This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Psychometry: An Omnifit Model for Image Reconstruction from Human Brain Activity

Ruijie Quan[1], Wenguan Wang[1]*, Zhibo Tian[2], Fan Ma[1], Yi Yang[1]
[1] ReLER, CCAI, Zhejiang University, [2]Lanzhou University
https://github.com/QUANRJ/Psychometry

Abstract

Reconstructing the viewed images from human brain activity bridges human and computer vision through the Brain-Computer Interface. The inherent variability in brain function between individuals leads existing literature to focus on acquiring separate models for each individual using their respective brain signal data, ignoring commonalities between these data. In this article, we devise Psychometry, an omnifit model for reconstructing images from functional Magnetic Resonance Imaging (fMRI) obtained from different subjects. Psychometry incorporates an omni mixture-of-experts (Omni MoE) module where all the experts work together to capture the inter-subject commonalities, while each expert associated with subject-specific parameters copes with the individual differences. Moreover, Psychometry is equipped with a retrieval-enhanced inference strategy, termed Ecphory, which aims to enhance the learned fMRI representation via retrieving from prestored subject-specific memories. These designs collectively render Psychometry omnifit and efficient, enabling it to capture both inter-subject commonality and individual specificity across subjects. As a result, the enhanced fMRI representations serve as conditional signals to guide a generation model to reconstruct high-quality and realistic images, establishing Psychometry as state-of-the-art in terms of both high-level and low-level metrics.

1. Introduction

Understanding the intricacies of brain activity entails extracting meaningful semantics from complex patterns of neural activity [4, 5, 45, 54]. In the context of visual stimuli, neural responses in the brain are commonly measured by monitoring changes in blood oxygenation using functional Magnetic Resonance Imaging (fMRI) [29]. Over time, techniques for understanding brain activity based on fMRI have evolved from fMRI classification [11, 27] to the more challenging fMRI-to-Image reconstruction [4, 55]. In neuroscience studies [15, 42], individual brains are typically normalized to a template in an attempt to identify common patterns of activation across a group that can be generalized to a given population. However, it is evident that brains vary considerably in terms of shape and functional organization among individuals—normalization cannot fully compensate for these differences [3, 20, 51]. Furthermore, even if anatomical features are perfectly aligned, the same functional region may not occupy the same anatomical region in different participants [18, 55].

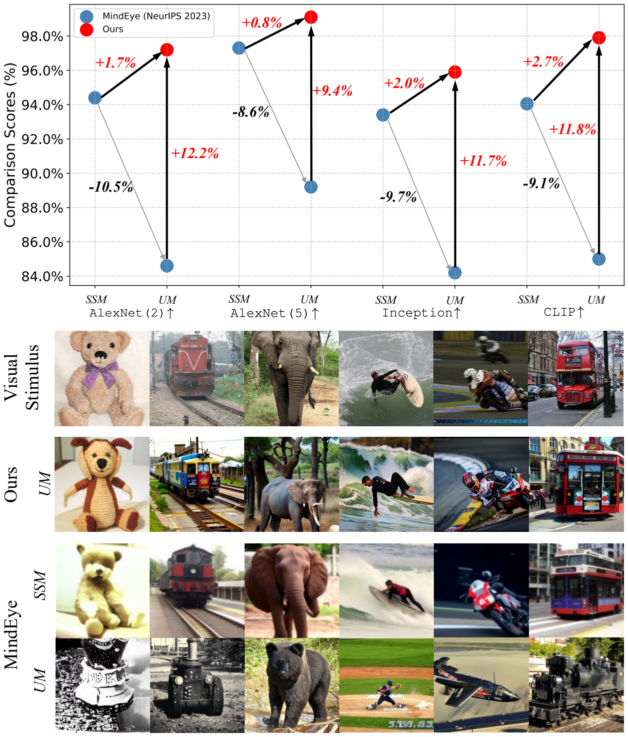

Figure 1. Current fMRI-to-Image methods (*e.g.*, MindEye [52]) train subject-specific models (*SSM*) on their respective data. They suffer obvious performance degradation when utilizing data from all the subjects to train a unified model (*UM*). Our *Psychometry* enables consistent performance improvements over MindEye by training one omnifit model on the amalgamated fMRI data.

*Corresponding author: *Wenguan Wang*.

The inherent variability in brain functioning across individuals adds complexity to interpreting brain activity. As a result, *all* existing fMRI-to-Image studies [9, 21, 32, 52, 62] delve into individual subject-specific characteristics by training separate models for each individual using their respective brain signal data. While these methods undeniably enhance the accuracy and semantic consistency of visual stimulus image reconstruction, they demand the development of individually tailored models for each subject. Not only does this consume substantial computational resources, but the specialized focus on individual differences may also potentially obscure the opportunity to uncover common patterns and similarities among subjects. Consequently, the exploration of broader inter-subject commonality largely remains uncharted territory. The most straightforward strategy is to amalgamate fMRI data from different subjects for training. Surprisingly, we find that state-of-the-art fMRI-to-Image methods [52] suffer obvious performance degradation when utilizing data from all the subjects to train a unified model, as illustrated in Figure 1. This divergence from the expected benefits of data scaling in improving deep learning model stability and performance [64] reveals the challenge of building a generalized model for diverse subjects, given their inherent individual differences.

To address this challenge, we propose *Psychometry*, an omnifit model for reconstructing images from fMRI data of various individuals. *Psychometry* can capture both the inter-subject commonalities and the individual variabilities through two essential components. **First**, drawing inspiration from the powerful concept of Mixture-of-Experts (MoE) [26, 53], *Psychometry* is equipped with an *Omni MoE* module, where all experts participate in the process of fMRI representation learning in order to capture the commonalities from fMRI data among subjects. Moreover, each expert is associated with subject-specific parameters aimed at addressing individual differences. In addition, Omni MoE adopts a *split-then-lump* mechanism with learnable splitting and lumping weights to maintain efficiency. **Second**, *Psychometry* employs a retrieval-enhanced inference strategy, termed *Ecphory*. This strategy retrieves the most relevant image or text CLIP [47] embedding from prestored training data (referred to as "memories") to enhance the learned fMRI representation via a mix-up approach. The enhanced representations serve as reliable conditional signals to guide a pretrained diffusion model in reconstructing high-quality and realistic images.

Psychometry enjoys a few attractive qualities: **First**, it significantly reduces the model size, training time, and computational resources required. This is achieved by the creation of an omnifit model that can handle fMRI data of different subjects, eliminating the need for separately training tailored models on subject-specific data. **Second**, *Omni MoE* along with the *split-then-lump* mechanism enables *Psychometry* to identify the inter-subject commonality and cope with the individual specificity in an efficient way. **Third**, with the help of *Ecphory*, *Psychometry* can further improve the fMRI embedding via incorporating the retrieved reliable information from the prestored subject-specific memories, leading to higher-quality image reconstructions from fMRI data.

