
Flexible Context-Driven Sensory Processing in Dynamical Vision Models

Lakshmi Narasimhan Govindarajan^{*1, 3, 4}, Abhiram Iyer^{*2, 3, 4},
Valmiki Kothare¹, and Ila Fiete^{1, 3, 4}

¹Department of Brain and Cognitive Sciences, MIT, Cambridge, MA

²Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA

³McGovern Institute for Brain Research, MIT, Cambridge, MA

⁴K. Lisa Yang Integrative Computational Neuroscience (ICoN), MIT, Cambridge, MA ,
{lakshmin, abiyer, valmiki, fiete}@mit.edu

Abstract

Visual representations become progressively more abstract along the cortical hierarchy. These abstract representations define notions like objects and shapes, but at the cost of spatial specificity. By contrast, low-level regions represent spatially local but simple input features. How do spatially non-specific representations of abstract concepts in high-level areas flexibly modulate the low-level sensory representations in appropriate ways to guide context-driven and goal-directed behaviors across a range of tasks? We build a biologically motivated and trainable neural network model of dynamics in the visual pathway, incorporating local, lateral, and feedforward synaptic connections, excitatory and inhibitory neurons, and long-range top-down inputs conceptualized as low-rank modulations of the input-driven sensory responses by high-level areas. We study this **Dynamical Cortical network** (*DCnet*) in a visual cue-delay-search task and show that the model uses its own cue representations to adaptively modulate its perceptual responses to solve the task, outperforming state-of-the-art DNN vision and LLM models. The model's population states over time shed light on the nature of contextual modulatory dynamics, generating predictions for experiments. We fine-tune the same model on classic psychophysics attention tasks, and find that the model closely replicates known reaction time results. This work represents a promising new foundation for understanding and making predictions about perturbations to visual processing in the brain.

1 Introduction

We readily use abstract cues to modulate our sensory perception. These include cues to attend to abstract high-level features (find Waldo; count the number of hoop shots, etc.) or low-level features (find the red items, vertically oriented bars, etc.). Such cue-based modulations of the visual pathway allow us to locate items of interest more rapidly and accurately, and to perform goal-directed computations.

Understanding how top-down modulatory cues are represented and then interact with bottom-up sensory-driven neural responses has been a longstanding goal in computational cognitive neuroscience [1–4]. Extensive psychophysical experiments [5] and studies showing that individual neurons whose receptive fields are aligned with the attentional cue exhibit a gain modulation [6, 7] shed light

^{*}Denotes equal contribution.

on the phenomenon. However, a fundamental circuit-level and conceptual problem remains open: what is the nature of the “*modulatory homunculus*” that knows, given various goals and context cues, which low-level representations to modulate, in which combination and which topographic part of the representational space in different processing layers? We lack a cohesive computational framework to link these two levels of representation. Simultaneously, we lack models of sensory processing that fully take into account the recurrent and temporally unfolding nature of computation in the brain; thus, they fall short of explaining phenomena like reaction time variations with task difficulty and the sharpening of perception as information is integrated within a trial.

We combine known architectures of visual cortex with advances in machine learning to introduce a biophysically-inspired model, the **Dynamical Cortical network** (*DCnet*), to solve cue-delay-visual search tasks (Figure. 1). The model is endowed with several relevant properties of biological circuits, including separate (tuned) excitatory and (weakly-tuned) inhibitory populations, lateral inhibition (intra-area recurrence), and neuron types with distinct learnable time constants. We operationalize the “*modulatory homunculus*” as multiplicative low-rank perturbations from higher-order cortical and thalamic areas. Specifically, these low-rank perturbations arise from the model’s own sensory responses from earlier times within the trial.

Contributions. In this work, we focus on analyzing and interpreting the internal dynamics and behavioral modes of our biophysically inspired model on a suite of visual cue-mediated tasks.

- We introduce *vis-count*, a novel, challenging, and parametrically generated cue-delay-visual search task. A visual cue (consisting of a color, a shape, or a color-shape conjunction) specifies which objects in a subsequently presented scene of geometric objects to count.
- We present a biologically realistic model of the brain’s visual system, the **Dynamical Cortical network** (*DCnet*), that is relatively shallow with local and top-down feedback and separate E/I neurons, which is capable of top-down attentional modulation based on previously presented cues and processes information over time. Our framework is among the first to link levels of analyses from physiology to behavior via computations and is a necessary first step toward hypothesis generation for neuroscience.
- Our model outperforms state-of-the-art standard DNNs and LLMs on the task, while being interpretable and having orders of magnitude fewer parameters.
- We perform *in silico* electrophysiology of the circuit’s population dynamics to show that cue-based modulation drives a divergence of the bottom-up responses from the uncued case.
- We can fine-tune the same model on new stimulus sets corresponding to classic human psychophysics attention tasks. Reaction time analogues in our model closely replicate experimental observations.

In sum, these contributions suggest that our approach is a promising framework for modeling the brain’s visual processing dynamics, one that replicates many key attentional phenomena and that can generate testable hypotheses and predictions about circuit mechanisms.

2 Related Work

Stimulus computable models of visual cognition. Modeling top-down contextual effects on bottom-up sensory processing during visual search is of longstanding interest in the vision sciences community. There are a large number of phenomenological models of visual search and associated reaction time findings [8–12], but only recently have models been able to operate directly on high dimensional sensory inputs [13–16]. The most promising approach involves augmenting pre-attentive ventral stream models with controllers either via attention maps [13] or multiplicative modulation factors on network activities [17]. However, these models cannot be used to faithfully study sensory processing dynamics because they are purely feedforward. Recurrent models of early sensory processing with lateral feedback and distinct excitatory and inhibitory populations [18–20], and top-down feedback [21] have shown promise in accounting for contextual visual computations and human reaction times on visual cognitive tasks. However, to our knowledge, stimulus-computable recurrent vision models compatible with *cue-delay-target* paradigms do not exist. More generally, models have tended to either be stimulus-computable or grounded in biological realism, but usually not both.

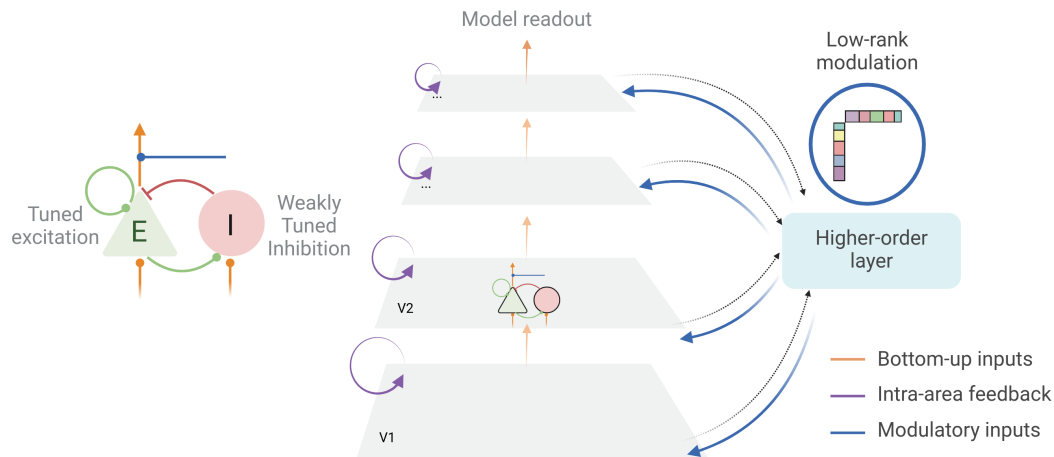


Figure 1: Low-rank modulations to drive context-aware processing. We present a biologically motivated, end-to-end trainable network model of dynamics in the visual pathway. Layers in the model are parameterized by recurrent (tuned) Excitatory (E) and (weakly tuned) Inhibitory (I) neural populations that interact bidirectionally with a higher-order layer in a low-rank manner. The low-rank modulatory factors serve to extract abstract context cues from sensory responses for subsequently driving neural dynamics into context-appropriate dynamical regimes.

Low-rank interactions. Neuronal networks in the brain can be involved in disparate computations simultaneously [22]. Computational neuroscience research has focused on understanding how such neural networks can represent task-specific information and, relatedly, its consequence on the system's overall behavior. Theoretical results suggest that the coexistence of structured low-rank connectivity and random connectivity in networks can enable multi-task computations and expand the dynamic range of the network's functional capability [23–27]. An alternate but non-orthogonal viewpoint is that low-rank control inputs from an external system can switch a network's input-output mappings context-dependently [28–30]. Most of the computational work in this regard is, however, rooted in the motor domain, with Schmid and Neumann being a recent exception.

