A. NSD Dataset

In this study, we leverage the largest publicly available fMRI-image dataset, the Natural Scenes Dataset (NSD) [2], which encompasses extensive 7T fMRI data collected from eight subjects while they viewed images from the COCO dataset. Each subject viewed each image for 3 seconds and indicated whether they had previously seen the image during the experiment. Our analysis focuses on data from four subjects (Sub-1, Sub-2, Sub-5, and Sub-7) who completed all viewing trials. The training dataset consists of 9,000 images and 27,000 fMRI trials, while the test dataset includes 1,000 images and 3,000 fMRI trials, with up to 3 repetitions per image. It is important to note that the test images are consistent across all subjects, whereas distinct training images are utilized.

We used preprocessed scans from NSD for functional data, with a resolution of 1.8 mm. Our analysis involved employing single-trial beta weights derived from generalized linear models, along with region-of-interest (ROI) data specific to early and higher (ventral) visual regions as provided by NSD. The ROI voxel counts for the respective four subjects are as follows: [15724, 14278, 13039, 12682].

A.1. Preprocessing

We perform **per-session z-score normalization** to centre each voxel to zero mean and scale it to unit variance within its respective session. All trials are retained individually for model training, while repeated presentations in the test set are averaged to yield a single, denoised beta pattern for each stimulus. To focus on visual cortical processing, we employ the official **nsdgeneral** ROI mask, which spans early to higher-order visual regions. Voxels within this mask are extracted and flattened into a one-dimensional sequence that serves as the fMRI input $x_{\text{fMRI}} \in \mathbb{R}^{1\times n}$ to the neural encoder.

A.2. Brain Functional Regions

High-level visual cortex contains multiple **category-selective regions** that respond preferentially to distinct types of visual stimuli. In humans, the most extensively studied functional regions are those selective for **faces, bodies, places, and food**, which exhibit reliable and dissociable responses across individuals. These regions provide a well-characterized framework for investigating category-specific neural representations and are commonly targeted in fMRI studies of visual processing.

Face-selective regions. These areas respond preferentially to faces and primarily include the fusiform face area (FFA). They are organized along the ventral visual pathway and support the perception of facial identity and expression.

Body-selective regions. Located adjacent to face-selective cortex, these regions respond strongly to images of human bodies, with the extrastriate body area (EBA) playing a key role in encoding body form, posture, and movement.

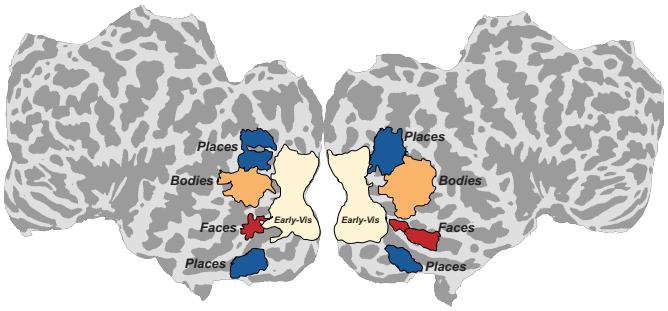

Figure 7. High-order functional regions in the visual cortex.

Place-selective regions. This network responds most strongly to scenes and environmental layouts. It comprises the parahippocampal place area (PPA) and the occipital place area (OPA), which collectively encode spatial structure and navigationally relevant information.

Food-selective regions. These regions are located within the ventral temporal cortex and show enhanced responses to edible items and food-related visual features.

B. Evaluation Metric

We evaluate visual encoding and decoding at the **semantic level** rather than at the voxel or pixel level, as lower-level analyses exhibit poor consistency between encoding and decoding, leading to blurry reconstructions.

i) Visual Decoding (fMRI→Image). Given fMRI signals, we generate images and assess their semantic fidelity against the original visual stimuli using multiple semantic metrics: i) **Incep:** A two-way comparison of the last pooling layer of InceptionV3; ii) **CLIP:** A two-way comparison of the output layer of the CLIP-Image model; iii) **Eff:** A distance metric gathered from EfficientNet-B1 model; iv) **SwAV:** A distance metric gathered from SwAV-ResNet50 model. A two-way comparison evaluates the accuracy percentage by determining whether the original image embedding aligns more closely with its corresponding brain embedding or with a randomly selected brain embedding.

