ActAlign: Action-Aware Representation Alignment for Generalizable Vision-Language-Action Models

Anonymous submission

Abstract

Inspired by the success of vision-language models (VLMs), vision-language-action models (VLAs) have been proposed to extend pretrained VLMs to embodied decision-making with action prediction heads by training on large data. However, generalization remains a key challenge for VLAs, particularly under out-of-domain distribution shifts. Existing methods mainly exploit the inherited visual-language understanding ability of VLMs for object- or semantic-level generalization, but often struggle with action-related generalization, such as variations in robot initial states and camera views. We argue that this limitation arises from the representation gap between VLM pretraining and VLA requirements: VLMs learn object- and semantic-centric representations, while VLA tasks require action-aware representations that capture discriminative action-related information beyond preserving object semantics. To bridge this gap, we propose ActAlign, an annotation-free method for action-aware representation learning. ActAlign leverages action embeddings as dense supervision to explicitly align VLM representations with action-aware structures through a decoder-based reconstruction objective. By sharing the decoder between the alignment module and the diffusion action head, ActAlign further couples representation alignment and action generation within a unified action-aware latent space for better generation. Extensive experiments on simulation and real-robot benchmarks show that ActAlign consistently improves VLA performance and generalization.

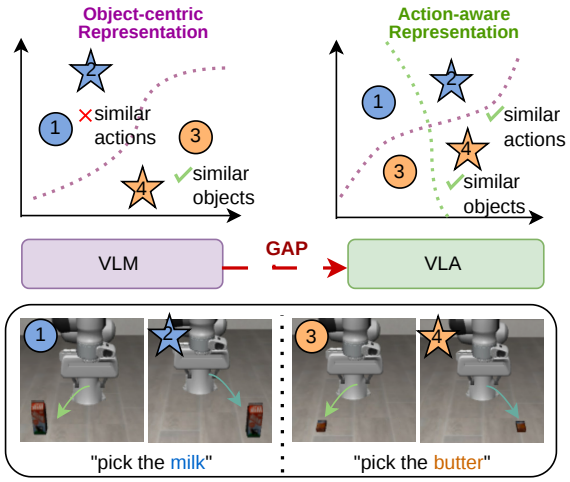

Figure 1: Illustration of the representation gap between VLMs and VLAs. VLMs mainly learn object-centric representations separated by object semantics (purple boundary), while VLA tasks require action-aware representations that distinguish both object categories (purple boundary) and action patterns (green boundary).

Introduction

Nowadays, vision-language models (VLMs) (Yang et al. 2025a) have achieved significant progress in aligning visual perception with natural language understanding, allowing models to recognize, describe, and reason about complex visual environments. Inspired by this success, vision-language-action models (VLAs) (Black et al. 2024; Bjorck et al. 2025; Kim et al. 2024) have been proposed to adapt such multimodal intelligence to embodied systems, which aim to predict executable actions conditioned on visual observations and language instructions, making them a key foundation for generalist robotic systems.

Despite their effectiveness in in-distribution settings, current VLA methods still struggle with generalization (Hung et al. 2025), especially under out-of-domain distribution (OOD) shifts. Existing studies (Cen et al. 2025; Hancock et al. 2025; Lian et al. 2026) mainly address this issue by preserving the inherited capabilities of pretrained VLMs for object- or scene-level generalization. However, they largely overlook action-related generalization, such as variations in robot initial states and camera views, which require understanding the robot state and manipulation dynamics beyond object-level semantic understanding. Consequently, current VLAs exhibit weaker performance under action-related distribution shifts (Fei et al. 2025), making robust action-related generalization a central challenge for current models.

Our key observation is that this limited action-related generalization of current VLAs stems from a representation gap between object semantic-centric VLM pretraining and action-aware VLA task requirements. VLMs are primarily optimized to encode object- and semantic-level information in representations, such as recognizing visual concepts, describing scenes, and aligning images with language, but these representations do not explicitly learn action-aware structures required for VLA tasks. As a result, VLAs that rely on

inherited VLM representations may overemphasize semantic similarity but fail to organize the representation space according to action similarity, even after fine-tuning on VLA data. For example, as illustrated in Fig. 1, tasks such as "move left to pick the milk" and "move right to pick the milk" involve the same object and may therefore produce similar VLM representations, even though they require substantially different action patterns for successful execution. Therefore, beyond maintaining semantic understanding, it is crucial to explicitly investigate and enhance action-aware representations for improving the robustness and generalization ability of VLAs in embodied control tasks.

To bridge this representation gap, we propose ActAlign, which learns action-aware VLM representations without relying on extra data resources or annotations. Our key idea is to use action embeddings as dense supervision signals for action-aware representation alignment. Unlike language-based supervision, which requires costly annotations and is limited by the granularity of textual descriptions, action embeddings can provide more fine-grained supervision for representation learning by capturing continuous action patterns and action-relevant variations that are difficult to describe with discrete language tokens. Since the action domain lacks a general pretrained encoder for extracting such embeddings, we learn them through a decoder-based reconstruction objective, where the learnable representations are optimized to reconstruct the corresponding ground-truth action trajectories, thereby encouraging them to capture action-aware structures. Furthermore, we share the action decoder between the action-aware representation alignment module and the diffusion action prediction head. This shared decoder encourages representation alignment and action generation to operate in the same action-aware latent space for better generation. Extensive experiments across multiple benchmarks demonstrate the effectiveness and generalization ability of our method. Our main contributions are summarized as follows:

- From a representation learning perspective, we find that the limited action-related generalization of current VLAs stems from their reliance on object-centric VLM representations without explicitly modeling action-aware structures.
- We propose ActAlign, an annotation-free method that aligns VLM representations with action-aware embeddings to improve generalization, and couples representation alignment and action generation within a unified action-aware latent space by sharing the action decoder for better action generation.
- Our proposed method achieves strong performance across multiple simulation and real-robot benchmarks, with particularly notable improvements in generalization and robustness under diverse out-of-distribution settings.

