Dynamics Based Neural Encoding with Inter-Intra Region Connectivity

Mai Gamal
German University in Cairo
Egypt
mai.tharwat@guc.edu.eg

Mohamed Rashad
Ain Shams University
Egypt
mohamed.rashad@cis.asu.edu.eg

Eman Ehab
Nile University
Egypt
e.ehab@nu.edu.eg

Seif Eldawlatly
American University in Cairo
Egypt
seldawlatly@aucegypt.edu

Mennatullah Siam
University of British Columbia
Canada
mennatullah.siam@ubc.ca

Abstract

Extensive literature has drawn comparisons between recordings of biological neurons in the brain and deep neural networks. This comparative analysis aims to advance and interpret deep neural networks and enhance our understanding of biological neural systems. However, previous work did not consider the time aspect and how video and dynamics (e.g., motion) modelling in deep networks relate to the biological neural systems within a large-scale comparison. Towards this end, we propose the first large-scale study focused on comparing video understanding models with respect to the visual cortex recordings using video stimuli. The study encompasses around half a million regression fits, examining image *vs.* video understanding, convolutional *vs.* transformer-based and fully *vs.* self-supervised models. We show that video understanding are better than image understanding models, convolutional models are better in the early-mid visual cortex regions than transformer based ones except for multiscale transformers, and that two-stream models are better than single stream. Furthermore, we propose a novel neural encoding scheme that is built on top of the best performing video understanding models, while incorporating inter-intra region connectivity across the visual cortex. Our neural encoding leverages the dynamics modelling from video stimuli, through utilizing two-stream networks and multiscale transformers, while taking connectivity priors into consideration. Our results show that merging both intra and inter-region connectivity priors increases the encoding performance over each one of them standalone or no connectivity priors. It also shows the necessity for encoding dynamics to fully benefit from such connectivity priors.

1 Introduction

There has been a recent increase in studies that compare how deep neural networks process input stimuli to the processing that occurs in the brain Zhou et al. (2022); Conwell et al. (2021); Schrimpf et al. (2018); Cichy et al. (2019, 2021). Recent benchmarks have been released to improve machine learning for neural encoding Schrimpf et al. (2018); Cichy et al. (2019, 2021); Gifford et al. (2023). One of the well-established benchmarks that studied how deep networks compare to biological neural systems is The Algonauts 2021 dataset and challenge that focused on video stimuli Cichy et al. (2021); Lahner et al. (2024). Recent works investigated the ability of deep networks to regress on the brain responses for video stimuli Zhou et al. (2022) from the aforementioned dataset. However,

NeurIPS 2024 Workshop on NeuroAI, Vancouver, Canada (NeuroAI 2024).

they mainly worked with single-image deep neural networks. Inspired by this approach, we focus on studying video understanding models to draw insights on how the brain understands actions and models dynamics. While some works in neuroscience studied the time aspect Zhuang et al. (2021); Nishimoto et al. (2011); Khosla et al. (2021); Nishimoto (2021); Lahner et al. (2024); Güçlü & Van Gerven (2017); Shi et al. (2018); Sinz et al. (2018); Huang et al. (2023), they did not focus on large-scale comparison. Our work focuses on the first study of state-of-the-art deep video understanding models from a neuroscience lens. Our study takes various properties into consideration where we study image *vs.* video understanding, convolutional *vs.* transformer based, single-stream vs two-stream, and fully supervised *vs.* self supervised ones. Our results show that video understanding models are better than image understanding ones in predicting the human visual cortex recordings. Specifically, two-stream convolutional models and multiscale transformers were the best. Interestingly, we show that multiscale transformers exhibit similar behaviour to convolutional models when encoding early-mid cortex regions unlike other transformer based models.

The brain is an interconnected system with local correlations within one region and global correlations across regions Genç et al. (2016); Li et al. (2022). Few recent works explored the potential of using cortical connectivity in neural encoding models Mell et al. (2021); Xiao et al. (2022). Nonetheless, previous voxels-to-voxels models are not designed to take stimulus as input and define source voxels in an ad hoc manner. Inspired by that direction, we propose a fully integrated model that learns a two-stage architecture, stimulus-to-voxels and voxels-to-voxels. Our approach takes into consideration voxels from all visual cortex regions and learns the weighting mechanism, instead of relying on ad hoc non learnable mechanism to define source voxels. Finally, we show the interplay of dynamics modelling and connectivity priors in improving the neural encoding of the visual cortex.

