Visual Anagrams Reveal Hidden Differences in Holistic Shape Processing Across Vision Models

Fenil R. Doshi
Dept. of Psychology
& Kempner Institute
Harvard University
fenil_doshi@fas.harvard.edu

Thomas Fel
Kempner Institute
Harvard University
tfel@g.harvard.edu

Talia Konkle
Dept. of Psychology
& Kempner Institute
Harvard University
talia_konkle@harvard.edu

George A. Alvarez
Dept. of Psychology
& Kempner Institute
Harvard University
alvarez@wjh.harvard.edu

Abstract

Humans are able to recognize objects based on both local texture cues and the configuration of object parts, yet contemporary vision models primarily harvest local texture cues, yielding brittle, non-compositional features. Work on shape-vs-texture bias has pitted shape and texture representations in opposition, measuring shape relative to texture, ignoring the possibility that models (and humans) can simultaneously rely on both types of cues, and obscuring the absolute quality of both types of representation. We therefore recast shape evaluation as a matter of absolute configural competence, operationalized by the **Configural Shape Score (CSS)**, which *(i)* measures the ability to recognize both images in *Object-Anagram pairs* that preserve local texture while permuting global part arrangement to depict different object categories. Across 86 convolutional, transformer, and hybrid models, CSS *(ii)* uncovers a broad spectrum of configural sensitivity with fully self-supervised and language-aligned transformers – exemplified by DINOv2, SigLIP2 and EVA-CLIP – occupying the top end of the CSS spectrum. Mechanistic probes reveal that *(iii)* high-CSS networks depend on long-range interactions: radius-controlled attention masks abolish performance showing a distinctive U-shaped integration profile, and representational-similarity analyses expose a mid-depth transition from local to global coding. A BagNet control, whose receptive fields straddle patch seams, remains at chance *(iv)*, ruling out any "border-hacking" strategies. Finally, *(v)* we show that configural shape score also predicts other shape-dependent evals (e.g.,foreground bias, spectral and noise robustness). Overall, we propose that the path toward truly robust, generalizable, and human-like vision systems may not lie in forcing an artificial choice between shape and texture, but rather in architectural and learning frameworks that seamlessly integrate both local-texture and global configural shape. [1]

1   Introduction

Human object recognition is remarkably robust: we can effortlessly identify objects across dramatic variations in texture, scale, viewpoint, and context because we can focus on aspects of global

---

[1]Project Page: https://www.fenildoshi.com/configural-shape/

A) Visual Anagram example
The same puzzle pieces can be configured to either look like a wolf or an elephant

B) Object Anagram Dataset samples
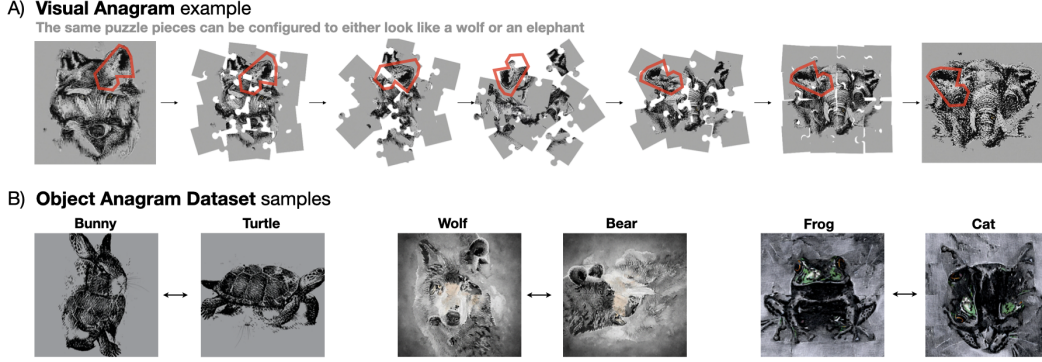Bunny — Turtle   Wolf — Bear   Frog — Cat

Figure 1: **Object-Anagram task: a probe of configural shape perception. (A)** visual-anagram example—an identical set of 16 square diffusion patches is spatially permuted to form two distinct objects, here a wolf and an elephant (one shared patch is outlined in red). **(B)** additional image pairs from the object-anagram benchmark. each pair comprises globally different objects built from the same unordered patch multiset, forcing any successful classifier to rely solely on the global arrangement of parts.

configuration that are stable across such local photometric quirks [1, 2, 3, 4, 5, 6]. By contrast, state-of-the-art vision networks still harvest local, high-frequency shortcuts [7, 8, 9]. This strategy achieves high ImageNet accuracy [10] but leaves models brittle under texture shifts, adversarial noise, and compositional out-of-distribution stresses [11, 12, 13, 14]. The failure arises because models often seize on spurious yet linearly separable features when multiple predictive cues are available [15, 16, 17]. These differences between models and humans are often studied using the shape–versus–texture bias diagnostic, which pits shape vs. texture using cue-conflict stimuli [8], but this metric is inherently relative: scores rise whenever shape coding strengthens *or* when texture coding weakens, rendering the absolute fidelity of global shape ambiguous [18, 19, 20]. Effective vision systems should exploit both cues when helpful [21, 22, 23], motivating an absolute assessment of shape and texture processing A.1.

We close this gap by recasting shape evaluation as an absolute test of configural competence. Building on "visual anagrams" [24], we synthesize image–pairs that share an identical multiset of local diffusion patches yet differ in their permutations (Fig. 1). Correctly classifying both views demands sensitivity to spatial relations alone. We formalise the task through the **Configural Shape Score (CSS)**, a joint two–image criterion whose chance level is below 2% and whose ceiling mandates perfect configural sensitivity. Our study benchmarks 86 convolutional, transformer, and hybrid checkpoints— from BagNet [25], stylized [8] and adversarially robust CNNs [26], to fully self-supervised ViTs like DINOv2 [27, 28], and language-aligned models such as SigLIP [29, 30] and EVA-CLIP [31]. Combining behavioral metrics with mechanistic probes yields five key findings:

- The nine-category *Object-Anagram* dataset enforces a stringent falsification test for holistic vision by permuting the global arrangement of an invariant multiset of local patches; any success therefore hinges on configural integration. The Configural Shape Score over this dataset gives an absolute score of configural shape, rigorously decoupling genuine shape inference from the artefactual gains that cue-conflict paradigms can achieve through mere texture suppression.

- Vision transformers optimized via self-supervised learning and language-alignment, notably DINOv2 [27], EVA-CLIP [31] and SigLIP2 [30], dominate the CSS spectrum; their global-consistency objectives appear uniquely effective at instilling holistic shape representations, whereas comparably accurate, purely supervised counterparts achieve lower configural shape scores.

- Mechanistic dissection reveals that high-CSS models leverage cross-patch communication spanning long-range interactions: performance collapses under radius-clipped attention masks with a *U-shaped* integration profile indicating that intermediate layers perform the key configural processing; representational-similarity analyses echo this profile, exposing a mid-depth pivot from local to global coding that is predicitive of overall CSS score.

2

- Architectures confined to local receptive fields, exemplified by BagNet, perform near chance, ruling out "border-hacking" and underscoring that authentic configural shape demands long-range integration.
- Models with higher configural shape scores also score high on other shape-dependent evals like foregorund-vs-background bias, robustness to noise, phase dependence and critical band masking.

By converting a long-standing theoretical critique into a falsifiable measurement and linking the resulting scores to identifiable computational mechanisms, this work advances the science of holistic shape perception and offers actionable design principles for future vision systems.

2 Related Work

Configural and Holistic Shape Processing in Human Vision: In humans, there is evidence that configural shape processing is multifaceted, and the term broadly encapsulates any computation where the precise arrangement of parts affects the representation of object appearance or identity [1, 4, 5]. There is even evidence that the appearance and recognition of local parts can be influenced by long-range interactions with other distal parts [32], indicating that contextual modulation is an important component of configural processing. These forms of configural shape processing can be distinguished from texture-based representations, where items can appear to have the same texture despite spatial shifts in local features or parts, as long as the key higher-order statistical properties are the same [33, 34].

Computational Approaches to shape sensitivity: Prior work has investigated shape representations and texture bias in vision models [8, 7, 35, 36, 37, 38, 39], often framing these issues in terms of shortcut learning driven by spurious correlations in the training data [15, 13]. Building on this, more recent studies have begun to probe whether models are sensitive to the spatial configuration of object parts [40, 41, 42, 43]. However, these efforts typically rely on synthetic datasets and/or require explicit fine-tuning, focusing on understanding whether an architecture is at all capable of supporting relational reasoning, rather than whether such sensitivities emerge naturally during training. A notable exception is work by Baker and Elder [9], who tested whether pretrained models could detect disruptions in object configuration. Their approach involved splitting silhouette images along the horizontal meridian, flipping the bottom half, and stitching the parts back together. This manipulation was intended to break global structure while preserving local part content. However, this manipulation has some limitations: it is ineffective for symmetric shapes, and the use of black-and-white silhouettes can obscure subtle configural differences.

