Mind Artist: Creating Artistic Snapshots with Human Thought

Jiaxuan Chen[1, 3] Yu Qi[2, 3, 1 †] Yueming Wang[4, 2] Gang Pan[3, 1, 2]
[1]College of Computer Science and Technology, Zhejiang University
[2]MOE Frontier Science Center for Brain Science and Brain-Machine Integration, Zhejiang University
[3]The State Key Lab of Brain-Machine Intelligence, Zhejiang University
[4]Qiushi Academy for Advanced Studies, Zhejiang University
{jiaxuan_chen, qiyu, ymingwang, gpan}@zju.edu.cn

Figure 1. Appreciating snapshot views of our mind in the style of your liking. Left: The proposed neural decoder Mind Artist supports not only the reconstruction of visual stimuli from fMRI signals but also acts as an *"artistic mind camera" without extra optimizations and image-to-image translation operations*. Right: We show the decoding results of our MindArt in oil, cartoon, Picasso and ink wash styles.

Abstract

We introduce Mind Artist (MindArt), a novel and efficient neural decoding architecture to snap artistic photographs from our mind in a controllable manner. Recently, progress has been made in image reconstruction with non-invasive brain recordings, but it's still difficult to generate realistic images with high semantic fidelity due to the scarcity of data annotations. Unlike previous methods, this work casts the neural decoding into optimal transport (OT) and representation decoupling problems. Specifically, under discrete OT theory, we design a graph matching-guided neural representation learning framework to seek the underlying correspondences between conceptual semantics and neural signals, which yields a natural and meaningful self-supervisory task. Moreover, the proposed MindArt, structured with multiple stand-alone modal branches, enables the seamless incorporation of semantic representation into any visual style information, thus leaving it to have multi-modal reconstruction and training-free semantic editing capabilities. By doing so, the reconstructed images of MindArt have phenomenal realism both in terms of semantics and appearance. We compare our MindArt with leading alternatives, and achieve SOTA performance in different decoding tasks. Importantly, our approach can directly generate a series of stylized "mind snapshots" w/o extra optimizations, which may open up more potential applications. Code is available at https://github.com/JxuanC/MindArt.

1. Introduction

As we venture into the frontiers of human creativity, one captivating question has surfaced: *Can our brains serve as hidden "artistic cameras"?* Let's envision a world where painting tools, and photographic skills are going out of fashion. All you need is to think or gaze upon a visual object or immerse yourself in the wonders of nature. Then, your brain will be turned into an artistic lens transforming your perceptions into astonishing works of art, perhaps resembling a Van Gogh-style snapshot with vibrant yellow and orange sunflowers. To actualize aforesaid goals, we present

†Corresponding author: Yu Qi.

a new way of thinking about the neural decoding problems.

Humans are adept at abstracting rich sensory information into simple concepts. For example, the photograph of a dog, the cartoon depiction of Snoopy, or even the word "dog" can activate similar conceptual representations in our mind [21, 38, 51, 64]. That, in turn, demonstrates concrete concepts for different contexts (*i.e.* personal experience) contribute to shaping sensory perception-derived representations [20, 48]. To incorporate this principle into computational models, we present a "1+N" neural decoding paradigm (Fig. 1 Left) to translate your thoughts into individual works of art, where "1" stands for amodal concepts (semantics), while "N" denotes different context conditions acted as prior knowledge like layout or style information.

There are all sorts of practices to evoke conceptual semantics, among them viewing natural images is an ideal choice. Neuroscience [20, 21, 33] has revealed that when people look at objective stimuli, the parts of their brains linked to thoughts and feelings about the knowledge of that object lighting up (Fig. 1 Left). For another, with the help of generative models [25, 36, 58] and self-supervised learning [9, 29, 55], recently, progress has been made in the task of reconstructing visual stimulus from brain recordings, especially using non-invasive neuroimaging methodology such as functional Magnetic Resonance Imaging (fMRI) [10, 60, 63]. However, as a non-invasive tool, fMRI is highly susceptible to various interferences like noise [39], and the acquisition is a cumbersome, costly process [11].

Therefore, it is vital to revisit generalizable semantic correspondences between visual stimuli and brain activities, which is also key to the success of the "1+N" decoding paradigm. Given a random set of neural patterns and images, a biological-valid model should assign the neural patterns to the images that are semantically closest to their true stimuli. Under the circumstances, we introduce a novel "mind reading" technique, termed Mind Artist (MindArt), which boils the neural decoding down to finding the optimal semantic matching between stimuli and brain activities by solving a generalized linear assignment problem. This target can be elegantly formulated as a graph-based optimal transport (OT) framework, aiming to shift the mass from one distribution to another while minimizing cost.

Specifically, we propose a self-supervised graph matching (GM) framework to predict the cost function of OT optimization by explicitly modelling dynamic graphs that describe more flexible relationships between image and fMRI entities. In doing so, the assignment structure of the predictions can steer us toward capturing the high-level semantic representations preserved within neural signals. Previous works [1, 5, 10, 40, 60, 63], by contrast, largely treat fMRI data independently, which neglects the rich connection among fMRI signals. Finally, following the "1+N" decoding paradigm, the MindArt is designed as a multi-branch architecture, which not only supports multi-modal generation, but also allows us to snap a photograph imbued with artistic ambiance from our own thoughts in a training-free prompt fashion. The key here is to incorporate prior knowledge into the GM-guided neural semantic representations by the collaborative use of diffusion models [31, 58] and large language models [3, 53, 54].

In a nutshell, our main contributions are threefold: **(i)** We introduce a novel self-supervision framework within an optimal transport perspective to tackle neural decoding by formalizing neural representation learning as a graph matching problem; **(ii)** We propose Mind Artist, a two-stream neural decoding structure, which not only supports multi-modal reconstruction, but also can be served as an *"artistic mind camera"* to snap stylized views (See Fig.1), without the need for extra optimization or fine-tuning; **(iii)** Compared with previous state-of-the-art methods, the reconstructed images (and the stylized snapshots) of MindArt are more faithful to the stimulus images with better preserved high visual quality, and high-level semantic fidelity.

2. Related Work

There is a large body of works in the literature focused on visual neural decoding problem, which can be broadly broken down into stimuli classification [7, 13, 20, 27, 65, 71], identification [28, 32, 34, 45], and reconstruction [18, 42, 45, 57, 62]. An exhaustive review covering all branches is out of the scope of this paper, thus in this section, we only summarize the pertinent background material of reconstruction tasks that puts our MindArt into context.