Neurophysiology of attentional control. Biological attention upmodulates task-relevant information by filtering sensory responses using “templates”. Empirical work shows that attentional templates are primarily represented in higher order cortical areas [31–34] and that such templates can be learned rapidly on a per-task basis [35]. Moreover, the nature of such attentional filtering is known to be cell-type and layer-specific [36], with theories emphasizing the role of top-down mediation of specific GABAergic interneurons [37]. In addition to cortical feedback inputs, higher-order thalamic nuclei are also known to convey contextual inputs to sensory regions either via direct projections [38, 39] or indirectly through the frontal cortices [40, 41]. Building biological sensory processing models that account for all these disparate findings is of primary importance to the Neuroscience community.

3 General methods

3.1 DCnet Model and training details

DCnet comprises two core components: a biologically motivated sensory perception stream and a higher-order area that interacts bidirectionally with it (Figure. 1). Each of the four sensory areas in our model are organized retinotopically as hypercolumns consisting of distinct excitatory and inhibitory neural subpopulations that obey Dale's law [42]. Excitatory pyramidal neurons receive bottom-up and recurrent lateral excitatory inputs as well as short-range lateral inhibitory inputs from interneurons in the same area. Interneurons receive bottom-up and lateral excitatory inputs. Finally, pyramidal neurons project in a feedforward manner to their downstream area. The ratio between excitatory and inhibitory neurons is 4 : 1, as observed in cortical areas [43].

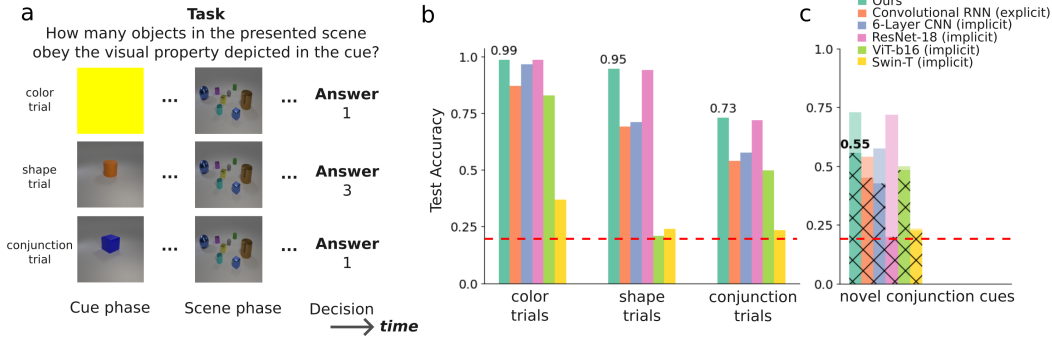


Figure 2: Explicit context-guided modulation of sensory dynamics is necessary for learning generalizable solutions. **a.** We introduce *vis-count*, a parametric, visually-cued, delayed search task. On each trial, models are cued with a visual attribute (either a color, shape, or a conjunction of the two), and after a delay, a scene is presented. The task is to count and report the number of cue-consistent objects in the scene. **b.** On each of the three types of trails, we find that our model consistently outperforms state-of-the-art standard DNNs (Section. 3.2; Baselines) on novel held-out scenes. *Implicit* models refer to the condition where cues and scenes are presented simultaneously. The red dashed line denotes chance performance. **c.** A harder test of generalization on novel scenes *and* cues reveal that our model is robust to such variations, unlike performant implicit models.

The higher-order layer receives pyramidal input from each area and, in turn, modulates inter-area projections in a low-rank manner operationalized as follows.

$$\begin{aligned} \mathbf{r}_1; \mathbf{r}_2 &= \xi(\mathbf{h}_t^l) \mathbf{W}_{l,1}^T + \mathbf{b}_1; \xi(\mathbf{h}_t^l) \mathbf{W}_{l,2}^T + \mathbf{b}_2 && \# \text{ Compute the linear projections} \\ \mathbf{m}_t^l &= \xi(\mathbf{h}_t^l) [\mathbf{r}_1 \otimes \mathbf{r}_2] && \# \text{ Compute the modulating factors} \\ \mathbf{h}_{t+\Delta}^l &= \mathbf{h}_{t+\Delta-\epsilon}^l \odot \mathbf{m}_t^l && \# \text{ Execute the modulation} \end{aligned}$$

Here, $\mathbf{h}_t^l \in \mathbb{R}^{C \times H \times W}$ is the excitatory population readout of area $l \in [1..4]$ at time t ; $\xi(\cdot)$ is the spatial average pooling operator; $\otimes$ denotes outer product and $\odot$ denotes pointwise scaling.

Neurons have learnable cell-type specific time constants. We dispense of traditional ML operations such as BatchNorm or LayerNorm, designed to impart stability during training. Our model has $\sim 1.8\text{M}$ parameters that we learn via gradient descent on a task-performance objective. A full mathematical specification of the model is provided in Appendix. A.1.

3.2 Baselines

We instantiate baseline models of two varieties. First, we consider a “Convolutional RNN” model ($\sim 8.8\text{M}$ params) that uses a traditional 6-Layer convolutional backbone feeding into a gated recurrent unit (GRU) [44] with $N = 2048$ neurons (Appendix. B). We construct this baseline to evaluate the benefits of an explicit modulatory mechanism such as the higher-order layer in our model. Second, we consider four standard deep feedforward neural network architectures. As these models do not operate on spatiotemporal inputs, we condense trials to a single time point by stacking cues and scenes together. Given the lack of an explicit cue followed by scene phase for these models, we term them *implicit*. Cues and scenes are presented at the same time. These baselines provide an upper bound on the expected performance of *DCnet* and verify that our chosen task is a non-trivial computational challenge. In our experiments, we consider the following implicit models: a 6-Layer CNN ($\sim 2.8\text{M}$ params); ResNet-18 [45] ($\sim 11.5\text{M}$ params); ViT-B/16 [46] ($\sim 86\text{M}$ params), a vision transformer with a patch size of 16px; Swin-T [47] ($\sim 28\text{M}$ params), a hierarchical vision transformer. Furthermore, we also ran experiments on the zero-shot generalization abilities of an LLM (Appendix. C).

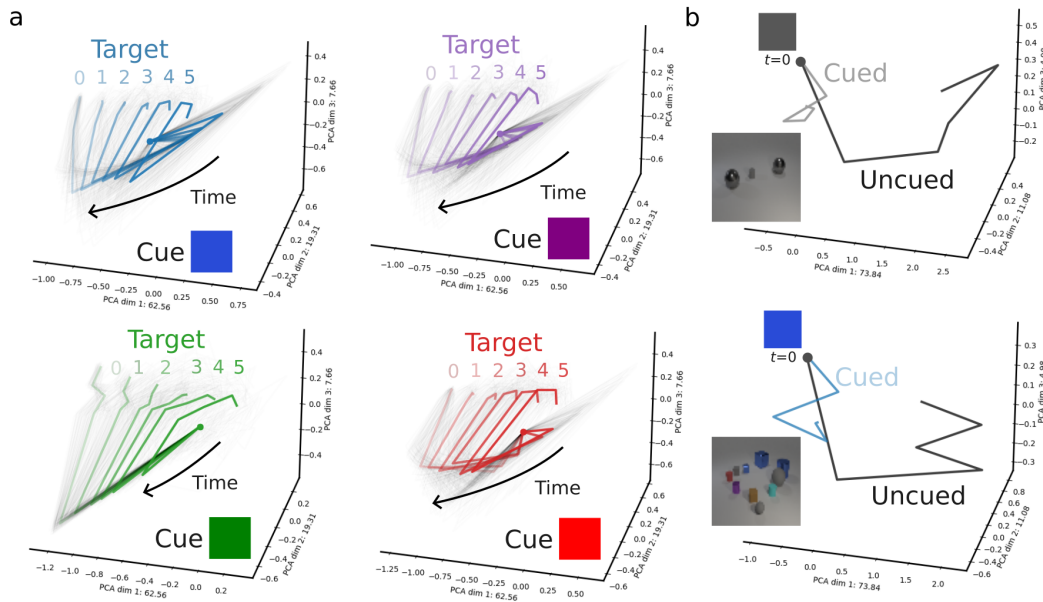


Figure 3: Neural network dynamics reveal context-dependent behavior. We perform dimensionality reduction on late layer model activities and visualize neural trajectories (Gray for individual trials; Dark-colored for trial averages) for different experimental manipulations. The solid dot indicates the trial start. **a.** For a fixed cue (color trials visualized here), network dynamics reflect the extraction and preservation of task-relevant information (object counts) while being invariant to task-irrelevant bottom-up responses from the different scenes. **b.** Matched cued vs. uncued trials for the same scene (inset) reveal a divergence of the bottom-up responses driven by the low-rank modulations.

