Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Action-aligned Representation Geometry in Vision–Language–Action Models

Han Qi

hqi@g.harvard.edu

Harvard University https://orcid.org/0009-0004-5788-2603

Huangyuan Su

Harvard University

Liuyixin Shao

Harvard University

Na Li

Harvard University

Heng Yang

Harvard University

Article

Keywords: Vision–Language–Action (VLA) models, embodied AI, neural collapse, latent geometry, robotic manipulation, representation learning, transformer models, action tokenization, robot learning

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10286824/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Action-aligned Representation Geometry in Vision–Language–Action Models

Han Qi[1*], Huangyuan Su[1], Liuyixin Shao[1], Na Li[1], Heng Yang[1]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA.

*Corresponding author(s). E-mail(s): hqi@g.harvard.edu;
Contributing authors: csu@g.harvard.edu; lshao@g.harvard.edu;
nali@seas.harvard.edu; hankyang@g.harvard.edu;

Abstract

Vision–Language–Action (VLA) models have achieved remarkable success in robotic manipulation, yet little is understood about how they organize multimodal representations to support action prediction. Here we investigate the latent geometry of transformer-based VLA models through the lens of neural collapse (NC). Across diverse architectures, datasets, and real-world robotic tasks, we find that VLA models consistently develop structured, action-aligned latent geometry during training. The strength of this geometry tracks task performance, emerges progressively across transformer layers, reflects hierarchical action semantics; and is functionally involved in action prediction: disrupting it impairs policy performance, whereas preserving it largely maintains performance. We further show that larger training datasets are accompanied by increasingly structured latent geometry, providing a representation-level perspective on the benefits of data scaling. Together, these results establish structured, action-aligned latent geometry as an organizing principle of representation learning in VLA models, providing an interpretable way for action prediction, model scaling, and generalization.

Keywords: Vision–Language–Action (VLA) models, embodied AI, neural collapse, latent geometry, robotic manipulation, representation learning, transformer models, action tokenization, robot learning

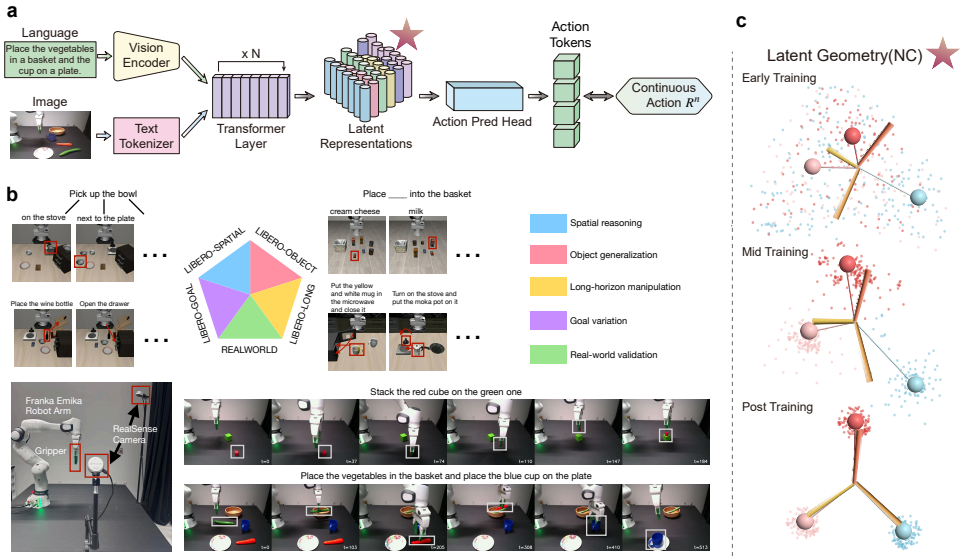


Fig. 1 Latent geometry in vision–language–action (VLA) models. a, Overview of the VLA policy-learning framework. Language instructions and visual observations are encoded by a transformer into latent representations, mapped to action tokens by an action-prediction head, and decoded into continuous robot actions through the corresponding action-tokenization scheme. **b,** Robotic manipulation settings analyzed in this study. LIBERO provides four simulation suites probing spatial reasoning, object generalization, goal-conditioned manipulation, and long-horizon planning; real-world experiments evaluate block stacking and table bussing with a Franka Panda robot. **c,** Schematic of neural-collapse (NC) geometry emerging during training. As optimization proceeds, latent representations associated with the same action token become more compact, while mean vectors from each representation group organize into a separated global configuration consistent with nearest-class-center decision making.

1 Introduction

Behavior cloning has become a dominant paradigm for embodied AI, achieving remarkable success in applications ranging from autonomous driving [1] to robotic manipulation [2–5]. Recent Vision–Language–Action (VLA) foundation models, including π-series [2], OpenVLA [6], and Octo [7], further extend this paradigm by learning robot policies directly from large-scale multimodal demonstrations. These models can execute diverse manipulation tasks from natural language instructions, yet they remain largely opaque: despite their impressive capabilities, little is understood about how multimodal observations are internally organized to support action prediction and generalization. This lack of mechanistic understanding limits our ability to diagnose failures, interpret learned behaviors, and systematically improve VLA models.

A fundamental question is therefore whether the latent representations learned by VLA models exhibit intrinsic geometric organization. If VLA models generalize across tasks, environments, and objects, their internal representations must capture task-relevant structure that links high-dimensional visual and linguistic observations to control actions. Understanding this latent geometry could provide a principled perspective on how modern robotic foundation models learn, generalize, and scale.

However, characterizing such geometry is challenging because VLA models operate on multimodal inputs, heterogeneous datasets, and continuous action spaces that are discretized through diverse action-tokenization strategies. Existing studies provide only partial answers: prior work has primarily focused on simplified control settings [8], probing analyses reveal only limited aspects of learned representations [9, 10], and representation alignment methods [11–14] do not explain how geometric structure naturally emerges during end-to-end policy learning.

The search for structured latent representations is not unique to robotics. Across biological neural systems, population activity is organized into geometric representations that support abstraction, efficient readout, and generalization across diverse sensory experiences [15–21]. Although biological and artificial systems differ substantially, these observations suggest a broader computational principle: intelligent systems may organize high-dimensional observations into structured latent geometries that facilitate decision making. This perspective naturally raises the question of whether analogous geometric organization also emerges in modern VLA models.

A natural framework for studying this question is *neural collapse* (NC) [22], a phenomenon originally discovered in deep classification networks. Neural collapse describes the emergence of highly structured latent geometry, characterized by within-class feature collapse, globally symmetric class means, alignment between representations and classifier weights, and nearest-class-center prediction. While NC originates from image classifcation [23–27], whether similar geometric principles govern representation learning in embodied control remains unknown.

Extending NC to VLA models is far from straightforward. Unlike conventional image classification, VLA models employ autoregressive transformer architectures to predict action tokens from multimodal observations during behavior cloning rather than fixed semantic labels. These action tokens are induced from continuous control trajectories through different tokenization strategies, exhibit substantial class imbalance, and are shared across diverse visual scenes, language instructions, and task contexts. These characteristics fundamentally distinguish VLA training from conventional classification, making it unclear whether structured latent geometry should emerge at all, let alone whether it should play a functional role in action prediction and generalization.

In this work, we investigate the latent geometry of transformer-based Vision–Language–Action models through the lens of neural collapse. Across diverse architectures, action tokenization strategies, datasets, and real-world robotic tasks, we find that VLA models consistently develop highly structured action-aligned latent geometry during training. This geometry strengthens as models improve, emerges progressively across transformer layers, and reflects hierarchical action semantics. Beyond establishing this phenomenon empirically, we demonstrate that the learned geometry is functionally involved in action prediction: perturbations that selectively disrupt the geometry impair prediction, whereas geometry-preserving interventions largely maintain performance. Finally, we show that improvements from data scaling are accompanied by increasingly structured latent geometry, suggesting that representation geometry provides a compact and interpretable lens for understanding learning dynamics, generalization, and scaling in VLA models.

