PAM: Predictive Attention Mechanism for Neural Decoding of Visual Perception

Thirza Dado, Lynn Le, Marcel van Gerven, Yağmur Güçlütürk, Umut Güçlü
Donders Institute for Brain, Cognition and Behaviour
Radboud University, Nijmegen, Netherlands
thirza.dado@donders.ru.nl
u.guclu@donders.ru.nl

Abstract

Attention mechanisms enhance deep learning models by focusing on the most relevant parts of the input data. We introduce predictive attention mechanisms (PAMs) – a novel approach that dynamically derives queries during training which is beneficial when predefined queries are unavailable. We applied PAMs to neural decoding, a field challenged by the inherent complexity of neural data that prevents access to queries. Concretely, we designed a PAM to reconstruct perceived images from brain activity via the latent space of a generative adversarial network (GAN). We processed stimulus-evoked brain activity from various visual areas with separate attention heads, transforming it into a latent vector which was then fed to the GAN's generator to reconstruct the visual stimulus. Driven by prediction-target discrepancies during training, PAMs optimized their queries to identify and prioritize the most relevant neural patterns that required focused attention. We validated our PAM with two datasets: the first dataset (B2G) with GAN-synthesized images, their original latents and multi-unit activity data; the second dataset (GOD) with real photographs, their inverted latents and functional magnetic resonance imaging data. Our findings demonstrate state-of-the-art reconstructions of perception and show that attention weights increasingly favor downstream visual areas. Moreover, visualizing the values from different brain areas enhanced interpretability in terms of their contribution to the final image reconstruction. Interestingly, the values from downstream areas (IT for B2G; LOC for GOD) appeared visually distinct from the stimuli despite receiving the most attention. This suggests that these values help guide the model to important latent regions, integrating information necessary for high-quality reconstructions. Taken together, this work advances visual neuroscience and sets a new standard for machine learning applications in interpreting complex data.

1 Introduction

Attention mechanisms in deep learning draw inspiration from the cognitive ability to selectively focus on specific aspects of the environment while neglecting others Kastner and Ungerleider (2000). These computational models dynamically weigh the importance of different input data segments to prioritize the most relevant information for the task at hand Bahdanau et al. (2014); Vaswani et al. (2017) – just like humans focus their attention on key details for understanding a scene or addressing a problem. In brief, an attention model derives three components from the input data: queries, keys and values. A query acts like a spotlight, shaped by a specific objective, to identify which parts of the input data are most pertinent (e.g., in a language translation model, the query could be the representation of a word in a sentence for which the model seeks the best equivalent in the target language). Keys are representations of the input data that embed contextual information (e.g., in

Preprint. Under review.

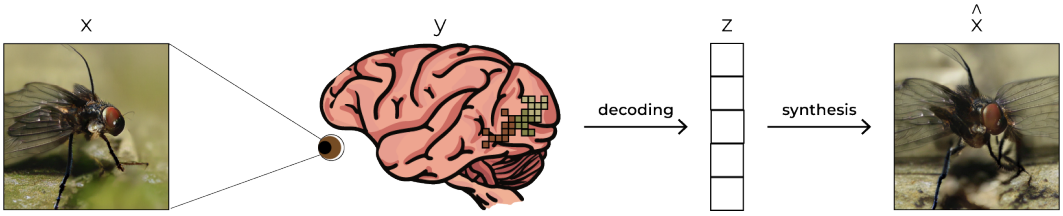

Figure 1: **Neural decoding.** This inverse problem seeks to infer the underlying stimulus that triggered the observed neural activity. It is common to divide this process into two stages: a "decoding" transformation that maps neural responses to an intermediate feature representation, and a more complex "synthesis" transformation that converts these features into an actual image.

a language translation model, a key associated with a particular word would also capture aspects of the surrounding words) so that the model can understand how each data segment fits into the larger picture. Keys are designed to be matched against queries to evaluate their relevance and their compatibility results in the corresponding attention weight. Values carry the actual output information and are aggregated according to the attention weights (e.g., in a language translation model, values could be potential translations for words or phrases). Through this mechanism, a model dynamically prioritizes the most relevant parts of the input by calculating an attention-weighted sum of the values.

Next, neural decoding involves the inverse problem of translating neural activity back into the features of a perceived stimulus that the brain was responding to (Figure 1). As such, this process seeks to find how the characteristics of a phenomenon are represented in the brain by classification Haxby et al. (2001); Kamitani and Tong (2005); Stansbury et al. (2013); Huth et al. (2016); Horikawa and Kamitani (2017), identification Mitchell et al. (2008); Kay et al. (2008); Güçlü and van Gerven (2017a,b) or reconstruction Thirion et al. (2006); Miyawaki et al. (2008); Naselaris et al. (2009); van Gerven et al. (2010); Nishimoto et al. (2011); Schoenmakers et al. (2013); Güçlü and van Gerven (2013); Cowen et al. (2014); Du et al. (2017); Güçlütürk et al. (2017); Shen et al. (2019); VanRullen and Reddy (2019); Dado et al. (2022, 2023).