In a nutshell, our contributions are three-fold:

- We propose *Psychometry*, an omnifit model designed to reconstruct images from fMRI data, representing a shift from separately trained models to a more comprehensive and generalized approach.
- *Psychometry* is integrated with an *Omni MoE* module, enabling all the experts to collectively identify the inter-subject commonalities and individual specificities among fMRI data from diverse subjects, along with a *split-then-lump* manner to ensure efficiency.
- *Psychometry* employs a retrieval-enhanced inference strategy, termed *Ecphory*, which accurately retrieves pertinent "memories" based on the acquired fMRI representation.

2. Related Work

Our work draws on existing literature in image reconstruction from fMRI and mixture-of-experts. For brevity, only the most relevant works are discussed.

Image Reconstruction from fMRI. Traditional fMRI-to-Image reconstruction methods [19, 28, 41] rely on fMRI-image paired data and utilize sparse linear regression to predict features from fMRI. In recent years, researchers have advanced the reconstruction from fMRI techniques by mapping brain signals to the latent space of generative adversarial networks (GANs) [32, 40, 44]. With the release of multimodal vision-language models [31, 34, 37, 38, 47, 61, 68], diffusion models [24, 50, 56–59, 71], and large-scale fMRI datasets [2, 7, 25, 66], image reconstruction from fRMI has reached an unprecedented level of quality [48]. These diffusion model-based methods [43, 52] explore mapping fMRI signals to both CLIP text and image embeddings by adopting individual regression models for each subject, subsequently utilizing the pre-trained diffusion model that accommodates multiple inputs for image reconstruction.

Though impressive, these methods primarily focus on individual subject analysis; they train specific models for different subjects on their respective data, thus ignoring the commonalities among these data. This highlights the need for a more universal and generalized framework, which is the core motivation behind this work. Furthermore, unlike previous methods that attempt to strictly align fMRI data to CLIP image or text embeddings, we introduce an inference-enhanced strategy named *Ecphory*. It retrieves the image or text CLIP embedding most relevant to the learned fMRI embedding from the pre-stored training data (memories) to enhance the learned fMRI representation as a reliable conditional signal. Moreover, *Ecphory* can effectively explore the individual specificity in the subject-specific memories.

Mixture of Experts (MoE). MoE initially suggests sharing certain experts at the lower levels and combining them through a gating network [26]. Recently, a sparse-MoE framework [53] was introduced, which routes each input to a subset of activated experts. This leads to a series of studies focusing on routing strategies within Sparse MoEs [39, 49]. In particular, [10, 14, 23, 69] introduce task-specific gating networks to choose different experts for processing information from each task. These methods demonstrate success across various applications, *e.g.*, recommendation system [39], natural language processing [16], and computer vision [1, 49], although the majority of existing works primarily focus on classification tasks [8, 13, 22].

This work represents the initial exploration of the application of MoE in the field of reconstructing images from fMRI data, with the aim of capturing inter-subject commonality and individual specificity across subjects. This concept is akin to multi-task learning, which involves utilizing a general model to handle diverse data. However, unlike multi-task learning MoEs that selectively activate experts to address task-specific attributes for different tasks, our focus is on exploring both the inter-subject commonalities and individual specificities present in the diverse individual fMRI inputs. To achieve this, we introduce an Omni MoE module, where all the experts work together to cooperatively learn the inter-subject commonality. Simultaneously, their associated subject-specific parameters enable different experts to capture the individual specificity.

3. Methodology

Task Setup and Notations. Our target is to reconstruct images from recorded fMRI data as the visual stimulus is presented to a healthy subject. The input fMRI data is usually preprocessed and extracted as a 1D vector of voxels. Formally, let $\boldsymbol{X}_{s,n} \in \mathbb{R}^d$ be the input preprocessed fMRI data as an image $I_n \in \mathbb{R}^{H \times W \times 3}$ was presented to the subject $s \in \{1, \cdots, S\}$, where d is the number of voxels and $n \in \{1, \cdots, N\}$. The latent representation of I_n and its corresponding caption text T_n are denoted as $\boldsymbol{I}_n \in \mathbb{R}^{v \times c}$ and $\boldsymbol{T}_n \in \mathbb{R}^{t \times c}$, respectively, which are obtained by feeding I_n and text T_n into CLIP [47]. v and c are the numbers of tokens of the CLIP image and text embeddings while c indicates their dimensions. Considering the individual variabilities across subjects, existing methods [35, 36] usually train separate models for each subject using their respective fMRI data, denoted as $\mathtt{f}_s^V : \boldsymbol{X}_{s,n} \rightarrow \mathbb{R}^{v \times c}$ and $\mathtt{f}_s^T : \boldsymbol{X}_{s,n} \rightarrow \mathbb{R}^{t \times c}$ to predict image- or text-aligned fMRI embeddings ($\tilde{\boldsymbol{I}}_n$ and $\tilde{\boldsymbol{T}}_n$) for each individual subject s. Those predicted features then serve as conditional signals for a diffusion-based model to reconstruct the viewed image $\tilde{I}_n$.

Method Overview. *Psychometry* is an omnifit model that can explain fMRI data from various subjects. Unlike existing methods necessitate the creation of S models and require training S times for every single modality, *Psychometry* only needs to be trained once using the amalgamated fMRI data, *i.e.*, $\mathtt{p}^V : \boldsymbol{X}_n \rightarrow \mathbb{R}^{v \times c}$ and $\mathtt{p}^T : \boldsymbol{X}_n \rightarrow \mathbb{R}^{t \times c}$. *Psychometry* involves two core modules: i) an *Omni MoE* layer (§ 3.1) that exploits inter-subject commonalities and individual specificities; and ii) a retrieval-enhanced inference strategy (*Ecphory*, §3.2). An overview of our complete pipeline can be found in Figure 2 and the detailed network architecture is presented in §3.3.

3.1. Omni MoE for Learning Inter-Subject Commonality and Individual Specificity

Omni MoE Layer. To achieve a full exploration of both the inter-subject commonality and individual specificity from the amalgamated fMRI data of various subjects, our *Psychometry* is equipped with a *Omni MoE* layer. Specifically, in *Omni MoE*, there are multiple experts who work in a collaborative manner to capture the commonalities, while each of these experts is assigned a set of subject-specific parameters so as to cope with the individual variabilities. Moreover, *Omni MoE* is empowered with a *split-then-lump* mechanism to ease the computational load and prohibit overfitting caused by learning with all experts. The above designs of our *Omni MoE* are encapsulated into a network layer and deeply embedded into the Transformer blocks.

