Vision-Language-Action Models

Supplementary Material

Abstract

This supplementary material expands upon our main paper by providing additional theoretical analysis, implementation details, experimental results, and discussions to further support the proposed method MixVLA.
▷ **Sec. A**: We analyze the compression effect of IB, the statistical properties of adaptive mixing, its conditional OOD benefit in an illustrative regression model, and local-linearization and kernel interpretations of MixVLA.
▷ **Sec. B**: We present additional implementation details, including the hardware setup, training time, and hyperparameter settings used in the mixing process.
▷ **Sec. C**: We report additional experiments and qualitative results, including ablations on mixing strategies, more visualizations on LIBERO, LIBERO-Plus, RoboTwin, and real-world demonstration videos.
▷ **Sec. D**: We discuss the limitations of the current framework and outline several possible directions for future research.

A Theoretical Analysis

This section analyzes the information-compression effect of the Information Bottleneck (IB), the statistical properties of adaptive mixing, and a local-linearization interpretation of MixVLA. We also present an illustrative regression model in which mixing improves OOD prediction under explicit conditions. These analyses characterize aspects of the proposed framework rather than establish unconditional guarantees for nonlinear VLA policies.

Throughout this section, $I(\cdot;\cdot)$ and $I(\cdot;\cdot \mid \cdot)$ denote mutual information and conditional mutual information, respectively. For brevity, we write $T = Ie$ and $N = NIe$. Importantly, the learned complementary representation N is not assumed to contain only nuisance information. All information quantities and moments used below are assumed finite. Language instructions are treated as fixed context and omitted from the notation.

A.1 Invariant Representation via Information Bottleneck

The IB objective balances input compression against predictive information retention (Tishby et al., 2000). Related analyses connect representation minimality with insensitivity to nuisance factors under suitable assumptions (Achille & Soatto, 2018). Here, we characterize the conditional nuisance information controlled by the compression objective.

Assume the observation is generated as

$$X = S(Z_{\text{inv}}, Z_{\text{ninv}}), \tag{12}$$

where Z_{inv} contains task-relevant latent factors and Z_{ninv} contains environment-dependent factors. Assume

$$A \perp X \mid Z_{\text{inv}}, \tag{13}$$

and let T be sampled from an encoder $p_\phi(T \mid X)$, whose sampling randomness is independent of the other variables conditional on X.

We consider the constrained IB objective

$$\min_{p_\phi(T|X)} I(X;T) \quad \text{s.t.} \quad I(T;A) \geq \kappa, \tag{14}$$

where $\kappa > 0$ is chosen so that the constraint is feasible.

Proposition A.1 (An upper bound on conditional nuisance information). Under the assumptions above,

$$\begin{aligned} I(T; Z_{\text{ninv}} \mid Z_{\text{inv}}) &= I(X;T) - I(T; Z_{\text{inv}}) \\ &\le I(X;T) - I(T;A). \end{aligned} \tag{15}$$

Consequently, any feasible encoder in Eq. equation 14 satisfies

$$I(T; Z_{\text{ninv}} \mid Z_{\text{inv}}) \le I(X;T) - \kappa. \tag{16}$$

Proof. Since X is a deterministic function of $(Z_{\text{inv}}, Z_{\text{ninv}})$ and the encoder depends on these factors only through X,

$$I(X;T) = I(Z_{\text{inv}}, Z_{\text{ninv}}; T). \tag{17}$$

Applying the chain rule gives

$$I(X;T) = I(Z_{\text{inv}}; T) + I(Z_{\text{ninv}}; T \mid Z_{\text{inv}}). \tag{18}$$

Moreover, the action-sufficiency assumption and the encoder construction imply the Markov chain $A \leftrightarrow Z_{\text{inv}} \leftrightarrow T$. The data processing inequality therefore yields

$$I(T;A) \le I(T; Z_{\text{inv}}). \tag{19}$$

Combining these relations proves the result. □

Interpretation. At a prescribed level of predictive information, reducing $I(X;T)$ tightens an upper bound on the information retained about environment-dependent factors beyond the task-relevant latent factors. This provides an information-theoretic motivation for IB regularization, but does not imply that a practical encoder eliminates all nuisance dependence or generalizes to arbitrary unseen environments.

The ideal latent factors Z_{inv} should also be distinguished from the learned representation T. Finite-capacity encoders and practical IB optimization need not make T sufficient for action prediction. A complementary representation can therefore retain useful information that is missing from T, motivating regularization rather than unconditional removal of the complementary branch.