Here, we focus on the visual reconstruction task, entailing the re-generation of a visual representation of a stimulus from brain data alone. To this end, we make use of generative adversarial networks (GANs) Goodfellow et al. (2014). This approach has been demonstrated to be highly effective for neural reconstruction tasks, as evidenced by previous research Dado et al. (2022, 2023). In brief, a decoder is trained to map neural responses to GAN latent vectors, which are then fed to the GAN's generator to reconstruct the corresponding images. The GAN latents of the training set can be acquired either through (1) the use of images already generated by the GAN so that the latents are accessible a priori or (ii) optimizing a latent vector such that its corresponding image matches the training stimulus in terms of perceptual similarity.

In this work, we have integrated an attention mechanism into neural decoding, enhancing predictive accuracy and shedding light on the relevance of specific brain regions in visual processing. Conventionally, queries in attention models are derived from the embedded input data. However, the opaque nature of neural data complicates this approach, as the potentially relevant neural features are not directly observable. To address this challenge, we introduce predictive attention mechanisms (PAMs), which employ learnable queries (Figure 2). This allows the model to dynamically discover and prioritize the features of the neural data most relevant to the specific task. Consequently, this innovative architecture significantly improves our ability to interpret and analyze brain activity through attention-based models.

2 Methods

2.1 Neural decoding with predictive attention mechanisms

The PAM architecture (Figure 2) is specifically designed to handle neural data as their relevant neural features are not directly observable (unlike, for instance, in text or image data). The input data, $Y = \{y_1, y_2, \ldots, y_c\}$, comprises neural data from c regions of interest (i.e., the number of attention

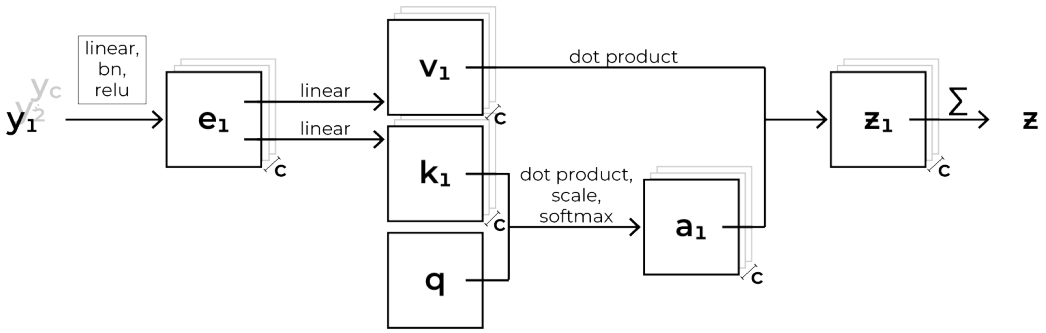

Figure 2: **Predictive attention mechanism (PAM)**. The input data, $Y = \{y_1, y_2, \ldots, y_c\}$ comprises neural data from c regions of interest (i.e., the number of attention heads) and the output z the decoded latent features of the stimulus. First, Y is transformed via n blocks, where each block consists of a linear layer, batch normalization, and ReLU activation to produce an embedded representation $E = \{e_1, e_2, \ldots, e_c\}$. Keys $K = \{k_1, k_2, \ldots, k_c\}$ and values $V = \{v_1, v_2, \ldots, v_c\}$ are derived from E using separate linear transformations. Note that each attention head has its own embedding-, key-, and value transformation. Unlike K and V, the queries q are learned during training. These queries interact with the keys through matrix multiplication, scaling and a softmax operation to compute attention weights, $A = \{a_1, a_2, \ldots, a_c\}$. Finally, the attention-weighted sum of V results in the predicted stimulus features, z.

heads), and the output z represents the decoded latent features of the stimulus. In our specific case, the model outputs are latent representations of a GAN which can be fed to the (pretrained) generator to synthesize corresponding images. We utilized StyleGAN-XL Sauer et al. (2022), trained on the ImageNet dataset Deng et al. (2009), to generate 512×512 pixel images from 512-dimensional feature-disentangled w-latents (rather than the z-latents).