4 vis-count: A parametric cued visual search task

4.1 Task and stimuli

We draw inspiration from Clevr [48], a synthetic dataset for language-mediated visual reasoning and construct *vis-count*, a parametric visually-cued, delayed search task. On each trial, we challenge models to count and report the number of geometric objects in a visual scene with features consistent with a cue provided earlier (Figure. 2a). Cues can be simple colors, geometric shapes, or a conjunction of the two. Scenes comprised 3 – 10 geometric 3D shapes of varying sizes, colors, and material properties. By construction, there can be 0–5 cue-consistent objects in the scene. We counterbalanced the dataset so that the distribution of target counts was uniform across trial types. Cues and scenes were rendered at 320×240 px and resized to 128×128 px for model training and evaluation. In total, our training (validation) dataset comprised of ~ 384 K (38K) trials. As with Clevr, we detect and discard scenes with fully occluded objects.

For *DCnet* and the Convolutional RNN baseline, cues and scenes are visual inputs provided to Layer 1 of these models. Cues are persistently provided to the network for the first T discrete time steps followed by scenes for the next T time steps. The output activities of the last layer at the last time step ($t = 2T$) are transformed into logits, and a supervision signal is provided via a cross-entropy loss, with ground truth labels counted from 0 to 5. For the “implicit” baseline models, cues and scenes are *stacked* together and presented simultaneously.

4.2 Results

We report the results of our model evaluations and comparisons to baselines in Figure. 2b-c. Our model consistently and significantly outperforms state-of-the-art DNN vision models when evaluated on trials with novel held-out scenes, achieving 99%, 95%, 73% accuracy on the color, shape, and conjunction trials, respectively. Additionally, we perform a harder generalization test by synthesizing trials with novel cues *and* novel scenes (e.g., we test our model on blue and green colors, cube

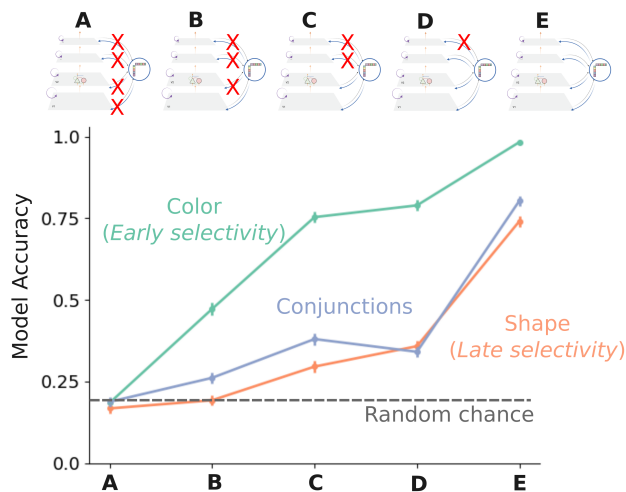


Figure 4: **A selectivity gradient emerges across the cortical layers for the different visual cues.** We sequentially lesion the modulatory synapses in a trained *DCnet* from the higher-order layer to each sensory area, starting from the top-most (D) to the bottom-most (A) sensory area, while documenting task performance on each trial type. (E) indicates the intact *DCnet*. While late lesions have a minimal effect on performance in color trials, the opposite is true for shape trials. Early areas in the *DCnet* exhibit color selectivity, while later areas exhibit shape selectivity.

shapes, and conjunctions thereof, all unseen in the training set). Our model demonstrated the best generalization performance (55%) while even the most performant baseline models (with orders of magnitude more parameters) dropped to chance (16.67%). Sample model results are visualized in Appendix. A.5.

Model errors, too, can be particularly enlightening. Consider the last example in Figure. A.5. When cued to find “blue cylinders”, the model mistakenly appears to include either the cyan cylinder or the occluded blue cube in its count, either of which would reflect an *illusory conjunction* [49]. Illusory conjunctions arise due to failures of feature-based attention and have been extensively studied in the human cognition literature. The presence of such pathologies in our model exposes an opportunity to gain potential insights into the neural implementations of feature-based attention.

In the following sections, we report analyses of our model’s learned dynamics and internal representations that support contextual guidance.

What to modulate? Low-rank structures support optimal cue representations. In daily life, “cues” are not specified to us as abstract rules but rather as high-dimensional sensory inputs. How do we construct a representational space where abstract rules (derived from sensory responses) and the sensory responses themselves are distinct yet coexist? Following insights from electrophysiology [29] and prior theoretical work [23, 24, 28], we hypothesized that learning to inject derived cue information back into sensory representations as low-rank perturbations will aid in optimally partitioning the sensory representational space.

DCnet’s task performance indicated that it had indeed learned task-optimal representations. Here, we perform dimensionality reduction on the activities of pyramidal neurons in the last layer of *DCnet* to probe how this optimality emerges.

First, we observe that cued vs. uncued dynamics on individual trials become progressively divergent and nearly orthogonal with time (Figure. 3b). This indicates that the bidirectional interactions likely promote the formation of low-dimensional activity subspaces embedded within the higher-dimensional ambient activity space. Second, we see that the bottom-up responses to the same set of scenes are differentially modulated to appropriate target subspaces based on the cue (Figure. 3a). We believe that this happens through the inactivation of context-irrelevant subspaces, resulting in invariance to current task-irrelevant features. As additional consideration, we note that training *DCnet* without this top-down feedback mechanism brings performance down to 38%, highlighting the importance of this interaction in our framework.

Where to modulate? A cortical gradient for feature selectivity. After extracting abstract context rules from sensory responses, a question remains: At what level of the representational hierarchy must the perturbation be applied? We perform lesion analysis on the trained *DCnet* to probe this question (Figure. 4).

We sequentially “turn off” modulations, area by area, and observe their detrimental effects on overall function as determined by task performance. We implement this by setting the modulating factors to

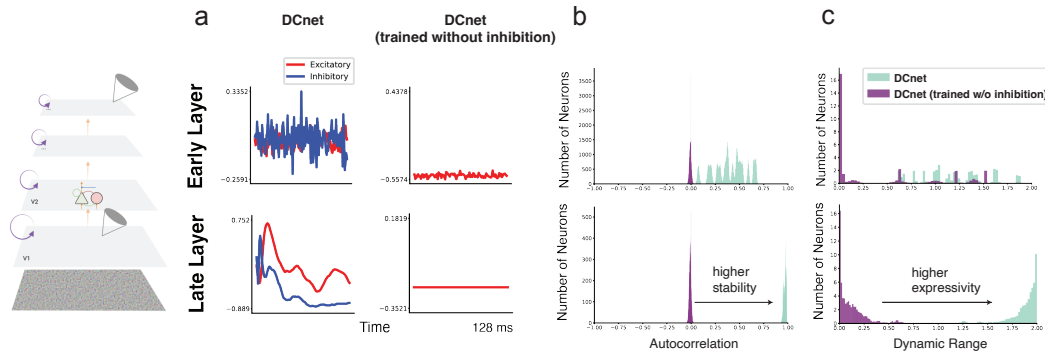


Figure 5: *In silico* electrophysiology sheds light on tuning properties and excitation-inhibition dynamics. We drive network activity in *DCnet* with uncorrelated, time-varying Gaussian noise inputs. (a) We depict a randomly chosen neuron in early and late sensory areas of *DCnet*. Each E/I neuron pair shown here has matched receptive fields. Learned time constants reflect fast/slow integration in early/late areas, revealing a macroscopic gradient. By contrasting *DCnet* trained with and without inhibition, we find that inhibition plays a crucial role in (b) imparting stability and (c) expanding the dynamic range of computations in the network. More quantitative details are presented in Appendix. A.4.

1. We hypothesized that if the modulation of pyramidal neuron activity in a given area is essential for the system's function, then lesioning this perturbation will result in a maximal drop in performance. Interestingly, this analysis yielded differential results per trial type (Figure. 4). For the color trials, we find that lesioning modulations in the early areas had the largest impact on performance, while for the shape trials, it was the late modulations that proved critical. In contrast, modulation strength seemed relatively uniform across the sensory areas.

