# Incorporating foveal sampling and integration to model 3D shape inferences

**Stephanie Fu**
UC Berkeley

**Tyler Bonnen**
UC Berkeley

**Trevor Darrell**
UC Berkeley

## ABSTRACT

**Human vision is inherently sequential. This is largely because of foveal constraints on the retina, which demand that we shift our gaze to collect high-resolution information from throughout the environment. Within the domain of visual object perception, the neural substrates that support these sequential visual inferences have been well characterized: ventral temporal cortex (VTC) rapidly extracts visual features at each spatial location, while medial temporal cortex (MTC) integrates over the sequential outputs of VTC. This neurocomputational motif is absent in contemporary deep learning models of human vision. Not surprisingly, contemporary models approximate the rapid visual inferences that depend on VTC, but not those behaviors that depend on MTC (e.g., novel 3D shape inference). Here we develop a modeling framework that embodies the sequential sampling/integration strategy emblematic of human vision. Given an image, this model first determines relevant locations to attend to, sequentially processes these locations as 'foveated' inputs using a VTC-like model, then integrates over these sequential visual features within a MTC-like model. Here we report preliminary results on the design choices that lead to stable model optimization and subsequent model behaviors.**

**Keywords:** ventral temporal cortex; medial temporal cortex; object perception; 3D vision; temporal dynamics; deep learning

## INTRODUCTION

There is temporal structure in how we perceive the world. Many visual attributes can be inferred 'at a glance' (Potter, 1975), an ability which depends on ventral temporal cortex (VTC) (DiCarlo, Zoccolan, & Rust, 2012). But not all visual inferences are possible 'at a glance' (Findlay & Gilchrist, 2003). This is due, in part, to constraints on the primate visual system (Van Essen & Anderson, 1990): high-acuity visual information is only maintained at the central visual field, i.e., the fovea (Hirsch & Curcio, 1989), and so to collect precise visual information we must shift the location of our gaze, roughly three times per second (Liversedge & Findlay, 2000). To support object-level visual inferences not possible 'at a glance', there are downstream neural structures that integrate over visuospatial sequences: lesion studies indicate that medial temporal cortex (MTC) supports perceptual inferences not supported by VTC alone (Murray & Bussey, 1999; Bussey, Saksida, & Murray, 2002; Barense, Gaffan, & Graham, 2007; Murray & Wise, 2012), and these MTC-dependent visual inferences rely on temporally-extended, sequential, visual processes (Bonnen, Wagner, & Yamins, 2023). Prevailing computational approaches to modeling human vision have centered on approximating VTC function (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014), but fail to approximate the sequential visual inferences that depend on MTC (Bonnen, Yamins, & Wagner, 2021). Here we develop a novel deep learning model that reflects the neurocomputational motif evident in MTC function. We train this model end-to-end on a large-scale dataset of 3D assets, using a biologically plausible training objective, reminiscent of 'mental rotation' from the cognitive science literature (Shepard & Metzler, 1971). We then evaluate on a 3D shape inference benchmark (Bonnen et al., 2024).

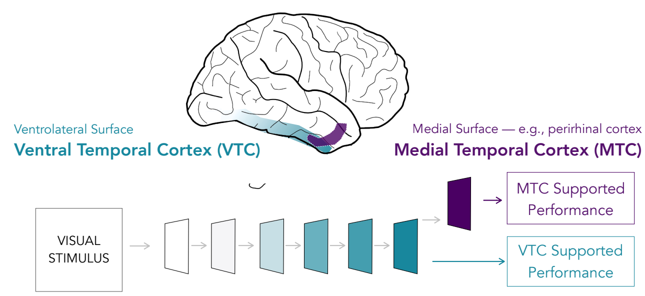


Figure 1: **Schematic of ventral and medial temporal cortex**. Ventral temporal cortex is a hierarchical network on the lateral and ventral surface of the temporal lobe (orange) shown to support rapid visual inferences through feedforward processing. Medial temporal cortex lies on the medial surface of the temporal lobe. Here, we visualize perirhinal cortex (green), an MTC structure at the apex of VTC.

## METHOD

Our architecture is comprised of three main components: the gazing module, which controls the locations to attend to; the feedforward vision encoder, which processes the attended locations using 'foveated' inputs; and an integration module, which processes the sequences of encoded features. We illustrate the architecture in Fig 2 and describe them below.

**Sequential sampling.** We determine locations of interest using a ConvNeXt block (Liu et al., 2022), which outputs a salience map that we interpret as a probability distribution over image patches (i.e., discretized locations). Using Gumbel top-k sampling so that the operation is differentiable, we sample the top 10% of patches and pass them to the VTC-like model.