ii) Visual Encoding (Image→fMRI→Image). Same as the decoding metrics above, but focuses on ***encoding-decoding consistency*** that assesses whether the neural representation of synthetic signals preserves critical semantic information for faithful image synthesis. Note that the encoding-only models (i.e., SynBrain [39] and MindSimulator [3]) need to employ the decoding-only model (i.e., MindEye2 [52]) as the frozen fMRI-to-image generator to evaluate the semantic quality of synthetic fMRI signals.

iii) Retrieval Metrics. We evaluate how well fMRI signals preserve semantic information by computing top-1 re-

Algorithm 1 NeuroVAE Architecture

Input: fMRI signal $x \in \mathbb{R}^{1\times n}$
Encoder:
  $x \leftarrow \texttt{Conv1D}(x,\ c\_out = 64)$
  $x \leftarrow \texttt{MLP}(x,\ h = 1024,\ d\_out = 6656)$
  $x \leftarrow \texttt{DownBlock}(x,\ num\_block = 2)$
Sampling:
  $[\mu,\ \log\sigma^2] = \texttt{Conv1d}(x,\ c\_out = 512)$
  $z_n \sim \mathcal{N}(\mu,\ \sigma^2)$ ▷ $z_n \in \mathbb{R}^{256\times 1664}$
Pre-Projector:
  $z_c = \texttt{Conv1d}(z_n,\ c\_out = 1)$ ▷ $z_c \in \mathbb{R}^{1\times 1664}$
Post-Projector:
  $x = \texttt{Conv1d}(z_c,\ c\_out = 256)$
Decoder:
  $x \leftarrow \texttt{UpBlock}(x,\ num\_block = 2)$
  $x \leftarrow \texttt{MLP}(x,\ h = 1024,\ d\_out = n)$
  $\hat{x} \leftarrow \texttt{Conv1D}(x,\ c\_out = 1)$
return $\hat{x} \in \mathbb{R}^{1\times n}$

trieval accuracy based on cosine similarity between neural latent representations ($z_n/\hat{z}_n$) and 300 candidate visual latent representations (z_v) extracted from test images, with one being the ground-truth visual stimulus for the fMRI data [39, 52]. Two signal sources are compared: (i) raw fMRI (Raw) $\rightarrow z_n$, and (ii) synthetic fMRI signals (Syn) $\rightarrow \hat{z}_n$. Retrieval performance is evaluated by calculating the average Top-1 retrieval accuracy (with a chance level of 1/300) and repeating the process 30 times to account for batch sampling variability.

Together, these metrics offer a comprehensive evaluation of semantic consistency across modalities, encompassing visual encoding, decoding, and the consistency between encoding and decoding.

C. Architecture Details

C.1. NeuroVAE Architecture

The overall architecture of NeuroVAE is summarized in Algorithm 1. Given an fMRI input $x_{\text{fMRI}} \in \mathbb{R}^{1\times n}$, the encoder first applies a 1×1 Conv1D, followed by an MLP and two downsampling blocks to produce intermediate features. A Conv1D layer then estimates the posterior parameters $[\mu, \log\sigma^2]$ for the latent representation $z_n \sim \mathcal{N}(\mu, \sigma^2)$. A pre-projector Conv1D compresses z_n into a channel-aggregated vector $z_c \in \mathbb{R}^{1\times d}$, which is mapped back through a post-projector Conv1D and two upsampling blocks, followed by an MLP and final Conv1D, to reconstruct the fMRI signal $\hat{x}_{\text{fMRI}}$. Here, c_out denotes the output number of channels, and d_out denotes the output number of feature dimensions. This design decouples the encoding and decoding pathways, resulting in a compact, probabilistic latent space that captures essential neural dynamics while supporting bidirectional cross-modal generative modeling.

