SMA-VLA: PROPRIOCEPTION-CONDITIONED SEMANTIC-MOTION ALIGNMENT FOR VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language-Action (VLA) models leverage pretrained Vision-Language Models (VLMs) to map visual observations and language instructions to robot actions, providing a promising paradigm for language-conditioned manipulation. However, strong in-distribution performance does not necessarily translate into robust behavior under distribution shifts, particularly when the robot starts from a different configuration, when an instruction is rephrased while preserving the same task semantics, or when visual layouts and distractors change. A key reason is that existing VLA models predominantly learn direct mappings from multimodal observations to expert actions, while how task semantics and current execution conditions influence the required motion remains insufficiently modeled. To address this problem, we propose **SMA-VLA**, a Vision-Language-Action framework centered on Proprioception-Conditioned Semantic–Motion Alignment (PC-SMA). The PC-SMA module incorporates the robot's current proprioceptive state into task conditioning and aligns the resulting task semantics with the coarse translational direction implied by expert demonstrations, thereby strengthening the otherwise implicit correspondence between task semantics and state-dependent motion. Multi-Level Context Enhancement integrates complementary information across VLM depths, providing richer semantic and spatial context for this alignment and reducing reliance on isolated layer representations. Patch-Focused Bridge uses the evolving action representation to emphasize action-relevant visual evidence, reducing interference from irrelevant scene factors and supporting more reliable motion prediction under visual shifts. Experiments on LIBERO, together with zero-shot transfer evaluations on LIBERO-Plus and LIBERO-Pro, show that SMA-VLA improves manipulation performance and enhances generalization under representative robot-state, instruction, and visual shifts.

1 INTRODUCTION

Vision-Language-Action (VLA) models extend pretrained Vision-Language Models (VLMs) to embodied control by mapping visual observations and language instructions to executable robot actions (Zitkovich et al., 2023; Kim et al., 2025b). Recent VLA models primarily follow two action-generation paradigms: autoregressive models that represent actions as discrete tokens and continuous-action models based on diffusion or flow matching (Qu et al., 2025; Liu et al., 2025; Bu et al., 2025). Despite their different generation mechanisms, most existing VLA models are trained with behavior cloning to reproduce expert actions conditioned on multimodal observations and language instructions (Black et al., 2025a; Chen et al., 2026). Scaling VLA models has substantially increased both model size and training cost. Conventional approaches typically rely on large VLM backbones and extensive robot datasets, using pretraining or end-to-end adaptation to learn the mapping from vision-language representations to robot actions (Luo et al., 2026b). VLA-Adapter instead introduces a lightweight bridge that leverages representations from multiple VLM layers and achieves competitive performance with a smaller model (Wang et al., 2026b). Beyond its efficiency, this result suggests that how hierarchical VLM representations are transferred to the action policy is itself important for control (Ni et al., 2026; Lin et al., 2026b).

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.

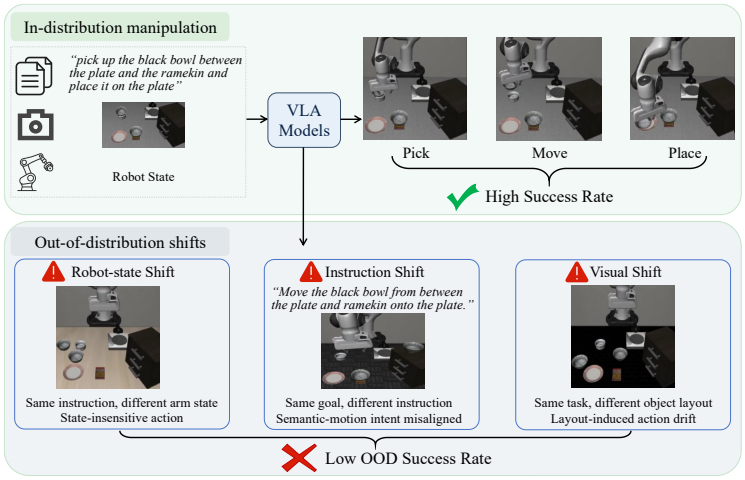

Figure 1: Representative out-of-distribution failures of VLA models under robot-state, instruction, and visual shifts.

Beyond model scaling, recent VLA research has increasingly focused on how pretrained multimodal representations are transferred and adapted for action generation (Kim et al., 2025a; Dalal et al., 2026; Xiao et al., 2026b). VLA-Adapter, for example, introduces a lightweight Policy Bridge that transfers representations from multiple VLM depths to the action policy, demonstrating that effective reuse of hierarchical VLM features can achieve competitive manipulation performance with a compact backbone (Wang et al., 2026b). Other approaches further improve representation adaptation through layer selection, representation preservation, or action-relevant visual modeling (Xiao et al., 2026b; Zhang et al., 2026b; Xiao et al., 2026a). These advances suggest that VLA performance depends not only on backbone capacity, but also on how pretrained multimodal representations are organized and adapted for control.

