Modeling Rapid Contextual Learning in the Visual Cortex with Fast-Weight Deep Autoencoder Networks

Yue Li[1], **Weifan Wang**[1], **Tai Sing Lee**[1*],

[1]Carnegie Mellon University, Pittsburgh, USA
{yueli4, weifanw, taislee}@andrew.cmu.edu

Abstract

Recent neurophysiological studies have revealed that the early visual cortex can rapidly learn global image context, as evidenced by a sparsification of population responses and a reduction in mean activity when exposed to familiar versus novel image contexts. This phenomenon has been attributed primarily to local recurrent interactions, rather than changes in feedforward or feedback pathways—supported by both empirical findings and circuit-level modeling. Recurrent neural circuits capable of simulating these effects have been shown to reshape the geometry of neural manifolds, enhancing robustness and invariance to irrelevant variations. In this study, we employ a Vision Transformer (ViT)-based autoencoder to investigate, from a functional perspective, how familiarity training can induce sensitivity to global context in the early layers of a deep neural network. We hypothesize that rapid learning operates via fast weights, which encode transient or short-term memory traces, and we explore the use of Low-Rank Adaptation (LoRA) to implement such fast weights within each Transformer layer. Our results show that: (1) The proposed ViT-based autoencoder's self-attention circuit is performing a manifold transform similar to a neural circuit developed for modeling the familiarity effect. (2) Familiarity training induces alignment of latent representation in early layers with the top layer that contains global context information. (3) Familiarity training makes self-attention pay attention to a broader scope details in the remembered image context, rather than just the critical features for object recognition. (4) These effects are significantly amplified by the incorporation of LoRA-based fast weights. Together, these findings suggest that familiarity training can introduce global sensitivity to earlier layers in a hierarchical network, and that a hybrid fast-and-slow weight architecture may provide a viable computational model for studying the functional consequences of rapid global context learning in the brain.

Code — https://github.com/ron7li/familiarity_training

Introduction

Visual perception operates in a world rich with ambiguity, where local features are interpreted in light of global context. The brain resolves this ambiguity by integrating information across spatial scales—leveraging long-range dependencies to disambiguate local signals. While global context is often modeled as top-down feedback—e.g., in analysis-by-synthesis or predictive coding frameworks —it can also reside within local circuits, dynamically modulated by global contextual signals (Rao and Ballard 1999; Lee and Mumford 2003; Gilbert and Li 2013; Angelucci et al. 2017; Coen-Cagli, Kohn, and Schwartz 2015). This division of labor would reduce the informational burden on the feedback pathways, allowing the brain to mediate context with architectural efficiency.

Contextual modulation has long been recognized as a key feature of early visual processing, reflecting the influence of experience (Gilbert and Li 2012) and statistical priors derived from natural scenes (Geisler 2008). Yet recent neurophysiological findings suggest an even more striking phenomenon: global image context can be rapidly learned and expressed in early visual areas such as V1 and V2 (Yan, Zhaoping, and Li 2018; Huang et al. 2018). Specifically, repeated exposure to a small set of natural images—typically 25 familiar images—over the first few days is sufficient to induce robust familiarity suppression in V2 (Huang et al. 2018). Such rapid plasticity raises a fundamental question: how does the system acquire new contextual information so quickly, without suffering catastrophic overwriting of the existing knowledge or priors? One compelling explanation is that the visual system forms local and long-range subnetworks that can be selectively activated by familiar global contexts. When activated, these circuits suppress competing circuits, thus reducing overall activity while sharpening population activities. This architecture would allow multiple contexts to be encoded in parallel, enhancing efficiency and specificity without requiring wholesale modification of feedforward or feedback weights.

Inspired by these findings, we investigate how similar forms of rapid contextual learning might be instilled in artificial neural networks, particularly within early layers, without disrupting previously acquired knowledge. We ask: what kind of architectural and learning mechanisms are required to enable such rapid, context-sensitive plasticity? And what are the functional consequences of this ability—for robustness, invariance, and specificity in visual representation?

To explore these questions, we develop a biologically motivated analysis-by-synthesis framework using autoencoder-based networks trained under self-supervised protocols that

*Corresponding author

mimic visual familiarity exposure. Within this architecture, we introduce a mechanism for temporary memory in the form of fast weights—transient synaptic changes that allow rapid adaptation without compromising the stability of slower, structural memory circuits. We implement these fast weights using low-rank adaptation (LoRA) (Hu et al. 2022), a method originally developed for parameter-efficient fine-tuning, and repurpose and modify it here as a computational analog of rapid, reversible plasticity. By embedding fast weights into the manifold-transform components of the network, we enable it to encode global contextual signals with minimal interference to core statistical priors.

We test the hypothesis that this form of familiarity training can induce global context sensitivity in early-layer representations. Our results show that these early representations become more aligned with the representation of the highest and most global layer, suggesting a functional reorganization of the representational geometry that reflects global sensitivity in the early layers. Furthermore, We found that networks augmented with LoRA exhibit faster and more robust alignment between early-layer representations and higher-level global representations. Moreover, their attention mechanisms display more coherent and consistent figure–ground segregation, reflecting improved contextual awareness. Finally, we demonstrate that the learned manifold transformations resemble those implemented by recurrent circuit models of familiarity in early visual cortex, compressing irrelevant variation while preserving object-discriminative dimensions of the neural manifold.

