Mind2Cloud: EEG-to-Point Cloud Generation with Two-Granularity Diffusion Decoding

Yongyi Lu[1], Xiongfeng Huang[1], and Zhijing Yang[1]*

Guangdong University of Technology, Guangzhou, China
{yylu, yzhj}@gdut.edu.cn, huangxiongfeng@mails.gdut.edu.cn

Abstract. Reconstructing 3D objects from brain signals offers a promising avenue for understanding human visual cognition. While prior work has shown initial success using EEG signals for 3D reconstruction, existing methods typically employ a uniform diffusion decoder, overlooking the evolving semantic granularity of both EEG representations and the diffusion denoising process. In this paper, we propose Mind2Cloud, a novel EEG-to-point-cloud generation framework based on two-granularity diffusion decoding. The core of Mind2Cloud is a time-aware decoder that integrates a global Transformer branch and a local Point-Voxel CNN (PVCNN) branch across diffusion timesteps through a learnable fusion mask. Specifically, Transformer layers are incorporated into the early upsampling stages to capture global object structure under high uncertainty, while PVCNN modules are used in later stages to refine local geometric details. Inspired by the hierarchical nature of EEG-based visual representations, this design dynamically adapts its spatial granularity in accordance with the coarse-to-fine trajectory of diffusion denoising. We further introduce an adversarial refinement module to enhance geometric realism and semantic consistency. Extensive experiments on the EEG-3D dataset across all 12 subjects demonstrate that Mind2Cloud outperforms prior work in both geometric accuracy and semantic alignment, setting a new benchmark for EEG-to-point-cloud generation. Our source code is available at `https://github.com/duasoi/Mind2Cloud`.

Keywords: EEG-3D · Point Cloud Generation · Diffusion Models

1 Introduction

Reconstructing 3D shapes from non-invasive brain signals is a fundamental challenge in neural decoding, as it requires mapping noisy, low-dimensional neural activity to structured, high-fidelity geometry representations. Prior studies have made notable progress in reconstructing 2D images from fMRI or EEG signals [6,24,41]. However, extending visual decoding from 2D images to 3D is substantially more challenging, due to the larger domain gap between neural signals and geometric structures. Recent work such as Neuro-3D [13] has demonstrated

* Corresponding author.

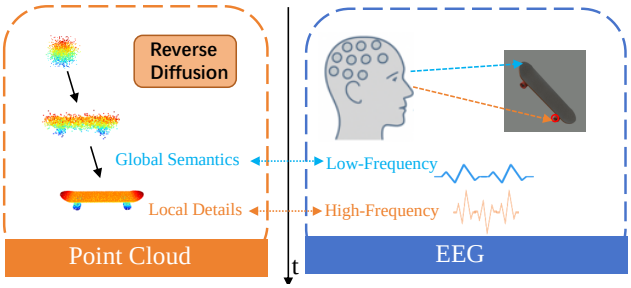


Fig. 1: Conceptual illustration of two-granularity diffusion decoding for EEG-to-point-cloud generation. Low-frequency EEG cues guide global semantic structure during early denoising, while high-frequency cues support local detail refinement in later steps.

the feasibility of EEG-to-3D shape reconstruction by combining an EEG encoder for static and dynamic signal learning with a diffusion-based point cloud generator. Nevertheless, existing methods typically employ a fixed decoding architecture throughout the entire generation process, overlooking the fact that the semantic granularity required for reconstruction evolves across both EEG representations and the diffusion denoising process.

Neuroscientific evidence suggests that EEG carries visual information at multiple levels of abstraction: low-frequency components are often associated with global object-level perception, while high-frequency components are more closely related to local visual details such as edges, contours and fine structures [7, 49]. A similar hierarchy naturally appears in diffusion-based point cloud generation. During early denoising steps, the model operates on highly noisy latent shapes and must recover coarse global structure. In later steps, it progressively refines local geometry and details, as illustrated in 1. This conceptual hierarchy serves as a key inspiration for designing a 3D visual decoding architecture from EEG that can adapt its spatial granularity over time, in alignment with the coarse-to-fine trajectory of diffusion generation.

To this end, we propose ***Mind2Cloud***, a two-granularity diffusion decoder for EEG-guided point cloud reconstruction. Our decoder integrates a global Transformer branch and a local Point-Voxel CNN (PVCNN) branch, fusing their outputs in a timestep-aware manner. Specifically, during the diffusion process, we embed Transformer blocks into the early upsampling stages of the decoder, where the model must synthesize coarse structure under high uncertainty. As denoising progresses and the generated point cloud becomes more structured, we gradually reduce the influence of the Transformer and rely on the PVCNN branch to refine spatial detail. A learnable fusion mask modulates the relative contribution of each branch at different timesteps, allowing the network to smoothly transition from global semantic reconstruction to local geometric refinement. Rather than a mere architectural combination, the key novelty of Mind2Cloud lies in its EEG-conditioned, timestep-adaptive global/local fusion mechanism. This design dynamically balances complementary decoding granularity across denoising

stages and implicitly aligns the hierarchical information encoded in EEG signals with the generative process.

To further enhance geometric realism, we introduce a lightweight adversarial refinement module that discriminates between reconstructed and real point clouds. The entire model is trained with a combination of contrastive alignment loss, geometric reconstruction loss and adversarial supervision. We validate Mind2Cloud on the EEG-3D dataset across all 12 subjects, demonstrating substantial improvements over prior work in both perceptual quality and reconstruction accuracy. The proposed decoder produces more structurally faithful 3D shapes, better preserves semantic correspondence with the neural signals without increasing point cloud resolution or inference cost.

