Movie reconstruction from mouse visual cortex activity

Joel Bauer[1,2] **Troy W. Margrie**[1] * **Claudia Clopath**[1,2] *

[1] *Sainsbury Wellcome Centre, University College London, UK*
[2] *Bioengineering Dept., Imperial College, UK*
*These authors contributed equally as last authors

joel.bauer@ucl.ac.uk

Abstract

The ability to reconstruct images represented by the brain has the potential to give us an intuitive understanding of what the brain sees. Reconstruction of visual input from human fMRI data has garnered significant attention in recent years. Comparatively less focus has been directed towards vision reconstruction from single-cell recordings, despite its potential to provide a more direct measure of the information represented by the brain. Here, we achieve high-quality reconstructions of natural movies presented to mice, from the activity of neurons in their visual cortex for the first time. Using our method of video optimization via backpropagation through a state-of-the-art dynamic neural encoding model we reliably reconstruct 10-second movies at 30 Hz from two-photon calcium imaging data. We achieve a pixel-level correction of 0.57 between the ground truth movie and the reconstructions from single-trial neural responses. We find that critical for high-quality reconstructions are the number of neurons in the dataset and the use of model ensembling. This paves the way for movie reconstruction to be used as a tool to investigate a variety of visual processing phenomena.

1 Introduction

One fundamental aim of neuroscience is to eventually gain insight into the ongoing perceptual experience of humans and animals. Reconstruction of visual perception directly from brain activity has the potential to give us a deeper understanding of how the brain represents visual information. Over the past decade, there have been considerable advances in reconstructing images and videos from human brain activity [Nishimoto et al., 2011, Shen et al., 2019a,b, Rakhimberdina et al., 2021, Ren et al., 2021, Takagi and Nishimoto, 2023, Ozcelik and VanRullen, 2023, Ho et al., 2023, Scotti et al., 2023, Chen et al., 2023, Benchetrit et al., 2023, Kupershmidt et al., 2022]. These advances have largely leveraged deep learning techniques to interpret fMRI or MEG recordings, taking advantage of the fact that spatially separated clusters of neurons have distinct visual and semantic response properties [Rakhimberdina et al., 2021]. Due to the low resolution of fMRI and MEG, relative to single neurons, the most successful models heavily rely on extracting semantic content and use diffusion models to generate semantically similar images and videos. Some approaches combine low-level perceptual (retinotopic) and semantic information in separate modules to achieve even better image similarity [Ren et al., 2021, Ozcelik and VanRullen, 2023, Scotti et al., 2023]. However, the pixel-level similarities are still relatively low. These methods are highly useful in humans, but their focus on semantic content may make them less useful when applied to non-human subjects or when using the reconstructed images to investigate visual processing.

Less attention has been given to image reconstruction from non-human brains. This is surprising given the advantages of large-scale single-cell-resolution recording techniques available in animal models, particularly mice. In the past, reconstructions using linear summation of receptive fields or

Gabor filters have shown some success using responses from retinal ganglion cells [Brackbill et al., 2020], thalamo-cortical neurons in lateral geniculate nucleus [Stanley et al., 1999], and primary visual cortex [Garasto et al., 2019, Yoshida and Ohki, 2020]. Recently, deep nonlinear neural networks have been used with promising results to reconstruct static images from mouse retina [Zhang et al., 2020, Li et al., 2023] and visual cortex [Cobos et al., 2022](bioRxiv), and in particular from monkey V4 extracellular recordings [Li et al., 2023, Pierzchlewicz et al., 2023].

Here, we present a method for the reconstruction of 10-second movie clips using two-photon calcium imaging data recorded in mouse V1 [Turishcheva et al., 2023, 2024]. Our method takes advantage of a state-of-the-art (SOTA) dynamic neural encoding model (DNEM) [Baikulov, 2023] which predicts neuronal activity based on video input as well as behavior. Our method allows us to successfully reconstruct videos despite the fact that V1 neuronal activity in awake mice is heavily modulated by behavioral factors such as running speed [Niell and Stryker, 2010] and pupil diameter (correlated with arousal; [Reimer et al., 2014]). We then quantify the spatio-temporal limits of this reconstruction approach and identify key aspects of our method necessary for optimal performance.

2 Video reconstruction using state-of-the-art dynamic neural encoding models

We used publicly available data provided by the Sensorium 2023 competition [Turishcheva et al., 2023, 2024]. The data included movies that were presented to mice and the evoked activity of V1 neurons along with pupil position, pupil diameter, and running speed. The neuronal activity was measured using two-photon imaging of GCaMP6s [Chen et al., 2013] fluorescence from 10 mice, with $\approx$ 8000 neurons from each mouse. In total, we reconstructed ten 10s natural movies from 5 mice.

We used the winning model of the Sensorium 2023 competition which achieved a score of 0.301 ([Baikulov, 2023, Turishcheva et al., 2024] single-trial correlation between predicted and ground truth neuronal activity; Figure 1A and Figure S1A-C). This state-of-the-art (SOTA) dynamic neural encoding model (DNEM) was composed of three parts: core, cortex and readout. The model takes the video as input with the behavioral data (pupil position, pupil diameter, and running speed) broadcast to four additional channels of the video. The original model weights were not used to avoid reconstructing movies the model was trained on. Instead, we retrained 7 instances of the model using the same training data, which did not include the movies reserved for reconstruction. Beyond this point the weights of the model were frozen, i.e. not influenced by future movie presentations.

To reconstruct the videos presented to mice we iteratively optimized an initially blank input video to the SOTA DNEM until the predicted activity in response to this input matched the ground truth recorded neuronal activity. In effect, we optimized the input video to be perceptually similar with respect to the recorded neurons. To achieve this we used an input optimization through gradient descent approach inspired by the optimization of maximally exciting images [Walker et al., 2019] and the reconstruction of static images [Cobos et al., 2022](bioRxiv) [Pierzchlewicz et al., 2023]. The input videos were initialized as uniform gray values and the behavioral parameters (Figure S1A) were added as additional channels, i.e. these were not reconstructed but given. The neuronal activity in response to the input video was predicted using the SOTA DNEM for a sliding window of 32 frames (1.067 sec) with a stride of 8 frames. We saw slightly better results with a stride of 2 frames but, in our case, this did not warrant the increase in training time. For each window, the difference between the predicted and ground truth responses was calculated and this loss backpropagated to the pixels of the input video to get the gradient of the loss with respect to each pixel. In effect, the input pixels were thus treated as if they were model weights. The gradients for each pixel were then averaged across all windows and the pixels of the input video updated accordingly (See Supplementary Algorithm 1).

The data from the Sensorium competition provided the activity of neurons within a 630 by 630 μm field of view for each mouse, i.e. covering roughly one-fifth of mouse V1. Due to the retinotopic organization of V1 we therefore did not expect to get good reconstructions of the entire video frame. However, gradients still propagated to the full video frame and produced non-sensical results along the periphery of the video frames (Figure S3). Inspired by previous work [Mordvintsev et al., 2018] [Willeke et al., 2023b](bioRxiv) we therefore decided to apply a mask during training and evaluation. To generate these masks, we optimized a transparency layer placed at the input to the SOTA DNEM. High values are given to pixels that contribute to the accurate prediction of neuronal activity and

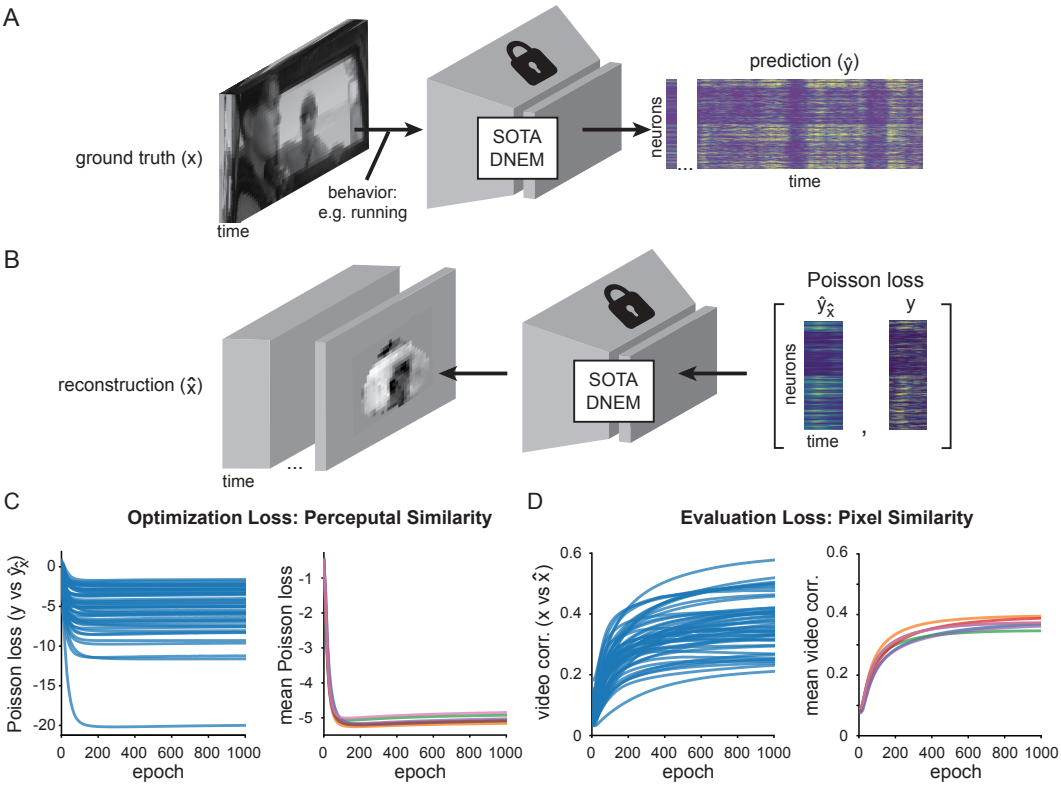


Figure 1: Video reconstruction from neuronal activity in mouse V1 (data provided by the Sensorium 2023 competition; [Turishcheva et al., 2023, 2024]) using a state-of-the-art (SOTA) dynamic neural encoding model (DNEM; [Baikulov, 2023]). A) DNEMs predict neuronal activity from mouse primary visual cortex, given a video and behavioural input. B) We use a SOTA DNEM to reconstruct part of the input video given neuronal population activity, using gradient descent to optimize the input. C) Poisson negative log likelihood loss across training steps between ground truth neuronal activity and predicted neuronal activity in response to videos. Left: all 50 videos from 5 mice for one model. Right: average loss across all videos for 7 model instances. D) Spatio-temporal (pixel-by-pixel) correlation between reconstructed video and ground truth video.

represent the collective receptive field of the neural population. None of the reconstructed movies were used in the optimization of this transparency mask. The transparency masks are aligned with with but not identical to the On-Off receptive field distribution maps using sparse-noise (Figure S2). This mask was applied during the optimization of the reconstructed movies (training mask: binarized with threshold $\alpha = 0.5$) and applied again to the final reconstruction (evaluation mask: binarized with threshold $\alpha = 1$) (See Supplementary Algorithm 2). Applying the mask in two stages first boosts the performance of reconstruction itself and separately allows evaluation of the reconstruction in a region of high confidence, given the neural population available (Figure S3).

As the loss between predicted (Figure S1D) and ground truth responses (Figure S1B) decreased, the similarity between the reconstructed and ground truth input video increased (Figure 1C-D). We generated 7 separate reconstructions from 7 neural encoding models (trained on the same data) and averaged them. Finally, we applied a 3D Gaussian filter with sigma 0.5 pixels to remove the remaining static noise (Figure S3) and applied the evaluation mask. When presenting videos in this paper we normalize the mean and standard deviation of the reconstructions to the average and standard deviation of the corresponding ground truth movie before applying the evaluation masks, but this is not done for quantification except in Figure S3D. The Gaussian filter was not applied when evaluating spatial or temporal resolution (Figure 4, Figure S6, Figure S5).

