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

Nikita Kachaev
AXXX & ITMO University
Moscow, Russia
kachaev@axxx.tech

Mikhail Kolosov
MIRAI
Moscow, Russia
kolosov.m@miriai.org

Daniil Zelezetsky
MIRAI
Moscow, Russia
zelezetsky.d@miriai.org

Alexey K. Kovalev
AXXX & MIRAI
Moscow, Russia
kovalev.a@miriai.org

Aleksandr I. Panov
AXXX & MIRAI
Moscow, Russia
panov@axxx.tech

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. Supplementary Material: **blind-vla-paper.github.io**

KEYWORDS

VLA; VLM; Representation Alignment; Robotics; Generalization

ACM Reference Format:
Nikita Kachaev, Mikhail Kolosov, Daniil Zelezetsky, Alexey K. Kovalev, and Aleksandr I. Panov. 2026. Don't Blind Your VLA: Aligning Visual Representations for OOD Generalization. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/PPER9186

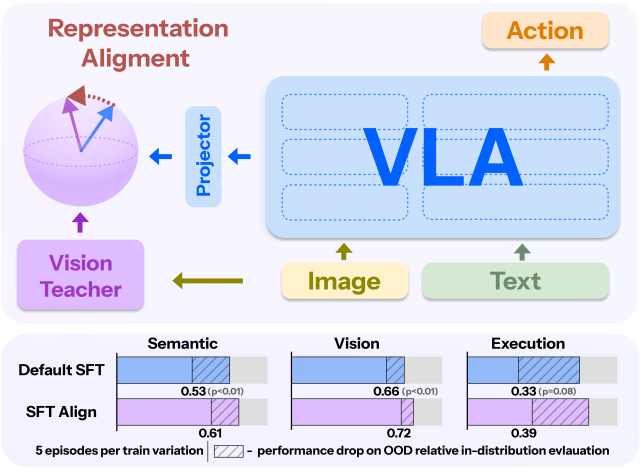

Figure 1: Visual alignment method overview. Mid-level VLA features are projected onto a normalized sphere and aligned with teacher embeddings, preserving visual semantics and improving OOD generalization. Bottom plots show comparison with standard SFT across three generalization axes on the Simpler-based benchmark [35].

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 [2–4, 6, 18, 48]. 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

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/PPER9186

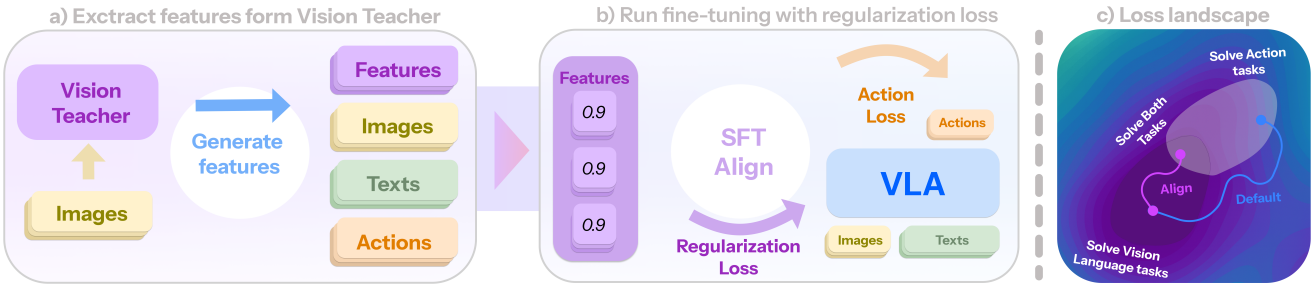

Figure 2: Overview of the proposed method. (a, b) Training pipeline with visual alignment loss – no extra overhead, only precomputed teacher features and a lightweight regularization term during SFT. (c) Conceptual illustration of the loss landscape for VL tasks: the core idea is to optimize the model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

studies [13, 17, 34, 38, 40] 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 [15, 16, 40, 43, 54].

