FreqSelect: Frequency-Aware fMRI-to-Image Reconstruction

Junliang Ye[1]
u7727288@anu.edu.au
Lei Wang[†, 2, 3]
l.wang4@griffith.edu.au
Md Zakir Hossain[1, 3, 4]
zakir.hossain1@curtin.edu.au

[1] Australian National University
[2] Griffith University
[3] Data61/CSIRO
[4] Curtin University

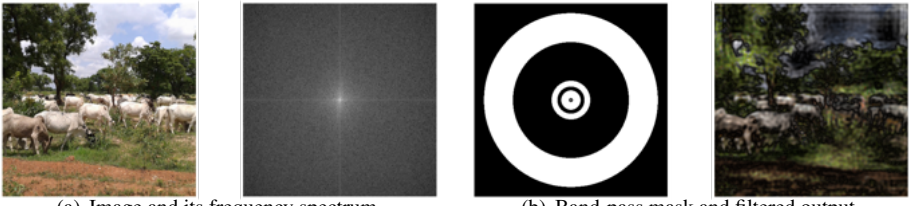
Abstract

Reconstructing natural images from functional magnetic resonance imaging (fMRI) data remains a core challenge in natural decoding due to the mismatch between the richness of visual stimuli and the noisy, low resolution nature of fMRI signals. While recent two-stage models, combining deep variational autoencoders (VAEs) with diffusion models, have advanced this task, they treat all spatial-frequency components of the input equally. This uniform treatment forces the model to extract meaning features and suppress irrelevant noise simultaneously, limiting its effectiveness. We introduce *FreqSelect*, a lightweight, adaptive module that selectively filters spatial-frequency bands before encoding. By dynamically emphasizing frequencies that are most predictive of brain activity and suppressing those that are uninformative, FreqSelect acts as a content-aware gate between image features and natural data. It integrates seamlessly into standard very deep VAE-diffusion pipelines and requires no additional supervision. Evaluated on the Natural Scenes dataset, FreqSelect consistently improves reconstruction quality across both low- and high-level metrics. Beyond performance gains, the learned frequency-selection patterns offer interpretable insights into how different visual frequencies are represented in the brain. Our method generalizes across subjects and scenes, and holds promise for extension to other neuroimaging modalities, offering a principled approach to enhancing both decoding accuracy and neuroscientific interpretability.

1 Introduction

Decoding visual experiences from functional magnetic resonance imaging (fMRI) is a central challenge at the intersection of neuroscience and machine learning. Natural images span a wide spectrum of spatial frequencies, from coarse structures to fine-grained textures, whereas fMRI signals are inherently noisy, temporally delayed, and spatially blurred due to the low resolution and hemodynamic nature of the measurements. This mismatch creates a fundamental obstacle: how can we recover the full richness of visual perception from signals that are both incomplete and noisy?

[†]Corresponding author.



(a) Image and its frequency spectrum. (b) Band-pass mask and filtered output.

Figure 1: (a) A sample image and its frequency-domain representation, where low-frequency components cluster near the center and high-frequency details appear toward the edges. (b) A circular band-pass mask highlighting intermediate frequencies (white = passed, black = suppressed), and the resulting spatial-domain reconstruction via inverse FFT, which preserves mid-scale structures while attenuating low-frequency shapes and high-frequency noise.

Recent advances have made notable progress by using two-stage generative pipelines. These approaches typically use a very deep variational autoencoder (VDVAE) [5] to produce a coarse reconstruction and then refine the semantic details using a latent diffusion model. Methods like Brain-Diffuser [19] and MindDiffuser [14] have shown impressive perceptual quality in image reconstruction tasks. However, they treat all spatial-frequency components of the visual input equally. This uniform treatment forces the model to simultaneously suppress noise and extract meaningful features across the entire frequency spectrum, leading to suboptimal use of model capacity and potential degradation in reconstruction fidelity.

In this work, we propose *FreqSelect*, a lightweight and adaptive frequency-selection module that filters spatial-frequency components of the input image prior to encoding. Rather than applying a fixed or uniform filter, FreqSelect dynamically adjusts the emphasis placed on different frequency bands based on their relevance to the underlying brain activity. It acts as a content-aware gate, suppressing uninformative frequencies while preserving those most predictive of neural responses. Crucially, FreqSelect integrates seamlessly into existing VDVAE–diffusion pipelines and introduces no need for additional supervision.

We evaluate FreqSelect on the Natural Scenes Dataset and show that it not only enhances reconstruction quality across both low- and high-level metrics, but also offers new insights into how different spatial frequencies are represented in the human brain. Our main contributions are:

i. **Adaptive frequency selection.** We introduce a dynamic, data-driven module that learns to gate spatial-frequency bands based on their informativeness for fMRI decoding, enabling better alignment between visual inputs and neural signals.

ii. **Improved reconstruction quality.** FreqSelect leads to consistent performance improvements in image reconstruction tasks by reducing the burden on downstream models to separate noise from signal.

iii. **Neuroscientific insight.** The learned frequency-selection patterns provide interpretable evidence about how the brain represents visual information at different spatial scales, opening new avenues for neuroscientific discovery.

In Appendix A.1, we review closely related work and highlight the key distinctions between our approach and prior methods. Below, we introduce our proposed method.