As shown in Figure 1, however, policies that perform well on in-distribution manipulation tasks can still fail when the robot configuration, instruction wording, or visual scene context changes (Fei et al., 2026; Zhou et al., 2025; Luo et al., 2026a). These failures reveal a limitation that is not fully addressed by representation transfer alone. Existing policies largely learn mappings from multimodal observations to expert actions within the training distribution, while how task semantics and current execution conditions influence the required motion remains insufficiently modeled. Under robot-state shifts, insufficient use of proprioceptive information can cause the same task intent to produce similar motion patterns even when different robot configurations require different motions. Under instruction shifts, different linguistic expressions should preserve the same underlying task semantics and maintain a consistent correspondence with the required motion. Under visual shifts, the task intent remains unchanged, but the policy must recover the action-relevant visual evidence needed to determine the appropriate motion despite changes in layout or distractors. Taken together, these cases expose a central bottleneck in VLA generalization: task semantics must be reliably aligned with the motion required under the current robot state, while remaining grounded in the visual evidence relevant to execution.

To address this problem, we propose **SMA-VLA**, a Vision-Language-Action framework centered on Proprioception-Conditioned Semantic–Motion Alignment. Its core mechanism, PC-SMA, incorporates the current proprioceptive state into task conditioning and explicitly encourages the resulting task representation to align with the coarse translational direction of demonstrated motion. To provide more reliable information for this alignment, Multi-Level Context Enhancement (MCE) integrates complementary semantic and spatial information across VLM depths while preserving layer-specific representations. Patch-Focused Bridge (PFB) further uses the evolving action representation to emphasize visual evidence relevant to the current motion and suppress task-irrelevant scene factors. In this way, PC-SMA provides an effective mechanism for state-conditioned semantic–motion alignment, while MCE and PFB provide complementary hierarchical context and action-relevant visual grounding to support it. Our contributions are summarized as follows:

- We propose **SMA-VLA**, a new Vision-Language-Action framework that strengthens the alignment between task semantics and the motion required under the robot's current execution state. Multi-Level Context Enhancement integrates complementary information across VLM depths, while Patch-Focused Bridge emphasizes visual evidence relevant to the evolving action.
- We introduce Proprioception-Conditioned Semantic–Motion Alignment, which incorporates the robot's current proprioceptive state into task conditioning and aligns the resulting task semantics with the motion direction implied by expert demonstrations. This design strengthens the otherwise implicit correspondence between task semantics and state-dependent motion, enabling the task semantics to better reflect the motion required under the current execution state and contributing to improved generalization.
- We evaluate **SMA-VLA** on LIBERO and further assess its zero-shot transfer on LIBERO-Plus and LIBERO-Pro. The results demonstrate consistent improvements in manipulation performance, with the largest gains observed under robot-state shifts and instruction rephrasings.

2 Related Work

Vision-Language-Action Models. Vision-Language-Action (VLA) models extend the multimodal understanding capabilities of pretrained Vision-Language Models (VLMs) to robot action generation. Early approaches mainly adopt autoregressive action prediction (Qu et al., 2025; Zhao et al., 2025; Hancock et al., 2026). RT-2 discretizes continuous actions into tokens and predicts them within the language-modeling space (Zitkovich et al., 2023), while OpenVLA scales this paradigm using large-scale robot demonstrations (Kim et al., 2025b). UniVLA further represents vision, language, and actions as discrete tokens within a unified autoregressive framework (Bu et al., 2025). Another line of work directly models continuous actions (Chi et al., 2023; Gao et al., 2026; Shukor et al., 2025). π_0 employs a flow-matching action expert to generate continuous action chunks (Black et al., 2025b), whereas OpenVLA-OFT combines action chunking with continuous action regression to improve efficiency and manipulation performance (Kim et al., 2025a). GR00T N1 couples vision-language understanding with continuous action generation to support learning across robot embodiments (Bjorck et al., 2025). These studies highlight the interface between vision-language representations and action generation as a central component of VLA design.

Multi-Level Representation Learning in VLA. Representations from different VLM depths provide complementary information for action generation. Prior studies suggest that intermediate representations tend to retain richer visual and spatial details, whereas deeper layers become increasingly biased toward abstract and language-aligned semantics, motivating the use of multi-level features for downstream control (Wang et al., 2026b; Lu et al., 2026; Luo et al., 2026c). VLA-Adapter transfers layer-wise VLM representations to the corresponding policy layers through Bridge Attention (Wang et al., 2026b). M^2-VLA instead employs a Mixture-of-Layers strategy to selectively extract task-relevant information across VLM depths (Xiao et al., 2026b). Action QFormer complements these approaches with instruction-conditioned queries that reorganize inherited multimodal information into action-oriented representations (Ji et al., 2026). DeepVision-VLA injects multi-level features from a vision expert into deeper VLA layers and uses action-guided pruning to retain relevant visual tokens (Luo et al., 2026c). While these methods improve the transfer and organization of hierarchical representations, explicit context integration across VLM depths remains relatively underexplored.