3 Object Anagram Dataset and Configural Shape Score (CSS)

Background and notation. Consider the classical supervised-learning paradigm, where $\mathcal{X}$ denotes the image space and $\mathcal{Y} = [C]$ the index set of C distinct categories. A classifier $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ maps an input image $\boldsymbol{x}$ to a predicted label $\boldsymbol{y}$. Our goal is to probe configural shape acuity: the ability to parse global part arrangements while remaining invariant to permutations of local texture elements. To do so we fix a grid of $K = 16$ equal-area square patches. Let $\mathfrak{S}_K$ be the symmetric group of the $K!$ possible permutations, and write $\boldsymbol{\pi} \in \mathfrak{S}_K$ for an element thereof. Given an ordered multiset of patches $\mathcal{P} = \{\boldsymbol{p}_k\}_{k=1}^{K}$ with $\boldsymbol{p}_k \in \mathbb{R}^{h \times w \times 3}$, we define the composition operator:

$$\boldsymbol{\Gamma}(\mathcal{P}, \boldsymbol{\pi}) = \begin{bmatrix} \boldsymbol{p}_{\pi(1)} & \cdots & \boldsymbol{p}_{\pi(4)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{p}_{\pi(13)} & \cdots & \boldsymbol{p}_{\pi(16)} \end{bmatrix} \in \mathcal{X},$$

which re-assembles the permuted patches into a 256×256 canvas. The permutation $\boldsymbol{\pi}$ therefore fully determines the global layout.

Object Anagram Dataset synthesis. The synthesis pipeline is directly adapted from [24]. For every ordered label pair $(y_1, y_2) \in \mathcal{Y}^2$ we prepare a text–layout tuple $\left(\boldsymbol{c}(y_j), \boldsymbol{\pi}_j\right)_{j=1,2}$, where $\boldsymbol{c}(y_j)$ encodes the prompt *"high-quality painting of a well-shown y_j with simple black paint texture on a grey background"* using a pretrained T5 encoder, and where $\boldsymbol{\pi}_1 = \mathrm{id}$ while $\boldsymbol{\pi}_2 \neq \boldsymbol{\pi}_1$ is drawn uniformly from $\mathfrak{S}_K$. Both tuples share a common Gaussian seed, ensuring identical low-level texture statistics, and are injected into the DeepFloyd-IF pipeline[2]. To maintain texture consistency while

---

[2]Available at : `https://github.com/deep-floyd/IF`.

supporting distinct global configurations, we use a permutation operator $\Pi(\boldsymbol{z}, \boldsymbol{\pi})$ that rearranges $\boldsymbol{z}$ according to $\boldsymbol{\pi}$. At each reverse-diffusion timestep t, we compute two denoising predictions: $\boldsymbol{\epsilon}^{(1)}$ for the canonical arrangement (y_1) and $\boldsymbol{\epsilon}^{(2)}$ for the permuted arrangement (y_2). Formally:

$$\boldsymbol{\epsilon}^{(1)} = \boldsymbol{\varepsilon_\theta}(\boldsymbol{z}_t,\, t,\, \boldsymbol{c}(y_1)), \quad \boldsymbol{\epsilon}^{(2)} = \boldsymbol{\varepsilon_\theta}(\Pi(\boldsymbol{z}_t, \boldsymbol{\pi}_2),\, t,\, \boldsymbol{c}(y_1)).$$

These are combined into a symmetrized target

$$\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}^{(1)} \;+\; \Pi^{-1}\!\big(\boldsymbol{\epsilon}^{(2)},\, \boldsymbol{\pi}_2\big),$$

where $\Pi^{-1}(\cdot, \boldsymbol{\pi})$ inverses the permutation so that the two predictions align in the canonical frame. The reverse-diffusion update is then

$$\boldsymbol{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(\boldsymbol{z}_t - \tfrac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}}\,\boldsymbol{\epsilon}_t\Big) + \sigma_t\,\boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \qquad \bar\alpha_t = \prod_{s=1}^{t} \alpha_s,$$

with a cosine noise schedule α_t and variance $\sigma_t^2 = 1 - \alpha_t$. This procedure jointly optimizes both category representations, using identical patch content but differing spatial arrangements, progressively refining a shared image. In each timestep, we obtain $\boldsymbol{z}_0$ at 64×64 resolution and after T steps the resulting image seeds a second diffusion at 256×256 resolution. From the final image we extract the patch multiset $\mathcal{P}$ by partitioning it into a 4×4 grid, yielding sixteen patches that share texture but differ in arrangement across the two views:

$$\boldsymbol{x}^{(1)} = \boldsymbol{\Gamma}(\mathcal{P}, \boldsymbol{\pi}_1), \qquad \boldsymbol{x}^{(2)} = \boldsymbol{\Gamma}(\mathcal{P}, \boldsymbol{\pi}_2),$$

with ground-truth labels $(y^{(1)}, y^{(2)})$. The critical property of these image pairs is texture invariance: $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$ share exactly the same patch multiset, and therefore identical first- and many higher-order texture statistics (color distributions, edge patterns, local frequencies) while differing solely in global configuration. Consequently, local cues alone are insufficient for classification, making this dataset a stringent test of a model's configural processing capabilities.

Configural Shape Score. Gathering N such pairs yields $\mathcal{A} = \{(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, y_i^{(1)}, y_i^{(2)})\}_{i=1}^{N}$. Each image is centre-cropped to $224{\times}224$, normalised by the training statistics of $\boldsymbol{f}$, and forwarded through the network. Mapping the resulting ImageNet logits to the nine object-anagram categories (Appendix A.3) we define

$$\mathrm{CSS}(\boldsymbol{f}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big(\boldsymbol{f}(\boldsymbol{x}_i^{(1)}) = y_i^{(1)} \wedge \boldsymbol{f}(\boldsymbol{x}_i^{(2)}) = y_i^{(2)}\big),$$

whose chance level is $1/C^2$. Suppressing texture alone cannot raise the score; only a genuinely holistic integration of global layout yields high CSS values.

4   Vision Models

To dissect the computational determinants of configural shape sensitivity we assembled a suite of 86 pretrained models and four randomly initialized baselines that together span the principal axes of modern visual representation learning. Standard convolutional networks trained with cross-entropy on ImageNet: ResNet-50[44], VGG-16[45], and AlexNet [46] establish a supervised point of comparison, while three targeted manipulations of this template probe whether amplifying shape-vs-texture bias alone suffices: Stylized models [8], Adversarially Robust models [26], and Top-k Sparse models [47]. Architecturally bio-inspired models, including the CORnet family[48, 49], Long-Range Modulatory CNNs [50], and an Edge-AlexNet trained exclusively on edge statistics, test the hypothesis that neural plausibility intrinsically fosters holistic processing. The role of sheer data exposure is examined through the BiT checkpoints [51, 52] together with SWSL-ResNet-50 and SSL-ResNet-50 trained on billion-image corpora [53]. Recent architectural refinements are covered by ConvNeXt architectures [54, 55] as well as by ResNet-50, ResNet-101 and ViT-B/16 checkpoints trained with rigorous augmentation pipelines [56, 57]. A second axis contrasts convolutional and transformer principles. To this end, we include supervised Vision Transformers[58] along with its self-supervised counterparts: BEiT and BEiTv2 ViTs [59, 60], MAE-ViTs and Hiera-MAE-ViTs [61, 62], and DINOv2-ViTs [27, 28]. We also add version with language aligned encoders such as

CLIP ViTs [63, 64], SigLIP ViTs and SigLIP2 ViTs [29, 30], and EVA-CLIP ViTs[31, 65]. Within each ViT family, we included several variants (e.g. S/M/L variations). These comparisons isolate the contribution of transformer and multimodal objectives. Finally, BagNets [25], whose receptive fields never exceed local neighborhoods, serves as explicit local-only controls.

Collectively, this curated but diverse cohort will allow us to identify which architectural, training, and data regimes are predictive of high Configural Shape Score (see A.11 for the complete list of all the evaluated models)..

5 Models differ substantially in reliance on configural information for recognition
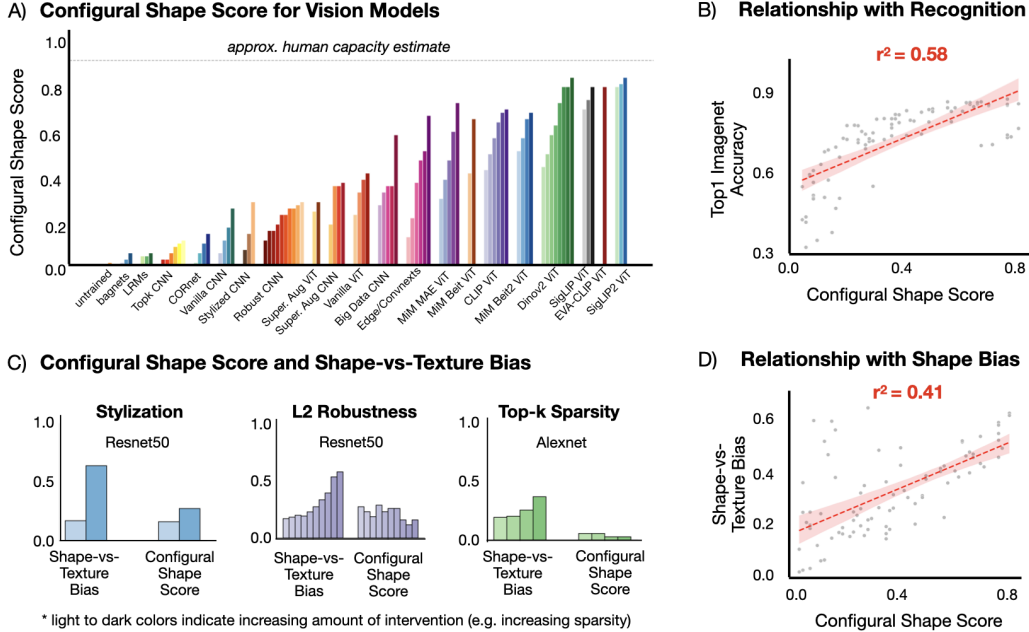

Figure 2: **Configural Shape Score (CSS) reveals variation across vision models matched in recognition performance and dissociates from imagenet accuracy and shape-vs-texture bias.** (A) CSS across 86 vision models, quantifying how accurately models recognize the distinct objects in each anagram pair. Human performance is shown as the dashed reference line. (B) Relationship between CSS and top-1 Imagenet Accuracy across all models. (C) CSS compared to shape-vs-texture bias for models trained with stylization, adversarial robustness, and Top-K sparsity. While these methods increase shape-vs-texture bias, they show modest-to-no gains in CSS. (D) Relationship between CSS and Shape-vs-Texture bias across all models.

