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 imagery 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 videos presented to mice, from the activity of neurons in their visual cortex. Using our method of video optimization via gradient descent 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 $\approx$ 2-fold increase in pixel-by-pixel correlation compared to previous reconstructions of static images from mouse V1, while also capturing temporal dynamics. We find that critical for high-quality reconstructions are the number of neurons in the dataset and the use of model ensembling.

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 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. 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 [4]. 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 [5, 7, 9]. However, the pixel level similarities are still quite low. These methods are highly useful, particularly 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 [12], thalamo-cortical neurons in lateral genicular nucleus [13], and primary visual cortex [14, 15]. Recently, deep

Preprint. Under review.

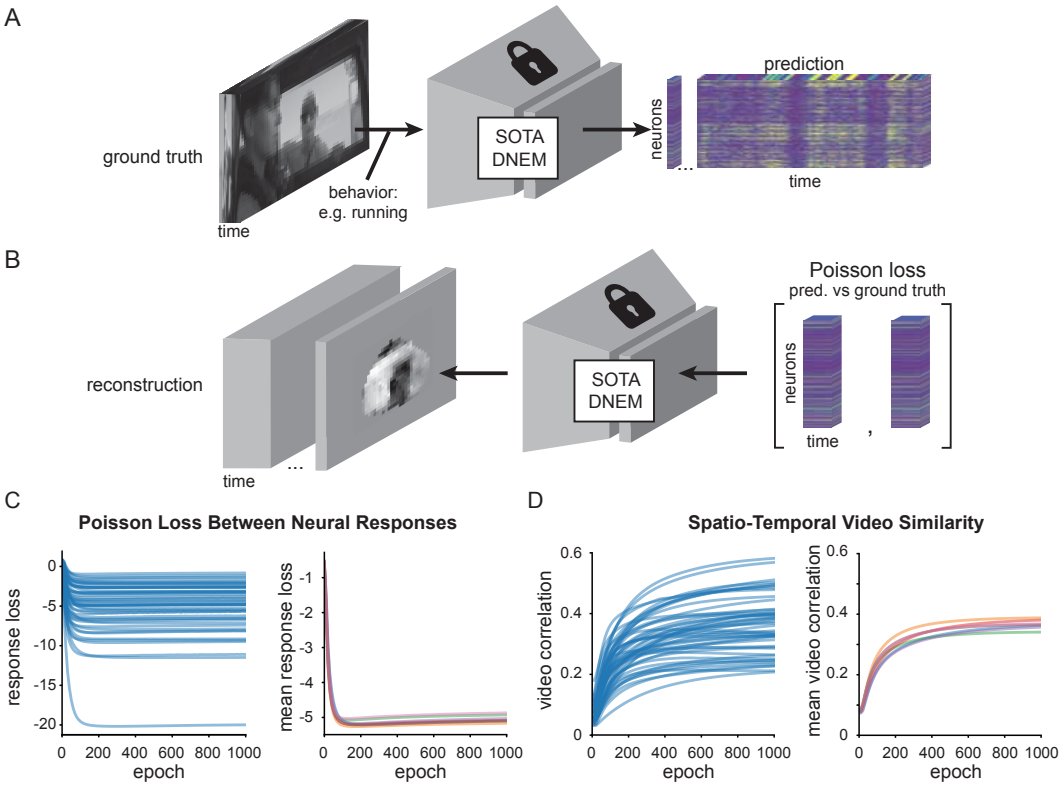


Figure 1: Video reconstruction from neuronal activity in mouse V1 (data provided by the Sensorium 2023 competition; [22]) using a state-of-the-art (SOTA) dynamic neural encoding model (DNEM; [19]). A) SOTA DNEMs predict neuronal activity from mouse primary visual cortex, given a video and behavioural input. B) We use a SOTA DNEM to reconstruct part of the input video given neural activity, using gradient descent on 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 5 identical models trained on different sets of data. D) Spatio-temporal (pixel-by-pixel) correlation between reconstructed video and ground truth video.

nonlinear neural networks have been used with promising results to reconstruct static images from retina [16, 17] and in particular monkey V4 extracellular recordings [17, 18].

