PURRTURBED BUT STABLE: HUMAN-CAT INVARIANT REPRESENTATIONS ACROSS CNNS, VITS AND SELF-SUPERVISED VITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cats and humans differ in ocular anatomy. Most notably, Felis Catus (domestic cats) have vertically elongated pupils linked to ambush predation; yet, how such specializations manifest in downstream visual representations remains incompletely understood. We present a unified, frozen-encoder benchmark that quantifies feline-human cross-species representational alignment in the wild, across convolutional networks, supervised Vision Transformers, windowed transformers, and self-supervised ViTs (DINO), using layer-wise Centered Kernel Alignment (linear and RBF) and Representational Similarity Analysis, with additional distributional and stability tests reported in the paper. Across models, DINO ViT-B/16 attains the most substantial alignment (mean CKA-RBF ≈ 0.814, mean CKA-linear ≈ 0.745, mean RSA ≈ 0.698), peaking at early blocks, indicating that token-level self-supervision induces early-stage features that bridge species-specific statistics. Supervised ViTs are competitive on CKA yet show weaker geometric correspondence than DINO (e.g., ViT-B/16 RSA ≈ 0.53 at block8; ViT-L/16 ≈ 0.47 at block14), revealing depth-dependent divergences between similarity and representational geometry. CNNs remain strong baselines but below plain ViTs on alignment, and windowed transformers underperform plain ViTs, implicating architectural inductive biases in cross-species alignment. Results indicate that self-supervision coupled with ViT inductive biases yields representational geometries that more closely align feline and human visual systems than widely used CNNs and windowed Transformers, providing testable neuroscientific hypotheses about where and how cross-species visual computations converge.

1 INTRODUCTION

Representation learning has reshaped computer vision, with convolutional networks (CNNs) and Vision Transformers (ViTs) delivering strong features for transfer and understanding (He et al., 2015; Huang et al., 2018; Tan & Le, 2020; Liu et al., 2022; Dosovitskiy et al., 2021; Liu et al., 2021). A growing body of work evaluates learned representations by their geometry and cross-domain stability, using tools such as Centered Kernel Alignment (CKA) and Representational Similarity Analysis (RSA) to compare models and brains (Kornblith et al., 2019; Kriegeskorte et al., 2008). However, despite interest in biological plausibility and cross-species comparisons (e.g., alignment studies between artificial and biological vision (Yamins et al., 2014; Schrimpf et al., 2021; 2018)), there is limited, systematic evidence on how modern encoders align visual representations across species differences that arise from distinct ocular and ecological constraints.

This work addresses that gap with a unified, frozen-encoder benchmark that quantifies cross-species representational alignment between cats and humans. We compare paired feline-human images across widely used CNN families (ResNet, DenseNet, EfficientNet, ConvNeXt, MobileNet) (Howard et al., 2017; Sandler et al., 2019; Howard et al., 2019), supervised ViTs (ViT-B/L), windowed transformers (Swin-T/S/B), and self-supervised ViTs (DINO and DINOv2/v3) (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). To ensure comparability and statistical rigor, we measure alignment layer-wise using CKA (linear and RBF) and RSA/Mantel, and we probe distributional differences via MMD, Energy distance, and 1-Wasserstein, alongside paired stability tests with Benjamini-Hochberg FDR control.

At a high level, the approach is simple: freeze encoders, extract per-layer features on paired feline-human images, vectorize features consistently (tokens for ViTs, global average pooling for CNN-like maps), and compute alignment and shift statistics per model and layer. This design isolates representational similarities attributable to the trained feature spaces themselves, rather than to downstream heads or fine-tuning protocols; by keeping models frozen, we test inherent alignment capacity without confounding effects of task-specific fine-tuning or domain adaptation.

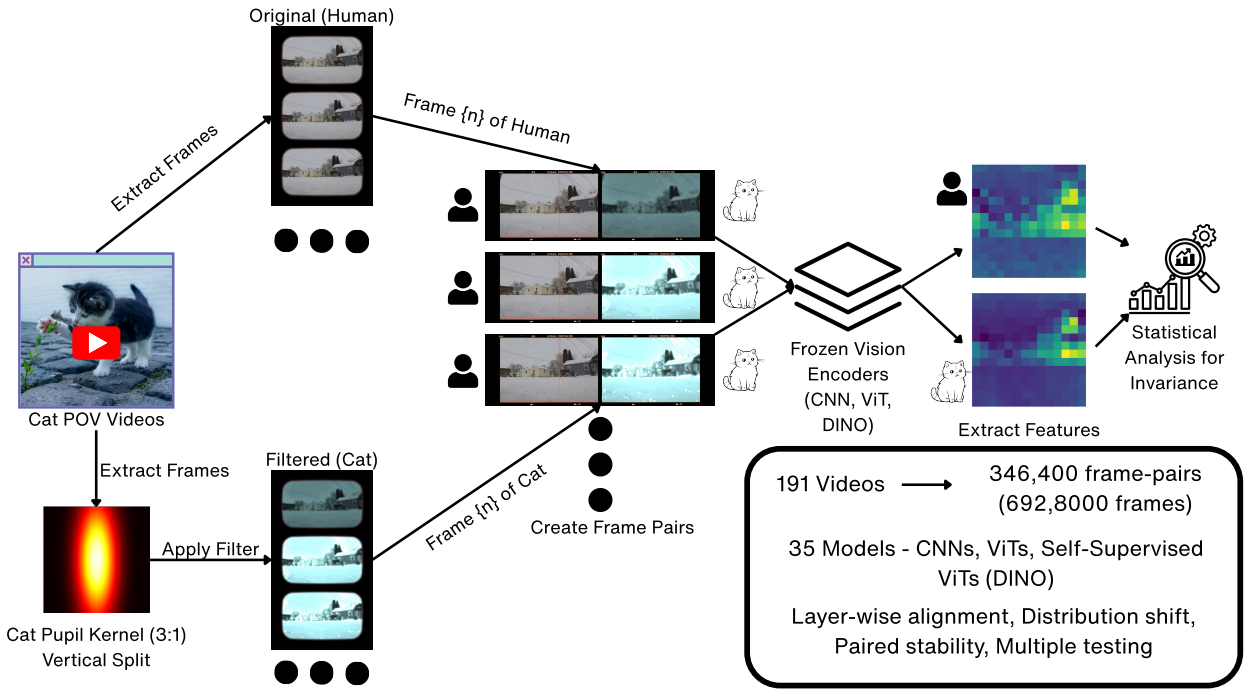

Figure 1: Process Overview for Measuring Human-Cat Invariant Representations in CNNs, ViTs and Self-Supervised ViTs. A total of 191 videos containing cat POV videos are sourced from the internet. Our biologically informed cat vision filter is applied to individual frames and we create pairs of original (human) vs. cat vision filtered frames which pass through a suite of frozen vision encoders and the extracted features are then subjected to statistical tests.

On our dataset, self-supervised ViTs exhibit the strongest cross-species alignment. In particular, DINO ViT-B/16 attains mean CKA-RBF 0.814, mean CKA-linear 0.745, and mean RSA (Mantel) 0.698 with peaks at early blocks; supervised ViTs are competitive on CKA-linear but show weaker geometric correspondence (e.g., ViT-B/16 RSA ≈ 0.53; ViT-L/16 ≈ 0.47). CNNs remain strong baselines yet generally trail plain ViTs on alignment (e.g., ResNet-50 mean CKA-linear 0.663, RSA 0.488), and windowed transformers (Swin) underperform plain ViTs on average. DINOv2/v3 variants show high early-layer maxima but lower means than DINO ViT-B/16. These findings suggest that token-level self-supervision plus ViT inductive biases yield early-stage features that better bridge species-specific statistics than widely used CNNs and windowed transformers.

In this work, we contribute (1) A unified, frozen-encoder benchmark and analysis pipeline for cross-species representational alignment spanning CNNs (ResNet, DenseNet, EfficientNet, ConvNeXt, MobileNet), supervised ViTs (ViT-B/L), windowed transformers (Swin-T/S/B), and self-supervised ViTs (DINO, DINOv2/v3). (2) A layer-wise, model-agnostic evaluation using CKA (linear/RBF) and RSA/Mantel with distribution-shift and paired-stability tests under BH-FDR, enabling precise comparisons across heterogeneous architectures. (3) Quantitative evidence that self-supervised ViTs, especially DINO ViT-B/16, achieve the strongest feline-human alignment with early-layer peaks, while supervised ViTs, CNNs, and Swin lag to varying degrees. (4) A new in-the-wild dataset of 191 videos yielding $\approx 300,000$ paired human-cat frames plus a *biologically informed cat-vision filter* (novel, first-of-its-kind implementation), enabling controlled cross-species benchmarking; we document curation and filtering details.

2 Related Work

Prior work spans four themes relevant to our study: (i) representation similarity and analysis methods, (ii) architectural families for vision and self-supervision, (iii) alignment of artificial and biological vision, and (iv) benchmarks probing invariances and frozen-encoder evaluations. We review each and situate our contribution: a unified, *frozen-encoder*, per-layer cross-species analysis across CNNs, supervised ViTs, and DINO-style ViTs with comprehensive statistics.

2.1 Representation Similarity and Analysis Methods

RSA compares internal representational geometries via RDMs across models, brains, and behavior (Kriegeskorte et al., 2008). Subspace methods align activations: SVCCA (PCA + CCA) (Raghu et al., 2017) and PWCCA (projection-weighted) (Morcos et al., 2018). CKA offers invariant, empirically stable similarity across architectures and runs (Kornblith et al., 2019). Orthogonal Procrustes supports rotation-constrained alignment for complementary perspectives (Schönemann, 1966). To capture distributional differences beyond similarity, kernel two-sample tests (MMD) (Gretton et al., 2012), energy distance (Rizzo & Székely, 2016), and Wasserstein distances quantify shifts between domains or species.

Reliability and inference are integral to RSA: the Mantel test assesses significance via matrix permutations (Mantel, 1967); noise ceilings and reliability checks contextualize absolute scores (Walther et al., 2016). For CKA, its HSIC foundation links representational similarity to RKHS dependence, guiding linear vs. RBF choices by expected nonlinearities (Gretton et al., 2008).

We use *frozen-encoder, per-layer* alignment with CKA (linear/RBF) and RSA/Mantel, plus MMD, Energy, and 1-Wasserstein, employing paired tests under BH-FDR to jointly quantify similarity and shift across species.

2.2 ConvNets, Vision Transformers, and Self-Supervised ViTs

CNN encoders include residual (He et al., 2015), dense (Huang et al., 2018), compound-scaled (Tan & Le, 2020), modernized (Liu et al., 2022), and mobile-efficient designs (Howard et al., 2017; Sandler et al., 2019; Howard et al., 2019). ViT introduces tokenized global attention (Dosovitskiy et al., 2021); Swin adds hierarchical shifted windows (Liu et al., 2021). Self-supervised ViTs (DINO, DINOv2/v3) yield strong token-level features without labels and scale effectively (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). Frozen-encoder transfer is competitive, yet *cross-species*, per-layer alignment remains underexplored; our evaluation compares CNNs, supervised ViTs, and DINO-style ViTs in a unified frozen setting.