Together, these findings point toward a principled mechanism by which global contextual memory can be rapidly and efficiently encoded in early visual representations, with implications for both understanding the brain and for the design of adaptive neural architectures.

Background and Related Works

Neuroscience of familiarity learning

Familiarity learning refers to the phenomenon—well documented in inferotemporal cortex (ITC) (Meyer et al. 2014; Fahy, Riches, and Brown 1993; Xiang and Brown 1998; Mruczek and Sheinberg 2007; Sobotka and Ringo 1993; Freedman et al. 2006; Woloszyn and Sheinberg 2012) and, more recently, in early visual cortex (Huang et al. 2018)—where repeated exposure to a set of stimuli produces a suppression of average population responses, a sharpening of single-neuron tuning curves, and a sparsification of population activity relative to responses to novel stimuli (Woloszyn and Sheinberg 2012; Freedman et al. 2006; Tang et al. 2018) (1B and 1C). Huang et al. 2018 further showed that neurons with localized receptive fields become sensitive to the global context of familiar images. The timing of these effects suggests that they are mediated primarily by recurrent circuitry within V2 rather than by feedback from higher visual areas. Together, these observations point to a rapid plasticity mechanism in early visual cortex that modifies local recurrent circuits in each visual area along the visual hierarchy to encode relatively global, at a scale appropriate at each level, context.

Neural circuit modeling and manifold transform

Wang et al. 2025 developed a V1-based neural circuit model, using Hebbian learning and standard V1 circuit motifs, that accounts for these familiarity effects (1A,B,C). The central idea is that repeated exposure to a particular global image context drives the formation of local excitatory subnetworks that encode that context via Hebbian plasticity. Such subnetworks support pattern completion when inputs are occluded or corrupted by noise. By analyzing the recurrent circuit's input–output neural activities, Wang et al. 2025 showed that the circuit implements a manifold transform: it compresses task-irrelevant (within-context) variability while preserving distinctions between different global contexts.

While Wang et al. 2025 revealed this manifold-transform view in a biologically grounded, single-area model, several questions remain. Can such circuits be generalized to hierarchical networks? How does sensitivity to global context evolve across layers, and where do similarities or divergences from biological circuits emerge? Are these phenomena specific to Hebbian-trained, biologically realistic circuits, or do they generalize to modern deep networks trained with backpropagation trained under similar conditions?

To address these questions, we construct a hierarchical autoencoder that uses the image encoder from CLIP (Radford et al. 2021) as a backbone—an architecture shown to align well with the hierarchical organization of human visual cortex. Functionally, the encoder–decoder pair mirrors (to a degree) feedforward and feedback pathways in the brain. This setup affords two advantages: (i) it leverages a powerful state-of-the-art encoder that can be trained or fine-tuned with backpropagation, and (ii) it allows us to introduce LoRA as a form of fast memory. If a deep network trained in this way exhibits analogous familiarity effects and manifold transforms layer-by-layer, this would support the generality of the underlying principles beyond biological circuits.

LoRA and fast weights

Familiarity training in cortex appears to be mediated by rapid plasticity: within a few days of repeated exposure, V2 neurons exhibit familiarity suppression that depends on global context (Huang et al. 2018). Rapidly learning new episodic information, however, risks catastrophic forgetting of existing knowledge. Two influential, complementary solutions have been proposed: the complementary learning systems framework (McClelland, McNaughton, and O'Reilly 1995) and fast vs. slow weights (Hinton and Plaut 1987). Here we will focus our study on the latter.

Fast and slow weights separate learning across timescales. Slow weights encode stable, semantic knowledge (synaptic consolidation), while fast weights support rapid, context-dependent, episodic-like adaptation without overwriting long-term memory. Ba et al. 2016 formalized fast weights in RNNs by introducing a secondary, rapidly updated, auto-associative weight matrix—functionally reminiscent of a Hopfield network layered atop the RNN (Hopfield 1982).

We propose an alternative implementation of fast weights using LoRA (Low-Rank Adaptation). LoRA is a parameter-efficient fine-tuning method in which low-rank, additive

adapters are trained while the base model remains frozen. This modularity allows different LoRA modules to be swapped for different tasks or contexts without retraining the entire model. Because LoRA adapters are low rank and explicitly decoupled from the slow (frozen) parameters, they confer both data and training efficiency when learning new contexts—precisely the desiderata of fast weights.

LoRA is commonly applied to the self-attention circuitry of transformer-based networks (Vaswani et al. 2017). Since transformers capture both local and long-range dependencies—and attention has been argued to be mathematically related to modern Hopfield networks in certain regimes (Ramsauer et al. 2020)—LoRA-equipped transformers offer a scalable substrate for fast-weight-like mechanisms. This motivates our use of LoRA within a ViT-based (Dosovitskiy et al. 2020) autoencoder to investigate the functional consequences of familiarity training in a hierarchical visual system. Unlike typical LoRA applications that freeze the entire pretrained model, we adopt a partial-freezing strategy: only the modules into which LoRA is inserted are frozen, while surrounding layers remain trainable. This design aims to better emulate local circuit-level plasticity in the brain and reflects the coexistence of slow, stable pathways and fast, adaptive components observed in cortical microcircuits.