Stimuli reconstruction is an exciting yet demanding task, which purposes to directly recover the perceived image from brain recordings. In the early stages, the image reconstruction works [42, 45] commonly leverage simple linear regression models to seek mappings between fMRI and handcrafted image descriptors. Though previous techniques successfully recovered simple low-level detail images, severely relying on manual configuration has limited the applicability, and the decoding results are also blurry. With the advent of deep learning, researchers have opted for leveraging generative models (*e.g.*, generative adversarial networks [4, 25] and diffusion models [31, 58]) or self-supervised learning [29] to address the perceived image reconstruction problem. Authors in [1, 24] introduced a self-supervision visual decoding architecture by stacking the image-to-fMRI encoder and fMRI-to-image decoder back-to-back. The symmetric design enables the model to be trained on more extensive unlabeled image and fMRI datasets. Based on a similar philosophy, [5] rethought self-supervised visual reconstruction problem via some neuroscience principles [15], and proposed a cross-modal inpainting framework, termed VQ-fMRI, to implement visual content completion mechanism, thereby avoiding the decoding

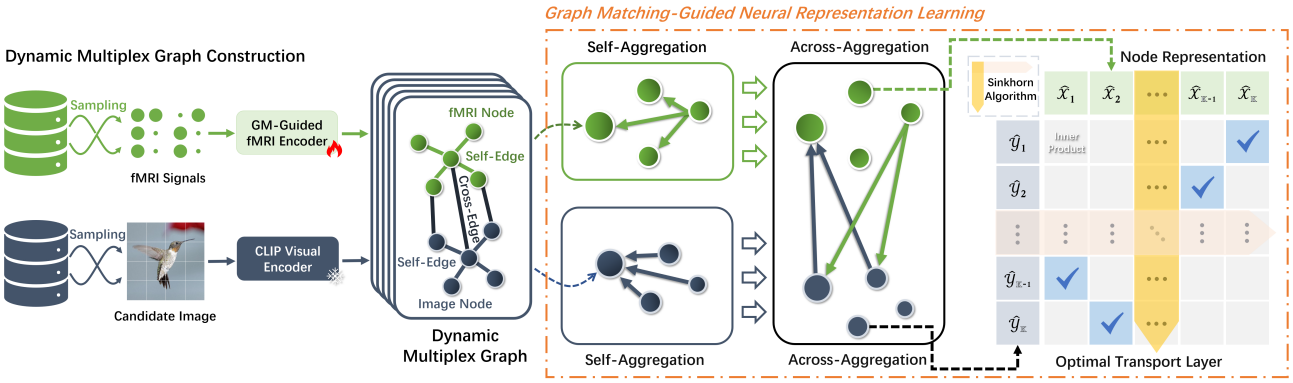

Figure 2. The proposed neural representation learning framework guided by graph matching (GM). This pipeline is made up of three major components: dynamic multiplex graph construction, alternating aggregation of node embeddings, seeking optimal transport (OT) plan.

of imperceptible local details. The decoded results of self-supervised models can match the original stimuli in terms of contours and poses, but lack distinguishable semantic information. To generate more plausible images, [10] explored a masked brain modeling with a latent diffusion model (LDM) [58], named MinD-Vis. A concurrent work [63] also used the LDM as the generative prior but adopted a ridge regression to predict latent representations. Additionally to the methods above, other representative studies include [19, 22, 37, 40, 43, 47, 60, 61]. The generative model-based reconstruction techniques have limitations in low-level detail control like shape and posture, albeit capable of yielding high-fidelity images. From a broader perspective, some efforts have demonstrated that sequential data, *e.g.*, video [11, 68] and language [6, 23, 41, 64], also can be generated from non-invasive brain recordings. What we wish to highlight is that our MindArt takes a step beyond, offering desirable properties: multi-modal reconstruction and training-free style transfer, which can be regarded as a supplement to the existing visual decoding methods.

3. The MindArt Approach

In the following, we detail our MindArt architecture, which contains two stages: a graph matching (GM) guided neural representation learning procedure (See Fig. 2), followed by a two-branch multi-modal decoding structure, as illustrated in Fig. 3. The formulation of the visual decoding problem from a perspective of optimal transport (OT) is first introduced, and then the concrete implementations successively.

3.1. Problem Formulation

Consider an image-fMRI dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i indicates an fMRI signal, and y_i denotes the stimulus image of x_i. For seeking a good neural representation, a prevailing practice is to fit a mapping f_θ^*,

$$s.t.\ f_\theta^* = \arg\min_{f_\theta} \mathbb{E}_{x_i, y_i \sim \mathcal{D}} ||\mathcal{T}(y_i) - f_\theta(x_i)||, \quad (1)$$

where $\mathcal{T}(\cdot)$ is normally a frozen network pre-trained on large image datasets. However, it is challenging to accurately align the potential representations in all dimensions, since the cardinality of $\mathcal{D}$ is relatively small [5]. To conquer this, we attempt to discover biologically valid cross-domain relations within a more principled framework germinating from recent advances in OT [8, 59, 67].

Formally, suppose we have sampled two random sets of entities from fMRI and image domains, defined as $\mathbb{X} = \{x_i\}_{i=1}^{n}$ and $\mathbb{Y} = \{y_i\}_{i=1}^{m}$, where n and m are the number of entities. Let the entity be a sample point, represented by a vector, in the complete separable metric spaces, $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ denote two vector variables, and then their discrete measures are formulated as $\boldsymbol{\mu} = \sum_{i=1}^{n} p_i \delta_{x_i}$ and $\boldsymbol{\nu} = \sum_{i=1}^{m} q_i \delta_{y_i}$, where δ_{x_i} is the Dirac function centered on x_i. The above formulation describes probability measures if the wights $[p_i]_{1:n} \in \Delta_n := \{\mathbf{p} \in \mathbb{R}_+^n : \mathbf{p}^{\mathrm{T}} \mathbb{1}_n\}$ and $[q_i]_{1:m} \in \Delta_m := \{\mathbf{q} \in \mathbb{R}_+^m : \mathbf{q}^{\mathrm{T}} \mathbb{1}_m\}$ belong the n- and m-dimensional simplex, respectively [50], where $\mathbb{1}_n$ is an n-dimensional all-one vector. OT aims to seek the least costly transport plan $\boldsymbol{\Gamma}$ between distribution $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, which can be written as