First, the model embeds the neural data via multiple attention heads, where each head corresponds to the input from a specific brain area. This embedding process involves transforming Y through n blocks, each consisting of a linear layer, batch normalization, and ReLU activation to produce an embedded representation, $E = \{e_1, e_2, \ldots, e_c\}$. Formally, for each region i, this transformation can be represented as:

$$e_i = \text{ReLU}(\text{BN}(\text{linear}(y_i)))$$

Keys $K = \{k_1, k_2, \ldots, k_c\}$ and values $V = \{v_1, v_2, \ldots, v_c\}$ are derived from E using separate linear transformations:

$$k_i = \text{linear}(e_i)$$
$$v_i = \text{linear}(e_i)$$

Each attention head has its own embedding, key, and value transformation. Unlike K and V, the queries q are not predefined but learned during training by minimizing the error between the predicted and target outputs. This allows the PAM to identify and emphasize the most salient neural features indicative of the perceived images based on the interactions between queries and transformed versions of the input. The keys then interact with these queries to determine the focus of attention through matrix multiplication, scaling, and a softmax operation (i.e., the scaled dot product) to compute attention weights, $A = \{a_1, a_2, \ldots, a_c\}$:

$$a_i = \text{softmax}\left(\frac{q \cdot k_i^T}{\sqrt{d_k}}\right)$$

where d_k is the dimension of the key vectors, and the softmax function ensures that the attention weights sum to one.

The final output is computed as the weighted sum of the value vectors and the attention weights. This process allows the PAM to dynamically identify and emphasize the neural signatures most indicative of the perceived images. The attention-weighted sum of V results in the predicted stimulus features, z:

$$z = \sum_{i=1}^{c} a_i v_i$$

The training objective was to minimize the mean squared error (MSE) between the predicted and target latents:

$$\mathcal{L} = \text{MSE}(z_{\text{predicted}}, z_{\text{target}})$$

We employed the Adam optimizer with default parameters. Model weights were initialized using Xavier uniform initialization to ensure stable gradient flow. We used a batch size of 32 and continued training the model until convergence was achieved.

2.2 Neural datasets

To decode perceived images from brain data, we utilized two datasets comprising naturalistic images and corresponding brain responses. The first dataset, "B2G", includes synthetic images generated by StyleGAN-XL with their associated latent vectors readily available, offering a controlled setup for evaluating the decoding process. This dataset features multi-unit activity (MUA) recordings from visual areas V1, V4 and IT in one macaque, as detailed in Dado et al. (2023). In total, B2G consists of 4000 and 200 training (1 repetition) and test (20 repetitions) examples, respectively. The preprocessed MUA data was taken from Figshare at DOI 25637856. The second dataset, "GOD", contains natural images from ImageNet paired with fMRI responses from seven visual areas (V1-V4, FFA, LOC, PPA) in three human participants, as detailed in Shen et al. (2019). GOD consists of 1200 and 50 training (5 repetitions) and test (24 repetitions) examples, respectively. The preprocessed fMRI data (GOD) was taken from Figshare at DOI: 7033577/13.

2.2.1 Preprocessing steps

fMRI recordings of the GOD dataset were hyperaligned per brain area to map the subject-specific responses to a shared common functional space Haxby et al. (2020). Hyperalignment adjusts for individual differences in brain anatomy and functional topology such that, after transforming each participant's data into this common space, the average response across the three participants could be computed which reflects the typical response pattern while eliminating the variability between subjects. By doing so, we enhanced the reliability of subsequent analyses. Next, we z-scored these averaged training and test responses based on the training set statistics.

For voxel selection, we fit a ridge regression model with cross-validation to predict voxel responses from the latent vectors of the training examples, using 10-fold cross-validation to evaluate the model.

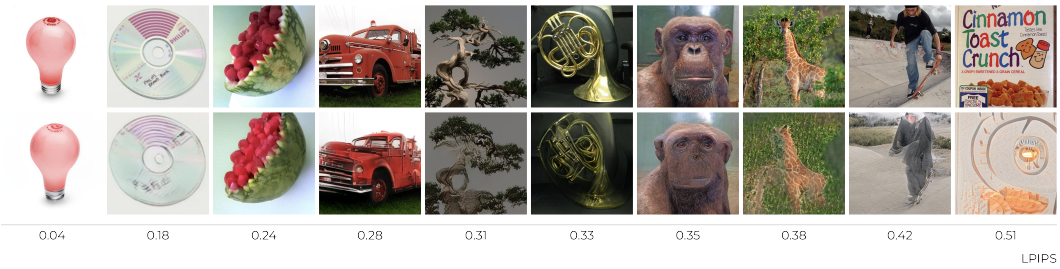

Figure 3: **Latent inversion from real photographs**. Ten arbitrary visual stimuli (top) from the GOD training set and their corresponding reconstructions from the inverted latents (bottom), with the LPIPS distance indicating the level of dissimilarity between them.