Here, we present a method for the reconstruction of 10-second movie clips using two-photon calcium imaging data recorded in mouse V1. We use a state-of-the-art (SOTA) dynamic neural encoding model (DNEM) [19] which predicts neuronal activity based on video input as well as behavior (Figure 1A), allowing us to reconstruct videos despite the fact that neuronal activity in mouse V1 is heavily modulated by behavioral factors such as running speed [20] and pupil diameter (correlated with arousal; [21]). We then quantify the spatio-temporal limits of the reconstruction approach and identify key aspects of our method necessary for good performance.

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

2.1 Source data

We reconstructed videos seen by mice based on the calcium activity of neurons measured using two-photon imaging [23] in V1. The data was provided by the Sensorium 2023 competition [22, 24] (`https://gin.g-node.org/pollytur/Sensorium2023Data` and `https://gin.g-node.org/pollytur/sensorium_2023_dataset`). In short, the data included the movies presented to

the mice, pupil position, pupil diameter, running speed and neural activity. The grayscale movies were 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 provided movies were spatially downsampled to 36 by 64 pixels, corresponding to a 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 of GCaMP6s [25] fluorescence, extracted and deconvolved using the CAIMAN pipeline [26]. For each of the 10 mice, the activity of ≈ 8000 neurons was provided. The different data types were resampled to 30 Hz.

