

Instruction Routing in Fine-Tuned Vision-Language-Action Models: A Cross-Architecture Analysis

Anonymous Author(s)

Affiliation

Address

email

Abstract: When vision-language models (VLMs) are fine-tuned for robotic action prediction, which representational changes are associated with training and which are inherited from the pretrained backbone? We compare trained and reference checkpoints of three Vision-Language-Action (VLA) models across four LIBERO benchmark task suites via layer-wise probes, attention profiling, and counterfactual perturbations. We find that VLM-to-VLA fine-tuning is associated with a consistent exit-layer collapse in action-probe decodability ($\Delta R^2 = 0.16\text{--}0.46$) that is absent in VLA-to-VLA transitions ($\Delta R^2 < 0.10$) and in three natural language processing (NLP) instruction-tuning controls ($\Delta R^2 = 0.000$), suggesting this reorganization is specific to action-oriented training rather than a generic fine-tuning artifact. Complementary analyses reveal that trained VLAs develop substantial sensitivity to instruction content while the dominant attention patterns at early layers largely pre-exist in pretrained backbones. These findings provide a cross-architecture decomposition of instruction routing in fine-tuned VLAs, disentangling training-associated representational changes from inherited structure.

Keywords: Vision-Language-Action models, representational analysis, attention routing, fine-tuning

1 Introduction

Vision-language-action models (VLAs) fine-tune pretrained language models on robot demonstrations to produce actions from visual observations and language instructions [1, 2], but which of the resulting representational changes reflect action training and which are inherited from the pretrained backbone? In these decoder-only architectures, image tokens precede instruction tokens, and the causal attention mask restricts multimodal integration to a single readout token [3, 4]; how representations at this critical position are reshaped by action fine-tuning remains unknown.

Behavioral studies show that VLAs exhibit visual dominance [5] and underweight instructions [6], but existing representational analyses target single architectures. No prior work tests whether instruction routing (the internal mechanisms by which instruction tokens influence action-relevant representations) is consistent across backbone families or decomposes training-associated effects from inherited base-model priors.

We address this gap with a cross-architecture analysis of three VLAs (OpenVLA [2], TraceVLA [7], and NORA [8]) across four LIBERO [9] task suites (§4.1). TraceVLA and NORA baselines are pure VLMs ($N = 2$ independent VLM-to-VLA transitions); OpenVLA’s baseline is the Open X-Embodiment (OXE)-pretrained [10] VLA (VLA-to-VLA transition).

Our analysis reveals three findings (Figure 1): (1) VLM-to-VLA fine-tuning is associated with substantial R^2 exit collapse ($\Delta R^2 = 0.16\text{--}0.46$), whereas VLA-to-VLA fine-tuning and reference models show only small fluctuations ($\Delta R^2 < 0.10$), with three NLP controls (all $\Delta R^2 = 0.000$) confirming

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

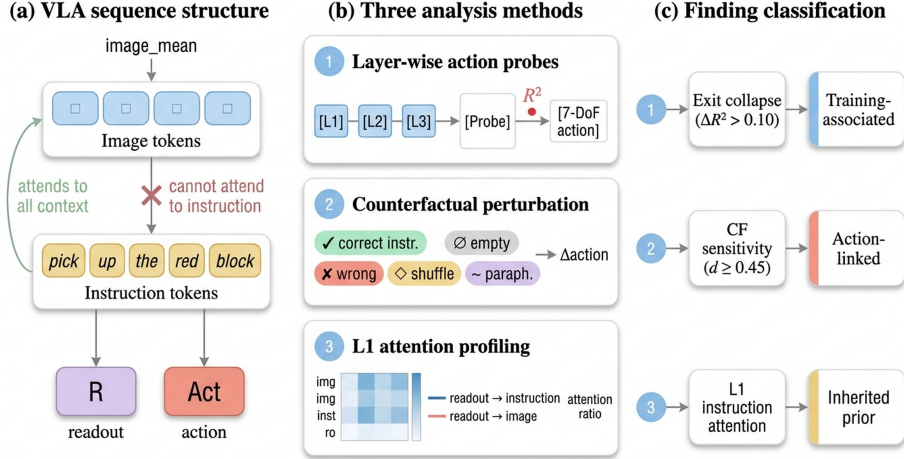


Figure 1: **Analysis framework.** We analyze hidden states at image-token averages and the readout token via layer-wise action probes, counterfactual instruction perturbations, and L1 attention profiling to distinguish training-associated changes from inherited backbone priors.

37 transition-type specificity; (2) all trained VLAs exhibit medium-to-large counterfactual sensitivity
 38 (Cohen’s $d = 0.45$ – 1.52); and (3) first-layer (L1) instruction attention reflects base-model priors,
 39 not VLA training (§4.3–4.5). While concurrent work evaluates VLA instruction-following behav-
 40 iorally [11, 12], we provide complementary representational analysis of how instruction information
 41 is routed internally.

42 Our contributions:

- 43 1. A cross-architecture representational analysis of three VLAs, combining trained/reference
- 44 controls with token-level probing across a 3×4 architecture-suite matrix.
- 45 2. Evidence that R^2 exit collapse is associated with VLM-to-VLA transitions ($\Delta R^2 > 0.10$
- 46 across two architectures) but not with VLA-to-VLA fine-tuning or NLP instruction tuning
- 47 of the same backbones.
- 48 3. A decomposition of VLA instruction routing into training-associated effects and inherited
- 49 base-model priors, based on cross-architecture trained/reference contrasts.

50 2 Related Work

51 **Vision-language-action models.** VLAs fine-tune large language models on robot demonstrations
 52 to produce actions from visual and language input [1, 2, 13, 14, 7, 8, 15]. NLP instruction
 53 tuning is well-characterized behaviorally [16, 17], but the representational effects of VLA fine-
 54 tuning remain unexamined. Fine-tuning operates in a low intrinsic dimension [18] and can dis-
 55 tort pretrained features [19]; recent VLA-specific work proposes representation-preserving strate-
 56 gies [20, 21]. As Table 1 summarizes, existing analyses each target a single architecture and none
 57 include reference-model controls; our study is the first to span multiple backbone families with
 58 systematic trained/reference comparisons.

59 **Mechanistic and behavioral analysis.** Mechanistic interpretability has revealed residual stream
 60 communication [22], induction heads [23], and task circuits [24, 25]; causal tracing localizes fac-
 61 tual recall [26]; attention probing shows heads encode syntactic relations [27] and most can be
 62 pruned [28]. In VLMs, recent work characterizes visual token redundancy [29], progressive visual-
 63 textual alignment [30, 31], and instruction tokens as shallow-layer modality arbitration anchors [32].
 64 Behavioral evaluations of VLAs reveal visual dominance [5, 33, 34]; mechanistic studies identify
 65 sparse semantic directions [6], interpretable features [12, 35], symbolic representations [36, 37],
 66 comprehension-action trade-offs [38], and task-specific attention heads for selective fine-tuning [39].

Table 1: **Comparison with concurrent VLA representational analyses.** Each prior study targets a single architecture and lacks a reference-model control; our work spans three backbone families with systematic trained/reference comparisons using probes, attention profiling, and counterfactual perturbation.

Work	Archs	Method	Ref. ctrl
Grant et al. [5]	1	feature import.	×
Haön et al. [6]	1	SAE / steering	×
Swann et al. [12]	1	SAE features	×
Buurmeijer et al. [35]	1	low-level probes	×
Fang et al. [11]	1	behav. bench.	×
Ours	3	probes + attn + CF	✓

Table 2: Architectures studied. The reference baselines differ in kind: OpenVLA’s is an OXE-pretrained VLA; TraceVLA and NORA references are pure VLMs. MHA = multi-head attention; GQA = grouped-query attention; DinoSigLIP = fused DINOv2+SigLIP vision encoder.

Arch.	Backbone	Layers	Attention	Vision Enc.	Reference
OpenVLA	Llama-2-7B	32	32 MHA	DinoSigLIP → MLP proj.	OXE-pret. VLA
TraceVLA	Phi-3-mini	32	32 MHA	SigLIP + spatial trace	Phi-3-Vision
NORA	Qwen2.5-VL-3B	36	16q/2kv GQA	Qwen2.5-VL native	Qwen2.5-VL

67 Concurrent continual-learning studies find that pretrained VLAs resist forgetting under sequential
68 fine-tuning [40], consistent with our observation that VLA-to-VLA transitions preserve exit-layer
69 structure. All prior representational analyses target a single architecture; our 3×4 cross-architecture
70 matrix tests whether mechanisms replicate across backbone families.

71 **Concurrent work.** Fang et al. [11] evaluate VLA instruction-following behaviorally; our analysis
72 measures representational change (§4.4). Buurmeijer et al. [35] probe low-level motor features in a
73 single VLA; we focus on instruction-level routing across three architectures. Concurrent work on
74 visual [41] and instruction information collapse [42] in VLAs is complementary; our ΔR^2 metric
75 offers a cross-architecture diagnostic.

76 **Probing and similarity methodology.** Representational similarity analysis (RSA) [43], centered
77 kernel alignment (CKA) [44], and linear probes [45, 46] provide complementary geometric and
78 functional evaluation. We adopt RSA and MLP probes (CKA in appendix); the two metrics can
79 diverge on the same data (Appendix A.5), and Cohen’s d vs. cosine distance are not comparable
80 across trained/reference conditions (§4.4).

81 3 Method

82 We investigate how action-relevant representations are organized across three VLA architectures
83 using complementary probing, attention, and perturbation analyses.

84 3.1 Architectures and Data

85 We study three VLAs spanning distinct backbone families (Table 2): **OpenVLA** [2, 13] (Llama-2-
86 7B [47], 32 layers, pretrained on Open X-Embodiment [10] [OXE; 970K episodes]), **TraceVLA** [7]
87 (Phi-3-mini [48], 32 layers, trained on an OXE subset with spatial trace conditioning), and
88 **NORA** [8] (Qwen2.5-VL-3B, 36 layers, grouped-query attention with 16 query heads). Each is
89 evaluated in trained and reference conditions. A critical asymmetry distinguishes the references:
90 OpenVLA’s is an OXE-pretrained VLA (already producing actions), while TraceVLA and NORA
91 references are pure VLMs (Phi-3-Vision, Qwen2.5-VL) with no action training; thus OpenVLA com-
92 parisons measure VLA-to-VLA transitions, while TraceVLA and NORA measure VLM-to-VLA
93 transitions. OpenVLA additionally provides Llama-2 base and random-weight conditions (§4.2).

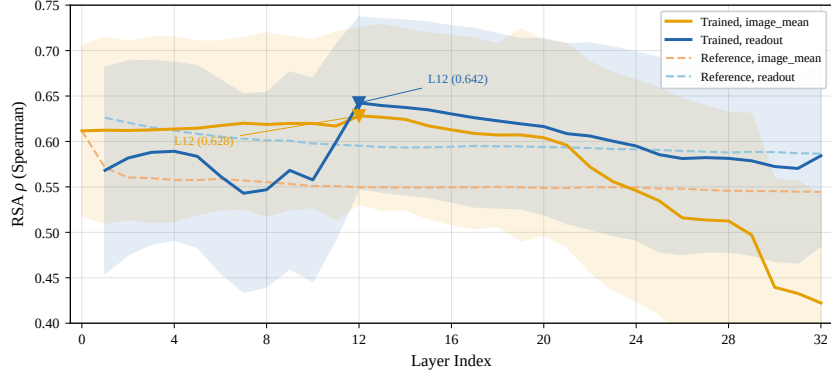


Figure 2: Per-layer RSA (Spearman ρ) for trained/reference OpenVLA. Trained: inverted-U peaking at L12; reference: monotonic decline. Readout and image_mean diverge after L12.

We evaluate on LIBERO [9], comprising four task suites (Goal, Spatial, Object, Long-Horizon) with 10 tasks and 50 demonstrations each. LIBERO provides a controlled evaluation framework: all suites share the same simulated tabletop environment and robot embodiment, isolating instruction complexity as the primary experimental variable. The four suites systematically vary instruction demands (Goal requires object identification, Spatial adds relative positioning, Object introduces distractors, and Long-Horizon chains multiple sub-tasks), enabling us to test whether routing patterns generalize across instruction complexity levels. The three architectures and four suites form a 3×4 matrix; 10 of 12 cells are filled (NORA Object and Long-Horizon lack available checkpoints).

3.2 Representation Extraction

From each architecture, we extract hidden states at every transformer layer from two token positions: (a) **image_mean**, the mean activation over image tokens, and (b) **readout**, the hidden state at the final pre-action token, which can attend to all preceding context. L0 denotes the embedding layer, L1–LN the post-transformer hidden states ($N = 32$ for OpenVLA/TraceVLA, 36 for NORA). We cross training status (trained vs. reference) $\times$ instruction presence (with vs. without) across all 10 cells. OpenVLA additionally provides DinoSigLIP projector, DINOv2, and CLIP representations for the single-architecture analyses (§4.2).

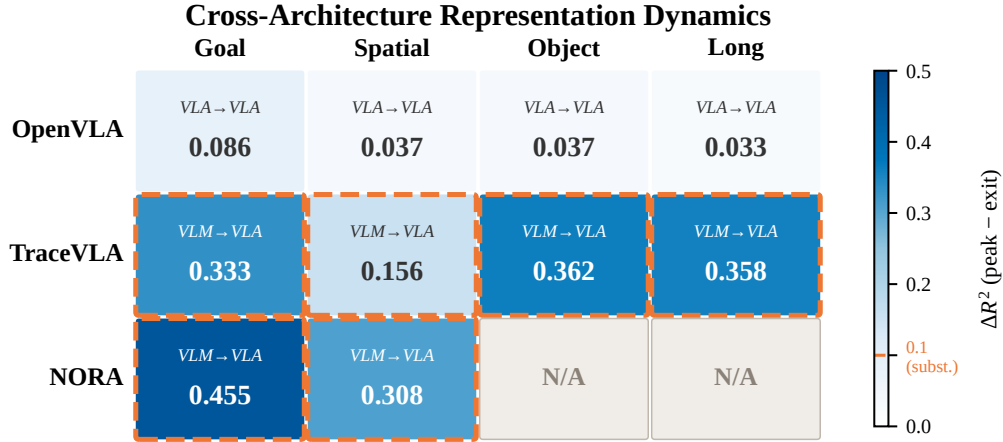
3.3 Cross-Architecture Analyses

Three analyses are applied to every available cell in the 3×4 matrix.

Action probes. Two-layer MLP probes [45] predict 7 degree-of-freedom (DoF) actions from frozen representations (details in Appendix A.2). Probe R^2 measures decodability, not causal usage. $\Delta R^2 = R^2_{\text{peak}} - R^2_{\text{exit}}$; we use > 0.10 as the threshold for substantial peak-to-exit reduction. Demo-level train/test splits and Monte Carlo cross-validation ($n = 1,000$) provide 95% confidence intervals (CIs; Appendix A.19); linear probes yield consistent patterns (Appendix A.14).

L1 attention head profiling. To determine whether attention heads preferentially route instruction information to the readout position, each L1 head is classified as instruction-dominant ($> 50\%$ of readout-token attention to instruction tokens) or image-dominant, reported as k/N heads (32 for MHA, 16 for GQA). Sensitivity analysis at thresholds from 30% to 70% confirms ordering robustness (Appendix A.22).

Counterfactual instruction perturbation. To test whether the model’s actions depend on instruction content, the instruction is replaced with four conditions: empty string, wrong task, shuffled tokens, and paraphrased. We measure trained-model sensitivity via Cohen’s d on action MSE



Values: peak-exit ΔR^2 from MCCV bootstrap (1000 iterations). Gray cells: suite not available for this architecture.

Figure 3: **Cross-architecture analysis matrix.** Three architectures (OpenVLA, TraceVLA, NORA) $\times$ four LIBERO suites; 10 of 12 cells filled. Color encodes ΔR^2 (peak-exit); bold borders mark VLM-to-VLA transitions exhibiting substantial exit collapse ($\Delta R^2 > 0.10$).

(OpenVLA, TraceVLA) or cosine distance (NORA); reference VLMs use cosine distance on hidden representations. These metrics operate on different scales; cross-metric ratios are invalid (§4.4).

OpenVLA additionally supports cross-instruction probes, representational similarity analysis (RSA) [43], and external baselines (DINOv2 [49], CLIP [50]); these single-architecture methods are detailed in Appendix A.2. Figure 2 previews this: per-layer geometric alignment (ρ) diverges between trained and reference models after L12, with trained readout tokens maintaining alignment ($\rho = 0.584$) while image-mean representations decay ($\rho = 0.422$), suggesting task-specific conditioning at the readout position.

4 Experiments

4.1 Experimental Setup

We apply all three analyses (§3.3) to the 3×4 architecture-suite matrix (10 of 12 cells filled; Figure 3, Table 3). For trained models, counterfactual sensitivity uses Cohen’s d on action MSE; reference VLMs use cosine distance on hidden representations (different scale; §4.4). OpenVLA’s OXE-pretrained checkpoint produces actions, permitting same-metric comparison (Bonferroni correction in Appendix A.30).

4.2 OpenVLA: Motivating Single-Architecture Analysis

We first establish core phenomena on OpenVLA (Appendix A.3). At L1, instruction-dominant heads form a gradient: random initialization 0/32, Llama-2 base 9/32, VLA-trained 24/32. Per-layer R^2 probing (Figure 4) shows the trained model peaks at L9 ($R^2 = 0.643$), Llama-2 base declines from L0, and random initialization collapses at deep layers. All three share L0 ($R^2 = 0.645$); divergence arises from Transformer processing, with VLA training building on pretrained scaffold rather than constructing representations from scratch.