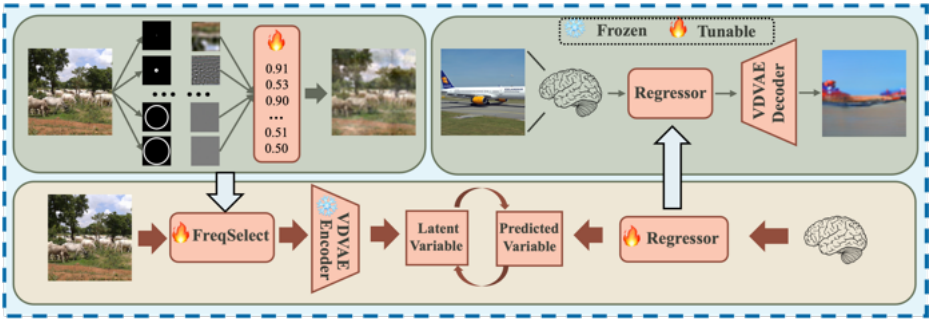


Figure 2: Overview of our fMRI-to-image reconstruction pipeline with FreqSelect (Stage 1). During training, the input image is first processed by *FreqSelect* (top left): An input image is decomposed into N frequency bands via circular masks in the Fourier domain. Each band is transformed back to the spatial domain, weighted, and summed to form a filtered image retaining informative frequencies. The resulting frequency-filtered image is then passed through a frozen very deep variational autoencode (VDVAE) encoder to obtain the "true" latent representation $\mathbf{z}_{\text{true}}$. In parallel, a ridge regressor maps fMRI signals to a predicted latent vector $\mathbf{z}_{\text{pred}}$, and the model is optimized by minimizing the latent-space loss $\|\mathbf{z}_{\text{true}} - \mathbf{z}_{\text{pred}}\|^2$. At inference time, the trained regressor (top right) generates $\mathbf{z}_{\text{pred}}$ directly from fMRI data, which is decoded by the frozen VDVAE decoder to reconstruct a low-level image. This output then serves as the input to Stage 2 of the pipeline.

2 Methodology

In this section, we present our complete fMRI-to-image reconstruction pipeline, with particular emphasis on the role and integration of the proposed *FreqSelect* module.

2.1 FreqSelect: Frequency-Aware Adaptive Filtering

Input and frequency transform. Given an input image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, where $C = 3$ denotes the RGB color channels and $H = W = 64$ is the image resolution, we begin by transforming each channel into the frequency domain via the 2D Discrete Fourier Transform (DFT):

$$\widehat{\mathbf{x}}_c(u,v) = \sum_{n=0}^{H-1}\sum_{m=0}^{W-1} \mathbf{x}_c(n,m)\, e^{-j2\pi\left(\frac{un}{H}+\frac{vm}{W}\right)}, \tag{1}$$

for each channel $c \in \{1,2,3\}$. To facilitate radial filtering, we apply an `fftshift` operation to center zero-frequency component. The radial distance from the center is then given by:

$$r = \sqrt{\left(u-\tfrac{H}{2}\right)^2 + \left(v-\tfrac{W}{2}\right)^2}, \tag{2}$$

so that small r values correspond to low frequencies and large r values to high frequencies. Figure 1(a) visualizes this transformation.

Band-pass decomposition. To isolate distinct frequency components, we divide the frequency space into N radial bands, using cutoff values $\{\nu_i\}_{i=0}^{N}$ with $\nu_0 = 0$ and ν_N equal to the Nyquist frequency. These values are linearly spaced across the radial frequency range.

For each band i, we define a binary mask:

$$M_i(u,v) = \begin{cases} 1, & \nu_{i-1} < r \leq \nu_i, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This mask selects a ring-shaped frequency band. Applying it to the DFT of the image and then performing an inverse DFT yields the corresponding band-limited image:

$$f_i(x) = \text{IDFT}\left(M_i \cdot \widehat{\mathbf{x}}\right). \tag{4}$$

This process produces a series of filtered images, each emphasizing specific frequency bands. Figure 1(b) illustrates an example mask and its effect on spatial content.

Adaptive frequency weighting and fusion. To prioritize the most informative frequency bands, we introduce a learnable scalar weight w_i for each band. These are passed through a sigmoid to ensure values between 0 and 1: $\alpha_i = \sigma(w_i) \in (0,1)$. The final filtered image is then computed as a weighted average of all band-pass outputs:

$$\widetilde{\mathbf{x}} = \frac{\sum_{i=1}^{N} \alpha_i f_i(\mathbf{x})}{\sum_{i=1}^{N} \alpha_i + \varepsilon}, \tag{5}$$

where ε is a small constant (*e.g.*, 10^{-10}) to prevent division by zero. This mechanism enables the model to adaptively emphasize frequency bands most predictive of neural responses. Figure 2 (top left) shows the entire FreqSelect pipeline.

2.2 Low-Level Image Reconstruction Using FreqSelect

Variational Autoencoders (VAEs) model complex data distributions by encoding an input $\mathbf{x}$ into a low-dimensional latent representation $\mathbf{z}$, typically under a Gaussian prior, and then decoding it to reconstruct $\mathbf{x}$. While conventional VAEs often fail to capture the full diversity and structure of natural scenes, VDVAE [5] addresses this limitation through a hierarchy of latent variables that progressively model finer-grained spatial details. Specifically, VDVAE defines a structured approximate posterior:

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = q_\phi(\mathbf{z}_0 \mid \mathbf{x})\, q_\phi(\mathbf{z}_1 \mid \mathbf{z}_0, \mathbf{x}) \ldots q_\phi(\mathbf{z}_K \mid \mathbf{z}_{K-1}, \mathbf{x}), \tag{6}$$

and a corresponding top-down generative prior:

$$p_\theta(\mathbf{z}) = p_\theta(\mathbf{z}_0)\, p_\theta(\mathbf{z}_1 \mid \mathbf{z}_0) \ldots p_\theta(\mathbf{z}_K \mid \mathbf{z}_{K-1}). \tag{7}$$

Here, $\mathbf{z}_0$ represents a coarse, low-resolution latent code, while deeper layers $\mathbf{z}_i$ ($i > 0$) refine the reconstruction by encoding increasingly fine-grained features. This hierarchical structure allows the model to generate rich, high-fidelity representations of natural images.

To maintain consistency with prior work, we adopt the exact VDVAE configuration used in brain-diffuser [19]. Specifically, we use a pretrained VDVAE trained on 64×64 ImageNet images, comprising 75 latent layers. However, following [19], we extract latents from only the first 31 layers, as deeper layers were found to contribute minimal additional benefit. These 31 latent outputs are concatenated into a single 91,168-dimensional feature vector.