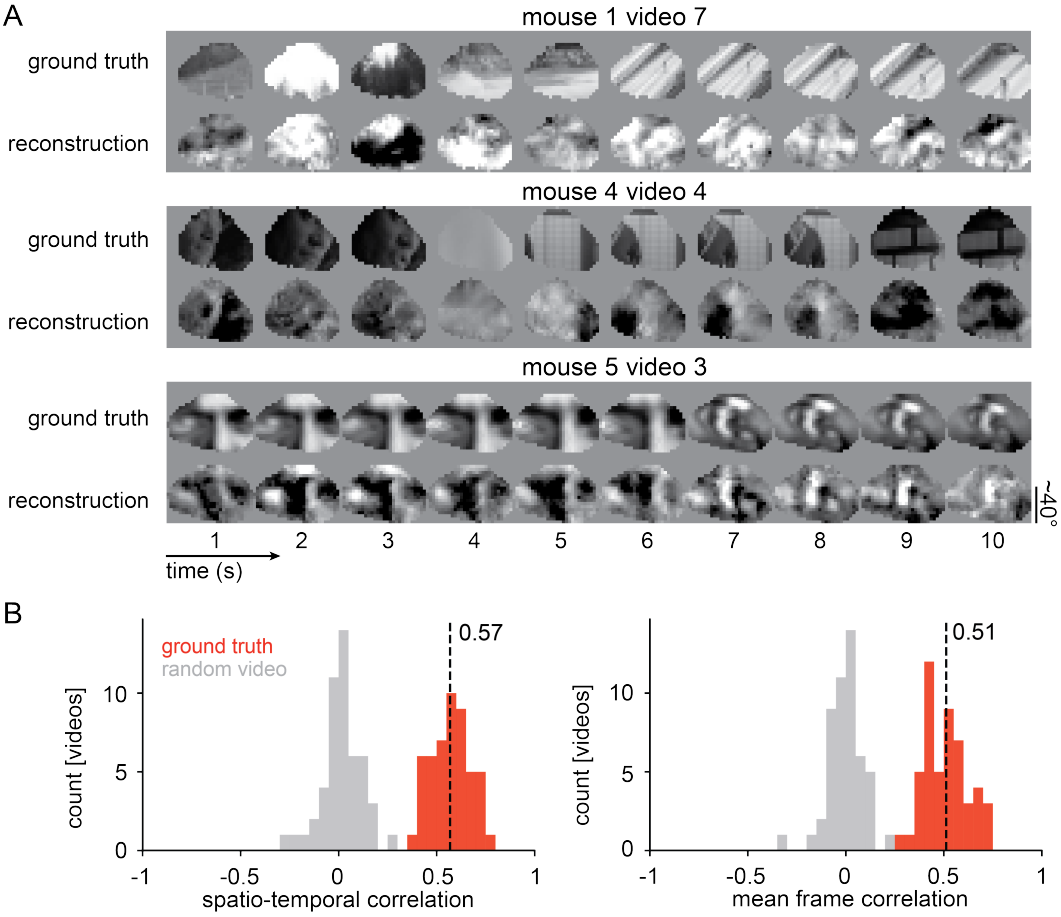


Figure 2: Reconstruction performance. A) Three reconstructions of 10s videos from different mice (see Supplementary Video 1 for full set: YouTube link). Reconstructions have been luminance (mean pixel value across video) and contrast (standard deviation of pixel values across video) matched to ground truth. B) The reconstructed videos have high correlation to ground truth in both spatio-temporal correlation (mean Pearson's correlation r = 0.569 with 95% CIs 0.542 to 0.596, t-test between ground truth and random video p = $6.69 * 10^{-49}$, n = 50 videos from 5 mice) and mean frame correlation (mean Pearson's correlation r = 0.512 with 95% CIs 0.481 to 0.543, t-test between ground truth and random video p = $4.29 * 10^{-45}$, n = 50 videos from 5 mice).

2.1 High-quality video reconstruction

As can be seen in Figure 2 and Supplementary Video 1, the reconstructed videos capture much of the spatial and temporal dynamics of the original input video. Because our optimization of the movies was based on a perceptual loss function, we were interested in how closely these movies matched the originals on the pixel level. To evaluate performance of the video reconstructions we therefore correlated either all pixels from all time points between ground truth and reconstructed videos (Pearson's correlation r = 0.569; to quantify temporal and spatial similarity), or the average correlation between all sets of frames (Pearson's correlation r = 0.512; to quantify just spatial similarity)(Figure 2B and Figure S1E). This represents a $\approx$ 2x higher pixel-level correlation over previous single-trial static image reconstructions from V1 in awake mice (image correlation 0.238 +/- 0.054 s.e.m for awake mice) [Yoshida and Ohki, 2020] over a similar retinotopic area ($\approx$ 43° x 43°) while also capturing temporal dynamics. However, we would like to stress that directly comparing static image reconstruction methods with movie reconstruction approaches is fundamentally problematic, as they rely on different data types both during training and evaluation (temporally averaged vs continuous neural activity, images flashed at fixed intervals vs continuous movies).

Reconstruction quality, however, was not consistent across movies (Figure 2B) or constant throughout the 10 second videos (Figure S1E). We therefore investigated what factors may cause these fluctuations by correlating video motion energy, contrast and luminance, as well as running speed, pupil diameter and eye movement with frame correlation. We found that contrast correlated with frame correlation, but only to a moderate degree. Video motion energy shows a trend but was not significant (Figure S4). We also found that the ability of the SOTA DNEM to predict neural activity correlated with reconstruction performance. This could be because some frames are harder to reconstruct due to their content (high temporal and spatial frequencies) or because neural activity in these moments is influenced by factors the model cannot take into account.

2.2 Ensembling

We found that the 7 instances of the SOTA DNEMs by themselves performed similarly in terms of reconstructed video correlation (Figure 1D), but that this correlation was significantly increased by taking the average across reconstructions from different models (Figure 3) – A technique known as bagging, and more generally ensembling [Breiman, 1996]. We averaged over 7 model instances, which gave a performance increase of 28.0%, but the largest gain in performance, 13.7%, came from averaging across just 2 models (Figure 3). Doubling the number of models to 4 increased the performance by another 8.32%. Individual models produced reconstructions with high-frequency noise in the temporal and spatial domains. We therefore think the increase in performance from ensembling is mostly an effect of averaging out this high-frequency noise. On-the-other-hand, it is possible that averaging over separately optimized reconstructions degrades high-frequency information. We therefore tested whether averaging pixel gradients from all models at each iteration rather than averaging the final movies yields higher performance, but we observed no improvement (Figure S3C). Overall, although ensembling over models trained on separate data splits is a computationally expensive method, it substantially improved reconstruction quality.

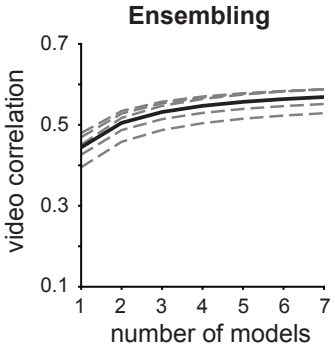


Figure 3: Model ensembling. Mean video correlation is improved when predictions from multiple models are averaged. Dashed lines are individual animals, solid line is mean. One-way repeated measures ANOVA p = $1.11 * 10^{-16}$. Bonferroni corrected paired t-test outcomes between consecutive ensemble sizes are all p < 0.001, n = 5 mice.

2.3 Not all spatial and temporal frequencies are reconstructed equally

While the reconstructed videos achieve high correlation to ground truth, it is not entirely clear if the remaining deviations are due to the limitations of the model or arise from the recorded neurons themselves. To assess the resolution limits of our reconstruction process, we assessed the model's ability to reconstruct synthetic stimuli at varying spatial and temporal resolutions in a noise-free scenario.

To quantify which spatial and temporal frequencies our reconstruction approach is able to capture we used a Gaussian noise stimulus set generated using a Gaussian process (`https://github.com/TomGeorge1234/gp_video`; Figure 4A). The dataset consisted of 49, 2 second, 36 by 36 pixel videos at 30 Hz, which varied in the spatial and temporal length constants. As we did not have ground truth neuronal activity in response to this stimulus set, we first predicted the neuronal responses given these videos using the ensembled SOTA DNEMs. We then used gradient descent to reconstruct the original input using these predicted neuronal responses as the target. In this way, we generated reconstructions in an ideal case with no biological noise and assuming the SOTA DNEM perfectly predicts neuronal activity (Figure 4B). This means the video reconstruction quality loss reflects the inefficiency of the reconstruction process itself without the additional loss or transformation of information by processes such as top-down modulation, e.g. predictive coding or selective feature

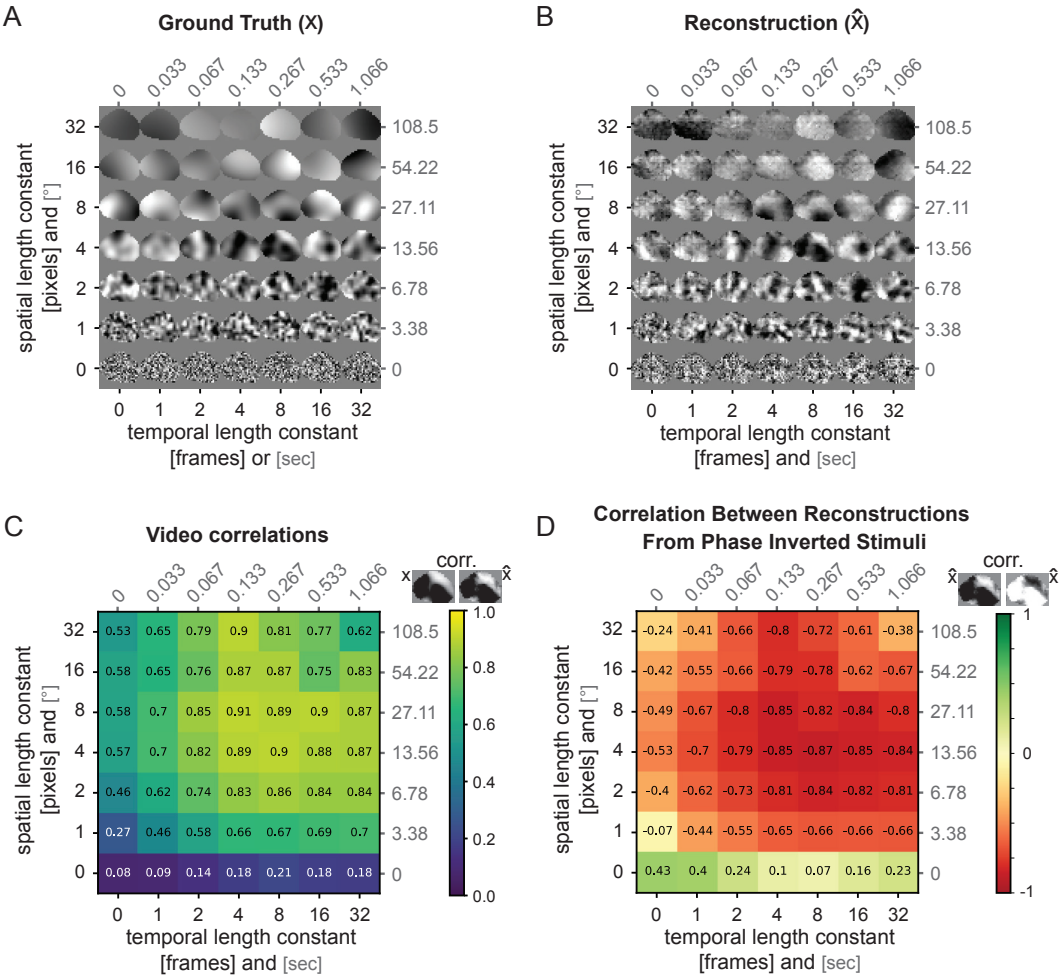


Figure 4: Reconstruction of Gaussian noise across the spatial and temporal spectrum using predicted activity. A) Example Gaussian noise stimulus set with evaluation mask for one mouse. Shown is the last frame of a 2 second video. B) Reconstructed Gaussian stimuli with SOTA DNEM predicted neuronal activity as the target (see also Supplementary Video 2: YouTube link). C) Pearson's correlation between ground truth (A) and reconstructed videos (B) across the range of spatial and temporal length constants. For each stimulus type the average correlation across 5 movies reconstructed from the SOTA DNEM of 3 mice is given. D) Pearson's correlation between reconstructions from phase-inverted Gaussian noise stimuli.

attention (see Discussion). We found that the reconstruction process failed at high spatial frequencies (< 1 pixel, or < 3.4° retinotopy) and performed worse at high temporal frequencies (< 1 frame, or > 30 Hz)(Figure 4C and Supplementary Video 2). We repeated this analysis using full-field high-contrast square gratings drifting in the four cardinal directions and similarly found that high-spatial and temporal frequencies were not reconstructed as well as low-spatial and temporal frequency gratings (Figure S6). We also found that beyond the spatial reconstruction limit, the reconstructions from phase-inverted Gaussian noise stimuli had higher correlation with each other than with their ground truth stimuli (Figure 4D). Never-the-less, even when the reconstructions were not captured on the pixel level, they did capture some of the spatial entropy and motion energy of the ground truth stimuli (Figure S5).