Table 1: **Voxel Selection.** The number of voxels pre- and post-FDR thresholding, which is used to eliminate less reliable responses. The voxels that remain post-FDR are considered more likely to be truly responsive to the visual stimuli. Notably, the voxel count in the PPA was reduced to zero following FDR adjustment, suggesting a lower reliability in the initial responses from this area.

	before FDR	after FDR
V1	2521	1643
V2	2687	1860
V3	2058	1451
V4	1283	968
LOC	2724	2134
FFA	2100	1210
PPA	900	0
Total	14273	9266

To optimize regularization, we explored a range of lambda values for the ridge parameter derived using singular value decomposition of the input matrix. Specifically, we filtered for non-zero singular values from the decomposition and used these to generate a set of lambda values. The range was determined by the square of the maximum and minimum non-zero singular values, generating five lambda values logarithmically spaced between these bounds. Based on the Pearson correlation between the predicted and target responses from the training set, we selected voxels using a false discovery rate (FDR) thresholding approach ($\alpha = 0.05$) per visual area to control the expected proportion of "false discoveries" (erroneously rejected null hypotheses) in multiple hypothesis testing (see Table 1).

Finally, for each training image, we optimized an input latent such that its corresponding image matched the stimulus in terms of VGG16 features by minimizing their learned perceptual image patch similarity (LPIPS) distance. Due to variability in approximation quality based on initial conditions, we repeated this ten times with a different seed and selected the latent that resulted in the lowest LPIPS distance with its corresponding image, ensuring the best match in perceptual similarity (see Figure 3).

2.3 Evaluation

To quantify the alignment between the original stimuli and their reconstructions, decoding performance was evaluated using three metrics, each based on cosine similarity, namely, learned perceptual image patch similarity, perceptual similarity and latent similarity. For LPIPS, we extracted feature representations from multiple layers of VGG16 pretrained for object recognition. For perceptual similarity, we also used feature representations of VGG16, but from five distinct levels following max pooling. As such, this resulted in five independent metrics that each reflected a different complexity level, with lower layers capturing more low-level image features and higher layers representing increasingly complex characteristics. Latent similarity measured the cosine similarity between the latent vectors of the original and reconstructed images.

2.4 Implementation details

All analyses were conducted using Python 3.10.4 on a cloud-based virtual machine equipped with an AMD EPYC 7F72 24-Core Processor (2.5 GHz - 3.2 GHz) and 96 cores, running a Linux kernel version 4.18.0-372.80.1.el8_6 on an x86_64 architecture. We employed the original PyTorch implementation of StyleGAN-XL and used VGG16 for object recognition to measure perceptual similarity during evaluation. The code to reproduce the main experimental results can be found on our anonymous GitHub repository.

3 Results

We trained two decoder models with PAM, tailored to the specific characteristics of the B2G and GOD datasets. For B2G, the embedding transformation consisted of five blocks to capture complex transformations from high-resolution MUA data to latent vectors. This contributed significantly to the superior performance of PAM over the baseline linear decoder, as evidenced by state-of-the-art