Taken together, these results offer two insights. First, opposing cortical gradients for color and shape selectivity emerge and coexist in the sensory areas despite the model only being supervised to "count". Second, leveraging this selectivity gradient, the higher-order area learns to apply appropriate low-rank modulations. We believe that the low-rank nature of the context-based modulation, not only within but also across areas, is fundamental to generalization.

Excitation-Inhibition Dynamics. The amplification of context-relevant sensory representations must be balanced to support stable dynamics [50]. The neural underpinnings of this balance and its computational role in feature gain modulation have previously only been studied phenomenologically. Here, we leverage *DCnet* to investigate how the circuit-level properties discovered through optimization can support overall network function.

We probe the cell-type specific neural dynamics in *DCnet* by driving network activity with uncorrelated, time-varying Gaussian noise inputs. First, despite fewer interneurons in the model, inhibitory interactions play a crucial role in imparting stability (Figure. 5b) and expanding the dynamic range of computations expressed by the network when compared to a version of *DCnet* trained without inhibition (Figure. 5c). Second, we detect the presence of co-tuned excitation and inhibition (Appendix. A.4). Empirically, co-tuning is known to be a common organizational principle across the sensory areas. These findings imply the critical role of interneurons in cortical amplification dynamics underlying context-dependent computations. Finally, we observe that neurons learn to integrate information faster in early vs. late areas, revealing the emergence of a macroscopic gradient in neural timescales (Figure. 5a, 8).

5 Model psychophysics on parametric cued feature searches

A feature search is a variant of the general visual search problem in which a "target" is defined by a single discriminative feature [5]. Parametric variations along (or orthogonal to) this discriminative feature axis help shed light on the mechanisms of top-down contextual guidance and its impact on behavior. Here, we consider two feature search tasks from the human psychophysics literature.

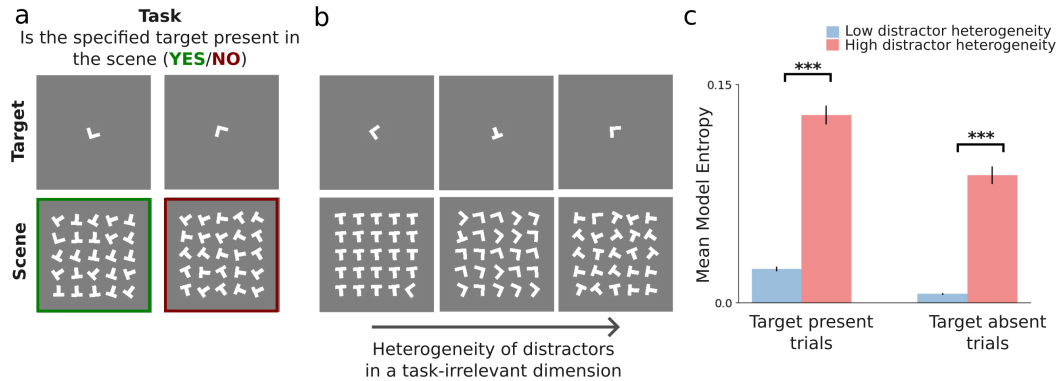


Figure 6: **Slower reaction times/larger output entropy with more heterogeneous distractors.** **a.** Task description. **b.** We parametrically vary the heterogeneity in distractor orientation (similar to [5]). **c.** In both target-present and target-absent trials, we find that the *DCnet*'s output entropy is significantly different between low distractor heterogeneity vs. high distractor heterogeneity trials. Error bars represent the standard error of the mean.

General methods. To perform model psychophysics, we start from the *DCnet* model trained on *vis-count*. We replace its readout to do binary classification (target present/absent) and fine-tune the model on each task below. Furthermore, to derive a measure of model "reaction time," we compute the entropy of *DCnet*'s decision outputs after evolving dynamics for the duration of each trial.

5.1 The role of distractor heterogeneity

Task and stimuli. A target feature's bottom-up salience (pop-out) does not survive variations in irrelevant distractor features. (Figure. 6b) [5]. To test this, we construct cued-feature search trials where models search for an L/T (target) in a grid with a fixed number of Ts/Ls respectively (scene). Targets and scenes were rendered on a 128×128 px canvas. Targets (L/T) were presented at the center and oriented uniformly at random between $[0, 2\pi)$. The distractor heterogeneity level (ω) for every trial was uniform random between $[0, 1]$. Consequently, distractor orientations in the scene were uniform random between $[\theta - \omega \frac{\pi}{2}, \theta + \omega \frac{\pi}{2}]$ with $\theta \in [0, 2\pi)$ being the mean distractor orientation. The target was present in 50% of the trials. Our training (test) dataset comprised of 32K (8K) trials.

Results. When fine tuned on this task, *DCnet* achieved an overall accuracy of 97% on the test trials. Moreover, faithful to the human psychophysics results, we find that the entropy of *DCnet* outputs is significantly different for low ($\omega \leq 0.5$) vs. high ($\omega > 0.5$) distractor heterogeneity for both target-present (Mann–Whitney $U = 302947$, $p < .001$) and target-absent trials (Mann–Whitney $U = 301376$, $p < .001$). While target-absent trials yield higher reaction times in humans when compared to target-present trials, we don't see this in our model's output entropy. This possibly reflects an imperfect choice for a model reaction time measure, an aspect of our work that we look to extend upon in future work.

5.2 The role of target-distractor feature differences in the presence of distractors

Task and stimuli. In a classic critical color difference task, human participants were required to detect a target that differed from distracting stimuli only in color [51]. Search times were measured for varying color differences as a function of display density (the number of distractors). We parametrically synthesize stimuli to recreate this task (Figure. 7a-b).

Target and distractor stimuli were chosen to be circular discs of radii 10px. Cues were rendered at the center of a 128×128 px canvas at a target color chosen at random from among four perceptually uniform color spaces. Search scenes consisted of 1 – 7 distractor stimuli whose color differed from the target color at one of 10 preset target-distractor difference levels chosen uniformly at random. The target was present in 50% of the trials. In total, our training (test) dataset comprised of 32K (8K) trials.

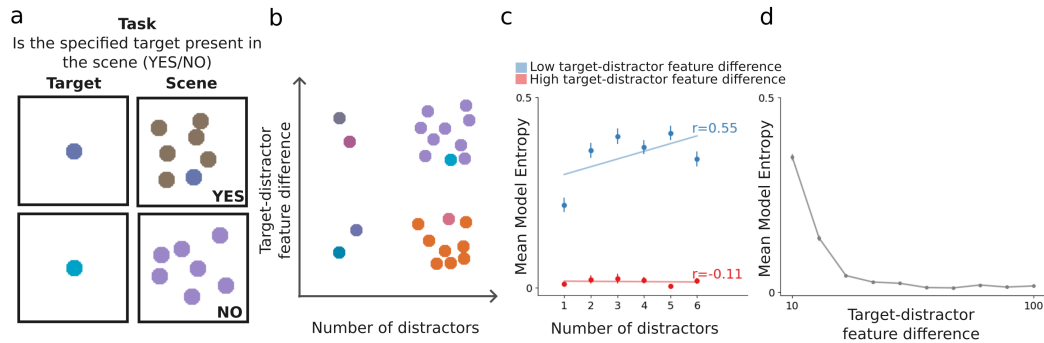


Figure 7: Differential reaction times/output entropy predictions for varying levels of target-distractor differences. **a.** Task description. **b.** We parametrically vary both the number of distracting stimuli in the scene and the target-distractor color difference [5]. **c.** *DCnet*'s output entropy is **invariant** (linear) to the number of distracting stimuli when the feature difference is **high** (low), capturing the psychophysical phenomenon of "popout." **d.** Model entropy as a function of target-distractor feature difference when marginalized over the number of distracting stimuli. Error bars represent the standard error of the mean.

Results. When fine tuned on this task, *DCnet* achieved an overall accuracy of 95% on the test trials. Moreover, *DCnet* recapitulates two key findings from the psychophysics literature.

We probe *DCnet*'s output entropy (a proxy for model RT) as a function of the number of distractors in the scene. We find that for trials in which the target-distractor color difference was high, model entropy was invariant to the number of distracting stimuli (Pearson's $r = -0.11$). On trials where the target-distractor color difference was low, model entropy increased linearly with the number of distracting stimuli in the scene (Pearson's $r = 0.55$). We find a similar result when we probe *DCnet*'s output entropy as a function of the target-distractor feature difference. These results faithfully replicate behavioral effects reported in prior literature [5, 51, 52].