Approach

This section outlines our approach to investigating the impact of familiarity training on neural representation, particularly in early visual areas of a hierarchical visual system or in the early layers of a neural network. We aim to evaluate two key hypotheses: (1) familiarity training introduces global contextual information into early layers, and (2) it alters the manifold transformation at each level to compress irrelevant variant dimensions while preserving distinctions between different image contexts.

We simulate familiarity training via passive exposure using an autoencoder with a reconstruction loss (1D). The encoder is a pretrained CLIP ViT-B/16 model (Radford et al. 2021), selected for its alignment with the hierarchical organization of the primate visual cortex (Yang, Gee, and Shi 2024). The decoder is a lightweight transformer with 8 layers, 8 attention heads, and a 512-dimensional embedding. It reconstructs the input image from the encoder's highest-level latent representation.

The reconstruction loss combines mean squared error (MSE) and L1 loss:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{L1}} \tag{1}$$

where $\lambda = 0.1$ controls the L1 term. Gradients from this loss are backpropagated through both encoder and decoder, enabling end-to-end training.

The model is trained for 100 epochs with a batch size of 4, a learning rate of 1×10^{-4}, and the AdamW optimizer. We assume that self-supervised learning under this reconstruction objective mimics familiarity learning under passive exposure, consistent with predictive coding and analysis-by-synthesis frameworks.

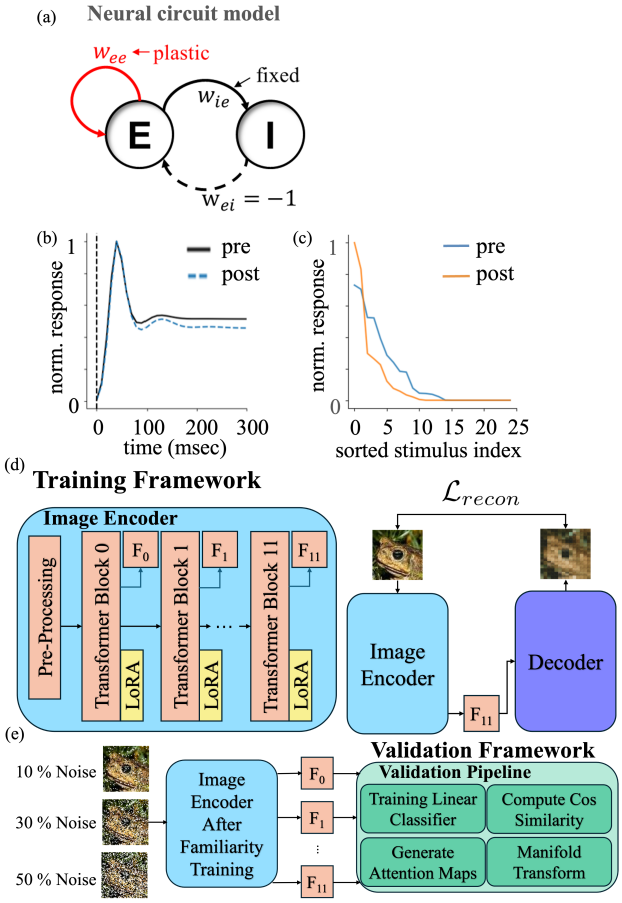


Figure 1: **(a)** Neural Recurrent Circuit Model for Familiarity training (Wang et al. 2025) **(b)** Familiarity Suppression – population-averaged neural responses to familiar images are reduced relative to novel images. (Huang et al. 2018) **(c)** Response sharpening for familiar images, indicating sparsification of neural representations. **(d)** ViT-based autoencoder framework for modeling familiarity learning in deep networks. **(e)** Testing and Evaluation Framework.

Since self-attention enables long-range dependencies within layers, we hypothesize that it can serve as a functional analog of biological recurrent circuits mediating contextual interactions. To test this, we compare the input-output mappings of each transformer block with those of recurrent circuits as described in Wang et al. 2025.

We performed four analytical evaluation on the networks' activities (1E) to assess: 1. Whether familiarity training is performing a manifold transform similar to the neural recurrent circuit. 2. Whether incorporating LoRA enhances the effects of familiarity on manifold geometry and context sensitivity. 3. Whether familiarity training induces global context sensitivity in the different self-attention layers in the ImageEncoder, particularly the early layers. 4. In what way does familiarity trainig change the self-attention computation in the network?

Following Wang et al. 2025, we will test the network with

familiarity training of 4 global image contexts randomly sampled from distinct categories in the Tiny ImageNet-200 dataset (Le and Yang 2015). Each image context can be corrupted by salt-and-pepper noise at 10%, 30%, or 50% levels, yielding 12 image-noise context combinations. The network is trained only on clean images but tested on all image-noise combination contexts, with 10 samples drawn from each context. In each validation run, we evaluate the network on 10 samples per context, using only one specific noise level across all image contexts.