Together, these results identify a common geometric phenomenon underlying representation learning in modern VLA models. By revealing the action-aligned representation space of VLA models, our work provides a geometric perspective on how multimodal observations are transformed into robot actions and offers a principled framework for analyzing, interpreting, and revealing the structure of the representation space supporting effective policies. An overview of the problem setting and key findings is shown in Fig. 1.

2 Methods

2.1 Vision–Language–Action models

We study transformer-based Vision–Language–Action (VLA) models that learn robot control policies from large-scale multimodal demonstrations through behavior cloning [2, 6, 7]. At each time step t, the policy receives a language instruction l_t together with recent visual observations $o_{t-H+1:t}$ and predicts either a single action a_t or a short action sequence $a_{t:t+T}$. Visual observations consist of RGB images from one or more cameras, while actions are continuous robot commands represented as end-effector poses and gripper states.

Although robot actions are continuous, most transformer-based VLA models formulate policy learning as autoregressive next-token prediction by discretizing actions into *action tokens*. Given an input sequence $x_{1:t}$ containing visual, language, and previously generated action tokens, the model predicts the next action token as

$$p(x_{t+1} \mid x_{1:t}) = \mathrm{softmax}(W h(x_{1:t})), \tag{1}$$

where $h(\cdot)$ denotes the latent representation immediately before the action prediction head and W is the output projection. The model is trained using the standard behavior-cloning objective

$$\mathcal{L} = -\sum_t \log p(x_{t+1} \mid x_{1:t}). \tag{2}$$

Although existing VLA models share this general formulation, they differ substantially in architecture and action tokenization. We evaluate four representative approaches spanning the major design choices in current VLA systems: OpenVLA [6], π_0-FAST [2, 28], BeT [29], and VQ-BeT [30]. These models differ in language backbones, visual encoders, and action-token representations, providing diverse settings for studying latent geometry.

Action tokenization.

A key distinction among VLA models lies in how continuous robot actions are converted into discrete tokens (Fig. 2a). We consider four representative tokenization strategies.

Uniform quantization, used in OpenVLA [6], independently discretizes each action dimension into fixed bins. This strategy is simple and preserves the structure of individual action dimensions.

Joint quantization, used in BeT [29], clusters the entire action vector using k-means and represents each action by its nearest cluster centroid, allowing dependencies between action dimensions to be captured.

Structured tokenization, used in π_0-FAST [28], first transforms short action sequences into a compact temporal representation (e.g., using the discrete cosine transform) before discretization, improving temporal efficiency for long action horizons.

Learned tokenization, used in VQ-BeT [30], employs vector quantization to learn a discrete codebook jointly with the policy, enabling data-adaptive action representations.

These tokenization strategies differ substantially in vocabulary size, token distribution, and semantic granularity, providing a broad testbed for evaluating whether latent geometry depends on the particular discretization of the action space.

Datasets and evaluation.

Unless otherwise specified, experiments are conducted on the LIBERO benchmark [31], which comprises four task suites evaluating spatial reasoning (LIBERO-SPATIAL), object manipulation (LIBERO-OBJECT), goal-conditioned manipulation (LIBERO-GOAL), and long-horizon planning (LIBERO-LONG). All models are evaluated using standard task success rate. To study scaling behavior, we additionally evaluate on RoboCasa [32], a large-scale household manipulation benchmark with diverse environments and object configurations. Finally, we validate our findings on real-world block stacking and table-bussing tasks using a 7-DoF Franka Panda robot equipped with two Intel RealSense cameras (Fig. 1b).

2.2 Quantifying latent geometry

Our central question is how transformer-based VLA models organize latent representations to support action prediction. During training, multimodal observations are mapped to discrete action tokens, suggesting that action supervision may progressively shape the geometry of the latent space. To characterize this organization, we analyze the latent representation immediately before the action prediction head using the framework of *neural collapse* (NC) [22].

Although neural collapse was originally discovered in image classification, it provides a natural framework for studying VLA models. From the perspective of behavior cloning, predicting the next action token can be viewed as a large-scale classification problem, where action tokens act as supervision signals that organize the latent space. Unlike conventional classification, however, these action tokens arise from discretizing continuous control trajectories, exhibit substantial class imbalance, and are induced by different tokenization strategies (as in Fig. 2b). Consequently, whether NC-like geometry emerges in VLA models is far from obvious.

Let V denote the action-token vocabulary and let $C = |V|$. For each action token $c \in V$, we compute the mean latent representation

$$\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} h_i, \tag{3}$$

where h_i denotes the latent representation of sample i immediately before the action prediction head and N_c is the number of samples assigned to token c. We further denote the global mean over all latent vectors by $\bar{\mu} = N^{-1} \sum_{i=1}^{N} h_i$ and the within-class variance by σ_c^2. These quantities provide a geometric summary of how different action tokens are organized in latent space.

Following the neural collapse framework, we characterize latent geometry through four complementary properties corresponding to local compactness, global organization, representation–classifier alignment, and geometry-driven prediction.

NC1: Within-class compactness.

NC1 measures whether representations associated with the same action token become increasingly concentrated. We compute the normalized ratio between within-class variation and the separation of class centers,

$$\mathrm{NC}_1 = \mathrm{Mean}_{c \neq c'} \frac{\sigma_c^2 + \sigma_{c'}^2}{2\|\mu_c - \mu_{c'}\|_2^{k+2}}, \tag{4}$$

where $k = 2$ throughout this work. Lower values indicate stronger within-class collapse.

NC2: Global organization.

While NC1 characterizes local clustering, NC2 measures whether action-token centers form a globally regular geometry. Following previous work, we quantify both the variability of class-center norms, denoted NC2-Norm in the figures, and the deviation of pairwise angles from the simplex equiangular tight frame (ETF), denoted NC2-Angle,

$$\mathrm{NC}_{2,\mathrm{norm}} = \mathrm{CoV}\left(\log \|\mu_c - \bar{\mu}\|_2\right), \tag{5}$$

$$\mathrm{NC}_{2,\mathrm{angle}} = \left| \mathrm{Mean}_{c \neq c'} \left(\langle \hat{\mu}_c, \hat{\mu}_{c'} \rangle + \frac{1}{C-1} \right) \right|, \tag{6}$$

where $\hat{\mu}_c = (\mu_c - \bar{\mu})/\|\mu_c - \bar{\mu}\|_2$ and CoV denotes the coefficient of variation. Lower values indicate stronger agreement with the ideal neural-collapse geometry. Because VLA datasets exhibit substantial action-token imbalance, we additionally report a relaxed hyperspherical uniformity metric, denoted GNC_2 in the text and GNC2 in the figures, in the Supplementary Information.

NC3: Representation–classifier alignment.

NC3 measures whether the learned latent geometry aligns with the directions used for action prediction, corresponding to the self-duality metric shown in the Fig. 1c. Specifically, we compute the average distance between normalized class centers and normalized classifier weights,

$$\mathrm{NC}_3 = \mathrm{Mean}_c \left\| \hat{\mu}_c - \hat{w}_c \right\|_2, \tag{7}$$

where $\hat{w}_c$ denotes the normalized classifier weight corresponding to action token c. Lower values indicate stronger alignment between representation geometry and the prediction head.

NC4: Geometry-driven prediction.

