Don't Blind Your VLA: Aligning Visual Representations for OOD Generalization

Nikita Kachaev [1 2] **Mikhail Kolosov** [3] **Daniil Zelezetsky** [3] **Alexey K. Kovalev** [1 3] **Aleksandr I. Panov** [1 3]

Abstract

The growing success of Vision-Language-Action (VLA) models stems from the promise that pretrained Vision-Language Models (VLMs) can endow agents with transferable world knowledge and vision-language (VL) grounding, laying a foundation for action models with broader generalization. Yet when these VLMs are adapted to the action modality, it remains unclear to what extent their original VL representations and knowledge are preserved. In this work, we conduct a systematic study of representation retention during VLA fine-tuning, showing that naive action fine-tuning leads to degradation of visual representations. To characterize and measure these effects, we probe VLA's hidden representations and analyze attention maps, further, we design a set of targeted tasks and methods that contrast VLA models with their counterpart VLMs, isolating changes in VL capabilities induced by action fine-tuning. We further evaluate a range of strategies for aligning visual representations and introduce a simple yet effective method that mitigates degradation and yields improved generalization to out-of-distribution (OOD) scenarios. Taken together, our analysis clarifies the trade-off between action fine-tuning and the degradation of VL representations and highlights practical approaches to recover inherited VL capabilities. Code is publicly available: **blind-vla-paper.github.io**

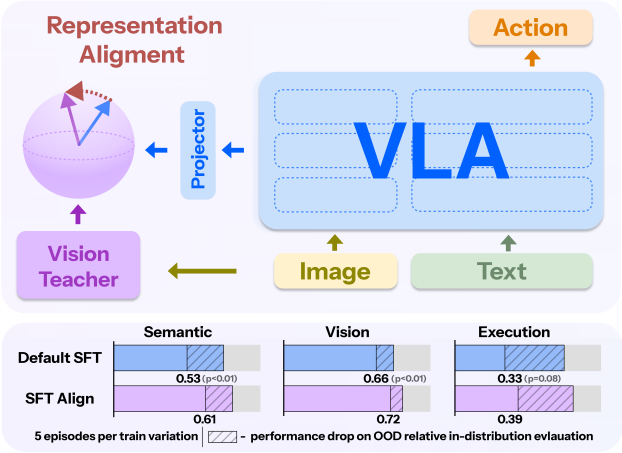

Figure 1. Visual alignment method overview. Mid-level VLA features are projected onto a normalized sphere and aligned with teacher embeddings, preserving visual semantics and improving OOD generalization. Bottom plots show comparison with standard SFT across three generalization axes on the Simpler-based benchmark (Liu et al., 2025).

1. Introduction

Vision–Language Models (VLMs) have demonstrated remarkable success due to their ability to integrate large-scale multimodal datasets, thereby acquiring semantic grounding and generalizable visual-language (VL) representations (Azzolini et al., 2025; Bai et al., 2025; Wang et al., 2025a; Beyer et al., 2024; Driess et al., 2023; Awadalla et al., 2023). When exposed to novel visual or linguistic contexts, such models exhibit robust cross-modal understanding and compositional perception – properties that underpin their strong zero and few-shot generalization beyond the training distribution. These advancements have naturally inspired the extension of VLMs toward embodied domains.

Vision–Language–Action (VLA) models represent a prominent direction in this research trajectory. They adapt pretrained VLMs to action prediction tasks in robotic settings, with the goal of leveraging the semantic priors and cognition abilities inherited from large-scale vision–language pretraining. The underlying hypothesis is that, if appropriately adapted, VLA models can transfer the visual–semantic representations of their initial VLM to the action domain, enabling generalization to previously unseen scenes, instructions, and scenarios. However, in practice, adapting VLMs to the action modality often introduces new challenges. Several recent studies (Pugacheva et al., 2025;

[1]AXXX, Moscow, Russia [2]ITMO University, Saint Petersburg, Russia [3]MIRAI, Moscow, Russia. Correspondence to: Nikita Kachaev <kachaev@axxx.tech>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

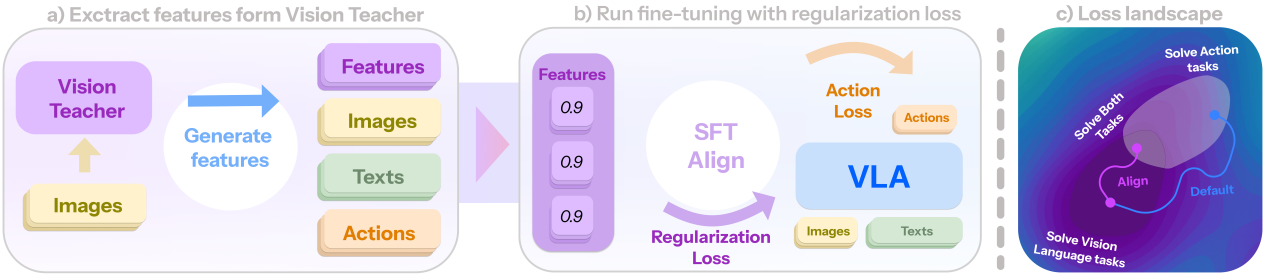

Figure 2. Overview of the proposed method. (a, b) Training pipeline with visual alignment loss – no extra overhead, only precomputed teacher features and a lightweight regularization term during SFT. (c) Conceptual illustration of the loss landscape for VL tasks: the core idea is to optimize the model with respect to the action objective while preserving performance on VL understanding.

Driess et al., 2025; Liu et al., 2023; Mees et al., 2022; Chen et al., 2023) have shown that current VLA models struggle to maintain generalization in visually and linguistically complex tasks, raising questions about whether strong VL capabilities of VLMs truly transfer to embodied settings. This issue becomes the most evident during task-specific fine-tuning, where limited data diversity and datasets frequently lead to overfitting (Zang et al., 2024; Staroverov et al., 2023; Ding et al., 2024; Pugacheva et al., 2025; Cherepanov et al., 2025b). During large-scale robotic pretraining, recent works have attempted to mitigate this degradation by preserving multimodal understanding capabilities. Prior strategies include incorporating auxiliary reasoning objectives (Chen et al., 2025), applying multimodal co-training on web-scale data (Yang et al., 2025), or freezing pretrained visual–language backbones to preserve VL representations and improve instruction following (Bjorck et al., 2025; Driess et al., 2025). While these approaches help retain vision–language knowledge and improve generalization, they often depend on heavy supervision, high computational cost, or constrained model architecture. Yet, despite these advances at the pretraining stage, there remain no effective methods to address representation degradation during task-specific supervised fine-tuning (SFT) – the critical phase where VLA models must adapt to certain robotic domains without losing their semantic grounding and VL abilities.

In this work, we adopt a realistic VLA deployment setting: starting from a pretrained VLA and adapting it with limited data for supervised fine-tuning in a chosen embodiment and domain. Under these constraints, we conduct a systematic investigation into the degradation of VL representations and multimodal understanding abilities in VLA models and ask a central question: **Can we design a simple yet effective method to recover the inherited VL representations during fine-tuning on robotic actions?** To answer this question, we first examined the attention maps and feature activations of the VLA model in comparison to VLM's across matched image-instruction pairs from the robotics domain. Our analysis of attention maps revealed that: while the pretrained VLM accurately focuses on task-relevant objects, the fine-tuned VLA models often produce diffuse or misplaced activations, failing to attend to key entities under out-of-distribution (OOD) conditions (Figure 4). Next, we conducted a t-SNE (Van der Maaten & Hinton, 2008) analysis of intermediate representations across VLM's and VLA's layers, which exposed a clear representation collapse (Barbero et al., 2024; Arefin et al., 2024) in VLA models – indicating that standard action fine-tuning compresses diverse internal features into a narrow representation space, reducing representational diversity and generalization capacity. Next, we propose VL-Think task suite (section 3) to assess transfer of VL knowledge from VLMs to VLA models, benchmark several strong VLMs and compare OpenVLA–7B (Kim et al., 2024) to its pretrained base (PrismaticVLM (Karamcheti et al., 2024)). We observe systematic, domain-specific forgetting after action fine-tuning, indicating that VLAs lose VL knowledge about domains absent from the robotics fine-tuning data. To address this representational degradation, we introduce a lightweight **Visual Representation Alignment** method inspired by the *Platonic Representation Hypothesis* (Huh et al., 2024). This hypothesis suggests that large vision and language models tend to converge toward a shared latent representation space that encodes general visual and semantic representations across generalist models. Our method explicitly constrains the visual representations of a VLA to remain aligned with a generalist vision model throughout fine-tuning. By maintaining this link, the VLA preserves semantic consistency while adapting its action policy to new tasks. The method adds negligible computational overhead and integrates seamlessly with SFT (Figure 2). Extensive experiments on different variations of Simpler (Li et al., 2024) benchmark demonstrates that this alignment consistently improves out-of-distribution generalization – yielding up to a 10% relative gain over naive SFT (Table 1).

