VoxelFormer: Parameter-Efficient Multi-Subject Visual Decoding from fMRI

Chenqian Le[1], Yilin Zhao[1], Nikasadat Emami[1], Kushagra Yadav[2], Xujin "Chris" Liu[1], Xupeng Chen[1], Yao Wang[1]

***Abstract*— Recent advances in fMRI-based visual decoding have enabled compelling reconstructions of perceived images. However, most approaches rely on subject-specific training, limiting scalability and practical deployment. We introduce VoxelFormer, a lightweight transformer architecture that enables multi-subject training for visual decoding from fMRI. VoxelFormer integrates a Token Merging Transformer (ToMer) for efficient voxel compression and a query-driven Q-Former that produces fixed-size neural representations aligned with the CLIP image embedding space. Evaluated on the 7T Natural Scenes Dataset, VoxelFormer achieves competitive retrieval performance on subjects included during training with significantly fewer parameters than existing methods. These results highlight token merging and query-based transformers as promising strategies for parameter-efficient neural decoding. The source code is available at `https://github.com/kushagrayadv/voxel-former`.**

Keywords: fMRI decoding, multi-subject learning, parameter efficiency, brain–computer interface, representation learning

I. Introduction

Decoding human visual perception from fMRI signals can transform brain–computer interfaces, clinical neuroimaging, and our understanding of how the visual cortex encodes complex scenes [1]–[3]. Most accurate decoders to date rely on massive subject-specific datasets or require extensive anatomical or functional alignment, hindering scalability and limiting practical deployment.

In this work, we ask: **Can we build a parameter-efficient visual decoder that leverages multi-subject training data effectively?** To address this, we propose **VoxelFormer**, a two-stage transformer architecture that (1) compresses and fuses raw voxel activations into a compact latent representation using a novel **ToMer encoder**, and (2) refines these features via a **Q-Former** to align with the CLIP image embedding [4] space while producing fixed-size representations across subjects.

We evaluate on the 7T Natural Scenes Dataset (NSD) [5] across eight participants. VoxelFormer achieves competitive performance for the subjects seen during training as well as lower parameter counts compare to other works.

Our main contributions are:

1) A **Token Merging Transformer (ToMer)** that dynamically reduces the fMRI token count via learned attention, lowering memory cost while preserving critical information.
2) A **query-driven Q-Former** that produces fixed-size latent representations enabling multi-subject training.
3) Demonstration of parameter-efficient multi-subject visual decoding achieving competitive performance with significantly reduced model size.

[1]Department of Electrical and Computer Engineering, New York University Tandon School of Engineering, Brooklyn, NY, USA. {cl6707, yz10381, ne2213, xl3942, xc1490, yaowang}@nyu.edu

[2]Department of Computer Science and Engineering, New York University Tandon School of Engineering, Brooklyn, NY, USA. ky2684@nyu.edu

II. Related Work

A. Subject-Specific Visual Decoding

Early fMRI-based visual decoders map each subject's voxel activations to image or feature representations. Shen *et al.* [6] trained deep generative models per subject to reconstruct images, requiring hundreds of images per individual. Scotti *et al.* [7] demonstrated image retrieval by fine-tuning a subject-specific encoder on CLIP features [8], achieving strong within-subject performance at the cost of per-user adaptation. The resulting model is referred to as MindEye1. MindEye1 maps 15000 visual-cortex voxels directly to the full 257×768 CLIP token matrix using a 4-block residual MLP that contains $\approx$ 940M parameters-over an order of magnitude larger than earlier linear or shallow-network decoders. Because this massive network is trained independently for each participant, it must see tens of thousands of stimulus-voxel pairs (30-40 hours of scans in the NSD) to avoid over-fitting to idiosyncratic voxel partterns and to learn a stable voxel-to-token mapping.

B. Cross-Subject Alignment

To mitigate the need for per-subject data, alignment methods project multiple brains into a shared space. Hyperalignment [6] align voxel responses across individuals with co-registration. While anatomical alignment is routinely performed in clinical MRI and may be necessary for meaningful multi-subject training (since the same voxel coordinate does not represent the same brain location across individuals), more recent networks [9], [10] focus on improving generalization through architectural innovations. The MindEye2 framework [9] uses subject-specific layers (a ridge regession layer) to map raw fMRI data from different training subjects into a common latent space, which is then transformed using a four-block residual MLP backbone to the features similar to CLIP features for the same input image. After pre-training the shared pipeline on seven NSD subjects ($\approx$ 250h of data in total), MindEye2 fine-tunes the entire model with as little as 1 hour of fMRI from a new subject, yet attains reconstruction quality comparable to a single-subject MindEye1 model trained on the full 40 hours scan set.