To test if model ensembling improves Gaussian noise reconstruction quality across all spatial and temporal length constants uniformly, we subtracted the average video correlation across the seven model instances from the video correlation of the average video (i.e. ensembled video reconstruction minus unensembled video reconstruction; Figure S5C). We found that, in particular, short temporal

and spatial length constant stimuli improved in correlation, supporting our hypothesis that ensembling mitigates the high-frequency noise we observed in the reconstruction from individual models.

2.4 Neuronal population size

In order to design future *in vivo* experiments to investigate visual processing using our video reconstruction approach, it would be useful to know how reconstruction performance scales with the number of recorded neurons. This is vital for prioritizing experimental parameters such as weighing between sampling density within a similar retinotopic area and retinotopic coverage to maximize both video reconstruction quality and visual coverage. We therefore performed an *in silico* ablation experiment, dropping either 50, 75% or 87.5% of the total recorded population of $\approx$ 8000 neurons per mouse by setting their activity to 0 (Figure 5). We found that dropping 50% of the neurons reduced the video correlation by only 9.96% while dropping 75% reduced the performance by 24.9%. We would therefore argue that $\approx$ 4000-8000 neurons within a 630 by 630 μm area ($\approx$ 10000-20000 neurons/mm^2) of mouse V1 would provide a balance when compromising between density and 2D coverage.

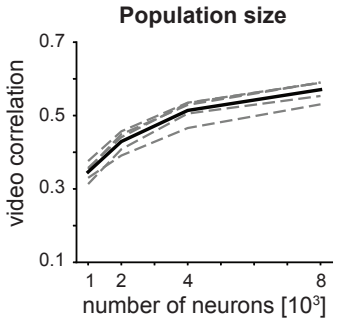


Figure 5: Video reconstruction using fewer neurons (i.e. population ablation) leads to lower reconstruction quality. Dashed lines are individual animals, solid line is mean. One-way repeated measures ANOVA p = $5.70 * 10^{-13}$. Bonferroni corrected paired t-test outcomes between consecutive drops in population size are all p < 0.001, n = 5 mice.

2.5 Visualization of Reconstruction Error

One advantage of stimulus reconstruction compared to stimulus identity decoding (i.e. classification) is that it is possible to visualize the deviation of the reconstructed stimuli from what is expected. This is interesting because reconstruction performance is not stable over time but fluctuates (Figure 6A), likely due to the fact that the DNEM does not have access to all possible factors that influence neural activity. When using our reconstruction method it is not the input stimulus similarity that is optimized, but the evoked activity of the stimulus. As a consequence, the predicted neural response from the reconstructed movie is more similar to the experimental neural response compared to the predicted neural response evoked by the original ground truth movie (Figure 6B). It is possible to visualize this deviation on a pixel level by subtracting the experimentally derived movie reconstruction (i.e., based on measured neural responses) from the *in silio* simulation derived movie reconstruction (i.e., first predict activity based on the ground truth video and then reconstruct the movie based on the resulting simulated neural activity)(Figure 6B-C). With the current dataset it is not possible to test if these deviations reflect failures of the encoding model to predict neural activity given the sensory stimulus or true deviations of the images represented by the neural population from the sensory stimulus, but this approach may be an interesting method for investigating when and why model predictions of neural activity deviate from the experimentally measured activity.

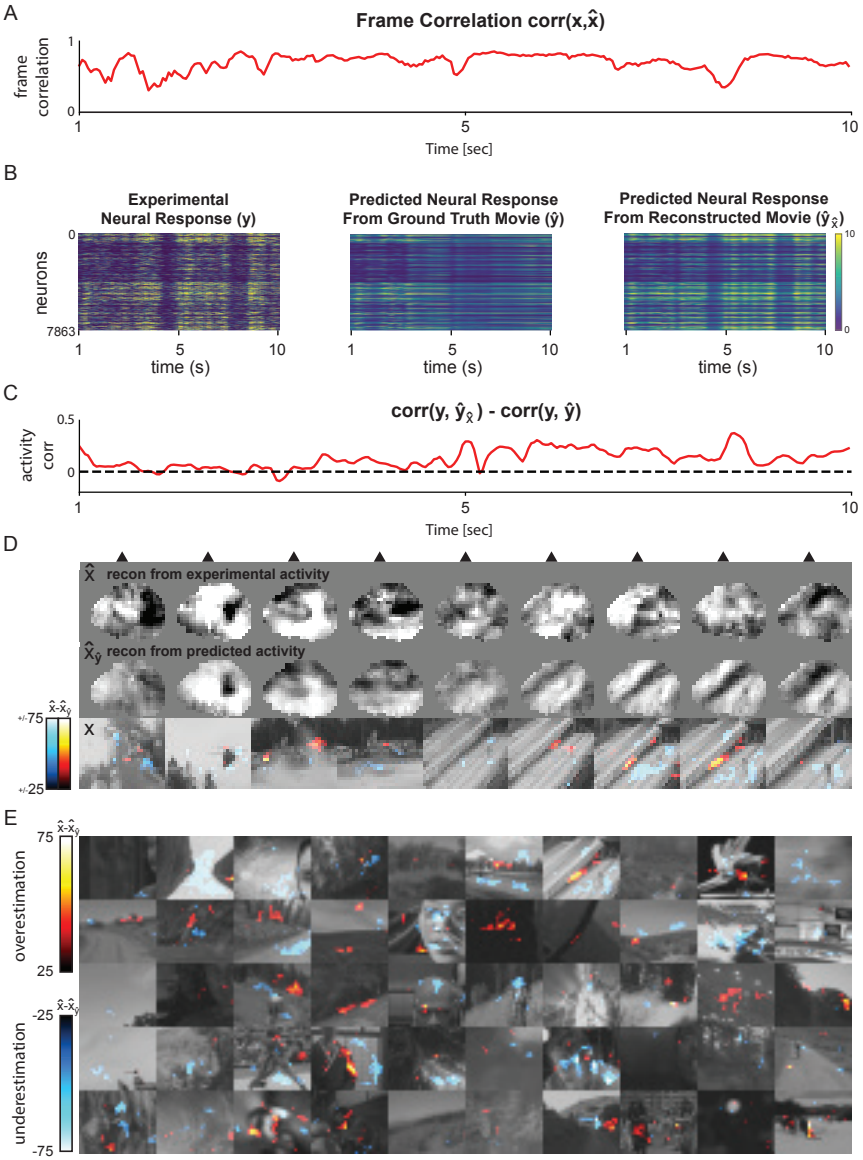


Figure 6: Comparison of reconstructions from experimental responses vs expected responses and their visualization as error maps. A) Frame-by-frame correlation between reconstructed and ground truth video for mouse 1 trial 7 (same as Figure S1). B) From left to right: experimental (ground truth) neural activity y, neural activity predicted by DNEM from ground truth video $\hat{y}$, neural activity predicted by DNEM based on reconstructed movie $\hat{y}_{\hat{x}}$. C) Difference between correlation of true neural response y with predicted neural response from the ground truth movie $\hat{y}$, and correlation of true neural response y with predicted neural response from the reconstructed movie $\hat{y}_{\hat{x}}$. D) 9 frames from mouse 1 trial 7. From top to bottom: reconstructed movie $\hat{x}$, reconstructed movie from predicted neural response to ground truth movie $\hat{x}_{\hat{y}}$, ground truth movie x with overlayed heatmap of the difference between $\hat{x}$ and $\hat{x}_{\hat{y}}$ (error map). E) Error map of one frame from all 50 movie clips. Each row is 10 trials from one mouse. Aee also Supplementary Video 4: YouTube link)

3 Discussion

3.1 Stimulus identification vs reconstruction

Stimulus identification, i.e. identifying the most likely stimulus from a constrained set, has been a popular approach for quantifying whether a population of neurons encodes the identity of a particular stimulus [Földiák, 1993, Kay et al., 2008]. This approach has, for instance, been used to decode frame identity within a movie [Deitch et al., 2021, Xia et al., 2021, Schneider et al., 2023, Chen et al., 2024]. Some of these approaches have also been used to reorder the frames of the ground truth movie [Schneider et al., 2023] based on the decoded frame identity. Importantly, stimulus identification methods are distinct from stimulus reconstruction where the aim is to recreate what the sensory content of a neuronal code is in a way that generalizes to new sensory stimuli [Rakhimberdina et al., 2021]. This is inherently a more demanding task because the range of possible solutions is much larger. Although stimulus identification is a valuable tool for understanding the information content of a population code, stimulus reconstruction could provide a more generalizable approach, because it can be applied to novel stimuli.

3.2 Comparison to other reconstruction methods

There has recently been a growing number of publications in the field of image reconstruction, primarily from fMRI data, and a comprehensive review of all the approaches is outside the scope of this paper. However, we will briefly summarize the most common approaches and how they relate to our own method. In general, image reconstruction methods can be categorized into one of four groups: direct decoding models, encoder-decoder models, invertible encoding models, and encoder model input optimization.

Direct decoders, directly decode the input image/videos from neuronal activity with deep neuronal networks [Shen et al., 2019a, Zhang et al., 2020, Li et al., 2023]. When training direct decoders, the decoders can be pretrained [Ren et al., 2021] or additional constraints can be added to the loss function to encourage the decoder to produce images that adhere to learned image statistics [Shen et al., 2019a, Kupershmidt et al., 2022]. A direct decoder approach has been used for video reconstruction in mice [Chen et al., 2024], but in that case, the training and test movies were the same, meaning it is unclear if out-of-training set generalization was achieved (a key distinction between sensory reconstruction and stimulus identification, see previous section).

In encoder-decoder models the aim is to combine separately trained brain encoders (brain activity to latent space) and decoders (latent space to image/video). Recently, this approach has become particularly popular because it allows the use of SOTA generative image models such as stable diffusion [Rombach et al., 2021, Takagi and Nishimoto, 2023, Scotti et al., 2023, Chen et al., 2023, Benchetrit et al., 2023]. The encoder part of the models are first trained to translate brain activity into a latent space that the pretrained generative networks can interpret. Because these latent spaces are often conditioned on semantic information, this lends itself to separate processing of low-level visual and high-level semantic information from brain activity [Scotti et al., 2023].

Invertible encoding models are encoding models which, once trained to predict neuronal activity, can implicitly be inverted to predict sensory input given brain activity. We would also include those models in this class which first compute the receptive field or preferred stimulus of neurons (or voxels) and reconstruct the input as the weighted sum of the receptive fields by their activity [Stanley et al., 1999, Thirion et al., 2006, Garasto et al., 2019, Brackbill et al., 2020, Yoshida and Ohki, 2020, Nishimoto et al., 2011]. The down-side of this approach is that invertible linear models generally under perform in terms of capturing the coding properties of neurons compared to more complex deep neural networks [Willeke et al., 2023a].

Encoder input optimization, also involves first training an encoder which predicts the activity of neurons or voxels given sensory input. Once trained the encoder is fixed and the input to the network is optimized using backpropagation until the predicted activity matches the observed activity [Pierzchlewicz et al., 2023]. Unlike with invertible encoding models any SOTA neuronal encoding model can be used. But like invertible models, the networks are not specifically trained to reconstruct images so they may be less likely to extrapolate information encoded by the brain by learning general image statistics. There is some evidence to support this, static image reconstructions which where optimized to evoke similar *in silico* predicted neural activity also evoke more similar neural

responses *in vivo* compared to other methods which optimized image similarity directly [Cobos et al., 2022](bioRxiv).

Although outlined here as 4 distinct classes these approaches can be combined. For instance, encoder input optimization can be combined with image diffusion [Pierzchlewicz et al., 2023] and in principle invertible models could also be combined in such a way.