In summary, our contributions are two fold: (i) We showcase the first large-scale study of deep video understanding models on responses from the human visual cortex where the models include convolutional *vs.* transformer-based, single *vs.* two stream, and fully *vs.* self-supervised. (ii) We propose a novel fully integrated encoding model with intra and inter-region connectivity priors with features extracted from video understanding models that learned to encode dynamics.

2 Method

Environment design. In this study, we focus on the question of "How do deep video understanding models families compare to biological neural systems?". Towards this, we study the identification across families of models when encoding the brain responses. Specifically, families are defined based on: (i) the input, whether models learned from single images or videos encouraging them to learn dynamics and motion, (ii) the supervision, whether they are trained fully-supervised or in a self-supervised manner using unlabeled data, and (iii) the architecture, whether it uses local convolutions or transformer-based global operations. While previous works Han et al. (2023) working on single image architectures focused on the architecture aspect, we argue it is even more important to look into whether the model is learning dynamics (e.g., motion) or simply using static information from a single image. Moreover, it is important to understand the impact of the supervision signal used to train the model. We use the public fMRI dataset from Mini-Algonauts Cichy et al. (2021). We perform cross-validation over four folds throughout all our experiments. The dataset provides fMRI recordings of ten subjects who watched 1000 short video clips of three seconds average duration. The videos clips were sampled from the Memento10k dataset Newman et al. (2020). The fMRI data were acquired with a 3 T Trio Siemens scanner and provided at TR one second and resolution of $2.5 \times 2.5 \times 2.5$ mm Lahner et al. (2024). We use the brain responses from nine regions of interest of the visual cortex, these are across two levels: (i) early and mid-level visual cortex (V1, V2, V3, and V4), and (ii) high-level visual cortex (EBA, FFA, STS, LOC, and PPA). We run our experiments on more than 35 models that are listed in Table 1, along with their model family and configurations. Video understanding models include C2D Li et al. (2019), CSN Tran et al. (2019), I3D Carreira & Zisserman (2017), R(2+1)D Tran et al. (2018), SlowFast, the Slow branch (3D ResNet-50) Feichtenhofer et al. (2019), X3D Feichtenhofer (2020), MViT Fan et al. (2021), and TimeSformer Bertasius et al. (2021). Self-supervised video understanding models, stMAE Feichtenhofer et al. (2022) and OmniMAE Girdhar et al. (2023) are used as well. Single image understanding models include ResNets He et al. (2016), ViTs Dosovitskiy et al. (2021), DINO Caron et al. (2021), and MAE He et al. (2022).

Neural encoding. Inspired by the recent work Zhou et al. (2022), we use a layer-weighted region of interest encoding that takes the hierarchical nature of deep networks into consideration. Initially, we

Table 1: List of the families of models, (their backbone/s, the training datasets, and the configuration as clip length, sampling rate). IN: ImageNet Deng et al. (2009), K: Kinetics-400 Kay et al. (2017), Ch: Charades Kay et al. (2017) and SSV2: Something-something v2 Sigurdsson et al. (2016).