Our key contributions are as follows:

1. We systematically demonstrate that naive VLA fine-tuning induces representation collapse and attention

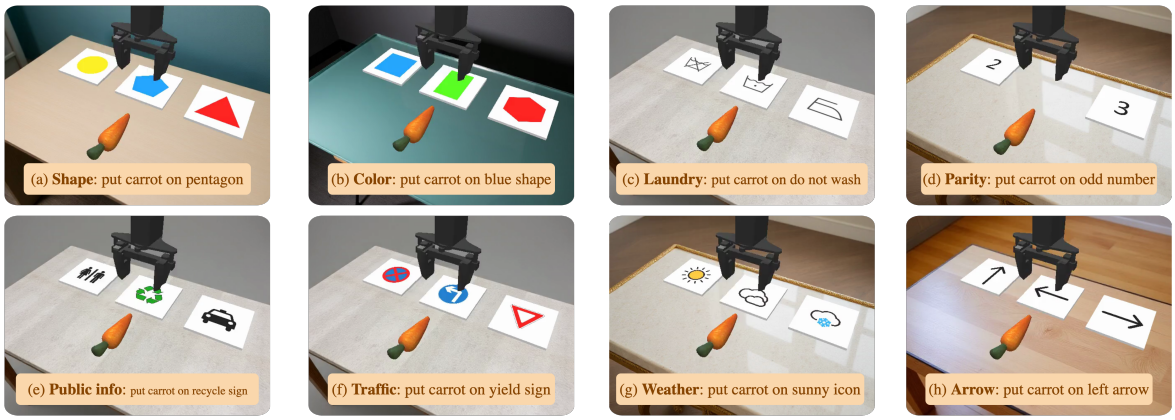

Figure 3. **VL-Think Task Suite examples.** Each panel illustrates a pick-and-place episode where the agent must place an object on the board matching the instructed concept (e.g., color, number, symbol, or category).

sink relative to their initial VLM.

2. We introduce VL-Think, a diagnostic task suite for assessing transfer of VL knowledge from VLMs across VLA models and show that VLA action fine-tuning lead to domain-specific forgetting.

3. We propose a simple and efficient visual alignment method that anchors the VLA's vision representations to strong visual teacher features, preserving multimodal understanding and improving OOD generalization without added complexity (Figure 2).

Taken together, our findings provide new insights into the trade-off between action fine-tuning and representation degradation in VLA models. They underscore the importance of maintaining visual-language alignment during fine-tuning and provide a practical recipe for building VLAs that do not "blind" the pretrained perceptual knowledge they rely upon.

2. Preliminaries

VLA architecture. Let the input multimodal token sequence to the VLM backbone be

$$x_{1:n} = [x_{1:k},\, x_{k+1:n}], \tag{1}$$

where $x_{1:k}$ correspond to visual tokens and $x_{k+1:n}$ correspond to textual instruction tokens. These tokens are obtained from two encoders:

$$x_{1:k} = E_{\text{img}}(I) \in \mathbb{R}^{k\times d_e}, x_{k+1:n} \quad = E_{\text{text}}(\ell) \in \mathbb{R}^{(n-k)\times d_e}. \tag{2}$$

Here, E_{img} and E_{text} denote the image and text encoders into the common embedding space of dimension d_e, and I and ℓ are the input image and textual instruction, respectively. The combined sequence $x_{1:n}$ is processed by a multimodal Transformer backbone $B_\theta : \mathbb{R}^{n\times d_e} \to \mathbb{R}^{n\times d_e}$ with L stacked layers. Denote the hidden states after layer i by $h^i_{1:n} \in \mathbb{R}^{n\times d_e}$. Each layer updates the hidden states using self-attention and a feed-forward network $h^0_{1:n} = x_{1:n}$:

$$h^i_{1:n} = \text{Attn}(h^{i-1}_{1:n}) + \text{FFN}(h^{i-1}_{1:n}), \qquad i = 1, \ldots, L. \tag{3}$$

3. VL-Think Task Suite

Current evaluations of VLA models (Liu et al., 2023; Mees et al., 2022; Cherepanov et al., 2025a) primarily emphasize task execution under distribution shifts – such as changes in objects, scenes, recall-based demands or textures but provide little insight into whether the VL capabilities and knowledge inherited from the pretrained VLM are preserved after action fine-tuning. To address this gap, we introduce the **VL-Think Task Suite**, a diagnostic suite designed to evaluate the transfer of VL capabilities from VLMs to VLAs independently of their low-level control performance. The suite focuses on testing whether a model continues to understand visual symbols, compositional cues, and categorical distinctions that are commonly evaluated in VLM datasets but underrepresented in robotics domain – rather than whether it can successfully execute grasp or placement actions. We intentionally minimize control complexity to ensure that any observed performance degradation reflects a loss of VL understanding, rather than action execution.

3.1. Evaluation protocol

To quantify the gap in VL capabilities, we perform evaluations across both VLA and VLM models. **VLA evaluation.** The agent observes RGB frames and language instructions. The success rate is recorded if a well-known object is placed on the correct target board. Since motion complexity is fixed, this directly measures the model's capacity to ground language in visual categories rather than its manipulation skills. **VLM evaluation.** To assess reasoning

in robotics setup without actions, the same scenes are presented as static initial images with the probe: *"Do you see the ¡board_name¿?". Answer 'yes' or 'no'. If yes, specify where: 'left', 'center', or 'right'"*. A response is counted as successful only if both the predicted board and its target location match the ground truth, yielding a success rate that serves as an action-free measure of semantic grounding.

3.2. VL-Think description

VL-Think reduces embodiment- and setup-specific bottlenecks by using the realistic Simpler (Li et al., 2024) WidowX-250S pick-and-place task. Each episode has one well-known source object, a carrot, placed for 100% grasp reliability, and several planar "boards" textured with abstract categories. The instruction names a target concept, such as shape, color, icon class, direction, or parity, and success requires placing the carrot on the matching board. Fixing the object and action complexity isolates VL skills while bounding execution complexity. The suite contains eight board-selection tasks (Figure 3): **Shape**, named geometric shapes; **Color**, named colors; **Traffic**, 24 common traffic signs; **Laundry care**, 17 standard laundry symbols; **Weather**, 9 common weather icons; **Directional arrow**, arrows pointing "up", "down", "left", or "right"; **Public information**, 14 public-information signs; and **Numeral parity**, numerals matching "odd" or "even" parity.

4. VL representations analysis

In this section, we ask: what happens to VL representations and knowledge in VLA models after action fine-tuning? Does knowledge transfer from VLMs actually occur, and is strong semantic grounding retained? To examine how strongly VL representations degrade in VLA models, we conduct complementary analyses. First, we use t-SNE (Van der Maaten & Hinton, 2008) visualization to assess whether the model preserves a structured and separable latent space for instruction-related tokens. Second, we analyze attention maps to evaluate how accurately the model focuses on objects referenced in the input instruction. Finally, using the VL-Think suite, we assess the transferability of VLM VL skills to VLA policies. Together, these methods provide intuitive and interpretable diagnostics of VL representation degradation and domain forgetting – revealing whether the model maintains focused visual grounding, coherent latent organization and erodes domain-specific knowledge after action fine-tuning.

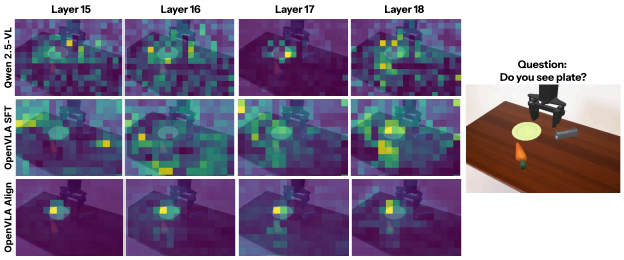

Figure 4. Attention map comparison: the strongest and most semantically grounded attention appears around middle layers. OpenVLA fine-tuned with our proposed method (OpenVLA Align) maintains object-aligned focus in attention maps, while default OpenVLA SFT shows diffused and noisy patterns, indicating loss of visual-language grounding (for more results see Appendix Figure 6).