Generalization in VLA. Recent work improves VLA generalization by learning action-relevant visual representations and regulating multimodal adaptation (Fei et al., 2026; Zhou et al., 2025; Zhang et al., 2026a). Afford-VLA aligns intermediate VLA features with task-conditioned affordance representations, encouraging the policy to encode functional interaction regions without introducing an additional inference module (Wang et al., 2026a). FocusVLA uses cascaded and focused attention to emphasize relevant visual regions while suppressing redundant patches and background interference (Zhang et al., 2026b). StableVLA improves robustness to unseen visual disturbances through an information-bottleneck adapter without requiring additional training data (Fu et al., 2026). From a complementary perspective, LA4VLA learns reusable language-conditioned

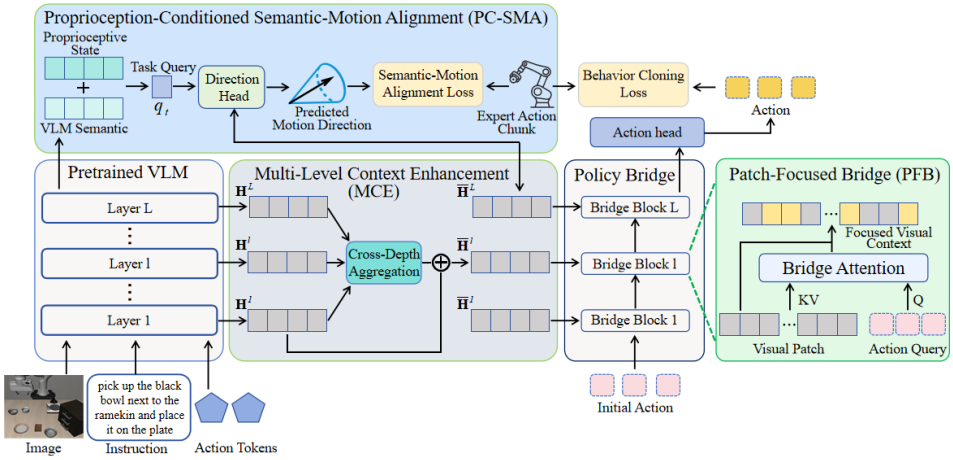

Figure 2: **Overall architecture of SMA-VLA.** MCE integrates position-aligned information across VLM depths. PC-SMA combines task semantics with proprioceptive state, conditions the visual hierarchy, and uses demonstrated motion direction as training-only supervision. PFB operates within the Policy Bridge to inject an action-conditioned focused visual residual for continuous action prediction. The motion-direction supervision branch is removed at inference.

action priors through language-action pretraining, reducing reliance on scene-specific visual cues (Lin et al., 2026a). Anchor-Align constrains representation drift through layer-wise anchoring and promotes language-action consistency using motion-direction supervision (Dalal et al., 2026). These approaches address complementary sources of distribution shift, but the coordinated modeling of hierarchical visual context, proprioceptive state, and continuous motion remains underexplored.

3 SMA-VLA

Overview. Building on VLA-Adapter, SMA-VLA retains its pretrained VLM, Policy Bridge, and continuous action head (Wang et al., 2026b), while modifying the information flow between these components. Given multi-view observations and a language instruction, the VLM produces hierarchical hidden states $\{\mathbf{H}_t^l\}_{l=1}^L$. MCE first enriches each layer with a shared cross-depth residual. PC-SMA then combines instruction semantics with the current proprioceptive state to form a task query $\mathbf{q}_t$, which conditions the MCE-enhanced visual tokens and is additionally supervised by demonstrated motion direction during training. The conditioned representations are passed to the corresponding Policy Bridge blocks, where PFB uses the evolving action hidden states to construct an action-conditioned visual residual. The final action head predicts continuous action chunks following the original VLA-Adapter formulation. The overall architecture of SMA-VLA is illustrated in Figure 2.

3.1 Multi-Level Context Enhancement

VLA-Adapter exposes representations from multiple VLM depths to the policy, but each layer is mainly transferred through its own Bridge interaction (Wang et al., 2026b; Xiao et al., 2026b). We instead make complementary information from the full hierarchy available before this transfer. Let $\mathbf{H}_t^l \in \mathbb{R}^{N\times D}$ denote the visual tokens at VLM layer l. MCE learns a depth prior and constructs the shared position-aligned context:

$$\mathbf{H}_t^{\mathrm{ctx}} = \sum_{l=1}^{L} \alpha_l \mathbf{H}_t^l, \tag{1}$$

where $\alpha_l = \mathrm{softmax}(p_l/\tau)$ is a learned layer weight. Because aggregation occurs only along depth, the spatial/token organization of every layer is preserved.

Simply replacing all layers with their mixture would erase the hierarchy we aim to exploit. We therefore use the final VLM layer as a semantic reference and inject only the shared cross-depth difference:

$$\widetilde{\mathbf{H}}_t^l = \mathbf{H}_t^l + g_{\mathrm{M}} \left(\mathbf{H}_t^{\mathrm{ctx}} - \mathbf{H}_t^L \right). \tag{2}$$

Here g_{M} is a bounded, zero-initialized scalar gate. Thus every layer receives the same position-aligned context residual while retaining its original layer-specific representation. The enhanced hierarchy $\{\widetilde{\mathbf{H}}_t^l\}$ is passed to PC-SMA.

3.2 Proprioception-Conditioned Semantic–Motion Alignment

MCE improves *what the visual hierarchy contains*; PC-SMA determines *how that information should be interpreted from the robot's current state*. The same instruction can demand different end-effector motion when the robot starts from different configurations, so task semantics should not be treated independently of proprioception (Dalal et al., 2026; Wang et al., 2026a).