Building on this foundation, NeuroVAE provides a variational backbone that constructs a semantically organized latent space for mapping between fMRI and visual embeddings. Compared with SynBrain [39], NeuroVAE introduces several key improvements motivated by biological plausibility, computational efficiency, and the requirements of generative neural modeling.

i) Biologically motivated feature processing. The encoder processes one-dimensional fMRI inputs using 1×1 Conv1D layers, which act as *channel-wise linear transformations* rather than spatial convolutions. Since ***voxel ordering in the 1D flattened fMRI vector does not encode meaningful spatial relationships***, this design avoids injecting artificial spatial inductive biases. SynBrain instead applies adaptive max pooling, which implicitly assumes spatial locality; NeuroVAE replaces this with an MLP projection to obtain a more biologically justified global transformation.

ii) Reduced attention dimensionality and improved computational efficiency. SynBrain computes attention over a $\mathbb{R}^{512\times 4096}$ representation, requiring four A100-40GB GPUs for training. NeuroVAE reduces the representation to $\mathbb{R}^{256\times 1664}$, which aligns with the CLIP visual feature space and allows training on a single A100-40GB GPU. This reduction preserves semantic alignment while substantially lowering memory and compute demands.

iii) Compact latent space enabling bidirectional modeling. SynBrain retains a $\mathbb{R}^{256\times 1664}$ latent tensor and supports only encoding. NeuroVAE aggregates channel-wise information into a single latent vector $z_c \in \mathbb{R}^{1\times 1664}$, cleanly separating the encoder and decoder pathways. This compact latent representation enables both neural reconstruction and generative fMRI synthesis through cross-modal flow matching.

iv) Cycle-consistency loss for semantically coherent fMRI synthesis. To ensure that reconstructed fMRI signals maintain semantic fidelity, NeuroVAE introduces ***a cycle-consistency loss*** that feeds synthetic signals back into the encoder and aligns their latent representations with the corresponding visual embeddings. This encourages generative fMRI signals to preserve semantic information instead of overfitting to voxel-level noise.

Together, these improvements allow NeuroVAE to construct a structured and probabilistic neural latent space that is both computationally efficient and better suited for semantic-level bidirectional modeling between visual and neural domains.

C.2. XFM Architecture

The Cross-modal Flow Matching (XFM) module is built on a SiT backbone [36] with temporal and positional embeddings. We use an 12-layer Transformer with 13 attention

heads per layer. The neural and visual latent representations share the same dimensionality, with 256 tokens and a feature dimension of 1664 (i.e., $z_n, z_v \in \mathbb{R}^{256\times1664}$), which is required for ***cosine interpolation*** when defining the continuous path between the two distributions. The patching and unpatching layers are removed since XFM operates directly in the high-dimensional latent space.

Bypassing the Gaussian noise distributions in standard implementations, we treat visual and neural distributions as the initial $(z_0, t = 0)$ and target distributions$(z_1, t = 1)$, and learn a reversibly consistent flow between them. This reframes the unidirectional denoising process into a unified cross-modal transport. We further remove the classifier guidance to facilitate a direct transport between the two distributions that aligns more closely with the biological process of visual encoding and decoding. For the first time, we reformulate visual encoding and decoding as ***a time-dependent, reversible process*** for unified modeling between neural and visual latent distributions. This formulation derives ***reversibility*** from the uniqueness of ordinary differential equation (ODE) solutions: the learned vector field can be integrated forward for visual encoding $z_{\mathrm{v}} \rightarrow z_{\mathrm{n}}$ or backward for visual decoding $z_{\mathrm{n}} \rightarrow z_{\mathrm{v}}$. Given that, ***encoding-decoding consistency*** is rigorously enforced by principles of flow matching. During inference, cross-modal translation is achieved by numerically solving the learned ODE with Euler updates parameterized by the learned vector field. In this way, NeuroFlow achieves a unified formulation for visual encoding and decoding, with ***the two processes distinguished solely by the temporal sampling direction***.

Ablation Study on Interpolation Schedule. We assessed the effect of different interpolation schedules on cross-modal flow performance. As reported in Tab. 4, NeuroFlow performs well under both linear and cosine schedules, demonstrating the robustness of our approach. The cosine schedule yields slightly better results, likely due to its smoother transition, which enables a more stable and accurate mapping between the visual and neural latent distributions. The smoother progression helps reduce abrupt changes in the latent space, facilitating better alignment and more reliable flow estimation. These findings indicate that while NeuroFlow is inherently robust, careful design of the interpolation path can further enhance performance.