We will compare two network architectures: 1. Without LoRA: the encoder is fine-tuned via full-parameter optimization. 2. With LoRA: only the original $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ in the self-attention blocks are frozen, while the LoRA modules and all other model parameters remain trainable.

In the LoRA-equipped network, the frozen projection matrices serve as slow weights that preserve long-term visual priors. Into these matrices, we insert low-rank ($r = 8$) LoRA adapters that act as fast synaptic adjustments, selectively modulating the attention mechanism during familiarity training. All other model parameters—including those outside the attention projections—remain trainable and are functionally integrated into a distributed fast-weight subsystem, supporting task-specific adaptation beyond the LoRA injection points.

To analyze layer-wise behavior, we extract the `[CLS]` token embedding from each layer as a summary of that layer's representational activity.

Results

Impact of Familiarity Training on Manifold Transformations

To evaluate how familiarity training alters representational geometry, we compare the manifold transformations induced by the recurrent circuit model and the transformer-based autoencoder. Following the methodology of Wang et al. (2025), we compute relative distances to characterize intra-context variation (within an image across noise levels) and inter-context separability (across different images).

Figure 2A illustrates this analysis. Each cone represents the manifold formed by samples of a given image across multiple noise levels. Ellipsoids within the cone correspond to distributions of representations at individual noise levels.

We define two key relative distance metrics:

- *Relative Level Distance* measures the distance between representations of the same image at adjacent noise levels, normalized by the average inter-image distance at the same noise level. - *Relative Residual Distance* measures the spread (variance) of samples within a single image-noise cluster, also normalized by the inter-image distance.

Let $\mathbf{z}_{l,n,k}$ denote the representation of sample k from image l at noise level n. These metrics are formalized as:

$$R_{\text{level}}(k,n,l) = \sqrt{\frac{\sum_{k'} \|\mathbf{z}_{k,n,l} - \mathbf{z}_{k',n-1,l}\|^2}{\sum_{m\neq l,j} \|\mathbf{z}_{k,n,l} - \mathbf{z}_{j,n,m}\|^2}}, \quad (2)$$

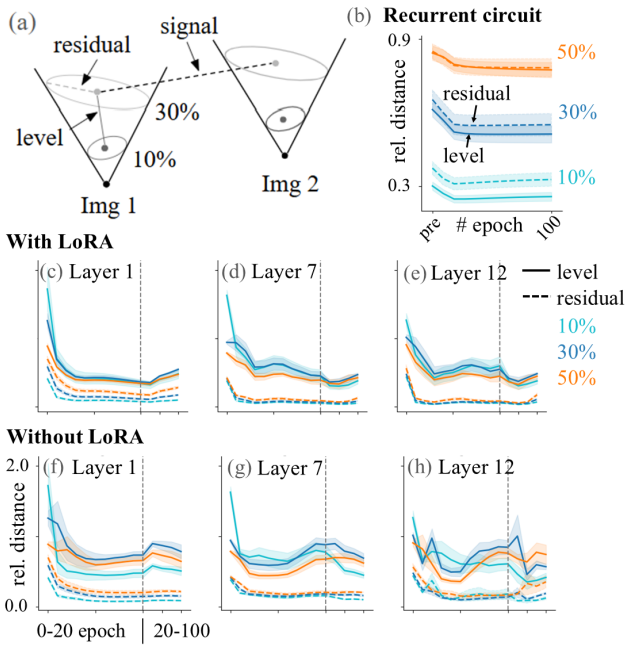


Figure 2: **(a)** Illustration of the manifold geometry for two image contexts and the associated noise-cone. **(b)** Learning curve of relative distances in the recurrent neural circuit model. **(c-e)** ViT encoder with LoRA. **(f-h)** ViT encoder without LoRA. Training reduces both relative level and residual distances, indicating manifold compression.

$$R_{\text{residual}}(k,n,l) = \sqrt{\frac{\sum_{i\neq k} \|\mathbf{z}_{k,n,l} - \mathbf{z}_{i,n,l}\|^2}{\sum_{m\neq l,j} \|\mathbf{z}_{k,n,l} - \mathbf{z}_{j,n,m}\|^2}}. \quad (3)$$

In the recurrent neural circuit model, familiarity training leads to a reduction in both relative level and residual distances, indicating compression of the variant manifold and improved representational efficiency (Figure 2B).

Figures 2C–E and F–H show that similar effects emerge in the transformer-based autoencoder. Across training epochs, we observe consistent decreases in relative distances, particularly in early layers of the encoder. The compression is more pronounced when LoRA is incorporated, suggesting that LoRA-enhanced fast weights amplify the effects of familiarity training.

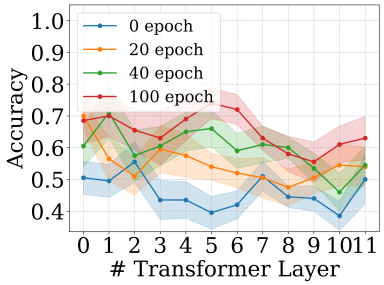
Overall, these findings indicate that familiarity training compresses task-irrelevant variation (e.g., noise-induced variability) while preserving task-relevant distinctions (e.g., image identity). This pattern mirrors the representational geometry observed in trained neural circuits in early visual areas.