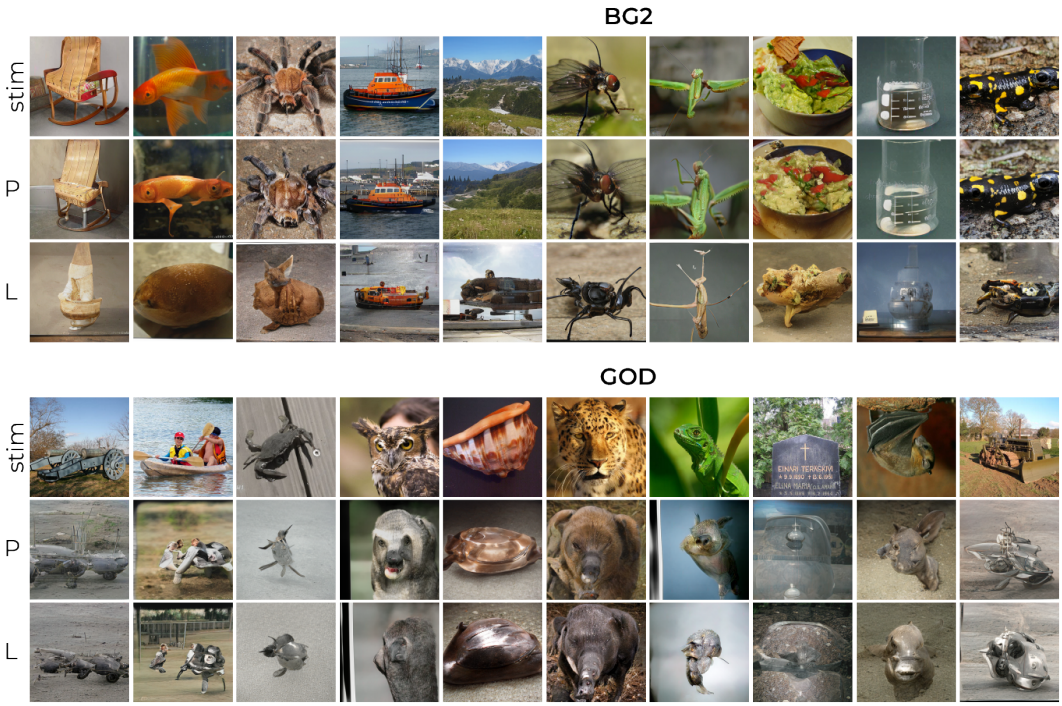

Figure 4: **Reconstruction results**. The upper and lower block show ten arbitrary yet representative examples from the B2G dataset (GAN-synthesized stimuli) and GOD dataset (natural stimuli), respectively. The top rows display the originally-perceived stimuli, the middle rows the reconstructions by PAM (P) and the bottom rows the reconstructions by the linear decoder baseline (L).

reconstructions (Figure 4) and quantitatively higher metrics (Table 2). Conversely, for GOD, which involves noisier, lower-resolution fMRI data and a smaller set of images, we limited the embedding layer to a single block to prevent overfitting and maintain model efficiency. While qualitative assessments suggest improved image reconstructions with PAM, the quantitative metrics indicate only marginal differences between PAM and the linear baseline. This is primarily due to the simpler model architecture used in the PAM for this particular dataset. Specifically, the embedding transformation in PAM consists of only one linear layer, just like the linear model. Consequently, both models are limited to extracting the same linear features from the data. Despite their similar performance, PAM

Table 2: **Reconstruction performance.** Results show reconstruction performance (mean $\pm$ standard error) across the B2G and GOD datasets, measured using seven metrics, including LPIPS, perceptual similarity at different levels of complexity (VGG 2/16, 4/16, 7/16, 10/16, 13/16) assessed through feature representations extracted from the VGG16 network, and latent similarity (Lat sim), which measures the cosine similarity between the original and predicted latent vectors. Note that the latent representations for the GOD dataset's real-world photographs are unavailable, so we could not include this as a metric. Results are shown for both PAM (P) and a baseline linear decoder (L). The last row in each block shows the p-values obtained from paired t-tests, indicating the statistical significance of the performance differences between PAM and the linear decoder: for B2G, PAM is significantly better than the linear decoder, but not for GOD, where the predictions were very similar.

		LPIPS sim	VGG 2/16	VGG 4/16	VGG 7/16	VGG 10/16	VGG 13/16	Lat sim
B2G	**P**	0.36 ± 0.007	0.43 ± 0.006	0.35 ± 0.004	0.29 ± 0.005	0.28 ± 0.009	0.36 ± 0.013	0.84 ± 0.006
	L	0.32 ± 0.005	0.41 ± 0.005	0.33 ± 0.004	0.26 ± 0.003	0.22 ± 0.004	0.25 ± 0.007	0.77 ± 0.003
		$p \ll 0.0001$	$p \ll 0.0001$	$p \ll 0.0001$	$p \ll 0.0001$	$p \ll 0.0001$	$p \ll 0.0001$	$p \ll 0.0001$
GOD	**P**	0.26 ± 0.009	0.38 ± 0.009	0.30 ± 0.006	0.23 ± 0.004	0.18 ± 0.005	0.17 ± 0.008	-
	L	0.25 ± 0.008	0.38 ± 0.009	0.30 ± 0.007	0.22 ± 0.005	0.18 ± 0.005	0.17 ± 0.007	-
		$p = 0.0546$	$p = 0.6709$	$p = 0.6600$	$p = 0.0678$	$p = 0.3340$	$p = 0.6634$	-