Ablation Study on Sampling Step. We examined the effect of different sampling steps on NeuroFlow's performance using the Euler solver. As shown in Tab. 5, the model remains robust across varying step counts, with 20 steps providing an optimal balance between accuracy and computational efficiency. Increasing to 30 steps offers only marginal gains, indicating that 20 steps are sufficient to achieve high-quality results while maintaining efficiency.

D. Baseline Models

D.1. Framework

As illustrated in Sec. 5.2, a nontrivial ***modality gap*** remains between neural and visual distributions after contrastive learning in the first stage (i.e., **NeuroVAE**, or w/o $\mathcal{L}_{\text{XFM}}$.), leading to distorted image reconstructions. To evaluate the contribution of the proposed XFM module in bridging this gap, we construct two baseline configurations that replace XFM with simpler mechanisms:

i) NeuroVAE + MSE. This variant substitutes XFM with a direct mean-squared error loss between the neural and visual latent codes. It represents a naïve regression-based alignment strategy and tests whether pointwise matching alone is sufficient to reduce the modality discrepancy.

ii) NeuroVAE + LRs. This variant replaces XFM with two *independent* linear projection networks mapping between neural and visual latent spaces. As a non-unified framework, it provides a direct comparison point for evaluating the advantage of our unified XFM formulation in bridging the modality gap and maintaining encoding–decoding consistency.

By contrasting these baselines with **NeuroFlow (NeuroVAE + XFM)**, which provides a *single, unified flow-based transformation* between neural and visual latent distributions, we can directly quantify the benefits of a shared latent structure for improving encoding–decoding consistency and for effectively bridging the neural–visual modality gap.

D.2. Results

We present the quantitative and qualitative comparisons in Tab. 6 and Fig. 8. The results reveal clear differences between the baseline configurations and demonstrate the advantages of the proposed cross-modal flow matching (XFM) framework.

As shown in Tab. 6, incorporating a simple MSE objective (**NeuroVAE + MSE**) does not improve performance; instead, it causes noticeable degradation across decoding, encoding, and retrieval metrics. This suggests that pointwise supervision in a shared latent space is inadequate for aligning neural and visual representations and may even exacerbate the existing modality gap.

Replacing XFM with two independent linear projection networks (**NeuroVAE + LRs**) produces a mixed pattern of results. While decoding performance decreases slightly, both encoding accuracy and retrieval improve substantially. This indicates that linear mappings provide a more flexible alignment mechanism than a pointwise MSE loss. However, ***a linear network is not sufficient for modeling the complexities inherent in the latent distributions***, and the

Table 4. Ablation experiments on linear and cosine interpolation schedules.

Schedule	Decoding				Encoding (→Decoding)				Retrieval	
	Incep↑	CLIP↑	Eff↓	SwAV↓	Incep↑	CLIP↑	Eff↓	SwAV↓	Raw↑	Syn↑
Linear	95.4%	94.4%	.682	.377	98.5%	98.4%	.618	.359	86.4%	96.0%
Cosine*	**95.9%**	**95.0%**	**.675**	**.370**	**98.5%**	**98.7%**	**.600**	**.347**	**86.4%**	**96.4%**

Table 5. Ablation experiments on the effect of different sampling steps.

Step	Decoding				Encoding (→Decoding)				Retrieval	
	Incep↑	CLIP↑	Eff↓	SwAV↓	Incep↑	CLIP↑	Eff↓	SwAV↓	Raw↑	Syn↑
10	95.0%	94.3%	.682	.385	98.0%	98.3%	.617	.356	86.4%	96.0%
20*	**95.9%**	95.0%	.675	**.370**	**98.5%**	**98.7%**	.600	.347	**86.4%**	**96.4%**
30	95.7%	**95.3%**	**.674**	**.372**	**98.5%**	**98.7%**	**.597**	**.345**	**86.4%**	**96.4%**

independence of the forward and backward transforms prevents them from operating as a unified cross-modal mapping. As a consequence, the modality gap is only partially reduced, and the resulting encoding–decoding consistency remains limited.