During training, each stimulus image is passed through the frozen VDVAE encoder to obtain its latent representation. A ridge regression model is then trained to predict this latent vector from the corresponding fMRI pattern. At inference time, the trained regressor

generates a latent vector from a novel fMRI input, which is decoded by the frozen VDVAE decoder to produce a 64×64 image reconstruction. This coarse reconstruction serves as the initial input to the second-stage refinement module, based on latent diffusion.

Integration of FreqSelect. Figure 2 provides an overview of incorporating FreqSelect. The fused image $\widetilde{\mathbf{x}}$ is permuted and normalized to match the input format of the pretrained VDVAE encoder. This encoder produces a hierarchical set of latent variables $\{\mathbf{z}_k\}_{k=0}^{K-1}$, which are flattened and concatenated into a single latent vector: $\mathbf{z}_{\text{true}} \in \mathbb{R}^D$. This representation serves as the supervision target for learning to decode brain activity.

Stage 1 training objective. We train *FreqSelect* and the fMRI-to-latent regressor jointly, minimizing the latent space mean squared error:

$$\mathcal{L}_{\text{s1}} = \frac{1}{B} \sum_{b=1}^{B} \left\| \mathbf{z}_{\text{true}}^{(b)} - \mathbf{z}_{\text{pred}}^{(b)} \right\|^2, \tag{8}$$

where B is the batch size. During training, the VDVAE encoder is frozen. Only the frequency weights $\boldsymbol{w} \in \mathbb{R}^N$ and regressor parameters are updated. This setup encourages the model to learn which frequency bands are most informative for fMRI decoding.

Stage 1 inference procedure. At test time, unseen fMRI samples are mapped through the trained regressor to predict latent codes $\mathbf{z}_{\text{pred}}$, which are then decoded by the frozen VDVAE decoder to reconstruct full images. This enables robust, frequency-aware brain-to-image decoding that uses the structure of the visual frequency spectrum.

2.3 High-Level Refinement via Latent Diffusion

While the VDVAE encoder produces a coherent low-level layout from fMRI, it is limited in capturing high-level semantic content and photorealistic textures. To address this shortcoming and ensure consistency with baseline methods, we adopt the Versatile Diffusion latent diffusion model [36] as a second-stage refinement module (see Fig. 3). This model is pretrained on the LAION-2B-en dataset [27] at 512×512 resolution, using CLIP-ViT/L-14 [21] to separately extract image and text embeddings.

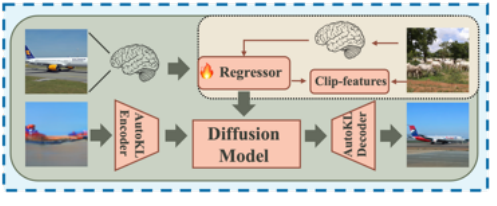


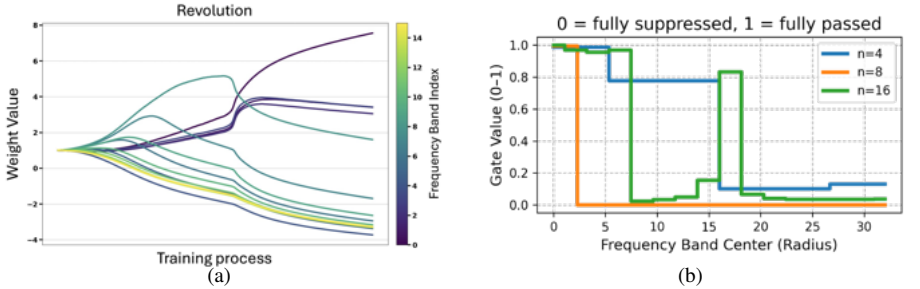
Figure 3: Stage 2: A latent diffusion model refines the VDVAE reconstruction using predicted CLIP features from fMRI, injecting high-level semantics and structure, for high-level image reconstruction.

Stage 2 training objective. During training, we freeze all components of the diffusion model, including U-Net [24] and AutoKL [23] modules. The training objective is to minimize standard noise-prediction loss:

$$\mathcal{L}_{\text{s2}} = \mathbb{E}_{t,\mathbf{z}_0,\varepsilon,y} \left\| \varepsilon - \varepsilon_\theta(\mathbf{z}_t, t, \tau_\phi(y)) \right\|^2, \tag{9}$$

where ε is used as ground truth noise that is added to $\mathbf{z}_0$, $\varepsilon_\theta(\cdot)$ is the model's prediction of noise. This loss compares the true injected noise ε to the predicted noise ε_θ, training the model to denoise $\mathbf{z}_t$ back toward $\mathbf{z}_0$. The noisy latent at timestep t is generated as:

$$\mathbf{z}_t = \sqrt{\bar{\beta}_t}\, \mathbf{z}_0 + \sqrt{1 - \bar{\beta}_t}\, \varepsilon, \tag{10}$$



(a) (b)

Figure 4: (a) Training dynamics of adaptive frequency band weights for $N = 16$. Each line corresponds to one of the 16 frequency bands, color-coded from low (purple) to high (yellow) frequencies. The trends show a progressive increase in low-frequency weights, transient peaks in mid-frequency bands, and gradual suppression of high-frequency components over the course of training. (b) Final learned pass-through rates for each frequency band under FreqSelect, shown for different band counts ($N = 4, 8, 16$). Values range from 0 to 1, indicating the relative degree of frequency preservation by each band.

Here, $\mathbf{z}_0$ is the initial latent representation obtained from AutoKL encoder, ε is a sample from a standard normal distribution: $\mathcal{N}(0, I)$, and $\tau_\phi(y)$ represents the CLIP-based conditioning features predicted from fMRI using a trained ridge regressor. $\bar{\beta}_t$ denotes the cumulative product of noise scaling factors up to timestep t, defined as $\bar{\beta}_t = \prod_{s=1}^{t} \beta_s$, and reflects the total signal retention after t diffusion steps. These embeddings can originate from either the target image or its caption, enabling semantic supervision from multiple modalities.