Formally, the *Omni MoE* layer contains a group of E experts $f_1, f_2, ..., f_E$. Given the input sequence tokens $\boldsymbol{O} \in \mathbb{R}^{m \times c}$ of a certain Transformer block, where m is the number of tokens and c is their feature dimension, *Omni MoE* works as follows:

- *MOE.* Basically, given the input $\boldsymbol{O}$, *Omni MoE* actively engages all the E experts to generate the output $\boldsymbol{P}$:

$$\boldsymbol{P} = \sum\nolimits_{e=1}^{E} f_e(\boldsymbol{O}), \tag{1}$$

  where the weights of the E experts $f_1, f_2, ..., f_E$ are not shared. Note that each expert needs to process inputs from all the subjects, hence the inter-subject commonalities are explored.
- *Subject-Specific Parameters.* In order to further capture individual variants, each expert f_e is associated with a c-dimension vector of parameters for each individual subject s. The combined parameters are called subject-specific parameters, denoted as $\{\boldsymbol{\alpha}^s \in \mathbb{R}^{c \times E}\}_{1:S}$, that are only trained with the data of individual subjects, *e.g.*, α^1 is only optimized by the gradient collected from the fMRI data of subject 1. Given $\boldsymbol{O}$ and subject-specific parameters $\{\boldsymbol{\alpha}^s\}_{1:S}$, the subject-specific features are obtained as $\boldsymbol{O} \cdot \boldsymbol{\alpha}^s \in \mathbb{R}^{m \times E}$. As such, despite letting every expert explore cross-subject patterns from amalgamated fMRI data (Eq. 1), these subject-specific parameters enable experts to address the unique aspects of different subjects.

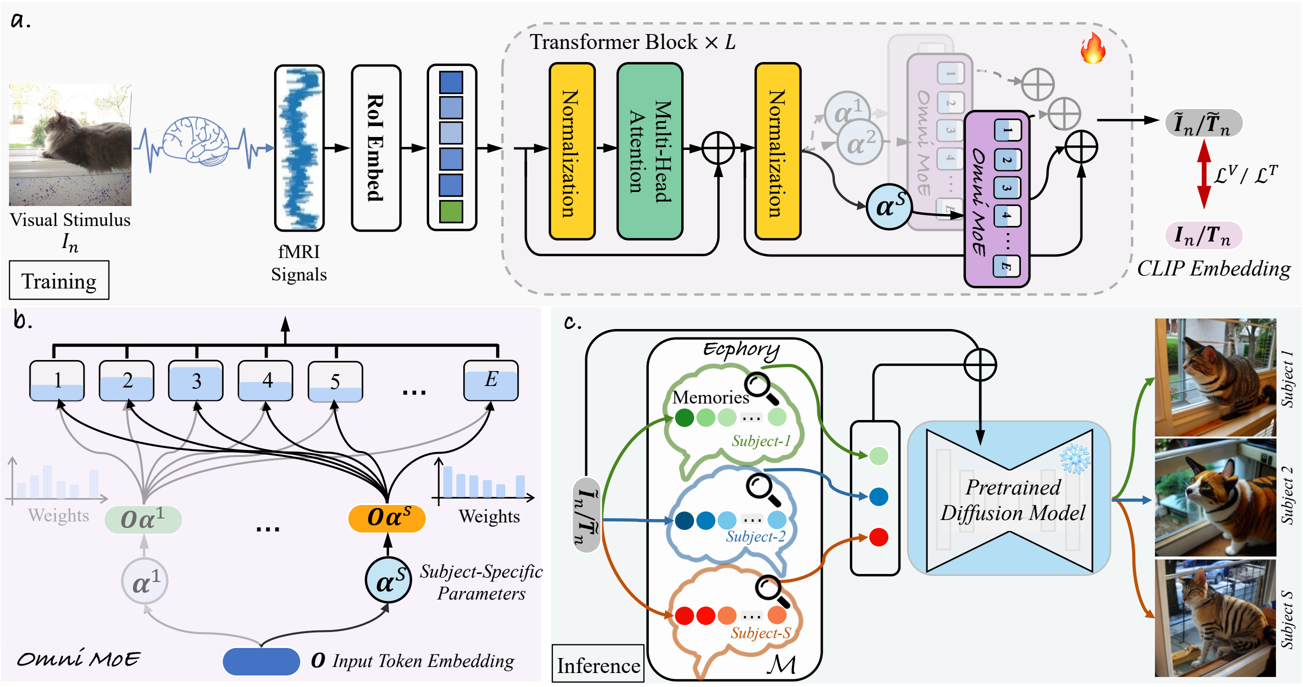

Figure 2. (a) Illustration of the *Psychometry* framework (§ 3.3). (b) *Omni MoE* engages all experts with subject-specific parameters to work together to capture the inter-subject commonality and individual specificity. The detailed illustration of the "split-then-lump" mechanism are presented in Eq. 2-Eq. 5. (c) *Ecphory* enhances the predicted fMRI embedding by incorporating the retrieved most pertinent "memories", serving as more dependable conditional signals to a pre-trained diffusion model. The reconstruction results for different subjects should align as closely as possible with the visual stimulus, while the inconsistency among the results of different subjects is caused by the individual specificity of each subject's fMRI data. Please refer to § 3 for more details.

- *Split-then-Lump: Split.* Instead of directly processing the original input with a large feature map through every expert, we adopt a *split-then-lump* mechanism in order to maintain computational efficiency. First, the splitting weights are computed $\boldsymbol{\omega}$ by applying a `softmax` function on all the m input tokens of the subject-specific features $\boldsymbol{O}\cdot\boldsymbol{\alpha}^s$. The splitting weights refer to the specific weights associated with each token, computed as:

$$\boldsymbol{\omega}^s_{je} = \frac{\exp\big((\boldsymbol{O}\cdot\boldsymbol{\alpha}^s)_{je}\big)}{\sum_{j'=1}^{m}\exp\big((\boldsymbol{O}\cdot\boldsymbol{\alpha}^s)_{j'e}\big)} \in \mathbb{R}^{m\times E}. \tag{2}$$

  This allows the input $\boldsymbol{O}$ to be compressed into token-wise feature $\boldsymbol{\omega}^{s\top}\boldsymbol{O} \in \mathbb{R}^{E\times c}$, suggesting all E experts collectively handle m tokens. Next, we assign the corresponding expert to tackle the splitted feature and compute the output $\boldsymbol{Q}^s$ as a convex combination of all m input tokens.