As we alluded to in Section 5.1, our reaction time (RT) metric is one among many possible choices. To take a step further, we implement and test another RT metric inspired by evidential learning (EDL) theory as proposed in [20]. We verify that our primary conclusions from this experiment hold even in the context of this new metric. Finetuned on the EDL objective, *DCnet* achieves 86% accuracy on this task. Model RTs, computed as a time-averaged uncertainty measure, increased linearly with the number of distractors in the low target-distractor color difference trials (Pearson's $r = 0.92$) and was invariant to the number of distractors on high target-distractor color difference trials (Pearson's $r = 0.2$).

6 Discussion

The advent of highly-performant ventral stream models of visual perception is rapidly revolutionizing visual neuroscience research. Stimulus computable models operating directly on high dimensional sensory inputs yield benefits as hypothesis generators for scientific discovery. [53–55]. However, there exist two fundamental axes of dissonance, which are potential rate limiters. First, the emphasis on building "bigger and better" models promotes a divergence from biological realism [56]. Second, it is evident that the extent of biological visual capabilities spans a space bigger than one that entails only feedforward perceptual modules [1, 57, 58]. It calls for accounting for the cooperative dynamics between perceptual and cognitive processes [59].

In this work, we aim to bridge this gap by introducing a computational framework that emphasizes biological realism and conceptualizes interactions between perceptual dynamics and abstract cognitive demands while being scalable and stimulus-computable. We present the **Dynamical Cortical network** (*DCnet*): a trainable neural network model of visual dynamics incorporating local, lateral, and feedforward synaptic connections, excitatory and inhibitory neurons, and long-range top-down inputs conceptualized as low-rank modulations of the input-driven sensory responses by high-level areas.

We start by studying the ability and behavior of *DCnet* to operate in a visually-cued search paradigm. *DCnet* not only outperforms state-of-the-art DNN models, but its population states over time reveal the computational structure and neural underpinnings of contextual modulatory dynamics, generating predictions for experiments. Furthermore, we fine-tune the same model to perform two classic 2AFC attention psychophysics tasks. Reaction time analogs from *DCnet* strikingly recapitulate core tenets of feature-based attentional modulation. We find that *DCnet* responses are sensitive to target-distractor feature differences, heterogeneity of irrelevant distractor features, and display density.

Overall, these contributions suggest that our approach is a promising framework for modeling the brain's visual cortical dynamics, one that replicates key neural and behavioral signatures of contextual attentional modulation.

Limitations and future directions. In this work, we take the first step towards building an overcomplete model of cortical circuitry. Palpable omissions from our current framework include long-range inter-area feedback and compartmental separation of feedforward and feedback inputs. We hope to continue building on this framework in the future to include these components that will open up novel avenues for computational neuroscience. Additionally, extracting reaction time measures from large-scale recurrent models is a discipline of its own [20]. We adopt a simple reaction time metric here that suffices for our current purpose, but we plan to include other sophisticated comparisons in future work. Lastly, we have restricted our purview to visual sensory processing. The concept of contextual modulation, however, is pervasive across sensory modalities. We are excited about extending our ideas to other sensory domains.

Broader impact. Artificially intelligent models with enhanced visual capabilities are now pervasive in our daily lives. The not-so-hidden cost we pay to enjoy these models' benefits is their carbon footprint on our environment. Building energy-, parameter-, and sample-efficient models that are also performant is non-negotiable going forward. Understanding context-aware behavior is of utmost importance for neuroscience research as its failure modes are associated with several psychiatric and neuropathologies. We do not anticipate any negative impact that our work would create.

Acknowledgments and Disclosure of Funding

We thank the McGovern Institute and the K. Lisa Yang Integrative Computational Neuroscience (ICoN) Center for supporting and funding this research. We are grateful to Mark Harnett, Jim DiCarlo, and Bob Desimone for enlightening conversations.

References

- [1] Victor AF Lamme and Pieter R Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 23(11):571–579, 2000.
- [2] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.
- [3] Christopher Summerfield and Floris P De Lange. Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 15(11):745–756, 2014.
- [4] Moshe Bar, Karim S Kassam, Avniel Singh Ghuman, Jasmine Boshyan, Annette M Schmid, Anders M Dale, Matti S Hämäläinen, Ksenija Marinkovic, Daniel L Schacter, Bruce R Rosen, et al. Top-down facilitation of visual recognition. *Proceedings of the national academy of sciences*, 103(2):449–454, 2006.
- [5] Jeremy M Wolfe and Todd S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, 5(6):495–501, 2004.
- [6] Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013.
- [7] Charles D Gilbert and Torsten N Wiesel. Receptive field dynamics in adult primary visual cortex. *Nature*, 356(6365):150–152, 1992.
- [8] Jeremy M Wolfe. Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, 28(4):1060–1092, 2021.

- [9] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3): 419, 1989.
- [10] Seth A Herd and Randall C O'Reilly. Serial visual search from a parallel model. *Vision Research*, 45(24): 2987–2992, 2005.
- [11] Sharat Chikkerur, Thomas Serre, Cheston Tan, and Tomaso Poggio. What and where: A bayesian inference theory of attention. *Vision research*, 50(22):2233–2247, 2010.
- [12] Kyle R Cave. The featuragate model of visual selection. *Psychological research*, 62(2):182–194, 1999.
- [13] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):3730, 2018.
- [14] Shashi Kant Gupta, Mengmi Zhang, Chia-Chien Wu, Jeremy Wolfe, and Gabriel Kreiman. Visual search asymmetry: Deep nets and humans share similar inherent biases. *Advances in neural information processing systems*, 34:6946–6959, 2021.
- [15] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Cocossearch18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776, 2021.
- [16] Zenglin Shi, Ying Sun, and Mengmi Zhang. Training-free object counting with prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 323–331, 2024.
- [17] Grace W Lindsay and Kenneth D Miller. How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, 7:e38105, 2018.
- [18] Drew Linsley, Alekh Karkada Ashok, Lakshmi Narasimhan Govindarajan, Rex Liu, and Thomas Serre. Stable and expressive recurrent vision models. *Advances in Neural Information Processing Systems*, 33: 10456–10467, 2020.
- [19] Vijay Veerabadran, Srinivas Ravishankar, Yuan Tang, Ritik Raina, and Virginia de Sa. Adaptive recurrent vision performs zero-shot computation scaling to unseen difficulty levels. *Advances in Neural Information Processing Systems*, 36, 2024.
- [20] Lore Goetschalckx, Lakshmi Narasimhan Govindarajan, Alekh Karkada Ashok, Aarit Ahuja, David Sheinberg, and Thomas Serre. Computing a human-like reaction time metric from stable recurrent vision models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [21] Junkyung Kim, Drew Linsley, Kalpit Thakkar, and Thomas Serre. Disentangling neural mechanisms for perceptual grouping. *arXiv preprint arXiv:1906.01558*, 2019.
- [22] Wael F Asaad, Gregor Rainer, and Earl K Miller. Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, 21(6):1399–1407, 1998.
- [23] Francesca Mastrogiuseppe and Srdjan Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.
- [24] Friedrich Schuessler, Alexis Dubreuil, Francesca Mastrogiuseppe, Srdjan Ostojic, and Omri Barak. Dynamics of random recurrent networks with correlated low-rank structure. *Physical Review Research*, 2(1): 013111, 2020.
- [25] Stefano Recanatesi, Ulises Pereira-Obilinovic, Masayoshi Murakami, Zachary Mainen, and Luca Mazzucato. Metastable attractors explain the variable timing of stable behavioral action sequences. *Neuron*, 110 (1):139–153, 2022.
- [26] Alfonso Renart, Rubén Moreno-Bote, Xiao-Jing Wang, and Néstor Parga. Mean-driven and fluctuation-driven persistent activity in recurrent networks. *Neural computation*, 19(1):1–46, 2007.
- [27] Lea Duncker, Laura Driscoll, Krishna V Shenoy, Maneesh Sahani, and David Sussillo. Organizing recurrent network dynamics by task-computation to enable continual learning. *Advances in neural information processing systems*, 33:14387–14397, 2020.
- [28] Laureline Logiaco, LF Abbott, and Sean Escola. Thalamic control of cortical dynamics in a model of flexible motor sequencing. *Cell reports*, 35(9), 2021.