$$\mathcal{D}_w(\boldsymbol{\mu}, \boldsymbol{\nu}) := \inf_{\gamma \in \Pi} \mathbb{E}_{(\mathbb{X}, \mathbb{Y}) \sim \gamma}[c(\mathbb{X}, \mathbb{Y})] = \min_{\boldsymbol{\Gamma} \in \Pi} \langle \boldsymbol{\Gamma}, \mathbf{C} \rangle, \quad (2)$$

where $\Pi(\mathbf{p}, \mathbf{q}) := \{\boldsymbol{\Gamma} \in \mathbb{R}_+^{n \times m} | \boldsymbol{\Gamma} \mathbb{1}_m = \mathbf{p}, \boldsymbol{\Gamma}^{\mathrm{T}} \mathbb{1}_n = \mathbf{q}\}$ indicates all the joint distributions with marginal $\mathbf{p}$ and $\mathbf{q}$, $\mathbf{C}$ is a cost matrix calculated by the function $c(\cdot, \cdot)$ such as cosine distance, and $\langle \cdot, \cdot \rangle$ denotes Frobenius inner product. Overall, Eq. 2 defines an OT distance (*a.k.a.* Wasserstein distance) that measures the discrepancy across two domains.

By modelling $\mathbb{X}$ and $\mathbb{Y}$ as complete graphs and performing GM, not only intra- and cross-domain relations can be captured via adding undirected edges, but also more flexible supervision signals can be exploited, *i.e.*

$$\mathbf{C}_{ij} = \min\{\mathbf{C}_{i1}, \cdots, \mathbf{C}_{ij}\} \text{ if } y_j = \arg\max_{y \in \mathbb{Y}} d(y, \bar{y}_i), \quad (3)$$

where $\bar{y}_i$ denotes the true visual stimulus of x_i, and $d(\cdot, \cdot)$

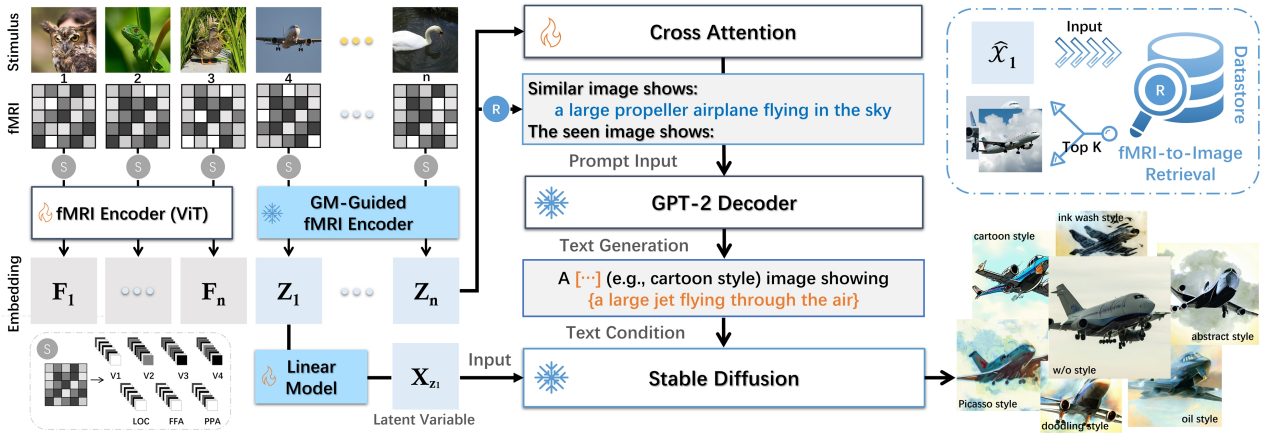

Figure 3. The illustration of the MindArt. In our architecture, the cross-attention is leveraged to bridge a frozen GPT-2 decoder with two fMRI encoders. Note that the fMRI signal is divided into fixed size voxel vectors, which is the only input to the ViT-based encoders.

returns the semantic similarity of the images. Note that there is not a one-to-one correspondence between $\mathbb{X}$ and $\mathbb{Y}$. Intuitively, Eq. 3 explicitly encourages models to find semantically similar nodes, rather than confined to limited ground-truth relations, which yields a natural and significative self-supervised task by representing image-fMRI annotation into dynamically-constructed graphs (See Sec. 3.2).

3.2. GM-Guided Neural Representation Learning

Dynamic Multiplex Graph Construction. To tackle the neural representation learning problem via adopting OT for GM, we start by taking a simple yet effective graph definition step. Given an image-fMRI dataset $\mathcal{D}$, we first randomly sample two subsets of $\mathbb{K}$ entities from the fMRI and image domains, respectively, and construct them into a single complete graph $\mathcal{G}(\mathcal{V}_x, \mathcal{V}_y, \mathcal{E}_{self}, \mathcal{E}_{cross})$, which has two types of initial graph node embeddings computed by a ViT-based fMRI encoder [17], and a CLIP visual encoder [55] (See Fig. 2 Left). The graph $\mathcal{G}$ also has two types of undirected edges (*a.k.a.* multiplex graph [44, 46]): self edges $\mathcal{E}_{self}$ connecting within the unilateral domain (namely, fMRI and image), and cross edges $\mathcal{E}_{cross}$ connecting one domain to another. Unlike raw image-fMRI pairs, various context information exists in the multiplex graphs, and seeking good potential correspondences between $\mathcal{V}_x$ and $\mathcal{V}_y$ is naturally formulated into a GM problem.