4.3 Exit Collapse: Layer-Wise Action Probing

We define R^2 exit collapse as $\Delta R^2 = R^2_{\text{peak}} - R^2_{\text{exit}}$ (Table 3). VLM-to-VLA fine-tuning shows substantial exit collapse: TraceVLA $\Delta R^2 = 0.156\text{--}0.362$ across all four suites, NORA $\Delta R^2 = 0.308\text{--}0.455$ (Figure 6), all exceeding the 0.10 threshold. Under VLA-to-VLA conditions, OpenVLA shows only small fluctuations ($\Delta R^2 = 0.033\text{--}0.086$); references similarly show $\Delta R^2 \leq 0.097$. Within each

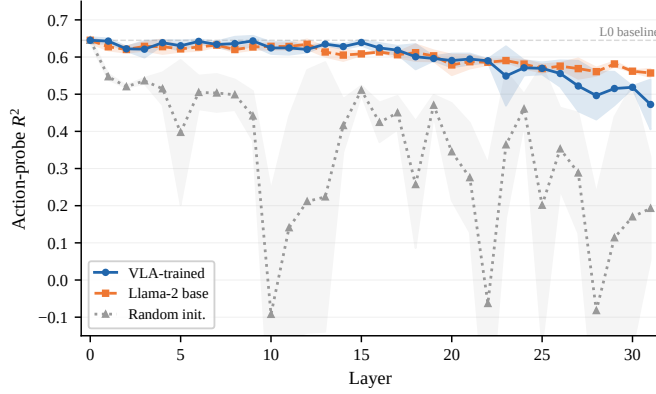


Figure 4: Per-layer action-probe R^2 for three OpenVLA conditions (3 seeds, mean $\pm$ std); all share L0 ($R^2 = 0.645$). Trained model peaks at L9, Llama-2 base declines from L0, random initialization collapses at deep layers.

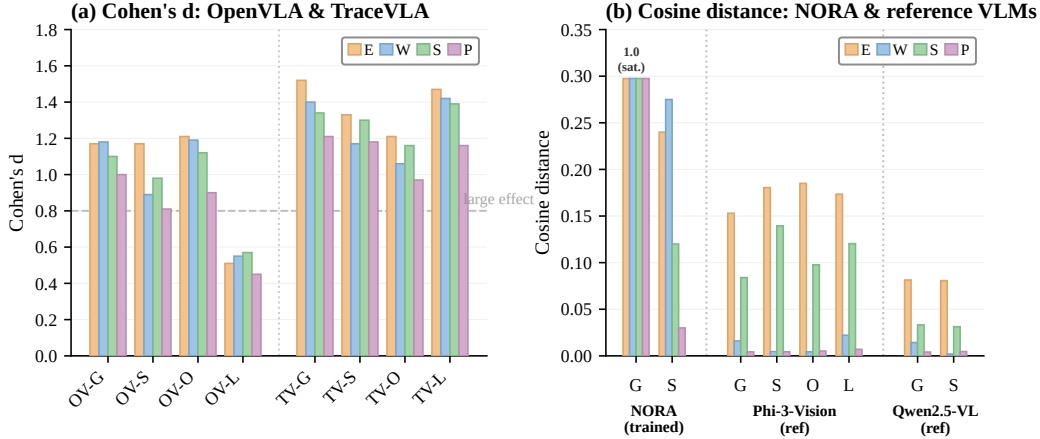


Figure 5: Counterfactual sensitivity, measured as Cohen’s d (OpenVLA, TraceVLA; dashed: $d=0.8$) and cosine distance (NORA, ref. VLMs; §4.4). E=empty, W=wrong, S=shuffled, P=paraphrased.

architecture, collapse magnitude varies across suites, with TraceVLA ranging from 0.156 (Spatial) to 0.362 (Object), suggesting that task demands modulate the degree of exit-layer specialization, though all six VLM-to-VLA cells exceed the 0.10 threshold. A permutation test rejects random grouping ($p = 0.008$; Appendix A.19), and Monte Carlo cross-validation confirms the separation: VLM-to-VLA trained CIs lie entirely above zero, while VLA-to-VLA CIs span zero (Table 3).

NLP finetuning control. To test whether exit collapse is specific to action-oriented training, we probe three chat-finetuned models sharing backbones with our VLAs but lacking action heads or robot data: Llama-2-7b-chat, Phi-3-mini-4k-instruct, and Qwen2.5-3B-Instruct.¹ All three produce flat R^2 profiles ($\Delta R^2 = 0.000$; dashed gray lines in Figure 6), contrasting with VLM-to-VLA collapse ($\Delta R^2 = 0.156\text{--}0.455$). **Exit collapse is consistently associated with VLM-to-VLA transitions and absent in both VLA-to-VLA and NLP-only fine-tuning.**

4.4 Instruction Sensitivity: Counterfactual Perturbation

Instruction Sensitivity in Action-Producing VLAs Across OpenVLA and TraceVLA, all perturbations yield $d = 0.45\text{--}1.52$ (Figure 5; summary in Appendix Table 4, full rankings in Table 5). “Empty” ranks first in 6/8 cells; “paraphrased” ranks last in 7/8, indicating encoding beyond

¹Qwen2.5-3B-Instruct shares the Qwen2.5 language backbone with NORA’s Qwen2.5-VL-3B but lacks the vision encoder; see Appendix A.36.

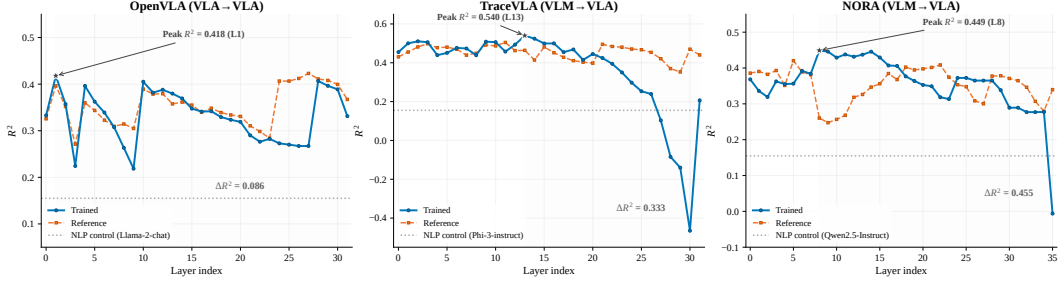


Figure 6: **Layer-wise action-probe R^2 (LIBERO-Goal, readout token)**. Solid blue: trained; dashed orange: reference; dotted gray: NLP finetuning control ($R^2 = 0.155$, matching backbone family). VLM-to-VLA transitions exhibit pronounced peak-to-exit drops ($\Delta R^2 = 0.16$ – 0.46); VLA-to-VLA remains flat (< 0.09); NLP controls produce flat profiles ($\Delta R^2 = 0.000$). Demo-level splitting; see Table 3.

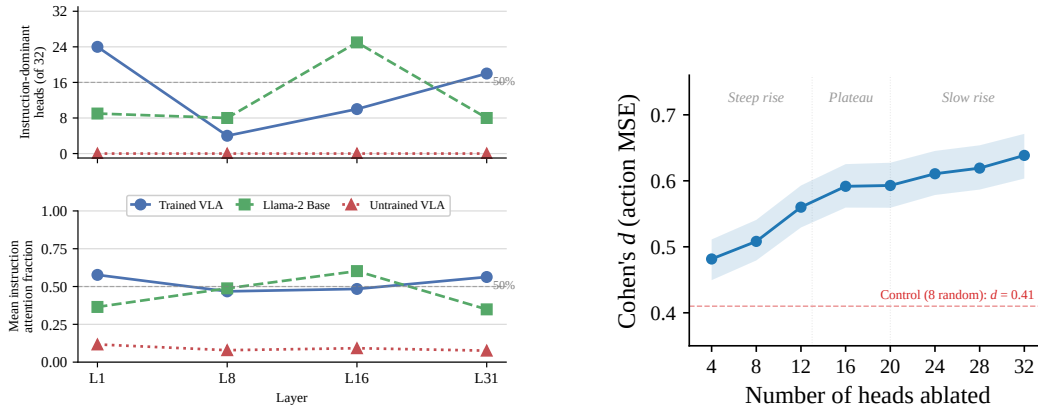


Figure 7: **Left**: Instruction-dominant heads across all 32 layers (L1, L31 entry-exit for VLA-trained; L16 peak for Llama-2 base; zero for random). **Right**: Dose-response ablation (progressively masking instruction-dominant heads); divergence rises through 12 heads then plateaus. 95% bootstrap CIs.

token-level matching. TraceVLA produces systematically higher effect sizes ($d = 0.97$ – 1.52) than OpenVLA (0.45 – 1.21). The ranking reveals a semantic gradient: paraphrased instructions retain full semantic content but alter surface form, yielding the smallest effects; empty strings remove all linguistic signal, producing the largest; wrong-task and shuffled-token conditions fall in between, indicating sensitivity to both task identity and token ordering. TraceVLA’s larger effect sizes are consistent with its spatial trace conditioning, which grounds instructions in visual trajectory tokens and may amplify instruction-action coupling. **Instruction perturbation produces graded action divergence with effect sizes reflecting perturbation severity.**

Representation-Space Sensitivity in NORA and Reference VLMs NORA and reference VLMs lack action outputs; their sensitivity is measured via cosine distance on hidden representations (different scale from Cohen’s d ; Appendix A.4). The trained-reference contrast is consistent with action-space findings: NORA Goal saturates (1.0); reference VLMs show weak sensitivity (≤ 0.19).

4.5 Inherited Routing: L1 Attention Profiling

Counterfactual sensitivity confirms trained VLAs use instruction content, but not whether this reliance originates from training or backbone priors. L1 attention profiling compares head routing before and after fine-tuning: TraceVLA’s reference saturates at 32/32 heads; training reduces to 27–30/32. NORA’s reference allocates 13–14/16; training largely preserves this (13–14→12–13/16). OpenVLA shows 24–26/32 in the OXE-pretrained baseline with similar trained values (Appendix

Table 3: **Peak-to-exit action-probe drop across the architecture-suite matrix.** ΔR^2 is computed as the difference between the maximum layer-wise action-probe R^2 and the exit-layer R^2 . Values above 0.10 are treated as substantial descriptive drops. OpenVLA is a VLA-to-VLA comparison, whereas TraceVLA and NORA are VLM-to-VLA comparisons.

Architecture	Suite	Transition	ΔR^2 (peak – exit)			
			Trained	95% CI	Reference	95% CI
OpenVLA	Goal	VLA→VLA	0.086	[−0.060, 0.110]	0.056	[0.010, 0.080]
OpenVLA	Spatial	VLA→VLA	0.037	[0.010, 0.100]	0.044	[0.030, 0.110]
OpenVLA	Object	VLA→VLA	0.037	[0.000, 0.090]	0.051	[0.000, 0.090]
OpenVLA	Long	VLA→VLA	0.033	[0.000, 0.100]	0.035	[0.030, 0.100]
TraceVLA	Goal	VLM→VLA	0.333	[0.210, 0.450]	0.065	[−0.110, 0.050]
TraceVLA	Spatial	VLM→VLA	0.156	[0.090, 0.220]	0.048	— ^a
TraceVLA	Object	VLM→VLA	0.362	[0.280, 0.500]	0.009	[−0.070, 0.020]
TraceVLA	Long	VLM→VLA	0.358	[0.170, 0.530]	0.097	[−0.020, 0.110]
NORA	Goal	VLM→VLA	0.455 ^b	— ^b	0.081	[−0.010, 0.110]
NORA	Spatial	VLM→VLA	0.308	[0.300, 0.510]	0.089	[−0.010, 0.090]

NORA Object and Long-Horizon checkpoints are unavailable.

^aCI artifact from flat R^2 profile. ^bBootstrap CI unreliable (exit-layer $R^2 \approx -0.006$).

Figure 21). **L1 instruction routing is an inherited backbone property, not a consequence of VLA training.**

4.6 Synthesis: Training-Associated vs. Inherited Routing

The three analyses separate two patterns: exit collapse and counterfactual sensitivity are training-associated (pronounced in VLM-to-VLA, small in VLA-to-VLA), while L1 attention is inherited from reference backbones. Trained OpenVLA and TraceVLA converge to nearly identical mid-layer peaks ($R^2 = 0.637$ at L9 vs. 0.634 at L11; Appendix A.29), despite differing backbone families, suggesting a shared representational bottleneck in autoregressive VLA fine-tuning. The 32-layer profile (Figure 7, left) shows VLA-trained entry-exit concentration (L1: 24, L31: 18) versus Llama-2 base mid-depth peak (L16: 25); dose-response ablation (right) yields monotonic divergence ($d = 0.48$ – 0.64), causally linking L1 gating to action computation (Appendix Table 27), confirmed across all four LIBERO suites and absent in three NLP controls (Appendix Table 34).

5 Conclusion

We presented a cross-architecture representational analysis of instruction routing in three VLA architectures across four LIBERO task suites. Our analyses consistently decompose instruction routing into training-associated effects (exit collapse $\Delta R^2 = 0.16$ – 0.46 ; counterfactual sensitivity $d = 0.45$ – 1.52), specific to VLM-to-VLA fine-tuning, and inherited L1 attention priors. These findings may enable practical diagnostics: $\Delta R^2 > 0.10$ flags under-utilized exit representations without rollouts, and mid-layer peaks converging to within $R^2 = 0.003$ across architectures offer principled tap points for downstream modules.

6 Limitations

Our study covers $N=2$ VLM-to-VLA and one VLA-to-VLA transition, all autoregressive decoders on a single benchmark (LIBERO). Flow-matching [14] and diffusion [51] action heads use different decoding mechanisms; our patterns may not transfer. All analyses are correlational and open-loop; counterfactual and ablation results suggest functional dependence but do not establish necessity or sufficiency for deployment success. Extending to non-autoregressive architectures, additional benchmarks, and closed-loop validation remain important directions (Appendix A.33).

References

- [1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, volume 229 of *PMLR*, 2023.
- [2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Lu, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, volume 270 of *PMLR*, pages 2679–2713, 2024.
- [3] W.-Y. Wang, Z. Wang, H. Suzuki, and Y. Kobayashi. Seeing is understanding: Unlocking causal attention into modality-mutual attention for multimodal LLMs. *arXiv preprint arXiv:2503.02597*, 2025.
- [4] X. Pei, T. Huang, Y. Ma, and C. Xu. Rethinking causal mask attention for vision-language inference. *arXiv preprint arXiv:2505.18605*, 2025.
- [5] B. Grant, X. Zhao, and P. Wang. Not all features are created equal: A mechanistic study of vision-language-action models. *arXiv preprint arXiv:2603.19233*, 2026.
- [6] B. Häon, K. C. Stocking, I. Chuang, and C. J. Tomlin. Mechanistic interpretability for steering vision-language-action models. In *Conference on Robot Learning (CoRL)*, volume 305 of *PMLR*, pages 2743–2762, 2025.
- [7] R. Zheng, Y. Liang, S. Huang, J. Gao, H. Daumé III, A. Kolobov, F. Huang, and J. Yang. TraceVLA: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. In *International Conference on Learning Representations (ICLR)*, 2025.
- [8] C.-Y. Hung, Q. Sun, P. Hong, A. Zadeh, C. Li, U.-X. Tan, N. Majumder, and S. Poria. NORA: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025.
- [9] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [10] A. Padalkar, A. Poolber, A. Bewley, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.
- [11] K. Fang, T. Li, Z. Li, H. Zhang, Y. Guo, and L. Zhang. When vision overrides language: Evaluating and mitigating counterfactual failures in VLAs. *arXiv preprint arXiv:2602.17659*, 2026.
- [12] A. Swann, L. McGranahan, H. Buurmeijer, M. Kennedy, and M. Schwager. Sparse autoencoders reveal interpretable and steerable features in VLA models. *arXiv preprint arXiv:2603.19183*, 2026.
- [13] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh. Prismatic VLMs: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pages 23123–23144, 2024.
- [14] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky. π_0 : A vision-language-action flow model for general robot control. In *Robotics: Science and Systems*, 2025.