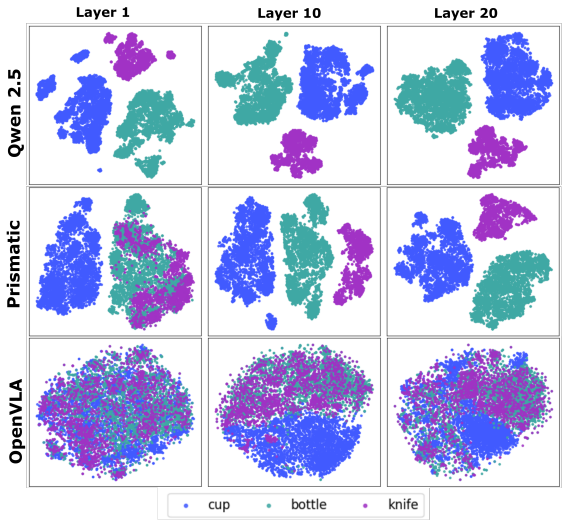

Figure 5. t-SNE visualization of token embeddings for Qwen2.5-VL, PrismaticVLM, and OpenVLA. While PrismaticVLM and Qwen2.5-VL maintains well-separated clusters for target objects, OpenVLA shows huge overlap across classes, indicating that action fine-tuning causes representations collapse.

4.1. Attention sink

We examine attention maps to assess whether fine-tuned VLAs preserve object-centric VL grounding. For each model, we visualize attention over visual patch embeddings from middle transformer layers, where prior work suggests vision–language fusion is most active (Zhang et al., 2025). As shown in Figure 4, layers 14–24 contain the strongest semantic attention patterns. Qwen2.5-VL attends clearly to the queried object, whereas OpenVLA produces diffuse, noisy maps weakly correlated with the target, often leaking to background regions or distractors, indicating attention sink (Kang et al., 2025; Lin et al., 2025b). Visual Representation Alignment mitigates this degradation: OpenVLA-Align recovers crisper, object-centric attention maps; additional examples are provided in subsection A.4.

4.2. Representations collapse

To assess how action fine-tuning alters internal VL representations in VLA models, we run a t-SNE probe on Qwen2.5-

VL (Bai et al., 2025), PrismaticVLM (Karamcheti et al., 2024), and OpenVLA (Kim et al., 2024), providing a qualitative view of semantic structure in latent space through action training. Using COCO (Lin et al., 2014), we sample images from three household classes (cup, bottle, knife) and query each model with *"Do you see ¡object_name¿?"*. We extract the *¡object_name¿* token embedding from transformer layers and project to 2D with t-SNE, coloring points by class. Figure 5 (middle layers) shows Qwen2.5-VL and PrismaticVLM yield cleanly separated class clusters, consistent with semantically organized VLM features, while OpenVLA produces blurred, overlapping clusters, suggesting robot-control fine-tuning disrupts inherited structure. This reduced separability resembles representation collapse (Barbero et al., 2024; Arefin et al., 2024), where distinct VL representations compress into less discriminative subspaces.

4.3. Domain forgetting in VLA models

Using the VL-Think task suite (section 3), we evaluate VL capabilities across several state-of-the-art VLMs: InternVL3.5 (Wang et al., 2025a), Ovis2.5 (Lu et al., 2025), Qwen2.5-VL (Bai et al., 2025) and focus on OpenVLA–7B (Kim et al., 2024) versus its pretrained base PrismaticVLM (Karamcheti et al., 2024), which we use as an approximate upper bound. This comparison probes how much VL knowledge and semantic grounding skills persist after action fine-tuning. Two clear trends emerge. First, strong VLMs achieve high success rate across all domains, reflecting robust semantic grounding. Second, action fine-tuning induces systematic, domain-specific forgetting in VLA models: relative to its pretrained counterpart, OpenVLA–7B exhibits substantial drops in nearly all domains, with the largest declines in symbolic and abstract categories (traffic, arrows, public information, weather). We hypothesize that VLA models lose knowledge about domains that are absent in robotics fine-tuning datasets. The single domain where transfer persists is *Color*: the success rate remains at the level of the initial VLM, likely because color cues are directly useful for control and are implicitly present in robotics datasets.

5. Method

Following the *Platonic Representation Hypothesis* (Huh et al., 2024), we assume that high-performing vision, language, and multimodal models tend to converge toward a shared latent representation space that captures general semantic and perceptual structure across different modalities. Each modality provides a distinct but compatible view of this shared space, encoding complementary aspects of the same underlying VL regularities. From this perspective, a VLA model can be regarded as a policy that grounds its decision-making in a subset of these multimodal representations. However, during task-specific fine-tuning, the policy's internal features may drift away from this generalized representation space, causing it to lose connection to broad, transferable semantics. To mitigate this effect, we introduce a Visual Representation Alignment objective that anchors the VLA's visual representations to a stable external reference encoding consistent, general-purpose visual semantics (Figure 1).

5.1. Visual representation alignment

We propose a lightweight visual alignment method that recover generalized and semantically consistent visual representations inside a VLA model by regularizing its internal embeddings to remain close to those of a frozen, pretrained vision teacher. In the Platonic interpretation, the teacher encoder provides a more stable and semantically precise projection of the generalized representation space, while the VLA's own representations form a task-adapted approximation of this space. By minimizing their discrepancy, the model is guided back toward a common semantic structure. Let $E^{\star}_{\text{img}}$ denote the frozen teacher encoder that produces patch-level features $z_{1:k} = E^{\star}_{\text{img}}(I) \in \mathbb{R}^{k\times d_t}$, where each patch embedding $z_{m-1:m}$ captures localized visual semantics within the teacher's high-level feature space. Within the VLA model, we select an internal layer $i^{\star}$ that carries semantically rich visual information and extract the corresponding vision tokens $h^{i^{\star}}_{1:k} \in \mathbb{R}^{k\times d_e}$. Since the dimensionalities differ, we propose a projector $P_{\varphi} : \mathbb{R}^{d_e} \to \mathbb{R}^{d_t}$ and define $u_{1:k} = P_{\varphi}\big(h^{i^{\star}}_{1:k}\big)$. We then compute a patch-wise similarity between the student's projected embeddings and the teacher's features:

$$\mathcal{L}_{\text{align}} = -\frac{1}{k}\sum_{j=1}^{k} \text{Sim}\big(u_j,\ z_j\big), \tag{4}$$

This objective encourages the hidden representations from the VLA's latent feature space to remain aligned with the teacher's generalized visual representations, helping preserve perceptual consistency across tasks and environments.

5.2. Objective

The total loss integrates the standard autoregressive action objective with the alignment term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VLA}} + \lambda\,\mathcal{L}_{\text{align}}, \lambda > 0. \tag{5}$$

Here, $\mathcal{L}_{\text{VLA}}$ supervises policy learning within the current environment, while $\mathcal{L}_{\text{align}}$ acts as a regularizer that limits representational drift away from generalized visual features. Gradients propagate through the VLA's visual encoder E_{img}, text encoder E_{text}, and transformer backbone B_{θ}, while the teacher encoder $E^{\star}_{\text{img}}$ remains frozen, serving as a fixed reference to stable perceptual structure. From the Platonic

Table 1. OOD generalization performance across evaluation environments (mean ± SD). The proposed alignment objective yields consistent gains over SFT and frozen-encoder baselines, indicating enhanced robustness to OOD domain shifts.

Method	Semantic					Vision					Execution		
	Carrot	Instruct	MultiCarrot	MultiPlate	Plate	VisionImg	Tex03	Tex05	Whole03	Whole05	Position	EEPose	PosChangeTo
OpenVLA	0.49±0.02	0.74±0.02	0.28±0.02	0.43±0.02	0.73±0.02	0.81±0.01	0.67±0.01	0.55±0.03	0.71±0.02	0.56±0.01	0.43±0.02	0.34±0.01	**0.23±0.01**
OpenVLA-Freeze	0.03±0.01	0.05±0.01	0.01±0.01	0.02±0.01	0.03±0.01	0.02±0.01	0.03±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.03±0.01	0.03±0.01	0.04±0.01
OpenVLA-Align	**0.61±0.01**	**0.83±0.03**	**0.35±0.02**	**0.49±0.02**	**0.75±0.01**	**0.86±0.02**	**0.70±0.02**	**0.67±0.02**	**0.79±0.02**	**0.60±0.02**	**0.58±0.02**	**0.38±0.02**	0.20±0.03

viewpoint, our method maintains an semantic prior to shared, generalized VL knowledge. Action fine-tuning alone narrows the model's perceptual space toward the statistics of a specific dataset or embodiment, causing the internal features to drift away from broad generalized representations. The alignment loss restores this balance by enforcing consistency between the student's intermediate features and those of a strong, pre-trained vision model that encodes more general visual–semantic relationships.