$$\boldsymbol{Q}^s = \sum\nolimits_{e=1}^{E} f_e(\boldsymbol{\omega}^{s\top}\boldsymbol{O}) \in \mathbb{R}^{E\times c}. \tag{3}$$

- *Split-then-Lump: Lump.* Then, the lumping weights $\mathbf{C}^s$ denote the results of applying a `softmax` function over the E experts. The lumping weights suggest the importance of different experts when lumping the features, formulated as:

$$\mathbf{C}^s_{je} = \frac{\exp\big((\boldsymbol{O}\cdot\boldsymbol{\alpha}^s)_{je}\big)}{\sum_{e'=1}^{E}\exp\big((\boldsymbol{O}\cdot\boldsymbol{\alpha}^s)_{je'}\big)} \in \mathbb{R}^{m\times E}. \tag{4}$$

  Finally, the output sequence tokens $\boldsymbol{P}^s$ for subject s is derived as a convex combination from all the E experts, utilizing the computed lumping weights:

$$\boldsymbol{P}^s = \mathbf{C}^s\boldsymbol{Q}^s \in \mathbb{R}^{m\times c}. \tag{5}$$

By utilizing the *split-then-lump* mechanism, *Omni MoE* layer enjoys an efficient approach via separate learning of tokens and dimensions. Specifically, *split-then-lump* distributes m tokens to E experts (Eq. 2 & Eq. 3), where $E \ll m$. Then, a comprehensive feature $\boldsymbol{P}^s \in \mathbb{R}^{m\times c}$ is derived from the compressed one $\boldsymbol{Q}^s \in \mathbb{R}^{E\times c}$ through the lumping operation (Eq. 4 & Eq. 5).

Discussion. Existing MoEs used in multi-task learning [10, 14] typically employs a routing network $\mathcal{R}$ to determine task routings via sparsely activating top-K experts [53] with the largest scores. Although these sparse MoEs can offer substantial computational savings, the discrete procedure may introduce biases in activating specific experts based on task-specific attributes, which contradicts our objective of capturing both commonalities and differences in fMRI data from different subjects. To address this challenge, *Omni MoE* layer engages all E experts to actively participate in the process of fMRI representation learning, *i.e.*, each expert f_e in MOE (Eq. 1) is required to process the fMRI data from all the subjects to capture inter-subject commonalities. Concurrently, their associated subject-specific parameters

$\{\boldsymbol{\alpha}^s\}_{1:S}$ cope with the individual variabilities by only being trained using the data from individual subjects. Note that traditional dense MoE methods also leverage all E experts for learning, but they suffer from intensive computational costs since each expert in dense MoE is responsible for processing every input, along with the burden of handling the extensive parameter size in the router. Quantitative analyses are later provided in §4.3.

3.2. Ecphory for Test-Time Reconstruction

In neurobiological research area [17, 60], Ecphory is an automatic memory retrieval process activated when a specific cue interacts with stored information gathered from training data, bringing forth recollections of past events [65]. Drawing inspiration from this concept, we integrate *Psychometry* with a retrieval-enhanced inference strategy named *Ecphory*, where the predicted fMRI embedding serves as the specific cue to interact with the prestored information. Specifically, this strategy is tailored to retrieve the most relevant CLIP image or text embedding used as reliable information to enhance the predicted fMRI representation $\tilde{\boldsymbol{I}}$ and $\tilde{\boldsymbol{T}}$ rather than directly using them as the conditional signals, *i.e.*, $\boldsymbol{X}_{s,n} \rightarrow \boldsymbol{I}_n$, $\boldsymbol{X}_{s,n} \rightarrow \boldsymbol{T}_n$. This approach is effective as it allows for obtaining a more reliable fMRI representation through a retrieval method, rather than aiming for an ideal alignment to a specific CLIP embedding, which is more challenging. The enhanced conditional signals are then utilized to guide the generation process of a latent diffusion-based model.

Prestored Subject-Specific Memories. Since each subject retains unique memories, a straightforward approach involves retrieving the relevant embedding from subject-specific memories. Specifically, we prestore the image CLIP embeddings $\boldsymbol{I}$ and text CLIP embeddings $\boldsymbol{T}$ from different subjects during training as the prestored subject-specific memories, *i.e.*, $\mathcal{M}^V = \{\boldsymbol{I}_n\}_{n=1}^N$ and $\mathcal{M}^T = \{\boldsymbol{T}_n\}_{n=1}^N$. Note that the memories are directly derived from the respective training data and stored before inference.

***Ecphory* Mechanism.** With the prestored subject-specific memories, the predicted fMRI embeddings $\tilde{\boldsymbol{I}}_n$ and $\tilde{\boldsymbol{T}}_n$ act as the specific cues to activate the memory retrieval process. In practice, they are used as queries to retrieve the most relevant CLIP embedding from the subject-specific memories based on their similarities. These similarities, denoted as, $sim(\tilde{\boldsymbol{I}}_n, \boldsymbol{m}^V), \forall \boldsymbol{m}^V \in \mathcal{M}^V$ and $sim(\tilde{\boldsymbol{T}}_n, \boldsymbol{m}^T), \forall \boldsymbol{m}^T \in \mathcal{M}^T$, are computed via a cosine similarity function. In order to utilize the retrieved embedding (denoted as $\boldsymbol{F}_n$), we employ a mixed-up approach to enrich the fMRI embedding by blending the retrieved embedding, *i.e.*, $\alpha \cdot \tilde{\boldsymbol{I}}_n + (1-\alpha) \cdot \boldsymbol{F}_n$, where α is a hyperparameter. Consequently, the mixed-up embeddings act as conditional signals to steer the reconstruction process of the pretrained Versatile Diffusion. Given that the diffusion model was initially trained with the CLIP embeddings, our retrieved CLIP embedding could offer more dependable information to the learned fMRI embedding. This process is akin to the ecphory psychology process, where the retrieved "memory" aims to evoke recollections of viewed images.

3.3. Detailed Network Architecture

We adopt the Vision Transformer architecture [12] as our backbone, in which *Omni MoE* layer is inserted into the transformer block. As in [30, 70], lower layers in deep neural networks tend to learn more generic information than higher layers, we can reduce computational overhead by applying the *Omni MoE* layer solely to the higher layers. In our experiments, we incorporate the *Omni MoE* layer into the last four out of the twelve transformer blocks.