A.2 Generalization Benefit of Adaptive Mixing

We first characterize the moments of hard mixing and then identify conditions under which it improves prediction in an illustrative regression model.

Moments of hard mixing. Let N_i and N_j be independent samples from the same source feature distribution, with mean μ and covariance Σ. Let $\lambda \sim \text{Beta}(\alpha, \alpha)$, $\alpha > 0$, be independent of the samples, and define

$$\widetilde{N}_i = \lambda N_i + (1-\lambda) N_j. \tag{20}$$

Write

$$m_\alpha = \mathbb{E}[\lambda] = \frac{1}{2}, \qquad q_\alpha = \mathbb{E}\big[\lambda^2 + (1-\lambda)^2\big] = \frac{\alpha+1}{2\alpha+1}. \tag{21}$$

Then

$$\mathbb{E}[\widetilde{N}_i] = \mu, \qquad \text{Cov}(\widetilde{N}_i) = q_\alpha \Sigma. \tag{22}$$

To obtain Eq. equation 22, note that the conditional mean given λ is μ, while independence gives

$$\text{Cov}(\widetilde{N}_i \mid \lambda) = \big[\lambda^2 + (1-\lambda)^2\big]\, \Sigma. \tag{23}$$

The result follows from the law of total covariance. Since $1/2 < q_\alpha < 1$, mixing contracts the across-sample covariance in this population model.

This contraction alone does not establish improved prediction: useful signals can also be attenuated. The following model makes the resulting trade-off explicit.

An illustrative regression model. Consider a scalar action

$$A = T + C, \tag{24}$$

where T is the contribution captured by the invariant representation and C is a complementary predictive signal not captured by T. Assume the complementary feature is

$$N_{\text{src}} = C + \epsilon_{\text{src}}, \qquad N_{\text{tgt}} = C + \epsilon_{\text{tgt}}. \tag{25}$$

The variables T, C, and the noise terms are mutually independent and centered, with

$$\text{Var}(C) = v > 0, \qquad \text{Var}(\epsilon_{\text{src}}) = \sigma^2_{\text{src}}, \qquad \text{Var}(\epsilon_{\text{tgt}}) = \sigma^2_{\text{tgt}}. \tag{26}$$

The distributions of T and C remain unchanged between source and target environments.

We restrict attention to predictors

$$f_w(T, N) = T + wN, \tag{27}$$

so that invariant-only prediction corresponds to $w = 0$. ERM minimizes population squared error using N_{src}. AMI instead minimizes squared error using $\widetilde{N}_i$ from Eq. equation 20, while retaining the original target A_i. The partner feature is sampled independently, and unmixed features are used at evaluation.

Define

$$V_{\text{src}} = v + \sigma^2_{\text{src}}, \qquad V_{\text{tgt}} = v + \sigma^2_{\text{tgt}}, \qquad r_\alpha = \frac{m_\alpha}{q_\alpha} = \frac{2\alpha + 1}{2\alpha + 2}. \tag{28}$$

Theorem A.2 (Conditional OOD benefit in the regression model). Under the assumptions above, the population-optimal coefficients are

$$w_{\text{ERM}} = \frac{v}{V_{\text{src}}}, \qquad w_{\text{AMI}} = r_\alpha w_{\text{ERM}}. \tag{29}$$

If

$$\frac{2V_{\text{src}}}{1 + r_\alpha} < V_{\text{tgt}} < \frac{2V_{\text{src}}}{r_\alpha}, \tag{30}$$

then AMI has strictly lower target squared error than both ERM and invariant-only prediction.

Proof. For the source distribution, $\mathbb{E}[N_{\text{src}}C] = v$ and $\mathbb{E}[N^2_{\text{src}}] = V_{\text{src}}$. Thus,

$$w_{\text{ERM}} = \frac{\mathbb{E}[N_{\text{src}}C]}{\mathbb{E}[N^2_{\text{src}}]} = \frac{v}{V_{\text{src}}}. \tag{31}$$

For mixed features, independence of the partner sample gives

$$\mathbb{E}[\widetilde{N}_i C_i] = m_\alpha v, \qquad \mathbb{E}[\widetilde{N}_i^2] = q_\alpha V_{\text{src}}. \tag{32}$$

This proves Eq. equation 29.

