Early Visual Degradation Scaffolds Hierarchical Representations

Jeonghwan Cheon[1] **Marin Vogelsang**[2] **Lukas Vogelsang**[2] **Pawan Sinha**[2]
[1]Independent Researcher, Seoul, Republic of Korea,
[2]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
{cheon, ozaki, lvogelsa, psinha}@mit.edu

Abstract

Curriculum learning typically relies on externally designed sequences of training examples. Learning in natural environments, however, rarely follows such explicit sequencing. Here, we propose that developmental visual constraints, such as the initially reduced acuity and color sensitivity characteristic of infant vision, can induce an implicit curriculum without externally prescribing the order in which categories are learned. First, using a multi-level semantic taxonomy, we show that simulating these constraints preferentially impairs discrimination of finer categories while largely preserving coarse ones. We then demonstrate that a bio-inspired curriculum transitioning from blurred, grayscale to clear, full-color images results in coarse-to-fine learning, with coarse distinctions developing more rapidly during early degraded training and finer distinctions improving after the transition to clear input. The resulting visual representations also exhibit stronger alignment with semantic taxonomic structure and human similarity judgments. This alignment benefit generalizes across multiple architectures, datasets, and even a multimodal learning setting. Ablation experiments further reveal that persistent blur reduces hierarchical alignment, whereas blur-to-clear transitions improve it when early input is grayscale, highlighting the importance of the temporal progression of visual input beyond degradation alone. Together, these findings suggest that developmental visual constraints can yield a coarse-to-fine curriculum, shaping both the emergence and hierarchical organization of visual category representations.

1 Introduction

Curriculum learning, in which training examples are presented in a structured order, typically from easy to difficult, has become a well-established strategy for training machine learning models [1–3]. For hierarchically organized tasks, this strategy can take the form of a coarse-to-fine curriculum, in which broad category distinctions are learned before progressively finer-grained ones [4, 5]. In humans, perceptual and conceptual development have likewise been described as the progressive differentiation and organization of representations through experience [6, 7]. Learning in natural environments, however, rarely follows an externally designed curriculum. Instead, learners acquire knowledge through continuous sensory experience without explicit learning stages [8, 9].

Previous work suggests that curricula may emerge naturally from learners' active interactions with their environments, as the content and structure of their sensory input change throughout development [10, 11]. For example, infants' developing sensorimotor abilities shape visual inputs [10], and active object manipulation creates moments of visual selectivity that facilitate word learning [11]. In these cases, a curriculum arises from learners' own actions. Here, we ask whether it can also arise directly from their senses. Specifically, can developmental changes in the quality of sensory input alone yield a coarse-to-fine learning trajectory, without prescribing the order in which category distinctions are learned?

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We hypothesize that the developmental trajectory of vision can provide such a structure [12, 13]. Infants begin visual experience with markedly reduced acuity and chromatic sensitivity, which mature gradually over the first months and years of life [12, 14, 15]. These early limitations restrict access to fine spatial detail and chromatic cues while plausibly preserving global structure sufficient for relatively coarse-grained visual discrimination.

Prior empirical and computational work has linked such low-spatial-frequency input to potential benefits in early category formation. The basic category level is the most inclusive level at which members share clusters of correlated attributes, including similar overall shape [16]. Silhouette similarity can computationally cluster objects into basic-level categories, whereas finer detail may be required at the subordinate level [17]. Consistent with this account, three- to four-month-old infants can categorically distinguish cats from dogs using silhouette information alone [18, 19]. As an early computational proof of principle, low-pass filtering that emphasizes coarse outline shape has been shown to improve basic-level discrimination between images of cats and dogs by reducing within-category variance relative to between-category variance [20].

Extending this analysis to large-scale object recognition, previous work compellingly demonstrated that initially training convolutional neural networks with blurred grayscale input benefits basic-level category learning and transfers to subordinate-level categorization [21]. More broadly, computational studies suggest that early visual degradation may benefit holistic face processing [22, 23], global shape biases [24, 25], abstraction [26], and generalization [27]. Low-pass filtering of visual input has also been shown to improve behavioral alignment [28]. Complementing these computational studies, individuals who were treated for congenital visual deprivation later in development, and who thus skipped the initial period of severely degraded vision, have been shown to exhibit lasting impairments in tasks such as configural face processing and recognition of grayscale imagery [22, 29, 30].

Despite this converging evidence on the functional significance of early visual degradation, it remains unclear how such degradation affects category learning across multiple levels of abstraction. The studies of category formation cited above have mainly examined basic-level and subordinate categories [17, 20, 21]. We here extend this line of work to distinctions above the basic level, such as humans vs. animals or animate vs. inanimate objects. Specifically, we test whether early visual degradation can induce a systematic coarse-to-fine learning trajectory across multiple levels of semantic abstraction, and whether it shapes the hierarchical organization of learned visual representations. We address these questions using bio-inspired visual curricula that introduce fine spatial detail and color information after an initial phase of blurred grayscale input.

Our results reveal that visual degradation differentially affects category discrimination across the semantic hierarchy, and that early exposure to degraded visual input induces a coarse-to-fine learning trajectory. Moreover, bio-inspired training enhances the alignment of learned visual representations with semantic taxonomic structure and human similarity judgments, across architectures, datasets, and a multimodal model.

2 Experimental Setup

2.1 Semantic hierarchy and datasets

We evaluate visual representations using a hand-crafted, four-level semantic taxonomy ranging from broad semantic domains to increasingly specific object categories [31]. The hierarchy is motivated by evidence that object representations are organized both by broad domain distinctions and by nested category relationships [16, 32].

At Depth 1, the taxonomy distinguishes *animate* from *inanimate* entities. Animacy provides an important axis for evaluating broad semantic organization, with response patterns in human and monkey inferior temporal cortex exhibiting distinct clusters for animate and inanimate objects [32]. At Depth 2, the animate branch separates *human* from *animal* categories, while the inanimate branch separates *vehicle* from *tool* categories. These domains are motivated by behavioral evidence for animal–vehicle categorization in infancy [33] and neural evidence for category-related representations of animals and tools [34]. Depth 3 introduces more specific groupings within each domain, such as *adult* and *child* within humans, *pet*, *equid*, *predator*, and *bird* within animals, and *air*, *rail*, *road*, and *water* within vehicles. Depth 4 distinguishes individual object categories within these groups. For example, the path from *animate* to *animal*, *pet*, and, finally, *dog* assigns the same image increasingly specific category labels.

Because the intermediate groupings combine biological, functional, and contextual criteria, the hierarchy serves as a framework for evaluating semantic organization rather than as a strictly biological classification. Its depths indicate relative specificity within the selected category set. Although they do not map one-to-one onto the classical superordinate, basic, and subordinate levels [16], Depth-4 categories mostly correspond to basic-level categories (e.g., dog, car, hammer), with Depths 1–3 capturing distinctions above the basic level. This structure allows us to examine both when category distinctions at different depths emerge during learning and how their relationships are reflected in representational geometry. The taxonomy contains 2, 4, 14, and 32 classes at Depths 1–4, respectively, with 16 animate and 16 inanimate leaf concepts; we selected 12 images per concept from THINGS [35], yielding 384 evaluation images (see Supplementary Methods C.1).

2.2 Developmentally inspired visual curriculum

Our visual curriculum is motivated by the initially limited visual acuity and chromatic sensitivity characteristic of early development [12, 14, 15]. Following previous computational work on the potential benefits of early visual degradation [21, 22, 30], we manipulate the availability of spatial detail and chromatic information through Gaussian blur and grayscale conversion. These manipulations capture simplified aspects of visual development rather than reconstructing its precise biological trajectory.

Standard training uses clear color images throughout. Our primary bio-inspired condition uses blurred grayscale images for the first half of training, followed by clear color images for the second half. The learning-trajectory and main schedule-ablation experiments used 200 epochs, whereas the architecture and dataset comparisons used 90 epochs (see Supplementary Methods C.2). The curriculum therefore changes the visual information available during training without explicitly prescribing the order in which category distinctions are learned. Additional conditions isolate changes in spatial detail and color information, maintain degraded inputs throughout training, or vary the initial blur intensity and transition timing. We denote color and grayscale conditions by C and G, respectively, with a numerical suffix indicating the Gaussian blur standard deviation in pixels. For example, G4→C0 denotes a transition from blurred grayscale images with $\sigma = 4$ to clear color images.

2.3 Evaluation of hierarchical representations

Category information and learning trajectories. We distinguish between category information accessible under degraded inputs and category information acquired through developmental training. For the former, we use a pretrained CLIP visual encoder [36], freeze its parameters, and independently train linear probes at each hierarchical depth. Each probe predicts the category labels defined at its corresponding depth, allowing us to assess how visual degradation affects broad and finer-grained category information.

For the developmental training experiments, we apply the same depth-specific probing approach to representations extracted at successive training checkpoints. Let $A_{d,s}(t)$ denote probing accuracy at hierarchical depth d, training condition s, and epoch t. To compare the temporal progression of learning across depths, we normalize accuracy as

$$\widetilde{A}_{d,s}(t) = \frac{A_{d,s}(t) - A_{d,s}(0)}{A_{d,s}(T) - A_{d,s}(0)}, \tag{1}$$

where T denotes the final training epoch. This normalization maps initial and final accuracy to zero and one, respectively, and expresses intermediate performance relative to the total accuracy gain within each depth and training condition.

Hierarchical alignment. We quantify hierarchical organization by comparing the geometry of learned visual representations with the semantic taxonomy. Let $\mathbf{z}_i$ denote the representation of category i. We construct a representational dissimilarity matrix (RDM) using pairwise cosine distances:

$$D^{\mathrm{rep}}_{ij} = 1 - \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}. \tag{2}$$

We also construct a taxonomy distance matrix D^{tax}, with pairwise distances determined by the deepest hierarchical level shared by each pair of categories. Categories sharing a deeper level receive smaller taxonomic distances.