We chose to pursue a pure encoder input optimization approach for single cell mouse visual cortex activity for two reasons. First, there have been considerable advances in the performance of neuronal encoding models for dynamic visual stimuli [Sinz et al., 2018, Wang et al., 2025, Turishcheva et al., 2024] and we aimed to take advantage of these developments. Second, the addition of a generative decoder trained to producing high quality images brings with it the risk of extrapolating information based on general image statistics rather than interpreting what the brain is representing. In some cases, the brain may not be encoding coherent images and in those cases we would argue image reconstruction should fail, rather than producing an image when only the semantic information is present.

3.3 Key contributions and limitations

We demonstrate high-quality video reconstruction from mouse V1 using SOTA DNEMs to iteratively optimize the input video to match the resulting predicted activity with the recorded neuronal activity. Key to achieving high-quality reconstructions is model ensembling and using a large enough number of recorded neurons over a given retinotopic area.

While we averaged the video reconstructions from several models, an alternative method would be to average the gradients calculated by multiple models at each epoch, as has been done for the generation of maximally exciting images in the past [Walker et al., 2019]. When using video models this can be an impractical solution due to the amount of GPU memory required, but in principle there might be situations in which averaging gradients yields better reconstructions. For instance, there may be multiple solutions for the activation pattern of a neural population, e.g. if their responses are translation/phase invariant [Ito et al., 1995, Tacchetti et al., 2018]. In such a case, averaging 'misaligned' reconstructions from multiple models might degrade overall quality. However, we observed no performance improvement when ensembling with gradients instead of ensembling with reconstructions.

The SOTA DNEM we used takes video data at an angular resolution of 3.4 °/pixels at the center of the screen which is about 3x worse than the visual acuity of mice ($\approx$ 0.5 cycles/° [Prusky and Douglas, 2004]). As our model can reconstruct Gaussian noise stimuli down to a spatial length constant of 1 pixel, and drifting gratings up to a spatial frequency of 0.071 cycles/°, there is still some potential for improving spatial resolution. To close this gap and achieve reconstructions equivalent to the limit of mouse visual acuity, a different dataset and model would likely need to be developed. However, the frame rate of the videos the SOTA DNEM takes as input (30 Hz) is faster than the flicker fusion frequency of mice (14 Hz [Nomura et al., 2019]) and our tests with Gaussian noise and drifting grating stimuli show that the temporal resolution of reconstruction is close to this expected limit. Future efforts should therefore focus on the spatial resolution of video reconstruction rather than the temporal resolution.

It is, however, unclear how closely the representation of vision by the brain is expected to match the actual input. There are a number of visual processing phenomena that have previously been identified which leads us to suspect that some deviations between video reconstructions and ground truth input are to be expected. One such phenomenon is predictive coding [Rao and Ballard, 1999, Fiser et al., 2016]. It is possible that the unexpected parts of visual stimuli are sharper and have higher contrast compared to the expected parts when reconstructed from neuronal activity. Alternatively, perceptual learning is a phenomenon where visual stimulus detection or discriminability is enhanced through prolonged training [Li, 2015] and is associated with changes in the tuning distribution of neurons in the visual system [Goltstein et al., 2013, Poort et al., 2015, Jurjut et al., 2017, Schumacher et al., 2022]. Similarly, selective feature attention can modulate the response amplitude of neurons that have a preference for the features that are currently being attended to [Kanamori and Mrsic-Flogel, 2022]. Visual task engagement and training could therefore alter the accuracy and biases of what features of a video can accurately be reconstructed from the neuronal activity. Visualizing differences between

movie reconstructions from experimentaly derived recordings to those from predicted activity, as we have done, may be an interesting approach.

Although fMRI-based reconstruction techniques are starting to be used to investigate visual phenomena in humans (such as illusions [Cheng et al., 2023] and mental imagery Shen et al. [2019b], Koide-Majima et al. [2024], Kalantari et al. [2025]), visual processing phenomena are likely difficult to investigate using existing fMRI-based reconstruction approaches, due to the low spatial and temporal resolution of the data. Additionally, many of these fMRI-based reconstruction approaches rely on the use of pretrained generative diffusion models to achieve more naturalistic and semantically interpretable images [Takagi and Nishimoto, 2023, Ozcelik and VanRullen, 2023, Scotti et al., 2023, Chen et al., 2023], but very likely at the cost of introducing information that may not be present in the actual neuronal representation. In contrast, our video reconstruction approach using single-trial single-cell resolution recordings, without a pretrained generative model, provides a more accurate method to investigate visual processing phenomena such as predictive coding, perceptual learning, and selective feature attention.

In conclusion, we reconstruct videos presented to mice based on single-trial activity of neurons in the mouse visual cortex. This paves the way to using movie reconstruction as a tool to investigate a variety of visual processing phenomena.

4 Methods

4.1 Source data

The data was provided by the Sensorium 2023 competition [Turishcheva et al., 2023, 2024] and downloaded from `https://gin.g-node.org/pollytur/Sensorium2023Data` and `https://gin.g-node.org/pollytur/sensorium_2023_dataset`. The data included grayscale movies presented to the mice at 30 Hz on a 31.8 by 56.5 cm monitor 15 cm from and perpendicular to the left eye. The movies were provided as spatially downsampled versions of the original screen resolution to 36 by 64 pixels, corresponding to an angular resolution of 3.4 °/pixel at the center of the screen. The pupil position and diameter were recorded at 20 Hz and the running at 100 Hz. The neuronal activity was measured using two-photon imaging [Denk et al., 1990] of GCaMP6s [Chen et al., 2013] fluorescence at 8 Hz, extracted and deconvolved using the CAIMAN pipeline [Giovannucci et al., 2019]. For each of the 10 mice, the activity of $\approx$ 8000 neurons was provided. The different data types were resampled to 30 Hz.

4.2 State-of-the-art dynamic neural encoding model

We used the winning model of the Sensorium 2024 competition, DwiseNeuro [Turishcheva et al., 2023, 2024]. The code for the SOTA DNEM was downloaded from `https://github.com/lRomul/sensorium`. This model achieved an average single-trial correlation between predicted and ground truth neural activity of 0.291 during the competition, this was later improved to 0.301. The competition benchmark models achieved 0.106, 0.164 and 0.197 single-trial correlation, while the second and third place models achieved 0.265 and 0.243. Across the models, a variety of architectural components were used, including 2D and 3D convolutional layers, recurrent layers, and transformers, to name just a few. The winning model consists of 3 main components: core, cortex, and readout. The core largely consisted of factorized 3D convolution blocks with residual connections, positional encoding [Vaswani et al., 2017] and SiLU activations [Elfwing et al., 2017] followed by spatial average pooling. The cortex consisted of three fully connected layers. The readout consisted of a 1D convolution for each mouse with a final Softplus nonlinearity, that gives activity predictions for all neurons of each mouse. The kernel of the input layer had size 16 with a dilation of 2 in the time dimension, so spanned 32 video frames.

The original ensemble of models consisted of 7 model instances trained on a 7-fold cross-validation split of all available Sensorium 2023 competition data ($\approx$ 1 hour of training data and $\approx$ 8 min of cross-validation data per fold from each mouse). Each model instance was trained on 6 of 7 data folds, with different validation data excluded from training for each model. To allow ensembled reconstructions of videos without test set contamination we instead retrained the models with a shared validation fold, i.e. we retrained the models leaving out the same validation data for all 7 model instances. The only other difference in the training procedure was that we retrained the models using a batch size of 24 instead of 32, this did not change the performance of neuronal response prediction on the withheld data folds (mean validation fold predicted vs ground truth response correlation for original weights: 0.293; and retrained weights: 0.291). We also did not use model distillation, while the original model did (see `https://github.com/lRomul/sensorium`).

We chose the first 10 movies in data fold 0 (assigned as part of the DNEM code using a video hashing function) for reconstructions. We additionally excluded 9 movies which were incorrectly assigned to fold 0 and replaced them with other movie clips from fold 0.

4.3 Additional visual stimuli

The Gaussian noise stimuli were downloaded from `https://github.com/TomGeorge1234/gp_video` and spanned a range of 0 to 32 pixels in spatial length constant and 0 to 32 frames in temporal length constant used in the Gaussian process. 5 separately generated movies of 2 seconds each were generated and combined with their phase-inverted versions to give a total of 10 trials.

The drifting grating stimuli were produced using PsychoPy [Peirce et al., 2019] and ranged from 0.5 to 0.062 cycles/degree and 0.5 to 0 cycles/second, with 2 seconds of movie for each cardinal direction. These ranges were chosen to avoid aliasing effects in the 36 by 64 pixel videos. The highest temporal frequency corresponds to a flicker stimulus.

The receptive field mapping stimulus, i.e. sparse noise stimulus, consisted of a pre-stimulus gray (gray value 127) screen period of 0.5 seconds, a 0.5 second stimulus period where one pixel was set to either 0 (Off) or 255 (On), and a 0.5 second post-stimulus gray screen period. The full stimulus set consisted of 4608 stimuli, one On and one Off stimulus for every pixel of the 36 by 64 movie.

4.4 Mask training

To generate the transparency masks we used an alpha blending approach inspired by [Mordvintsev et al., 2018] [Willeke et al., 2023b](bioRxiv). A transparency layer was placed at the input to the SOTA DNEM. This transparency layer was used to alpha blend the true video V with another randomly selected background video BG from the data:

$$V_{BG} = V * \alpha + V_{BG} * (1 - \alpha) \tag{1}$$

where α is the 2D transparency mask and V_{BG} is the blended input video. This mask was optimized using stochastic gradient descent (for 1000 epochs with learning rate 10) with mean squared error (MSE) loss between the true responses y and the predicted responses $\hat{y}$ scaled by the average weight of the transparency mask $\bar{\alpha}$:

$$MSE(y, \hat{y}) = \sum_{i=1}^{n} (y - \hat{y})^2 \tag{2}$$

$$Loss = \text{MSE}(y, \hat{y} * (1 - \bar{\alpha})) \tag{3}$$

where n is the total number of neurons. The mask was initialized as uniform noise between 0 and 0.05. At each epoch the neuronal activity in response to a randomly selected 32 frame video segment from the training set was predicted and the gradients of the loss (Equation 3) with respect to the pixels in the transparency mask α were calculated for each video frame. The gradients were normalized by their matrix norm, clipped to between -1 and 1 and averaged across frames. The gradients were smoothed with a 2D Gaussian kernel of $\sigma = 5$ and subtracted from the transparency mask. The transparency mask was only calculated using one SOTA DNEM instance using its validation fold. See Supplementary Algorithm 2.

The transparency mask was thresholded and binarized at 0.5 for the masked gradients ∇_{masked} or 1 for the masked videos for evaluation V_{eval}:

$$\nabla_{masked} = \nabla * (\alpha > 0.5) \tag{4}$$

$$V_{eval} = V * (\alpha > 1) \tag{5}$$

where ∇ is the gradients of the loss with respect to each pixel in the video and V is the reconstructed video before masking. These masks were trained independently for each mouse using one model instance with the original weights of the model `https://github.com/lRomul/sensorium`, not the retrained models used in the rest of this paper to reconstruct the videos.

4.5 Video reconstruction

To reconstruct the input video we initialized the video as uniform gray values and concatenated the ground truth behavioral parameters. The SOTA DNEM took 32 frames at a time and we shifted this window by 8 frames until all frames of the whole 10s video were covered. For each 32-frame window, the Poisson negative log-likelihood loss between the predicted and true neuronal responses was calculated:

$$Loss(y, \hat{y}) = \sum \hat{y}_{neuron} - y_{neuron} \log(\hat{y}_{neuron} + 10^{-8}) \tag{6}$$

where $\hat{y}$ are the predicted responses and y are the ground truth responses. The gradients of the loss with respect to each pixel of the input video were calculated for each window of frames and averaged across all windows. The gradients for each pixel were normalized by the matrix norm across all gradients and clipped to between -1 and 1. The gradients were masked (Equation 4) and applied to

the input video using Adam without second order momentum [Kingma and Ba, 2014] ($\beta_1 = 0.9$) for 1000 epochs and a learning rate of 1000, with a learning rate warm-up for the first 10 epochs. After each epoch, the video was clipped to between 0 and 255. The optimization was run for 1000 epochs. 7 reconstructions from 7 model instances were averaged, denoised with a 3D Gaussian filter $\sigma = 0.5$ (unless specified otherwise), and masked with the evaluation mask. See Supplementary Algorithm 1. Optimizing each 10-second video with one model instance for 1000 epochs took $\approx$ 60 min using a desktop with an RTX4070 GPU.