At evaluation, the target risk is

$$R_{\text{tgt}}(w) = \mathbb{E}\big[(C - wN_{\text{tgt}})^2\big] = v - 2vw + V_{\text{tgt}}w^2. \tag{33}$$

Let $w_0 = w_{\text{ERM}}$ and $r = r_\alpha$. The risk differences satisfy

$$R_{\text{tgt}}(rw_0) - R_{\text{tgt}}(w_0) = (1 - r)w_0\left[2v - (1 + r)w_0 V_{\text{tgt}}\right], \tag{34}$$

$$R_{\text{tgt}}(rw_0) - R_{\text{tgt}}(0) = rw_0\left[rw_0 V_{\text{tgt}} - 2v\right]. \tag{35}$$

Since $0 < r < 1$ and $w_0 = v/V_{\text{src}}$, both differences are negative under Eq. equation 30. □

Interpretation and scope. In this model, AMI shrinks reliance on the complementary feature without removing it. When target noise increases within the stated range, this balances the cost of using a noisy signal against the cost of discarding a useful signal. The result establishes a setting in which regularized complementary information is beneficial, rather than a universal ranking of AMI, ERM, and invariant-only prediction.

The model assumes independent pairing, population optimization, a linear predictor, and a particular noise-variance shift. It does not establish that real VLA features satisfy these assumptions, and the benefit for the proposed policies is evaluated empirically.

A.3 Special Cases of AMI

The two mixing strategies operate on different quantities. Hard mixing interpolates complete embeddings, whereas soft mixing interpolates their feature statistics.

Hard mixing. Equation equation 20 places the mixed representation on the line segment joining the two sampled embeddings. Under independent sampling, its covariance satisfies Eq. equation 22. This result concerns the source feature distribution; it does not imply averaging over unseen environments or preservation of every action-relevant feature. The OOD benefit in Theorem A.2 applies only under the additional regression assumptions stated there.

Soft mixing via adaptive normalization. Following the feature-statistic formulation of adaptive normalization and statistic mixing (Huang & Belongie, 2017; Zhou et al., 2021), let

$$\mu_i = \mu(N_i), \qquad \sigma_i = \sigma(N_i), \qquad R_i = \frac{N_i - \mu_i}{\sigma_i + \varepsilon}, \tag{36}$$

where the statistics are computed along the specified aggregation dimension and arithmetic is elementwise. Soft mixing forms

$$\mu_{\text{mix}} = \lambda\mu_i + (1-\lambda)\mu_j, \tag{37}$$

$$\sigma_{\text{mix}} = \lambda\sigma_i + (1-\lambda)\sigma_j, \tag{38}$$

and reconstructs

$$\widetilde{N}_i^{\text{soft}} = \sigma_{\text{mix}} \odot R_i + \mu_{\text{mix}}. \tag{39}$$

Define the statistic vector

$$s_i = \begin{bmatrix} \mu_i \\ \sigma_i \end{bmatrix}. \tag{40}$$

Its mixed counterpart is $s_{\text{mix}} = \lambda s_i + (1-\lambda)s_j$. If s_i and s_j are independent samples with mean μ_s and covariance Σ_s, the same moment calculation gives

$$\mathbb{E}[s_{\text{mix}}] = \mu_s, \qquad \text{Cov}(s_{\text{mix}}) = q_\alpha \Sigma_s. \tag{41}$$

Thus, soft mixing contracts the across-sample covariance of the statistic vectors in this idealized setting, while retaining the normalized source pattern R_i in the reconstruction. This does not imply that $\text{Cov}(\widetilde{N}_i^{\text{soft}})$ equals $q_\alpha \,\text{Cov}(N_i)$, because R_i can be statistically dependent on the source statistics. Preservation of the normalized pattern also does not, by itself, guarantee preservation of task semantics.

Soft and hard mixing therefore provide distinct forms of feature regularization. Their effectiveness for action prediction depends on the structure of the learned representations and is examined in our experiments.

A.4 NTK Interpretation of MixVLA

We distinguish a local interpretation of weight arithmetic from a kernel interpretation of feature mixing. The analysis uses the tangent-space perspective (Lee et al., 2020), but does not assume that invariant and non-invariant information occupy exactly separated parameter blocks.

Local linearization and weight arithmetic. Consider one scalar output of a network $f(x; W)$. Around a common reference point W_0, assume the network is twice differentiable with bounded Hessian in the relevant neighborhood. For a sufficiently small displacement Δ,

$$f(x; W_0 + \Delta) = f(x; W_0) + J_0(x)^\top \Delta + O(\|\Delta\|^2), \tag{42}$$

where $J_0(x) = \nabla_W f(x; W_0)$.

