# Hierarchical VAEs provide a normative account of motion processing in the primate brain

**Hadi Vafaii**[1] vafaii@umd.edu
**Jacob L. Yates**[2] yates@berkeley.edu
**Daniel Butts**[1] dab@umd.edu

[1]University of Maryland, College Park [2]UC Berkeley

## Abstract

The relationship between perception and inference, as postulated by Helmholtz in the 19th century, is paralleled in modern machine learning by generative models like Variational Autoencoders (VAEs) and their hierarchical variants. Here, we evaluate the role of hierarchical inference and its alignment with brain function in the domain of motion perception. We introduce a novel synthetic data framework, Retinal Optic Flow Learning (ROFL), which enables control over motion statistics and their causes. We introduce a new hierarchical VAE and test it against alternative models on two downstream tasks: (i) predicting ground truth causes of retinal optic flow (e.g., self-motion); and (ii) predicting the responses of neurons in the motion processing pathway of primates. We manipulate the model architectures (hierarchical versus non-hierarchical), loss functions, and the causal structure of the motion stimuli. We found that hierarchical latent structure in the model leads to several improvements. First, it improves the linear decodability of ground truth variables and does so in a sparse and disentangled manner. Second, our hierarchical VAE outperforms previous state-of-the-art models in predicting neuronal responses with a performance gain of over 2x and exhibits sparse latent-to-neuron relationships. Finally, these results depend on the causal structure of the world, indicating that alignment between brains and artificial neural networks depends not only on architecture but also on matching ecologically relevant stimulus statistics. Collectively, these results support the notion that hierarchical Bayesian inference underlies the brain's understanding of the world, and hierarchical VAEs offer an effective means of capturing this understanding.

## 1 Introduction

Intelligent interactions with the world require representation of its underlying composition. This inferential process has long been postulated to underlie human perception [1–9], and is paralleled in modern machine learning by generative models [10–17], which learn latent representations of the generative process. The question of what constitutes a "good" representation has no clear answer [18, 19], but several desirable features have been proposed. In the field of neuroscience, studies focused on object recognition have suggested that effective representations "*untangle*" the various factors of variation in the input, rendering them linearly decodable [20, 21]. This intuitive notion of linear decodability has independently emerged in the machine learning community under different names such as "*informativeness*" [22] or "*explicitness*" [23]. Additionally, it has been suggested that "*disentangled*" representations are desirable, wherein distinct, informative factors of variations in the data are separated [19, 24–29]. ANNs are also increasingly evaluated based on their alignment with biological neural processing [30–38], because of the shared goals of artificial neural networks (ANNs) and the brain's sensory processing [25, 39]. Such alignment also provides the possibility of gaining insights into the brain by understanding the operations within an ANN [40–45].

Preprint. Under review.

In this work, we investigate how the combination of (i) model architecture, (ii) loss function, and (iii) training set, affects learned representations, and whether this is related to the brain-alignment of the ANN [43]. Notably, our work contributes to 3 out of 4 core ingredients of the "*neuroconnectionist*" research programme [40], which describes the necessary ingredients for using ANNs as a computational language for expressing falsifiable theories about brain computation.

We focus specifically on understanding the representation of motion because large sections of the visual cortex are devoted to processing motion [34] and the causes of retinal motion (moving objects and self-motion [46]) can be manipulated systematically. Crucially, motion in an image can be described irrespective of the identity and specific visual features that are moving, just as the identity of objects is invariant to how they are moving. This separation of motion and object processing mirrors the division of primate visual processing into dorsal (motion) and ventral (object) streams [47, 48].

We designed a *naturalistic* motion simulation based on distributions of ground truth variables corresponding to the location and depth of objects, motion of these objects, motion of the observer, and their direction of gaze (i.e., the fixation point; Figure 1a). Using our simulated retinal flow, we then trained and evaluated an ensemble of autoencoder-based models on our simulation. We based our evaluation on (1) whether the models untangle and disentangle the ground truth variables in our simulation; and (2) the degree to which their latent spaces could be directly related to neural data recorded in the dorsal stream of primates (area MT).

We introduce a novel hierarchical variational autoencoder (cNVAE) based on compressing NVAE [49]. The cNVAE exhibited superior performance compared to other models in all of our evaluation metrics. First, it discovered latent factors that accurately captured the ground-truth variables in the simulation in a more disentangled manner than other models. Second, it achieved significant improvements in predicting neural responses compared to the previous state-of-the-art model [34], with a performance gain of over 2x and sparse mapping from its latent space to neural responses.

Taken together, these observations demonstrate the power of the synthetic data framework and show that a single inductive bias—hierarchical latent structure—leads to many desirable features of representations, including brain alignment.

## 2 Background & Related Work

**Neuroscience and VAEs.** It has long been argued that perception and inference are tightly linked. This concept is commonly referred to as "perception as unconscious inference", initially proposed by Helmholtz in the 19th century [1]. However, similar ideas can be traced back more than a millennium to Alhazen [2], and possibly even earlier. More recently, the Bayesian Brain Hypothesis suggested that brains contain an internal generative model of the world [6, 8, 50]. Furthermore, the hierarchical architecture of the cortex, paired with the larger volume of feedback connections relative to feedforward connections, inspired Mumford [3] to conjecture that brains engage in hierarchical Bayesian inference to comprehend the world [3, 5]. A similar idea underlies modern generative models [17, 51], especially hierarchical variants of VAEs [49, 52–54].

Nouveau VAE (NVAE) [49] and very deep VAE (vdvae) [52] demonstrated that deep hierarchical VAEs generate realistic high-resolution images, overcoming the limitations of their non-hierarchical predecessors. However, neither work evaluated how the hierarchical latent structure changed the quality of learned representations. Additionally, the NVAE and vdvae both have an undesirable property: their convolutional latents result in a latent space that is several orders of magnitude larger than the input space, defeating the main purpose of autoencoders: compression. Indeed, Hazami et al. [55] showed that only a tiny subset (around 3%) of the vdvae latent space is necessary. Here, we demonstrate that it is possible to compress hierarchical VAEs and focus on investigating their latent representations with applications to neuroscience data.

**Evaluating ANNs on predicting biological neurons.** Several studies have focused on evaluating ANNs on their performance in predicting brain responses, but almost entirely on describing static ("ventral stream") image processing [30, 33, 36]. In contrast, motion processing (corresponding to the dorsal stream) has only been considered thus far in Mineault et al. [34], who used a 3D ResNet (DorsalNet) to extract ground truth variables about self-motion from drone footage ("AirSim", [56]) in a supervised manner. DorsalNet learned representations with receptive fields that matched known

features of the primate dorsal stream and achieved state-of-the-art on predicting neural responses on the dataset that we consider here. In addition to model architecture and training set, a fundamental difference between our approach and Mineault et al. [34] is that they train their models using direct supervision. As such, their models have access to the ground truth variables at all times. Here, we demonstrate that it is possible to obtain ground-truth variables "for free", in a completely unsupervised manner, while achieving much better performance in predicting responses of biological neurons.

**Using synthetic data to train ANNs.** A core component of a reductionist approach to studying the brain is to characterize neurons based on their selectivity to a particular subset of pre-identified visual "features", usually by presenting sets of "feature-isolating" stimuli [57]. In the extreme, stimuli are designed that remove all other features except the one under investigation [58]. While these approaches can inform how pre-selected feature sets are represented by neural networks, it is often difficult to generalize this understanding to more natural stimuli, which are not necessarily well-described by any one feature set. As a result, here we generate synthetic data representing a *naturalistic* distribution of natural motion stimuli. Such synthetic datasets allow us to manipulate the causal structure of the world, in order to make hypotheses about what aspects of the world matter for the representations learned by brains and ANNs [59]. Like previous work on synthesized textures [15], here we specifically manipulate the generative structure to contain factors of variation due to known ground truth variables.

## 3 Approach: Data & Models

**Retinal Optic Flow Learning (ROFL).** Our synthetic dataset framework, ROFL, generates the resulting optic flow from different world structures, self-motion trajectories, and object motion (Fig. 1a). This can be seen as an extension of previous work [60].

ROFL can be used to generate *naturalistic* flow fields that share key elements with those experienced in navigation through 3-D environments. Specifically, each frame contains global patterns that are due to self-motion, including rotation that can arise due to eye or head movement [61, 62]. In addition, local motion patterns can be present due to objects that move independently of the observer [46]. The overall flow pattern is also affected by the observer's direction of gaze (fixation point, Fig. 1d).

| Category | Description | Dimensionality |
|---|---|---|
| `fixate-1` | A moving observer maintains fixation on a background point. In addition, the scene contains one independently moving object. | $11 = 2 + 3 + 6$ |
| `fixate-0` | Same as `fixate-1` but without the object. | $5 = 2 + 3$ |
| `obj-1` | A single moving object, stationary observer. | $8 = 2 + 6$ |

Table 1: ROFL categories used in this paper. Ground truth variables include fixation point $(+2)$; velocity of the observer when self-motion is present $(+3)$; and, object position & velocity $(+6)$. Figure 1b showcases a few example frames for each category. The stimuli can be rendered at any given spatial scale $N$, yielding an input shape of $2 \times N \times N$. Here we work with $N = 17$.

ROFL generates flow vectors that are instantaneous in time, representing the velocity across the visual field resulting from the spatial configuration of the scene and motion vectors of self and object. Ignoring the time-evolution of a given scene (which can arguably be considered separably [63]) dramatically reduces the input space from $[3 \times H \times W \times T]$ to $[2 \times H \times W]$, and allows a broader sampling of configurations without introducing changes in luminance and texture. As a result, we can explore the role of different causal structures in representation learning in ANNs.

From an observer's perspective, the retinal flow patterns generated by a moving object depend on the observer's self-motion and the rotation of their eyes as they maintain fixation in the world. One such interaction is visualized in Fig. 1c. In this example, the observer is moving forward, and the object is moving to the right. As a result of interactions between self and object motion, an object on the left side will have its flow patterns distorted, while an object on the right will have its flow patterns largely unaffected because its flow vectors are parallel with that of the self-motion. In summary, ROFL allows us to simulate retinal optic flow with a known ground truth structure driven by object and self-motion.

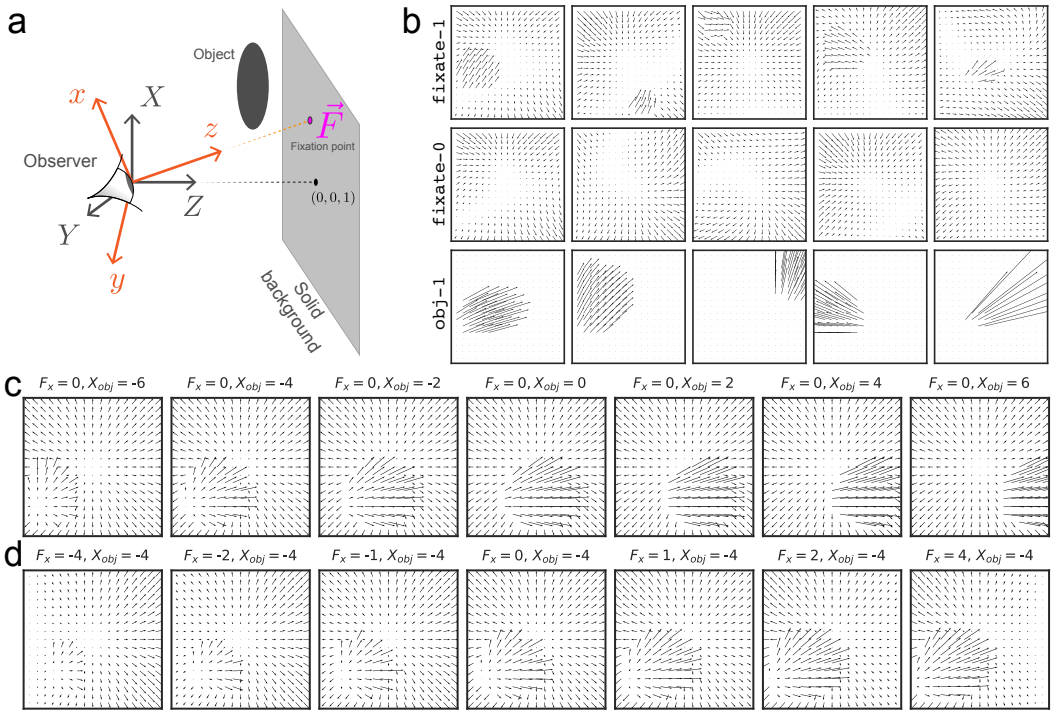


Figure 1: Introducing Retinal Optic Flow Learning (ROFL), a simulation platform for synthesizing naturalistic optic flow patterns. **(a)** The general setup includes a moving or stationary observer and a solid background, with optional moving object(s) in the scene. More details are provided in the appendix (section 12). **(b)** Example frames showcasing different categories (see Table 1 for definitions). **(c, d)** Demonstrating the causal effects of varying a single ground truth variable while keeping all others fixed. **(c)** $X_{obj}$, the $x$ component of object position (measured in retinal coordinates, orange). **(d)** $F_x$, the $X$ component of the fixation point (measured in fixed coordinates, gray).

**The compressed NVAE (cNVAE).** Our model architecture is shown in Fig. 2. The cNVAE closely follows the NVAE [49] with one important difference: the original NVAE latent space is convolutional, and ours is not. We modified the *sampler* layers such that their receptive field sizes match the spatial scale they operate on, which effectively integrates over spatial information before sampling from the approximate posterior. To project each latent back to convolution space, we introduce *expand* modules, which are a simple deconvolution step. Further details about the processing of the sampler and expand layers are provided in supplementary section 11.1.

Our modification of the NVAE serves two purposes. First, it decouples spatial information from the functionality of latent groups, allowing them to capture abstract features that are invariant to particular spatial locations. Second, it has the effect of compressing the input space into a lower-dimensional latent code. A convolutional NVAE latent group operating at spatial scale $s$ has $k \times s \times s$ latent dimensions, where $k$ is the number of latents per group. In contrast, each latent group in our model has just $k$ dimensions. Thus, a cNVAE with 21 latent groups, each with 20 latent dimensions, has a 420 dimensional latent space (see below), which is smaller than the input dimensionality of 578 ($2 \times 17 \times 17$). In contrast, the NVAE would have a $17,520$-dimensional latent space using otherwise similar settings as above.

**Structure of cNVAE latent space.** The latent space of cNVAE contains a hierarchy of latent groups sampled sequentially: "top" latents are sampled first, all the way down to "bottom" latents that are closest to the stimulus (Fig. 2). Child [52] demonstrated that model depth, in this statistical sense, was responsible for the success of hierarchical VAEs. Here, we increased the depth of cNVAE architecture while ensuring that training remained stable. Our resulting model had the following structure: 3 latent groups operating at the scale of $2 \times 2$; 6 groups at the scale of $4 \times 4$; and 12 groups at the scale of

| Model | Architecture | Loss | KL term |
|---|---|---|---|
| cNVAE | Hierarchical | EPE $+\beta * \mathrm{KL}$ | $\mathrm{KL} = \sum_{\ell=1}^{L} \mathbb{E}_{q(\mathbf{z}_{<\ell}\|\mathbf{x})}\left[\mathrm{KL}_\ell\right]$, where $\mathrm{KL}_\ell := D_{\mathrm{KL}}\left[q\left(\mathbf{z}_\ell \mid \mathbf{x}, \mathbf{z}_{<\ell}\right) \| p\left(\mathbf{z}_\ell \mid \mathbf{z}_{<\ell}\right)\right]$ |
| VAE | Non-hierarchical | EPE $+\beta * \mathrm{KL}$ | $\mathrm{KL} = D_{\mathrm{KL}}\left[q\left(\mathbf{z} \mid \mathbf{x}\right) \| p\left(\mathbf{z}\right)\right]$ |
| cNAE | Hierarchical | EPE | - |
| AE | Non-hierarchical | EPE | - |

Table 2: Model details. Here, *hierarchical* means that there are parallel pathways for information to flow from the encoder to the decoder (Fig. 2), which is slightly different from the conventional notion. For variational models, this implies hierarchical dependencies between latents in a statistical sense [52]. This hierarchical dependence is reflected in the KL term for the cNVAE, where $L$ is the number of hierarchical latent groups. All models have an equal # of latent dimensions, approximately the same # of convolutional layers, and # of parameters ($\sim 24\ M$). EPE, endpoint error.