Finally, NC4 measures whether action prediction can be explained directly by the learned geometry, corresponding to the classifier-agreement metric shown in the figures. Specifically, we compare the model prediction with the nearest-class-center prediction,

$$\mathrm{NC}_4 = \Pr\left[\arg\max_c w_c^\top h = \arg\max_c \left(2h^\top \mu_c - \|\mu_c\|_2^2\right)\right]. \tag{8}$$

Higher NC4 indicates that action prediction increasingly follows a nearest-class-center decision rule.

Together, these complementary metrics characterize how VLA models organize latent representations during training: NC1 captures local compactness; NC2-Norm, NC2-Angle, and GNC_2 describe global geometric organization; NC3 measures the alignment between representations and the prediction head; and NC4 evaluates whether the learned geometry directly explains action prediction. This framework provides a unified geometric lens for analyzing representation learning across different VLA architectures, action tokenization strategies, and robotic tasks.

2.3 Why action-token prediction may induce neural-collapse geometry

Although VLA models operate on multimodal observations and continuous control signals, they are trained by predicting discrete action tokens. Consequently, latent representations are optimized not only to encode observations but also to organize them according to the target actions. We hypothesize that this optimization naturally encourages neural-collapse (NC)-like geometry in the latent space.

For a latent representation h immediately before the action prediction head, the gradient of the cross-entropy loss is

$$\nabla_h \mathcal{L} = \sum_{c \in V} p_c w_c - w_y, \tag{9}$$

where V is the action-token vocabulary, w_c denotes the row of W corresponding to action token c, p_c is its predicted probability, and y is the target action token.

This gradient has three geometric consequences. First, representations sharing the same action token are repeatedly attracted toward a common classifier direction, producing increasingly compact within-class clusters (NC1). Second, competing action tokens exert repulsive forces that enlarge the separation between class centers, leading to a more regular global organization (NC2-Norm, NC2-Angle, and GNC_2). Third, because classifier weights and latent representations are optimized jointly, the two become progressively aligned (NC3). Together, these effects naturally give rise to nearest-class-center prediction (NC4).

Although similar optimization dynamics have been analyzed theoretically in simplified classification settings, where neural collapse can be shown to emerge under idealized assumptions such as balanced classes, linear classifiers, or simplified feature-learning models [26, 33, 34], these assumptions do not directly apply to modern VLA models. However, these theoretical settings differ substantially from modern VLA models. VLA policies employ large transformer architectures operating over multi-modal observations, predict action tokens induced from continuous control trajectories, and are trained on highly heterogeneous datasets with substantial action-token imbalance. Consequently, the assumptions underlying existing NC theory do not directly apply. Consequently, we do not claim a formal theoretical explanation for neural collapse in VLA models. Instead, the gradient analysis above provides a qualitative intuition for why action-token prediction may encourage NC-like geometry, which we investigate empirically in the following sections.

3 Results

3.1 Neural-collapse geometry consistently emerges in VLA models

We first ask whether neural-collapse (NC) geometry consistently emerges during VLA training. Unlike semantic labels in image classification, action tokens are obtained by discretizing continuous control trajectories and therefore exhibit substantial variation in vocabulary size, token frequency, and class imbalance. As shown in Fig. 2, π_0-FAST produces hundreds to thousands of action tokens, whereas BeT and VQ-BeT typically employ fewer than one hundred, with complete token-count statistics across LIBERO suites reported in Extended Data Fig. 1. These markedly different token distributions provide a stringent test of whether the observed geometry reflects a particular tokenization strategy or a more general consequence of action-token prediction.

Simulation benchmarks.

We begin by analyzing the evolution of latent geometry during fine-tuning on the four LIBERO task suites (LIBERO-SPATIAL, LIBERO-OBJECT, LIBERO-GOAL, and LIBERO-LONG). Approximately 30 checkpoints are collected throughout training, allowing us to jointly track task success and NC metrics.

Fig. 3(a) shows a representative example on LIBERO-SPATIAL. Across all architectures, improved task performance is consistently accompanied by stronger neural-collapse geometry: NC1, NC2-Norm, NC2-Angle, GNC_2, and NC3 decrease, whereas NC4 increases. The same trends are observed across all LIBERO task suites and VLA models (Extended Data Fig. 2), demonstrating that the emergence of NC geometry is robust to both model architecture and action-tokenization strategy.

The NC metrics reveal a coherent progression of representation learning. During optimization, representations associated with the same action token become increasingly compact (NC1), action-token centers organize into a more regular global geometry (NC2-Norm, NC2-Angle, and GNC_2), classifier weights progressively align with class centers (NC3), and prediction approaches a nearest-class-center rule (NC4). Although exact simplex equiangular tight frame (ETF) geometry is not expected

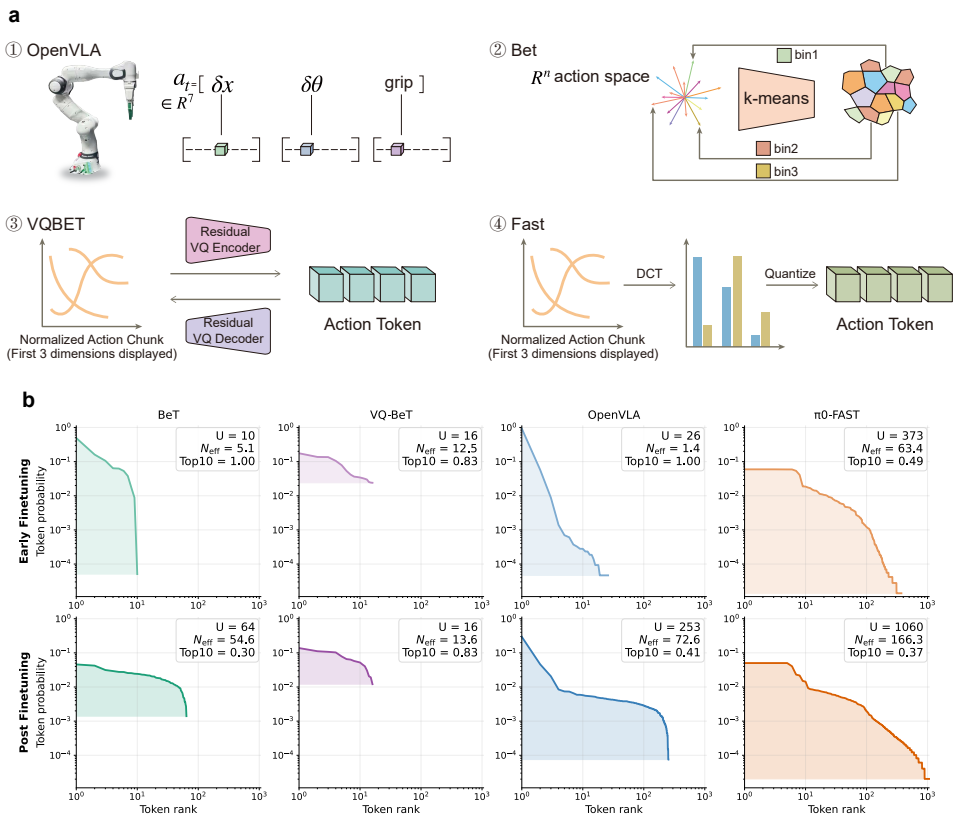


Fig. 2 Action tokenization induces diverse action-token distributions across VLA models. a, Representative action-tokenization strategies used by OpenVLA [6], BeT [29], VQ-BeT [35], and π_0-FAST [2]. Continuous robot actions are discretized through independent quantization, vector clustering, learned vector quantization, or frequency-domain trajectory tokenization, respectively. **b,** Action-token frequency distributions before (top) and after (bottom) fine-tuning on LIBERO-SPATIAL. Frequencies are plotted against token rank on logarithmic axes, showing substantial differences in vocabulary size, token imbalance, and token usage across models and training stages.

because of severe action-token imbalance, GNC_2 exhibits the same monotonic trend, indicating increasingly uniform organization on the hypersphere.