State-conditioned task representation. We obtain a compact language condition $\mathbf{c}_t^{\mathrm{txt}}$ by masked mean pooling and projecting the instruction-associated VLM states. The current proprioceptive state $\mathbf{s}_t$ is separately projected into $\mathbf{c}_t^s$. A learned sample-dependent gate z_t controls the state contribution:

$$\mathbf{c}_t = \mathrm{LN} \left(\mathbf{c}_t^{\mathrm{txt}} + \rho_s z_t \mathbf{c}_t^s \right). \tag{3}$$

The fused condition is projected and normalized to obtain the task query $\mathbf{q}_t$. Language remains the semantic anchor, while proprioception adjusts the condition to the current physical configuration.

Task-conditioned visual relevance. For each MCE-enhanced visual token we compute a normalized key $\mathbf{k}_{t,n}^l$. The task query induces a within-layer relevance distribution:

$$r_{t,n}^l = \mathrm{softmax}_n \left(\mathbf{q}_t^\top \mathbf{k}_{t,n}^l \right). \tag{4}$$

We use this relevance to apply a bounded multiplicative modulation:

$$\overline{\mathbf{H}}_{t,n}^l = \left[1 + g_{\mathrm{T}} \tanh\!\big(N r_{t,n}^l - 1 \big) \right] \widetilde{\mathbf{H}}_{t,n}^l. \tag{5}$$

A uniform relevance distribution leaves the representation unchanged. PC-SMA therefore emphasizes task- and state-relevant evidence without deleting or reordering visual tokens. The resulting $\{\overline{\mathbf{H}}_t^l\}$ are transferred to the corresponding Policy Bridge blocks.

Training-time semantic–motion alignment. The task query should not only encode semantic intent; it should also reflect the coarse direction in which the robot must move. From the de-normalized expert action chunk, we form the demonstrated translational direction:

$$\mathbf{u}_t = \mathrm{Norm}_\epsilon \left(\sum_{i=0}^{K-1} \Delta \mathbf{x}_{t+i} \right). \tag{6}$$

Not all action chunks provide equally reliable directional supervision. We therefore estimate motion coherence from the ratio between net displacement and accumulated step-wise displacement. Coherent trajectories receive a narrower admissible cone; curved or nearly cancelling trajectories receive a wider cone and lower weight. A lightweight direction head predicts $\widehat{\mathbf{u}}_t$ from $\mathbf{q}_t$, and we penalize only predictions that fall outside this coherence-adaptive cone:

$$\mathcal{L}_{\mathrm{SMA}} = \frac{\sum_{t \in \mathcal{B}} w_t \left[\max\!\big(0,\, \cos\theta_t - \widehat{\mathbf{u}}_t^\top \mathbf{u}_t \big) \right]^2}{\max\!\big(\sum_{t \in \mathcal{B}} w_t, \epsilon \big)}. \tag{7}$$

Here θ_t is determined by motion coherence and w_t additionally downweights near-static chunks. The objective constrains only coarse translational direction; magnitude, rotation, gripper state, and intra-chunk dynamics remain supervised by the original action loss. Gradients from $\mathcal{L}_{\mathrm{SMA}}$ are stopped at the VLM representations, and the direction head is removed at inference.

3.3 PATCH-FOCUSED BRIDGE

After PC-SMA, the visual hierarchy is task- and state-conditioned, but the Policy Bridge can still aggregate visual evidence that is irrelevant to the current action phase (Hanyu et al., 2026; Xiao et al., 2026a). PFB therefore uses the evolving action representation itself to select visual evidence inside each eligible Bridge block.

At block b, action hidden states $\mathbf{X}_t^b$ are projected to queries, while conditioned visual tokens $\overline{\mathbf{H}}_{t,c}^b$ from view c are projected to keys and values. PFB computes action-to-patch attention:

$$\mathbf{A}_{t,c}^b = \mathrm{softmax}_n \left(\frac{\mathbf{Q}_t^b (\mathbf{K}_{t,c}^b)^\top}{\sqrt{d_h}} \right). \tag{8}$$

The resulting focused visual context is:

$$\mathbf{F}_t^b = \frac{1}{M} \sum_{c=1}^{M} \mathbf{A}_{t,c}^b \mathbf{V}_{t,c}^b. \tag{9}$$

To isolate focused evidence rather than simply increasing visual magnitude, we subtract a uniform reference $\mathbf{U}_t^b$ obtained by averaging valid visual values within each view. The focused residual augments the original Bridge context:

$$\widetilde{\mathbf{C}}_t^b = \mathbf{C}_t^b + g_b^{\mathrm{P}} \left(\mathbf{F}_t^b - \mathbf{U}_t^b \right). \tag{10}$$

The gate g_b^{P} is bounded and zero-initialized, so PFB begins from the original Bridge computation and progressively learns which visual patches are useful to the current action representation.

Training and inference. SMA-VLA preserves the original behavior-cloning objective and adds only the PC-SMA auxiliary loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{BC}} + \lambda_{\mathrm{SMA}} \mathcal{L}_{\mathrm{SMA}}. \tag{11}$$

During training, the demonstrated action chunk supervises the direction head in addition to the original action target. At inference, the demonstrated-motion construction, direction head, and $\mathcal{L}_{\mathrm{SMA}}$ are removed. MCE, the state-conditioned task query, PC-SMA visual modulation, PFB, and the original action head remain active.