Moving to a unified transformation framework, **NeuroVAE + XFM** (NeuroFlow) achieves the best performance across all evaluation metrics, including decoding, encoding-decoding consistency, and retrieval. The substantial gains highlight the effectiveness of XFM in ***bridging the neural–visual modality gap through a coherent flow-based alignment, while simultaneously preserving strong encoding–decoding consistency***.

These trends are also evident qualitatively in Fig. 8. NeuroFlow produces reconstructions that better preserve global structure and fine-grained visual semantics in both decoding (fMRI→Image) and encoding (Image→fMRI→Image), whereas **NeuroVAE**, **NeuroVAE + MSE** , and **NeuroVAE + LRs** baselines show *varying degrees of blurriness, structural distortion, or semantic drift*. Together, these results demonstrate that unified flow-matching alignment is essential for achieving high-fidelity cross-modal generation.

E. Additional Results

E.1. Subject-specific Results

Tab. 7 presents quantitative visual encoding and decoding results for four subjects (Sub1, Sub2, Sub5, Sub7). NeuroFlow consistently achieves high performance across decoding metrics (Inception, CLIP, Eff, SwAV), encoding-decoding consistency, and retrieval accuracy, demonstrating robust performance on different subjects. While absolute scores vary due to individual neural differences, overall trends remain stable, highlighting reliable model performance.

Qualitative results in Fig. 9 show visual reconstructions for Sub1, Sub2, Sub5, and Sub7, including direct decoding from fMRI and encoding-decoding cycles (image → fMRI → image). These examples confirm that NeuroFlow preserves semantic content and fine-grained visual details across subjects, maintaining strong encoding-decoding consistency and further demonstrating the robustness and generalizability of the model.

E.2. Brain Functional Analysis

As illustrated in Fig. 10, we present subject-specific fMRI activations generated by NeuroFlow along with their corresponding fMRI-to-image reconstructions across different functional domains, including face, body, place, and food-related regions. Despite substantial inter-subject variability in the spatial distribution of brain activations, NeuroFlow consistently focuses on the appropriate functional areas, such as the fusiform face area (FFA) for face stimuli, while generating semantically coherent reconstructions.

Furthermore, Fig. 11 shows subject-specific category-selective fMRI activations for Faces, Bodies, Places, and Food. These results reinforce that NeuroFlow reliably captures functional specificity across individuals, attending to the corresponding brain regions for each stimulus category and preserving consistent semantic content in the generated images.

Table 6. Quantitative comparison between NeuroFlow (NeuroVAE+XFM*) and baseline models.

Method	Decoding				Encoding→Decoding				Retrieval	
	Incep↑	CLIP↑	Eff↓	SwAV↓	Incep↑	CLIP↑	Eff↓	SwAV↓	Raw↑	Syn↑
NeuroVAE	87.7%	83.7%	.788	.497	60.3%	58.1%	.953	.644	86.4%	14.1%
NeuroVAE+MSE	85.2%	81.4%	.832	.536	51.0%	52.0%	.978	.671	86.4%	2.8%
NeuroVAE+LRs	86.8%	83.0%	.794	.496	90.3%	87.0%	.783	.493	86.4%	90.8%
NeuroVAE+XFM*	**95.9%**	**95.0%**	**.675**	**.370**	**98.5%**	**98.7%**	**.600**	**.347**	**86.4%**	**96.4%**

Table 7. Quantitative subject-specific visual encoding and decoding results.

Method	Decoding				Encoding→Decoding				Retrieval	
	Incep↑	CLIP↑	Eff↓	SwAV↓	Incep↑	CLIP↑	Eff↓	SwAV↓	Raw↑	Syn↑
Sub1	95.9%	95.0%	.675	.370	98.5%	98.7%	.600	.347	86.4%	96.4%
Sub2	95.4%	93.3%	.677	.362	98.5%	98.5%	.595	.344	81.3%	97.8%
Sub5	97.0%	95.9%	.645	.345	98.8%	99.0%	.569	.329	84.7%	98.3%
Sub7	93.9%	92.7%	.699	.378	98.5%	98.5%	.598	.344	69.8%	95.5%

