Do Dynamics Matter for Neural Alignment? A Comparative study of Video and Static Vision Models

Khaled Jedoui Al-Karkari (thekej@stanford.edu)
Stanford University
Stanford, CA

Yingtian Tang (yingtian.tang@epfl.ch)
Ecole Polytechnique Federale de Lausanne (EPFL),
1015 Lausanne, Switzerland

Martin Schrimpf (martin.schrimpf@epfl.ch)
Ecole Polytechnique Federale de Lausanne (EPFL),
1015 Lausanne, Switzerland

Daniel L.K. Yamins (yamins@stanford.edu)
Stanford University
Stanford, CA

Abstract

Understanding how the brain constructs and updates visual representations from dynamic input is central to our comprehension of perception and cognition. While deep learning has achieved impressive performance in visual tasks, the extent to which these models capture the computational principles of biological vision is unclear. In this paper, we investigate the alignment between representations learned by vision Artificial Intelligence (AI) models and neural activity in biological brains. Using brain fMRI data, we benchmark a diverse range of models, trained on various tasks, in their ability to predict brain responses to image and video stimuli. Our results demonstrate a clear advantage for video-based representations over static image representations across all analyzed brain regions. Our findings suggest that temporal modeling is a key component in the development of models that better align with biological vision, providing new insights into computational modeling of vision.

Keywords: Artificial Intelligence; Vision: Video Representation; Neural Predictivity; Model-Brain Alignment;Intuitive Physics Understanding.

Introduction

Understanding how the brain perceives and predicts the dynamic world remains a fundamental challenge across neuroscience, psychology, and artificial intelligence. Humans are able to effortlessly process continuous streams of information, extracting not just static features (shapes and colors) but also crucial dynamic information such as motion, object interactions, and causal relationships. This ability to perceive and anticipate events happening over time is fundamental to our ability to navigate and interact with the world. While deep learning has achieved impressive performance in vision tasks, it is unclear to what extent these models capture the computational principles of biological dynamic vision.

Traditional computer vision models (Krizhevsky, Sutskever, & Hinton, 2012; He, Zhang, Ren, & Sun, 2016), often trained on static images, struggle to capture the richness and complexity of dynamic scenes. They may be able to recognize objects in a scene, but fail to predict its trajectory, or understand the consequences of its interactions with another object. This limitation highlights a crucial gap to human visual perception, which seamlessly integrates spatial and temporal information. Recent advances in video understanding models offer a promising avenue for exploring this challenge with models that take advantage of video data (Carreira & Zisserman, 2017; Feichtenhofer, Fan, Malik, & He, 2019; Bertasius, Wang, & Torresani, 2021a). These models, incorporating mechanisms like 3D convolutions or self-attention, are designed to learn representations that capture temporal dependencies and spatio-temporal patterns. However, it remains an open question whether these learned representations align with the neural mechanisms underlying dynamic vision in biological brains.

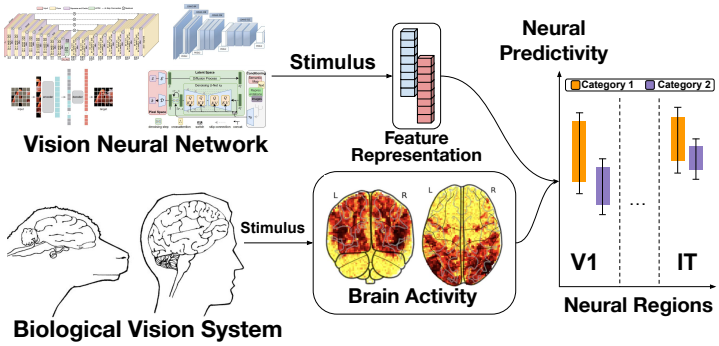

Figure 1: We illustrate our evaluation workflow for neural alignment comparison between models and primate observers. We systematically evaluate a comprehensive set of video and image models on multiple neural benchmarks to establish the importance of temporal dynamics modeling in close to biological visual processing.

This work attempts to bridge this gap by investigating the representational alignment between state-of-the-art video AI models and neural activity in the primate visual cortex. We investigate whether video models, trained to understand and/or predict temporal dynamics, show neural representations more similar to those in biological brains compared to static-image models. To address this questions, we systematically benchmark a diverse range of video and image models, trained on various tasks, against neural data recorded using fMRI. We employ established neural predictivity metrics (Schrimpf et al., 2020; Yamins et al., 2014) to quantify the similarity between model activations and neural responses to both static images and dynamic video stimuli.