4 EXPERIMENTS

We evaluate SMA-VLA from four complementary perspectives: standard manipulation performance on LIBERO (Liu et al., 2023), zero-shot transfer under controlled distribution shifts on LIBERO-Plus and LIBERO-Pro, the incremental contribution of each component, and representative rollout behavior (Fei et al., 2026; Zhou et al., 2025).

4.1 OVERALL PERFORMANCE

Benchmarks and baselines. LIBERO (Liu et al., 2023) contains four suites covering spatial reasoning, object manipulation, goal-conditioned behavior, and long-horizon tasks. We compare SMA-VLA with representative methods across three model scales. The large-model baselines include OpenVLA (Kim et al., 2025b), CoT-VLA (Zhao et al., 2025), MolmoAct (Lee et al., 2026), OpenVLA-OFT (Kim et al., 2025a), and UniVLA (Bu et al., 2025); the small-model baselines include SpatialVLA (Qu et al., 2025), SmolVLA (Shukor et al., 2025), VLA-0 (Goyal et al., 2025), $\pi_{0.5}$-KI (Black et al., 2025a; Driess et al., 2026), GR00T N1 (Bjorck et al., 2025), and π_0 (Black et al., 2025b); and the tiny-model baselines include Diffusion Policy (Chi et al., 2023), VLA-OS (Gao et al., 2026), and VLA-Adapter (Wang et al., 2026b). Among them, VLA-Adapter serves as our direct architectural baseline because SMA-VLA retains its backbone and Policy Bridge. Following the standard LIBERO protocol, we train a separate policy for each suite and evaluate every task over 50 episodes, yielding 500 rollouts for each 10-task suite. We report the task success rate and its arithmetic mean across the four suites. A rollout is considered successful only when it satisfies the task-specific completion criterion.

Table 1: Success rates (%) on the LIBERO benchmark. Methods are grouped by their backbone scale. Avg. is the arithmetic mean over the four LIBERO suites. **Bold** and underlined values denote the best and second-best result in each column, respectively.

Method	Publication	Params	Spatial	Object	Goal	Long	Avg.
Large models							
OpenVLA	CoRL 2024	7B	84.7	88.4	79.2	53.7	76.50
CoT-VLA	CVPR 2025	7B	87.5	91.6	87.6	69.0	83.93
MolmoAct	ICRA 2026	7B	87.0	95.4	87.6	77.2	86.80
OpenVLA-OFT	RSS 2025	7B	94.3	95.2	91.7	86.5	91.93
UniVLA	RSS 2025	7B	96.5	96.8	95.6	**92.0**	95.23
Small models							
SpatialVLA	RSS 2025	4B	88.2	89.9	78.6	55.5	78.05
SmolVLA	arXiv 2025	2.2B	93.0	94.0	91.0	77.0	88.75
VLA-0	arXiv 2025	3B	93.6	96.0	95.6	87.6	93.20
$\pi_{0.5}$-KI	NeurIPS 2025	3B	96.6	97.2	94.6	85.8	93.55
GR00T N1	arXiv 2025	2B	94.4	97.6	93.0	<u>90.6</u>	93.90
π_0	RSS 2025	3B	<u>96.8</u>	98.8	95.8	85.2	94.15
Tiny models							
Diffusion Policy	RSS 2023	–	78.3	92.5	68.3	50.5	72.40
VLA-OS	NeurIPS 2025	0.5B	87.0	96.5	92.7	66.0	85.60
VLA-Adapter	AAAI 2026	0.5B	96.0	**99.8**	<u>96.0</u>	89.0	<u>95.20</u>
SMA-VLA (Ours)	–	0.5B	**98.8**	99.2	**97.2**	89.6	**96.20**

Implementation details. We train SMA-VLA with distributed data parallelism on four NVIDIA H800 GPUs. The implementation uses Python 3.10.16, PyTorch 2.2.0, CUDA 12.1, FlashAttention 2.5.5, Transformers 4.40.1, LIBERO 0.1.0, robosuite 1.4.1, and MuJoCo 3.3.0. The model combines Prismatic-VLM, a Qwen2.5-0.5B language backbone, and the VLA-Adapter Policy Bridge. It receives two RGB views, a language instruction, and the current proprioceptive state, and predicts an eight-step continuous action chunk. We fine-tune the VLM using rank-64 LoRA. In the first stage, we train for 60K steps with a per-GPU batch size of 12, a global batch size of 48, and a learning rate of 2×10^{-4}. We then initialize PFB refinement from this checkpoint and continue training for 20K steps with the same global batch size.

Comparative experimental results. Table 1 shows that SMA-VLA achieves an average success rate of 96.2% with a 0.5B backbone, outperforming VLA-Adapter by 1.0 percentage point. SMA-VLA performs best on LIBERO-Spatial at 98.8% and LIBERO-Goal at 97.2%, while VLA-Adapter leads on LIBERO-Object and UniVLA on LIBERO-Long. The gains are thus more pronounced on spatial and goal-conditioned manipulation tasks. The relatively lower performance on LIBERO-Long may partly reflect the limited capacity of the 0.5B VLM backbone for long-horizon reasoning and sequential decision making.

4.2 Out-of-Distribution Generalization

For both LIBERO-Plus and LIBERO-Pro, all methods are trained only on the original LIBERO-Spatial training split and evaluated directly under distribution shifts without additional adaptation, following a zero-shot transfer protocol.