- [29] Kaushik J Lakshminarasimhan, Marjorie Xie, Jeremy D Cohen, Britton A Sauerbrei, Adam W Hantman, Ashok Litwin-Kumar, and Sean Escola. Specific connectivity optimizes learning in thalamocortical loops. *Cell Reports*, 43(4), 2024.
- [30] Daniel Schmid and Heiko Neumann. Thalamo-cortical interaction for incremental binding in mental contour-tracing. *bioRxiv*, pages 2023–12, 2023.
- [31] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.
- [32] Narcisse P Bichot, Matthew T Heard, Ellen M DeGennaro, and Robert Desimone. A source for feature-based attention in the prefrontal cortex. *Neuron*, 88(4):832–844, 2015.
- [33] Andrew F Rossi, Narcisse P Bichot, Robert Desimone, and Leslie G Ungerleider. Top-down attentional deficits in macaques with lesions of lateral prefrontal cortex. *Journal of Neuroscience*, 27(42):11306–11314, 2007.
- [34] Yang Zhou and David J Freedman. Posterior parietal cortex plays a causal role in perceptual and categorical decisions. *Science*, 365(6449):180–185, 2019.
- [35] Caroline I Jahn, Nikola T Markov, Britney Morea, Nathaniel D Daw, R Becket Ebitz, and Timothy J Buschman. Learning attentional templates for value-based decision-making. *Cell*, 2024.
- [36] Georgios Spyropoulos, Marius Schneider, Jochem van Kempen, Marc Alwin Gieselmann, Alexander Thiele, and Martin Vinck. Distinct feedforward and feedback pathways for cell-type specific attention effects in macaque v4. *bioRxiv*, pages 2022–11, 2022.
- [37] Siyu Zhang, Min Xu, Tsukasa Kamigaki, Johnny Phong Hoang Do, Wei-Cheng Chang, Sean Jenvay, Kazunari Miyamichi, Liqun Luo, and Yang Dan. Long-range and local circuits for top-down modulation of visual cortex processing. *science*, 345(6197):660–665, 2014.
- [38] Wenzhi Sun, Zhongchao Tan, Brett D Mensh, and Na Ji. Thalamus provides layer 4 of primary visual cortex with orientation-and direction-tuned inputs. *Nature neuroscience*, 19(2):308–315, 2016.
- [39] Morgane M Roth, Johannes C Dahmen, Dylan R Muir, Fabia Imhof, Francisco J Martini, and Sonja B Hofer. Thalamic nuclei convey diverse contextual information to layer 1 of visual cortex. *Nature neuroscience*, 19(2):299–307, 2016.
- [40] Domenico G Guarino, Andrew P Davison, Yves Frégnac, and Ján Antolík. The cortico-thalamic loop attunes competitive lateral interactions across retinotopic and orientation preference maps. *BioRxiv*, pages 2022–12, 2022.
- [41] Anthony D Lien and Massimo Scanziani. Tuned thalamic excitation is amplified by visual cortical circuits. *Nature neuroscience*, 16(9):1315–1323, 2013.
- [42] John C Eccles, P Fatt, and K Koketsu. Cholinergic and inhibitory synapses in a pathway from motor-axon collaterals to motoneurons. *The Journal of physiology*, 126(3):524, 1954.
- [43] D Fitzpatrick, JS Lund, DE Schmechel, and AC Towles. Distribution of gabaergic neurons and axon terminals in the macaque striate cortex. *Journal of Comparative Neurology*, 264(1):73–91, 1987.
- [44] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [46] 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.
- [47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [48] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

- [49] Anne Treisman and Hilary Schmidt. Illusory conjunctions in the perception of objects. *Cognitive psychology*, 14(1):107–141, 1982.
- [50] Jagruti J Pattadkal, Boris V Zemelman, Ila Fiete, and Nicholas J Priebe. Primate neocortex performs balanced sensory amplification. *Neuron*, 112(4):661–675, 2024.
- [51] Allen L Nagy and Robert R Sanchez. Critical color differences determined with a visual search task. *JOSA A*, 7(7):1209–1217, 1990.
- [52] Allen L Nagy and Robert R Sanchez. Chromaticity and luminance as coding dimensions in visual search. *Human Factors*, 34(5):601–614, 1992.
- [53] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.
- [54] Hidenori Tanaka, Aran Nayeibi, Niru Maheswaranathan, Lane McIntosh, Stephen Baccus, and Surya Ganguli. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *Advances in neural information processing systems*, 32, 2019.
- [55] Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B Issa, and James J DiCarlo. Evidence that recurrent circuits are critical to the ventral stream’s execution of core object recognition behavior. *Nature neuroscience*, 22(6):974–983, 2019.
- [56] Drew Linsley, Ivan F Rodriguez Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *Advances in Neural Information Processing Systems*, 36, 2024.
- [57] Gabriel Kreiman and Thomas Serre. Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, 1464(1):222–241, 2020.
- [58] Caspar M Schwiedrzik and Winrich A Freiwald. High-level prediction signals in a low-level area of the macaque face-processing hierarchy. *Neuron*, 96(1):89–97, 2017.
- [59] Petra Vetter and Albert Newen. Varieties of cognitive penetration in visual perception. *Consciousness and cognition*, 27:62–75, 2014.

A Dynamical Cortical network (*DCnet*)

A.1 Architectural specification

We instantiate *DCnet* as a recurrent convolutional neural network model following the primate visual cortex's anatomical and neurophysiological properties. *DCnet* implements local, lateral, and feedforward synaptic connections, excitatory and inhibitory neurons, and long-range top-down inputs conceptualized as low-rank modulations of the input-driven sensory responses by high-level areas.

Specified below are the governing dynamics for the neurons in our model. Neurons in our model are either excitatory (E) or inhibitory (I). l denotes the network layer (brain area). α^l and β^l are indices used to describe excitatory and inhibitory cell types, respectively, in layer l . (x, y) describes a specific spatial location. $z^{(l)}$ is the feedforward input to layer l . h refers to a neuron's instantaneous state.

$$\tau_{E_{\alpha^l}} \frac{dh_{E_{\alpha^l}}^{(l,x,y)}}{dt} = -h_{E_{\alpha^l}}^{(l,x,y)} + g_E(z^{(l)}, h_E^{(l)}, h_I^{(l)}, \alpha^l, x, y)$$

$$g_E(z^{(l)}, h_E^{(l)}, h_I^{(l)}, \alpha^l, x, y) = f([W_{\text{input} \rightarrow E}^{(l)} z^{(l)} + [W_{E \rightarrow E}^{(l)}]_+ h_E^{(l)} + [W_{I \rightarrow E}^{(l)}]_- h_I^{(l)}]_{\alpha^l, x, y} + b_{E_{\alpha^l}})$$

$$\tau_{I_{\beta^l}} \frac{dh_{I_{\beta^l}}^{(l,x,y)}}{dt} = -h_{I_{\beta^l}}^{(l,x,y)} + g_I(z^{(l)}, h_E^{(l)}, h_I^{(l)}, \beta^l, x, y)$$

$$g_I(z^{(l)}, h_E^{(l)}, h_I^{(l)}, \beta^l, x, y) = f([W_{\text{input} \rightarrow I}^{(l)} z^{(l)} + [W_{E \rightarrow I}^{(l)}]_+ h_E^{(l)}]_{\beta^l, x, y} + b_{I_{\beta^l}})$$

$\tau_{E_{\alpha^l}}$ and $\tau_{I_{\beta^l}}$ are cell-type specific learnable neural time constants. Synaptic connections $\mathbf{W}$'s are sparse matrices on which we impose translational invariance (details below). These are, in practice, realized as convolutions. $b_{E_{\alpha^l}}$ and $b_{I_{\beta^l}}$ are excitatory and inhibitory cell-type specific learnable thresholds. $f(\cdot)$ is a non-linear activation function. We use the hyperbolic tangent as our activation function f . An average pooling operation Pool is applied to layer pyramidal outputs (h_E) to increase the receptive field size by a factor of two.

$$z^{(l+1)}[t] = \begin{cases} \text{Pool}(h_E^{(l)}[t]) \odot \Gamma(\xi(h_E^{(l)}[t - T])) & t \geq T \\ \text{Pool}(h_E^{(l)}[t]) & t < T \end{cases}$$

We make discrete time approximations to train our model. The cue is presented first to the network ($z^{(1)}[0]$) and the dynamics are unrolled for T steps followed by scene presentation ($z^{(1)}[T]$) for another T steps. While the scene is presented, inputs to each layer are modulated as follows, where $\xi(\cdot)$ is a pooling operator that computes the average activity per cell type across all (x, y) . $\Gamma(\mathbf{e})$ is the low-rank modulation function defined as follows:

$$\Gamma(\mathbf{e}) = \mathbf{e} \odot \sigma([\mathbf{W}_{l,1} \mathbf{e}^T + \mathbf{b}_1] \otimes [\mathbf{W}_{l,2} \mathbf{e}^T + \mathbf{b}_2])$$

Here, $\mathbf{W}_{l,1}$, $\mathbf{W}_{l,2}$ are learnable linear projections and $\mathbf{b}_1$, $\mathbf{b}_2$ are learnable biases. $\otimes$ denotes outer product and $\odot$ denotes pointwise scaling. By construction, the output of $[\mathbf{W}_{l,1} \mathbf{e}^T + \mathbf{b}_1] \otimes [\mathbf{W}_{l,2} \mathbf{e}^T + \mathbf{b}_2]$ is a low-rank matrix.