The correlation analysis in Fig. 3(b) further confirms these observations. Across all datasets and architectures, NC1, NC2-Norm, NC2-Angle, GNC_2, and NC3 are strongly negatively correlated with task success, whereas NC4 is positively correlated. Moreover, more challenging tasks exhibit systematically weaker geometry. In particular, LIBERO-LONG achieves lower success rates than the other task suites and correspondingly exhibits weaker within-class compactness (NC1) and nearest-class-center consistency (NC4) (Extended Data Fig. 3(b)). These observations suggest that increasingly structured latent geometry accompanies successful policy learning across diverse manipulation tasks.

Real-world robotic manipulation.

We next investigate whether similar geometric organization emerges outside simulation. OpenVLA and π_0-FAST are fine-tuned on real-world block-stacking and table-bussing tasks using a Franka Panda robot (Fig. 3(c) and Extended Data Fig. 4). Because validation success is unavailable during training, we compare NC metrics with training loss.

The same geometric trends are consistently observed: as training loss decreases, NC1, NC2-Norm, NC2-Angle, GNC_2, and NC3 decrease while NC4 increases, closely mirroring the behavior observed on LIBERO. Despite substantially smaller datasets and different sensing conditions, latent representations progressively become more compact, better organized, and increasingly aligned with the action prediction head. These results demonstrate that the emergence of neural-collapse geometry is not restricted to simulation benchmarks but also accompanies learning in real-world robotic systems.

Emergence across transformer layers.

Neural collapse is typically characterized using the final representation layer, but its formation throughout the network remains largely unexplored. We therefore evaluate the representation-only metrics NC1, NC2-Norm, NC2-Angle, and GNC_2 across intermediate transformer layers in OpenVLA and π_0-FAST. Unlike these metrics, NC3 and NC4 are defined through the final linear action-prediction head, requiring direct alignment between latent representations and the output classifier. They are therefore only well-defined for the final representation layer. As shown in Extended Data Fig. 3(a), latent geometry develops progressively with network depth: early layers exhibit weak organization, whereas deeper layers display increasingly compact within-class representations and more regular global geometry. Interestingly, strong NC structure starts to appear in intermediate layers for OpenVLA but in the deepest few layers for π_0-FAST, suggesting that different architectures organize latent representations through distinct layerwise dynamics.

Control experiments.

Finally, we verify that the observed geometry is specifically induced by action-token supervision rather than generic clustering. We replace action-token labels with random labels, frequency-matched random labels, and task-level labels. As shown in Extended Data Fig. 3(c), none of these alternative labeling schemes produces meaningful geometric organization. Furthermore, clustering directly in the latent space using k-means yields only marginal improvements over the learned action-token clusters, indicating that the representation geometry learned through action-token prediction is already close to the intrinsic clustering structure of the latent space.

Taken together, these results demonstrate that neural-collapse geometry consistently emerges during VLA training across diverse architectures, action-tokenization strategies, transformer layers, simulation benchmarks, and real-world robotic tasks. This robust emergence motivates the following sections, where we investigate what this geometry represents and whether it plays a functional role in action prediction.

3.2 The learned geometry reflects action semantics

Having established that neural-collapse geometry consistently emerges during VLA training, we next ask what this geometry represents. If the learned latent space is organized by action semantics rather than merely discrete token identities, then geometrically nearby representations should correspond to semantically related actions. We investigate this question using two complementary action-tokenization schemes that expose different notions of action similarity.

Hierarchical organization in VQ-BeT.

VQ-BeT employs a hierarchical action representation in which coarse-level tokens describe high-level action patterns and fine-level tokens capture detailed variations. This structure provides a natural test of whether the learned latent geometry reflects the hierarchy of action semantics.

As shown in Fig. 4(a,b), the latent space develops a clear hierarchical organization during training. Representations first cluster according to coarse-level action tokens, and progressively separate into finer subclusters corresponding to more detailed action variations. This organization is evident both from the NC metrics and from PCA visualizations of the latent space. Similar hierarchical structures are consistently observed across different datasets (Extended Data Fig. 5 and Extended Data Fig. 6), indicating that the learned geometry naturally reflects the hierarchical organization imposed by the action tokenizer.

Semantic organization in OpenVLA.

Unlike VQ-BeT, OpenVLA discretizes each action dimension independently, resulting in action tokens with explicit semantic relationships—nearby token indices correspond to similar continuous motions. We therefore ask whether these semantic relationships are preserved in the learned latent geometry.

To this end, we progressively merge fine-grained action tokens into coarser semantic groups (e.g., from 256 bins to 32 or 16 groups) and examine the resulting latent representations. As shown in Fig. 4(c,d), the latent space exhibits increasingly clear semantic organization throughout training. Early checkpoints display little geometric structure, whereas later checkpoints form well-separated clusters that closely follow the hierarchical grouping of action tokens. Class centers simultaneously evolve toward increasingly symmetric configurations, indicating that the learned geometry captures not only discrete token identities but also their underlying semantic relationships. The same behavior is consistently observed across additional datasets (Extended Data Fig. 7 and Extended Data Fig. 8).

Taken together, these results demonstrate that neural-collapse geometry encodes meaningful action semantics rather than merely separating discrete action tokens. Across two fundamentally different tokenization strategies, the learned latent space preserves both hierarchical organization and semantic similarity, suggesting that the observed geometry reflects the underlying structure of the action space itself.

3.3 Latent geometry is functionally involved in action prediction

The previous sections establish that neural-collapse (NC) geometry consistently emerges during VLA training and reflects meaningful action semantics. We next ask whether this geometry plays a functional role in action prediction.

To answer this question, we directly intervene on the learned latent representations using four complementary perturbations. We first evaluate a well-trained policy to obtain the latent representations together with the corresponding class centers and class-mean subspace estimated from the clean latent space. During a second evaluation rollout, latent representations are perturbed online immediately before the action prediction head while all model parameters remain fixed. The resulting policy is then evaluated using standard closed-loop task success. In parallel, we perform an offline representation analysis using the saved latent representations. The same perturbations are applied to the recorded embeddings, after which NC metrics are recomputed using the original action-token labels and action-token prediction accuracy is measured using the perturbed predictions. This protocol enables us to separately evaluate the effect of perturbations on robot behavior, latent geometry, and action-token decoding.

We consider four complementary perturbations targeting different aspects of the learned geometry: (i) norm-preserving isotropic perturbation, which injects Gaussian noise while preserving the norm of each latent representation; (ii) within-class destructive perturbation, which inflates the residual relative to the corresponding class center to selectively weaken NC1; (iii) between-class destructive perturbation, which suppresses the component of the latent representation lying in the class-mean subspace to disrupt the geometry; and (iv) geometry-preserving projection, which moves latent representations toward their corresponding class centers while preserving the overall class-center geometry.

Geometry-destroying versus geometry-preserving perturbations.

Fig. 5 summarizes the effects of the four perturbation strategies on closed-loop task success, action-token prediction, and latent geometry. Perturbations that explicitly disrupt the learned geometry consistently degrade both policy performance and action prediction. Although the three destructive perturbations affect different geometric components, they all weaken the NC structure and lead to progressively poorer performance as the intervention strength increases.