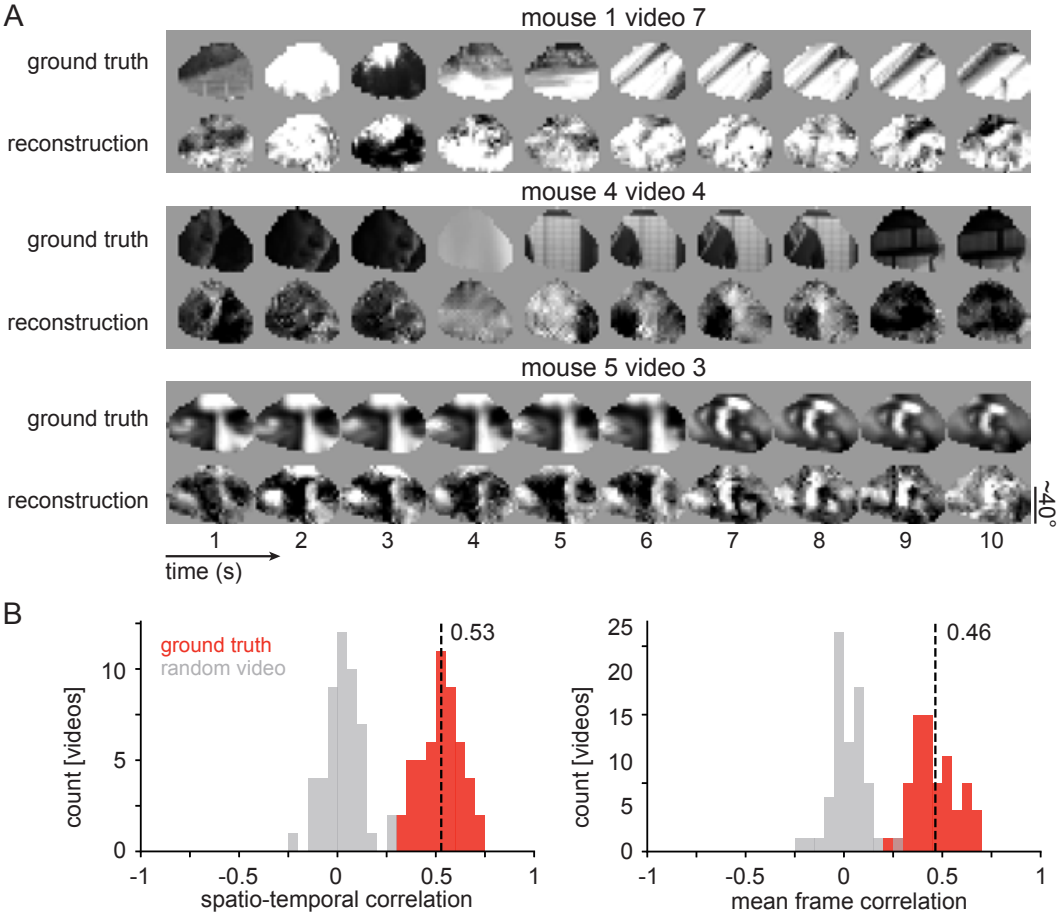


Figure 2: Reconstruction performance. A) Three reconstructions of 10s videos from different mice (see Supplementary Video 1 for full set: YouTube link). Reconstructions have been luminance (mean pixel value across video) and contrast (standard deviation of pixel values across video) matched to ground truth. B) The reconstructed videos have high correlation to ground truth in both spatio-temporal correlation (mean Pearson's correlation r = 0.528 with 95% CIs 0.498 to 0.557, t-test between ground truth and random video p = $6.19 * 10^{-45}$, n = 50 videos from 5 mice) and mean frame correlation (mean Pearson's correlation r = 0.465 with 95% CIs 0.433 to 0.496, t-test between ground truth and random video p = $1.27 * 10^{-40}$, n = 50 videos from 5 mice).

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

We used the winning model of the Sensorium 2023 competition which achieved a score of 0.301 ([24] single trial correlation between predicted and ground truth neural activity; Figure 1A and Figure 3A-C). The code was downloaded from `https://github.com/lRomul/sensorium` [19]. 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 core consisted of

factorized 3D convolution layers, the cortex consisted of three fully connected layers and the readout consisted of a 1D convolution 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 and the output neural activity prediction was 32.

The original ensemble of models consisted of 7 model instances trained on a non-overlapping 7-fold cross-validation split of all available Sensorium 2023 competition data (around 8 min of data per fold). Each model instance was trained on 6 of 7 data folds, with a different data fold excluded from training and reserved for validation 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 data fold 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). Figure 5 (reconstruction of Gaussian noise) and supplementary Figure S1 (reconstruction of drifting gratings) was done using the original model weights (this will be changed in a future version of this preprint). Note that we did not use the distilled version of the models (see `https://github.com/lRomul/sensorium`) in order to avoid "contamination" of the held-out dataset.

2.3 Video reconstruction via gradient descent

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. However, due to the fully connected layers within the SOTA DNEM, gradients still propagated to the full video frame and produced non-sensical results along the periphery of the video frames. Inspired by [27, 28] 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. 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 $\hat{y}$ and the predicted responses y scaled by the average weight of the transparency mask $\bar{\alpha}$:

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

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

where n is the total number of neurons. The mask was initialized as uniform noise between 0 and 0.05. At each epoch the neural 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.

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.