$8 \times 8$ (Fig. 2). Therefore, the model had a total of $3 + 6 + 12 = 21$ hierarchical latent groups. Each latent group had 20 latents, which resulted in an overall latent dimensionality of $21 \times 20 = 420$.

**Alternative models.** We evaluated a range of unsupervised models alongside cNVAE. These models include standard (non-hierarchical) VAEs [11, 12], a hierarchical autoencoder with identical architecture as the cNVAE but trained only with reconstruction loss (cNAE), and an autoencoder (AE) counterpart for the VAE. All models had the same latent dimensionality, and approximately the same number of parameters and number of convolutional layers. We used endpoint error as our measure of reconstruction loss, which is the Euclidean norm of the difference between actual and reconstructed flow vectors. This metric works well with optical flow data. [64]. See Table 2 for a summary.

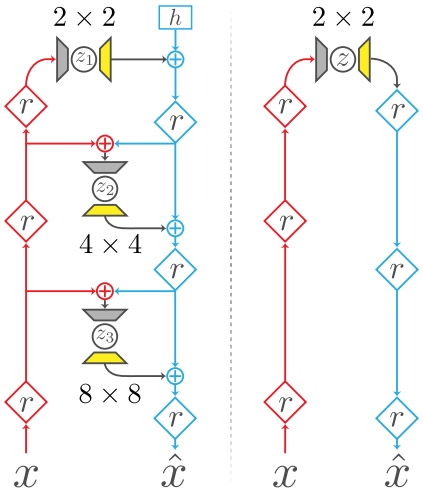


Figure 2: Architecture comparison. Left, compressed NVAE (cNVAE); right, non-hierarchical VAE. We modified the NVAE *sampler* layer (grey trapezoid) and introduced a deconvolution *expand* layer (yellow trapezoid). The encoder (inference) and decoder (generation) pathways are depicted in red and blue, respectively. $r$, residual block; $h$, trainable parameter; $+$, feature combination.

**Model representations.** We define a model's internal representation to be either the mean of each Gaussian for variational models (i.e., samples drawn from $q\left(\mathbf{z}|\mathbf{x}\right)$ at zero temperature), or the bottleneck activations for autoencoders.

**Training details.** Models were trained for 160 $k$ steps at an input scale of $17 \times 17$, requiring slightly over a day on Quadro RTX 5000 GPUs. Please refer to the supplementary section 11.2 for additional information and a comprehensive list of hyperparameters.

**Disentanglement and $\beta$-VAEs.** A critical decision when optimizing VAEs involves determining the weight assigned to the KL term in the loss function compared to the reconstruction loss. Prior research has demonstrated that modifying a single parameter, denoted as $\beta$, which scales the KL term, can lead to the emergence of disentangled representations [65, 66]. Most studies employing VAEs for image reconstruction typically optimize the standard evidence lower bound (ELBO) loss, where $\beta$ is fixed at a value of 1 [11, 49, 52]. However, it should be noted that due to the dependence of the reconstruction loss on the input size, any changes in the dimensionality of the input will inevitably alter the relative contribution of the KL term [65], and thus the "effective" $\beta$.

Furthermore, a recent study by Higgins et al. [16] has established a strong correspondence between the generative factors discovered by $\beta$-VAEs and the factors encoded by inferotemporal (IT) neurons in the primate ventral stream. The alignment between these factors and IT neurons exhibits a linear relationship with the value of $\beta$. In light of these findings, we explicitly manipulate the parameter $\beta$ within a range spanning from $0.01$ to $10$ to investigate the extent to which our results depend on its value.

# 4 Results

## 4.1 Functional specialization emerges in hierarchical VAEs

Our approach is based on the premise that the visual world contains a hierarchical structure. We use a simulation containing a hierarchical structure (ROFL, described above) and a hierarchical VAE (the cNVAE, above) to understand how this architectural choice affects latent representations. While we are using a relatively simple simulation generated from a small number of ground-truth variables, $\mathbf{g}$, we do not specify how $\mathbf{g}$ should be represented in our model or include $\mathbf{g}$ in the loss. Rather, we allow the model to develop its own latent representation in a purely unsupervised manner.

For our first experiment, we trained and tested hierarchical and non-hierarchical VAEs on `fixate-1` (see Table 1; throughout this work, `fixate-1` is used unless stated otherwise). We then extracted latent representations from each model and estimated the mutual information (MI) between the representations and ground truth variables such as self-motion, etc. For `fixate-1`, each data sample is uniquely determined using 11 ground truth variables (Table 1), and the models have latent dimensionality of $420$. Thus, the resulting MI matrix has shape $11 \times 420$, where each entry shows how much information is contained in that latent variable about a given ground truth variable.

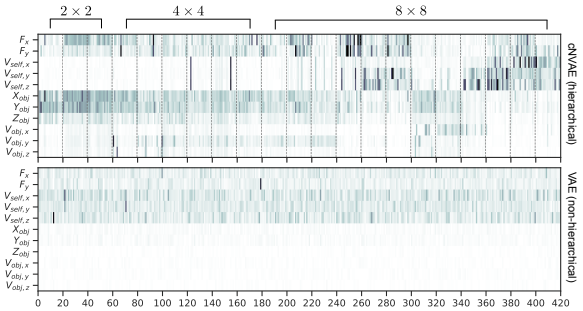


Figure 3: Mutual information between latent variables (x-axis) and ground truth variables (y-axis) is shown for cNVAE (top) and VAE (bottom). Dashed lines indicate 21 hierarchical latent groups of 20 latents each, comprising a $420$-dimensional latent space. These groups operate at three different spatial scales, as indicated. In contrast, the VAE latent space lacks such grouping and operates solely at the spatial scale of $2 \times 2$ (see Fig. 2).

Figure 3 shows the MI matrix for the latent space of cNVAE (top) and VAE (bottom). While both models achieved a good reconstruction of validation data (Fig. S7), the MI matrix for cNVAE exhibits clusters corresponding to distinct ground-truth variables at different levels of the hierarchy. Specifically, object-related factors of variation are largely captured at the top $2 \times 2$ scale, while information about fixation point can be found across the hierarchy, and self-motion is largely captured by $8 \times 8$ latent groups. In contrast, non-hierarchical VAE has no such structure. This suggests that the inductive bias of hierarchy enhances the quality of latent spaces, which we will quantify next.

**Evaluating the latent code.** To quantitatively measure the clear relationship between ground truth variables and latent representations discovered by the cNVAE visible in Fig. 3, we apply three metrics referred to as "untangling", "disentengling", and "brain-alignment". We discuss each of these in detail in the following sections.

**Untangling: the cNVAE untangles factors of variation.** One desirable feature of a latent representation is whether it makes information about ground truth factors easily (linearly) decodable [20, 21, 67]. This concept has been introduced in the context of core object recognition as "*untangling*". Information about object identity that is "tangled" in the retinal input is untangled through successive nonlinear transforms, thus making it linearly available for higher brain regions to extract [20]. This concept is closely related to the "*informativeness*" metric of Eastwood and Williams [22] and "*explicitness*" metric of Ridgeway and Mozer [23].

To assess the performance of our models, we evaluated the linear decodability of the ground truth variables, $\mathbf{g}$, from their respective latent codes, $\mathbf{z}$. Based on the $R^2$ scores obtained by predicting $\mathbf{g}$ from $\mathbf{z}$ using linear regression (Fig. 4), the cNVAE greatly outperforms competing models, faithfully capturing all ground-truth variables. In contrast, the non-hierarchical VAE fails to capture object-related variables. Notably, the cNVAE can recover the fixation point location $(F_X, F_Y)$ in physical space almost perfectly. The fixation location has a highly nontrivial effect on the flow patterns, and varying it causes both global and local changes in the flow patterns (Fig. 1d).

Furthermore, cNVAE is the only model that reliably captures object position and velocity: especially note $V_{obj,z}$; last column in Fig. 4. To compute object motion, a model must extract self-motion and understand how it influences flow patterns globally: only then can the model subtract self-motion from global flow vectors to obtain object motion. In vision science, this is known as the "*flow-parsing hypothesis*" [68–71]. Such flow-parsing is achieved by the cNVAE but none of the other models. See supplementary section 8 for further discussion of this result and its implications.

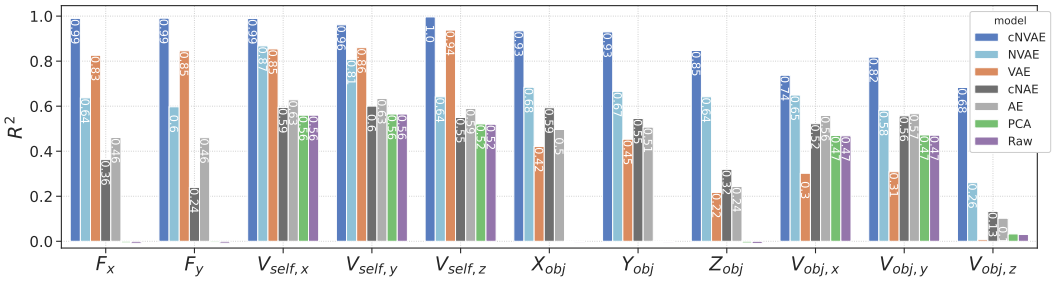


Figure 4: Hierarchical VAE untangles underlying factors of variation in data. The linear decodability of ground truth factors (x-axis) from different latent codes is shown. Untangling scores averaged across all ground truth variables are cNVAE $= 0.898$; NVAE $= 0.639$; VAE $= 0.548$; cNAE $= 0.456$; AE $= 0.477$; PCA $= 0.236$; Raw $= 0.235$. For variational models, the best performing $\beta$ value was selected: cNVAE, $\beta = 0.15$; VAE, $\beta = 1.5$. See Fig. 5 for the full range of $\beta$ values explored.

**Disentanglement: the cNVAE produces more disentangled representations.** The pursuit of disentanglement in neural representations has garnered considerable attention [23, 65, 72–80]. Prior efforts have predominantly focused on various forms of regularization, such as $\beta$-VAE [65] that demonstrated that increasing the weight of the KL loss (indicated by $\beta$) promotes disentanglement in VAEs. However, despite the success of these approaches, Locatello et al. [19] established theoretically that learning fully disentangled representations is fundamentally impossible without inductive biases. Recent work demonstrated that simple biologically inspired constraints such as non-negativity and energy constraints encourage disentanglement [72]. Here, we hypothesize that another biological inductive bias, hierarchy in the latent space, will improve the representations learned by VAEs.

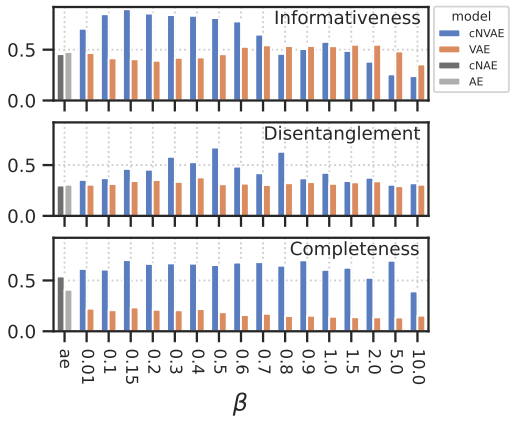


Figure 5: Evaluating the learned latent codes using the DCI framework [22]. Larger values are better for all metrics. Note that *informativeness* is closely related to *untangling* [20, 21].

To evaluate the role of hierarchy, we adopted the DCI framework [22] which offers a well-rounded evaluation of latent representations. The approach involves training a simple decoder (e.g., lasso regression) that predicts data generative factors $\mathbf{g}$ from a latent code $\mathbf{z}$; followed by computing a matrix of relative (e.g., based on lasso weights) which is then used to evaluate different aspects of the code quality: *Informativeness*—measures whether $\mathbf{z}$ contains easily accessible information about $\mathbf{g}$ (similar to untangling from above).

*Disentanglement*—measures whether individual latents correspond to individual generative factors. *Completeness*—measures how many $z_i$ are required to capture any single $g_j$. If a single latent contributes to $g_j$'s prediction, the score will be 1 (complete). If all latent variables equally contribute to $g_j$'s prediction, the score will be 0 (maximally overcomplete). Note that "*completeness*" is also referred to as "*compactness*" [23]. See ref. [81] for a review, and ref. [82] for a recent extension of the DCI framework.

Here, we follow the methods outlined by Eastwood and Williams [22] with two modifications: (1) we replaced lasso with linear regression to avoid the strong dependence on the lasso coefficient that we observed, and (2) we estimate the matrix of relative importances using a feature permutation-based algorithm (sklearn.inspection.permutation_importance), which measures the relative amount of performance drop based on the shuffling of a given feature.

We found that cNVAE outperforms competing models across all metrics for a broad range of $\beta$ values (Fig. 5). The observed pattern of an inverted U shape is consistent with previous work [65], which suggested there is a sweet spot for $\beta$ that can be found empirically. In our case, cNVAE with $\beta = 0.5$ achieved the best average DCI score. Further, we found that VAEs lacking hierarchical structure learn highly overcomplete codes, such that many latents contribute to predicting a single ground truth variable. In conclusion, the simple inductive bias of hierarchy in the latent space led to a substantial improvement in VAE performance across all components of the DCI metric.

**Brain-alignment: the cNVAE aligns more closely with MT neurons.** To evaluate the performance of models at predicting the activity of neurons in response to motion stimuli, we used a dataset of $N = 141$ previously recorded MT neurons [83, 84]. A subset of these neurons ($N = 84$) are publicly available on crcns.org, and were recently used in Mineault et al. [34] that we compare against.

To measure neuron alignment, we first determined the mapping between each model's internal representations and neural data (Fig. 6a). We define *representation* to be the mean of each Gaussian for VAEs, and the bottleneck activations for AEs. We learn a linear mapping from each model's representation to neural firing rates using ridge regression. Figure 6b shows the average firing rate and model predictions for an example neuron. Because sensory neurons have a nonzero response latency, for each neuron we determine its response latency that maximizes cross-validated performance. The distribution of best-selected latencies (Fig. 6c) peaked around $100\ ms$, which is consistent with known MT latencies [83]. We also optimize over 20 orders of magnitude of ridge coefficients to ensure each neuron has its best fit. Figure 6d shows that the models capture the receptive field properties of MT neurons as measured by the spike-triggered average stimulus. To evaluate performance, we follow methods established by Mineault et al. [34]: whenever repeated trials were available, we report Pearson's $R$ on that held-out data, normalized by maximum explainable variance [85]. When repeats were not available, we performed 5-fold cross-validation and report the held-out performance using Pearson's $R$ between model prediction and spike trains.

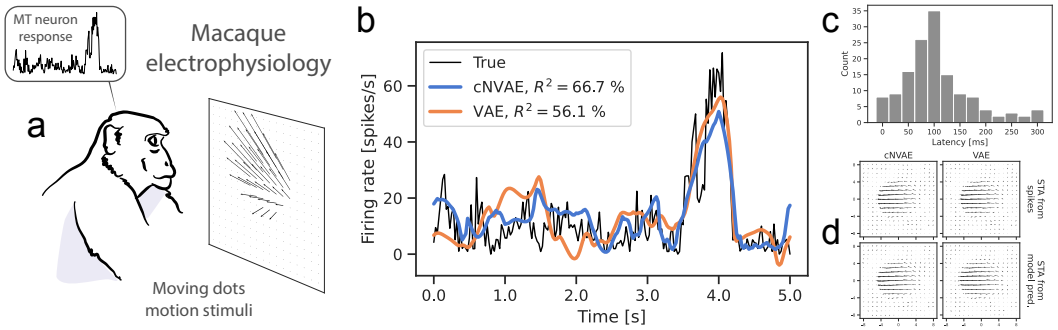


Figure 6: **(a)** Experimental setup form [83, 84]. **(b)** Both models explain MT neural variability well. **(c)** Distribution of best estimated latencies. **(d)** Spike-triggered averages (STA) are shown.