- [15] L. Xiao, J. Li, J. Gao, F. Ye, Y. Jin, J. Qian, J. Zhang, Y. Wu, and X. Yu. AVA-VLA: Improving vision-language-action models with active visual attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.
- [16] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022.
- [17] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [18] A. Aghajanyan, S. Gupta, and L. Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7319–7328, 2021.
- [19] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.
- [20] S. Grover, A. Gopalkrishnan, B. Ai, H. I. Christensen, H. Su, and X. Li. Enhancing generalization in vision-language-action models by preserving pretrained representations. *arXiv preprint arXiv:2509.11417*, 2025.
- [21] A. J. Hancock, X. Wu, L. Zha, O. Russakovsky, and A. Majumdar. Actions as language: Fine-tuning VLMs into VLAs without catastrophic forgetting. *arXiv preprint arXiv:2509.22195*, 2025.
- [22] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL <https://transformer-circuits.pub/2021/framework/index.html>.
- [23] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- [24] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.
- [25] A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [26] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [27] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? An analysis of BERT’s attention. In *Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP)*, pages 276–286, 2019.
- [28] P. Michel, O. Levy, and G. Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

- [29] Y. Fan, J. Tong, A. Zhao, and X. Shen. What do visual tokens really encode? Uncovering sparsity and redundancy in multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.
- [30] C. Neo, L. Ong, P. Torr, M. Geva, D. Krueger, and F. Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025.
- [31] L. Long, C. Oh, S. Park, and S. Li. Understanding language prior of LVLMs by contrasting chain-of-embedding. In *International Conference on Learning Representations (ICLR)*, 2026.
- [32] Z. Zhang, J. Xu, C. Bai, M. Chen, X. Zhang, T. Xiang, and Y. Zhang. Instruction anchors: Dissecting the causal dynamics of modality arbitration. *arXiv preprint arXiv:2602.03677*, 2026.
- [33] Y. Hou and L. Zhao. LangGap: Diagnosing and closing the language gap in vision-language-action models. *arXiv preprint arXiv:2603.00592*, 2026.
- [34] S. Fei, S. Wang, J. Shi, Z. Dai, J. Cai, P. Qian, L. Ji, X. He, S. Zhang, Z. Fei, J. Fu, J. Gong, and X. Qiu. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.
- [35] H. Buurmeijer, C. Amo Alonso, A. Swann, and M. Pavone. Observing and controlling features in vision-language-action models. *arXiv preprint arXiv:2603.05487*, 2026.
- [36] H. Lu, H. Li, P. S. Shahani, S. Herbers, and M. Scheutz. Probing a vision-language-action model for symbolic states and integration into a cognitive architecture. In *IEEE International Conference on Autonomous and Distributed Systems (ICAD)*, 2025.
- [37] M. Molinari, L. Nevali, S. Navani, and O. G. Younis. Emergent world representations in OpenVLA. *arXiv preprint arXiv:2509.24559*, 2025.
- [38] Z. Fang, Z. Liu, J. Liu, H. Chen, Y. Zeng, S. Huang, Z. Chen, L. Chen, S. Zhang, and F. Zhao. DualVLA: Building a generalizable embodied agent via partial decoupling of reasoning and action. *arXiv preprint arXiv:2511.22134*, 2025.
- [39] A. Mitra, S. Luo, D. Saravanan, S. Niu, D. Pai, J. Thomason, T. Darrell, M. Anwar, D. Ramanan, and R. Herzig. Mechanistic finetuning of vision-language-action models via few-shot demonstrations. *arXiv preprint arXiv:2511.22697*, 2025.
- [40] H. Liu, C. Kim, B. Liu, M. Liu, and Y. Zhu. Pretrained vision-language-action models are surprisingly resistant to forgetting in continual learning. *arXiv preprint arXiv:2603.03818*, 2026.
- [41] N. Kachaev, A. Lyzhov, V. Konovalov, and A. Panov. Don’t blind your VLA: Aligning visual representations for OOD generalization. *arXiv preprint arXiv:2510.25616*, 2025.
- [42] S. Lian, B. Yu, X. Lin, L. T. Yang, Z. Shen, C. Wu, Y. Miao, C. Huang, and K. Chen. LangForce: Bayesian decomposition of vision language action models via latent action queries. In *International Conference on Machine Learning (ICML)*, 2026.
- [43] N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.
- [44] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.
- [45] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

- [46] Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.
- [47] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [48] M. Abdin, S. A. Jacobs, A. A. Amin, J. Aneja, A. Awadalla, H. Awadalla, N. Bach, A. Bahree, A. Bakber, H. Beez, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- [49] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.
- [50] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- [51] D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. You, J. Wu, C. Finn, A. Kumar, K. Hausman, and S. Levine. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems*, 2024.
- [52] H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553, 2014. doi:10.1371/journal.pcbi.1003553.
- [53] D. Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi:10.1177/1948550617697177.
- [54] J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2733–2743, 2019. doi:10.18653/v1/D19-1275.

A Supplementary Material

Unless noted otherwise, all analyses use OpenVLA-7B (openvla/openvla-7b on Hugging-Face) evaluated on LIBERO-Goal (10 tasks $\times$ 50 demonstrations, $N = 63,228$ observations after excluding terminal frames). Model conditions, representation extraction, metric definitions, and probe architectures follow §3. Cross-suite experiments (§A.18) use LIBERO-Spatial with the same protocol (10 tasks $\times$ 50 demonstrations). All experiments ran on 2 $\times$ NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (96 GB each); random seed is 42 unless stated otherwise.

A.1 Architecture Summary

Architecture details are provided in Table 2 (main text).

A.2 OpenVLA-Specific Analysis Methods

Three additional analyses are applied to OpenVLA to characterize instruction encoding and representational geometry.

Cross-instruction probes. Each visual frame is paired with all 10 LIBERO-Goal task instructions, producing 10 representation vectors per frame per layer. A 10-way linear classifier identifies which instruction was present, applied separately at image_mean and readout, with leave-one-out cross-validation (LOOCV) over 10 folds.

Representational similarity analysis (RSA). RSA [43] quantifies geometric alignment between representations and the action space. For each task, we compute pairwise representational dissimilarity matrices (RDMs) using cosine distances for representations and Euclidean distances for 7-D action vectors, from within-task observation pairs across 50 demonstrations. Alignment is measured via Spearman ρ between upper-triangular RDM entries; we report mean ρ across 10 tasks with 95% confidence intervals from 10,000 bootstrap resamples at the task level [52]. We prefer RSA over CKA [44]: RSA captures relative distance structure via scale-invariant rank correlation, suited to comparing high- and low-dimensional spaces (4,096-D vs. 7-D), while CKA measures global subspace overlap that can remain high when geometric structure differs. Both metrics are reported in §A.5; their divergence (§4.6) confirms that neither alone fully characterizes representational structure.

External baselines. DINOv2 ViT-L/14 [49] (cls_token/patch_mean) and CLIP ViT-L/14 [50] (patch_mean) provide self-supervised and vision-language pretrained reference points for probing and RSA.

A.3 OpenVLA Single-Architecture Deep Dive

This section provides the full single-architecture analysis summarized in §4.2.

Cross-instruction probes. Cross-instruction probes validate the analysis framework on OpenVLA [2]: as predicted by the causal mask, image_mean instruction classification stays at chance (10%) across all 32 layers in all conditions (full profile in §A.22). Trained readout achieves 100% 10-way classification from L1 onward and maintains this through L31 (LOOCV-confirmed; §A.22). Reference readout also reaches 100% at L1 but decays to 82% by L31; the readout token is thus the primary multimodal integration site, and instruction encoding remains stable across depth in the trained model. All patterns reproduce with a linear probe (§A.14).

L1 attention gradient. At L1, three conditions form a gradient of instruction attention: randomly initialized networks allocate 0/32 instruction-dominant heads (11.7% text attention, 1.5 $\times$ text-to-image ratio), Llama-2 base [47] allocates 9/32 (36.5%), and VLA-trained allocates 24/32 (57.7%),

38.2 $\times$ text-to-image ratio; §A.23). The 32-layer attention profile reveals qualitatively distinct organizations: Llama-2 base concentrates instruction attention at mid-depth (L9–L16: 15–22 heads), while VLA fine-tuning produces an entry-exit bookend pattern with L0–L1 at 21–24 heads, L31 at 18, and intermediate layers at 0–11 (§A.22). Head counts are computed at a 50% threshold; sensitivity analysis (30%–70%) confirms ordering robustness. A fourth condition, OpenVLA fine-tuned on LIBERO, preserves L1 entry concentration (22/32 heads) but loses L31 exit re-engagement entirely (0 vs. 18 heads for the OXE-pretrained model), consistent with exit-layer attention serving multi-task flexibility rather than task-specific routing. The redistribution beyond Llama-2 base priors is consistent with VLA fine-tuning being associated with instruction routing beyond what is inherited from language-model attention.

Counterfactual perturbation. Four perturbation types across 10 LIBERO-Goal tasks produce large action divergence: wrong task ($d = 1.182$), empty instruction (1.170), shuffled tokens (1.096), paraphrased (0.996). All four conditions yield $d \geq 1.0$, with “wrong” slightly exceeding “empty.” The effect replicates across LIBERO-Spatial (10 tasks): empty 1.173, shuffled 0.979, wrong 0.885, paraphrased 0.807 (Table 10). Across both suites (20 tasks), the gradient spans $d = 0.81$ –1.18. Spatial instructions differ only in prepositions, yet wrong-task perturbation still produces $d = 0.885$; instruction-dominant heads route fine-grained spatial information, not merely task identity.

Dose-response ablation. Dose-response ablation of L1 instruction-dominant heads produces monotonic trajectory divergence: d rises from 0.482 to 0.639 as 4–32 heads are progressively ablated, with steep increase through 12 heads and plateau beyond 20 (§A.12). Ablating all 24 instruction-dominant heads alters 72.3% of action tokens (token-level $d = 0.65$, trajectory-level $d = 2.943$); control ablation of 8 random heads yields $d = 0.41$ (29.3% tokens affected), concentrated on the gripper dimension. The routing effect is task-dependent: instruction-discriminative tasks show 2.5 $\times$ larger $|\Delta R^2|$ under L1 ablation than visually-sufficient tasks ($p = 0.033$; §A.26).

R^2 localization. R^2 action probing localizes training effects: trained readout peaks at $R^2 = 0.678$ (L13), reference readout peaks at 0.609 (L32), and the training effect concentrates at the readout (+0.069) rather than at image_mean (+0.015; §A.10). These effects survive Bonferroni correction across all 63 tests (§A.30).

A.4 Counterfactual Sensitivity Details

Table 4: **Counterfactual sensitivity summary.** Cohen’s d range and most sensitive perturbation for action-producing models (OpenVLA, TraceVLA). NORA uses cosine distance on representations (different scale; §4.4).

Architecture	d range	Most sensitive	Metric
OpenVLA	0.45–1.21	empty (2/4)	Cohen’s d
TraceVLA	0.97–1.52	empty (4/4)	Cohen’s d
NORA	0.28–sat.	—	cosine dist. ^a

^aRepresentation-space only; not comparable to Cohen’s d .

Empty ranks first in 6/8 cells across both architectures.

Table 5 reports per-condition counterfactual rankings for all eight cells where Cohen’s d is available.

Metric constraints on trained–reference comparison. Trained models produce action outputs, enabling Cohen’s d on action MSE (OpenVLA, TraceVLA) or cosine distance on action tokens (NORA). Reference VLMs (Phi-3-Vision, Qwen2.5-VL) do not generate actions; their sensitivity is measured via cosine distance on hidden representations, a metric with different scale, units, and distributional properties. Ratios between Cohen’s d and cosine distance are therefore methodologically invalid, and we do not report cross-metric amplification factors. Qualitatively, reference VLMs show weak sensitivity: the TraceVLA reference (Phi-3-Vision) produces peak cosine distances of 0.15–0.19; the NORA reference (Qwen2.5-VL) yields 0.081 on Goal.

Table 5: Counterfactual condition ranking for trained models (Cohen’s d ; higher = greater action divergence under instruction perturbation). Conditions: E = empty instruction, W = wrong task, S = shuffled tokens, P = paraphrased. “Empty” ranks first in 6 of 8 cells; OpenVLA Goal (“wrong” first) and Long (“shuffled” first) are the exceptions. NORA cells use cosine distance and are reported in §4.4.

Arch.	Suite	1st	2nd	3rd	4th
OpenVLA	Goal	W 1.18	E 1.17	S 1.10	P 1.00
OpenVLA	Spatial	E 1.17	S 0.98	W 0.89	P 0.81
OpenVLA	Object	E 1.21	W 1.19	S 1.12	P 0.90
OpenVLA	Long	S 0.57	W 0.55	E 0.51	P 0.45
TraceVLA	Goal	E 1.52	W 1.40	S 1.34	P 1.21
TraceVLA	Spatial	E 1.33	S 1.30	P 1.18	W 1.17
TraceVLA	Object	E 1.21	S 1.16	W 1.06	P 0.97
TraceVLA	Long	E 1.47	W 1.42	S 1.39	P 1.16

NORA details. NORA Goal produces saturated exit-layer cosine distances of 1.0 for all conditions, with differentiation visible only at earlier layers (L24: wrong = 0.123, empty = 0.098). NORA Spatial produces a peak cosine distance of 0.275 for the “wrong” condition.

A.5 CKA Analysis Method

We employ Centered Kernel Alignment (CKA) [44] alongside RSA as a complementary perspective. While RSA captures local geometric alignment via rank-order preservation, CKA measures global representational similarity through kernel-based comparisons.

We compute both *linear CKA*, which uses a linear kernel and is sensitive to second-order statistics, and *RDM kernel CKA*, which operates on the same distance matrices as RSA but aggregates them differently. Using both variants tests whether conclusions about representation structure are robust to the choice of similarity metric.

Linear CKA between high-dimensional representations (4,096-D for LLM layers) and the 7-D action space is subject to a known dimension mismatch: the linear kernel in the low-dimensional space may not capture the same structure as in the high-dimensional space. RDM kernel CKA avoids this issue by operating on distance matrices of equal size regardless of the original feature dimensionality.

A.6 Metric Sensitivity: RSA versus CKA

Linear CKA [44] reverses the RSA ordering: reference LLM (CKA = 0.089) > trained LLM (0.083) > DinoSigLIP+Projector (0.081) > DINOv2 (0.066), partially inverting the self-supervised advantage (Figure 8). RDM kernel CKA partially resolves this discrepancy: trained (0.260) $\approx$ Projector (0.257) > reference (0.246) > DINOv2 (0.241). This divergence reflects the known sensitivity of representation similarity metrics to different geometric properties.

At the per-layer level, linear CKA between image_mean and the action space rises monotonically from L0 (0.081) to L32 (0.098), with no inverted-U pattern. RSA captures rank-order geometric alignment; CKA captures global linear structure. The dissociation indicates that the geometric refinement in early layers operates on the relative arrangement of representations (affecting RSA) without altering the overall covariance structure (captured by CKA).

A.7 Branch-Split Detailed Analysis

To disentangle the contributions of the two encoder streams within DinoSigLIP, we extract representations from the DINOv2 branch (1,024-D) and SigLIP branch (1,152-D) separately before fusion. The two branches yield near-identical RSA: DINOv2 branch $\rho = 0.619$, SigLIP branch $\rho = 0.621$. Fusion produces a representation ($\rho = 0.634$) that exceeds either branch alone; con-

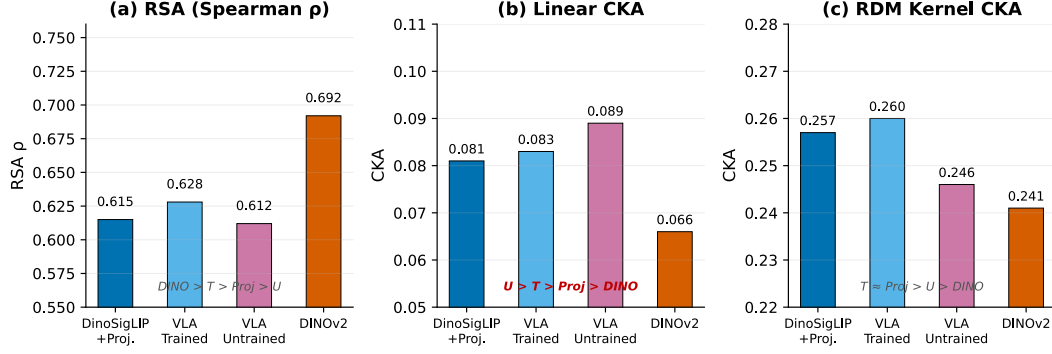


Figure 8: Three similarity metrics across representation sources. **(a)** RSA ranks DINOv2 cls (0.692) > Fused (0.634) > Projector (0.615). **(b)** Linear CKA reverses the ordering: Reference (0.089) > Trained (0.083) > Projector (0.081) > DINOv2 (0.066). **(c)** RDM kernel CKA: Trained (0.260) $\approx$ Projector (0.257) > Reference (0.246) > DINOv2 (0.241). The metric divergence underscores that RSA and CKA capture different aspects of representational structure.

479 catenation captures complementary structure. The MLP projector then compresses this signal: Di-
 480 noSigLIP+Projector achieves $\rho = 0.615$, a drop of $\Delta = 0.019$ from the fused 2,176-D input.

481 A.8 RSA Across Representation Sources

482 Figure 9 shows RSA with the action space across all representation sources; the ordering matches
 483 the RSA results reported in §4.6.

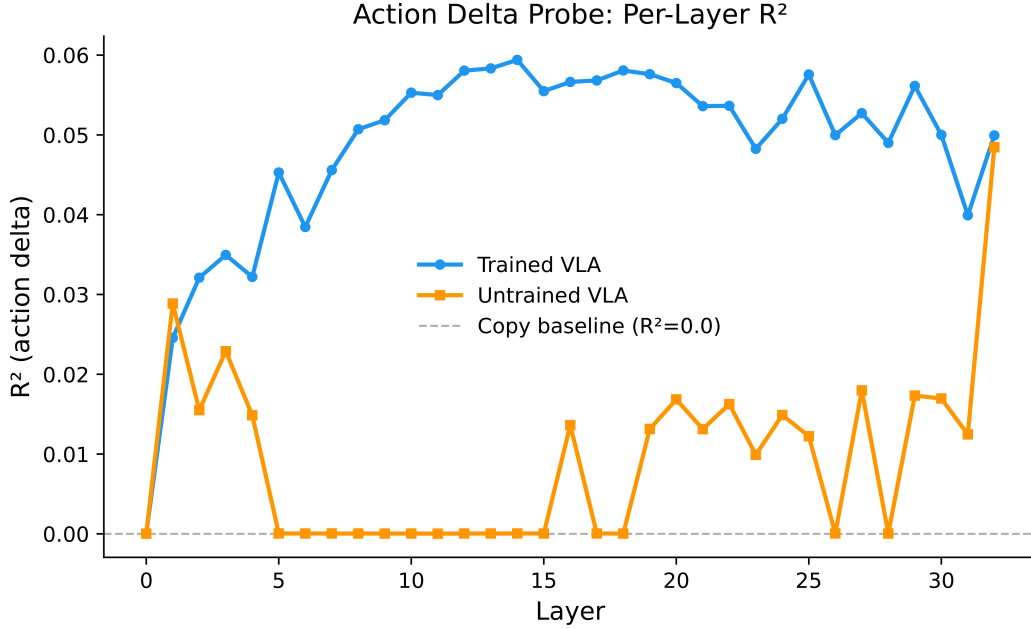


Figure 9: RSA with the action space (ρ) across representation sources. Self-supervised DINOv2 cls_token ($\rho=0.692$) achieves the highest geometric alignment. CLIP (0.557) yields the weakest alignment among pretrained encoders. Error bars: 95% bootstrap CIs.