4.6 Reconstruction quality assessment

To evaluate the similarity between reconstructed and ground truth videos, we used the mean Pearson's correlation between pixels of corresponding frames to evaluate spatial similarity:

$$\text{mean frame correlation} = \frac{1}{f}\sum_{i=1}^{f}\frac{cov(x_i, \hat{x}_i)}{\sigma_{xi}\sigma_{\hat{x}i}} \tag{7}$$

where f is the number of frames, and x_i and $\hat{x}_i$ are the ground truth and reconstructed frames. To evaluate temporal and spatial similarity between ground truth and reconstructed videos we used the Pearson's correlation between all pixels of the whole movie:

$$\text{video correlation} = \frac{cov(x, \hat{x})}{\sigma_x \sigma_{\hat{x}}} \tag{8}$$

To calculate the Shannon entropy, we first computed the intensity histogram of the pixels inside the evaluation mask for every frame (25 bins between 0 and 255). Shannon entropy of one frame (H_f) was then calculated as:

$$H_f = -\sum_{k=1}^{n} p_k \log_2 p_k \tag{9}$$

where p_k is the normalized histogram count of bin k (only including non-zero bins). For each movie the average Shannon entropy across frames is taken. n is the total number of non-zero bins.

The motion energy of a frame (E_f) is calculated as:

$$E_f = \frac{1}{n}\sum_{i=1}^{n} |V_{f,i} - V_{f-1,i}| \tag{10}$$

where $V_{f,i}$ is the intensity value for one pixel i inside the evaluation mask at frame f, n is the total number of pixels inside the mask.

4.7 Retinotopic mapping

To calculate the receptive fields of neurons *in silico* we predicted each neurons response to the full sparse noise stimulus set using the ensembles prediction of 7 SOTA DNEM instances. The response map across pixels for each neuron ($OnR_{h,w,n}$) was defined as:

$$OnR_{h,w,n} = \frac{(R_{stim} - R_{pre})}{R_{pre}} \tag{11}$$

where h and w denote the position of the pixel on the screen, and n the neuron. R_{stim} is the predicted response of the neuron during the stimulus period and R_{pre} during the pre-stimulus period. $OnR_{h,w,n}$ was thresholded at 0.1. The same procedure was done to calculate $OffR_{h,w,n}$. The OnR and $OffR$ maps where smoothed using a 2D Gaussian filter with σ = 2 and then normalized by the maximum value for each neuron. The On and Off receptive field centers were defined as the pixel with the maximum value for each neuron. We calculate the On-Off receptive fields for example neurons as:

$$\text{On-Off receptive fields} = \frac{OnR_n - OffR_n}{Max(OnR_n - OffR_n)} \tag{12}$$

and calculate the population On or Off response as:

$$\text{population On or Off response} = \frac{OnR + OffR}{Max(|OnR + OffR|)} \tag{13}$$

4.8 Reconstruction area calculation

To calculate the retinotopic diameter of a mask we first computed the retinotopic area of each pixel of the movie based on the screen size (31.8 cm by 56.5 cm) and distance from the mouse eye (15 cm). Strictly speaking this is the visuotopic area as it does not take eye position into account, but we refer to it as retinotopic for simplicity. We then take the sum of all pixel areas for an evaluation mask with a given α threshold. Then we define the retinotopic diameter of this area (A) as:

$$\text{retinotopic diameter} = 2\sqrt{\frac{A}{\pi}} \tag{14}$$

4.9 Error map calculation

To calculate the error maps we reconstruct movie clips either from the experimental neural responses or the predicted neural responses given the ground truth movie and took the difference:

$$\text{positive error} = \hat{x} - \hat{x}_{\hat{y}} > 0 \tag{15}$$

$$\text{negative error} = \hat{x} - \hat{x}_{\hat{y}} < 0 \tag{16}$$

where x is the ground truth video, y is the experimental neural activity, $\hat{x}$ is the reconstructed movie from y, $\hat{y}_{\hat{x}}$ is the predicted neural activity from $\hat{x}$, $\hat{y}$ is the predicted neural activity from x, and $\hat{x}_{\hat{y}}$ is the reconstructed movie from $\hat{y}$. Using Fiji [Schindelin et al., 2012] the positive error (LUT: red hot, range 25 to 75), negative error (LUT: cyan hot, range -25 to -75), and ground truth video (LUT: gray scale, range 0 to 255) were then combined into a composite image.

Acknowledgments and Disclosure of Funding

We would like to thank Emmanuel Bauer, Sandra Reinert and the anonymous reviewers for useful input and discussions, and Tom George for the Gaussian noise stimulus set. T.W.M. is funded by The Wellcome Trust (219627/Z/19/Z; 306384/Z/23/Z) and Gatsby Charitable Foundation (GAT3755) and J.B. is funded by EMBO (ALTF 415-2024).

References

Ruslan Baikulov. Solution for Sensorium 2023 Competition (v23.11.22). Zenodo, 2023. URL `https://github.com/lRomul/sensorium`. https://doi.org/10.5281/zenodo.10155151, nov 2023.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv*, 2023. doi: 10.48550/arxiv.2310.19812.

Nora Brackbill, Colleen Rhoades, Alexandra Kling, Nishal P Shah, Alexander Sher, Alan M Litke, and EJ Chichilnisky. Reconstruction of natural images from responses of primate retinal ganglion cells. *eLife*, 9: e58516, 2020. doi: 10.7554/elife.58516.

Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996. ISSN 0885-6125. doi: 10.1007/bf00117832.

Tsai-Wen Chen, Trevor J. Wardill, Yi Sun, Stefan R. Pulver, Sabine L. Renninger, Amy Baohan, Eric R. Schreiter, Rex A. Kerr, Michael B. Orger, Vivek Jayaraman, Loren L. Looger, Karel Svoboda, and Douglas S. Kim. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300, 2013. ISSN 0028-0836. doi: 10.1038/nature12354.

Ye Chen, Peter Beech, Ziwei Yin, Shanshan Jia, Jiayi Zhang, Zhaofei Yu, and Jian K. Liu. Decoding dynamic visual scenes across the brain hierarchy. *PLOS Computational Biology*, 20(8):e1012297, 2024. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1012297.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity. *arXiv*, 2023. doi: 10.48550/arxiv.2305.11675.

Fan L Cheng, Tomoyasu Horikawa, Kei Majima, Misato Tanaka, Mohamed Abdelhack, Shuntaro C Aoki, Jin Hirano, and Yukiyasu Kamitani. Reconstructing visual illusory experiences from human brain activity. *Science Advances*, 9(46):eadj3906, 2023.

Erick Cobos, Taliah Muhammad, Paul G Fahey, Zhiwei Ding, Zhuokun Ding, Jacob Reimer, Fabian H Sinz, and Andreas S Tolias. It takes neurons to understand neurons: Digital twins of visual cortex synthesize neural metamers. *bioRxiv*, pages 2022–12, 2022.

Daniel Deitch, Alon Rubin, and Yaniv Ziv. Representational drift in the mouse visual cortex. *Current Biology*, 31(19):4327–4339.e6, 2021. ISSN 0960-9822. doi: 10.1016/j.cub.2021.07.062.

Winifried Denk, James H. Strickler, and Watt W. Webb. Two-Photon Laser Scanning Fluorescence Microscopy. *Science*, 248(4951):73–76, 1990. ISSN 0036-8075. doi: 10.1126/science.2321027.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *arXiv*, 2017. doi: 10.48550/arxiv.1702.03118.

Aris Fiser, David Mahringer, Hassana K Oyibo, Anders V Petersen, Marcus Leinweber, and Georg B Keller. Experience-dependent spatial expectations in mouse visual cortex. *Nature Neuroscience*, 19(12):1658–1664, 2016. ISSN 1097-6256. doi: 10.1038/nn.4385.

Peter Földiák. Computation and Neural Systems. pages 55–60, 1993. doi: 10.1007/978-1-4615-3254-5_9.

Stef Garasto, Wilten Nicola, Anil A. Bharath, and Simon R. Schultz. Neural Sampling Strategies for Visual Stimulus Reconstruction from Two-photon Imaging of Mouse Primary Visual Cortex. *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, 00:566–570, 2019. doi: 10.1109/ner.2019.8716934.

Andrea Giovannucci, Johannes Friedrich, Pat Gunn, Jérémie Kalfon, Brandon L Brown, Sue Ann Koay, Jiannis Taxidis, Farzaneh Najafi, Jeffrey L Gauthier, Pengcheng Zhou, Baljit S Khakh, David W Tank, Dmitri B Chklovskii, and Eftychios A Pnevmatikakis. CaImAn an open source tool for scalable calcium imaging data analysis. *eLife*, 8:e38173, 2019. doi: 10.7554/elife.38173.

Pieter M. Goltstein, Emily B. J. Coffey, Pieter R. Roelfsema, and Cyriel M. A. Pennartz. In Vivo Two-Photon Ca2+ Imaging Reveals Selective Reward Effects on Stimulus-Specific Assemblies in Mouse Visual Cortex. *The Journal of Neuroscience*, 33(28):11540–11555, 2013. ISSN 0270-6474. doi: 10.1523/jneurosci.1341-12.2013.

Jun Kai Ho, Tomoyasu Horikawa, Kei Majima, Fan Cheng, and Yukiyasu Kamitani. Inter-individual deep image reconstruction via hierarchical neural code conversion. *NeuroImage*, 271:120007, 2023. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2023.120007.

M. Ito, H. Tamura, I. Fujita, and K. Tanaka. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73(1):218–226, 1995. ISSN 0022-3077. doi: 10.1152/jn.1995.73.1.218.

Ovidiu Jurjut, Petya Georgieva, Laura Busse, and Steffen Katzner. Learning Enhances Sensory Processing in Mouse V1 before Improving Behavior. *The Journal of Neuroscience*, 37(27):6460–6474, 2017. ISSN 0270-6474. doi: 10.1523/jneurosci.3485-16.2017.

Fatemeh Kalantari, Karim Faez, Hamidreza Amindavar, and Soheila Nazari. Improved image reconstruction from brain activity through automatic image captioning. *Scientific Reports*, 15(1):4907, 2025.

Takahiro Kanamori and Thomas D. Mrsic-Flogel. Independent response modulation of visual cortical neurons by attentional and behavioral states. *Neuron*, 110(23):3907–3918, 2022. ISSN 0896-6273. doi: 10.1016/j.neuron.2022.08.028.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. ISSN 0028-0836. doi: 10.1038/nature06713.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv*, 2014. doi: 10.48550/arxiv.1412.6980.

Naoko Koide-Majima, Shinji Nishimoto, and Kei Majima. Mental image reconstruction from human brain activity: Neural decoding of mental imagery via deep neural network-based bayesian estimation. *Neural Networks*, 170:349–363, 2024.

Ganit Kupershmidt, Roman Beliy, Guy Gaziv, and Michal Irani. A Penny for Your (visual) Thoughts: Self-Supervised Reconstruction of Natural Movies from Brain Activity. *arXiv*, 2022. doi: 10.48550/arxiv.2206.03544.

Wenyi Li, Shengjie Zheng, Yufan Liao, Rongqi Hong, Chenggang He, Weiliang Chen, Chunshan Deng, and Xiaojian Li. The brain-inspired decoder for natural visual image reconstruction. *Frontiers in Neuroscience*, 17:1130606, 2023. ISSN 1662-4548. doi: 10.3389/fnins.2023.1130606.

Wu Li. Perceptual Learning: Use-Dependent Cortical Plasticity. *Annual Review of Vision Science*, 2(1):1–22, 2015. ISSN 2374-4642. doi: 10.1146/annurev-vision-111815-114351.

Alexander Mordvintsev, Nicola Pezzotti, Ludwig Schubert, and Chris Olah. Differentiable image parameterizations. *Distill*, 3(7), July 2018. ISSN 2476-0757. doi: 10.23915/distill.00012. URL `http://dx.doi.org/10.23915/distill.00012`.