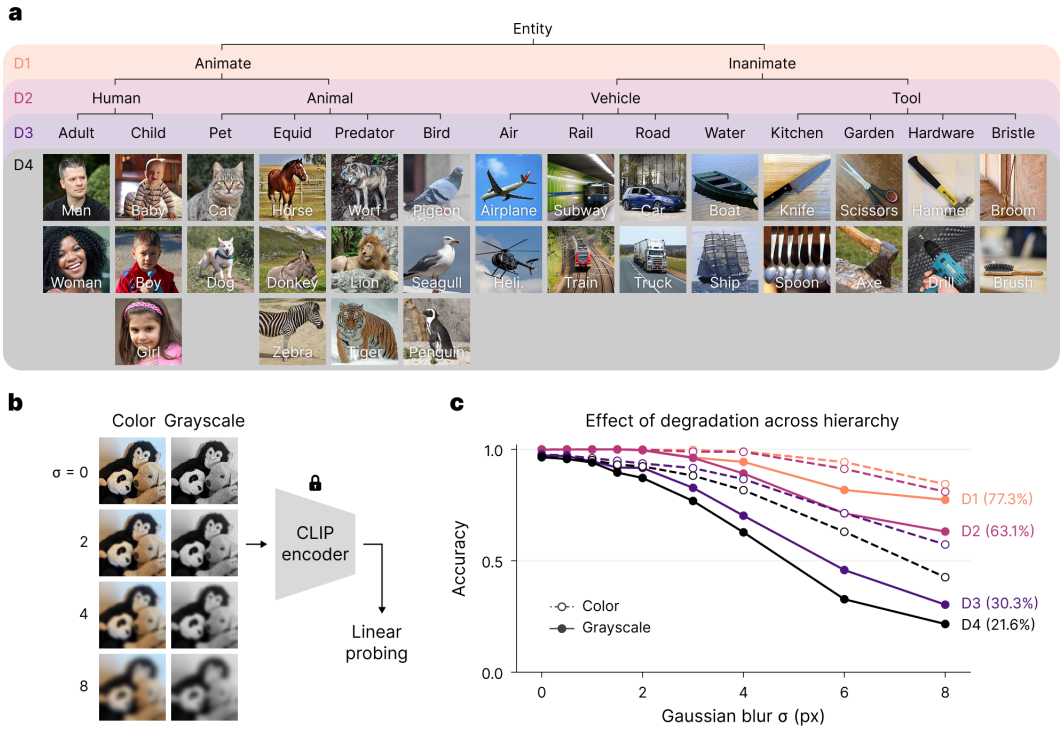


Figure 1: **Hierarchy-dependent information loss under visual degradation.** **a,** Four-level semantic hierarchy (Depths 1–4), ranging from broad entity distinctions to finer object categories. **b,** Experimental setup for classification under Gaussian blur and grayscale conversion using a frozen CLIP encoder and linear probing. Separate linear probes are trained at each hierarchical level to distinguish among categories within that level. **c,** Linear probing accuracy at each hierarchical level as a function of Gaussian blur intensity, evaluated separately for color and grayscale images. Solid and dashed lines indicate grayscale and color conditions, respectively.

Hierarchical alignment is measured as the Spearman correlation between corresponding off-diagonal entries of the two matrices:

$$\rho_{\text{hier}} = \text{Corr}_{\text{Spearman}}\left(\{D_{ij}^{\text{rep}}\}_{i<j}, \{D_{ij}^{\text{tax}}\}_{i<j}\right). \tag{3}$$

Using the upper-triangular entries excludes self-distances and avoids counting symmetric category pairs twice. Higher correlation indicates a stronger correspondence between representational geometry and semantic taxonomic structure.

3 Results

3.1 Visual degradation differentially affects representations across the semantic hierarchy

To investigate how visual degradation affects category discrimination at different levels of semantic abstraction, we examined hierarchical classification under degraded visual inputs, using our hand-crafted, four-level semantic taxonomy described in Section 2 (Figure 1a) [31]. The taxonomy ranges from animate–inanimate distinctions (Depth 1), through broader object families (Depths 2–3), to largely basic-level object categories (Depth 4).

We evaluated hierarchical classification using a pretrained CLIP visual encoder [36] whose parameters were frozen throughout the analysis (Figure 1b). To simulate degraded visual inputs, images were subjected to Gaussian blur of varying intensity and converted to grayscale. We independently trained linear probes on the resulting visual representations at each hierarchical depth, with each probe distinguishing categories defined at that level. This allowed us to test category discrimination at each depth without modifying the underlying visual encoder.

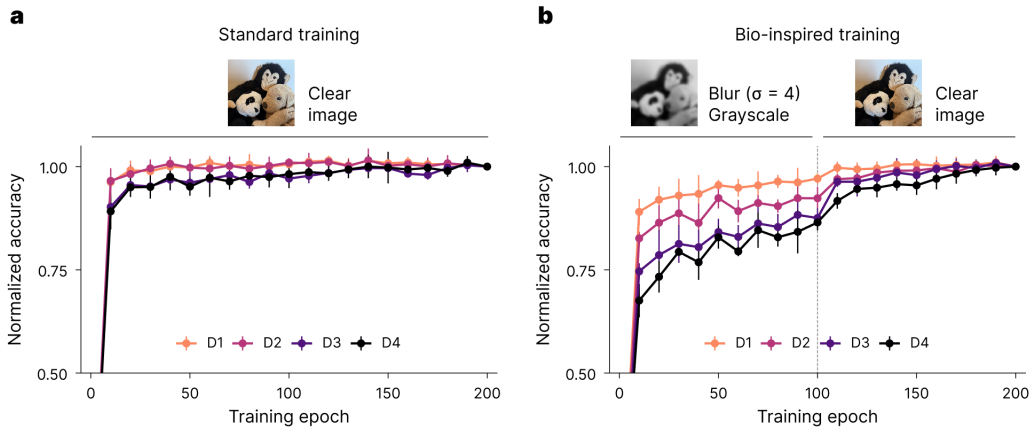


Figure 2: **Developmental trajectories of hierarchical classification under standard and bio-inspired training. a,** Linear probing accuracy at each hierarchical depth (Depths 1–4) as a function of training epochs, for the standard model trained on clear color images. **b,** Linear probing accuracy under bio-inspired training, where blurred grayscale images during the first 100 epochs are followed by clear color images for the remaining 100 epochs. The vertical dashed line indicates the transition between training phases. For both panels, accuracy is normalized at each hierarchical depth, such that initial and final performance correspond to 0 and 1, respectively.

We found that visual degradation differentially affected classification accuracy across the semantic hierarchy (Figure 1c). As Gaussian blur increased, accuracy declined more rapidly for finer category distinctions, whereas coarse-grained distinctions remained relatively robust. In particular, broad distinctions between animate and inanimate entities (Depth 1) were largely preserved under substantial blur, while discrimination among more closely related object categories (Depth 4) deteriorated markedly. Similar hierarchical differences were observed for both color and grayscale images, although removing color information further accentuated these differences. This hierarchy-dependent decline persisted after correcting accuracy for the different chance levels at each depth (Supplementary results A.1). In contrast, a control with shuffled taxonomic groupings showed much less separation among depths under strong blur, indicating that the preserved coarse distinctions depend on the semantic organization of the taxonomy rather than on class counts alone (Supplementary results A.1). Together, these results suggest that degraded visual inputs preferentially impair finer category discrimination while leaving coarse-grained discrimination largely intact.

3.2 Early visual degradation induces a coarse-to-fine developmental trajectory

Having established that visual degradation preferentially impairs finer-grained category discrimination, we next investigated whether such constraints during early training could induce a coarse-to-fine trajectory in category learning. We compared standard training on clear color images with our bio-inspired regimen, in which blurred grayscale images during the first half of training are followed by clear color images during the second half (Figure 2a,b). This two-stage schedule serves as a rough proxy for developmental visual maturation.

To track the emergence of category discrimination throughout training, we evaluated visual representations at successive training epochs using independent linear probes at each hierarchical depth. Classification accuracy was normalized separately for each depth, such that performance at the beginning and end of training corresponded to 0 and 1, respectively. Under standard training, classification performance increased rapidly across all hierarchical depths, with relatively similar learning trajectories for broad and finer category distinctions (Figure 2a).

In contrast, bio-inspired training produced a more distinct coarse-to-fine developmental trajectory (Figure 2b). During the initial phase of blurred grayscale exposure, classification performance increased more rapidly for broad distinctions (Depths 1–2), whereas finer-grained distinctions (Depths 3–4) developed more slowly. Following the transition to clear color images, classification performance at finer hierarchical depths increased, progressively approaching its final level. These findings suggest that early visual degradation can temporally separate the emergence of category distinctions across hierarchical levels, inducing an implicit coarse-to-fine learning trajectory.

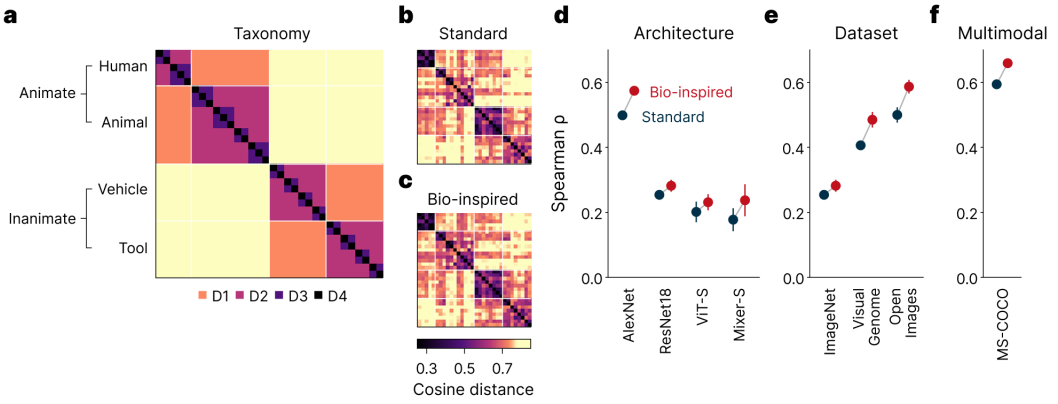


Figure 3: **Hierarchical organization of visual representations under standard and bio-inspired training. a,** Taxonomy distance matrix constructed from the four-level semantic hierarchy (Depths 1–4), with pairwise distances determined by the deepest shared hierarchical level. **b,c,** Representational dissimilarity matrices of models trained under standard (**b**) and bio-inspired (**c**) conditions. Pairwise cosine distances between category representations are computed and arranged according to the semantic hierarchy. **d-f,** Hierarchical alignment across model architectures (**d**), datasets (**e**), and a multimodal model (**f**). Alignment is quantified by Spearman correlation between the taxonomy distance matrix and model-derived representational dissimilarity matrices. Blue and red markers indicate standard and bio-inspired training, respectively.

To test whether this coarse-to-fine trajectory depended on the semantic grouping of concepts, we shuffled the assignment of concepts to taxonomy leaves while preserving the number of classes and branch sizes at each depth (Supplementary results A.2). Under the shuffled taxonomy, the depth-wise trajectories became more similar, and the early advantage of coarse distinctions seen with the original taxonomy was no longer evident. This control suggests that the original trajectory reflects the semantic grouping of the taxonomy rather than differences in class counts alone.

3.3 Bio-inspired training enhances the hierarchical organization of visual representations

We next investigated whether developmental visual training shapes the hierarchical organization of learned representations. To quantify hierarchical organization, we constructed a taxonomy distance matrix from our four-level semantic hierarchy, assigning pairwise distances according to the deepest hierarchical level shared by each pair of categories (Figure 3a). We then computed representational dissimilarity matrices (RDMs) from the learned visual representations using pairwise cosine distances between category representations. Hierarchical alignment was quantified by the Spearman correlation between the taxonomy distance matrix and the model-derived RDMs, measuring how closely the geometry of learned representations reflects the semantic taxonomy.

