LEARNING FROM NATURAL IMAGES UNTANGLES NUMBER SENSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Numerical cognition is a simple yet fundamental ability that remains challenging for modern vision-language models. Number sense is already present in numerically naïve animals, suggesting that it can emerge without explicit numerical learning. Deep neural networks provide a related observation, with numerosity-selective units appearing after training on unrelated natural-image tasks and even before any training. However, these findings leave unresolved whether visual learning contributes beyond the spontaneous emergence of numerical selectivity. Human development provides a clue, as children become better at making judgments in which numerosity and proportion conflict. Here, we introduce a geometric framework to trace how representations of these two quantities change across learning. Although both untrained and pretrained networks exhibited unit-level numerosity selectivity and similar basic comparison performance, their population-level organization differed substantially. Numerosity and proportion representations overlapped strongly before training but became separated after natural-image learning, supporting more robust readout under conflict. Importantly, this reorganization did not require supervised category learning, as self-supervised learning produced comparable separation while shuffled-label training did not. Together, our results suggest that rudimentary number sense may emerge spontaneously, while structured visual learning reorganizes latent quantity information into representations that enable more independent and robust decoding.

Keywords number sense; numerical cognition; representation learning; representational geometry

TL;DR Natural-image learning enables robust number sense by separating numerosity and proportion representations.

1 INTRODUCTION

Numerical cognition supports estimation, comparison, and counting, and provides a foundation for more complex quantitative reasoning (Nieder & Dehaene, 2009). Yet despite broad visual pretraining, modern vision-language models remain unreliable at object counting and related quantitative judgments (Paiss et al., 2023; Sengupta et al., 2026). In biological systems, however, sensitivity to numerical quantity emerges without explicit numerical instruction. Numerosity-selective neurons are observed in numerically naïve monkeys, and human infants can discriminate ratios between visual quantities (Viswanathan & Nieder, 2013; McCrink & Wynn, 2007). These observations raise a fundamental question of how robust quantity representations emerge without direct numerical training.

Artificial neural networks provide a controlled system for studying how learning experience contributes to the emergence and refinement of number sense. Previous work has shown that numerosity-selective units emerge after training on unrelated object-recognition tasks (Nasr et al., 2019), and even in randomly initialized networks that have never been trained (Kim et al., 2021). Untrained networks can also support comparisons of other visual quantities, including proportion and difference (Lee et al., 2023). These findings establish that basic quantity sensitivity does not require numerical supervision. What remains unclear is whether ordinary visual learning simply preserves this numerical selectivity or reorganizes it to support more robust population-level decoding.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

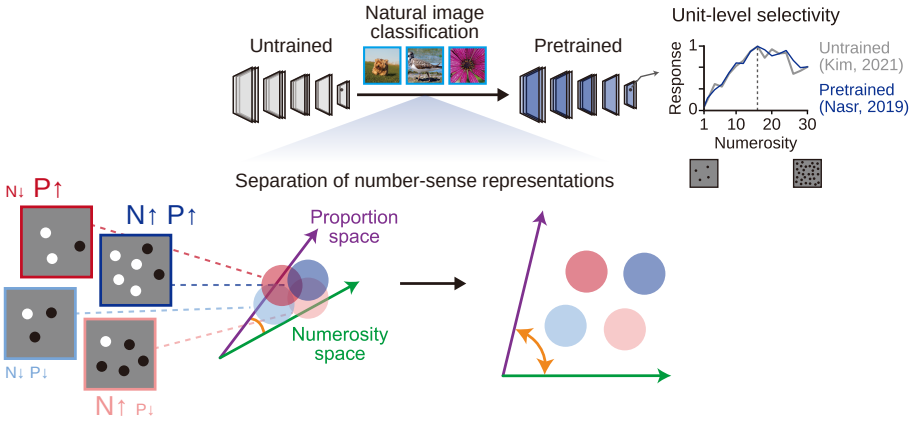

Figure 1: Natural-image pretraining reorganizes number-sense representations at the population level. Although similar numerosity selectivity has been reported at the unit level in untrained and pretrained DNNs, we show that visual learning substantially reorganizes population geometry, separating numerosity and proportion representations for more independent readout.

Human development suggests a more demanding test of this distinction. In proportional reasoning, absolute counts can interfere with proportion judgments when the two cues conflict, particularly for discrete quantities (Jeong et al., 2007; Boyer et al., 2008). Importantly, children become better able to resolve conflicting quantity cues over development, suggesting a developmental refinement of the representations supporting numerical judgments (Hamamouche & Cordes, 2023). Resolving such conflicting cues requires one quantity to be independently extracted despite variation in the other quantity, suggesting that the two quantities need to be represented along separable dimensions at the population level beyond single-unit selectivity.

Recent work provides a geometric framework for addressing this question. Rather than characterizing representations solely by single-unit tuning, population-level analyses quantify how responses to different input conditions are arranged in a multidimensional space. Measures such as the separation, overlap, distance, and relative orientation of condition-specific representations have been used to track how decodable structure emerges and how well it generalizes across conditions (Rigotti et al., 2013; Bernardi et al., 2020; Cheng et al., 2025; St-Yves et al., 2025; Wakhloo et al., 2026). Inspired by this perspective, we examine not only whether numerosity and proportion information is present at the level of individual units, but how their relative geometry changes with visual learning and whether this reorganization supports more robust decoding.

Here, we trace the representational geometry of numerosity and proportion across learning. Although both quantities are already represented at the unit level in untrained networks, visual learning without an explicit numerical objective separates their population representations and improves decoding when the cues conflict. Similar reorganization emerges through self-supervised learning, developmentally restricted visual input, and structured synthetic visual experience, whereas shuffled-label training degrades it. As a functional validation of this geometric account, greater representational separation was consistently associated with improved performance across decoding, cross-condition generalization, and cross-quantity ablation tests. Our main contributions are as follows:

- We show that natural-image pretraining refines number sense by enabling conflicting quantity cues to be represented and read out more independently.
- We identify a population-level geometric reorganization in which numerosity and proportion representations become more separated through supervised and self-supervised learning on natural images, and even through learning on synthetic images.
- We functionally validate this geometric account by linking greater representational separation to more robust and generalizable quantity readout.

2 EXPERIMENTAL SETUP

2.1 NETWORKS AND VISUAL EXPERIENCE

We used AlexNet (Krizhevsky et al., 2012) to examine how visual experience shapes quantity representations without numerical training. We compared randomly initialized networks, ImageNet-1K supervised pretraining, self-supervised learning with SimCLR (Chen et al., 2020) and RotNet (Gidaris et al., 2018), shuffled-label training, and a Developmental Visual Diet (DVD; Lu et al., 2026). We additionally evaluated untrained and ImageNet-pretrained ResNet-50 models to test whether the geometric reorganization generalized beyond AlexNet.

We also trained self-supervised networks on synthetic distributions ranging from noise and sinusoidal gratings to Spectrum + Color + WMM images (Baradad et al., 2021) and procedural Shaders (Baradad et al., 2022). Synthetic datasets were matched to the ImageNet-1K training-set size. Training details and hyperparameters are provided in Appendix A.

2.2 QUANTITY STIMULI AND REPRESENTATIONAL GEOMETRY

The evaluation set contained 18,000 images spanning 495 combinations of white- and black-dot counts (W, B) under three regimes controlling dot size, total area, and spatial extent. Numerosity ranged from 1 to 30, with numerosity and proportion defined as

$$N = W + B, \qquad P = \frac{W}{W + B}. \tag{1}$$

Frozen post-ReLU conv5 responses were averaged across variation in the other quantity and subjected to separate PCA. We retained one numerosity component $\mathbf{u}_N$ and two proportion components $\mathbf{V}_P$, for all networks, which captured more than 90% of the condition-dependent variance in most network states (Figure S3). The principal angle between the two subspaces was measured as their smallest principal angle, defined as

$$\theta = \cos^{-1}\left(\left\|\mathbf{V}_P^\top \mathbf{u}_N\right\|_2\right). \tag{2}$$

In the combined three-dimensional subspace, Gaussian-fitted condition distributions were compared using the Bhattacharyya coefficient,

$$BC_{ij} = \exp\left[-\frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right] \sqrt{\frac{\sqrt{|\boldsymbol{\Sigma}_i||\boldsymbol{\Sigma}_j|}}{|\boldsymbol{\Sigma}|}}, \qquad \boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}. \tag{3}$$

Representational overlap was averaged across all condition pairs and held-out folds; lower values indicate more distinct representations.