In summary, our main contributions are as follows:

- We show that selectively incorporating Transformer modules into early upsampling stages improves EEG-to-point-cloud generation by better matching the coarse-to-fine nature of both neural visual perception and diffusion decoding.
- We introduce a two-granularity point cloud decoder featuring an EEG-conditioned, timestep-adaptive fusion mechanism that adaptively fuses global Transformer and local PVCNN features across diffusion timesteps.
- Mind2Cloud sets a new benchmark for EEG-to-point-cloud generation on the EEG-3D dataset under full-subject evaluation across all 12 subjects, achieving improved geometric accuracy and semantic fidelity.

2 Related Work

2.1 Neural Decoding of 2D Visual Stimuli

Decoding visual stimuli from neural signals [4–6, 21, 41, 49] has emerged as a key research direction bridging computer vision and neuroscience, aiming to reveal how the brain encodes visual information. Early efforts leveraged CNNs and GANs to map fMRI or EEG signals to 2D images [2,12,15,24,31,42], demonstrating the viability of reconstructing perceptually coherent visuals from brain activity. Recent advances in diffusion models [14,30] and vision-language frameworks such as CLIP [22] have further enhanced decoding fidelity. Many approaches align neural representations with pre-trained visual or textual embeddings via contrastive learning, enabling conditional generation from brain signals [1,36,45]. Despite success in static 2D and short video reconstruction [6,20,43], these methods remain constrained to 2D, limiting their ability to capture the spatial depth of 3D perception.

2.2 3D Reconstruction from Brain Signals

Reconstructing 3D objects from neural activity is an emerging direction in neural decoding, with promising applications in brain–computer interfaces (BCIs) and understanding spatial perception. Early work mainly used fMRI due to its high

spatial resolution to recover coarse 3D structures [10, 11]. Projects like Mind-3D [11] and Mind-3D++ [9] introduced paired datasets and diffusion-based generative models for geometry inference. However, fMRI methods are costly, immobile, and have poor temporal resolution [33], limiting real-time or large-scale BCI use. They also often miss perceptual details like color and texture, important for semantic fidelity. To address this, Neuro-3D [13] proposed the first EEG-based 3D reconstruction dataset and framework, allowing geometry and appearance estimation from non-invasive EEG signals. EEG is affordable, portable, and temporally responsive, better suited for interactive real-time BCIs [17, 46]. Despite this, EEG-to-3D reconstruction accuracy and completeness remain limited, especially under noisy, low-SNR conditions. This motivates further research on improving neural representations, model robustness, and cross-modal alignment to advance reliable, semantically rich 3D brain decoding systems.

2.3 Diffusion Models

Diffusion models have become a powerful generative paradigm capable of synthesizing high-quality visual content by reversing a noise corruption process [14]. Their applications span 2D image generation [18, 37, 39, 40], editing [48, 50], and 3D tasks including point cloud modeling [44, 52], multi-view synthesis [16, 26, 27, 53], and text-to-3D reconstruction [3, 23, 29, 34]. To enhance efficiency and detail fidelity, latent diffusion models compress inputs via autoencoders [39], while Transformer-based architectures like DiT improve scalability and generation quality [32]. In neuroscience, diffusion models have been leveraged for fMRI-based visual decoding, with frameworks like TIGER [38] combining convolutional and Transformer modules to reconstruct dynamic 3D shapes. Building on these insights, we extend diffusion-based generation to EEG-driven 3D reconstruction, establishing a new direction for neural-guided generative modeling.

3 Methodology

3.1 Overview

We propose an end-to-end framework for reconstructing 3D shapes from EEG signals, as illustrated in Figure 2. The framework consists of three main components: 1) **Two-stream EEG Encoder**, which adaptively integrates static and dynamic EEG signals to extract complementary neural representations. Motivated by the Neuro-3D [13] paradigm, this design enhances the expressiveness of EEG features for 3D object understanding. To ensure semantic alignment across modalities, we introduce InfoNCE and MSE losses between the EEG embeddings and CLIP-derived visual embeddings; 2) **Point Cloud Decoder**, which reconstructs 3D shapes conditioned on the EEG embedding by leveraging a latent point cloud Transformer. This decoder fuses global representations from the Transformer and local geometric features extracted via shallow PVCNN modules [25] using a timestep-aware weighting strategy; and 3) **Adversarial**

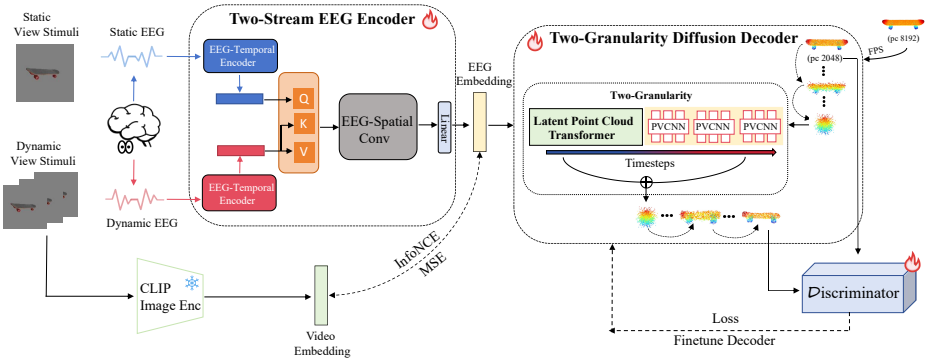


Fig. 2: Overview of Mind2Cloud. A two-stream EEG encoder extracts complementary features from static and dynamic EEG responses and aligns them with CLIP image embeddings. The resulting EEG embedding guides a two-granularity diffusion decoder to reconstruct 3D point clouds, while adversarial refinement improves geometric fidelity.

Refinement, where a discriminator is employed to distinguish generated point clouds from ground-truth samples. This adversarial objective fine-tunes the decoder, guiding it to produce more realistic and structurally coherent 3D outputs.

3.2 Two-stream EEG Encoder

To extract robust neural representations from temporally and spatially distributed EEG signals, we propose a dual-branch EEG encoder that separately processes static and dynamic modalities and integrates them via temporal attention and spatial convolution.