Configural Shape Score varies widely across the full suite of models tested (Fig 2A), with the highest-scoring models approaching human-level scores (see A.4 for human experiment details), and the lowest scoring models demonstrating little-to-no configural shape sensitivity at all. Models that showed the highest CSS were either self-supervised ViTs(DINOv2s) or language-aligned ViTs (SigLIP and EVA-CLIP models). It is also notable that models with similar, generally high levels of ImageNet top-1 accuracy vary markedly in their CSS scores. For example, a supervised ViT-B/16 (ImageNet top-1: 76.35%; 23.61% CSS) and a language-aligned SigLIP ViT-B/16(ImageNet top-1: 74.96%; 77.78% CSS have near equivalent ImageNet top-1 accuracy, but achieve this through different reliance on configural shape information. Thus, achieving high accuracies on ImageNet is not enough to obtain a high Configural Shape Score, and ImageNet accuracy alone does not determine the Configural Shape Score (Fig. 2B).

Configural Shape Score also dissociates from Shape-vs-Texture bias. As shown in Fig. 2C, CSS is unaffected by three key strategies known to enhance shape-vs-texture bias: stylization-based training [8], adversarial training [26], and top-k activation pruning [47]. In stylization training, object textures are decorrelated from object identity during training, forcing models to rely more on shape than the

object's texture. Fig. 2C (left) shows that models trained with stylization (dark blue) have much higher shape-bias than models trained without stylization (light blue), but there's little to no effect of stylization on configural shape score. Adversarial training optimizes models for robustness to worst-case perturbations, varying strength of the adversarial attack with an epsilon parameter. Fig. 2C (center) shows that shape-bias increases with epsilon (purple), while CSS is unaffected by this manipulation. Finally, Top-k activation pruning restricts forward propagation to the highest-activating units within each layer, and increasing sparsity via this pruning (green bars) increases shape-bias but has no effect on configural shape score (2C right). Across these manipulations, we find that gains in CSS were modest-to-none compared to the substantial increases in shape bias. Finally, overall we find that the correlation between CSS and shape-vs-texture bias is moderate (r=0.64) indicating that only about 41% of the variance in CSS is accounted for by shape-vs-texture bias and vice versa (2D).

Taken together, these results indicate that the Configural Shape Score varies widely across models and dissociates from both ImageNet accuracy and Shape-vs-Texture bias. To gain deeper mechanistic insight into how models achieve high CSS, we next performed attentional ablation and representational similarity analyses.

6 Long-range Contextual Interactions lead to higher Configural Shape Scores in Vision Transformers
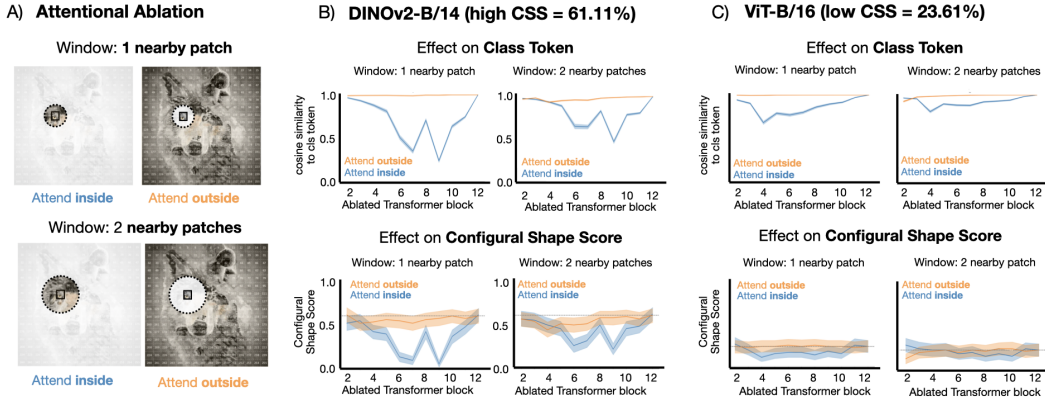

Figure 3: **Long-range Contextual Interactions leads to higher Configural Shape Score.** (A) Ablating self-attention in DINOv2-B/14 by selectively restricting each patch to attend only inside (blue) or outside (orange) a local window.Ablations are applied over windows with 1 or 2 nearby patches. (B) Effect of attentional ablation on the class token representation and configural shape score for high CSS model (Dinov2-B/14). Restricting attention to short-range interactions ("attend inside" condition - blue line) changes class tokens and disrupts CSS, most strongly at intermediate blocks. This effect is minimal when restricting attention to long-range interactions ("attend outside" condition - orange line). Dashed line shows CSS in unablated condition. (C) Effect of attentional ablation on the class token representation and configural shape score for low CSS model (ViT-B/16). Disruption for short-range interactions have reduced in this model.

Vision transformers provide a unique opportunity to examine the mechanisms of configural processing, because standard ViTs divide the image into a grid of patches and any configural processing (interactions between patch representations) must be performed via self-attention mechanisms. Thus, by targeting self-attention mechanisms with ablations, we can determine the relative impact of both short-range and long-range contextual interactions. Here we examined how intermediate attention mechanisms influence representational dynamics within DINOv2-B/14, a self-supervised ViT with 61.11% CSS and 84.1% top-1 ImageNet recognition. DINOv2-B/14 processes an input image of size 224×224 pixels by dividing it into a grid of 16×16 patches (each 14×14 pixels). For comparison, we also performed the ablation study on ViT-B/16, which achieved high top-1 ImageNet accuracy (76.35%) but had a low CSS (23.61%). We performed attentional ablations during inference at different intermediate stages of these models by selectively restricting each patch's attention within a targeted attention block.

6

To determine the relative impact of short-range and long-range attentional interactions, we defined two distinct attention masking conditions (Fig. 3A): (i) "attend inside," where each query patch attends only to patches within a specified Manhattan radius, and (ii) "attend outside," where attention is restricted to patches outside this radius. The class token was always allowed unrestricted attention in both conditions. We measured the cosine similarity of the original class token (unmasked) to the class token in both ablation conditions, as well as the CSS scores. If the class token representation depends mostly on short-range attention, then it should be unchanged for the "attend inside" masks, but should drop when "attending outside" (excluding critical short-range interactions). In contrast, if long-range interactions are most important, then the representation should be unchanged when attending outside, and should drop when attending inside (excluding the critical long-range interactions). We defined "short-range" attention interactions as those within a radius of 1 or 2 patches.

The results of these ablations for the high-CSS DINOv2 are shown in Fig. 3B. Attending only to very local neighborhoods (radius of 1 or 2 patches) substantially disrupted both class token representations and CSS (blue line, "attend inside" drops dramatically in middle layers). In contrast, attending only to more distant patches (orange curves, "attend outside") resulted in little-to-no change in class-token representations or CSS. Thus, it appears that attention interactions beyond at least 2 patches are necessary and sufficient to determine the class token representation and CSS score. Fig. 3C shows that a vision transformer (ViT-B/16) with lower CSS shows a reduced dependence on long-range interactions, suggesting that long-range attentional interactions are crucial for obtaining high-CSS.

Finally, these results show a U-shaped trend, suggesting that long-range interactions are particularly important in intermediate layers, and less important at early and late layers. These results suggest that early layers process patches relatively locally, then intermediate layers reinterpret and modify these local patch representations based on context, and then later stages aggregate locally over these contextually-modulated patch representations en route to the final model output. This observation is broadly consistent with work on LLMs, which suggest that early layers process text at the local token-level, followed by syntatic and broader contextual processing, and then finally return to more token-specific processing focusing on task-specific predictions and output generation [66, 67].

To further test the critical role of these intermediate layers, especially in deeper architectures where single-block ablations can be too subtle, we designed a more stringent cumulative ablation experiment on the full DINOv2 family. In this analysis, we measured the irrecoverable contribution of intermediate layers by systematically ablating long-range interactions in all blocks before or after a selected layer, and measuring it's impact on the class token's representation from the last block. The full methodology and results are detailed in A.5. The biggest drop was seen in the intermediate blocks across all model sizes, confirming that the mid-depth layers were indeed the locus of the most critical computations for configural processing in these models.

7 Relational Positional Encodings Boost Configural Shape Scores

While long-range attention is necessary for configural processing, the underlying mechanism for encoding spatial positions is also critical. To isolate this factor, we evaluated five variants of the supervised ViT-B/16 architecture, each with a different positional encoding (PE) scheme (Table 1). While the standard model using learned absolute PEs achieved a

Table 1: Configural Shape Scores (CSS) for ViT-B/16 variants.

Model Variant (ViT-B/16)	CSS (%)
Relative Positional Embeddings	38.89
RoPE (Rotary Position Embedding)	51.38
RoPE (Rotary Position Embedding) + Absolute PE	51.38
Mixed RoPE	48.61
Mixed RoPE + Absolute PE	52.77

CSS of 23.61%, models equipped with relative PEs, such as Rotary Position Embeddings (RoPE), more than doubled this score to over 51%. This result demonstrates that a model's ability to process the relational spatial structure of an image is critical for holistic shape perception, and that CSS rewards this capacity.