**Evaluating brain alignment.** We use two measures of brain alignment: the success at predicting the neural response ($R$, Fig. 7, Table 3); and, the "*alignment*" between neurons and individual model latents (Fig. 8). These mirror the untangling and completeness metrics described above (more details are provided below).

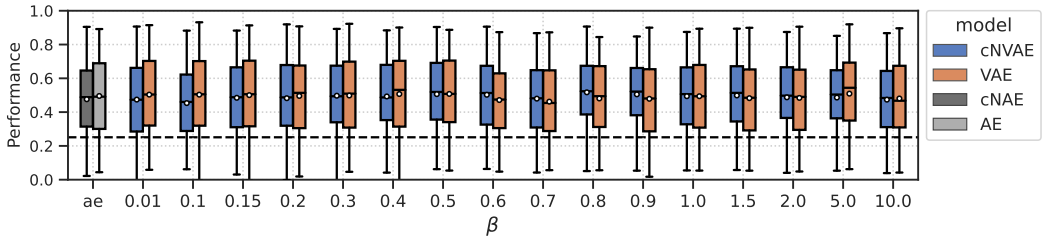


Figure 7: All models (pretrained on `fixate-1`) perform comparably in predicting MT neuron responses. Dashed line corresponds to the previous state-of-the-art on this data [86].

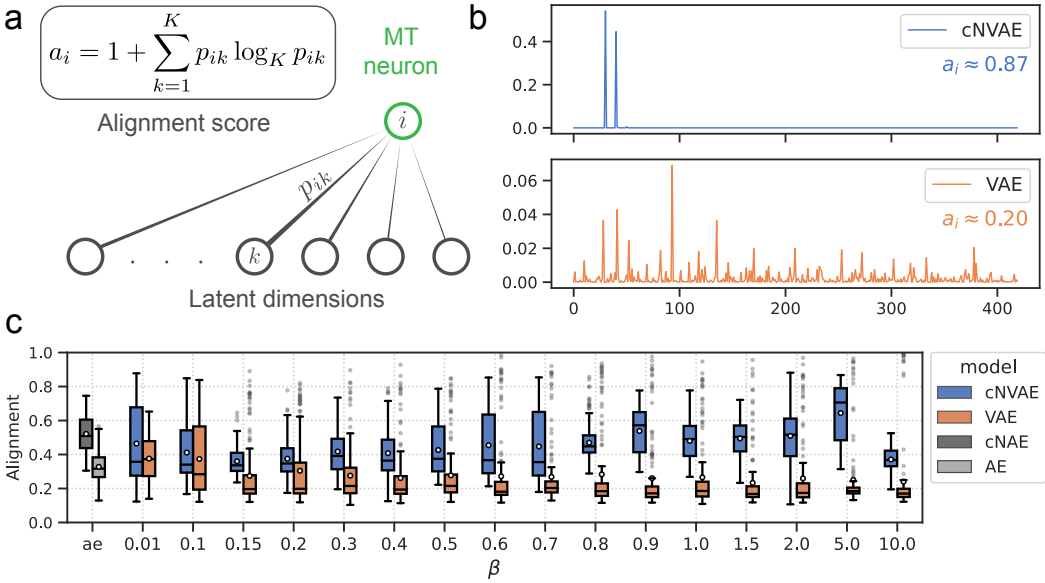


Figure 8: Hierarchical models (cNVAE, cNAE) are more aligned with MT neurons since they enable sparse latent-to-neuron relationships. **(a)** Alignment score measures the sparsity of permutation feature importances. $a_i = 0$ when all latents are equally important in predicting neuron $i$; and, $a_i = 1$ when a single latent predicts the neuron. **(b)** Feature importances are plotted for an example neuron (same as in Fig. 6b). cNVAE predicts this neuron's response in an extremely sparse manner, in sharp contrast to non-hierarchical VAE. **(c)** Alignment across $\beta$ values, and autoencoders (ae).

**All models predict MT responses well.** After training a large ensemble of unsupervised models on `fixate-1` and learning the neural mapping, we found that both hierarchical (cNVAE & cNAE) and non-hierarchical (VAE & AE) variants performed roughly equally well at predicting neural responses (Fig. 7). We did observe some dependence on the loss function with the variational loss outperforming simple autoencoder reconstruction loss (Table 3).

**Hierarchical VAEs are more aligned with MT neurons.** We next tested how these factors affect neural alignment, i.e., how closely neurons are related to individual latents in the model. Figure 8a demonstrates what we mean by "alignment": a sparse latent-to-neuron relationship means larger alignment. To formalize this notion, we use feature permutation importance applied to the ridge regression coefficients. This yields a 420-dimensional vector per neuron. Each dimension of this vector captures the importance of a given latent variable in predicting the responses of the neuron. We normalize these vectors and interpret them as the probability of importance. We then define alignment score $a_i$ of neuron $i$ as $a_i = 1 + \sum_{k=1}^{K} p_{ik} \log_K p_{ik}$, where $p_{ik}$ is interpreted as the importance of $k-$th latent variable in predicting neuron $i$ (Fig. 8a). This concept is closely related to the "*completeness*" score from the DCI framework [22] (see above), and it was also employed by Higgins et al. [16], although they used lasso regression with the magnitude of nonzero coefficients as their feature importance.

| Model | Pretraining dataset | Performance, $R$ $(\mu \pm se;\ N = 141)$ | | | |
|---|---|---|---|---|---|
| | | $\beta = 0.5$ | $\beta = 0.8$ | $\beta = 1$ | $\beta = 5$ |
| cNVAE | `fixate-1` | $\mathbf{0.506 \pm 0.018}$ | $\mathbf{0.517 \pm 0.017}$ | $0.494 \pm 0.018$ | $0.486 \pm 0.016$ |
| | `fixate-0` | $0.428 \pm 0.018$ | $0.450 \pm 0.019$ | $0.442 \pm 0.019$ | $0.469 \pm 0.018$ |
| | `obj-1` | $0.471 \pm 0.018$ | $0.465 \pm 0.018$ | $0.477 \pm 0.017$ | $0.468 \pm 0.018$ |
| VAE | `fixate-1` | $\mathbf{0.508 \pm 0.019}$ | $0.481 \pm 0.018$ | $0.494 \pm 0.018$ | $\mathbf{0.509 \pm 0.018}$ |
| cNAE | `fixate-1` | $0.476 \pm 0.018$ | | | |
| AE | `fixate-1` | $0.495 \pm 0.019$ | | | |
| CPC | AirSim | $0.250 \pm 0.020$ (Mineault et al. [34]) | | | |
| DorsalNet | AirSim | $0.251 \pm 0.019$ (Mineault et al. [34]) | | | |

Table 3: Both cNVAE and VAE perform well in predicting MT neuron responses, surpassing previous state-of-the-art models by more than a twofold improvement. Moreover, the clear gap between `fixate-1` and other categories highlights the importance of pretraining data.

For almost all $\beta$ values, the cNVAE exhibited a greater brain alignment than non-hierarchical VAE (cNVAE > VAE, paired $t-$test; see Fig. S9 and Table S1). Similarly, for the autoencoders, we found that the hierarchical variant outperformed the non-hierarchical one (cNAE > AE). Based on these observations, we conclude that higher brain alignment is primarily due to hierarchical latent structure. However, note that hierarchy in the traditional sense did not matter: all these models had approximately the same number of convolutional layers and parameters.

**Factors leading to brain-alignment.** To test the effect of the training dataset (i.e., category of ROFL) on model performance, we trained cNVAE models using `fixate-0`, `fixate-1`, and `obj-1` categories (Table 1), while also exploring a variety of $\beta$ values. We found that `fixate-1` clearly outperformed the other two ROFL categories (Table 3), suggesting that both global (e.g., self-motion) and local (e.g., object motion) sources of variation are necessary for learning MT-like representations. The effect of loss function was also visible: some beta values led to more alignment. But this effect was small compared to the effect of hierarchical architecture (Fig. 8c).

## 5 Discussion

We introduced a new framework for understanding and evaluating the representation of visual motion learned by artificial and biological neural networks. This framework provides a way to manipulate causes in the world and evaluate whether learned representations untangle and disentangle those causes. In particular, our framework enabled testing the influence of architecture (Fig. 2), loss function (Table 2), and training set (Table 1) on the learned representations, encompassing 3 out of the 4 core components of a recently proposed neuroconnectionist research programme [40]. In sum, our framework brings hypothesis-testing to understand [biological] neural processing of vision and provides an interpretive framework to understand neurophysiological data.

**Limitations & considerations.** Our current approach is necessarily simplified, as it focuses on establishing this framework and demonstrating its potential. As such, it currently has several limitations:

First, our simulation generates velocity fields, rather than pixels. This necessarily avoids the complexities of extracting flow patterns from real spatiotemporal movies and thus, does not require the models to capture the intricacies of the early stages of processing that are thought to do exactly this [87, 88]. Notably, this allows for direct comparison with recorded neural data in MT and MST using random dot kinematograms [83, 84, 86], but would likely fail at explaining V1 responses, which necessarily requires a pixel-computable model as in previous work (e.g., DorsalNet [34]). Likewise, our $> 2\times$ improvement in performance in explaining MT neural data over DorsalNet is likely due in part to their network being based on spatiotemporal movies, while the neural data used optic flow

patterns as we do. It is possible that a hierarchical VAE trained and tested on video stimuli would align more poorly than the model from Mineault et al. [34]. Future work in this space will involve rendering images in simulations and using image-computable models for a fair comparison.

Second, we did not model fully natural optic flow, nor the true complexity of 3-D natural environments and their interaction with eye movements and self-motion as has recently been measured [63, 89]. The simplification of our simulation still demonstrates the importance of including such elements in understanding neural representations and provides a framework for incorporating real eye-tracking and scene data [63, 89] into future work with ROFL.

Finally, we found that hierarchy in the latent space leads to many desirable features of latent representations, including increasing the alignment to biological representations. Here, we tested this on only one dataset from area MT, which leaves the question of whether this is a general principle of brain computation. Addressing this requires testing our approach on more data from other brain areas, such as MST. From previous work [86, 90], we expect that hierarchical computation is even more necessary for MST, which we leave as an open question to address in future work.

**Interpreting brain-alignment.** We measured the alignment between ANN models and MT neurons using both linear predictive power (Fig. 7), and an alternative measure of alignment that is sensitive to the sparsity of latent-to-neuron relationships ("alignment-score", Fig. 8a). Linear regression has been used extensively to measure similarity, or alignment, between pairs of representations [30, 31, 34, 36], but often results in a degenerate [30, 36, 38], and unreliable [91], measure of representational alignment. Our application of linear regression found that it was not a particularly effective signal in differentiating between models: although cNVAE produced the single best model in terms of neuronal prediction, we found that both hierarchical and non-hierarchical VAEs performed similarly in predicting MT neuron responses.

In contrast, the alignment score (Fig. 8a) was much more consistent in distinguishing between models (see Fig. S9), and revealed that hierarchical models (both cNVAE and cNAE) had much sparser latent-to-neuron relationships. The alignment score measures whether a model has learned a similar representational "form" to the brain, which would enable the sparsity of latent-to-neuron relationships. However, this alignment score also has limitations as it is not a proper *metric* in the mathematical sense [92]. Future work will consider more sophisticated metrics for brain alignment [92–95].

**Conclusions.** We used synthetic data to test how causal structure in the world affects the representations learned by autoencoder-based models and evaluated the learned representations based on how they represent ground truth variables and how well they align with biological brains. We found that a single inductive bias, hierarchical latent structure, leads to desirable representations and increased brain alignment.

## 6 Acknowledgments

This work was supported by NSF IIS-2113197 (HV and DAB), NSF DGE-1632976 (HV), and NIH R00EY032179 (JLY).

## References

[1] Hermann Von Helmholtz. *Handbuch der physiologischen Optik*. Vol. 9. Voss, 1867.

[2] Ibn al-Haytham. *Book of optics (Kitab Al-Manazir)*. 1011–1021 AD.

[3] David Mumford. "On the computational architecture of the neocortex: II The role of cortico-cortical loops". In: *Biological cybernetics* 66.3 (1992), pp. 241–251.

[4] Rajesh PN Rao and Dana H Ballard. "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". In: *Nature neuroscience* 2.1 (1999), pp. 79–87.

[5] Tai Sing Lee and David Mumford. "Hierarchical Bayesian inference in the visual cortex". In: *JOSA A* 20.7 (2003), pp. 1434–1448.

[6] David C Knill and Alexandre Pouget. "The Bayesian brain: the role of uncertainty in neural coding and computation". In: *TRENDS in Neurosciences* 27.12 (2004), pp. 712–719.

[7] Alan Yuille and Daniel Kersten. "Vision as Bayesian inference: analysis by synthesis?" In: *Trends in cognitive sciences* 10.7 (2006), pp. 301–308.

[8] Karl Friston. "A theory of cortical responses". In: *Philosophical transactions of the Royal Society B: Biological sciences* 360.1456 (2005), pp. 815–836.

[9] Andy Clark. "Whatever next? Predictive brains, situated agents, and the future of cognitive science". In: *Behavioral and brain sciences* 36.3 (2013), pp. 181–204.

[10] Peter Dayan et al. "The helmholtz machine". In: *Neural computation* 7.5 (1995), pp. 889–904.

[11] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[12] Danilo Jimenez Rezende et al. "Stochastic backpropagation and approximate inference in deep generative models". In: *International conference on machine learning*. PMLR. 2014, pp. 1278–1286.

[13] Lukas Schott et al. "Towards the first adversarially robust neural network model on MNIST". In: *arXiv preprint arXiv:1805.09190* (2018).

[14] Ilker Yildirim et al. "Efficient inverse graphics in biological face processing". In: *Science advances* 6.10 (2020), eaax5979.

[15] Katherine R Storrs et al. "Unsupervised learning predicts human perception and misperception of gloss". In: *Nature Human Behaviour* 5.10 (2021), pp. 1402–1417.

[16] Irina Higgins et al. "Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons". In: *Nature communications* 12.1 (2021), p. 6456.

[17] Joseph Marino. "Predictive coding, variational autoencoders, and biological connections". In: *Neural Computation* 34.1 (2022), pp. 1–44.

[18] Irina Higgins et al. "Towards a definition of disentangled representations". In: *arXiv preprint arXiv:1812.02230* (2018).

[19] Francesco Locatello et al. "Challenging common assumptions in the unsupervised learning of disentangled representations". In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[20] James J DiCarlo and David D Cox. "Untangling invariant object recognition". In: *Trends in cognitive sciences* 11.8 (2007), pp. 333–341.

[21] James J DiCarlo et al. "How does the brain solve visual object recognition?" In: *Neuron* 73.3 (2012), pp. 415–434.

[22] Cian Eastwood and Christopher KI Williams. "A framework for the quantitative evaluation of disentangled representations". In: *International Conference on Learning Representations*. 2018.

[23] Karl Ridgeway and Michael C Mozer. "Learning deep disentangled embeddings with the f-statistic loss". In: *Advances in neural information processing systems* 31 (2018).

[24] Yoshua Bengio et al. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[25] Brenden M Lake et al. "Building machines that learn and think like people". In: *Behavioral and brain sciences* 40 (2017), e253.

[26] Jonas Peters et al. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[27] Yann LeCun et al. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[28] Jürgen Schmidhuber. "Learning factorial codes by predictability minimization". In: *Neural computation* 4.6 (1992), pp. 863–879.