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 [12], applying multimodal co-training on web-scale data [52], or freezing pretrained visual–language backbones to preserve VL representations and improve instruction following [7, 17]. 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 [47] analysis of intermediate representations across VLM's and VLA's layers, which exposed a clear representation collapse [1, 5] 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 4) to assess transfer of VL knowledge from VLMs to VLA models, benchmark several strong VLMs and compare OpenVLA–7B [28] to its pretrained base (PrismaticVLM [27]). 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* [24]. 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 [30] 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 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

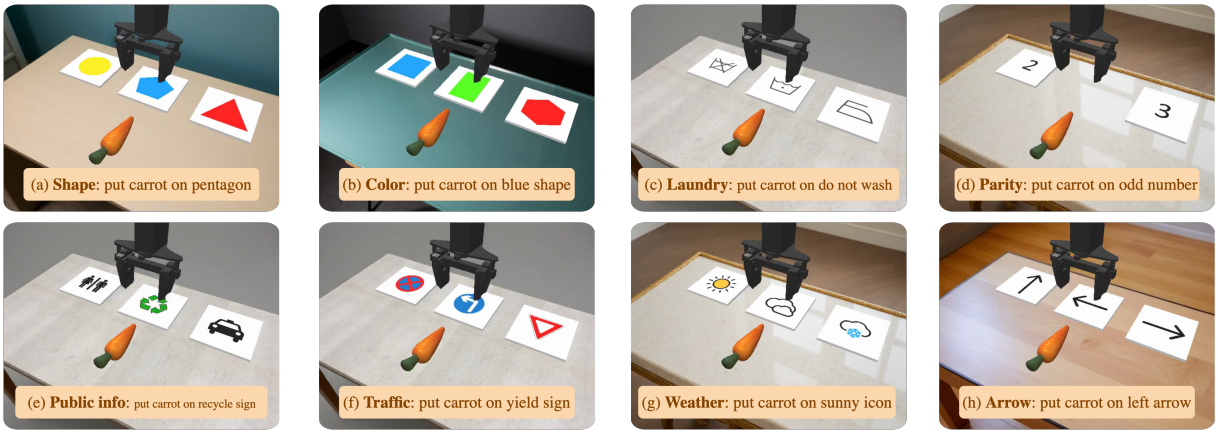

Figure 3: VL-Think Task Suite examples. Each panel illustrates a pick-and-place episode where the agent must place an object on the board matching the instructed concept (e.g., color, number, symbol, or category).

recipe for building VLAs that do not "blind" the pretrained perceptual knowledge they rely upon.

2 RELATED WORKS

2.1 Vision-Language-Action models

VLA models aim to unify perception, reasoning, and control through large-scale multimodal learning. Early approaches such as RT-1 [10] and RT-2 [57] demonstrated that scaling VL pretraining to robot data enables generalization across diverse manipulation tasks. Subsequent works – including OpenVLA [28], Octo [45], MolmoAct [29], OneTwoVLA [31], and π_0 [9] – 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.

2.2 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 [24] 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 [19, 19, 37].

Recent representation learning methods reinforce this trend: REPA [53] aligns diffusion hidden states to strong image encoders (faster training, better ImageNet quality), OLA-VLM [25] distills multi-teacher targets into intermediate LLM layers via predictive embedding losses, 3DRS [23] injects 3D-aware supervision with multi-view correspondence, and Geometry Forcing [51] aligns video-diffusion features with a 3D backbone via angular/scale objectives for temporally consistent generations.

3 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:

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

where E_{image} and E_{text} denote the image and text encoders into the common embedding space of dimension d_e of the VLA model, 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 standard self-attention with $h^0_{1:n} = x_{1:n}$.:

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

4 VL-THINK TASK SUITE

Current evaluations of VLA models [14, 34, 38] 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.

4.1 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.