Stage 2 inference procedure At inference, the 64×64 VDVAE output from Stage 1 is upsampled to 512×512 and encoded by the frozen AutoKL encoder to obtain $\mathbf{z}_0$. We apply forward diffusion for $T_{\text{init}} = 37$ steps ($\sim$75% of the 50-step schedule), yielding a noisy latent $\mathbf{z}_{T_{\text{init}}}$. Reverse diffusion is then performed with cross-attention to predicted CLIP embeddings, and the denoised latent is decoded by AutoKL to produce a high-fidelity 512×512 reconstruction enriched with semantic detail, color, and structure, significantly surpassing VDVAE alone. We present our experiments and evaluations below.

3 Experiment

3.1 Experimental Setup

Dataset. We conduct all experiments on the Natural Scenes Dataset [1]. a 7 Tesla fMRI collection in which participants viewed images from COCO. From the original eight subjects, we selected the four (sub1, sub2, sub5, and sub7) who completed every trial for our analyses. Each image was presented for three seconds while subjects performed a continuous recognition task. The training set comprises 8,859 unique images and 24,980 fMRI trials (each image shown up to three times), and the test set comprises 982 images with 2,770 trials. we applied the General ROI mask, extracting responses from different voxels for each subjects respectively. For our FreqSelect module we vary the number of bands $N \in \{4, 8, 16\}$, spacing the cutoff frequencies uniformly over $[0, 32]$. All band-weight parameters $\boldsymbol{w} \in \mathbb{R}^N$ are initialized to $\mathbf{1}$, so that $\sigma(w_i) = \text{sigmoid}(1) \approx 0.73$ at start.

Metric. In order to comprehensively examine the reconstruction performance of the

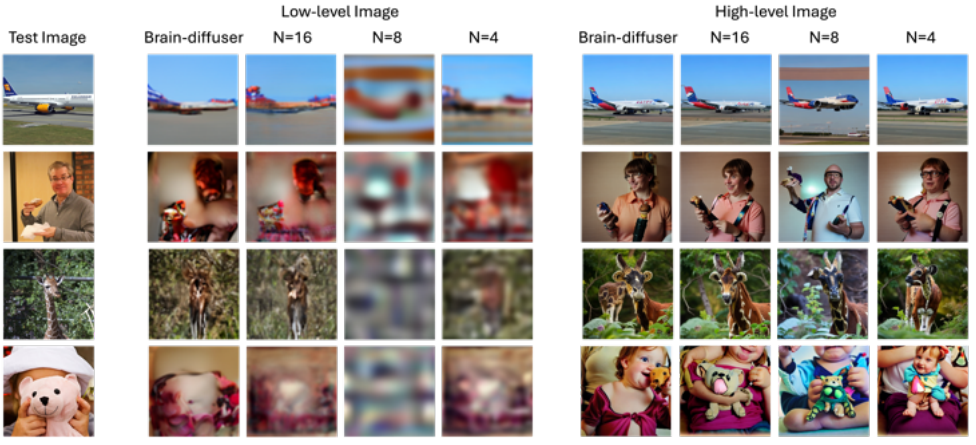


Figure 5: Comparison of fMRI reconstructions from Brain-Diffuser [19] and our FreqSelect. The first column shows ground-truth images. Each panel compares Brain-Diffuser with FreqSelect using $N = 4$, 8, and 16 frequency bands.

model at different levels, we use the following evaluation indicators: PixCorr quantifies pixel-level fidelity by computing the Pearson correlation between each reconstructed image and its ground-truth [20], whereas SSIM evaluates perceptual similarity through comparisons of luminance, contrast, and structural components [34]. Mid-level alignment is measured via AlexNet(2) and AlexNet(5), which correlate feature activations from the 2nd and 5th convolutional layers of AlexNet, respectively [11]. At the semantic level, Inception Score gauges both diversity and classifiability using the final pooling outputs of Inception-v3 [26], while CLIP Score computes the cosine similarity between CLIP-Vision image embeddings and their corresponding text embeddings [21]. Finally, reconstruction quality in learned feature spaces is assessed by the mean squared error in EfficientNet-B's feature maps [32] and by cosine similarity of SwAV-ResNet50 representations [3].

Below, we present both qualitative and quantitative results.

3.2 Qualitative Evaluation

FreqSelect as a dynamic spectral gate. We present two complementary visualizations of FreqSelect's behavior: the training-time dynamics shown in Figure 4(a), and the final learned pass-through rates across different band counts in Figure 4(b). In Figure 4(a), the lowest-frequency curves (dark purple) increase steadily throughout training, indicating that the model progressively relies on coarse, low-frequency components to represent global contours and basic structure. In contrast, mid-frequency bands exhibit a temporary rise, peaking mid-training, before being de-emphasized, suggesting transient reliance on texture and edge details. The highest-frequency bands (bright green to yellow) are consistently suppressed, often reaching negative values, which aligns with our goal of filtering out fMRI-induced high-frequency noise and preventing overfitting.

Functionally, FreqSelect acts as a trainable, differentiable pre-encoder gate that modulates the spectral content entering the VDVAE–diffusion pipeline. Figure 4(b) shows the

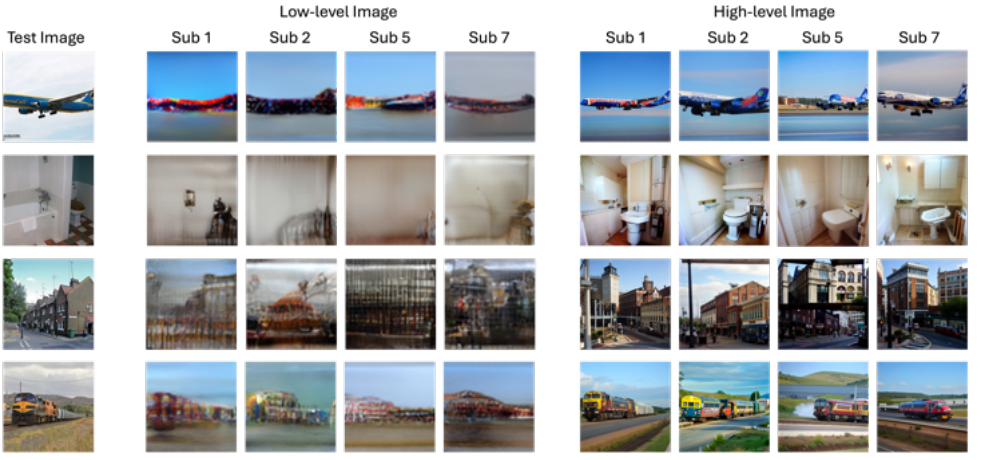


Figure 6: Example fMRI reconstructions using our FreqSelect module. The first column shows the ground-truth test image. In both the low- and high-level reconstruction panels, each subsequent column corresponds to an individual subject (Sub1, Sub2, Sub5, Sub7).