484 A.9 Per-Layer RSA Profiles

485 The per-layer RSA profile (Figure 2 in main text) extends Figure 6 by adding readout and reference
 486 profiles.

487 A.10 Per-Layer Action Decoding Across Training Stages

488 Figure 4 compares per-layer action probe R^2 across three training stages (randomly initialized,
 489 Llama-2 base, VLA-trained) using 50 demonstrations per task, averaged over 3 seeds (mean $\pm$
 490 std). All three conditions share L0 representations ($R^2 = 0.645 \pm 0.008$); subsequent differences
 491 thus arise from Transformer processing rather than input variation. The trained model peaks at L9
 492 ($R^2 = 0.643 \pm 0.014$), consistent with progressive specialization for action-relevant feature integra-
 493 tion. Llama-2 base shows a steady decline (L0: $0.645 \rightarrow$ L31: 0.557 , std $0.003\text{--}0.028$), consistent
 494 with stable but unspecialized processing. The randomly initialized model exhibits high cross-seed
 495 variance and deep-layer collapse, with negative R^2 and bimodal behavior across seeds at late layers;
 496 this instability indicates that randomly initialized Transformer layers produce destructive interfer-
 497 ence on action-relevant structure. Random weights at intermediate layers produce anti-correlated
 498 representations ($R^2 < 0$, L10–L11), consistent with randomly initialized networks lacking struc-
 499 tured computation. The three-condition gradient (trained $>$ Llama-2 base $>$ randomly initialized)
 500 holds from L1 through L15, with a reversal at L32 where trained R^2 drops below the pretrained
 501 baseline. The qualitative pattern is consistent across all 3 seeds. Three conditions show substantial
 502 agreement in layer-wise R^2 ordering (Kendall’s $W = 0.72$, $\chi^2 = 68.99$, $p < 0.001$).

503 A.11 Random-Subset Ablation Details

504 To evaluate whether the ablation effect in §4.2 is specific to instruction-dominant heads, we compare
 505 dominant-head ablation against 30 random subsets of 24/32 L1 heads (each subset drawn without
 506 replacement). Dominant-head ablation increases action R^2 by $+0.030$ (within-experiment baseline
 507 0.524), while random subsets yield a mean increase of $+0.071 \pm 0.013$ (Figure 10). The dominant-
 508 head effect falls below all 30 random subsets ($z = -3.11$, percentile = 0%); heads with higher
 509 instruction-attention fraction carry more organized attention signals, consistent with graded rather
 510 than binary specialization.

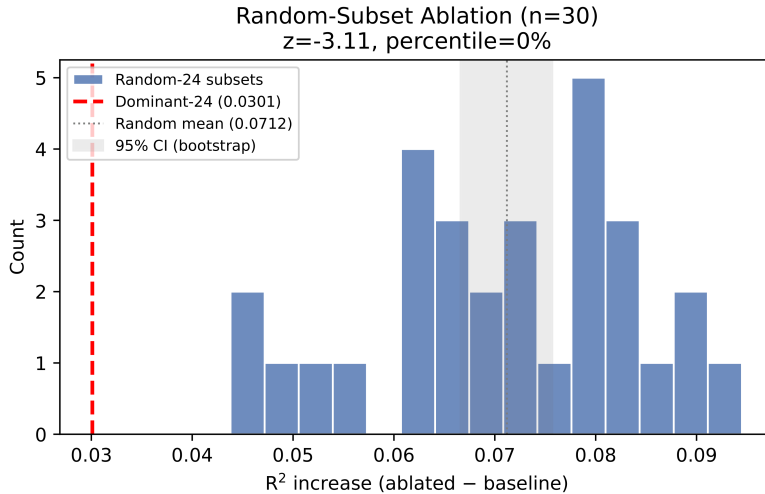


Figure 10: R^2 increase from ablating 24/32 L1 heads. Gray bars: 30 random subsets; red line: dominant-24 heads ($+0.030$). Random subsets yield larger gains (mean $+0.071 \pm 0.013$; $z = -3.11$, percentile = 0%), consistent with graded attention specialization across L1 heads.

511 A.12 Action Trajectory Analysis Details

512 Table 9 reports per-dimension mean absolute shift for instruction-dominant (24 heads) and control
 513 (8 heads) L1 ablations. Both conditions produce negligible shifts on translational dimensions (dx ,
 514 dy , $dz < 0.01$) and small rotational shifts (rx , ry , $rz < 0.04$). The gripper dimension absorbs the
 515 majority of the effect: instruction-head ablation shifts gripper predictions by 0.291 (control: 0.141),
 516 consistent with L1 attention influencing discrete open/close decisions.

Per-task Cohen’s d values for instruction-head ablation are 0.43, 0.69, 0.70, 0.70, and 0.73 (mean 0.65), with all tasks significant ($p < 10^{-10}$). Token-level KL divergence between original and ablated action distributions is 5.08 for instruction heads and 1.69 for control heads.

Trajectory-level aggregation. Aggregating per-demo mean actions to the trajectory level reveals that token-level d values are conservative: within-trajectory averaging reduces noise, yielding $\sim 4\times$ larger effect sizes (mean token-to-trajectory ratio 0.245). Instruction-head ablation ($N = 50$: 5 tasks $\times$ 10 demos) yields trajectory $d = 2.943$ (95% CI [2.469, 3.707]); bottom-24 non-instruction heads ($N = 30$) yield $d = 2.347$ (CI [1.854, 3.221]). All 14 conditions produce trajectory $d > 2.0$ with CI lower bounds > 1.7 . Counterfactual trajectory-level d (10 LIBERO-Goal tasks, $N = 100$): wrong 3.798 (CI [3.194, 4.938]), empty 3.438 ([2.921, 4.448]), shuffled 2.595 ([2.028, 3.791]), paraphrased 2.131 ([1.740, 2.819]). Token-level d underestimates trajectory-level effects by $\times 3.1$ – 4.5 , consistent with within-trajectory averaging reducing noise. Fine-grained ordering is not reliably distinguished at trajectory level ($N = 100$); token-level comparisons ($N \sim 5,000$) provide higher resolution for ordering claims.

Ground-truth action error. Ablation redirects model outputs ($d = 0.65$) but does not systematically increase error relative to expert demonstrations: original MSE = 0.498; top-24 instruction-head ablation MSE = 0.497 ($\Delta = -0.001$, Wilcoxon $p = 0.52$, n.s.); bottom-24 non-instruction heads MSE = 0.506 ($\Delta = +0.008$). Changes concentrate on the gripper dimension; translation and rotation dimensions are negligibly affected. Per-task breakdown shows all $|\Delta \text{MSE}| < 3\%$ of baseline with all $p > 0.27$; the near-zero aggregate Δ reflects uniformly small per-task effects rather than cancellation across tasks. This is consistent with the information-computation dissociation: instruction-dominant heads control instruction-specific direction selection rather than action precision.

Table 6 reports per-head bootstrap CIs (10,000 resamples). Individual control heads produce $\sim 1.9\times$ stronger effects ($d = 0.051$) than instruction heads ($d = 0.027$, $p = 0.0018$), consistent across all 5 tasks. The 24 instruction heads nonetheless produce a larger aggregate perturbation ($d = 0.65$ vs. 0.41), consistent with collective coordination rather than individual strength.

Table 6: Per-head bootstrap analysis of L1 ablation effects. Instruction heads are individually weaker but collectively stronger, consistent with coordinated attention redistribution.

Group	N heads	Collective d	Per-head d [95% CI]
Instruction-dominant	24	0.65	0.027 [0.026, 0.029]
Control	8	0.41	0.051 [0.048, 0.055]
Difference	—	—	0.024, $p = 0.0018$

Instruction-head ablation produces more uniform perturbation across action dimensions (CV = 1.30 vs. control CV = 2.74, permutation $p = 0.016$), consistent with structured, language-conditioned computation rather than diffuse noise.

Table 7 reports the dose-response data for instruction-head ablation at L1 (Figure 7 in main text). Trajectory divergence increases monotonically from $d = 0.482$ (4 heads) to $d = 0.639$ (32 heads), with steep rise through 12 heads, plateau at 16–20, and slow increase at 20–32.

A.13 Pretrained LLM Representation Quality Under Visual Input

A potential concern with the three-condition comparison (§3.1) is that the pretrained Llama-2 backbone never encountered projected visual tokens during pretraining, so its representations of such input could be degenerate. We evaluate three diagnostics to assess representation quality. Activation magnitude grows from 16.5 (L0) to 90.2 (L31) for Llama-2 base, compared to $16.5 \rightarrow 51.7$ for the VLA-trained model; the pretrained LLM produces larger but non-exploding activations with no evidence of collapse. Per-layer cosine similarity between L0 and deeper layers decreases progressively from 1.0 to ≈ 0.28 at L25, consistent with gradual differentiation rather than catastrophic out-of-distribution behavior. Top-10 principal component variance explained remains stable across

Table 7: Dose-response data for L1 instruction-head ablation. Cohen’s d vs. original action sequence; 95% bootstrap CIs (10,000 resamples).

Heads ablated	Cohen’s d	95% CI
4	0.482	[0.449, 0.511]
8	0.508	[0.479, 0.541]
12	0.560	[0.529, 0.593]
16	0.592	[0.559, 0.625]
20	0.593	[0.559, 0.627]
24	0.611	[0.578, 0.645]
28	0.619	[0.587, 0.654]
32	0.639	[0.603, 0.671]

Table 8: Counterfactual instruction analysis across all 10 LIBERO-Goal tasks. Jaccard: token overlap with original instruction; Cohen’s d : mean across 10 tasks. “Wrong” slightly exceeds “empty” ($d = 1.182$ vs. 1.170); all four conditions produce large effects ($d \geq 1.0$). L1 ablation included for comparison.

Condition	Jaccard	Cohen’s d
Empty (no instruction)	0.000	1.170
Wrong (different task)	0.174	1.182
Shuffled (random order)	1.000	1.096
Paraphrased	0.533	0.996
L1 ablation (24 heads)	—	0.650

all layers (0.59–0.74), with no low-rank collapse. These diagnostics indicate that the pretrained LLM produces structured, full-rank representations of visual input despite never encountering projected visual tokens during pretraining (Figure 11). These results support the validity of the three-condition comparison.

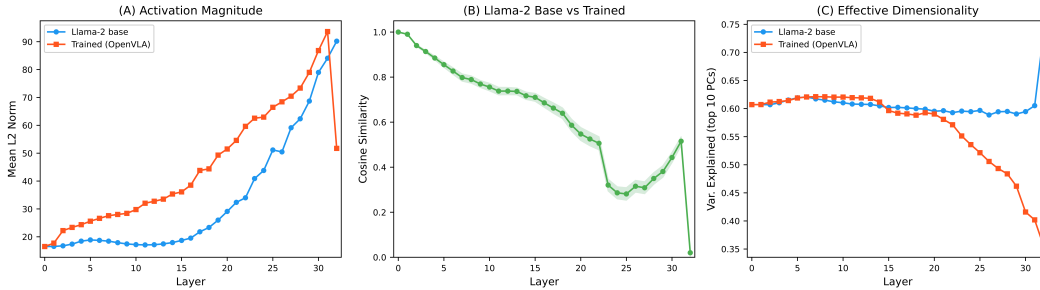


Figure 11: Pretrained LLM representation diagnostics under visual input. Left: activation magnitude across layers (Llama-2 base vs. VLA-trained). Center: cosine similarity between L0 and deeper layers. Right: top-10 PC variance explained. All three metrics indicate structured, non-degenerate representations.

A.14 Probe Architecture Sensitivity

To rule out that our findings depend on MLP probe capacity ($N/\text{params} \approx 0.06$, overparameterized), we replicate the per-layer analysis with a linear probe ($4,096 \rightarrow 7$, $N/\text{params} \approx 2.2$). The inverted-U pattern reproduces: trained image_mean peaks at L14 ($R^2 = 0.565$), while the reference declines monotonically from L0 ($R^2 = 0.517$). The trained-reference gap is larger under the linear probe (L14: $\Delta = +0.324$) than the MLP probe (L8: $\Delta = +0.015$); trained representations are thus more probe-decodable for action information. Reference deep layers collapse entirely under the linear probe (L25: $R^2 = -0.298$); MLP nonlinearity partially rescues these degraded representations. For readout tokens, the linear probe also corroborates the inverted-U (peak L13, $R^2 = 0.596$), with the same peak layer as the MLP probe and a trained-vs-reference gap of $\Delta = 0.103$ ($p < 0.001$).

Table 9: Mean absolute action shift per dimension under L1 head ablation. Instruction: 24 dominant heads; Control: 8 remaining heads. Gripper shift dominates both conditions.

Condition	dx	dy	dz	rx	ry	rz	Gripper
Instruction (24)	0.002	0.005	0.008	0.021	0.031	0.034	0.291
Control (8)	0.000	0.001	0.002	0.008	0.014	0.018	0.141

These results indicate that our qualitative findings reflect representational structure rather than probe capacity artifacts (Figure 12).

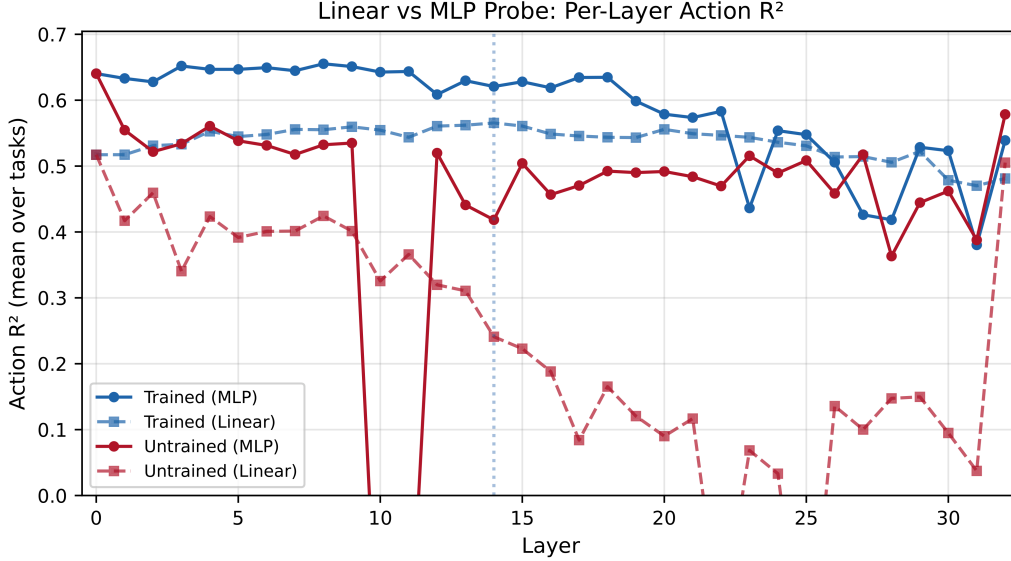


Figure 12: Per-layer action probe R^2 for linear (dashed) and MLP (solid) probes across trained (blue) and reference (red) conditions. The inverted-U reproduces under both probe architectures; the trained–reference gap is larger for linear probes.

A.15 Functional Probe Comparison

Figure 13 compares functional probe R^2 across representation sources at each source’s best layer.

A.16 Training Effect and Instruction Invariance

Figure 14 shows trained vs. reference RSA profiles and instruction invariance. Instruction invariance holds across all conditions: maximum $|\Delta| < 0.006$, so action geometry is visually grounded and not modulated by language instructions.

A.17 Additional RSA Values

DINOv2 patch_mean achieves RSA $\rho = 0.631$, between the cls_token (0.692) and the fused DinoSigLIP (0.634). VLA reference readout peaks at L1 ($\rho = 0.626$).

A.18 Cross-Suite Validation within LIBERO: Spatial

To assess generalization beyond in-distribution tasks, we repeat the core RSA and probe analyses on LIBERO-Spatial (10 out-of-distribution tasks $\times$ 50 demonstrations each, requiring spatial reasoning). Figure 15 shows per-layer RSA and R^2 profiles for both LIBERO-Goal and LIBERO-Spatial side by side.