8 Disentangling the influence of object category and anagram puzzle pieces on configural shape score

The ablation experiments on vision transformers demonstrated the emergence of configural representations in the intermediate model layers. However, our ablation method only applies to models with
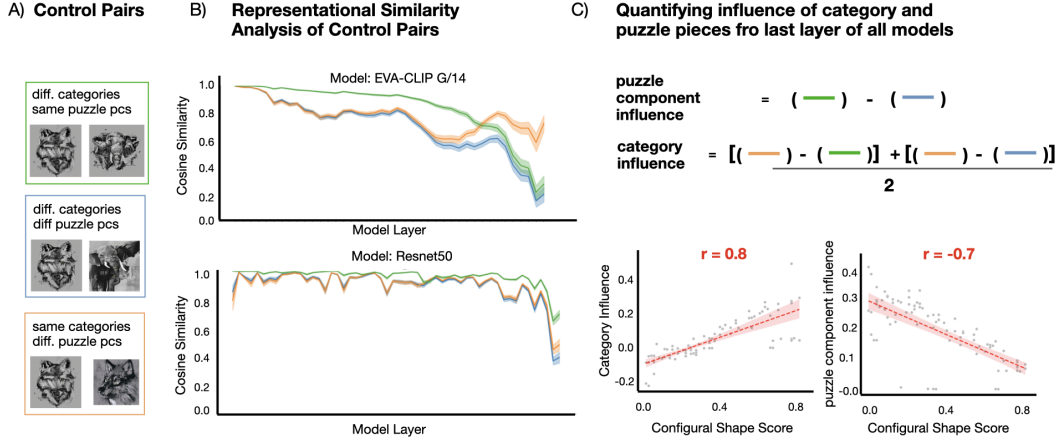
Figure 4: (A) Control pairs to tease apart category-level and component-level influence in model representations. (B) Cosine similarity across layers for each control pair type in EVA-CLIP G/14 and ResNet50. (C) Quantifying influence of object category vs. puzzle component from final layer embeddings. Models with higher Configural Shape Score (CSS) show stronger category influence and weaker component influence

self-attention mechanisms. To examine this transition from local to configural representations more generally for all model classes, we conducted a representational similarity analysis using a subset of carefully controlled image pairs from the Object Anagram Dataset. The purpose of this analysis was to determine at which point, if ever, different models transition from locally-driven representations dominated by puzzle-piece similarity to globally-driven representations dominated by categorical similarity. To disentangle the contributions of puzzle-piece similarity and configural-shape similarity, we measured the cosine-similarity between representations for three types of image pairs (Fig. 4A): (1) Same-parts/different-category: anagram pair composed of images sharing identical puzzle pieces but representing different global categories (e.g., wolf vs. elephant); (2) Different-parts/different-category: with containing different puzzle pieces and an equivalent categorical difference as a matched anagram pair (e.g., wolf vs. elephant); and (3) Different-parts/same-category pairs composed of different puzzle pieces but the same category (e.g., both wolves). We evaluated 60 image pairs of each type (180 total). If cosine-similarity depends on shared puzzle pieces, we would expect a higher correlation for the anagram pair (same-parts, different-category) than for either of the other pairs (which all have different parts). If cosine-similarity depends only on similarity in global-configuration (category), then it should be higher for the Different-Parts/Same-Category pairs and equally low for the the other pairs (which all have different category). Based on the ablation study, we expect high-CSS models to show configural effects emerging by middle-to-late model layers.

We quantified representational similarity using cosine similarity between image embeddings at each intermediate layer of a model. Fig. 4B shows layer-by-layer results for one selected high-CSS model (EVA-CLIP G/14 (CSS=77.78%), and one-selected low-CSS model (Resnet50, CSS=16.67%). Focusing first on the high-CSS model (top), several patterns emerge that are consistent with the idea that configural shape representations emerge in intermediate layers and dominate the final output of the model. First, in early model layers, local-similarity dominates: image pairs with shared parts (green) are more similar to each other than image pairs with different parts (orange and blue). Second, just beyond the midpoint, the effect of category similarity emerges: images with the different-parts/same-category (orange) begin to show greater similarity than the different-parts/different-category pairs (blue), and by the later layers the same-category pairs actually show *greater* similarity than the anagram pair (green). Indeed, by the final layers, the green/blue lines have collapsed together, indicating that having the same puzzle pieces is irrelevant by that point, and only the configural/category-level similarity matters. The results for the low-CSS Resnet50 are markedly different, and suggest that the ResNet50 model never shows a transition to more configuration-based processing. As shown in Fig 4B (bottom), the ResNet50 model shows greater part-based similarity (green) across all layers, including the final layers, and there is only a slight increase in similarity for

the different-part/same-category pairs (orange) relative to the different-part/different-category pairs (blue) at the final layer.

We formalized these qualitative observations by computing two metrics (Fig. 4C) – Puzzle component influence and Category influence – over the last and penultimate layer of all models. Puzzle component influence is the difference in cosine similarity between pairs with identical puzzle pieces (anagram pairs) and those with different puzzle pieces and the matched category differences (different-parts/different-category). Category Influence is the average similarity advantage for same-category pairs over different-category pairs, irrespective of local puzzle pieces. We then measure whether the configural shape scores can predict these metrics across all the models. Fig. 5C shows this relationship for the last layer. Higher Configural Shape Score corresponded to lower Puzzle Component Influence (negative correlations: $r = -0.70$ at the last layer, $r = -0.57$ at the penultimate layer), suggesting that higher-CSS models focus less on details of the local part appearance. Conversely, Configural Shape Score correlated positively with Category Influence ($r = 0.80$ at the last layer, $r = 0.83$ at the penultimate layer), indicating that high-CSS models encode representations that are shared between images within a category, while discriminating between categories. This trend is observed even when considering representations from DINOv2 backbones, which are fully self-supervised and have no pressure to form abstract category representations ($r = 0.94$ between CSS and Category Influence and $r = -0.86$ between CSS and Puzzle Component Influence).

Taken together with the results of the ablation study, these results suggest that long-range contextual interactions enable high-CSS models to transition from representations that are initially dominated by local parts, to representations that are dominated by a holistic view that depends on the configuration of parts — i.e., encodes the image as more than the sum of its parts, specifically in terms of relationships between those parts.

9 BagNets Provide Evidence Against a "Border Hacking" Solution

Can a local strategy be used to recognize both pairs of images in a visual anagram? When rearranging and rotating the puzzle pieces, what if features that emerge at the intersection between abutting pieces are sufficient to identify the global category of each image in a pair, yielding successful classification through non-configural "border-hacking"? BagNets [25] provide some evidence against the viability of a local solution for anagram recognition. These models have a ResNet-style architecture, except that the receptive field of the intermediate units is highly restricted (at max 9, 17, or 33 pixels throughout), making them very sensitive to local features and incapable of any kind of long-range spatial/contextual interaction. While they are competent in ImageNet recognition (with top1 accuracy: Bagnet9 at 41.38%, Bagnet17 at 55.08%, Bagnet33 at 61.28%), they all have very low CSS (Bagnet9 at 2.78%, Bagnet17 at 1.38%, Bagnet33 at 5.5%). Overall, the near-chance CSS of Bagnets underscores that fine-scale junction statistics alone are insufficient for anagram disambiguation, strengthening the interpretation of CSS as a global-configuration probe.

10 From Configural Shape Score to Broad Shape-Dependent Performance
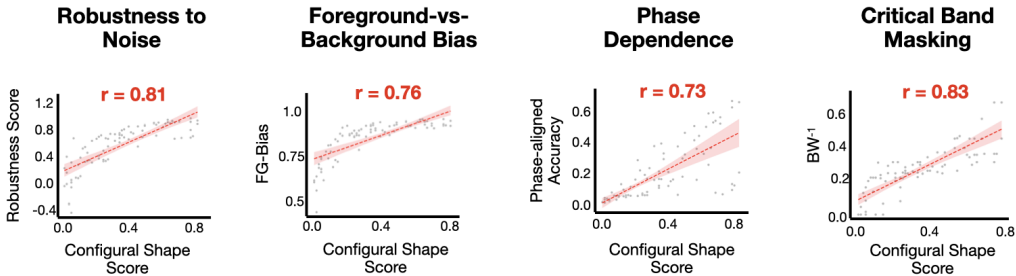

Figure 5: **Configural Shape Score (CSS) predicts model performance across a range of benchmarks.** CSS is positively correlated with foreground-vs-background bias, robustness to noise, phase dependence and critical band masking bandwidth

To what degree does having a high CSS predict other representational benefits and qualities? As shown in Fig. 5, we found that Configural Shape Score was significantly correlated with scores from several evals, including: 1) Robustness to Noise (r=0.81), testing each model's performance across varying severity levels for five distinct noise types, as described in [11]. 2) Foreground-vs-Background

9

Bias (r=0.76), tested using ImageNet-9 dataset from [68], quantifying the extent to which a model relies primarily on the foreground object rather than background information for classification. 3) Phase Dependence (r=0.73), assessed by swapping phase information in Fourier space between images and then measuring top-1 accuracy, quantifying the model's reliance on phase information [69]. 4) Critical band masking strategy outlined by [14] (r=0.83), used to determine the bandwidth of spatial frequencies essential for accurate object recognition. In contrast, the shape-vs-texture bias score across these models showed weaker relationships to these evals (r=0.62 with Robustness to Noise; r=0.32 with Foreground-vs-Background Bias; r=0.52 with Phase Dependence and r=0.55 with Critical Band Masking). Statistical comparison using Williams's test confirmed that CSS was a significantly better predictor of these metrics than shape-vs-texture bias (all p<0.001; see A.7 for test statistics and details). These results suggest that models with better configural shape scores also have other favorable and human-like perceptual qualities. See A.6 for more information about these evaluations, and A.8 and A.9 for feature attributions to qualitatively compare low- and high-CSS models on challenging stimuli from each benchmark.