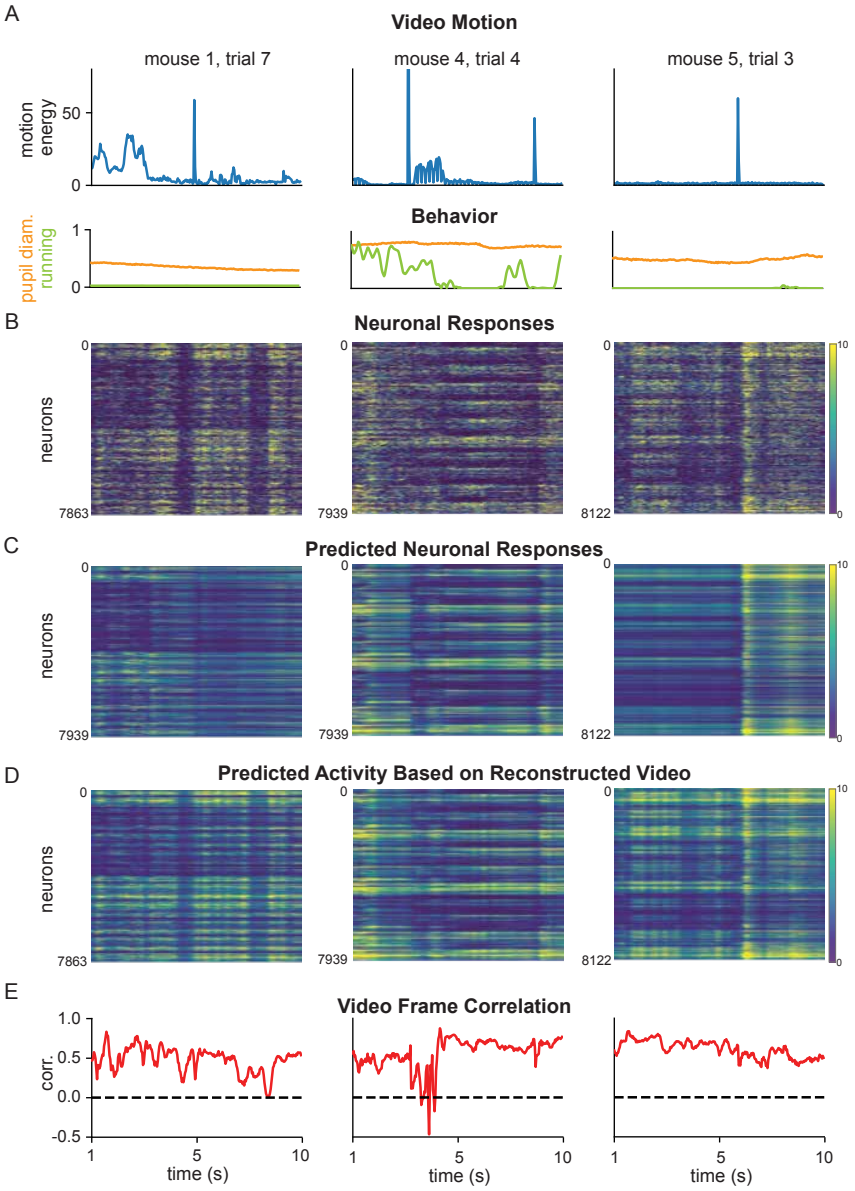


Figure 3: Summary ethogram of SOTA DNEM inputs, output predictions, and video reconstruction over time for three videos from three mice (same as Figure 2A). A) Top: motion energy of the input video. Bottom: pupil diameter and running speed of the mouse during the video. B) Ground truth neuronal activity. 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.

The input optimization through gradient descent approach we use was largely inspired by the optimization of maximally exciting images [29] and the reconstruction of static images from monkey V4 extracellular recordings [18]. The input videos were initialized as uniform grey values. The behavioral parameters (Figure 3A) were added as additional channels, i.e. these were not reconstructed but given. The neuronal activity in response to the input video and behavioral parameters was predicted using the SOTA DNEM for a sliding window of 32 frames with a stride of 8 frames. We saw slightly better results with a stride of 2 frames but this did not warrant the increase in training time in our case. For each batch of frames, 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 y are the predicted responses and $\hat{y}$ are the ground truth responses. The gradients of the loss with respect to each pixel of the input video were calculated for each batch of frames and averaged across batches. 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 [30] ($\beta_1 = 0.9$) 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.

As the loss between predicted (Figure 3D) and ground truth responses (Figure 3B) 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 separate data) and averaged them to get the final reconstruction. Optimizing each 10-second video for 1000 epochs took 60 min using a desktop with an RTX4070 GPU.

2.4 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. To evaluate performance of the video reconstructions we correlated either all pixels from all time points between ground truth and reconstructed videos (Pearson's correlation r = 0.53; to quantify temporal and spatial similarity), or the average correlation between all sets of frames (Pearson's correlation r = 0.46; to quantify just spatial similarity)(Figure 2B and Figure 3E). Importantly, this represents a $\approx$ 2x improvement over previous static image reconstructions from V1 in awake mice (0.33 +/- 0.015 s.e.m for anaesthetized mice and 0.23 +/- 0.02 s.e.m for awake mice) [15] over a similar retinotopic area ($\approx 60°$ diameter) while also capturing temporal dynamics.

2.5 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 4). A technique known as bagging, and more generally ensembling [31]. 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.

In this paper we averaged over 7 model instances which gave a performance increase of 44.1%, but the largest gain in performance, 19.5%, came from averaging across just 2 models (Figure 4). Doubling the number of models to 4 increased the performance by another 8.12%. Overall, although ensembling over models trained on separate data splits is a computationally expensive method, it substantially improved reconstruction quality.