Related Work

Vision-Language-Action (VLA) Models. With the rapid progress of vision-language models (Yang et al. 2025a), recent works have increasingly adapted VLMs into VLAs for robotic control. These methods (Team et al. 2024; O'Neill et al. 2024; Wen et al. 2025; Liu et al. 2025; Black et al. 2024; Intelligence et al. 2025; Bjorck et al. 2025; Kim et al. 2024; Kim, Finn, and Liang 2025; Brohan et al. 2022; Zitkovich et al. 2023; Wang et al. 2026; Chi et al. 2025; Liang et al. 2025; Zhong et al. 2026) typically build upon pretrained VLMs for visual-language understanding and reasoning, attach with an additional action head to generate executable actions in an end-to-end manner.

To further enhance VLM representations for VLA tasks, recent works have explored various adaptation strategies, including improving reasoning and planning abilities (Cen et al. 2025; Zhao et al. 2025; Huang et al. 2026a; Bai et al. 2026; Huang et al. 2026b), and strengthening visual understanding (Zheng et al. 2024; Qu et al. 2025; Song et al. 2026; Fan et al. 2025; Liu et al. 2026; Li et al. 2026). However, these approaches typically rely on task-specific auxiliary supervision and extra training data, which may introduce additional biases and limit their general applicability (Zhang et al. 2026a; Lian et al. 2026). In contrast, our method investigates the gap between VLMs and VLAs from the perspective of action-aware representation and provides a solution without extra data or annotations.

Generalization in VLA. Generalization, particularly under out-of-domain distribution shifts, is one of the most important capabilities of foundation models, enabling them to adapt to unseen tasks and scenarios without further fine-tuning. Similarly, VLAs aim to learn generalist embodied policies that can generalize across diverse environments and settings. Existing works in VLAs (Zhang et al. 2024; Kim et al. 2024; Cen et al. 2025; Hancock et al. 2025; Lian et al. 2026) mainly focus on object-level or semantic generalization by preserving and leveraging the visual-language understanding ability of pretrained VLMs, such as generalization to novel objects, background changes, and instruction variations. While effective, these methods still struggle with action-related generalization, such as changes in initial robot states and camera views, which require capturing robot states and manipulation dynamics beyond object semantics (Fei et al. 2025). Our method fills this gap by introducing an action-aware representation learning framework that explicitly aligns VLM representations with action-aware structures, thereby improving robustness and generalization under various distribution shifts.

Motivation

Current methods (Kim et al. 2024; Cen et al. 2025; Hancock et al. 2025) mainly rely on the representations inherited from pretrained VLMs to support visual-language understanding and action prediction. Although effective for object- and semantic-level generalization, such as language, background, and object changes, these methods still struggle with action-related distribution shifts. As shown in Fig. 2, robot initial state and camera view generalization consistently achieve lower scores than language- and background-level semantic generalization, indicating that these action-relevant settings remain more challenging for current VLA models. In this section, we further analyze why existing methods struggle under such settings and highlight the importance of learning action-aware representations.

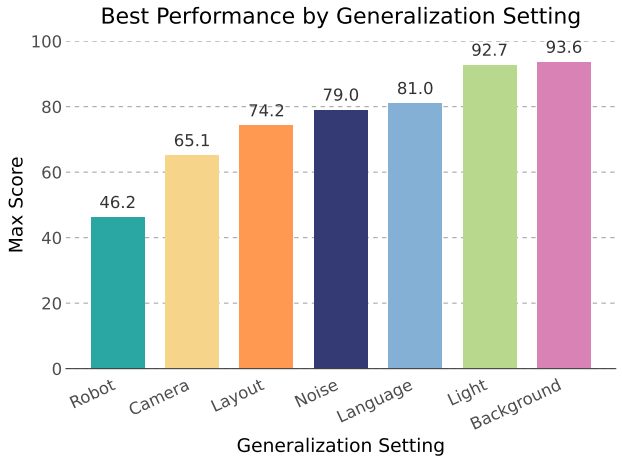

Figure 2: Best achieved performance among various VLAs across LIBERO-Plus generalization variants. Robot initial-state and camera-view shifts are the most challenging settings for current VLAs (lower scores indicate harder settings).

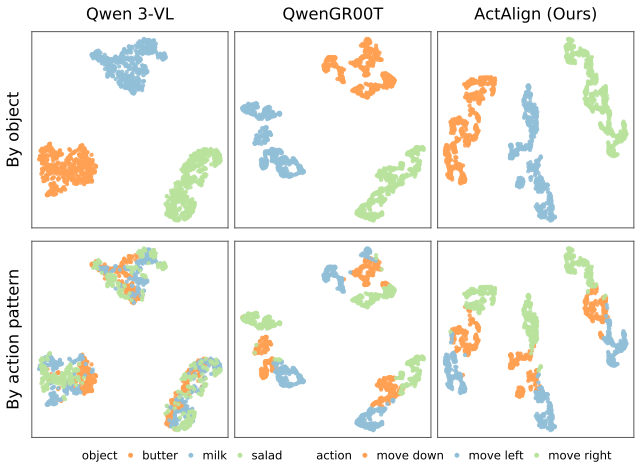

Figure 3: The t-SNE visualization of Qwen3-VL, QwenGR00T and our ActAlign representations. Qwen3-VL features cluster mainly by object categories rather than action patterns. QwenGR00T with VLA data tuning forms small action-aware sub-clusters within each object group, but still fails to align similar action patterns across different objects. Our ActAlign learns action-aware representations, better grouping samples with similar action patterns beyond object semantics.