In contrast, geometry-preserving projection maintains nearly identical task success and action prediction while further improving the NC metrics, particularly NC1 and NC4. The close correspondence between changes in NC metrics and changes in policy performance demonstrates that the learned geometry is not merely correlated with action prediction but is functionally involved in it. Together, these intervention experiments provide causal evidence that preserving the learned latent geometry preserves policy behavior, whereas disrupting it consistently impairs both representation quality and control performance.

3.4 Latent geometry provides a representation-level perspective on data scaling

Data scaling is one of the most effective strategies for improving the performance of modern VLA models, yet the representation-level mechanisms underlying these gains remain poorly understood. Given the strong relationship between latent geometry and policy performance established above, we ask whether improvements from data scaling are accompanied by systematic changes in neural-collapse geometry.

To investigate this question, we evaluate π_0-FAST on RoboCasa [32], a large-scale benchmark for household manipulation containing diverse environments, object configurations, and mobile manipulation tasks. For four representative tasks, we fine-tune the model using different fractions of the available demonstrations and evaluate both task success and latent geometry under identical evaluation protocols.

As expected, increasing the amount of training data consistently improves task performance (Fig. 6). More importantly, the learned latent geometry becomes progressively more structured with increasing data scale. To summarize the overall quality of the learned geometry, we aggregate the normalized NC metrics into a single geometry score. This score exhibits a strong monotonic relationship with task success across all data scales, indicating that models trained with more demonstrations not only perform better but also develop increasingly organized latent representations.

These observations suggest that improvements from data scaling are accompanied by the progressive formation of action-aligned latent geometry. Rather than merely increasing data coverage, larger training datasets appear to facilitate the emergence of structured representations that support more accurate and robust action prediction. This result further highlights latent geometry as a compact representation-level indicator of learning dynamics and generalization in Vision–Language–Action models.

4 Discussion

Understanding how Vision–Language–Action (VLA) models organize multimodal observations into executable robot actions is central to explaining their emerging capabilities in robotic manipulation. Our results identify *action-aligned representation geometry* as a robust organizing principle in modern VLA policies. Rather than treating latent representations as opaque intermediate features, we show that they acquire a structured geometry shaped by autoregressive action-token prediction during behavior cloning, strengthens with training, emerges across transformer layers, and is consistently associated with downstream control performance across simulation benchmarks and real-world robotic systems.

Neural collapse provides a useful geometric vocabulary for characterizing learned representations, but the phenomenon studied here differs substantially from its classical setting in supervised image classification. Existing studies primarily consider feedforward classification networks trained to predict fixed semantic labels. In contrast, modern VLA models employ large autoregressive transformer architectures that jointly process visual observations, language instructions, and action histories to predict the next action token during behavior cloning. Moreover, these supervisory targets

are not natural semantic categories but action tokens induced from continuous control trajectories. Their definition depends on the chosen action-tokenization scheme, they are often highly imbalanced, may encode different levels of temporal and motor abstraction, and are shared across diverse visual scenes, language instructions, objects, and task contexts. It is therefore far from obvious that coherent collapse-like geometry should emerge in such a complex multimodal setting, let alone that it should be functionally relevant for control. The consistent geometric organization observed across OpenVLA, BeT, VQ-BeT, and π_0-FAST suggests that autoregressive action prediction exerts a strong representational pressure, organizing multimodal observations into an action-aligned latent geometry despite substantial differences in architecture and action tokenization.

A key finding is that this geometry is not merely a by-product of token separation. The latent space preserves meaningful *action semantics*. For hierarchical tokenization schemes, representations reflect coarse-to-fine structure in the action codebook; for independently discretized action dimensions, geometrically related centers correspond to semantically related continuous motions. Thus, the learned geometry organizes observations according to what the robot should do, rather than simply according to what the robot sees. This distinguishes VLA representation learning from visual semantic classification and points to a different form of abstraction: multimodal inputs that differ in appearance, language, or object configuration can become geometrically aligned when they imply similar robot behavior.

The perturbation experiments further show that action-aligned geometry is functionally tied to policy execution. Interventions that selectively disrupt within-token compactness, between-token organization, or the class-mean subspace degrade action-token prediction and closed-loop task success, whereas geometry-preserving projections largely maintain policy behavior. These results do not provide a complete theory of why the geometry emerges during optimization, but they go beyond correlation: they show that the learned geometric structure is part of the computational substrate used by the policy to select actions. In this sense, latent geometry provides a mechanistic bridge between representation learning and embodied control.

Our scaling experiments suggest that this perspective can also clarify how additional data improves VLA policies. Across RoboCasa tasks, larger training sets lead not only to higher task success but also to more coherent action-aligned geometry. This indicates that data scaling may improve policies in part by enabling the model to organize diverse multimodal experiences into a more stable and separable action space. Latent geometry therefore offers a representation-level view of scaling: it reveals how broader data coverage is internalized by the model, rather than only measuring the behavioral outcome after training.

More broadly, the framework developed here complements standard policy evaluation. Task success measures whether a robot completes a task, but it provides limited insight into how the policy represents the relationship between observations, language, and actions. Action-aligned geometry exposes this internal organization. Because the proposed metrics are architecture-independent, inexpensive to compute, and predictive of downstream behavior, they can serve as diagnostic tools for comparing architectures, evaluating action-tokenization strategies, monitoring training dynamics, and

identifying representation failures before or alongside behavioral evaluation. This may be especially useful as robotic foundation models become larger, more heterogeneous, and more difficult to interpret through task success alone.

Several limitations remain. First, our analysis focuses on transformer-based behavior-cloning policies trained with discrete action-token prediction. Whether analogous action-aligned geometry emerges in diffusion policies, reinforcement-learning agents, or world-model-based architectures remains an important question. Second, although our perturbation results demonstrate a functional link between latent geometry and policy behavior, a full optimization-level theory explaining when and why this geometry emerges is still lacking. Third, our experiments focus on manipulation; extending geometric analyses to broader embodied domains such as autonomous driving and humanoid control may reveal whether action-aligned representation geometry is a general principle of embodied decision making.

Together, these findings suggest that modern Vision–Language–Action models do not simply learn mappings from multimodal observations to robot actions. Instead, they organize observations into a structured action-aligned latent space, where geometrically nearby representations correspond to similar control behaviors and progressively support more effective action prediction. This structured geometry provides a compact, interpretable, and functionally meaningful perspective for understanding how robotic foundation models learn, generalize, and scale.

References

[1] Yang, B., Su, H., Gkanatsios, N., Ke, T.-W., Jain, A., Schneider, J., Fragkiadaki, K.: Diffusion-es: Gradient-free planning with diffusion for autonomous driving and zero-shot instruction following. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024)

[2] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., *et al.*: π_0: A vision-language-action flow model for general robot control. In: Robotics: Science and Systems (RSS) (2025)

[3] Huang, W., Chen, C., Qi, H., Lv, C., Du, Y., Yang, H.: Motvla: A vision-language-action model with unified fast-slow reasoning. arXiv preprint arXiv:2510.18337 (2025)

[4] Qi, H., Yin, H., Zhu, A., Du, Y., Yang, H.: Strengthening generative robot policies through predictive world modeling. arXiv preprint arXiv:2502.00622 (2025)

[5] Qi, H., Chen, C., Yang, H.: Compose by focus: Scene graph-based atomic skills. In: IEEE International Conference on Robotics and Automation (ICRA) (2026)

[6] Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., *et al.*: Openvla: An open-source vision-language-action model. In: Conference on Robot Learning (CoRL) (2025)

[7] Octo Model Team, Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., *et al.*: Octo: An open-source generalist robot policy. In: Robotics: Science and Systems (RSS) (2024)