2.3 FUNCTIONAL VALIDATION AND STATISTICAL ANALYSIS

Baseline single-unit selectivity and numerosity comparison followed Kim et al. (2021). We used four tests to relate representational separation to quantity readout and independence. Linear readouts decoded N, P, and joint (N, P) classes from frozen representations; pairwise classifiers were evaluated on incongruent trials where the quantities favored opposite choices.

To test generalization across changes in the other quantity, an N decoder was trained at selected P conditions and evaluated at unseen P conditions. Functional independence was assessed by ablating the numerosity direction,

$$\mathbf{x}_{\text{ablated}} = \mathbf{x} - \mathbf{u}_N \mathbf{u}_N^\top \mathbf{x}, \tag{4}$$

and measuring damage to P decoding without retraining.

Networks were independent replication units, and untrained–pretrained comparisons used two-sided paired t-tests. Geometry–performance associations used Pearson correlations across matched network-state means; full readout, ablation, and statistical procedures are provided in Appendix A.

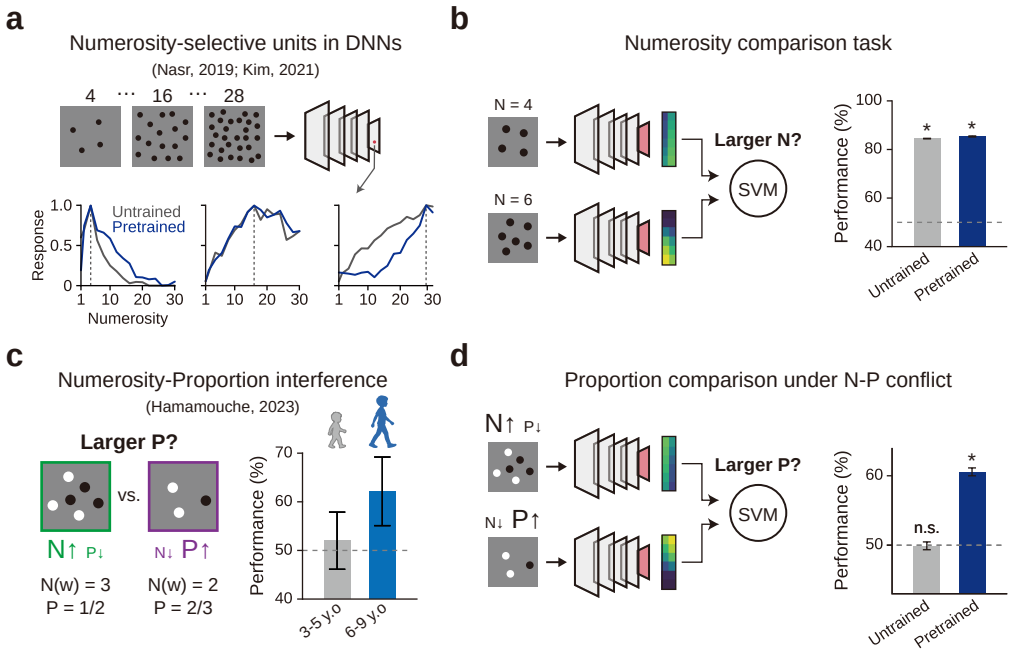

Figure 2: Basic numerosity sensitivity is similar across untrained and pretrained networks, whereas quantity conflict reveals a functional difference. (a) Both untrained and ImageNet-pretrained AlexNet contain conv5 units selectively tuned to numerosity across single-color dot arrays. (b) A linear SVM comparing which of two arrays has the larger numerosity achieves similarly high performance in both network states. (c) Human developmental studies show that proportion judgments become more robust with age when absolute count and proportion provide conflicting cues. (d) We adapted this conflict paradigm to two-color dot arrays, asking which array has the larger proportion when numerosity and proportion favor opposite choices. The untrained network performs near chance, whereas the pretrained network remains above chance.

3 Natural-image pretraining refines number sense under conflict

We first asked whether the standard unit-level criterion of number sense distinguishes a trained network from an untrained one. Following Nasr et al. (2019) and Kim et al. (2021), we presented single-color dot arrays spanning numerosities from 1 to 30 to an ImageNet-pretrained AlexNet and to the same architecture with randomly initialized weights. Figure 2a shows that both untrained and pretrained networks contained conv5 units selectively tuned to numerosity across all three stimulus sets controlling for dot size, total area, and convex hull, consistent with previous reports (Nasr et al., 2019; Kim et al., 2021). Although pretraining sharpened the tuning of number-selective units (selectivity index, untrained vs. pretrained: 0.569 vs. 0.948) while slightly reducing their fraction (7.7% vs. 5.7%), it remained unclear whether these unit-level changes translated into improved population-level decoding.

We therefore assessed population-level numerosity readout using a comparison task following Nasr et al. (2019) and Kim et al. (2021). A linear classifier trained on conv5 activity predicted which of two dot arrays contained the larger number of dots. Figure 2b shows that both untrained and pretrained networks performed well above chance (50%), with only a small increase after pretraining (84.5% vs. 85.4%). This motivated a more demanding test to reveal how number sense develops beyond basic numerical sensitivity.

A clue came from human developmental research on proportional reasoning, in which absolute counts can interfere with judgments of proportion. Hamamouche & Cordes (2023) examined how children aged three to nine compared the relative proportions of different colors in two images. Figure 2c illustrates their developmental finding: when proportion and absolute count conflicted, younger children performed near chance, whereas older children were better able to resolve the conflict. These

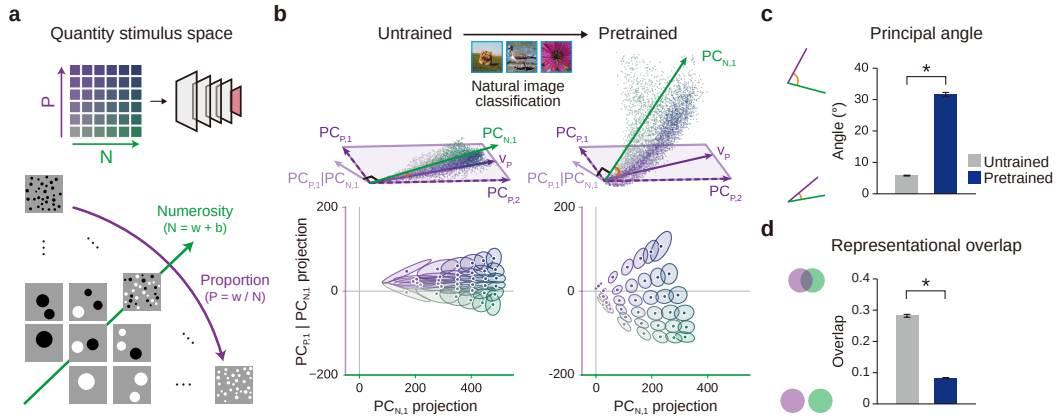

Figure 3: Natural-image pretraining separates numerosity and proportion representations at the population level. (a) Stimulus design spanning numerosity and proportion across two-color dot arrays. (b) PCA of conv5 responses reveals a one-dimensional numerosity subspace and a two-dimensional proportion subspace, which become more separated after pretraining. (c,d) This reorganization is quantified by an increased principal angle between the two subspaces and reduced representational overlap among (N,P) conditions. Together, these results show that visual learning without an explicit numerical objective produces a more separable quantity geometry.

findings suggest that numerical cognition becomes more robust over development, marking a more mature form of numerical cognition.

Inspired by this paradigm, we compared untrained and pretrained networks on a more demanding task involving conflicting numerosity and proportion cues. Using pairs of two-color dot arrays, we trained a linear readout on conv5 activity to identify the array with the higher proportion of white dots. One array could therefore have both a larger numerosity and a higher proportion (congruent), or a higher proportion despite a smaller numerosity (incongruent). On incongruent trials, the untrained network performed near chance (49.9%), whereas the pretrained network reached 60.6% accuracy. This increase in performance suggests that understanding how number sense develops requires examining not only numerosity information at the unit level, but also how distinct quantity cues are organized and independently read out at the population level.

4 Pretraining separates the numerosity and proportion subspaces

To characterize this population-level organization, we compared the geometry of numerosity and proportion representations between untrained and ImageNet-pretrained AlexNet. Figure 3a shows the two-color stimulus space, spanning numerosity values from N = 1 to 30 and proportion values from P = 0 to 1 in steps of 1/6. For each quantity, conv5 responses were first averaged across variation in the other quantity and then subjected to PCA. We used a one-dimensional numerosity subspace and two-dimensional proportion subspace across network state, which captured more than 90% of the condition-dependent variance in most network states (Figure S3).

If the improved conflict performance after pretraining reflects a more separable population representation, numerosity- and proportion-related variation should become more distinct in representational space. Figure 3b illustrates two complementary aspects of representational separation: the relative orientation of the N and P subspaces and the overlap among individual (N, P) response distributions. We quantified these using the principal angle and Gaussian distributional overlap, respectively.