Data-efficient supervised ViTs (DeiT) leverage distillation/augmentation (Touvron et al., 2021); masked image modeling (BEiT) improves token-level pretexts (Bao et al., 2022). Contrastive and non-contrastive SSL (MoCo v2, BYOL) establish strong frozen-transfer baselines (Chen et al., 2020b; Grill et al., 2020), and token-centric iBOT extends masked prediction with instance/patch tasks (Zhou et al., 2022). These trends position self-supervised ViTs as promising for invariant representations, aligning with our findings.

2.3 Aligning Artificial and Biological Vision

Performance-optimized hierarchical models predict high-level ventral stream responses (Yamins et al., 2014); Brain-Score aggregates neural and behavioral benchmarks (Schrimpf et al., 2018; 2021). RSA unifies representational geometry comparisons across species and modalities (Kriegeskorte et al., 2008), and CKA offers desirable invariances for encoder comparisons (Kornblith et al., 2019). Time-resolved RSA with MEG/EEG links model layers to temporal dynamics (Cichy et al., 2014).

Goal-driven modeling ties task optimization to cortical predictivity (Yamins & DiCarlo, 2016), extending to early visual areas under naturalistic stimuli (Cadena et al., 2019), supporting learned encoders as testable hypotheses about neural coding.

Ecological factors shape species vision; vertical slit pupils in ambush predators reflect niche-specific constraints (Banks et al., 2015), motivating cross-species comparisons on in-the-wild data and biologically informed prefilters. We use a cat-vision filter to approximate species-specific inputs and assess frozen, layer-wise alignment.

Unlike brain-predictivity or within-species behavior studies, we measure *cross-species representational alignment* directly on feline-human imagery, combining per-layer RSA/CKA with distributional tests.

2.4 Benchmarks, Invariances, and Frozen-Encoder Evaluations

Benchmarks expose invariance and robustness gaps: ImageNet-trained CNNs often favor texture over shape (Geirhos et al., 2022); Stylized-ImageNet encourages shape bias (Geirhos et al., 2022). Corruption benchmarks (CIFAR-10-C, ImageNet-C) quantify robustness (Hendrycks & Dietterich, 2019). ImageNet-A/O/R probe natural adversarial, OOD, and rendition shifts (Hendrycks et al., 2021). Broader efforts measure real-world distribution shift (Taori et al., 2020; Koh et al., 2021). SSL shows strong frozen transfer (SimCLR, BYOL, MAE) (Chen et al., 2020a; Grill et al., 2020; He et al., 2021).

We complement this literature by targeting *cross-species* invariance with a unified frozen-encoder, layer-wise pipeline across CNNs, supervised ViTs, and DINO/DINOv2/v3, quantifying RSA/CKA and MMD, Energy, and 1-Wasserstein with paired stability tests and BH-FDR; our in-the-wild feline-human dataset and cat-vision filter add ecological validity.

Distinct from within-domain/model comparisons or single-metric studies, we analyze paired cross-species data across heterogeneous frozen encoders with RSA/CKA and distributional tests under FDR control, enabled by our dataset and biologically informed filter.

3 METHODOLOGY

We study cross-species representational invariance by comparing layer-wise activations from diverse vision encoders on paired views of the same scenes captured in human and feline domains. All encoders are kept frozen to isolate the inductive biases of the pretrained features, enabling apples-to-apples comparisons across architectures and training paradigms.

Our methodology comprises six stages: (1) constructing a strictly paired dataset with filename-level alignment and integrity checks; (2) standardizing inputs using each encoder's canonical preprocessing; (3) extracting intermediate activations at semantically comparable layers along the processing hierarchy; (4) vectorizing activations to a common representation to remove architectural idiosyncrasies; (5) quantifying alignment and shift with complementary metrics such as Centered Kernel Alignment (CKA) (Kornblith et al., 2019), Representational Similarity Analysis (RSA) and Mantel permutation testing (Kriegeskorte et al., 2008; Mantel, 1967), plus Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), Energy distance (Rizzo & Székely, 2016), and a projected 1-Wasserstein distance under false-discovery-rate control; and (6) reporting layer-wise summaries and diagnostic visualizations to contextualize effects across families. Design choices reflect established best practices in representation analysis: frozen-encoder evaluation to avoid confounds from fine-tuning; layer-wise probes to localize where alignment emerges; vectorization via global average pooling for map features and class-token or token-mean for tokenized features (Dosovitskiy et al., 2021; Liu et al., 2021); and complementary geometric and distributional tests to capture both similarity and shift.

3.1 DATASET

We curate a paired image dataset to enable controlled cross-species comparisons on the same scenes. POV(point-of-view) videos of domestic cats with a camera strapped to their neck are collected from the internet and temporally aligned and decomposed into frames; images are then paired at the filename level to ensure one-to-one correspondences. Pairs with missing counterparts, corrupted files, or non-RGB formats are excluded. Each pair is assigned a stable identifier for reproducibility and joined consistently across outputs. Our curated dataset turns out to be 191 videos large, consisting of over 300,000 frame-pairs. To minimize confounds, we adopt three practices: (i) mirroring directory structures across domains so that every human frame has at most one feline counterpart; (ii) enforcing identical preprocessing pipelines per model family (resolution, normalization) to avoid input-induced artifacts in representational geometry; and (iii) evaluating only exact pairs for which both domains are present. This conservative pairing mitigates label leakage or scene-mismatch issues common in out-of-distribution evaluations (Taori et al., 2020; Koh et al., 2021).

We report all results with per-layer sample counts to make coverage explicit and avoid over-interpreting layers with few matched pairs. No personal identifiers are present, and images are drawn from public, in-the-wild recordings with standard research use; we analyze representational statistics and do not perform identity recognition. For each encoder, images are resized and normalized according to its canonical evaluation recipe (e.g., ImageNet-style center crop for convolutional networks; patch embedding resolutions for transformer encoders). Analyses are performed on the full paired corpus; for ablations we optionally downsample with a fixed random seed to a "golden" subset to test stability of conclusions. We avoid fine-tuning to isolate encoder invariances. Paired cross-domain datasets enable stronger causal interpretations of representational alignment than independent draws, as they hold scene content fixed while varying the domain. This design complements robustness benchmarks that evaluate distribution shifts without strict pairing (Hendrycks & Dietterich, 2019; Taori et al., 2020; Koh et al., 2021) and helps localize where along the processing hierarchy cross-species invariances are preserved.

We situate our corpus relative to existing resources involving animal or egocentric visual data. To our knowledge, public datasets rarely provide strictly paired, same-scene human-animal imagery; most focus on single-species egocentric capture or unpaired wildlife monitoring. Table 1 summarizes representative datasets and highlights how our paired, cross-species framing differs.

Table 1: Representative related datasets. Unlike these resources, our dataset provides strictly paired, same-scene images across human and feline domains to enable per-layer cross-species alignment analyses.

Dataset (year/venue)	Species	Modality	Paired same-scene?	Key difference vs. ours
CatCam (Betsch et al., 2004)	Cat	Egocentric video	No	Natural egocentric feline videos; no human counterpart; used for natural video statistics.
DogCentric Activity (Iwashita et al., 2014)	Dog	Egocentric video	No	First-person animal activity recognition; single species, unpaired with humans.
EgoPet (Bar et al., 2024)	Pet animals	Egocentric video	No	Animal egocentric interactions/locomotion benchmarks; not cross-species paired.
EGO4D (Grauman et al., 2022)	Human	Egocentric video	No	Large-scale human egocentric dataset; provides human-only perspective and tasks.
WILDS iWildCam (Koh et al., 2021)	Wildlife	Camera traps	No	Distribution shifts across camera locations and time; not paired same-scene, different sensing modality.

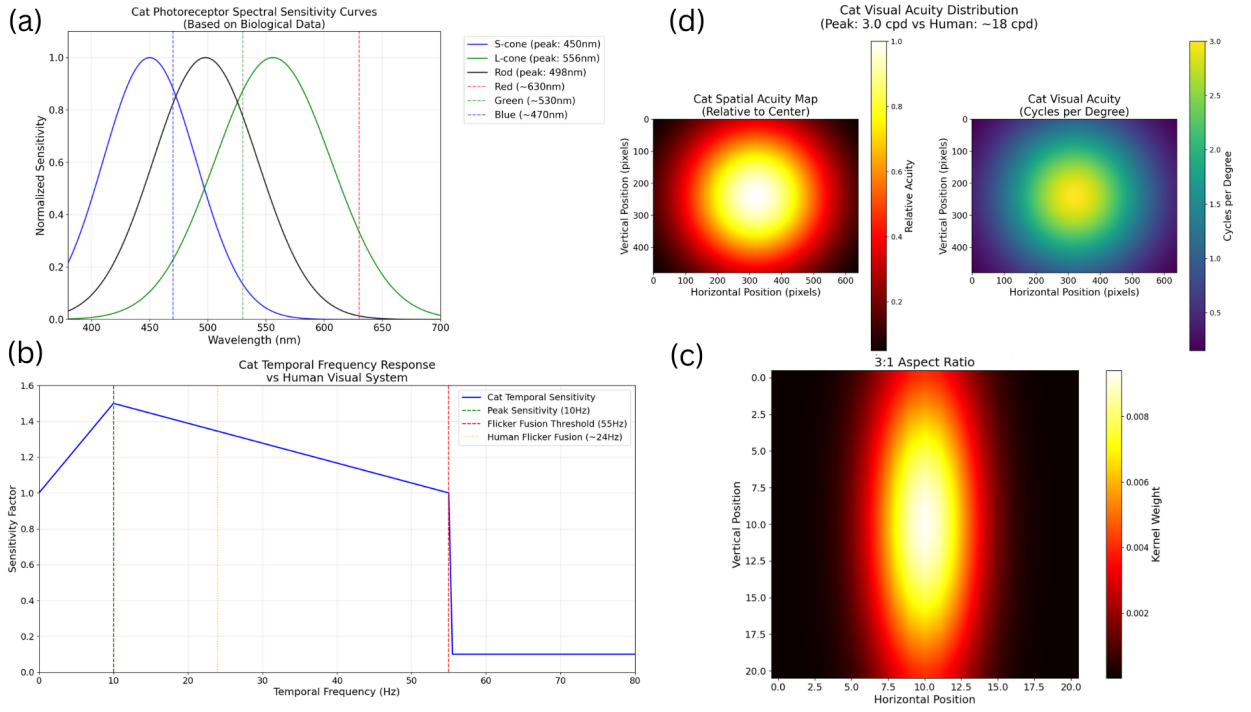

Figure 2: Moving counter-clockwise, Panel (a) depicts the cat's photoreceptor spectral sensitivity curves based on biological data, Panel (b) depicts the temporal frequency response of cat vs. human visual system, Panel (c) shows the cat's vertical slit pupil kernel in 3:1 aspect ratio. (d) shows the cat's spectral and visual acuity map based on our biologically informed implementation of cat's pupil, and Panel

3.2 Biologically Informed Cat-Vision Filter

To probe whether architectural invariances persist under species-specific optics and early vision, we apply a biologically informed image transformation that approximates key feline visual characteristics prior to feature extraction. The transformation models (i) spectral sensitivity with rod dominance and reduced long-wavelength sensitivity; (ii) spatial acuity and peripheral falloff; (iii) extended field-of-view distortions; (iv) temporal sensitivity and elevated flicker fusion; (v) motion sensitivity with horizontal bias; (vi) vertical-slit pupil optics; and (vii) a tapetum lucidum low-light enhancement. We interleave each component with its mathematical definition and refer back to equations where relevant in the Appendix A. We instantiate the filter with the following defaults unless otherwise stated: pupil aspect

Table 2: Summary of reported metrics and statistical controls.