Cristopher M. Niell and Michael P. Stryker. Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron*, 65(4):472–479, 2010. ISSN 0896-6273. doi: 10.1016/j.neuron.2010.01.033.

Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, 21(19):1641–1646, 2011. ISSN 0960-9822. doi: 10.1016/j.cub.2011.08.031.

Yuichiro Nomura, Shohei Ikuta, Satoshi Yokota, Junpei Mita, Mami Oikawa, Hiroki Matsushima, Akira Amano, Kazuhiro Shimonomura, Yasuhiro Seya, and Chieko Koike. Evaluation of critical flicker-fusion frequency measurement methods using a touchscreen-based visual temporal discrimination task in the behaving mouse. *Neuroscience Research*, 148:28–33, 2019. ISSN 0168-0102. doi: 10.1016/j.neures.2018.12.001.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. doi: 10.1038/s41598-023-42891-8.

Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51:195–203, 2019.

Pawel Pierzchlewicz, Konstantin Willeke, Arne Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, et al. Energy guided diffusion for generating neurally exciting images. *Advances in Neural Information Processing Systems*, 36:32574–32601, 2023.

Jasper Poort, Adil G. Khan, Marius Pachitariu, Abdellatif Nemri, Ivana Orsolic, Julija Krupic, Marius Bauza, Maneesh Sahani, Georg B. Keller, Thomas D. Mrsic-Flogel, and Sonja B. Hofer. Learning Enhances Sensory and Multiple Non-sensory Representations in Primary Visual Cortex. *Neuron*, 86(6):1478–1490, 2015. ISSN 0896-6273. doi: 10.1016/j.neuron.2015.05.037.

G.T. Prusky and R.M. Douglas. Characterization of mouse cortical spatial vision. *Vision Research*, 44(28):3411–3418, 2004. ISSN 0042-6989. doi: 10.1016/j.visres.2004.09.001.

Zarina Rakhimberdina, Quentin Jodelet, Xin Liu, and Tsuyoshi Murata. Natural Image Reconstruction From fMRI Using Deep Learning: A Survey. *Frontiers in Neuroscience*, 15:795488, 2021. ISSN 1662-4548. doi: 10.3389/fnins.2021.795488.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. ISSN 1097-6256. doi: 10.1038/4580.

Jacob Reimer, Emmanouil Froudarakis, Cathryn R. Cadwell, Dimitri Yatsenko, George H. Denfield, and Andreas S. Tolias. Pupil Fluctuations Track Fast Switching of Cortical States during Quiet Wakefulness. *Neuron*, 84(2):355–362, 2014. ISSN 0896-6273. doi: 10.1016/j.neuron.2014.09.033.

Ziqi Ren, Jie Li, Xuetong Xue, Xin Li, Fan Yang, Zhicheng Jiao, and Xinbo Gao. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228:117602, 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117602.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv*, 2021. doi: 10.48550/arxiv.2112.10752.

Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676–682, 2012.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 12(1):5170, 2023. ISSN 0028-0836. doi: 10.1038/s41586-023-06031-6.

Joseph W. Schumacher, Matthew K. McCann, Katherine J. Maximov, and David Fitzpatrick. Selective enhancement of neural coding in V1 underlies fine-discrimination learning in tree shrew. *Current Biology*, 2022. ISSN 0960-9822. doi: 10.1016/j.cub.2022.06.009.

Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A Norman, and Tanishq Mathew Abraham. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. *arXiv*, 2023. doi: 10.48550/arxiv.2305.18274.

Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-End Deep Image Reconstruction From Human Brain Activity. *Frontiers in Computational Neuroscience*, 13:21, 2019a. ISSN 1662-5188. doi: 10.3389/fncom.2019.00021.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1):e1006633, 2019b. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1006633.

Fabian Sinz, Alexander S Ecker, Paul Fahey, Edgar Walker, Erick Cobos, Emmanouil Froudarakis, Dimitri Yatsenko, Zachary Pitkow, Jacob Reimer, and Andreas Tolias. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *Advances in neural information processing systems*, 31, 2018.

Garrett B Stanley, Fei F Li, and Yang Dan. Reconstruction of Natural Scenes from Ensemble Responses in the Lateral Geniculate Nucleus. *The Journal of Neuroscience*, 19(18):8036–8042, 1999. ISSN 0270-6474. doi: 10.1523/jneurosci.19-18-08036.1999.

Andrea Tacchetti, Leyla Isik, and Tomaso A. Poggio. Invariant Recognition Shapes Neural Representations of Visual Input. *Annual Review of Vision Science*, 4(1):403–422, 2018. ISSN 2374-4642. doi: 10.1146/annurev-vision-091517-034103.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104–1116, 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.06.062.

Polina Turishcheva, Paul G Fahey, Laura Hansel, Rachel Froebe, Kayla Ponder, Michaela Vystrčilová, Konstantin F Willeke, Mohammad Bashiri, Eric Wang, Zhiwei Ding, Andreas S Tolias, Fabian H Sinz, and Alexander S Ecker. The Dynamic Sensorium competition for predicting large-scale mouse visual cortex activity from videos. *arXiv*, 2023. doi: 10.48550/arxiv.2305.19654.

Polina Turishcheva, Paul Fahey, Michaela Vystrčilová, Laura Hansel, Rachel Froebe, Kayla Ponder, Yongrong Qiu, Konstantin Willeke, Mohammad Bashiri, Ruslan Baikulov, et al. Retrospective for the dynamic sensorium competition for predicting large-scale mouse primary visual cortex activity from videos. *Advances in Neural Information Processing Systems*, 37:118907–118929, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv*, 2017. doi: 10.48550/arxiv.1706.03762.

Edgar Y. Walker, Fabian H. Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G. Fahey, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S. Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 22(12):2060–2065, 2019. ISSN 1097-6256. doi: 10.1038/s41593-019-0517-x.

Eric Y Wang, Paul G Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, et al. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640(8058):470–477, 2025.

Konstantin F Willeke, Paul G Fahey, Mohammad Bashiri, Laura Hansel, Christoph Blessing, Konstantin-Klemens Lurz, Max F Burg, Santiago A Cadena, Zhiwei Ding, Kayla Ponder, et al. Retrospective on the sensorium 2022 competition. In *NeurIPS 2022 Competition Track*, pages 314–333. PMLR, 2023a.

Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization. *bioRxiv*, page 2023.05.12.540591, 2023b. doi: 10.1101/2023.05.12.540591.

Ji Xia, Tyler D. Marks, Michael J. Goard, and Ralf Wessel. Stable representation of a naturalistic movie emerges from episodic activity with gain variability. *Nature Communications*, 12(1):5170, 2021. doi: 10.1038/s41467-021-25437-2.

Takashi Yoshida and Kenichi Ohki. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications*, 11(1):872, 2020. doi: 10.1038/s41467-020-14645-x.

Yichen Zhang, Shanshan Jia, Yajing Zheng, Zhaofei Yu, Yonghong Tian, Siwei Ma, Tiejun Huang, and Jian K. Liu. Reconstruction of natural visual scenes from neural spikes with deep neural networks. *Neural Networks*, 125:19–30, 2020. ISSN 0893-6080. doi: 10.1016/j.neunet.2020.01.033.

5 Code

The code is available at `https://github.com/Joel-Bauer/movie_reconstruction_code`.

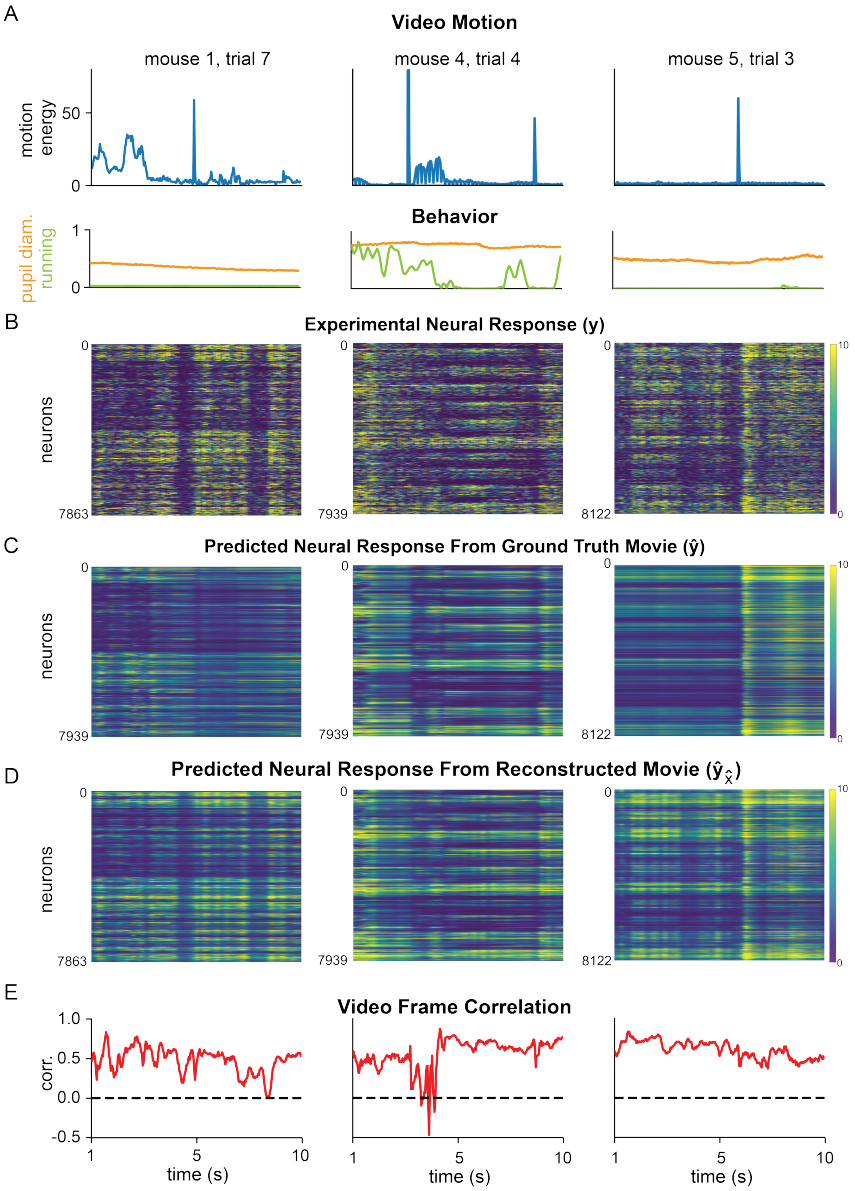
A Appendix / supplemental material

Supplementary Video 1: Reconstructed natural videos from mouse brain activity. Odd rows are ground truth (GT) movie clips presented to mice. Even rows are the reconstructed movies from the activity of ≈8000 V1 neurons. Reconstructed movies are smoothed ($\sigma = 0.5$ pixels), masked, and contrast (std) and luminance (mean) matched to ground truth movies.