Write

$$W_{\text{Normal}} = W_0 + \Delta_{\text{Normal}}, \qquad W_{\text{Inv}} = W_0 + \Delta_{\text{Inv}}. \tag{43}$$

If both parameter sets remain in this neighborhood,

$$f(x; W_{\text{Normal}}) - f(x; W_{\text{Inv}}) = J_0(x)^\top (W_{\text{Normal}} - W_{\text{Inv}}) + O\big(\|\Delta_{\text{Normal}}\|^2 + \|\Delta_{\text{Inv}}\|^2\big). \tag{44}$$

Equation equation 44 relates a parameter difference to a difference in network outputs within a common local approximation. It does not establish

$$f(x; W_{\text{Normal}} - W_{\text{Inv}}) = f(x; W_{\text{Normal}}) - f(x; W_{\text{Inv}}), \tag{45}$$

nor does it prove that the resulting branch contains only non-invariant information. Weight subtraction in MixVLA is therefore a construction of a candidate complementary branch, whose utility is assessed empirically.

A linear-head kernel view of mixing. For a batch of B samples, let $U \in \mathbb{R}^{B\times d_I}$ and $V \in \mathbb{R}^{B\times d_N}$ contain fixed invariant and complementary features, respectively. Consider a trainable linear head

$$f_{a,b}(T, N) = a^\top T + b^\top N. \tag{46}$$

Its parameter-gradient kernel on the batch is

$$K_{\text{base}} = UU^\top + VV^\top = K_{\text{inv}} + K_{\text{noninv}}. \tag{47}$$

Let P be the pairing matrix, where $P_{i,j_i} = 1$ selects the partner of sample i and all other entries in that row are zero. Let $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_B)$ and define

$$M = \Lambda + (I_B - \Lambda)P. \tag{48}$$

Hard mixing produces the feature matrix $\widetilde{V} = MV$, while leaving U unchanged. The resulting linear-head kernel is exactly

$$K_{\text{mixed}} = K_{\text{inv}} + MK_{\text{noninv}}M^\top. \tag{49}$$

Consequently,

$$\mathbb{E}[K_{\text{mixed}}] = K_{\text{inv}} + \mathbb{E}\big[MK_{\text{noninv}}M^\top\big]. \tag{50}$$

The expectation depends on the pairing distribution, products of mixing coefficients, and cross-sample feature similarities. It is not generally equivalent to $K_{\text{inv}} + \mathbb{E}[\lambda]K_{\text{noninv}}$.

This restricted kernel view shows how mixing modifies the similarity structure contributed by the complementary branch while preserving the invariant feature block. It does not describe the complete NTK of a nonlinear VLA policy or establish an unconditional improvement in generalization.

B Implementation Details

Backbones and fine-tuning. We implement MixVLA on OpenVLA-OFT for LIBERO/LIBERO-Plus and on $\pi_{0.5}$ for RoboTwin and real-world manipulation. For OpenVLA-OFT, we adapt the pretrained 7B model using LoRA with rank 32, scaling parameter 16, and zero dropout, targeting linear layers. The base weights remain frozen, while the adapters, continuous action head, and proprioceptive projector are trainable. The model receives an external camera image, a wrist camera image, the language instruction, and proprioceptive state. It predicts chunks of eight 7-dimensional actions using an L_1 regression objective. Training uses image augmentation and quantile-based action and proprioception normalization.

For $\pi_{0.5}$, we fully fine-tune the policy, including the visual encoder, with the flow-matching objective. Inputs comprise three RGB views resized to 224×224, the instruction, and robot state; The action horizon is 50. Normalization statistics are computed from the corresponding training split. We use bfloat16 backbone computation, gradient checkpointing, and a training seed of 42. The OpenVLA experiments use eight NVIDIA A100 GPUs with 80 GB memory each. Training is distributed across eight GPUs. The global batch size is 64 for $\pi_{0.5}$; the OpenVLA configuration uses 64 for BC/IB and 48 for AMI, without gradient accumulation.

Optimization and schedules. Normal Net and Invariant Net are independently initialized from the pretrained backbone. AMI initializes its main branch from the selected Invariant Net checkpoint and uses a newly initialized fusion module and optimizer. OpenVLA uses AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and weight decay 10^{-2}. Its Normal Net and Invariant Net schedules have no warmup and reduce the learning rate by a factor of ten at step 100k. The AMI learning rate remains constant during its short training stage. For $\pi_{0.5}$, AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 10^{-10}, and gradient-norm clipping at 1.0. Linear warmup is followed by cosine decay to one tenth of the peak learning rate at the end of each run.