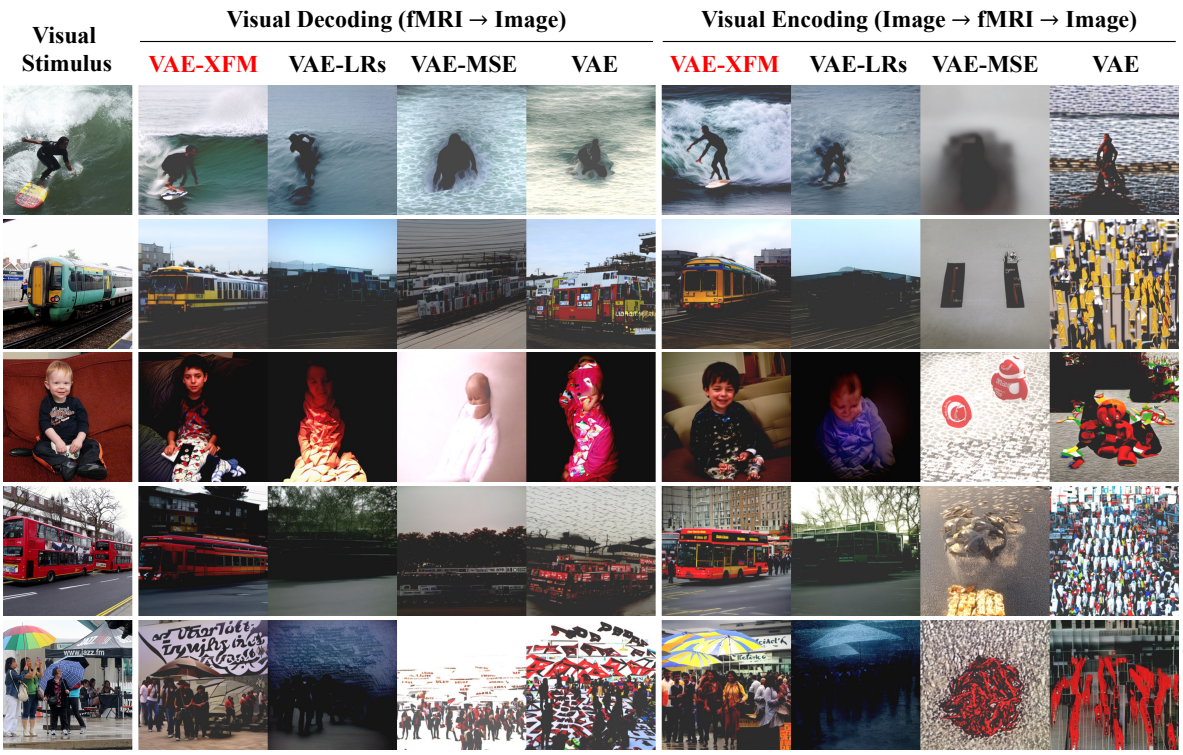

Figure 8. Qualitative comparisons between NeuroFlow (NeuroVAE-XFM) and baseline models.