**VTC-like encoding.** We use DINOv2 (a large vision transfomer (Oquab et al., 2024)) as a computational proxy for VTC. Given the locations selected by the gaze module, we use 3x3 patch receptive fields (42x42 px) centered on these locations; this provides a coarse operationalization for foveated visual inputs. Critically, we modify the attention masking in each layer to only allow for self-attention within these 3x3 blocks. This localized attention mask mimics this per-fixation processing by limiting integration to nearby regions. Moreover, this strategy disregards between-fixation VTC dynamics, consistent with recent findings (Xiao, Sharma, Kreiman, & Livingstone, 2024).

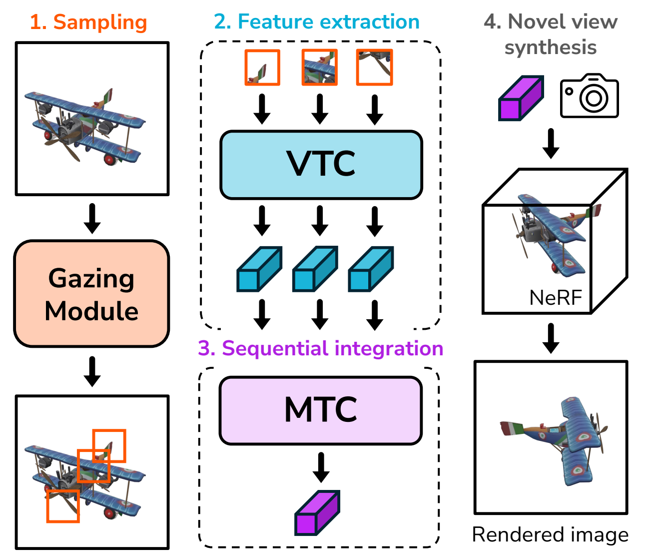


Figure 2: **Left:** The gazing module determines the locations of interest within the input image. **Center:** the VTC-like model encodes a region around the attended location (i.e., pseudo-foveated input), then passes this sequence of features to an MTC-like model to integrate. **Right:** we train our model using a biologically plausible objective, akin to 'mental rotation': novel view synthesis. By decoding the MTC output into a 3D representation, we can render novel views and optimize a reconstruction loss between rendered and ground truth views.

**MTC-like integration.** We sequentially pass these embeddings generated by the VTC-like module to a causal transformer, which only allows patch $n$ to attend to previous patches $\{0, 1, ..., n-1\}$. This attentional setting resembles the integration dynamics of a recurrent system (i.e., only attending to previous states and current inputs) while maintaining the desirable properties of transformers. We accumulate the information output the transformer with a single global embedding, which can attend to all patch embeddings. We use this global embedding as the final VTC+MTC encoding of the input image.

**Training details.** We train our model using a biologically plausible training objective: given an image, we predict what the object would look like from a different viewpoint. This novel view synthesis (NVS) approach is akin to 'mental rotation' (Shepard & Metzler, 1971) and leverages information corresponding to visual (pixels) and proprioceptive (camera pose) signals. While this objective does not fully capture how humans learn, it provides a training signal better suited to learn the 3D properties of objects than more common objectives (e.g., classification).

Our training setup follows the general approach of LRM (Hong et al., 2024) (Fig. 2, right). Specifically, we use the MTC-like output and camera pose to predict a triplane representation of the object. We then decode a NeRF predicting point color and density. For each object, we then render four images from varying camera poses and optimize a reconstruction loss (L1 + LPIPS (Zhang, Isola, Efros, Shechtman, & Wang, 2018)). We train on 450,000 objects from Objaverse, sampling four novel viewpoints per object. To encourage exploration and avoid early collapse, we anneal temperature in the top-k sampling and add entropy regularization over gazing probabilities.

## RESULTS

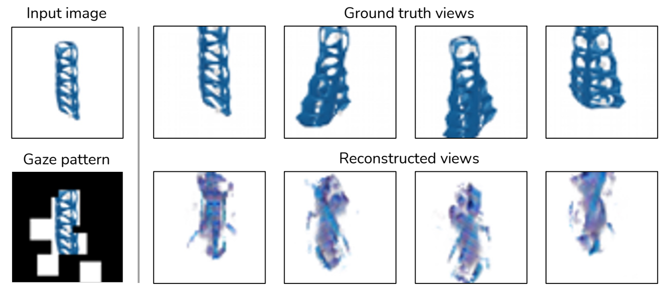


Figure 3: Reconstructed views of a held-out object. In the bottom left, we visualize the 3x3 patches that the VTC-like model encodes.)

First, we visualize renderings of an unseen object (Fig 3). While the results are not pixel-perfect, they show that our model captures meaningful structural and color information. We anticipate that a larger integration model can capture more nuanced information about the object and improve reconstruction fidelity.