Gap Between VLM and VLA. VLMs mainly focus on object-centric understanding, as they are optimized for recognizing visual concepts and describing scenes. In contrast, VLA models should understand not only what object is involved, but also how the robot should act. This means that effective VLA representations should capture action-relevant factors, such as robot states and manipulation dynamics, rather than relying solely on object semantics. For instance, as shown in Figure 1, "move left to pick the milk" and "move left to pick the butter" require similar action patterns even when applied to different objects, while the same object may require different actions under different settings. Therefore, directly relying on VLM representations may lead the model to focus excessively on object semantics while overlooking the underlying action structure.

Experiments Analysis To further investigate this issue, we conduct an empirical analysis by clustering learned representations according to object categories and action patterns. Specifically, we perform a t-SNE representation probing study on the LIBERO-Plus benchmark (Fei et al. 2025), comparing Qwen 3-VL (Yang et al. 2025a) with QwenGR00T (Community 2026), which implements GR00T using Qwen3-VL as the backbone. We select samples involving three object categories (butter, milk, and salad dressing), placed at different spatial locations to induce different action patterns (moving down, left, and right). As shown in Fig. 3, the representations of Qwen3-VL tend to cluster primarily according to object or semantic attributes rather than action similarity, suggesting that the inherited VLM features remain biased toward object-centric information. In contrast, QwenGR00T, which fine-tunes the VLM with an action head on robot data, exhibits small action-aware sub-clusters within each object cluster, demonstrating the importance of action-aware representation learning in VLA tasks. However, similar action patterns across different objects are still not well aligned in the representation space. This suggests that existing VLA adaptation remains largely constrained by object-centric representations and does not sufficiently capture action-aware structure across semantic variations, which may explain their weaker performance under initial-pose and camera-view generalization.

Method

As shown in Fig. 4, the overall framework of our proposed ActAlign consists of two main components: a diffusion-based action prediction module that generates executable continuous action trajectories, and an action-aware representation alignment module that enforces the VLM representations toward an action-aware latent space. We introduce the details of each component in the following sections.

Action Prediction

To adapt VLMs to VLAs for action prediction, we follow previous works (Intelligence et al. 2025; Bjorck et al. 2025) by introducing a diffusion-based action head on top of the VLM backbone for more efficient and continuous action generation. Given the image observation $\boldsymbol{o}_t$ and a language instruction l at t time step, the VLM f_θ first encodes the visual and textual inputs through multimodal attention layers to obtain a task-conditioned representation $\boldsymbol{h}_t = f_\theta(\boldsymbol{o}_t, l)$. The diffusion action head f_ϕ then takes $\boldsymbol{h}_t$ as the condition and learns to generate a continuous action trajectory over a horizon of H steps $\{\boldsymbol{a}_t, ..., \boldsymbol{a}_{t+H}\}$. Specifically, we adapt the flow-matching (Lipman et al. 2022) process during training, where Gaussian noise ϵ is added to the ground-truth action sequence at a randomly sampled diffusion step k, obtaining

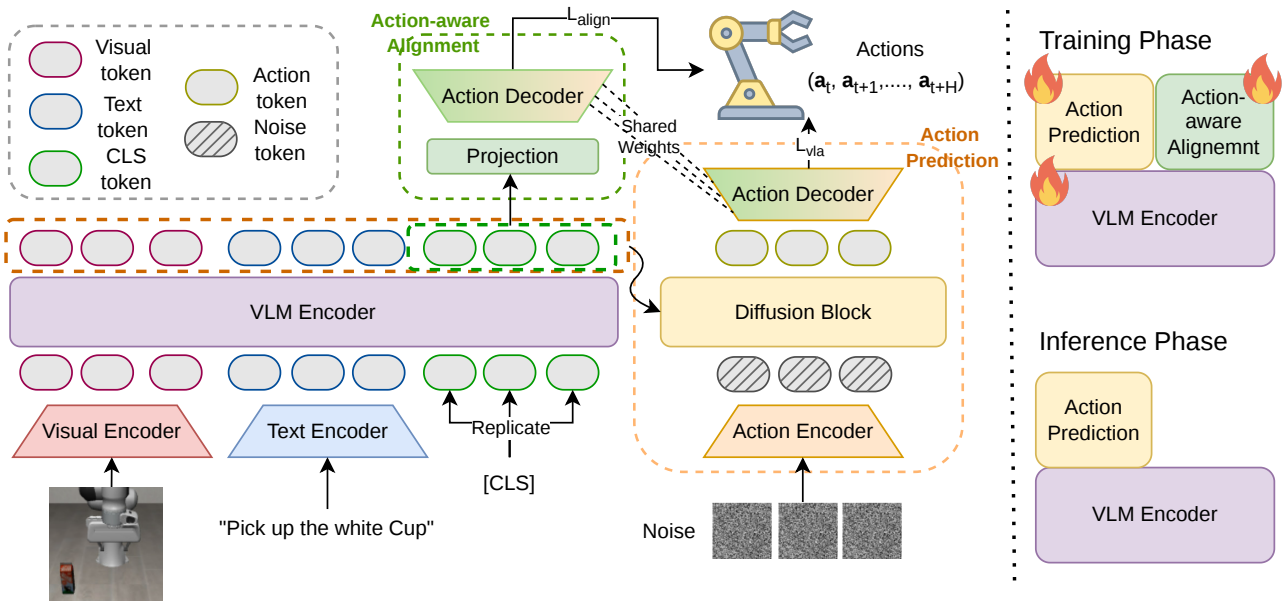

Figure 4: Overall framework of ActAlign. The model consists of a VLM encoder, an action-aware alignment module, and an action prediction module. During training, after obtaining representations via the VLM encoder, the alignment module uses action embeddings as dense supervision to learn action-aware representations via reconstruction loss $\mathcal{L}_{align}$, and the action prediction module generates continuous action trajectories via diffusion loss $\mathcal{L}_{vla}$ with a shared decoder to operate in the same action-aware latent space. During inference, only the VLM encoder and action prediction module are used for action generation.