Multiplex Graph Neural Network. With the above definition, huge amounts of graph-structured data can be easily generated, and this construction process is of "dynamic" nature due to the update of encoder's parameters during training. To propagate information along both self edges $\mathcal{E}_{self}$ and cross edges $\mathcal{E}_{cross}$, we leverage a multiplex attention graph neural network (MAGNN) [59] for aggregating messages, which starts with the initial node state, and calculates an updated node representation at each layer. Specifically, let $\mathcal{X}_i^{\ell}$ be the intermediate representation at layer ℓ for node i in $\mathcal{V}_x$. The residual message passing scheme is:

$$\mathcal{X}_i^{\ell+1} = \mathcal{X}_i^{\ell} + \mathrm{MLP}\Big(\Big[\mathcal{X}_i^{\ell} || \mathbf{m}_{\mathcal{E}\to i}\Big]\Big), \tag{4}$$

where $[\cdot||\cdot]$ is a concatenation operation, and $\mathbf{m}_{\mathcal{E}\to i}$ denotes the result of the alternate aggregation (*i.e.* $\mathcal{E} = \mathcal{E}_{self}$ if ℓ is odd, otherwise, $\mathcal{E} = \mathcal{E}_{cross}$) with multi-head attention (MHA) [66] from all nodes. The MHA weights enable the network to selectively focus on a subset of $\mathcal{V}_x$ that potentially shares similar neural semantic patterns. An analogous update procedure can be simultaneously performed for $\mathcal{V}_y$.

Optimal Matching. From the OT formulation (Eq. 2), we see that the OT plan $\mathbf{\Gamma}$ (*a.k.a.* coupling matrix) can be inferred by providing a cost matrix $\mathbf{C}$, and then solving a linear assignment problem. Here, we express the pairwise cosine similarity of final node representations as a score matrix to reflect the matching cost:

$$\mathbf{C}_{ij} = 1 - \frac{\hat{\mathcal{X}}_i^{\mathrm{T}} \hat{\mathcal{Y}}_j}{||\hat{\mathcal{X}}_i||_2 ||\hat{\mathcal{Y}}_j||_2}, \quad 1 \le i, j \le \mathbb{K}, \tag{5}$$

where $\hat{\mathcal{X}}_i \in \mathbb{R}^{\cdot \times 1}$ and $\hat{\mathcal{Y}}_j \in \mathbb{R}^{\cdot \times 1}$ are the final representations for node i in $\mathcal{V}_x$ and node j in $\mathcal{V}_y$, respectively. The complexity of minimizing $\langle \mathbf{\Gamma}, \mathbf{C} \rangle$ under the constraint $\mathbf{\Gamma} \in \Pi$ is $\mathcal{O}(N^3 log N)$ [49, 67]. To reduce the complexity, we leverage entropy-regularized OT, which is a strictly convex optimization problem, to find the OT distance, *i.e.*

$$\mathcal{D}_s(\boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{\Gamma}) := \min_{\mathbf{\Gamma} \in \Pi} \langle \mathbf{\Gamma}, \mathbf{C} \rangle - \varepsilon H(\mathbf{\Gamma}), \tag{6}$$

where $\varepsilon > 0$ is a trade-off parameter (by making ε higher, $\mathbf{\Gamma}$ will be smoother), and $H(\mathbf{\Gamma}) = -\sum_{ij} \mathbf{\Gamma}_{ij} log \mathbf{\Gamma}_{ij}$ is an entropy constriaint term. Under this entropic regularization, the solution [12] reads $\mathbf{\Gamma}^* = \mathrm{diag}(\mathbf{a}) \mathbf{K} \mathrm{diag}(\mathbf{b})$, where $\mathbf{K} = e^{-(1/\varepsilon)\mathbf{C}} \in \mathbb{R}_+^{n \times m}$ is a Gibbs kernel, $\mathbf{a} \in \mathbb{R}_+^n$

and $\mathbf{b} \in \mathbb{R}_+^m$ can be computed by Sinkhorn iterations [12]. Since the iterations are differentiable, it is straightforward to backpropagate within deep learning models.

As mentioned in Sec.3.1, rather than seeking a precise correspondence between fMRI and raw stimulus, we pursue an assignment structure that allocates neural patterns to images that share visual or semantic similarities with the ground truth stimuli. For this purpose, a reasonable practice is to consider images of the same category as potential matching objects. This also lets us for data augmentation via adding candidate images from ImageNet [14]. For the fMRI domain, we generate virtual neural signals by performing linear interpolation on the fMRIs that evoked by the identical class of images, which shares similarities with mixup [72], but the difference is our labels are from a random candidate set. In addition, to handle unmatched cases (namely, *w/o* identical class), we adopt a dustbin technique widely employed in GM [16, 59], which allows us to explicitly assign unmatched entities (nodes) to the extra bin by augmenting the cost matrix $\mathbf{C} \in \mathbb{R}^{n\times m}$ to $\hat{\mathbf{C}} \in \mathbb{R}^{(n+1)\times(m+1)}$, filled with learnable parameters. Finally, given potential correspondence labels $\mathcal{M}$ and unmatched indexes $\mathcal{R}$, our optimization goal is to minimize the negative log-likelihood of the coupling matrix $\mathbf{\Gamma}^*$:

$$\mathcal{L}_{graph} = -\sum_{(i,j)\in\mathcal{M}} \log \mathbf{\Gamma}^*_{i,j} - \sum_{(i,j)\in\mathcal{R}} \log \mathbf{\Gamma}^*_{i,j}. \quad (7)$$

This supervision explicitly encourages models to capture shared high-level semantic embedding across $\mathbb{X}$ and $\mathbb{Y}$.

3.3. Multimodal Reconstruction Architecture

The semantic representation for human brains has its origin in various perceptual-cognitive systems: it is supramodal in nature [2, 21, 51, 64]. Psychological evidence suggests that humans acquire language by grounding meanings to knowledge about the world [26, 52, 73]. That meant verbal language is a proper neurosemantic proxy. Inspired by this, the proposed decoding model MindArt is configured as a two-stream structure (See illustration in Fig. 3). Technically, our MindArt is a "1+2" decoding instance, which incorporates the GM-guided fMRI encoder and two stand-alone vision and language branches, where visual sub-module provides visual generative prior (*e.g.* shape and position), and linguistic branch guides the learned neural representations toward a desired high-level linguistic semantic direction. More importantly, the "divide and conquer" strategy allows us to exactly control semantics by prompt-based text editing, thereby opening a door for adding style information prior in a training-free fashion.

