MINDSHAPE: Superquadric-Constrained High-Fidelity 3D Reconstruction from fMRI

Xiaoquan Shen[1,2,5], **Ming Li**[1,2,4], **Jianxiong Gao**[3]
Jiaxuan Chen[1,5,†], **Xiangru Huang**[4], **Yanwei Fu**[2,3,†,*], **Gang Pan**[1,5]

[1]College of Computer Science and Technology, Zhejiang University
[2]Shanghai Innovation Institute
[3]Fudan University [4]Westlake University
[5]State Key Lab of Brain-Machine Intelligence, Zhejiang University

Abstract

Recent advances in neural decoding have enabled the reconstruction of 3D visual content from functional magnetic resonance imaging (fMRI) signals, opening a route toward brain-conditioned 3D generation. Nevertheless, existing approaches largely rely on semantic or multi-view priors without explicitly modeling 3D geometric constraints, often leading to structural distortions and degraded geometric fidelity. To address this limitation, we propose MINDSHAPE, a geometry-aware neural decoding framework for fMRI-conditioned 3D reconstruction. Unlike previous methods, MINDSHAPE decodes superquadric geometry as an explicit, low-dimensional structural constraint and jointly conditions the 3D generation process on both geometric primitives and semantic descriptions, enabling semantically consistent and geometrically faithful reconstructions. Experiments show that MINDSHAPE improves reconstruction quality in terms of geometric consistency and semantic fidelity over existing baselines. These results suggest that explicit geometric primitives provide an effective structured interface between noisy neural measurements and controllable 3D generation.

1 Introduction

Reconstructing 3D objects from human functional magnetic resonance imaging (fMRI) responses is an important problem at the intersection of visual neuroscience and generative 3D content creation. In recent years, brain decoding has made substantial progress with the help of large-scale vision-language models [41, 7, 9, 8, 10, 57] and diffusion-based generative models [42, 4]. More broadly, advances in neural decoding are increasingly enabling brain signals to guide the extraction and refinement of task-relevant perceptual representations [18]. In particular, 2D image and video reconstruction methods can already recover rich semantic information and visual appearance from fMRI [11, 6, 5, 43, 24, 47, 23, 48]. However, the human visual system does not merely process flat images; it also forms an understanding of the 3D world through spatial layout and object structure [20, 49]. Compared with 2D reconstruction, fMRI-to-3D reconstruction places greater emphasis on recovering global object shape and part layout. Furthermore, the inherent complexity of 3D visual processing makes its computational modeling directly from brain representations challenging.

Recent studies [21, 22, 56, 29] have begun to explore fMRI-driven 3D reconstruction, gradually moving from semantically recognizable geometry generation toward high-quality textured 3D asset

*Yanwei Fu is with the School of Data Science, Fudan University, and Fudan ISTBI–ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University, Jinhua, China.
†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

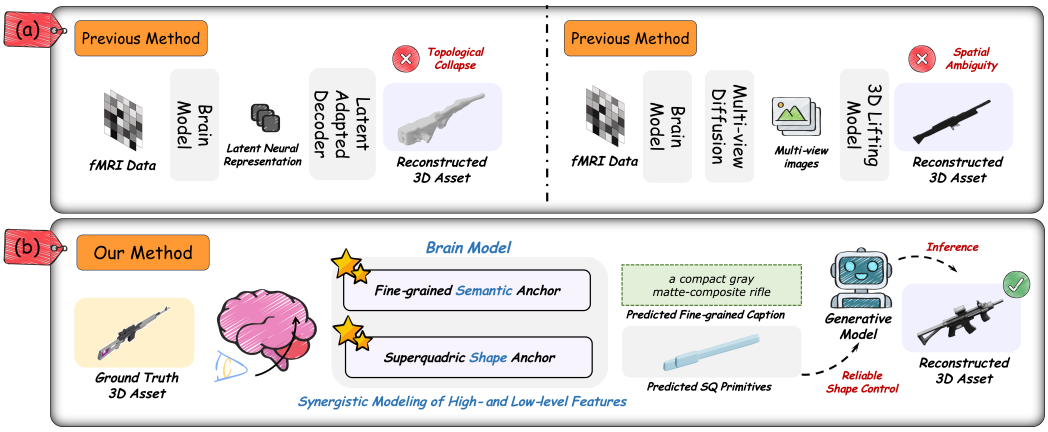

Figure 1: **Comparison of fMRI-to-3D decoding paradigms.** (a) Existing approaches lack explicit 3D constraints. Direct implicit decoding (left) may cause topological collapse, while cascaded multi-view 3D lifting (right) introduces spatial ambiguity. (b) By decoupling fMRI decoding into explicit superquadric primitives and fine-grained semantic anchors, our framework imposes robust spatial priors to resolve structural distortions for high-fidelity 3D reconstruction.

reconstruction. These methods demonstrate the feasibility of recovering 3D visual content from brain signals, and show that large-scale visual generative priors can substantially improve reconstruction quality. Unlike 2D reconstruction, fMRI-to-3D reconstruction is closely tied to how the human visual system represents object shape and spatial structure. Visual processing is known to be hierarchical [13, 36], with the brain progressively abstracting local curvature, boundaries, part composition, global shape, and high-level semantic categories rather than representing objects purely as pixels or textures [15, 25, 55, 1, 31]. Neuroscientific evidence suggests that object recognition relies on stable shape and structural representations, not merely category labels or surface texture [12, 28]. This perspective suggests that fMRI-to-3D reconstruction should not only recover semantic appearance, but also preserve an intermediate representation of object shape and part structure.

However, existing methods largely leave this mid-level geometric structure implicit. As illustrated in Fig. 1a, their geometry is typically constrained through 2D multi-view features or generative latent spaces. Under the inherently low-SNR conditions of fMRI signals, both indirect conditionings are highly brittle. Specifically, direct decoding into latent spaces frequently accumulates geometric errors, leading to structural drift such as redundant parts, missing structures, or scale deviations. On the other hand, constraining geometry through multi-view intermediate representations introduces severe spatial ambiguity, as spatial uncertainty cascades into the subsequent 3D generation phase, making low-level shape topologies difficult to preserve reliably. This bottleneck suggests that decoding 3D objects from brain activity requires a more direct, explicit geometric representation to enforce interpretable and reliable shape constraints.

Inspired by this view, we introduce superquadrics [2] (SQs) as explicit mid-level geometric proxies. SQs describe spheres, box-like shapes, cylinders, and their continuous variants with only a few parameters, and can approximate complex object structures through combinations of primitives. Compared with reconstructed images, or high-dimensional implicit latents, SQs are low-dimensional, interpretable, compositional, and able to directly encode spatial constraints such as volume, scale, pose, and part layout.