A.2 Implementational details

This section provides specific details on model parameters detailed in Section A.1. All layers have the following convolutional kernels ($\mathbf{W}$ s specified in the governing equations): Input to excitation ($W_{\text{input} \rightarrow E}$), excitation to excitation ($W_{E \rightarrow E}$), excitation to inhibition ($W_{E \rightarrow I}$), inhibition to excitation ($W_{I \rightarrow E}$), and input to inhibition ($W_{\text{input} \rightarrow I}$). The dimensionality of these convolutions is listed layer-wise below, along with the kernel and padding shapes for all convolutions in that layer.

	Input size	# exc. cell types (N_E^l)	# inh. cell types (N_I^l)	Kernel size, Padding
Layer 1	$3 \times 128 \times 128$	16	4	(5, 5), (2, 2)
Layer 2	$16 \times 64 \times 64$	32	8	(5, 5), (2, 2)
Layer 3	$33 \times 32 \times 32$	64	16	(5, 5), (2, 2)
Layer 4	$128 \times 16 \times 16$	128	32	(3, 3), (1, 1)

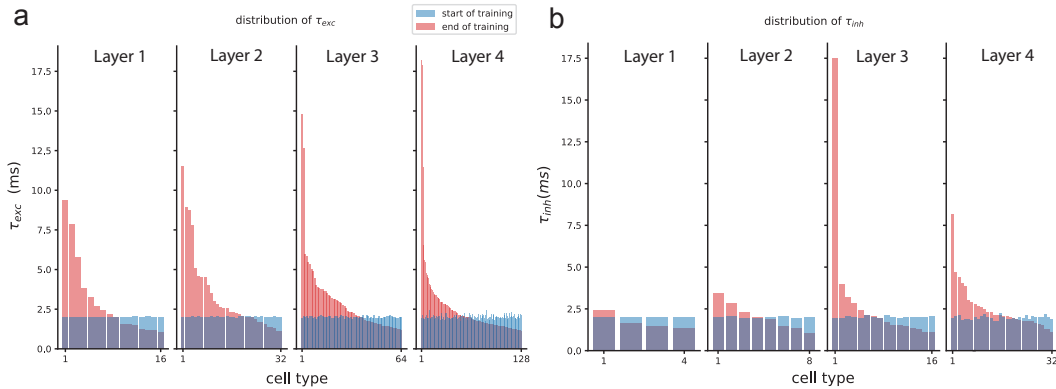


Figure 8: An emergent macroscopic gradient in neuron time constants. Cell-type specific integration time constants in *DCnet* are learnable parameters. While we initialize the τ 's uniformly (blue) pre-training, we observe layer-dependence post-training for both the (a) excitatory, and (b) inhibitory cell types.

The outputs from the last layer are transformed via a linear readout parameterized as a fully connected layer (1024×6) into logits, which we supervise with a CrossEntropy loss. The model does not include standard embellishments like explicit normalization layers, model ensembling, test-time augmentations, etc.

A.3 Training details

All models were trained on A100 GPUs for 100 epochs each. An experimental run took anywhere from 4-8 hours to complete. We used an AdamW optimizer (momentum=0.9, $\beta_1 = 0.9$, $\beta_2 = 0.999$), a one-cycle learning rate scheduler with a warm-up period of 30 epochs and a maximum learning rate of $4e - 4$. *DCnet* was 4 layers deep (~ 1.8 M learnable parameters) and was trained with batches of 256 samples. Code and datasets can be found here: Project repository.

A.4 Mechanistic interpretability analyses

In the interest of mechanistic interpretability, we perform a series of experiments on *DCnet*. We drive activity by uncorrelated, time-varying Gaussian noise inputs. We then compute:

1. The Lag-1 autocorrelation as a measure of **stability** of the excitatory neurons (Fig. 5b). We find that *DCnet* excitatory neurons are significantly more stable than excitatory neurons in the *DCnet* (trained w/o inhibition) model. A Kolmogorov–Smirnov test confirmed the significant difference (statistic=1.0, $p < .001$).
2. The **Dynamic Range (DR)** of excitatory cells in each layer of *DCnet* and compare that to the DR of corresponding excitatory cells from *DCnet* (trained w/o inhibition) (Fig. 5c). DR is computed as the Interquartile range (a measure of statistical dispersion) of a neuron's activity over a time period of 128ms when driven by noise. We take the mean over 64 trials for each neuron and plot this distribution per layer in Fig. 5c. We find that excitatory neurons in the *DCnet* model have a significantly higher DR across layers compared to *DCnet* (trained w/o inhibition), suggesting the role of inhibitory interactions in expanding the range of computations carried out by each neuron. A Kolmogorov–Smirnov test confirmed these significant differences (Layer 1 (statistic=0.667, $p < .001$); Layer 2 (statistic=0.99, $p < .001$); Layer 3 (statistic=1.0, $p < .001$); Layer 4 (statistic=0.97, $p < .001$)).
3. The E-I correlation coefficient as a measure of **co-tuning** in *DCnet*. We find that the average (across neurons) E-I correlation is as follows: -0.076 (Layer 1), 0.766 (Layer 2), 0.699 (Layer 3), 0.535 (Layer 4).

Additionally, we also highlight a macroscopic gradient in learned time constants across the layers of *DCnet* (Fig. 8).

A.5 Sample model outputs

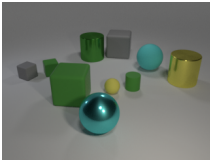
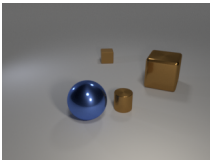
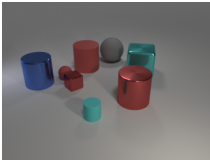
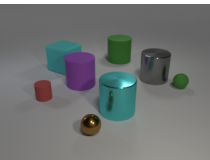
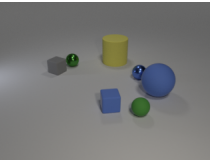
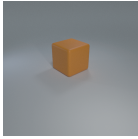
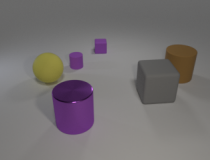
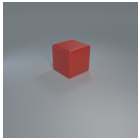
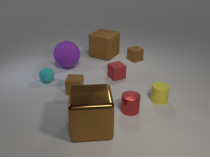
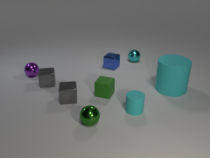
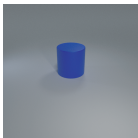
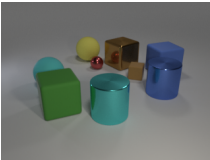
	Cues	Scenes	Answer	Model
Color Trials			4	4
			3	3
			4	3
Shape Trials			5	5
			4	4
			2	1
Conjunction Trials			1	1
			1	1
			1	2

Figure 9: **Model outputs.** Visualizing example *DCnet* predictions on out-of-distribution (held-out) scenes across trial types.

A.6 *DCnet* on visual object recognition

In a follow-up experiment, we train the sensory backbone of *DCnet* on a visual object recognition task. On CIFAR-10, we report a test accuracy of 84.79%. We highlight that our model does not include standard embellishments used in machine learning, including explicit normalization layers, model ensembling, test-time augmentations, etc. This is a first step that strongly demonstrates the potential of our framework to scale up.