We first compared the representational geometry of models trained under standard and bio-inspired conditions (Figure 3b,c). Both models exhibited a hierarchical structure in their representations, with smaller distances between semantically related categories and larger distances between more distant categories. The bio-inspired AlexNet [37] showed stronger taxonomic alignment than the standard AlexNet (mean Spearman $\rho = 0.574$ versus 0.499 across three training seeds; Figure 3d and Supplementary Table S1). Together with the trajectories in Section 3.2, this suggests that early visual degradation influences both the temporal emergence of category distinctions and their organization in the learned representation. Layer-wise analysis localized the larger mean alignment difference to later representations, particularly FC6 and FC7, while the earlier convolutional layers showed smaller differences (Supplementary results A.3).

We next examined whether this improvement generalizes across different model architectures, datasets, and learning paradigms. We evaluated hierarchical alignment in AlexNet, ResNet-18 [38], ViT-S [39], and Mixer-S [40] architectures, observing higher alignment under bio-inspired training across all models examined (Figure 3d). We further extended the analysis to ImageNet [41], Visual Genome [42], and Open Images [43], where bio-inspired training consistently improved alignment with the semantic taxonomy (Figure 3e). Finally, we examined a multimodal model trained on

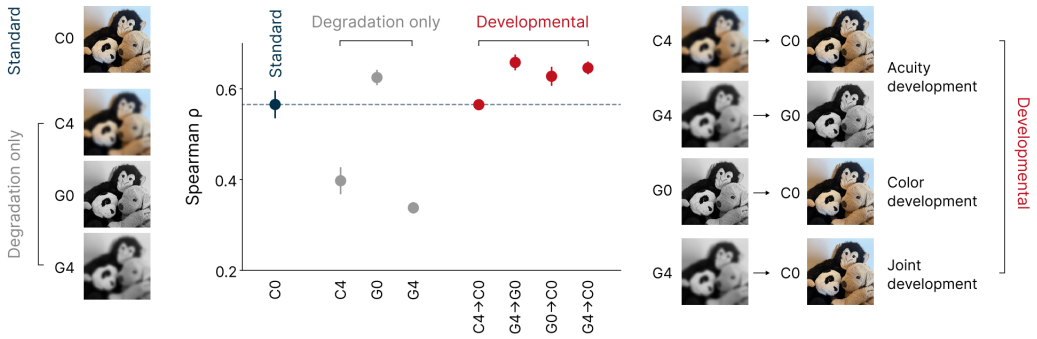


Figure 4: **Effects of visual degradation and developmental training schedules on hierarchical alignment.** Hierarchical alignment across standard training (C0), degradation-only conditions (C4, G0, and G4), and developmental training schedules (C4→C0, G4→G0, G0→C0, and G4→C0). C and G denote color and grayscale images, respectively, with the number indicating Gaussian blur intensity (σ, in pixels). Developmental conditions involve a transition between visual input conditions during training. The horizontal dashed line indicates the standard training baseline (C0).

MS-COCO [44] and observed a similar improvement in hierarchical alignment (Figure 3f). Together, these results indicate that the effect of early visual degradation on hierarchical organization is not restricted to a single architecture, dataset, or learning paradigm.

We also compared model dissimilarities with human behavioral similarity judgments represented by SPoSE embeddings [45]. SPoSE is a 66-dimensional sparse positive similarity embedding fitted to human odd-one-out judgments of THINGS objects [46]; human alignment is the Spearman correlation between model distances and one minus the cosine similarity of SPoSE vectors across 496 concept pairs. Mean alignment with this human reference was higher after bio-inspired training across the examined architectures, datasets, and multimodal setting (Supplementary results A.4). Thus, the representational difference extends beyond correspondence with the semantic taxonomy to the tested measure of human conceptual similarity.

We then asked whether the taxonomic alignment benefit was uniform across concept domains. Bio-inspired training showed a consistent mean advantage within the animate subset across the tested architectures, datasets, and multimodal setting, whereas the inanimate subset showed smaller, absent, or reversed differences (Supplementary results A.5). Animate concepts also had slightly higher within-domain similarity in a frozen CLIP feature space, although raw-pixel similarities showed little domain separation (Supplementary Figure S6). These observations are consistent with a larger benefit for concepts that are more coherent in visual feature space. Because the analysis compares only two domains, it does not establish that animacy itself uniquely accounts for the difference.

3.4 Developmental training conditions shape hierarchical alignment

To investigate which aspects of developmental training contribute to hierarchical alignment, we systematically varied the visual input conditions and their temporal progression during training. We first examined whether visual degradation alone could account for the improvement in hierarchical alignment observed. To this end, we compared standard training with clear color images (C0) against three degradation-only conditions: blurred color images (C4), clear grayscale images (G0), and blurred grayscale images (G4) (Figure 4). Here, C and G denote color and grayscale inputs, respectively, and the numbers indicate Gaussian blur intensity (σ, in pixels). We found that persistent blur substantially reduced hierarchical alignment, with both C4 and G4 exhibiting lower mean correlations with the semantic taxonomy than standard training. In contrast, grayscale training without blur (G0) yielded higher mean alignment than C0. These results show that the effects of sustained visual degradation depend on which visual information is removed.

We next examined the role of the developmental progression of visual input quality. We considered several training schedules that independently manipulate the maturation of visual acuity and color sensitivity (Figure 4). Acuity development was simulated by transitioning from blurred to clear images while maintaining color information (C4→C0) or grayscale input (G4→G0). Color development was

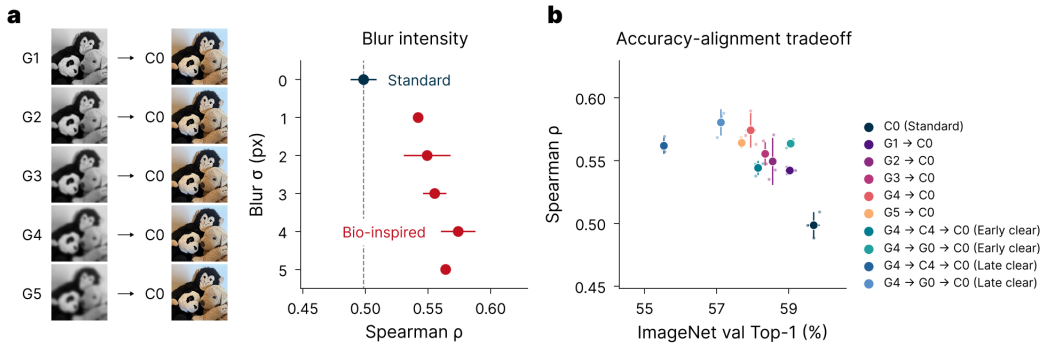


Figure 5: **Effects of developmental conditions on hierarchical alignment and classification accuracy. a,** Hierarchical alignment under bio-inspired training with varying initial Gaussian blur intensities (σ = 1–5 px), compared with standard training (C0). **b,** ImageNet top-1 classification accuracy vs. hierarchical alignment across standard training, varying initial blur intensities, and developmental schedules with different transition timings.

simulated by transitioning from grayscale to color images without reducing spatial detail (G0→C0), while joint development involved transitioning from blurred grayscale to clear color images (G4→C0). The grayscale blur-to-clear schedule G4→G0 yielded higher mean taxonomic alignment than constant G0 (0.658 versus 0.625), whereas the color blur-to-clear schedule C4→C0 matched the C0 baseline (0.565 for both; Supplementary Table S4).

This contrast is consistent with the finding of Jinsi et al. [21] that blur reduction benefited category-learning accuracy with grayscale images but offered little benefit with color images, although our outcome is representational alignment rather than classification accuracy. The G0→C0 and G4→C0 schedules also exceeded the C0 baseline (0.628 and 0.646 versus 0.565), suggesting that later color availability can accompany improved hierarchical alignment. Neither schedule, however, clearly exceeded its grayscale counterpart (G0 and G4→G0, respectively). These comparisons therefore do not establish an additional benefit of introducing color itself.

We next investigated how the severity of initial visual degradation influences hierarchical alignment. To this end, we varied the initial Gaussian blur intensity from σ = 1 to 5 pixels before transitioning to clear color images (Figure 5a). Hierarchical alignment exceeded the standard training baseline across all tested blur intensities, although the magnitude of improvement varied with initial blur strength. In particular, moderate-to-strong initial blur produced higher alignment than weaker blur conditions, suggesting that the degree of early visual degradation modulates the organization of learned representations. We also examined how the order and timing of color and acuity transitions affect taxonomic alignment (Supplementary results A.6). When color and acuity changed in separate phases, the schedules that reached clear color input later had higher mean alignment than their corresponding early-clear schedules. The simultaneous transition and late-clear acuity-first schedule yielded the highest means in this comparison. These results suggest that the timing and sequence of visual changes may modulate the alignment benefit.

Finally, we examined the relationship between hierarchical alignment and image classification performance under different developmental conditions (Figure 5b). Standard training achieved relatively high ImageNet top-1 accuracy but lower hierarchical alignment, whereas several bio-inspired conditions exhibited stronger alignment with modest reductions in classification accuracy. We further compared developmental schedules with earlier and later transitions to clear inputs, observing distinct accuracy–alignment profiles depending on the timing and sequence of visual maturation. Together, these results demonstrate that developmental training conditions, including the type, intensity, and timing of visual degradation, systematically shape hierarchical alignment and its relationship with classification accuracy.

4 Discussion

Our findings suggest that a developmentally inspired visual curriculum can scaffold the emergence of hierarchical perceptual representations. Temporarily restricting access to color and fine spatial detail

during the early stages of training produces a clearer coarse-to-fine learning trajectory, with resulting representations also corresponding more closely to taxonomic structure and human behavioral similarity judgments. The alignment benefits generalize across multiple neural architectures, learning objectives, and training corpora. Together, these results suggest that the temporal staging of visual experience can influence when distinctions at different taxonomic levels emerge and how those distinctions are organized within the learned representational space.

This work builds on rich developmental literature showing that the initially degraded visual input characteristic of a newborn's visual experience [14, 15] can confer important benefits to category learning and perceptual organization. Specifically, coarse shape and silhouette information can support the organization of basic-level categories [16–20]. Moreover, developmentally inspired, low-pass-filtered input has been linked to enhanced holistic processing, global shape bias, abstraction, generalization, and robustness in deep neural networks [22, 24–27, 47, 48], although such benefits do not always exceed those of training on mixed blurred and clear images [49, 50]. Empirically, individuals treated for congenital blindness late in life, who begin their post-operative experience with a markedly more mature visual system, exhibit lasting perceptual deficits [22, 29, 30]. Most directly relevant for the present study, previous work [21] demonstrated that initial training with blurred grayscale images improves basic-level category learning and also facilitates subsequent finer-grained categorization.