Based on these observations, we propose MINDSHAPE, a decoupled fMRI-to-3D reconstruction framework (Fig. 1b). MINDSHAPE first aligns fMRI signals with a visual-text latent space via contrastive learning to establish a unified brain representation. From this shared foundation, the framework decouples into two synergistic pathways, utilizing a geometric branch to decode explicit superquadric primitives for constraining macroscopic topology, alongside a semantic branch that decodes fine-grained captions to govern appearance attributes. By jointly conditioning a pretrained 3D generative prior on both SQ proxies and textual anchors, MINDSHAPE effectively mitigates latent structural distortions, yielding 3D assets with improved geometric consistency and semantic fidelity.

We evaluate MINDSHAPE on the public fMRI-Shape dataset [21], where it outperforms the strong MinD-3D++ baseline [22] and achieves the best reported performance under the same evaluation protocol across semantic and geometric metrics. Our contributions are summarized as follows:

- **Explicit SQ geometric constraints.** We introduce superquadrics into fMRI-to-3D reconstruction as compact, interpretable, and controllable mid-level geometry, enabling the decoding process to explicitly constrain object shape, scale, and part layout rather than relying only on semantic or image-level intermediate features.
- **Fine-grained captions as semantic anchors.** We construct structured fine-grained captions that organize category, color, material, and shape attributes into stable semantic anchors, providing lightweight yet effective semantic conditions for downstream 3D generation.
- **An integrated geometry-semantic 3D generation paradigm.** We propose MINDSHAPE, a framework that jointly learns geometry decoding and semantic decoding during training, and conditions a pretrained 3D generative model on decoded SQs and semantic descriptions during inference, achieving improved semantic and geometric performance on fMRI-to-3D reconstruction.

2 Related Work

Brain-to-3D Reconstruction. While neural decoding has achieved remarkable success [24, 43, 34, 45] in reconstructing 2D images and videos, extending this capability to 3D visual representations remains a challenging frontier. Unlike 2D visual reconstruction, fMRI-to-3D reconstruction requires recovering 3D structure, shape semantics, and textured appearance from neural responses. MinD-3D [21] first constructed the fMRI-Shape dataset and combined fMRI decoding with 3D generative models, demonstrating the feasibility of recovering 3D object geometry from brain signals. MinD-3D++ [22] further uses brain-decoded multi-view images as intermediate representations to guide the generation of high-quality textured 3D meshes. Beyond fMRI, Neuro-3D [26] reconstructs 3D point clouds from EEG and augments them with a coarse object-level dominant color, with its appearance feature supervised by color-category labels. Its EEG-derived geometry and appearance representations remain implicit. In contrast, MINDSHAPE decodes fMRI into an explicit, parameterized set of SQ primitives that can be directly visualized, evaluated at the surface level, and used as spatial control, while an independently decoded structured caption conditions the frozen generator on semantics and appearance.

Controllable 3D Generation. Native 3D generative models have recently achieved substantial progress, enabling the rapid generation of high-quality textured meshes [27, 51, 50, 54, 53]. However, these models typically rely on text or images as prompts, which often fall short of precisely representing the geometric details of 3D objects. Consequently, several studies [16, 14, 44] have explored the use of more granular, 3D-aware conditions to guide the generative process. Inspired by these works, we design a coarse-to-fine fMRI-to-3D pipeline that first retrieves coarse 3D control signals and captions, then leverages a 3D generative model to produce geometrically refined and semantically consistent results. Specifically, as SpaceControl [16] is training-free and offers an effective trade-off between geometric faithfulness and visual realism, we adopt superquadrics [2] as our intermediate representation and follow a similar methodology to inject these geometric constraints into the Trellis [51] generation process.

3 Method

Preliminaries: Superquadrics. Superquadrics [2] (SQs) provide a compact parametric family for representing volumetric shapes, including spheres, ellipsoids, cylinders, box-like objects, and continuous variants between them. In a normalized object coordinate system, an axis-aligned SQ centered at the origin is defined by

$$\left(\left| \frac{x}{s_x} \right|^{\frac{2}{\epsilon_2}} + \left| \frac{y}{s_y} \right|^{\frac{2}{\epsilon_2}} \right)^{\frac{\epsilon_2}{\epsilon_1}} + \left| \frac{z}{s_z} \right|^{\frac{2}{\epsilon_1}} = 1, \tag{1}$$

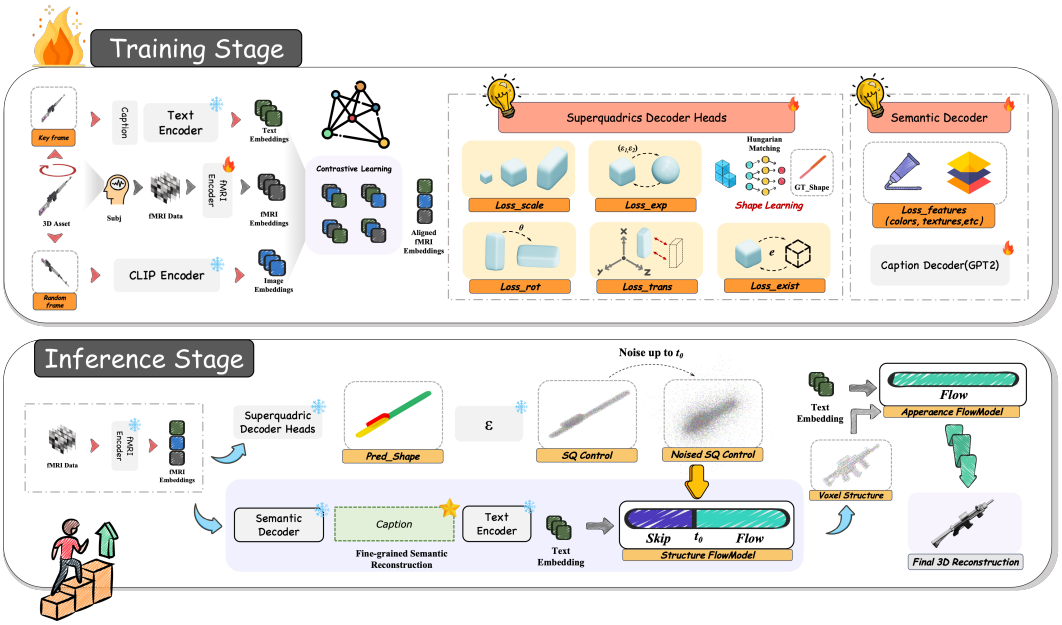

Figure 2: **Overview of MINDSHAPE framework.** (i) During training, fMRI responses are encoded into shared brain representations and decoded into two complementary conditions: explicit SQ geometry and fine-grained semantic descriptions (Secs. 3.3 and 3.4). (ii) During inference, the decoded SQ geometry and semantic caption jointly condition the pretrained 3D generation pipeline to produce the final reconstruction (Sec. 3.5).

where $\mathbf{s} = (s_x, s_y, s_z) \in \mathbb{R}^3_+$ denotes the scale along the three principal axes, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2) \in (0, 2)^2$ controls the smoothness and boxiness of the shape. A posed SQ primitive is represented as

$$y_k = (e_k, \boldsymbol{\epsilon}_k, \mathbf{s}_k, \mathbf{t}_k, \mathbf{R}_k), \tag{2}$$

where e_k is the existence indicator, $\mathbf{t}_k \in \mathbb{R}^3$ is the translation in the normalized object coordinate system, and $\mathbf{R}_k \in SO(3)$ is a valid 3D rotation. A 3D object can then be approximated by a set of SQ primitives $\mathcal{Y} = \{y_k\}_{k=1}^K$, yielding a low-dimensional and interpretable geometry representation for global shape and part layout.