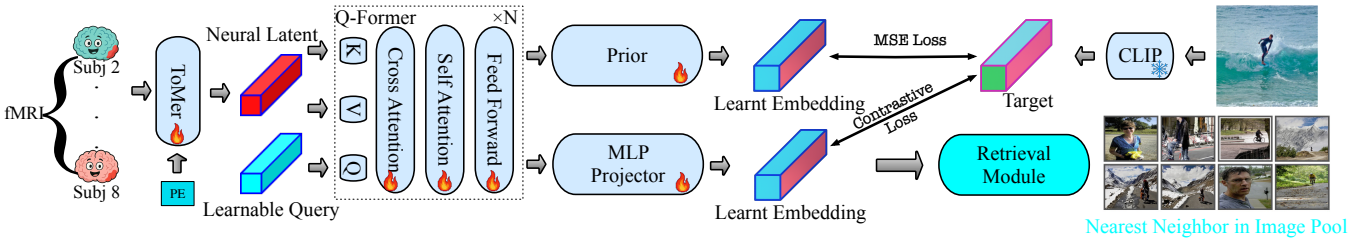

Fig. 1: Overview of the proposed **VoxelFormer** pipeline for cross-subject fMRI-to-image decoding. Multi-subject fMRI volumes are first encoded using a token-merging encoder (**ToMer**) with coordinate-based positional embeddings (PE) to generate compact neural latents. These latents interact with a small set of learnable queries within a Query-Former (**Q-Former**), composed of repeated cross-attention, self-attention, and feed-forward layers, to produce subject-invariant embeddings. The resulting representation branches into two decoding heads: (i) a Prior transformer trained with mean squared error (MSE) loss to regress frozen CLIP image embeddings, and (ii) an MLP projector trained with a contrastive loss for image retrieval via a nearest-neighbor search module. Trainable modules are denoted with a fire symbol, while the CLIP encoder remains frozen. In this work, only the retrieval branch is evaluated.

C. Transformer-Based Token Compression & Query Encoders

Token-merging techniques for vision transformers, such as Token Merging (ToMe) [11] and Tokens-to-Token (T2T) [12], reduce compute by merging redundant patches via attention. Perceiver [13] and Q-Former [14] architectures use learned queries to distill variable-size inputs into a fixed latent. However, these strategies have not been fully explored for fMRI decoding across subjects.

VoxelFormer integrates dynamic token merging and query-based encoding to achieve parameter-efficient multi-subject training for neural decoding.

III. Method

A. Dataset

We use the 7T Natural Scenes Dataset (NSD) [5], which comprises whole-brain, high-resolution fMRI from eight adults, each exposed to thousands of natural scene images from Microsoft COCO [15] over 30–40 sessions. This dataset is well-suited for evaluating non-invasive brain-based visual decoding. In our case, we use S2-S7 together to train the model.

B. ToMer Encoder

We introduce a novel Transformer-based encoder called **Token Merging Transformer (ToMer)** to efficiently process high-dimensional fMRI data. As depicted in Fig. 2, ToMer first tokenizes the input voxel data using a 1×1 convolutional layer, followed by the addition of sinusoidal positional embeddings derived from voxel coordinates via a SiREN [16] module. Subsequently, a self-attention mechanism captures the relationships among tokens, yielding latent neural representations and corresponding attention matrices.

Leveraging the learned attention scores, the ToMer encoder dynamically merges pairs of highly correlated tokens through the Token Merging operation [11]. In the original ToMe formulation, this attention-guided merging is applied only at inference time to accelerate forward passes by reducing the effective token count. In contrast, we integrate Token Merging directly into the training loop, merging tokens on-the-fly as gradients propagate. This yields a compressed, progressively coarsened representation throughout both learning and evaluation, substantially lowering computational and memory complexity during training without sacrificing task performance.