Information bottleneck. The OpenVLA implementation uses the Donsker–Varadhan MINE estimator, with an exponential-moving-average coefficient of 0.01 and three ReLU hidden layers of width 400. The critic receives pooled visual-prefix hidden states and flattened action-token hidden states. The policy IB coefficient is 10^{-3}. The critic is optimized separately with Adam at a constant learning rate of 5×10^{-6}, with its loss multiplied by 0.1.

The $\pi_{0.5}$ implementation uses a Jensen–Shannon critic with two SiLU hidden layers of width 512. Its inputs are the masked mean of context tokens and the mean of action-expert tokens. Marginal pairs are formed within each GPU's minibatch. The critic maximizes the bound, while gradient reversal encourages the policy to minimize it. The selected configuration uses an IB coefficient of 0.01 and a constant critic learning rate of 10^{-4}. The critic uses AdamW with the same coefficients and clipping threshold as the policy, with one critic update per policy update.

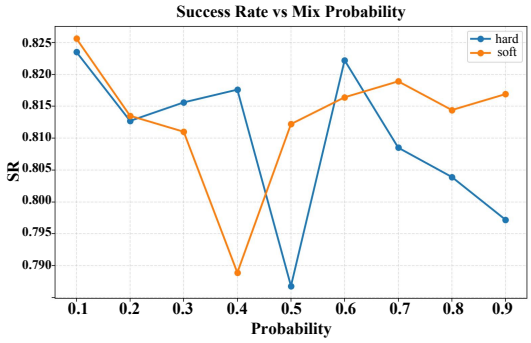


Figure 6: **Ablation of mixing probability.**

C More Experiments

Table 6: **Comparison of hard mixing and soft mixing on LIBERO-Plus.** We compare two variants of Adaptive Mixing of Non-Invariant Information (AMI): *hard mixing*, which performs direct interpolation between non-invariant embeddings, and *soft mixing*, which mixes feature statistics through adaptive normalization. Results are reported as task success rates (%) under different perturbation types.

Hard	Soft	Camera	Robot	Language	Light	Background	Noise	Layout	Total
✓		49.0	61.7	82.3	95.5	**93.9**	82.1	**76.8**	77.3
	✓	**49.6**	**63.3**	**82.6**	**96.1**	93.2	**82.3**	76.1	**77.6**

Ablation of mixing strategies. We further investigate the impact of different mixing strategies used in AMI. Specifically, we compare two variants: *hard mixing*, which directly interpolates non-invariant embeddings, and *soft mixing*, which performs statistic-level mixing through adaptive normalization.

As shown in Table 6, both strategies improve robustness under distribution shifts on the LIBERO-Plus benchmark. Hard mixing introduces stronger cross-sample coupling by explic-