3.1 MINDSHAPE Overview

We introduce MINDSHAPE, a framework for reconstructing semantically accurate and shape-faithful 3D objects from human cortical fMRI responses. Instead of directly regressing an implicit 3D representation from brain activity, MINDSHAPE decomposes fMRI-conditioned 3D reconstruction into two control signals used at inference time: an explicit geometric structure represented by superquadrics and a fine-grained semantic description decoded from the same brain response. The former provides an interpretable spatial constraint, while the latter supplies category, appearance, and local attribute guidance. Through this geometry-semantic decomposition, MINDSHAPE converts noisy and low-SNR neural measurements into more stable conditions for 3D generation.

The overall framework is illustrated in Fig. 2. It consists of three components: fMRI representation encoding (Sec. 3.3), geometry-semantic decoding (Sec. 3.4), and SQ/text-conditioned 3D generation at inference time (Sec. 3.5).

3.2 Geometry-Semantic Condition Construction

MINDSHAPE is built around two intermediate targets: SQ geometry for constraining 3D structure, and fine-grained captions for specifying category and appearance. These two targets correspond to the spatial control signal and the textual condition used by the final 3D generation stage.

Superquadric Geometry Construction. Based on the SQ representation introduced in the preliminaries, a set of SQ primitives serves as an explicit spatial condition for downstream 3D generation.

Compared with pure semantic prompts or 2D intermediate representations, SQs directly constrain object scale, position, and orientation in 3D space, thereby reducing geometric drift, structural collapse, and within-category shape homogenization. SQs are used as a primitive-based geometric prior rather than a replacement for the final 3D asset; fine texture and local geometric details are completed by the downstream generative model.

We construct SQ-GT offline on the stimulus side. For each 3D object stimulus, we first sample a point cloud from its mesh or equivalent 3D representation and normalize it into a unified coordinate system. Following the idea of SUPERDEC-style [17] object decomposition, the input point cloud is decomposed into a compact set of SQ primitives. Such methods predict multiple SQ parameters and their existence probabilities, and constrain the decomposition with bidirectional geometric discrepancies between the point cloud and sampled SQ surfaces; lightweight optimization can be further used to refine primitive parameters. The resulting decomposition is organized into a fixed-size set of at most K SQ labels, including existence, shape exponents, scales, translations, and rotations. Inactive primitive slots are marked as non-existing. Thus, SQ-GT is a geometry-derived pseudo-label rather than a manual semantic annotation, and is used to supervise the decoding of explicit geometry from fMRI (details are illustrated in Appendix A).

Fine-grained Caption Generation. In addition to explicit geometry, the final 3D generation stage requires a textual condition that describes the category, appearance, and local attributes. Category names alone are stable but often insufficient to distinguish intra-class variations, such as differences in shape, color, or material among rifles, chairs, or vehicles. We therefore construct one structured fine-grained caption for each stimulus using the template

$$c = \text{“a <shape> <color> <material-texture> <category>”}. \tag{3}$$

The `shape` slot describes the main geometric form or local structure, `color` is selected from a controlled color vocabulary, `material-texture` specifies the material and surface appearance, and `category` uses the canonical name of the corresponding ShapeNet-13 category [3].

This templated caption provides richer instance-level semantics than a category name while maintaining a fixed word order and a constrained vocabulary, making the target distribution more learnable for the GPT-2 [40] TextDecoder. Caption-GT is also constructed offline before model training (introduced in Appendix A). We feed a key frame of each 3D object into the Qwen3-VL-Plus vision-language model and prompt it to generate a fine-grained caption following the template above. A lightweight rule-based validator then checks whether the generated caption satisfies the prescribed format. The resulting structured caption forms the fine-grained semantic ground truth used in this work.

3.3 MINDSHAPE Brain Representation Encoding

This section describes the brain encoding model in MINDSHAPE. Its role is to encode cortical fMRI responses into conditional representations that can be consumed by the subsequent geometry and semantic decoders. Since fMRI signals are noisy, high-dimensional, and subject to strong inter-subject variability, learning such an encoder from scratch on limited 3D stimulus data can lead to unstable representations. We therefore initialize the backbone with a pretrained fMRI encoder from LEA [39] and further adapt it to our 3D object stimulus data.

The overall model takes fMRI responses as the neural input for both geometry and semantic decoding. Let $\mathbf{X} \in \mathbb{R}^{T\times H\times W}$ denote the multi-frame fMRI response sequence elicited by a 3D object stimulus, where T is the number of sampled fMRI frames and $H \times W$ is the spatial resolution of each preprocessed fMRI map. The encoder maps $\mathbf{X}$ into a CLIP-aligned visual-token space:

$$\mathbf{Z}_v = E_\theta(\mathbf{X}). \tag{4}$$

Here, $\mathbf{Z}_v$ denotes the brain-derived visual condition representation. We further apply a semantic projection head to map $\mathbf{Z}_v$ into the CLIP text space:

$$\mathbf{z}_t = \phi_{\text{text}}(\mathbf{Z}_v). \tag{5}$$

During training, CLIP image features from randomly sampled image frames and text features are used to align these brain-derived representations, providing a stable cross-modal semantic basis. This design is consistent with recent language-anchored multimodal vision–brain alignment studies that emphasize robust semantic correspondence across modalities [19]. The encoding stage

adopts MixCo [30] data augmentation with bidirectional contrastive learning to improve cross-modal alignment under limited fMRI samples (detailed loss formulation is provided in the Appendix B). In the main text, we summarize the objective as

$$\mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{CLIP}_v} + \lambda_t \mathcal{L}_{\text{CLIP}_t}. \tag{6}$$

Here, $\mathcal{L}_{\text{CLIP}_v}$ aligns fMRI embeddings with CLIP image embeddings from randomly sampled image frames, while $\mathcal{L}_{\text{CLIP}_t}$ aligns the projected fMRI semantic representation with CLIP text embeddings.

3.4 MINDSHAPE Geometry-Semantic Decoding

Given the shared brain-conditioned representations, MINDSHAPE uses two decoders to predict geometry and semantics separately: an SQ decoder for explicit primitive sets, and a TextDecoder for structured fine-grained captions. To reduce interference across heterogeneous supervision signals, the decoding stage keeps the fMRI encoder, dimensional adapter, and semantic projection head fixed by default, preserving the category-level semantics learned during encoding, while the SQ decoder and GPT-2 TextDecoder are optimized.