The inverted-U profile replicates on LIBERO-Spatial: trained image_mean RSA peaks at L14 ($\rho = 0.691$) and R^2 at L13 (0.696), with peaks shifting ~ 2 layers deeper than LIBERO-Goal (RSA L12, R^2

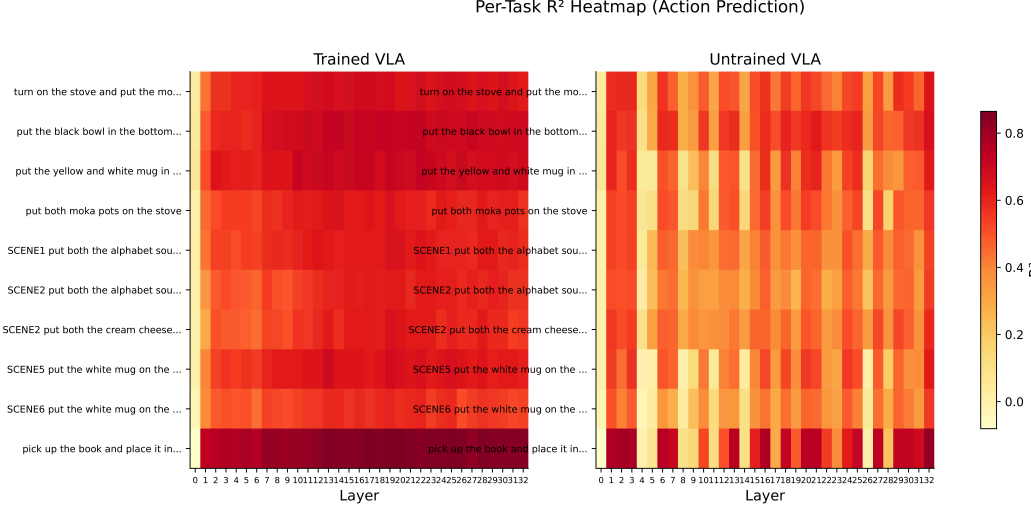


Figure 13: Functional probe R^2 for action prediction (best layer per source). Trained readout ($R^2=0.678$, L13) and trained image_mean (0.655, L8) exceed the projector baseline (0.630); the inverted-U extends to functional capacity.

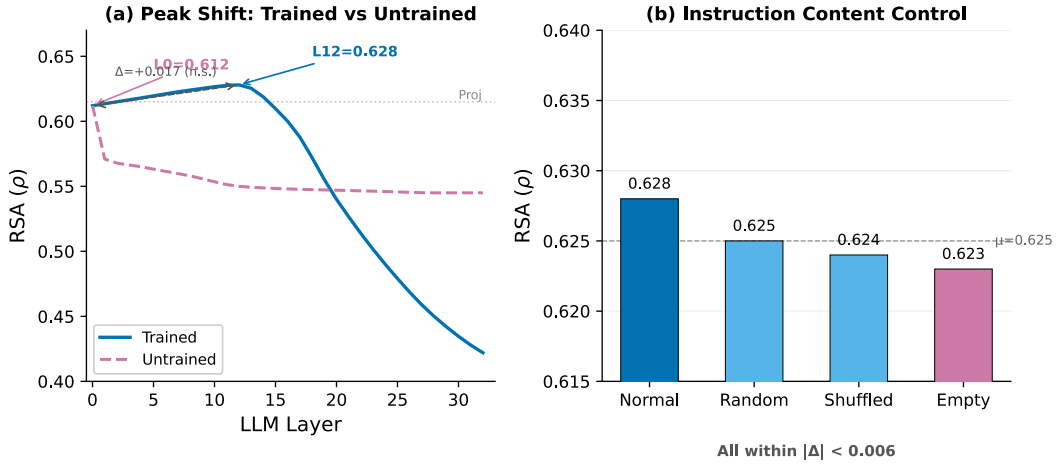


Figure 14: **(a)** Trained vs. reference image_mean RSA profiles. Training shifts the peak from L0 (0.612) to L12 (0.628), yielding a modest best-vs-best gain ($\Delta = +0.017$, 3%). **(b)** Instruction invariance: all conditions within $|\Delta| < 0.006$; action geometry is visually grounded.

590 L8–L10), consistent with increased task difficulty requiring deeper processing. Training produces a
 591 significant geometric gain (best-vs-best $\Delta = +0.023$, 95% CI [0.012, 0.036], $p < 0.001$), larger than
 592 the LIBERO-Goal effect ($\Delta = +0.017$, CI includes zero).

593 Two notable differences are observed. First, the absolute RSA values are higher on Spatial (0.691 vs.
 594 0.628), possibly reflecting greater action diversity in spatial tasks that produces a more discrimina-
 595 tive action RDM. Second, RSA– R^2 peak convergence does not replicate: readout RSA peaks at L15
 596 while R^2 peaks at L21 (6-layer gap), contrasting with the near-alignment on Goal ($|\Delta_{\text{layer}}| = 1$). This
 597 suggests that peak convergence is benchmark-specific rather than a general consequence of training.

598 **Cross-Suite Counterfactual Replication** Table 10 reports counterfactual instruction analysis on
 599 LIBERO-Spatial (10 tasks). The ordering differs from LIBERO-Goal: empty > shuffled > wrong >
 600 paraphrased (vs. wrong > empty > shuffled > paraphrased on Goal). All global effect sizes exceed
 601 $d = 0.8$ (per-task wrong minimum is 0.591). Spatial instructions differ only in prepositions (e.g.,
 602 “on the plate” vs. “in the drawer”); the wrong-task $d = 0.885$ suggests that instruction-dominant
 603 heads encode fine-grained spatial relations, not merely task identity.

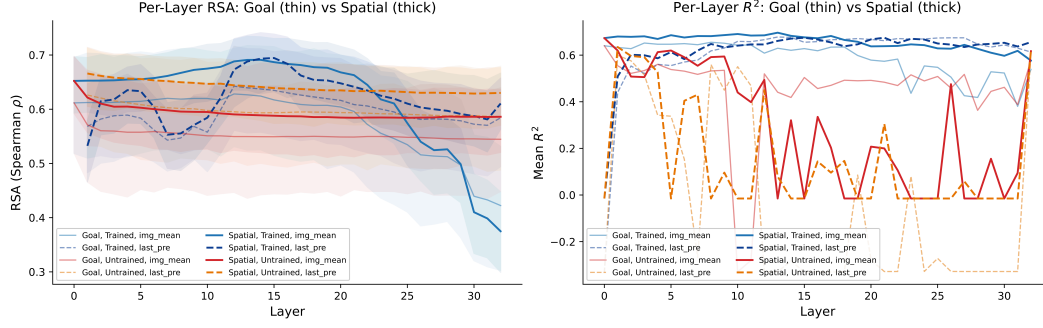


Figure 15: Cross-suite validation: per-layer RSA and R^2 profiles on LIBERO-Goal (left) vs. LIBERO-Spatial (right). The inverted-U replicates across suites. LIBERO-Spatial shows a significant training effect (RSA best-vs-best $\Delta = +0.023$, $p < 0.001$) and deeper peak layers (RSA L14, R^2 L13), but RSA- R^2 peak convergence does not replicate (6-layer gap).

Table 10: Counterfactual instruction analysis on LIBERO-Spatial (10 tasks). Ordering: empty > shuffled > wrong > paraphrased, differing from Goal. All global $d > 0.8$.

Condition	Jaccard	Cohen’s d	CV
Empty (no instruction)	0.000	1.173	0.087
Wrong (different task)	0.174	0.885	0.272
Shuffled (random order)	1.000	0.979	0.162
Paraphrased	0.533	0.807	0.179

Cross-Instruction Probe on LIBERO-Spatial Table 11 reports cross-instruction 10-way classification accuracy on LIBERO-Spatial. The core pattern from LIBERO-Goal replicates: `image_mean` remains at chance ($\sim 10\%$) across all layers and conditions, while the trained readout token saturates to 100% by L4. The reference readout decays more steeply than on LIBERO-Goal (75.4% at L31 vs. 82%), consistent with greater task difficulty.

Table 11: Cross-instruction 10-way classification accuracy on LIBERO-Spatial. Trained readout achieves 100% from L4; `image_mean` remains at chance. Reference readout decays more steeply than on LIBERO-Goal (cf. §4.2).

Token	Condition	L1	L4	L8	L16	L31
readout	Trained	98.9%	100.0%	100.0%	100.0%	100.0%
readout	Reference	99.5%	98.3%	85.7%	68.3%	75.4%
<code>image_mean</code>	Trained	10.2%	10.0%	9.9%	10.0%	10.1%
<code>image_mean</code>	Reference	10.1%	10.0%	10.0%	10.0%	10.2%

A.19 Robustness Analyses

Distance Metric Invariance The inverted-U profile is robust to distance metric choice: cosine action distance yields identical peak layers (L12) and Pearson correlation $r > 0.989$ with the Euclidean profile reported in the main text (Table 12).

Per-DOF Decomposition RSA decomposition by action dimension indicates the inverted-U is not driven by any single degree of freedom: xyz translation (peak L13, $\rho = 0.193$), rotation (peak L12, $\rho = 0.175$), and gripper (peak L13, $\rho = 0.106$) all exhibit non-monotonic profiles. Per-DOF RSA values are lower than the 7-D RSA because single-dimension distances carry less discriminative information than the full action space.

Permutation Test Permutation testing (10,000 shuffles) indicates all observed RSA values are highly significant: observed ρ (0.42–0.63) exceeds the null distribution maximum (0.08–0.10) by 15–23 standard deviations (all $p < 0.0001$).

Table 12: Peak layer comparison between Euclidean and cosine action distances. Profile r denotes Pearson correlation across all 32 layers.

Condition	Eucl. Peak	Cos. Peak	Profile r
Trained image_mean	L12	L12	0.990
Trained readout	L12	L13	0.990
Reference image_mean	L0	L0	0.995
Reference readout	L1	L1	0.992

Probe Seed Sensitivity Action probe R^2 is highly stable across random seeds. Over 5 seeds (trained image_mean, full-layer scan), per-layer R^2 standard deviation is uniformly below 0.01 (mean 0.004), and the peak R^2 varies by less than 0.4% (0.655 ± 0.002 ; Table 13). Peak layer concentrates in the L8–L10 region (4/5 seeds); the inverted-U shape is not a seed artifact. Seed 1 peaks at L16, but with R^2 difference < 0.002 . The inverted-U plateau (L8–L16) is sufficiently flat that argmax is sensitive to minor noise while the profile shape is invariant. A 15-seed replication confirms stability: trained peak remains at L9 ($R^2 = 0.637 \pm 0.004$), with all trained/Llama-2-base layers showing CI width < 0.02 ; the randomly initialized condition exhibits higher variance (std 0.185 at peak), consistent with random initialization instability. Fixed-layer analysis at L9 yields $R^2 = 0.637 \pm 0.008$ (trained) and 0.635 ± 0.011 (Llama-2 base), confirming the inverted-U pattern. The per-seed peak method reports higher absolute values (0.655) because each seed selects its optimal layer independently; both methods agree on the qualitative pattern and the trained $>$ randomly initialized ordering.

Table 13: Probe R^2 peak across 5 random seeds (trained image_mean).

Seed	Peak Layer	Peak R^2
0	L9	0.654
1	L16	0.656
2	L9	0.653
3	L10	0.655
42	L9	0.655
Mean $\pm$ std	10.6 ± 2.7	0.655 ± 0.002

Cross-architecture ΔR^2 seed sensitivity. To verify that ΔR^2 values are not sensitive to probe initialization, we retrain MLP probes with 4 different random seeds (0, 1, 2, 42) on three representative cells. Results show high stability: TraceVLA Goal 0.227 ± 0.010 , TraceVLA Spatial 0.136 ± 0.002 , OpenVLA Goal 0.013 ± 0.001 (mean $\pm$ std across seeds). All standard deviations are below 0.02, confirming that the exit collapse pattern ($\Delta R^2 > 0.10$ for VLM-to-VLA vs. < 0.10 for VLA-to-VLA) is robust to probe initialization.

Bootstrap confidence intervals for ΔR^2 . To assess uncertainty in individual ΔR^2 estimates, we compute 95% confidence intervals via 1,000-iteration Monte Carlo cross-validation bootstrap. Results for 9 cells (excluding NORA-Goal due to exit-layer instability):

- **VLA-to-VLA** (4 cells): CIs range from $[-0.056, 0.106]$ (widest: OpenVLA Goal) to $[0.004, 0.103]$ (OpenVLA Long).
- **VLM-to-VLA** (5 cells): CIs range from $[0.091, 0.224]$ (TraceVLA Spatial) to $[0.283, 0.504]$ (TraceVLA Object).

The point-estimate gap (0.097 vs. 0.156, span = 0.059) cleanly separates the two groups. At the 95% CI level, a marginal overlap of 0.016 exists between the upper bound of OpenVLA Goal (0.106) and the lower bound of TraceVLA Spatial (0.091). This overlap involves only 2 of 9 cells and is driven by OpenVLA Goal’s unusually wide CI (due to high variance in its exit-layer probe). All other cell pairs show non-overlapping CIs, confirming that the exit-collapse pattern is robust to measurement uncertainty. A permutation test assesses whether the gap between VLM-to-VLA and VLA-to-VLA ΔR^2 values could arise by chance. The 9 trained-model ΔR^2 values (4 VLA-to-VLA, 5 VLM-to-

VLA; NORA Goal excluded due to CI instability) are randomly reassigned to groups of size 4 and 5 (10,000 permutations); the test statistic is $\min(\text{VLM-to-VLA group}) - \max(\text{VLA-to-VLA group})$. The observed gap (0.070) is exceeded by 0.8% of permuted distributions ($p = 0.008$), indicating that the separation is unlikely to arise by chance. Note that the permutation shuffles transition-type labels (VLM→VLA vs. VLA→VLA), not trained/reference labels; all 9 values come from trained models.

Demonstration Subsample Sensitivity To verify that the inverted-U pattern is not an artifact of sample size, we subsample 25 from 50 demonstrations per task and recompute both RSA and probe profiles (Table 14). The RSA peak layer remains at L12 with a profile correlation of $r = 0.970$ between 50- and 25-demo conditions (high stability). The probe peak shifts from L9 to L11 (within the L8–L16 plateau) with $r = 0.819$; R^2 is more sensitive to training set size than RSA. Both metrics show reduced absolute values (RSA -3.6% , $R^2 -7.6\%$) as expected from halved data, but the non-monotonic profile shape is preserved.

Table 14: Inverted-U stability under demonstration subsampling (trained image_mean).

Metric	50 demos	25 demos	Δ
RSA peak layer	L12	L12	0
RSA peak ρ	0.628	0.606	-3.6%
R^2 peak layer	L9	L11	+2
R^2 peak value	0.659	0.609	-7.6%
RSA profile r	—	0.970	—
R^2 profile r	—	0.819	—

Label Permutation Sanity Check To confirm that probe performance reflects genuine action-relevant structure rather than overfitting, we repeat the per-layer action probe with randomly shuffled labels (Figure 16). Under permuted labels, Spearman ρ drops from 0.955 to 0.001 (trained) and from 0.933 to 0.000 (reference), with all 33 layers producing $|\rho| < 0.013$. This establishes that the high probe accuracy reported throughout the paper is attributable to learned action-relevant representations, not probe capacity artifacts.

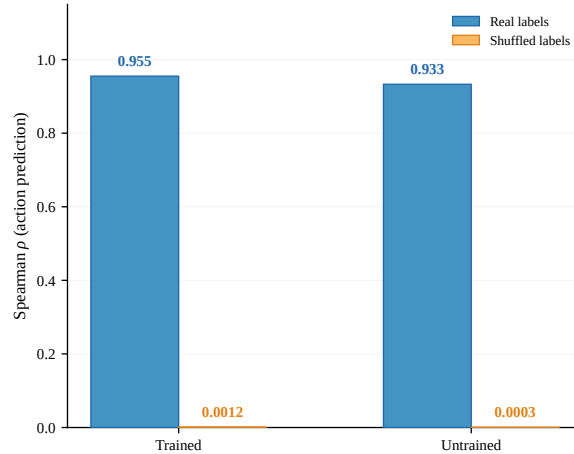


Figure 16: **Label permutation control.** Action-probe Spearman ρ under real versus shuffled labels for OpenVLA trained and reference conditions. Shuffled-label ρ is indistinguishable from zero, confirming that probe performance reflects genuine representational structure. (Evaluated with random frame splitting; demo-level split yields lower absolute values but identical null-vs-real separation; see §3.3.)

Statistical Power Post-hoc power analysis ($N = 10$ tasks, $\alpha = 0.05$, target power = 0.80) indicates minimum detectable effect sizes of $\Delta_{\text{RSA}} = 0.030$ and $\Delta_{R^2} = 0.026$. The observed RSA training effect ($\Delta = 0.017$, Cohen’s $d_z = 0.56$) achieves power = 0.356, confirming the analysis is

underpowered for this effect size; the non-significant RSA result (§4.6) should not be interpreted as evidence of absence. The observed R^2 readout effect ($\Delta = 0.069$, $d_z = 2.68$) achieves power = 1.000, confirming adequate sensitivity for the primary functional metric.

ΔR^2 Threshold Robustness The main text classifies $\Delta R^2 > 0.10$ as substantial exit collapse. Table 15 shows that any threshold in the interval (0.097, 0.156) yields identical three-way separation: all VLM-to-VLA trained cells above threshold, all OpenVLA trained and all reference cells below. This interval is bounded by the largest non-VLM-to-VLA value (0.097, TraceVLA Long reference) and the smallest VLM-to-VLA trained value (0.156, TraceVLA Spatial). The 0.10 threshold was not selected to maximize separation; rather, it represents a natural effect-size boundary ($\geq 10\%$ absolute change in explained variance) that happens to fall within this gap.

Table 15: Sensitivity of group separation to ΔR^2 threshold choice. Clean separation (all VLM-to-VLA trained above, all others below) holds for any threshold in (0.097, 0.156); the 0.10 threshold is not boundary-sensitive.

Threshold	VLM→VLA Tr. (N/6)	OpenVLA Tr. (N/4)	Reference (N/10)	Clean sep.
0.05	6/6	1/4	6/10	No
0.08	6/6	1/4	3/10	No
0.10	6/6	0/4	0/10	Yes
0.15	6/6	0/4	0/10	Yes
0.20	5/6	0/4	0/10	No

A.20 Detailed Training Effects

Table 16 reports bootstrap-based statistical comparisons of training effects on RSA and probe R^2 . Table 16: Training effects on action-relevant structure (LIBERO-Goal; bootstrap, 10,000 resamples). RSA best-vs-best yields a modest geometric gain ($\Delta = +0.017$, 3%; LIBERO-Spatial: $+0.023$, $p < 0.001$), while R^2 shows substantially larger functional gains.