The resulting geometry differed markedly between untrained and pretrained networks. Figures 3c and 3d show that pretraining increased the principal angle from 5.8° to 31.6° and reduced distributional overlap from 0.28 to 0.083, respectively (both $p < 0.001$, paired t-tests). Thus, learning from natural images in the absence of any numerical objective reorganized the quantity subspaces, providing a population-level account of the improved conflict performance observed in Figure 2. This geometric separation also emerged in ResNet-50, extending beyond AlexNet (Figure S5c).

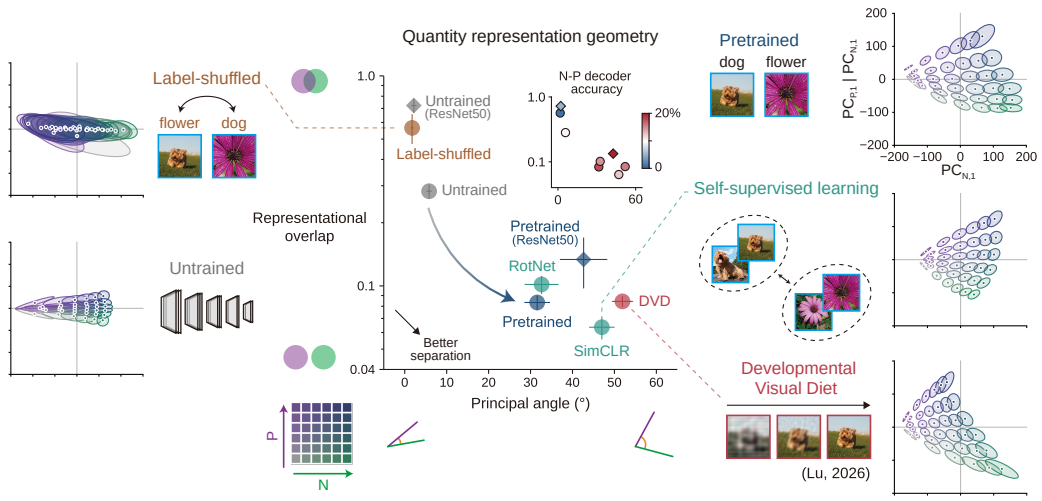

Figure 4: Quantity separation emerges without category labels but not under arbitrary supervision. Self-supervised learning produces a similar reorganization to supervised pretraining, whereas shuffled-label training reduces separation below the untrained baseline. Developmentally restricted visual input also supports strong quantity separation.

5 Quantity separation extends beyond conventional supervised pretraining

Having found that natural-image pretraining reorganizes numerosity and proportion representations, we next asked whether additional information beyond the images themselves, such as category labels, was necessary. Keeping natural-image exposure, we found that quantity subspaces still separated under label-free self-supervised learning and a developmental visual diet that restricted visual information early in training.

We first examined networks trained with self-supervised learning (SSL) on natural images using SimCLR (Chen et al., 2020) and RotNet (Gidaris et al., 2018). Figure 4 shows that self-supervised networks exhibited a similar geometry to supervised pretraining, with larger N–P principal angles and lower overlap than the untrained reference. The presence of this geometry in label-free models indicates that explicit category labels are not a prerequisite for quantity separation.

At the opposite extreme, we asked whether incorrect learning could impair the development of number sense. Figure 4 also shows the opposite effect of shuffled-label training: the principal angle decreased and representational overlap increased beyond the untrained baseline. This indicates that maladaptive learning can disrupt the emergence of separable quantity representations even when natural-image exposure is preserved.

We next asked whether quantity separation could still emerge under visual conditions that better approximate early development, when information such as contrast and acuity is substantially limited. We therefore examined a Developmental Visual Diet (DVD) model, in which visual acuity, contrast sensitivity, and chromatic sensitivity gradually mature over training (Lu et al., 2026). Figure 4 further shows that the DVD-trained network retained this reorganization despite early visual restrictions, with larger subspace angles and lower condition-wise overlap than the untrained state. Thus, these results suggest that the reorganization of quantity representations is compatible with a developmental pathway that more closely approximates the constraints of biological visual experience.

Together, these comparisons show that quantity separation is not confined to conventional supervised pretraining. It was observed without explicit category labels and after training with restricted early input, whereas arbitrary category assignments were associated with reduced separation. These findings support visual learning as a candidate route to refining quantity representations without numerical instruction.

Table 1: Representational geometry across network states: subspace orientation, overlap, distance, and volume.

State	PA (°)	NP Overlap	N Overlap	P Overlap	Normalized distance	Normalized volume (95%)
Untrained	5.8	0.282	0.547	0.491	92.8	5.28×10^5
Pretrained	31.6	0.083	0.242	0.216	138.8	4.69×10^5
DVD	51.9	0.084	0.289	0.191	141.6	4.21×10^5
Label-shuffled	1.8	0.565	0.889	0.809	118.3	2.25×10^7
SimCLR	46.8	0.065	0.253	0.168	118.7	2.06×10^5
RotNet	32.6	0.102	0.263	0.281	102.0	1.91×10^5
Noise	1.6	0.547	0.915	0.740	134.3	2.44×10^7
Grating	35.5	0.211	0.431	0.444	123.1	1.12×10^6
Statistical	39.9	0.062	0.292	0.148	101.4	1.07×10^5
Shaders	58.3	0.052	0.216	0.145	128.7	1.10×10^5

Means across networks. Centroid distance and volume are normalized within each network as D/a and V/a^3 before averaging, where a is the centered RMS activation scale across the full relu5 population. PA and BC are invariant to this scalar normalization. See Appendix A.4 for geometry estimation procedures.

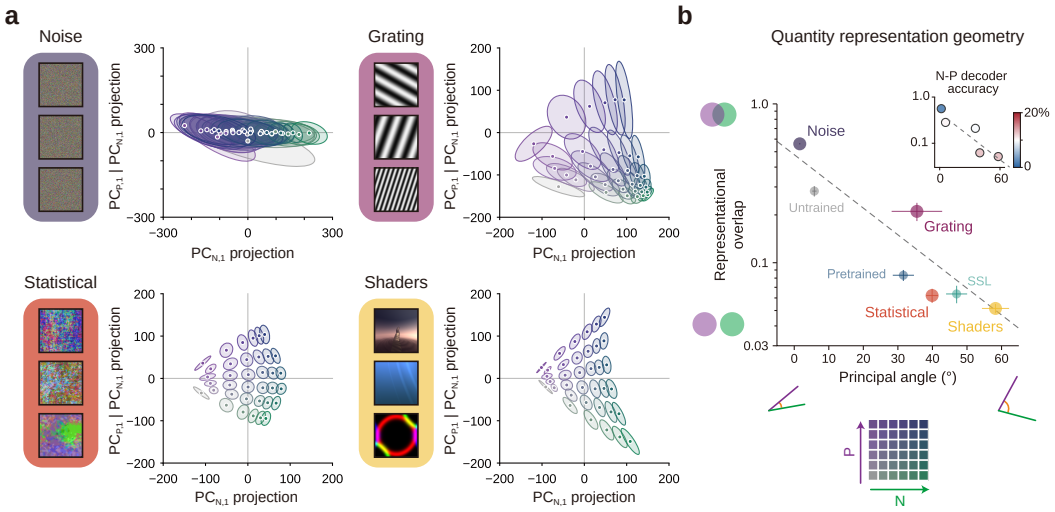

Figure 5: Structured synthetic visual experience is sufficient to separate quantity representations. (a) Self-supervised learning on synthetic images ranging from noise to shader-based patterns produces progressively more organized quantity representations. (b) Representational separation increases with richer visual structure, showing that natural images are not uniquely required for quantity reorganization.

6 More naturalistic synthetic images induce greater quantity separation

Having found that category labels were not required for quantity separation, we next asked whether natural images themselves were necessary. Figure 5a shows the synthetic training distributions, ranging from noise and gratings to statistical images containing diverse frequency and color information and Shader images containing varied geometric patterns. Figure 5b shows that quantity separation increased across increasingly structured visual inputs, following the order Noise < Grating < Statistical < Shader. Notably, Shader-based training produced even stronger separation than self-supervised learning on natural images. Thus, natural images are not uniquely required for quantity reorganization; rather, sufficiently rich and structured visual experience appears to be enough to disentangle numerosity and proportion representations.