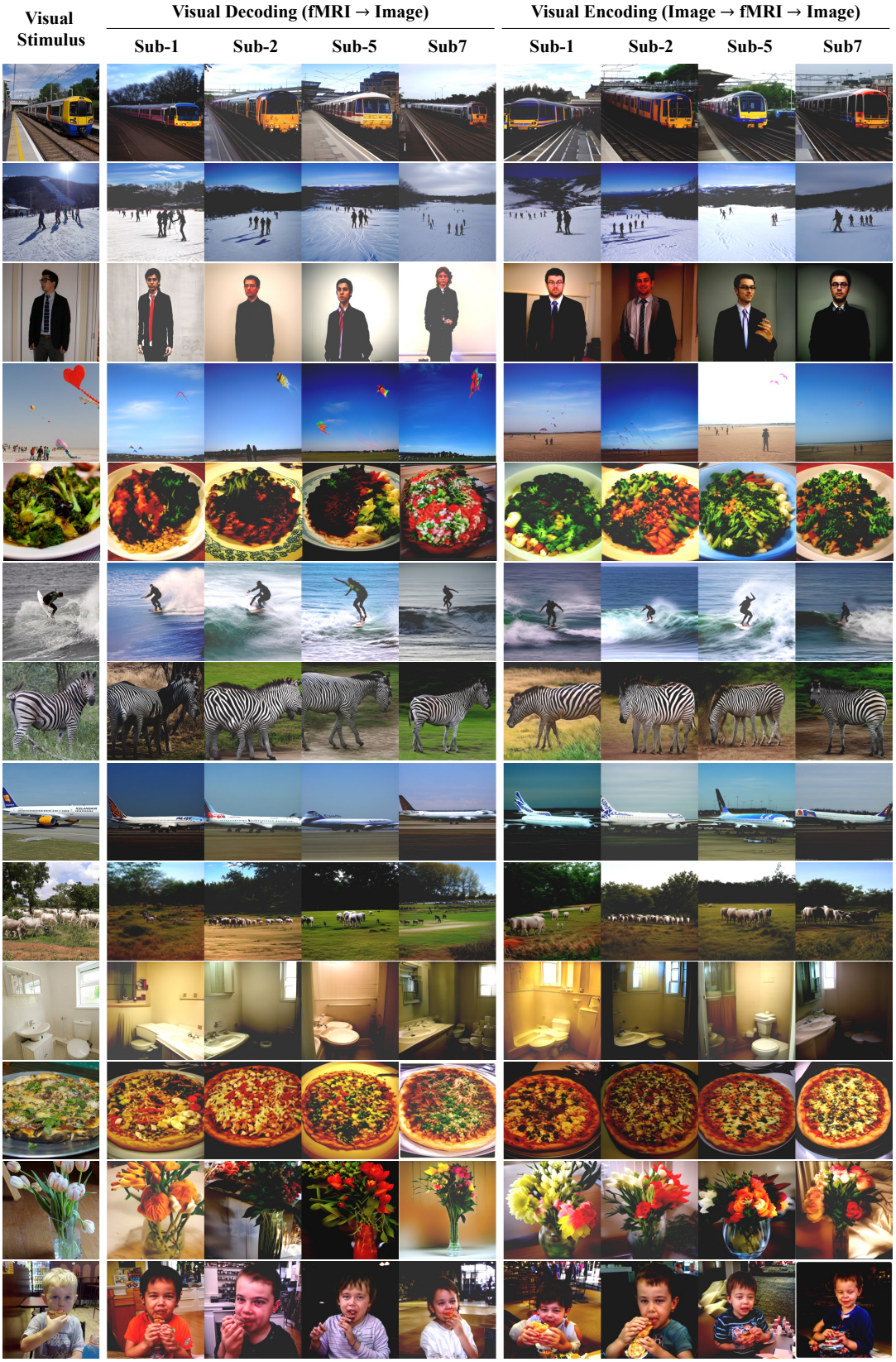

Figure 9. Qualitative subject-specific visual encoding and decoding results.