Report	Contents	Notes
Layer-wise alignment	CKA (linear, RBF); RSA Spearman; Mantel r/p	500 permutations for Mantel
Distribution shift	MMD statistic/p; Energy statistic/p; 1-Wasserstein	Median-heuristic bandwidth for MMD
Paired stability	Mean cosine, mean L_2; test statistic/p; test type	Shapiro-Wilk for test selection
Multiple testing	Raw p, q-value, rejection	Benjamini-Hochberg at 0.05

ratio 3:1 (vertical), rod:cone weighting 25:1, spectral peaks at (450, 556, 498) nm for S/L/rod, spatial acuity reduction equivalent to $\sim$1/6 of human high-contrast acuity (behavioral estimates in cats often range 3-9 cpd, with higher optical potential (Clark & Clark, 2013), field of view $\sim$200° horizontal/140° vertical, temporal sensitivity peaking near 10 Hz with reduced gain beyond 50 Hz, and enhanced weighting of horizontally directed motion.

3.3 Models, Feature Extraction, Analyses and Statistical Procedures

All encoders are used with their canonical pretrained weights and are not fine-tuned, in line with frozen-transfer evaluations that emphasize representation quality independent of adaptation. For convolutional encoders, we probe a small set of semantically comparable stages spanning early, middle, and late processing, plus the global pooling stage when available. For transformer encoders, we probe block-wise token representations (ViT) and stage outputs (Swin). To render representations commensurate across families, we standardize activations to vectors: global average pooling for feature maps and class token (or mean over tokens if no class token is defined) for tokenized outputs (Dosovitskiy et al., 2021; Liu et al., 2021). This adheres to common practice in transfer and representational analysis.

Self-supervised models trained with knowledge distillation and masked/teacher-student objectives (e.g., DINO/DINOv2) are included due to their strong frozen-transfer performance and emergent invariances (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). For completeness, we include both base and large capacity variants where available to assess whether capacity modulates cross-species alignment. All images are normalized according to each encoder's canonical evaluation recipe (e.g., ImageNet mean/variance for CNNs; processor-defined normalization for transformers). We batch inputs per domain to ensure identical batch statistics across human and feline images and report per-layer sample counts. Optional ablations operate on a fixed-size, deterministically sampled subset to measure stability. We quantify alignment between human and feline representations at each layer using complementary geometric and statistical tools, and we control for multiple comparisons across models, layers, and metrics. We interleave each method with its mathematical definition and refer to equations in Appendix B and t-SNE, UMAP visualizations (see Figure 4, 5, 6) in Appendix C, and provide a brief summary table of reported metrics and statistical controls in Table 2. Layer-wise and block-wise feature maps for the best performing models across all the model families are presented as Figures 7, 8 and 9 in Appendix C

4 Results

We evaluate cross-species representational alignment on frozen encoders across three families: CNNs (ResNet, MobileNet, DenseNet, EfficientNet, ConvNeXt), supervised transformers (ViT, Swin), and self-supervised transformers (DINO, DINOv2/v3). Following the methodology, we report layer-wise alignment aggregated per model via CKA (linear and RBF), RSA/Mantel, and distribution-shift and paired metrics, and we identify the best-performing layer or block by CKA.

Table 3 summarizes model-level aggregates from the three familes: CNNs, Transformers(ViTs) and DINOv1/v2/v3. We use mean RBF-CKA as the primary selection criterion, as it is sensitive to certain nonlinear correspondences while remaining stable in practice; linear-CKA and RSA trends are consistent.

Across families, the self-supervised Vision Transformer **DINO ViT-B/16** achieves the highest mean RBF-CKA (**0.8144**), closely followed by supervised **ViT-L/16** (**0.8057**). Among CNNs, **EfficientNet-B3** yields the strongest mean RBF-CKA (**0.7017**). These results indicate that transformer-based encoders, particularly self-supervised ViTs, preserve stronger cross-species invariances under our paired design.

Table 3: Model-level aggregates from overall summaries. For selection we prioritize mean RBF-CKA; best within each family and overall are bolded. Values are taken from the three overall summary CSVs.

Family	Model	Best layer/block	CKA-RBF (mean)	CKA-Linear (mean)	RSA (mean)	Mean cosine
CNN	EfficientNet-B3	stage5	**0.7017**	0.6371	0.5344	0.6308
CNN	ResNet-50	layer3	0.6902	0.6628	0.4876	0.6022
CNN	DenseNet-169	db3	0.6853	0.6166	0.5417	0.7036
CNN	EfficientNet-B1	stage5	0.6838	0.6389	0.5107	0.4939
CNN	ConvNeXt-L	stage1	0.5599	0.5355	0.5428	0.8292
Transformer (sup.)	ViT-L/16	block14	**0.8057**	0.7050	0.4647	0.5960
Transformer (sup.)	ViT-B/16	block8	0.7755	0.6840	0.5266	0.6943
Transformer (sup.)	Swin-B	stage3	0.4688	0.4269	0.3818	0.6110
Self-sup. (DINO)	DINO ViT-B/16	block0	**0.8144**	0.7446	0.6980	0.7995
Self-sup. (DINO)	DINO ViT-S/16	block0	0.7682	0.6991	0.6668	0.8384
Self-sup. (DINOv2)	DINOv2-Base	block0	0.7232	0.6082	0.5669	0.8454
Overall best	**DINO ViT-B/16**	block0	**0.8144**	0.7446	0.6980	0.7995

We make the following family-specific observations:

1. **CNNs.** EfficientNet variants perform strongly, with B3 (CKA-RBF mean 0.7017) leading, followed closely by ResNet-50 (0.6902) and DenseNet-169 (0.6853). Best alignment typically occurs at later blocks (e.g., EfficientNet-B3 stage5; ResNet-50 layer3), consistent with hierarchical convergence.
2. **Supervised transformers.** ViT-L/16 (0.8057) outperforms Swin variants by a wide margin; best alignment arises at deeper transformer blocks (block14 for ViT-L/16; block8 for ViT-B/16).
3. **Self-supervised transformers.** DINO ViT-B/16 achieves the highest overall alignment (0.8144); DINOv2-Base is strong (0.7232), while DINOv3 pretrain variants show moderate alignment in our setting.

Mantel permutation tests, MMD, and Energy distance frequently reject the null across layers (see summary CSV counts), confirming measurable distributional differences even when alignment is high. Nevertheless, the leading models maintain robust cross-domain alignment by CKA and RSA, suggesting shape/semantic consistency despite domain shifts.

Complete per-model tables for each family, including all reported aggregates are provided in Tables 4, 5, 6 in Appendix D. We also investigate layers with most dissimilarities. See Tables 7, 8, 9 in Appendix E Beyond alignment, we systematically localize layers with strongest dissimilarity signals using low CKA/RSA and high projected 1D Wasserstein. Three robust patterns emerge across families: (i) *lowest alignment concentrates early*: initial CNN convolutions (e.g., ResNet/conv1) and the earliest blocks in ViT/DINO variants tend to have the lowest CKA-Linear and RSA; (ii) *distributional shift can peak late*: deeper EfficientNet stages and late ViT blocks exhibit the largest Wasserstein shifts while still retaining moderate alignment; (iii) *self-supervised giants can decouple geometry and distribution*: DINOv3 large models show very high late-block Wasserstein despite competitive CKA/RSA. Taken together, early layers are the most geometry-dissimilar, whereas certain deep layers carry the largest magnitude/distribution differences.

4.1 Discussion

We investigated cross-species representational invariance using frozen encoders on strictly paired human-feline views, triangulating alignment (CKA linear/RBF, RSA/Mantel), distributional shift (MMD, Energy, projected 1D Wasser-

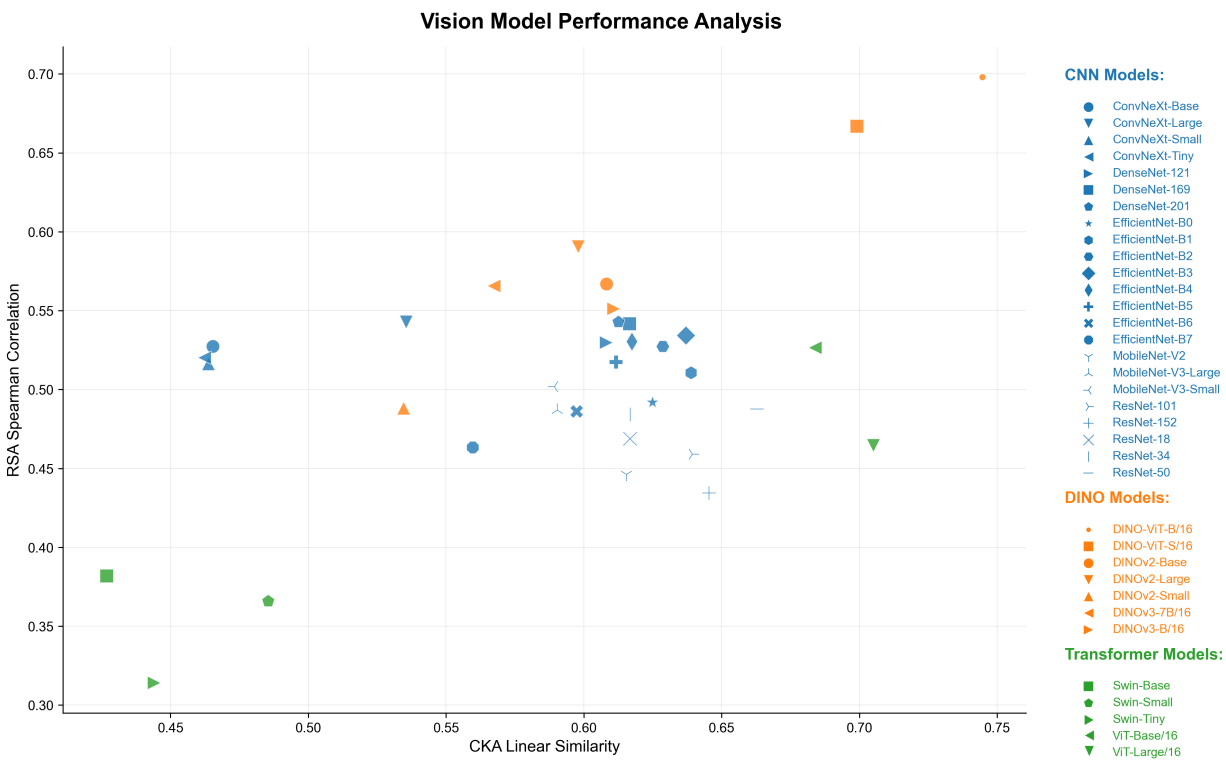

Figure 3: DINO models cluster in the upper-right region (high performance on both metrics), with DINO-ViT-B/16 achieving the highest RSA Spearman performance (0.698). **CNN** models form a tight cluster in the middle region, with EfficientNet variants generally performing better than ResNet and DenseNet models on both metrics. **Transformer** models show high variability, with ViT models (ViT-B/16, ViT-L/16) achieving high CKA Linear performance but moderate RSA Spearman scores, while Swin transformers cluster in the lower-left region.

stein), and paired similarity (cosine, L2) to characterize architectural behavior across domains. We summarize findings, likely causes, implications, scope choices, and future directions.