Superquadric geometry decoding. The geometry branch takes the brain-derived visual representation and semantic condition as input, and predicts an unordered set of K SQ primitives using a Transformer set decoder. Each prediction slot corresponds to a candidate primitive and outputs its existence, shape, scale, translation, and rotation parameters. Since the primitive set is unordered, we use Hungarian matching to establish the optimal assignment between active predicted primitives and ground-truth primitives.

Let $\hat{\mathcal{Y}} = \{\hat{y}_i\}_{i=1}^{K}$ denote the predicted primitive slots and $\mathcal{Y} = \{y_j\}_{j=1}^{M}$ denote active ground-truth primitives, where $M \leq K$. We compute the pairwise matching cost between prediction i and target j as

$$\mathcal{C}_{ij} = \alpha_{\text{exist}} \mathcal{C}_{ij}^{\text{exist}} + \alpha_{\epsilon} \mathcal{C}_{ij}^{\epsilon} + \alpha_{\text{scale}} \mathcal{C}_{ij}^{\text{scale}} + \alpha_{\text{trans}} \mathcal{C}_{ij}^{\text{trans}} + \alpha_{\text{rot}} \mathcal{C}_{ij}^{\text{rot}}, \tag{7}$$

where the terms respectively measure existence confidence, exponent discrepancy, scale discrepancy, translation discrepancy, and rotation disagreement. The optimal assignment is obtained by solving

$$\sigma^{\star} = \arg \min_{\sigma \in \mathfrak{S}_{K,M}} \sum_{j=1}^{M} \mathcal{C}_{\sigma(j),j}, \tag{8}$$

where $\mathfrak{S}_{K,M}$ is the set of one-to-one assignments from the M active ground-truth primitives to K prediction slots. Given $\sigma^{\star}$, the matched SQ loss is

$$\mathcal{L}_{\text{sq}} = \lambda_{\text{exist}} \mathcal{L}_{\text{exist}} + \lambda_{\epsilon} \mathcal{L}_{\epsilon} + \lambda_{\text{scale}} \mathcal{L}_{\text{scale}} + \lambda_{\text{trans}} \mathcal{L}_{\text{trans}} + \lambda_{\text{rot}} \mathcal{L}_{\text{rot}}. \tag{9}$$

The existence term supervises whether a primitive is active, the exponent term constrains SQ shape, scale and translation constrain size and position, and the rotation term constrains pose. We compute rotation discrepancy with a matrix-cosine form [58] based on a continuous 6D rotation representation:

$$\mathcal{L}_{\text{rot}} = 1 - \cos\theta, \qquad \cos\theta = \frac{\text{tr}(\mathbf{R}_{\text{pred}}^{\top} \mathbf{R}_{\text{gt}}) - 1}{2}. \tag{10}$$

This avoids the gradient instability of directly optimizing an $\arccos(\cdot)$ angular loss.

Fine-grained semantic decoding. The semantic branch adopts GPT-2 prefix decoding to generate fine-grained structured captions from the brain-derived semantic representation. Specifically, the semantic embedding $\mathbf{z}_t$ is first mapped by a small MLP into a sequence of GPT-2 prefix tokens:

$$\mathbf{P} = \psi(\mathbf{z}_t). \tag{11}$$

GPT-2 then autoregressively generates the fine-grained caption under this prefix condition, trained with the standard negative log-likelihood objective:

$$\mathcal{L}_{\text{cap}} = -\frac{1}{T} \sum_{j=1}^{T} \log p_{\eta}(w_j \mid w_{<j}, \mathbf{P}). \tag{12}$$

Overall, the second stage is decoder-centric: the geometry branch and semantic caption branch learn their respective targets on top of the shared brain representation, reducing interference between heterogeneous supervision signals.

3.5 Inference Flow

At inference time, the input is the test fMRI response, and the output is a textured 3D asset. The fMRI response is first passed through the trained brain encoding model to obtain a representation for conditional decoding. The geometry branch predicts a set of SQ primitives, while the semantic branch generates a fine-grained textual description c. The final 3D synthesis stage relies only on these two decoded conditions: predicted SQ geometry and a fine-grained semantic description.

Inspired by the training-free test-time spatial control paradigm of SpaceControl [16], we inject the brain-decoded SQ geometry as an explicit spatial condition into the structure generation process of a pretrained 3D generative model. The predicted SQ set is converted into a spatial control signal and mapped into the structure latent space (details are illustrated in Appendix D); the fine-grained text description is encoded as a semantic condition that guides the denoising process of the structure flow model. After the resulting structure latent is decoded into a 3D occupancy structure, the generative model further completes local appearance features through an appearance flow model, yielding a renderable 3D asset.

This inference mechanism explicitly integrates brain-decoded spatial control and semantic guidance within a unified 3D generation process. SQ geometry provides a faithfulness constraint on global shape, scale, and part layout, while the fine-grained description provides semantic guidance for category, color, material, and shape attributes. Together, they regulate the realism-faithfulness trade-off and allow MINDSHAPE to transform fMRI responses into high-fidelity 3D reconstructions with both perceptual semantic consistency and geometric reliability.

4 Experiments

Dataset. We evaluate MINDSHAPE on the core set of fMRI-Shape [22] and use the publicly released preprocessed fMRI data from this benchmark. The core set contains fMRI responses from eight participants viewing 3D object stimuli, together with the corresponding 3D object ground truth for each stimulus. The stimulus objects are drawn from 13 commonly used categories in ShapeNetCore [3], with 100 training objects and 8 test objects per category. Each object is rendered as an 8-second 360-degree rotating video, allowing participants to observe the same 3D object from continuous viewpoints and eliciting cortical responses related to object category and structure.

In our experiments, we follow the standard training and testing protocol of fMRI-Shape and use the official preprocessing pipeline [22], which converts the functional responses into unified 256×256 2D fMRI maps as model inputs. For each test sample, MINDSHAPE first decodes SQ geometry and a fine-grained semantic description from fMRI, and then conditions a pretrained 3D generative model on these two signals to obtain the final textured 3D reconstruction. All methods are compared under the same test stimuli and evaluation protocol to ensure a fair comparison.

Evaluation Metrics. We evaluate reconstruction quality from semantic consistency and geometric fidelity. At the semantic level, following standard practice in prior fMRI-based visual reconstruction studies, we use N-way top-K accuracy to assess the category discriminability of reconstructed objects and report 2-way top-1 accuracy and 10-way top-1 accuracy. These metrics compare reconstructed results with the ground-truth stimulus in a semantic feature space and measure whether the model preserves category-level information consistent with the original stimulus.

At the geometric level, we measure structural fidelity with surface-voxel IoU at 32^3 resolution, Chamfer Distance (CD), Earth Mover's Distance (EMD), and F-score at thresholds $\tau \in \{0.02,\ 0.05\}$ [33, 52, 46]. IoU compares the 32^3 surface-voxel occupancies of the two meshes to summarise volumetric agreement; CD and EMD quantify per-instance discrepancy on $2{,}048$ surface points via nearest-neighbor distance and optimal transport; F-score reports the harmonic mean of precision and recall within τ as a threshold-based coverage measure. For readability, IoU, CD and EMD are reported after multiplying by 100.