Contrastive Learning. In practice, *Psychometry* model is trained via treating fMRI as an additional modality, aiming to pull the fMRI embeddings closer to the CLIP space. Given image and text embeddings $\boldsymbol{I}_n$ and $\boldsymbol{T}_n$ extracted by CLIP, the training objective is to minimize the embedding distances of $(\tilde{\boldsymbol{I}}_n, \boldsymbol{I}_n)$ and $(\tilde{\boldsymbol{T}}_n, \boldsymbol{T}_n)$. Formally, the employed bidirectional contrastive learning loss is formulated as:

$$\begin{aligned}
\mathcal{L}^T &= -\log \frac{\exp\left(\tilde{\boldsymbol{T}}_n^\top \boldsymbol{T}_n / \tau\right)}{\sum_{j=0}^{J} \exp\left(\tilde{\boldsymbol{T}}_n^\top \boldsymbol{T}_j / \tau\right)} - \log \frac{\exp\left(\tilde{\boldsymbol{T}}_n^\top \boldsymbol{T}_n / \tau\right)}{\sum_{j=0}^{J} \exp\left(\tilde{\boldsymbol{T}}_j^\top \boldsymbol{T}_n / \tau\right)}, \\
\mathcal{L}^V &= -\log \frac{\exp\left(\tilde{\boldsymbol{I}}_n^\top \boldsymbol{I}_n / \tau\right)}{\sum_{j=0}^{J} \exp\left(\tilde{\boldsymbol{I}}_n^\top \boldsymbol{I}_j / \tau\right)} - \log \frac{\exp\left(\tilde{\boldsymbol{I}}_n^\top \boldsymbol{I}_n / \tau\right)}{\sum_{j=0}^{J} \exp\left(\tilde{\boldsymbol{I}}_j^\top \boldsymbol{I}_n / \tau\right)},
\end{aligned} \tag{6}$$

where τ is a temperature hyperparameter. The sum for each term is over one positive and J negative samples. The sum over samples of the batch size is omitted for brevity.

4. Experiments

4.1. Experimental Setup

Datasets. Natural Scenes Dataset (NSD) [2] comprises fMRI data collected from 8 participants who viewed a total of 73,000 RGB images. This dataset has been widely utilized [9, 21, 32, 52, 62] to reconstruct perceived images from fMRI. Following the standard setting, we use the data from subjects 1, 2, 5, and 7, who completed all the designed trials, *i.e.*, viewed 10,000 natural scene images and repeated 3 times. We train and evaluate our method using the exact same data split as previous studies. Specifically, the `train` set for each subject contains 8,859 image stimuli and 24,980 fMRI trials. The `test` set includes 982 image stimuli and 2,770 fMRI trials. All images and captions are sourced from MS-COCO database [33]. Different from previous methods which separately train the network for each subject, our proposed method jointly learns the training set for all subjects.

Evaluation Metrics. Both qualitative and quantitative evaluations are conducted in our experiments. For qualitative

	Methods	Low-Level				High-Level			
		PixCorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	EffNet-B↓	SwAV↓
SSM	Mind-Reader [32] [NeurIPS2022]	–	–	–	–	78.2%	–	–	–
	Mind-Vis [9] [CVPR2023]	.080	.220	72.1%	83.2%	78.8%	76.2%	.854	.491
	Takagi *et al.* [62] [CVPR2023]	–	–	83.0%	83.0%	76.0%	77.0%	–	–
	Gu *et al.* [21] [MIDL2023]	.150	.325	–	–	–	–	.862	.465
	MindEye [52] [NeurIPS2023]	**.309**	.323	94.7%	97.8%	93.8%	94.1%	.645	.367
UM	Mind-Reader [32] [NeurIPS2022]	–	–	–	–	66.5%	–	–	–
	Mind-Vis [9] [CVPR2023]	.067	.196	67.7%	74.2%	67.9%	69.3%	.898	.513
	Takagi *et al.* [62] [CVPR2023]	–	–	74.0%	75.1%	67.3%	69.0%	–	–
	Gu *et al.* [21] [MIDL2023]	.103	.264	–	–	–	–	.892	.508
	MindEye [52] [NeurIPS2023]	.129	.255	84.2%	89.2%	84.1%	85.0%	.812	.487
	PSYCHOMETRY (ours)	.297	**.340**	**96.4%**	**98.6%**	**95.8%**	**96.8%**	**.628**	**.345**

Table 1. Quantitative comparison results (§ 4.2) on NSD [2] `test`. *UM* denotes a unified model trained on the amalgamated fMRI data from all subjects, while *SSM* indicates that subject-specific models are trained on subjects' respective data.