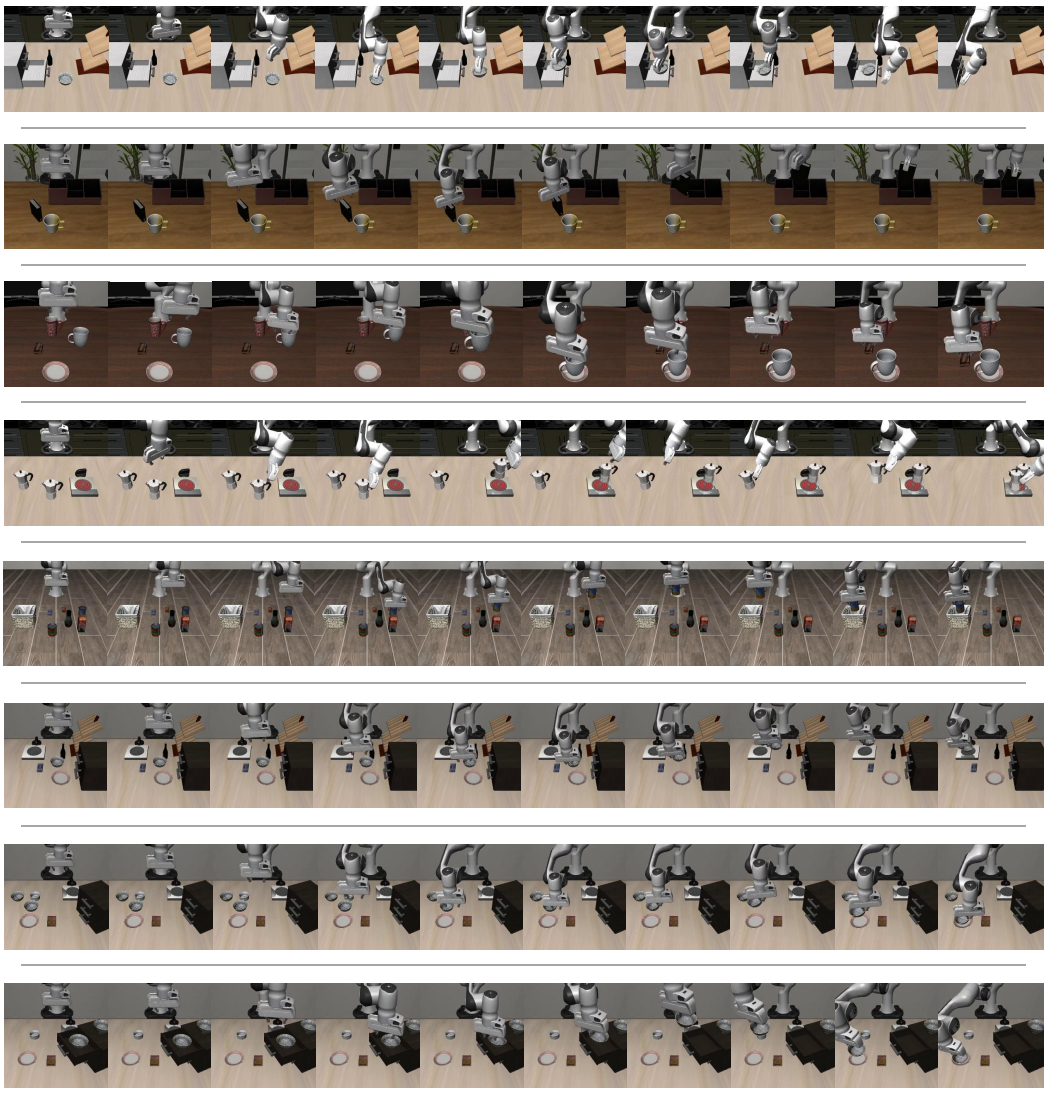


Figure 7: **Visualization of manipulation trajectories on the LIBERO benchmark.** Each row presents a sequence of key frames from a representative task execution, illustrating the manipulation progress from the initial state to task completion. Different rows correspond to different manipulation tasks and scene configurations, demonstrating the policy behavior under the standard in-domain setting.

itly interpolating latent representations, which helps suppress environment-specific feature modes. In contrast, soft mixing stabilizes feature distributions by aligning feature statistics across samples, leading to slightly better overall performance.

Overall, the results indicate that both strategies are effective for regularizing non-invariant representations, while soft mixing achieves the best average performance across perturbation types.

Effect of mixing strategies and probability. We compare hard and soft mixing on a LIBERO-Plus subset containing 2,318 evaluation episodes, varying the mixing probability from 0.1 to 0.9. As shown in Figure 6, both strategies exhibit a non-monotonic response to the mixing probability, with pronounced performance drops at intermediate values of 0.4 for soft mixing and 0.5 for hard mixing. Soft mixing exhibits smaller fluctuations in the higher-probability range of 0.6–0.9, whereas hard mixing shows a more substantial decline.

Both strategies achieve their highest observed success rates at a mixing probability of 0.1. These results indicate that increasing the mixing probability does not necessarily improve robustness. Instead, the effectiveness of non-invariant feature regularization depends on the interaction between the mixing strategy and its probability, highlighting the importance of considering these choices jointly.

Visualization on LIBERO benchmark. As shown in 7, we visualize representative manipulation trajectories of the model on the LIBERO benchmark, illustrating the policy behavior during task execution under the standard in-domain setting.

Visualization on LIBERO-plus benchmark. As shown in 9, we visualize representative manipulation trajectories of the model on the LIBERO-Plus benchmark, illustrating the policy behavior under diverse observation perturbations.

Visualization on RoboTwin 2.0 benchmark. As shown in Figure 10, we visualize representative observations from RoboTwin 2.0 under the randomized evaluation setting used in our clean-to-randomized (C2R) experiments.



Figure 8: **Three-camera observations for real-world manipulation.** Representative views from one wrist-mounted camera and two side-view cameras used in our Franka experiments.

Real-world demo **Warning.** The middle segment of the supplementary video contains high-frequency flashing lights that may cause discomfort. Viewers with epilepsy or sensitivity to flashing lights should avoid this segment.