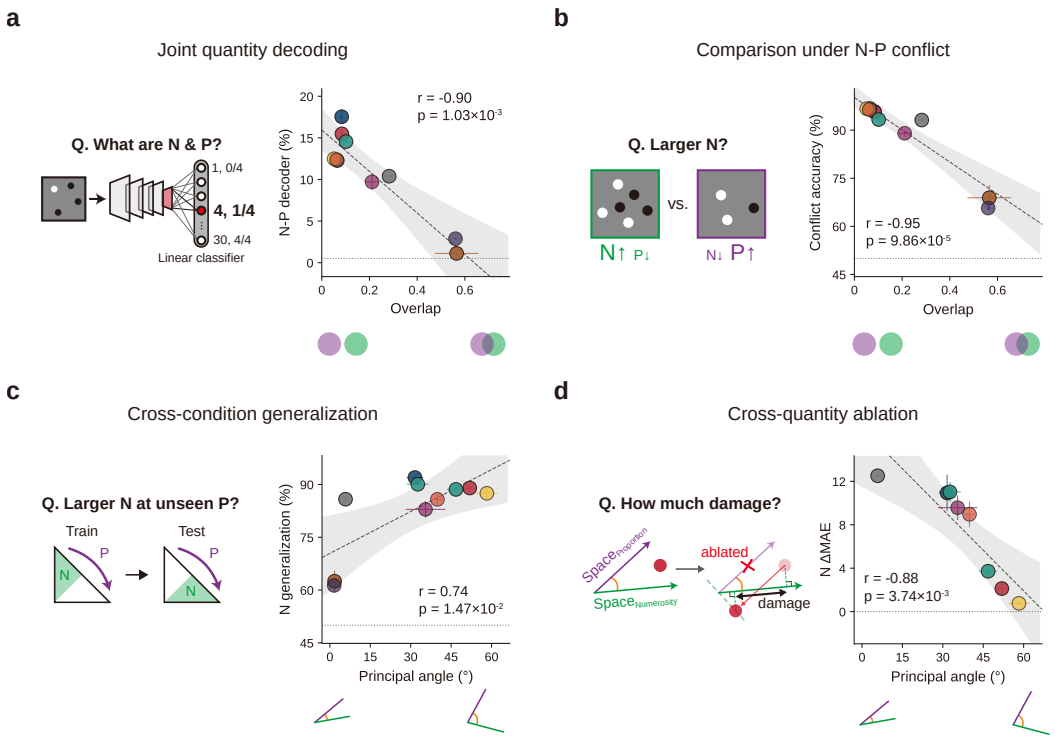

Figure 6: Functional validation of quantity representation geometry. (a,b) Lower representational overlap is associated with better joint quantity decoding and higher accuracy under N–P conflict. (c,d) Larger principal angles are associated with better generalization to unseen quantity conditions and less damage following cross-quantity ablation. Colors indicate the corresponding network states in Figures 4 and 5.

7 Representational separation predicts robust and independent quantity readout

Finally, we asked whether the geometric separation of numerosity and proportion was reflected in functional number sense across network states. We first attached linear readouts to the final feature representation to decode N and P separately, as well as jointly, and evaluated the comparison task specifically on incongruent trials, where numerosity and proportion favored opposite choices. Figures 6a and 6b show that lower condition-wise overlap was associated with better joint quantity decoding and conflict resolution across network states ($r = -0.90$, $p = 0.001$ and $r = -0.95$, $p < 0.001$, respectively). Thus, overlap captured how distinctly quantity information could be read out when different quantity conditions had to be discriminated.

We next tested whether subspace orientation predicted the ability to preserve one quantity across changes in the other. An N decoder trained at selected P conditions was evaluated on previously unseen P conditions. Figure 6c shows that larger N–P principal angles were associated with better cross-condition generalization ($r = 0.74$, $p = 0.015$). We then ablated the P subspace and measured the resulting damage to N decoding. Figure 6d shows that larger principal angles were associated with less cross-quantity ablation damage ($r = -0.88$, $p = 0.0037$), indicating greater functional independence between the two quantity representations. Together, these results link reduced representational overlap to robust quantity readout and increased subspace angle to generalization and functional independence, providing a functional validation of the geometric reorganization observed across learning.

Table 2: Compact summary of functional performance across network states.

State	Linear NP (%)	Damage N (MAE)	Damage P (MAE)	Hard Congruent (%)	Hard Incongruent (%)	Generalization N (%)	Generalization P (%)
Untrained	10.4	12.51	1.03	76.9	49.9	85.8	86.3
Pretrained	17.5	10.94	0.47	78.9	60.6	92.0	87.9
DVD	15.5	2.12	0.44	78.9	59.8	89.0	86.9
Label-shuffled	1.1	2.99	0.84	54.6	59.3	62.5	67.1
SimCLR	12.3	3.72	0.36	82.8	61.6	88.6	89.1
RotNet	14.6	11.02	0.35	74.8	52.6	90.1	88.8
Noise	2.9	4.74	1.16	52.3	72.5	61.2	73.4
Grating	9.7	9.57	0.20	70.5	65.9	82.9	80.1
Statistical	12.4	8.96	0.40	84.4	65.9	85.8	90.7
Shaders	12.5	0.76	0.59	82.9	59.6	87.5	89.3

State means; SDs and all outcomes appear in Appendix Table 3. Damage is the increase in ordinal MAE. Hard is the W/P proportion-comparison task; congruency is defined by white-dot count. Generalization averages both transfer directions.

8 Conclusion

Number sense has traditionally been studied through numerosity-selective responses in biological and artificial systems, yet such selectivity alone does not explain whether quantity information can be used robustly when multiple cues interact. Recent work on multidimensional population geometry has similarly emphasized that computation depends not only on what information is encoded, but also on how variables are organized in representational space (Rigotti et al., 2013; Bernardi et al., 2020; Cheng et al., 2025; St-Yves et al., 2025; Wakhloo et al., 2026). Our results extend this perspective to number sense, showing that visual learning without an explicit numerical objective can transform initially available quantity sensitivity into more separable representations that support robust and independent readout.

A broader implication is that this reorganization is not tied to a particular training objective or to natural photographs themselves. Numerical variables that are never specified by the training objective can nevertheless become more separable as a network learns regularities in visual input. Importantly, this separation does not appear to depend on a single form of supervision or on natural photographs as a unique source of experience. Together, these findings suggest that structured learning may reorganize incidental information into representations better suited for downstream use.

More generally, these findings suggest that learning can substantially reorganize representations of variables that are not directly related to the training objective. This provides a possible geometric perspective on transfer learning, in which source-task experience can improve downstream capabilities by shaping latent representations before the downstream variable is ever explicitly trained (Yosinski et al., 2014; Neyshabur et al., 2020; Ericsson et al., 2021). Rather than evaluating training data and objectives only by performance on the source task, it may therefore be useful to consider how they organize representations that can support generalization to unseen tasks and domains. In this view, the emergence of separable latent geometry may be one factor underlying the broader benefits of pretraining and transfer.

These results also motivate extensions to both biological and multimodal systems. In biological vision, developmental recordings could test how ordinary visual experience shapes the population geometry of quantity representations over time. In VLMs, future work could trace how quantity representations evolve from visual encoders through multimodal and language layers, identifying where numerical information becomes entangled, inaccessible, or lost during processing. Such analyses could help explain persistent counting failures and reveal representational bottlenecks that may be targeted by future training strategies. More broadly, number sense provides a compact setting for studying how experience transforms latent information into representations that can be flexibly and independently used.

AI USE STATEMENT

Generative AI tools were used to provide feedback on experimental methodology and interpretation of results, and to assist with manuscript organization, editing, and the presentation of analyses. They were also used to aid and polish writing and for retrieval and discovery of relevant literature and related work. All AI-assisted suggestions were reviewed by the authors, and all reported analyses and results were independently verified. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Detailed descriptions of the network architectures, training procedures, stimulus generation, representational analyses, functional readouts, statistical tests, and replication settings are provided in Section 2, Experimental Setup, and the Appendix. The number of independent runs and evaluation procedures are reported for each condition to facilitate reproduction of the results.

REFERENCES

Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/14f2ebeab937ca128186e7ba876faef9-Abstract.html.

Manel Baradad, Richard Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, and Phillip Isola. Procedural image programs for representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2a25d9d873e9ae6d242c62e36f89ee3a-Abstract-Conference.html.

Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4): 954–967.e21, 2020. doi: 10.1016/j.cell.2020.09.031.

Ty W. Boyer, Susan C. Levine, and Janellen Huttenlocher. Development of proportional reasoning: Where young children go wrong. *Developmental Psychology*, 44(5):1478–1490, 2008. doi: 10.1037/a0013110.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Yu-Ang Cheng, Mehdi Sanayei, Xing Chen, Ke Jia, Sheng Li, Fang Fang, Takeo Watanabe, Alexander Thiele, and Ru-Yuan Zhang. A neural geometry approach comprehensively explains apparently conflicting models of visual perceptual learning. *Nature Human Behaviour*, 9(5):1023–1040, 2025. doi: 10.1038/s41562-025-02149-x.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5414–5423, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Ericsson_How_Well_Do_Self-Supervised_Models_Transfer_CVPR_2021_paper.html.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1803.07728.