This study makes several key contributions: (1) It provides a comprehensive comparison of a diverse set of video and image models in terms of their neural alignment with biological visual processing. (2) It identifies specific brain regions that exhibit the strongest alignment with video models, providing insights into the neural substrates of dynamic scene understanding and their functional specialization.

Ultimately, this work aims to provide a deeper understanding of how current video AI models represent and reason about the dynamic visual world, highlighting the gap between their capabilities and human-level understanding, and guiding the development of more human-aligned AI systems.

Tasks and Datasets

This study systematically evaluates the neural alignment of a comprehensive set of video and image models against a diverse suite of neuroimaging datasets. For each dataset, we extract features from multiple layers of vision models corresponding to the stimuli. We apply ridgeCV regression to map model activations to neural activity, quantifying neural predictivity using the correlation between predicted and actual neural responses. This analysis was performed separately for each model and brain region. We report the maximum neural

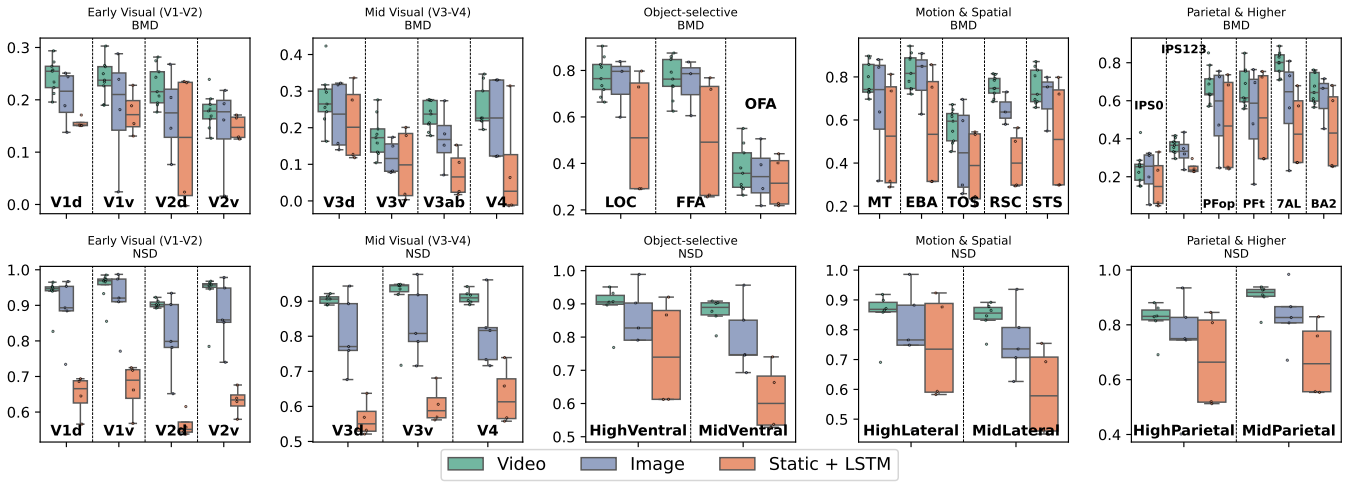

Figure 2: We illustrate the neural predictivity performance of different representation modalities on brain fMRI data captured on dynamic and static stimuli. We find that video-based representations consistently outperform static image models across multiple brain regions, highlighting the crucial role of temporal dynamics in predicting neural responses to visual stimuli.

alignment score across a given model's layers.

The Natural Scenes Dataset (NSD) (Allen et al., 2022) is a high-resolution fMRI dataset capturing brain responses from humans viewing natural images. We analyze preprocessed data from four subjects, focusing on 1,000 shared COCO dataset images (Lin et al., 2014) with three trials per image.

The BOLD Moments in Time (BMD) dataset (Lahner et al., 2024) provides whole-brain fMRI responses from ten human subjects viewing 1,102 short (3-second) naturalistic video clips. Sampled from the Memento10k dataset (Newman et al., 2020)., these videos include rich metadata such as object labels, scene descriptions, and memorability scores. We use both training (1,000 stimuli) and testing (102 stimuli) sets.

Results and Discussion

Our investigation into the neural alignment of vision models reveals several key findings. We compare three distinct model classes: dedicated video models (Morgado, Vasconcelos, & Misra, 2020; Bertasius, Wang, & Torresani, 2021b; Tong, Song, Wang, & Wang, 2022; Bardes et al., 2024; Fan et al., 2021), static image models (He et al., 2021; Tewari et al., 2023; Oquab et al., 2023; Dosovitskiy et al., 2020; Yu, Ye, Tancik, & Kanazawa, 2021), and LSTM-augmented image models.