Input	Sup.	Arch.	Network (Backbone/s - Dataset/s - Config.)
Vid.	Full	Conv.	C2D (R50-K-$8,8$) Li et al. (2019)
			CSN (R101-K-$32,2$) Tran et al. (2019)
			I3D (R50-K-$8,8$) Carreira & Zisserman (2017)
			R(2+1)D (R50-K-$16,4$) Tran et al. (2018)
			SlowFast (R50,101-K,Ch,SSV2-8, $8/4, 16$) Feichtenhofer et al. (2019)
			3DResNet (R18,50-K,Ch,SSV2-8, $8/4, 16$) Feichtenhofer et al. (2019)
			X3D (XS,S,M,L-K-Matched Sampling) Feichtenhofer (2020)
		Transf.	MViT (B-K-16, $4/32, 3$) Fan et al. (2021)
			TimeSformer (B-K,SSV2-$8,8$) Bertasius et al. (2021)
			OmniMAE finetuned (B-SSV2-$8,8$) Girdhar et al. (2023)
	Self	Transf.	stMAE (L-K-$8,8$) Feichtenhofer et al. (2022)
			OmniMAE (B,L-IN/SSV2-$8,8$) Girdhar et al. (2023)
Img.	Full	Conv.	ResNet (R152,101,50,34,18-IN-$8,8$) He et al. (2016)
		Transf.	ViT (B16,32,L16,32-IN-$8,8$) Dosovitskiy et al. (2021)
	Self	Transf.	DINO (B-IN-$8,8$) Caron et al. (2021)
			MAE (B-IN-$8,8$) He et al. (2022)

pre-process the input features from the different layers of a candidate model through averaging the features on the temporal dimension. This is followed by performing sparse random projection Li et al. (2006) for dimensionality reduction and computational efficiency reasons. Assume input features for layer, l, after dimensionality reduction as, $X_l \in \mathbb{R}^{C\times 1}$, with C features. We learn the weights of one fully connected layer to provide the predictions of the voxels of one region of interest in the visual cortex as, $\hat{Y}_l = W_l X_l$. Where $W_l \in \mathbb{R}^{N\times C}$, $\hat{Y} \in \mathbb{R}^{N\times 1}$ and N is the number of voxels in the region of interest. We learn a weighted sum of the predictions of all layers and use the following loss,

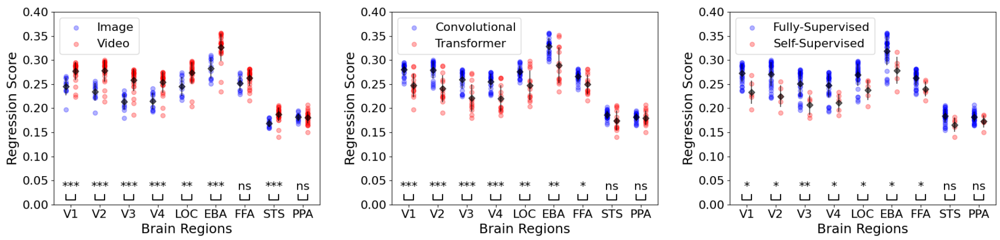
$$\mathcal{L} = \|Y - \sum_{l=1}^{L} \omega_l \hat{Y}_l\|_2^2 + \beta_1 \sum_{l=1}^{L} \|W_l\|_2 + \beta_2 \|\omega\|_1, \tag{1}$$

where ω_l is a learnable scalar weight for layer, l, and, ω, is the vector of weights. Each ω_l, controls the contribution of layer, l, to the final regression, and β_1, β_2 are hyper-parameters of the regularization. We use L1 regularization for the layer weights to enforce sparsity. This encoding avoids unnecessary assumptions that there is a one-to-one alignment between layers and visual brain regions.

Inter-intra region connectivity priors. We present a novel encoding scheme on top of the best-performing video understanding models by fully integrating the neural encoding with inter- and intra-region voxel connectivity priors. The input video stimuli goes through the source video understanding model to extract multiple layers features, followed by the connectivity module which takes the concatenated voxels of the nine visual regions as input. This module consists of two fully-connected layers with L2 regularization and dropout, outputting the predicted voxels of one region. We train our model in a two-stage fashion, where we train the standard neural encoding scheme, followed by training the connectivity module. In the training phase, the main target is to learn the connectivity between the voxels of all the regions and the target region including intra-connectivity between the voxels of the target region itself and the inter-connectivity between voxels of the target region and the other visual regions. During training, the input to the connectivity is the groundtruth voxel activations. During inference, the input to the connectivity is the predicted activations.