EEG-Temporal Encoder. Given preprocessed EEG signals $x_s \in \mathbb{R}^{C\times T_s}$ and $x_d \in \mathbb{R}^{C\times T_d}$ corresponding to static and dynamic visual stimuli, we employ two Transformer encoders to model their temporal structures. Each input sequence is enriched with learnable positional encodings and encoded as:

$$\begin{aligned} \mathbf{z}_{\text{dyn}} &= \text{Transformer}(\text{PosEnc}(x_d)) \\ \mathbf{z}_{\text{stc}} &= \text{Transformer}(\text{PosEnc}(x_s)) \end{aligned} \tag{1}$$

To fuse temporal cues from both branches, we apply a cross-attention mechanism where static features serve as the query and dynamic features as key–value pairs:

$$\mathbf{z}_{\text{fuse}} = \text{CrossAttn}(Q = \mathbf{z}_{\text{stc}},\ K = \mathbf{z}_{\text{dyn}},\ V = \mathbf{z}_{\text{dyn}}), \tag{2}$$

yielding a unified temporal representation that integrates static and motion-related information.

EEG-Spatial Conv. To model inter-channel dependencies, we apply a series of 1D convolutional layers with residual connections to $\mathbf{z}_{\text{fuse}}$, followed by a linear projection to generate the final EEG embedding. This spatial encoding enhances the topographical structure of EEG signals and stabilizes training.

To align EEG and visual representations, we adopt a symmetric InfoNCE loss:

$$\mathcal{L}_{\text{clip}} = \frac{1}{2}\left[\text{CE}\left(\frac{\mathbf{z}_{\text{fuse}}\mathbf{f}_v^\top}{\tau}, y\right) + \text{CE}\left(\frac{\mathbf{f}_v\mathbf{z}_{\text{fuse}}^\top}{\tau}, y\right)\right], \tag{3}$$

where τ is a temperature parameter, and an additional regression term for direct embedding alignment:

$$\mathcal{L}_{\text{img}} = \|\mathbf{z}_{\text{fuse}} - \mathbf{f}_v\|_2^2. \tag{4}$$

The combined loss is defined as:

$$\mathcal{L}_{\text{EEG}} = \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{clip}}, \tag{5}$$

encouraging EEG embeddings that are both semantically aligned and discriminative for downstream 3D generation.

3.3 Two-Granularity Diffusion Decoder

To reconstruct high-fidelity 3D shapes from noisy EEG-derived embeddings, we design a **time-aware Two-Granularity point cloud decoder** that adaptively adjusts its spatial granularity across diffusion timesteps. While Transformer and CNN architectures are standard modules, their complementary inductive biases remain highly effective for structured 3D generation tasks. In our design, the Transformer branch captures long-range dependencies and global semantic structure, whereas the CNN-based PVCNN branch focuses on localized geometric refinement and high-frequency spatial details. Importantly, the innovation of our framework lies not in architectural reinvention, but in the EEG-conditioned diffusion step–adaptive fusion mechanism. This mechanism dynamically balances global and local representations across denoising stages while being guided by neural embeddings extracted from brain activity. By coupling EEG semantics with timestep-aware feature modulation, the decoder progressively translates abstract neural representations into structured geometric details, achieving semantically grounded and structurally coherent 3D generation. This design is motivated by the coarse-to-fine trajectory of diffusion-based generation and the hierarchical abstraction levels present in EEG signals. Drawing inspiration from the TIGER architecture [38], our decoder integrates a local CNN-based branch for geometric refinement, a Transformer-based branch for modeling global structure, and a learnable fusion mechanism that dynamically balances the two throughout the denoising process. This enables the decoder to progressively transition from global shape synthesis to fine-grained spatial detail, aligning feature representation with the generative stage.

Local Geometric Refinement. Given a noisy point cloud $\mathbf{X}_t \in \mathbb{R}^{N\times 3}$ at diffusion timestep t, we first concatenate it with EEG-derived latent embeddings to form a joint representation. To extract high-frequency spatial details and local semantic cues, we adopt a hierarchical feature extractor built upon the Point-Voxel Convolutional Neural Network (PVCNN) framework.

Specifically, we employ a multi-stage Set Abstraction (SA) architecture that combines point-based and voxel-based processing. In each stage, point-wise MLPs capture fine-grained variations by aggregating features from local k-NN neighborhoods, while voxel-based convolutions introduce geometric regularity and reduce sensitivity to spatial noise. This hybrid design allows the model to simultaneously leverage unstructured geometric richness and structured volumetric priors, making it well-suited for EEG-guided generation where signal uncertainty is high.

The extracted local features are progressively downsampled and aggregated, preserving intermediate coordinates and activations for later upsampling via Feature Propagation (FP) layers. This enables effective gradient flow and spatial correspondence between encoder and decoder stages. Overall, the PVCNN branch focuses on capturing localized structures—such as edges, surface curvature, and fine object parts—that are critical for reconstructing semantically meaningful and geometrically detailed 3D shapes. When integrated with the Transformer-based global branch, it provides complementary inductive bias that enhances both structural fidelity and robustness during generation.