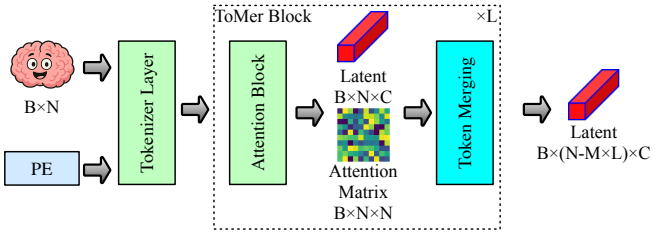

Fig. 2: **ToMer Encoder Architecture** The ToMer encoder processes input fMRI data (BN, where B is batch size and N is the number of voxels in the visual cortex) by first applying a Tokenizer Layer and Positional Embedding (PE). The tokenized features are then passed through an attention block, producing latent representations (BNC) and an attention matrix (BNN). A Token Merging operation reduces the number of tokens by merging those with the highest attention similarity. This encoder structure can be stacked L times and serves as the neural feature extractor in our VoxelFormer framework. If M tokens are reduced in each stage, the compressed latent representation has a shape B×(N–M×L)×C

The ToMer block can be stacked multiple times, progressively condensing the neural representation into a compact and informative latent space. This adaptive compression strategy is critical for enabling scalable and efficient decoding across subjects, making it well-suited for the cross-subject fMRI decoding problem tackled by our VoxelFormer model.

C. Q-Former

To enable robust multi-subject training and produce fixed-size representations regardless of the original number of

voxels, we propose to use [17], a query-based transformer module that produces consistent-sized neural embeddings (Fig. 1). Specifically, the Q-Former accepts compressed neural latent features from the ToMer encoder and utilizes learnable queries to flexibly attend and aggregate the most salient information from individual subject's brain data into a representation similar to a chosen learnt feature space. In our case, we used the CLIP features [8].

Specifically, the Q-Former employs a cross-attention mechanism where a fixed set of trainable query tokens repeatedly attend to the variable number of token features produced by the ToMer. The resulting embeddings from the query tokens at the last stage provide a consistent representation size that facilitates multi-subject training and alignment with visual features in the CLIP embedding space.

Note that the proposed pipeline in Fig. 1) enables training with multiple subject data without using subject-specific layers as in MindEye2.

D. Loss Function

To facilitate stable training and ensure compatibility with downstream tasks, we follow a dual-pathway training strategy inspired by recent work [9]. The output embeddings from the Q-Former branch into two distinct modules: (1) a *prior transformer*, which aligns embeddings to CLIP-derived visual embeddings via mean squared error (MSE) loss, theoretically enabling potential use as conditioning signals for diffusion-based image generation ; and (2) an *MLP projector*, trained with a contrastive loss, which directly supports robust image retrieval from a visual database through nearest-neighbor search. While our current experiments focus primarily on image retrieval performance, the prior transformer pathway could potentially be used for image reconstruction using a diffusion model. Training both branches together enables the shared modules (the ToMer and Q-Former) to produce good features for both branches.

Specifically, we use a two-phase training schedule: for the first one-third of the epochs, the MLP projector is trained with the BiMixCo contrastive loss (combined with MSE on the prior branch), and for the remaining two-thirds of training we replace BiMixCo with the SoftCLIP loss (while continuing to optimize the MSE term).

Mean Squared Error (MSE) Loss: This loss aligns the prior transformer embeddings $\mathbf{z}_i^{\text{prior}}$ with the CLIP-generated visual embeddings $\mathbf{z}_i^{\text{CLIP}}$ for the same i-th image stimulus using the MSE loss formula:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{B}\sum_{i=1}^{B} \|\mathbf{z}_i^{\text{prior}} - \mathbf{z}_i^{\text{CLIP}}\|^2 \tag{1}$$

Contrastive Loss with MixCo and Soft CLIP: The MLP projector embeddings $\mathbf{z}_i^{\text{MLP}}$ utilize a combination of InfoNCE contrastive loss and Mixup data augmentation, collectively referred to as BiMixCo. We further combine BiMixCo loss with SoftCLIP loss to enhance the discriminative power of the embeddings, while aligning the resulting features with the CLIP features.

BiMixCo Loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{BiMixCo}} = &-\sum_{i=1}^{N}\left[\lambda_i \cdot \log\left(\frac{\exp\left(\frac{p_i^* \cdot t_i}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{p_i^* \cdot t_m}{\tau}\right)}\right)\right.\\ &\left.+(1-\lambda_i)\cdot\log\left(\frac{\exp\left(\frac{p_i^* \cdot t_{k_i}}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{p_i^* \cdot t_m}{\tau}\right)}\right)\right]\\ &-\sum_{j=1}^{N}\left[\lambda_j \cdot \log\left(\frac{\exp\left(\frac{p_j^* \cdot t_j}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{p_m^* \cdot t_j}{\tau}\right)}\right)\right.\\ &\left.+\sum_{\{l|k_l=j\}}(1-\lambda_l)\cdot\log\left(\frac{\exp\left(\frac{p_l^* \cdot t_j}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{p_m^* \cdot t_j}{\tau}\right)}\right)\right]\end{aligned} \tag{2}$$