[8] Qi, H., Yin, H., Yang, H.: Control-oriented clustering of visual latent representation. In: International Conference on Learning Representations (ICLR) (2025)

[9] Lu, H., Li, H., Shahani, P.S., Herbers, S., Scheutz, M.: Probing a vision-language-action model for symbolic states and integration into a cognitive architecture. In: IEEE International Conference on AI and Data Analytics (ICAD), pp. 1–8 (2025). IEEE

[10] Buurmeijer, H., Alonso, C.A., Swann, A., Pavone, M.: Observing and controlling features in vision-language-action models. arXiv preprint arXiv:2603.05487 (2026)

[11] Qi, H., Su, Y., Kumar, A., Levine, S.: Data-driven offline decision-making via invariant representation learning. In: Conference on Neural Information Processing Systems (NeurIPS), vol. 35, pp. 33803–33817 (2022)

[12] Bai, S., Lyu, J., Zhou, W., Li, Z., Wang, D., Xing, L., Zhao, X., Wang, P., Wang, Z., Chi, C., et al.: Latent reasoning vla: Latent thinking and prediction for vision-language-action models. arXiv preprint arXiv:2602.01166 (2026)

[13] Jiang, G., Liang, Y., Ye, J., Huang, J.-Y., Jing, C., Duan, R., Abbeel, P., Wang, X., Zou, X.: Cross-hand latent representation for vision-language-action models. arXiv preprint arXiv:2603.10158 (2026)

[14] Zhang, Z., Chen, X.-H., Wang, Y., Sun, Y., Luo, W., Ren, H., Lin, H., Yu, Y.: On the representation degradation in vision-language-action models. OpenReview preprint (2026)

[15] Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., Fried, I.: Invariant visual representation by single neurons in the human brain. Nature **435**(7045), 1102–1107 (2005)

[16] Kriegeskorte, N., Mur, M., Bandettini, P.A.: Representational similarity analysis–connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience **2**, 4 (2008)

[17] Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., Harris, K.D.: High-dimensional geometry of population responses in visual cortex. Nature **571**(7765), 361–365 (2019)

[18] Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., Moser, E.I.: Microstructure of a spatial map in the entorhinal cortex. Nature **436**(7052), 801–806 (2005)

[19] Constantinescu, A.O., O'Reilly, J.X., Behrens, T.E.: Organizing conceptual knowledge in humans with a gridlike code. Science **352**(6292), 1464–1468 (2016)

[20] Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., Fusi, S.: The importance of mixed selectivity in complex cognitive tasks. Nature **497**(7451), 585–590 (2013)

[21] Bernardi, S., Benna, M.K., Rigotti, M., Munuera, J., Fusi, S., Salzman, C.D.: The geometry of abstraction in the hippocampus and prefrontal cortex. Cell **183**(4), 954–967 (2020)

[22] Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. In: International Conference on Learning Representations (ICLR) (2025)

[23] Jiang, J., Zhou, J., Wang, P., Qu, Q., Mixon, D.G., You, C., Zhu, Z.: Generalized neural collapse for a large number of classes. In: International Conference on Machine Learning (ICML), pp. 22010–22041 (2024). PMLR

[24] Wu, R., Papyan, V.: Linguistic collapse: Neural collapse in (large) language models. In: Conference on Neural Information Processing Systems (NeurIPS) (2021)

[25] Rangamani, A., Lindegaard, M., Galanti, T., Poggio, T.A.: Feature learning in deep classifiers through intermediate neural collapse. In: International Conference on Machine Learning (ICML), pp. 28729–28745 (2023). PMLR

[26] Fang, C., He, H., Long, Q., Su, W.J.: Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. Proceedings of the National Academy of Sciences **118**(43), 2103091118 (2021)

[27] Han, X., Papyan, V., Donoho, D.L.: Neural collapse under mse loss: Proximity to and dynamics on the central path. arXiv preprint arXiv:2106.02073 (2021)

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

[29] Shafiullah, N.M., Cui, Z., Altanzaya, A.A., Pinto, L.: Behavior transformers: Cloning k modes with one stone. In: Conference on Neural Information Processing Systems (NeurIPS), vol. 35, pp. 22955–22968 (2022)

[30] Lee, S., Wang, Y., Etukuru, H., Kim, H.J., Shafiullah, N.M.M., Pinto, L.: Behavior generation with latent actions. In: International Conference on Machine Learning (ICML), pp. 26991–27008 (2024). PMLR

[31] Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. In: Conference on Neural Information Processing Systems (NeurIPS) (2023)

[32] Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., Zhu, Y.: Robocasa: Large-scale simulation of household tasks for generalist robots. In: Robotics: Science and Systems (RSS) (2024)

[33] Mixon, D.G., Parshall, H., Pi, J.: Neural collapse with unconstrained features. Sampling Theory, Signal Processing, and Data Analysis **20**(2), 11 (2022)

[34] Tirer, T., Bruna, J.: Extended unconstrained features model for exploring deep neural collapse. In: International Conference on Machine Learning, pp. 21478–21505 (2022). PMLR

[35] Haldar, S., Peng, Z., Pinto, L.: Baku: An efficient transformer for multi-task policy learning, 2024. In: Conference on Neural Information Processing Systems (NeurIPS) (2024)