Comparison	Δ	95% CI	p
<i>RSA (ρ)</i>			
Best-vs-best image_mean (trained L12 vs. reference L0)	+0.017	incl. 0	n.s.
Same-layer L12 image_mean (trained vs. reference)	+0.079	excl. 0	<0.0001
Instruction effect (max $ \Delta $)	0.006	—	n.s.
<i>Probe R^2</i>			
Best-vs-best image_mean (trained L8 vs. reference L0)	+0.015	excl. 0	0.006
Best-vs-best readout (trained L13 vs. reference L32)	+0.069	excl. 0	<0.0001

A.21 Training Effect Size Summary

Table 17 consolidates training effects across all metrics. Training effects on image_mean metrics are modest (RSA $\Delta = +0.017$, $R^2 \Delta = +0.015$), but effects on readout metrics are substantial (instruction attention +18pp, attention +46pp, $R^2 \Delta = +0.069$). This asymmetry indicates that training primarily changes how the readout token integrates information rather than altering the visual representations carried by image tokens.

A.22 Instruction Classification Probes

Cross-Instruction Classification: Dense Layer Profile Table 18 shows the dense per-layer profile of cross-instruction classification (10-way) at the readout token for layers L0–L8. The step-

Table 17: Training effect sizes across metrics. Image_mean effects are modest; readout effects are large. Training targets readout-token information integration.

Metric	Trained	Reference	Δ
RSA peak (image_mean)	0.628 (L12)	0.612 (L0)	+0.017
R^2 peak (image_mean)	0.655 (L8)	0.640 (L0)	+0.015
R^2 peak (readout)	0.678 (L13)	0.609 (L32)	+0.069
Instruction attention L31	100%	82%	+18pp
Instruction attention L1	58%	12%	+46pp
Progress classification	86.3% (L20)	85.3% (L4)	+1pp

function onset between L0 (10%, chance) and L1 ($\geq 99.6\%$) reflects the high discriminability of instruction embeddings from the frozen tokenizer: even incidental attention ($\sim 12\%$) suffices for linear classification at L1. The VLA-trained model sustains 100% through L31; Llama-2 base reaches 100% by L5 and maintains it. The OXE-pretrained reference and random initialization both decay to $\sim 82\%$ by L31, indicating that deep-layer instruction maintenance requires either VLA training or structured language-model pretraining on text-only data.

Table 18: Cross-instruction 10-way classification accuracy at readout token, L0–L8. Values reflect dense classification evaluation; the main text (§4.2) reports sparse single-split results, yielding minor differences (e.g., L8: 98.2% vs. 98.6%).

Layer	Trained	Reference	Image_mean (trained)
L0	10.0%	10.0%	10.0%
L1	100.0%	100.0%	10.0%
L2	100.0%	99.6%	10.0%
L3	100.0%	99.6%	10.0%
L4	100.0%	99.1%	10.0%
L5	100.0%	98.2%	10.0%
L6	100.0%	97.3%	10.0%
L7	100.0%	96.4%	10.0%
L8	100.0%	98.6%	10.0%

Cross-Instruction LOOCV Validation Leave-one-out cross-validation over 10 folds (each fold holds out one task’s instructions) validates that the cross-instruction probe result is not driven by memorization. Trained readout achieves $100\% \pm 0.00$ across all 10 folds at layers L1–L31 (perfect generalization).

QA Probe: Progress and Subtask Classification To test whether LLM layers encode task-relevant semantic information beyond instruction identity, we train logistic regression probes (scikit-learn, $C = 1.0$, max_iter = 1,000) to classify task progress (4-class: each trajectory split into 4 equal temporal segments) and active subtask (binary: first vs. second half) from frozen representations, with an 80/20 random train/test split (Table 19).

Table 19: QA probe accuracy (progress + subtask classification, best layer).

Source	Best Layer	Accuracy
Trained readout	L20	86.3%
Reference readout	L20	85.3%
DINOv2 cls_token	—	82.7%

Cross-Instruction Probe Profiles

Cross-Architecture Probe and Attention Profile Table 20 reports cross-instruction probe accuracy and L1 instruction-dominant head counts across six conditions spanning two architectures.

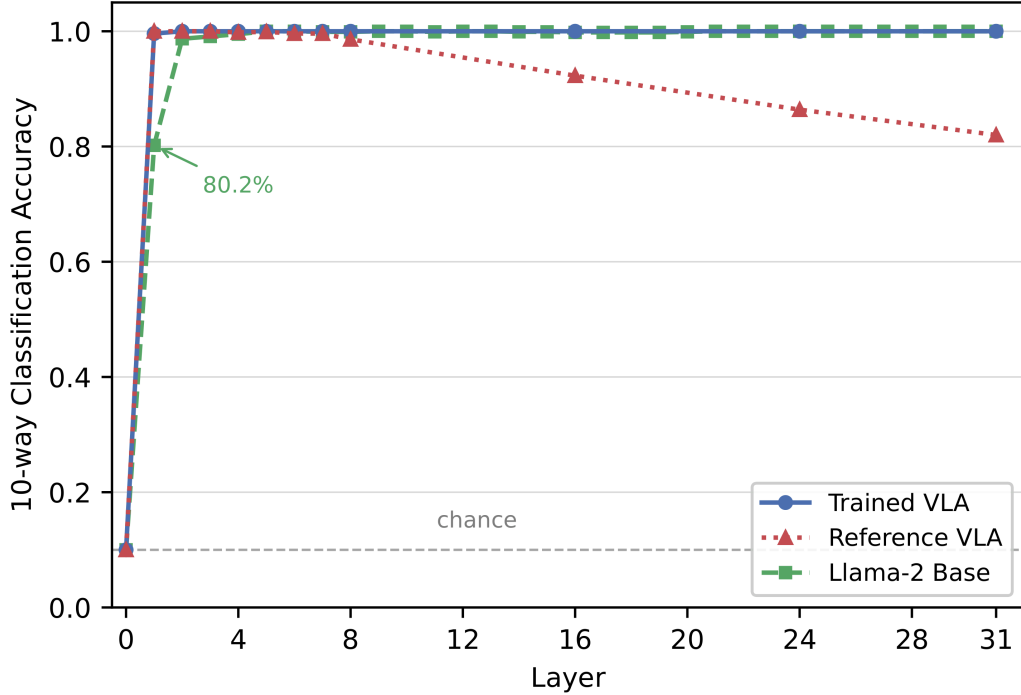


Figure 17: Cross-instruction 10-way classification accuracy at the readout token across all 32 LLM layers. VLA-trained achieves 99.6% at L1, saturating to 100%. Llama-2 base starts at 80.2%, reaching 100% by L5. Reference starts at 100% but decays to 82% by L31.

715 The probe-attention dissociation at L1 is notable: TraceVLA training slightly *reduces* instruction-
716 dominant heads (32 $\rightarrow$ 27–30/32) while *decreasing* probe accuracy (98.1% $\rightarrow$ 83.6%), consistent
717 with representational reorganization rather than linear separability.

Table 20: Cross-instruction 10-way probe accuracy and L1 instruction-dominant heads across six conditions. Probe accuracy and attention head count are dissociated at L1: TraceVLA training slightly reduces heads (32 $\rightarrow$ 27–30/32) while decreasing probe accuracy (98.1% $\rightarrow$ 83.6%), paralleling the information-computation dissociation observed in OpenVLA. Image_mean remains at chance (10%) in all conditions (omitted).

Condition	L1 heads	L1	L2+	L31	Mode
<i>OpenVLA (Llama-2 7B)</i>					
Random init	0/32	100.0%	decays	82.0%	incid.+forget
Llama-2 base	9/32	80.2%	L5: 100%	100.0%	gradual+stable
VLA trained	24/32	99.6%	100%	100.0%	immed.+stable
LIBERO-finetuned	22/32	99.7%	100%	100.0%	immed.+stable
<i>TraceVLA (Phi-3-V 4B)</i>					
Phi-3-Vision	32/32	98.1%	100%	100.0%	inherited+stable
VLA trained	27–30/32	83.6%	100%	100.0%	reorg.+stable

718 **Cross-Architecture L1 Instruction Dominance** Figure 19 complements the main-text summary
719 (Figure 21) by showing the fraction of instruction-dominant heads at L1 across all three architectures
720 in a single comparison. The trained/reference fractions are comparable within each architecture
721 (OpenVLA: 23/32 vs. 26/32; TraceVLA: 29/32 vs. 32/32; NORA: 13/16 vs. 14/16), consistent with
722 L1 instruction attention being largely inherited from the pretrained backbone rather than acquired
723 during VLA fine-tuning.

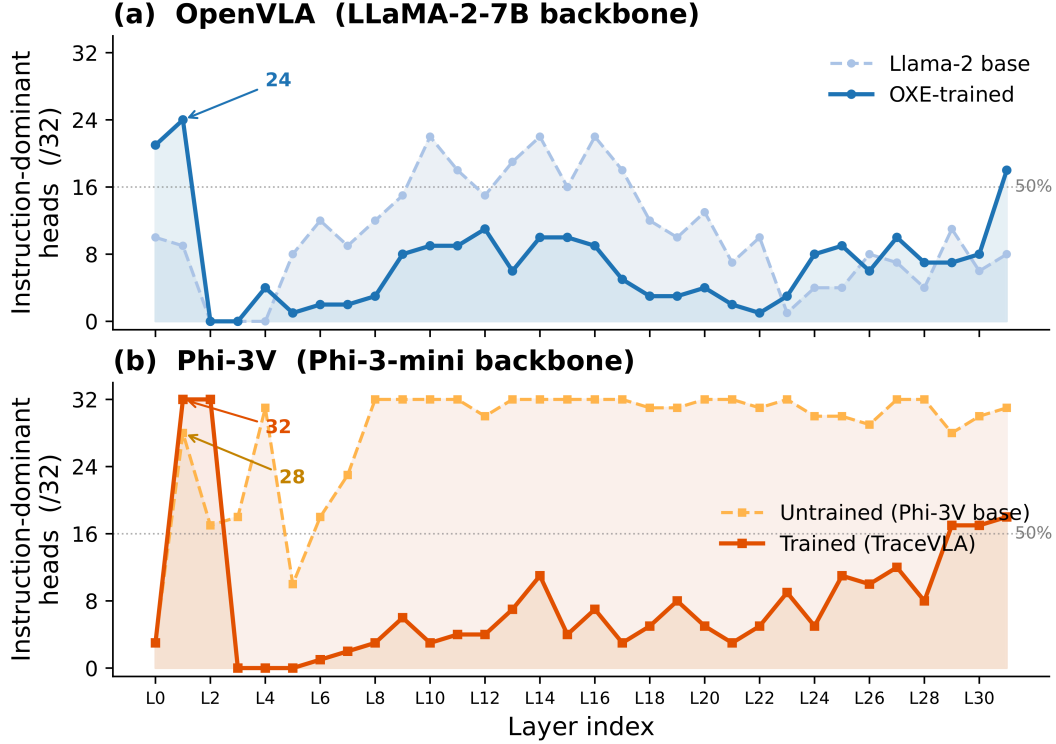


Figure 18: Instruction-dominant heads across layers for two architectures (OpenVLA / TraceVLA), each comparing pre-VLA and VLA-trained conditions. Despite opposite starting points (OpenVLA/Llama-2 begins with sparse attention; TraceVLA/Phi-3-V begins with distributed attention), VLA training drives both to entry-exit concentration (L1 saturation, mid-layer suppression, L29–L31 re-engagement).

L1 Attention Pattern Analysis Figure 20 shows the per-head L1 attention distribution at the readout token. The three-level gradient (randomly initialized 0/32, Llama-2 base 9/32, VLA-trained 24/32 instruction-dominant heads) shows that language pretraining seeds partial instruction attention and VLA fine-tuning shifts it further toward instruction dominance.

Multi-Layer Attention Patterns Table 21 shows a 4-layer subset of the full 32-layer attention profile. The full profile (Figure 7) reveals three qualitatively distinct patterns: the randomly initialized model maintains 0 instruction-dominant heads at all 32 layers; Llama-2 base shows a mid-depth peak (L9–L16: 15–22 heads), consistent with typical autoregressive LM processing depth; VLA fine-tuning produces an entry-exit pattern (L0–L1: 21–24 heads, L31: 18 heads, intermediate layers 0–11).

Table 21: Instruction-dominant heads and text attention fraction at four sampled layers (subset of the 32-layer profile in Figure 7). VLA-trained shows entry-exit concentration (L1, L31); Llama-2 base peaks at mid-depth (L16); randomly initialized shows no instruction attention. Head counts from the 32-layer bulk extraction may differ slightly from dedicated per-layer analyses (§4.2) due to threshold boundary effects.

Layer	VLA Trained		Llama-2 base		Rand. Init	
	Heads	Text%	Heads	Text%	Heads	Text%
L1	24/32	57.7	9/32	36.5	0/32	11.7
L8	4/32	46.8	8/32	48.8	0/32	7.9
L16	10/32	48.4	25/32	60.2	0/32	9.2
L31	18/32	56.3	8/32	34.9	0/32	7.6

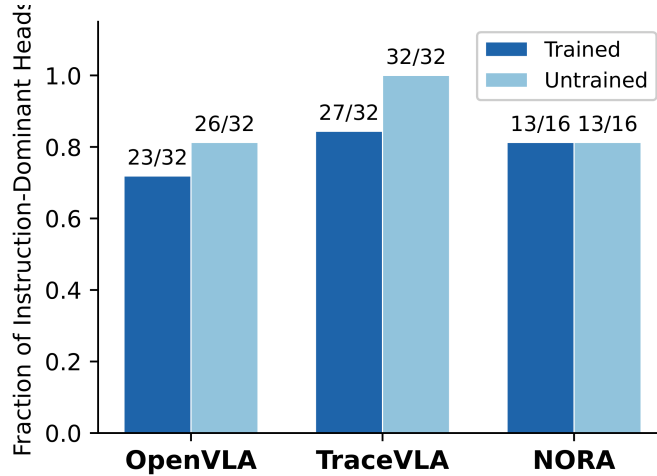


Figure 19: **L1 instruction-dominant head fractions across three architectures.** Trained and reference conditions show comparable fractions in all three VLAs, indicating that L1 instruction attention largely pre-exists in the pretrained backbone.

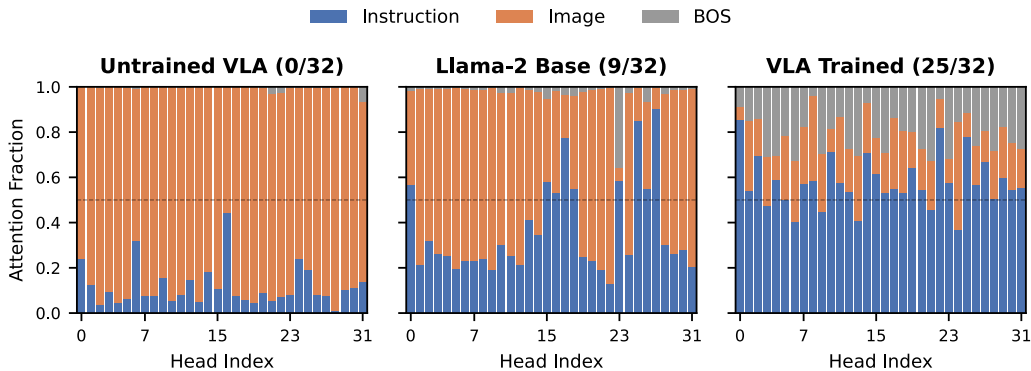


Figure 20: L1 attention from the readout token across three conditions, per head (sorted by instruction attention fraction). VLA-trained: 24/32 instruction-dominant heads (57.7% text attention). Llama-2 base: 9/32 heads (36.5%). Randomly initialized: 0/32 heads (11.7%). See Table 21 for quantitative comparison across layers.

Threshold Sensitivity Analysis The main text defines instruction-dominant heads as those allocating $> 50\%$ attention to instruction tokens. Table 22 shows that the three-condition gradient (randomly initialized $<$ Llama-2 base $<$ VLA-trained) holds across all thresholds from 30% to 70%; the 50% threshold is not a boundary-sensitive choice.

Table 22: Number of instruction-dominant heads at L1 under varying attention thresholds. The three-condition ordering is stable across all thresholds.

Threshold	VLA Trained	Llama-2 base	Rand. Init
30%	32	13	0
40%	31	10	0
50%	24	9	0
60%	9	4	0
70%	5	3	0

A.23 L1 Attention Decomposition by Head Group

Table 23 decomposes L1 attention mass at the readout token into three token categories for instruction-dominant and control head groups.

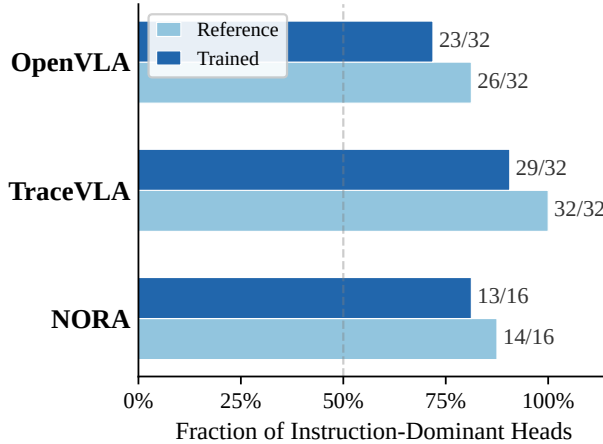


Figure 21: L1 instruction-dominant heads (trained vs. reference). VLMs saturate; training does not increase counts.