LIBERO-Plus. LIBERO-Plus (Fei et al., 2026) evaluates seven controlled shifts: camera viewpoint, robot initial state, instruction wording, lighting, background texture, sensor noise, and object layout. We compare SMA-VLA with OpenVLA-OFT (Kim et al., 2025a), VLA-Adapter (Wang et al., 2026b), and Anchor-Align VLA (Dalal et al., 2026). SMA-VLA achieves an overall success rate of 90.76% in Table 2, outperforming VLA-Adapter by 5.66 percentage points and achieving the highest overall success rate among the compared methods. The largest gains over VLA-Adapter occur under robot-state shifts at 64.86%, improving by 12.26 percentage points, and instruction shifts at 91.79%, improving by 6.69 percentage points. SMA-VLA also achieves the best performance under lighting changes at 99.66% and background changes at 99.61%, while matching the best score under object-layout shifts at 97.40%. Although Anchor-Align VLA provides a strong

Table 2: Success rates (%) on LIBERO-Plus under zero-shot transfer. Overall denotes the success rate computed over all evaluation trials across the seven distribution shifts.

Method	Camera	Robot	Language	Light	Background	Noise	Layout	Overall
OpenVLA-OFT	88.30	40.00	80.50	98.30	97.30	96.30	93.90	84.00
VLA-Adapter	92.60	52.60	85.10	93.20	90.70	89.50	93.20	85.10
Anchor-Align VLA	**96.30**	59.10	87.20	99.00	99.60	**96.90**	**97.40**	90.30
SMA-VLA (Ours)	91.22	**64.86**	**91.79**	**99.66**	**99.61**	93.73	**97.40**	**90.76**

generalization-oriented baseline, SMA-VLA attains a higher overall success rate. Its lower results under camera and noise perturbations suggest that robustness gains remain shift-dependent rather than uniform.

Table 3: Success rates (%) on LIBERO-Pro under zero-shot transfer.

Method	Lang. Reph.	Object Swap	Pos. Swap
MolmoAct	77.8	82.4	0.0
OpenVLA-OFT	74.4	95.2	0.0
VLA-Adapter	91.1	89.6	2.3
Anchor-Align VLA	**97.0**	96.2	**22.6**
SMA-VLA (Ours)	96.8	**96.4**	2.6

Table 4: Component ablations on LIBERO-Plus under zero-shot transfer. All variants are trained only on LIBERO-Spatial and directly evaluated under the seven LIBERO-Plus shifts. Overall denotes the success rate computed over all evaluation trials across the seven shifts.

Variant	MCE	PC-SMA	PFB	Camera	Robot	Language	Light	Background	Noise	Layout	Overall
baseline				92.60	52.60	85.10	93.20	90.70	89.50	93.20	85.10
MCE	✓			94.68	53.43	84.10	98.97	99.22	91.17	94.55	87.43
MCE + PC-SMA	✓	✓		89.36	**64.86**	90.51	99.32	98.84	94.59	96.88	90.17
MCE + PFB	✓		✓	**96.01**	61.43	81.54	99.32	98.45	**97.44**	**97.66**	89.76
SMA-VLA	✓	✓	✓	91.22	**64.86**	**91.79**	**99.66**	**99.61**	93.73	97.40	**90.76**

LIBERO-Pro. LIBERO-Pro (Zhou et al., 2025) separately evaluates instruction rephrasing, object replacement, and positional changes. We report the three dimensions individually and compare SMA-VLA against MolmoAct (Lee et al., 2026), OpenVLA-OFT (Kim et al., 2025a), VLA-Adapter (Wang et al., 2026b), and Anchor-Align VLA (Dalal et al., 2026). SMA-VLA improves over VLA-Adapter by 5.7 percentage points under instruction rephrasing and 6.8 percentage points under object replacement, as shown in Table 3, demonstrating stronger generalization to instruction and object-level variations. In contrast, SMA-VLA achieves only 2.6% under positional swaps, substantially below Anchor-Align VLA at 22.6%. This result indicates that SMA-VLA remains limited in handling severe spatial rearrangements.

4.3 Ablation Studies

We evaluate incremental variants under the same LIBERO-Spatial training and zero-shot transfer protocol. Starting from VLA-Adapter, we first add MCE, then combine MCE separately with PC-SMA or PFB, and finally evaluate the complete SMA-VLA model. All variants use the same backbone, optimizer family, action horizon, total number of optimization steps, and evaluation settings. Because PC-SMA and PFB are evaluated on top of MCE, the ablation measures their additional contributions beyond MCE. As shown in Table 4, MCE raises the overall success rate from 85.10% to 87.43%, showing the benefit of introducing cross-depth context before state or action conditioning. Adding PC-SMA further increases the overall success rate to 90.17%. Its largest gains over MCE occur under robot-state shifts, from 53.43% to 64.86%, and instruction rephrasings, from 84.10% to 90.51%, consistent with the role of state-grounded semantic–motion alignment. MCE+PFB reaches