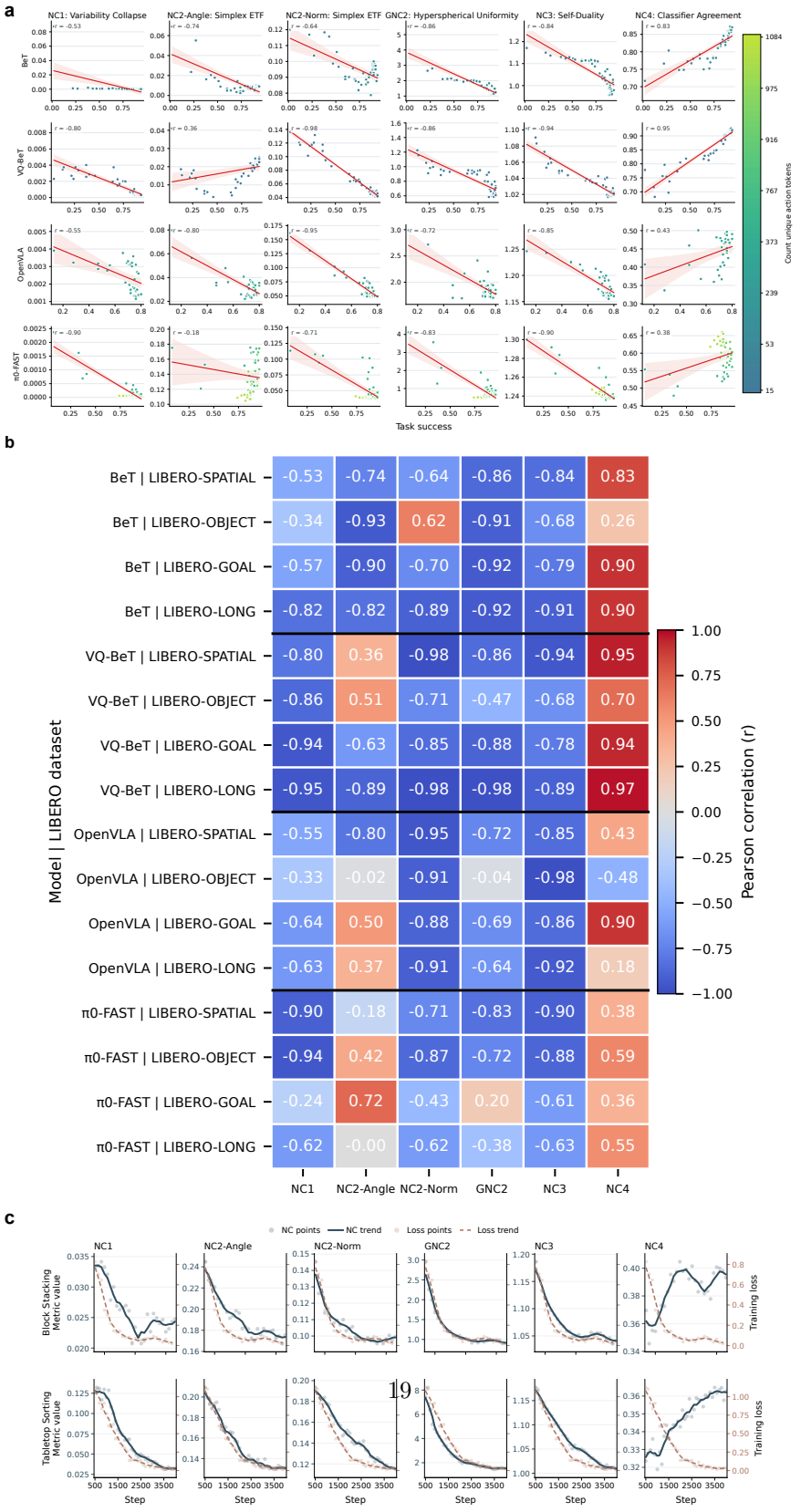


Fig. 3 Neural-collapse geometry strengthens as VLA task performance improves. a, Relationship between NC metrics and task success during fine-tuning on LIBERO. Each point denotes a model checkpoint, coloured by the number of unique action tokens; solid lines denote linear fits. Higher task success is accompanied by stronger NC geometry, reflected by lower NC1, NC2-Norm, NC2-Angle, GNC_2, and NC3, and by higher NC4. **b,** Pearson correlations between NC metrics and task success across VLA architectures and LIBERO task suites. NC1, NC2-Norm, NC2-Angle, GNC_2, and NC3 are consistently negatively correlated with performance, whereas NC4 is positively correlated with performance, indicating that stronger latent geometric organization accompanies better policy generalization. **c,** Evolution of NC metrics during fine-tuning on real-world manipulation tasks. As training loss decreases, latent representations progressively develop stronger NC geometry, showing that structured latent representations also emerge outside simulation.

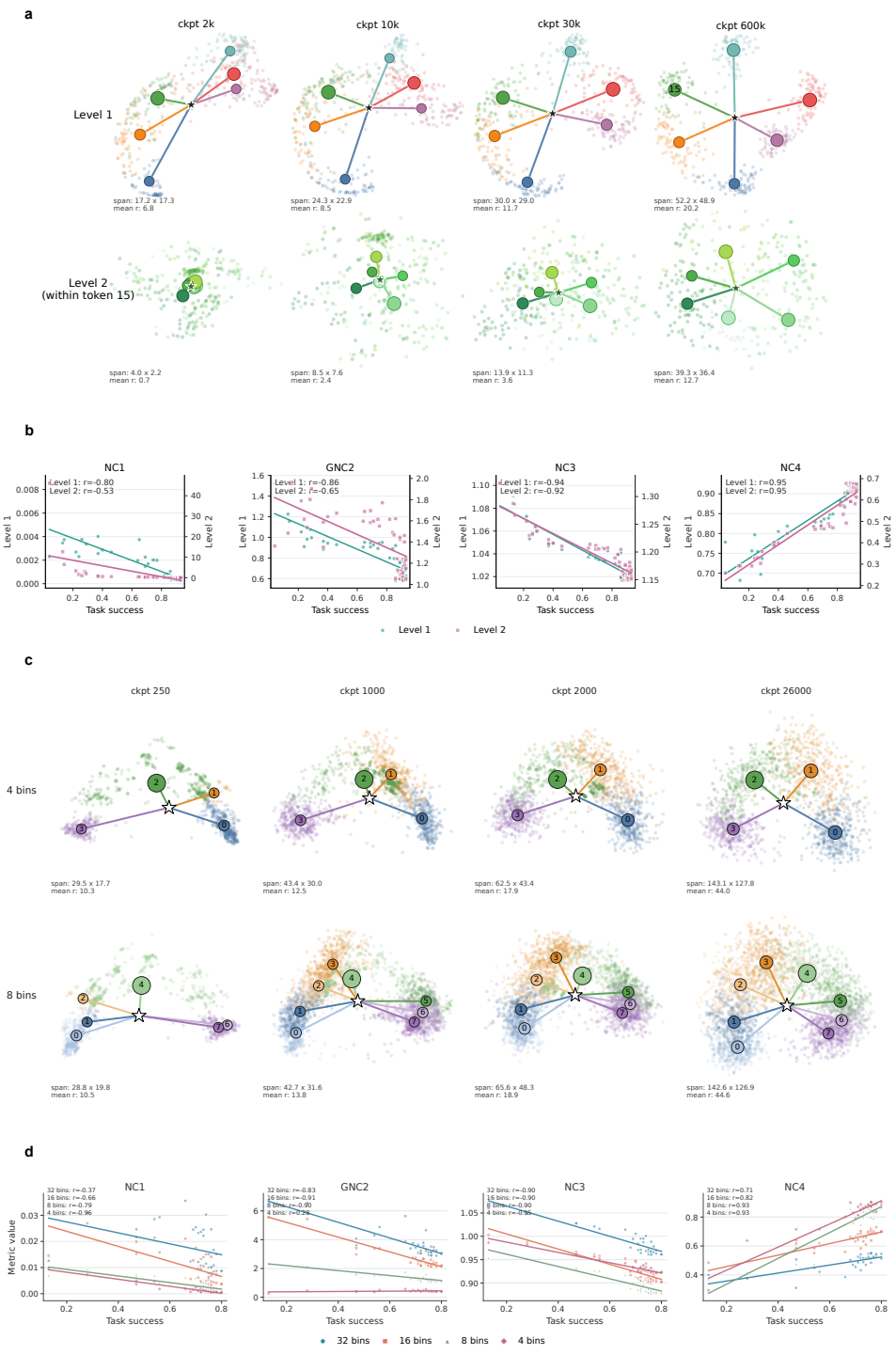


Fig. 4 Latent geometry reflects the hierarchical organization of action semantics. a, Evolution of VQ-BeT latent representations during training. Coarse- and fine-level action groups become progressively more compact and better separated, revealing a hierarchical organization of the latent space. **b,** Relationship between NC metrics and task success for coarse- and fine-level VQ-BeT action tokens. Consistent trends across both hierarchy levels indicate that increasingly structured latent geometry is associated with improved policy performance. **c,** Evolution of OpenVLA latent representations under different action-token grouping granularities. During training, initially weakly organized representations develop into separated clusters aligned with coarser semantic action groups. **d,** NC metrics across token-grouping granularities. Coarser groupings exhibit stronger neural-collapse geometry, consistent with a hierarchy in which global semantic structure emerges before finer action distinctions.