learned pass-through values (after sigmoid) for different band counts ($N = 4, 8, 16$). Across all cases, FreqSelect assigns the highest weights to the lowest bands, reflecting the importance of global structural information in fMRI-based decoding. At $N = 4$, the model relies almost entirely on the lowest band, discarding others. With $N = 8$, although more bands are available, the model still suppresses most beyond the first few, indicating that the useful signal remains concentrated in low frequencies. At $N = 16$, the finer spectral resolution allows a more nuanced behavior: lower bands are fully passed, middle bands receive moderate weights, and higher bands are effectively gated out. These learned pass-through rates are bounded between 0 and 1, providing smooth, differentiable control over frequency inclusion. Critically, FreqSelect learns these rates end-to-end, adapting to spectral profile of fMRI signal without relying on static, hand-crafted filters. This flexibility allows practitioners to choose N to balance frequency resolution and model complexity, ensuring that informative spectral bands are retained while noisy ones are suppressed.

Visual and subject-level impact of band selection. Although FreqSelect reliably preserves the lowest-frequency band regardless of N, increasing N improves spectral resolution at the cost of model complexity. However, poor choices, especially $N = 8$, can overly prune informative mid-frequency bands, degrading reconstruction quality. We analyze this trade-off in Figure 5, which compares fMRI-based reconstructions across different N values (4, 8, 16) and against Brain-Diffuser [19]. At $N = 4$ and $N = 8$, reconstructions exhibit over-smoothing, lacking texture and detail relative to Brain-Diffuser and $N = 16$. In particular, the $N = 8$ condition discards more mid-frequency content, yielding outputs with distorted contrast and incorrect colors. For instance, in the third row of Figure 5, Brain-Diffuser reconstructs a deer image that mismatches the test image, whereas FreqSelect, benefiting from better spectral filtering, correctly captures the number of deer. In the fourth row, Brain-Diffuser erroneously inserts a face behind a teddy bear, while FreqSelect correctly omits it, demonstrating its capacity to suppress misleading high-frequency artifacts.

We further assess generalization across subjects in Figure 6, where the same test im-

Method	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Brain-Diffuser [19]	**0.304**	0.293	**96.84%**	**97.48%**	**88.6%**	92.5%	**0.761**	**0.410**
Ours (N=4)	0.2902	0.2914	94.94%	96.80%	88.11%	91.85%	0.7725	0.4176
Ours (N=8)	0.0738	0.2633	86.82%	93.13%	86.00%	91.81%	0.7986	0.4520
Ours (N=16)	0.2734	**0.2961**	96.25%	97.46%	88.36%	**92.65%**	0.7672	0.4141

Table 1: Quantitative comparison of fMRI reconstructions across different N of our method and Brain-Diffuser. Best scores are bolded. Higher is better (↑) for PixCorr, SSIM, AlexNet(2/5), Inception, and CLIP; lower is better (↓) for EffNet-B and SwAV distances.

ages are reconstructed for four individuals (Sub1, Sub2, Sub5, Sub7). While inter-subject variability, such as in contrast or fine detail, is evident, semantic content and spatial layout are well preserved across all reconstructions. In low-level outputs (first row of each subject panel), structural features like airplane fuselage contours, sharp bathroom fixtures, building facades, and parallel train tracks are highly consistent. This consistency suggests that FreqSelect robustly captures task-relevant spatial-frequency components, even under substantial neural variability across individuals.

Characterization of reconstruction failures. Reconstruction errors reveal systematic challenges in capturing fine-grained semantic details despite faithful low-frequency structure recovery. Figure 7 illustrates typical failure cases: regardless of the original food type, be it a colorful bento, hamburger platter, or buffet, the reconstructions degrade into a generic "tower-shaped" arrangement of similar snack cups and green decorations. The model loses the diverse shapes, colors, and material textures, retaining only a universal "food combination" template. In contrast, for natural scenes like waves, the model successfully reproduces low-frequency features such as the texture and color gradients of seawater and the overall wave structure, demonstrating a strong capacity for capturing large-scale background patterns. However, finer details in the foreground, like human postures, facial expressions, and surfboard features, are heavily distorted or blurred. Semantic information is lost, resulting in imprecise outlines and ambiguous object identities.

Figure 7: Failure cases from fMRI reconstructions using FreqSelect.

3.3 Quantitative Evaluation

FreqSelect boosts reconstruction fidelity. Table 1 presents a quantitative comparison between Brain-Diffuser [19] and our proposed FreqSelect on subject 1, evaluated across three spectral settings ($N \in \{4, 8, 16\}$). Brain-Diffuser serves as a strong baseline, delivering high scores on both low-level metrics (PixCorr, SSIM, AlexNet) and high-level metrics (Inception, CLIP, EfficientNet-B, SwAV), as detailed in Section 3.