To evaluate whether our model captures MTC-like integration, we test on MOCHI (Bonnen et al., 2024); success on this benchmark depends on the ability to integrate VTC encodings into unified object representations-a capacity known to depend on MTC (Bonnen et al., 2023). We compare gaze from three sources: attention maps from the VTC-like model, gazing log-probs from our model, and human participants (Fig 4). Overall, we find that our maps more confidently highlight regions similar to what humans focus on. However, they are generally noisier and over-weight the background. Further work is needed to learn more human-aligned maps for sequential sampling.


Figure 4: Salience maps from the VTC-like model (DINOv2 attention map), MTC-like model (gaze log probabilities), and humans.

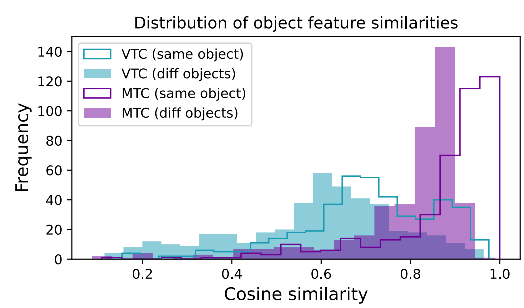


Figure 5: Distribution of feature similarities for the same object (outlined) and different objects (filled) for both models.

Finally, we analyze how zero-shot feature similarity is affected on the unseen objects in MOCHI: in Fig 5, we plot the distribution of feature similarities between views of the same object versus views of different objects. In aggregate, both models exhibit higher similarity among same-object views, but the VTC+MTC model produces more cleanly separated distributions. These results indicate that sequential integration can help disentangle object identity across views and supports the formation of more coherent object representations.

## References


Barense, M. D., Gaffan, D., & Graham, K. S. (2007). The human medial temporal lobe processes online representations of complex objects. *Neuropsychologia*, *45*, 2963–2974. Retrieved from 10.1016/j.neuropsychologia.2007.05.023

Bonnen, T., Fu, S., Bai, Y., O'Connell, T., Friedman, Y., Kanwisher, N., ... Efros, A. (2024). Evaluating multiview object consistency in humans and image models. *Advances in Neural Information Processing Systems*, *37*, 43533–43548.

Bonnen, T., Wagner, A. D., & Yamins, D. L. (2023). Perirhinal cortex supports object perception by integrating over visuospatial sequences. *bioRxiv*, 2023–09.

Bonnen, T., Yamins, D. L., & Wagner, A. D. (2021). When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, *109*(17), 2755–2766.

Bussey, T. J., Saksida, L. M., & Murray, E. A. (2002). Perirhinal cortex resolves feature ambiguity in complex visual discriminations. *Eur. J. Neurosci.*, *15*(2), 365–374. Retrieved from 10.1046/j.0953-816x.2001.01851.x

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. Retrieved from 10.1016/j.neuron.2012.01.010

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing* (No. 37). Oxford University Press.

Hirsch, J., & Curcio, C. A. (1989). The spatial resolution capacity of human foveal retina. *Vision research*, *29*(9), 1095–1101.

Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., ... Tan, H. (2024). *Lrm: Large reconstruction model for single image to 3d.* Retrieved from https://arxiv.org/abs/2311.04400

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A convnet for the 2020s.* Retrieved from https://arxiv.org/abs/2201.03545

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in cognitive sciences*, *4*(1), 6–14.

Murray, E. A., & Bussey, T. J. (1999). Perceptual–mnemonic functions of the perirhinal cortex. *Trends. Cogn. Sci.*, *3*(4), 142–151. Retrieved from 10.1016/s1364-6613(99)01303-0

Murray, E. A., & Wise, S. P. (2012). Why is there a special issue on perirhinal cortex in a journal called hippocampus? the perirhinal cortex in historical perspective. *Hippocampus*, *22*(10), 1941–1951. Retrieved from 10.1002/hipo.22055

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... Bojanowski, P. (2024). *Dinov2: Learning robust visual features without supervision.* Retrieved from https://arxiv.org/abs/2304.07193

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*(4180), 965–966.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Van Essen, D. C., & Anderson, C. H. (1990). Information processing strategies and pathways in the primate retina and visual cortex. In *An introduction to neural and electronic networks* (pp. 43–72).

Xiao, W., Sharma, S., Kreiman, G., & Livingstone, M. S. (2024). Feature-selective responses in macaque visual cortex follow eye movements during natural vision. *Nature neuroscience*, *27*(6), 1157–1166.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA.*, *111*(23), 8619–8624. Retrieved from 10.1073/pnas.1403112111

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). *The unreasonable effectiveness of deep features as a perceptual metric.* Retrieved from https://arxiv.org/abs/1801.03924