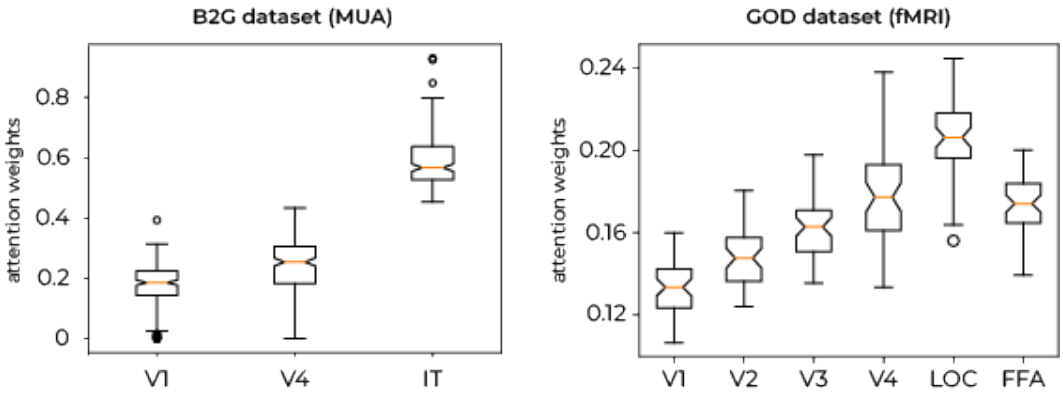

Figure 5: **Distribution of the attention weights.** The left panel illustrates the box plots of attention weights for the B2G dataset, derived from intracranial MUA recordings, across three regions of interest: V1, V4, and IT. The right panel displays the distribution of attention weights for the GOD dataset, obtained from fMRI recordings, across six ROIs: V1, V2, V3, V4, LOC and FFA. Each box plot shows the median (orange line), interquartile range (box), and the range excluding outliers denoted by circles. Notably, for B2G, V4 received the lowest and IT the highest attention. This trend contrasts with the GOD dataset, where attention is more evenly distributed across the areas, with slight peaks in regions specialized for higher-order processing such as LOC and FFA.

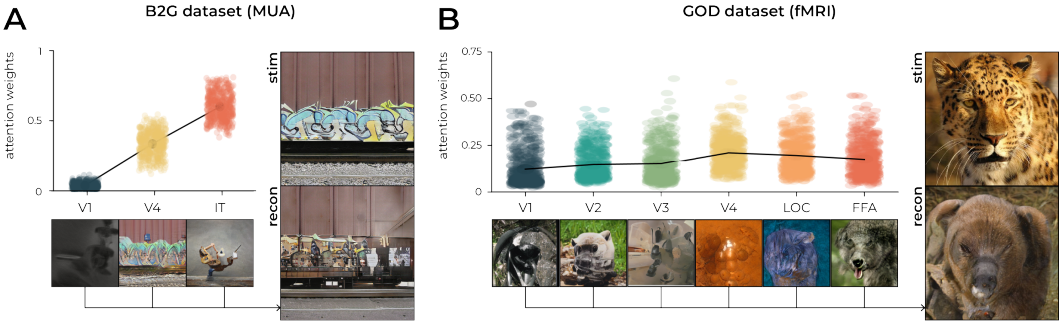

Figure 6: **Attention-weighted values by PAM.** The graphs visualize the distribution of 512-dimensional attention weights across the visual areas (V1, V4 and IT for B2G; V1, V2, V3, V4, LOC, and FFA for GOD) for two stimulus examples ('stim'; on the right of the graph). The black lineplot denotes the mean attention per neural area. We can notice a gradual increase of attention from up- to downstream visual areas (more subtle for GOD). Below each label in the graph (x-axis), we visualized the visual information from the corresponding values by feeding them to the generator of the GAN. We then took a weighted combination of the values and the attention weights to obtain the final latent corresponding to the final reconstruction ('recon'; displayed on the right, below the stimulus). For this example from B2G, particularly V4's visualized value seems to resemble the stimulus. And also for the example from GOD, the warm colors and the dotted pattern from the panther seem to be reflected in the reconstructed value of V4 but not necessarily in the final reconstruction itself.

still has significant interpretive advantages: it allows for the visualization of how attention is allocated across different brain areas but also provides insights into the specific contributions of these areas to the reconstructed images. As such, the distribution of attention weights across different visual areas revealed that more downstream areas (IT in the B2G dataset and LOC in the GOD dataset) generally received higher attention (Figure 5). This observation aligns with the existing finding that w-latents of StyleGAN-XL mainly capture high-level visual features relevant to high-level neural activity Dado et al. (2023).

Figure 6 shows how PAM is utilized to decode and reconstruct individual images from brain activity for the B2G and GOD datasets. The attention mechanism weighs the extracted values from the neural