Comparison with State-of-the-Art. We compare MINDSHAPE with existing fMRI-to-3D reconstruction methods (implementation details are introduced in Appendix E). Table 1 reports quantitative results from semantic and geometric perspectives, while Fig. 3 presents representative qualitative examples. All methods are evaluated on the same test set under the same protocol and metrics are averaged over eight subjects. We select several previous methods and the current state-of-the-art method MinD-3D++ [22] as the main baselines.

Table 1: **Quantitative comparison.** ↑ indicates higher is better, and ↓ indicates lower is better. The best results are shown in **bold**, and the second-best results are underlined. (See Appendix C for individual subject results.)

Method	Semantic-level		Geometry-level				
	Acc@2 ↑	Acc@10 ↑	IoU ↑	CD ↓	EMD ↓	F@0.02 ↑	F@0.05 ↑
LEA-3D [37]	0.767	0.361	6.804	7.011	8.596	0.122	0.293
fMRI-PTE-3D [38]	0.809	0.407	7.235	6.218	7.233	0.139	0.310
MinD-3D [21]	0.829	0.436	7.705	5.276	6.191	0.152	0.341
MinD-3D++ [22]	0.872	0.591	8.114	4.886	5.736	0.160	0.363
MINDSHAPE	**0.901**	**0.643**	**8.511**	**4.485**	**5.388**	**0.167**	**0.378**

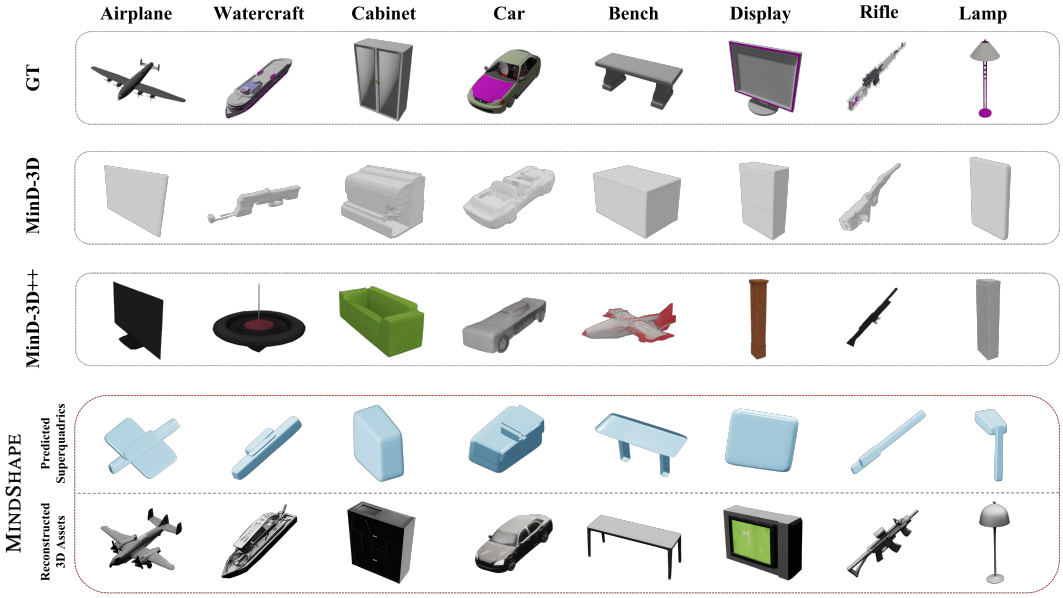

Figure 3: **Qualitative comparison.** The upper rows show the ground truth and reconstructions from existing methods, while the lower rows show the SQ geometry decoded by MINDSHAPE and its final 3D reconstruction. (See Appendix C for more results.)

Quantitative Analysis. Table 1 compares MINDSHAPE against previous methods across semantic and geometric dimensions. MINDSHAPE attains the best Acc@2 and Acc@10, surpassing MinD-3D++ by 3.3% (from 0.872 to 0.901) and 8.8% (from 0.591 to 0.643), confirming that the fine-grained semantic branch yields highly discriminative priors that steer the downstream 3D reconstruction. Beyond semantics, MINDSHAPE improves over prior fMRI-to-3D methods across all reported spatial metrics. As IoU rises from 8.114 to 8.511 (+4.9%), CD and EMD drop by 8.2% and 6.1% relative to MinD-3D++, and F@0.02 / F@0.05 rise from 0.160 / 0.363 to 0.167 / 0.378. Joint gains on occupancy-, distance- and threshold-based criteria indicate broader geometric coverage rather than improvement on a few easy samples, supporting the value of decoding SQ geometry constraint.

Qualitative Analysis. Fig. 3 shows representative reconstruction results across multiple categories. MinD-3D can generate coarse shapes related to the target category, but because its geometry is mainly governed by an implicit latent constraint, the results often exhibit structural omissions, geometric redundancy, or abnormal local structures. For example, in the watercraft, car, and rifle examples, the reconstructions show missing contours or redundant local geometry, indicating that latent constraints do not stably encode the scale, position, and compositional relationships of object parts.

MinD-3D++ first decodes multi-view images and then generates a 3D representation from these views. This strategy improves texture appearance, but the final 3D structure is still affected by image-level

semantic errors and cross-view inconsistencies. In Fig. 3, the airplane and watercraft examples show severe semantic misidentification, leading to generated objects that deviate from the ground-truth categories. Even for car and rifle samples where the category is broadly correct, inconsistent appearance and structural information across views lead to rough surfaces, local distortions, or structural discontinuities in the final 3D results.

In contrast, MINDSHAPE decomposes fMRI decoding into explicit SQ geometry and fine-grained semantic descriptions, which respectively constrain 3D structure and appearance semantics. The decoded SQ geometry in Fig. 3 captures the wing-body layout of the airplane, the elongated hull of the watercraft, the slender body of the rifle, and the box-like structure of the cabinet. With these geometric constraints, the final reconstructions preserve category semantics while exhibiting more stable global contours and part layouts. This observation is consistent with our motivation in the introduction, namely that image-level intermediates or implicit latent priors alone are insufficient to reliably carry 3D shape information, whereas superquadrics provide an interpretable and compositional geometric representation that supplies more direct structural constraints for brain-driven 3D generation. Meanwhile, fine-grained semantic descriptions improve texture controllability, and the combination of the two signals enhances the overall quality of 3D object reconstruction.

5 Ablation Studies

In this section, we compare MINDSHAPE against three baselines: a base model lacking both conditions, a semantic-only variant, and a geometry-only variant (Table 2 and Fig. 4).