11 Limitations and Discussion

Although the Object Anagram Dataset and the accompanying Configural Shape Score (CSS) provide a quantitative measure of holistic processing, several caveats warrant mention. First, the stimuli we generate are constrained by the priors of the diffusion model, and may explore only a subspace of configural encoding relationships. Furthermore, we use a uniform "black paint texture" to ensure local cues are perfectly matched, though this raises the possibility of out-of-distribution effects. However, our analyses show that even models with poor CSS can classify single anagram images successfully, suggesting the primary failure is in configural processing, not in handling the texture itself. Second, this work targets whole-object configurations and therefore does not directly probe part-based compositionality, an orthogonal facet of shape reasoning that future work should address. Third, despite containing thousands of composites, the dataset is modest in scale compared with modern billion-image corpora, suggesting that larger or more ecologically varied stimuli could reveal subtler effects. To that end, we created a 20x larger set of 1440 anagram pairs to test the robustness of our main findings. Re-evaluating all 86 models confirmed that the CSS metric is highly stable: the relative model rankings were preserved, and the scores from the original and expanded datasets showed a correlation of r=0.99 (see Appendix A.10).

Within these bounds our contributions are threefold. First, we formalized configural shape sensitivity as an interpretable metric distinct from texture reliance. Second, by charting CSS across 86 pretrained networks, we showed that holistic competence is neither fully explained by ImageNet accuracy nor canonical shape-vs-texture bias. Third, attention ablation and representational similarity analyses revealed that CSS relies on intermediate-stage, long-range interactions enabling recognition based on the configuration of parts and contextual relations. Finally, we demonstrated that models with higher CSS also perform better across other shape-dependent evaluations. These findings highlight configural shape processing as a critical yet underexplored dimension of visual intelligence and invite future work in advancing vision models toward human-like holistic representations.

Our results suggest that high configural competence stems primarily from training objectives that enforce local-to-global consistency. For instance, DINOv2's teacher-student objective, which trains a student network on local image crops to match the output distribution of a teacher network seeing global crops, directly rewards the integration of long-range information. This view-consistency objective leads to a more powerful signal than standard data augmentations alone or purely reconstructive approaches like MAE, whose local pixel-prediction task is insufficient on its own to produce the mid-layer "configural flip" characteristic of high-CSS models. Language supervision (BEiT-v2, EVA-CLIP, CLIP, SigLIP) would also enforce alignment between standard local-image crops and a global text caption, and this form of local-global consistency objective could also place an emphasis on multiple distant concepts in the image. Together, these findings point to a clear roadmap for building more shape-aware systems, which should prioritize (1) global-local consistency objectives, paired with (2) architectures capable of long-range integration, and (3) evaluated with explicit shape metrics like CSS.

References

[1] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[2] Andrew W Young, Deborah Hellawell, and Dennis C Hay. Configurational information in face perception. *Perception*, 16(6):747–759, 1987.

[3] Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.

[4] Daphne Maurer, Richard Le Grand, and Catherine J Mondloch. The many faces of configural processing. *Trends in cognitive sciences*, 6(6):255–260, 2002.

[5] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger Von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

[6] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

[7] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12): e1006613, 2018.

[8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

[9] Nicholas Baker and James H Elder. Deep learning models fail to capture the configural nature of human shape perception. *Iscience*, 25(9), 2022.

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[11] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[12] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[13] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

[14] Ajay Subramanian, Elena Sizikova, Najib Majaj, and Denis Pelli. Spatial-frequency channels, shape bias, and adversarial robustness. *Advances in neural information processing systems*, 36: 4137–4149, 2023.

[15] Katherine L Hermann, Hossein Mobahi, Thomas Fel, and Michael C Mozer. On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228*, 2023.

[16] Gaurav Malhotra, Marin Dujmović, and Jeffrey S Bowers. Feature blindness: A challenge for understanding and modelling visual object recognition. *PLOS Computational Biology*, 18(5): e1009572, 2022.

[17] Thomas Fel, Louis Bethune, Andrew Kyle Lampinen, Thomas Serre, and Katherine Hermann. Understanding visual feature reliance through the lens of complexity. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[18] Fenil R Doshi, Talia Konkle, and George A Alvarez. Quantifying the quality of shape and texture representations in deep neural network models. *Journal of Vision*, 24(10):1263–1263, 2024.

[19] Ziqi Wen, Tianqin Li, Zhi Jing, and Tai Sing Lee. Does resistance to style-transfer equal global shape bias? measuring network sensitivity to global shape configuration. *arXiv preprint arXiv:2310.07555*, 2023.

[20] Christian Jarvers and Heiko Neumann. Shape-selective processing in deep networks: integrating the evidence on perceptual integration. *Frontiers in Computer Science*, 5:1113609, 2023.

[21] Akshay V Jagadeesh and Justin L Gardner. Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, 119(17):e2115302119, 2022.

[22] Jenelle Feather, Guillaume Leclerc, Aleksander Mądry, and Josh H McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, 2023.

[23] Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. *arXiv preprint arXiv:2412.09606*, 2024.

[24] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24154–24163, 2024.

[25] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[28] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

[29] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[30] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[31] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023.

[32] James W Tanaka and Joseph A Sengco. Features and their configuration in face recognition. *Memory & cognition*, 25(5):583–592, 1997.

[33] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40:49–70, 2000.

[34] Benjamin J Balas. Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision research*, 46(3):299–309, 2006.

[35] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[36] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*, 2023.

[37] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Can we talk models into seeing the world differently? In *Thirteenth International Conference on Learning Representations*. OpenReview. net, 2025.

[38] Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[39] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[40] Hojin Jang, Pawan Sinha, and Xavier Boix. Configural processing as an optimized strategy for robust object recognition in neural networks. *Communications Biology*, 8(1):386, 2025.

[41] Fenil R Doshi, Talia Konkle, and George A Alvarez. A feedforward mechanism for human-like contour integration. *bioRxiv*, pages 2024–06, 2024.

[42] Drew Linsley, Junkyung Kim, Vijay Veerabadran, Charles Windolf, and Thomas Serre. Learning long-range spatial dependencies with horizontal gated recurrent units. *Advances in neural information processing systems*, 31, 2018.

[43] Valerio Biscione, Dong Yin, Gaurav Malhotra, Marin Dujmovic, Milton L Montero, Guillermo Puebla, Federico Adolfi, Rachel F Heaton, John E Hummel, Benjamin D Evans, et al. Mindset: Vision. a toolbox for testing dnns on key psychological experiments. *arXiv preprint arXiv:2404.05290*, 2024.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[45] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[47] Tianqin Li, Ziqi Wen, Yangfan Li, and Tai Sing Lee. Emergence of shape bias in convolutional neural networks through activation sparsity. *Advances in Neural Information Processing Systems*, 36:71755–71766, 2023.

[48] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, page 408385, 2018.

[49] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

[50] Talia Konkle and George Alvarez. Cognitive steering in deep neural networks via long-range modulatory feedback connections. *Advances in Neural Information Processing Systems*, 36: 21613–21634, 2023.

[51] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision– ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[52] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10925–10934, 2022.

13

[53] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[54] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[55] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023.

[56] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021.

[57] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[58] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[59] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[60] Z Peng, L Dong, H Bao, Q Ye, and F Wei. Beit v2: masked image modeling with vector-quantized visual tokenizers. arxiv 2022. *arXiv preprint arXiv:2208.06366*, 2, 2022.

[61] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[62] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, pages 29441–29454. PMLR, 2023.

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[64] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

[65] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[66] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

[67] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.

[68] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

[69] Thomas Garity, Thomas Fel, George A Alvarez, and Thomas Serre. The role of frequency in shaping features from artificial vision models. *Cognitive Computational Neuroscience*, 2024.

[70] Alexa R Tartaglini, Wai Keen Vong, and Brenden M Lake. A developmentally-inspired examination of shape versus texture bias in machines. *arXiv preprint arXiv:2202.08340*, 2022.

A Appendix

A.1 Shape-vs-Texture Bias: A Useful but Incomplete Metric for Assessing Shape Representations