[29] Michael Tschannen et al. "Recent advances in autoencoder-based representation learning". In: *arXiv preprint arXiv:1812.05069* (2018).

[30] Martin Schrimpf et al. "Brain-score: Which artificial neural network for object recognition is most brain-like?" In: *BioRxiv* (2018), p. 407007.

[31] Daniel LK Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8619–8624.

[32] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. "Deep supervised, but not unsupervised, models may explain IT cortical representation". In: *PLoS computational biology* 10.11 (2014), e1003915.

[33] Daniel LK Yamins and James J DiCarlo. "Using goal-driven deep learning models to understand sensory cortex". In: *Nature neuroscience* 19.3 (2016), pp. 356–365.

[34] Patrick Mineault et al. "Your head is there to move you around: Goal-driven models of the primate dorsal pathway". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 28757–28771.

[35] Eric Elmoznino and Michael F Bonner. "High-performing neural network models of visual cortex benefit from high latent dimensionality". In: *bioRxiv* (2022), pp. 2022–07.

[36] Colin Conwell et al. "What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?" In: *BioRxiv* (2022), pp. 2022–03.

[37] Nicholas J Sexton and Bradley C Love. "Reassessing hierarchical correspondences between brain and deep networks through direct interface". In: *Science Advances* 8.28 (2022), eabm2219.

[38] Greta Tuckute et al. "Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions". In: *bioRxiv* (2023). DOI: `10.1101/2022.09.06.506680`.

[39] Anthony Zador et al. "Catalyzing next-generation Artificial Intelligence through NeuroAI". In: *Nature Communications* 14.1 (2023), p. 1597.

[40] Adrien Doerig et al. "The neuroconnectionist research programme". In: *Nature Reviews Neuroscience* (2023), pp. 1–20.

[41] Nancy Kanwisher et al. "Using artificial neural networks to ask 'why'questions of minds and brains". In: *Trends in Neurosciences* (2023).

[42] Blake Richards et al. "The application of artificial intelligence to biology and neuroscience". In: *Cell* 185.15 (2022), pp. 2640–2643.

[43] Blake A Richards et al. "A deep learning framework for neuroscience". In: *Nature neuroscience* 22.11 (2019), pp. 1761–1770.

[44] David GT Barrett et al. "Analyzing biological and artificial neural networks: challenges with opportunities for synergy?" In: *Current opinion in neurobiology* 55 (2019), pp. 55–64. DOI: `10.1016/j.conb.2019.01.007`.

[45] Nikolaus Kriegeskorte. "Deep neural networks: a new framework for modeling biological vision and brain information processing". In: *Annual review of vision science* 1 (2015), pp. 417–446.

[46] James J Gibson. "The visual perception of objective motion and subjective movement." In: *Psychological review* 61.5 (1954), p. 304.

[47] Leslie G Ungerleider. "Two cortical visual systems". In: *Analysis of visual behavior* (1982), pp. 549–586.

[48] Melvyn A Goodale and A David Milner. "Separate visual pathways for perception and action". In: *Trends in neurosciences* 15.1 (1992), pp. 20–25.

[49] Arash Vahdat and Jan Kautz. "NVAE: A deep hierarchical variational autoencoder". In: *Advances in neural information processing systems* 33 (2020), pp. 19667–19679.

[50] Bruno A Olshausen et al. "Perception as an inference problem". In: *The cognitive neurosciences* (2014), p. 295. URL: `http://rctn.org/bruno/papers/perception-as-inference.pdf`.

[51] Yujia Huang et al. "Neural networks with recurrent generative feedback". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 535–545.

[52] Rewon Child. "Very deep vaes generalize autoregressive models and can outperform them on images". In: *arXiv preprint arXiv:2011.10650* (2020).

[53] Casper Kaae Sønderby et al. "Ladder variational autoencoders". In: *Advances in neural information processing systems* 29 (2016).

[54] Lars Maaløe et al. "Biva: A very deep hierarchy of latent variables for generative modeling". In: *Advances in neural information processing systems* 32 (2019).

[55] Louay Hazami et al. "Efficientvdvae: Less is more". In: *arXiv preprint arXiv:2203.13751* (2022).

[56] Shital Shah et al. "Airsim: High-fidelity visual and physical simulation for autonomous vehicles". In: *Field and Service Robotics: Results of the 11th International Conference*. Springer. 2018, pp. 621–635.

[57] Nicole C Rust and J Anthony Movshon. "In praise of artifice". In: *Nature neuroscience* 8.12 (2005), pp. 1647–1650.

[58] Bela Julesz. "Foundations of cyclopean perception." In: (1971).

[59] Tal Golan et al. "Controversial stimuli: Pitting neural networks against each other as models of human cognition". In: *Proceedings of the National Academy of Sciences* 117.47 (2020), pp. 29330–29337.

[60] Michael Beyeler et al. "3D visual response properties of MSTd emerge from an efficient, sparse population code". In: *Journal of Neuroscience* 36.32 (2016), pp. 8399–8415.

[61] James J Gibson. "The perception of the visual world." In: (1950).

[62] William H Warren Jr and Daniel J Hannon. "Direction of self-motion is perceived from optical flow". In: *Nature* 336.6195 (1988), pp. 162–163.

[63] Jonathan Samir Matthis et al. "Retinal optic flow during natural locomotion". In: *PLOS Computational Biology* 18.2 (2022), e1009575.

[64] Eddy Ilg et al. "Flownet 2.0: Evolution of optical flow estimation with deep networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2462–2470.

[65] Irina Higgins et al. "beta-vae: Learning basic visual concepts with a constrained variational framework". In: *International conference on learning representations*. 2017.

[66] Christopher P Burgess et al. "Understanding disentangling in $\beta-$VAE". In: *arXiv preprint arXiv:1804.03599* (2018).

[67] Nikolaus Kriegeskorte and Jörn Diedrichsen. "Peeling the onion of brain representations". In: *Annual review of neuroscience* 42 (2019), pp. 407–432.

[68] Simon K Rushton and Paul A Warren. "Moving observers, relative retinal motion and the detection of object movement". In: *Current Biology* 15.14 (2005), R542–R543.

[69] Paul A Warren and Simon K Rushton. "Optic flow processing for the assessment of object movement during ego movement". In: *Current Biology* 19.18 (2009), pp. 1555–1560.

[70] Paul A Warren and Simon K Rushton. "Perception of object trajectory: Parsing retinal motion into self and object movement components". In: *Journal of vision* 7.11 (2007), pp. 2–2.

[71] Nicole E Peltier et al. "Optic flow parsing in the macaque monkey". In: *Journal of vision* 20.10 (2020), pp. 8–8.

[72] James CR Whittington et al. "Disentanglement with biological constraints: A theory of functional cell types". In: *The Eleventh International Conference on Learning Representations*. 2023.

[73] Sebastien Lachapelle et al. "Synergies between Disentanglement and Sparsity: Generalization and Identifiability in Multi-Task Learning". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 18171–18206. URL: `https://proceedings.mlr.press/v202/lachapelle23a.html`.

[74] Abhishek Kumar et al. "VARIATIONAL INFERENCE OF DISENTANGLED LATENT CONCEPTS FROM UNLABELED OBSERVATIONS". In: *International Conference on Learning Representations*. 2018. URL: `https://openreview.net/forum?id=H1kG7GZAW`.

[75] Hyunjik Kim and Andriy Mnih. "Disentangling by factorising". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2649–2658.

[76] Ricky TQ Chen et al. "Isolating sources of disentanglement in variational autoencoders". In: *Advances in neural information processing systems* 31 (2018).

[77] Michal Rolinek et al. "Variational autoencoders pursue pca directions (by accident)". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12406–12415.

[78] Sjoerd van Steenkiste et al. "Are Disentangled Representations Helpful for Abstract Visual Reasoning?" In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: `https://proceedings.neurips.cc/paper_files/paper/2019/file/bc3c4a6331a8a9950945a1aa8c95ab8a-Paper.pdf`.

[79] Andrea Dittadi et al. "On the Transfer of Disentangled Representations in Realistic Settings". In: *International Conference on Learning Representations*. 2021. URL: `https://openreview.net/forum?id=8VXvj1QNRl1`.

[80] W Jeffrey Johnston and Stefano Fusi. "Abstract representations emerge naturally in neural networks trained to perform multiple tasks". In: *Nature Communications* 14.1 (2023), p. 1040.

[81] Marc-André Carbonneau et al. "Measuring disentanglement: A review of metrics". In: *IEEE transactions on neural networks and learning systems* (2022).

[82] Cian Eastwood et al. "DCI-ES: An Extended Disentanglement Framework with Connections to Identifiability". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: `https://openreview.net/forum?id=462z-gLgSht`.

[83] Yuwei Cui et al. "Diverse suppressive influences in area MT and selectivity to complex motion features". In: *Journal of Neuroscience* 33.42 (2013), pp. 16715–16728.

[84] Yuwei Cui et al. "Inferring cortical variability from local field potentials". In: *Journal of Neuroscience* 36.14 (2016), pp. 4121–4135.

[85] Maneesh Sahani and Jennifer Linden. "How linear are auditory cortical responses?" In: *Advances in neural information processing systems* 15 (2002).

[86] Patrick J Mineault et al. "Hierarchical processing of complex motion along the primate dorsal visual pathway". In: *Proceedings of the National Academy of Sciences* 109.16 (2012), E972–E980.

[87] Edward H Adelson and James R Bergen. "Spatiotemporal energy models for the perception of motion". In: *Josa a* 2.2 (1985), pp. 284–299.

[88] Shinji Nishimoto and Jack L Gallant. "A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies". In: *Journal of Neuroscience* 31.41 (2011), pp. 14551–14564.

[89] Karl S Muller et al. "Retinal motion statistics during natural locomotion". In: *Elife* 12 (2023), e82410.

[90] Benedict Wild and Stefan Treue. "Primate extrastriate cortical area MST: a gateway between sensation and cognition". In: *Journal of neurophysiology* 125.5 (2021), pp. 1851–1882.

[91] Yena Han et al. "System identification of neural systems: If we got it right, would we know?" In: *International Conference on Machine Learning*. PMLR. 2023, pp. 12430–12444.

[92] Alex H Williams et al. "Generalized shape metrics on neural representations". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 4738–4750.

[93] Lyndon R Duong et al. "Representational dissimilarity metric spaces for stochastic neural networks". In: *arXiv preprint arXiv:2211.11665* (2022).

[94] Max Klabunde et al. "Similarity of Neural Network Models: A Survey of Functional and Representational Measures". In: *arXiv preprint arXiv:2305.06329* (2023).

[95] Abdulkadir Canatar et al. *A Spectral Theory of Neural Prediction and Alignment*. 2023. arXiv: `2309.12821 [q-bio.NC]`.

[96] Andre M Bastos et al. "Canonical microcircuits for predictive coding". In: *Neuron* 76.4 (2012), pp. 695–711.

[97] Fabian A Mikulasch et al. "Where is the error? Hierarchical predictive coding through dendritic error computation". In: *Trends in Neurosciences* 46.1 (2023), pp. 45–59.

[98] Beren Millidge et al. *Predictive Coding: Towards a Future of Deep Learning beyond Backpropagation?* 2022. arXiv: `2202.09467 [cs.NE]`.

[99] David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

[100] Iris Vilares and Konrad Kording. "Bayesian models: the structure of the world, uncertainty, behavior, and the brain". In: *Annals of the New York Academy of Sciences* 1224.1 (2011), pp. 22–39.

[101] Sabyasachi Shivkumar et al. "A probabilistic population code based on neural samples". In: *Advances in neural information processing systems* 31 (2018).

[102] Ishaan Gulrajani et al. "PixelVAE: A Latent Variable Model for Natural Images". In: *International Conference on Learning Representations*. 2017. URL: `https://openreview.net/forum?id=BJKYvt5lg`.

[103] Ferenc Csikor et al. "Top-down effects in an early visual cortex inspired hierarchical Variational Autoencoder". In: *SVRHM 2022 Workshop @ NeurIPS*. 2022. URL: `https://openreview.net/forum?id=8dfboOQfYt3`.

[104] Eleni Miliotou et al. "Generative Decoding of Visual Stimuli". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 24775–24784. URL: https://proceedings.mlr.press/v202/miliotou23a.html.

[105] Gregory C. DeAngelis and Dora E. Angelaki. "Visual–Vestibular Integration for Self-Motion Perception". In: *The Neural Bases of Multisensory Processes* (2012), pp. 629–644.

[106] Eduard Von Holst. "Relations between the central nervous system and the peripheral organs." In: *British Journal of Animal Behaviour* (1954).

[107] Paul R MacNeilage et al. "Vestibular facilitation of optic flow parsing". In: *PLoS One* 7.7 (2012), e40264.

[108] Kathleen E Cullen and Omid A Zobeiri. "Proprioception and the predictive sensing of active self-motion". In: *Current opinion in physiology* 20 (2021), pp. 29–38.

[109] Constance S Royden and Ellen C Hildreth. "Human heading judgments in the presence of moving objects". In: *Perception & psychophysics* 58 (1996), pp. 836–856.

[110] William H Warren Jr and Jeffrey A Saunders. "Perceiving heading in the presence of moving objects". In: *Perception* 24.3 (1995), pp. 315–331.

[111] Edward AB Horrocks et al. "Walking humans and running mice: perception and neural encoding of optic flow during self-motion". In: *Philosophical Transactions of the Royal Society B* 378.1869 (2023), p. 20210450.

[112] Denis N Lee. "The optic flow field: The foundation of vision". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 290.1038 (1980), pp. 169–179.

[113] Markus Lappe et al. "Perception of self-motion from visual flow". In: *Trends in cognitive sciences* 3.9 (1999), pp. 329–336.

[114] Jean-Paul Noel et al. "Causal inference during closed-loop navigation: parsing of self-and object-motion". In: *Philosophical Transactions of the Royal Society B* 378.1886 (2023), p. 20220344.

[115] David Ha and Jürgen Schmidhuber. "World models". In: *arXiv preprint arXiv:1803.10122* (2018).

[116] Christopher R Fetsch et al. "Dynamic reweighting of visual and vestibular cues during self-motion perception". In: *Journal of Neuroscience* 29.49 (2009), pp. 15601–15612.

[117] John S Butler et al. "Bayesian integration of visual and vestibular signals for heading". In: *Journal of vision* 10.11 (2010), pp. 23–23.

[118] Mattia Rigotti et al. "The importance of mixed selectivity in complex cognitive tasks". In: *Nature* 497.7451 (2013), pp. 585–590.

[119] Irina Higgins et al. "Symmetry-based representations for artificial and biological general intelligence". In: *Frontiers in Computational Neuroscience* 16 (2022), p. 836498. DOI: 10.3389/fncom.2022.836498.

[120] Fabio Anselmi et al. "On invariance and selectivity in representation learning". In: *Information and Inference: A Journal of the IMA* 5.2 (2016), pp. 134–158.

[121] Michael M Bronstein et al. "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges". In: *arXiv preprint arXiv:2104.13478* (2021).

[122] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[123] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 1988.

[124] Jie Hu et al. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[125] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.

[126] Tim Salimans and Durk P Kingma. "Weight normalization: A simple reparameterization to accelerate training of deep neural networks". In: *Advances in neural information processing systems* 29 (2016).

[127] Prajit Ramachandran et al. "Searching for activation functions". In: *arXiv preprint arXiv:1710.05941* (2017).

[128] Stefan Elfwing et al. “Sigmoid-weighted linear units for neural network function approximation in reinforcement learning”. In: *Neural Networks* 107 (2018), pp. 3–11.

[129] Samuel R Bowman et al. “Generating sentences from a continuous space”. In: *arXiv preprint arXiv:1511.06349* (2015).

[130] Hao Fu et al. “Cyclical annealing schedule: A simple approach to mitigating kl vanishing”. In: *arXiv preprint arXiv:1903.10145* (2019).