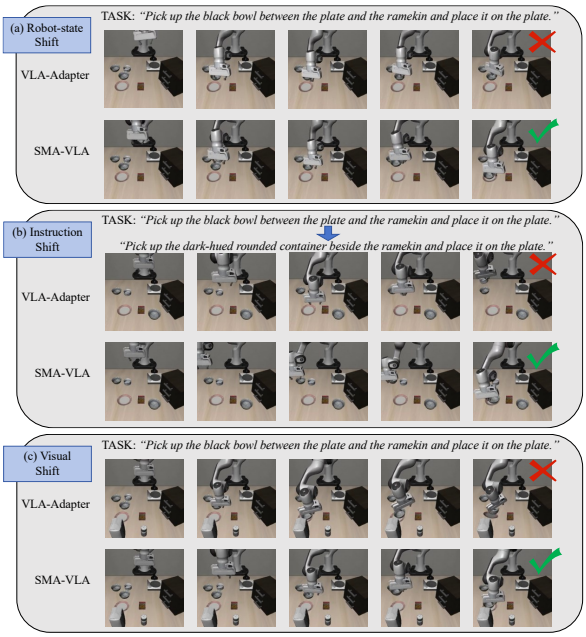

Figure 3: Qualitative rollouts. Each block shows the same manipulation task under a robot-state shift, an instruction rephrasing, or a visual shift. SMA-VLA successfully completes the illustrated trials, whereas VLA-Adapter fails.

89.76% and performs best among the partial variants under camera, noise, and layout changes, indicating the benefit of action-conditioned access to visual evidence. The complete model achieves the highest overall success rate of 90.76% and the strongest results under language, lighting, and background shifts. Overall, these results demonstrate that the three modules are complementary and that their combination yields the strongest overall generalization performance.

4.4 Qualitative Generalization Analysis

Figure 3 presents illustrative paired rollouts under robot-state, instruction, and visual distribution shifts. Although VLA-Adapter performs strongly under in-distribution conditions, its behavior becomes less reliable when the initial robot state, task wording, or visual scene changes. In contrast, SMA-VLA adapts its approach from a changed initial state, follows a rephrased instruction while preserving the original task goal, and completes the illustrated trial under distractors and layout changes. These qualitative results are consistent with the quantitative trends reported in Table 2 and Table 3.

5 Conclusion

We presented SMA-VLA, a Vision-Language-Action framework centered on proprioception-conditioned semantic–motion alignment. By incorporating the robot's current proprioceptive state into task conditioning and aligning task semantics with demonstrated motion, SMA-VLA strengthens the correspondence between task intent and state-dependent motion, while Multi-Level Context Enhancement and Patch-Focused Bridge provide complementary hierarchical context and action-relevant visual grounding. SMA-VLA achieves 96.2% average success on LIBERO and 90.76% on LIBERO-Plus, with the largest gains under robot-state shifts and instruction rephrasings. LIBERO-Pro further shows improved generalization to instruction rephrasing and object replacement, while severe positional swaps remain challenging. Overall, these results support semantic–motion alignment as an effective direction for improving VLA generalization.

AI USE STATEMENT

Generative AI tools were used only for literature exploration, reference checking, and language polishing of selected sentences in the final version of the manuscript. The research ideas, overall logic, methodological design and description, and experimental analysis were developed independently by the authors. All technical content, citations, equations, implementations, experimental results, and scientific claims were manually reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Our model and experiments are reproducible based on the methodological and experimental details provided in this paper. Section 3 specifies the architecture and mathematical formulations of MCE, PC-SMA, and PFB, while Section 4 reports the implementation details, training configurations, benchmark protocols, and evaluation settings used in our experiments. All experiments are conducted on the publicly available LIBERO, LIBERO-Plus, and LIBERO-Pro benchmarks following the stated evaluation protocols, and the reported results are obtained from the corresponding experimental runs. Upon acceptance, we will release the source code, configuration files, and evaluation scripts to facilitate independent reproduction of our results.

REFERENCES

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

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, pp. 17–40, 2025a.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, 2025b.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. In *Proceedings of Robotics: Science and Systems*, 2025.

Jiayi Chen, Wenxuan Song, Pengxiang Ding, Ziyang Zhou, Han Zhao, Barrett Tang, Donglin Wang, and Haoang Li. Unified diffusion vla: Vision-language-action model via joint discrete denosing diffusion process. In *International Conference on Learning Representations*, pp. 139291–139311, 2026.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, pp. 1684–1704, 2023.

Dwip Dalal, Shivansh Patel, Chahit Jain, Jeonghwan Kim, Utkarsh Mishra, Alex Baratian, Hyeonjeong Ha, Heng Ji, Svetlana Lazebnik, and Unnat Jain. Generalizable vla finetuning via representation anchoring and language-action alignment. *arXiv preprint arXiv:2607.13429*, 2026.

Danny Driess, Jost Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. In *Advances in Neural Information Processing Systems*, volume 38, pp. 102867–102888, 2026.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: A progressive robustness benchmark for visual-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38574–38583, 2026.

Yiyang Fu, Chubin Zhang, Shukai Gong, Yufan Deng, Kaiwei Sun, Qiyang Min, Qibin Hou, Yansong Tang, Jianan Wang, and Daquan Zhou. Stablevla: Towards robust vision-language-action models without extra data. In *International Conference on Machine Learning*, 2026.