Karina Hamamouche and Sara Cordes. Winning or losing: Children's proportional reasoning across motivational contexts. *Cognitive Development*, 68:101390, 2023. doi: 10.1016/j.cogdev.2023.101390.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yoonkyung Jeong, Susan C. Levine, and Janellen Huttenlocher. The development of proportional reasoning: Effect of continuous versus discrete quantities. *Journal of Cognition and Development*, 8(2):237–256, 2007. doi: 10.1080/15248370701202471.

Gwangsu Kim, Jaeson Jang, Seungdae Baek, Min Song, and Se-Bum Paik. Visual number sense in untrained deep neural networks. *Science Advances*, 7(1):eabd6127, 2021. doi: 10.1126/sciadv.abd6127.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012. URL https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Hyeonsu Lee, Woochul Choi, Dongil Lee, and Se-Bum Paik. Comparison of visual quantities in untrained neural networks. *Cell Reports*, 42(8):112900, 2023. doi: 10.1016/j.celrep.2023.112900.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Zejin Lu, Sushrut Thorat, Radoslaw M. Cichy, and Tim C. Kietzmann. Adopting a human developmental visual diet yields robust and shape-based AI vision. *Nature Machine Intelligence*, 8:735–748, 2026. doi: 10.1038/s42256-026-01228-6.

Koleen McCrink and Karen Wynn. Ratio abstraction by 6-month-old infants. *Psychological Science*, 18(8):740–745, 2007. doi: 10.1111/j.1467-9280.2007.01969.x.

Khaled Nasr, Pooja Viswanathan, and Andreas Nieder. Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances*, 5(5):eaav7903, 2019. doi: 10.1126/sciadv.aav7903.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 512–523, 2020. URL https://papers.nips.cc/paper/2020/hash/0607f4c705595b911a4f3e7a127b44e0-Abstract.html.

Andreas Nieder and Stanislas Dehaene. Representation of number in the brain. *Annual Review of Neuroscience*, 32:185–208, 2009. doi: 10.1146/annurev.neuro.051508.135550.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Paiss_Teaching_CLIP_to_Count_to_Ten_ICCV_2023_paper.html.

Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013. doi: 10.1038/nature12160.

Saurav Sengupta, Nazanin Moradinasab, Jiebei Liu, and Donald E. Brown. Can vision-language models count? A synthetic benchmark and analysis of attention-based interventions. *arXiv preprint arXiv:2511.17722v3*, 2026. URL https://arxiv.org/abs/2511.17722v3. Revised April 2, 2026; first posted in 2025.

Ghislain St-Yves, Kendrick Kay, and Thomas Naselaris. Variation in the geometry of concept manifolds across human visual cortex. *PLOS Computational Biology*, 21(9):e1013416, 2025. doi: 10.1371/journal.pcbi.1013416.

Pooja Viswanathan and Andreas Nieder. Neuronal correlates of a visual "sense of number" in primate parietal and prefrontal cortices. *Proceedings of the National Academy of Sciences*, 110 (27):11187–11192, 2013. doi: 10.1073/pnas.1308141110.

Albert J. Wakhloo, Will Slatton, and SueYeon Chung. Neural population geometry and optimal coding of tasks with shared latent structure. *Nature Neuroscience*, 29:682–692, 2026. doi: 10.1038/s41593-025-02183-y.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, pp. 3320–3328, 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/532a2f85b6977104bc93f8580abbb330-Abstract.html.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. URL https://arxiv.org/abs/1708.03888.

A Detailed methods

A.1 Networks and visual training

We used AlexNet with matched He-normal initializations (He et al., 2015): 20 networks each for Untrained, Pretrained, Label-shuffled, and DVD, and five each for SimCLR, RotNet, Grating, Shaders, Statistical, and Noise. Training datasets contained 1,281,167 images, matching ImageNet-1K (Deng et al., 2009), and supplied no numerical supervision. Geometry and task analyses used matched checkpoints: epoch 0 for Untrained, 90 for supervised, synthetic, and natural-image SimCLR conditions, and 30 for RotNet.

Supervised training used SGD for 90 epochs (batch size 512, initial learning rate 0.02, momentum 0.9, weight decay 5×10^{-4}). Label-shuffled training used a fixed random permutation of labels across training images, preserving the empirical class-frequency distribution. DVD (Lu et al., 2026) retained this protocol while scheduling acuity, color, and contrast over 0–300 simulated months.

SimCLR (Chen et al., 2020) used a $4096 \to 4096 \to 128$ projection head and LARS (You et al., 2017) (batch size 256, learning rate 0.3, momentum 0.9, weight decay 10^{-6}, temperature 0.5). A 10-epoch linear warmup preceded cosine decay over 90 epochs. Two 224×224 views used random resized crops, color jitter, grayscale, Gaussian blur, and horizontal flips. Synthetic datasets comprised sinusoidal Gratings, Spectrum + Color + WMM images (Statistical; Baradad et al., 2021), procedural Shaders (Baradad et al., 2022), or fixed independent Gaussian pixel/channel samples converted to RGB (Noise).

RotNet (Gidaris et al., 2018) predicted $0°$, $90°$, $180°$, and $270°$ rotations with a linear $4096 \to 4$ head, averaging cross-entropy across views. Training used source batches of 192, SGD momentum 0.9, weight decay 5×10^{-4}, and learning rates 0.01, 0.001, and 0.0001 over successive ten-epoch blocks. Quantity analyses used frozen backbone features.

A.2 Stimuli and frozen population responses

The common evaluation set contained 18,000 white/black-dot images covering 495 integer (W, B) conditions with $1 \le W + B \le 30$, where $N = W + B$ and $P = W/(W + B)$. Three stimulus regimes controlled dot size, total dot area, or spatial extent. All networks received the same 227×227 three-channel images with ImageNet input normalization. Analyses used all $d = 43{,}264$ flattened post-ReLU conv5 features in raw activation units. Proportion used seven bins:

$$b(P) = \sum_{k=0}^{5} \mathbb{I}\left[P \ge \frac{2k+1}{12}\right], \qquad b(P) \in \{0, \dots, 6\}. \tag{A1}$$

Centers were $0, 1/6, \dots, 1$, with midpoint boundaries assigned to the upper bin.

A.3 Quantity subspaces and principal-angle search

Responses were averaged within each of the 30 N levels or seven P bins, pooling over the other quantity. After centering across levels, separate PCA retained the first N component $\mathbf{u}_N$ and first two P components V_P for every network. The closest direction in the P subspace was obtained by

$$\cos\theta = \max_{\|\mathbf{a}\|_2 = 1} \left|\mathbf{u}_N^\top V_P \mathbf{a}\right| = \left\|V_P^\top \mathbf{u}_N\right\|_2, \tag{A2}$$

$$\mathbf{a}_\star = \frac{V_P^\top \mathbf{u}_N}{\|V_P^\top \mathbf{u}_N\|_2}, \qquad \mathbf{v}_P = V_P \mathbf{a}_\star, \qquad \theta = \arccos\left(\|V_P^\top \mathbf{u}_N\|_2\right). \tag{A3}$$

Network-level principal angles used the full evaluation set.

A.4 Held-out Gaussian overlap, distance, and volume

Overlap was evaluated on held-out responses using three-fold cross-validation. Responses projected into the training-derived three-dimensional N1/P2 union space were fitted per cell c and fold f with

Gaussian mean $\boldsymbol{\mu}_{cf}$ and Ledoit–Wolf covariance Σ_{cf} (Ledoit & Wolf, 2004). For $K = 495$ cells, NP overlap was

$$\mathrm{BC}_{NP} = \frac{1}{3}\sum_{f=1}^{3}\frac{2}{K(K-1)}\sum_{c<d}\mathrm{BC}\big(\mathcal{N}(\boldsymbol{\mu}_{cf},\Sigma_{cf}),\mathcal{N}(\boldsymbol{\mu}_{df},\Sigma_{df})\big), \tag{A4}$$

using the Gaussian BC defined in Section 2.