[131] Arash Vahdat et al. “Dvae++: Discrete variational autoencoders with overlapping transformations”. In: *International conference on machine learning*. PMLR. 2018, pp. 5035–5044.

[132] Xi Chen et al. “Variational lossy autoencoder”. In: *arXiv preprint arXiv:1611.02731* (2016).

[133] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[134] Ilya Loshchilov and Frank Hutter. “Sgdr: Stochastic gradient descent with warm restarts”. In: *arXiv preprint arXiv:1608.03983* (2016).

Supplementary material for:

# Hierarchical VAEs provide a normative account of motion processing in the primate brain

## 7 Background & Related Work (extended)

**Neuroscience and VAEs.** The centuries-old notion of "perception as unconscious inference" from Helmholtz (1860s, [1]) and similar ideas from Alhazen (1011–1021 AD, [2]), have been influential in theoretical/cognitive neuroscience. A recent extension of these ideas stemming from the hierarchical organization of the sensory cortex, and in particular, the larger volume of feedback connections relative to feedforward connections, led Mumford [3] to put forth a bold conjecture: that the cortex engages in hierarchical Bayesian inference to comprehend the world.

There has been much influential work following different aspects of how this idea connects to neuroscience, including predictive coding [4, 9, 96–98], the Bayesian Brain hypothesis [6, 99–101], Analysis-by-Synthesis [7], and the Free Energy Principle [8]. Notably, these ideas are closely paralleled by VAEs whose loss function can be derived from the requirement of effective posterior inference—a congruence that resonates well with Helmholtz's original conjecture.

The deep connections between VAEs and neuroscience is explored in a review by Marino [17], but work looking at the synergy between neuroscience and VAEs has thus far been limited to the following recent papers.

Higgins et al. [16] trained $\beta$-VAE [65] models on face images and evaluated the models on predicting responses of individual inferotemporal neurons from Macaque face patch. They found that $\beta$-VAE discovers individual latents that are aligned with IT neurons and that this alignment linearly increases with increased beta values.

Storrs et al. [15] trained PixelVAE [102] models on synthesized images of glossy surfaces and found that the model spontaneously learned to cluster images according to underlying physical factors like material reflectance and illumination, despite receiving no explicit supervision. Most notably, PixelVAE mimicked human perceptual judgment of gloss in that it made errors that were similar to those made by humans.

Csikor et al. [103] investigated the properties of representations and inference in a biologically inspired hierarchical VAE called Topdown VAE that is capable of reproducing some key properties of representations emerging in V1 and V2 of the visual cortex. The authors showed that architectures that feature top-down influences in their recognition model give rise to a richer representation, such that specific information that is not present in mean activations becomes linearly decodable from posterior correlations.

Finally, Miliotou et al. [104] learned mapping from the latent space of a hierarchical VAE to fMRI voxel space. This allowed the use of the trained decoder to reconstruct images directly from brain responses. Importantly, they perform ablation experiments and find hierarchy is an essential component of their models.

**Hierarchical VAEs.** An early influential work that introduced hierarchy to the latent space of VAEs was Ladder VAE (LVAE, Sønderby et al. [53]), which improved upon standard VAEs by sharing information between the inference and generative networks, allowing LVAEs to learn deeper and more distributed latent representations compared to regular VAEs. Building on LVAE, Maaløe et al. [54] introduced Bidirectional-Inference Variational Autoencoder (BIVA), a model that uses skip-connections to further enhance the flow of information among latent variables and better avoids inactive units. Both the LVAE and BIVA made valuable contributions towards enabling VAEs to effectively leverage deep stochastic hierarchies, through innovations in the inference model design and its relationship to the generative network.

More recent variants further improved VAE performance. The Nouveau VAE (NVAE, [49]), is a VAE equipped with a hierarchical architecture that achieved state of the art in several benchmarks and generated high-quality faces. Very deep VAE (vdvae, [52]) achieved a similarly impressive performance on complex image benchmarks. Additionally, vdvae paper demonstrated that hierarchical VAEs perform well because they are deep in a stochastic sense (i.e., several stages of sampling vs. just once for the vanilla VAE). This key conceptual advance provided an explanation of why non-hierarchical VAEs previously struggled to compete against autoregressive models. In sum, both NVAE and vdvae achieved impressive performance in difficult benchmarks, while a central motivation of NVAE was to improve the quality of generated samples. However, neither work focused on evaluating how the hierarchical latent structure alters and potentially enhances the quality of learned representations in downstream applications.

Despite their success, NVAE and vdvae have an undesirable property: the convolutional nature of their latents results in an overly redundant latent space that is several orders of magnitude larger than the number of variables in the input space. This is at odds with the principle of "parsimony" (promoting compact and efficient representations) and defeats one of the main purposes of the autoencoder family. Hazami et al. [55] showed that for vdvae, only a tiny subset $(\sim 3\%)$ of its latent space is necessary, and the rest is there for marginal improvements on log-likelihood performance. Here, we demonstrated that it is possible to compress hierarchical VAEs and focus on investigating the functionality of individual latents. As far as we know, ours is the first study focused on the evaluation of representations in hierarchical VAEs with applications to neuroscience data.

# 8 Supplementary Discussion

## 8.1 Visual-only cues and unsupervised learning are sufficient for untangling optic flow causes

Gibson [61] was the first to recognize how an observer's movement through the environment results in certain radial expansion patterns, which he called optic flow. Furthermore, he contemplated the intricate interplay between translational and rotational self-motion components [61, 62], and their impact on resulting optic flows. Gibson then went on to propose that the translational and rotational components of the optic flow can be decomposed visually, although he did not elaborate on how this could be accomplished [46, 61]. Other theories have proposed that, to estimate self-motion, the brain relies on a combination of visual and non-visual cues [105], such as efference copies of motor commands [106], vestibular inputs [107], and proprioception [108].

Further complications arise when independently moving objects are added to the mix [46, 109, 110]. How can the brain understand the motion of an object when self-motion is non-trivially affecting the entire flow field [111]? This is a critical motion processing problem [112], manifested in diverse ecological scenarios such as chasing prey, crossing the street, or playing soccer. Are retinal visual-only cues sufficient to extract information about different self-motion components, as well as object motion? Vision scientists have debated this for decades, and still, there is no clear answer (see [113] for a review).

Notably, this ongoing debate underscores a fundamental divergence between our approach and that of Mineault et al. [34]. Namely, our work is based on the premise that retinal-only signals contain sufficient information about their ground truth causes and that these factors can be extracted using unsupervised learning alone. In contrast, Mineault et al. [34] trained their DorsalNet by providing dense supervision on self-motion components, arguing that such supervised training is biologically plausible since these signals are approximately available to the observer from vestibular inputs and other such extra-retinal sources.

Our results in Fig. 4 provide proof of concept that it is possible to untangle ground truth causes of retinal optic flow in a completely unsupervised manner, using the raw optic flow patterns alone. Comparing the performance of alternative models revealed two essential ingredients for the success of our approach: (1) a properly designed architecture—hierarchical, like the sensory cortex [3]; and, (2) an appropriate loss function—inference, [1].

## 8.2 ROFL in an active, multimodal setting

A recent study by Noel et al. [114] examined the role of causal inference in attributing retinal motion to self and object motion during closed-loop navigation. The study reports that in the active self-

motion condition, self-motion is estimated using both optic flow and an efference copy of the motor commands from hands controlling a joystick that drove virtual velocity. In contrast, during passive self-motion, there was no efference copy, which led to more uncertain self-motion estimates and greater misattribution between self and object motion.

In our own study, we found that cNVAE was particularly successful in untangling fixation and self-motion-related variables, almost reaching the ceiling value of $R^2 \sim 1.0$. However, it did not reach this maximum value for the object-related variables (Fig. 4). Based on these observations and the findings of Noel et al. [114], we speculate that addition of some extra-retinal signal as weak supervision in an active setting would help cNVAE learn even better representations of its retinal inputs. Training hierarchical VAEs in active settings [115], using multimodal signals [105], while characterizing the interplay between different sensory modalities [116, 117] is an exciting topic for future work.

### 8.3 Comparing cNVAE to NVAE, its non-compressed counterpart

In Fig. 4, we also investigated the performance of the original non-compressed NVAE with an approximately matched latent dimensionality ($440$, slightly larger than $420$ for cNVAE). We found that although NVAE did better than non-hierarchical VAE, it still underperformed cNVAE. This is because trying to match the latent dimensionality of NVAE with cNVAE necessitates reducing the number of hierarchical latent groups in the NVAE since most of its convolutional latents capture spatial dimensions rather than abstract features like cNVAE. From previous work [52], we know that the "stochastic depth" of hierarchical VAEs is the key reason behind their effectiveness; therefore, it was expected that reduced depth was going to hurt an NVAE with a matched number of latents.

It is worth noting that cNVAE and NVAE had a balanced performance across all ground truth variables (Fig. 4), in that they captured both global (self-motion-related) and local (object-related) aspects well. In contrast, non-hierarchical VAE completely ignored object-related variables and focused solely on the global scale determined by fixation point and self-motion. This suggests that hierarchical architecture is crucial for a more comprehensive understanding of multi-scale datasets such as ROFL.

### 8.4 Disentanglement is in the eyes of the beholder

The conventional definition of disentanglement in machine learning relies on predefining data-generative factors, a choice that is often guided by heuristics. This subjective aspect raises fundamental questions about the feasibility and value of persistently pursuing disentangled representations [19, 118], particularly when framed in terms of the ambiguous concept of "*data generative factors*" [18].

Throughout our experiments, we observed that cNVAE robustly identified one or more latent variables that exhibited equally high mutual information with all components of self-motion velocity in Cartesian coordinates. Latent dimensions 154 and 122 in Fig. 3 exemplify this phenomenon. To elucidate the function of these latents, we transformed the velocity vectors into polar coordinates and found that they solely encoded the magnitude of self-motion, irrespective of its direction (Fig. S1). These seemingly highly entangled latent variables would have achieved almost perfect disentanglement scores had we employed polar coordinates instead of Cartesian coordinates to represent the ground-truth velocities ($d_i \approx 0.38 \rightarrow d_i \approx 0.81$).

Our findings shed light on the fact that disentanglement cannot be considered an inherent property of objective reality; rather, it is significantly influenced by our a priori determination of independent factors of variation within the data. Given the inherent complexity and richness of natural datasets, it becomes challenging to unequivocally determine the true factors of variation a priori, pointing to the need for a more systematic approach.

An alternative and potentially more promising approach lies in focusing on the symmetry transformations inherent in the world [18, 119]. By seeking representations that respect the underlying geometrical and group-theoretical structures of the world [120, 121], and comparing them in principled ways [92–94], we may uncover new perspectives that provide valuable insights beyond the conventional and subjective notions of disentanglement.

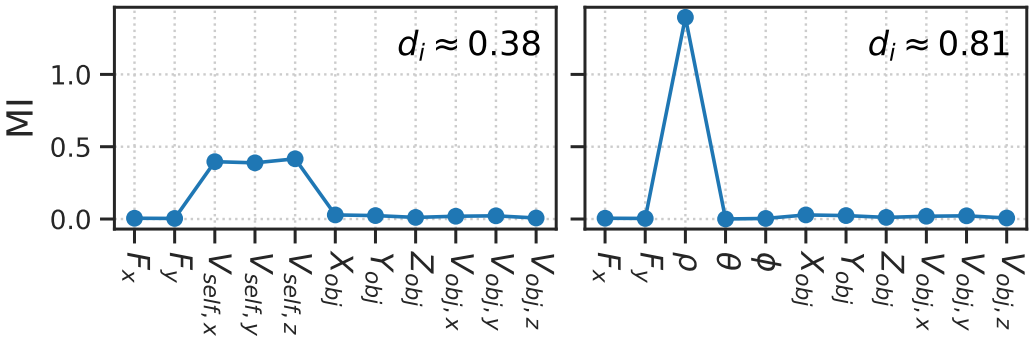


Figure S1: This latent variable cares about the magnitude of self-motion only. Its disentanglement score depends strongly on what we choose a priori as our independent factors of variation in data.

## 9 Further evidence for the superiority of cNVAE over VAE

In this section, we present additional results demonstrating the advantages of hierarchical VAEs versus non-hierarchical ones. All models explored in this section were trained on `fixate-1`. For both cNVAE and VAE, we chose $\beta$ values that maximized their informativeness (see Fig. 5). Specifically, we set $\beta = 0.15$ for cNVAE and $\beta = 1.5$ for VAE. These same models trained using the respective $\beta$ values were also used for results presented in Fig. 3 and Fig. 4 in the main text.

### 9.1 Individual cNVAE latents exhibit greater linear correlation with ground truth factors

In Fig. 4 we asked whether a simple linear regression model can extract ground truth factors $\mathbf{g}$, from model latent codes, $\mathbf{z}$, which are 420-dimensional vectors. Here, we ask whether there are individual latents $z_i$ that are highly (Pearson) correlated with individual factors $g_j$. The results are shown in Fig. S2. For some ground truth variables, we find that cNVAE learns latents that are almost perfectly correlated with them. For instance, $r = -0.91$ for $F_x$ and $r = 0.87$ for $F_y$, compared to modest correlation values of $r = 0.34$ and $r = 0.38$ for VAE. The higher linear correlation is consistently observed for cNVAE over VAE, especially we note that the VAE does not capture object-related variables at all, consistent with other results in the main paper (Fig. 3 and Fig. 4).

For the cNVAE, the indices for the selected highly correlated latents are as follows: $[296, 248, 393, 284, 368, 92, 206, 105, 338, 60, 63]$, ordered left-to-right in the same manner as in Fig. S2. Here, an index of 0 corresponds to the very top latent operating at the spatial scale of $2 \times 2$, and an index of 419 corresponds to the very bottom one at scale $8 \times 8$ (see Fig. 2). By considering the hierarchical grouping of these latents (see dashed lines in Fig. 3) we see that the majority of the highly correlated latents operate at the spatial scale of 8, with the exception of object-related ones corresponding to $X_{obj}$, $Z_{obj}$, $V_{obj,y}$, and $V_{obj,z}$ which operate at $4 \times 4$.

In conclusion, Fig. S2 revealed that cNVAE learns individual latents that are highly correlated with individual ground truth factors, which provides complementary insights as to why cNVAE is so successful in untangling (Fig. 4). However, we note that a high correlation with individual latents is not the best way to capture the overall functional organization, which is better reflected for example in the mutual information heatmap of Fig. 3.

### 9.2 Generated samples: non-hierarchical VAE completely neglects objects

We examine generated samples from both the cNVAE and VAE models, as illustrated in Fig. S3 and Fig. S4, respectively. For comparison, we also include more real samples randomly drawn from the `fixate-1` dataset in Fig. S5. The cNVAE samples appear indistinguishable from true samples; however, VAE fails to capture and generate objects, corroborating previous results (see Fig. 3, Fig. 4, and Fig. S2). This striking disparity indicates that a hierarchical latent space is necessary for effectively modeling datasets with a multi-scale hierarchical structure.

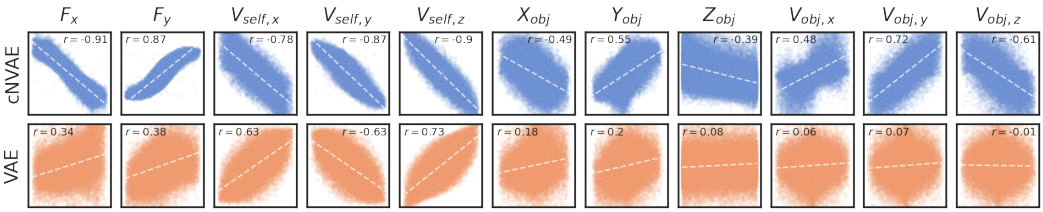


Figure S2: For each ground truth variable and model, we identified a single latent variable that displayed the highest Pearson correlation with that variable. The ground truth factors exhibited greater linearity within the latent space of cNVAE compared to VAE.