Quantitative Analysis. Table 2 shows a clear separation of roles. Adding SQ geometry substantially improves spatial accuracy, reducing CD from 5.223 to 4.579, which confirms that explicit geometric constraints are effective. However, the geometry-only variant also lowers Acc@2 to 0.862, suggesting that over-emphasizing shape regularity may hurt semantic reconstruction when appearance cues are absent. Conversely, the semantic-only variant improves semantic accuracy but provides weaker geometric fidelity than the geometry-only model. The full model raises Acc@10 from 0.650 to 0.660 compared with the semantic-only variant while retaining most of the geometric gain introduced by SQ. This balance is more meaningful as the final goal is to reconstruct 3D assets that are both recognizable and structurally faithful.

Table 2: **Ablation study on semantic and geometric conditions.** Sem.: Fine-grained Semantic branch; Geo.: Geometric primitives.

Method	Sem.	Geo.	Acc@2↑	Acc@10↑	CD↓	EMD↓
Base	✗	✗	0.870	0.571	5.223	5.743
+Geo.(only)	✗	✓	0.862	0.560	**4.579**	**5.402**
+Sem.(only)	✓	✗	**0.907**	0.650	4.824	5.612
MINDSHAPE	✓	✓	**0.907**	**0.660**	4.639	5.511

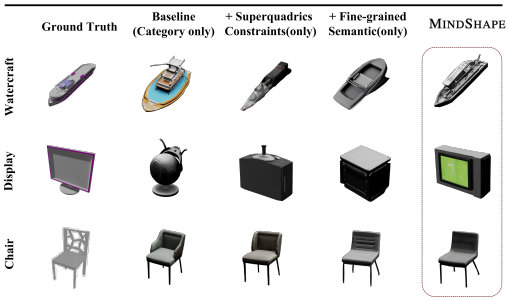

Figure 4: **Qualitative ablation study on** MINDSHAPE. Isolating the effects of semantic guidance and SQ geometric conditions.

Qualitative Analysis. Visual comparisons in Fig. 4 corroborate these quantitative findings. The base model synthesizes generic, unstable shapes. Relying solely on SQ geometry (Col. 3) stabilizes the global structure but yields featureless, semantically ambiguous assets (e.g., the simplified watercraft lacking discriminative appearance cues). Adding only semantic anchors (Col. 4) recovers textures (e.g., the screen of the display) but fails to resolve noticeable shape distortions. By harmonizing both pathways, MINDSHAPE (Col. 5) faithfully reconstructs assets that adhere to both complex geometric topologies (e.g., the thin frame of the chair) and specific categorical details, demonstrating the indispensable synergy of our decoupled framework.

6 Conclusion

We presented MINDSHAPE, a geometry-semantic co-decoding framework for high-fidelity 3D reconstruction from fMRI signals. The framework decomposes brain decoding into two complementary conditions: explicit SQ geometry, which provides spatial shape constraints, and fine-grained semantic descriptions, which supplement texture-related attributes such as color and material. By jointly inject-

ing these two conditions into a pretrained 3D generative model, MINDSHAPE improves geometric reliability in global shape and part layout while preserving semantic consistency. Experimental results show that MINDSHAPE outperforms existing methods on fMRI-to-3D reconstruction, offering a new design paradigm for brain-driven 3D generative models with explicit geometric control.

Limitations and future work. MINDSHAPE is evaluated on object-centric ShapeNet stimuli from fMRI-Shape, and its generalization to natural scenes, complex interactions, and unseen categories remains to be studied. SQ primitives provide compact global shape constraints but may miss highly non-rigid or fine-grained structures, while reconstruction quality still depends on the reliability of decoded geometry and captions under noisy, subject-variable fMRI signals.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 624B2127), and by the Postdoctoral Fellowship Program and the China Postdoctoral Science Foundation (Grant No. BX2026288).

References

[1] Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

[2] Alan H Barr. Superquadrics and angle-preserving transformations. *IEEE Computer graphics and Applications*, 1(1):11–23, 1981.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022.

[5] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the semantic latent space between brain and machine: Similarity is all you need. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11302–11310, 2024.

[6] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mind artist: Creating artistic snapshots with human thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27207–27217, 2024.

[7] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the gap between brain and machine in interpreting visual semantics: Towards self-adaptive brain-to-text decoding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21938–21948. IEEE, 2025.

[8] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings. *IEEE Transactions on Image Processing*, 34:3281–3293, 2025.

[9] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Human-like cognitive generalization for large models via mental representation-guided supervision. *Nature Communications*, 2026.

[10] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Reducing semantic mismatch in brain-to-text decoding through personalized multimodal masking. In *International Conference on Learning Representations*, volume 2026, pages 68777–68793, 2026.

[11] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023.

[12] David D Coggan, David M Watson, Ao Wang, Robert Brownbridge, Christopher Ellis, Kathryn Jones, Charlotte Kilroy, and Timothy J Andrews. The representation of shape and texture in category-selective regions of ventral-temporal cortex. *European Journal of Neuroscience*, 56 (3):4107–4120, 2022.

[13] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[14] Wenqi Dong, Bangbang Yang, Lin Ma, Xiao Liu, Liyuan Cui, Hujun Bao, Yuewen Ma, and Zhaopeng Cui. Coin3d: Controllable and interactive 3d assets generation with proxy-guided conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024.

[15] James H Elder. Shape from contour: Computation and representation. *Annual review of vision science*, 4(1):423–450, 2018.

[16] Elisabetta Fedele, Francis Engelmann, Ian Huang, Or Litany, Marc Pollefeys, and Leonidas Guibas. Spacecontrol: Introducing test-time spatial control to 3d generative modeling. *arXiv preprint arXiv:2512.05343*, 2025.

[17] Elisabetta Fedele, Boyang Sun, Leonidas Guibas, Marc Pollefeys, and Francis Engelmann. Superdec: 3d scene decomposition with superquadrics primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24625–24635, 2025.

[18] Zehui Feng, Yangge Yang, Chenqi Zhang, Junxuan Li, and Ting Han. A novel hierarchical attention-guided refinement method with eeg assistance for enhancing target speech in a multi-speaker competing environment. *Advanced Engineering Informatics*, 65:103363, 2025.

[19] Zehui Feng, Chenqi Zhang, Mingru Wang, Minuo Wei, Shiwei Cheng, Cuntai Guan, and Ting Han. Unveiling deep semantic uncertainty perception for language-anchored multi-modal vision-brain alignment. *arXiv preprint: 2511.04078*, 2025.

[20] Nonie J Finlayson, Xiaoli Zhang, and Julie D Golomb. Differential patterns of 2d location versus depth decoding along the visual hierarchy. *NeuroImage*, 147:507–516, 2017.

[21] Jianxiong Gao, Yuqian Fu, Yun Wang, Xuelin Qian, Jianfeng Feng, and Yanwei Fu. Mind-3d: Reconstruct high-quality 3d objects in human brain. In *European Conference on Computer Vision*, pages 312–329. Springer, 2024.