Our contribution is therefore not the introduction of developmental visual degradation, but the characterization of how this established training principle shapes learning dynamics and representational geometry across multiple abstraction depths. Whereas previous work [21] focused on the relationship between basic-level and finer-grained categorization, we examine distinctions above the basic level and how visual categories are organized across multiple levels of abstraction. Bio-inspired training alters the temporal emergence of distinctions across these abstraction levels and produces a representational geometry that more closely reflects both formal taxonomic relations and human similarity structure [45]. This distinction is important because machine–human visual alignment can vary substantially across abstraction levels, and these differences may be obscured by measures computed at a single level or aggregated across levels [51]. Our results therefore extend prior work on category-learning performance by showing that temporally structured visual experience may also shape the broader representational hierarchy in which categories at different levels of abstraction are embedded. Because the alignment benefit also emerged in a vision–language model, it will be interesting to test whether linguistic and other multimodal input further enrich this structure. Given that infants begin to map words onto objects while their vision is still maturing [52], such examination could also help test how perceptual and linguistic development interact.

An interpretation of our findings is that early degradation creates a temporary bottleneck that changes which levels of structure are most accessible during learning. This interpretation is consistent with starting-small, coarse-to-fine curriculum, and adaptive-initial-degradation accounts [1, 4, 13, 53]. Theoretical work has further shown that deep linear networks can progressively differentiate hierarchical semantic structure, with broader distinctions emerging before finer ones as a consequence of the statistical structure of the learning environment and the dynamics of learning [54]. By limiting color and fine spatial detail, this bottleneck may increase the relative usefulness of broad shape-based regularities shared across diverse instances. As the bottleneck is relaxed, representations may become increasingly differentiated while retaining aspects of the broader organization established during earlier stages of learning. This account could explain both the observed coarse-to-fine learning trajectories and the stronger correspondence with hierarchical taxonomic structure. Since the manipulation constrains the perceptual information available to the model rather than providing explicit semantic or taxonomic structure, hierarchical organization may emerge indirectly as the visual features available at different stages of learning change over time.

More broadly, our findings connect to the idea of an implicit curriculum, in which structured learning trajectories emerge without explicitly prescribing the order of skill acquisition. Work on language model pretraining has shown that different capabilities emerge in a consistent order across models [55]. Building also on earlier proposals that developmental limitations can facilitate learning [53, 56, 57], our results suggest that an implicit curriculum can emerge from developmentally inspired constraints on perceptual access, without explicitly sequencing training examples by semantic complexity. Overall, our findings indicate that developmentally inspired visual constraints offer a potential design principle for inducing implicit learning curricula in artificial neural networks, in line with broader proposals that principles of human development can enrich machine learning [58].

5 Broader impacts

This work suggests that the temporal organization of sensory input can help shape hierarchical visual representations, offering a potential design principle for developmentally inspired machine learning. Such approaches may support the development of models whose representational organization better reflects selected semantic relationships. However, stronger correspondence with a semantic taxonomy or human similarity judgments does not by itself establish broader understanding, fairness, or safety. These measures reflect the assumptions and coverage of the taxonomies, datasets, and behavioral judgments used for evaluation, and optimizing them may reinforce those assumptions. Downstream applications should therefore assess performance across relevant categories and contexts, particularly given the observed trade-offs between hierarchical alignment and classification accuracy.

Our findings also provide a computational perspective on how early perceptual constraints may influence learning. The training manipulations are simplified analogues of visual development, however, and do not establish causal mechanisms in infants or justify recommendations to alter children's visual experience or clinical care. Our work's primary contribution is to generate testable hypotheses for future computational and developmental research.

6 Limitations

We would like to note several important limitations. First, the coarse-to-fine trajectory observed in our simulations should be interpreted relative to the formal hierarchy used in our probing analysis, rather than as a reconstruction of the order in which children acquire particular categories. Human category development does not consist of a simple progression from superordinate to basic and then subordinate categories. Instead, perceptually grounded basic-level distinctions such as cat vs. dog are available early in infancy [18, 19], whereas some broader, definitionally organized categories may emerge later [59]. Our claim is that the curriculum temporally separates distinctions defined at different depths of this probing hierarchy, not that the model recapitulates exact human category-acquisition order.

The developmental analogy should also be interpreted cautiously. Different aspects of visual function, such as acuity, contrast sensitivity, and chromatic sensitivity, do not mature according to a single shared timeline. For instance, acuity does not reach adult levels until well after age 3, and contrast sensitivity matures over an even longer period [60]. Our training schedules simulate reduced acuity and chromatic sensitivity but do not explicitly model reduced contrast sensitivity, as done in other work [25], and they do not reconstruct the precise biological progression of these functions [12, 14, 15, 60, 61]. Moreover, Gaussian blur and color removal capture only simplified aspects of early visual limitations, and static training images and epoch-based schedules do not correspond directly to infants' continuous, actively sampled visual experience. More naturalistic tests would require longitudinal egocentric visual and auditory streams, active visual sampling, and perceptual schedules calibrated to empirical developmental trajectories [9, 62–65]. We therefore describe our method as a bio-inspired visual curriculum rather than as a fully faithful model of human visual development. Notwithstanding these caveats, our results suggest that initial perceptual degradations can influence not only which visual distinctions are learned, but also when they emerge and how they are organized across abstraction levels.

Finally, our analyses rely on a hand-crafted semantic taxonomy designed to examine learning curricula and representational structure along controlled, interpretable semantic axes. Although this framework enables systematic comparisons across abstraction depths, its category selection and grouping reflect specific design choices. Generalization to substantially larger and more diverse sets of object categories remains untested. We also did not directly compare the learned representations with neural representations in infants or adults. Future work could extend the evaluation to broader category spaces and alternative semantic hierarchies, and compare model learning trajectories and representational geometry with developmental and adult neural data.

7 Code availability

Python 3.12 (Python software foundation) with PyTorch 2.12 is used to perform the simulation. The code used in this work is available at `https://github.com/jeonghwan-cheon/dev-hierarchy`.

Acknowledgments and Disclosure of Funding

The authors acknowledge the MIT Office of Research Computing and Data for providing high performance computing resources that have contributed to the research results reported within this paper. Marin Vogelsang was supported by the Japan Society for the Promotion of Science (JSPS) Overseas Research Fellowship and the Yamada Science Foundation, and Lukas Vogelsang was supported by a grant from the Simons Foundation International to the Simons Center for the Social Brain at MIT.

References

[1] Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 41–48 (2009).

[2] Kumar, M., Packer, B. & Koller, D. Self-paced learning for latent variable models. *Advances in Neural Information Processing Systems* **23** (2010).

[3] Platanios, E. A., Stretcu, O., Neubig, G., Poczos, B. & Mitchell, T. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 1162–1172 (2019).

[4] Stretcu, O., Platanios, E. A., Mitchell, T. & Póczos, B. Coarse-to-fine curriculum learning for classification. In *International Conference on Learning Representations (ICLR) Workshop on Bridging AI and Cognitive Science (BAICS)* (2020).

[5] Ren, X. *et al.* Generalized coarse-to-fine visual recognition with progressive training. *arXiv preprint arXiv:1811.12047* (2018).

[6] Gibson, E. J. Principles of perceptual learning and development. (1969).

[7] Mandler, J. M. Perceptual and conceptual processes in infancy. *Journal of Cognition and Development* **1**, 3–36 (2000).

[8] Gibson, E. J. & Pick, A. D. *An ecological approach to perceptual learning and development* (Oxford University Press, 2000).

[9] Sullivan, J., Mei, M., Perfors, A., Wojcik, E. & Frank, M. C. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind* **5**, 20–29 (2021).

[10] Smith, L. B., Jayaraman, S., Clerkin, E. & Yu, C. The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences* **22**, 325–336 (2018).

[11] Yu, C. & Smith, L. B. Embodied attention and word learning by toddlers. *Cognition* **125**, 244–262 (2012).

[12] Braddick, O. & Atkinson, J. Development of human visual function. *Vision Research* **51**, 1588–1609 (2011).

[13] Vogelsang, L., Vogelsang, M., Pipa, G., Diamond, S. & Sinha, P. Butterfly effects in perceptual development: A review of the 'adaptive initial degradation' hypothesis. *Developmental Review* **71**, 101117 (2024).

[14] Atkinson, J., Braddick, O. & Moar, K. Development of contrast sensitivity over the first 3 months of life in the human infant. *Vision Research* **17**, 1037–1044 (1977).

[15] Morrone, M. C., Burr, C. D. & Fiorentini, A. Development of contrast sensitivity and acuity of the infant colour system. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **242**, 134–139 (1990).

[16] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. Basic objects in natural categories. *Cognitive Psychology* **8**, 382–439 (1976).

[17] Cutzu, F. & Tarr, M. J. Representation of three-dimensional object similarity in human vision. In *Human Vision and Electronic Imaging II*, vol. 3016, 460–471 (SPIE, 1997).

[18] Eimas, P. D. & Quinn, P. C. Studies on the formation of perceptually based basic-level categories in young infants. *Child Development* **65**, 903–917 (1994).

[19] Quinn, P. C., Eimas, P. D. & Tarr, M. J. Perceptual categorization of cat and dog silhouettes by 3-to 4-month-old infants. *Journal of Experimental Child Psychology* **79**, 78–94 (2001).

[20] French, R. M., Mermillod, M., Chauvin, A., Quinn, P. C. & Mareschal, D. The importance of starting blurry: Simulating improved basic-level category learning in infants due to weak visual acuity. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (Routledge, 2002).

[21] Jinsi, O., Henderson, M. M. & Tarr, M. J. Early experience with low-pass filtered images facilitates visual category learning in a neural network model. *PLOS One* **18**, e0280145 (2023).

[22] Vogelsang, L. *et al.* Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences* **115**, 11333–11338 (2018).

[23] Jang, H. & Tong, F. Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *Journal of Vision* **21**, 6–6 (2021).

[24] Vogelsang, M., Vogelsang, L., Pipa, G., Diamond, S. & Sinha, P. Potential role of developmental experience in the emergence of the parvo-magno distinction. *Communications Biology* **8**, 987 (2025).

[25] Lu, Z., Thorat, S., Cichy, R. M. & Kietzmann, T. C. Adopting a human developmental visual diet yields robust and shape-based ai vision. *Nature Machine Intelligence* 1–14 (2026).

[26] Cheon, J. & Paik, S.-B. Gradual sensory maturation promotes abstract representation learning. *bioRxiv* 2025–06 (2025).

[27] Cai, Y., Lin, Q., Nunna, B. S. & Zhang, M. Learning to see through a baby's eyes: Early visual diets enable robust visual intelligence in humans and machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13766–13780 (2026).

[28] Wolff, M., Klein, T., Rusak, E., Wichmann, F. & Brendel, W. Low-pass filtering improves behavioral alignment of vision models. In *International Conference on Learning Representations*, vol. 2026 (2026).

[29] Le Grand, R., Mondloch, C. J., Maurer, D. & Brent, H. P. Early visual experience and face processing. *Nature* **410**, 890–890 (2001).

[30] Vogelsang, M. *et al.* Impact of early visual experience on later usage of color cues. *Science* **384**, 907–912 (2024).

[31] Cheon, J., Vogelsang, M., Vogelsang, L. & Sinha, P. Spectral stratification of semantic abstraction in vision-language models. In *Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning* (2026).

[32] Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).