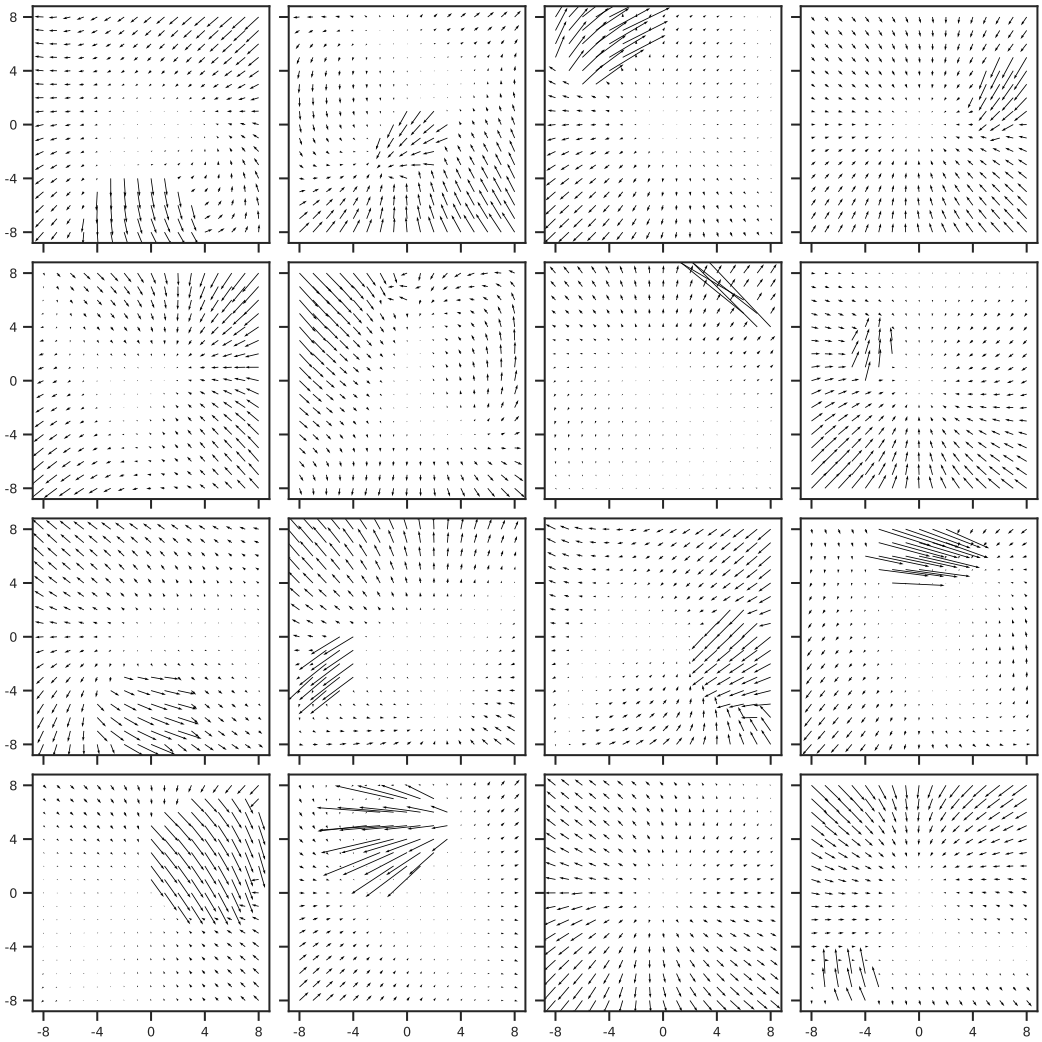


Figure S3: Generated samples from cNVAE closely resemble real data (cf. Fig. S5). Samples were generated at a temperature of $t = 1$ without cherry-picking.

### 9.3 Latent traversal reveals multi-scale understanding in cNVAE

Figure S6 provides visual examples of latent traversal in cNVAE, focusing on a latent dimension that effectively captures $Y_{obj}$. These examples demonstrate intriguing interactions between object velocity patterns and self-motion within the scene. Notably, in some examples such as rows 0 and 6, relocating the object from the bottom to the top of the scene induces a significant change in its

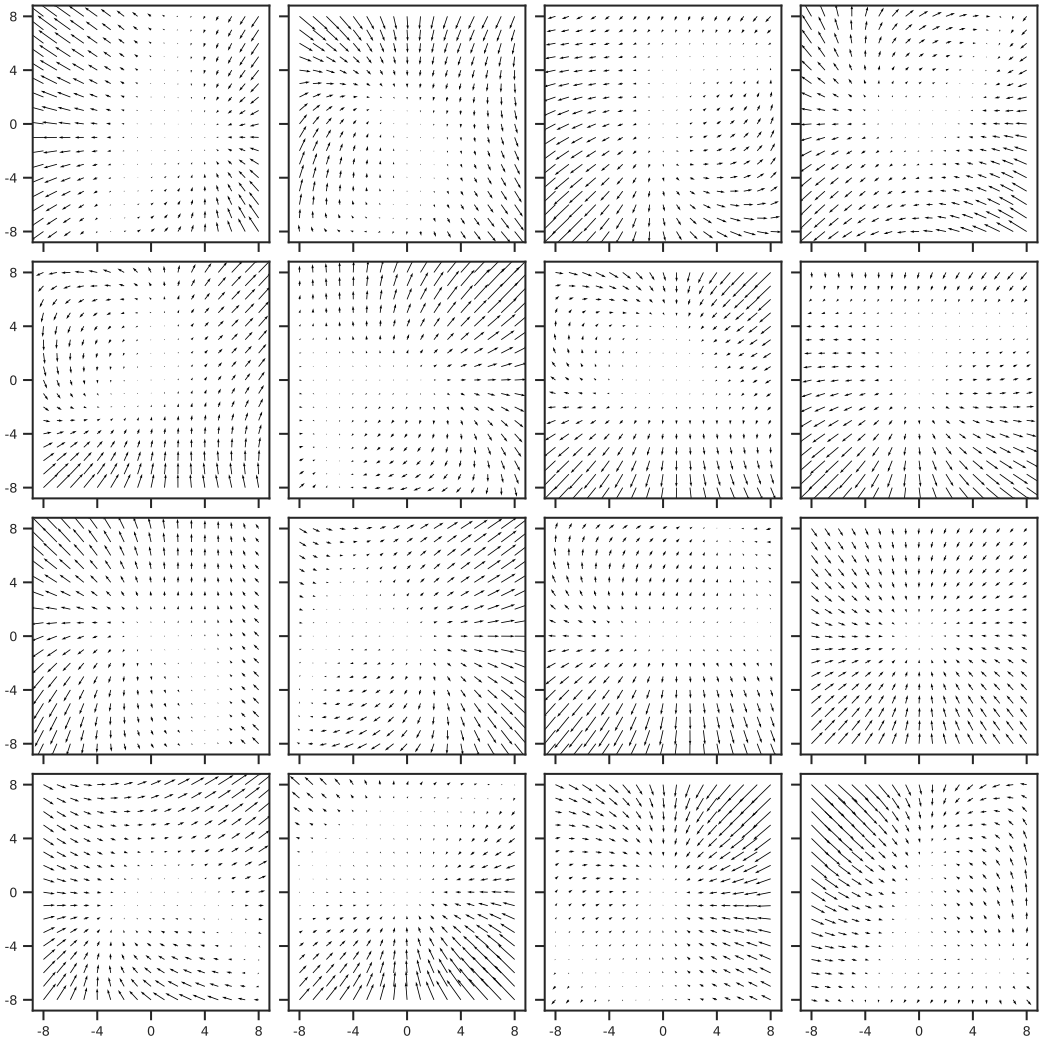


Figure S4: Generated samples from VAE illustrate the failure of non-hierarchical architecture in understanding object motion, further supported by Fig. 3 and Fig. 4 in the main text.

linear velocity along the y-axis. This phenomenon arises due to the influence of self-motion, as seen in the background. These illustrations collectively showcase that flow vectors in the visual scene are shaped both by local variables, such as the object's physical location, and their interactions with global-influencing variables like self-motion. By comprehending data across scales, cNVAE is able to capture such hierarchical and multi-scale properties inherent in naturalistic optic flow patterns. In contrast, non-hierarchical VAEs exhibit a complete failure in this regard.

### 9.4 Robust reconstruction performance across all models

Next, we shift our focus to evaluating the models' performance on held-out data (details provided below). Figure S7 illustrates the dependence of reconstruction loss (EPEPD, endpoint error per dim) and negative evidence lower bound (NELBO) for different values of $\beta$ and model types. Both hierarchical and non-hierarchical VAE models perform well, with cNVAE exhibiting slightly superior performance for $\beta < 0.8$.

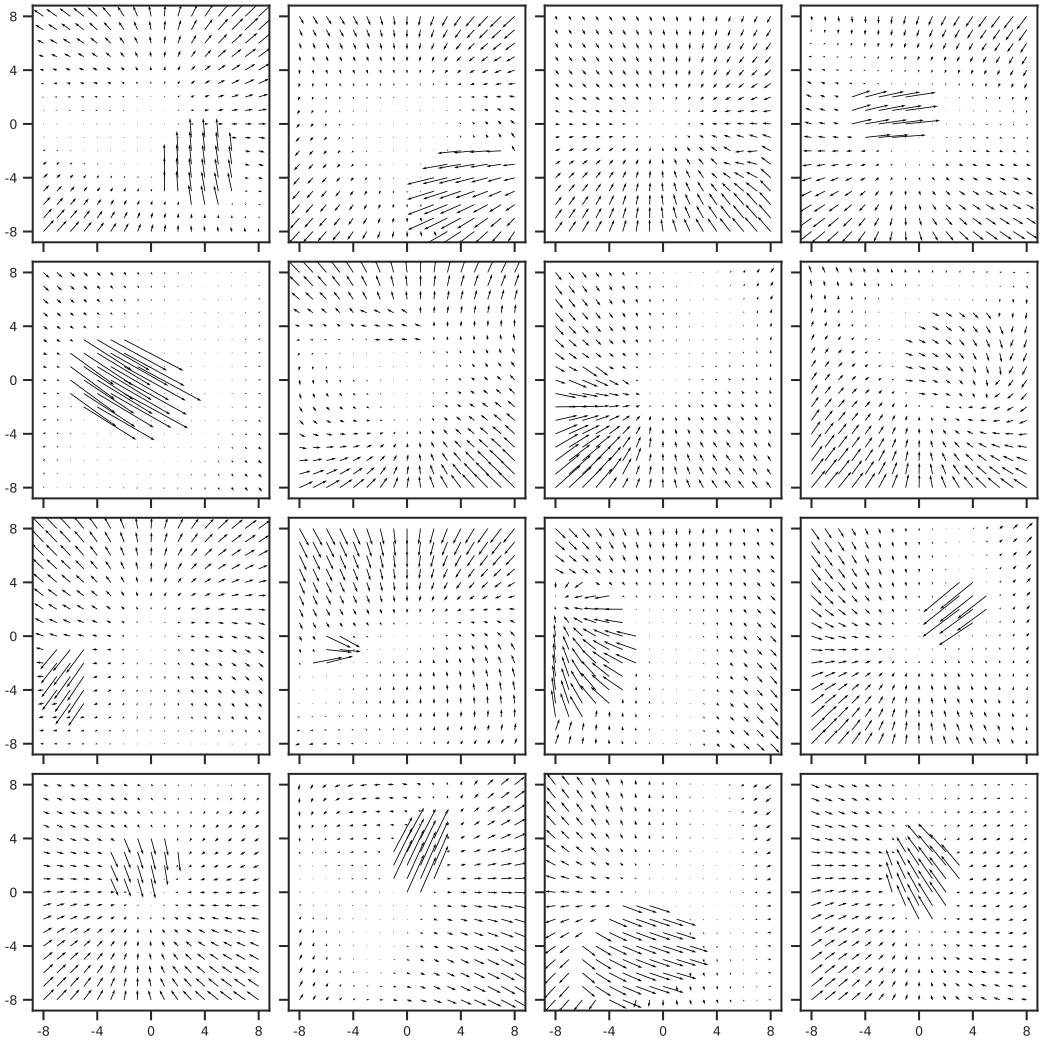


Figure S5: Additional examples showcasing the `fixate-1` dataset. Compare to Fig. S3 for cNVAE-generated samples and Fig. S4 for VAE-generated samples.

### 9.5 Untangling and reconstruction loss are strongly anti-correlated for cNVAE but not VAE

A desirable characteristic of a representation is its ability to encode ground-truth variables in a linearly decodable manner, commonly referred to as "*untangling*" within the neuroscience community [20, 21]. In the field of machine learning, similar concepts have been explored under terms such as "*informativeness*" [22] and "*explicitness*" [18, 23]. Evaluating the linear decodability or informativeness requires access to ground truth labels [22], which poses limitations when working with real-world datasets that typically lack such labels.

To address this, we investigated the relationship between informativeness and reconstruction loss, as well as NELBO, depicted in Fig. S8. Strikingly, we observe a significant and strong negative linear relationship between reconstruction loss and informativeness in cNVAE ($r = -0.91, p = 8.9 \times 10^{-7}$, $t-$test). Conversely, no such correlation is observed for VAE ($r = 0.10, p = 0.72$, $t-$test). Notably, this relationship is absent for both models when comparing informativeness with NELBO (Fig. S8). These findings suggest that reconstruction loss—exclusively in the case of cNVAE—can serve as a viable proxy for informativeness, even in the absence of ground-truth variables in real-world datasets.

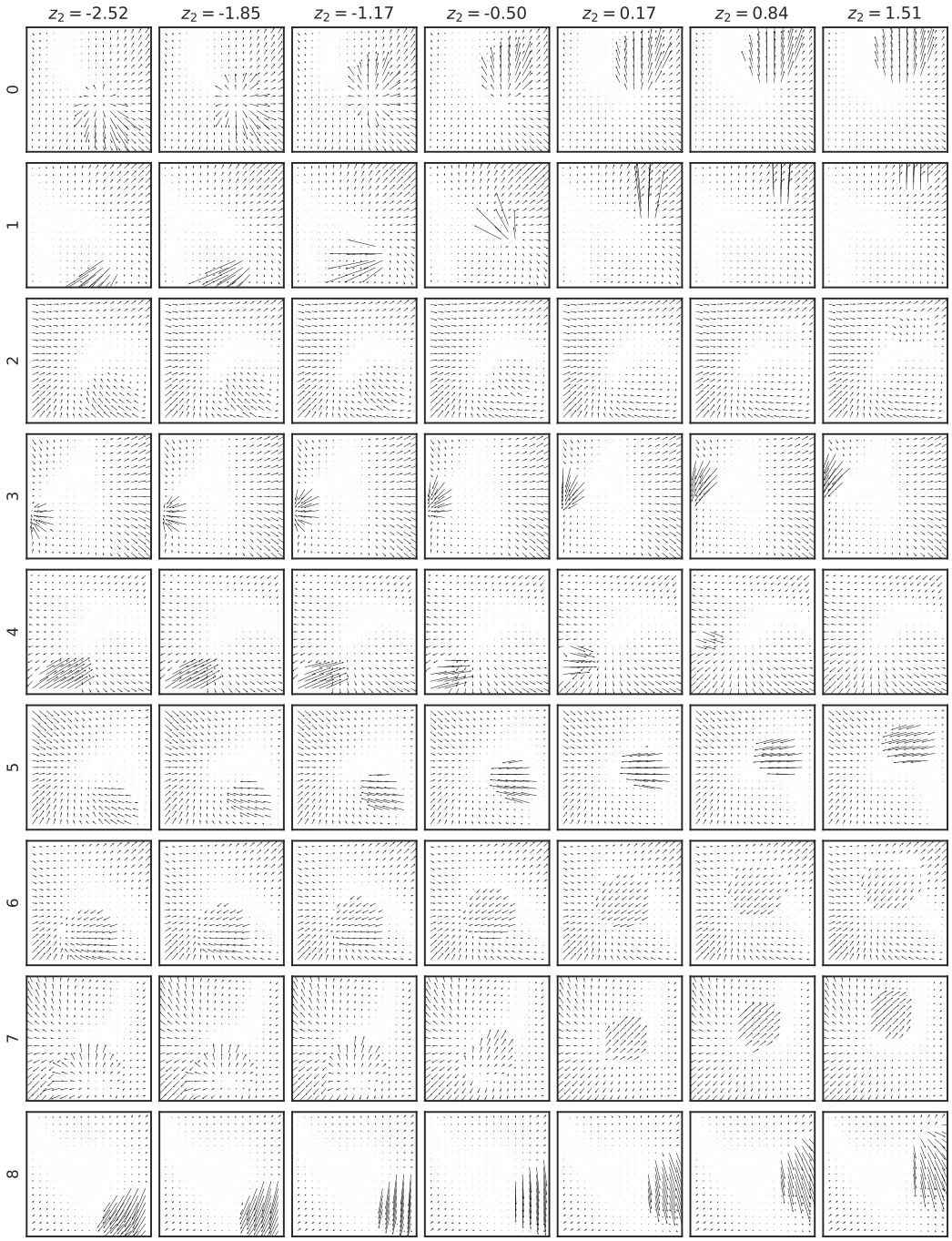


Figure S6: Latent traversal performed using a latent variable that effectively captures $Y_{obj}$. This variable belongs to a latent group operating at the spatial scale of $2 \times 2$ (see Fig. 3).