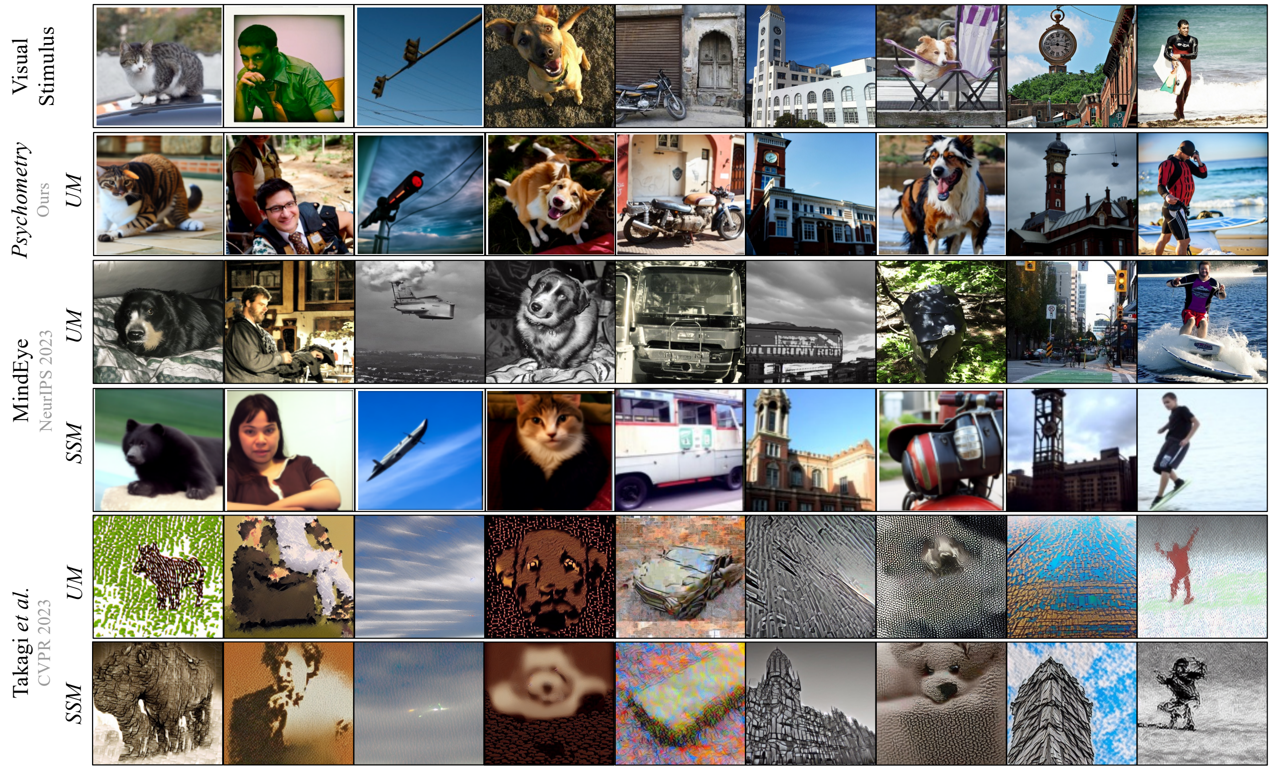

Figure 3. Visual comparison on NSD `test`. *Psychometry* trains only one unified model (*UM*) for once on the amalgamated fMRI data but generates more accurate reconstructions, even compared to two recent methods [52, 62] that train subject-specific models (*SSM*) on their respective data. See § 4.2 for more detailed discussion.

evaluation, we visually compare our reconstructed images with the ground truth images and the results of the state-of-the-art methods in Figure 3. For quantitative evaluation, we employ eight metrics for high-level and low-level evaluation following established research [21, 52, 62]. Specifically, high-level metrics consist of the latent distance of EffNet-B [63] and SwAV [6], which quantifies the similarity between artificial neural networks and the brain's mechanisms for core object recognition. Low-level metrics include the classical Structural Similarity (SSIM) and pixel-wise correlation (PixCorr).

Reproducibility. Our model is implemented in PyTorch and trained on one NVIDIA RTX A6000 GPU with a 48GB memory. Testing is conducted on the same machine.

4.2. Comparison to State-of-the-Arts

Quantitative Results. We compare *Psychometry* with five state-of-the-art methods, namely Mind-Reader [32], Mind-Vis [9], Takagi *et al*. [62], Gu *et al*. [21], and MindEye [52]. As shown in Table 1, *Psychometry* trained on the amalgamated data from all subjects demonstrates competitive results compared to all other baselines. In particular, we ob-

#	*Omni MoE*	*Subject-Specific Parameters*	*Ecphory*	Low-Level				High-Level			
				PixCorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	EffNet-B↓	SwAV↓
1				.163	.238	74.7%	85.2%	80.8%	81.6%	.856	.471
2	✔		✔	.237	.287	89.5%	90.9%	86.7%	87.2%	.794	.423
3	✔	✔		.279	.317	94.9%	97.0%	93.7%	94.8%	.647	.365
4	✔	✔	✔	**.297**	**.340**	**96.4%**	**98.6%**	**95.8%**	**96.8%**	**.628**	**.345**

Table 2. Ablation study on NSD [2] test. See related analysis in §4.3.

serve that existing methods show a noticeable performance decrease when their models are trained on the amalgamated data from all subjects. For instance, MindEye [52] shows a significant decrease **58.3**%/**21.1**%/**10.5**%/**8.6**% in all low-level metrics, suggesting that these methods severely suffer from the individual specificities across subjects. However, our method earns **12.2**%, **9.4**%, **11.7**%, and **11.8**% performance gains over MindEye [52], which is current state-of-the-art, in terms of AlexNet(2), AlexNet(5), Inception, and CLIP respectively. Compared to Takagi [62], our method significantly lifts the scores by **28.5**% and **27.8**% on two high-level metrics. Note that the results of "*SSM*" in Table 1 are averaged from four subject-specific models, each of which is trained on the subject's respective data. As demonstrated, *Psychometry* can still provide notable performance gains when compared to these methods and sets a new state-of-the-art. For instance, *Psychometry* promotes MindEye [52] by **2.0**%/**2.7**%/**2.6**%/**6.0**% and Mind-Vis [9] by **17.0**%/**20.6**%/**26.5**%/**29.7**% over the four high-level metrics. These improvements are particularly impressive considering that our method only has to train a single model once on the amalgamated data.

Qualitative Results. As depicted in Figure 3, the qualitative results are consistent with the numerical findings, demonstrating that our approach produces superior quality and more realistic reconstructions compared to the other methods. In particular, current state-of-the-art, *i.e.*, MindEye [52], suffers from a noticeable performance decrease when its models are trained on the amalgamated data from all subjects. This decrease in performance is evident in the reconstructed images, *e.g.*, when it reconstructs a truck while the visual stimulus was a motorbike. In contrast, the reconstructed images generated by *Psychometry* maintain a high level of consistency with the visual stimuli in terms of semantics, appearance, and structure. This indicates that *Psychometry* can effectively capture inter-subject commonality and individual specificity across subjects, resulting in high-quality image reconstructions from fMRI data.

4.3. Diagnostic Experiment

To thoroughly demonstrate how each component in *Psychometry* contributes to the performance, a series of ablation experiments are conducted on NSD test set. All variants are based on ViT [12] backbone ('#1' in Table 2).

Omni MoE Layer. We first investigate the effectiveness of the Omni MoE layer which consists of subject-specific parameters and a *split-then-lump* mechanism. As shown in Table 2, the Omni MoE layer bings noticeable performance boost (*e.g.*, 0.163/0.238/74.7%/85.2% → 0.279/0.317/94.9%/97.0% on all low-level metrics). This suggests that a single shared MLP layer in the baseline backbone is far from enough to capture the inter-subject commonality and tackle the individual specificity, and proves the effectiveness of our Omni MoE layer. In addition, we derive two more variants that replace the omni MoE layer with a sparse MoE (*i.e.*, top$K$ [53]) and a classical dense MoE. The comparison results in Figure 4 suggest that, although these two variants also boost the performance over the baseline ('#1' in Table 2), Omni MoE layer outperforms them obviously.

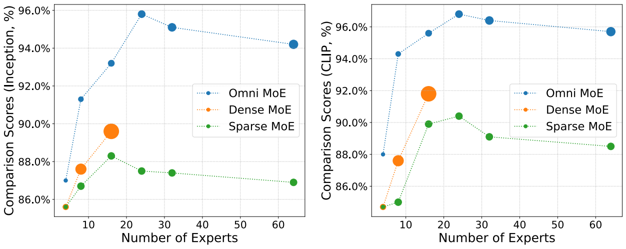

Figure 4. The comparison scores (Inception and CLIP) and the model parameters vary as the number of experts increases. The size of the marker depends on the model size. See §4.3 for details.