[33] Mandler, J. M. & McDonough, L. Concept formation in infancy. *Cognitive Development* **8**, 291–318 (1993).

[34] Chao, L. L., Haxby, J. V. & Martin, A. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience* **2**, 913–919 (1999).

[35] Hebart, M. N. *et al.* THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS One* **14**, e0223792 (2019).

[36] Radford, A. *et al.* Learning transferable visual models from natural language supervision. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 8748–8763 (PMLR, 2021).

[37] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, vol. 25 (2012).

[38] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

[39] Dosovitskiy, A., Beyer, L., Kolesnikov, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).

[40] Tolstikhin, I. O., Houlsby, N., Kolesnikov, A. *et al.* MLP-Mixer: An all-MLP architecture for vision. In *Advances in Neural Information Processing Systems*, vol. 34, 24261–24272 (2021).

[41] Russakovsky, O., Deng, J., Su, H. *et al.* ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* **115**, 211–252 (2015).

[42] Krishna, R., Zhu, Y., Groth, O. *et al.* Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* **123**, 32–73 (2017).

[43] Kuznetsova, A., Rom, H., Alldrin, N. *et al.* The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision* **128**, 1956–1981 (2020).

[44] Lin, T.-Y., Maire, M., Belongie, S. *et al.* Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, 740–755 (2014).

[45] Hebart, M. N., Zheng, C. Y., Pereira, F. & Baker, C. I. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour* **4**, 1173–1185 (2020).

[46] Hebart, M. N. *et al.* THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife* **12**, e82580 (2023).

[47] Jang, H. & Tong, F. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nature Communications* **15**, 1989 (2024).

[48] Vogelsang, M., Vogelsang, L. & Sinha, P. Computationally probing the role of time-limited neuronal plasticity in early visual development. *Developmental Science* **29**, e70099 (2026).

[49] Avberšek, L. K., Zeman, A. & de Beeck, H. O. Training for object recognition with increasing spatial frequency: A comparison of deep learning with human vision. *Journal of Vision* **21**, 14–14 (2021).

[50] Yoshihara, S., Fukiage, T. & Nishida, S. Does training with blurred images bring convolutional neural networks closer to humans with respect to robust object recognition and internal representations? *Frontiers in Psychology* **14**, 1047694 (2023).

[51] Muttenthaler, L. *et al.* Aligning machine and human visual representations across abstraction levels. *Nature* **647**, 349–355 (2025).

[52] Bergelson, E. & Swingley, D. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences* **109**, 3253–3258 (2012).

[53] Elman, J. L. Learning and development in neural networks: The importance of starting small. *Cognition* **48**, 71–99 (1993).

[54] Saxe, A. M., McClelland, J. L. & Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences* **116**, 11537–11546 (2019).

[55] Liu, E. *et al.* What do language models learn and when? the implicit curriculum hypothesis. In *Conference On Language Models*, vol. 3 (2026).

[56] Newport, E. L. Maturational constraints on language learning. *Cognitive Science* **14**, 11–28 (1990).

[57] Turkewitz, G. & Kenny, P. A. Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology* **15**, 357–368 (1982).

[58] Zaadnoordijk, L., Besold, T. R. & Cusack, R. Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence* **4**, 510–520 (2022).

[59] Keil, F. C. & Batterman, N. A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior* **23**, 221–236 (1984).

[60] Leat, S. J., Yadav, N. K. & Irving, E. L. Development of visual acuity and contrast sensitivity in children. *Journal of Optometry* **2**, 19–26 (2009).

[61] Priamikov, A., Narayan, V., Shi, B. E. & Triesch, J. The role of contrast sensitivity in the development of binocular vision: A computational study. In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 33–38 (IEEE, 2015).

[62] Schaumlöffel, T., Aubret, A., Roig, G. & Triesch, J. Caregiver talk shapes toddler vision: A computational study of dyadic play. In *2023 IEEE International Conference on Development and Learning (ICDL)*, 67–72 (IEEE, 2023).

[63] Long, B. L. *et al.* The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. In *8th Annual Conference on Cognitive Computational Neuroscience* (2025).

[64] Vong, W. K., Wang, W., Orhan, A. E. & Lake, B. M. Grounded language acquisition through the eyes and ears of a single child. *Science* **383**, 504–511 (2024).

[65] Lin, D. *et al.* EgoBabyVLM: Benchmarking cross-modal learning from naturalistic egocentric video data. *arXiv preprint arXiv:2605.19130* (2026).

A Supplementary results

A.1 Effect of taxonomic grouping on classification under visual degradation

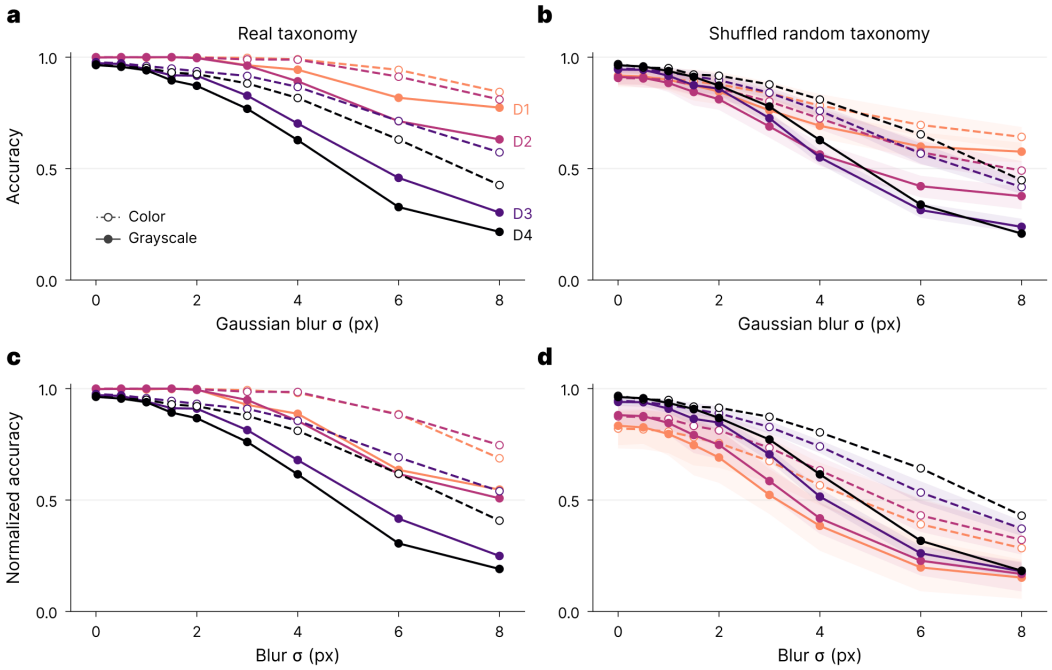


Figure S1: **Hierarchical classification under visual degradation with real and shuffled taxonomies. a,b,** Linear-probe accuracy across Gaussian blur intensities at taxonomy depths D1–D4 using the original taxonomy (**a**) or 50 random assignments of whole concepts to taxonomy leaves (**b**). **c,d,** The corresponding accuracies after correcting for the number of classes at each depth using $(a - 1/K)/(1 - 1/K)$, where a is accuracy and K is the number of classes. Chance accuracy therefore corresponds to zero, and perfect accuracy corresponds to one. Solid lines and filled markers indicate grayscale images; dashed lines and open markers indicate color images. Original-taxonomy curves are averaged over three cross-validation split seeds; shuffled-taxonomy curves show the mean and shaded 5th–95th percentile range across 50 permutations using one split seed. Permuting leaf assignments preserves the D4 concept labels.

To assess whether the hierarchy-dependent effect of visual degradation depended on the semantic grouping of concepts, we compared the original taxonomy with random assignments of the same concepts to its leaf positions (Supplementary Figure S1a,b). Under the original taxonomy, broad distinctions remained relatively robust as blur increased, while accuracy at finer depths fell more sharply (Supplementary Figure S1a). With shuffled taxonomic groupings, accuracy also declined under blur, but the robustness of the broad levels was reduced (Supplementary Figure S1b). D4 labels were unaffected by the reassignment of concepts to leaves.

Because the depths contain different numbers of classes, we also expressed accuracy relative to each depth's chance level (Supplementary Figure S1c,d). The original taxonomy retained a marked coarse-to-fine separation at strong blur after this correction, whereas the shuffled taxonomy showed much less separation among depths, particularly for grayscale images. These controls support the interpretation that the preserved broad distinctions under degraded input depend in part on the semantic organization of the taxonomy, rather than only on differences in class counts.

A.2 Hierarchical classification trajectories under permuted taxonomies

To examine whether the developmental trajectories in Figure 2 depend on the semantic organization of the taxonomy, we repeated the depth-wise probe analysis after permuting the assignment of concepts to taxonomic groups while preserving the number of classes at each depth. Under standard training, normalized accuracy rose rapidly and followed nearly overlapping trajectories across depths

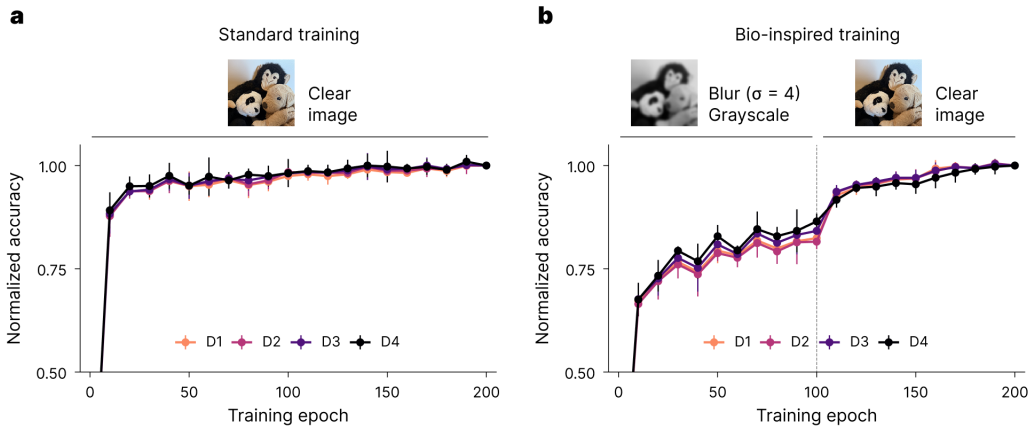


Figure S2: **Classification trajectories under permuted taxonomic groupings. a,** Normalized linear-probe accuracy at each taxonomy depth (D1–D4) during standard training with clear color images for 200 epochs. **b,** The same analysis for bio-inspired training with blurred grayscale images ($\sigma = 4$ px) for the first 100 epochs and clear color images for the remaining 100 epochs. The vertical dashed line marks the input transition. At D1–D3, accuracy is averaged over 50 random assignments of concepts to taxonomy leaves; D4 retains the original concept labels. Accuracy is normalized separately for each depth and training seed to its initial and final values. Points and error bars show the mean and standard deviation across five training seeds.