SoftCLIP Loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{SoftCLIP}} = -\sum_{i=1}^{N}\sum_{j=1}^{N}&\left[\frac{\exp\left(\frac{t_i \cdot t_j}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{t_i \cdot t_m}{\tau}\right)}\right.\\ &\left.\cdot\log\left(\frac{\exp\left(\frac{p_i \cdot t_j}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{p_i \cdot t_m}{\tau}\right)}\right)\right]\end{aligned} \tag{3}$$

Here, $p_i = \lambda_i\,\mathbf{z}_i^{\text{MLP}} + (1-\lambda_i)\,\mathbf{z}_{k_i}^{\text{MLP}}$ denotes the (possibly mixup-augmented) MLP-projector output for sample i; $t_i = \mathbf{z}_i^{\text{CLIP}}$ is the frozen CLIP image embedding for the same sample; and $t_m = \mathbf{z}_m^{\text{CLIP}}$ $(m = 1, \ldots, N)$ are all CLIP embeddings in the batch used as negatives (including $m = i$ in the denominator of each softmax).

Total Loss: The overall training objective is expressed as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{contrastive}}\mathcal{L}_{\text{contrastive}} \tag{4}$$

Here, λ_{MSE} and $\lambda_{\text{contrastive}}$ are hyperparameters that balance the trade-off between reconstruction alignment and discriminative retrieval performance. In our case, we use $\lambda_{\text{MSE}} = 30$ and $\lambda_{\text{contrastive}} = 1$. This dual-phase loss strategy, beginning with MixCo and transitioning to Soft CLIP loss after training for the first $\frac{1}{3}$ epochs, supports robust generalization and stable model convergence, facilitating both effective zero-shot retrieval and potential image reconstruction.

IV. Results

We evaluate VoxelFormer on the 7T NSD, following the standard top-1 retrieval protocol [7] with a candidate pool of 300 images. Image retrieval is performed by computing cosine similarity between brain-derived embeddings and CLIP image embeddings, then selecting the closest match. Forward retrieval measures accuracy when using brain embeddings to retrieve the correct image from the pool, while backward retrieval measures accuracy when using image embeddings to retrieve the correct brain response. Chance level performance is 0.33% (1/300).

a) Subject-Wise Performance: Table I summarizes retrieval accuracy for individual subjects. For our model, 7 subjects were used during training (subjects 2-7), with subject 1 was held out for evaluation of zero-shot retrieval performance. Compared to MindEye1 and MindEye2, which are state-of-the-art subject-specific and aligned models with substantially larger parameter counts, VoxelFormer achieves competitive performance on subjects included in the training set. It is important to recall that MindEye1 has a separate trained model for each subject, whereas MindEye2 trained a single model using subjects 2-7, with subject specific input layers plus a shared module. VoxFormer has a shared ToMer module and a shared Q-Former module, without any subject-specific layers. Despite the absence of the subject-specific layers, VoxelFormer achieved consistently above 66% accuracy on all evaluated subjects, underlining the robustness of our query-based representation.

TABLE I: Top-1 retrieval accuracy (%) by subject for Subjects within Training Data.

Subject	Method	Fwd Acc. (%)	Bwd Acc. (%)
2	MindEye1	97.1	93.9
	MindEye2	99.88	99.84
	Ours	86.54	85.78
3	MindEye1	90.7	85.7
	Ours	74.97	74.17
4	MindEye1	89.4	85.9
	Ours	75.15	73.36
5	MindEye2	98.39	96.94
	Ours	73.03	71.62
6	Ours	74.93	74.16
7	MindEye2	96.89	96.53
	Ours	68.65	67.46

TABLE II: Mean retrieval accuracy (Top-1) across training subjects and Model Size.

Method	Fwd Acc. (%)	Bwd Acc. (%)	Param
MindEye1 [7] (S1-S4)	93.6	90.1	940M
MindEye2 [9] (S1,2,5,7)	**98.3**	**98.3**	469M
Brain Diffuser [18] (S1-S4)	21.1	30.3	–
Ours (S2–S7)	74.3	73.1	**39M**

b) Parameter Efficiency: A major contribution of VoxelFormer is its parameter efficiency. Table II compares mean top-1 retrieval accuracy across training subjects for all methods, along with the model size. While MindEye2 achieves the highest mean retrieval accuracy, it uses over 469M parameters (counting one subject-specific layer for the S1, plus the shared module). MindEye2 achieves slightly worse performance, even though each subject specific model is twice the size the MindEye2. In contrast, VoxelFormer attains reasonable mean accuracy with only 39M parameters—a 12× reduction in model size compared to MindEye2 and 24× reduction when compared to MindEye1. This demonstrates that our proposed architecture can achieve competitive performance with significantly fewer parameters. Our findings suggest that architectural design—specifically attention-guided token merging and query-based feature distillation—can compensate for reduced capacity, offering an efficient path forward for future neural decoders, particularly in resource-constrained settings.