the noisy action $\boldsymbol{a}_{t:t+H}^{k,\epsilon} = k * \boldsymbol{a}_{t:t+H} + (1-k)\epsilon$. The action head is trained to predict the flow vector field $\epsilon - \boldsymbol{a}_t$:

$$\mathcal{L}_{vla} = \mathbb{E}_{\mathcal{D},k,\epsilon}[||\epsilon - \boldsymbol{a}_{t:t+H} - f_\phi(a_{t:t+H}^{k,\epsilon}, \boldsymbol{h}_t)||^2], \quad (1)$$

where $\mathcal{D}$ denotes the dataset and $\hat{y}_{t:t+H} = f_\phi(a_{t:t+H}^{k,\epsilon}, \boldsymbol{h}_t)$ denotes the output from the diffusion action head conditioned on the VLM representation $\boldsymbol{h}_t$.

Action-aware Representation Alignment

As analyzed in Motivation, learning action-aware representations is essential for improving VLA generalization. One straightforward approach is to construct action-understanding tasks with annotated language supervision, such as visual question answering (VQA) (Chen et al. 2025). However, this requires additional data and annotations, making it costly and difficult to scale. In addition, language-based action descriptions are often sparse and limited, as finite words cannot fully describe continuous and fine-grained robot manipulations. To address this, we propose to use action embeddings as dense action descriptions for self-supervised learning. Since action trajectories are naturally available in VLA training data, they can provide rich supervision for aligning visual-language representations with action-aware structures, without introducing extra data or annotations.

Although action embeddings can provide dense self-supervision for action-aware representation learning, obtaining such embeddings is non-trivial. Unlike the image domain (Oquab et al. 2023), the action domain does not have a widely adopted large-scale pretrained encoder for extracting general action representations. To solve this, we learn action embeddings with a decoder-based objective. Specifically, we introduce an action decoder to reconstruct the ground-truth actions from the learnable representations. This forces the representations to encode sufficient information about the underlying action structures. Additionally, inspired by previous works (Yao, Yang, and Wang 2025) showing that aligning diffusion features with discriminative features benefits generation, we share the action decoder between the alignment module and the diffusion action head. With this design, the alignment module and the action head are coupled through the same action-aware decoding space, facilitating the transfer of action-discriminative information to the diffusion action head for improved action generation.

More specifically, we repeat the CLS tokens to match the action chunk length. After feeding them into the VLM, we obtain the hidden states $\boldsymbol{h}_t$ and extract the CLS tokens' representations $\boldsymbol{h}_t^{cls}$, which aggregate information from both visual and textual tokens to capture the task-conditioned context. We then project $\boldsymbol{h}_t^{cls}$ into the action-aware latent space using a linear layer parameterized by $\boldsymbol{W}$ and $\boldsymbol{b}$, and feed the projected representation into the shared action decoder from the diffusion action head to reconstruct the corresponding ground-truth action trajectory. In this way, the VLM representation is explicitly aligned with an action-aware latent space for action-aware representation learning, and the alignment loss is formulated as:

$$\mathcal{L}_{align} = \mathbb{E}_{\mathcal{D}}[||f_{\phi'}(\boldsymbol{W}\boldsymbol{h}_t^{cls} + \boldsymbol{b}) - \boldsymbol{a}_{t:t+H}||^2], \quad (2)$$

where $\mathcal{D}$ denotes the dataset and $f_{\phi'}$ denotes the shared action decoder.

Methods	Camera	Robot	Language	Light	Background	Noise	Layout	Average
Trained on LIBERO-Plus (In-domain Adaptation)								
OpenVLA-OFT	92.8	30.3	85.8	94.9	93.9	89.3	77.6	79.6
Trained on LIBERO (Generalization)								
OpenVLA	0.8	3.5	23.0	8.1	34.8	15.2	28.5	15.6
OpenVLA-OFT	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
OpenVLA-OFT_w	10.4	38.7	70.5	76.8	93.6	49.9	69.9	55.8
NORA	2.2	37.0	65.1	45.7	58.6	12.8	62.1	39.0
WorldVLA	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
UniVLA	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
π_0	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
π_0-Fast	**65.1**	21.6	61.0	73.2	73.2	74.4	68.8	61.6
RIPT-VLA	55.2	31.2	77.6	88.4	91.6	73.5	74.2	68.4
OpenVLA-OFT_m	55.6	21.7	81.0	92.7	91.0	78.6	68.7	67.9
QwenGR00T	53.1	55.5	90.6	93.1	**94.6**	68.2	75.5	74.1
ActAlign (Ours)	59.0	**60.2**	**91.2**	**96.2**	93.3	**79.1**	**79.7**	**78.4**

Table 1: Simulation results on LIBERO-Plus (out-of-domain performance). The **best**, second-best, and third-best results are highlighted. Our method achieves the best performance on most of the tasks.

Overall Objective. ActAlign is designed to improve continuous action prediction by enhancing action-aware representation learning for better generalization, and the overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{vla} + \lambda \mathcal{L}_{align}, \tag{3}$$

where $\mathcal{L}_{vla}$ denotes the diffusion-based action prediction loss in Equation (1), $\mathcal{L}_{align}$ denotes the action-aware representation alignment loss in Equation (2), and λ is a weighting coefficient that balances the two objectives.

Training and Inference

Training paradigm. We train the overall model end-to-end with the loss defined in Equation (3). Unlike previous works (Bjorck et al. 2025; Kim et al. 2024) that rely on large-scale pretraining on robot and general vision-language data before benchmark-specific fine-tuning, we follow works (Community 2026) to directly fine-tune the pre-trained VLM on smaller robot-only datasets for each benchmark. This paradigm is more data-efficient and practical to enable effective adaptation of VLM for action prediction.