A widely used benchmark for assessing the degree to which models rely on shape is shape bias, introduced by Geirhos et al. (2019). In this paradigm, each stimulus is a hybrid image that contains the shape defined by one category (e.g., the shape of a cat) with a conflicting texture from a different category (e.g., elephant skin). Humans show strong shape preference in this task, performing near ceiling (95%).In contrast, standard deep net models such as ResNet-50 typically exhibit a strong texture bias, favoring the incongruent texture on 70–80% of trials. While this paradigm reflects the model's relative preference between two competing cues—shape or texture—it is ambiguous if a high score is attained by suppressing textural information or enhancing shape representations. For all shape-vs-texture bias we follow the updated method used in [18] that adjusts for baseline shape accuracy, providing a more principled measure of shape quality using the following equation:

$$\text{Shape-vs-Texture Bias (accuracy-corrected)} = \sqrt{\frac{\#\text{Correct Shape Decisions}}{\#\text{Correct (Shape + Texture)}}} \times \sqrt{\frac{\#\text{Correct Shape Decisions}}{\text{Total Trials}}}$$
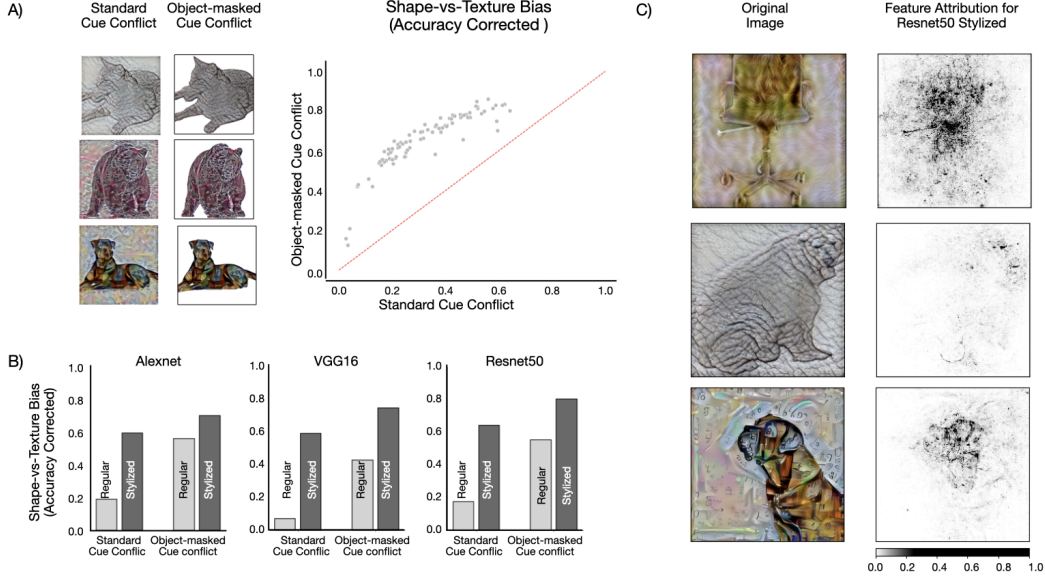

Figure 6: (A) (Left) Cue-conflict stimuli illustrating hybrid images composed of shape from one category and texture from another. Object-masked cue-conflict stimuli has texture removed from the background (Right) Shape-vs-Texture bias (accuracy corrected) across models using standard and object-masked cue conflict stimuli. (B) Shape-vs-Texture bias (accuracy corrected) for original and stylized models. (C) Feature attribution on ResNet50-Stylized reveals that model decisions are still driven by local texture-rich patches, not the full object extent, suggesting that stylization does not completely enhance shape processing.

In Fig. 6A we measure shape-vs-texture bias for a variant of the cue conflict dataset developed [70], in which the conflicting textures are masked out in the background, preserving only the object silhouette. If shape-vs-texture bias truly reflects shape representations, this manipulation should not significantly alter bias scores. However, across a broad range of well-trained deep networks (n = 86), we observed consistent and substantial increases in bias scores under the object-masked condition. To contextualize these findings, we compared the improvements achieved by background masking with those achieved through stylization-based training. As shown in Fig. 6B, stylized models evaluated on the standard cue conflict task showed as much shape-vs-texture bias gains as it would have shown when vanilla (non-stylized) architectures were tested on object-masked cue-conflict stimuli, suggesting that the shape bias measure is confounded not only by texture within

15

the object but also by surrounding local image statistics that lie outside the object's boundary. In Fig. 6C, we applied attribution-based analyses on ResNet50-Stylized to assess which parts of the image were most influential for the model's decision. The results show that model activations often remained highly localized—focusing on small, texture-rich fragments—rather than spanning the full extent of the object, consistent with findings from [19]. In other words, these results suggest that merely suppressing texture either during training (i.e. via stylization) or removing texture footprint in images via silhouette masking, all while keeping the shape information intact, can drive a model's shape-vs-texture bias scores up. Together, these findings suggest that while shape bias remains a valuable comparative diagnostic for assessing model preference between competing shape and texture cues, it should not be interpreted as the only single evidence of good quality shape representation.

A.2 Compute Details

All models were analyzed on an internal computer cluster with 24 cores, 384GB of system RAM, and a NVIDIA H100 GPU with upto 80GB memory. The object Anagram Dataset was generated using a single NVIDIA A100 GPU with 40 GB memory.

A.3 Extracting 1000-way logits and mapping to 9 categories from the Object Anagram Dataset

For self-supervised models like BeITs, MAEs, CLIP, and EVA-CLIP models we used the finetuned linear classifier head provided via the timm library and for DinoV2 we used the full-4 classifier head provided with the model backbone. For SigLIP models we analyzed zero-shot predictions extracted by probing the outputs of the vision encoder with embeddings from text encoder using the category prompts and the given image as inputs.

To map the 1000-way ImageNet logits to our nine target categories in the Object Anagram Dataset, we used the category-to-ImageNet class mapping used in [9] (see below). For each target category, we collected logits corresponding to each category's ImageNet class indices and then took the maximum value from those indices. Once a logit value was computed for each target category, we applied a softmax to get a 9-way probability vector. The predicted label was set to the category with the highest probability.

Category	ImageNet Class Indices
bear	[294, 295, 296, 297]
bunny	[330, 331, 332]
cat	[281, 282, 283, 284, 285]
elephant	[101, 385, 386]
frog	[30, 31, 32]
lizard	[38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48]
tiger	[286, 287, 288, 289, 290, 291, 292, 293]
turtle	[33, 34, 35, 36, 37]
wolf	[269, 270, 271, 272, 273, 274, 275]

Table 2: Mapping between 9 target categories on Object Anagram Dataset and their corresponding ImageNet class indices.

A.4 Human Configural Shape score Estimate

We measured human Configural Shape Score (CSS) using a behavioral experiment implemented in jsPsych. Participants first completed informed consent and viewed instructions explaining the task and rationale. Each trial began with a fixation display followed by an image from the dataset presented centrally for 750 milliseconds. Immediately afterward, a noise mask consisting of randomly generated grayscale pixels appeared for 500 milliseconds to disrupt visual persistence. Following the mask, participants selected the object's category from nine visually presented icons (bear, bunny, cat, elephant, frog, lizard, tiger, turtle, or wolf). Participants completed all 144 images (72 anagram pairs), presented in randomized order, with their category selections and response times recorded. The resulting human CSS served as an approximate baseline for evaluating the configural shape sensitivity of the computational vision models. This study was approved by the IRB of the corresponding author's home institution.

A.5   Cumulative Ablation Analysis on DINOv2 Models

To more precisely characterize the role of intermediate layers for configural processing, we performed a cumulative ablation study on the DINOv2 model family (S/14, B/14, L/14, and G/14) and the ViT-B/16 baseline. We hypothesized that the single-block ablation presented in Section 6 might be too subtle to reveal the full processing dynamics in deeper models. To address this, we designed an ablation experiment with two conditions: (1) ablating all blocks up to block 'n' to isolate the contribution of the network up to that point, and (2) ablating all blocks after block 'n' to test for recovery by later layers. The ablation follows the same strategy of masking the attention in the targeted blocks using a mask window of 2 nearby patches for all the patches, hence blocking long-range interactions.

Fig. 7 shows the effect of these ablations on the final class token representation, measured by its cosine similarity to the token from the original, unablated model. In the "Ablate all before" condition (blue line), the similarity starts high and drops steeply as intermediate blocks are removed, while the "Ablate all after" condition (orange line) shows the opposite trend, creating the characteristic crossover point in the intermediate layers of each model. The results demonstrate that across all DINO models tested, the drop in CSS is highest and most irrecoverable in the intermediate layers. This confirms that even for the large and giant DINO variants, the mid-depth layers are the critical locus for integrating local features into a global, configural shape representation.
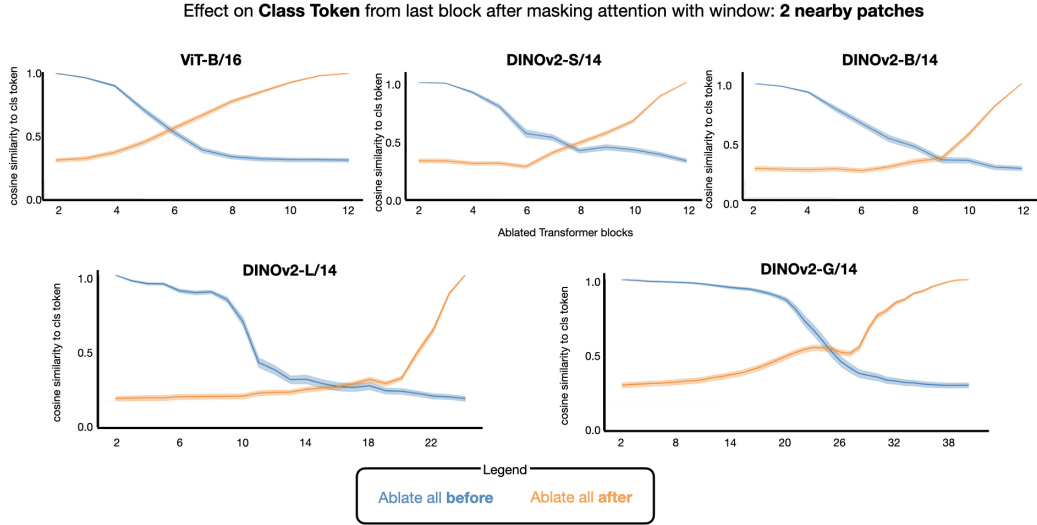

Figure 7: **Effect of cumulative ablation on class token representation.** The y-axis shows the cosine similarity between the final class token of the original model and the ablated model. The blue line (*Ablate all before*) shows the condition when ablating long-range interactions that is 'accumulated' in all blocks before block n. The orange line (*Ablate all after*) shows the condition when ablating long-range interactions that could be 'recovered' in all blocks after block n.