A consequence of variant manifold compression is improved discriminability between different image-noise contexts. To assess this, we trained a linear decoder to classify joint image-noise contexts based on representations from each encoder layer.

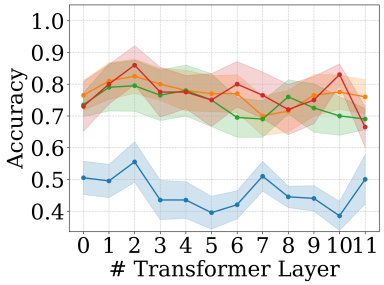
Figure 3 compares discriminability for the LoRA and non-LoRA conditions. The LoRA-enhanced model achieves

consistently higher classification accuracy, supporting the hypothesis that fast-weight mechanisms improve representational separation of the different familiar image contexts.

Interestingly, we find that average discriminability is relatively stable across layers, suggesting that while familiarity training compresses variance within noise cones, the separation between context manifolds is preserved throughout the hierarchy.



(a) Without LoRA



(b) With LoRA

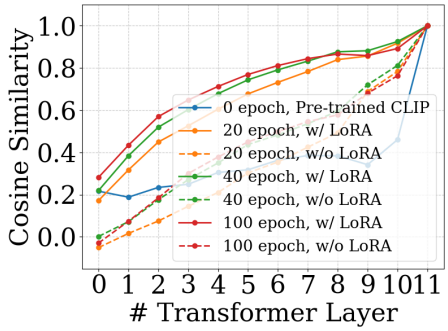
Figure 3: Evolution of linear-decoder classification accuracy across transformer layers during familiarity training. Test images were remembered contexts corrupted with 30 % salt and pepper noise.

Impact of Familiarity Training on Global Context Sensitivity

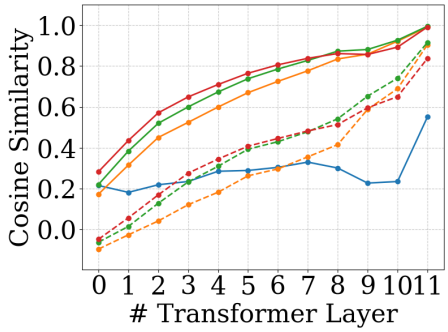
We next examine whether familiarity training induces global context sensitivity in the early layers of the autoencoder network. We operationalize global context as the representation in the top layer (layer 11) of the image encoder, which we assume encodes the most abstract and holistic features of the input. For each sample from an image-noise context, we compute the cosine similarity between its layer-wise latent representations and the corresponding clean image's top-layer representation, hereafter referred to as the "global context representation."

Figure 4 shows how these similarities evolve over the course of familiarity training. As training progresses, we observe an increasing alignment between early-layer representations and the global context representation. This indicates that the early layers of the network progressively acquire global contextual sensitivity.

This effect is consistently stronger and emerges more rapidly in the LoRA-enhanced model compared to the non-



(a) Noise Ratio = 0.0 (Clean)



(b) Noise Ratio = 0.5

Figure 4: Cosine similarity between the class token embedding of every layer and the top (i.e. the 11th) layer of the Image Encoder which reflects global context.

LoRA baseline, across all noise levels. Notably, the alignment is robust to noise perturbations up to 50%, suggesting that LoRA facilitates the formation of globally aligned, noise-invariant features throughout the network hierarchy.

These findings support the hypothesis that familiarity training enhances the global context sensitivity of early layers, and that LoRA accelerates and amplifies this effect via a fast-weight plasticity mechanism.

Interestingly, this increase in global alignment closely tracks the compression of variant manifolds described earlier. By epoch 20, manifold compression (as reflected by reduced level and residual distances) has largely stabilized, coinciding with the plateauing of global alignment, particularly in higher layers. Additionally, the alignment is consistently stronger in the LoRA condition, paralleling the stronger manifold compression observed with LoRA.

These correlations suggest a potential mechanistic link between variant manifold compression and the emergence of global context sensitivity. However, further experimentation is needed to establish causality and directionality between these phenomena.

Impact of Familiarity Training on Self-Attention

To better understand why LoRA enhances familiarity-driven manifold compression, untangling, and global alignment, we investigate the effect of familiarity training on the network's self-attention mechanisms. Attention score maps de-

rived from the [CLS] token embedding serve as a window into how information is aggregated into global representations and which image patches contribute most prominently to this process.

Figures 5a and 5b show layer-wise attention maps from the pretrained CLIP encoder (prior to familiarity training) in response to clean (0% noise) and noisy (30% noise) inputs. In the clean case, early layers focus attention sharply on object regions, consistent with CLIP's training for object-level recognition. In contrast, noise causes attention to diffuse, reflecting a breakdown in localized feature selection.

After familiarity training, this pattern shifts. As shown in Figures 5c and 5d, models trained without LoRA begin to spread attention across more regions of the image, reflecting increased sensitivity to global context. Rather than focusing solely on task-relevant (object) features, the network learns to attend to broader aspects of the image context.