Inference. During inference, our method follows the standard VLA inference pipeline, where the observation and instruction are first encoded by the VLM into task-conditioned representations. The diffusion action head then predicts the action sequence by progressively denoising conditioned on the VLM representations. The action-aware representation alignment module is only used during training and is discarded at test time. As a result, our method introduces no additional inference cost or architectural overhead.

Experiment

Experimental Setup

Benchmarks. We evaluate our method on both simulation and real-robot benchmarks. For simulation experiments, we use LIBERO (Liu et al. 2023), LIBERO-Plus (Fei et al. 2025), and SimplerEnv (Li et al. 2024b) to assess in-domain adaptation, OOD generalization, and cross-embodiment performance, and report the success rate as a metric. For real-robot experiments, we conduct pick-and-place tasks on a 7-DoF UR robot arm to evaluate in-domain performance and object, robot initial state, and camera view generalization, and report the number of successful trials per task.

Baselines and Implementation. We compare ActAlign with various baselines that leverage inherited VLM generalization ability: 1) large pretrained models OpenVLA (Kim et al. 2024), OpenVLA-OFT (Kim, Finn, and Liang 2025), Octo (Team et al. 2024), GR00T (Bjorck et al. 2025), π_0 (Black et al. 2024), π_0-Fast (Pertsch et al. 2025), RT-1 (Brohan et al. 2022), RT-1-X (O'Neill et al. 2024), RT-2-X (Zitkovich et al. 2023), CogACT (Li et al. 2024a) and UniVLA (Bu et al. 2025); 2) efficient adaptation models NORA (Hung et al. 2025), Diffusion Policy (Chi et al. 2025), SmolVLA (Shukor et al. 2025), PD-VLA (Song et al. 2025), VLA-Adapter (Wang et al. 2026) and QwenGR00T (Community 2026); 3) reasoning/planning-enhanced models CoT-VLA (Zhao et al. 2025), ThinkAct (Huang et al. 2026b), VLA-OS (Gao et al. 2026), MolmoAct (Lee et al. 2025) and FlowVLA (Zhong et al. 2025); 4) visual-understanding-enhanced models TraceVLA (Zheng et al. 2024), SpatialVLA (Qu et al. 2025), 4D-VLA (Zhang et al. 2026b) and GraspVLA (Deng et al. 2025); 5) world-knowledge-enhanced models WorldVLA (Cen et al. 2025), Unified-VLA (Wang et al. 2025) and Magma (Yang et al. 2025b); and 6) reinforcement learning-based models RIPT-VLA (Tan et al. 2025), VLA-RL (Lu et al. 2025) and GRAPE (Zhang et al. 2024). We use Qwen3-VL-4B-Instruct (Yang et al. 2025a) as the VLM backbone with a diffusion-based action head and a two-layer MLP action decoder. We set $\lambda = 0.1$ and train all learnable components from scratch under the end-to-end objective. More details are in the Appendix.

Methods	Spatial	Object	Goal	Long	Avg
Diffusion Policy	78.3	92.5	68.3	50.5	72.4
TraceVLA	84.6	85.2	75.1	54.1	74.8
Octo	78.9	85.7	84.6	51.1	75.1
OpenVLA	84.7	88.4	79.2	53.7	76.5
SpatialVLA	88.2	89.9	78.6	55.5	78.1
GRAPE	87.6	91.2	82.2	55.8	79.2
VLA-RL	90.2	91.8	82.2	59.8	81.0
CoT-VLA	87.5	91.6	87.6	69.0	81.1
WorldVLA	87.6	96.2	83.4	60.0	81.8
ThinkAct	88.3	91.4	87.1	70.9	84.4
π_0-FAST	96.4	96.8	88.6	60.2	85.5
VLA-OS	87.0	96.5	92.7	66.0	85.6
MolmoAct	87.0	95.4	87.6	77.2	86.6
NORA	92.2	95.4	89.4	74.6	87.9
FlowVLA	93.2	95.0	91.6	72.6	88.1
4D-VLA	88.9	95.2	90.9	79.1	88.6
SmolVLA	93.0	94.0	91.0	77.0	88.8
GraspVLA	–	94.1	91.2	82.0	89.1
GR00T N1	94.4	97.6	93.0	90.6	93.9
π_0	96.8	98.8	95.8	85.2	94.2
PD-VLA	95.5	96.7	94.9	91.7	94.7
UniVLA	96.5	96.8	95.6	92.0	95.2
UnifiedVLA	95.4	98.8	93.6	94.0	95.5
OpenVLA-OFT	97.6	98.4	97.9	94.5	97.1
VLA-Adapter	**97.8**	99.2	97.2	95.0	**97.3**
QwenGR00T	95.0	99.0	97.6	93.8	96.4
ActAlign (Ours)	94.4	**99.4**	**98.0**	**95.2**	96.8

Table 2: Simulation results on LIBERO (in-domain performance). The **best**, second-best, and third-best results are highlighted. Our method achieves the best performance on Object, Goal, and Long tasks.

Main Results

Results on LIBERO and LIBERO-Plus. We train ActAlign from scratch on the four LIBERO task suites jointly, and evaluate it on LIBERO for in-domain adaptation and LIBERO-Plus for out-of-domain generalization. As shown in Table 1, ActAlign achieves the best average performance on out-of-domain tasks compared with existing baselines, demonstrating the effectiveness of the proposed action-aware representation learning. Notably, our method obtains the highest performance under robot initial-state generalization, with a 14% improvement over previous methods, indicating that explicitly enhancing action-aware representations can better handle action-related distribution shifts. This result supports our motivation that robust VLA generalization requires going beyond semantic understanding and capturing the underlying action structure. For in-domain tasks, as shown in Table 2, ActAlign achieves comparable average performance to methods that rely on large-scale pretraining data, while obtaining the best results on the Object, Goal, and Long task suites. These results further demonstrate that ActAlign can effectively improve action prediction and task adaptation without requiring costly large-scale robot pretraining, highlighting its data efficiency and general applicability.