In this work, the visual stream includes a latent diffusion model (LDM) [58] and a linear model, conditioning on the pre-trained fMRI representation z_i. Following [63], we first leverage L2-regularized linear regression to predict a latent embedding of the ground truth stimulus $\bar{y}_i$, where the target variable is the latent representation of $\bar{y}_i$ compressed by the auto-encoder (AE) of LDM, and the weights of linear mapping are estimated from all the training data. Second, the predicted embedding is fed into the decoder of AE to generate an intermediate image $\mathbf{X}_{z_i}$, which is then resized to 512×512. On the other hand, the language stream is built on GPT-2 [54]. To bridge two fMRI encoders and a GPT-2 decoder, we use multi-head cross-attention mechanism, leaving each layer of the GPT decoder attends to the outputs of the fMRI encoders [66]. It is noteworthy that we freeze the GPT-2 decoder and GM-guided fMRI encoder, which is why we need two fMRI encoders. To harness the inherent contextual language capabilities of GPT-2, we leverage fMRI-to-image retrieval to access image captions, which are then used to populate the placeholder * in a prompt template like *"Similar image shows *. The seen image shows"*. In this context, the goal of the stand-alone language stream can be boiled down to an fMRI-to-text translation problem that can be approached by minimizing the cost function:

$$\begin{aligned}\mathcal{L}_{text} = -\sum_{i=1}^{M} \log P\Big(s_i | s_{<i}, [\mathbf{E}_{\mathcal{G}}(x) || \mathbf{E}_{\Phi}(x)]; \Theta\Big) \\ +\lambda \Big|\Big| [\mathbf{E}_{clip}(y)]_0 - [\mathbf{E}_{\Phi}(x)]_0 \Big|\Big|_2^2,\end{aligned} \quad (8)$$

where Θ is the parameters of cross-attention modules, $[s_i]_{1:M}$ is a visual captioning (pseudo-labels) of stimulus $\bar{y}$ generated from [56], $\mathbf{E}_{\mathcal{G}}$, $\mathbf{E}_{\Phi}$ and $\mathbf{E}_{clip}$ denote the frozen GM-guided encoder, the ViT-based encoder with trainable weights Φ, and the frozen CLIP encoder, respectively. The first term is the sum of the negative log-likelihood conditioned on the fMRI embeddings and the previous tokens, the second is a mean-squared loss to constrain $\mathbf{E}_{\Phi}$, and $\lambda = 10$ is a hyper-parameter weighting these items.

To fulfill stylized visual reconstruction, now we only need to integrate the two information flows into the LDM, and provide extra style prompts. This controllable decoding mechanism can be described as:

$$\hat{\mathbf{x}} = \text{StableDiffusion}\Big(\mathbf{X}_z; \text{concat}([o_i]_{1:}, [\hat{s}_i]_{1:M})\Big), \quad (9)$$

where $[\hat{s}_i]_{1:M}$ is the predicted word sequence, and $[o_i]_{1:}$ indicates an extra conditional prompt such as *"A Van Gogh-style painting showing"*, concatenated in front of $[\hat{s}_i]_{1:M}$.

4. Experimental Results

4.1. Implementation Details

The architecture of our MindArt comprises three off-the-shelf sub-models CLIP-B/32, GPT-2_{Base}, and Stable Diffusion (version 1.4), which can be available on HuggingFace [69]. Other configurations of MindArt are summarized as follows. All fMRI encoders within MindArt adopt ViT

Method	DS	Visual Perception ↑			Semantic Perception ↑	
		SSIM	$\text{CLIP}_{\text{Score}}$	FID ↓	CLIP_{T}@10	CLIP_{T}@50
Takagi *et al.* [63]	DIR	$0.152_{\pm 0.11}$	$0.572_{\pm 0.07}$	17.4	$24.6\%_{\pm 3.7\%}$	10.0%
VQ-fMRI [5]	DIR	$\mathbf{0.433}_{\pm 0.12}$	$0.520_{\pm 0.09}$	36.8	$24.3\%_{\pm 4.1\%}$	12.0%
MindGPT [6]	DIR	$0.177_{\pm 0.11}$	$0.575_{\pm 0.12}$	2.70	$31.6\%_{\pm 3.8\%}$	14.0%
Ours	DIR	$0.223_{\pm 0.12}$	$\mathbf{0.622}_{\pm 0.14}$	**2.67**	$\mathbf{40.3\%}_{\pm 3.8\%}$	**26.0%**
Takagi *et al.* [63]	GOD	$0.182_{\pm 0.11}$	$0.602_{\pm 0.08}$	18.5	$29.8\%_{\pm 4.0\%}$	10.0%
MinD-Vis [10]	GOD	$0.251_{\pm 0.15}$	$0.625_{\pm 0.10}$	2.69	$39.7\%_{\pm 5.6\%}$	12.0%
VQ-fMRI [5]	GOD	$\mathbf{0.423}_{\pm 0.11}$	$0.548_{\pm 0.09}$	37.0	$24.8\%_{\pm 4.4\%}$	10.0%
GESS [22]	GOD	$0.267_{\pm 0.15}$	$0.620_{\pm 0.10}$	5.05	$43.1\%_{\pm 5.3\%}$	12.0%
Ours	GOD	$0.242_{\pm 0.13}$	$\mathbf{0.631}_{\pm 0.13}$	**2.55**	$\mathbf{43.6\%}_{\pm 3.9\%}$	**24.0%**

Table 1. Quantitative comparison of six reconstruction methods (↑ denotes the higher the better, and **Bold** is the optimal value).

models with 8-head self-attention. During pre-training, we use an embedding size of 512 and 12 network layers. For the MAGNN, we use 6 layers of alternating 4-head self- and cross-attention with an embedding size of 512, $\mathbb{K} = 1024$ nodes of the dynamic graphs, and perform 100 Sinkhorn iterations. In the fine-tuning phase, the embedding size and layer number of encoder are 768 and 16, respectively. The 12-head cross-attention layer is added to each of the 12 layers of GPT-2 decoder. The MindArt is optimized using Adam solver [35] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 1e-4 and learning rate of 1e-4 until convergence. Moreover, the batch size is 256, and the candidate images are selected from ImageNet including 200 categories totaling 273.4k images. The MindArt is implemented by Pytorch, and trained on 4 NVIDIA GeForce RTX3090 GPUs.

4.2. Dataset and Evaluation Metrics

Two public image-fMRI datasets are used to verify the decoding performance of our MindArt: DIR dataset [62], and GOD dataset [32]. In DIR and GOD datasets, eight subjects were required to see 1250 natural images involving 200 categories, and simultaneously fMRI signals were recorded using a 3.0-Tesla Siemens MAGNETOM Verio scanner. The visual stimuli involved in the image presentation experiments for both DIR and GOD datasets are identical, which are selected from ImageNet [14], where 1200 images across 150 categories are utilized for training sessions, and 50 images from 50 categories for test sessions. Note that the training/test split has no overlapping classes.