Transformers, especially self-supervised ViTs show the strongest alignment: DINO ViT-B/16 has the highest mean RBF-CKA (0.8144), followed by supervised ViT-L/16 (0.8057); the top CNN is EfficientNet-B3 (0.7017). Linear CKA means mirror this (DINO ViT-B/16: 0.7446; ViT-L/16: 0.7050; EfficientNet-B3: 0.6371), as do RSA/Mantel (0.698; 0.465; 0.534). Three factors plausibly drive these trends: (i) token-based global attention supports shape and part-whole consistency; (ii) invariance-focused self-supervision (DINO/DINOv2/v3) avoids human-label biases; and (iii) capacity/depth (e.g., ViT-L/16) enables abstractions aligning semantic manifolds. DINO ViT-B/16's RBF-CKA edge over supervised ViTs underscores the role of the objective beyond architecture.

Best layers differ by family. Supervised ViTs peak deeper (ViT-L/16 at block14; ViT-B/16 at block8). DINO often peaks at block0 for linear CKA, suggesting early, domain-general structure under our preprocessing. CNNs peak late (EfficientNet-B3 at stage5; ResNet-50 at layer3), consistent with hierarchical abstraction.

Despite high CKA/RSA, shift tests frequently reject after BH-FDR. CNNs show large Energy means (EfficientNet-B3: 20.0463) and projected 1D Wasserstein (EfficientNet-B3: 26.6422); supervised ViTs show moderate projected Wasserstein (ViT-B/16: 6.4436; ViT-L/16: 34.3499); DINO ranges widely (DINO ViT-B/16: 11.1702; DINOv3 ViT-7B/16 pretrain: 263.1593). MMD/Energy rejections are common (CNNs: 7-9; supervised ViTs: up to 25; DINO/DINOv3: teens or higher). Geometry metrics (CKA/RSA) are invariant to uniform scaling/affine transforms, whereas MMD/Energy/Wasserstein probe absolute distribution; cosine is magnitude-invariant, L2 is not. ConvNeXt variants exemplify this (convnext_large mean cosine: 0.8292 with larger L2), indicating similar directions but norm shifts, likely from domain-specific factors (e.g., illumination).

EfficientNet-B3 is the top CNN by mean RBF-CKA (0.7017), edging ResNet-50 (0.6902) and DenseNet-169 (0.6853); late-layer alignment (stage5/layer3) is typical. CNNs often have larger Energy/Wasserstein, consistent with mass/magnitude shifts. MobileNet variants are competitive for size but trail EfficientNet/ResNet. ViT-L/16 (0.8057)

surpasses ViT-B/16 (0.7755) and Swin (∼0.46-0.50). Best alignment at deeper blocks (block14 for ViT-L/16) supports semantics-driven consistency. Swin's windowed attention is efficient but less aligned cross-species than global ViTs here. DINO ViT-B/16 (0.8144) leads overall; DINOv2-Base (0.7232) is competitive; DINOv3 pretrains show moderate alignment but sometimes very large shift statistics (e.g., projected Wasserstein 263.1593 for ViT-7B/16 pretrain), indicating substantial distributional differences despite geometric agreement.

Three implications follow. First, training objective matters: self-supervised ViTs consistently top mean RBF-CKA (e.g., DINO ViT-B/16), yielding cross-species alignment without labels. Second, hierarchy matters: deeper blocks/stages align best, guiding feature extraction for cross-domain stability. Third, geometry and distribution are complementary: prioritize geometric alignment, then apply lightweight calibration (feature normalization, small adapters) to reduce residual MMD/Energy/Wasserstein shifts without altering geometry.

We freeze all encoders and use strictly paired human-feline views to isolate architectural/pretraining biases, report aggregates (means/medians; best layers/blocks) for stability, and include projected 1D Wasserstein alongside MMD/Energy for interpretability. These choices localize where alignment arises along the hierarchies. Together, these results indicate that training objective, architectural hierarchy, and capacity jointly shape cross-species representational invariance, and that aligning geometry and distribution requires complementary tools.

A primary reason for investigation of the feline visual system was the apparent difference in the retinal structure between humans and cats and also the initial works by Hubel and Wiesel on cat neuronal fields that ushered in the study of the visual system (Hubel & Wiesel, 1959). Neuroscientifically one would assume that given the differences in the retinal structure, gray matter volume, brain folding patterns, layer differences amongst animal and human models (Krubitzer, 2009; Yin et al., 2025), condensed brain hierarchy in small mammalian models (Lanfranchi et al., 2025), we would see a striking difference in cat representation of the world but apparently that's not the case. The hierarchical organization of the visual system results in invariant and stable representations of the perceived world which highlights the convergent evolutionary characteristics of the visual model. Future work can build directly on this foundation in several directions that are complementary to our goals here. (i) *Species generalization.* Extending the paired protocol to additional animals (e.g., canine, nonhuman primate, avian) would test how architectural and objective choices scale with ecological and optical diversity and evolutionary goals; this creates opportunities to connect to systems neuroscience by comparing model RDMs to neural data across species and cortical hierarchies. (ii) *Visual deficiencies.* We can use such an approach to investigate visual deficiences like partial blindness, myopia and hypermetropia and how corrections at different stages of visual development can impact visual processing and detection.(iii) *Non-visual sensory modalities.* This line of work could be extended to incorporate differences in auditory, tactile and olfactory senses across animal models.

Code of Ethics

We, the authors of this paper, hereby acknowledge and commit to the ICLR Code of Ethics. We strive to contribute to society and human well-being; uphold high standards of scientific excellence with transparent, reproducible methods; avoid harm to people, society, and the environment; act with honesty and transparency; and promote fairness and non-discrimination. We respect privacy, confidentiality, and the work of others; our datasets have faces blurred with no publicly identified information; and have steps to mitigate any unintended harms to the best of our knowledge.

Reproducibility statement

For our submission we provide the complete code with all the commands to run and reproduce the results we got as part of our study. Since the code execution might be too heavy compute-wise, we even provide a golden-subset of our dataset such that the results reported in the paper closely match without running the complete extraction and analysis of 191 videos converted into roughly 346,400 frame-pairs (692,8000 frames). The complete source code is provided in the '.zip' folder with a README clearly laying out instructions to setup the environment and commands to run step-by step. On acceptance the full dataset (approx. 110 GB), along with the source code shall be made public.

References

D. P. Andrews and P. Hammond. Mesopic increment threshold spectral sensitivity of single optic tract fibres in the cat: cone—rod interaction. *The Journal of Physiology*, 209(1):65–81, July 1970. ISSN 1469-7793. doi: 10.1113/jphysiol.1970.sp009156. URL http://dx.doi.org/10.1113/jphysiol.1970.sp009156.

Martin S. Banks, William W. Sprague, Jürgen Schmoll, Jared A. Q. Parnell, and Gordon D. Love. Why do animal eyes have pupils of different shapes? *Science Advances*, 1(7), August 2015. ISSN 2375-2548. doi: 10.1126/sciadv.1500391. URL http://dx.doi.org/10.1126/sciadv.1500391.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. URL https://arxiv.org/abs/2106.08254.

Amir Bar, Arya Bakhtiar, Danny Tran, Antonio Loquercio, Jathushan Rajasegaran, Yann LeCun, Amir Globerson, and Trevor Darrell. Egopet: Egomotion and interaction data from an animal's perspective, 2024. URL https://arxiv.org/abs/2404.09991.

Belinda Y. Betsch, Wolfgang Einhäuser, Konrad P. Kröding, and Peter König. The world from a cat's perspective? statistics of natural videos. *Biological Cybernetics*, 90(1):41–50, January 2004. ISSN 1432-0770. doi: 10.1007/s00422-003-0434-6. URL http://dx.doi.org/10.1007/s00422-003-0434-6.

Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLOS Computational Biology*, 15(4):e1006897, April 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006897. URL http://dx.doi.org/10.1371/journal.pcbi.1006897.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

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*, ICML'20. JMLR.org, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020b. URL https://arxiv.org/abs/2003.04297.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, January 2014. ISSN 1546-1726. doi: 10.1038/nn.3635. URL http://dx.doi.org/10.1038/nn.3635.

Daria L. Clark and Robert A. Clark. The effects of time, luminance, and high contrast targets: Revisiting grating acuity in the domestic cat. *Experimental Eye Research*, 116:75–78, November 2013. ISSN 0014-4835. doi: 10.1016/j.exer.2013.08.004. URL http://dx.doi.org/10.1016/j.exer.2013.08.004.

J. A. Coles. Some reflective properties of the tapetum lucidum of the cat's eye. *The Journal of Physiology*, 212(2):393–409, January 1971. ISSN 1469-7793. doi: 10.1113/jphysiol.1971.sp009331. URL http://dx.doi.org/10.1113/jphysiol.1971.sp009331.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022. URL https://arxiv.org/abs/1811.12231.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. URL https://arxiv.org/abs/2110.07058.

Arthur Gretton, Karsten Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander J. Smola. A kernel method for the two-sample problem, 2008. URL https://arxiv.org/abs/0805.2368.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

E Guenther and E Zrenner. The spectral sensitivity of dark- and light-adapted cat retinal ganglion cells. *The Journal of Neuroscience*, 13(4):1543–1550, April 1993. ISSN 1529-2401. doi: 10.1523/jneurosci.13-04-01543.1993. URL http://dx.doi.org/10.1523/jneurosci.13-04-01543.1993.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples, 2021. URL https://arxiv.org/abs/1907.07174.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3, 2019. URL https://arxiv.org/abs/1905.02244.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. URL https://arxiv.org/abs/1704.04861.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. URL https://arxiv.org/abs/1608.06993.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, Oct 1959. ISSN 0022-3751. doi: 10.1113/jphysiol.1959.sp006308.