Methods	PC	MV	OD	OTD	Avg
RT-1	89.8	50.0	32.3	2.6	43.7
RT-1-X	49.0	32.3	29.4	10.1	30.2
RT-2-X	82.3	79.2	35.3	20.6	54.4
OpenVLA	60.8	67.7	28.8	0.0	39.3
CogACT	89.6	80.8	28.3	**46.6**	61.3
SpatialVLA	88.0	**82.5**	41.8	-	**70.7**
π_0	75.2	63.7	25.6	-	54.8
π_0-FAST	77.6	68.2	31.3	-	59.0
GR00T N1.5	69.3	68.7	35.8	4.0	44.5
Magma	68.8	65.7	**53.4**	18.5	51.6
QwenGR00T	92.5	59.2	50.6	33.9	59.1
ActAlign (Ours)	**93.3**	68.4	39.5	38.8	60.0

Table 3: Simulation results on the SimplerEnv Google Robot benchmark, including Pick Coke Can (PC), Move Near (MN), Open Drawer (OD), and Open Top Drawer (OTD). The **best**, second-best, and third-best results are highlighted. Our method achieves competitive performance compared with baselines requiring extra pre-training data.

Methods	Pre-Train	In-Domain	OOD Generalization		
			Object	Robot	Camera
$\pi_{0.5}$	✓	30/30	7/10	**6/10**	8/10
QwenGR00T	✗	26/30	7/10	3/10	5/10
ActAlign (Ours)	✗	29/30	**8/10**	**6/10**	**9/10**

Table 4: Real robot results. Each method is fairly assessed over 10 trials on each task.

Results on SimplerEnv. We train ActAlign from scratch on the Bridge-v2 and Fractal datasets and evaluate it on SimplerEnv. The variant-aggregated results are reported in Table 3. Compared with pretrained baselines, ActAlign achieves comparable overall performance and obtains the best result on the Pick Coke Can task, demonstrating its effectiveness in transferring learned action-aware representations across different embodiments and environments. To provide a more detailed comparison across different variants, we further average the results of the Pick Coke Can and Move Near tasks under each variant, as shown in Fig. 5. Compared with the baseline QwenGR00T, our method achieves better performance in both the base and variant settings, demonstrating improved robustness to environmental changes.

Results on Real Robot. We use a 7-DoF UR robot arm and design pick-and-place tasks with three training objects: a red cube, a cup, and a tape roll. For each object, we collect 20 demonstrations for training. During evaluation, we test both in-domain performance and OOD generalization, including novel-object generalization with a white cube, robot initial state generalization, and camera view generalization. As shown in Table 4, for in-domain evaluation, ActAlign achieves performance comparable to the large-scale pretrained $\pi_{0.5}$ model (Intelligence et al. 2025), while outperforming QwenGR00T, which uses the same backbone architecture. For OOD generalization, ActAlign achieves the

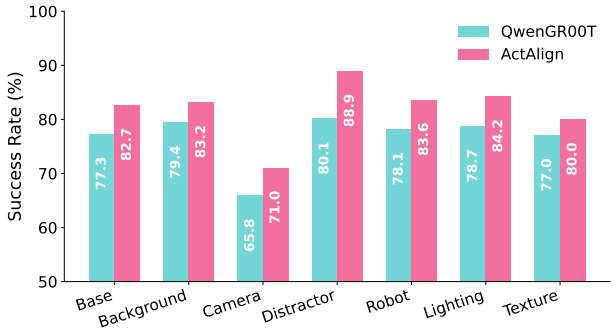

Figure 5: Variant-aggregated results on the SimplerEnv Google Robot benchmark. Our method improves performance across variants over the same-architecture baseline.

Figure 6: Real robot results on object generalization. ActAlign distinguishes the new white cube from the training red cube, and successfully completes the task.

best overall performance, even surpassing $\pi_{0.5}$. In particular, compared with QwenGR00T, ActAlign brings substantial improvements (nearly 50% relative gains) under action-related shifts, such as robot initial state and camera view variations. We also provide qualitative examples in Fig. 6 (with more results in the Appendix), further demonstrating ActAlign's robustness and generalization in different settings.

Analysis

Ablaiton. As shown in Table 5, we ablate the proposed action-aware alignment module, including the alignment loss and decoder sharing with the diffusion action head. Using both components achieves the best performance, demonstrating the effectiveness of jointly aligning VLM representations with action-relevant information and coupling this alignment with action generation. Notably, the alignment loss brings the largest gains on OOD generalization, suggesting that action-aware representation learning is critical for handling distribution shifts. We further analyze the weighting coefficient λ between the alignment loss and the action prediction loss, as shown in Table 6. The best performance is achieved with $\lambda = 0.1$, suggesting that a moderate alignment weight provides useful action-aware alignment while preserving the action generation capability of the diffusion head.

$\mathcal{L}_{align}$	Share Decoder	In-Domain	Out-of-Domain
✗	✗	96.4	74.1
✓	✗	96.3	77.4
✓	✓	**96.8**	**78.4**

Table 5: Ablation study of design choices of ActAlign.

λ	0.1	0.3	0.5
OOD Avg	**78.37**	75.02	76.93

Table 6: Ablation study of loss weight λ on LIBERO-Plus.

Figure 7: Cross-attention visualization between the VLM and action prediction head. ActAlign focuses more on action-relevant regions (e.g., movable areas and robot end-effector), beyond the target object.