6. Experiments

6.1. Evaluation setup

We evaluate our approach in Simpler-based robotics environments (Li et al., 2024; Tao et al., 2024) using the proposed VL-Think suite (section 3), diverse suite of long-horizon, language-conditioned manipulation tasks LIBERO (Liu et al., 2023) and the RL4VLA (Liu et al., 2025) benchmark, which measures VLA generalization along three axes. **Vision**: foreground/background changes via dynamic textures and image-level noise, testing robustness to weak and strong visual perturbations. **Semantics**: unseen objects/receptacles, paraphrased instructions, and multi-object or distractor setups that stress compositional reasoning. **Execution**: randomized initial poses and mid-episode object repositioning, probing action-level robustness. OOD evaluation holds out at least one factor per axis (9 novel objects, 16 unseen receptacles, 5 new textures, 16 distractor backgrounds). We also run linear probing on ImageNet-100 (Tian et al., 2020) to assess representation quality. Each variant is evaluated over 128 seeds; we report mean success ± SD. In subsection A.2, we use a paired one-sided Wilcoxon signed-rank test (Wilcoxon, 1945) and report p-values. All models use identical epochs and hyperparameters for fair comparison.

Table 2. Success rate on LIBERO. Representation Alignment improves performance over the corresponding naive SFT baseline across most suites and all evaluated VLA models.

Method	Spatial	Object	Goal	Long
OpenVLA	85.2	89.0	90.4	76.8
OpenVLA-Align	**93.2**	**96.4**	**95.6**	**89.4**
$\pi_{0.5}$	94.4	92.4	91.6	92.0
$\pi_{0.5}$-Align	**96.8**	**95.0**	**92.4**	**93.2**
SmolVLA	89.0	91.8	90.2	63.4
SmolVLA-Align	**93.8**	**95.2**	**92.8**	**81.0**

6.2. Training setup and baselines

For RL4VLA (Liu et al., 2025) supervised fine-tuning, we collect 1400 MPLib-generated expert demonstrations (Guo et al., 2025) across 16 tables, 16 objects, and multiple pose perturbations, with ∼5 episodes per training variation. LoRA adapters (Hu et al., 2022) are applied to all linear VLA layers. We evaluate OpenVLA (Kim et al., 2024), $\pi_{0.5}$ (Black et al., 2025a), and SmolVLA (Shukor et al., 2025) with three variants: **Default** SFT using cross-entropy on demonstrations, **Freeze** with the visual encoder frozen and **Align**, which adds our Visual Representation Alignment loss from subsection 5.1 to anchor VLA visual features to a pretrained generalist vision teacher.

6.3. Results: OOD Evaluation

Results in Table 1 show that our Visual Alignment method yields consistent improvements across all evaluation axes highlighting the effectiveness of Visual Representation Alignment for robustness to visual shifts, text instruction variations, texture changes, and background perturbations common in real-world scenarios. These gains suggest that aligning internal visual-language embeddings both stabilizes perception and strengthens semantic grounding. Conversely, the Freeze baseline fails across all categories (as also observed in (Wang et al., 2025b)), yielding near-zero performance. Without joint optimization, frozen features become mismatched with evolving action components, severely degrading perception and control. Overall, these results shows that Visual Alignment enabling the model to recover general-purpose visual semantics while adapting to new robotic environments.

6.4. Results: LIBERO

Using LIBERO task suite, we compare OpenVLA, $\pi_{0.5}$ and SmolVLA naive SFT baselines to their aligned counterparts trained using Visual Representation Alignment. Table 2 shows that alignment consistently improves performance over default SFT for most VLA models on LIBERO. This shows that Visual Representation Alignment is model and benchmark agnostic and applies across multiple VLA archi-

Table 3. VL-Think VLM results across eight domains. The benchmark reveals a strong correlation between VL understanding and model scale: larger VLMs achieve higher overall success. However, OpenVLA–7B fine-tuned for action shows clear VL degradation: its performance drops markedly compared to the original PrismaticVLM across all domains except color, where VL skills remain largely preserved.

Model	Arrow	Color	Laundry	Parity	PublicInfo	Shape	Traffic	Weather
InternVL3.5-4B	0.76±0.01	0.91±0.01	0.26±0.02	0.60±0.03	**0.88±0.02**	0.78±0.02	0.71±0.02	0.71±0.01
InternVL3.5-8B	0.73±0.05	0.86±0.02	0.21±0.02	0.53±0.05	0.77±0.00	0.76±0.02	0.54±0.03	0.73±0.03
Ovis2.5-2B	0.86±0.01	**0.96±0.01**	0.34±0.02	**0.74±0.04**	0.78±0.02	0.92±0.02	0.73±0.01	**0.95±0.01**
Ovis2.5-9B	**0.99±0.01**	0.92±0.02	**0.51±0.01**	0.56±0.04	0.84±0.03	**0.94±0.01**	**0.76±0.02**	**0.95±0.00**
Qwen2.5-7B	0.58±0.03	0.73±0.02	0.16±0.02	0.49±0.04	0.34±0.02	0.70±0.06	0.42±0.04	0.48±0.01
Prismatic-DS-7B	0.47±0.03	0.69±0.03	0.37±0.03	0.45±0.03	0.62±0.03	0.59±0.03	0.48±0.03	0.62±0.03
OpenVLA-7B	0.26±0.02	0.69±0.02	0.30±0.03	0.43±0.02	0.24±0.02	0.40±0.02	0.29±0.02	0.32±0.03
OpenVLA-7B-Align	0.24±0.02	0.82±0.02	0.29±0.03	0.42±0.03	0.30±0.03	0.48±0.02	0.28±0.03	0.27±0.02

tectures and task suites. Notably, $\pi_{0.5}$ is initialized from a knowledge-insulated (Driess et al., 2025) checkpoint, yet alignment still improves success rate over SFT. This indicates our method remains effective alongside pretraining-time techniques that mitigate representation degradation, making the two approaches complementary.

6.5. Results: Linear probing

We assess representation quality via ImageNet-100 linear probing (Tian et al., 2020). Following standard practice (Yu et al., 2024; Huh et al., 2024), we freeze C-RADIOv3 (Heinrich et al., 2025) and OpenVLA variants, extract patch embeddings from the final teacher layer or intermediate VLA visual layers, and train a linear classifier on frozen features. Table 4 shows that C-RADIOv3 performs best overall, while OpenVLA-Align outperforms both pretrained OpenVLA and naive SFT. The SFT drop confirms representational degradation; alignment mitigates it and improves semantic separability.

6.6. Results: VL-Think

Following subsection 4.3, we evaluate OpenVLA fine-tuned with our visual representation alignment (OpenVLA-7B-Align) under identical data, budget, and evaluation settings. Table 3 shows SFT-Align partially mitigates domain forgetting compared to default SFT: *Color* and *Shape* improve, even surpassing the PrismaticVLM upper bound, while other domains remain largely unchanged. This highlights both the promise and limits of representation alignment under constrained settings. We hypothesize that limited SFT data breadth and LoRA expressivity are insufficient to recover rarer VL concepts underrepresented in robotics data. Expanding data diversity and relaxing parameter-efficiency constraints may yield broader gains, which represent promising direction for future work.

Table 4. Linear probing results on ImageNet-100

Model	Accuracy (%)
C-RADIOv3	**87.31**
OpenVLA Align	<u>82.13</u>
OpenVLA Pretrained	79.88
OpenVLA SFT	77.48

6.7. Ablation

We conduct extensive ablations of the alignment design, with full detailed results reported in subsection A.2 subsection A.7. The best configuration aligns middle-layer VLA backbone features to the final patch embeddings of a strong visual teacher, C-RADIOv3, using a frozen MLP projector and cosine loss with $\lambda = 0.2$. These results indicate that stronger teacher representations, middle-layer alignment, and preventing the projector from absorbing the alignment objective are key to effective Visual Representation Alignment.

7. Conclusion

In this work, we examined how fine-tuning VLA models on robotic tasks leads to degradation of VL understanding and representation quality. To analyze this effect, we introduced the VL-Think diagnostic suite and interpretability probes, including attention map analyses and linear probing, which reveal how VL skills degrade during action fine-tuning. To address this issue, we proposed a lightweight Visual Alignment method that anchors the VLA to its pretrained visual teacher, consistently improving OOD generalization across diverse domains including novel objects, unseen scene compositions, texture and lighting variations, and instruction paraphrases. Due to compute constraints, our study focused on fine-tuning rather than full-scale pretraining. We hope this study guides future efforts toward scalable robotic pretraining and systematic evaluation of how VLAs inherit and retain VL knowledge from VLMs.

Impact Statement

This work aims to improve the reliability and generalization of vision-language-action (VLA) models by studying how action fine-tuning affects the visual-language representations inherited from pretrained vision-language models. The proposed diagnostics and alignment method may help researchers better understand, evaluate, and mitigate representation degradation in embodied agents, which is important for developing more robust robotic systems under visual, semantic, and instruction-level distribution shifts.