To make comprehensive quantitative comparisons with existing approaches, we initially consider visual perception metrics including SSIM that calculating the similarity of local spatial pixels, FID [30] and CLIP visual score ($\text{CLIP}_{\text{score}}$) [55], which reflect the high-level visual similarity between images. Besides, we employ 10- and 50-way CLIP image-text scores (CLIP_{T}@10 and CLIP_{T}@50) to evaluate the semantic fidelity of the reconstructed images. Specifically, we use the names of all the categories in the test stimulus images as the set of potential text, and focus on identifying the class name of the ground truth stimulus image among candidate texts (one being the actual ground truth and the other selected from the remaining test set).

Method	Language Similarity Metrics ↑				
	B@1	B@4	ROUGE	METEOR	SPICE
MindGPT [6]	37.9	15.7	35.9	12.8	10.3
Ours	**38.5**	**16.0**	**37.2**	**13.1**	**11.3**

Table 2. Quantitative comparison on language reconstruction (↑ denotes the higher the better, and **Bold** is the optimal value).

4.3. Visual Reconstruction with Style Control

This section focuses on the performance of our MindArt in perceptual image reconstruction. To intuitively demonstrate the decoding capabilities of MindArt under the various style prompts, we consider 7 prompt-based conditional inputs: default (*w/o* style prompt), cartoon, Picasso, oil, ink wash, abstract, and doodling. Qualitative results of our MindArt (as well as leading techniques including Takagi *et al.* [63], MinD-Vis [10], VQ-fMRI [5], MindGPT [6] and GESS [22]) on one subject are presented in Fig. 4 (See Appendix for full samples). The first column represents the raw visual stimuli, and the rest columns are the reconstruction images from different methods. From the results, we see that our MindArt produces detail-rich images with high-level semantic fidelity. Compared with recently published methods, the proposed MindArt exhibits superior capabilities in recovering shapes, tones, and overall layouts, while maintaining relatively faithful semantic attributes. The stylized snapshots also demonstrate satisfying quality in terms of low-level details and high-level semantic information, which conform with the stimulus image in most cases. The reconstructions from VQ-fMRI [5] excel in pixel-level layout, however, the generated images tend to be blurry.

We also present a comprehensive quantitative evaluation of our MindArt on subject 3 of the DIR and GOD, as detailed in Tab. 1. We outperform the state-of-the-art methods on three criteria. Specifically, concerning CLIP_{T}@10 and CLIP_{T}@50, reflecting the semantic fidelity of reconstructed images, MindArt achieves the highest recognition accuracy in both DIR and GOD datasets, which surpasses that of competitors by a margin ranging from 0.5% to 15.7%. In terms of visual perception criteria, our method also attains an optimal CLIP visual score. Consistent with qualitative findings, the pixel-level reconstruction method VQ-fMRI excel in low-level similarity metric like SSIM.

4.4. Describing What You See

In order to understand the language semantics decoding capacity of the proposed MindArt, we provide some generation examples from subject 3 of the DIR dataset, and compare them with MindGPT [6], as shown in Fig. 5. We observe that MindArt can generate semantically satisfying text

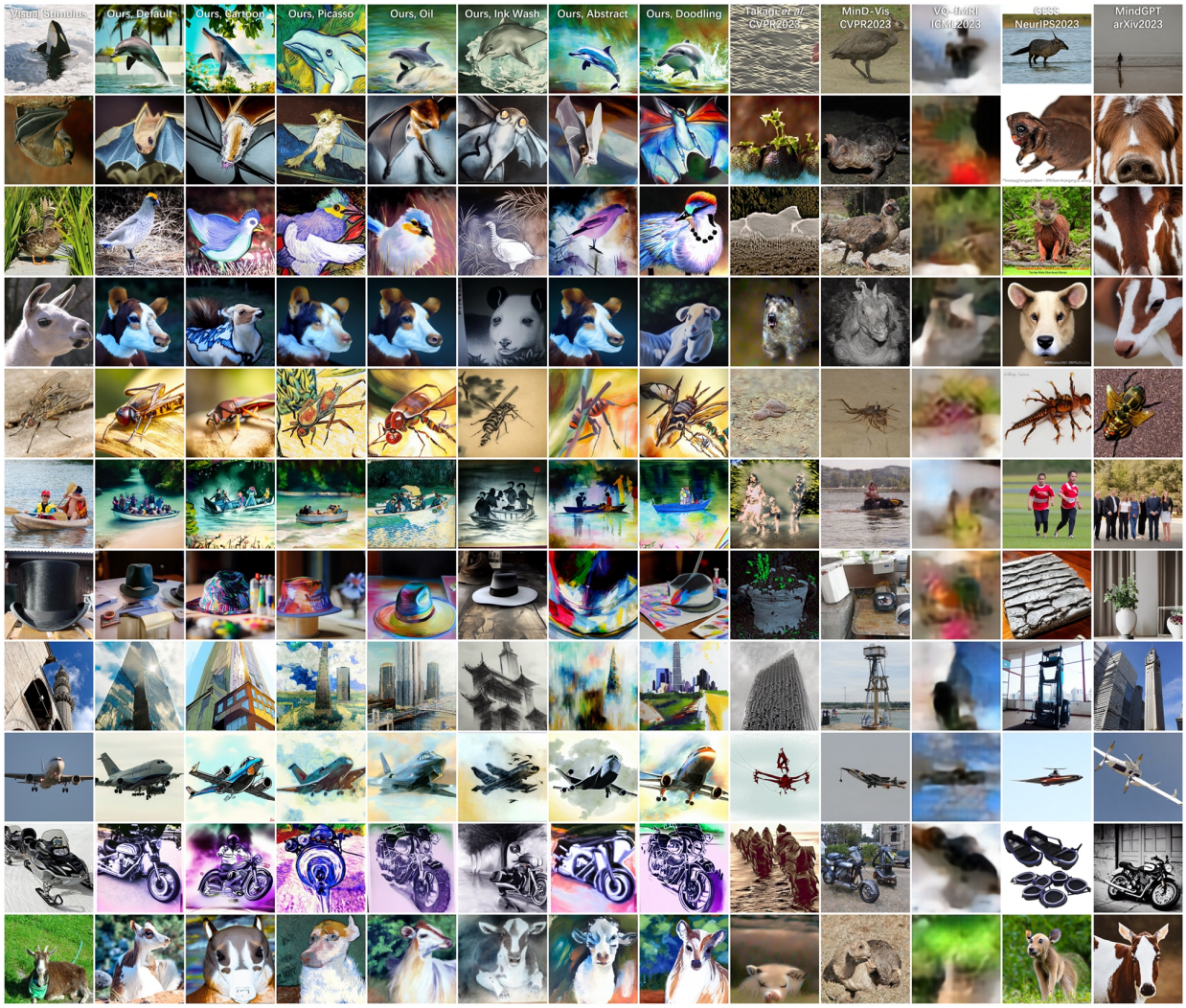

Figure 4. Reconstruction results. We present seven stylized reconstructions of the MindArt. The first column provides ground truth images.

sequences, extracting not only the precise class names of the raw stimuli (e.g., *"boat"*, *"hat"*, and *"bat"*), but often even details like *"in the water"*, *"through the air"* and *"a black and white photo"*. The quantitative results for all test data, calculated based on the pseudo labels generated by the image captioning method SMALLCAP [56], are also summarized in Tab. 2. The results indicate that our method surpasses MindGPT in all five language similarity metrics.

4.5. Ablation Studies

Architecture Variants. We conduct the first ablation experiments to assess the influence of various architectural configuration choices. In what follows, we employ abbreviated notation to denote the model size. Specifically, MindArt-S, MindArt-B, and MindArt-L mean variants with encoders of 4, 8, and 16 layers. The quantitative results of our MindArt on different model configurations are illustrated in Tab. 3. The biggest model, MindArt-L, outperforms the smaller models MindArt-B and MindArt-S in both visual and semantic perception metrics, especially for the challenging CLIP_T@50. We also note that the proposed GM-guided neural representation learning module effectively boosts MindArt's capability to capture semantic information, which is reflected in the significant improvement in semantic perception.

Performance of Different Brain Areas. To examine the potential contributions of different brain areas to the visual reconstruction task, we repeatedly run quantitative experiments using voxels from various ROIs (including VC, LVC, and HVC). Here, VC represents the entire visual cortex, LVC comprises voxels from V1-V3, while voxels from FFA, PPA, and LOC form the HVC. Tab. 4 Top shows the results. We discover that decoding from the HVC resulted in the highest performance in semantic perception metrics.

Figure 5. Qualitative comparison on text reconstruction. For each group, the left is the stimulus. Red denotes obvious semantic deviations.

GM	Model	Visual Perception ↑		Semantic Perception ↑	
		SSIM	$\text{CLIP}_{\text{Score}}$	CLIP_{T}@10	CLIP_{T}@50
✓	MindArt-S	$0.232_{\pm 0.12}$	$0.601_{\pm 0.13}$	$39.9\%_{\pm 4.0\%}$	18.0%
	MindArt-B	$0.233_{\pm 0.14}$	$0.619_{\pm 0.11}$	$40.1\%_{\pm 3.8\%}$	20.0%
	MindArt-L	$\mathbf{0.242}_{\pm 0.13}$	$\mathbf{0.631}_{\pm 0.13}$	$\mathbf{43.6\%}_{\pm 3.9\%}$	**24.0%**
×	MindArt-S	$0.193_{\pm 0.14}$	$0.581_{\pm 0.15}$	$30.2\%_{\pm 4.3\%}$	10.0%
	MindArt-B	$\mathbf{0.199}_{\pm 0.13}$	$0.589_{\pm 0.15}$	$31.1\%_{\pm 4.2\%}$	**12.0%**
	MindArt-L	$0.198_{\pm 0.13}$	$\mathbf{0.591}_{\pm 0.14}$	$\mathbf{32.0\%}_{\pm 4.2\%}$	**12.0%**

Table 3. Quantitative Results with different variants of our MindArt. GM indicates the graph matching-guided fMRI encoder.

Conversely, using the LVC tends to improve the similarity in low-level visual perception (e.g. SSIM) between the reconstructed images and the raw stimuli. The results support a well-accepted theory in neuroscience, the hierarchical nature of visual information propagation [32, 70].

Figure 6. The impact of $\mathbf{X}_z$ in visual reconstruction.

Analyzing the Impact of Latent Representations. Using only high-level semantic information might be inadequate to govern the layout details (e.g., position and orientation) of reconstructed images. Our key idea is incorporating low-level visual information into the reconstruction results by the predicted latent representation $\mathbf{X}_z$. To verify whether that strategy working, we conduct quantitative experiments to evaluate the importance of $\mathbf{X}_z$ in our MindArt, as reported in Tab. 4 Bottom. From the results, it can be observed that $\mathbf{X}_z$ can bring a 22.8% performance gain in the SSIM metrics. Fig. 6 provides examples to visually demonstrate the impact of latent representations on the reconstructed images. In the first example, we observe that $\mathbf{X}_z$ not only conveys positional information of the bat but also nullifies the language semantics related to the phrase "on a table", which might be due to the absence of the outline features representing a table in the latent representation. In the second case, despite the inability to capture fine-grained semantic information, the reconstructed image is still faithful to the stimulus in terms of shape and layout.

Model	Input	Visual Perception ↑		Semantic Perception ↑	
		SSIM	$\text{CLIP}_{\text{Score}}$	CLIP_{T}@10	CLIP_{T}@50
MindArt-L	LVC	$0.239_{\pm 0.14}$	$0.597_{\pm 0.15}$	$39.9\%_{\pm 4.0\%}$	14.0%
	HVC	$0.220_{\pm 0.15}$	$0.628_{\pm 0.13}$	$\mathbf{44.9\%}_{\pm 3.9\%}$	**26.0%**
	VC	$\mathbf{0.242}_{\pm 0.13}$	$\mathbf{0.631}_{\pm 0.13}$	$43.6\%_{\pm 3.9\%}$	24.0%
MindArt-L	*w/o* $\mathbf{X}_z$	$0.197_{\pm 0.15}$	$0.629_{\pm 0.14}$	$\mathbf{43.7\%}_{\pm 3.9\%}$	**24.0%**
	with $\mathbf{X}_z$	$\mathbf{0.242}_{\pm 0.13}$	$\mathbf{0.631}_{\pm 0.13}$	$43.6\%_{\pm 3.9\%}$	**24.0%**

Table 4. Top: Results of visual reconstruction on different brain areas. Bottom: Results of ablation experiments on $\mathbf{X}_z$. The best and worst are highlighted in **Bold** and red, respectively.