4.2 VL-Think description

To reduce the embodiment and setup-specific adaptation bottlenecks, VL-Think Task Suite is based on the realistic Simpler [30] benchmark with WidowX-250S arm pick-and-place task. Each episode spawns a single source well-known object (carrot) positioned to yield 100% grasp reliability and multiple planar "boards" textured with abstract categories (e.g., icons, shapes, numerals). A language instruction specifies a single target concept (shape, color, icon class, direction, or parity). The agent succeeds if it places the carrot on the board that matches the instructed concept. By keeping the objects and action complexity fixed, the evaluation isolates VL skills while bounding execution complexity. The VL-Think suite consists of eight board-selection tasks that probe different aspects of knowledge (see Figure 3). In each task, the agent must place the object on the board that matches the instructed concept: **Shape** – the board whose graphic is the named geometric shape; e.g., "Put the object on the star."), **Color** – the board whose shape has the named color; e.g., "Put the object on the blue shape", **Traffic** – the board depicting one of 24 common traffic signs; e.g., "the yield sign", **Laundry care** – the board depicting one of 17 standard laundry symbols, e.g., "Do not bleach", **Weather** – the board depicting one of 9 common weather icons; e.g., "sunny", "cloudy", **Directional arrow** – the board whose arrow points in the named direction: "up", "down", "left", "right", **Public information** – the board depicting one of 14 public-information signs; e.g., "no dogs allowed", and **Numeral parity** – the board whose printed numeral matches the requested parity ("odd" or "even"); e.g., "Put the object on the odd number".

5 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 [47] 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.

5.1 Attention sink

To further investigate how fine-tuning affects the VL grounding capabilities of VLA models, we examine their attention maps, which reveal how effectively the model focuses on the object referenced in a textual instruction. This analysis provides a direct probe into how well the model maintains connection between visual and language features. For each model, we visualize the attention maps for visual patch embeddings from the middle layers. Following prior studies [56], we observe (Figure 4) that the strongest and most semantically meaningful attention patterns typically emerge in the middle transformer layers (layers 14–24), where vision–language fusion is the most active. Among the evaluated models, Qwen2.5-VL exhibits clear and relevant object-aligned attention, indicating that its attention is precisely localized on the queried object with minimal spatial noise. In contrast, OpenVLA displays substantial degradation in attention quality: the maps become diffuse, noisy, and weakly correlated with the target object indicating attention sink [26, 33]. Instead of concentrating on relevant image regions, the OpenVLA's attention maps frequently leak into irrelevant background regions or concentrate on distractor objects (for more results see **??**). By contrast, our proposed Visual Representation Alignment approach remedies this issue: OpenVLA (Align) trained with it produces crisp, object-centric attention maps (see **??** for details).

5.2 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 [4], PrismaticVLM [27], and OpenVLA [28], providing a qualitative view of semantic structure in latent space through action training. Using COCO [32], 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

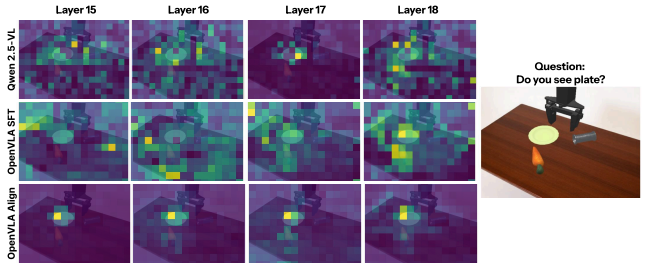

Figure 4: Attention map comparison: the strongest and most semantically grounded attention appears around middle layers. OpenVLA fine-tuned with our proposed method (OpenVLA Align) maintains object-aligned focus in attention maps, while default OpenVLA SFT shows diffused and noisy patterns, indicating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

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 [1, 5], where distinct VL representations compress into less discriminative subspaces.