At the same time, improved generalization in robotic policies can increase the practical capability of autonomous systems, including systems that act in physical environments. If deployed without sufficient safety validation, such systems may cause physical harm, misinterpret human instructions, or behave unreliably under unseen conditions. Our experiments are conducted in simulated and benchmarked settings, and the reported improvements should not be interpreted as sufficient evidence for safe real-world deployment. We also note that VLA models can inherit biases, spurious correlations, and failure modes from their pretrained vision-language backbones and teacher models. We therefore view our method as a step toward more diagnosable and robust embodied models, rather than a replacement for task-specific safety testing, human oversight, and real-world validation.

LLM/Agent Usage Disclosure

The authors used large language models only for language polishing and LaTeX formatting. All scientific content, experiments, analyses, and conclusions were produced and verified by the authors.

References

Arefin, M. R., Subbaraj, G., Gontier, N., LeCun, Y., Rish, I., Shwartz-Ziv, R., and Pal, C. Seq-vcr: Preventing collapse in intermediate transformer representations for enhanced reasoning. *arXiv preprint arXiv:2411.02344*, 2024.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Azzolini, A., Bai, J., Brandon, H., Cao, J., Chattopadhyay, P., Chen, H., Chu, J., Cui, Y., Diamond, J., Ding, Y., et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., and et al. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Barbero, F., Banino, A., Kapturowski, S., Kumaran, D., Madeira Araújo, J., Vitvitskyi, O., Pascanu, R., and Veličković, P. Transformers need glasses! information over-squashing in language tasks. *Advances in Neural Information Processing Systems*, 37:98111–98142, 2024.

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

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

Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., Galliker, M. Y., Ghosh, D., Groom, L., Hausman, K., ichter, b., Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Ren, A. Z., Shi, L. X., Smith, L., Springenberg, J. T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Yu, L., and Zhilinsky, U. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In Lim, J., Song, S., and Park, H.-W. (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 27–30 Sep 2025a. URL https://proceedings.mlr.press/v305/black25a.html.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L. X., Smith, L., Tanner, J., Vuong, Q., Walling, A., Wang, H., and Zhilinsky, U. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025b. doi: 10.15607/RSS.2025.XXI.010.

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

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Chen, W., Belkhale, S., Mirchandani, S., Mees, O., Driess, D., Pertsch, K., and Levine, S. Training strate-

gies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243*, 2025.

Chen, Z., Kiami, S., Gupta, A., and Kumar, V. Genaug: Retargeting behaviors to unseen situations via generative augmentation. *arXiv preprint arXiv:2302.06671*, 2023.

Cherepanov, E., Kachaev, N., Kovalev, A. K., and Panov, A. I. Memory, benchmark & robots: A benchmark for solving complex tasks with reinforcement learning, 2025a. URL https://arxiv.org/abs/2502.10550.

Cherepanov, E., Kovalev, A. K., and Panov, A. I. Elmur: External layer memory with update/rewrite for long-horizon rl, 2025b. URL https://arxiv.org/abs/2510.07151.

Ding, C., Gao, X., Dong, S., He, Y., Wang, Q., Kot, A., and Gong, Y. Lobg: less overfitting for better generalization in vision-language model. *arXiv preprint arXiv:2410.10247*, 2024.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., et al. Palm-e: An embodied multimodal language model. 2023.

Driess, D., Springenberg, J. T., Ichter, B., Yu, L., Li-Bell, A., Pertsch, K., Ren, A. Z., Walke, H., Vuong, Q., Shi, L. X., et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv preprint arXiv:2505.23705*, 2025.

Fan, D., Tong, S., Zhu, J., Sinha, K., Liu, Z., Chen, X., Rabbat, M., Ballas, N., LeCun, Y., Bar, A., et al. Scaling language-free visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 370–382, 2025.

Guo, Y., Zhang, J., Chen, X., Ji, X., Wang, Y.-J., Hu, Y., and Chen, J. Improving vision-language-action model with online reinforcement learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15665–15672. IEEE, 2025.

Heinrich, G., Ranzinger, M., Yin, H., Lu, Y., Kautz, J., Tao, A., Catanzaro, B., and Molchanov, P. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22487–22497, 2025.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Huang, X., Wu, J., Xie, Q., and Han, K. Mllms need 3d-aware representation supervision for scene understanding. *arXiv e-prints*, pp. arXiv–2506, 2025.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Jain, J., Yang, Z., Shi, H., Gao, J., and Yang, J. Elevating visual perception in multimodal llms with visual embedding distillation. *arXiv preprint arXiv:2412.09585*, 2024.

Kang, S., Kim, J., Kim, J., and Hwang, S. J. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025.

Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., and Sadigh, D. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024.

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

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

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

Lin, F., Nai, R., Hu, Y., You, J., Zhao, J., and Gao, Y. Onetwovla: A unified vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2505.11917*, 2025a.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Lin, Z., Lin, M., Lin, L., and Ji, R. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5334–5342, 2025b.

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

Liu, J., Gao, F., Wei, B., Chen, X., Liao, Q., Wu, Y., Yu, C., and Wang, Y. What can rl bring to vla generalization? an empirical study. *arXiv preprint arXiv:2505.19789*, 2025.

Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., Wei, Z., Li, Y., Duan, L., Zhao, J., et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Maniparambil, M., Akshulakov, R., Djilali, Y. A. D., El Amine Seddik, M., Narayan, S., Mangalam, K., and O'Connor, N. E. Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14334–14343, 2024.

Mees, O., Hermann, L., Rosete-Beas, E., and Burgard, W. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

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

Pugacheva, D., Moskalenko, A., Shepelev, D., Kuznetsov, A., Shakhuro, V., and Tutubalina, E. Bring the apple, not the sofa: Impact of irrelevant context in embodied ai commands on vla models. *arXiv preprint arXiv:2510.07067*, 2025.

Shang, J., Schmeckpeper, K., May, B. B., Minniti, M. V., Kelestemur, T., Watkins, D., and Herlant, L. Theia: Distilling diverse vision foundation models for robot learning. *arXiv preprint arXiv:2407.20179*, 2024.

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

Staroverov, A., Gorodetsky, A. S., Krishtopik, A. S., Izmesteva, U. A., Yudin, D. A., Kovalev, A. K., and Panov, A. I. Fine-tuning multimodal transformer models for generating actions in virtual and real environments. *Ieee Access*, 11:130548–130559, 2023.

Tao, S., Xiang, F., Shukla, A., Qin, Y., Hinrichsen, X., Yuan, X., Bao, C., Lin, X., Liu, Y., Chan, T.-k., et al. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

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

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Wang, Y., Ding, P., Li, L., Cui, C., Ge, Z., Tong, X., Song, W., Zhao, H., Zhao, W., Hou, P., et al. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. *arXiv preprint arXiv:2509.09372*, 2025b.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., and Bian, J. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

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

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., and Xie, S. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Zang, Y., Goh, H., Susskind, J., and Huang, C. Overcoming the pitfalls of vision-language model finetuning for ood generalization. *arXiv preprint arXiv:2401.15914*, 2024.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Zhang, W., Huang, Y., Xu, Y., Huang, J., Zhi, H., Ren, S., Xu, W., and Zhang, J. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture. *arXiv preprint arXiv:2509.02359*, 2025.

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

A. Appendix

This appendix provides additional technical details, extended results, and supplementary materials that support and complement the main findings presented in the paper. We include comprehensive ablations on alignment strategies, projector architectures, teacher models, and layer selection and alignment coefficient. These materials aim to enhance the transparency, reproducibility, and interpretability of our proposed method.

A.1. Related Works

A.1.1. Vision-Language-Action models

VLA models aim to unify perception, reasoning, and control through large-scale multimodal learning. Early approaches such as RT-1 (Brohan et al., 2022) and RT-2 (Zitkovich et al., 2023) demonstrated that scaling VL pretraining to robot data enables generalization across diverse manipulation tasks. Subsequent works – including OpenVLA (Kim et al., 2024), Octo (Team et al., 2024), MolmoAct (Lee et al., 2025), OneTwoVLA (Lin et al., 2025a), and π_0 (Black et al., 2025b) – explored large scale robotic pretraining, compact diffusion-based policies, modular reasoning architectures, token-based decision sequencing, and continuous flow-matching policies. Across these models, the shared goal is to couple semantic grounding with low-level motor control in a unified policy, while maintaining efficiency and generalization in real-world settings. A central challenge remains the preservation and retention of VL understanding capabilities during robot fine-tuning.

A.1.2. Representation alignment

Recent studies reveal a consistent pattern: as models scale in parameters, data, and tasks, their representations increasingly align across architectures and modalities. The Platonic Representation Hypothesis (Huh et al., 2024) frames this as convergence to a shared statistical model of reality, independently trained vision and language encoders show semantically compatible spaces, and large language-free visual models reach CLIP-level performance while naturally aligning with text (Maniparambil et al., 2024; Fan et al., 2025;?).