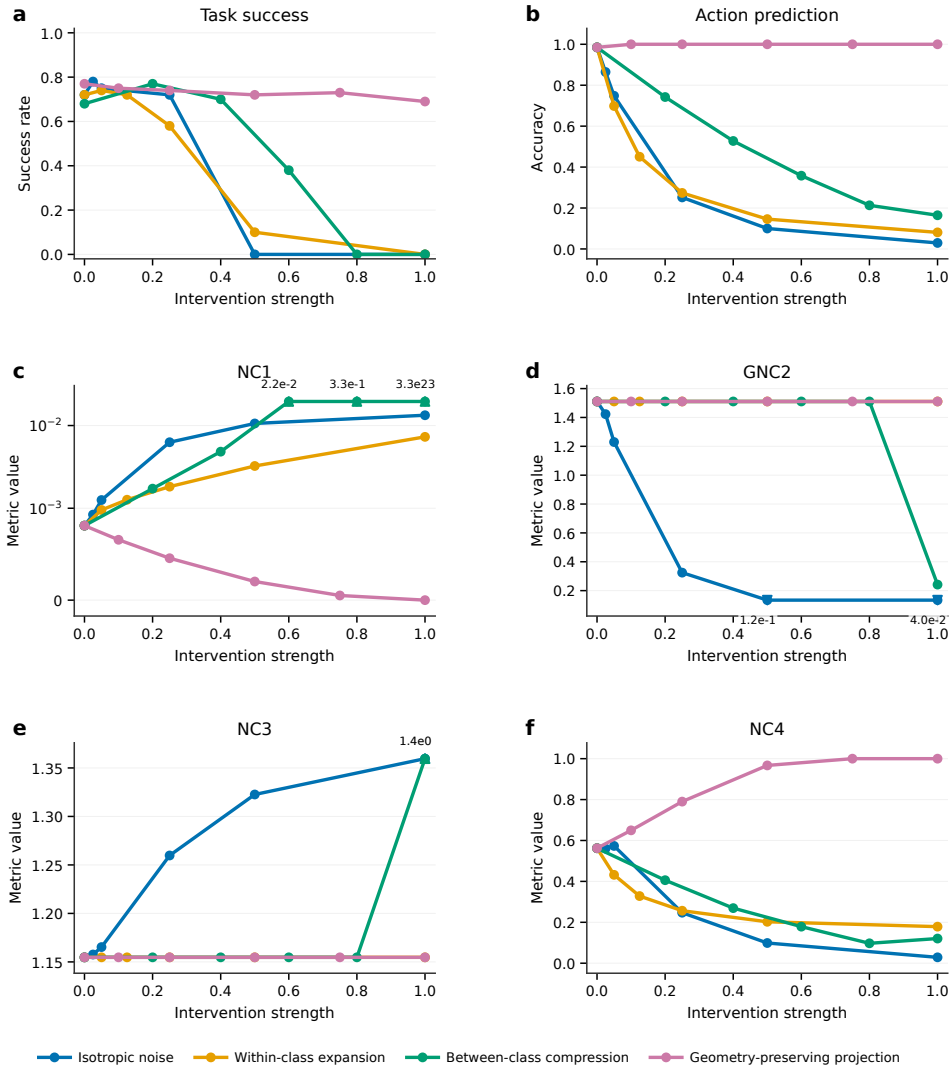


Fig. 5 Perturbing latent geometry reveals its functional role in action prediction. **a,** Effects of latent-space perturbations on closed-loop task success. **b,** Effects of the same perturbations on action-token prediction accuracy. **c–f,** Effects of perturbation strategies on NC metrics. Geometry-disrupting perturbations, including random noise, within-class expansion, and between-class compression, weaken neural-collapse geometry and degrade both task performance and action-token prediction. In contrast, *geometry-preserving projection* toward class centers maintains, and in some cases modestly improves, both policy performance and latent geometric structure.

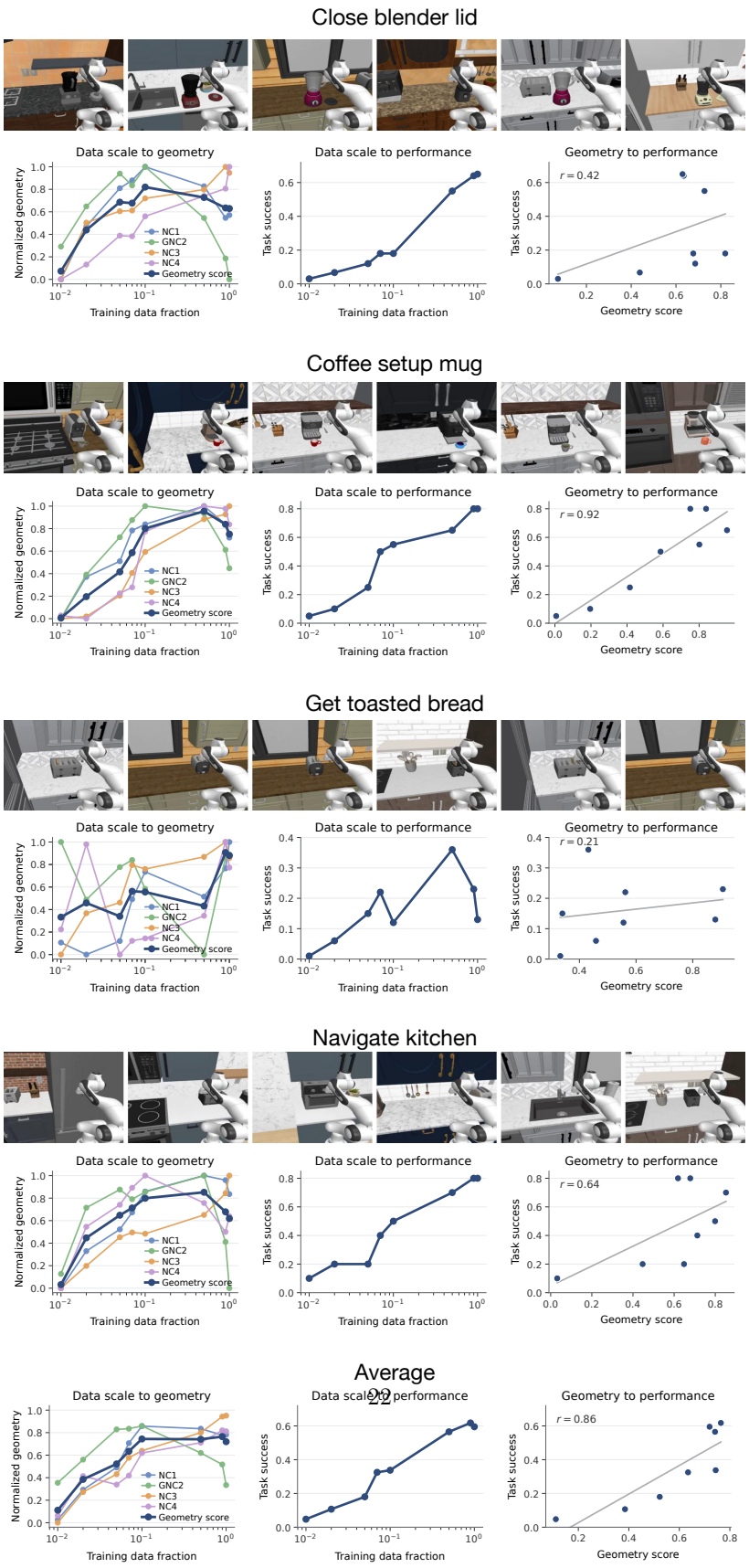


Fig. 6 Data scaling strengthens latent geometry and improves policy performance. Results for four representative RoboCasa tasks (*Close blender lid*, *Coffee setup mug*, *Get toasted bread*, and *Navigate kitchen*) and their average. For each task, representative observations are shown above three analyses: training data scale versus latent geometry (left), training data scale versus task success (middle), and latent geometry score versus task success (right). Increasing the number of demonstrations consistently strengthens latent geometry and improves downstream performance. Across tasks, stronger latent geometry is associated with higher task success, suggesting that data scaling promotes generalization by facilitating structured latent representations. The latent geometry score aggregates normalized NC metrics, with higher values indicating stronger neural-collapse geometry.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf