Single-Stage fMRI-to-3D Reconstruction via Viewpoint-Aware Embedding and Hierarchical Guidance

Xun Zhang[1], Weihao Xia[2], Yulong Liu[3], Bo Yang[4], Alessandro Bozzon[1], Pan Wang[1]*

[1]Delft University of Technology, The Netherlands
[2]University of Cambridge, United Kingdom
[3]The Hong Kong University of Science and Technology, Hong Kong SAR
[4]The Hong Kong Polytechnic University, Hong Kong SAR
X.Zhang-16@tudelft.nl, wx258@cam.ac.uk, yliuom@connect.ust.hk, bo.yang@polyu.edu.hk, A.Bozzon@tudelft.nl, P.Wang-2@tudelft.nl

Abstract

Understanding the neural basis of three-dimensional (3D) perception is a fundamental objective in cognitive neuroscience. Despite advances in decoding 2D visual stimuli from neural data, reconstructing high-fidelity 3D objects with detailed texture and geometry remains largely unexplored. In this work, we introduce **NeuroSculptor3D**, the first single-stage, end-to-end framework for reconstructing textured 3D shapes directly from brain activity. NeuroSculptor3D integrates a viewpoint-aware brain embedding module that captures fine-grained spatial variations across visual perspectives, and a hierarchical guidance mechanism that aligns brain-derived features with perceptual, semantic, and structural priors. Together, these components facilitate the generation of consistent multi-view embeddings, which are then decoded via TRELLIS to produce high-quality textured 3D reconstructions. Experiments on the fMRI-Shape dataset demonstrate that NeuroSculptor3D outperforms existing baselines across multiple settings, achieving significant improvements in both structural accuracy and semantic consistency. Code will be released to facilitate further research.

Introduction

Understanding neural mechanisms underlying visual perception remains a central challenge in cognitive neuroscience. Given that humans navigate a 3D world, the brain's capacity to perceive and interpret complex 3D environments is critical for tasks such as spatial reasoning and scene understanding (Tarr and Bülthoff 1995; Epstein and Kanwisher 1998). While functional magnetic resonance imaging (fMRI) has enabled notable progress in decoding 2D images and videos from brain activity (Shen et al. 2019; Takagi and Nishimoto 2023), reconstructing 3D objects from neural signals remains underexplored. Advancing fMRI-based 3D reconstruction would provide deeper insight into visual cognition and support the development of brain-computer interfaces capable of operating in 3D environments.

Recent efforts in fMRI-based 3D reconstruction (Yang et al. 2024) primarily rely on datasets featuring 2D visual stimuli, such as the Natural Scenes Dataset (NSD) (Allen et al. 2022). However, generating 3D shapes from 2D stimuli diverges from the brain's intrinsic mechanisms for perceiving depth and spatial structure (Yang et al. 2024). To address this, recent datasets like the fMRI-Shape dataset (Gao et al. 2024) have introduced paired fMRI recordings and 3D object data. Despite this advancement, the MinD-3D (Gao et al. 2024) model averages multi-view features (see Fig. 1), which can dilute semantic content and obscure fine-grained structural cues. This mismatch hinders compatibility with

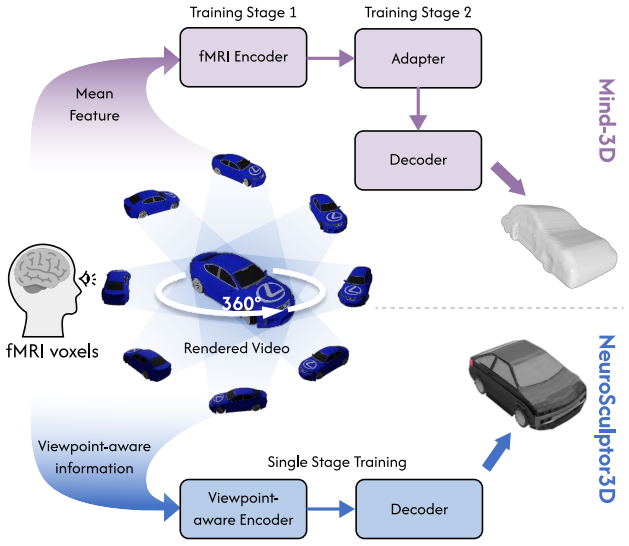

Figure 1: Comparison between existing fMRI-to-3D framework and our proposed NeuroSculptor3D. Prior work (Gao et al. 2024) disregards viewpoint-specific variations by averaging features across views, causing semantic dilution and structural degradation. The two-stage training increases training complexity. Due to model limitations, they produce textureless shapes (top). In contrast, **NeuroSculptor3D** decodes viewpoint-aware brain embeddings aligned with 3D generative priors, enabling single-stage training and high-fidelity textured 3D reconstruction (bottom).

*Corresponding author.

modern 3D generative models and necessitates a two-stage training pipeline, ultimately reducing training efficiency. Moreover, MinD-3D (Gao et al. 2024) focuses solely on reconstructing object geometry, omitting color, texture, and other high-frequency visual attributes. As a result, its reconstructions lack the realism and visual fidelity necessary for accurate and perceptually aligned 3D modeling.

To address these limitations, we draw inspiration from human spatial cognition, where the brain integrates visual input across multiple viewpoints to form internal 3D representations (Kourtzi and Kanwisher 2000; Yamins and DiCarlo 2016). Specifically, we introduce multi-viewpoint feature decoding to emulate the brain's capacity to synthesize spatial information and reconstruct 3D structure. Furthermore, building on the hierarchical human visual system (Felleman and Van Essen 1991) and recent brain decoding studies (Scotti et al. 2024; Ozcelik and VanRullen 2023; Xia et al. 2024a), we incorporate two complementary guidance signals: high-level semantic features from object-category text embeddings and low-level geometric priors that provide coarse yet informative structural constraints.

Building upon these neuroscientific insights, we propose **NeuroSculptor3D**, a novel single-stage training framework for reconstructing textured 3D objects from fMRI signals (cf. Fig. 1). Our approach comprises two key components. First, we introduce a viewpoint-aware brain embedding module that explicitly models viewpoint-specific visual features by conditioning on learnable viewpoint embeddings. This design preserves intra-object variation and facilitates alignment with DINOv2 (Oquab et al. 2025) visual representations through a diffusion prior (Ramesh et al. 2022). Second, we incorporate hierarchical guidance mechanisms that jointly align brain-derived embeddings with complementary priors: high-level semantic information from CLIP text embeddings (Radford et al. 2021) and low-level geometric structure captured by a 3D shape autoencoder (Xiang et al. 2025). This multi-level alignment ensures the simultaneous optimization of perceptual fidelity, semantic consistency, and spatial accuracy. As the reconstruction backbone, we adopt TRELLIS (Xiang et al. 2025), a state-of-the-art generative model capable of producing high-fidelity textured 3D assets from multi-view image features, thereby enabling faithful and generalizable 3D reconstruction.

Experiments on the fMRI-Shape dataset (Gao et al. 2024) demonstrate that NeuroSculptor3D achieves state-of-the-art performance in 3D reconstruction across both semantic and structural evaluation metrics. We assess the model with three distinct generalization settings: (a) Same-Subject Same-Category (SS-SC), which evaluates performance under subject-dependent training; (b) New-Subject Same-Category (NS-SC), which tests generalization to unseen subjects while retaining object categories seen during SS-SC training; and (c) New-Subject New-Category (NS-NC), which further examines generalization to both novel subjects and novel object categories. Across all settings, NeuroSculptor3D consistently outperforms prior methods by a substantial margin. In contrast to previous approaches (Gao et al. 2024) that produce textureless 3D meshes, NeuroSculptor3D generates textured 3D reconstructions with enhanced surface details, significantly improving perceptual realism. In summary, our contributions are threefold:

- We present NeuroSculptor3D, a single-stage end-to-end training framework for reconstructing textured 3D objects directly from fMRI signals, eliminating the need for multi-stage pipelines and improving training efficiency.
- NeuroSculptor3D integrates a viewpoint-aware brain embedding module and a hierarchical guidance strategy that jointly leverage perceptual, semantic, and structural cues to decode viewpoint-specific features and enforce multi-level consistency in the reconstruction process.
- Experiments on the fMRI-Shape benchmark demonstrate that NeuroSculptor3D outperforms prior methods in both accuracy and generalizability across different settings.

Related Work

fMRI Decoding Methods

Existing fMRI decoding has primarily focused on reconstructing 2D images and videos from human brain activity (Shen et al. 2019; Horikawa and Kamitani 2017; Lu et al. 2023; Scotti et al. 2023; Xia and Öztireli 2025), achieving significant progress over recent decades. Standard approaches utilize an encoder to extract relevant features from fMRI signals, coupled with a pretrained decoder, such as a generative adversarial network (GAN) or variational autoencoder (VAE), to reconstruct 2D visual representations (Ozcelik and VanRullen 2023; Scotti et al. 2024; Xia et al. 2024b). Recent advancements have incorporated diffusion-based generative models (Dhariwal and Nichol 2021; Rombach et al. 2022) and multimodal contrastive models, such as CLIP (Radford et al. 2021), to enhance reconstruction. Several studies have further improved fidelity by disentangling high-level semantic and low-level pixel information (Scotti et al. 2023, 2024; Gong et al. 2025), enabling robust decoding in both domains. Despite these advances, existing methods remain limited to 2D stimuli, constraining their ability to capture the volumetric structures inherent to human three-dimensional (3D) perception.

Early efforts to extend fMRI decoding to 3D representations (Yang et al. 2024) have relied on 2D stimulus datasets (Allen et al. 2022), such as the Natural Scenes Dataset (NSD), where decoded semantic features are fed into 3D or multi-viewpoint generative models. However, these approaches, grounded in 2D visual inputs, struggle to extract robust 3D-specific features, compromising reconstruction accuracy. MinD-3D (Gao et al. 2024) represents a pioneering effort to directly reconstruct 3D objects from fMRI signals, utilizing a neuro-fusion encoder to aggregate features and a fine-tuned adapter to align with the feature space of a 3D generative model Argus (Qian et al. 2024). Despite this progress, its reliance on averaging multi-frame visual features from video inputs leads to semantic dilution and structural degradation, hindering alignment with standard 3D generative priors. These limitations underscore the need for approaches that better emulate human visuospatial integration for robust 3D decoding. In contrast, our model is

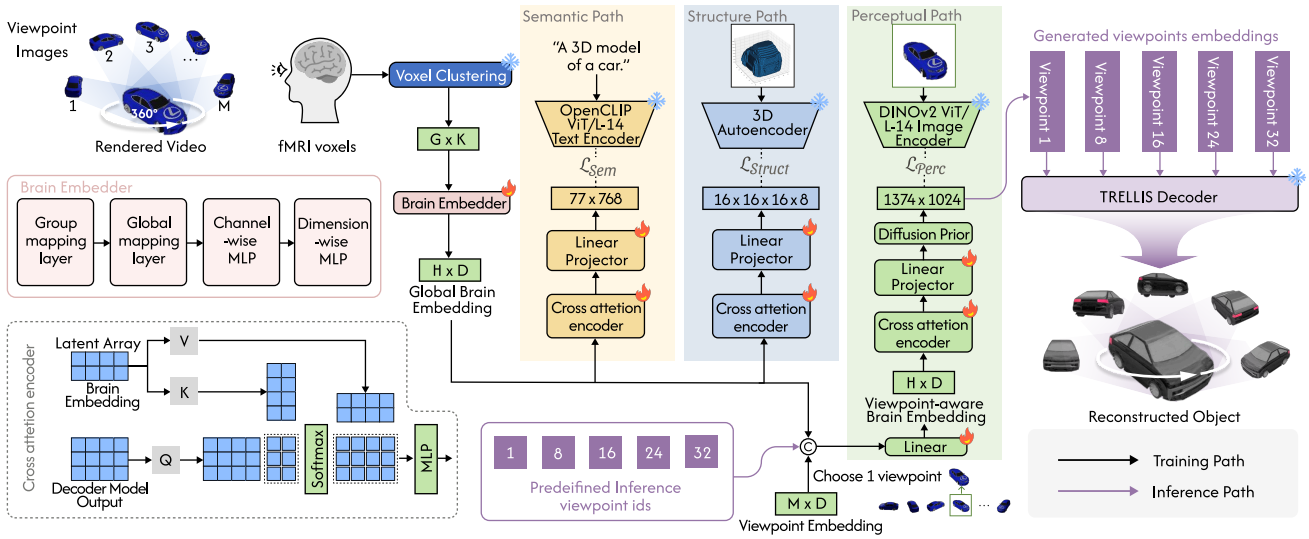

Figure 2: Overview of NeuroSculptor3D. NeuroSculptor3D begins with a voxel clustering module that partitions the fMRI volume into fixed-size clusters of spatially adjacent voxels. A brain embedder then aggregates the clustered signals to extract a global brain embedding. To incorporate viewpoint information, a learnable viewpoint embedding is initialized, from which a specific viewpoint vector is sampled and concatenated with the global embedding during training. This viewpoint-aware embedding is then projected into the DINOv2 feature space of the corresponding frame via a cross-attention module and a learned diffusion prior, serving as the perceptual path. In parallel, two additional guidance streams are employed: a semantic path, which projects the global brain embedding into the CLIP text space, and a geometric path, which maps it to the latent space of a 3D shape autoencoder. Both utilize cross-attention modules followed by linear projection layers. The three pathways, perceptual, semantic, and geometric, together constitute a hierarchical guidance mechanism that supervises the brain-to-feature decoding process. At inference time, predefined viewpoint IDs are used to generate corresponding brain-derived viewpoint-specific embeddings, which are then fed into the pretrained TRELLIS model to synthesize high-fidelity, textured 3D reconstructions.

grounded in human visuospatial cognition and decodes independent multi-view visual features from fMRI signals to support accurate and reliable 3D reconstruction.

3D Generative Models

Early 3D generative methods utilized Generative Adversarial Nets (GANs) (Goodfellow et al. 2014) to model 3D distributions (Chan et al. 2022; Deng et al. 2022; Gao et al. 2022; Wu et al. 2016), such as voxel grids and meshes, but encounter problems while facing diverse scenarios. More recent methods have adopted diffusion models (Ho, Jain, and Abbeel 2020) to generate representations like point clouds (Luo and Hu 2021; Nichol et al. 2022), 3D Gaussians (He et al. 2024; Zhang et al. 2024), and triplanes (Chen et al. 2023a; Wang et al. 2023), offering improved stability and quality. An alternative line of work applies 2D diffusion to generate multi-viewpoint object representations, which are then aggregated to construct 3D shapes (Hong et al. 2024; Liu et al. 2023). Others employ autoregressive models for mesh generation (Chen et al. 2025; He et al. 2024). To improve generation efficiency and modularity, two-stage pipelines often separate shape modeling and texture synthesis (Li et al. 2024; Ren et al. 2024), while single-stage methods (Xia and Xue 2023; Xiong et al. 2025) directly incorporate appearance cues into latent spaces for holistic 3D generation. In this work, we adopt TRELLIS (Xiang et al. 2025), a two-stage 3D generative model that supports multi-view image-guided generation, as the reconstruction backbone for decoding 3D objects from corresponding fMRI signals.

Methodology

NeuroSculptor3D is a novel single-stage framework for reconstructing textured 3D objects from fMRI data using multi-level guidance. It comprises three key components: (1) **viewpoint-aware brain embedding extraction**, which captures viewpoint-specific visual features using learnable viewpoint embeddings; (2) **hierarchical guidance training**, which aligns brain-derived representations with visual, semantic, and geometric priors; and (3) **multi-viewpoint-conditioned reconstruction**, which uses the TRELLIS model to synthesize high-fidelity textured 3D shapes. The overall architecture is illustrated in Fig. 2, and each component is described in detail in the following subsections.

Notations

Our objective is to improve the accuracy of 3D object reconstruction from fMRI data. The dataset $\mathcal{D}$ comprises tuples of (fMRI, 3D shapes, multi-view frames, and captions) as

$$\mathcal{D} = \left\{ \left(X_s(n), V_s(n), \{F_s(n,m)\}_{m=1}^{M}, C_s(n) \right) \right\}_{s=1,n=1}^{S,N} \quad (1)$$

where $s \in S$ indexes subjects and $n \in N$ indexes individual samples, with S and N representing the total numbers of subjects and samples. The fMRI signal for the n-th trial of subject s is denoted as $X_s(n) \in \mathbb{R}^{1\times D}$. Correspondingly,

$V_s(n)$ represents the 3D stimulus in video format, and $C_s(n)$ is the associated 3D object caption. From each video, we extract a set of M frames $F_s = \{1, \ldots, M\}$, each corresponding to a distinct viewpoint. Given the 3D shape structure correspond to stimulus $V_s(n)$, we extract a coarse structural latent $L_s(n) \in \mathbb{R}^{16\times16\times16\times8}$ using a pretrained 3D autoencoder (Xiang et al. 2025). This representation encodes low-level geometry and serves as geometric guidance for shape reconstruction.

Viewpoint-Aware Brain Embedding Extraction

To mitigate inter-individual variability in voxel dimensions and improve cross-subject generalization, we adopt training-free adaptive voxel clustering (Wang et al. 2024). This partitions voxel responses into semantically coherent groups, preserving local spatial structure while standardizing dimensions across subjects. The resulting fMRI representation is denoted as $X'_s(n) \in \mathbb{R}^{G\times K}$, where G is the number of clusters and each cluster contains K spatially adjacent voxels. The clustered $X'_s(n)$ is then fed into a brain embedder to derive a global brain embedding $X^g_s(n) \in \mathbb{R}^{H\times D}$. The brain embedder includes a group mapping layer and a global mapping layer, both implemented as single linear layers. These are followed by two MLP modules for channel-wise and dimension-wise representation learning, each consisting of four sequential MLP blocks with linear transformation, GELU activation, and dropout.

To explicitly incorporate viewpoint-specific information and preserve fine-grained variations across different perspectives, we enable viewpoint-aware brain embedding decoding by introducing a learnable viewpoint embedding module that captures the viewpoint identity of each frame within the stimulus video. Specifically, for a frame with viewpoint index m, the corresponding viewpoint embedding is retrieved as $\mathbf{e}_m \in \mathbb{R}^{1\times D}$. The viewpoint-aware brain embedding $X^v_s(n, m)$ is obtained by concatenating the viewpoint embedding $\mathbf{e}_m$ with the global brain embedding $X^g_s(n)$ along the feature dimension, followed by a linear projection to maintain consistency, yielding:

$$X^v_s(n, m) = \text{Linear}(\text{Concat}(X^g_s(n), \mathbf{e}_m)) \in \mathbb{R}^{H\times D}. \quad (2)$$

Hierarchical Guidance Training

Our model incorporates three complementary hierarchical guidance paths to jointly decode viewpoint-specific visual features, category-level semantic representations, and coarse structural priors directly from neural activity.

Perceptual Path. The perceptual path is viewpoint-aware. Specifically, for each view m, the viewpoint-aware brain embedding $X^v_s(n, m)$, hereafter simplified as $\bar{X}$, is fed into a cross-attention module, where it is projected into key (K) and value (V) matrices. A set of learnable query tokens Q, with dimensions matching the target representation (i.e. 1374×1024), is randomly initialized and iteratively updated through multi-head attention over these brain-derived representations. The cross-attention operation is formulated as:

$$\text{CrossAttn}(Q, \bar{X}) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \quad (3)$$

This enables brain-conditioned latent features to align with the target latent space (e.g. image embedding) in a viewpoint-aware manner. The resulting embedding, denoted as $\hat{Y}^v_s(n, m)$, corresponds to the predicted latent feature for the m-th frame and is projected to match the DINOv2 token space. During training, we enforce perceptual consistency between brain and image features using a diffusion prior loss (Scotti et al. 2024). To enhance the brain-visual alignment, we further adopt the two-stage training strategy (Scotti et al. 2024), which initially use BiMixCo, a mixup-based augmentation method that mitigates overfitting and compensates for limited data, and transition to a contrastive SoftCLIP loss in later epochs. This combined loss is denoted as $\mathcal{L}_{\text{Latent}}$. The total loss for the perceptual path is:

$$\mathcal{L}_{\text{Perc}} = \lambda_{\text{Prior}}\mathcal{L}_{\text{Prior}} + \lambda_{\text{Latent}}\mathcal{L}^{\text{Perc}}_{\text{Latent}}. \quad (4)$$

Semantic and Geometric Guidance Paths. To enrich the model with multi-level context, we introduce two parallel guidance branches that project the global brain embedding $X^g_s(n)$ into the semantic and structural latent spaces using two cross-attention modules and two linear projectors identical to the previous perceptual path. The semantic branch learns to align the brain features with CLIP text representations (i.e. 77×768), while the geometric branch guide the decoding toward a low-resolution 3D latent space (i.e. $16 \times 16 \times 16 \times 8$). The semantic path supervision is applied using element-wise reconstruction loss (i.e. mean squared error, MSE) and a $\mathcal{L}_{\text{BiMixCo|SoftCLIP}}$ (Scotti et al. 2024) for alignment with CLIP text features. The loss for the semantic path is as follows:

$$\mathcal{L}_{\text{Sem}} = \lambda_{\text{MSE}}\mathcal{L}^{\text{Sem}}_{\text{MSE}} + \lambda_{\text{Latent}}\mathcal{L}^{\text{Sem}}_{\text{Latent}}. \quad (5)$$

To provide stronger geometric supervision, we feed the predicted structural latent produced by the geometric path and its ground truth counterpart $L_s(n)$ into the same 3D decoder, resulting in a predicted voxel $\hat{Y}^o_s(n)$ and a target voxel $Y^o_s(n)$. To encourage voxel-level alignment between the two, we apply a Dice loss (Xiang et al. 2025) that maximizes spatial overlap:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \cdot \sum_v \hat{Y}^o_s(n) \cdot Y^o_s(n) + \epsilon}{\sum_v \hat{Y}^o_s(n) + \sum_v Y^o_s(n) + \epsilon}, \quad (6)$$

where v indexes voxel positions and ϵ is a smoothing term. This loss penalizes mismatches in voxel-level occupancy and encourages structurally faithful reconstruction. The final loss for the geometric path is defined as:

$$\mathcal{L}_{\text{Geo}} = \lambda_{\text{MSE}}\mathcal{L}^{\text{Geo}}_{\text{MSE}} + \lambda_{\text{Latent}}\mathcal{L}^{\text{Geo}}_{\text{Latent}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}}. \quad (7)$$

Overall Objective. The final training loss combines all supervisory signals from the three paths:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Perc}} + \mathcal{L}_{\text{Sem}} + \mathcal{L}_{\text{Geo}}. \quad (8)$$

3D Reconstruction with Multi-viewpoint Guidance

During inference, we select d viewpoint indices uniformly sampled from the M available camera views in the stimulus video. For each selected index k, the trained model decodes a brain-derived visual feature embedding conditioned on the

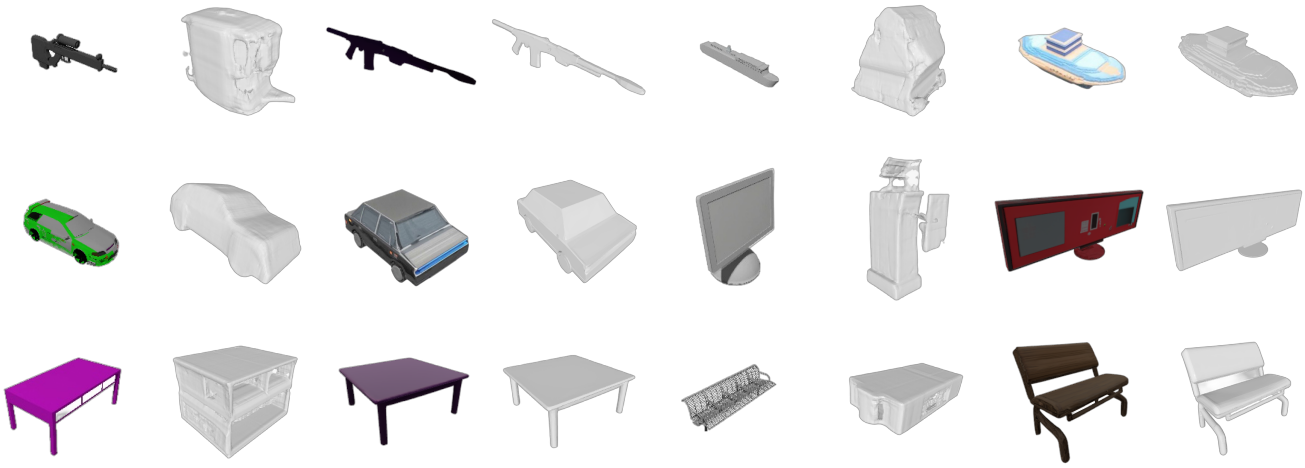

Figure 3: Qualitative results of NeuroSculptor3D on the fMRI-Shape dataset under Same-Subject Same-Category (SS-SC) setting. We show ground truth 3D shapes, mesh outputs from MinD-3D (Gao et al. 2024), and both the textured and mesh outputs from our method NeuroSculptor3D. Compared to MinD-3D, NeuroSculptor3D achieves notably higher structural accuracy and semantic consistency, producing 3D objects that more closely resemble the original targets.

fMRI data and the corresponding viewpoint identity. The resulting set of the d viewpoint-aware embeddings is then aggregated and passed into the pre-trained TRELLIS generator $\mathcal{G}_{\text{TRELLIS}}$ to synthesize the final 3D reconstruction $\hat{S}$:

$$\hat{S}_s(n) = \mathcal{G}_{\text{TRELLIS}}\left(\left\{\hat{Y}_s^v(n,k)\right\}_{k=1}^{d}\right), \tag{9}$$

where $\hat{Y}_s^v(n,k)$ denotes the predicted visual feature for k-th viewpoint. The rightmost panel in Fig. 2 illustrates the inference process. This enables our model to integrate multiple viewpoint-aware neural features derived from the input fMRI, facilitating more accurate 3D reconstruction while maintaining training efficiency by circumventing the requirement to train with all the available multi-view frames.

Experiment

Experimental Setup

Dataset. The fMRI-Shape dataset (Gao et al. 2024) is the first dataset specifically collected for decoding 3D objects from brain activity. It contains fMRI recordings from 12 participants viewing 8-second video clips of 3D objects sampled from ShapeNet (Chang et al. 2015). Each video depicts a continuous 360-degree rotation at a fixed pitch angle, offering rich multi-view visual information. The dataset includes three splits: a Same-Subject Same-Category (SS-SC) set for subject-dependent training and in-distribution evaluation, and two out-of-distribution (OOD) sets, New-Subject Same-Category (NS-SC) and New-Subject New-Category (NS-NC), are designed to evaluate model generalization across unseen subjects and novel object categories.

Data Preprocessing. Departing from the 2D cortical projection strategy employed in (Gao et al. 2024), which represents each 8-second trial as 10 full-brain surface maps, we instead preprocess the raw data in volumetric space using fMRIPrep (Esteban et al. 2019) to preserve spatial fidelity and enables decoding in visually responsive voxels, reducing noise from non-informative regions. To constrain decoding to regions most relevant to visual perception, we adopt the NSDGENERAL mask from NSD (Allen et al. 2022), which aggregates manually annotated, visually responsive regions across eight subjects. Although this mask originates from a different subject pool, prior studies (Güçlü and Van Gerven 2015; Fischl et al. 1999; Klein et al. 2010; Gordon et al. 2017) has shown that population-level functional ROIs can be reliably mapped onto new individuals via non-linear anatomical registration. Following this approach, we construct a group-level mask from the voxel-wise union of the eight NSDGENERAL masks and project it into each participant's native space using ANTs (Avants et al. 2009), ensuring anatomical alignment and compatibility with individual voxel geometries. We then apply GLMsingle (Prince et al. 2022) to the ROI-restricted data to estimate voxel-wise beta responses for each stimulus trial. This procedure yields denoised and temporally deconvolved activation patterns, which serve as input to our decoding model.

Evaluation Metrics. Following prior work (Gao et al. 2024), we evaluate reconstruction quality from both semantic and structural perspectives. For semantic evaluation, we adopt n-way top-k classification accuracy (Chen et al. 2023b; Chen, Qing, and Zhou 2023) to access recognizability, reporting 2-way top-1 and 10-way top-1 accuracies. In addition, we compute the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) to quantify perceptual similarity between reconstructed and ground-truth shapes. For both metrics, 3D objects are rendered into 2D images from multiple viewpoints (every 60 degrees), and scores are averaged across views to obtain the final re-

Method	Semantic-Level			Structure-Level			
	2-way↑	10-way↑	LPIPS↓	SSIM↑	FPD↓	CD↓	EMD↓
LEA-3D (Qian et al. 2023a)	0.803	0.386	0.451	0.622	9.891	10.231	23.753
fMRI-PTE-3D (Qian et al. 2023b)	0.826	0.401	0.339	0.695	9.163	9.512	19.437
MinD-3D (Gao et al. 2024)	0.857	0.483	0.153	0.892	8.137	8.931	15.563
NeuroSculptor3D	**0.909**	**0.638**	**0.073**	**0.951**	**6.163**	**7.012**	**13.639**
w/o semantic path	0.836	0.519	<u>0.141</u>	<u>0.931</u>	<u>6.861</u>	7.798	<u>14.523</u>
w/o geometric path	<u>0.871</u>	<u>0.549</u>	0.194	0.903	7.250	<u>7.319</u>	15.391
w/o both	0.822	0.491	0.208	0.831	7.518	8.291	16.725

Table 1: Quantitative comparison against state-of-the-art methods (top) and ablation study on hierarchical guidance training (bottom) under the Same-Subject Same-Category (SS-SC) setting. Results are averaged across subjects. The **best** and <u>second-best</u> performance are highlighted. ↑ indicates higher is better, while ↓ indicates the opposite.

Method	NS-SC			NS-NC		
	FPD↓	CD↓	EMD↓	FPD↓	CD↓	EMD↓
LEA-3D (Qian et al. 2023a)	19.142	21.377	29.631	24.381	28.402	36.572
fMRI-PTE-3D (Qian et al. 2023b)	16.914	18.620	23.118	21.937	25.123	31.108
MinD-3D (Gao et al. 2024)	12.457	15.281	19.316	17.884	21.645	27.364
NeuroSculptor3D	**10.637**	**12.815**	**16.934**	**13.982**	**15.072**	**20.956**

Table 2: Quantitative comparison under New-Subject Same-Category (NS-SC) and New-Subject New-Category (NS-NC) settings. Following MinD-3D (Gao et al. 2024), for NS-SC, the model is trained on subject 1 and tested on subject 9 with *shared* object categories, and for NS-NC, the model is trained on subject 1 and tested on subject 11 with *unseen* object categories.

sults. On the structural level, we evaluate the geometric consistency using three point-based metrics widely adopted in 3D vision (Xu et al. 2019; Qian et al. 2024; Xiang et al. 2025): Fréchet Point Cloud Distance (FPD), Chamfer Distance (CD), and Earth Mover's Distance (EMD). These metrics are computed between point clouds (2048 points) sampled from predicted and ground-truth meshes. We further report the Structural Similarity Index Measure (SSIM) (Wang et al. 2004), averaged across all rendered viewpoints, to assess frame-level appearance alignment.

Implementation Details. For preprocessing the fMRI-shape (Gao et al. 2024) dataset, we extract 32 frames from each stimulus video and resize them to 224×224. During training, one frame and its corresponding frame ID are randomly sampled per batch. For inference, we use a set of d viewpoint indices (defaulted to 5), uniformly distributed across the viewpoint ID range, to balance accuracy and efficiency. The input fMRI signals are clustered into $G = 512$ voxel groups, each comprising $K = 32$ neighboring voxels. We use AdamW (Loshchilov and Hutter 2019) as the optimizer with a cyclic learning rate scheduler (Smith and Topin 2019), where the maximum learning rate is set to 3.0×10^{-5}. The model is trained for 300 epochs with a batch size of 4. Loss weights are empirically set as follows: $\lambda_{\text{Prior}} = 30$, $\lambda_{\text{Latent}} = 1$, $\lambda_{\text{MSE}} = 10000$, and $\lambda_{\text{Dice}} = 2$. All experiments are conducted on a single NVIDIA H100 GPU with 94 GB of memory.

Experimental Results

For evaluation, we compare NeuroSculptor3D against state-of-the-art methods (Qian et al. 2023a,b; Gao et al. 2024) across three benchmark settings defined in the fMRI-Shape dataset (Gao et al. 2024): (a) **Same-Subject Same-Category** (SS-SC), where models are trained and tested per subject on shared object categories; (b) **New-Subject Same-Category** (NS-SC), training on SS-SC training data and tested on unseen subjects with the same categories; and (c) **New-Subject New-Category** (NS-NC), training on SS-SC and testing on new subjects with novel, disjoint object categories. Quantitative results are shown in Tab. 1 and Tab. 2.

Evaluation on SS-SC. We first evaluate our model on the SS-SC split. Baseline results, including MinD-3D (Gao et al. 2024), LEA-3D (Qian et al. 2023a) and fMRI-PTE-3D (Qian et al. 2023b), are re-calculated using their official codebases for fair comparison. As shown in Tab. 1, our method outperforms all baselines on both semantic and structural metrics, demonstrating superior semantic fidelity and more structurally accurate 3D reconstructions.

Qualitative comparisons in Fig. 3 show that, compared to MinD-3D, our model produces reconstructions that are more accurate and visually faithful, closely resembling the target objects. In contrast, MinD-3D often fails to recover correct semantic attributes and structural details, and generates textureless meshes, whereas our method NeuroSculptor3D reconstructs richly textured 3D shapes. These gains largely stem from our multi-viewpoint guidance, which supplies diverse visual cues to the 3D generation module, enhancing realism and fidelity of the final reconstructions.

Evaluation on NS-SC and NS-NC. To evaluate generalization ability of NeuroSculptor3D under out-of-distribution (OOD) conditions, we further test it on the NS-SC and NS-NC settings. For NS-SC, the model is trained only on data

#Viewpoint	Semantic-Level			Structure-Level				Runtime
	2-way↑	10-way↑	LPIPS↓	SSIM↑	FPD↓	CD↓	EMD↓	(s)
1	0.783	0.436	0.195	0.841	6.928	7.251	15.494	**8**
3	0.854	0.519	0.116	0.885	6.130	7.061	14.751	12
5	0.909	**0.638**	0.073	0.951	6.163	7.012	13.639	15
9	**0.916**	0.617	0.073	**0.958**	5.531	6.933	13.955	34
18	0.913	0.622	**0.069**	0.952	**5.341**	**6.926**	**12.849**	62
MinD-3D	0.857	0.483	0.153	0.892	7.856	7.723	14.412	112

Table 3: Ablation study on the number of input viewpoints during inference. We evaluate the impact of varying the number of decoded viewpoint features d on semantic-level and structure-level metrics, as well as the inference runtime. Results indicate that increasing viewpoint diversity improves both semantic accuracy and structural fidelity, with $i = 5$ achieving an optimal balance between reconstruction quality and computational cost. MinD-3D (Gao et al. 2024) is included as a reference baseline.

Loss			Semantic-Level			Structure-Level			
Latent	MSE	Dice	2-way↑	10-way↑	LPIPS↓	SSIM↑	FPD↓	CD↓	EMD↓
✓			0.614	0.489	0.336	0.703	9.415	11.264	21.574
	✓		0.761	0.552	0.219	0.782	8.311	9.059	17.643
✓	✓		0.892	0.619	0.092	0.942	6.528	7.711	14.032
✓	✓	✓	**0.909**	**0.638**	**0.073**	**0.951**	**6.163**	**7.012**	**13.639**

Table 4: Ablation study on the impact of different loss components under the Same-Subject Same-Category (SS-SC) setting. We evaluate the contribution of latent contrastive loss, MSE loss, and voxel-wise Dice loss to mdoel performance. Results show that each component contributes to improved reconstruction quality, with the full model achieving the best overall performance. All metrics are averaged across subjects; ↑ indicates higher is better, ↓ indicates lower is better.

from subject 1 and tested on subject 9, both sharing the same test object categories. For NS-NC, training is on subject 1, while testing is on subject 11 with unseen categories.

Results are summarized in Tab. 2. As expected, performance on OOD data decreases relative to the in-distribution SS-SC setting. This decline can be attributed to inter-individual variability in brain responses to visual stimuli, influenced by personal, physiological, and contextual factors. Nevertheless, our model consistently outperforms existing methods across both OOD settings, highlighting its robustness and demonstrating the efficacy of hierarchical guidance in enhancing generalizability under distribution shifts.

Ablation Study

Ablation on Hierarchical Guidance. To assess the impact of each component within our hierarchical guidance training, we perform an ablation study under the same-subject same-category setting, with results summarized in Tab. 1. Models with guidance paths (semantic, geometric, or both) removed exhibit declines in both semantic and structural performance metrics, underscoring the importance of each guidance pathway for optimal reconstruction quality.

Ablation on Number of Viewpoints for Inference. To evaluate the effect of multi-viewpoint visual guidance during inference, we conduct an ablation study by varying the number of selected viewpoints $d \in 1, 3, 5, 9, 18$ used to decode brain-derived visual features. As shown in Tab. 3, performance improves consistently across both semantic and structural metrics as the number of viewpoints increases, demonstrating the value of incorporating diverse visual perspectives for more complete and accurate 3D reconstruction. However, the gains plateau beyond $d = 5$, indicating a saturation point where additional views offer diminishing returns relative to computational cost. We therefore set d to 5 in all main experiments to balance quality and efficiency.

Ablation on Loss Functions To evaluate the contribution of each training loss, we perform an ablation by selectively removing losses from the full objective. As shown in Tab. 4, using only the MSE loss moderately degrades performance, whereas relying solely on the latent contrastive loss results in a significant drop in both semantic accuracy and structural consistency. Combining both losses improves alignment between brain-derived features and latent target embeddings. Further adding the Dice loss enhances structural supervision through voxel-level guidance, similar to the low-level pathway in prior work (Scotti et al. 2024; Ozcelik and VanRullen 2023). These results highlight the complementary benefits of perceptual, semantic, and geometric paths for accurate 3D reconstruction from brain signals.

Conclusion

In this work, we present NeuroSculptor3D, a novel single-stage framework for reconstructing high-fidelity, textured 3D objects from fMRI signals. Our method integrates viewpoint-aware embeddings with hierarchical supervision to align brain-derived features across perceptual, semantic, and structural spaces. Multi-viewpoint guidance enhances the accuracy and detail of decoded 3D shapes, while the generative capacity of TRELLIS enables realistic texture synthesis. Experiments on the fMRI-Shape dataset demonstrate that NeuroSculptor3D achieves state-of-the-art performance across in-distribution and out-of-distribution settings.

References

Allen, E. J.; St-Yves, G.; Wu, Y.; Breedlove, J. L.; Prince, J. S.; Dowdle, L. T.; Nau, M.; Caron, B.; Pestilli, F.; Charest, I.; et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1): 116–126.

Avants, B. B.; Tustison, N.; Song, G.; et al. 2009. Advanced normalization tools (ANTS). *Insight j*, 2(365): 1–35.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 16123–16133.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, H.; Gu, J.; Chen, A.; Tian, W.; Tu, Z.; Liu, L.; and Su, H. 2023a. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2416–2425.

Chen, Y.; He, T.; Huang, D.; Ye, W.; Chen, S.; Tang, J.; Chen, X.; Cai, Z.; Yang, L.; Yu, G.; et al. 2025. Meshanything: Artist-created mesh generation with autoregressive transformers. In *ICLR*.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2023b. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *CVPR*, 22710–22720.

Chen, Z.; Qing, J.; and Zhou, J. H. 2023. Cinematic mindscapes: High-quality video reconstruction from brain activity. In *NeurIPS*, volume 36, 24841–24858.

Deng, Y.; Yang, J.; Xiang, J.; and Tong, X. 2022. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 10673–10683.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, 8780–8794.

Epstein, R.; and Kanwisher, N. 1998. A cortical representation of the local visual environment. *Nature*, 392(6676): 598–601.

Esteban, O.; Markiewicz, C. J.; Blair, R. W.; Moodie, C. A.; Isik, A. I.; Erramuzpe, A.; Kent, J. D.; Goncalves, M.; DuPre, E.; Snyder, M.; et al. 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods*, 16(1): 111–116.

Felleman, D. J.; and Van Essen, D. C. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1): 1–47.

Fischl, B.; Sereno, M. I.; Tootell, R. B.; and Dale, A. M. 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8(4): 272–284.

Gao, J.; Fu, Y.; Wang, Y.; Qian, X.; Feng, J.; and Fu, Y. 2024. Mind-3d: Reconstruct high-quality 3d objects in human brain. In *ECCV*, 312–329.

Gao, J.; Shen, T.; Wang, Z.; Chen, W.; Yin, K.; Li, D.; Litany, O.; Gojcic, Z.; and Fidler, S. 2022. Get3d: A generative model of high quality 3d textured shapes learned from images. In *NeurIPS*, volume 35, 31841–31854.

Gong, Z.; Zhang, Q.; Bao, G.; Zhu, L.; Xu, R.; Liu, K.; Hu, L.; and Miao, D. 2025. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *AAAI*, volume 39, 14247–14255.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*, volume 27.

Gordon, E. M.; Laumann, T. O.; Gilmore, A. W.; Newbold, D. J.; Greene, D. J.; Berg, J. J.; Ortega, M.; Hoyt-Drazen, C.; Gratton, C.; Sun, H.; et al. 2017. Precision functional mapping of individual human brains. *Neuron*, 95(4): 791–807.

Güçlü, U.; and Van Gerven, M. A. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014.

He, X.; Chen, J.; Peng, S.; Huang, D.; Li, Y.; Huang, X.; Yuan, C.; Ouyang, W.; and He, T. 2024. Gvgen: Text-to-3d generation with volumetric representation. In *ECCV*, 463–479. Springer.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, 6840–6851.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2024. Lrm: Large reconstruction model for single image to 3d. In *ICLR*.

Horikawa, T.; and Kamitani, Y. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1): 15037.

Klein, A.; Ghosh, S. S.; Avants, B.; Yeo, B. T.; Fischl, B.; Ardekani, B.; Gee, J. C.; Mann, J. J.; and Parsey, R. V. 2010. Evaluation of volume-based and surface-based brain image registration methods. *Neuroimage*, 51(1): 214–220.

Kourtzi, Z.; and Kanwisher, N. 2000. Cortical regions involved in perceiving object shape. *Journal of Neuroscience*, 20(9): 3310–3318.

Li, W.; Liu, J.; Yan, H.; Chen, R.; Liang, Y.; Chen, X.; Tan, P.; and Long, X. 2024. Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*.

Liu, M.; Xu, C.; Jin, H.; Chen, L.; Varma T, M.; Xu, Z.; and Su, H. 2023. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*, volume 36, 22226–22246.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *ICLR*.

Lu, Y.; Du, C.; Zhou, Q.; Wang, D.; and He, H. 2023. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *ACM MM*, 5899–5908.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2837–2845.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2025. Dinov2: Learning robust visual features without supervision. *TMLR*.

Ozcelik, F.; and VanRullen, R. 2023. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1): 15666.

Prince, J. S.; Charest, I.; Kurzawski, J. W.; Pyles, J. A.; Tarr, M. J.; and Kay, K. N. 2022. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *Elife*, 11: e77599.

Qian, X.; Wang, Y.; Fu, Y.; Xue, X.; and Feng, J. 2023a. Semantic neural decoding via cross-modal generation. *arXiv preprint arXiv:2303.14730*.

Qian, X.; Wang, Y.; Huo, J.; Feng, J.; and Fu, Y. 2023b. fmri-pte: A large-scale fmri pretrained transformer encoder for multi-subject brain activity decoding. *arXiv preprint arXiv:2311.00342*.

Qian, X.; Wang, Y.; Luo, S.; Zhang, Y.; Tai, Y.; Zhang, Z.; Wang, C.; Xue, X.; Zhao, B.; Huang, T.; et al. 2024. Pushing auto-regressive models for 3d shape generation at capacity and scalability. *arXiv preprint arXiv:2402.12225*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ren, X.; Huang, J.; Zeng, X.; Museth, K.; Fidler, S.; and Williams, F. 2024. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *CVPR*, 4209–4219.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Scotti, P.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Dempster, A.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K.; et al. 2023. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *NeurIPS*, 36.

Scotti, P. S.; Tripathy, M.; Villanueva, C. K. T.; Kneeland, R.; Chen, T.; Narang, A.; Santhirasegaran, C.; Xu, J.; Naselaris, T.; Norman, K. A.; et al. 2024. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. In *ICML*.

Shen, G.; Horikawa, T.; Majima, K.; and Kamitani, Y. 2019. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1): e1006633.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, 369–386.

Takagi, Y.; and Nishimoto, S. 2023. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, 14453–14463.

Tarr, M. J.; and Bülthoff, H. H. 1995. Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*.

Wang, T.; Zhang, B.; Zhang, T.; Gu, S.; Bao, J.; Baltrusaitis, T.; Shen, J.; Chen, D.; Wen, F.; Chen, Q.; et al. 2023. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 4563–4573.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4): 600–612.

Wang, Z.; Zhao, Z.; Zhou, L.; and Nachev, P. 2024. UniBrain: A Unified Model for Cross-Subject Brain Decoding. *arXiv preprint arXiv:2412.19487*.

Wu, J.; Zhang, C.; Xue, T.; Freeman, B.; and Tenenbaum, J. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NeurIPS*, volume 29.

Xia, W.; de Charette, R.; Oztireli, C.; and Xue, J.-H. 2024a. Dream: Visual decoding from reversing human visual system. In *WACV*, 8226–8235.

Xia, W.; de Charette, R.; Oztireli, C.; and Xue, J.-H. 2024b. Umbrae: Unified multimodal brain decoding. In *ECCV*, 242–259.

Xia, W.; and Xue, J.-H. 2023. A survey on deep generative 3d-aware image synthesis. *ACM Computing Surveys*, 56(4): 1–34.

Xia, W.; and Öztireli, C. 2025. Exploring the Visual Feature Space for Multimodal Neural Decoding. In *ICCV*.

Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2025. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, 21469–21480.

Xiong, B.; Wei, S.-T.; Zheng, X.-Y.; Cao, Y.-P.; Lian, Z.; and Wang, P.-S. 2025. Octfusion: Octree-based diffusion models for 3d shape generation. *Computer Graphics Forum*.

Xu, Q.; Wang, W.; Ceylan, D.; Mech, R.; and Neumann, U. 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, volume 32.

Yamins, D. L.; and DiCarlo, J. J. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3): 356–365.

Yang, Y.; Zhang, L.; Xie, Z.; Yuan, Z.; Feng, J.; Zhu, X.; and Jiang, Y.-G. 2024. Brain3D: Generating 3D Objects from fMRI. *arXiv preprint arXiv:2405.15239*.

Zhang, B.; Cheng, Y.; Yang, J.; Wang, C.; Zhao, F.; Tang, Y.; Chen, D.; and Guo, B. 2024. Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. In *NeurIPS*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.