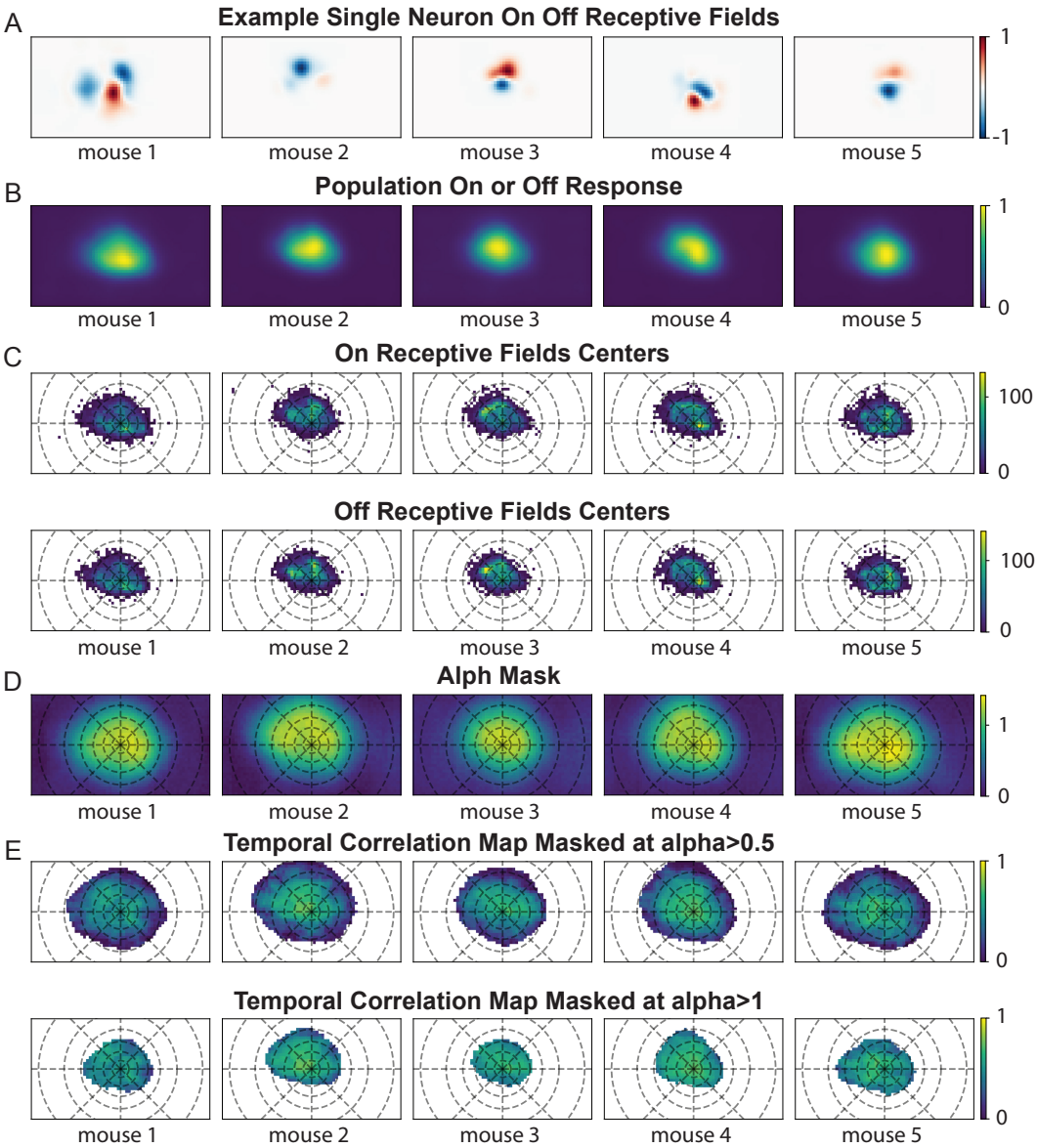
Supplementary Video 2: Gaussian noise stimuli and reconstructions. Odd rows are ground truth (GT) video inputs to the model. Even rows are the reconstructed videos from the predicted neuronal activity for 1 mouse. Reconstructed movies are masked, and contrast (std) and luminance (mean) matched to ground truth videos.

Supplementary Video 3: Drifting grating stimuli and reconstructions. Odd rows are ground truth (GT) video inputs to the model. Even rows are the reconstructed videos from the predicted neuronal activity for 1 mouse. Reconstructed movies are masked, and contrast (std) and luminance (mean) matched to ground truth videos.

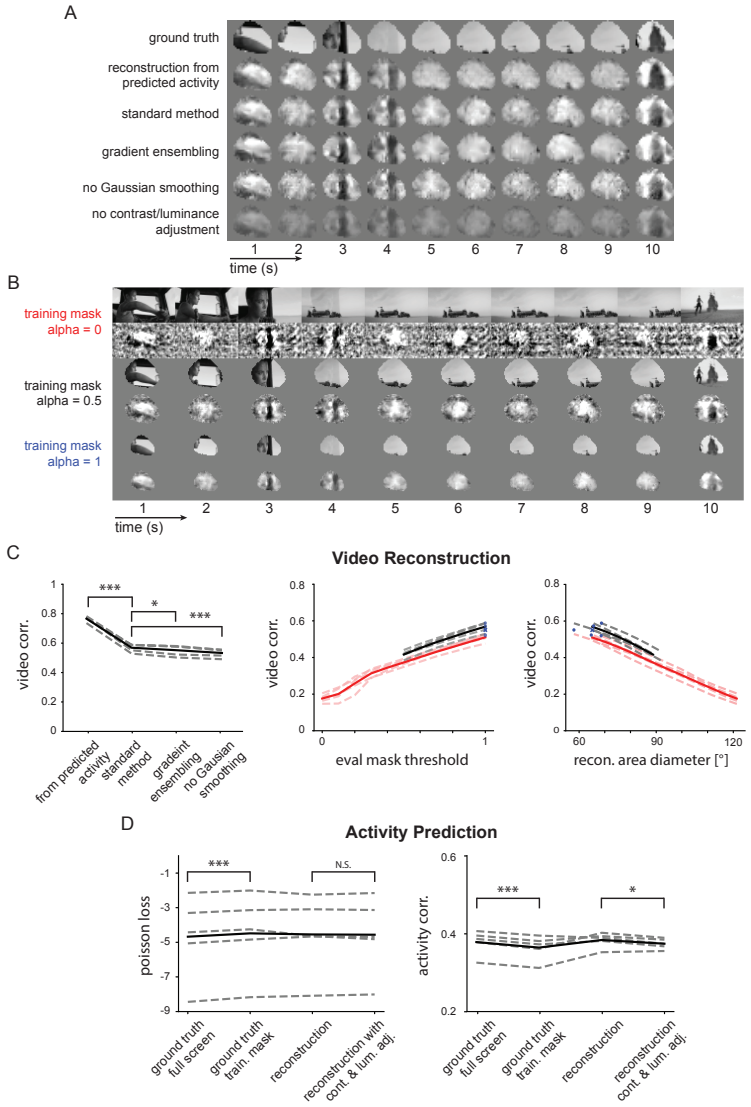
Supplementary Video 4: Reconstruction error maps. Pixel error = reconstructions from experimental neural activity – reconstructions from expected neural activity. Over- & underestimations of pixel values as hot & cold heat maps, respectively. Ground truth movies in gray. Each row is 10 trials from 1 mouse played at half speed.



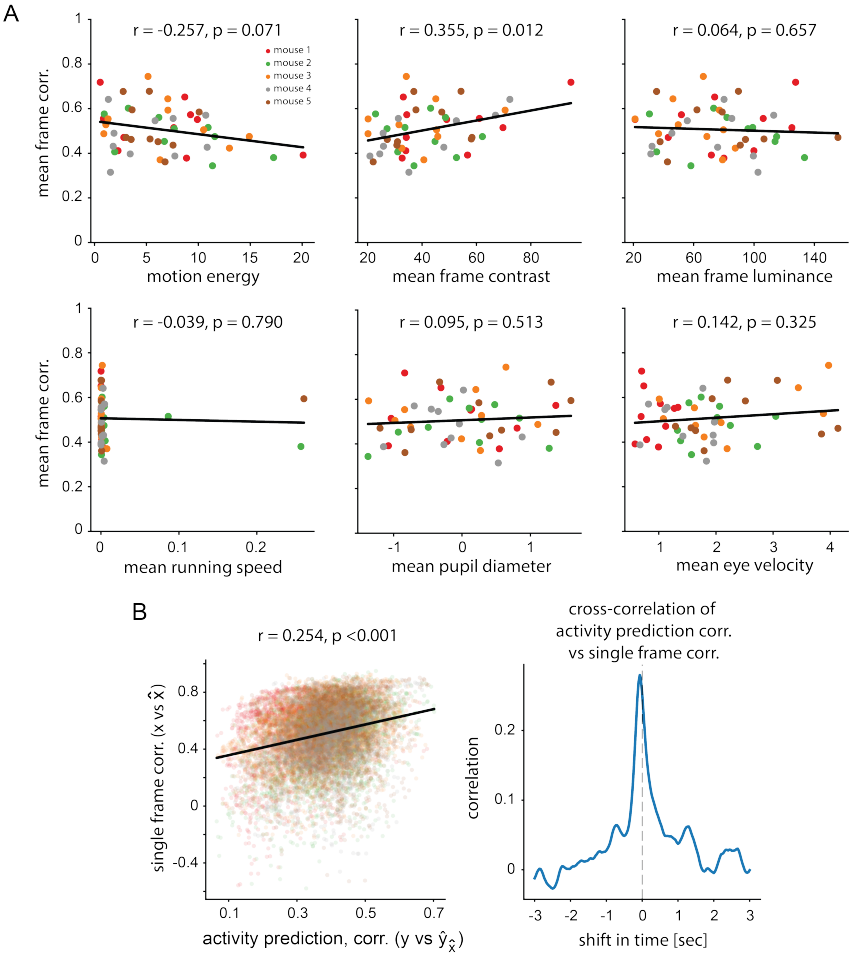
Supplementary Figure S1: Summary ethogram of SOTA DNEM inputs, output predictions, and video reconstruction over time for three videos from three mice (same as Figure 2A). A) Top: motion energy of the input video. Bottom: pupil diameter and running speed of the mouse during the video. B) Ground truth neuronal activity. C) Predicted neuronal activity in response to input video and behavioural parameters. D) predicted neuronal activity given reconstructed video and ground truth behaviour as input. E) Frame by frame correlation between reconstructed and ground truth video.



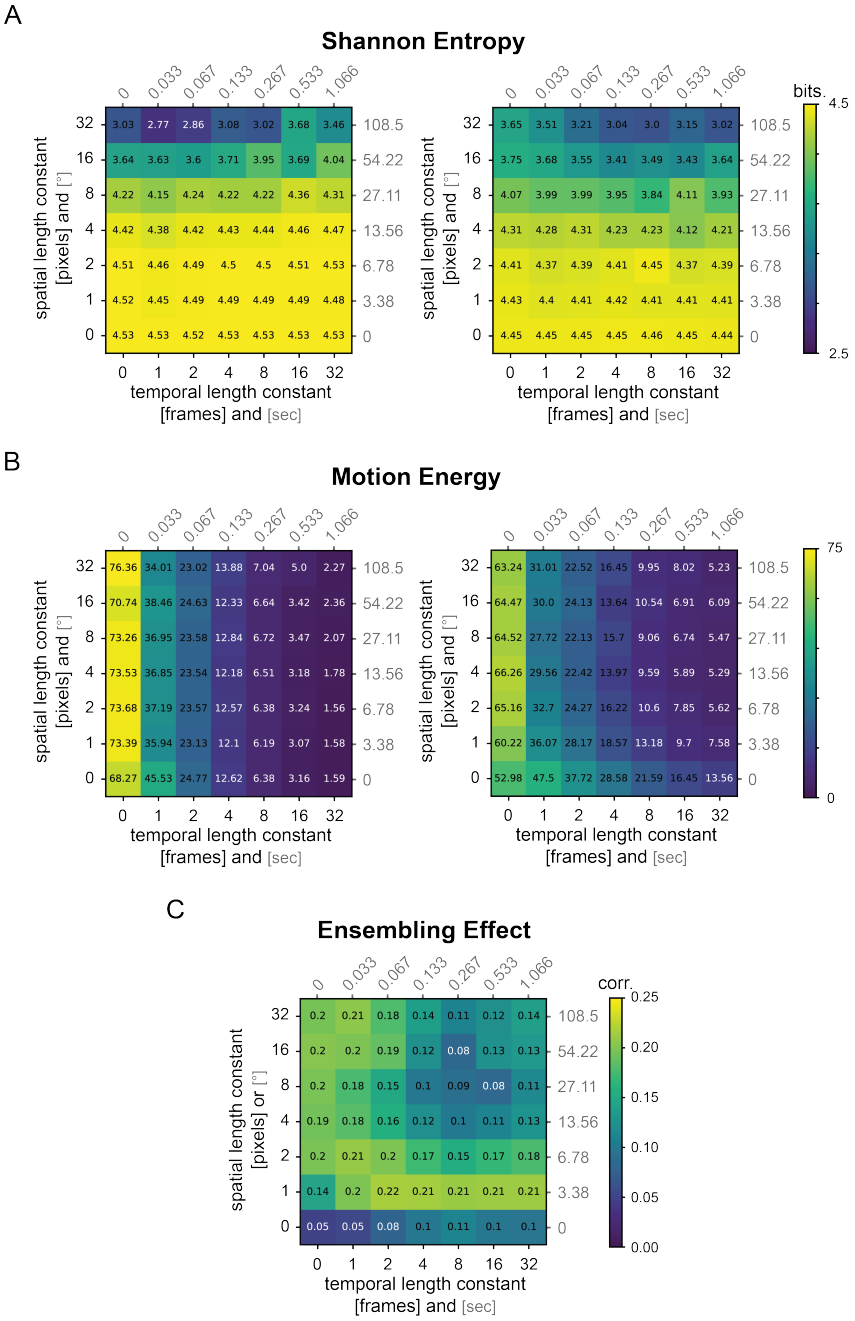
Supplementary Figure S2: Receptive fields and transparency masks. A) One example receptive field for one neuron from each mouse mapped using on & off patch stimuli *in silico*. B) Average population receptive fields from each mouse. C) Distribution of on and off receptive field centers for each moue. D) Unthresholded alpha masks, i.e. transparency masks, for each mouse. E) Pixel-wise temporal correlation between ground truth and reconstructed videos with either the training or the evaluation mask applied. Dashed lines in C-E indicate retinotopic eccentricity in steps of 10°. Plot limits correspond to screen size.