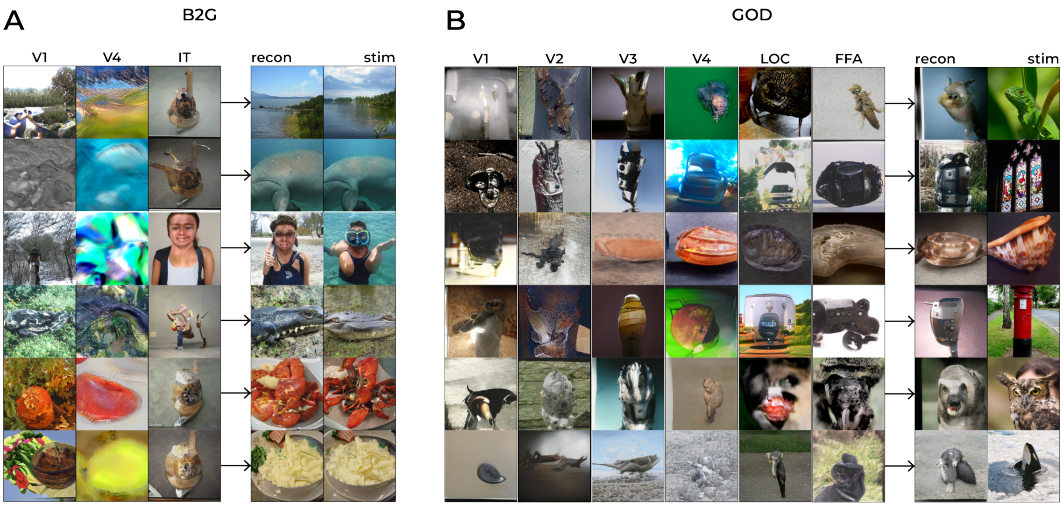

Figure 7: **Reconstructed values.** We visualized the information about the stimulus from each neural area by feeding their corresponding value to the generator of the GAN. For B2G, the reconstructed values from V1 seem to match the stimulus in basic outline, from V4 in color information and from IT in faces although the other reconstructions from this area seem rather meaningless despite the high assigned attention. Note that these stimuli are computer-generated such that the people in the third column do not really exist. For GOD, the reconstructed values from V1-2 seem to match the stimulus basic outlines as well, and V3 and V4 in shape and color information, respectively. The reconstructions from LOC and FFA seem to match the stimulus in terms of faces and contextual information.

data which carry information about the visual stimuli. We reconstructed these area-specific values to see what specific stimulus properties each brain area is processing (Figures 6 and 7). For B2G, the reconstructed values from V1 seem mostly similar to stimuli in terms of basic outlines, those from V4 capture the color information, and those from IT, while generally not revealing clear, meaningful features, notably include reconstructions of faces and animal faces (e.g., the diver in Figure 7A). However, the weighted sum of values and attention weights clearly resulted in a latent vector that integrated these region-specific contributions holistically as the final reconstructions resembled the original stimuli very closely. For GOD, the distribution of the attention weights across the neural areas seemed more uniform but still showed a slight increase in attention from early areas like V1 to V4, LOC, with a noticeable dip at FFA. The visualized values below the graphs show that areas V1-2 predominantly capture basic outlines. V3 captures more defined shapes while V4 captures color and textural information. The higher-order areas LOC and FFA seemed to capture more contextual information. As in the B2G dataset, the integration of values and attention weights into a single latent vector produced reconstructions that, while not as high-quality as those from MUA data, resemble the stimuli in their specific characteristics.

4 Discussion

In this work, we introduced PAMs as a powerful tool for handling complex input data where predefined queries are inaccessible. By applying PAMs to neural decoding of perceived images, we leveraged attention mechanisms to dynamically prioritize the most informative features of neural data. This approach not only achieved state-of-the-art reconstructions but also enhanced interpretability through the analysis of attention weights and values for each stimulus. The insights from this work hold promise for advancing brain-computer interfaces (BCIs) and neuroprosthetics, particularly for individuals with sensory impairments. By identifying the relevant brain areas for specific stimuli, we can refine BCIs for improved sensory processing. Furthermore, analyzing how attention is distributed could help to customize clinical interventions (e.g., improved treatments for visual disorders through targeted neural stimulation or improved neurofeedback paradigms for neurotherapeutics).

The reconstructions from the B2G dataset, which utilizes MUA data, appear superior to those from the GOD dataset. MUA data captures rapid and localized neural activity with high temporal and spatial resolution, providing clear, detailed signals with a high signal-to-noise ratio. This allows for the presentation of many images in one session, resulting in a larger dataset and subsequently for training a more complex model architecture (the embedding layer consisted of five blocks) without overfitting. Together with the fact that the latents underlying the images were readily available to train a decoder model (the images were generated from these latents by StyleGAN-XL), this dataset represents the optimal scenario for achieving the best possible reconstructions. In contrast, the GOD dataset used fMRI data, which measures slower hemodynamic responses that indirectly reflect neural activity, offering broader, less detailed information with lower temporal resolution and greater susceptibility to noise. Additionally, this dataset consists of real-world images without pre-existing latent representations, requiring us to post-hoc approximate these latents from the training stimuli to train our model. This lack of access to the precise latents further compromises reconstruction accuracy. This scenario presents a less optimal condition, reflecting the more challenging end of what can be achieved when using non-invasive fMRI data and real photographs. So, while MUA's invasive nature limits its use, fMRI's non-invasiveness sacrifices some precision and detail in neural activity, which compromises reconstruction quality.