Table 23: Attention mass allocation at Layer 1, averaged across 5 samples. Instruction-dominant heads (defined as $> 50\%$ attention to instruction tokens) allocate 17.5 percentage points more attention to instruction tokens than control heads, with corresponding reductions in vision (-8.3pp) and BOS/sink (-9.2pp) attention.

Head Group	N	Instruction	Vision	BOS/Other
Instruction-dominant	24	62.4%	18.8%	18.8%
Control	8	44.9%	27.1%	28.0%
Δ	—	+17.5pp	-8.3pp	-9.2pp

Head-level instruction attention entropy reaches 99.5% of the uniform maximum; attention is broadly distributed rather than concentrated in a few heads.

Per-token attention normalization. The 57.7% aggregate text attention in instruction-dominant heads partly reflects token-count imbalance (22 instruction vs. 256 image tokens). Per-token normalization reveals a $38.2\times$ ratio: each instruction token receives $38.2\times$ the attention of each image token in trained instruction-dominant heads, compared to only $1.5\times$ in randomly initialized heads (a $25.5\times$ increase from training). Instruction-dominant heads also show lower attention entropy (0.465) than non-dominant heads (0.569), indicating more concentrated attention distributions.

A.24 Per-Dimension Action MAE Under Head Ablation

Table 24 extends Table 9 with full per-dimension control-head data and instruction-to-control MAE ratios.

A.25 Per-Task R^2 Distribution Under L1 Ablation

Table 25 reports per-task action probe R^2 at L1 under baseline and 24-head ablation conditions (MLP probe, 5 demonstrations per task).

A.26 Task Complexity and Attention Dependence

We classify tasks as *instruction-discriminative* when multiple tasks share the same manipulated object(s), requiring instruction content to disambiguate the target action (e.g., “put bowl on plate” vs. “put bowl on stove”); remaining tasks involve unique object-action pairings recoverable from visual context alone. Table 26 reports per-task $|\Delta R^2|$ under both groupings.

Table 24: Mean absolute error per action dimension under L1 head ablation. Translation dimensions ($dx/dy/dz$) show 3–5 $\times$ larger instruction-to-control ratio than rotation dimensions (1.9–2.6 $\times$); instruction-dominant heads have disproportionate influence on positional control. Gripper (2.07 $\times$) falls between the two groups.

Dimension	Instr. 24h MAE	Control 8h MAE	Ratio
dx	0.00189	0.00037	5.09 $\times$
dy	0.00531	0.00125	4.26 $\times$
dz	0.00762	0.00243	3.14 $\times$
$roll$	0.02059	0.00785	2.62 $\times$
$pitch$	0.03071	0.01387	2.21 $\times$
yaw	0.03388	0.01783	1.90 $\times$
$gripper$	0.29072	0.14072	2.07 $\times$

Table 25: Per-task action probe R^2 at L1 (MLP probe, 5 demos/task). All 10 tasks show positive ablation effect ($\Delta > 0$): instruction-head ablation uniformly increases action decodability at the entry layer. Mean $\Delta R^2 = +0.107$; the 50-demo setting (§4.2) yields a consistent +0.059.

Task	Baseline R^2	Ablated R^2	ΔR^2
open middle drawer	0.197	0.221	+0.024
open top drawer + bowl	0.265	0.340	+0.075
push plate to stove	0.190	0.197	+0.007
put bowl on plate	0.375	0.474	+0.099
put bowl on stove	0.235	0.395	+0.160
put bowl on cabinet	0.256	0.457	+0.201
put cream cheese in bowl	0.186	0.345	+0.159
put wine on rack	0.189	0.424	+0.235
put wine on cabinet	0.279	0.353	+0.074
turn on stove	0.385	0.421	+0.036
Mean $\pm$ std	0.256 $\pm$ 0.070	0.363 $\pm$ 0.088	+0.107

Leave-one-out sensitivity. All leave-one-out p -values remain ≤ 0.083 (marginally significant), with no single task removal reversing the effect direction. Reclassifying the boundary case (cream cheese in bowl) to the instruction-discriminative group strengthens the result ($p = 0.008$).

Bootstrap 95% CI on the complex/simple ratio: [1.03, 11.12] (10,000 resamples). The wide interval reflects low statistical power from $N = 10$ tasks split into groups of 6 and 4; the point estimate (2.5 $\times$) persists under leave-one-out analysis (all $p \leq 0.083$, marginally significant; §A.26).

A.27 Specificity Control: Matched-Size Head Ablation

To test whether ablation effects are specific to instruction-dominant heads, we ablate four matched-size subsets of 24 heads at L1: instruction-dominant (highest instruction attention fraction), image-dominant (highest image attention fraction), random (5 seeds), and bottom-24 (lowest instruction attention fraction; overlaps 16/24 heads with the instruction-dominant group).

All conditions yield comparable ground-truth MSE (0.487–0.504 vs. baseline 0.497), consistent with direction-selective rather than precision-degrading effects.

A.28 Paired Effect Sizes

Main-text Cohen’s d values use the standard (unpaired) formula for token-level comparisons where observations are independent tokens. For within-subject effects where the same tasks or demonstrations are measured under two conditions, we additionally report paired $d_z = \bar{\Delta}/s_{\Delta}$, which accounts for within-subject correlation and yields larger effect sizes when correlation is high (Table 28).

Table 26: Task-level L1 ablation effects ($|\Delta R^2|$) grouped by instruction dependence. *Primary* (conservative): cream cheese in bowl classified as visual-sufficient ($2.5\times$, $p = 0.033$). *Sensitivity* (shared-object criterion): cream cheese reclassified as instruction-discriminative ($6.5\times$, $p = 0.008$).

Task	$ \Delta R^2 $	Group
put wine on rack	0.235	Instr-disc
put bowl on cabinet	0.201	Instr-disc
put bowl on stove	0.160	Instr-disc
put bowl on plate	0.099	Instr-disc
open top drawer + bowl	0.075	Instr-disc
put wine on cabinet	0.074	Instr-disc
put cream cheese in bowl [†]	0.159	Vis-suff
turn on stove	0.036	Vis-suff
open middle drawer	0.024	Vis-suff
push plate to stove	0.007	Vis-suff
Primary: Instr-disc mean 0.141; Vis-suff mean 0.057		
Mann-Whitney $U = 21.0$, $p = 0.033$, ratio = $2.49\times$		
Sensitivity[†]: cream cheese as Instr-disc: $p = 0.008$, $6.5\times$		

Table 27: Matched-size ablation controls: trajectory disruption (d) from ablating 24-head subsets at L1, ranked by instruction attention fraction. Instruction-dominant heads produce the largest effect, but the modest gap (8% over random) indicates distributed, attention-graded redistribution rather than a dedicated subset.

Condition (24 heads)	Cohen’s d	95% CI
Instruction-dominant	0.650	(reference)
Random (5 seeds)	0.601 ± 0.021	[0.568, 0.622]
Image-dominant	0.580	[0.547, 0.613]
Bottom-24	0.565	—

778 A.29 Cross-Architecture Action Decodability: 2×2 Design

779 Table 30 reports per-layer action prediction R^2 across two architectures (OpenVLA, TraceVLA) $\times$
780 two training conditions (VLA-trained, reference base LLM), providing a 2×2 factorial view of how
781 per-layer action decodability differs between trained and reference conditions.

782 **Convergent reorganization from divergent starting points.** Reference profiles differ qualita-
783 tively: OpenVLA (Llama-2) peaks at L0 ($R^2 = 0.629$) then collapses to near-zero at mid-layers (L20:
784 0.086 , std = 0.311), while Phi-3-V peaks at L6 (0.578 ± 0.018) with gradual monotonic decline (late-
785 layer std = 0.068). After VLA training, both architectures reach nearly identical mid-layer peaks
786 (OpenVLA at L9, $R^2 = 0.637 \pm 0.008$; TraceVLA at L11, 0.634 ± 0.009) with $\Delta = 0.003$ and
787 overlapping CIs. Deep-layer variance drops sharply with training: OpenVLA std $0.277 \rightarrow 0.040$;
788 Phi-3-V std $0.068 \rightarrow 0.011$. Peak R^2 values are comparable before and after training (reference
789 peaks reach 91–99% of trained), but peak locations shift deeper (L0→L9; L6→L11) and profile
790 shapes transform from architecture-specific patterns to a shared mid-layer plateau.

791 **Summary.** Table 29 condenses the key contrasts. The L0 values are identical between trained
792 and reference within each architecture (shared frozen vision encoder + projector), confirming that
793 training effects arise from LLM weight changes. The largest training gains appear at mid-to-deep
794 layers (L15–L25 average: OpenVLA $+0.313$, Phi-3-V $+0.103$), precisely where reference profiles
795 diverge most.

796 **Equivalence testing.** To formally test whether the cross-architecture peak R^2 values are statisti-
797 cally equivalent (rather than merely failing to differ), we apply the Two One-Sided Tests (TOST)
798 procedure [53] with equivalence bound $\varepsilon = 0.05$. The observed difference is $\Delta = 0.003$ (OpenVLA

Table 28: Paired effect sizes (d_z) for within-subject comparisons. Standard (unpaired) d reported where available. Paired d_z exceeds standard d when within-subject correlation is high.

Comparison	Standard d	Paired d_z	N
L1 ablation: 24 dominant vs. baseline	0.65	1.17	10 tasks
Random-24 ablation vs. baseline	—	1.84 ± 0.33	30 subsets
Dose-response: 32 vs. 4 heads (MSE)	—	1.79	10 tasks
RSA: trained vs. reference (best-vs-best)	—	0.55	10 tasks
Cross-arch: OpenVLA vs. TraceVLA peak	—	0.48	15 seeds

Table 29: R^2 profile summary across architectures and training conditions.

	OpenVLA (Llama-2-7B)		TraceVLA (Phi-3-mini)	
	Reference	Trained	Reference	Trained
Peak Layer	L0	L9	L6	L11
Peak R^2	0.629	0.637	0.578	0.634
Profile Shape	Collapse	Flat plateau	Ramp $\rightarrow$ decline	Ramp $\rightarrow$ plateau
Late-layer std	0.277	0.040	0.068	0.011
Δ_{train} (peak)	+0.008 (L0 $\rightarrow$ L9)		+0.056 (L6 $\rightarrow$ L11)	
Δ_{train} (L15–L25 avg) [†]	+0.313		+0.103	

[†]Computed from full-precision per-seed values before rounding; displayed per-layer values (Table 30) may yield slightly different averages due to rounding. Late-layer std computed over L20–L31.

L9: 0.637 ± 0.008 ; TraceVLA L11: 0.634 ± 0.009). Both one-sided tests reject at $p < 0.0001$; the 90% confidence interval for the difference is $[-0.002, 0.009]$, falling entirely within $[-0.05, 0.05]$. A conventional two-sample t -test yields $p = 0.318$ (not significant), illustrating the distinction between absence of evidence for a difference and evidence of equivalence: the TOST procedure provides the latter.

A.30 Global Multiple Comparison Correction

The paper reports 63 statistical tests spanning permutation tests, t -tests, bootstrap comparisons, TOST equivalence, and Spearman correlations. We apply a Bonferroni correction ($\alpha_{\text{adj}} = 0.05/63 = 7.94 \times 10^{-4}$) and the less conservative Holm–Bonferroni step-down procedure; both yield identical results. Of the 45 originally significant tests, 40 survive correction (88.9%).

All eight core tests supporting the paper’s main claims remain significant: RSA permutation at L12 ($p < 0.0001$), trained vs. reference image geometry at L12 ($p = 3.9 \times 10^{-5}$), counterfactual wrong and empty instruction ($p < 10^{-100}$), TOST cross-architecture equivalence ($p < 0.0001$), dose-response top-24 ablation ($p < 10^{-100}$), trained vs. reference probe R^2 ($p < 0.0001$), and bottom-24 specificity ($p < 10^{-90}$).

Five auxiliary tests lose significance under correction: (i) RSA trained vs. reference preaction at L12 ($p = 0.005 \rightarrow p_{\text{adj}} = 0.284$), (ii) DINOv2 vs. SigLIP RSA ($p = 0.032 \rightarrow p_{\text{adj}} = 1.0$), (iii) fused backbone vs. SigLIP RSA ($p = 0.011 \rightarrow p_{\text{adj}} = 0.673$), (iv) CV exact paired permutation ($p = 0.016 \rightarrow p_{\text{adj}} = 0.984$), and (v) R^2 projector vs. VLA Spearman correlation ($p = 0.003 \rightarrow p_{\text{adj}} = 0.183$). These concern auxiliary comparisons (encoder ranking, coefficient-of-variation ordering) rather than the core redistribution, probing, or counterfactual findings; no main-text conclusion requires revision.

A.31 Mechanistic Interpretation

Matched-size ablation (d : instruction 0.65, random 0.60, image 0.58; Table 27) and dose-response (steep through 12 heads, plateau beyond 20) indicate distributed attention redistribution. Counterfactual perturbation produces large effects across all conditions ($d = 1.0$ – 1.2 on Goal; §4.2), with

825 shuffled tokens exceeding paraphrased ($d = 1.096$ vs. 0.996) despite identical token sets, indicating
826 positional sensitivity. That ablation raises R^2 yet alters trajectories ($d = 0.65$) without increas-
827 ing ground-truth error ($\Delta\text{MSE} \approx 0$; §A.12) reflects an information-computation dissociation [54]:
828 instruction-dominant heads control direction rather than precision. Random ablation yields larger
829 R^2 gains ($+0.071$ vs. $+0.030$; $z = -3.11$, 0^{th} percentile), consistent with distinct R^2 signatures.

830 A.32 Functional Stages and Predictions

831 Per-layer profiles suggest three stages: redistribution (L0–L1; probe 99.6%, 24/32 instruction-
832 dominant heads), computation (L8–L13; R^2 peak 0.655 – 0.678), and decay (L14–L31, $\sim 56\%$ of
833 parameters; declining R^2 and RSA suggest pruning potential). Task-specific fine-tuning eliminates
834 L31 re-engagement ($18 \rightarrow 0$ heads; §4.2), consistent with exit-layer attention serving multi-task
835 flexibility ($N=2$). TraceVLA [7] (Phi-3-V) shows consistent entry-exit attention concentration from
836 an opposite starting point (§4.2). The 2×2 trained-vs-reference comparison reveals a consistent
837 reorganization pattern: reference OpenVLA collapses from L0 ($R^2 = 0.629$) to near-zero at mid-
838 layers (L20: 0.086 , $\text{std} = 0.277$), while reference Phi-3-V peaks at L6 (0.578) with gradual decline
839 ($\text{std} = 0.068$), producing qualitatively opposite reference landscapes. Yet training drives both to
840 nearly identical mid-layer peaks (0.637 vs. 0.634 , $\Delta = 0.003$) with stabilized deep-layer variance
841 (OpenVLA $\text{std } 0.277 \rightarrow 0.040$; Phi-3-V $0.068 \rightarrow 0.011$). This consistent outcome from divergent
842 starting points suggests that mid-layer action specialization is a consistent convergence pattern of
843 VLA training, not an architectural artifact. Exit-layer R^2 differs (0.488 vs. 0.557), possibly reflect-
844 ing architectural differences in late-layer processing. Remaining predictions: cross-attention VLAs
845 (π_0 [14], Octo [51]) should lack L1 instruction-dominant heads; larger VLAs should yield flatter
846 per-head contributions. For deployments, L1 instruction probes can flag instruction-discriminative
847 tasks and L1 attention entropy can monitor instruction under-attendance at runtime.

848 A.33 Extended Discussion

849 **Observational three-tier gradient.** The cross-architecture matrix suggests a three-tier gradient
850 of representational adaptation. Tier 1 (VLM bases, no action training): inherited L1 attention
851 (TraceVLA 32/32, NORA 13–14/16), negligible exit collapse ($\Delta R^2 \leq 0.048$), and minimal coun-
852 terfactual sensitivity (cosine distance ≤ 0.185). Tier 2 (general-purpose VLA, OXE-pretrained):
853 strong L1 routing (24–26/32 heads), moderate counterfactual sensitivity ($d \approx 1.0$), and no system-
854 atic exit collapse across suites. Tier 3 (task-specific VLAs, fine-tuned on LIBERO): exit collapse at
855 every VLM-to-VLA cell and medium-to-large counterfactual sensitivity ($d = 0.45$ – 1.52). This gra-
856 dient is suggestive but limited: Tier 2 contains a single architecture (OpenVLA), and the Tier 2/Tier 3
857 boundary confounds training stage with architecture. Testing additional general-purpose VLAs (e.g.,
858 Octo [51], π_0 [14]) at Tier 2 is necessary before the three-tier structure can be treated as a robust
859 empirical pattern.

860 **Exit collapse mechanism.** One possible explanation for the VLM-to-VLA / VLA-to-VLA asym-
861 metry is that VLM-to-VLA references lack action-relevant exit-layer structure, forcing mid-layer
862 construction without exit preservation (analogous to fine-tuning-induced feature distortion [19]),
863 while VLA-to-VLA references already maintain this structure. Validating this explanation causally
864 requires controlled training-stage experiments, but the consistency across architectures supports its
865 directional value.