[22] Jianxiong Gao, Yanwei Fu, Yuqian Fu, Yun Wang, Xuelin Qian, and Jianfeng Feng. Mind-3d++: Advancing fmri-based 3d reconstruction with high-quality textured mesh generation and a comprehensive dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[23] Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. Cinebrain: A large-scale multi-modal audiovisual brain dataset for brain-conditioned video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 36224–36234, 2026.

[24] Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, et al. Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. *Advances in Neural Information Processing Systems*, 37: 51655–51683, 2024.

[25] Kalanit Grill-Spector, Zoe Kourtzi, and Nancy Kanwisher. The lateral occipital complex and its role in object recognition. *Vision research*, 41(10-11):1409–1422, 2001.

[26] Zhanqiang Guo, Jiamin Wu, Yonghao Song, Weijian Mai, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Neuro-3d: Towards 3d visual decoding from eeg signals. *arXiv preprint arXiv:2411.12248*, 2024.

[27] Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025.

[28] Akshay V Jagadeesh and Justin L Gardner. Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, 119(17):e2115302119, 2022.

[29] Haodong Jing, Dongyao Jiang, Jixin Wang, Junhao Jia, Yanshu Li, Yongqiang Ma, and Nanning Zheng. More natural, more real: Object-aware gaussian splatting for 3d visual decoding from human brain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19033–19044, June 2026.

[30] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. *arXiv preprint arXiv:2010.06300*, 2020.

[31] Zoe Kourtzi, Michael Erb, Wolfgang Grodd, and Heinrich H Bülthoff. Representation of the perceived 3-d object shape in the human lateral occipital complex. *Cerebral cortex*, 13(9): 911–920, 2003.

[32] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

[33] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2022.

[34] Ruijie Lu, Yixin Chen, Junfeng Ni, Baoxiong Jia, Yu Liu, Diwen Wan, Gang Zeng, and Siyuan Huang. Movis: Enhancing multi-object novel view synthesis for indoor scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26767–26778, 2025.

[35] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[36] Mujeeb Ur Rehman Parrey. From cats to the cortex: Unravelling the hierarchical processing system of vision and brain plasticity. *Cureus*, 16(9), 2024.

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

[38] Xuelin Qian, Yun Wang, Jingyang Huo, Jianfeng Feng, and Yanwei Fu. fmri-pte: A large-scale fmri pretrained transformer encoder for multi-subject brain activity decoding. *arXiv preprint arXiv:2311.00342*, 2023.

[39] Xuelin Qian, Yikai Wang, Xinwei Sun, Yanwei Fu, Xiangyang Xue, and Jianfeng Feng. Lea: Learning latent embedding alignment model for fmri decoding and encoding. *Transactions on Machine Learning Research*, 2024.

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

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[43] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

[44] Etai Sella, Gal Fiebelman, Noam Atia, and Hadar Averbuch-Elor. Spice· e: Structural priors in 3d diffusion using cross-entity attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[45] Xiaoquan Shen, Jiajun Li, Jiaxuan Chen, and Gang Pan. Cortexvideo: A semantic-spatial dual-anchor framework for high-fidelity fmri-to-video reconstruction. In *European Conference on Computer Vision*, pages 676–693. Springer, 2026.

[46] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3405–3414, 2019.

[47] Haonan Wang, Qixiang Zhang, Lehan Wang, Xuanqi Huang, and Xiaomeng Li. Neurons: Emulating the human visual cortex improves fidelity and interpretability in fmri-to-video reconstruction. *arXiv preprint arXiv:2503.11167*, 2025.

[48] Yujie Wei, Chenglong Ma, Jianxiong Gao, Chenhui Wang, Shiwei Zhang, Biao Gong, Shuai Tan, Hangjie Yuan, and Hongming Shan. Bridging brain and semantics: A hierarchical framework for semantically enhanced fmri-to-video reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28211–28223, June 2026.

[49] Puti Wen, Michael S Landy, and Bas Rokers. Identifying cortical areas that underlie the transformation from 2d retinal to 3d head-centric motion signals. *Neuroimage*, 270:119909, 2023.

[50] Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025.

[51] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21469–21480, 2025.

[52] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems*, 32, 2019.

[53] Chongjie Ye, Cheng Cao, Chuanyu Pan, Yiming Hao, Yihao Zhi, Yuanming Hu, and Xiaoguang Han. Omni123: Exploring 3d native foundation models with limited 3d data by unifying text to 2d and 3d generation. *arXiv preprint arXiv:2604.02289*, 2026.

[54] Junliang Ye, Zhengyi Wang, Ruowen Zhao, Shenghao Xie, and Jun Zhu. Shapellm-omni: A native multimodal llm for 3d generation and understanding. *arXiv preprint arXiv:2506.01853*, 2025.

[55] Astrid A Zeman, J Brendan Ritchie, Stefania Bracci, and Hans Op de Beeck. Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Scientific reports*, 10(1):2453, 2020.

[56] Xun Zhang, Weihao Xia, Yulong Liu, Bo Yang, Alessandro Bozzon, and Pan Wang. Single-stage fmri-to-3d reconstruction via viewpoint-aware embedding and hierarchical guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18037–18045, 2026.

[57] Yu Zhang, Qi Zhang, Zixuan Gong, Yiwei Shi, Yepeng Liu, Duoqian Miao, Yang Liu, Ke Liu, Kun Yi, Wei Fan, et al. Mlip: Efficient multi-perspective language-image pretraining with exhaustive data utilization. *arXiv preprint arXiv:2406.01460*, 2024.

[58] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.

A Condition Construction Details

A.1 Fine-Grained Caption Construction

To establish robust and low-variance semantic targets for the TextDecoder, we construct a single fine-grained caption for each 3D stimulus. To prevent the decoding distribution from becoming overly diffuse, the captions strictly adhere to a standardized template with closed-set vocabularies:

$$c = \text{“a <shape> <color> <material-texture> <category>”}. \tag{13}$$

Specifically, the $\langle category \rangle$ slot is restricted to the 13 canonical ShapeNet categories utilized in fMRI-Shape. The $\langle color \rangle$ and $\langle material - texture \rangle$ attributes are explicitly constrained to predefined vocabularies comprising 12 base colors and 10 base materials (optionally modified by hyphenated texture prefixes, e.g., *matte* or *brushed*). The $\langle shape \rangle$ slot is limited to a 1–2 word geometric descriptor (e.g., *twin-engine*, *boxy*).

We employ a vision-language model, Qwen3-VL-Plus, to automatically generate these structured captions based on the rendered 3D frames. The VLM is prompted to focus solely on the centered object and adhere strictly to the prescribed format. A representative prompt is provided below:

> This image shows a centered 3D rendering of a “`{category}`”. Describe this object as a single concise English noun phrase in the EXACT format: “a <shape> <color> <material-texture> `{category}`”. The <shape> is a 1–2 word geometric or structural descriptor. The <color> must be exactly one word from the closed color list. The <material-texture> is a base material with an optional hyphenated finish prefix. The category at the end must be exactly `{category}`. Start with “a” and output only the noun phrase without any punctuation.