This trend is even more pronounced in LoRA-trained models (Figures 5e and 5f), which exhibit rich, spatially distributed attention patterns. This may stem from the low-rank constraints imposed by LoRA modules, which force the network to efficiently encode contextual information through limited additional capacity.

Quantifying Figure–Ground Sensitivity in Attention. A qualitative inspection of LoRA-trained attention maps suggests improved figure–ground discrimination. To test this, we compute a layer-wise figure–ground Intersection over Union (fg-IoU) between binarized attention maps and reference segmentation masks.

Reference masks are obtained using the Grounding-SAM pipeline (Ren et al. 2024), which leverages Grounding DINO (Liu et al. 2024) for object detection and SAM (Kirillov et al. 2023) for mask generation. Although our autoencoder is trained without labels, it may still exhibit emergent semantic segmentation behaviors.

We define fg-IoU as:

$$\text{fg-IoU} = \max\left(\text{IoU}(\mathbf{A} > t, \mathbf{M}),\ \text{IoU}(\mathbf{A} > t, 1 - \mathbf{M})\right) \tag{4}$$

where $\mathbf{A}$ is the attention map, $\mathbf{M}$ is the binary reference mask, and $t = 0.5$ is the binarization threshold.

As shown in Figure 6a, LoRA-equipped models achieve substantially higher fg-IoU scores across layers—surpassing 0.6 in some cases—whereas non-LoRA models remain near zero. This indicates that LoRA facilitates attention structures that respect figure–ground organization.

Attention Stability Across Noise. Another noteworthy observation is the increased similarity between attention maps for clean and noisy inputs after familiarity training, particularly in the LoRA condition. We quantify this stability using cosine similarity between attention maps across noise levels.

Figure 6b shows that LoRA-trained models yield significantly more consistent attention across layers than both non-LoRA and untrained models. This suggests that LoRA enhances the noise-invariance of attention dynamics.

Interpretation. Taken together, these results suggest that familiarity training encourages the network to encode detailed global context through diffuse attention patterns. This shift supports more robust representations under noise and occlusion, but also moves the network away from sparse, task-driven saliency. LoRA mitigates this risk by confining familiarity-driven plasticity to low-rank subspaces, allowing the network to potentially modulate or suppress context memories based on task demands.

Thus, LoRA not only improves representational efficiency but also offers a mechanism for compartmentalizing memory and attention—retaining the benefits of global familiarity while preserving task flexibility.