As N increases, FreqSelect gains representational flexibility. With $N = 4$, it matches Brain-Diffuser in capturing coarse image structure. At $N = 16$, it outperforms Brain-Diffuser on several metrics by finely suppressing high-frequency components dominated by noise, while preserving informative low-frequency content. This pattern highlights the value of fine-grained, adaptive band weighting in reconstructing images from fMRI signals. How-

Method	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Lin *et al.* [13]	-	-	-	-	78.2%	-	-	-
Takagi *et al.* [31]	-	-	83.0%	83.0%	76.0%	77.0%	-	-
Gu *et al.* [8]	0.150	0.325	-	-	-	-	0.775	0.423
Ferrante *et al.* [7]	**0.353**	0.287	89.00%	97.00%	84.00%	90.00%	-	-
Ozcelikfu *et al.* [19]	0.304	0.293	**96.84%**	**97.48%**	**88.6%**	92.5%	**0.761**	**0.410**
Ours (with *FreqSelect*)	0.2734	**0.2961**	96.25%	97.46%	88.36%	**92.65%**	0.7672	0.4141

Table 2: Quantitative comparison of fMRI reconstructions. Best scores are in bold.

ever, when $N = 8$, performance drops significantly. The model converges to retain only the lowest-frequency band, discarding mid- and high-frequency components that carry crucial structural and semantic cues. This failure stems from the coarse and uniform partitioning of the frequency domain: each band spans too wide a frequency range, mixing signal and noise. As a result, the model suppresses entire bands rather than selectively filtering noise, compromising overall reconstruction quality.

FreqSelect achieves the best balance of detail and semantics. Table 2 presents qualitative comparisons between FreqSelect and five state-of-the-art methods. Except for Brain-Diffuser, whose reconstructions are taken from [19], all other results are reported from their respective papers. Lin *et al.* [13] first used the Natural Scenes Dataset for fMRI-to-image reconstruction, using a StyleGAN2 generator. Takagi *et al.* [31] used a latent diffusion model to produce recognizable outlines from fMRI data. Gu *et al.* [8] used an instance-conditioned GAN trained on ImageNet to generate semantically coherent reconstructions. Ferrante *et al.* [7] introduced a multimodal alignment approach that co-reconstructs images and captions, yielding plausible silhouettes but lacking in low-level detail and texture.

Among all models, Brain-Diffuser [19] performs consistently well across most metrics. Its two-stage architecture uses a VDVAE encoder for robust global layout and a diffusion model for photorealistic refinement. Our FreqSelect module builds on this architecture by introducing adaptive frequency-band weighting to retain task-relevant signals and suppress irrelevant noise. This refinement addresses Brain-Diffuser's uniform frequency treatment, resulting in improved SSIM and CLIP scores. The SSIM gains stem from better preservation of structural edges and contrast, while the improved CLIP scores reflect enhanced semantic fidelity through retention of frequency content most aligned with vision–language embeddings. See Appendix A.2 for further discussion and Appendix A.3 for future work.

4 Conclusion

We presented *FreqSelect*, an adaptive frequency-selection module that enhances fMRI-to-image reconstruction by selectively filtering spatial-frequency bands before encoding. Integrated seamlessly into VDVAE-diffusion pipelines, FreqSelect improves reconstruction quality by emphasizing frequency components most predictive of neural activity while suppressing noise. Across multiple quantitative and qualitative benchmarks on the Natural Scenes Dataset, our approach consistently outperforms existing methods in both structural fidelity and semantic alignment. Moreover, the learned frequency patterns reveal interpretable insights into how the brain processes visual information at different spatial scales. FreqSelect offers a general, lightweight framework applicable across subjects and potentially extendable to other neural recording modalities, paving the way toward more accurate and interpretable neural decoding.

Acknowledgments

Junliang Ye conducted this research under the supervision of Lei Wang and Md Zakir Hossain as part of his final year master's research project at ANU. This work was supported by computational resources provided by the Pawsey Supercomputing Centre, a high-performance computing facility funded by the Australian Government. We sincerely thank the anonymous reviewers for their invaluable insights and constructive feedback, which have greatly contributed to improving our work.

References

[1] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J.B. Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x.

[2] S. Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience*, 2017.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924, 2020.

[4] Xueyan Chi, Pinxuan Lei, and Enze Xie. Fast fourier convolution for full-image tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3009–3018, 2020. doi: 10.1109/CVPR42600.2020.00307.

[5] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations (ICLR)*, 2021. OpenReview: https://openreview.net/forum?id=RLRXCV6DbEJ, code: https://github.com/openai/vdvae.

[6] Russell L. De Valois, David G. Albrecht, and Lars G. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22(5):545–559, 1982. doi: 10.1016/0042-6989(82)90113-4.

[7] Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Generative multimodal decoding: Reconstructing images and text from human fmri. In *Deep Generative Models for Health (DGM4H) Workshop, NeurIPS 2023*, December 2023. URL https://dgm4h.github.io/NeurIPS2023/Poster3.pdf. Poster; Submission Number: 3.

[8] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert R. Sabuncu. Decoding natural image stimuli from fmri data with a surface-based convolutional network. In *Medical Imaging with Deep Learning*, volume 227, pages 107–118. PMLR, 2023.

[9] Linda Henriksson, Niina SalminenVaparanta, Heikki Railo, and Simo Vanni. Retinotopic maps, spatial tuning, and locations of human visual areas in surface coordinates

characterized with multifocal and blocked fmri designs. *PLoS ONE*, 7(5):e36859, 2012. doi: 10.1371/journal.pone.0036859.

[10] Jie Huang and et al. Li. Adaptive frequency filters as efficient global token mixers. In *ICCV*, 2023. arXiv:2307.14008.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[13] Shiqi Lin, Zhipeng Huang, and et al. Wang. Deep frequency filtering for domain generalization. *CVPR*, pages XXXX–XXXX, 2023.

[14] Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, pages 5899–5908, Ottawa, ON, Canada, October 2023. doi: 10.1145/3581783.3613832.

[15] Damien J. Mannion, Daniel J. Kersten, and Cheryl A. Olman. Spatial frequency tuning in human retinotopic visual areas. *Journal of Vision*, 15(6):14, 2015. doi: 10.1167/15.6.14.

[16] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C. Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008. doi: 10.1016/j.neuron.2008.11.004.

[17] M. M. Murray, C. M. Michel, and ... Eeg mapping of neural responses to visual stimuli. *NeuroImage*, 2005.

[18] Thomas Naselaris, Ryan J. Prenger, Kendrick N. Kay, Michael D. Oliver, and Jack L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009. doi: 10.1016/j.neuron.2009.08.003.

[19] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion, 2023. Scientific Reports 13:15666 (2023). DOI:10.1038/s41598-023-42891-8. Supplied as additional material `brain-diffuser.pdf`.