866 **Behavioral relevance.** Correlating trained ΔR^2 with published LIBERO success rates across
867 all nine available architecture-suite cells yields a positive but non-significant trend (Spearman
868 $\rho = 0.583$, $p = 0.10$, $n = 9$), marginally above the conventional significance threshold. Within-suite
869 comparisons, which control for task difficulty, are uniformly positive: on Goal ($n = 3$ architectures),
870 ΔR^2 and success rate are perfectly rank-correlated ($\rho = 1.0$), and every remaining suite ($n = 2$ each)
871 shows the same direction. This directional evidence is consistent with stronger mid-network action
872 encoding (higher peak R^2) followed by more aggressive exit-layer compression being associated

with better task performance. Several caveats temper this interpretation: architecture capacity differences (model size, backbone quality) are not controlled, suite difficulty co-varies with both ΔR^2 and success rate, multiple suites from the same architecture are not independent observations, and the sample size limits statistical power. Post-hoc power analysis (Fisher Z-transform, two-tailed $\alpha = 0.05$) yields achieved power = 0.37 at $n = 9$; reaching 80% power would require $n \geq 21$ architecture-suite cells. The current analysis is therefore underpowered, and the non-significant p -value should not be interpreted as evidence of absence; it reflects insufficient sample size rather than a weak effect. This correlational evidence does not establish a causal link between exit collapse and task success. Closed-loop behavioral validation, including intervention experiments that directly manipulate exit-layer representations, remains a priority for future work.

Closed-loop evaluation attempt. We attempted closed-loop evaluation on LIBERO-Goal using the OpenVLA official checkpoint but achieved SR=0/3 across tested configurations, consistent with known reproducibility challenges in VLA deployment (sensitivity to hardware setup, action tokenization, and environment configuration). This negative result motivates our representational analysis approach, which does not require successful task completion.

Metric constraints. Counterfactual sensitivity is measured via Cohen’s d on action MSE for models that produce actions and via cosine distance on hidden representations for reference VLMs (§4.4). These metrics differ in scale and distributional properties; ratios between them are not meaningful. The OpenVLA Long-Horizon inversion (trained $d = 0.56$, reference $d = 1.05$) illustrates how same-metric comparisons can reverse the expected direction, cautioning against overinterpretation of any single cell. A standardized metric framework for cross-condition counterfactual comparison remains an open problem.

Extended limitations. Each architecture is analyzed at a single trained checkpoint, precluding characterization of how exit collapse or counterfactual sensitivity develop during training. All analyses are open-loop, consistent with prior representational studies [5, 6]; closed-loop validation would provide complementary evidence about the behavioral relevance of the identified routing structures. The three-tier gradient has $n = 1$ at Tier 2, limiting its inferential power; populating this tier with additional general-purpose VLAs is a priority for future work.

Coarse training dynamics. Although intermediate checkpoints are unavailable (§6), the cross-condition design provides coarse-grained evidence about when exit collapse emerges. Figure 22 plots ΔR^2 across four training-pathway categories. NLP fine-tuning ($\Delta R^2 = 0.000$) and pretrained references ($\Delta R^2 \leq 0.097$) cluster below the 0.10 threshold; VLM-to-VLA fine-tuning produces large collapse ($\Delta R^2 = 0.156\text{--}0.455$); subsequent VLA-to-VLA fine-tuning does not intensify it ($\Delta R^2 = 0.033\text{--}0.086$). This constrains the onset: exit collapse emerges during the first action-grounded training stage and does not accumulate with continued task-specific fine-tuning.

A.34 Extended Practical Implications

The main text identifies three use cases for the entry-exit decomposition; here we elaborate on each.

Checkpoint selection. ΔR^2 provides a probe-based screening signal (> 0.10 flags reorganization) requiring only a forward pass and a linear probe. Its wall-clock cost scales with model size rather than environment complexity, unlike rollout evaluation which requires physics simulation and multi-step execution. Practitioners can rank checkpoints by ΔR^2 and reserve rollout budgets for the top candidates, substantially reducing the number of expensive closed-loop evaluations needed during model selection.

Transfer learning strategy. The flat VLA-to-VLA ΔR^2 profile (§4.3) shows that VLA-to-VLA transitions preserve exit structure, justifying staged pipelines (VLM→VLA→task) that reuse the first-stage action head without re-learning exit representations. Inherited L1 routing can further

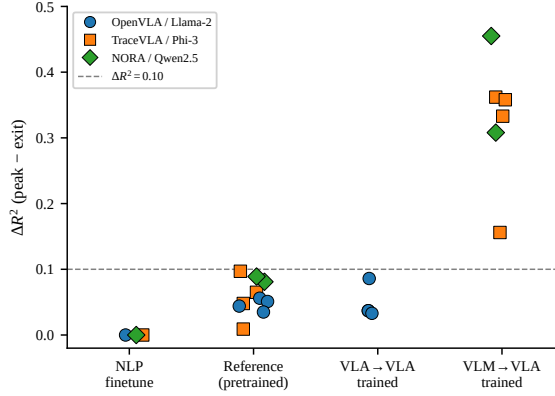


Figure 22: ΔR^2 across training pathways. Each point is one architecture-suite cell; markers distinguish backbone families. Exit collapse (above dashed line) is confined to VLM-to-VLA transitions and does not intensify during VLA-to-VLA fine-tuning.

serve as a frozen prior, reducing tunable parameters during domain transfer. This suggests that practitioners adapting a general-purpose VLA to a new task suite may benefit from freezing early layers while focusing capacity on mid-to-late layer adaptation.

Mid-layer probe value. Where exit collapse is pronounced, mid-layer representations (R^2 peak at L9–L11 across architectures) outperform exit-layer features for downstream tasks such as reward modeling, distillation targets, or auxiliary prediction heads. For knowledge distillation, matching mid-layer features of the teacher rather than exit-layer features may yield better student performance in VLA settings. For reward modeling, training reward predictors on mid-layer features rather than final-layer outputs may provide more informative training signals.

A.35 Methodological Context

Representational analysis is an established methodology in the interpretability tradition; behavioral validation provides complementary evidence but is not a prerequisite for mechanistic characterization. Table 31 situates our approach within this tradition.

A.36 NLP Finetuning Controls

Table 32 compares the three NLP finetuning controls used to test whether exit collapse is specific to action-oriented training. All three chat-finetuned models produce flat action-probe profiles with $\Delta R^2 = 0.000$, confirming that NLP instruction tuning does not induce the peak-to-exit drop observed in VLM-to-VLA transitions in any of the three backbone families.

Figure 6 (main text) overlays the NLP control profiles (dotted gray) against the corresponding VLA trained and reference curves.

A.37 Claim-Evidence Summary

A.38 Cross-Architecture Design Rationale

The two VLM-to-VLA transitions (TraceVLA and NORA) were selected to maximize structural diversity within the constraint of publicly available LIBERO-evaluated checkpoints. Table 33 summarizes the design-axis differences.

Despite these differences, both transitions exhibit the same exit-collapse pattern ($\Delta R^2 = 0.156$ – 0.455), while their respective reference VLMs show only small fluctuations ($\Delta R^2 \leq 0.097$). The probability of observing consistent collapse in two structurally divergent architectures under a null

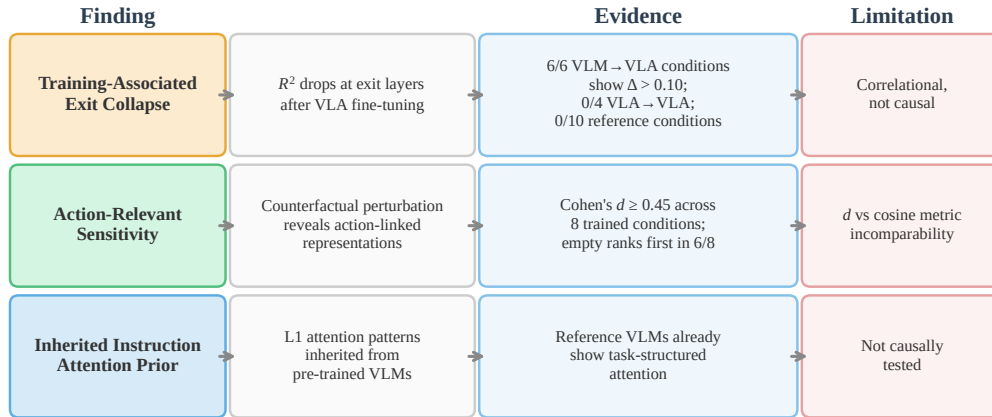


Figure 23: **Claim–evidence–scope summary.** Each main finding is paired with its supporting analysis and its main limitation. The figure distinguishes directly observed representational patterns from stronger causal interpretations that require interventional or training-dynamics evidence.

947 hypothesis of random association is low (permutation test $p = 0.008$; §A.19). Structural diversity
 948 thus provides evidential weight beyond what the raw $N = 2$ count suggests.

949 A.39 Convergent Evidence Summary

950 Three methodologically distinct approaches triangulate instruction routing in VLAs. Table 34 sum-
 951 marizes what each method measures, its key finding, evidence type, and acknowledged limitation.

952 The three methods form a complementary chain: probing identifies *where* instruction information
 953 concentrates across layers, perturbation confirms *whether* instruction content functionally affects
 954 behavior, and ablation demonstrates *that* the identified attention structures carry relevant signals.
 955 No method alone supports causal claims about instruction routing mechanisms; their convergence
 956 provides stronger evidence than any single approach.

957 B Use of Generative AI

958 Generative AI tools (Claude, GPT-4) assisted with code implementation and manuscript editing; all
 959 scientific content was produced by the authors.

Table 30: Per-layer action prediction R^2 : training effect across architectures (mean $\pm$ std over 15 seeds). Peak values per condition in **bold**.

Layer	OpenVLA (Llama-2-7B)		TraceVLA (Phi-3-mini)	
	Reference	Trained	Reference	Trained
L0	0.629 $\pm$ 0.016	0.629 $\pm$ 0.016	-0.330 $\pm$ 0.003	-0.330 $\pm$ 0.002
L1	0.539 $\pm$ 0.023	0.625 $\pm$ 0.017	0.429 $\pm$ 0.034	0.447 $\pm$ 0.013
L2	0.517 $\pm$ 0.047	0.622 $\pm$ 0.010	0.458 $\pm$ 0.023	0.522 $\pm$ 0.014
L3	0.448 $\pm$ 0.216	0.624 $\pm$ 0.014	0.498 $\pm$ 0.036	0.579 $\pm$ 0.011
L4	0.496 $\pm$ 0.057	0.628 $\pm$ 0.014	0.572 $\pm$ 0.021	0.592 $\pm$ 0.009
L5	0.475 $\pm$ 0.101	0.627 $\pm$ 0.019	0.573 $\pm$ 0.020	0.602 $\pm$ 0.010
L6	0.422 $\pm$ 0.119	0.631 $\pm$ 0.012	0.578 $\pm$ 0.018	0.609 $\pm$ 0.012
L7	0.459 $\pm$ 0.093	0.632 $\pm$ 0.012	0.569 $\pm$ 0.023	0.616 $\pm$ 0.014
L8	0.347 $\pm$ 0.283	0.635 $\pm$ 0.013	0.560 $\pm$ 0.025	0.615 $\pm$ 0.010
L9	0.445 $\pm$ 0.127	0.637 $\pm$ 0.008	0.564 $\pm$ 0.021	0.617 $\pm$ 0.013
L10	0.353 $\pm$ 0.224	0.624 $\pm$ 0.014	0.551 $\pm$ 0.027	0.625 $\pm$ 0.007
L11	0.356 $\pm$ 0.122	0.623 $\pm$ 0.016	0.561 $\pm$ 0.025	0.634 $\pm$ 0.009
L12	0.409 $\pm$ 0.203	0.626 $\pm$ 0.011	0.557 $\pm$ 0.026	0.632 $\pm$ 0.005
L13	0.341 $\pm$ 0.227	0.624 $\pm$ 0.013	0.547 $\pm$ 0.028	0.633 $\pm$ 0.008
L14	0.374 $\pm$ 0.211	0.623 $\pm$ 0.015	0.550 $\pm$ 0.023	0.631 $\pm$ 0.009
L15	0.297 $\pm$ 0.225	0.619 $\pm$ 0.012	0.550 $\pm$ 0.022	0.633 $\pm$ 0.010
L16	0.323 $\pm$ 0.226	0.619 $\pm$ 0.011	0.534 $\pm$ 0.018	0.631 $\pm$ 0.010
L17	0.395 $\pm$ 0.109	0.609 $\pm$ 0.023	0.519 $\pm$ 0.024	0.633 $\pm$ 0.007
L18	0.214 $\pm$ 0.307	0.613 $\pm$ 0.017	0.502 $\pm$ 0.047	0.629 $\pm$ 0.012
L19	0.241 $\pm$ 0.258	0.613 $\pm$ 0.013	0.499 $\pm$ 0.058	0.627 $\pm$ 0.008
L20	0.086 $\pm$ 0.311	0.611 $\pm$ 0.014	0.482 $\pm$ 0.048	0.628 $\pm$ 0.011
L21	0.330 $\pm$ 0.216	0.589 $\pm$ 0.023	0.476 $\pm$ 0.036	0.624 $\pm$ 0.007
L22	0.207 $\pm$ 0.283	0.586 $\pm$ 0.013	0.408 $\pm$ 0.140	0.616 $\pm$ 0.008
L23	0.264 $\pm$ 0.227	0.578 $\pm$ 0.028	0.455 $\pm$ 0.074	0.613 $\pm$ 0.007
L24	0.173 $\pm$ 0.275	0.577 $\pm$ 0.013	0.446 $\pm$ 0.062	0.610 $\pm$ 0.008
L25	0.130 $\pm$ 0.305	0.553 $\pm$ 0.053	0.479 $\pm$ 0.039	0.609 $\pm$ 0.010
L26	0.251 $\pm$ 0.254	0.535 $\pm$ 0.046	0.411 $\pm$ 0.098	0.611 $\pm$ 0.008
L27	0.099 $\pm$ 0.278	0.552 $\pm$ 0.027	0.366 $\pm$ 0.197	0.604 $\pm$ 0.011
L28	0.099 $\pm$ 0.307	0.544 $\pm$ 0.034	0.358 $\pm$ 0.218	0.596 $\pm$ 0.015
L29	0.088 $\pm$ 0.275	0.530 $\pm$ 0.034	0.356 $\pm$ 0.271	0.585 $\pm$ 0.010
L30	0.220 $\pm$ 0.234	0.511 $\pm$ 0.029	0.424 $\pm$ 0.104	0.574 $\pm$ 0.008
L31	0.069 $\pm$ 0.284	0.488 $\pm$ 0.054	0.332 $\pm$ 0.262	0.557 $\pm$ 0.016
L32	0.584 $\pm$ 0.011	0.535 $\pm$ 0.015	0.410 $\pm$ 0.185	0.550 $\pm$ 0.007

Table 31: **Methodological precedent.** Representative interpretability studies and their use of behavioral evaluation. Partial behavioral eval includes correlational or indirect evidence.

Paper	Domain	Method	Behav. eval?
Belinkov [46]	NLP	Probing survey	No
Olsson et al. [23]	LM	Induction heads	No
Wang et al. [24]	LM	IOI circuit	No
Meng et al. [26]	LM	Causal tracing (localization)	No
Grant et al. [5]	VLA	Visual dominance	Partial
Haon et al. [6]	VLA	Sparse directions	No
Ours	VLA	ΔR^2 + attn. profiling	Partial ($\rho=0.583$)

Table 32: **NLP finetuning control comparison.** All three models are instruction-tuned for conversational use without action heads or robot demonstration data. None exhibits exit collapse.

Model	VLA backbone	ΔR^2	Mean R^2	Layers
Llama-2-7b-chat	OpenVLA (Llama-2-7B)	0.000	0.155 ± 0.001	32
Phi-3-mini-4k-instruct	TraceVLA (Phi-3-mini)	0.000	0.155 ± 0.000	32
Qwen2.5-3B-Instruct*	NORA (Qwen2.5-VL-3B)	0.000	0.155 ± 0.000	36

*Qwen2.5-3B-Instruct shares the Qwen2.5 language backbone with NORA’s Qwen2.5-VL-3B but lacks the vision encoder and cross-attention modules.

Table 33: **Structural diversity between the two VLM-to-VLA transitions.** The two architectures differ on nearly every major design axis, strengthening the generality of shared findings.

Design Axis	TraceVLA	NORA
Language backbone	Phi-3-mini (3.8B)	Qwen2.5-VL (3B)
Backbone family	Microsoft	Alibaba
Attention mechanism	Multi-Head Attention	Grouped-Query Attention
Number of layers	32	36
Vision encoder	SigLIP (shared w/ base)	Native ViT (dynamic resolution)
Vision integration	Projector (late fusion)	Token concatenation (native)
Pretraining data	English-centric web	Multilingual web + code
Action head design	Linear projection	MLP
Training data	OXE + LIBERO	OXE + LIBERO + custom

Table 34: **Convergent evidence across three analytical methods.** Each method addresses a different aspect of instruction routing; no single method is sufficient alone, but their convergence strengthens the overall characterization.

Method	What it measures	Key finding	Evidence type	Limitation
Layer-wise probing (§4.3)	Where action-relevant information is decodable	VLM-to-VLA: peak-to-exit collapse ($\Delta R^2 = 0.16\text{--}0.46$); VLA-to-VLA: flat	Correlational (decodability)	Does not establish causal necessity
Counterfactual perturbation (§4.4)	Whether instruction content affects model outputs	All trained VLAs: graded divergence ($d = 0.45\text{--}1.52$) ordered by perturbation severity	Behavioral (input-output)	Does not localize which layers are responsible
Attention head ablation (Appendix A.12)	Whether identified structures carry functional signals	Dose-response: monotonic divergence ($d = 0.48\text{--}0.64$) upon progressive head removal	Interventional (causal contribution)	Single architecture (OpenVLA); contribution $\neq$ necessity