Global Semantic Reconstruction. To effectively model global spatial relationships within latent point clouds, we introduce a Transformer-based module operating at an intermediate resolution. Specifically, given latent point cloud features $\hat{\mathbf{X}}_t \in \mathbb{R}^{M\times d}$, we first encode them into initial token embeddings using a dual normalization [19] and intermediate MLP scheme:

$$\mathbf{T}_0 = \mathrm{LN}\big(\mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{X}}_t))\big) \in \mathbb{R}^{M\times D}. \tag{6}$$

We further enrich these embeddings with explicit positional information by applying a continuous, multi-granularity 3D positional embedding [38] $\mathrm{Pemb} \in \mathbb{R}^{M\times D}$. This encoding scheme effectively captures fine-grained spatial differences and maintains continuity across spatial dimensions, essential for modeling irregular and continuous 3D data. Inspired by [51], rather than directly adding positional embeddings to token representations, which could diminish positional information across deeper layers, we integrate spatial context through a learned spatial interaction matrix (Fig. 3):

$$H = \mathrm{Softmax}\left((\mathrm{Pemb}W_p)(\mathrm{Pemb}W_p)^\top\right), \tag{7}$$

where $W_p \in \mathbb{R}^{D\times Z}$ is a projection matrix mapping the positional embeddings into an attention-specific space. This learned interaction explicitly preserves positional relationships, enabling consistent incorporation of relative spatial structure throughout successive Transformer layers.

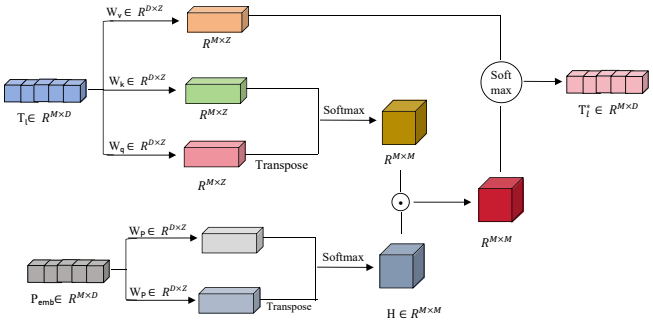


Fig. 3: Position-aware attention mechanism with a learned spatial interaction matrix H, integrating positional embeddings into Transformer attention for enhanced 3D structural modeling.

Formally, we stack L Transformer layers, where at each layer $l \in \{0, 1, ..., L-1\}$, the token embedding $\mathbf{T}_l \in \mathbb{R}^{M \times D}$ is iteratively refined. The position-aware attention computation within each Transformer layer is defined as:

$$\mathbf{T}_l^* = \text{Softmax}\left(\frac{(\mathbf{T}_l W_q)(\mathbf{T}_l W_k)^\top \odot H}{\sqrt{Z}}\right)(\mathbf{T}_l W_v), \tag{8}$$

where $W_q, W_k, W_v \in \mathbb{R}^{D \times Z}$ are the query, key, and value projection matrices, and Z denotes the dimensionality of the projected query and key vectors. A feed-forward network follows each attention layer to further refine the token embeddings. After L iterations, we obtain the final representation $\mathbf{F}_{\text{tr}}$, which encodes high-level semantic structure and serves as a global prior for downstream reconstruction.

This Transformer-based branch is particularly effective during early diffusion steps, where global semantics dominate and spatial uncertainty remains high. By capturing object-level structure and integrating geometric context over long distances, it complements the local refinement pathway and enhances the fidelity of 3D shape reconstruction.

Time-aware Feature Fusion. To dynamically integrate coarse global structure with fine local details across the generative trajectory, we design a **time-aware feature fusion module** that adaptively modulates the contributions of the Transformer and CNN branches throughout the diffusion process. This design reflects the observation that early timesteps favor semantic abstraction, while later stages benefit from localized refinement.

Specifically, the diffusion timestep t is encoded using sinusoidal positional embeddings and fed into an MLP to generate a learnable fusion mask $M_t \in \mathbb{R}^D$. The mask is constrained to $(0, 1)$ via a sigmoid activation and applied channel-wise, with broadcasting along spatial dimensions.

The fused feature is computed as:

$$F_{\text{out}} = M_t \cdot \text{Conv}(F_{\text{conv}}) + (1 - M_t) \cdot \text{TF}(F_{\text{tr}}), \tag{9}$$

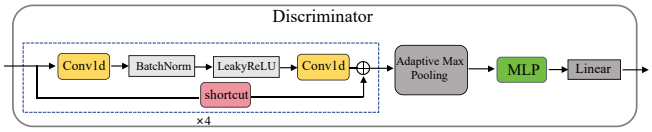


Fig. 4: Discriminator architecture. The input point cloud is processed by four stacked residual Conv1D blocks, followed by global pooling and an MLP classifier.

where $\mathrm{Conv}(\cdot)$ and $\mathrm{TF}(\cdot)$ denote 1×1 learnable convolutions that project both branches into a shared latent space. The mask M_t allows the network to emphasize global structural information during early denoising stages (high noise), and progressively shift toward local detail enhancement as uncertainty decreases over time.

The fused representation F_{out} is further processed by normalization and residual blocks, followed by a point-wise MLP that predicts the noise residual $\hat{\epsilon}$ for each 3D point. This output is used to guide the denoising process, refining the noisy point cloud $\mathbf{X}_t$ toward the target shape $\mathbf{X}_0$. The diffusion reconstruction is supervised via a standard noise prediction loss:

$$\mathcal{L}_{\text{diffuse}} = \|\hat{\epsilon} - \epsilon\|_2^2, \tag{10}$$

where ϵ is the Gaussian noise added during the forward diffusion step.

To jointly optimize geometric accuracy and EEG-image alignment, we combine this diffusion loss with the EEG supervision loss $\mathcal{L}_{\text{EEG}}$ (as introduced in Section 3.2), which includes an InfoNCE contrastive term $\mathcal{L}_{\text{clip}}$ and a regression term $\mathcal{L}_{\text{img}}$. The overall training objective is defined as:

$$\mathcal{L}_{\text{joint}} = \alpha \cdot \mathcal{L}_{\text{diffuse}} + (1 - \alpha) \cdot \mathcal{L}_{\text{EEG}}, \tag{11}$$

where $\alpha \in [0, 1]$ trades off between accurate shape reconstruction and cross-modal representation learning.

3.4 Adversarial Refinement

To enhance the geometric realism of reconstructed shapes, we introduce a lightweight point cloud discriminator that distinguishes real samples from generated ones. Given an input point cloud $\mathbf{X} \in \mathbb{R}^{N \times 3}$, the discriminator outputs a scalar authenticity score.

As shown in Figure 4, the discriminator leverages stacked residual 1D convolutional blocks to model local geometry and channel dependencies. A global max pooling followed by normalization produces a compact representation, which is mapped to a binary prediction.

During training, adversarial learning is applied using binary cross-entropy loss on logits (BCEWithLogitsLoss). Given real samples $\mathbf{X}_0$ and generated ones $\tilde{\mathbf{X}}$, the discriminator loss is:

$$\mathcal{L}_{\text{D}} = \mathrm{BCE}(D(\mathbf{X}_0), 1) + \mathrm{BCE}(D(\tilde{\mathbf{X}}), 0), \tag{12}$$

where $D(\cdot)$ denotes the discriminator output before activation.

Meanwhile, the generator is trained with an adversarial objective that encourages indistinguishability from real shapes:

$$\mathcal{L}_{\mathrm{G}} = \mathrm{BCE}(D(\tilde{\mathbf{X}}), 1). \tag{13}$$

This adversarial term is combined with the primary loss $\mathcal{L}$—which includes diffusion-based reconstruction and EEG alignment—yielding the final training objective:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{joint}} + \lambda_{\mathrm{adv}} \cdot \mathcal{L}_{\mathrm{G}}, \tag{14}$$

where λ_{adv} balances the influence of adversarial feedback. This joint formulation promotes structurally consistent and perceptually realistic 3D outputs.

4 Experiments

4.1 Datasets

We adopt the same dataset as the Neuro-3D framework, a multimodal benchmark tailored for EEG-driven 3D reconstruction. The dataset comprises EEG recordings from 12 participants, each exposed to both static images and 360-degree rotating videos of 3D objects drawn from 72 categories in the Objaverse dataset [8,47]. Each object is annotated with shape category, textual description, and color style labels. EEG signals were recorded using a 64-channel EASYCAP system at a sampling rate of 1000 Hz. A controlled stimulus paradigm was employed, presenting each object with a 0.5-second static image followed by a 6-second rotation video to elicit rich visual responses. Each participant completed approximately 5.5 hours of EEG acquisition, including resting-state, image-evoked, and video-evoked sessions. EEG preprocessing was conducted using the MNE toolbox. For more details, we refer readers to the original Neuro-3D study [13].

4.2 Implementation Details

All experiments were performed on a single NVIDIA A6000 GPU, with total training time of approximately 2.5 days. We employed the AdamW optimizer with an initial learning rate of 1×10^{-3}. The visual encoder produced feature vectors of dimension 1024. In training, the loss weight α in Equation (11) was set to 0.95 to emphasize precise geometric reconstruction, while the adversarial loss weight λ_{adv} in Equation (14) was set to 0.05 to incorporate moderate discriminator guidance for improving surface fidelity. The Point Cloud Decoder utilized an 8-layer Transformer architecture. To maintain manageable memory usage and computational requirements, we applied Farthest Point Sampling (FPS) [28] to downsample each raw point cloud from 8192 to 2048 points before training. Specifically, FPS is performed dynamically on the ground-truth meshes to ensure uniform spatial distribution. For semantic evaluation, since the original baseline

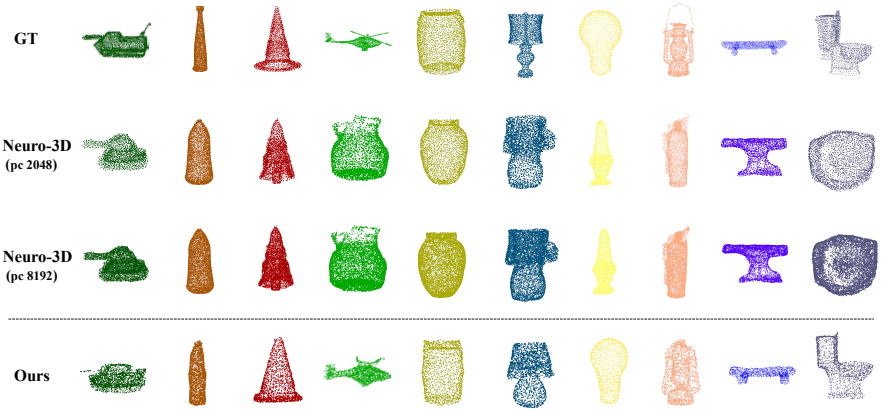
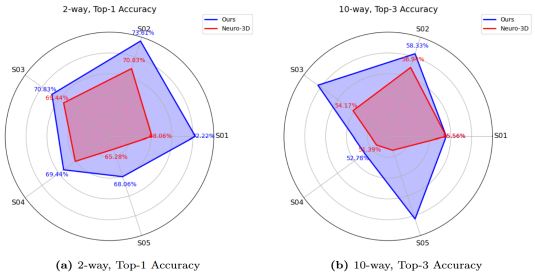


Fig. 5: Qualitative comparison of reconstructed point clouds for Subject 1. Rows show ground-truth point clouds, FPS-downsampled Neuro-3D outputs with 2048 points, original Neuro-3D outputs with 8192 points, and our 2048-point reconstructions. Despite using fewer points, Mind2Cloud better preserves object structure and local geometry.



(a) 2-way, Top-1 Accuracy **(b)** 10-way, Top-3 Accuracy

Fig. 6: Subject-wise radar plots of the 2-way/top-1 accuracy and 10-way/top-3 settings for S01–S05. Mind2Cloud consistently outperforms Neuro-3D under both evaluations.

classifier was not publicly released, we retrained our own PointNet++ classifier [35] following their evaluation protocol. To guarantee a fair evaluation, all test-set-related samples were strictly excluded during its training. This classifier was trained for 200 epochs using an Adam optimizer with a batch size of 32 and a learning rate of 1×10^{-4} to ensure stable convergence.

4.3 Results

Visualization Results. To qualitatively evaluate reconstruction performance, we visualize representative 3D point clouds from Subject 1 in Figure 5. The first row displays the ground-truth (GT) shapes, while the second and third rows show results from the Neuro-3D baseline, originally trained with 8192-point clouds. For fair comparison, we apply Farthest Point Sampling (FPS) to downsample these outputs to 2048 points, matching our input resolution. Both predicted

Datasets	Method	CD ↓	EMD ↓
EEG	Neuro-3D [13]	4.32	16.31
	Ours (w/o Discriminator)	3.26	15.93
	Ours (w/o Two-Granularity)	3.20	15.93
	Ours (w/o PVCNN)	3.71	17.10
	Ours (Fixed Fusion)	3.00	15.42
	Ours	**2.53**	**14.02**
	Ours (S01-S12)	2.65	14.43

Table 1: Comparison of reconstruction fidelity using Chamfer Distance (CD) and Earth Mover's Distance (EMD), both multiplied by $\times 10^2$. Lower values indicate better geometric accuracy.

Method	Average		Top-1 of 5 samples	
	2-way/top-1	10-way/top-3	2-way/top-1	10-way/top-3
Neuro-3D [13]	50.80%	31.72%	68.33%	53.89%
Ours (w/o Discriminator)	50.65%	31.80%	66.94%	52.22%
Ours (w/o Two-Granularity)	50.12%	31.05%	67.22%	51.95%
Ours (w/o PVCNN)	49.58%	29.58%	50.56%	37.50%
Ours (Fixed Fusion)	50.69%	32.36%	66.39%	52.22%
Ours	**52.92%**	**32.47%**	**70.83%**	**56.67%**
Ours (S01-S12)	51.68%	31.30%	70.53%	52.84%

Table 2: Semantic reconstruction accuracy under N-way Top-k classification. Results are reported for the 2-way/top-1 and 10-way/top-3 settings, including average accuracy and Top-1 accuracy over five sampled candidates. Higher values indicate better semantic preservation.

shapes and corresponding ground-truth point clouds are uniformly resampled to 2048 points before metric computation to ensure fair comparison across methods. The last row presents results from our method trained and tested directly with 2048 points. All point clouds are rendered via Open3D with automatically assigned colors. Despite the reduced point density, our approach reconstructs finer geometric details with sharper contours and fewer deformations. In particular, in complex categories such as *lightbulb*, *helicopter*, and *skateboard*, our model better preserves global structure and part connectivity. These results highlight the effectiveness of our framework in generating high-fidelity 3D shapes from EEG signals with improved efficiency.

Quantitative Evaluation. Since EEG-driven 3D point cloud reconstruction remains a relatively emerging research direction, publicly available and reproducible works are still limited. Among existing approaches, Neuro-3D is a representative and recently proposed method with publicly released models and code, making it suitable for direct and fair comparison. Therefore, we adopt Neuro-3D as our primary baseline and follow its evaluation protocol to ensure consistency and reproducibility. To quantitatively evaluate the geometric fidelity and semantic alignment of reconstructed 3D shapes, we report Chamfer Distance (CD) and Earth Mover's Distance (EMD) in Table 1, and N-way Top-k classification accuracy in Table 2. We follow Neuro-3D and report results on the same five subjects

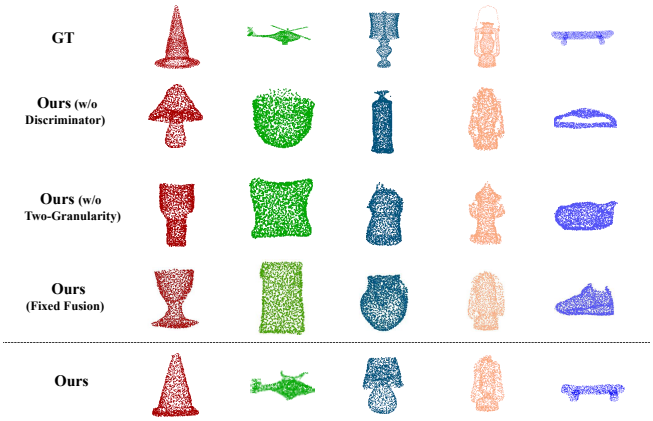


Fig. 7: Qualitative ablation comparison. From top to bottom: ground truth, Mind2Cloud without the adversarial discriminator, Mind2Cloud without the two-granularity decoder, Mind2Cloud with fixed fusion, and the full model.

to ensure direct comparability with the prior work. While the full dataset contains 12 participants, we adopt this evaluation protocol for fair benchmarking. Our method significantly outperforms the EEG-based baseline Neuro-3D on both CD and EMD, demonstrating superior reconstruction quality.

Furthermore, to demonstrate cross-subject robustness, we extend our quantitative evaluation to the remaining subjects (S06–S12) and report the overall averages across all 12 subjects (S01–S12) at the bottom of Table 1 and Table 2. The stable and consistent performance across the entire participant pool demonstrates that our framework generalizes well to unseen subjects, successfully mitigating the inherent variability of neural signals.

For semantic evaluation, we retrain a PointNet++ [35] classifier (84% accuracy) since the original Neuro-3D classifier (90%) is not publicly available. As shown in Table 2, our method surpasses Neuro-3D in both 2-way and 10-way classification, indicating better semantic preservation. Ablation studies further verify the contribution of each component. Figure 6 visualizes per-subject accuracy of 2-way/top-1 and 10-way/top-3 settings, our method consistently outperforms Neuro-3D under both evaluations.

Ablation Study. We conduct ablation studies to evaluate the contributions of key components in our framework. As reported in Table 1 and Table 2, removing the discriminator (*w/o Discriminator*) leads to increased CD/EMD scores and a drop in both average and Top-1 N-way classification accuracy, suggesting that adversarial supervision enhances both geometric fidelity and semantic alignment. To evaluate our timestep-aware design, we also test a variant with a fixed equal-weight fusion across all diffusion steps (*Fixed Fusion*). As shown in Table 1 and Table 2, the fixed fusion yields suboptimal performance, indicating that our performance gains stem from time-varying global/local feature modulation

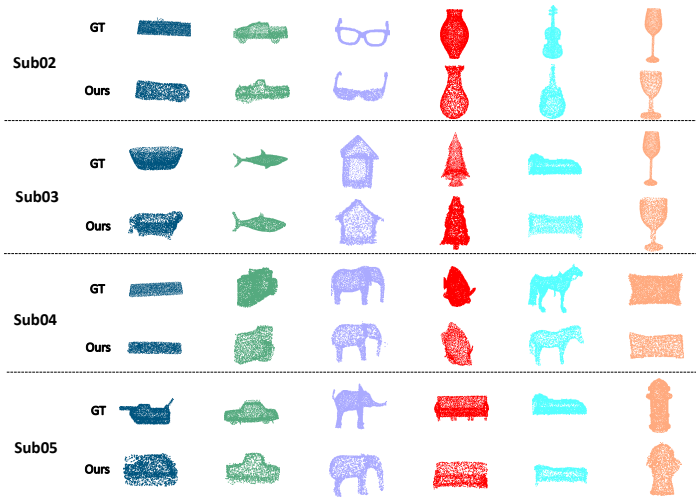


Fig. 8: Subject-wise visualization results on Subject 2 to Subject 5. For each subject, the top row shows ground-truth point clouds and the bottom row shows Mind2Cloud reconstructions generated from EEG signals.

rather than a static combination. Similarly, removing our two-granularity fusion design altogether (*w/o Two-Granularity*) results in further degradation across all metrics. Visual comparisons in Figure 7 further reveal less complete and structurally coherent shapes in both ablated variants. These results collectively confirm that both adversarial fine-tuning and the two-granularity decoder are essential for generating accurate and semantically meaningful reconstructions.

Subject-wise Visualization Results. To evaluate cross-subject robustness, we present qualitative reconstructions for four subjects (Sub02–Sub05) in Figure 8. Each example shows the ground-truth point clouds and the corresponding outputs generated solely from EEG signals. Despite subject-specific variations in neural patterns, our method produces geometrically plausible and semantically consistent shapes across subjects. The reconstructed objects preserve key structural characteristics for diverse categories such as glasses, vases, animals, and furniture. Although some fine-grained surface details remain missing due to the inherently low spatial resolution of EEG signals, the results demonstrate that our framework can reliably capture meaningful 3D structure from noisy inputs.

Subject-wise EEG Activation Mapping. To investigate inter-subject variability in cortical activity, we adopt the standard 64-channel EEG scalp partitioning into seven functional regions

Figure 9 (left). This schematic groups the electrodes into seven principal lobes: left/right frontal, central, parietal, left/right temporal, and occipital. Such an anatomical division enables region-specific interpretation of EEG patterns

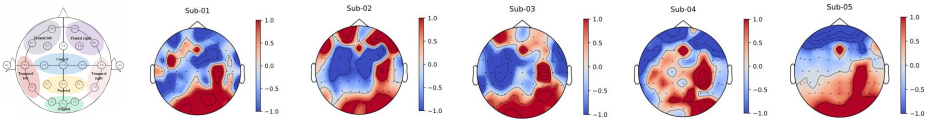


Fig. 9: Normalized EEG spatial activation maps for five subjects (Sub01–Sub05). The left schematic shows the 64-channel scalp partitioning, and each subject map shows the average potential distribution across trials and time points. Red indicates higher positive voltage, while blue indicates negative or suppressed activity.

and provides a foundation for our subsequent subject-wise analysis by supporting spatial localization of neural activation.

Figure 9 (right) illustrates normalized activation maps for five subjects, averaged over trials and time. Sub01 and Sub05 exhibit a center-suppressed, periphery-enhanced pattern, with diminished frontal-central activity and elevated occipital-temporal responses—consistent with their superior decoding performance (e.g., 72.22% Top-1 for Sub01), suggesting more structured neural representations. By contrast, Sub03 and Sub04 show localized occipital-temporal activation with weaker global suppression, while Sub02 displays polarity shifts across central-frontal regions, indicating unstable engagement and moderate accuracy. These subject-specific patterns reflect inherent variability in perceptual strategies and neural encoding, reinforcing the importance of personalized modeling in EEG-based 3D reconstruction.

5 Conclusion

We propose an end-to-end framework that reconstructs 3D shapes from EEG signals using a two-stream EEG encoder and a time-aware two-granularity diffusion decoder. Guided by a learnable fusion strategy, the model balances global and local features throughout denoising. An adversarial point cloud discriminator further enhances geometric fidelity. While achieving strong 3D reconstruction and classification results, the model does not yet reconstruct photorealistic appearance and leaves room for stronger cross-subject generalization. Future work will focus on color reconstruction via vision-language priors and improving robustness across subjects through domain adaptation and contrastive learning, advancing semantic brain decoding and practical non-invasive BCIs.

Acknowledgments

This work is supported by the Guangdong Pearl River Talent Program (No. 2023QN10X721), and the Natural Science Foundation of Guangdong Province (No. 2025A1515010454).