For N or P overlap, cells were pooled by N level or P bin with equal cell weights. Suppressing the fold index, the class moments were

$$\boldsymbol{\mu}_g = \frac{1}{K_g}\sum_{c\in g}\boldsymbol{\mu}_c, \tag{A5}$$

$$\Sigma_g = \frac{1}{K_g}\sum_{c\in g}\left[\Sigma_c + (\boldsymbol{\mu}_c - \boldsymbol{\mu}_g)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_g)^\top\right]. \tag{A6}$$

BC was averaged over all class pairs and folds. Distance and 95% ellipsoid volume used the same held-out cell statistics:

$$D = \frac{1}{3}\sum_{f}\frac{2}{K(K-1)}\sum_{c<d}\|\boldsymbol{\mu}_{cf} - \boldsymbol{\mu}_{df}\|_2, \tag{A7}$$

$$V_{95} = \frac{1}{3}\sum_{f}\frac{1}{K}\sum_{c}\frac{4\pi}{3}\left[\chi^2_3(0.95)\right]^{3/2}\sqrt{\det\Sigma_{cf}}. \tag{A8}$$

BC is dimensionless; distance and volume retain raw activation units and cubic units, respectively.

A.5 Two-dimensional representation displays

Bases fitted on folds B and C displayed fold A responses. We orthonormalized either the N-PC1/P-PC1 directions or the closest principal-vector pair $(\mathbf{u}_N, \mathbf{v}_P)$. Plots used six N groups of width five and the seven P bins in Eq. A1; ellipses indicate one-standard-deviation covariance contours. A network-specific activation scale normalized display coordinates only.

A.6 Functional readouts and comparison tasks

Single-unit and comparison baseline. Following Kim et al. (2021), single-color dot arrays and two-way ANOVA identified units with a numerosity effect at $p < 0.01$ and stimulus-regime and interaction effects at $p > 0.01$ (uncorrected). A linear SVM compared numerosity using 256 selected units; unit-draw results were averaged within networks.

Direct N, P, and NP decoding. Affine softmax readouts without hidden layers decoded 30 N classes, seven P bins, or 195 observed NP classes. A shared stratified 80/10/10% split defined training, validation, and test sets. AdamW (Loshchilov & Hutter, 2019) used learning rate and weight decay 10^{-5}, batch size 128, at most 200 epochs, and validation-loss early stopping with patience 15. Performance was exact-class test accuracy.

Standard N/P pairwise comparison. Separate LinearSVC readouts predicted larger N or P from concatenated raw responses $[\mathbf{x}_a; \mathbf{x}_b]$ using identical pairs and ordering. Disjoint image pools supplied 800 training and 800 test pairs, including 400 N/P-congruent and 400 N/P-incongruent test pairs; congruency was $\mathrm{sign}(N_a - N_b) = \mathrm{sign}(P_a - P_b)$. Settings were $C = 1$, hinge loss, L2 penalty, and at most 20,000 iterations, with no feature scaling or PCA. Restricted tests used small N gaps or adjacent P bins.

Hard W/P comparison. A separate SVM predicted larger P for pairs one P bin apart, with congruency defined by absolute white-dot count:

$$\mathrm{congruent}_{WP}(a,b) \iff \mathrm{sign}(W_a - W_b) = \mathrm{sign}(P_a - P_b). \tag{A9}$$

The test set contained 640 congruent and 160 incongruent pairs, giving $A_{\mathrm{total}} = 0.8A_{\mathrm{congruent}} + 0.2A_{\mathrm{incongruent}}$.

Bidirectional range generalization. N comparisons transferred between $P < 0.5$ and $P > 0.5$; P comparisons transferred between $N = 1$–15 and $N = 16$–30. Each raw-response SVM used 800 training and 800 test pairs balanced by congruency, training within one domain and testing in the other. Scores averaged the two transfer directions.

A.7 Cross-domain ablation and damage

Bases fitted on the full evaluation set removed the opposite quantity's subspace from raw responses without recentering or refitting the readout:

$$\mathbf{x}_{\setminus P} = (I - V_P V_P^\top)\mathbf{x}, \qquad \mathbf{x}_{\setminus N} = (I - \mathbf{u}_N \mathbf{u}_N^\top)\mathbf{x}. \tag{A10}$$

Damage N removed P2 and evaluated the fixed N decoder; Damage P removed N1 and evaluated the fixed P decoder. For test set T and opposite quantity $\bar{q}$, damage was

$$\Delta\mathrm{MAE}_q = \frac{1}{|T|} \sum_{i \in T} \left(|\widehat{y}_q(\mathbf{x}_{i,\setminus \bar{q}}) - y_{qi}| - |\widehat{y}_q(\mathbf{x}_i) - y_{qi}| \right). \tag{A11}$$

Here $\widehat{y}_q$ is the predicted class index. Positive damage denotes increased error, measured in N steps or P-bin steps.

A.8 Replication and statistical analysis

Network initializations were the independent replication units. Geometry and tasks were checkpoint-matched; folds and repeated unit draws were averaged within networks. Tables and endpoint geometry maps show sample SD, while paired comparison bars show SEM. Untrained–Pretrained comparisons used two-sided paired t-tests across 20 matched initializations; the introductory comparison baseline also used a one-sided test against 50% chance. Cross-state Pearson correlations and two-sided tests used equally weighted state means, with sample size equal to the included state count. Trend lines show ordinary least-squares fits with 95% confidence intervals; state inclusion followed each analysis.

B COMPLETE TASK PERFORMANCE

Table 3 reports all functional outcomes, retaining network-level variability and the same checkpoint-matched cohort as the compact main-text summary. The pairwise comparison protocols and the direction of each ablation are defined in Appendix A.6.

Table 3: Task performance across the checkpoint-matched appendix state cohort.

A. Linear decoding and cross-domain ablation damage

State	Linear N (%)	Linear P (%)	Linear NP (%)	Damage N (MAE)	Damage P (MAE)
Untrained	21.4 ± 0.7	69.2 ± 1.3	10.4 ± 0.5	12.51 ± 0.09	1.03 ± 0.02
Pretrained	29.2 ± 1.0	82.9 ± 0.7	17.5 ± 0.8	10.94 ± 1.67	0.47 ± 0.07
DVD	25.4 ± 0.9	83.2 ± 0.9	15.5 ± 0.5	2.12 ± 0.64	0.44 ± 0.06
Label-shuffled	4.9 ± 0.7	23.9 ± 5.2	1.1 ± 0.2	2.99 ± 1.57	0.84 ± 0.12
SimCLR	21.4 ± 0.9	78.7 ± 1.5	12.3 ± 0.8	3.72 ± 0.59	0.36 ± 0.05
RotNet	25.1 ± 0.8	74.7 ± 1.1	14.6 ± 0.2	11.02 ± 0.91	0.35 ± 0.07
Noise	5.9 ± 0.2	27.5 ± 0.4	2.9 ± 0.1	4.74 ± 0.24	1.16 ± 0.04
Grating	18.7 ± 1.6	57.1 ± 2.3	9.7 ± 1.0	9.57 ± 1.13	0.20 ± 0.06
Statistical	21.1 ± 1.5	77.8 ± 1.3	12.4 ± 0.6	8.96 ± 1.20	0.40 ± 0.04
Shaders	20.1 ± 1.1	81.7 ± 0.6	12.5 ± 0.7	0.76 ± 0.12	0.59 ± 0.06

B. Pairwise discrimination and cross-context generalization

State	SVM N (%)	SVM P (%)	Hard Total (%)	Hard Congruent (%)	Hard Incongruent (%)	Generalization N (%)	Generalization P (%)
Untrained	93.2 ± 0.5	90.0 ± 0.5	71.5 ± 1.1	76.9 ± 1.4	49.9 ± 2.5	85.8 ± 1.0	86.3 ± 0.7
Pretrained	95.8 ± 0.4	90.3 ± 0.5	75.2 ± 1.3	78.9 ± 1.3	60.6 ± 2.5	92.0 ± 0.6	87.9 ± 0.9
DVD	95.7 ± 0.3	90.2 ± 0.4	75.1 ± 1.0	78.9 ± 1.0	59.8 ± 2.1	89.0 ± 0.7	86.9 ± 0.5
Label-shuffled	68.9 ± 4.0	65.0 ± 8.0	55.5 ± 2.9	54.6 ± 6.0	59.3 ± 17.1	62.5 ± 3.1	67.1 ± 8.8
SimCLR	96.7 ± 0.4	90.9 ± 0.4	78.6 ± 1.2	82.8 ± 1.2	61.6 ± 2.4	88.6 ± 1.1	89.1 ± 0.9
RotNet	93.3 ± 0.5	89.7 ± 0.7	70.3 ± 1.5	74.8 ± 1.2	52.6 ± 2.9	90.1 ± 0.5	88.8 ± 1.0
Noise	65.8 ± 0.5	71.2 ± 0.4	56.4 ± 0.2	52.3 ± 0.2	72.5 ± 0.7	61.2 ± 0.3	73.4 ± 0.1
Grating	89.0 ± 0.8	86.6 ± 1.4	69.6 ± 1.8	70.5 ± 1.3	65.9 ± 4.1	82.9 ± 2.4	80.1 ± 1.5
Statistical	96.4 ± 0.3	92.0 ± 0.2	80.7 ± 1.1	84.4 ± 0.9	65.9 ± 2.0	85.8 ± 1.5	90.7 ± 0.6
Shaders	96.7 ± 0.2	90.5 ± 0.6	78.3 ± 0.5	82.9 ± 1.0	59.6 ± 4.2	87.5 ± 1.7	89.3 ± 0.7

Mean $\pm$ sample SD across checkpoint-matched networks. All accuracies are percentages; damage is the increase in ordinal MAE after P2 removal for N or N1 removal for P. Hard denotes W/P proportion comparison (640 congruent and 160 incongruent test pairs). Generalization averages both range-transfer directions. Protocols and endpoint epochs are detailed in Appendix A.