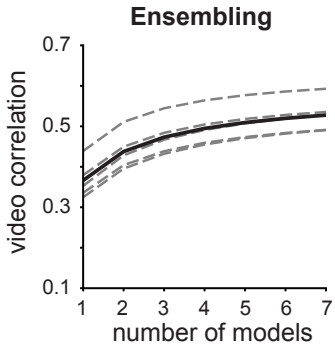


Figure 4: Model ensembling. Mean video correlation is improved when predictions from multiple models are averaged. Dashed lines are individual animals, solid line is mean. One-way repeated measures ANOVA p = $1.14 * 10^{-4}$. Bonferroni corrected paired t-test outcomes between consecutive ensemble sizes are all p < 0.001, n = 5 mice.

3 Assessing reconstruction quality limitations

3.1 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. We therefore assessed the model's ability to reconstruct synthetic stimuli at varying spatial and temporal resolutions in a noise-free scenario.

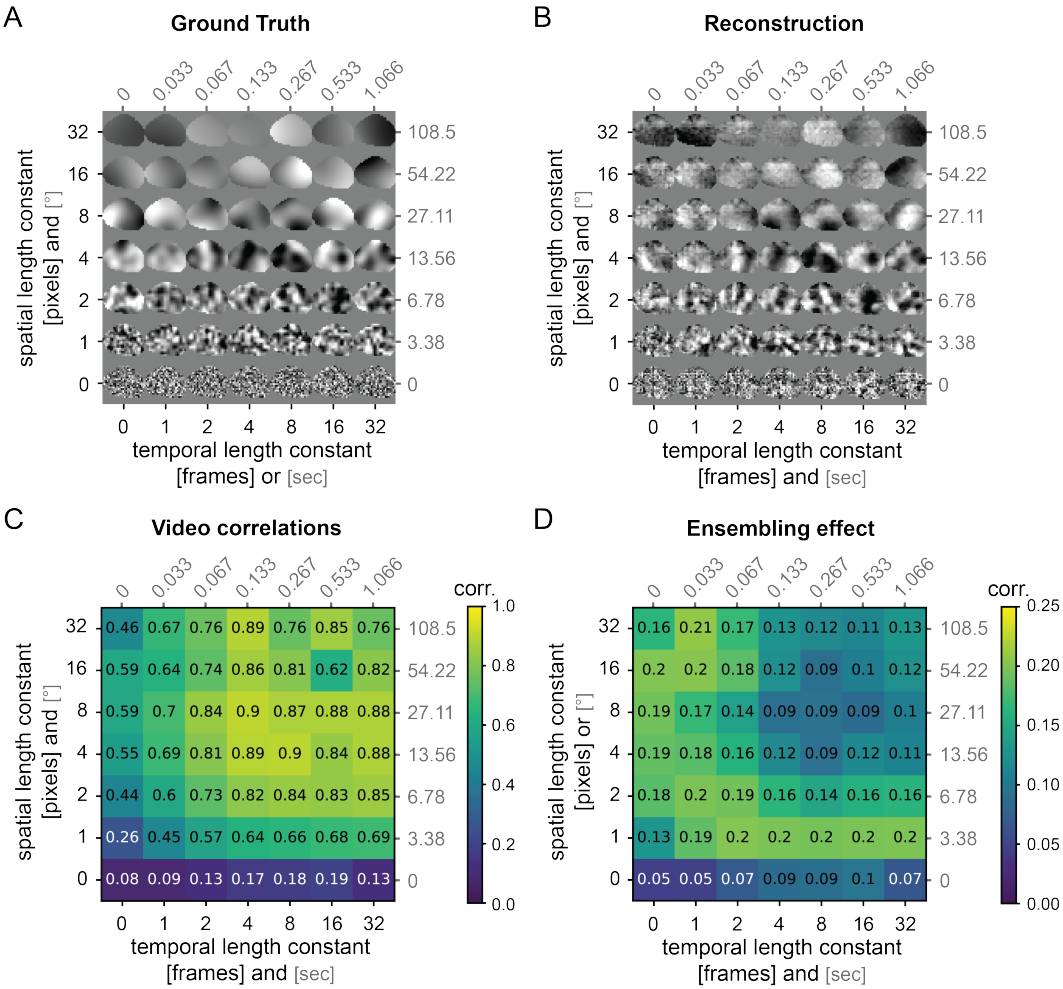


Figure 5: Reconstruction of Gaussian noise across the spatial and temporal spectrum using predicted activity. A) Example Gaussian noise stimulus set with evaluation mask for one mouse. Shown is the last frame of a 2 second video. B) Reconstructed Gaussian stimuli with SOTA DNEM predicted neuronal activity 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) Ensembling effect for each stimulus. Video correlation for ensembled prediction (average videos from 6 model instances) minus the mean video correlation across the 6 individual model instances.