Our results demonstrated a consistent trend where more downstream areas received increasingly more attention. There still was some variability in the allocation of attention across individual examples, which demonstrates that a PAM dynamically adapts to the unique characteristics of the neural data associated with each stimulus based on their match with the learned queries. Specifically, for GOD, attention increased progressively across higher-order visual areas but dipped again at the FFA (still receiving more attention than early areas like V1-V3). This could be attributed to the nature of the stimuli: the dataset included various animal faces (e.g., a panther and an owl) and human figures engaged in activities (e.g., a person playing the harp and two people in a canoe) but lacked prominent close-up human faces that are typical triggers for strong FFA activation. The absence of these specific face stimuli likely explains the reduced focus on FFA. Note that while PAMs enhance interpretability in some respects, the reasons why certain weights are assigned can still be opaque, which can, in turn, make it challenging to fully understand the underlying mechanisms guiding its performance.

Visualizing the values associated with each brain region adds insight into the model's decoding process: V2 and V3 capture basic outlines and shapes, while V4 captures colors and textures. Earlier visual regions correlate with low-level features (e.g., shape, color), whereas downstream areas encode higher-level attributes (e.g., object identities, contextual relationships) that integrate basic sensory features. Interestingly, while more downstream areas received higher attention from PAM, their reconstructed values often showed less visual similarity to the original stimuli than those from other areas (e.g., area V4). However, the final reconstructions are remarkably accurate when all values are integrated with the attention weights. This suggests that the values from deeper areas might not necessarily represent static features but rather act as directional vectors in latent space that guide the model toward specific regions necessary for reconstructing the overall perceptual quality of a stimulus. Therefore, these regions are considered very relevant to the model despite the disparity between their reconstructed values and the visual stimuli. As such, we believe that assigning greater weight to the more complex features leverages the richer semantic and contextual representations processed by higher-order regions of the brain's visual hierarchy.

The potential of PAMs extends far beyond their current application in neural decoding. Future studies should explore this further by integrating them across various domains and with other complex modalities where the queries cannot be predefined. Such research could revolutionize our understanding of how complex information is processed and interpreted.

Broader Impact

This research advances neural decoding and brain-computer interfaces, with significant potential to improve neuroprosthetics and sensory impairment therapies. While promising for clinical applications, it also raises concerns about mental privacy and the potential misuse of technology. Note that our models are trained specifically on the neural datasets used, which rely on full and constant subject cooperation, and cannot be reliably applied to data from other subjects. Further, these models can only reconstruct images that were externally perceived but not imagery or dreams. We are committed to transparency and reproducibility by providing open access to our code under appropriate licenses.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Cowen, A. S., Chun, M. M., and Kuhl, B. A. (2014). Neural portraits of perception: reconstructing face images from evoked brain activity. *NeuroImage*, 94:12–22.

Dado, T., Güçlütürk, Y., Ambrogioni, L., Ras, G., Bosch, S., van Gerven, M., and Güçlü, U. (2022). Hyperrealistic neural decoding for reconstructing faces from fmri activations via the gan latent space. *Scientific reports*, 12(1):1–9.

Dado, T., Papale, P., Lozano, A., Le, L., Wang, F., van Gerven, M., Roelfsema, P., Güçlütürk, Y., and Güçlü, U. (2023). Brain2gan: Feature-disentangled neural encoding and decoding of visual perception in the primate brain. *bioRxiv*, pages 2023–04.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE.

Du, C., Du, C., and He, H. (2017). Sharing deep generative representation for perceived image reconstruction from human brain activity. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1049–1056. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Güçlü, U. and van Gerven, M. (2013). Unsupervised learning of features for bayesian decoding in functional magnetic resonance imaging. In *Belgian-Dutch Conference on Machine Learning*.

Güçlü, U. and van Gerven, M. (2017a). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336.

Güçlü, U. and van Gerven, M. (2017b). Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in computational neuroscience*, 11:7.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., and van Gerven, M. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in neural information processing systems*, 30.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., and Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, 9:1–26.

Horikawa, T. and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15.

Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T., and Gallant, J. L. (2016). Decoding the semantic content of natural movies from human brain activity. *Frontiers in systems neuroscience*, 10:81.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685.

Kastner, S. and Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, 23:315–341.

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185):352–355.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., Sadato, N., and Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646.

Sauer, A., Schwarz, K., and Geiger, A. (2022). Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10.

Schoenmakers, S., Barth, M., Heskes, T., and Van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83:951–961.

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633.

Stansbury, D. E., Naselaris, T., and Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5):1025–1034.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., and Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104–1116.

van Gerven, M., de Lange, F. P., and Heskes, T. (2010). Neural decoding with hierarchical generative models. *Neural computation*, 22(12):3127–3142.

VanRullen, R. and Reddy, L. (2019). Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, 2(1):1–10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.