Y. Iwashita, A. Takamine, R. Kurazume, and M. S. Ryoo. First-person animal activity recognition from egocentric videos. In *International Conference on Pattern Recognition (ICPR)*, Stockholm, Sweden, August 2014.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2021. URL https://arxiv.org/abs/2012.07421.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019. URL https://arxiv.org/abs/1905.00414.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, Volume 2 - 2008, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008.

Leah Krubitzer. In search of a unifying theory of complex brain evolution. *Annals of the New York Academy of Sciences*, 1156:44–67, Mar 2009. ISSN 0077-8923. doi: 10.1111/j.1749-6632.2009.04421.x.

Frank F. Lanfranchi, Joseph Wekselblatt, Daniel A. Wagenaar, and Doris Y. Tsao. A compressed hierarchy for visual form processing in the tree shrew. *Nature*, August 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09441-w. URL https://doi.org/10.1038/s41586-025-09441-w.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL https://arxiv.org/abs/2103.14030.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. URL https://arxiv.org/abs/2201.03545.

M S Loop, C L Millican, and S R Thomas. Photopic spectral sensitivity of the cat. *The Journal of Physiology*, 382(1): 537–553, January 1987. ISSN 1469-7793. doi: 10.1113/jphysiol.1987.sp016383. URL http://dx.doi.org/10.1113/jphysiol.1987.sp016383.

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27 (2_Part_1):209–220, 02 1967. ISSN 0008-5472.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, September 2018. ISSN 2475-9066. doi: 10.21105/joss.00861. URL http://dx.doi.org/10.21105/joss.00861.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a7a3d70c6d17a73140918996d03c014f-Paper.pdf.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, 2017. URL https://arxiv.org/abs/1706.05806.

Maria L. Rizzo and Gábor J. Székely. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8, 2016. URL https://api.semanticscholar.org/CorpusID:128185426.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019. URL https://arxiv.org/abs/1801.04381.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10, 1966. URL https://api.semanticscholar.org/CorpusID:121676935.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2018. URL https://api.semanticscholar.org/CorpusID:91917265.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2105646118.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.11946.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021. URL https://arxiv.org/abs/2012.12877.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200, 2016. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.12.012. URL https://www.sciencedirect.com/science/article/pii/S1053811915011258.

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, February 2016. ISSN 1546-1726. doi: 10.1038/nn.4244. URL http://dx.doi.org/10.1038/nn.4244.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111. URL https://www.pnas.org/doi/abs/10.1073/pnas.1403112111.

Sifan Yin, Chunzi Liu, Gary P. T. Choi, Yeonsu Jung, Katja Heuer, Roberto Toro, and L. Mahadevan. Morphogenesis and morphometry of brain folding patterns across species. *bioRxiv*, 2025. doi: 10.1101/2025.03.05.641692. URL https://www.biorxiv.org/content/early/2025/03/10/2025.03.05.641692.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022. URL https://arxiv.org/abs/2111.07832.

E Zrenner and P Gouras. Blue-sensitive cones of the cat produce a rodlike electroretinogram. *Investigative Ophthalmology & Visual Science*, 18(10):1076–1081, 10 1979. ISSN 1552-5783.

A Math Behind Cat Vision Filter Implementation

Spectral sensitivity is modeled by smooth sensitivity curves for short- and long-wavelength cones and rods, with peaks near 450 nm, 550-556 nm, and ≈498-501 nm, respectively, consistent with electrophysiology of feline retina and ganglion cells (Andrews & Hammond, 1970; Loop et al., 1987; Guenther & Zrenner, 1993). We define Gaussian-like sensitivity functions $S(\lambda), L(\lambda), R(\lambda)$ (equation 1,equation 2,equation 3) and compute cone/rod activations per pixel, then combine them with rod-dominant weights (equation 7). This blending reflects nocturnal/crepuscular behavior and attenuated red sensitivity (Zrenner & Gouras, 1979).

$$S(\lambda) = \exp\Big(-\tfrac{(\lambda-\mu_S)^2}{2\sigma_S^2}\Big), \quad \mu_S \approx 450\,\text{nm}, \tag{1}$$

$$L(\lambda) = \exp\Big(-\tfrac{(\lambda-\mu_L)^2}{2\sigma_L^2}\Big), \quad \mu_L \approx 556\,\text{nm}, \tag{2}$$

$$R(\lambda) = \exp\Big(-\tfrac{(\lambda-\mu_R)^2}{2\sigma_R^2}\Big), \quad \mu_R \approx 498{-}501\,\text{nm}. \tag{3}$$

With nominal RGB wavelengths $(\lambda_R, \lambda_G, \lambda_B)$, per-pixel activations are

$$C_S(x) = w_B\, S(\lambda_B)\, B(x) + w_G\, S(\lambda_G)\, G(x) + w_R\, S(\lambda_R)\, R(x), \tag{4}$$

$$C_L(x) = w_B\, L(\lambda_B)\, B(x) + w_G\, L(\lambda_G)\, G(x) + w_R\, L(\lambda_R)\, R(x), \tag{5}$$

$$C_R(x) = w_B\, R(\lambda_B)\, B(x) + w_G\, R(\lambda_G)\, G(x) + w_R\, R(\lambda_R)\, R(x). \tag{6}$$

Rod dominance is enforced by

$$I_{spec}(x) = \beta_S C_S(x) + \beta_L C_L(x) + \beta_R C_R(x), \quad \beta_R : \beta_S + \beta_L \approx 25:1, \tag{7}$$

and the tapetum lucidum (Coles, 1971) is approximated downstream via a luminance-dependent gain with a blue-green tint (Eq. equation 15).

Spatial acuity is reduced with a frequency-domain low-pass filter to approximate lower feline grating acuity. Specifically, we use a Gaussian transfer function $H(u, v)$ (Eq. equation 8) and reconstruct the filtered image by inverse Fourier transform, choosing σ_{lp} to achieve ≈1/6 of human high-contrast acuity.

$$H(u,v) = \exp\Big(-\tfrac{u^2+v^2}{2\sigma_{lp}^2}\Big), \qquad I_{sp}(x) = \mathcal{F}^{-1}\{H \cdot \mathcal{F}\{I_{spec}\}\}(x). \tag{8}$$

Geometric optics are approximated by a vertical-slit pupil and a broadened field of view. Vertical slits are linked to ambush predation and depth-of-field control (Banks et al., 2015). We include a mild barrel distortion and a center-surround acuity mask (Eqs. equation 9-equation 10) to emulate wider effective field and peripheral blur.

$$r' = r\,(1 + k_1 r^2 + k_2 r^4), \quad \tilde{x}' = \tfrac{r'}{\max(r,\,\epsilon)}\,\tilde{x}, \quad x' = \Pi(\tilde{x}'), \tag{9}$$

$$A(r) = \frac{1}{1 + \exp\big(\gamma\,(r - r_0)\big)}, \qquad I_{fov}(x) = A(r(x))\,I_{sp}(x'). \tag{10}$$

Temporal processing emphasizes motion sensitivity in the ∼10 Hz band, with reduced gain beyond ≈50-60 Hz. We apply a temporal band-pass gain $G(f)$ in the Fourier domain (Eq. equation 11), with peak $f_0 \approx 10\,\text{Hz}$ and flicker-fusion cutoff $f_{ff} \approx 55\,\text{Hz}$.

$$I_{tmp}(t) = \mathcal{F}_t^{-1}\big\{G(f)\,\widehat{I}(f)\big\}(t), \qquad G(f) = \exp\Big(-\tfrac{(f-f_0)^2}{2\sigma_f^2}\Big)\,\mathbf{1}[f \le f_{ff}]. \tag{11}$$

Motion sensitivity includes a horizontal bias. We estimate optical flow (u, v) via Lucas-Kanade (Eq. equation 12) and modulate flow magnitude by $\eta(\theta) = 1 + \kappa|\cos\theta|$ (Eq. equation 13) before blending into the image stream.

$$(u,v) = \arg\min_{u,v} \sum_{y \in W(x)} \big(I_x(y)u + I_y(y)v + I_t(y)\big)^2, \tag{12}$$

$$M(x) = \eta(\theta(x))\,\sqrt{u(x)^2 + v(x)^2}, \qquad I_{mot}(x) = I_{tmp}(x) + \lambda_M\,\mathrm{norm}(M(x)). \tag{13}$$

Finally, low-light enhancement consistent with the tapetum lucidum is implemented as a logistic gain with a blue-green tint (Eq. equation 15).

$$g(I) = 1 + \alpha\,\sigma\big(\beta(\tau - \bar{I})\big), \qquad \bar{I} = \tfrac{1}{3}(R{+}G{+}B), \tag{14}$$

$$I_{tap}(x) = T\big(\,g(I_{mot}(x)) \cdot I_{mot}(x)\,\big), \qquad T = \mathrm{diag}(t_R, t_G, t_B),\ t_G \gtrsim t_B \ge t_R, \tag{15}$$

with T imparting a mild blue-green bias; the final output is $I_{cat} = I_{tap}$.

We acknowledge certain limitations and present the transformation as an engineering approximation, not a full optical-retinal-cortical simulator. It omits wavelength-dependent blur, detailed retinal sampling mosaics, chromatic aberrations, and dynamic pupil control. As such, it should be interpreted as a near-accurate biologically motivated stressor for invariance analyses rather than a physiologically complete forward model.

B Statistical Procedures

Representational geometry. We report Centered Kernel Alignment (CKA) in both its linear form and with an RBF kernel. Linear CKA is invariant to orthogonal transformations and isotropic scaling and has been shown to be stable across architectures and training runs (Kornblith et al., 2019). Given paired samples $X = \{x_i\}_{i=1}^n$, $Y = \{y_i\}_{i=1}^n$, linear CKA normalizes HSIC between centered Gram matrices (Eq. equation 16).

$$\mathrm{CKA}_{\text{lin}}(X,Y) = \frac{\mathrm{HSIC}(K,L)}{\sqrt{\mathrm{HSIC}(K,K)\ \mathrm{HSIC}(L,L)}}, \quad K = HXX^\top H,\ L = HYY^\top H, \tag{16}$$

with $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $\mathrm{HSIC}(K,L) = \frac{1}{(n-1)^2}\,\mathrm{tr}(KL)$. RBF CKA replaces K, L by RBF kernels (Eq. equation 17).

$$K_{ij} = \exp\Big(-\tfrac{\|x_i - x_j\|^2}{2\sigma^2}\Big), \quad L_{ij} = \exp\Big(-\tfrac{\|y_i - y_j\|^2}{2\sigma^2}\Big), \quad \mathrm{CKA}_{\text{rbf}} \text{ as in Eq. } equation\ 16. \tag{17}$$

In parallel, we conduct Representational Similarity Analysis (RSA) by computing cosine-based representational dissimilarity matrices (RDMs) per domain (Eq. equation 18) and correlating the upper triangles using Spearman rank correlation. Significance is assessed with the Mantel permutation test (Eq. equation 19).

$$D^X_{ij} = 1 - \frac{x_i^\top x_j}{\|x_i\|\,\|x_j\|}, \quad D^Y_{ij} = 1 - \frac{y_i^\top y_j}{\|y_i\|\,\|y_j\|}. \tag{18}$$

Let $u(\cdot)$ extract upper-triangular entries; the Mantel statistic is $r = \mathrm{corr}(u(D^X), u(D^Y))$, with p-values from permutations of indices (Eq. equation 19).

$$r = \frac{\sum_k (u(D^X)_k - \bar{u}_X)(u(D^Y)_k - \bar{u}_Y)}{\sqrt{\sum_k (u(D^X)_k - \bar{u}_X)^2 \, \sum_k (u(D^Y)_k - \bar{u}_Y)^2}}, \;\; p\text{-value by } \pi\text{-permutations}. \tag{19}$$

Distributional shift tests. To detect shifts beyond geometric alignment, we compute the Maximum Mean Discrepancy (MMD; RBF kernel), the Energy distance, and a projected 1-Wasserstein distance. For MMD we use the unbiased estimator with an RBF kernel (Eq. equation 20); p-values are obtained via label permutations. Energy distance is computed from pairwise Euclidean distances (Eq. equation 21). For interpretability, we project onto the first principal component and compute the 1D W_1 between the projected empirical distributions (Eq. equation 22).