Visualization Analysis. We first compare t-SNE visualizations by clustering representations according to object categories and action patterns. As shown in Fig. 3, ActAlign learns more discriminative action-aware representations while maintaining object-level separability, indicating improved action-aware structure without degrading semantic understanding. We further visualize cross-attention maps between the VLM and action prediction head in Fig. 7. The baseline QwenGR00T, even after VLA tuning, still mainly attends to object-semantic regions (robot and the target object), while ActAlign also focuses on action-relevant regions (e.g., movable areas and robot end-effector). This aligns with the t-SNE results and further shows ActAlign learns action-aware representation learning for robust action prediction.

Conclusion

In this work, we identify the representation gap between VLM pretraining and VLA task requirements as a key factor limiting action-related generalization in current VLAs. While VLMs mainly learn object- and semantic-centric representations from visual-language tasks, VLAs require action-aware representations that capture robot states and manipulation dynamics beyond semantic understanding. To address this gap, we propose ActAlign, an annotation-free method that uses action embeddings to align VLM representations through a reconstruction objective and shares the decoder with the diffusion action head for improved action generation. Our work offers a representation-learning perspective for analyzing VLAs and takes a step toward more action-aware and generalizable embodied models.

References

Bai, S.; Lyu, J.; Zhou, W.; Li, Z.; Wang, D.; Xing, L.; Zhao, X.; Wang, P.; Wang, Z.; Chi, C.; et al. 2026. Latent Reasoning VLA: Latent Thinking and Prediction for Vision-Language-Action Models. *arXiv preprint arXiv:2602.01166*.

Bjorck, J.; Castañeda, F.; Cherniadev, N.; Da, X.; Ding, R.; Fan, L.; Fang, Y.; Fox, D.; Hu, F.; Huang, S.; et al. 2025. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*.

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; et al. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.

Bu, Q.; Yang, Y.; Cai, J.; Gao, S.; Ren, G.; Yao, M.; Luo, P.; and Li, H. 2025. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*.

Cen, J.; Yu, C.; Yuan, H.; Jiang, Y.; Huang, S.; Guo, J.; Li, X.; Song, Y.; Luo, H.; Wang, F.; et al. 2025. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*.

Chen, K.; Xie, S.; Ma, Z.; Sanketi, P. R.; and Goldberg, K. 2025. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. *arXiv preprint arXiv:2505.15517*.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2025. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11): 1684–1704.

Community, S. 2026. StarVLA: A Lego-like Codebase for Vision-Language-Action Model Developing. *arXiv preprint arXiv:2604.05014*.

Deng, S.; Yan, M.; Wei, S.; Ma, H.; Yang, Y.; Chen, J.; Zhang, Z.; Yang, T.; Zhang, X.; Zhang, W.; et al. 2025. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*.

Fan, C.; Jia, X.; Sun, Y.; Wang, Y.; Wei, J.; Gong, Z.; Zhao, X.; Tomizuka, M.; Yang, X.; Yan, J.; et al. 2025. Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions. *arXiv preprint arXiv:2505.02152*.

Fei, S.; Wang, S.; Shi, J.; Dai, Z.; Cai, J.; Qian, P.; Ji, L.; He, X.; Zhang, S.; Fei, Z.; et al. 2025. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*.

Gao, C.; Liu, Z.; Chi, Z.; Huang, J.; Fei, X.; Hou, Y.; Zhang, Y.; Lin, Y.; Fang, Z.; and Shao, L. 2026. Vla-os: Structuring and dissecting planning representations and paradigms in vision-language-action models. *Advances in Neural Information Processing Systems*, 38: 136705–136736.

Hancock, A. J.; Wu, X.; Zha, L.; Russakovsky, O.; and Majumdar, A. 2025. Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting. *arXiv preprint arXiv:2509.22195*.

Huang, C.-P.; Man, Y.; Yu, Z.; Chen, M.-H.; Kautz, J.; Wang, Y.-C. F.; and Yang, F.-E. 2026a. Fast-ThinkAct: Efficient Vision-Language-Action Reasoning via Verbalizable Latent Planning. *arXiv preprint arXiv:2601.09708*.

Huang, C.-P.; Wu, Y.-H.; Chen, M.-H.; Wang, F.; and Yang, F.-E. 2026b. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *Advances in Neural Information Processing Systems*, 38: 82782–82802.

Hung, C.-Y.; Sun, Q.; Hong, P.; Zadeh, A.; Li, C.; Tan, U.; Majumder, N.; Poria, S.; et al. 2025. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*.

Intelligence, P.; Black, K.; Brown, N.; Darpinian, J.; Dhabalia, K.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; et al. 2025. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*.

Kim, M. J.; Finn, C.; and Liang, P. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Lee, J.; Duan, J.; Fang, H.; Deng, Y.; Liu, S.; Li, B.; Fang, B.; Zhang, J.; Wang, Y. R.; Lee, S.; et al. 2025. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*.

Li, Q.; Liang, Y.; Wang, Z.; Luo, L.; Chen, X.; Liao, M.; Wei, F.; Deng, Y.; Xu, S.; Zhang, Y.; et al. 2024a. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*.

Li, W.; Zhang, R.; Shao, R.; Fang, Z.; Zhou, K.; Tian, Z.; and Nie, L. 2026. Semanticvla: Semantic-aligned sparsification and enhancement for efficient robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18397–18405.

Li, X.; Hsu, K.; Gu, J.; Pertsch, K.; Mees, O.; Walke, H. R.; Fu, C.; Lunawat, I.; Sieh, I.; Kirmani, S.; et al. 2024b. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*.

Lian, S.; Yu, B.; Lin, X.; Yang, L. T.; Shen, Z.; Wu, C.; Miao, Y.; Huang, C.; and Chen, K. 2026. LangForce: Bayesian Decomposition of Vision Language Action Models via Latent Action Queries. *arXiv e-prints*, arXiv–2601.