(Supplementary Figure S2a). Under bio-inspired training, progress was slower during blurred grayscale exposure, but the permuted depths likewise showed similar trajectories rather than a consistent early advantage for coarse levels; all depths increased after the transition to clear color images (Supplementary Figure S2b). This suggests that the coarse-to-fine separation in the original taxonomy reflects its semantic grouping rather than class counts alone.

A.3 Layer-wise taxonomic alignment of visual representations

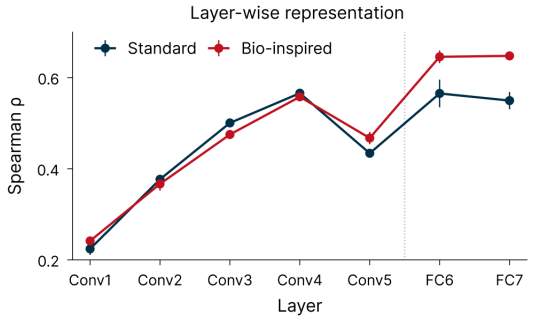


Figure S3: **Layer-wise taxonomic alignment under standard and bio-inspired training.** Spearman correlation between the taxonomy distance matrix and model representational dissimilarities from convolutional layers Conv1–Conv5 and fully connected layers FC6–FC7. Blue and red curves correspond to standard training with clear color images and bio-inspired training with blurred grayscale images followed by clear color images, respectively. Points and error bars show the mean and standard deviation across five training seeds per condition.

To determine where the taxonomic alignment difference emerges within the visual network, we measured Spearman alignment at successive layers for both standard and bio-inspired training (Supplementary Figure S3). Alignment increased from the early convolutional layers toward later representations under both training conditions. The two conditions were similar through Conv4, with small differences in either direction, whereas bio-inspired training showed higher mean alignment at Conv5 and a larger advantage at FC6 and FC7. Thus, the final-layer result in Figure 3 is accompanied by differences in later representations, rather than constituting a uniform advantage at every layer.

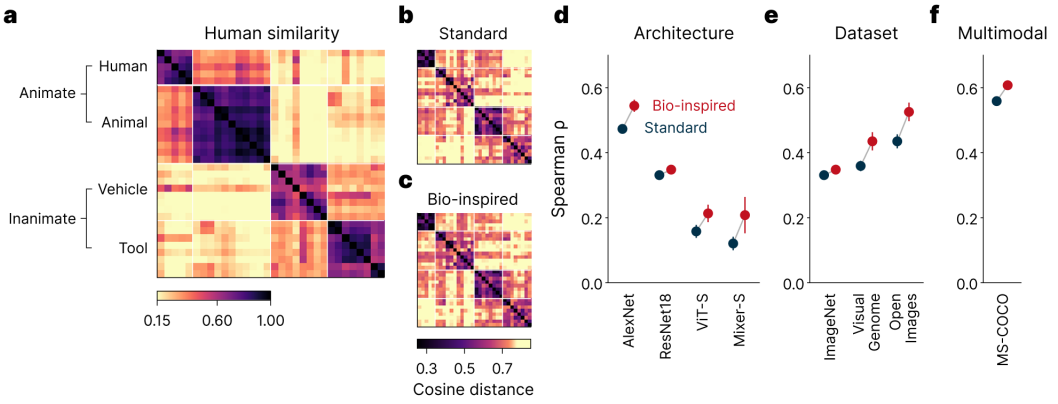


Figure S4: **Human similarity alignment under standard and bio-inspired training. a,** Human behavioral cosine-similarity matrix for the evaluated concepts, derived from SPoSE embeddings [45] and ordered by animate and inanimate groups. **b,c,** Model representational dissimilarity matrices under standard (**b**) and bio-inspired (**c**) training; entries are pairwise cosine distances between category representations. **d–f,** Spearman correlation between model dissimilarities and human behavioral dissimilarities (one minus the similarity shown in **a**) across architectures (**d**), datasets (**e**), and multimodal training on MS-COCO (**f**). Blue and red markers denote standard and bio-inspired training, respectively; error bars show the standard deviation across three training seeds.

A.4 Alignment of visual representations with human behavioral similarity

We next asked whether the representational organization associated with bio-inspired training also corresponds to human similarity judgments. The behavioral similarity matrix shows strong within-group structure among the evaluated concepts (Supplementary Figure S4a). The standard model dissimilarity matrix shows structure across these groups (Supplementary Figure S4b), as does the bio-inspired model matrix (Supplementary Figure S4c). We therefore compared the rank ordering of model-derived distances with the corresponding human behavioral dissimilarities, using the same Spearman-based alignment approach as for the taxonomic reference.

Across AlexNet, ResNet-18, ViT-S, and Mixer-S, mean human alignment was higher following bio-inspired than standard training (Supplementary Figure S4d). The same direction held for ImageNet, Visual Genome, and Open Images (Supplementary Figure S4e), as well as for the multimodal model trained on MS-COCO (Supplementary Figure S4f). These comparisons indicate that the alignment effect extends to the tested behavioral similarity reference across the examined architectures and training settings.

A.5 Taxonomic alignment within animate and inanimate subsets

We next assessed whether the taxonomic alignment results in Figure 3 were shared across concept domains. When alignment was computed only among animate concepts, bio-inspired training yielded higher mean Spearman correlations across all four architectures (Supplementary Figure S5a), all three datasets (Supplementary Figure S5b), and the MS-COCO multimodal comparison (Supplementary Figure S5c). The size of the difference varied across settings, but its direction was consistent in this subset.

The inanimate subset, in contrast, showed a different pattern. Architecture comparisons included a small increase for AlexNet and lower mean alignment for ResNet-18, ViT-S, and Mixer-S under bio-inspired training (Supplementary Figure S5d). Mean alignment was also lower for bio-inspired training in the ImageNet, Visual Genome, and Open Images comparisons (Supplementary Figure S5e), while the MS-COCO values were nearly unchanged (Supplementary Figure S5f). Thus, the advantage of bio-inspired training was more pronounced for the animate subset and not uniform across the tested concept domains. Although animacy is an important axis in human object representations and early infant categorization [32, 33], these results do not establish that animacy itself uniquely explains the difference.

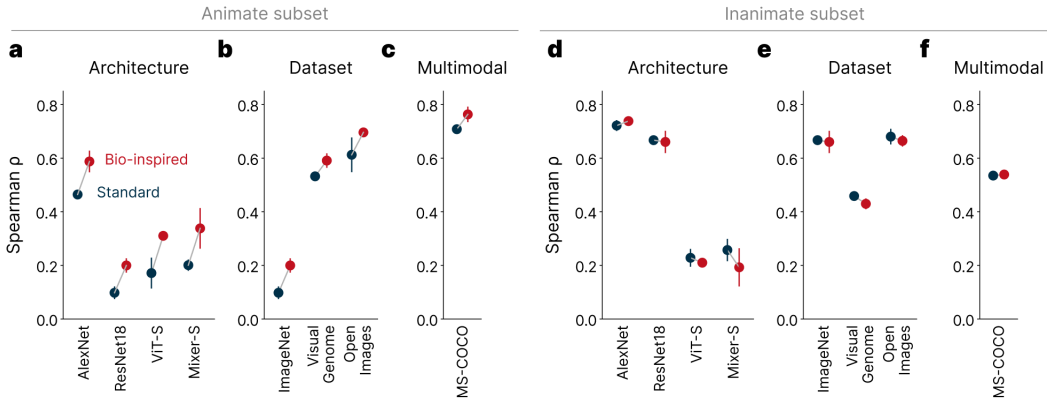


Figure S5: **Taxonomic alignment within animate and inanimate concept subsets. a–c,** Spearman correlation between taxonomy distances and model representational dissimilarities among animate concepts, compared across architectures (**a**), datasets (**b**), and multimodal training on MS-COCO (**c**). **d–f,** Corresponding comparisons among inanimate concepts across architectures (**d**), datasets (**e**), and MS-COCO (**f**). Blue and red markers denote standard and bio-inspired training, respectively; error bars show the standard deviation across three training seeds.

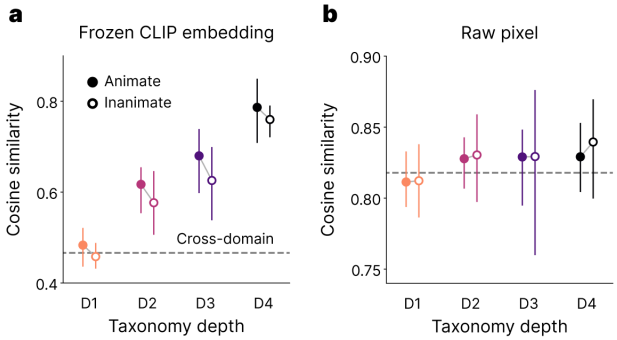


Figure S6: **Within-domain visual similarity across taxonomic depths. a,** Mean cosine similarity between image pairs from animate (filled markers) or inanimate (open markers) concepts at each exclusive taxonomy depth (D1–D4), measured in frozen CLIP image embeddings. **b,** The same pairwise analysis using raw image pixels. Each pair contributes to the depth of its deepest shared taxonomic level. The horizontal dashed line shows the mean similarity of cross-domain animate–inanimate pairs. Error bars indicate 95% bootstrap confidence intervals.

To characterize the visual structure available at different taxonomic depths, we compared image-pair similarities separately within animate and inanimate concepts. In frozen CLIP embeddings, mean within-domain cosine similarity increased from broad to fine shared taxonomic levels for both domains, with animate pairs showing slightly higher mean similarity than inanimate pairs at every depth; the cross-domain baseline was below the D2–D4 means (Supplementary Figure S6a). Raw-pixel similarities were higher overall and changed much less across depths, with overlapping uncertainty intervals for many levels and little separation between the two domains (Supplementary Figure S6b).

Together with the subset-specific alignment results, this pattern is consistent with a larger benefit of bio-inspired training for concepts that are more coherent in visual feature space. The comparison involves only two predefined domains, however, and does not separate visual coherence from animacy or other domain properties. We therefore interpret the association as a possible explanation for the uneven alignment benefit, rather than as evidence that animacy plays a unique role in the training effect.