Recent representation learning methods reinforce this trend: REPA (Yu et al., 2024) aligns diffusion hidden states to strong image encoders (faster training, better ImageNet quality), OLA-VLM (Jain et al., 2024) distills multi-teacher targets into intermediate LLM layers via predictive embedding losses, 3DRS (Huang et al., 2025) injects 3D-aware supervision with multi-view correspondence, and Geometry Forcing (Wu et al., 2025) aligns video-diffusion features with a 3D backbone via angular/scale objectives for temporally consistent generations.

A.2. Ablations

In this section, we conduct a systematic ablation study to analyze how different design choices affect the performance of our visual alignment method. We examine the impact of the teacher model used for alignment, the alignment strategy and target layers, the projector type and the loss functions. Together, these experiments provide insights into which components are most critical for effective alignment of visual representations.

A.2.1. Visual teacher models

A key question is teacher choice for reference representations. From a Platonic perspective, each vision foundation encoder projects generalizable visual knowledge differently, and aligning to a stronger teacher better preserves transferable representations in the VLA during fine-tuning. We test whether encoders trained on large-scale, diverse, multi-view data improve alignment and transfer, evaluating DINOv2 (Oquab et al., 2023), SigLIP (Zhai et al., 2023), C-RADIOv3 (Heinrich et al., 2025), and Theia (Shang et al., 2024). Table 5 shows C-RADIOv3 performs best overall, suggesting that more capable models trained on semantically rich, multimodal data provide more stable, generalizable features for alignment that guide the VLA toward transferable, semantically consistent representations and improved robustness across tasks and domains.

A.2.2. Alignment method

We evaluate which VLA level benefits most from visual representation alignment by comparing two paradigms: (i) **Backbone2Enc**, aligning the VLA transformer backbone to the teacher visual encoder's final-layer features, and (ii) **Enc2Enc**, aligning the VLA visual encoder to the teacher's visual encoder's final-layer embeddings. Our experiments (Table 9) show Backbone2Enc consistently performs better, indicating degradation primarily arises in the middle-to-late

fusion layers where VL fusion are most active. Regularizing these middle representations is key to preserving visual–semantic consistency while letting lower layers adapt to domain-specific low-level cues.

A.2.3. Projector type

To evaluate how different projection mappings affect representation alignment, we compare projectors mapping from VLA hidden states $\mathbb{R}^{d_e}$ to the teacher space $\mathbb{R}^{d_t}$ via P_φ. All projectors share identical input–output dimensions but differ in their internal transformation $P_\varphi : \mathbb{R}^{d_e} \to \mathbb{R}^{d_t}$. We examine multiple projection strategies, including linear, cosine-similarity–based, orthogonal, spectral-normalized, FiLM-conditioned, Whitening–affine, and MLP-based mappings. Across all evaluations, a frozen MLP yields the most reliable alignment. Freezing is crucial: when trainable, the model mainly lowers loss by adapting the projector especially with limited alignment data and a large gap ($d_t = 768$, $d_e = 4096$) quickly approximating the teacher space and bypassing representational change. Freezing blocks this shortcut, forcing updates into the student hidden states and producing more semantically grounded, transferable alignment.

A.2.4. Alignment layers

We further investigate which layers within the VLA transformer's backbone should be aligned to achieve the most effective representation recovery. Prior literature on VLM interpretability (Zhang et al., 2025) and our own analyses (subsection 4.1) suggest that middle layers are primarily responsible for VL fusion and semantic grounding, whereas early layers encode low-level features and later layers specialize in action prediction. Accordingly, we perform experiments aligning different types of layers: Early, Middle, Late. The results (Table 8) confirm that the middle layers play a central role in semantic grounding and and aligning them yields the most substantial improvements across generalization axes.

A.2.5. Loss functions and alignment coefficient

Finally, we assess the impact of the alignment loss and its weighting coefficient. We test several variants, including cosine similarity (Cosine), L2, and contrastive NT-Xent (Chen et al., 2020) losses, across alignment coefficients $\lambda = \{0.2, 0.5, 1.0, 3.0\}$. The results demonstrate (Table 6) that Cosine loss achieves the most stable and consistent improvements, particularly when the auxiliary weight is set to $\lambda = 0.2$. This setting effectively constrains representation drift without overpowering the task objective.

A.3. Training Hyperparameters

Table 14 presents the training parameters used during the visual alignment fine-tuning. All other training parameters remained unchanged across both experiments and model variants.

A.4. Attention maps visualization

To further validate the qualitative effect of our alignment objective, we visualize attention maps for Qwen2.5-VL, OpenVLA SFT, and OpenVLA Align (ours) across middle layers of the internal transformer backbone. These layers correspond to the region of strongest vision–language fusion, where attention patterns most directly reflect the model's visual grounding quality.

As shown in Figure 6, the default OpenVLA SFT exhibits diffuse and spatially inconsistent attention, often extending beyond the queried object. In contrast, our OpenVLA Align model restores sharp, localized focus on task-relevant regions.

Table 5. Comparison of pretrained Vision Teachers. Values represent mean within each dimension and p-value (for more details see Table 10 from Appendix).

Teacher	Semantic	Vision	Execution
C-RADIOv3	**0.61**	**0.72**	**0.39**
DINOv2	0.57 (p=0.05)	<u>0.69</u> (p=0.12)	<u>0.37</u> (p=0.43)
SigLIP	0.54 (p=0.01)	0.65 (p=0.03)	0.35 (p=0.09)
Theia	0.56 (p=0.03)	0.67 (p=0.05)	0.36 (p=0.15)

This confirms that the proposed visual alignment effectively mitigates attention sink introduced by naive fine-tuning and preserves coherent object-centered attention.

A.5. t-SNE visualization

In subsection 4.2 we show t-SNE (Van der Maaten & Hinton, 2008) of internal representations for the VLM models and OpenVLA. To keep comparisons strict, we use an out-of-the-box t-SNE implementation with no tuning: perplexity $= 30$, max iterations $= 1000$, fixed random seed $= 42$ (all other parameters at library defaults). These plots are illustrative only (t-SNE distorts global geometry); quantitative conclusions come from linear probing in subsection 6.5.

A.6. Linear probing

Table 13. Linear probing configuration on ImageNet-100 with a frozen backbone and mean-pooled token features.

Parameter	Value
Feature pooling	Mean over visual embs.
Optimizer	SGD (momentum 0.9)
Weight decay	0.0
Learning rate	0.1
Epochs	40
Batch size (train/val/test)	128

For reproducibility and fair comparison, we evaluate representational quality with a frozen-feature linear probe under a single, fixed configuration (see Table 13). Concretely, we extract patch embeddings (mean-pooled to a single vector) from the final C-RADIOv3 teacher and from intermediate visual layers of each OpenVLA variant. A single linear classifier is trained on top of these frozen features. All hyperparameters are held constant across models and layers, and the same random seed and data split are used for every run. Due to computational constraints, we operate on a reduced ImageNet-100. We report top-1 accuracy on the evaluation split, without per-any tuning, so any differences reflect only the underlying representations rather than linear probe tuning changes.

A.7. Additional Ablations

This section provides the complete ablation results that underlie the analyses in subsection A.2, Vision teachers (Table 10), Alignment projectors (Table 9), Alignment layers (Table 14), Alignment coefficients (Table 12). In the ablation studies presented in subsection A.2, we test the hypothesis that a given model variant (denoted B) outperforms the baseline variant (A) in terms of success rate. For each pairwise comparison, we fix all other parameters altering only the component under investigation (e.g., alignment objective, layer depth, projection type). This ensures that any observed performance difference can be attributed solely to the ablated design choice.

To assess statistical significance, we use the paired Wilcoxon signed-rank test (Wilcoxon, 1945), a non-parametric test suited for comparing two matched samples that do not necessarily follow a normal distribution. The unit of analysis is the per-seed success rate over matched trials, evaluated independently for each environment type (Semantic, Vision, Execution). The test uses a one-sided alternative hypothesis (H_1: B ¿ A), corresponding to our directional research question and report the

Table 6. Comparison of different loss functions across generalization dimensions (mean across dimensions, p-value).

Objective	Semantic	Vision	Execution
Cosine	**0.61**	**0.72**	**0.39**
L2	0.54 (p¡0.01)	0.63 (p¡0.01)	0.34 (p=0.05)
InfoNCE	0.57 (p=0.05)	0.64 (p=0.04)	<u>0.36</u> (p=0.21)

Table 7. Comparison of alignment paradigms across generalization dimensions reported as mean across dimensions and p-value.

Method	Semantic	Vision	Execution
Backbone2Enc	**0.61**	**0.72**	**0.39**
Enc2Enc	0.55 (p=0.01)	0.66 (p=0.04)	0.38 (p=0.64)

Table 8. Comparison of different layers for alignment across generalization dimensions (mean across dimensions, p-value) (for detailed results see Table 11 from Appendix).