C SUPPLEMENTARY FIGURES

These figures expand the stimulus design, geometry–function comparisons, subspace dimensionality, label-correspondence controls, learning trajectories, synthetic-image distributions, and three-dimensional visualization used in this study. Analysis definitions are provided in Appendix A.

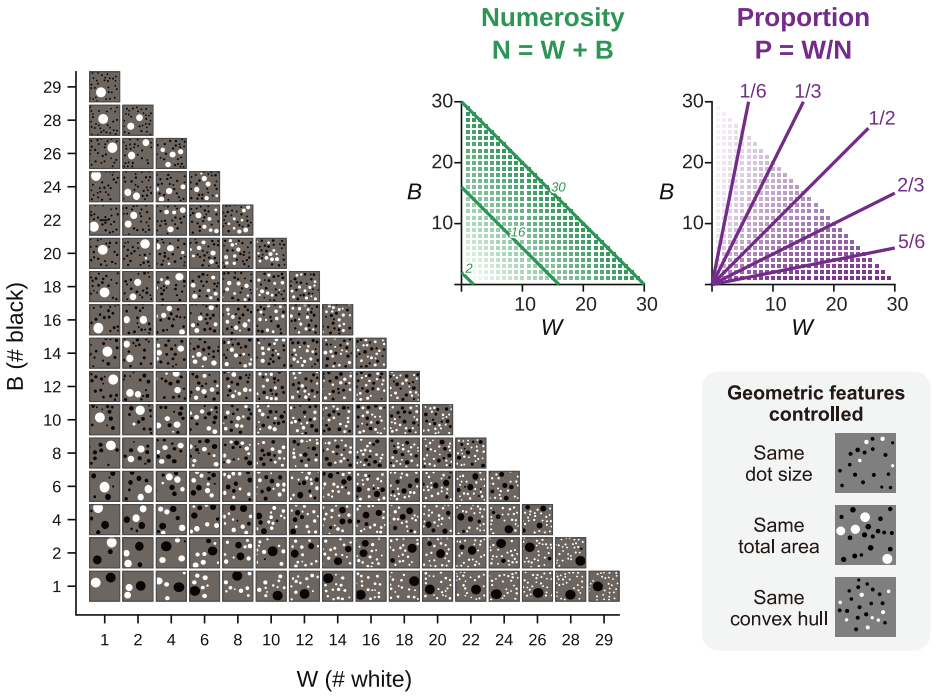

Figure S1: Systematic variation of numerosity and proportion with geometric controls. White-dot count W and black-dot count B jointly vary total numerosity $N = W + B$ and proportion $P = W/N$ across the stimulus grid. Constant-N and constant-P directions reveal the systematic organization of the two quantities. Three stimulus regimes control dot size, total dot area, and spatial extent, enabling quantity representations to be examined across these geometric properties.

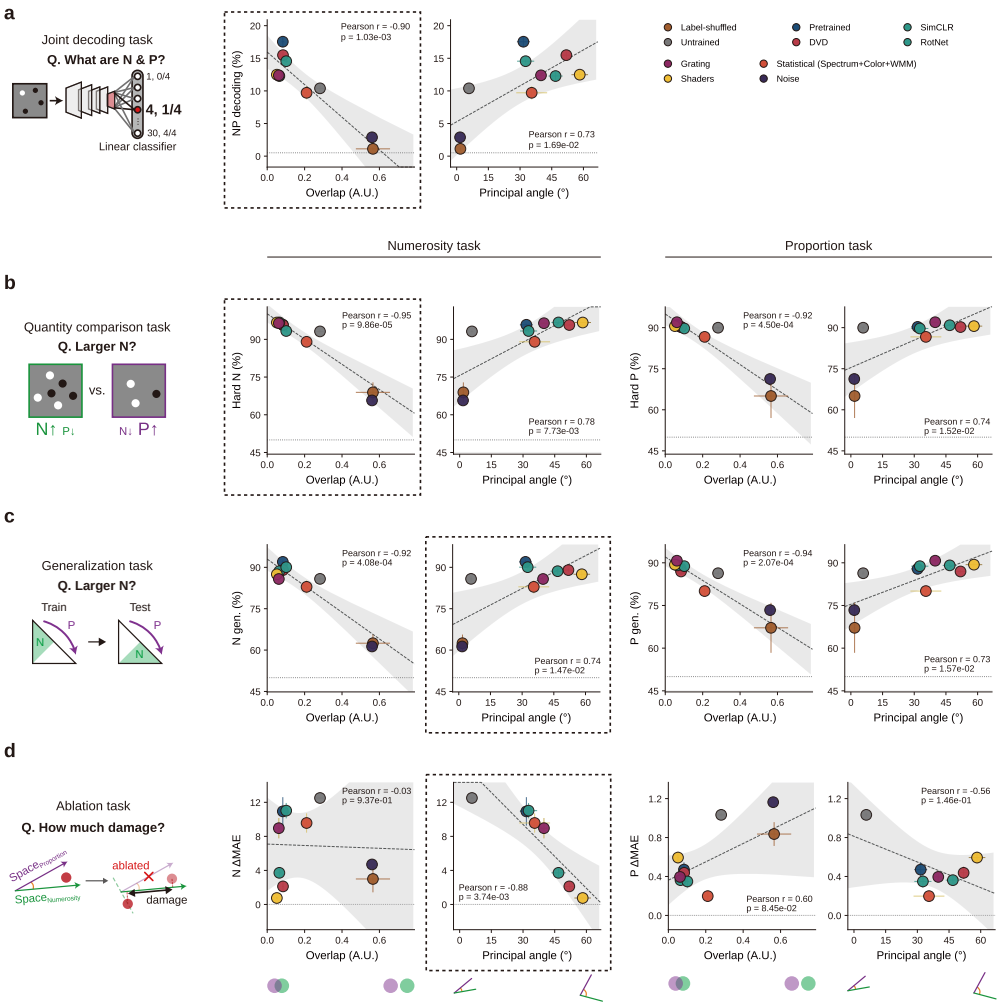

Figure S2: Geometry–function correlation matrix. (a) Exact joint NP decoding. (b) Restricted pairwise N and P comparisons. (c) Cross-context generalization of N and P comparison rules. (d) Cross-domain ablation damage, measured as the change in ordinal MAE. Each functional readout is shown against condition-wise overlap and principal angle. Points represent state means; the annotated Pearson correlations describe cross-state associations. Dashed fits and shaded bands show ordinary least-squares trends and 95% confidence intervals. Boxed panels correspond to the selected relationships in Figure 6.

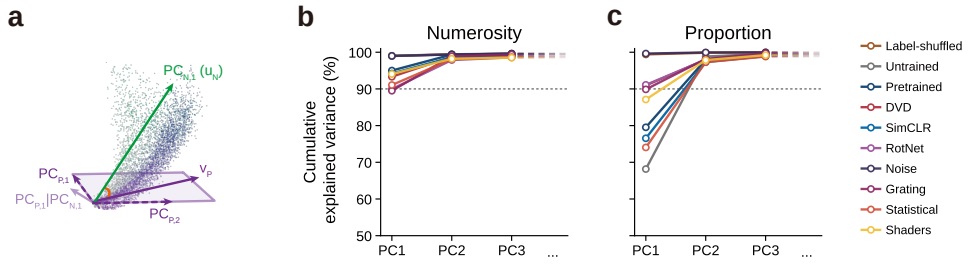

Figure S3: Rationale for the fixed N1/P2 subspace dimensions. (a) Schematic of the one-dimensional numerosity subspace and two-dimensional proportion subspace. (b,c) Cumulative explained variance of quantity-conditioned responses for N and P. One N component and two P components capture more than 90% of the variance in most states (dotted threshold). Retaining these dimensions throughout provides a common criterion for comparing the two quantities and their geometry across network states.