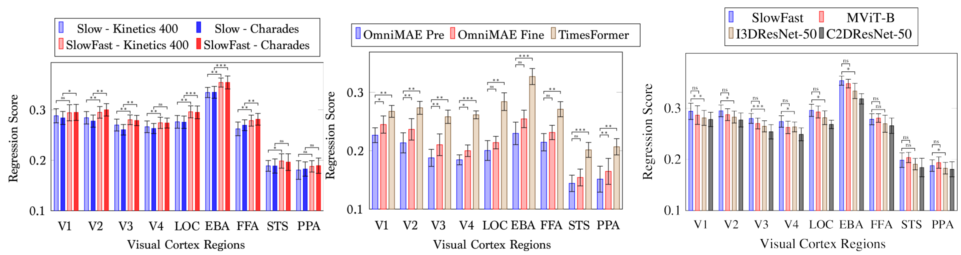
3 Experimental results

Implementation details. In the case of both video and image understanding, we sample a clip from the input video to extract features for that clip. The input clips are constructed based on the sampling rate used during the model training for video understanding models. As for image understanding models, we use sampling rate eight. Before training the regressor, a hyperparameter tuning for β_1, β_2 is conducted using two-fold cross-validation on the training set of the first subject, following previous work Zhou et al. (2022). Moreover, an early-stopping strategy is employed. We report



(a) (b) (c)

Figure 1: Experiments showing regression scores as Pearson's correlation coefficient for model families. (a) Comparison of image *vs.* video understanding models, (b) comparison of convolutional *vs.* transformer-based models and (c) comparison of fully supervised *vs.* self-supervised models.



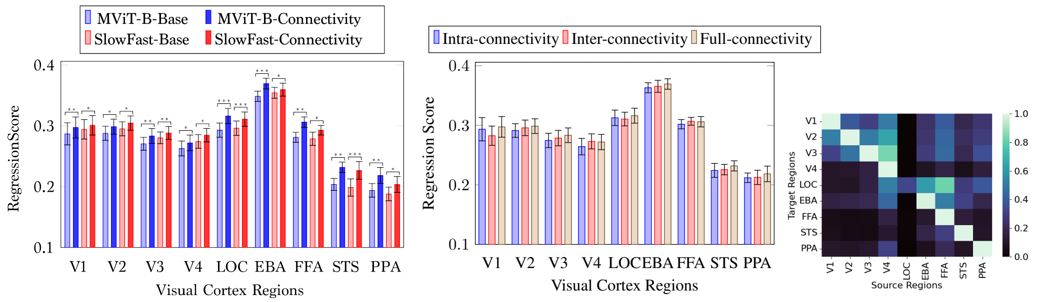
(a) (b) (c)

Figure 2: Fine-grained analysis showing the Pearson's correlation coefficient as the regression scores. (a) Single *vs.* two stream SlowFast. (b) OmniMAE pre-trained with self-supervision, TimeSformer, and OmniMAE fine-tuned with full supervision. (c) Top-2 video understanding models w.r.t others.

the average Pearson's correlation coefficient across all voxels within a specific region in the brain. All results are averaged over the subjects. We conduct experiments on four folds and report the average and standard deviation. In each fold, the 1,000 videos are split into training and testing sets as 90% and 10%, respectively. We ran statistical significance across families of models using Welch's t-test. We show statistical significance as 'ns' not significant, '$*, **, ***$' significant with p-values $< 0.05, 0.01, 0.001$, resp.