A.6 Timing of color and acuity transitions in developmental training

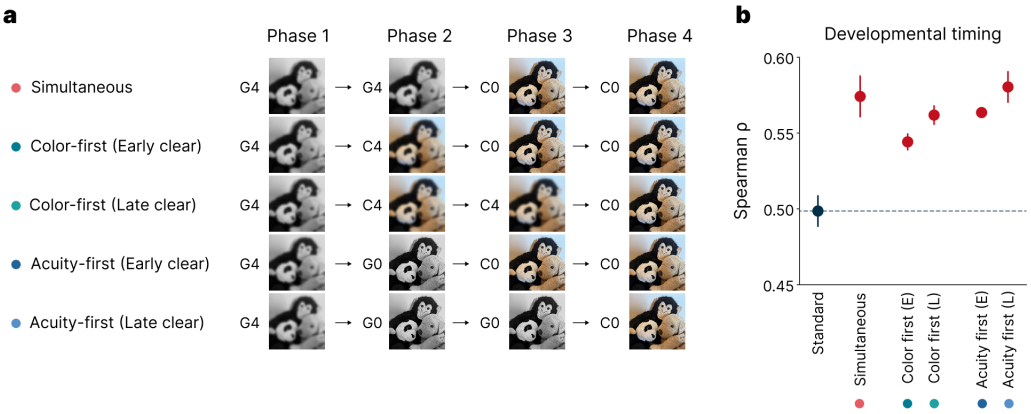


Figure S7: **Effects of developmental transition timing on taxonomic alignment. a,** Visual input sequences for simultaneous, color-first, and acuity-first developmental schedules across four training phases. C and G denote color and grayscale images, and 4 or 0 denotes Gaussian blur intensity (σ, in pixels). Early-clear schedules reach clear color input after two phases; late-clear schedules reach it in the final phase. **b,** Spearman correlation between the taxonomy distance matrix and final model representational dissimilarities under standard training and the five developmental schedules. Points and error bars show the mean and standard deviation across three training seeds; the horizontal dashed line indicates the standard-training mean.

Finally, we examined whether the order and timing of color and acuity development affect taxonomic alignment. The schedules in Supplementary Figure S7a begin with blurred grayscale input and then either introduce color before removing blur, remove blur before introducing color, or change both properties together. We also varied whether the staggered schedules reached clear color input midway through training or only in the final phase.

All five developmental schedules had higher mean Spearman alignment than standard training in this comparison (Supplementary Figure S7b). The simultaneous transition and the late-clear acuity-first schedule had the highest means, whereas early-clear color-first training had the lowest mean among the developmental schedules. Within both the color-first and acuity-first pairs, later introduction of clear color input was associated with higher mean alignment; the two orders also produced different values. These results indicate that the structure of the developmental schedule modulates the magnitude of taxonomic alignment.

B Numerical values for main results

The tables below report Spearman and Pearson correlations between model representational dissimilarities and either taxonomic distances (*Taxonomy*) or human behavioral dissimilarities (*Human*; derived from SPoSE embeddings [45]). All values use the full set of 32 concepts and its 496 unique concept pairs. Entries are the mean ± sample standard deviation across training seeds.

Table S1: Taxonomic and human alignment across visual architectures (Figure 3d; Supplementary Figure S4d). All models were evaluated after 90 training epochs ($n = 3$ seeds per condition).

		Taxonomy		Human	
Architecture	Training	Spearman (ρ)	Pearson (r)	Spearman (ρ)	Pearson (r)
AlexNet	Standard	0.499 ± 0.010	0.600 ± 0.008	0.473 ± 0.015	0.548 ± 0.010
	Bio-inspired	0.574 ± 0.014	0.664 ± 0.011	0.544 ± 0.019	0.615 ± 0.013
ResNet-18	Standard	0.254 ± 0.006	0.389 ± 0.008	0.331 ± 0.011	0.387 ± 0.016
	Bio-inspired	0.282 ± 0.019	0.417 ± 0.024	0.348 ± 0.015	0.409 ± 0.025
ViT-S	Standard	0.201 ± 0.032	0.390 ± 0.012	0.158 ± 0.020	0.319 ± 0.014
	Bio-inspired	0.231 ± 0.025	0.415 ± 0.016	0.213 ± 0.027	0.361 ± 0.019
Mixer-S	Standard	0.177 ± 0.035	0.383 ± 0.011	0.121 ± 0.021	0.310 ± 0.016
	Bio-inspired	0.237 ± 0.049	0.442 ± 0.022	0.208 ± 0.056	0.385 ± 0.028

Table S2: **Taxonomic and human alignment across training datasets** (Figure 3e; Supplementary Figure S4e). ResNet-18 models were evaluated after 90 training epochs ($n = 3$ seeds per condition).

		Taxonomy		Human	
Dataset	Training	Spearman (ρ)	Pearson (r)	Spearman (ρ)	Pearson (r)
ImageNet	Standard	0.254 ± 0.006	0.389 ± 0.008	0.331 ± 0.011	0.387 ± 0.016
	Bio-inspired	0.282 ± 0.019	0.417 ± 0.024	0.348 ± 0.015	0.409 ± 0.025
Visual Genome	Standard	0.406 ± 0.014	0.513 ± 0.020	0.359 ± 0.016	0.468 ± 0.015
	Bio-inspired	0.485 ± 0.024	0.571 ± 0.014	0.435 ± 0.028	0.526 ± 0.015
Open Images	Standard	0.500 ± 0.024	0.595 ± 0.031	0.435 ± 0.022	0.517 ± 0.037
	Bio-inspired	0.587 ± 0.021	0.657 ± 0.018	0.525 ± 0.029	0.584 ± 0.018

Table S3: **Taxonomic and human alignment for multimodal training on MS-COCO** (Figure 3f; Supplementary Figure S4f). The bio-inspired condition transitions from blurred grayscale input ($\sigma = 2$ px) to clear color input; values summarize $n = 3$ training seeds.

	Taxonomy		Human	
Training	Spearman (ρ)	Pearson (r)	Spearman (ρ)	Pearson (r)
Standard	0.594 ± 0.012	0.665 ± 0.004	0.559 ± 0.007	0.630 ± 0.008
Bio-inspired	0.659 ± 0.007	0.703 ± 0.007	0.608 ± 0.008	0.663 ± 0.007

Table S4: **Taxonomic and human alignment across visual input conditions and schedules after 200 training epochs** (Figure 4). C and G indicate color and grayscale input; the number gives Gaussian blur intensity σ in pixels. Arrows denote transitions between 100-epoch training phases. Values summarize $n = 5$ training seeds per condition.

	Taxonomy		Human	
Condition	Spearman (ρ)	Pearson (r)	Spearman (ρ)	Pearson (r)
C0 (standard)	0.565 ± 0.030	0.656 ± 0.019	0.537 ± 0.031	0.614 ± 0.021
C4	0.397 ± 0.030	0.393 ± 0.019	0.386 ± 0.019	0.401 ± 0.009
G0	0.625 ± 0.017	0.682 ± 0.006	0.589 ± 0.011	0.630 ± 0.006
G4	0.338 ± 0.013	0.291 ± 0.027	0.309 ± 0.019	0.279 ± 0.035
C4→C0	0.565 ± 0.007	0.670 ± 0.007	0.523 ± 0.010	0.627 ± 0.005
G4→G0	0.658 ± 0.017	0.715 ± 0.014	0.599 ± 0.015	0.667 ± 0.012
G0→C0	0.628 ± 0.021	0.686 ± 0.014	0.599 ± 0.017	0.642 ± 0.014
G4→C0	0.646 ± 0.014	0.710 ± 0.009	0.599 ± 0.015	0.667 ± 0.008

Table S5: **Taxonomic and human alignment by initial blur intensity** (Figure 5a). Bio-inspired models received blurred grayscale input at the indicated Gaussian blur intensity for the first 45 epochs, followed by clear color input for 45 epochs. The standard model received clear color input throughout; each condition has $n = 3$ training seeds.

	Taxonomy		Human	
Initial blur	Spearman (ρ)	Pearson (r)	Spearman (ρ)	Pearson (r)
C0 (standard)	0.499 ± 0.010	0.600 ± 0.008	0.473 ± 0.015	0.548 ± 0.010
$\sigma = 1$ px	0.542 ± 0.003	0.634 ± 0.005	0.512 ± 0.006	0.581 ± 0.003
$\sigma = 2$ px	0.549 ± 0.019	0.643 ± 0.016	0.517 ± 0.012	0.592 ± 0.011
$\sigma = 3$ px	0.555 ± 0.009	0.650 ± 0.008	0.524 ± 0.008	0.600 ± 0.006
$\sigma = 4$ px	0.574 ± 0.014	0.664 ± 0.011	0.544 ± 0.019	0.615 ± 0.013
$\sigma = 5$ px	0.564 ± 0.004	0.657 ± 0.003	0.536 ± 0.004	0.613 ± 0.004

Table S6: **Taxonomic and human alignment across developmental transition schedules after 90 training epochs** (Supplementary Figure S7b). Simultaneous training changes from blurred grayscale (G4) to clear color (C0) input halfway through training. Color-first and acuity-first schedules vary whether clear color input is introduced midway (early clear) or in the final phase (late clear). Values summarize $n = 3$ training seeds per condition.

	Taxonomy		Human	
Condition	Spearman (ρ)	Pearson (r)	Spearman (ρ)	Pearson (r)
C0 (standard)	0.499 ± 0.010	0.600 ± 0.008	0.473 ± 0.015	0.548 ± 0.010
G4→C0 (simultaneous)	0.574 ± 0.014	0.664 ± 0.011	0.544 ± 0.019	0.615 ± 0.013
Color-first, early clear	0.544 ± 0.006	0.645 ± 0.002	0.518 ± 0.003	0.599 ± 0.005
Color-first, late clear	0.562 ± 0.006	0.663 ± 0.006	0.525 ± 0.010	0.613 ± 0.008
Acuity-first, early clear	0.563 ± 0.003	0.654 ± 0.002	0.536 ± 0.001	0.605 ± 0.000
Acuity-first, late clear	0.580 ± 0.010	0.665 ± 0.009	0.546 ± 0.006	0.610 ± 0.005

C Methods

C.1 Semantic taxonomy and evaluation sets

Our analyses use a four-level semantic hierarchy containing 2 classes at Depth 1, 4 at Depth 2, 14 at Depth 3, and 32 leaf concepts at Depth 4. The 32 concepts are divided equally between animate and inanimate domains. For each concept, we selected 12 images from the THINGS image database, yielding a fixed evaluation set of 384 images. The same concept identities and selected images were used across visual models, training schedules, and the multimodal evaluation. Concept paths and image identities were fixed in a common manifest, and the 32 concepts were linked to the corresponding human behavioral embeddings for the reference analysis.

The supervised visual training experiments used ImageNet; dataset replications used Visual Genome and Open Images with ResNet-18. These training corpora were separate from the fixed THINGS evaluation images. The multimodal experiment used MS-COCO image–caption pairs for training and the same THINGS concept set for the final representational analysis. We evaluated the learned visual representations on THINGS images except where an analysis explicitly tested degraded inputs or measured a checkpoint during the degraded training phase.

C.2 Visual degradation and developmental schedules

Visual degradation was implemented by converting RGB input to grayscale, applying Gaussian blur, or combining both operations. Grayscale input was replicated across the three image channels, and blur strength was specified by the Gaussian standard deviation σ in pixels. We use C and G to denote color and grayscale input, respectively, followed by blur strength; C0 is clear color input, G0 is clear grayscale input, and G4 is grayscale input with $\sigma = 4$ px. For each matched comparison, the training condition changed the visual input schedule while keeping the model and training configuration unchanged.