A.6   Other Shape-dependent Evals

Evaluation	Stimuli	# Images	# Categories	Ref.
Robustness to Noise	Imagenette2	98,125	10	Hendrycks & Dietterich, 2019 (and fastai) [11]
Foreground Bias	Imagenet9	4050	9	Xiao et al., 2020 [68]
Shape-vs-Texture Bias	Cue Conflict	1200	16	Geirhos et al., 2018 [8, 18]
Critical Band Masking	Imagenet	1050	16	Subramanian et al., 2023 [14]
Phase-Dependence	Imagenet	50k	1000	Garity et al., 2024 [69]

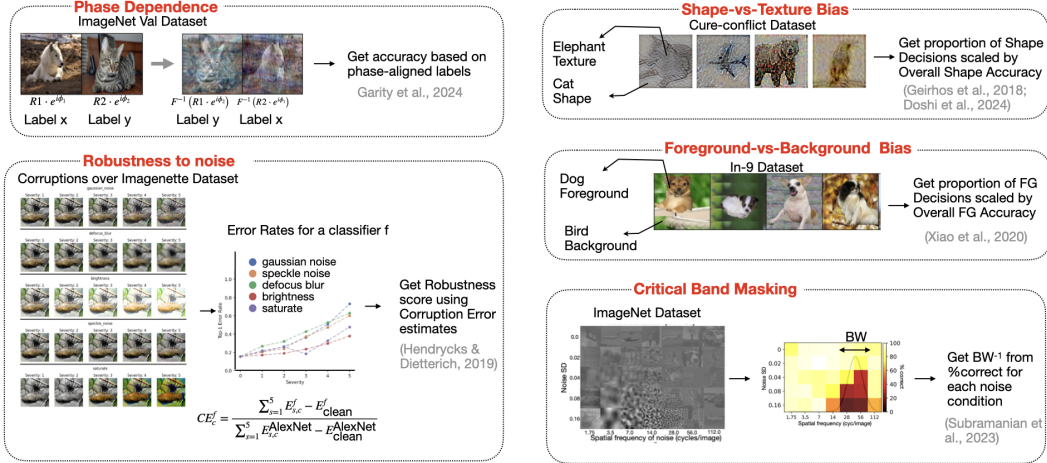Table 3: Overview of evaluation metrics, stimuli, and dataset statistics.

Figure 8: **Schematic of other Shape-dependent Evals.**

A.7 Statistical Comparison of Predictive Strength: CSS vs. Shape-vs-Texture Bias

To evaluate whether Configural Shape Score (CSS) better predicts other shape-dependent evals than the traditional Shape-vs-Texture Bias, we used Williams's test for dependent correlations with one variable in common (i.e., each eval score). The test compares two correlation coefficients (corr(CSS, Eval) and corr(Shape-vs-Texture Bias, Eval)) that share a common outcome variable, accounting for the correlation between the two predictors. We tested this for each of the four evals: Robustness to Noise, Foreground-vs-Background Bias, Phase Dependence, and Critical Band Masking Bandwidth. We used a one-tailed Williams test with n = 86 models, reflecting the directional hypothesis that CSS should better predict eval performance than Shape-vs-Texture Bias. All tests were statistically significant at p < 0.01, indicating that Configural Shape Score is a significantly stronger predictor of these eval benchmarks than Shape-vs-Texture Bias. Below are the results:

Eval Benchmark	r(CSS, Eval)	r(Shape-vs-Texture Bias, Eval)	t-value	p-value
Robustness to Noise	0.81	0.62	3.4116	0.0005
Foreground-vs-Background Bias	0.76	0.32	7.618	<0.0001
Phase Dependence	0.73	0.52	3.39	0.00053
Critical Band Masking	0.83	0.55	5.47	<0.0001

Table 4: Comparison of Configural Shape Score (CSS) and Shape-vs-Texture Bias.

A.8 Feature Attribution for Challenging stimuli
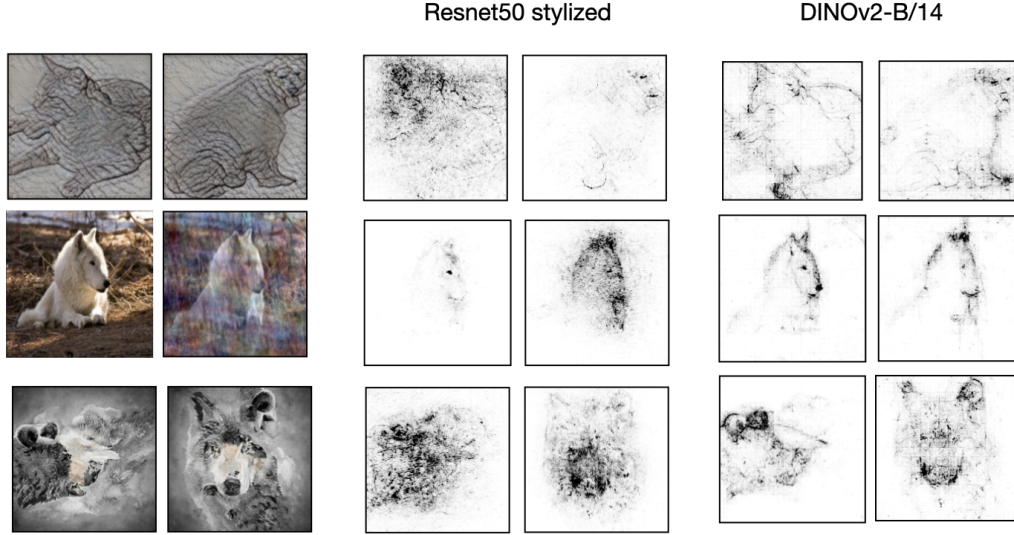
Resnet50 stylized   DINOv2-B/14

Figure 9: **Feature Attribution Maps for challenging stimuli.** Maps generated using Integrated Gradients for cue-conflict stimuli, phase swapped stimuli, and visual anagrams.

A.9 Feature Attributions of Anagrams in High-CSS model (DINOv2-B/14) are not Anagrams themselves
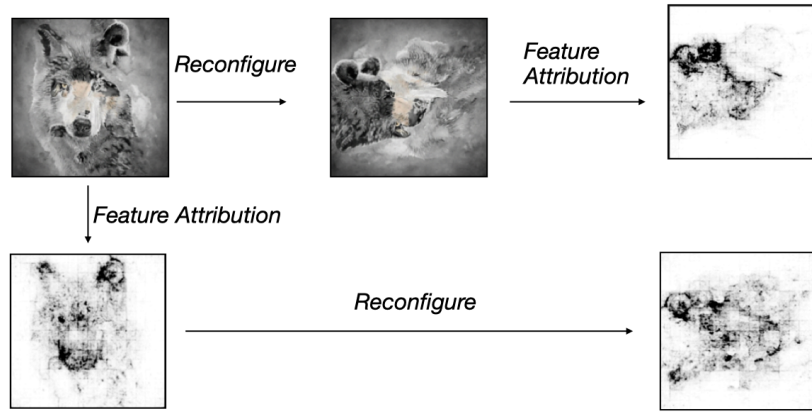

Figure 10: **Applying permutation transformation on feature attribution maps of visual anagrams.**

A.10 Validation on Expanded Anagram Dataset

A potential limitation of our initial study was the dataset size (72 anagram pairs). To address this, we expanded the dataset by 20x to include 1440 visual-anagram pairs and re-evaluated all 86 vision models. The strong correlation (r=0.99) between the scores on the original and expanded datasets shown in Fig. 11 demonstrate that the Configural Shape Score is a highly stable measurement and the conclusions drawn in the main paper are robust.
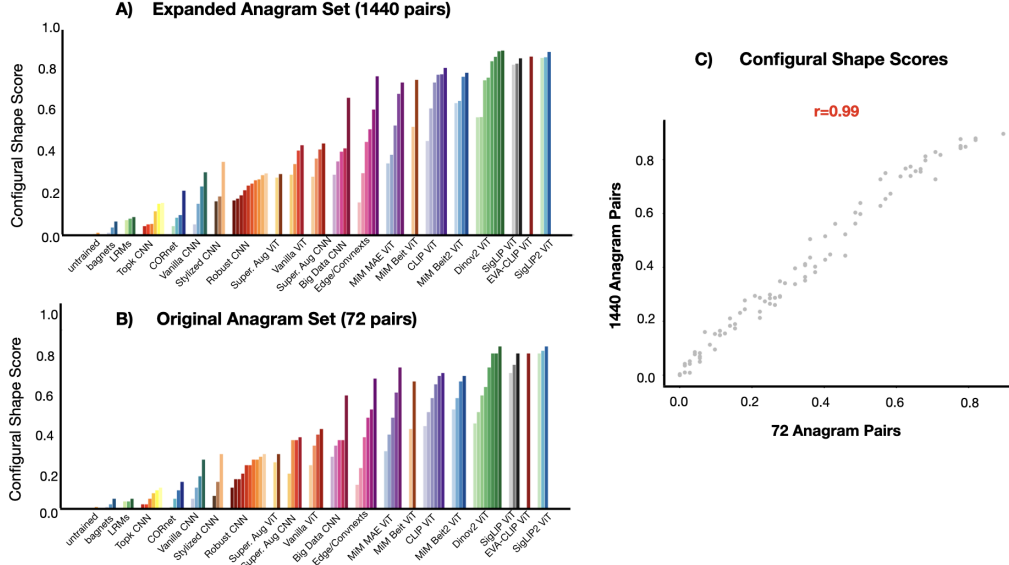

Figure 11: **Validation of Configural Shape Scores (CSS) on an expanded dataset.** (A) CSS scores for all models on the expanded 1440-pair anagram set. (B) The original CSS scores on the 72-pair set. (C) A scatter plot between the scores from the two datasets.