Liang, Z.; Li, Y.; Yang, T.; Wu, C.; Mao, S.; Pei, L.; Nian, T.; Zhou, S.; Yang, X.; Pang, J.; et al. 2025. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv:2508.20072*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; and Stone, P. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791.

Liu, C.; Zhang, J.; Li, C.; Zhou, Z.; Wu, S.; Huang, S.; and Duan, H. 2026. Ttf-vla: Temporal token fusion via pixel-attention integration for vision-language-action models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18452–18459.

Liu, J.; Chen, H.; An, P.; Liu, Z.; Zhang, R.; Gu, C.; Li, X.; Guo, Z.; Chen, S.; Liu, M.; et al. 2025. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*.

Lu, G.; Guo, W.; Zhang, C.; Zhou, Y.; Jiang, H.; Gao, Z.; Tang, Y.; and Wang, Z. 2025. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

O'Neill, A.; Rehman, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlekar, A.; Jain, A.; et al. 2024. 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)*, 6892–6903. IEEE.

Pertsch, K.; Stachowicz, K.; Ichter, B.; Driess, D.; Nair, S.; Vuong, Q.; Mees, O.; Finn, C.; and Levine, S. 2025. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*.

Qu, D.; Song, H.; Chen, Q.; Yao, Y.; Ye, X.; Ding, Y.; Wang, Z.; Gu, J.; Zhao, B.; Wang, D.; et al. 2025. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*.

Shukor, M.; Aubakirova, D.; Capuano, F.; Kooijmans, P.; Palma, S.; Zouitine, A.; Aractingi, M.; Pascal, C.; Russi, M.; Marafioti, A.; et al. 2025. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*.

Song, W.; Chen, J.; Ding, P.; Zhao, H.; Zhao, W.; Zhong, Z.; Ge, Z.; Li, Z.; Wang, D.; Ma, J.; et al. 2025. PD-VLA: Accelerating Vision-Language-Action Model Integrated with Action Chunking via Parallel Decoding. *arXiv preprint arXiv:2503.02310*.

Song, W.; Zhou, Z.; Zhao, H.; Chen, J.; Ding, P.; Yan, H.; Huang, Y.; Tang, F.; Wang, D.; and Li, H. 2026. Reconvla: Reconstructive vision-language-action model as effective robot perceiver. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18549–18557.

Tan, S.; Dou, K.; Zhao, Y.; and Krähenbühl, P. 2025. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*.

Team, O. M.; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Kreiman, T.; Xu, C.; et al. 2024. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*.

Wang, Y.; Ding, P.; Li, L.; Cui, C.; Ge, Z.; Tong, X.; Song, W.; Zhao, H.; Zhao, W.; Hou, P.; et al. 2026. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, 18638–18646.

Wang, Y.; Li, X.; Wang, W.; Zhang, J.; Li, Y.; Chen, Y.; Wang, X.; and Zhang, Z. 2025. Unified vision-language-action model. *arXiv preprint arXiv:2506.19850*.

Wen, J.; Zhu, Y.; Zhu, M.; Tang, Z.; Li, J.; Zhou, Z.; Liu, X.; Shen, C.; Peng, Y.; and Feng, F. 2025. Diffusionvla: Scaling robot foundation models via unified diffusion and autoregression. In *Forty-second International Conference on Machine Learning*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, J.; Tan, R.; Wu, Q.; Zheng, R.; Peng, B.; Liang, Y.; Gu, Y.; Cai, M.; Ye, S.; Jang, J.; et al. 2025b. Magma: A foundation model for multimodal ai agents. In *Proceedings of the computer vision and pattern recognition conference*, 14203–14214.

Yao, J.; Yang, B.; and Wang, X. 2025. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15703–15712.

Zhang, J.; Chen, X.; Wang, Q.; Li, M.; Guo, Y.; Hu, Y.; Zhang, J.; Bai, S.; Lin, J.; and Chen, J. 2026a. VLM4VLA: Revisiting Vision-Language-Models in Vision-Language-Action Models. *arXiv preprint arXiv:2601.03309*.

Zhang, J.; Chen, Y.; Xu, Y.; Huang, Z.; Zhou, Y.; Yuan, Y.-J.; Cai, X.; Huang, G.; Quan, X.; Xu, H.; et al. 2026b. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration. *Advances in Neural Information Processing Systems*, 38: 33914–33937.

Zhang, Z.; Zheng, K.; Chen, Z.; Jang, J.; Li, Y.; Han, S.; Wang, C.; Ding, M.; Fox, D.; and Yao, H. 2024. Grape: Generalizing robot policy via preference alignment. *arXiv preprint arXiv:2411.19309*.

Zhao, Q.; Lu, Y.; Kim, M. J.; Fu, Z.; Zhang, Z.; Wu, Y.; Li, Z.; Ma, Q.; Han, S.; Finn, C.; et al. 2025. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1702–1713.

Zheng, R.; Liang, Y.; Huang, S.; Gao, J.; Daumé III, H.; Kolobov, A.; Huang, F.; and Yang, J. 2024. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*.

Zhong, Y.; He, Y.; Yang, Z.; Tian, P.; Huang, Y.; Huang, Q.; Zhu, X.; and Ma, Y. 2026. From Noise to Intent: Anchoring Generative VLA Policies with Residual Bridges. *arXiv preprint arXiv:2604.21391*.

Zhong, Z.; Yan, H.; Li, J.; Liu, X.; Gong, X.; Zhang, T.; Song, W.; Chen, J.; Zheng, X.; Wang, H.; et al. 2025. Flowvla: Visual chain of thought-based motion reasoning for vision-language-action models. *arXiv preprint arXiv:2508.18269*.

Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2165–2183. PMLR.