Method	Semantic	Vision	Execution
Middle	**0.61**	**0.72**	0.39
Early	0.51 (p¡0.01)	0.66 (p=0.04)	0.38 (p=0.85)
Late	0.54 (p=0.03)	0.69 (p=0.83)	0.36 (p=0.52)

exact p-value. All comparisons are conducted over 128 shared random seeds, ensuring that each seed–environment pair is identical across the methods being compared. This careful experimental control allows us to draw meaningful conclusions about the contribution of each individual design choice.

A.8. Different projection approaches

Below, we provide detailed formulations of the various projectors used in subsubsection A.2.3 of our experiments.

Cosine Projector. A normalized linear projection that preserves angular similarity:

$$z = \frac{Wh}{\|Wh\|_2}, \qquad W \in \mathbb{R}^{d_z \times d_{\text{hidden}}}. \tag{6}$$

Orthogonal Projector. A fixed linear transform with orthonormal columns:

$$W^\top W = I_{d_z}, \quad z = Wh. \tag{7}$$

Random Fourier Feature (RFF) Projector. A fixed randomized mapping that implicitly approximates a kernel feature space:

$$z = \sqrt{\frac{2}{D}} \cos(Wh + b), \quad W_{ij} \sim \mathcal{N}(0, \gamma^{-2}),\ b_i \sim \mathcal{U}(0, 2\pi). \tag{8}$$

Whitening–Affine Projector. Combines feature whitening and affine normalization:

$$z = \Lambda^{-1/2}(h - \mu) + b, \tag{9}$$

Spectral–Norm Projector. A constrained linear mapping enforcing bounded operator norm:

$$z = Wh, \qquad \|W\|_2 \leq 1. \tag{10}$$

Table 9. Comparison on projection methods, table shows performance across environments (mean ± SD).

Method	Freeze	Semantic					Vision					Execution		
		Carrot	Instruct	MultiCarrot	MultiPlate	Plate	VisionImg	Tex03	Tex05	Whole03	Whole05	Position	EEPose	PosChangeTo
MLP	✗	0.49±0.05	0.79±0.02	0.27±0.01	0.46±0.02	0.72±0.01	0.85±0.01	0.73±0.02	0.58±0.03	0.77±0.03	0.66±0.02	0.53±0.05	0.31±0.01	0.21±0.04
MLP	✓	**0.61±0.01**	0.83±0.03	**0.35±0.02**	0.49±0.02	**0.75±0.01**	0.86±0.02	0.70±0.02	**0.67±0.02**	**0.80±0.02**	0.60±0.02	0.58±0.02	0.38±0.02	0.20±0.03
Cosine	✓	0.49±0.04	0.82±0.03	0.30±0.06	0.45±0.05	0.73±0.03	0.75±0.03	0.71±0.04	0.54±0.05	0.77±0.04	0.60±0.04	0.51±0.01	0.32±0.05	0.29±0.07
Cosine	✗	0.53±0.02	0.79±0.01	0.34±0.04	**0.55±0.02**	0.69±0.05	0.80±0.01	0.73±0.03	0.60±0.03	0.75±0.03	0.61±0.03	0.55±0.09	0.35±0.03	0.25±0.04
FILM	✗	0.53±0.04	**0.84±0.03**	0.35±0.04	0.44±0.03	0.68±0.01	0.83±0.03	0.72±0.04	0.61±0.02	0.75±0.06	0.60±0.07	0.58±0.09	0.31±0.03	0.26±0.03
FILM	✓	0.52±0.02	0.77±0.02	0.34±0.03	0.43±0.05	0.67±0.03	0.79±0.03	0.75±0.03	0.53±0.04	0.76±0.01	0.64±0.04	0.57±0.02	0.27±0.03	0.24±0.02
OrthProj	✓	0.54±0.01	0.78±0.03	0.31±0.03	0.44±0.06	0.71±0.04	0.84±0.02	0.73±0.01	0.58±0.01	0.72±0.05	0.60±0.03	0.57±0.04	0.29±0.02	0.20±0.02
OrthProj	✗	0.49±0.04	0.77±0.02	0.33±0.02	0.48±0.08	0.70±0.04	0.86±0.03	**0.77±0.01**	0.58±0.05	0.74±0.06	0.63±0.03	0.55±0.08	0.38±0.03	0.23±0.03
RFF	✓	0.46±0.02	0.74±0.03	0.31±0.03	0.43±0.04	0.70±0.02	0.84±0.04	0.72±0.03	0.61±0.03	0.75±0.05	0.56±0.04	0.56±0.02	0.32±0.04	0.20±0.03
RFF	✗	0.56±0.03	0.80±0.02	0.30±0.04	0.48±0.02	0.70±0.02	0.78±0.02	0.72±0.02	0.56±0.02	0.72±0.03	0.62±0.02	0.57±0.01	0.27±0.03	0.27±0.04
Whitening	✓	0.46±0.03	0.74±0.04	0.21±0.02	0.33±0.02	0.64±0.02	0.80±0.05	0.72±0.03	0.57±0.04	0.71±0.02	0.57±0.04	0.41±0.03	0.18±0.01	0.12±0.01
Whitening	✗	0.52±0.03	0.79±0.02	0.32±0.02	0.48±0.02	0.72±0.02	0.81±0.03	0.75±0.05	0.64±0.04	0.77±0.01	**0.68±0.01**	**0.63±0.03**	**0.39±0.02**	**0.32±0.05**
Spectral	✓	0.53±0.02	0.81±0.03	0.34±0.07	0.51±0.04	0.74±0.05	**0.91±0.03**	0.73±0.03	0.54±0.06	0.75±0.03	0.61±0.03	0.64±0.04	0.32±0.03	0.23±0.04
Default	-	0.49±0.02	0.74±0.02	0.28±0.02	0.43±0.02	0.73±0.02	0.81±0.01	0.67±0.01	0.55±0.03	0.71±0.02	0.56±0.01	0.43±0.02	0.34±0.01	0.23±0.01

Table 10. Comparison on vision teacher models, table shows performance across environments (mean ± SD).

Teacher	Semantic					Vision					Execution		
	Carrot	Instruct	MultiCarrot	MultiPlate	Plate	VisionImg	Tex03	Tex05	Whole03	Whole05	Position	EEPose	PosChangeTo
C-RADIOv3-ViT-L	**0.61±0.01**	0.83±0.03	**0.35±0.02**	**0.49±0.02**	**0.75±0.01**	**0.86±0.02**	0.70±0.02	**0.67±0.02**	**0.80±0.02**	0.60±0.02	0.58±0.02	0.38±0.02	0.20±0.03
DINOv2-ViT-L	0.49±0.02	0.74±0.04	0.31±0.02	0.46±0.03	0.73±0.03	0.80±0.01	0.72±0.04	0.57±0.04	0.72±0.01	**0.65±0.02**	0.55±0.03	0.33±0.02	0.21±0.03
DINOv2-ViT-G	0.55±0.04	**0.84±0.01**	0.32±0.05	0.42±0.03	0.74±0.01	0.83±0.02	**0.72±0.01**	0.60±0.03	0.72±0.02	0.58±0.02	**0.58±0.04**	0.34±0.02	0.20±0.04
Theia	0.52±0.03	0.76±0.04	0.29±0.03	0.39±0.01	0.70±0.01	0.74±0.04	0.70±0.03	0.53±0.05	0.72±0.08	0.60±0.02	0.53±0.02	**0.41±0.03**	0.21±0.04
SigLIP	0.36±0.03	0.65±0.01	0.15±0.02	0.26±0.06	0.57±0.06	0.69±0.05	0.58±0.05	0.47±0.02	0.64±0.05	0.48±0.04	0.52±0.04	0.32±0.02	0.18±0.04
Default	0.49±0.02	0.74±0.02	0.28±0.02	0.43±0.02	0.73±0.02	0.81±0.01	0.67±0.01	0.55±0.03	0.71±0.02	0.56±0.01	0.43±0.02	0.34±0.01	**0.23±0.01**

Table 11. Comparison on different aligning layers, table shows performance across environments (mean ± SD).