A.11 Evaluated Models

Architecture	Notes	Type
AlexNet	Untrained baseline	Standard Convolutional Networks (TorchVision)
VGG-16	Untrained baseline	Standard Convolutional Networks (TorchVision)
ResNet-50	Untrained baseline	Standard Convolutional Networks (TorchVision)
ResNet-101	Untrained baseline	Standard Convolutional Networks (TorchVision)
AlexNet	Standard AlexNet	Standard Convolutional Networks (TorchVision)
VGG-16	Standard VGG-16	Standard Convolutional Networks (TorchVision)
ResNet-50	Supervised baseline	Standard Convolutional Networks (TorchVision)
ResNet-101	Supervised baseline	Standard Convolutional Networks (TorchVision)
ViT-B/16	Base	Standard supervised Vision Transformers (TorchVision)
ViT-L/16	Large	Standard supervised Vision Transformers (TorchVision)
ViT-B/32	Base	Standard supervised Vision Transformers (TorchVision)
ViT-L/32	Large	Standard supervised Vision Transformers (TorchVision)
AlexNet	Shape-biased AlexNet	Stylized Models
ResNet-50	Shape-biased ResNet-50	Stylized Models
VGG-16	Shape-biased VGG-16	Stylized Models
ResNet-50	L2 $\varepsilon = 0$	Adversarially Robust Models
ResNet-50	L2 $\varepsilon = 0.01$	Adversarially Robust Models
ResNet-50	L2 $\varepsilon = 0.03$	Adversarially Robust Models
ResNet-50	L2 $\varepsilon = 0.05$	Adversarially Robust Models
ResNet-50	L2 $\varepsilon = 0.1$	Adversarially Robust Models
ResNet-50	L2 $\varepsilon = 0.25$	Adversarially Robust Models
ResNet-50	L2 $\varepsilon = 0.5$	Adversarially Robust Models
ResNet-50	L2 $\varepsilon = 1.0$	Adversarially Robust Models
ResNet-50	L2 $\varepsilon = 3.0$	Adversarially Robust Models
ResNet-50	L2 $\varepsilon = 5.0$	Adversarially Robust Models
Alexnet (top-k=80%)	Sparse activation variant	Top-k Sparse

Architecture	Notes	Type
Alexnet (top-k=60%)	Sparse activation variant	Top-k Sparse
Alexnet (top-k=40%)	Sparse activation variant	Top-k Sparse
VGG-16 (top-k=80%)	Sparse activation variant	Top-k Sparse
VGG-16 (top-k=60%)	Sparse activation variant	Top-k Sparse
VGG-16 (top-k=40%)	Sparse activation variant	Top-k Sparse
Resnet50-BagNet-9	Local Receptive field limited to 9 pixels	BagNets
Resnet50-BagNet-17	Local Receptive field limited to 17 pixels	BagNets
Resnet50-BagNet-33	Local Receptive field limited to 33 pixels	BagNets
ResNet-50	Instagram-1B → ImageNet	Data-scale ResNets (SWSL / SSL)
ResNet-50	YFCC100M → ImageNet	Data-scale ResNets (SWSL / SSL)
Resnet-50v2	BiT	Data-scale ResNets (BiT)
Resnet-50v2	BiT-distilled (ResNet-101x1)	Data-scale ResNets (BiT)
Resnet-101v2	BiT	Data-scale ResNets (BiT)
ResNet-50	Supervised with strong augmentation	Strong Augmentation Baselines (Timm)
ResNet-101	Supervised with strong augmentation	Strong Augmentation Baselines (Timm)
ViT-S/16	Supervised with strong augmentation (Small)	Strong Augmentation Baselines (Timm)
ViT-B/16	Supervised with strong augmentation (Base)	Strong Augmentation Baselines (Timm)
ConvNeXt-Base	Modern CNN	ConvNeXt
ConvNeXtV2-Base	FCMAE→ImageNet	ConvNeXt
ConvNeXtV2-Huge	FCMAE→ImageNet	ConvNeXt
EfficientFormer-L1	Efficient transformer, SnapDist	EfficientFormer
EfficientNet-B0	JFT Pretraining	EfficientNet
Edge-AlexNet	Trained on edge-filtered images	Edge-trained Alexnet
CORnet-Z	CORnet Family	Bio-Inspired
CORnet-R	CORnet Family	Bio-Inspired
CORnet-RT	CORnet Family	Bio-Inspired
CORnet-S	SCORnet Family	Bio-Inspired
Alexnet-LRM-Pass1	Long-Range Modulatory CNNs 1	Bio-Inspired
Alexnet-LRM-Pass2	Long-Range Modulatory CNNs 1	Bio-Inspired
Alexnet-LRM-Pass3	Long-Range Modulatory CNNs 1	Bio-Inspired
BEiT-B/16	ImageNet-22K → ImageNet	Self-supervised ViT BEiT
BEiT-L/16	ImageNet-22K → ImageNet	Self-supervised ViT BEiT
BEiTv2-B/16	ImageNet	Self-supervised ViT BEiTv2
BEiTv2-L/16	ImageNet	Self-supervised ViT BEiTv2
BEiTv2-B/16	ImageNet-22K → ImageNet	Self-supervised ViT BEiTv2
BEiTv2-L/16	ImageNet-22K → ImageNet	Self-supervised ViT BEiTv2
Hiera-MAE-T	Hierarchical MAE (tiny)	Self-supervised ViT MAE
Hiera-MAE-S	Hierarchical MAE (small)	Self-supervised ViT MAE
Hiera-MAE-B	Hierarchical MAE (base)	Self-supervised ViT MAE
Hiera-MAE-L	Hierarchical MAE (large)	Self-supervised ViT MAE
Hiera-MAE-H	Hierarchical MAE (huge)	Self-supervised ViT MAE
DINOv2-ViT-S/14	DINO2 (small)	Self-supervised ViT DINOv2
DINOv2-ViT-B/14	DINO2 (base)	Self-supervised ViT DINOv2
DINOv2-ViT-L/14	DINO2 (large)	Self-supervised ViT DINOv2
DINOv2-ViT-G/14	DINO2 (giant)	Self-supervised ViT DINOv2
DINOv2-ViT-S/14	DINO2 (small) + 4 Registers	Self-supervised ViT DINOv2
DINOv2-ViT-B/14	DINO2 (base) + 4 Registers	Self-supervised ViT DINOv2
DINOv2-ViT-L/14	DINO2 (large) + 4 registers	Self-supervised ViT DINOv2
DINOv2-ViT-G/14	DINO2 (giant) + 4 Registers	Self-supervised ViT DINOv2
CLIP ViT-B/16	OpenAI CLIP (base)	Language-aligned ViT CLIP
CLIP ViT-B/32	OpenAI CLIP (base)	Language-aligned ViT CLIP
CLIP ViT-L/14	OpenAI CLIP (large)	Language-aligned ViT CLIP
CLIP ViT-L/14	OpenAI CLIP (large) → ImageNet-12K	Language-aligned ViT CLIP
CLIP ViT-H/14	LAION-2B CLIP (huge)	Language-aligned ViT CLIP
CLIP ViT-H/14	LAION-2B CLIP (huge) → ImageNet-12K	Language-aligned ViT CLIP
SigLIP ViT-B/16	SigLIP (base), Image size 224, zeroshot	Language-aligned ViT SigLIP
SigLIP ViT-B/16	SigLIP (base), Image size 256, zeroshot	Language-aligned ViT SigLIP
SigLIP ViT-L/16	SigLIP (large), Image size 256, zeroshot	Language-aligned ViT SigLIP
SigLIP2 ViT-B/16	SigLIP2 (base), Image size 224, zeroshot	Language-aligned ViT SigLIP2
SigLIP2 ViT-B/16	SigLIP2 (base), Image size 256, zeroshot	Language-aligned ViT SigLIP2
SigLIP2 ViT-L/16	SigLIP2 (large), Image size 256, zeroshot	Language-aligned ViT SigLIP2
EVA-CLIP-G/14-g	EVA02 CLIP (giant)	Language-aligned ViT EVA-CLIP

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification:
   Guidelines:
   - The answer NA 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 NA 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:
   Guidelines:
   - The answer NA 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: [Yes] ,
   Justification:
   Guidelines:
   - The answer NA 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:
   Guidelines:
   - The answer NA 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: [No]
   Justification: We plan to release the complete dataset and codebase, along with detailed reproduction instructions, upon acceptance and in time for the camera-ready submission.
   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification:
   Guidelines:
   - The answer NA 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:

Guidelines:

- The answer NA 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: Details in Appendix

Guidelines:

- The answer NA 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:

Guidelines:

- The answer NA 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: [NA]

Justification: No direct positive/negative societal impacts are expected, so these are not mentioned in the pape

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?
Answer: [NA]
Justification:
Guidelines:
- The answer NA 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:
Guidelines:
- The answer NA 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: [No]
We plan to make entire data and codebase publicly available with the camera-ready version
Guidelines:
- The answer NA 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: [Yes]
Justification: We conducted a small-scale in-lab behavioral experiment to estimate human performance on the Object Anagram task. All participants gave informed consent to participate in the study. Images were shown under controlled timing (750 ms stimulus, 500 ms noise mask), after which participants selected a label from nine options. No personal data was collected, and the study qualifies as exempt under the home institution's IRB guidelines.
Guidelines:
- The answer NA 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: [Yes]
Justification: This study was approved by the IRB of the corresponding author's home institution.
Guidelines:
- The answer NA 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 rigorousness, or originality of the research, declaration is not required.
Answer: [NA]
Justification: Core method development in this research does not involve LLMs

Guidelines:
- The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.