5.3 Domain forgetting in VLA models

Using the VL-Think task suite (section 4), we evaluate VL capabilities across several state-of-the-art VLMs: InternVL3.5 [48], Ovis2.5 [36], Qwen2.5-VL [4] and focus on OpenVLA–7B [28] versus its pretrained base PrismaticVLM [27], 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.

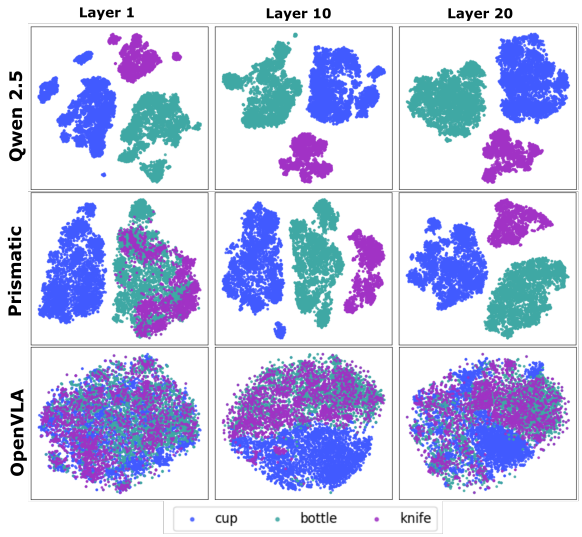

Figure 5: t-SNE visualization of token embeddings for Qwen2.5-VL, PrismaticVLM, and OpenVLA. While PrismaticVLM and Qwen2.5-VL maintains well-separated clusters for target objects, OpenVLA shows huge overlap across classes, indicating that action fine-tuning causes representations collapse.

6 METHOD

Following the *Platonic Representation Hypothesis* [24], 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).

6.1 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

$$z_{1:k} = E^{\star}_{\text{img}}(I) \in \mathbb{R}^{k \times d_t}, \tag{4}$$

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

$$u_{1:k} = P_{\varphi}\big(h^{i^{\star}}_{1:k}\big). \tag{5}$$

We then compute a patch-wise similarity between the student's projected embeddings and the teacher's features:

$$\mathcal{L}_{\text{align}} = -\frac{1}{k}\sum_{j=1}^{k} \text{Sim}\big(u_j,\ z_j\big), \tag{6}$$

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.

6.2 Objective

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

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VLA}} + \lambda\, \mathcal{L}_{\text{align}}, \lambda > 0. \tag{7}$$

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

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

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 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.

7 EXPERIMENTS

7.1 Evaluation setup

We evaluate our approach in Simpler-based robotics environments [30, 44] using the proposed VL-Think suite (section 4), diverse suite of long-horizon, language-conditioned manipulation tasks LIBERO [34] and the RL4VLA [35] 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 [46] to assess representation quality. Each variant is evaluated over 128 seeds; we report mean success ± SD. In section 8, we use a paired one-sided Wilcoxon signed-rank test [50] and report p-values. All models use identical epochs and hyperparameters for fair comparison.

7.2 Training setup

For supervised fine-tuning on RL4VLA [35] benchmark, we collect 1400 expert demonstration trajectories using the MPLib motion planner [20]. Training randomization spans 16 tables, 16 objects (yielding on average ∼5 episodes per training variation), and multiple pose perturbations. During all fine-tuning runs, LoRA adapters [22] are applied to all linear layers of the VLA. For LIBERO we use standard training setup, see Appendix (**??**) for details.

7.3 Baselines

Using a widely adopted open-source OpenVLA [28], $\pi_{0.5}$ [8], and SmolVLA [42] models, we compare our proposed alignment method against several fine-tuning baselines.

- **Default**: standard supervised fine-tuning (SFT) using cross-entropy loss on demonstration data, serving as the primary baseline.
- **Freeze**: SFT with the VLA's visual encoder weights frozen during training, this setup tests the hypothesis that frozen representations might help with generalization.
- **Align**: SFT combined with our auxiliary Visual Representation Alignment loss, described in subsection 6.1, which explicitly anchors the VLA's vision encoder to a pretrained generalist vision teacher.