## 10 Brain alignment requires more than just linear regression performance

In the field of neuroscience, linear regression $R^2$ scores have commonly been employed to measure alignment between biological and artificial neural networks [30, 31, 34, 36]. However, we found that relying solely on regression performance provided a weak signal for model selection, which is consistent with recent work [36–38, 91, 92]. In contrast, our alignment score defined in Fig. 8a provided a much stronger signal for differentiating between models (compare Fig. 7 and Fig. 8c).

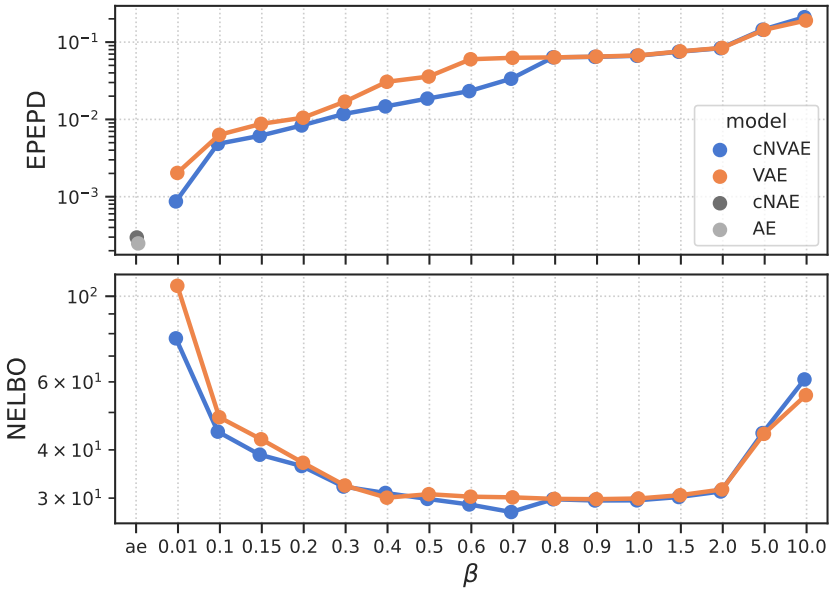


Figure S7: Both models demonstrate robust reconstruction performance, with cNVAE exhibiting a slight improvement. EPEPD, endpoint error per dim (lower is better). NELBO, negative evidence lower bound (lower is better).

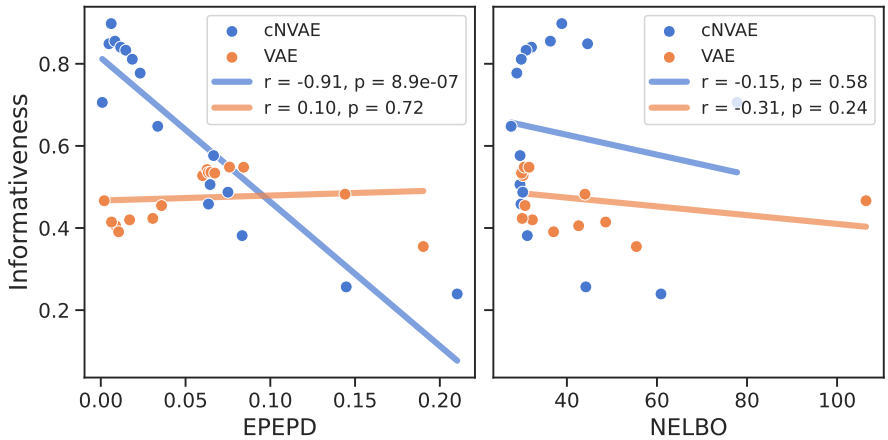


Figure S8: Strong anti-correlation between reconstruction loss and informativeness (i.e., untangling) observed only for cNVAE. Each scatter point corresponds to a model trained with a different $\beta$ value. EPEPD, endpoint error per dim (lower is better). NELBO, negative evidence lower bound (lower is better).

To investigate this further, we conducted statistical tests, comparing pairs of models based on either their performance in explaining the variance of MT neurons or their brain-alignment scores. The results are presented in Table S1. We observed that ridge regression performance was not particularly effective in distinguishing between models. We measured the effect sizes between distributions of $N = 141$ neurons, measuring how well ridge regression performance (Fig. 7) or alignment scores (Fig. 8c) discriminated between models. We found that the effect sizes were substantially larger for the alignment scores (Fig. S9). These findings suggest that while model performance is necessary, it is not sufficient for accurately measuring alignment between artificial and biological representations. In conclusion, refined and more comprehensive approaches are required to properly capture brain alignment [92–95].

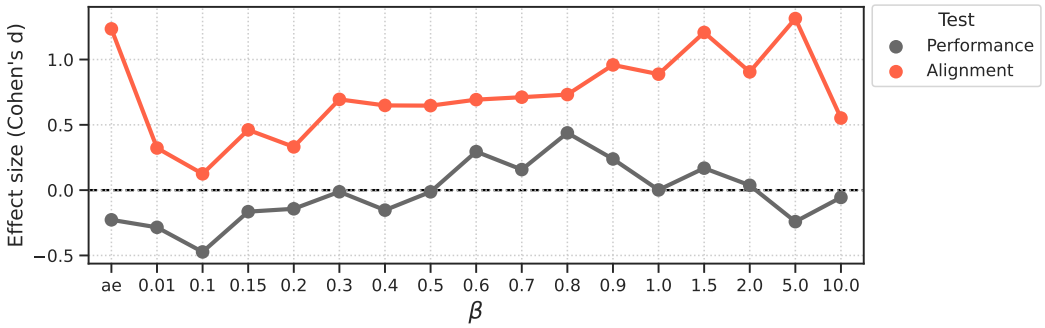


Figure S9: Illustrating the effect sizes for the two different approaches in model comparisons: using ridge regression *Performance* (related to Fig. 7), versus, *Alignment* scores (related to Fig. 8c). Alignment demonstrates a substantially larger effect size, leading to a more pronounced differentiation among the models. Positive values indicate cNVAE > VAE (or cNAE > AE). For different $\beta$ values, *Performance* fluctuates between positive and negative values; whereas, *Alignment* is consistently positive, indicating a greater brain alignment for hierarchical models.

| $\beta$ | Performance (Figure 7) | | | Alignment scores (Figure 8c) | | |
|---|---|---|---|---|---|---|
| | effect size | $p-$value | significant? | effect size | $p-$value | significant? |
| ae | -0.23 | 0.01 | ✓ | 1.2 | 8e-29 | ✓ |
| 0.01 | -0.28 | 0.002 | ✓ | 0.32 | 0.0004 | ✓ |
| 0.1 | -0.47 | 2e-07 | ✓ | 0.12 | 0.2 | ✗ |
| 0.15 | -0.16 | 0.07 | ✗ | 0.46 | 4e-07 | ✓ |
| 0.2 | -0.14 | 0.1 | ✗ | 0.33 | 0.0003 | ✓ |
| 0.3 | -0.012 | 0.9 | ✗ | 0.69 | 4e-13 | ✓ |
| 0.4 | -0.15 | 0.09 | ✗ | 0.65 | 7e-12 | ✓ |
| 0.5 | -0.013 | 0.9 | ✗ | 0.65 | 7e-12 | ✓ |
| 0.6 | 0.29 | 0.001 | ✓ | 0.69 | 4e-13 | ✓ |
| 0.7 | 0.16 | 0.08 | ✗ | 0.71 | 1e-13 | ✓ |
| 0.8 | 0.44 | 1e-06 | ✓ | 0.73 | 4e-14 | ✓ |
| 0.9 | 0.24 | 0.008 | ✓ | 0.96 | 1e-20 | ✓ |
| 1.0 | 0.0016 | 1 | ✗ | 0.89 | 1e-18 | ✓ |
| 1.5 | 0.17 | 0.07 | ✗ | 1.2 | 3e-28 | ✓ |
| 2.0 | 0.037 | 0.7 | ✗ | 0.91 | 3e-19 | ✓ |
| 5.0 | -0.24 | 0.008 | ✓ | 1.3 | 7e-31 | ✓ |
| 10.0 | -0.056 | 0.6 | ✗ | 0.55 | 3e-09 | ✓ |

Table S1: Hierarchical architectures are significantly more aligned with MT neurons compared to non-hierarchical ones. Paired $t-$tests were performed, and $p-$values were adjusted for multiple comparisons using the Benjamini-Hochberg correction method [122]. Effect sizes were estimated using Cohen's $d$ [123], where positive values indicate cNVAE > VAE (or cNAE > AE, first row).

## 11 Architecture and training details

Our model architecture and training procedure are largely inspired by the work of Vahdat and Kautz [49] who originally introduced NVAE, with some important differences. In this section, we summarize the key differences between our work and NVAE.

### 11.1 Key architectural differences between NVAE and cNVAE

Our main contribution to architecture design lies in the modification of the NVAE "*sampler*" layer and the introduction of the "*expand*" layer, which are simple deconvolutions. In the NVAE, the purpose of a sampler layer is to map encoder features to mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$ vectors that are then used to construct approximate posterior distributions $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$. In the NVAE, all sampler layers are convolutional with a fixed kernel size of $3 \times 3$ and padding of 1. Thus, the application of a sampler layer does not alter the spatial dimension of inputs, resulting in a convolutional latent space. In contrast, for the cNVAE, we chose different kernel sizes for sampler layers from different parts of the network. Specifically, we chose the sampler kernel sizes according to the spatial scale of each hierarchical latent group. For example, if a latent group operates at the scale of $s$, then the sampler layer will have a kernel size of $s \times s$, effectively integrating over space before the sampling step. Similarly, the deconvolution in an expand layer at the scale of $s$ has a kernel size of $s \times s$, which projects latents back to the convolution space.

Our approach has pros and cons. It achieved a much smaller latent space compared to NVAE, which allowed us to examine the function of individual latents, decoupled from their spatial influence. However, one drawback of our approach is that it makes the architecture input-size-dependent. Scaling up the input increases the parameter count of cNVAE. This did not pose serious problems in the present work, as we generated ROFL datasets in a relatively small spatial scale of $17 \times 17$. However, this could be a problem for larger datasets. To mitigate this, one possibility is to use depthwise separable convolutions for sampler and expand layers.

**Other details.** Because of the smaller scale of our experiments, we were able to use the same regular residual cells for both the encoder and the decoder, without the need to use depthwise separable convolutions (Figure 3b in Vahdat and Kautz [49]). Similar to NVAE, we found that the inclusion of squeeze & excitation [124] helped, but contrary to the suggestion of NVAE [49] and others [53], batch norm [125] destabilized training in our case. Instead, we found that weight normalization (WN) [126] and swish activation [127, 128] were both instrumental in training. We used WN without data-dependent initialization. Furthermore, we could obtain good results without the need for spectral regularization, perhaps because our dataset is synthetic and has a much smaller dimensionality compared to real-world datasets such as faces. Finally, similar to NVAE, we also found that residual Normal parameterization of the approximate posterior was helpful.

**Model hyperparameters.** It is known that model depth, in a statistical sense, is responsible for the success of hierarchical VAEs [52]. We increased the depth of our architecture while making sure that it was still stable to train. Thus, we ended up with a model that had 21 hierarchical latent groups: 12 groups operating at the scale of $8 \times 8$, 6 groups at the scale of $4 \times 4$, and 3 groups at the scale of $2 \times 2$ (see Fig. 2 and Fig. 3). Each latent group has 20 latents, which yields a $21 \times 20 = 420$ dimensional latent space. We observed that various aspects of model performance such as reconstruction loss and DCI scores increased with depth, corroborating previous work [52]. We set the initial number of channels to 32 in the first layer of the encoder and doubled the channel size every time we downsample the features spatially. This resulted in a width of 256 in the final layer. Please refer to our code for more information.

### 11.2 Model training details

We generated $750k$ samples for each ROFL category, with a $600k/75k/75k$ split for train/validation/test datasets. Model performance (reconstruction loss, NELBO) was evaluated using the validation set (Fig. S7 and S8). Similarly, we used the validation set to estimate mutual information between latents, $\mathbf{z}$, and ground-truth variables, $\mathbf{g}$, as shown in Fig. 3. To evaluate the latent code, we used the validation set to train linear regressors to predict $\mathbf{g}$ from $\mathbf{z}$. We then test the performance using the test set. This includes results presented in Fig. 4 and Fig. 5.

**KL annealing and balancing.** We annealed the KL term during the first half of the training, which is known to be an effective trick in training VAEs [49, 53, 129, 130]. Additionally, during the annealing period, we employed a KL balancing mechanism implemented by Vahdat and Kautz [49] which ensures an equal amount of information is encoded in each hierarchical latent group [49, 131, 132].

**Gradient norm clipping.** During training, we occasionally encountered updates with large gradient norms. Similar to the work of Child [52], we used gradient clipping, with an empirically determined clip value of $250$. However, we did not skip the updates with large gradients, as we found that the model usually gets stuck and does not recover after skipping an update.

**Training hyperparameters.** We used batch size of $600$, learning rate of $0.002$, and trained models for $160$ epochs (equivalent to $160k$ steps). Each training session took roughly 24 hours to conclude on a single Quadro RTX 5000 GPU. Similar to NVAE, we used the AdaMax optimizer [133] with a cosine learning rate schedule [134] but without warm restarts.

## 12 ROFL: derivations and detailed description

ROFL aims to generate synthetic optical flow patterns that closely resemble those arising from ecologically relevant types of motion. This entails the consideration of both self-motion and object-motion components. Specifically, the self-motion component encompasses various subtypes such as translational and rotational motions induced by eye or head movements. Furthermore, it is crucial to have control over the relative prominence of these two types of motion. Depending on the specific configuration, the resulting flow pattern can exhibit an *object-dominated* characteristic when an abundance of objects is incorporated, or conversely, a *self-motion-dominated* nature if fewer or no objects are present. By carefully manipulating these factors, ROFL enables the generation of optical flow patterns that faithfully capture the dynamics of real-world scenarios involving an interplay between self-motion and object motion.

### 12.1 Setup

Self-motion is meaningful if the observer is moving relative to a static background (i.e., the environment). We consider a solid background wall at a distance $Z = 1$ from the observer which extends to infinity in the $X - Y$ plane. Define the following coordinate frames

1. **Fixed coordinates**, represented with capital letters: $(X, Y, Z)$.
2. **Observer's coordinates**, represented with lowercase letters: $(x, y, z)$.