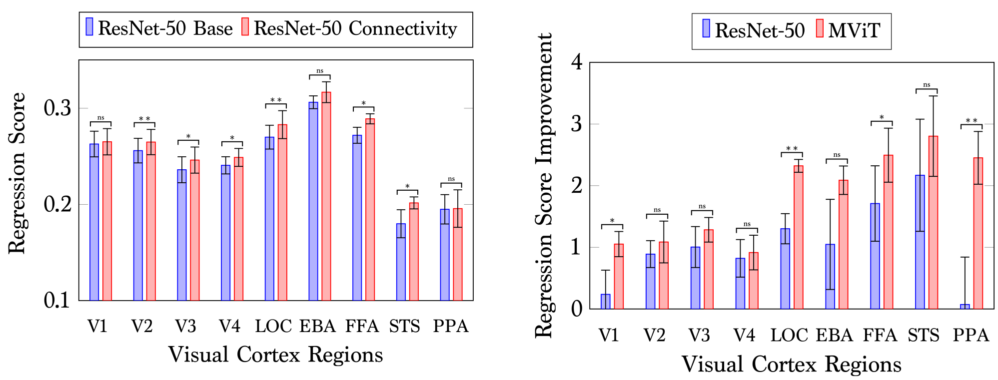
Neural encoding results. We compare families of video understanding models to the human visual cortex. We conduct three comparisons; single image *vs.* video understanding families of models, convolutional-based *vs.* transformer-based models, and fully-supervised *vs.* self-supervised models. Figure 1a demonstrates that across most brain regions, video understanding models have better capability to model the visual cortex responses than single image architectures. This is aligned with the dynamic nature of visual processing in the brain given that humans process the world in motion Hegdé (2008). Figure 1b shows the comparison between transformer-based and convolutional-based models. It shows that convolutional models have higher regression scores across early-mid regions in the visual cortex with relatively high statistical significance. This difference decrease as we go to higher level regions until it becomes insignificant. This might be tied to recent works showing transformers lacking the ability to capture high-frequency components Bai et al. (2022), while early layers in convolutional models are better in capturing such high-frequency components. Interestingly, we notice transformers equipped with multiscale processing, MViTs, tend to behave similar to convolutional ones in early-mid regions unlike other transformers. Figure 1c also shows that fully-supervised models are better able to predict most of the regions than self-supervised models.

Fine-grained analysis. We conduct a fine-grained analysis that goes beyond families of models. We start with studying two stream *vs.* single stream architectures across three datasets. Figure 2a shows that the two stream architectures have better ability to model the visual cortex than single stream ones in the low level regions and are either better or on-par in the high level regions. We then discuss the self-supervised learning results that showed worse regression scores in comparison to full supervision. Towards this end, we investigate OmniMAE variants (i.e., self supervised and finetuned) and TimeSformer. Figure 2b confirms that fully supervised models give better scores than the self-supervised ones across all regions. Furthermore, Figure 2c shows that both SlowFast, a two-stream architecture, and MViT, a multiscale vision transformer, are the best in neural encoding.



(a) (b) (c)

Figure 3: (a) Comparison of base model accuracies of MViT-B-16x4 and SlowFast and their accuracies after incorporating the intra-region and inter-region voxel connectivity. (b) Comparison of performance enhancement by incorporating the intra-region and inter-region voxel connectivity together or each of them separately. (c) Average weights per region contributing to the accuracy enhancement of each target visual region.



(a) (b)

Figure 4: (a) Comparison of ResNet-50 with and without connectivity. (b) Comparison of improvement in the regression scores in ResNet-50 *vs.* MViT w/ Connectivity Priors. Difference in regression scores is shown after multiplying by 100 for visualization.

Inter-intra region connectivity. We show the results of our improved neural encoding that builds upon the best video understanding models (MViT-B and SlowFast) while incorporating intra- and inter-region connectivity. Figure 3a shows the statistically significant enhancements in prediction accuracy after incorporating our connectivity priors. As an ablation study, we examined the performance enhancement in MViT-B in the case of intra- or inter-connectivity separately. Figure 3b shows that the full-connectivity (i.e., combining both) is either superior or on-par with intra- or inter-connectivity standalone. To better understand the directional connectivity between the regions, we analyzed the average learned weights of each region as shown in Fig. 3c. It shows: i) the effect of one region on another is not symmetric but directional, ii) early-mid regions are the highest contributors to the accuracy enhancement of other early-mid regions, and the same for late-regions, iii) V4 is contributing to both early-mid and late regions, and iv) the contributions of late regions on early regions (V1, V2) are stronger than contribution of early on late ones which could be attributed to the top-down influence of feedback-pathways Gilbert & Li (2013). Additionally we confirm the benefit of connectivity priors in single image understanding models in Fig. 4a, yet the gain from connectivity is higher in video understanding models than single image ones as shown in Fig. 4b. It confirms that connectivity priors are maximally beneficial with dynamics based models.