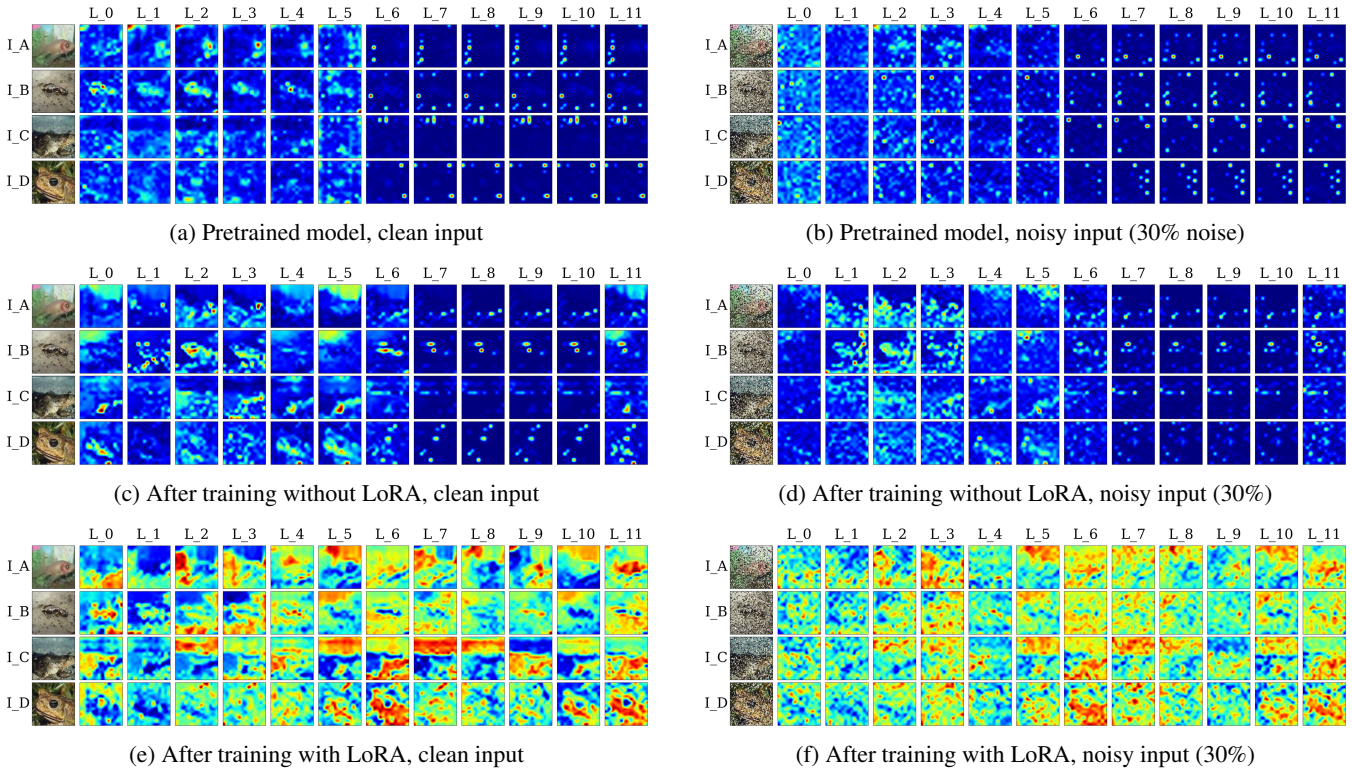
Discussion

In this paper, we propose that LoRA provides an effective framework for modeling fast weights. This approach helps to preserve the model's core generic statistical priors of natural scenes, while allowing adaptation to learning new contexts. By focusing updates on fewer parameters, LoRA more efficiently learns and represents specific familiar contexts.

A key finding of this paper is that familiarity training brings early and intermediate encoder layers into closer alignment with the top-layer representation that encodes global contextual information. Through the autoencoder's reconstruction loss, bottom-up features are fine-tuned to reflect top-down global structure, effectively integrating global context into early-stage representations. This process resembles analysis-by-synthesis or predictive coding: early layers increasingly mirror the global memory stored in deeper layers, supported by self-attention mechanisms that approximate flexible gating and multiplicative interactions.

However, familiarity training alone does not yield coherent figure–ground sensitive attention. Without LoRA, attention remains diffuse and lacks spatial organization. LoRA's low-rank updates serve as an effective regularizer, enforcing spatial coherence and producing emergent grouping across regions, analogous to hypercolumn-like organization in cortex. From a neuroscience perspective, this reflects how recurrent or lateral interactions compress noise dimensions of the manifold and induce correlated activity patterns. The LoRA-enhanced model achieves stronger and more biologically consistent manifold compression than the model without LoRA, which might be a consequence of operations in the low-dimensional manifold.

In this work, we freeze only the specific modules where LoRA adapters are inserted, yielding a micro-circuit with two distinct time scales: (i) the frozen full-rank weights constitute *slow pathways* that preserve long-term visual priors; (ii) the low-rank LoRA adapters act as *fast, transient synaptic adjustments*, rapidly encoding task-specific information. The surrounding layers remain trainable, enabling global representations to reorganize around these locally stabilized modules. This "slow-core/fast-shell" arrangement accords with neurophysiological evidence that cortical plasticity is both localized and circuit-specific (Poort et al. 2015; Gilbert and Li 2012), while still supporting co-existing slow and fast synaptic components (Abbott and Regehr 2004). By combining the stability of slow pathways with the adaptability of fast weights, our partial-freezing strategy attains biologi-



(a) Pretrained model, clean input

(b) Pretrained model, noisy input (30% noise)

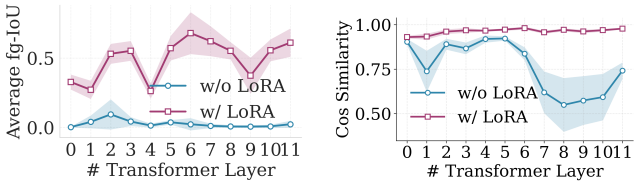
(c) After training without LoRA, clean input

(d) After training without LoRA, noisy input (30%)

(e) After training with LoRA, clean input

(f) After training with LoRA, noisy input (30%)

Figure 5: Layer-wise attention maps before and after familiarity training at 0% (clean) and 30 % noisy levels.



(a) Layer-wise fg-IoU with reference masks

(b) Cosine similarity between clean and noisy attention maps

Figure 6: **a** Quantitative assessment of the alignment of attention map against figure-background segmentation mask. **b** The alignment of different layers' attention score maps between the responses to clean images and that of the noisy images with and without LoRA.

cally inspired plasticity without the over-constraint imposed by fully frozen backbones.

The low-dimensional manifold induced by LoRA imposes a limit on how many global memories can be encoded. We found that increasing the number of stimuli to be represented necessitates a corresponding increase in the dimensionality of LoRA. An alternative strategy is to deploy multiple LoRA modules, each encoding a subset of global memories within its own low-dimensional manifold. Selection among these modules could be achieved through image-dependent routing or mixture mechanisms. The mechanisms for implementing such selective or compositional use of LoRA remain an open question for future research.

The autoencoder does not exhibit tuning-curve sharpening after familiarity training, with or without LoRA. Introducing L1 regularization to encourage sparsity attenuated global context sensitivity, indicating that simple sparsity constraints do not capture the operative principles. This divergence from neurophysiological evidence points to a deeper mechanistic distinction between recurrent neural circuit and ViT-based encoders; the manifold transform in the former is driven by selective amplification and strong normalization, which may not be aligned with the inductive bias of ViT. Clarifying how sparsity, manifold transforms, and context sensitivity interact in these models represents a tractable and important direction for future work.