[20] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[22] Nadav Rahaman, Dev Arpit, Aristide Baratin, Yan Acosta, Yannis Lin, ..., and Yoshua Bengio. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5301–5310, 2019.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[25] Nikhil Saldanha, Silvia L. Pintea, Jan C. van Gemert, and Nergis Tomen. Frequency learning for structured cnn filters with gaussian fractional derivatives. *BMVC*, 2021. arXiv:2111.06660.

[26] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, pages 2234–2242, 2016.

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

[28] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1): e1006633, 2019. doi: 10.1371/journal.pcbi.1006633.

[29] K. D. Singh, A. T. Smith, and M. W. Greenlee. Spatio-temporal frequency and direction sensitivities of human visual areas measured using fmri. *NeuroImage*, 12:550–564, 2000. doi: 10.1006/nimg.2000.0649.

[30] Haozhong Sun, Yuze Li, Zhongsen Li, Runyu Yang, Ziming Xu, Jiaqi Dou, Haikun Qi, and Huijun Chen. Fourier convolution block with global receptive field for mri reconstruction. *Medical Image Analysis*, 85:102349, 2024. doi: 10.1016/j.media.2024.103349.

[31] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[32] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6105–6114, 2019. doi: 10.5555/3327144.3327196.

[33] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS) 30*, pages 6306–6315, 2017.

[34] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[35] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *CVPR*, 2020. doi: 10.1109/CVPR42600.2020.01062.

[36] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7754–7765, October 2023. Open access version and code at https://github.com/SHI-Labs/Versatile-Diffusion.

A Appendix

A.1 Related Work

Neuroscientific insights on spatial frequency representation. Visual images can be decomposed into spatial-frequency bands, where low frequencies capture global structure and luminance contrast, while high frequencies convey fine textures and sharp contours [6]. Foundational studies using sinusoidal gratings in fMRI experiments have revealed that early visual areas exhibit selective, band-pass responses to specific spatial frequencies [9, 29]. In particular, V1 (primary visual cortex) responds most strongly to intermediate spatial frequencies that support edge detection and local detail [15]. Extrastriate areas such as V2 and V3 integrate signals from V1 into larger receptive fields, which tend to broaden frequency tuning and slightly shift preference toward lower frequencies [15].

While these neuroscientific findings highlight the brain's selective sensitivity to specific frequency ranges, existing neural decoding models have largely ignored this property by treating all frequencies uniformly. In contrast, our work is directly informed by these principles: *FreqSelect* incorporates a learnable, lightweight module that dynamically emphasizes spatial-frequency bands most predictive of neural responses. By aligning model inductive biases with cortical frequency preferences, FreqSelect not only improves decoding performance but also yields interpretable maps that reflect how different spatial scales are encoded in the brain.

Two-stage fMRI reconstruction with VAEs and diffusion models. Early fMRI-to-image decoding work relied on linear regression and Bayesian inference to reconstruct coarse stimulus representations from voxel responses [16, 18] . Subsequent methods introduced deep generator priors, matching DNN feature activations via GANs or DGN priors [28]. However, these single-stage methods either have difficulty capturing high-level semantics or robustly suppressing noise. Two-stage generative models have become the dominant architecture for fMRI-based image reconstruction. These methods typically involve an initial low-resolution approximation followed by refinement with a powerful generative prior. For instance, the Very Deep VAE (VDVAE) introduced by Child *et al*. [5] uses a deep hierarchy of latent variables to capture multi-scale image structure, from coarse layout to fine details. Similarly, VQ-VAE [33] uses vector quantization to discretize the latent space, which enhances feature preservation and semantic control.

Diffusion models have further improved reconstruction quality by using expressive priors in a compressed latent space. Stable Diffusion [23] and Versatile Diffusion [36] perform denoising steps in low-dimensional latent spaces, reducing computational cost while maintaining image fidelity. Building on these components, recent two-stage decoding pipelines, *e.g*., Brain-Diffuser [19] and MindDiffuser [14] have set the state of the art on the Natural Scenes Dataset. Brain-Diffuser maps fMRI signals to VDVAE latents to reconstruct coarse image layouts, which are then refined with Versatile Diffusion to add color, texture, and object-level details. MindDiffuser instead predicts VQ-VAE codes and CLIP embeddings to guide Stable Diffusion, combining structural accuracy with semantic consistency.

However, both models uniformly process all spatial-frequency components, treating high- and low-frequency information as equally informative. This design fails to account for the fact that fMRI signals contain noise that is not evenly distributed across frequency bands. *FreqSelect* addresses this shortcoming by learning to gate frequency components before encoding, enabling the model to suppress irrelevant frequencies and focus reconstruction capacity on informative ones, an approach grounded in the actual statistics of brain responses.

Frequency filtering in vision models. Frequency-domain processing has gained increasing traction in vision models as a means to improve efficiency, robustness, and generalization [4, 22]. Traditional convolutional neural networks (CNNs) operate in the spatial domain, implicitly learning filters with fixed receptive fields that impose certain frequency biases [12]. Studies on spectral bias have shown that deep networks tend to prioritize low-frequency components during training, which can hinder the representation of fine details [22, 35]. To mitigate this, some approaches use learnable Gaussian filters [25] or explicitly transform inputs via the Fast Fourier Transform (FFT) to retain specific frequencies while reducing spatial resolution [35].

Recent work has further integrated frequency-domain operations into model architectures. Fast Fourier Convolution (FFC) [4] processes features in both spatial and Fourier domains to capture long-range dependencies, while Deep Frequency Filtering (DFF) [13] applies adaptive frequency masks to enhance cross-domain generalization. In medical imaging, the Fourier Convolution Block (FCB) [30] improves MRI reconstruction by enhancing the effective receptive field using global frequency convolutions. Similarly, Adaptive Frequency Filters (AFF) [10] apply learned gating over frequency-transformed token representations in vision transformers, achieving strong performance–efficiency trade-offs.