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 5A). The dataset consisted of 49, 2 second, 36 by 36 pixel videos at 30 Hz, which varied in the spatial and temporal length constants. The stimuli were centered on the average evaluation masks of all mice. As we did not have ground truth neural 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 5B). 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 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 5C and Supplementary Video 2).

We repeated this analysis using full-field high-contrast square gratings drifting in four directions (up, down, left right), and similarly found that high-spatial and temporal frequencies were not reconstructed as well (Supplementary Figure S1).

To test if model ensembling improves video reconstruction quality across all spatial and temporal length constants uniformly, we subtracted the average video correlation across the six model instances from the video correlation of the average video (i.e. ensembled video reconstruction minus unensembled video reconstruction; Figure 5D). 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.

3.2 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 6). We found that dropping 50% of the neurons reduced the video correlation by only 10.9% while dropping 75% reduced the performance by 26.1%. 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 be a sweet spot when compromising between density and 2D coverage.

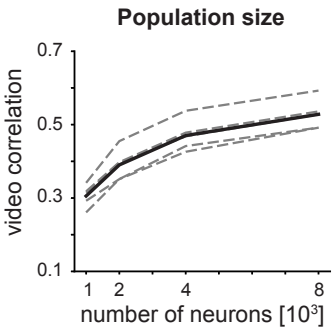


Figure 6: Video reconstruction using fewer neurons (i.e. population ablation) leads to lower reconstruction quality. Dashed lines are individual animals, solid line is mean. One-way repeated measures ANOVA $p = 3.52 * 10^{-12}$. Bonferroni corrected paired t-test outcomes between consecutive drops in population size are all $p < 0.001$, n = 5 mice.

4 Discussion

In summary, 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 [29]. However, this requires a large amount of GPU memory when using video models and was not practical with our hardware limitations. However, 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 [32, 33]. In such a case, averaging 'misaligned' reconstructions from multiple models might degrade overall quality.

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/° [34]). 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 [35]) 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 [36, 37]. 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 [38] and is associated with changes in the tuning distribution of neurons in the visual system [39, 40, 41, 42]. Similarly, selective feature attention can modulate the response amplitude of neurons that have a preference for the features that are currently being attended to [43]. 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.

Such 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 [6, 7, 9, 10], 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-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 the activity of neurons in the mouse visual cortex, paving the way to use this method to investigate a variety of visual processing phenomena.

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. and J.B. are funded by The Wellcome Trust (219627/Z/19/Z; 214333/Z/18/Z) and Gatsby Charitable Foundation (GAT3755).

References

[1] 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.

[2] 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, 2019.

[3] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1):e1006633, 2019.

[4] 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.

[5] 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.

[6] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, page 2022.11.18.517004, 2023.

[7] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

[8] 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.

[9] 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.

[10] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity. *arXiv*, 2023.

[11] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv*, 2023.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Paweł A. Pierzchlewicz, Konstantin F. Willeke, Arne F. Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, Andreas S. Tolias, and Fabian H. Sinz. Energy Guided Diffusion for Generating Neurally Exciting Images. *bioRxiv*, page 2023.05.18.541176, 2023.

[19] Ruslan Baikulov. Solution for Sensorium 2023 Competition (v23.11.22). Zenodo. https://doi.org/10.5281/zenodo.10155151, nov 2023.

[20] Cristopher M. Niell and Michael P. Stryker. Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron*, 65(4):472–479, 2010.

[21] 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.

[22] 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.

[23] Winifried Denk, James H. Strickler, and Watt W. Webb. Two-Photon Laser Scanning Fluorescence Microscopy. *Science*, 248(4951):73–76, 1990.