In this work, the autoencoder augmented with LoRA fast weights serves as a surrogate for the hierarchical visual system equipped with recurrent circuits. The self-supervised learning process, driven by reconstruction loss, functionally approximates the rapid plasticity mechanisms thought to underlie familiarity learning in recurrent cortical connections. Notably, the manifold transformation, specifically the compression of noise dimensions, observed in both neural circuit models and the autoencoder (particularly without LoRA) appears strikingly similar. This suggests that the approximation is plausible at the computational and algorithmic levels, though it diverges at the biological implementation level. This work thus provides insights into the computational processes and functional implications of the familiarity learning phenomena observed in the visual cortex.

Acknowledgements

This work was supported by NSF CISE RI 2420348 and NIH R01 EY030226-01A1.

References

Abbott, L. F.; and Regehr, W. G. 2004. Synaptic computation. *Nature*, 431(7010): 796–803.

Angelucci, A.; Bijanzadeh, M.; Nurminen, L.; Federer, F.; Merlin, S.; and Bressloff, P. C. 2017. Circuits and mechanisms for surround modulation in visual cortex. *Annual review of neuroscience*, 40(1): 425–451.

Ba, J.; Hinton, G. E.; Mnih, V.; Leibo, J. Z.; and Ionescu, C. 2016. Using fast weights to attend to the recent past. *Advances in neural information processing systems*, 29.

Coen-Cagli, R.; Kohn, A.; and Schwartz, O. 2015. Flexible gating of contextual influences in natural vision. *Nature neuroscience*, 18(11): 1648–1655.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fahy, F.; Riches, I.; and Brown, M. 1993. Neuronal activity related to visual recognition memory: long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. *Experimental Brain Research*, 96: 457–472.

Freedman, D. J.; Riesenhuber, M.; Poggio, T.; and Miller, E. K. 2006. Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cerebral Cortex*, 16(11): 1631–1644.

Geisler, W. S. 2008. Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.*, 59(1): 167–192.

Gilbert, C. D.; and Li, W. 2012. Adult visual cortical plasticity. *Neuron*, 75(2): 250–264.

Gilbert, C. D.; and Li, W. 2013. Top-down influences on visual processing. *Nature reviews neuroscience*, 14(5): 350–363.

Hinton, G. E.; and Plaut, D. C. 1987. Using fast weights to deblur old memories. In *Proceedings of the ninth annual conference of the Cognitive Science Society*, 177–186.

Hopfield, J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8): 2554–2558.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, G.; Ramachandran, S.; Lee, T. S.; and Olson, C. R. 2018. Neural correlate of visual familiarity in macaque area V2. *Journal of Neuroscience*, 38(42): 8967–8975.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Lee, T. S.; and Mumford, D. 2003. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7): 1434–1448.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

McClelland, J. L.; McNaughton, B. L.; and O'Reilly, R. C. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3): 419.

Meyer, T.; Walker, C.; Cho, R. Y.; and Olson, C. R. 2014. Image familiarization sharpens response dynamics of neurons in inferotemporal cortex. *Nature neuroscience*, 17(10): 1388–1394.

Mruczek, R. E.; and Sheinberg, D. L. 2007. Context familiarity enhances target processing by inferior temporal cortex neurons. *Journal of Neuroscience*, 27(32): 8533–8545.

Poort, J.; Khan, A. G.; Pachitariu, M.; Nemri, A.; Orsolic, I.; Krupic, J.; Bauza, M.; Sahani, M.; Keller, G. B.; Mrsic-Flogel, T. D.; et al. 2015. Learning enhances sensory and multiple non-sensory representations in primary visual cortex. *Neuron*, 86(6): 1478–1490.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ramsauer, H.; Schäfl, B.; Lehner, J.; Seidl, P.; Widrich, M.; Adler, T.; Gruber, L.; Holzleitner, M.; Pavlović, M.; Sandve, G. K.; et al. 2020. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Rao, R. P.; and Ballard, D. H. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1): 79–87.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Sobotka, S.; and Ringo, J. L. 1993. Investigation of long term recognition and association memory in unit responses from inferotemporal cortex. *Experimental Brain Research*, 96: 28–38.

Tang, S.; Zhang, Y.; Li, Z.; Li, M.; Liu, F.; Jiang, H.; and Lee, T. S. 2018. Large-scale two-photon imaging revealed super-sparse population codes in the V1 superficial layer of awake monkeys. *Elife*, 7: e33370.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W.; Niu, X.; Liang, L.; and Lee, T.-S. 2025. Manifold transform by recurrent cortical circuit enhances robust encoding of familiar stimuli. *PLOS Computational Biology*, 21(10): e1013587.

Woloszyn, L.; and Sheinberg, D. L. 2012. Effects of long-term visual experience on responses of distinct classes of single units in inferior temporal cortex. *Neuron*, 74(1): 193–205.

Xiang, J.-Z.; and Brown, M. 1998. Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology*, 37(4-5): 657–676.

Yan, Y.; Zhaoping, L.; and Li, W. 2018. Bottom-up saliency and top-down learning in the primary visual cortex of monkeys. *Proceedings of the National Academy of Sciences*, 115(41): 10499–10504.

Yang, H.; Gee, J.; and Shi, J. 2024. Brain decodes deep nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23030–23040.