For the real-world experiments, we deploy $\pi_{0.5}$ and MixVLA on a Franka robot using three cameras: one wrist-mounted camera and two side-view cameras. Figure 8 shows representative observations from the three camera views. The task requires the robot to pick up the sixth slice of bread and place it on the nearest or farthest plate, as specified by the language instruction. We consider two lighting perturbations shown in the supplementary video: low-frequency white flashing and high-frequency multicolored flashing. Each policy is evaluated over five trials per condition, for a total of ten trials. Under low-frequency white flashing, MixVLA succeeds in all five trials, whereas $\pi_{0.5}$ succeeds in two. High-frequency multicolored flashing is substantially more challenging for both policies: MixVLA succeeds in only one of five trials, while $\pi_{0.5}$ fails in all five. Aggregated across the two conditions, MixVLA achieves a success rate of 60%, compared with 20% for $\pi_{0.5}$.

The final segment of the video additionally presents our earlier implementation of the framework on Diffusion Policy (DP), referred to as *MixDP*, deployed on an xArm 6 robot. These executions provide a qualitative illustration of the framework's applicability to Diffusion Policy. Our investigation was initially motivated by early DP deployments, during which we observed frequent execution failures when ambient lighting differed from that during demonstration collection. These observations motivated us to investigate training strategies for improving policy robustness to visual distribution shifts.