The primary 200-epoch comparison used C0 throughout or G4 for 100 epochs followed by C0 for 100 epochs. The 200-epoch ablations also included constant C4, G0, and G4 conditions, as well as the C4→C0, G4→G0, G0→C0, and G4→C0 transitions, each with a 100-epoch first phase. In the 90-epoch comparisons, the standard condition used C0 throughout, while the simultaneous bio-inspired condition used G4 for 45 epochs and C0 for the remaining 45 epochs. The blur-strength sweep replaced the initial G4 phase with G1, G2, G3, or G5. The staggered schedules introduced color before removing blur (G4→C4→C0) or removed blur before introducing color (G4→G0→C0). In these schedules, the first transition occurred approximately one quarter of the way through training; clear color input began either halfway through training or in the final quarter. The multimodal bio-inspired condition used G2 for the first half of its optimization steps and C0 thereafter.

C.3 Visual model architectures and training

Supervised image classifiers were trained with the same standard or bio-inspired input conditions within each model comparison. The primary 200-epoch experiments used an AlexNet-style classifier, with checkpoints saved every 10 epochs for the learning-trajectory analysis (Results 3.2). The architectural comparisons used AlexNet, ResNet-18, ViT-S/16, and Mixer-S/16 trained for 90 epochs on ImageNet. Dataset comparisons used ResNet-18 trained for 90 epochs on ImageNet, Visual Genome, or Open Images. Each architecture and dataset comparison used three independent training seeds per condition (Results 3.3). The 200-epoch schedule comparisons used five available seeds per condition (Results 3.4).

The corresponding standard and bio-inspired runs used the same architecture, classification objective, optimizer settings, and training duration; only the visual input schedule differed. After training, models were evaluated on the fixed taxonomy dataset to obtain representational alignment scores. ImageNet top-1 accuracy was measured on clean validation images at the same final checkpoint used for the alignment calculation.

The 200-epoch AlexNet trajectory runs used a batch size of 128, SGD with momentum 0.9 and Nesterov acceleration, a constant learning rate of 10^{-3}, no weight decay, and cross-entropy loss without label smoothing. Images were resized to 256×256 pixels, randomly cropped to 227×227 pixels, and horizontally flipped with probability 0.5. For the 90-epoch classifier training, the batch

size was 256, and cross-entropy label smoothing was 0.1. ResNet-18 used SGD with momentum 0.9, learning rate 0.1, and weight decay 10^{-4}; ViT-S/16 and Mixer-S/16 used AdamW with learning rate 5×10^{-4} and weight decay 0.05. AlexNet used SGD with momentum 0.9, learning rate 0.01, and weight decay 5×10^{-4}. These 90-epoch runs used a cosine learning-rate schedule, with five warm-up epochs for ViT-S/16 and Mixer-S/16 and none for ResNet-18 or AlexNet.

C.4 Multimodal model and training

The multimodal comparison used a CLIP-style model with a ResNet-18 image encoder, attention pooling, a causal text Transformer, and a 512-dimensional joint embedding space. Following the CLIP visual encoder design [36], we retained the spatial ResNet feature map and used attention pooling to form the image embedding. A spatial-mean token with a learned positional embedding attends to the feature-map positions, and the pooled output is projected into the shared embedding space. Image and text embeddings were L2-normalized and trained with a symmetric image–text contrastive loss on MS-COCO image–caption pairs. Standard and bio-inspired runs used the same architecture and optimization setup. Training lasted 30,000 steps with a batch size of 1,024, AdamW (learning rate 2×10^{-3}; weight decay 0.1), a cosine learning-rate schedule decaying to zero, and 6,000 warm-up steps. Three independent seeds were evaluated for each condition.

The standard model received clear color images throughout training. In the bio-inspired condition, grayscale conversion and Gaussian blur with $\sigma = 2$ px were applied during the first half of training, followed by clear color input. For representational comparison, the final 30,000-step checkpoints were used to encode images from the taxonomy dataset used for the supervised visual models. Alignment was computed from their image embeddings, without using the text embeddings in the RDM calculation.

C.5 Hierarchical classification and learning trajectories

To measure category information under degraded input, we used the frozen image encoder of the pretrained `openai/clip-vit-base-patch32` model [36]. The images were evaluated in color and grayscale at Gaussian blur strengths $\sigma \in \{0, 0.5, 1, 1.5, 2, 3, 4, 6, 8\}$ px. A separate linear probe was trained and tested at each input condition and each taxonomy depth. Probes standardized features using training-fold statistics and used ridge classification with regularization coefficient 1. We used four-fold cross-validation stratified by leaf concept and repeated the splits with three seeds. Each fold held out three of the 12 images per concept; reported accuracies were averaged over the split seeds.

For the developmental trajectories, the 200-epoch standard and G4→C0 models were evaluated at epochs 0, 10, …, 200 using features from the fifth convolutional activation layer. The same concept-stratified four-fold probe and three split seeds were applied at each checkpoint, for five training seeds per condition. Images were presented in the input condition active at the evaluated checkpoint: blurred grayscale through epoch 100 of the bio-inspired run and clear color thereafter. Probe accuracy at each depth was normalized within each training seed and condition by subtracting its epoch-0 accuracy and dividing by its epoch-0-to-200 gain, as defined in Eq. 1. Curves show the mean across training seeds, and error bars show the sample standard deviation.

C.6 Permuted-taxonomy control

For the frozen-CLIP degradation analysis, we additionally assigned whole concepts to 32 randomly permuted taxonomy leaf positions, retaining the original number of classes and branch sizes at each depth. We refitted the D1–D3 linear probes for 50 such assignments using the same images, input conditions, and four-fold concept-stratified cross-validation protocol as for the original taxonomy; this control used one cross-validation split seed. D4 concept labels were unchanged by the permutations. For each input condition, we report the mean accuracy and 5th–95th percentiles across permutations. We also report chance-adjusted accuracy $(a - 1/K)/(1 - 1/K)$ for both taxonomies, where K is the number of classes at the evaluated depth.

To test whether the trajectory depended on the assignment of concepts to the hierarchy, we randomly permuted whole concepts among the 32 taxonomy leaf slots while keeping the branch sizes and the number of classes at each depth fixed. All images of a concept inherited the same reassigned path, and the same set of 50 permutations was used for all training seeds and checkpoints. The

frozen network features, evaluated images, probe splits, and checkpoint epochs matched the original trajectory analysis. Probes were refitted for the permuted Depth 1–3 labels; Depth 4 retained the original concept labels. For each permutation, accuracy was normalized by its own initial and final values before averaging over permutations and then across the five training seeds.

C.7 Representational and reference dissimilarity matrices

For each evaluated checkpoint, we first L2-normalized the representation of each clear color image and averaged the 12 image representations belonging to a concept. Pairwise cosine distances between the resulting 32 concept centroids formed the model representational dissimilarity matrix (RDM). The architecture comparisons used ReLU6 for AlexNet, average-pooled features for ResNet-18, the final normalized class-token feature for ViT-S, and the final normalized pooled feature for Mixer-S; the multimodal comparison used the final image embedding. Layer-wise analysis repeated the RDM construction at five convolutional and two fully connected activations of the 200-epoch visual model. All model RDMs used the fixed concept ordering of the evaluation manifest.

The taxonomy reference RDM assigned distance $D_{ij}^{\mathrm{tax}} = 4 - L_{ij}$ to concepts i and j, where L_{ij} is the depth of their deepest shared taxonomic node below the root. The human reference RDM used one minus the cosine similarity between the corresponding 66-dimensional SPoSE behavioral embeddings [46]. Both references used the same concept ordering as the model RDMs.

C.8 Spearman alignment and statistical aggregation

We measured alignment by Spearman rank correlation between the upper-triangular, off-diagonal entries of a model RDM and a reference RDM, as in Eq. 3. The full 32-concept analysis contained 496 unique concept pairs. The animate-only and inanimate-only analyses each contained 16 concepts and 120 pairs. Taxonomy and human alignment were scored separately, with average ranks assigned to tied distances.

For each condition, we calculated the mean and sample standard deviation of the correlation across independently trained model seeds. The 90-epoch visual and multimodal comparisons used three seeds per condition, whereas the 200-epoch visual schedule comparisons used five available seeds per condition. Error bars in the alignment figures and the $\pm$ values in the numerical tables represent these across-seed standard deviations; they are descriptive and do not denote confidence intervals or hypothesis tests.

C.9 Visual similarity analysis

We evaluated similarity among the same images from the taxonomy dataset independently of the trained visual encoders. We compared two feature spaces: L2-normalized embeddings from a frozen pretrained CLIP ViT-B/32 image encoder and L2-normalized raw RGB pixels resized to 128×128 and flattened. Within each space, cosine similarity was computed for image pairs from concepts in the same animate or inanimate domain. Pairs were assigned to an exclusive taxonomy depth according to their deepest shared taxonomic level; a separate animate–inanimate pair mean provided the cross-domain baseline.

Within-domain means pooled all eligible image pairs at each depth. Uncertainty intervals were calculated with 10,000 bootstrap replicates that resampled leaf concepts with replacement within each domain and recomputed the pooled similarity. The plotted intervals are the 2.5th and 97.5th percentiles of these replicates. This analysis characterizes visual coherence in fixed feature spaces and does not itself test a causal effect of animacy or feature coherence on training outcomes.

D Data availability

The image sets used in this study are available from the following sources:

- ImageNet-1k (ILSVRC 2012) [41]: `https://www.image-net.org/challenges/LSVRC/2012/2012-downloads.php`.
- Visual Genome [42]: `https://visualgenome.org/api/v0/api_home`.
- Open Images V7 [43]: `https://storage.googleapis.com/openimages/web/download_v7.html`.
- MS-COCO 2017 [44]: `https://cocodataset.org`.
- THINGS [35]: `https://osf.io/jum2f/`.

The human similarity reference was derived from the publicly released 66-dimensional SPoSE behavioral embedding [46], available with the THINGS behavioral similarity data at `https://osf.io/f5rn6/`. The fixed concept-to-taxonomy mapping and image selection are specified in the analysis manifest accompanying this work. Original images remain subject to the access and reuse conditions of their respective source datasets.

E Computing resources

Model training and GPU-based evaluation were conducted on the MIT ORCD Engaging computing cluster using NVIDIA A100 or H100 GPUs. The models were implemented in PyTorch, with batched computations performed on the GPUs and input loading parallelized across CPU worker processes through PyTorch.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction clearly state the paper's main claims, including its contributions, assumptions, and limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: We include a separate "Limitations" section in the paper, where we discuss the scope of our claims and assumptions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: We describe detailed methods and instructions for reproducing the experimental results in the corresponding sections of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes].

Justification: We provide open access to the data and code via a GitHub repository.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: We provide experimental details in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: We report error bars for all results shown in the figures, as well as detailed numerical statistical values.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: We provide information on the compute resources used for the experiments in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: We reviewed the NeurIPS Code of Ethics, and our research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: We discuss both the potential positive and negative societal impacts of the work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: Since our paper is basic academic research, it does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: We cite the original papers that provide the publicly available assets and describe how to access them in the Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: We provide detailed documentation, such as README files, alongside the assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method development in this research does not involve LLMs as important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.