Both coordinate systems are centered around the observer. The difference is that $\hat{Z}$ is always perpendicular to the background plane, while $\hat{z}$ is defined such that it is parallel to the fixation point.

All points in the space are initially expressed in the fixed coordinates. For example, any given point like $\vec{P} = (X, Y, Z)$ has its coordinates defined in the fixed system. The same is true for fixation point, $\vec{F} = (X_0, Y_0, Z)$, and the observer's position $\vec{O} = (0, 0, 0)$, and so on. See Fig. S10.

#### 12.1.1 Relating the two coordinate systems

How do we go from the fixed coordinates to the observer's coordinates? We need to find a rotation matrix that maps the $\hat{Z}$ direction to the direction defined by the fixation point. By definition, this will determine $\hat{z}$ and all the other directions in the observer's coordinates. First, let us express $\vec{F}$ using its polar coordinates:

$$\vec{F} = (X_0, Y_0, Z) \equiv \left( \|\vec{F}\|, \Theta_0, \Phi_0 \right), \tag{1}$$

where $\|\vec{F}\| = \sqrt{X_0^2 + Y_0^2 + Z^2}$, $\cos \Theta_0 = Z/\|\vec{F}\|$, and $\tan \Phi_0 = Y_0/X_0$. It turns out what we are looking for is a rotation by an amount $\Theta_0$ about a unit vector $\hat{u}$ that depends only on $\Phi_0$. That is,

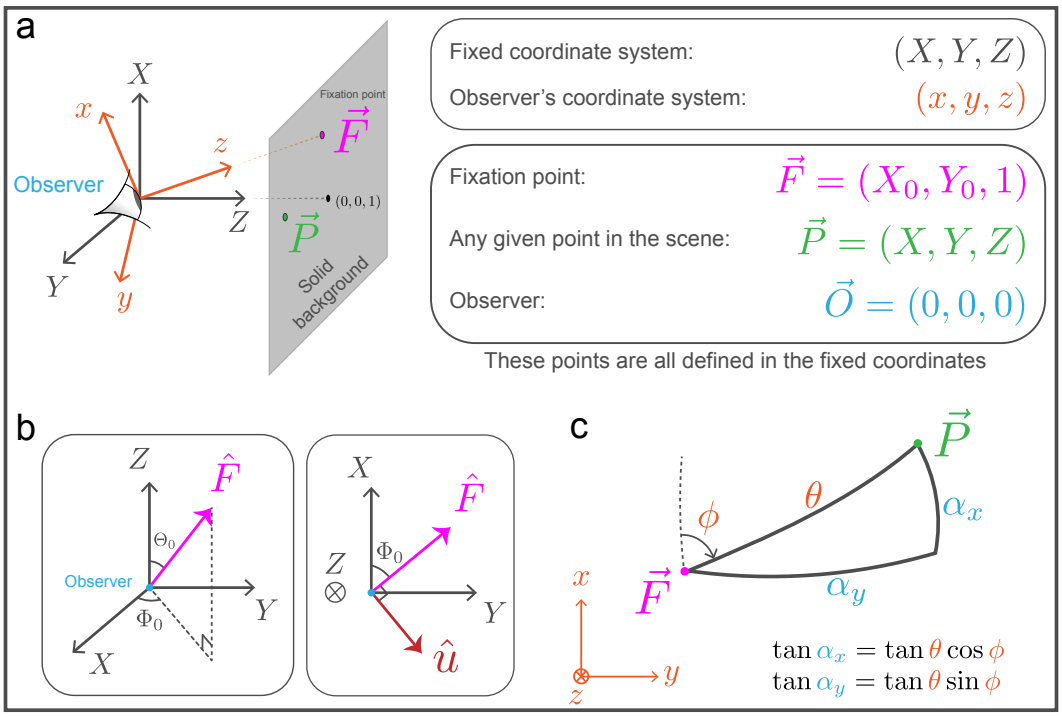


Figure S10: Setup. **(a)** Background and coordinate systems. **(b)** The rotation of the fixed coordinates $\left(\hat{X}, \hat{Y}, \hat{Z}\right)$ onto the observer-centric coordinates $(\hat{x}, \hat{y}, \hat{z})$ can be accomplished by applying a rotation of angle $\Theta_0$ around the unit vector $\hat{u}$ as defined in Equation 2. Note that by definition, $\hat{z}$ is always parallel to the fixation point. See the text for more details. **(c)** Perspective from the observer's point of view. The retinotopic coordinates of point $\vec{P}$ are denoted by $\alpha_x$ and $\alpha_y$, measured in radians.

$$\hat{u}(\Phi_0) = \begin{pmatrix} -\sin \Phi_0 \\ \cos \Phi_0 \\ 0 \end{pmatrix}, \tag{2}$$

and the rotation matrix is given by $R_{\hat{u}}(\Theta_0) = I + (\sin \Theta_0)\, L(\hat{u}) + (1 - \cos \Theta_0)\, L(\hat{u})^2$.

Here, $L(\vec{\omega})$ is the skew-symmetric matrix representation of a vector $\vec{\omega}$, that is

$$L(\vec{\omega}) := \vec{\omega} \cdot \vec{L} = \omega_X L_X + \omega_Y L_Y + \omega_Z L_Z, \tag{3}$$

where $L_i$ are the familiar basis matrices for $\mathfrak{so}(3)$, the Lie algebra of $SO(3)$:

$$L_X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad L_Y = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad L_Z = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{4}$$

The rotation matrix in its full form is expressed as

$$R := R_{\hat{u}}(\Theta_0) = \begin{pmatrix} s^2 + c^2 \cos \Theta_0 & -cs\,[1 - \cos \Theta_0] & c \sin \Theta_0 \\ -cs\,[1 - \cos \Theta_0] & c^2 + s^2 \cos \Theta_0 & s \sin \Theta_0 \\ -c \sin \Theta_0 & -s \sin \Theta_0 & \cos \Theta_0 \end{pmatrix}, \tag{5}$$

where we have defined $c := \cos \Phi_0$ and $s := \sin \Phi_0$.

Any point $\vec{P} = (X, Y, Z)$ defined in the fixed coordinate system can be mapped to a point $\vec{p} = R^T \vec{P}$ with components $\vec{p} = (x, y, z)$ given in the observer's system. For example, it can be shown that $R^T \vec{F}$ is parallel to $\hat{z}$, which is how we defined $R$ in the first place. See Fig. S10.

In conclusion, we have a rotation matrix $R$ that is fully determined as a function of the fixation point $\vec{F}$. Applying $R$ rotates the fixed coordinate system onto the observer's system

$$\begin{aligned} \hat{x} &\coloneqq R\hat{X}, \\ \hat{y} &\coloneqq R\hat{Y}, \\ \hat{z} &\coloneqq R\hat{Z}. \end{aligned} \tag{6}$$

This way, the observer's coordinates $(x, y, z)$ are obtained ensuring that the fixation point is always parallel to $\hat{z}$.

### 12.1.2 Adding motion

Consider a point $\vec{P} = (X, Y, Z)$. In the previous section, we showed that the rotation matrix in Equation 5 can be used to relate $\vec{P}$ to its observer-centric coordinates

$$\vec{P} = R\vec{p}. \tag{7}$$

Differentiate both sides to get

$$\vec{V} \coloneqq \frac{d\vec{P}}{dt} = \frac{dR}{dt}\vec{p} + R\frac{d\vec{p}}{dt}, \tag{8}$$

where $\vec{V}$ is the velocity of the point relative to the observer, with components measured in the fixed coordinate system.

### 12.1.3 Derivative of the rotation matrix

Here we will derive a closed form expression for $dR/dt$ in terms of the observer's angular velocity. This is important, because it essentially captures the effects of eye movement.

Let us use $\vec{\omega} = (\omega_X, \omega_Y, \omega_Z)$ to denote the angular velocity of observer's coordinates, with components expressed in fixed coordinates. The time derivative of $R$ is given as follows

$$\frac{dR}{dt} = L(\vec{\omega})R, \tag{9}$$

where $L(\vec{\omega})$ is given in Equation 3. Taken together, equations 3, 4, 8, and 9 yield

$$\boxed{\begin{aligned} \frac{d\vec{p}}{dt} &= R^T\vec{V} - \left[R^T L(\vec{\omega})R\right]\vec{p} \\ &= R^T\vec{V} - \left[R^T \begin{pmatrix} 0 & -\omega_Z & \omega_Y \\ \omega_Z & 0 & -\omega_X \\ \omega_Y & \omega_X & 0 \end{pmatrix} R\right]\vec{p} \end{aligned}} \tag{10}$$

This equation allows relating the velocity of a moving point expressed in fixed coordinates, $\vec{V}$, to its observer-centric equivalent, $\vec{v} \coloneqq d\vec{p}/dt$. Importantly, Equation 10 is a general relationship valid for any combination of velocities and rotations.

### 12.1.4 Motion perceived by the observer

Before discussing specific applications, we need to solve one remaining piece of the puzzle: what is $\vec{v}$, and how is it related to the motion perceived by the observer? To address this, we need to map the retinotopic coordinates of the observer (measured in radians), to points in the environment (measured in the observer-centric coordinates).

Denote the retinotopic coordinates using $(\alpha_x, \alpha_y)$. The goal is to relate each $(\alpha_x, \alpha_y)$ pair to a point in the world $\vec{p}$ with coordinates $(x, y, z)$. It can be shown that

$$\begin{aligned} \tan\alpha_x &= x/z = \tan\theta\cos\phi, \\ \tan\alpha_y &= y/z = \tan\theta\sin\phi, \end{aligned} \tag{11}$$

where $\theta$ and $\phi$ are the polar coordinates of $\vec{p}$. See Fig. S10c for an illustration.

Differentiate both sides of Equation 11 and rearrange terms to obtain the time derivative of $(\alpha_x, \alpha_y)$ as a function of $\vec{v} = (\dot{x}, \dot{y}, \dot{z})$ and $\vec{p}$. That is,

$$\boxed{\begin{aligned} \dot{\alpha_x} &= \frac{\dot{x}z - \dot{z}x}{z^2 + x^2} \\ \dot{\alpha_y} &= \frac{\dot{y}z - \dot{z}y}{z^2 + y^2} \end{aligned}} \tag{12}$$

Thus, we obtained a relationship between retinotopic velocity $(\dot{\alpha_x}, \dot{\alpha_y})$ (measured in $radians/s$) and velocity of objects in the world (measured in $m/s$). In conclusion, equations 10 and 12 enable us to compute the observer's perceived retinotopic velocity for any type of motion.

### 12.2 Example: maintaining fixation

We are now ready to apply the formalism developed in the preceding section to create optical flow patterns in realistic settings. In this section, we will consider a specific example: self-motion while maintaining fixation on a background point. This is an important and ecologically relevant type of motion, as we often tend to maintain fixation while navigating the environment—think about this next time you are moving around. In fact, one can argue that fixation occurs more naturally during behaviors such as locomotion, as opposed to typical vision experiments where the animals are stationary and passively viewing a screen.

What does it mean to maintain fixation while moving? It means the eyes rotate to cancel out the velocity of the fixation point. Assume the observer is moving with velocity $\vec{V}_{self}$. Equivalently, we can imagine the observer is stationary while the world is moving with velocity $\vec{V} = -\vec{V}_{self}$ (this is the same $\vec{V}$ that appears in Equation 10).

Any eye movement can be represented mathematically as a rotation of the observer's coordinates with respect to the fixed coordinates. Use $\vec{\omega}$ to denote the angular velocity of the rotation. We can determine $\vec{\omega}$ by requiring that the rotation cancels out the velocity of the fixation point. To this aim, we first compute the normal component of $\vec{V}$ with respect to the fixation point $\vec{F}$ as follows

$$\vec{V}_\perp := \vec{V} - \frac{\vec{V}\cdot\vec{F}}{\|\vec{F}\|^2}\vec{F}. \tag{13}$$

Use the above equation to compute the angular velocity

$$\vec{\omega} = \frac{\vec{F}\times\vec{V}_\perp}{\|\vec{F}\|^2}. \tag{14}$$

Plug Equation 14 in Equation 10 to find $\vec{v}$, and use Equation 12 to find $(\dot{\alpha_x}, \dot{\alpha_y})$. This way, we just calculated the motion perceived by the observer during self-motion while maintaining fixation.

### 12.3 The algorithm

Start by choosing a fixation point $\vec{F} = (X_0, Y_0, Z)$, field-of-view extent ($fov$), and retinal resolution ($res$). For example, $fov = 45°$ and $res = 5.625°$. Use this to create a mesh grid for $(\alpha_x, \alpha_y)$. The grid covers $(-fov, +fov)$ degrees in each dimension and has shape $2 * fov/res + 1$, or $17 \times 17$ in our example. Use Equation 11 to compute $(\theta, \phi)$ pair for each point on the $(\alpha_x, \alpha_y)$ grid.

Because the simulation is scale-invariant, we can always fix the distance between the observer and the background wall. Assume $Z = cte. = 1$, and use equations 5 and 7 to find $\vec{p} = (x, y, z)$ per grid point. Note that we assume infinite resolution for the points in the environment.

So far, we partitioned the retina into a grid and found the observer-centric coordinates of a point in the real world that falls on each grid point. Now we can sample a random self-motion velocity $\vec{V}_{self}$ and use the results described in the previous sections to find $(\dot{\alpha_x}, \dot{\alpha_y})$. This provides us with the $2D$ retinal velocity vector on the grid and concludes the algorithm for the simple case of `fixate-0` (i.e., no objects).

Please refer to our code for additional details, including instructions on how to incorporate independently moving objects into the scene.

### 12.3.1 Valid fixation points

The observer-wall distance is constrained to be a positive value. That is, the observer is always situated on the left side of the wall ($Z > 0$ for all points; Fig. S10). This constraint, combined with a given $fov$ value and Equation 5 results in a theoretical bound on which fixation points are valid. To derive this bound, we start by considering the $\hat{Z}$ component of Equation 7

$$(R\vec{p})_3 = R_{31}x + R_{32}y + R_{33}z = Z > 0. \tag{15}$$

Divide both sides by $z$ and use Equation 11 to get

$$R_{31}\tan\alpha + R_{32}\tan\beta + R_{33} > 0. \tag{16}$$

Plug the matrix entries from Equation 5 and rearrange to get

$$\begin{aligned} &\sin\Theta_0(\cos\Phi_0\tan\alpha + \sin\Phi_0\tan\beta) < \cos\Theta_0, \\ &\Rightarrow \quad X_0\tan\alpha + Y_0\tan\beta < Z. \end{aligned} \tag{17}$$

For a given $fov$ value, we have $-\tan(fov) \leq \tan\alpha, \tan\beta \leq \tan(fov)$. Assuming $Z = 1$, we find that the $(X_0, Y_0)$ coordinates of a valid fixation point should satisfy the following inequality

$$|X_0| + |Y_0| < \frac{1}{\tan(fov)}. \tag{18}$$

In other words, the $L_1$ norm of $(X_0, Y_0)$ should be less than $1/\tan(fov)$. For example, when $fov = 45°$, the upper bound is 1. This result can be used to choose an appropriate $fov$: smaller values allow a larger phase space to explore for fixation points.

## 12.4 Conclusions

We introduced Retinal Optic Flow Learning (ROFL), a novel synthetic data framework to simulate realistic optical flow patterns. The main result is given in Equation 10, which shows the simulation has the potential to generalize to various other types of motion. For instance, the observer might engage in smooth pursuit of a moving object instead of maintaining fixation, a scenario that can be modeled simply by adjusting the choice of $\vec{\omega}$ in Equation 10. This flexibility enables us to generate a diverse repertoire of realistic optical flow patterns, which can be used in training and evaluating unsupervised models, among other applications.

We believe that our framework will prove valuable not only to neuroscientists studying motion perception but also to machine learning researchers interested in disentangled representation learning. By bridging these domains, we aim to foster interdisciplinary collaboration and advance our understanding of both biological and artificial information processing systems.