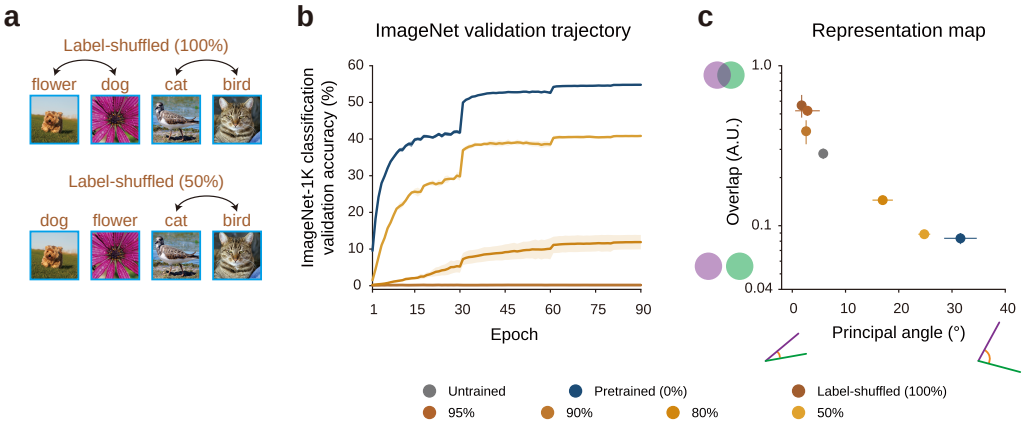

Figure S4: Correct-label correspondence links visual learning to quantity separation. (a) Schematic of complete and partial label reassignment. (b) ImageNet validation accuracy increases as the shuffled-label fraction decreases, demonstrating more effective learning from stronger label correspondence. (c) The conditions span a continuum from fully Label-shuffled through the vicinity of Untrained to Pretrained in principal-angle–overlap space. Decreasing the shuffled-label fraction increases the principal angle and reduces overlap, linking label correspondence to improved quantity separation. Percentages denote shuffled-label fractions, from 0% (Pretrained) to 100% (fully shuffled).

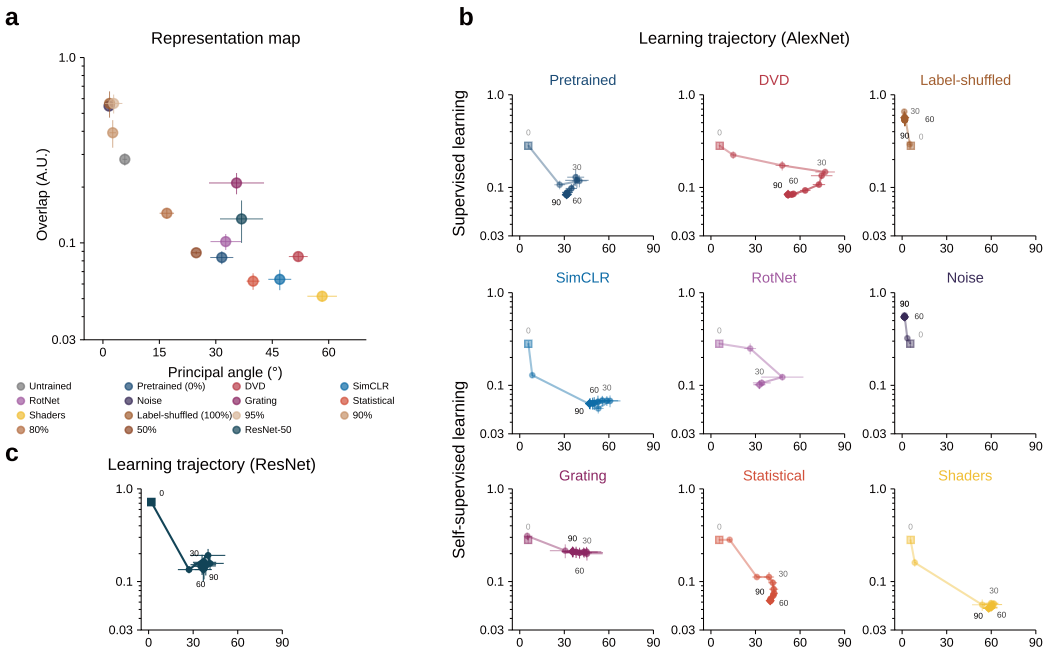

Figure S5: Learning trajectories in principal-angle–overlap space. (a) Summary representation map of the network states and label-correspondence conditions. (b) State-specific AlexNet trajectories during supervised and self-supervised learning. (c) The corresponding trajectory for ResNet-50. Connected points track geometry across training checkpoints; numeric annotations indicate epochs. Rightward and downward movements indicate larger principal angles and lower overlap, respectively. The trajectories show how different visual-learning conditions reshape quantity geometry relative to their initial representations.

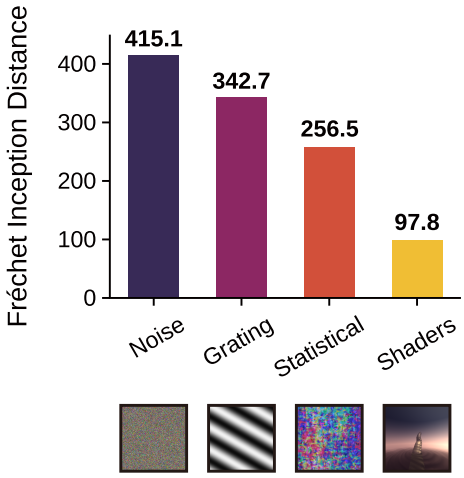

Figure S6: Synthetic-image similarity to ImageNet-1K. Fréchet Inception Distance (FID; Heusel et al., 2017) evaluates the similarity of Noise, Grating, Statistical, and Shaders image distributions to ImageNet-1K; lower values indicate greater similarity. Example images are shown below the bars. Across these visual diets, lower FID generally accompanies larger principal angles and lower overlap after training (Figure 5 and Figure S5). This pattern supports the main-text account that learning sufficiently rich visual structure, rather than natural-image origin alone, promotes quantity separation.

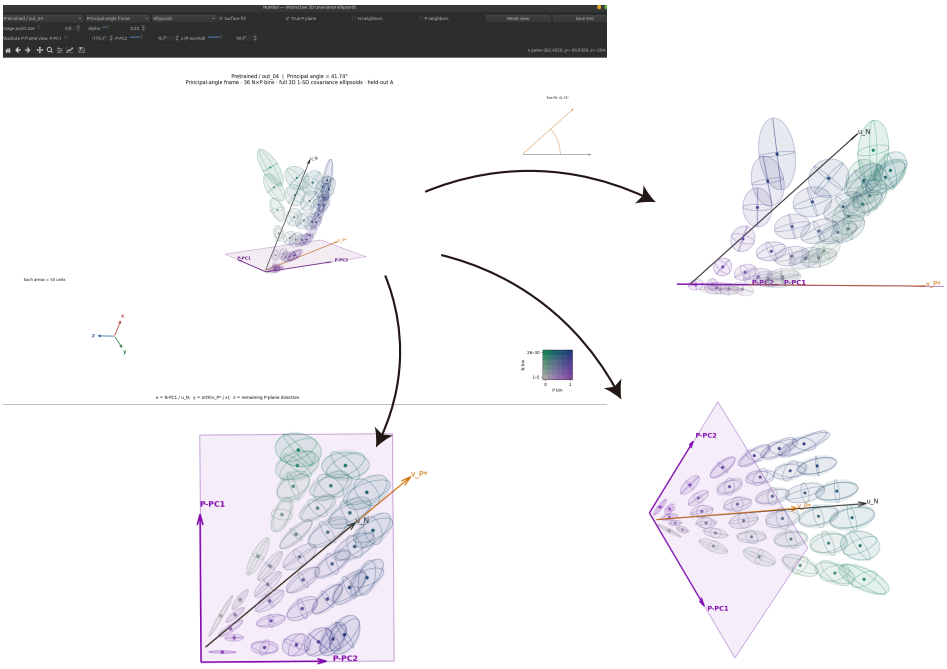

Figure S7: Interactive three-dimensional visualization of quantity geometry. The viewer displays state-specific representations in the joint N1/P2 space, with an example Pretrained network shown from multiple viewpoints. The N axis, P plane, and closest principal-vector pair allow the principal angle to be inspected directly. Condition centroids and covariance ellipsoids reveal how N- and P-conditioned representations are arranged, distributed, and intertwined. Rotating the same map provides complementary views of subspace orientation, centroid separation, distributional spread, and overlap, making the geometric measures intuitively interpretable.