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{m(m-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j), \tag{20}$$

$$k(u, v) = \exp\Big(- \tfrac{\|u-v\|^2}{2\sigma^2} \Big).$$

$$\widehat{\mathcal{E}}^2 = \frac{2}{mn} \sum_{i,j} \|x_i - y_j\|_2 - \frac{1}{m^2} \sum_{i,j} \|x_i - x_j\|_2 - \frac{1}{n^2} \sum_{i,j} \|y_i - y_j\|_2. \tag{21}$$

$$W_1 = \frac{1}{K} \sum_{k=1}^{K} \big| s_{(k)} - t_{(k)} \big|, \quad s_i = w^\top x_i,\; t_j = w^\top y_j,\; w = \text{first PC}. \tag{22}$$

Paired similarity. For each layer we compute per-pair cosine similarity and Euclidean distance between human and feline vectors. To test for mean shifts, we project differences onto the first principal component w of $\{x_i - y_i\}$ and test $H_0 : \mathbb{E}[d] = 0$ on $d_i = w^\top (x_i - y_i)$. If Shapiro-Wilk fails to reject normality, we report the paired t-statistic (Eq. equation 23); otherwise we report the Wilcoxon signed-rank statistic.

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}, \quad \bar{d} = \frac{1}{n} \sum_i d_i,\; s_d^2 = \frac{1}{n-1} \sum_i (d_i - \bar{d})^2. \tag{23}$$

Multiple testing. We aggregate all collected p-values and apply the Benjamini-Hochberg procedure (FDR level 0.05) to obtain q-values and rejection decisions (Eq. equation 24). This controls false discoveries across the large grid of model-layer-metric combinations. We also report q-values via the standard step-up estimator.

$$k = \max \Big\{ i : \; p_{(i)} \leq \tfrac{i}{M}\, q \Big\}, \qquad \text{reject } H_{(1)}, \ldots, H_{(k)}, \quad q_{(i)} = \min_{j \geq i} \frac{M}{j}\, p_{(j)}. \tag{24}$$

C VISUALIZATION

We employ qualitative and quantitative visualization to contextualize the statistical findings and to diagnose failure modes.

First, we project high-dimensional layer-wise representations to two dimensions using t-SNE (van der Maaten & Hinton, 2008) and UMAP McInnes et al. (2018). For each model and layer, we co-embed human and feline vectors and inspect whether clusters corresponding to domains separate or overlap. t-SNE perplexity is adapted to the sample size, and UMAP neighborhoods scale with data density; these settings provide stable 2D summaries while avoiding over-interpretation of local distances.

Second, we visualize intermediate activations for paired images. For convolutional layers, we render channel-wise activation maps after min-max normalization; for transformer blocks, we compute token saliency maps by averaging

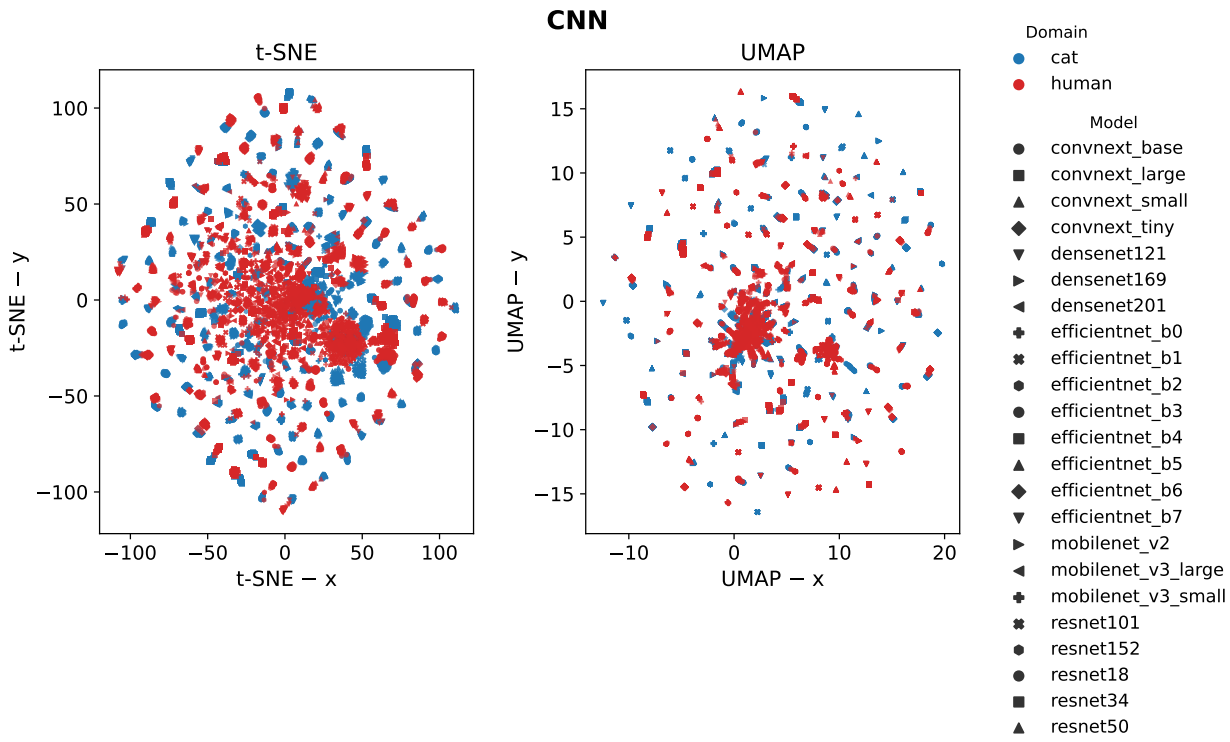

Figure 4: CNN embeddings with t-SNE (left) and UMAP (right). Colors encode domains (human vs. cat) and marker shapes encode models within the family. These panels are intended to assess domain-level overlap by visual inspection: color mixing indicates cross-domain similarity, while separated color clusters indicate stronger domain-specific structure; shape differences reveal whether such trends are consistent across CNN variants.

over the embedding dimension and reshaping to the token grid (dropping class tokens when appropriate). We display human and feline activations side-by-side to highlight convergences and divergences at the same layer.

These visualizations serve as diagnostic aids rather than formal tests: they illustrate patterns suggested by the alignment metrics (e.g., strong overlap in 2D projections when CKA is high) and reveal layer-wise phenomena such as texture vs. shape preferences or spatial pooling differences that may underlie quantitative effects.

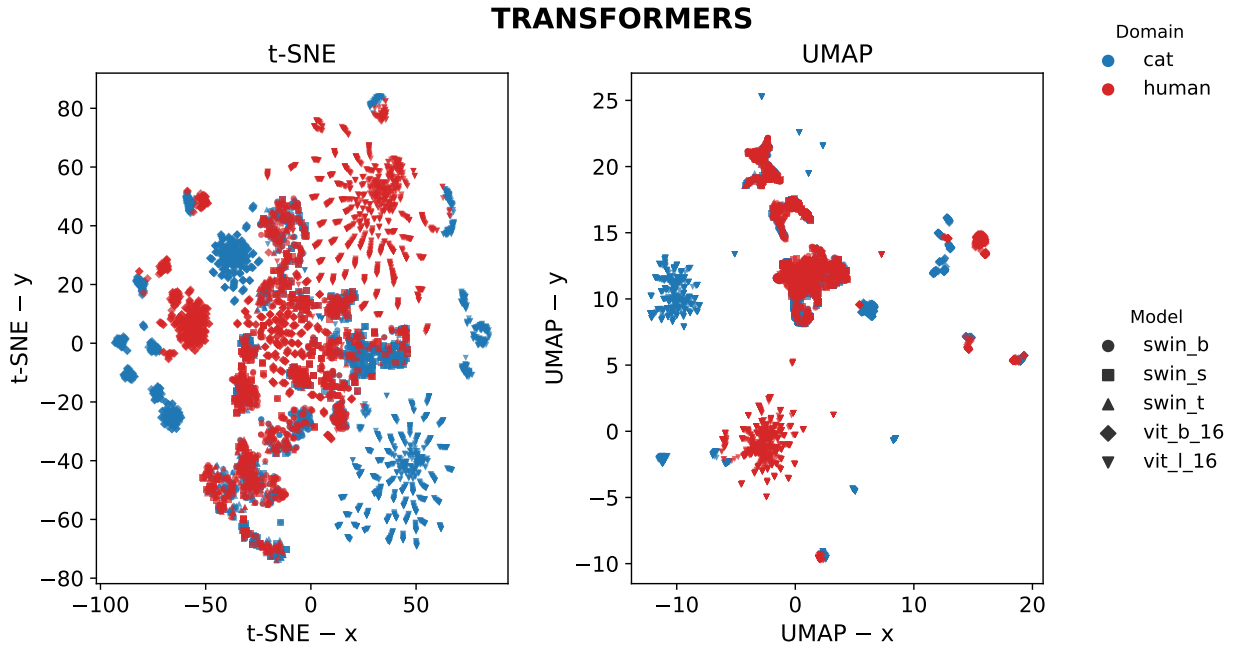

Figure 5: Transformer embeddings with t-SNE (left) and UMAP (right). Colors (domains) and marker shapes (models) follow Figure 4. Showing both t-SNE and UMAP allows a robustness check: consistent patterns across methods lend confidence, while differences may reflect method-specific neighborhood preservation.