References

1. Bao, G., Zhang, Q., Gong, Z., Zhou, J., Fan, W., Yi, K., Naseem, U., Hu, L., Miao, D.: Wills aligner: Multi-subject collaborative brain visual decoding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 14194–14202 (2025)
2. Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., Irani, M.: From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. Advances in Neural Information Processing Systems **32** (2019)
3. Chen, Y., Zhang, C., Yang, X., Cai, Z., Yu, G., Yang, L., Lin, G.: It3d: Improved text-to-3d generation with explicit view synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1237–1244 (2024)
4. Chen, Y., Ren, K., Song, K., Wang, Y., Wang, Y., Li, D., Qiu, L.: Eegformer: Towards transferable and interpretable large-scale eeg foundation model. arXiv preprint arXiv:2401.10278 (2024)
5. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: 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. pp. 22710–22720 (2023)
6. Chen, Z., Qing, J., Zhou, J.H.: Cinematic mindscapes: High-quality video reconstruction from brain activity. Advances in Neural Information Processing Systems **36**, 24841–24858 (2023)
7. Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A.: Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. arXiv preprint arXiv:1601.02970 (2016)
8. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
9. Gao, J., Fu, Y., Fu, Y., Wang, Y., Qian, X., Feng, J.: Mind-3d++: Advancing fmri-based 3d reconstruction with high-quality textured mesh generation and a comprehensive dataset. arXiv preprint arXiv:2409.11315 (2024)
10. Gao, J., Fu, Y., Wang, Y., Qian, X., Feng, J., Fu, Y.: fmri-3d: A comprehensive dataset for enhancing fmri-based 3d reconstruction. arXiv preprint arXiv:2409.11315 **3** (2024)
11. Gao, J., Fu, Y., Wang, Y., Qian, X., Feng, J., Fu, Y.: Mind-3d: Reconstruct high-quality 3d objects in human brain. In: European Conference on Computer Vision. pp. 312–329. Springer (2024)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
13. Guo, Z., Wu, J., Song, Y., Bu, J., Mai, W., Zheng, Q., Ouyang, W., Song, C.: Neuro-3d: Towards 3d visual decoding from eeg signals. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23870–23880 (2025)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Horikawa, T., Kamitani, Y.: Generic decoding of seen and imagined objects using hierarchical visual features. Nature communications **8**(1), 15037 (2017)
16. Hu, Y., Ye, S., Zhao, W., Lin, M., He, Y., Wen, Y.H., He, Y., Liu, Y.J.: Oˆ 2-recon: completing 3d reconstruction of occluded objects in the scene with a pre-trained 2d diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2285–2293 (2024)

17. Jin, M., Du, C., He, H., Cai, T., Li, J.: Pgcn: Pyramidal graph convolutional network for eeg emotion recognition. IEEE Transactions on Multimedia **26**, 9070–9082 (2024)
18. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2426–2435 (2022)
19. Kumar, M., Dehghani, M., Houlsby, N.: Dual patchnorm. arXiv preprint arXiv:2302.01327 (2023)
20. Kupershmidt, G., Beliy, R., Gaziv, G., Irani, M.: A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. arXiv preprint arXiv:2206.03544 (2022)
21. Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., Gifford, A.T., Pan, B., Jin, S., Ratan Murty, N.A., et al.: Modeling short visual events through the bold moments video fmri dataset and metadata. Nature communications **15**(1), 6241 (2024)
22. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
23. Li, Y., Dou, Y., Shi, Y., Lei, Y., Chen, X., Zhang, Y., Zhou, P., Ni, B.: Focaldreamer: Text-driven 3d editing via focal-fusion assembly. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 3279–3287 (2024)
24. Lin, S., Sprague, T., Singh, A.K.: Mind reader: Reconstructing complex images from brain activities. Advances in Neural Information Processing Systems **35**, 29624–29636 (2022)
25. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. Advances in neural information processing systems **32** (2019)
26. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9970–9980 (2024)
27. Melas-Kyriazi, L., Rupprecht, C., Vedaldi, A.: Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12923–12932 (2023)
28. Moenning, C., Dodgson, N.A.: Fast marching farthest point sampling. Tech. rep., University of Cambridge, Computer Laboratory (2003)
29. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
30. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021)
31. Ozcelik, F., Choksi, B., Mozafari, M., Reddy, L., VanRullen, R.: Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In: 2022 international joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2022)
32. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
33. Polimeni, J., Fischl, B., Greve, D., Wald, L.: Laminar analysis of 7t bold using an imposed activation pattern in human v1. NeuroImage **52**, 1334–46 (05 2010). https://doi.org/10.1016/j.neuroimage.2010.05.005

34. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
35. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
37. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. PMLR (2021)
38. Ren, Z., Kim, M., Liu, F., Liu, X.: Tiger: Time-varying denoising model for 3d point cloud generation with diffusion process. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9462–9471 (2024)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
40. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
41. Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., et al.: Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. Advances in Neural Information Processing Systems **36**, 24705–24728 (2023)
42. Shen, G., Horikawa, T., Majima, K., Kamitani, Y.: Deep image reconstruction from human brain activity. PLoS computational biology **15**(1), e1006633 (2019)
43. Sun, J., Li, M., Chen, Z., Moens, M.F.: Neurocine: Decoding vivid video sequences from human brain activities. arXiv preprint arXiv:2402.01590 (2024)
44. Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K., et al.: Lion: Latent point diffusion models for 3d shape generation. Advances in Neural Information Processing Systems **35**, 10021–10039 (2022)
45. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
46. Willett, F.R., Avansino, D.T., Hochberg, L.R., Henderson, J.M., Shenoy, K.V.: High-performance brain-to-text communication via handwriting. Nature **593**(7858), 249–254 (2021)
47. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. In: European Conference on Computer Vision. pp. 131–147. Springer (2024)
48. Yu, Z., Li, H., Fu, F., Miao, X., Cui, B.: Accelerating text-to-image editing via cache-enabled sparse diffusion inference. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 16605–16613 (2024)
49. Zhang, K., He, L., Jiang, X., Lu, W., Wang, D., Gao, X.: Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 14486–14493 (2025)

50. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
51. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021)
52. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5826–5835 (2021)
53. Zou, Z., Cheng, W., Cao, Y.P., Huang, S.S., Shan, Y., Zhang, S.H.: Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 7900–7908 (2024)