7.4 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 [49]), 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.

Table 2: Success rate on LIBERO. Representation Alignment method 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**

Table 3: 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

7.5 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 architectures and task suites. Notably, $\pi_{0.5}$ is initialized from a knowledge-insulated [17] 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.

7.6 Results: Linear probing

We evaluate representational quality via linear probing on ImageNet-100 [46]. Specifically, we extract patch embeddings from the final C-RADIOv3 [21] teacher layer and intermediate visual layers of OpenVLA variants. Following standard practice [24, 53], we freeze models and train a linear classifier on frozen features to measure semantic separability, directly testing linear separability after action fine-tuning. Table 3 shows that C-RADIOv3 attains the highest accuracy. Among VLA variants, our Visual Representation Alignment outperforms both the pretrained checkpoint and naive SFT, indicating improved representations during action fine-tuning. Naive SFT sharply reduces accuracy versus pretrained, confirming representational degradation. Our aligned model mitigates this and surpasses pretrained, suggesting the alignment loss strengthens semantic consistency and yields more transferable visual features.

7.7 Results: VL-Think

Following subsection 5.3, we evaluate OpenVLA fine-tuned with our visual representation alignment (OpenVLA-7B-Align) under identical data, budget, and evaluation settings. Table 7 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: Comparison of pretrained Vision Teachers. Values represent mean within each dimension and p-value (for more details see ?? 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)

Table 5: 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)	<u>0.38</u> (p=0.64)

8 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.

8.1 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 [39], SigLIP [55], C-RADIOv3 [21], and Theia [41]. Table 4 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.

Table 6: Comparison of different layers for alignment across generalization dimensions (mean across dimensions, p-value) (for detailed results see ?? from Appendix).

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

Table 7: 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

8.2 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 (**??**) 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.

8.3 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.

Table 8: 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)

8.4 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 [56] and our own analyses (subsection 5.1) 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 6) confirm that the middle layers play a central role in semantic grounding and and aligning them yields the most substantial improvements across generalization axes.

8.5 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 [11] losses, across alignment coefficients $\lambda = \{0.2, 0.5, 1.0, 3.0\}$. The results demonstrate (Table 8) 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.

9 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.

REFERENCES

[1] Md Rifat Arefin, Gopeshh Subbaraj, Nicolas Gontier, Yann LeCun, Irina Rish, Ravid Shwartz-Ziv, and Christopher Pal. 2024. Seq-VCR: Preventing collapse in intermediate transformer representations for enhanced reasoning. *arXiv preprint arXiv:2411.02344* (2024).

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390* (2023).

[3] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558* (2025).

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, and et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923

[5] Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João Madeira Araújo, Oleksandr Vitvitskyi, Razvan Pascanu, and Petar Veličković. 2024. Transformers need glasses! information over-squashing in language tasks. *Advances in Neural Information Processing Systems* 37 (2024), 98111–98142.

[6] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726* (2024).

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

[8] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. 2025. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. In *Proceedings of The 9th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 305)*, Joseph Lim, Shuran Song, and Hae-Won Park (Eds.). PMLR, 17–40. https://proceedings.mlr.press/v305/black25a.html

[9] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. 2025. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*. LosAngeles, CA, USA. https://doi.org/10.15607/RSS.2025.XXI.010

[10] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817* (2022).

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[12] William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. 2025. Training strategies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243* (2025).

[13] Zoey Chen, Sho Kiami, Abhishek Gupta, and Vikash Kumar. 2023. Genaug: Retargeting behaviors to unseen situations via generative augmentation. *arXiv preprint arXiv:2302.06671* (2023).