5. Conclusion

This paper proposes a novel double-stream neural decoding architecture, termed MindArt, for multi-modal reconstruction. For the first time, we cast traditional visual neural decoding into an OT-based graph matching problem by representing entities in both fMRI and image domains as multiplex graphs. In contrast to previous works, this practice explicitly encourages models to capture potential dependencies between neural patterns, thereby resulting in a natural and meaningful self-supervisory task. Furthermore, our method elegantly handles the coupling problem of visual and linguistic semantics within a unified two-stream model. This "1+N" decoding paradigm allows us to fulfill style transfer (with no optimization required) through prompt-based linguistic semantic editing. By integrating multi-modal information flows into appropriate models, the MindArt can reconstruct not only images, but even videos, 3D models, etc. We leave it to the future work.

Acknowledgments. This work was supported in part by the STI 2030 Major Projects (2021ZD0200400), the Key Research and Development Program of Zhejiang Province in China (2020C03004), the Zhejiang Provincial Natural Science Foundation of China (LR24F020002), and the Natural Science Foundation of China (NSFC) (61925603, U1909202, and 62276228).

References

[1] 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. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[2] Jeffrey R Binder and Rutvik H Desai. The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11): 527–536, 2011. 5

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 2

[4] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021. 2

[5] Jiaxuan Chen, Yu Qi, and Gang Pan. Rethinking visual reconstruction: Experience-based content completion guided by visual cues. In *Proceedings of the 40th International Conference on Machine Learning*, pages 4856–4866. PMLR, 2023. 2, 3, 6

[6] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings. *arXiv preprint arXiv:2309.15729*, 2023. 3, 6

[7] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the semantic latent space between brain and machine: Similarity is all you need. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11302–11310, 2024. 2

[8] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020. 3

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023. 2, 3, 6

[11] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *arXiv preprint arXiv:2305.11675*, 2023. 2, 3

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013. 4, 5

[13] Saudamini Roy Damarla and Marcel Adam Just. Decoding the representation of numerical values from brain activation patterns. *Human Brain Mapping*, 34(10):2624–2634, 2013. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 5, 6

[15] Robert Desimone, John Duncan, et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. 2

[16] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

[18] Changde Du, Changying Du, Lijie Huang, and Huiguang He. Reconstructing perceived images from human brain activities with bayesian deep multiview learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(8): 2310–2323, 2018. 2

[19] Changde Du, Changying Du, Lijie Huang, Haibao Wang, and Huiguang He. Structured neural decoding with multitask transfer learning of deep neural network representations. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):600–614, 2022. 3

[20] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2023. 2

[21] Scott L Fairhall and Alfonso Caramazza. Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience*, 33(25):10552–10558, 2013. 2, 5

[22] Tao Fang, Qian Zheng, and Gang Pan. Alleviating the semantic gap for generalized fmri-to-image reconstruction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3, 6

[23] Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023. 3

[24] Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254: 119121, 2022. 2

[25] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, page 2672–2680. MIT Press, 2014. 2

[26] Fritz Günther, Tri Nguyen, Lu Chen, Carolin Dudschig, Barbara Kaup, and Arthur M Glenberg. Immediate sensorimotor grounding of novel concepts learned from language alone. *Journal of Memory and Language*, 115:104172, 2020. 5

[27] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001. 2

[28] John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nature reviews neuroscience*, 7(7):523–534, 2006. 2

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[32] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):1–15, 2017. 2, 6, 8

[33] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of Neural Science*. McGraw-hill New York, 2000. 2

[34] 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

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[37] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022. 3

[38] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023. 2

[39] Thomas T Liu. Noise contributions to the fmri signal: An overview. *NeuroImage*, 143:141–151, 2016. 2

[40] 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*, pages 5899–5908, 2023. 2, 3

[41] Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*, 2023. 3

[42] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masaaki 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. 2

[43] Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 3

[44] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010. 4

[45] 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

[46] Vincenzo Nicosia, Ginestra Bianconi, Vito Latora, and Marc Barthelemy. Growing multiplex networks. *Physical review letters*, 111(5):058701, 2013. 4

[47] 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 *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022. 3

[48] Allan Paivio. *Imagery and verbal processes*. Psychology Press, 2013. 2

[49] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 460–467. IEEE, 2009. 4

[50] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 3

[51] Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636, 2021. 2, 5

[52] Friedemann Pulvermüller. How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17(9):458–470, 2013. 5

[53] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 5

[55] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4, 6

[56] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023. 5, 7

[57] Ziqi Ren, Jie Li, Xuetong Xue, Xin Li, Fan Yang, Zhicheng Jiao, and Xinbo Gao. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228:117602, 2021. 2

[58] 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*, pages 10684–10695, 2022. 2, 3, 5

[59] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 3, 4, 5

[60] 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. *arXiv preprint arXiv:2305.18274*, 2023. 2, 3

[61] 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*, page 21, 2019. 3

[62] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):1–23, 2019. 2, 6

[63] 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. 2, 3, 5, 6

[64] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, pages 1–9, 2023. 2, 3, 5

[65] Marcel AJ Van Gerven, Botond Cseke, Floris P De Lange, and Tom Heskes. Efficient bayesian multivariate fmri analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50(1):150–161, 2010. 2

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 4, 5

[67] Gang Wang. Lpot: Locality-preserving gromov–wasserstein discrepancy for nonrigid point set registration. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3, 4

[68] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12):4136–4160, 2018. 3

[69] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. 5

[70] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. 8

[71] Elahe' Yargholi and Gholam-Ali Hossein-Zadeh. Brain decoding-classification of hand written digits from fmri data employing bayesian networks. *Frontiers in human neuroscience*, 10:351, 2016. 2

[72] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 5

[73] Yizhen Zhang, Minkyu Choi, Kuan Han, and Zhongming Liu. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. *Advances in Neural Information Processing Systems*, 34:18513–18526, 2021. 5