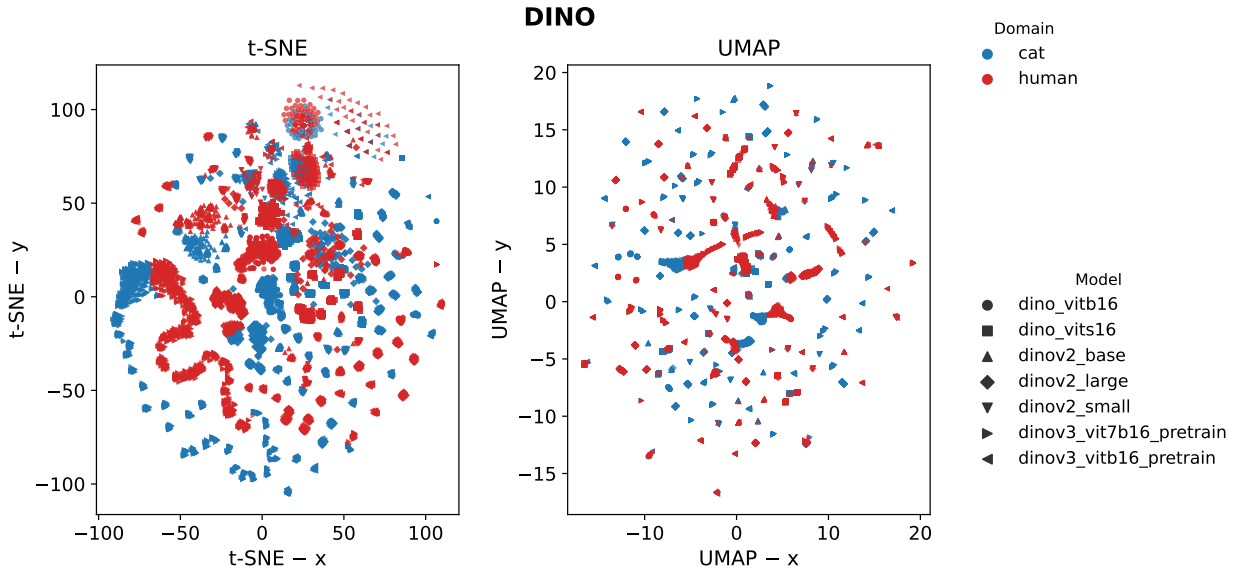

Figure 6: DINO embeddings with t-SNE (left) and UMAP (right). Colors denote domains; marker shapes denote DINO variants. Self-supervised representations often yield distinct geometry; these panels enable visual examination of domain separation vs. overlap and whether patterns are consistent across DINO variants.

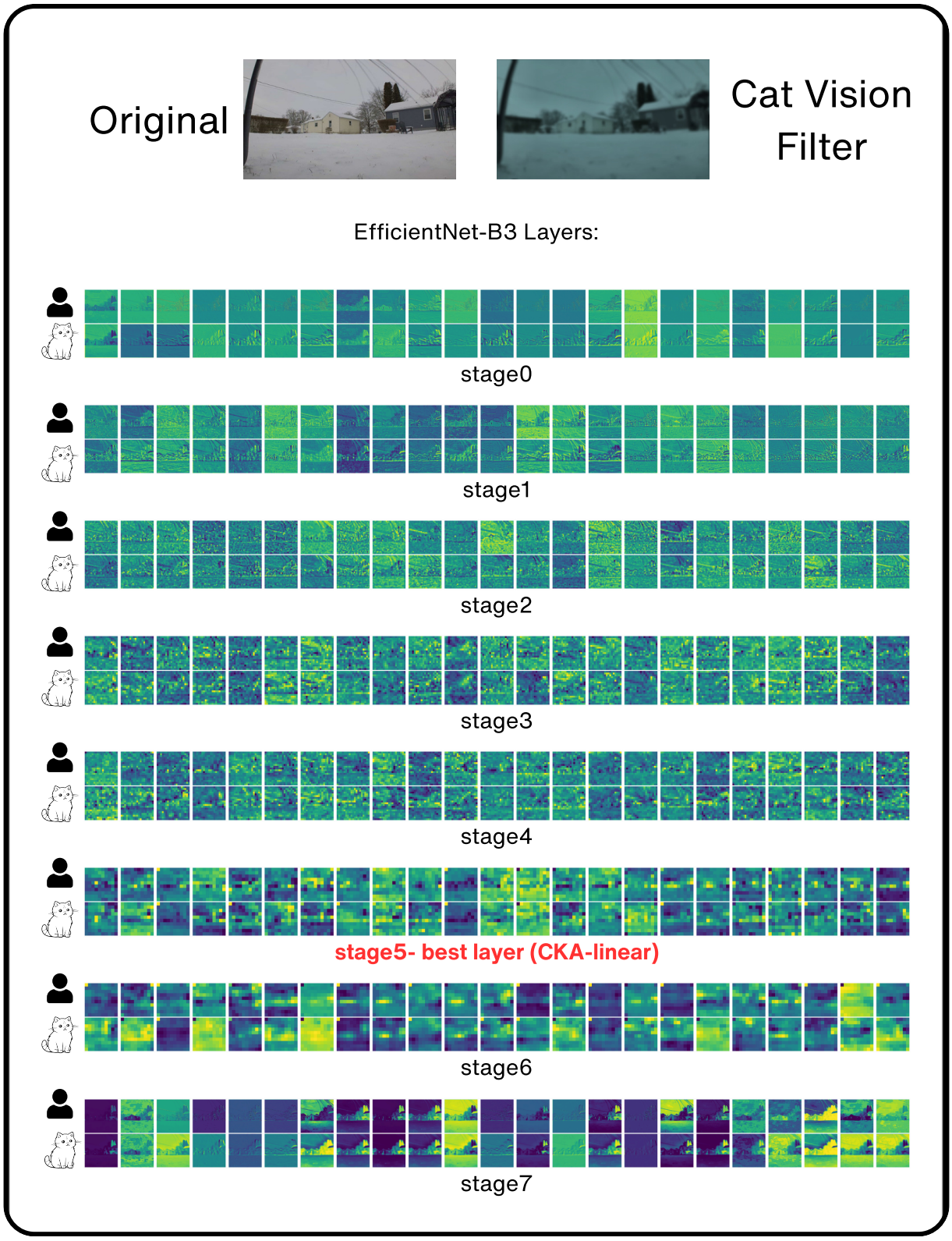

Figure 7: Layer-wise feature maps for the best performing model in CNN family; EfficientNet-B3 with best layer by CKA-linear: stage5

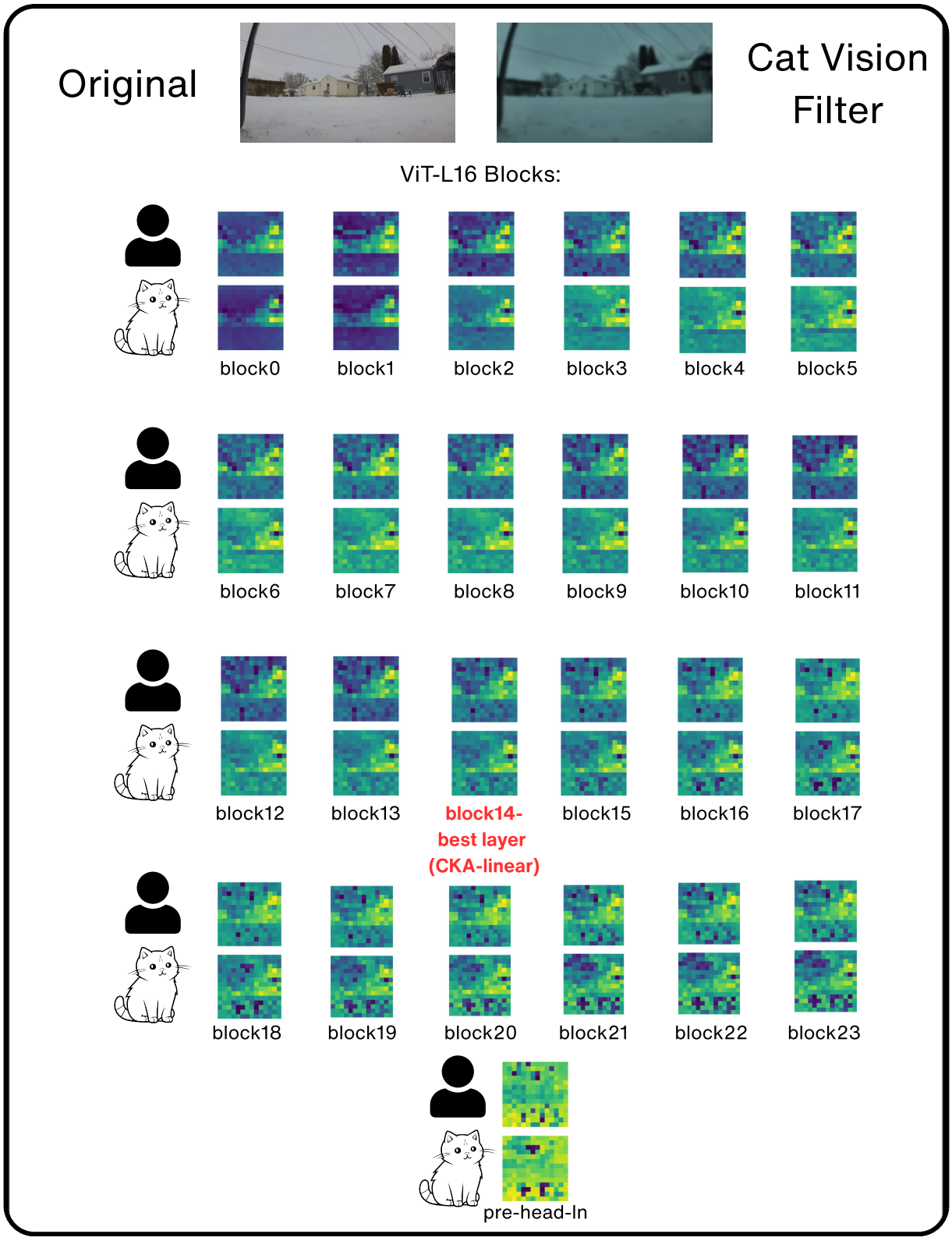

Figure 8: Block-wise visualization for the best performing model in Transformers family; ViT l16; best block by CKA-linear: block14