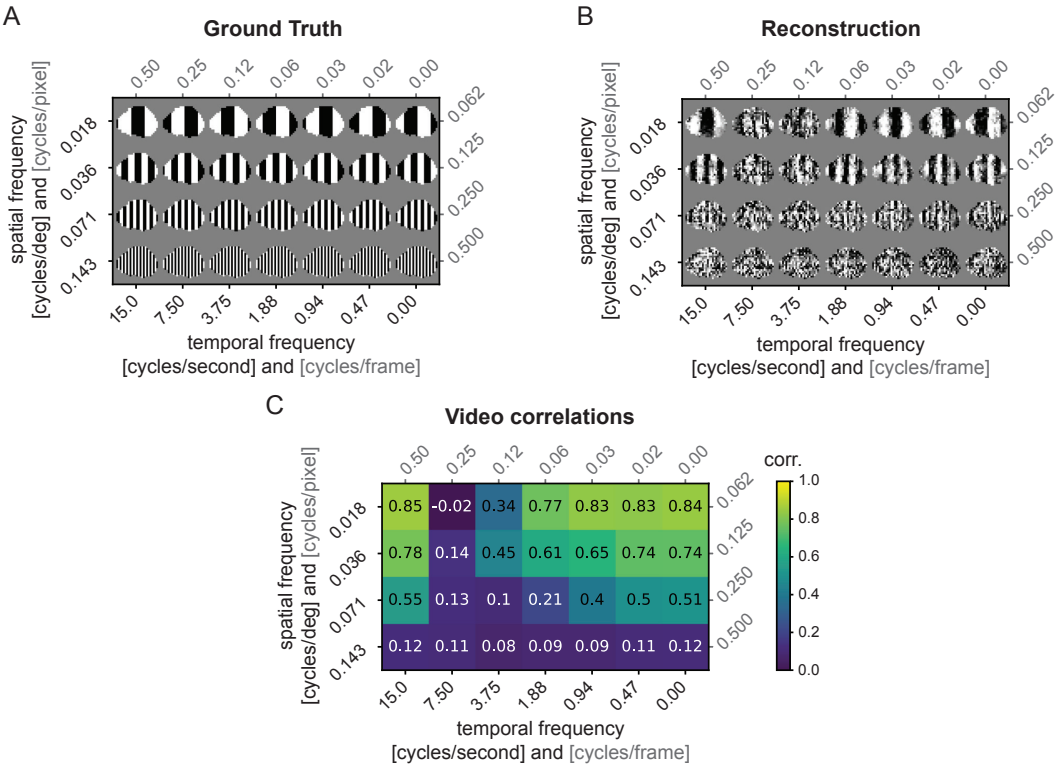
Supplementary Figure S3: Variations on reconstruction method. A) Example movie frames (from mouse 4 trial 1) using variations of the reconstruction method. B) Same as A but using different training mask thresholds. No evaluation mask is applied here. C) Left: reconstruction performance for different reconstruction versions. Version without contrast and luminance adjustment is not included because video correlation is always calculated before contrast and luminance adjustment. Reconstruction from predicted vs standard method (34.7% increase; paired t-test p = $1.5 * 10^{-}5$ n = 5 mice), standard method vs gradient ensembling (3.00% decrease; paired t-test p = 0.0198, n = 5 mice), standard method vs no Gaussian smoothing (6.28% decrease; paired t-test p = $6.26 * 10^{-6}$, n = 5 mice). Middle: reconstruction performance with different evaluation mask thresholds compared across the three training mask thresholds shown in B. Right: same as middle but plotting the mask diameter for each evaluation mask threshold. D) Neural activity prediction performance for different movie inputs. Left: Poisson loss, used to train the DNEM and movie reconstruction. Predicted activity from full video vs masked with alpha = 0.5 (5.02% increase, t-test p = $3.71 * 10^{-}6$, n = 5 mice), reconstruction vs reconstruction after contrast & luminance matching to ground truth video (0.227% increase, paired t-test p = 0.776, n = 5 mice). Right: correlation across all neurons and frames (note this is a different metric to the one used in the Sensorium competition). Predicted activity from full video vs masked with alpha = 0.5 (3.73% increase, t-test p = $2.49 * 10^{-}5$, n = 5 mice), reconstruction vs reconstruction after contrast & luminance matching to ground truth video (2.49% increase, paired t-test p = 0.0195, n = 5 mice). In C-D dashed lines are single mice, solid lines are means across mice.



Supplementary Figure S4: Reconstruction performance correlates with frame contrast but not with behavioral parameters. A) Pearson's correlation between mean frame correlation per movie, and 3 movie parameters and 3 behavioral parameters. Linear fit as black line. B) Left: Pearson's correlation between activity prediction accuracy and movie reconstruction accuracy. Right: cross-correlation plot of frame-by-frame activity prediction accuracy and video frame correlation. In other words, the more predictable the neural activity, the better the reconstruction performance.



Supplementary Figure S5: Gaussian Noise reconstruction. A) Average frame Shannon entropy (a measure of variance in the spatial domain) across Gaussian noise stimuli with various spatial and temporal Gaussian length constants. Left: ground truth stimuli. Right: reconstructions from predicted activity. B) Same as A but for motion energy (a measure of variance in the temporal domain). C) Ensembling effect for each stimulus. Video correlation for ensembled prediction (average videos from 7 model instances) minus the mean video correlation across the 7 individual model instances.



Supplementary Figure S6: Reconstruction of drifting grating stimuli with different spatial and temporal frequencies using predicted activity. A) Example drifting grating stimuli (rightwards moving) masked with the evaluation mask for one mouse. Shown is the 31st frame of a 2 second video. B) Reconstructed drifting grating stimuli with SOTA DNEM predicted neuronal activity as the target (see also Supplementary Video 3: YouTube link). C) Pearson's correlation between ground truth (A) and reconstructed videos (B) across the range of spatial and temporal frequencies. For each stimulus type the average correlation across 4 directions (up, down, left, right) reconstructed from the SOTA DNEM of 1 mouse is given. Interestingly, video correlation at 15 cycles/second (half the video frame rate of 30 Hz) is much higher than 7.5 cycles/second. This is an artifact of using predicted responses rather than true neural responses. The DNEM input layer convolution has dilation 2. The predicted activity is therefore based on every second frame, with the effect that the activity is predicted as the response of 2 static images which are then interleaved.

Algorithm 1: Movie reconstruction

Parameters:
Dynamic Neural Encoding Model (DNEM), Ground truth neuronal response y, predicted neuronal responses $\hat{y}$, predicted input video $\hat{x}$, video width $w = 64$, video height $h = 36$, number of frames $n = 300$, transparency mask α, sliding window size $k = 32$, sliding window stride $s = 8$, total number of windows $N \leftarrow \lfloor 2 + (n-k)/s \rfloor$, learning rate $lr = 1000$, $\beta_1 = 0.9$, loss function $Loss(y, \hat{y}) = \sum \hat{y}_{neuron} - y_{neuron} \log(\hat{y}_{neuron} + 10^{-8})$, number of epochs $epochs = 1000$, number of model instances $q = 7$

Objective: $F(y, \hat{x}) = Loss(y, DNEM(\hat{x}))$

Initialize variables:
$\hat{x}_{f,\text{h},\text{w}} \leftarrow 125$
$G_{i,f,\text{h},\text{w}} \leftarrow 0$
$g_{f,\text{h},\text{w}} \leftarrow 0$
$m_{f,\text{h},\text{w}} \leftarrow 0$

for iteration $t = 1, \ldots, epochs$ **do**

  Gradients across all windows:
  for iteration $i = 1, \ldots, N$ **do**
    if $i < N$ **then**
      $f \leftarrow [s(i-1), \ldots, k + s(i-1)]$
    else if $i = N$ **then**
      $f \leftarrow [n-k, \ldots, n]$
    end if

    $G_{i,f} \leftarrow \nabla F(y_f, \hat{x}_f)$
    $G_{i,f} \leftarrow \frac{G_{i,f}}{\|G_{i,f}\| + 10^{-6}}$
  end for

  Average gradients across windows: $g \leftarrow \frac{1}{N} \sum_{i=1}^{N} G_i$

  Clip and mask gradients:
  $g \leftarrow \text{clip}(g, -1, 1)$
  $g_{\text{f,h,w}} \leftarrow g_{\text{f,h,w}} \odot (\alpha_{\text{h,w}} > 0.5)$

  Update movie:
  if $t < 10$ **then**
    $lr_{\text{current}} \leftarrow lr \frac{t}{10}$
  else if $t > 10$ **then**
    $lr_{\text{current}} \leftarrow lr$
  end if

  $lr_{\text{current}} \leftarrow lr_{\text{current}} \sqrt{1 - 0.999^t} / (1 - \beta_1^t)$
  $m \leftarrow \beta_1 \cdot m + (1 - \beta_1) \cdot g$
  $\hat{m} \leftarrow \frac{m}{1 - \beta_1^t}$
  $\hat{x} \leftarrow \hat{x} - lr_{\text{current}} \cdot \hat{m}$

  Clip movie: $\hat{x} \leftarrow \text{clip}(\hat{x}, 0, 255)$

end for

Ensembling: $\hat{x} \leftarrow \frac{1}{q} \sum_{i=1}^{q} \hat{x}$

Denoise (optional): $\hat{x} \leftarrow GaussianBlur3D(\hat{x}, \sigma = (0.5, 0.5, 0.5))$

Mask movie: $\hat{x}_{\text{f,h,w}} \leftarrow \hat{x}_{\text{f,h,w}} \odot (\alpha_{\text{h,w}} > 1)$

Algorithm 2: Mask training

Parameters:
Dynamic Neural Encoding Model (DNEM), Ground truth neuronal response y, predicted neuronal responses $\hat{y}$, ground truth input video x, background video $background$, video width $w = 64$, video height $h = 36$, frames f, number of frames $n = 32$, transparency mask α, learning rate $lr = 10$, loss function $Loss(y, \hat{y}, \alpha) = MSE(y, \hat{y}(1 - \bar{\alpha}))$, number of epochs $epochs = 1000$, alpha blending function $Blend(x, background, \alpha) = x \odot \alpha + background \odot (1 - \alpha)$

Objective:
$F(y, x, background, \alpha) = Loss(y, DNEM(Blend(x, background, \text{clip}(\alpha, -1, 1))), \alpha)$

Initialize variables:
$\alpha_{\text{h,w}} \leftarrow U(0, 0.5)$

for iteration $t = 1, \ldots, epochs$ **do**

  $x \leftarrow$ get random video
  $background \leftarrow$ get different random video
  $background \leftarrow Flip(background, dim = 1)$
  $background \leftarrow Flip(background, dim = 2)$
  $background \leftarrow Flip(background, dim = 3)$

  Gradients with respect to transparency mask:
  $G \leftarrow \frac{\partial F(y, x, \text{background}, \alpha)}{\partial \alpha}$
  $G \leftarrow \frac{G}{\|G\| + 10^{-6}}$

  Average gradients across frames:
  $g \leftarrow \frac{1}{n} \sum_{f=1}^{n} G_f$

  Clip and mask gradients:
  $g \leftarrow \text{clip}(g, -1, 1)$
  $g \leftarrow GaussianBlur2D(g, \sigma = (5, 5))$

  Update mask:
  $\alpha \leftarrow \alpha - lr \cdot g$

end for