We begin by examining the neural predictivity of different model categories across various brain regions using the BMD dataset. As shown in Figure 2, video models consistently demonstrate the highest neural alignment across virtually all recorded brain regions. For instance, in motion-sensitive area MT, video models score between 0.696 (vjepa (Bardes et al., 2024)) and 0.897 (i3d-nonlocal (Fan et al., 2021)), generally surpassing image models and significantly outperforming most LSTM-augmented models. This pattern holds true across Early Visual (V1-V2), Mid Visual (V3ab and V3v), Object-selective (LOC, OFA and FFA), Motion & Spatial (MT, EBA, STS, TOS, RSC and STS), and Parietal & Higher areas (7AL, IPS123, PFt, PFop, BA2). While some high-performing image models like dinov2 (Oquab et al., 2023) achieve scores competitive with lower-performing video models in certain regions, the average performance and the peak performance within the video category consistently exceed those of the image category.

We further validate our previous findings using NSD responses on static images. Our results align with expectations regarding domain specificity, but also offer interesting insights. Image models, particularly dinov2 (Oquab et al., 2023), tend to achieve the highest peak scores across many regions (e.g., Mid Parietal and High Ventral). This confirms that models trained on static images generally provide great alignment for brain responses to static images. However, video models demonstrate remarkably strong and competitive performance on the NSD dataset. Their scores often surpass those of other image models (including vit (Dosovitskiy et al., 2020), mae (He et al., 2021), pixelnerf (Yu et al., 2021)) and consistently outperform the LSTM-augmented category by a significant margin (e.g. V1-4 and Mid Ventral/Lateral/Parietal).

On the other hand, the LSTM-augmented image models consistently exhibit the lowest neural alignment scores on both datasets across the visual hierarchy. This suggests that merely adding a recurrent layer (LSTM) to a static feature extractor is insufficient to capture the neural representations necessary for complex dynamic stimuli, performing worse even than models with no explicit temporal processing.

In summary, our experiments provide strong quantitative support that explicit temporal modeling, as found in dedicated video models, is crucial for aligning AI models with brain activity during dynamic visual experiences. These models significantly outperform static image architectures on video tasks. While specialized image models are great at predicting responses to static images, the robust performance of video models even in the static domain highlights their potential for learning powerful, generalizable visual representations.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... Kay, K. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. Retrieved from https://doi.org/10.1038/s41593-021-00962-x doi: 10.1038/s41593-021-00962-x

Bardes, A., Garrido, Q., Ponce, J., Rabbat, M., LeCun, Y., Assran, M., & Ballas, N. (2024). Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*.

Bertasius, G., Wang, H., & Torresani, L. (2021a). Is space-time attention all you need for video understanding? In *Proceedings of the international conference on machine learning.*

Bertasius, G., Wang, H., & Torresani, L. (2021b). *Is space-time attention all you need for video understanding?*

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the ieee conference on computer vision and pattern recognition.*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.*

Fan, H., Murrell, T., Wang, H., Alwala, K. V., Li, Y., Li, Y., ... Feichtenhofer, C. (2021). PyTorchVideo: A deep learning library for video understanding. In *Proceedings of the 29th acm international conference on multimedia.* (https://pytorchvideo.org/)

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the ieee international conference on computer vision.*

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked autoencoders are scalable vision learners. *arXiv:2111.06377*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems.*

Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., ... Cichy, R. (2024). Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature Communications*, *15*(1), 6241. Retrieved from https://doi.org/10.1038/s41467-024-50310-3 doi: 10.1038/s41467-024-50310-3

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Morgado, P., Vasconcelos, N., & Misra, I. (2020). Audio-visual instance discrimination with cross-modal agreement. In *arxiv preprint arxiv:2004.12943.*

Newman, A., Fosco, C., Casser, V., Lee, A., McNamara, B. A., & Oliva, A. (2020). Multimodal memorability: Modeling effects of semantics and decay on video memorability. *CoRR*, *abs/2009.02568*. Retrieved from https://arxiv.org/abs/2009.02568

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... others (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Schrimpf, M., et al. (2020). Brain-score: Which artificial neural network for object recognition is most brain-like? In *Advances in neural information processing systems.*

Tewari, A., Yin, T., Cazenavette, G., Rezchikov, S., Tenenbaum, J. B., Durand, F., ... Sitzmann, V. (2023). Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *NeurIPS*.

Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, *35*, 10078–10093.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, D. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021). pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4578–4587).