[14] Egor Cherepanov, Nikita Kachaev, Alexey K. Kovalev, and Aleksandr I. Panov. 2025. Memory, Benchmark & Robots: A Benchmark for Solving Complex Tasks with Reinforcement Learning. arXiv:2502.10550 [cs.LG] https://arxiv.org/abs/2502.10550

[15] Egor Cherepanov, Alexey K. Kovalev, and Aleksandr I. Panov. 2025. EMMUR: External Layer Memory with Update/Rewrite for Long-Horizon RL. arXiv:2510.07151 [cs.LG] https://arxiv.org/abs/2510.07151

[16] Chenhao Ding, Xinyuan Gao, Songlin Dong, Yuhang He, Qiang Wang, Alex Kot, and Yihong Gong. 2024. LOBG: less overfitting for better generalization in vision-language model. *arXiv preprint arXiv:2410.10247* (2024).

[17] Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. 2025. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv preprint arXiv:2505.23705* (2025).

[18] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. (2023).

[19] David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. 2025. Scaling language-free visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 370–382.

[20] Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. 2025. Improving vision-language-action model with online reinforcement learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 15665–15672.

[21] Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. 2025. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 22487–22497.

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr* 1, 2 (2022), 3.

[23] Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 2025. Mllms need 3d-aware representation supervision for scene understanding. *arXiv e-prints* (2025), arXiv–2506.

[24] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987* (2024).

[25] Jitesh Jain, Zhengyuan Yang, Humphrey Shi, Jianfeng Gao, and Jianwei Yang. 2024. Elevating Visual Perception in Multimodal LLMs with Visual Embedding Distillation. *arXiv preprint arXiv:2412.09585* (2024).

[26] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321* (2025).

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

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

[29] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. 2025. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917* (2025).

[30] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. 2024. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941* (2024).

[31] Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. 2025. Onetwovla: A unified vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2505.11917* (2025).

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[33] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5334–5342.

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

[35] Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. 2025. What can rl bring to vla generalization? an empirical study. *arXiv preprint arXiv:2505.19789* (2025).

[36] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. 2025. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737* (2025).

[37] Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E O'Connor. 2024. Do vision and language encoders represent the world similarly?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14334–14343.

[38] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. 2022. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters* 7, 3 (2022), 7327–7334.

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[40] Daria Pugacheva, Andrey Moskalenko, Denis Shepelev, Andrey Kuznetsov, Vlad Shakhuro, and Elena Tutubalina. 2025. Bring the apple, not the sofa: Impact of irrelevant context in embodied ai commands on vla models. *arXiv preprint arXiv:2510.07067* (2025).

[41] Jinghuan Shang, Karl Schmeckpeper, Brandon B May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. 2024. Theia: Distilling diverse vision foundation models for robot learning. *arXiv preprint arXiv:2407.20179* (2024).

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

[43] Aleksei Staroverov, Andrey S Gorodetsky, Andrei S Krishtopik, Uliana A Izmesteva, Dmitry A Yudin, Alexey K Kovalev, and Aleksandr I Panov. 2023. Fine-tuning multimodal transformer models for generating actions in virtual and real environments. *Ieee Access* 11 (2023), 130548–130559.

[44] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, et al. 2024. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425* (2024).

[45] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. 2024. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213* (2024).

[46] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems* 33 (2020), 6827–6839.

[47] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[48] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265* (2025).

[49] Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, et al. 2025. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. *arXiv preprint arXiv:2509.09372* (2025).

[50] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin* 1, 6 (1945), 80–83.

[51] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. 2025. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982* (2025).

[52] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. 2025. Magma: A foundation model for multimodal ai agents. In *Proceedings of the computer vision and pattern recognition conference*. 14203–14214.

[53] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. 2024. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940* (2024).

[54] Yuhang Zang, Hanlin Goh, Josh Susskind, and Chen Huang. 2024. Overcoming the pitfalls of vision-language model finetuning for ood generalization. *arXiv preprint arXiv:2401.15914* (2024).

[55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.

[56] Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. 2025. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture. *arXiv preprint arXiv:2509.02359* (2025).

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