V. Discussion

We present **VoxelFormer**, a lightweight transformer framework that combines token-merging for voxel compression with a query-driven alignment module, enabling parameter-efficient multi-subject visual decoding from fMRI. While retrieval accuracy remains below state-of-the-art subject-specific approaches, VoxelFormer demonstrates that competitive performance can be achieved with significantly fewer parameters through careful architectural design.

Crucially, VoxelFormer is far more compact than recent baselines—39M parameters versus over 469M in MindEye2—while remaining competitive for subjects included in training. This demonstrates that token reduction and query-based transformers are promising strategies for parameter-efficient neural decoders that can be trained using data for multiple subjects.

Future work will explore improved cross-subject architectures, anatomical alignment strategies, larger pretraining datasets, and joint optimization for image reconstruction, with the goal of further closing the performance gap while maintaining parameter efficiency. VoxelFormer provides a foundation for parameter-efficient neural decoding that could be valuable in resource-constrained settings.

References

[1] D. L. K. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neuroscience*, vol. 19, no. 3, pp. 356–365, Mar. 2016, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nn.4244

[2] J. I. Glaser, A. S. Benjamin, R. H. Chowdhury, M. G. Perich, L. E. Miller, and K. P. Kording, "Machine Learning for Neural Decoding," *eNeuro*, vol. 7, no. 4, Jul. 2020, publisher: Society for Neuroscience Section: Research Article: Methods/New Tools. [Online]. Available: https://www.eneuro.org/content/7/4/ENEURO.0506-19.2020

[3] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, "Encoding and decoding in fMRI," *NeuroImage*, vol. 56, no. 2, pp. 400–410, May 2011.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021, arXiv:2103.00020 [cs]. [Online]. Available: http://arxiv.org/abs/2103.00020

[5] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, J. B. Hutchinson, T. Naselaris, and K. N. Kay, "A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence," *Nature Neuroscience*, vol. 25, no. 1, pp. 116–126, 2021. [Online]. Available: https://www.nature.com/articles/s41593-021-00962-9

[6] G. Shen, T. Horikawa, K. Majima, and Y. Kamitani, "Deep image reconstruction from human brain activity," *PLoS computational biology*, vol. 15, no. 1, p. e1006633, 2019.

[7] P. Scotti, S. Banerjee, A. Goode, P. Shabalin, A. Nguyen, A. Cohen, A. Dempster, C. Verlinde, E. Yundler, S. M. Weisberg, K. A. Norman, and A. Abraham, "Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, arXiv:2305.18274. [Online]. Available: https://arxiv.org/abs/2305.18274

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[9] P. Scotti, R. Tripathy, J. Torrico, E. Kneeland, Y. Chen, R. Narang, P. Santhirasegaran, T. Xu, T. Naselaris, K. A. Norman, and A. Abraham, "Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024, arXiv:2403.11207. [Online]. Available: https://arxiv.org/abs/2403.11207

[10] S. Wang, S. Liu, Z. Tan, and X. Wang, "MindBridge: A Cross-Subject Brain Decoding Framework," Apr. 2024, arXiv:2404.07850 [cs]. [Online]. Available: http://arxiv.org/abs/2404.07850

[11] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your ViT but faster," in *International Conference on Learning Representations*, 2023.

[12] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, J. Zhang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *ICCV*, 2021.

[13] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General Perception with Iterative Attention," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 4651–4664, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v139/jaegle21a.html

[14] J. Li, R. Li, C. Xiao, C. Fang, and J. Lu, "Blip-2: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2023.

[15] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," Feb. 2015, arXiv:1405.0312 [cs]. [Online]. Available: http://arxiv.org/abs/1405.0312

[16] V. Sitzmann, J. N. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *arXiv*, 2020.

[17] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 12 888–12 900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[18] F. Ozcelik and R. VanRullen, "Natural scene reconstruction from fmri signals using generative latent diffusion," 2023. [Online]. Available: https://arxiv.org/abs/2303.05334