Chongkai Gao, Zixuan Liu, Zhenghao Chi, Junshan Huang, Xin Fei, Yiwen Hou, Yuxuan Zhang, Yudi Lin, Zhirui Fang, and Lin Shao. Vla-os: Structuring and dissecting planning representations and paradigms in vision-language-action models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 136705–136736, 2026.

Ankit Goyal, Hugo Hadfield, Xuning Yang, Valts Blukis, and Fabio Ramos. Vla-0: Building state-of-the-art vlas with zero modification. *arXiv preprint arXiv:2510.13054*, 2025.

Asher Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting. In *International Conference on Learning Representations*, pp. 75333–75360, 2026.

Taisei Hanyu, Nhat Chung, et al. Slotvla: Towards modeling of object-relation representations in robotic manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Yufeng Ji, Wenhao Tang, Haoyi Niu, Koushil Sreenath, Yi Wu, and Zhongyu Li. Action qformer: Structured representation shaping under action supervision in vision-language-action models. *arXiv preprint arXiv:2607.14635*, 2026.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. In *Proceedings of Robotics: Science and Systems*, 2025a.

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. In *Proceedings of The 8th Conference on Robot Learning*, pp. 2679–2713, 2025b.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. In *IEEE International Conference on Robotics and Automation*, 2026.

Tao Lin, Yuxin Du, Yiran Mao, Zewei Ye, Yilei Zhong, Bing Cheng, Yiming Wang, Jiting Liu, Yang Tian, Junchi Yan, et al. La4vla: Learning to act without seeing via language-action pretraining. *arXiv preprint arXiv:2606.27295*, 2026a.

Tao Lin, Yilei Zhong, Yuxin Du, Jingjing Zhang, Jiting Liu, Yinxinyu Chen, Encheng Gu, Ziyan Liu, Hongyi Cai, Yanwen Zou, et al. Evo-1: Lightweight vision-language-action model with preserved semantic alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13397–13406, 2026b.

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

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

Zheng Lu, Haoran Liao, Wanqi Zhong, Yunhe Ni, Lijie Wang, Xingjie Fan, Zhisheng Chen, Yantang Qu, Meijia Chen, Tianyu Xin, et al. Layerroute: Action-conditioned mixture-of-layers routing for vision-language-action policies. *arXiv preprint arXiv:2609.06079*, 2026.

Jingzhou Luo, Yifan Wen, Yongjie Bai, Xinshuai Song, Yang Liu, and Liang Lin. Rovla: Multi-consistency constraints for robust vision-language-action models. *arXiv preprint arXiv:2605.19678*, 2026a.

Yuankai Luo, Woping Chen, Tong Liang, Baiqiao Wang, and Zhenguo Li. Simvla: A simple vla baseline for robotic manipulation. *arXiv preprint arXiv:2602.18224*, 2026b.

Yulin Luo, Hao Chen, Zhuangzhe Wu, Bowen Sui, Jiaming Liu, Chenyang Gu, Zhuoyang Liu, Qiuxuan Feng, Jiale Yu, Shuo Gu, et al. Look before acting: Enhancing vision foundation representations for vision-language-action models. *arXiv preprint arXiv:2603.15618*, 2026c.

Chaojun Ni, Cheng Chen, Xiaofeng Wang, Zheng Zhu, Wenzhao Zheng, Boyuan Wang, Tianrun Chen, Guosheng Zhao, Haoyun Li, Zhehao Dong, et al. Swiftvla: Unlocking spatiotemporal dynamics for lightweight vla models at minimal overhead. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13474–13485, 2026.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Jiayuan Gu, Zhigang Wang, Yan Ding, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models. In *Proceedings of Robotics: Science and Systems*, 2025.

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

Runze Wang, Yuqian Fu, Yu Li, Tao Lin, Tianwen Qian, Mohamed Elhoseiny, Bo Zhao, Yanwei Fu, Yu-Gang Jiang, and Xiangyang Xue. Afford-vla: Action-aligned visual planning via internalized affordance. *arXiv preprint arXiv:2605.24203*, 2026a.

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*, pp. 18638–18646, 2026b.

Lei Xiao, Jifeng Li, Juntao Gao, Feiyang Ye, Yan Jin, Jingjing Qian, Jing Zhang, Yong Wu, and Xiaoyuan Yu. Ava-vla: improving vision-language-action models with active visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13453–13463, 2026a.

Siyao Xiao, Yuhong Zhang, Zhifang Liu, Zihan Gao, Jingye Zhang, Sinwai Choo, Dake Zhong, Mengzhe Wang, Xiao Lin, Xianfeng Zhou, et al. $m2$-vla: Boosting vision-language models for generalizable manipulation via layer mixture and meta-skills. *arXiv preprint arXiv:2604.24182*, 2026b.

Borong Zhang, Jiahao Li, Jiachen Shen, Yuhao Zhang, Yishuai Cai, Yuanpei Chen, Juntao Dai, Jiaming Ji, and Yaodong Yang. Vla-arena: An open-source framework for benchmarking vision-language-action models. In *International Conference on Machine Learning*, 2026a.

Yichi Zhang, Weihao Yuan, Yizhuo Zhang, Xidong Zhang, and Jia Wan. Focusvla: Focused visual utilization for vision-language-action models. *arXiv preprint arXiv:2603.28740*, 2026b.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713, 2025.

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 *Proceedings of The 7th Conference on Robot Learning*, pp. 2165–2183, 2023.