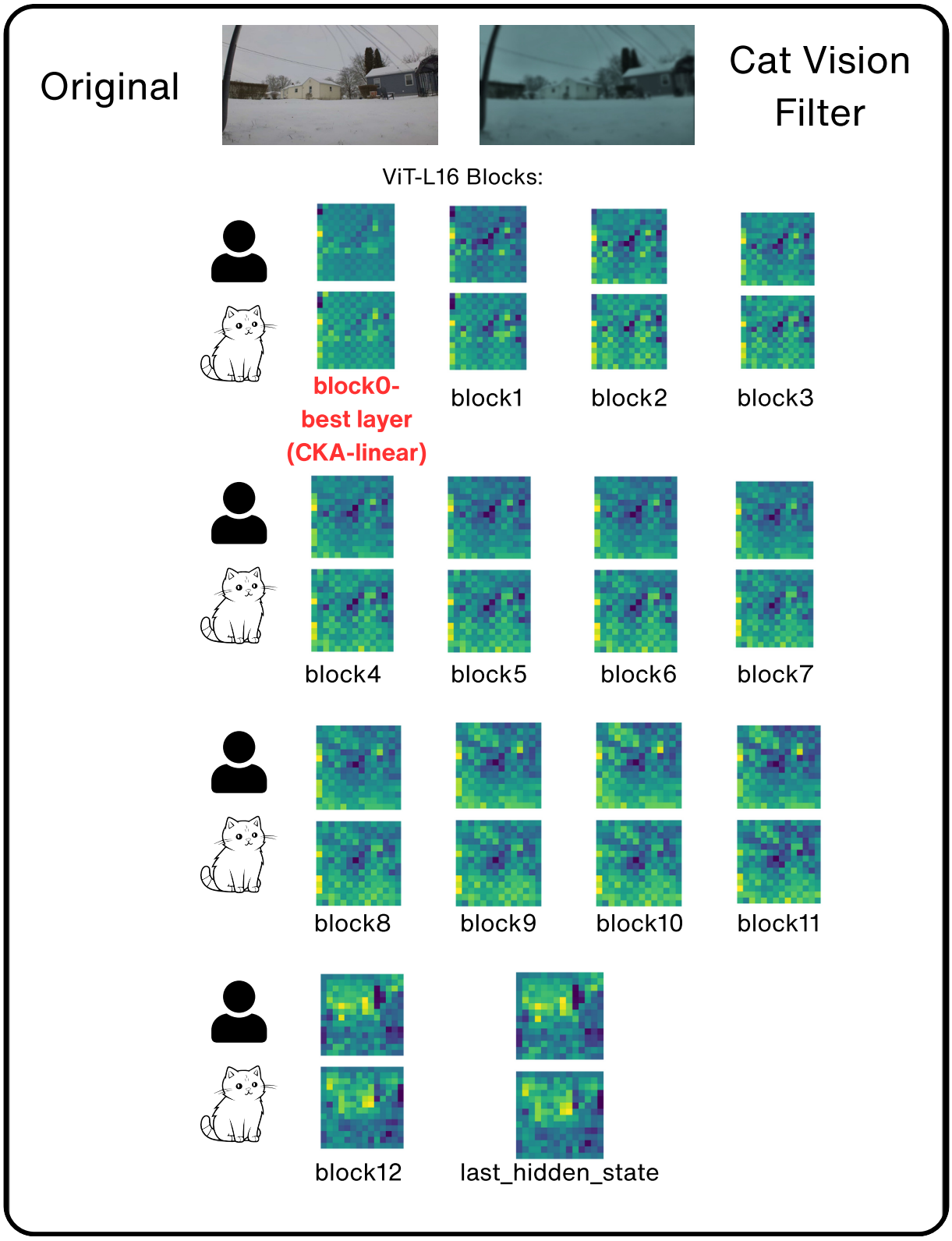

Figure 9: Block-wise visualization for the best performing model in DINO family; DINO ViT b16; best block by CKA-linear: block0

D Per Model Summary Tables

Table 4: EfficientNet-B3 achieved the highest mean RBF-CKA among CNNs (0.7017), with best alignment at stage5.

Model	Best layer	CKA-RBF mean	CKA-RBF max	CKA-Linear mean	RSA mean	Mean cosine	Mean L2
convnext_base	stage1	0.5212	0.6015	0.4653	0.5274	0.8612	1.3488
convnext_large	stage1	0.5599	0.6743	0.5355	0.5428	0.8292	2.1653
convnext_small	stage2	0.5142	0.6178	0.4638	0.5163	0.8108	0.8129
convnext_tiny	stage1	0.5473	0.6780	0.4624	0.5202	0.7473	0.9895
DenseNet-121	tr2	0.6776	0.8606	0.6081	0.5299	0.7090	4.6149
DenseNet-169	db3	0.6853	0.8721	0.6166	0.5417	0.7036	4.8734
DenseNet-201	tr3	0.6825	0.8870	0.6125	0.5429	0.7132	4.8342
EfficientNet-B0	stage4	0.6627	0.8560	0.6249	0.4920	0.4418	19.6399
EfficientNet-B1	stage5	0.6838	0.8813	0.6389	0.5107	0.4939	21.9126
EfficientNet-B2	stage4	0.6706	0.8743	0.6287	0.5273	0.5645	20.2296
EfficientNet-B3	stage5	0.7017	0.8743	0.6371	0.5344	0.6308	26.6422
EfficientNet-B4	stage5	0.6625	0.8591	0.6175	0.5304	0.6339	44.0907
EfficientNet-B5	stage3	0.6794	0.8393	0.6117	0.5175	0.7026	20.0516
EfficientNet-B6	stage6	0.6607	0.8553	0.5974	0.4862	0.7057	20.0586
EfficientNet-B7	stage3	0.6292	0.7974	0.5597	0.4634	0.7156	21.8801
mobilenet_v2	mid3	0.6517	0.8263	0.6154	0.4463	0.2946	16.1492
mobilenet_v3_large	final	0.6198	0.8597	0.5903	0.4871	0.3545	12.4279
mobilenet_v3_small	mid3	0.6664	0.8369	0.5893	0.5018	0.4325	4.4553
ResNet-18	layer2	0.6759	0.8370	0.6167	0.4689	0.7260	8.2918
ResNet-34	layer3	0.6730	0.8609	0.6169	0.4841	0.7324	9.1681
ResNet-50	layer3	0.6902	0.8988	0.6628	0.4876	0.6022	12.8899
ResNet-101	layer3	0.6787	0.9055	0.6394	0.4591	0.6153	16.1519
ResNet-152	layer3	0.6868	0.9081	0.6455	0.4344	0.6253	17.8186

Table 5: ViT-L/16 achieved the highest mean RBF-CKA among supervised transformers (0.8057), with best alignment at block14.

Model	Best block/stage	CKA-RBF mean	CKA-RBF max	CKA-Linear mean	RSA mean	Mean cosine	Mean L2
Swin-T	stage3	0.4624	0.5789	0.4440	0.3142	0.6217	0.4581
Swin-S	stage3	0.4993	0.5802	0.4854	0.3660	0.6723	0.5027
Swin-B	stage3	0.4688	0.6038	0.4269	0.3818	0.6110	0.3683
ViT-B/16	block8	0.7755	0.9251	0.6840	0.5266	0.6943	6.4436
ViT-L/16	block14	0.8057	0.9338	0.7050	0.4647	0.5960	34.3499

Table 6: DINO ViT-B/16 achieved the highest mean RBF-CKA among self-supervised transformers (0.8144), with best alignment at block0.

Model	Best block	CKA-RBF mean	CKA-RBF max	CKA-Linear mean	RSA mean	Mean cosine	Mean L2
DINO ViT-B/16	block0	0.8144	1.0000	0.7446	0.6980	0.7995	11.1702
DINO ViT-S/16	block0	0.7682	1.0000	0.6991	0.6668	0.8384	7.7151
DINOv2-Base	block0	0.7232	1.0000	0.6082	0.5669	0.8454	8.7336
DINOv2-Large	block0	0.7029	1.0000	0.5980	0.5906	0.8435	8.3610
DINOv2-Small	block0	0.6464	1.0000	0.5346	0.4881	0.9004	5.3112
DINOv3 ViT-B/16 (pretrain)	block0	0.7092	1.0000	0.6107	0.5513	0.8552	43.1978
DINOv3 ViT-7B/16 (pretrain)	block0	0.6969	1.0000	0.5673	0.5658	0.9360	263.1593

E Layers with Most Dissimilarity Across Families

We report layers/blocks exhibiting the strongest cross-domain dissimilarities, grounded in the per-layer analyses: (i) lowest alignment by CKA-Linear and RSA Spearman; and (ii) largest distributional shift by projected 1D Wasserstein (with MMD/Energy concordant).

Criterion	Model.Layer	Value
Lowest CKA-Linear	efficientnet_b3.stage1	0.2448
Lowest CKA-Linear	densenet201.conv0	0.2451
Lowest CKA-Linear	efficientnet_b7.stage1	0.2614
Lowest CKA-Linear	efficientnet_b2.stage1	0.2694
Lowest RSA Spearman	resnet34.conv1	0.0750
Lowest RSA Spearman	resnet18.conv1	0.0759
Lowest RSA Spearman	resnet50.conv1	0.1013
Largest W1	efficientnet_b4.stage4	74.662
Next largest W1	efficientnet_b4.stage6	58.455
Next largest W1	efficientnet_b3.stage6	30.629

Table 7: CNN layers with lowest alignment and highest shift.

Criterion	Model.Layer	Value
Lowest CKA-Linear	swin_b.stage7	0.0660
Lowest CKA-Linear	swin_t.stage2	0.3724
Lowest CKA-Linear	vit_l_16.block22	0.2814
Lowest RSA Spearman	swin_b.norm	0.0636
Lowest RSA Spearman	swin_t.stage2	0.0922
Lowest RSA Spearman	swin_b.stage7	0.0990
Largest W1	vit_l_16.block21-23	≈30.25-30.42

Table 8: Supervised transformer layers with lowest alignment and highest shift.

Across families, early convolutional layers (e.g., ‘conv1‘, ‘conv0‘, ‘stage1‘) show the lowest RSA/CKA, while later transformer blocks (ViT-L/16 blocks 19-23) combine high CKA with sizable distributional shifts by Wasserstein. DINOv3 ViT-7B/16 pretrain exhibits extreme late-block shifts, despite nontrivial geometric alignment; an informative dissociation to consider when selecting layers for analysis or adaptation.

Criterion	Model.Layer	Value
Lowest CKA-Linear	dinov2_small.block1	0.0983
Lowest CKA-Linear	dinov3_vit7b16_pretrain.block1	0.1757
Lowest CKA-Linear	dinov2_base.block1	0.2032
Lowest RSA Spearman	dinov2_small.block1	0.1498
Lowest RSA Spearman	dinov3_vitb16_pretrain.block1	0.2293
Lowest RSA Spearman	dinov2_base.block1	0.3557
Largest W1	dinov3_vit7b16_pretrain.block39	704.448
Next largest W1	dinov3_vit7b16_pretrain.block38	578.675
Next largest W1	dinov3_vit7b16_pretrain.block37	572.599

Table 9: Self-supervised transformer layers with lowest alignment and highest shift.

F LLM USAGE POLICY

As responsible authors of this paper, we disclose the usage of LLMs purely as a general-purpose assist tool for this paper. We make use of Grammarly's AI to assist in polishing our content. We use it in the form of LLM-as-a-judge fashion, where we pass certain sections of our content to the tool to have it evaluated for logical errors and grade our work and suggest improvements. We double check the suggested polishes and incorporate only those which deem necessary.