Subject-Specific Parameters. Table 2 also investigates the impact of the subject-specific parameters in the Omni MoE layer. When these parameters are not used (labeled as '#2'), each Omni MoE layer adopts shared parameters across subjects, without considering how to tackle the individual differences across subjects. Doing so leads to worse performance, *i.e.*, 96.8%→87.2% over the CLIP scores.

Computational Efficiency. We further investigate the efficiency gains facilitated by our *split-then-lump* mechanism (Eq. 2 - Eq. 5). As depicted in Figure 4, the *Omni MoE* exhibits cost-effective computational overhead, despite involving all its experts in the fMRI representation learning process. This results in a substantial reduction in model size when compared to the variant, dense MoE (Eq. 1), even when comprising the same number of experts. Furthermore, *Psychometry* equipped with *Omni MoE* demonstrates comparable parameters while achieving superior performance, even when compared to the variant using sparse MoE.

Ecphory Inference Strategy. We then proceeded to as-

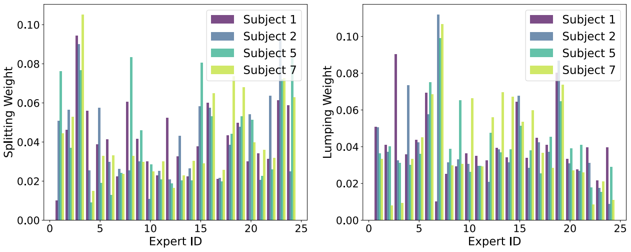

Figure 5. Splitting weights (Eq. 2) and lumping weights (Eq. 4) across experts for all four subjects. See related analysis in §4.3.

sess the effectiveness of our Ecphory inference strategy. The corresponding results are summarized in Table 2. In the absence of this strategy (labeled as '#3'), we directly use the predicted embeddings, *i.e.*, $\tilde{\boldsymbol{I}}_n$ and $\tilde{\boldsymbol{T}}_n$, as the conditional signals for the pretrained diffusion model. Consequently, the performance drops $\mathbf{0.018}/\mathbf{0.023}/\mathbf{1.5}\%/\mathbf{1.6}\%$ and $\mathbf{2.1}\%/\mathbf{2.0}\%/\mathbf{0.019}/\mathbf{0.02}$ across all low-level and high-level metrics, respectively. This evidences that leveraging such a retrieval-enhanced strategy during inference leads to more reliable condition signals and supports our motivation that directly mapping fMRI embeddings to the CLIP image or text embeddings falls short.

Splitting and Lumping Weights. We visualize the splitting weights and lumping weights by summing them across the token dimension, presented in Figure 5. We observe that some experts in our *Omni MoE* layer have high weights for all subjects, *e.g.*, 3rd and 16th expert in splitting weights, while others vary significantly, providing valuable insights into the model's behavior. This suggests that certain experts are adept at capturing common patterns across all subjects, while others excel at capturing subject-specific nuances. This aligns with the design of our model, where subject-specific parameters enable experts to focus on individual specificity, while the collaborative nature of the Omni MoE layer facilitates the capture of inter-subject commonalities. This balance between commonality and specificity is crucial for the model to effectively learn and generalize from the fMRI data across different subjects.

Number of Experts. As the number of experts increases, the computational cost of the model also rises. We conduct experiments by increasing the number of experts in all three variants and training these models for the same duration to determine the best-performing model. As depicted in Figure 4, we discontinue the use of $E > 16$ for Dense MoE (Eq. 1) due to memory constraints exceeding the computational limits of our hardware. Sparse MoE does not yield performance gains with the increased number of experts. On the other hand, Omni MoE achieves its peak performance when $E = 24$. However, increasing E above 24 provides marginal or even negative gain. This may be because too many experts would find some insignificant patterns that are trivial or harmful to decision-making. Therefore, we use $E = 24$ in all other experiments.

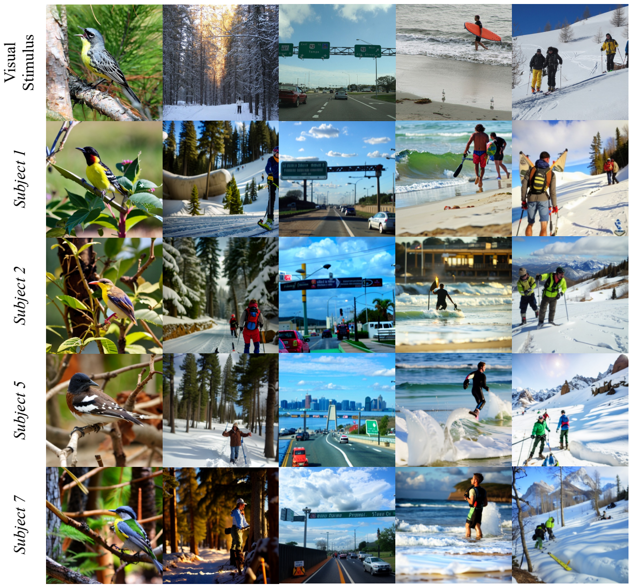

Figure 6. Reconstruction results of *Psychometry* for different subjects with the same visual stimuli. See related analysis in §4.3.

Inter-Subject Commonality and Individual Specificity. Figure 6 reveals the semantic coherence and visual discrepancies among the reconstruction results of different subjects when exposed to the same visual stimuli. This consistency underscores the proficiency of *Psychometry* in capturing shared patterns across subjects, while its use of a single model to generate subject-specific outcomes further validates its effectiveness in fMRI-based image reconstruction. However, the inconsistencies in the results accentuate the individual specificity of each subject's fMRI data.

5. Conclusion and Discussion

In this paper, we introduce *Psychometry*, an omnifit model for fMRI representation learning which marks a significant departure from previous separate training approaches. By leveraging the powerful concept of MoE in an efficient Omni MoE and introducing *Ecphory*, a retrieval-enhanced inference strategy, *Psychometry* can efficiently and effectively capture inter-subject commonality and individual specificity across subjects, resulting in high-quality and realistic image reconstructions from fMRI data. Moving forward, the development of *Psychometry* presents new challenges, particularly in the area of fMRI data privacy protection when amalgamating fMRI data from various subjects for training. Given the rapid advancements in related techniques, we anticipate a surge of innovation towards addressing this promising direction in the field of image reconstruction from human brain activity.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China (62293554), Natural Science Foundation of Zhejiang Province (DT23F020008), and CCF-Tencent Open Fund.