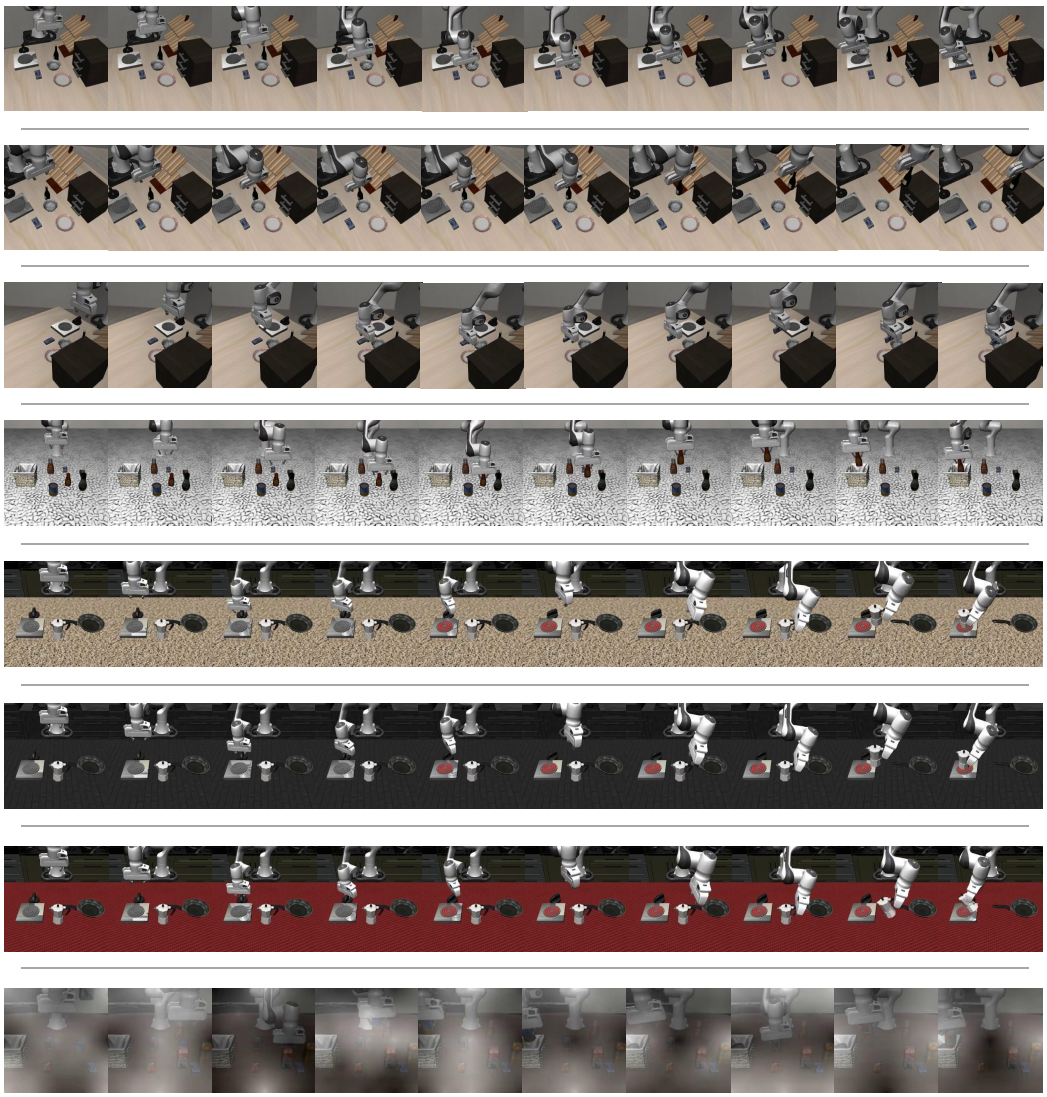


Figure 9: **Visualization of manipulation trajectories under LIBERO-Plus perturbations.** Each row presents a sequence of key frames from a representative task execution in the LIBERO-Plus benchmark. The scenes contain various observation perturbations such as changes in background textures, lighting conditions, and visual noise. The visualization illustrates how the policy performs manipulation under out-of-distribution environments from the initial state to task completion.

D Discussion

Limitations under Large Geometric Shifts. Limitations arise under perturbations that induce large geometric changes in observations, such as significant camera viewpoint shifts. As shown in our experiments, the improvement under the *Camera* perturbation is less pronounced compared with other appearance-related perturbations. This may be because viewpoint changes introduce discontinuous variations in the observation space, where latent-space vicinal smoothing becomes less effective. Addressing such geometric shifts may require incorporating explicit spatial reasoning or multi-view representations.

Future Research Directions. Several directions may further extend the proposed framework. First, more advanced mixing strategies could be explored to better capture the struc-

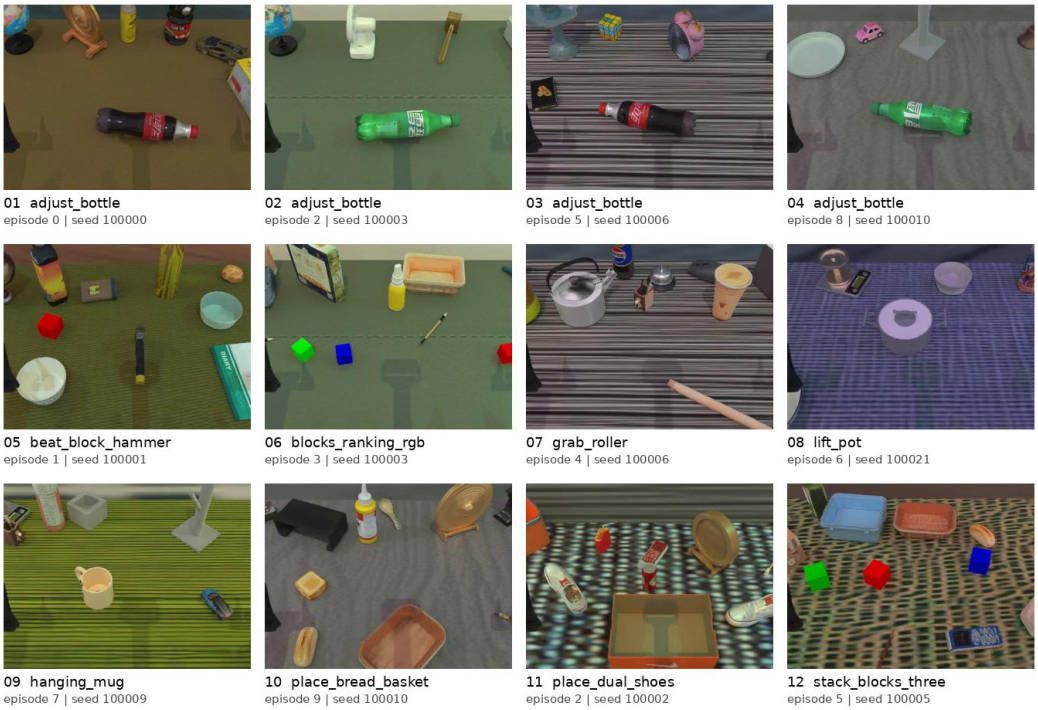


Figure 10: **Visualization of RoboTwin.**

ture of non-invariant representations. For instance, structure-aware or attention-guided mixing may provide finer control over how non-invariant information is regularized.

Second, integrating AMI with test-time adaptation or online representation updates may further improve robustness in highly dynamic environments.

Third, Self-MixVLA motivates further exploration of the model's own optimization trajectory for improving generalization. Future work could investigate how to identify and reuse parameter-update directions across training checkpoints to construct more effective complementary representations. Adaptive checkpoint selection and trajectory-based weight extrapolation may help narrow the robustness gap to full MixVLA while avoiding a separate Normal Net training run.