These methods, however, are designed for classification, segmentation, or supervised reconstruction tasks with full access to image labels. Crucially, none are tailored for fMRI-to-image decoding, where the input signals are noisy, indirect reflections of the original image content. Our proposed *FreqSelect* is the first frequency-filtering module specifically optimized for neural decoding. It learns frequency gating via supervision from brain-predicted VDVAE latents, aligning its filters with neural evidence rather than image-based labels. Its modularity enables plug-and-play integration into existing pipelines without retraining large generative backbones, and the learned frequency profiles correspond to known band-pass tuning in early visual cortex, bridging computational modeling and cognitive neuroscience.

A.2 Discussion

Frequency band insights. FreqSelect dynamically adapts frequency emphasis during training, enhancing reconstruction fidelity by balancing noise suppression and detail preservation. As shown in Figure 4(a), the model initially prioritizes low-frequency bands, transiently boosts mid-frequency bands to capture texture, and suppresses high-frequency noise. At the optimal number of bands $N = 16$, FreqSelect achieves competitive improvements in SSIM and CLIP scores, especially in selective reconstruction scenarios (*e.g.*, correctly omitting occluded faces unlike Brain-Diffuser [19]). This demonstrates the model's ability to learn interpretable pass-through rates $\alpha_i = \sigma(w_i)$, modulating frequency bands adaptively to improve reconstruction quality.

Figure 4(b) further reveals the impact of band granularity: (i) At $N = 4$, only coarse low-frequency bands survive, producing oversmoothed outputs. (ii) At $N = 16$, FreqSelect finely balances mid- and low-frequency preservation against high-frequency noise, surpassing Brain-Diffuser in key metrics. (iii) At $N = 8$, equal-width banding over-prunes critical mid-frequency components, degrading performance across all metrics. This underscores the importance of choosing N based on the underlying fMRI power spectrum or exploring non-uniform frequency partitioning in future work.

Qualitative comparisons (Figures 5, 6) confirm that FreqSelect consistently produces sharper object contours, better color fidelity, and improved multi-object layouts, while generalizing robustly across subjects. Remaining failures (Figure 7) highlight the limitations

of overemphasizing low frequencies at the expense of high-frequency semantic details, a challenge compounded by MSE-driven training and fMRI's limited spatial resolution.

Future improvements should include auxiliary losses (*e.g.*, perceptual, adversarial) to encourage retention of mid- and high-frequency information, adaptive or data-driven band boundaries instead of uniform splits, and extension to other modalities such as EEG [17] and MEG [2], where frequency-specific signal quality varies. These advances could further close the gap to human-level decoding fidelity and deepen our understanding of how the brain encodes visual information across spatial scales.

A.3 Future Work

While FreqSelect demonstrates clear benefits in fMRI-to-image reconstruction, the reviews highlight several open challenges and opportunities for future research. We outline these directions below.

Choosing N and designing better band partitions. Our experiments revealed instabilities at intermediate band counts, most notably the performance dip at $N = 8$, likely caused by overly coarse, uniform frequency partitions that entangle signal and noise. Future work will explore *non-uniform, data-driven partitions* (*e.g.*, logarithmic spacing, k-means on the power spectrum, or differentiable cutoffs learned end-to-end) as well as *orientation-aware bands* (elliptical or steerable masks). To prevent the model from collapsing onto a single dominant band, we plan to incorporate *anti-collapse regularizers* such as entropy penalties, Dirichlet priors, or band-dropout. We will also provide practical guidelines for selecting N across data regimes, for example by relating N to voxel SNR, ROI coverage, or dataset size, and reporting iso-compute frontiers to help practitioners balance accuracy and cost.

Beyond uniform frequency gating. Currently, frequency gates are global and radial. An important next step is to design *spatially adaptive* gates (low-resolution gate maps predicted from the image or neural input) and *multi-scale variants* that operate at different encoder resolutions. Such designs may preserve critical mid-frequency content (textures and edges) without re-introducing high-frequency noise, improving the model's ability to capture structural detail.

Stronger and broader baselines. Our present implementation uses a frozen VDVAE encoder for comparability. Future work will examine unfreezing early VDVAE layers or inserting parameter-efficient adapters (*e.g.*, LoRA, bottleneck adapters) to allow end-to-end fine-tuning with FreqSelect. We will also evaluate alternative generative backbones such as VQ-VAE/VQ-GAN or MAE, and perform ablations on the number of latent layers used. This will test the robustness of FreqSelect across different encoders and strengthen the claim of generality.

To contextualize FreqSelect, we will compare against (i) fixed, hand-crafted band-pass filters, (ii) FFT-attention or adaptive frequency convolution networks, and (iii) no-gate baselines matched for compute. These experiments will clarify the unique advantages of adaptive, brain-informed gating over existing frequency-aware modules.

Objective functions for mid-/high-frequency retention. To address cases where fine semantic details collapse despite good low-frequency reconstructions, we will incorporate auxiliary objectives such as perceptual feature losses, frequency-domain consistency losses, or lightweight adversarial losses. These objectives explicitly encourage retention of mid- and high-frequency information while maintaining low-frequency denoising.

Generalization across datasets, modalities, and subjects. Thus far, FreqSelect has been validated only on NSD. We will extend evaluation to other naturalistic fMRI datasets,

more diverse stimulus sets, and even additional neural modalities such as EEG and MEG, where spectral SNR differs from fMRI. We also plan to examine cross-subject transfer, for example via subject-agnostic gates combined with subject-specific adapters, and analyze ROI-wise band preferences to link interpretability with known visual neuroscience findings.

Toward real-time feasibility. Although real-time decoding was not a focus of this work, we will explore low-latency variants of FreqSelect by precomputing masks, streaming FFTs, and implementing early-exit inference in the diffusion stage guided by fMRI-predicted CLIP features. We will quantify the accuracy-latency trade-off to assess online applicability.

Collectively, these directions address how to (i) design better partitions, (ii) balance spectral resolution with computational cost, (iii) generalize across models, datasets, and modalities, and (iv) mitigate mid-band suppression and failure cases. Pursuing these avenues will ensure that FreqSelect remains both practical for neural decoding applications and valuable for neuroscientific insight.