References

[1] Karim Ahmed, Mohammad Haris Baig, and Lorenzo Torresani. Network of experts for large-scale image categorization. In *ECCV*, 2016. 3

[2] 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, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. 2, 5, 6, 7, 1

[3] Katrin Amunts and Karl Zilles. Architectonic mapping of the human brain beyond brodmann. *Neuron*, 88(6):1086–1107, 2015. 1

[4] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. In *NeurIPS*, 2019. 1

[5] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. 1

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 6

[7] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):49, 2019. 2

[8] Ke Chen, Lei Xu, and Huisheng Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*, 12(9):1229–1252, 1999. 3

[9] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *CVPR*, 2023. 2, 5, 6, 7, 1

[10] Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *CVPR*, 2023. 3, 4

[11] David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fmri)"brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, 19(2):261–270, 2003. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5, 7

[13] Markus Enzweiler and Dariu M Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *IEEE Transactions on Image Processing*, 20(10):2967–2979, 2011. 3

[14] Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, Zhangyang Wang, et al. M^{3}vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. In *NeurIPS*, pages 28441–28457, 2022. 3, 4

[15] Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194, 2010. 1

[16] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022. 3

[17] Paul W Frankland, Sheena A Josselyn, and Stefan Köhler. The neurobiological foundation of memory retrieval. *Nature Neuroscience*, 22(10):1576–1585, 2019. 5

[18] Martin A Frost and Rainer Goebel. Measuring structural–functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. *Neuroimage*, 59(2):1369–1381, 2012. 1

[19] Yusuke Fujiwara, Yoichi Miyawaki, and Yukiyasu Kamitani. Modular encoding and decoding models derived from bayesian canonical correlation analysis. *Neural Computation*, 25(4):979–1005, 2013. 2

[20] Evan M Gordon, Timothy O Laumann, Babatunde Adeyemo, Adrian W Gilmore, Steven M Nelson, Nico UF Dosenbach, and Steven E Petersen. Individual-specific features of brain systems identified with resting state functional correlations. *Neuroimage*, 146:918–939, 2017. 1

[21] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Decoding natural image stimuli from fmri data with a surface-based convolutional network. In *MIDL*, 2023. 2, 5, 6, 1

[22] Srinivas Gutta, Jeffrey RJ Huang, P Jonathon, and Harry Wechsler. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE Transactions on Neural Networks*, 11(4):948–960, 2000. 3

[23] Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In *NeurIPS*, 2021. 3

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2

[25] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037, 2017. 2

[26] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. 2, 3

[27] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685, 2005. 1

[28] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. 2

[29] Kenneth K Kwong, John W Belliveau, David A Chesler, Inna E Goldberg, Robert M Weisskoff, Brigitte P Poncelet,

David N Kennedy, Bernice E Hoppel, Mark S Cohen, and Robert Turner. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679, 1992. 1

[30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 5

[31] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting. In *CVPR*, 2023. 2

[32] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. In *NeurIPS*, 2022. 2, 5, 6

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5, 1

[34] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *ICCV*, 2023. 2

[35] Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding from fmri. *arXiv preprint arXiv:2302.12971*, 2023. 3

[36] 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 *ACM MM*, 2023. 3

[37] Yu Lu, Ruijie Quan, Linchao Zhu, and Yi Yang. Zero-shot video grounding with pseudo query lookup and verification. *IEEE Transactions on Image Processing*, 33:1643–1654, 2024. 2

[38] Fan Ma, Xiaojie Jin, Heng Wang, Jingjia Huang, Linchao Zhu, Jiashi Feng, and Yi Yang. Temporal perceiving video-language pre-training. In *AAAI*, 2023. 2

[39] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *ACM SIGKDD*, pages 1930–1939, 2018. 3

[40] Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *IJCNN*, pages 1–8, 2020. 2

[41] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009. 2

[42] Alfonso Nieto-Castañón and Evelina Fedorenko. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage*, 63 (3):1646–1669, 2012. 1

[43] Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. *arXiv preprint arXiv:2303.05334*, 2023. 2

[44] Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *IJCNN*, pages 1–8, 2022. 2

[45] Nikhil Parthasarathy, Eleanor Batty, William Falcon, Thomas Rutten, Mohit Rajpal, EJ Chichilnisky, and Liam Paninski. Neural networks for efficient bayesian decoding of natural images from retinal neurons. In *NeurIPS*, 2017. 1

[46] Xuelin Qian, Yikai Wang, Yanwei Fu, Xinwei Sun, Jianfeng Feng, and Xiangyang Xue. Joint fmri decoding and encoding with latent embedding alignment. *arXiv preprint arXiv:2303.14730*, 2023. 1

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[48] Zarina Rakhimberdina, Quentin Jodelet, Xin Liu, and Tsuyoshi Murata. Natural image reconstruction from fmri using deep learning: A survey. *Frontiers in Neuroscience*, 15:795488, 2021. 2

[49] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *NeurIPS*, 2021. 3

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[51] Joseph J Salvo, Ania M Holubecki, and Rodrigo M Braga. Correspondence between functional connectivity and task-related activity patterns within the individual. *Current Opinion in Behavioral Sciences*, 40:178–188, 2021. 1

[52] Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *NeurIPS*, 2023. 1, 2, 5, 6, 7

[53] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. 2, 3, 4, 7

[54] Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, 13:21, 2019. 1

[55] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1): e1006633, 2019. 1

[56] Xiaolong Shen, Jianxin Ma, Chang Zhou, and Zongxin Yang. Controllable 3d face generation with conditional style code diffusion. In *AAAI*, 2024. 2

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015.

[58] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 32, 2019.

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2020. 2

[60] Sarah Steinvorth, Suzanne Corkin, and Eric Halgren. Ecphory of autobiographical memories: an fmri study of recent and remote memory retrieval. *Neuroimage*, 30(1):285–298, 2006. 5

[61] Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *CVPR*, 2024. 2

[62] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, 2023. 2, 5, 6, 7, 1

[63] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 6

[64] Luke Taylor and Geoff Nitschke. Improving deep learning with generic data augmentation. In *2018 IEEE symposium series on computational intelligence (SSCI)*, pages 1542–1547, 2018. 2

[65] Endel Tulving. Ecphoric processes in episodic memory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 302(1110):361–371, 1983. 5

[66] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013. 2

[67] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *CVPR*, 2023. 1

[68] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models. *arXiv preprint arXiv:2401.08392*, 2024. 2

[69] Hanrong Ye and Dan Xu. Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In *ICCV*, 2023. 3

[70] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014. 5

[71] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *CVPR*, 2024. 2