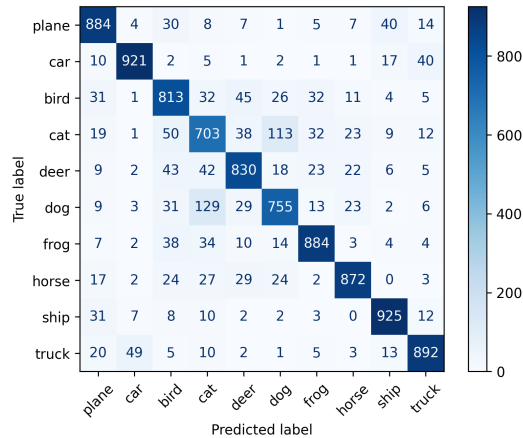


Figure 10: ***DCnet* performance on visual object recognition.** Confusion matrix of *DCnet* outputs on the CIFAR-10 test set. Overall model performance is 84.79%.

B Convolutional RNN Baseline

Layer	Input Size	Operations
Layer 1	$3 \times 128 \times 128$	Conv2d ($8 \times 3 \times 5 \times 5$), ReLU, AvgPool2d (5×5 , stride 2)
Layer 2	$8 \times 64 \times 64$	Conv2d ($16 \times 8 \times 5 \times 5$), ReLU, AvgPool2d (5×5 , stride 2)
Layer 3	$16 \times 32 \times 32$	Conv2d ($32 \times 16 \times 5 \times 5$), ReLU, AvgPool2d (3×3 , stride 2)
Layer 4	$32 \times 16 \times 16$	Conv2d ($64 \times 32 \times 3 \times 3$), ReLU, AvgPool2d (3×3 , stride 2)
Layer 5	$64 \times 8 \times 8$	Conv2d ($128 \times 64 \times 3 \times 3$), ReLU, AvgPool2d (2×2 , stride 2)
Layer 6	$128 \times 4 \times 4$	Conv2d ($128 \times 128 \times 3 \times 3$), ReLU, AvgPool2d (1×1 , stride 1)
Projection		Fully connected (2048 inputs, 2048 outputs), ReLU, GRU (2048 inputs, 2048 outputs)
Readout		Fully connected (2048 inputs, 6 outputs)

Table 1: Baseline model details

Convolution filters are mentioned as: (output channels $\times$ input channels $\times$ kernel height $\times$ kernel width). Pooling layers are mentioned as: (kernel height $\times$ kernel width, stride). Inputs to every layer were normalized with LayerNorm. The convolutional backbone serves as a feature extractor for the GRU network. We train this baseline model by passing in the cue for the first T time steps followed by the scene for the next T timesteps. For our experiments, we set $T = 3$. A cross-entropy loss is used for the readout activities at the final time step.

C Zero-shot testing of LLM

Prompt Here is a task that takes two images, a cue and a scene. Here are the instructions:
 1. If the cue is a color, return a count of how many objects in the scene have that color.
 2. If the cue is an orange colored shape, return a count of all objects in the scene that have that shape.
 3. If the cue is a non-orange colored shape, return a count of all objects in the scene that have that shape and color.
 You will only need to follow 1 of the 3 instructions above. Don't write code.
 There is no need to explain the logic of your counting.
 Simply provide the final answer.

GPT4 Sure, I can help with that. Please upload the two images you have: the cue image and the scene image, and I'll proceed according to your instructions.

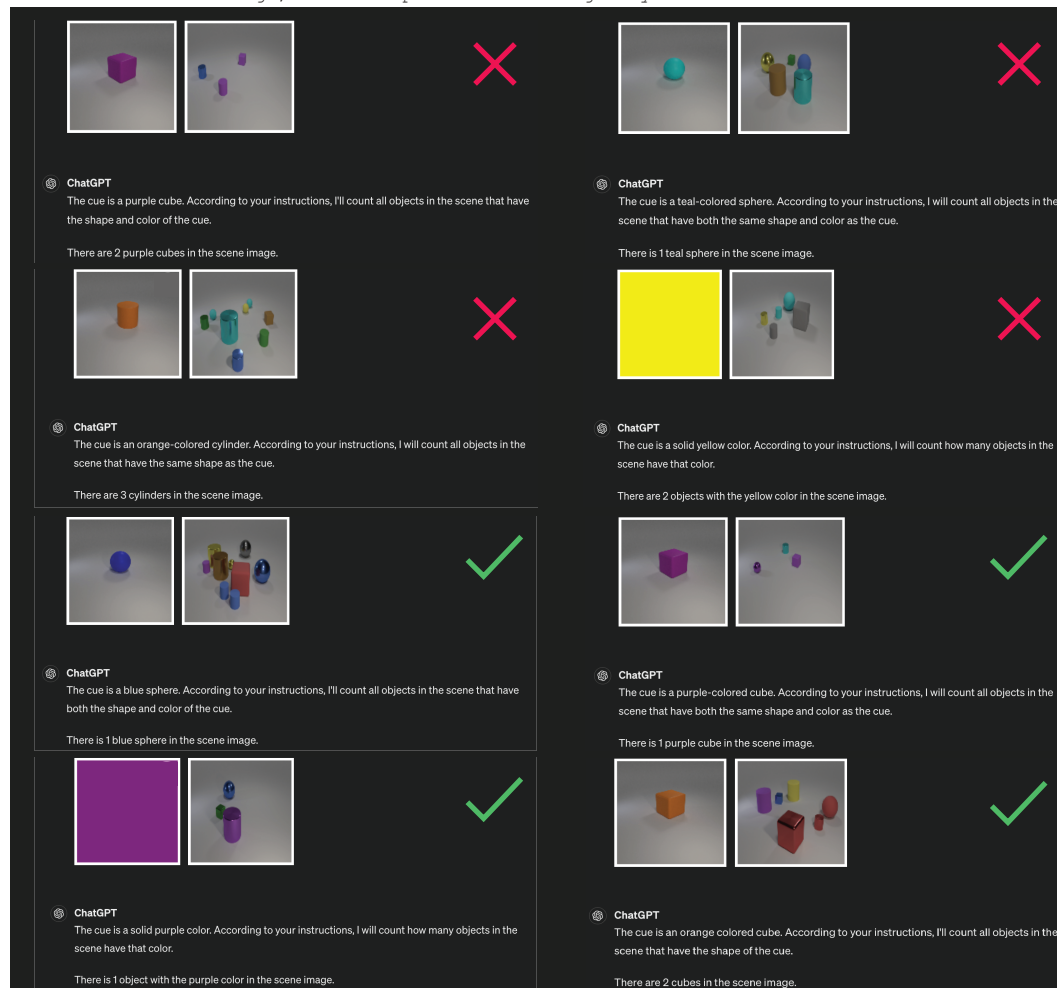


Figure 11: **ChatGPT4 experiments** We prompted GPT4 to solve solve *vis-count*. Given a color, shape, or conjunction cue, we asked the LLM to identify how many objects in the provided scene match the cue's properties. Out of 30 trials, GPT4 achieved a success rate of 37%.

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.

1. Claims

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

Answer: [Yes]

Justification: Yes, all claims made in the abstract and introduction reflect the paper's contributions. We make it clear the strengths of our work and also the limitations.

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: Yes, the limitations of our work is discussed in the Discussion section of our paper.

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: [NA]

Justification: This paper does not contain any purely mathematical or theoretical results. There are no mathematical proofs.

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: Yes, the details about our experimental setup are provided in the Supplementary Section.

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: [\[Yes\]](#)

Justification: Yes, we provide a link to our code that is anonymized. We ensure that the code can be used to reproduce our main experimental results.

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: Hyperparameters are provided in our Supplementary Section. Further details regarding our experiments can also be found in our code, which we fully disclose in our Supplementary Section as well.

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: We report all statistically significant measures for our experiments. These can be seen in our Results section.

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: Discussion about the resources used to train our model are discussed in the Supplementary Section.

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: We have verified that our work follows the NeurIPS Code of Ethics fully and completely.

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: Our work does not have a clear societal impact and thus we do not discuss this in our paper. This work is with regards to building biologically plausible vision models that can serve as hypothesis generators for real world experiments. These models will not be used by the average everyday user or affect everyday society in any regard.

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: Our model poses no such risks. We do not train large models that have collected user data, nor do we present such models in our work.

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: Baselines implemented by other research papers have been fully credited. All code is properly documented and cited, as well.

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: [\[Yes\]](#)

Justification: We provide the code for exactly reproducing the datasets that we generate. This reproduction process is well documented and is shared within the code that we share.

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: [NA]

Justification: Our paper does not involve crowdsourcing research or collecting humans in any way.

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: [NA]

Justification: This research does not involve crowdsourcing or conducting experiments/research with human subjects.

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.