4 Conclusion

This paper has provided a large-scale study of video understanding models from a neuroscience perspective. We show that convolutional models predict better the early-mid regions than transformer based ones except with multiscale transformers. Then we demonstrate a better neural encoding scheme that utilizes both dynamics modelling and inter-intra region connectivity.

References

Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. Improving vision transformers by revisiting high-frequency components. In *Proceedings of the European Conference on Computer Vision*, pp. 1–18. Springer, 2022.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*, volume 2, pp. 4, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Radoslaw Martin Cichy, Gemma Roig, and Aude Oliva. The algonauts project. *Nature Machine Intelligence*, 1(12):613–613, 2019.

Radoslaw Martin Cichy, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Polina Iamshchinina, M Graumann, Alex Andonian, NAR Murty, K Kay, Gemma Roig, et al. The algonauts project 2021 challenge: How the human brain makes sense of a world in motion. *arXiv preprint arXiv:2104.13714*, 2021.

Colin Conwell, Jacob S Prince, George A Alvarez, and Talia Konkle. What can 5.17 billion regression fits tell us about artificial models of the human visual system? In *Shared Visual Representations in Human and Machine Intelligence workshop at Conference on Neural Information Processing Systems*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, 2021.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 203–213, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in Neural Information Processing Systems*, 35:35946–35958, 2022.

Erhan Genç, Marieke Louise Schölvinck, Johanna Bergmann, Wolf Singer, and Axel Kohler. Functional connectivity patterns of visual cortex reflect its anatomical organization. *Cerebral cortex*, 26 (9):3719–3731, 2016.

Alessandro T Gifford, Benjamin Lahner, Sari Saba-Sadiya, Martina G Vilas, Alex Lascelles, Aude Oliva, Kendrick Kay, Gemma Roig, and Radoslaw M Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*, 2023.

Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013.

Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10406–10417, 2023.

Umut Güçlü and Marcel AJ Van Gerven. Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in computational neuroscience*, 11:7, 2017.

Yena Han, Tomaso A Poggio, and Brian Cheung. System identification of neural systems: If we got it right, would we know? In *International Conference on Machine Learning*, pp. 12430–12444. PMLR, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Jay Hegdé. Time course of visual perception: coarse-to-fine processing and beyond. *Progress in neurobiology*, 84(4):405–439, 2008.

Liwei Huang, ZhengYu Ma, Huihui Zhou, and Yonghong Tian. Deep recurrent spiking neural networks capture both static and dynamic representations of the visual cortex under movie stimuli. *arXiv preprint arXiv:2306.01354*, 2023.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Meenakshi Khosla, Gia H Ngo, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7(22):eabe7547, 2021.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature communications*, 15(1):6241, 2024.

Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Collaborative spatiotemporal feature learning for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7872–7881, 2019.

Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 287–296, 2006.

Yuanning Li, Huzheng Yang, and Shi Gu. Upgrading voxel-wise encoding model via integrated integration over features and brain networks. *BioRxiv*, pp. 2022–11, 2022.

Maggie Mae Mell, Ghislain St-Yves, and Thomas Naselaris. Voxel-to-voxel predictive models reveal unexpected structure in unexplained variance. *NeuroImage*, 238:118266, 2021.

Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pp. 223–240. Springer, 2020.

Shinji Nishimoto. Modeling movie-evoked human brain activity using motion-energy and space-time vision transformer features. *BioRxiv*, pp. 2021–08, 2021.

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.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Junxing Shi, Haiguang Wen, Yizhen Zhang, Kuan Han, and Zhongming Liu. Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. *Human brain mapping*, 39(5):2269–2282, 2018.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision*, pp. 510–526. Springer, 2016.

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.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5552–5561, 2019.

Wulue Xiao, Jingwei Li, Chi Zhang, Linyuan Wang, Panpan Chen, Ziya Yu, Li Tong, and Bin Yan. High-level visual encoding model framework with hierarchical ventral stream-optimized neural networks. *Brain Sciences*, 12(8):1101, 2022.

Qiongyi Zhou, Changde Du, and Huiguang He. Exploring the brain-like properties of deep neural networks: a neural encoding perspective. *Machine Intelligence Research*, 19(5):439–455, 2022.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.