Teacher	Layer (L)	Semantic					Vision					Execution		
		Carrot	Instruct	MultiCarrot	MultiPlate	Plate	VisionImg	Tex03	Tex05	Whole03	Whole05	Position	EEPose	PosChangeTo
C-RADIOv3-ViT-L	8	0.49±0.02	0.79±0.02	0.27±0.04	0.45±0.01	0.76±0.04	0.84±0.03	0.72±0.03	0.60±0.02	0.76±0.04	0.59±0.07	0.59±0.08	**0.39±0.03**	0.23±0.06
C-RADIOv3-ViT-L	16	**0.61±0.01**	**0.83±0.03**	**0.35±0.02**	0.49±0.02	0.75±0.01	0.86±0.02	0.70±0.02	**0.67±0.02**	**0.80±0.02**	0.60±0.02	0.58±0.02	0.38±0.02	0.20±0.03
C-RADIOv3-ViT-L	20	0.54±0.02	0.81±0.01	0.31±0.02	0.51±0.02	0.72±0.04	**0.89±0.02**	0.70±0.03	0.63±0.01	0.79±0.01	**0.66±0.03**	0.63±0.02	0.36±0.02	0.23±0.01
C-RADIOv3-ViT-L	22	0.54±0.01	0.77±0.01	0.32±0.02	**0.52±0.04**	**0.77±0.02**	0.79±0.01	0.74±0.01	0.61±0.01	0.72±0.02	0.60±0.01	0.59±0.02	0.32±0.03	0.19±0.01
C-RADIOv3-ViT-L	26	0.54±0.01	0.79±0.01	0.31±0.02	0.46±0.03	0.77±0.03	0.87±0.01	0.76±0.04	0.64±0.01	0.80±0.02	0.61±0.03	0.60±0.03	0.32±0.01	0.25±0.02
C-RADIOv3-ViT-L	30	0.54±0.01	0.79±0.01	0.31±0.02	0.46±0.02	0.77±0.04	0.87±0.01	**0.76±0.04**	0.64±0.01	0.79±0.01	0.61±0.02	**0.60±0.03**	0.32±0.01	**0.25±0.02**
C-RADIOv3-ViT-L	8-12	0.43±0.03	0.80±0.03	0.29±0.05	0.40±0.01	0.71±0.02	0.83±0.03	0.69±0.01	0.54±0.01	0.77±0.03	0.58±0.03	0.53±0.02	0.32±0.03	0.21±0.02
C-RADIOv3-ViT-L	12-16	0.49±0.02	0.74±0.03	0.29±0.03	0.43±0.01	0.73±0.02	0.82±0.01	0.73±0.04	0.65±0.03	0.73±0.01	0.62±0.02	0.52±0.06	0.32±0.02	0.11±0.02
C-RADIOv3-ViT-L	16-22	0.48±0.03	0.82±0.03	0.28±0.03	0.46±0.04	0.73±0.01	0.82±0.05	0.73±0.03	0.55±0.04	0.71±0.02	0.57±0.01	0.60±0.03	0.31±0.03	0.19±0.02

Table 12. Comparison on different aligment coefficients, table shows performance across environments (mean ± SD).

Teacher	Coeff. (C)	Semantic					Vision					Execution		
		Carrot	Instruct	MultiCarrot	MultiPlate	Plate	VisionImg	Tex03	Tex05	Whole03	Whole05	Position	EEPose	PosChangeTo
C-RADIOv3-ViT-L	0.2	**0.61±0.01**	**0.83±0.03**	**0.35±0.02**	**0.49±0.02**	0.75±0.01	**0.86±0.02**	0.70±0.02	**0.67±0.02**	**0.80±0.02**	**0.60±0.02**	**0.58±0.02**	**0.38±0.02**	**0.20±0.03**
C-RADIOv3-ViT-L	0.5	0.54±0.02	0.76±0.01	0.32±0.01	0.45±0.03	0.72±0.02	0.80±0.03	0.72±0.02	0.54±0.02	0.74±0.03	0.58±0.01	0.55±0.03	0.35±0.02	0.18±0.03
C-RADIOv3-ViT-L	1.0	0.51±0.04	0.77±0.02	0.31±0.05	0.47±0.02	**0.79±0.03**	0.80±0.02	0.70±0.03	0.55±0.02	0.73±0.01	0.61±0.03	0.58±0.06	0.32±0.07	0.14±0.06
C-RADIOv3-ViT-L	2.0	0.58±0.03	0.78±0.01	0.32±0.04	0.46±0.08	0.79±0.04	0.86±0.03	0.69±0.05	0.58±0.07	0.77±0.04	0.61±0.06	0.56±0.01	0.33±0.03	0.16±0.06
C-RADIOv3-ViT-L	3.0	0.52±0.10	0.78±0.03	0.24±0.03	0.36±0.02	0.73±0.03	0.80±0.01	**0.79±0.02**	0.66±0.05	0.77±0.04	0.57±0.03	0.55±0.02	0.37±0.03	0.15±0.01

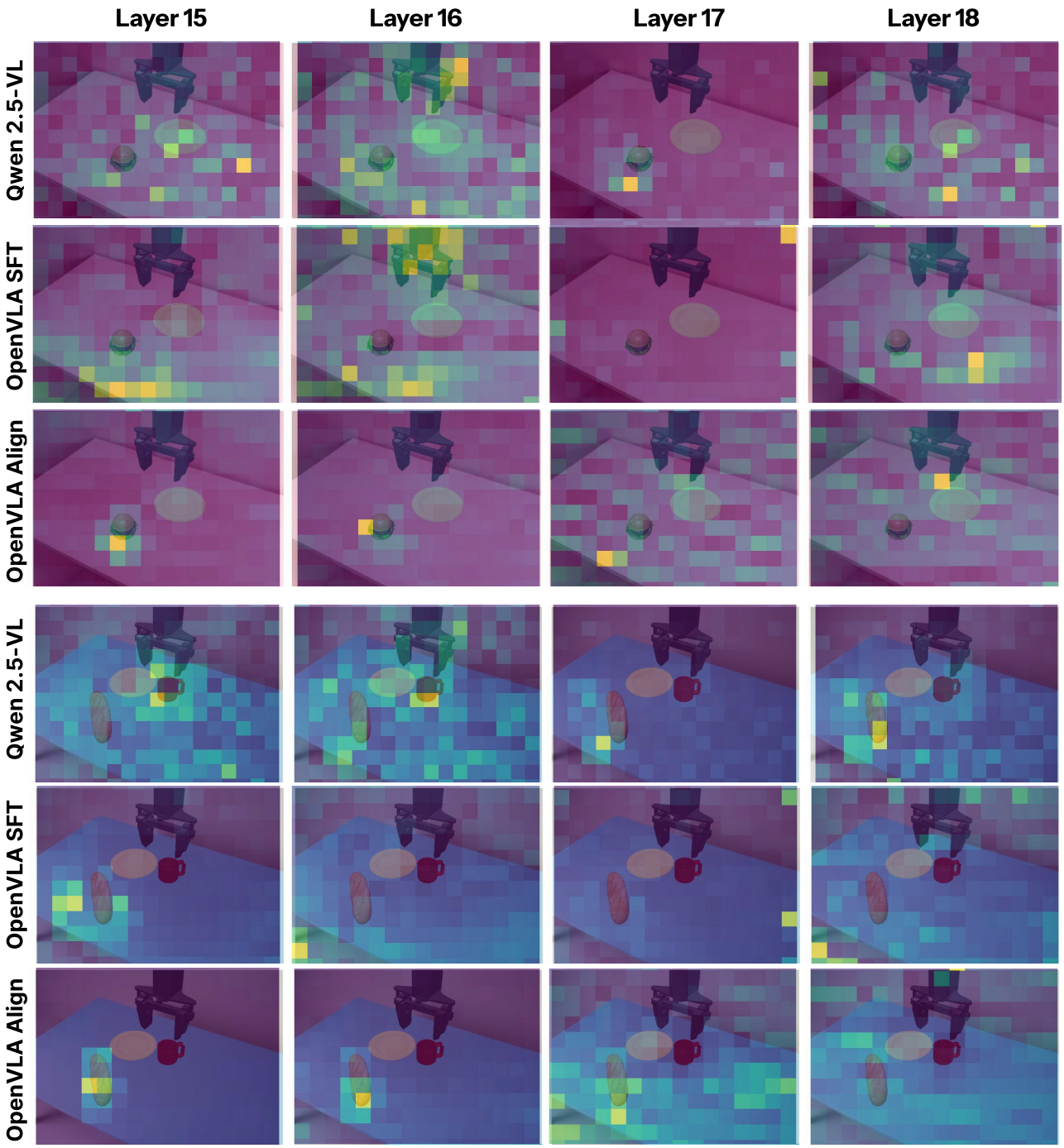

Figure 6. Attention maps across middle layers Qwen2.5-VL, OpenVLA SFT, and OpenVLA Align. The proposed alignment method restores sharp, object-centered attention patterns, improving visual grounding degraded by standard fine-tuning. Question: *"Do you see hamburger—baguette?"*.

Table 14. Best training configuration for OpenVLA fine-tuning with visual alignment. All other methods were trained with identical hyperparameters except for the alignment-specific settings.

Parameter	**Value**
Fine-tuning steps	60000
Batch size	8
Gradient accumulation steps	1
Learning rate	5×10^{-4}
LoRA rank	32
Alignment coefficient	0.2
Alignment projector	MLP
Alignment method	Backbone2Enc
Aligned layers	16
Mode	`alig`
Projector dimension	2048