[24] Polina Turishcheva, Paul G Fahey, Michaela Vystrčilová, Laura Hansel, Rachel Froebe, Kayla Ponder, Yongrong Qiu, Konstantin F Willeke, Mohammad Bashiri, Ruslan Baikulov, Yu Zhu, Lei Ma, Shan Yu, Tiejun Huang, Bryan M Li, Wolf De Wulf, Nina Kudryashova, Matthias H Hennig, Nathalie L Rochefort, Arno Onken, Eric Wang, Zhiwei Ding, Andreas S Tolias, Fabian H Sinz, and Alexander S Ecker. Retrospective for the Dynamic Sensorium Competition for predicting large-scale mouse primary visual cortex activity from videos. *arXiv*, 2024.

[25] 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.

[26] 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.

[27] Alexander Mordvintsev, Nicola Pezzotti, Ludwig Schubert, and Chris Olah. Differentiable image parameterizations. *Distill*, 3(7), July 2018.

[28] 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, 2023.

[29] 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.

[30] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv*, 2014.

[31] Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

[32] 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.

[33] 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.

[34] G.T. Prusky and R.M. Douglas. Characterization of mouse cortical spatial vision. *Vision Research*, 44(28):3411–3418, 2004.

[35] 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.

[36] 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.

[37] 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.

[38] Wu Li. Perceptual Learning: Use-Dependent Cortical Plasticity. *Annual Review of Vision Science*, 2(1):1–22, 2015.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

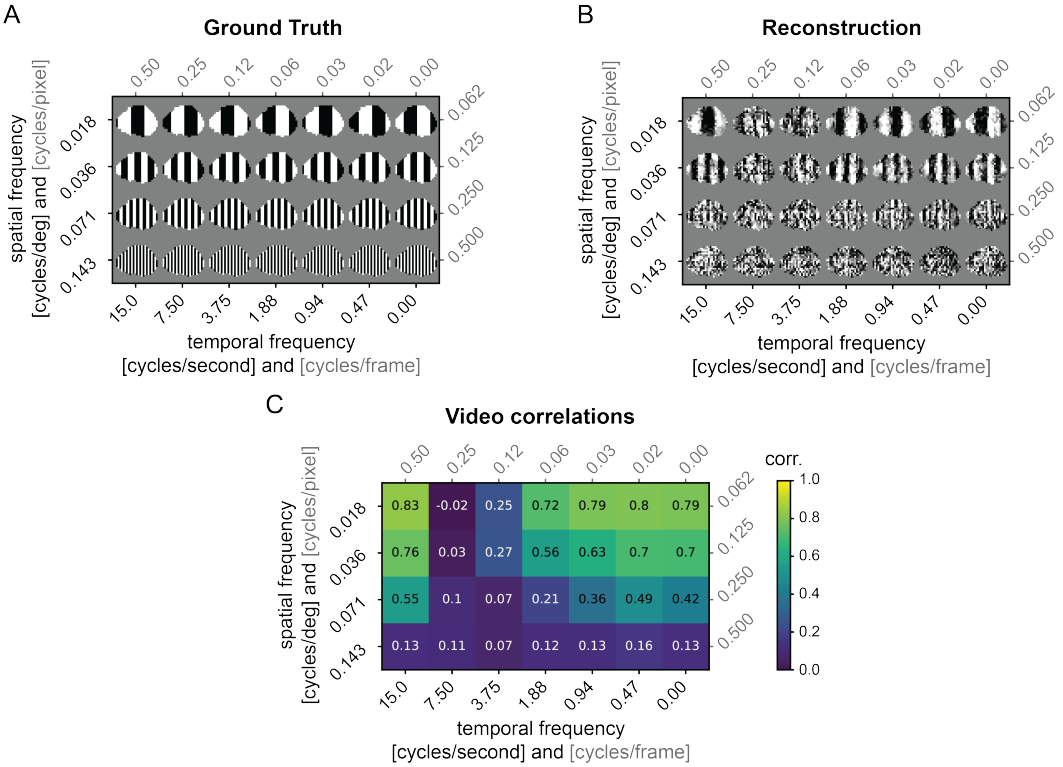
[43] Takahiro Kanamori and Thomas D. Mrsic-Flogel. Independent response modulation of visual cortical neurons by attentional and behavioral states. *Neuron*, 110(23):3907–3918, 2022.

A Appendix / supplemental material

Supplementary Video 1: Reconstructed 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 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 Figure S1: 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.

B Code

Code will be made available upon publication in a peer-reviewed journal.