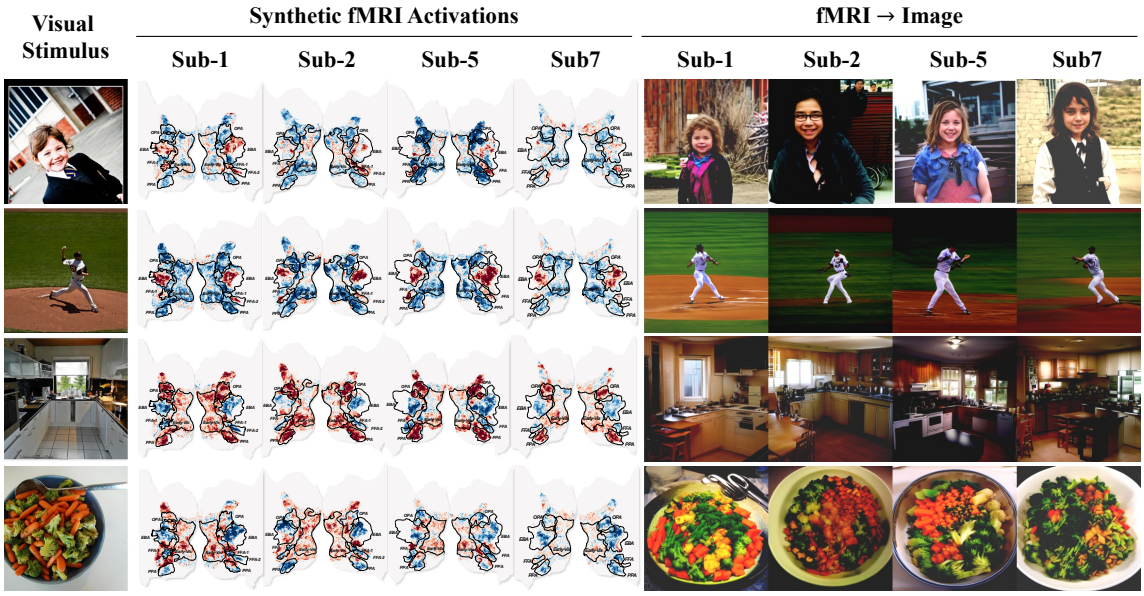

Figure 10. Subject-specific fMRI activations and corresponding image reconstructions.

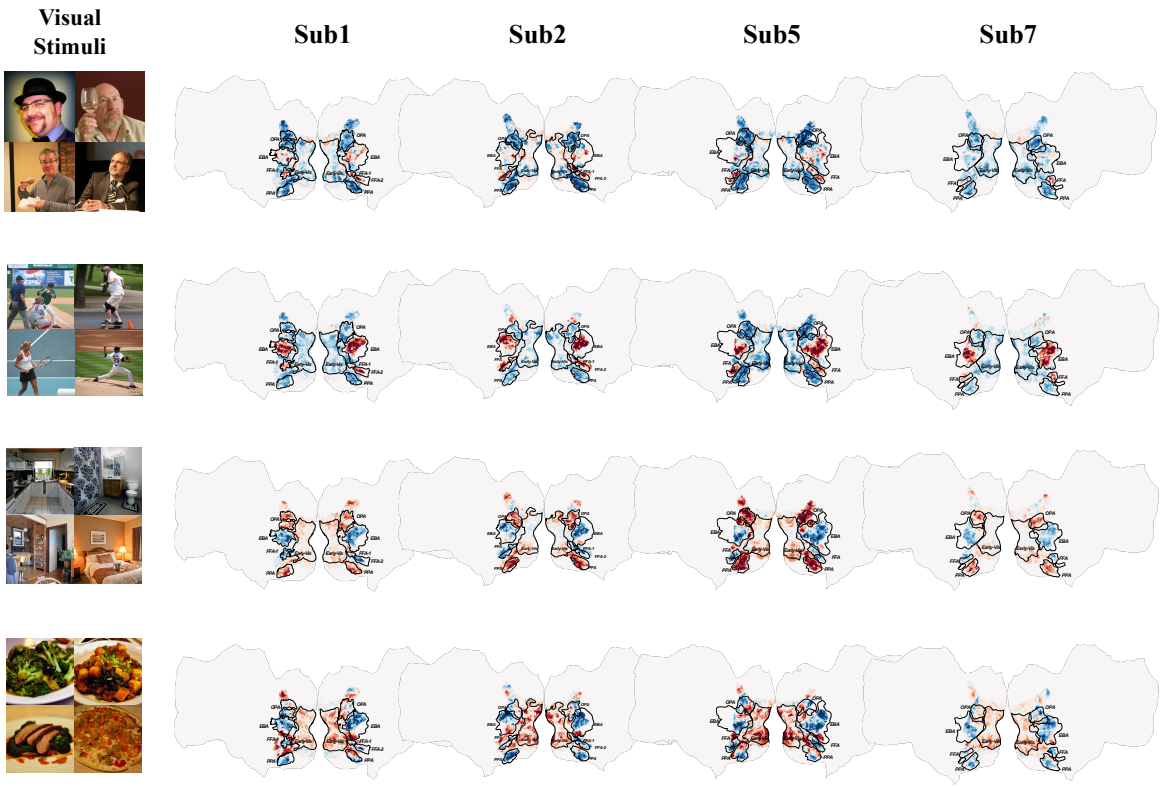

Figure 11. Subject-specific category-selective fMRI activations: Faces, Bodies, Places, and Food.