To ensure absolute structural consistency, all VLM outputs are subjected to a rigorous rule-based validator. This validator enforces syntactic correctness (e.g., proper length of 4–10 tokens, starting with “a”) and strictly filters out any out-of-vocabulary attributes (e.g., unlisted colors or size descriptors).

Finally, the validated captions are encoded using a pre-trained OpenCLIP model (`ViT-bigG-14`) to extract L_2-normalized 1280-dimensional textual embeddings. These embeddings serve as structured, vocabulary-controlled semantic conditions for the GPT-2 decoding module and downstream texture guidance.

A.2 Superquadric Label Construction

To establish a mathematically rigorous and robust geometry-derived pseudo ground truth for supervising our SQ decoder, we construct SQ labels offline from the 3D stimulus geometries. Following the state-of-the-art 3D decomposition paradigm proposed in SUPERDEC [17], the construction process consists of point cloud extraction, neural parameter prediction, and geometric optimization.

Point Cloud Normalization and Representation. For each 3D object mesh, we first sample a dense point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$ using Farthest Point Sampling (FPS) with $N = 4096$, which is subsequently centered and uniformly rescaled to fit within a sphere of radius 0.5.

Neural Primitive Prediction. To parse the object into explicit primitives, we process $\mathcal{P}$ through a Transformer-based architecture. A PVCNN encoder extracts dense point features $\mathcal{F}_{PC} \in \mathbb{R}^{N\times H}$, while a Transformer decoder refines P learnable superquadric queries $\mathcal{F}_{SQ} \in \mathbb{R}^{P\times H}$. The network utilizes two prediction heads. The superquadric head predicts 12 parameters Θ_j for each j-th primitive: 3 for scale (s_x, s_y, s_z), 2 for shape exponents (ϵ_1, ϵ_2), 6 for 3D pose (translation and rotation), and 1 for existence probability α_j. Simultaneously, the segmentation head predicts a soft assignment matrix $M \in \mathbb{R}^{N\times P}$. The assignment probability of the i-th point to the j-th superquadric is defined as:

$$m_{ij} = \sigma(\phi(\mathcal{F}_{PC}) \cdot \mathcal{F}_{SQ})_{ij}, \tag{14}$$

where $\phi(\cdot)$ is a learned projection and σ denotes the softmax function.

Geometric Refinement via LM Optimization. The canonical surface of a superquadric is governed by the implicit function:

$$f(x) = \left(\left(\frac{x}{s_x} \right)^{\frac{2}{\epsilon_2}} + \left(\frac{y}{s_y} \right)^{\frac{2}{\epsilon_2}} \right)^{\frac{\epsilon_2}{\epsilon_1}} + \left(\frac{z}{s_z} \right)^{\frac{2}{\epsilon_1}} = 1. \tag{15}$$

To guarantee maximum geometric fidelity for our pseudo ground truth, the initial parameters Θ_j are strictly refined using Levenberg-Marquardt (LM) optimization [32, 35]. The LM algorithm iteratively minimizes the residuals $r_{ij} = m_{ij}\tilde{d}_j(x_i)$. Here, $\tilde{d}_j(x_i)$ represents the continuous radial distance from point x_i to the j-th superquadric surface, computed in closed form as:

$$\tilde{d}_j(x_i) = |x_i| \cdot \left| 1 - f(x_i)^{-\epsilon_1/2} \right|. \tag{16}$$

By executing this optimization, the generated SQ parameter sets inherently encapsulate precise macroscopic structural constraints. These highly accurate parameter sets (Θ) are ultimately utilized as the direct supervision targets for our fMRI-to-3D geometry decoding module.

B Supplementary Loss Formulations

B.1 MixCo Augmentation

Given the inherently low signal-to-noise ratio of fMRI measurements, we employ a MixCo-style [30] data augmentation strategy during the encoding stage. Let $\{\mathbf{X}_i\}_{i=1}^{B}$ denote a mini-batch of fMRI responses, where $\mathbf{X}_i \in \mathbb{R}^{T\times H\times W}$ represents a multi-frame sequence elicited by a 3D object stimulus. We generate a mixed fMRI input $\mathbf{X}_i^*$ by blending $\mathbf{X}_i$ with a paired sample $\mathbf{X}_{m_i}$, where m_i is a randomly permuted index within the mini-batch:

$$\mathbf{X}_i^* = \lambda_i \mathbf{X}_i + (1-\lambda_i)\mathbf{X}_{m_i}, \tag{17}$$

with the mixing coefficient sampled as $\lambda_i \sim \mathrm{Beta}(\alpha, \alpha)$. This mixed neural response is subsequently encoded into a unified brain-derived representation and aligned against two corresponding positive visual targets. Specifically, the CLIP image targets are extracted from randomly sampled rendered frames of the associated 3D object stimuli.

B.2 Bidirectional CLIP Alignment

To maintain consistency with the main text, let $\mathbf{z}_{v,i}$ denote the fMRI visual embedding obtained from the mixed response $\mathbf{X}_i^*$, and let $\mathbf{v}_i$ be the pre-trained CLIP image embedding of the sampled frame for the i-th stimulus. Following the encoding pipeline, the mixed signal $\mathbf{X}_i^*$ is encoded by the LEA-initialized ViT backbone followed by a lightweight dimensional adapter that projects it into the CLIP embedding space, i.e., $\mathbf{z}_{v,i} = \mathcal{E}(\mathbf{X}_i^*)$. The InfoNCE-based cosine similarity, scaled by a learnable temperature parameter τ, is defined as:

$$s(\mathbf{a}, \mathbf{b}) = \frac{\cos(\mathbf{a}, \mathbf{b})}{\tau}. \tag{18}$$

To enforce robust cross-modal alignment, we formulate the visual contrastive objective as a symmetric loss comprising both forward and backward retrieval terms:

$$\begin{aligned}\mathcal{L}_{\mathrm{CLIP}_v} = &- \frac{1}{2B}\sum_{i=1}^{B}\left[\lambda_i \log \frac{\exp(s(\mathbf{z}_{v,i}, \mathbf{v}_i))}{\sum_{k=1}^{B}\exp(s(\mathbf{z}_{v,i}, \mathbf{v}_k))} + (1-\lambda_i)\log \frac{\exp(s(\mathbf{z}_{v,i}, \mathbf{v}_{m_i}))}{\sum_{k=1}^{B}\exp(s(\mathbf{z}_{v,i}, \mathbf{v}_k))}\right] \\ &- \frac{1}{2B}\sum_{j=1}^{B}\left[\lambda_j \log \frac{\exp(s(\mathbf{z}_{v,j}, \mathbf{v}_j))}{\sum_{k=1}^{B}\exp(s(\mathbf{z}_{v,k}, \mathbf{v}_j))} + \sum_{l:m_l=j}(1-\lambda_l)\log \frac{\exp(s(\mathbf{z}_{v,l}, \mathbf{v}_j))}{\sum_{k=1}^{B}\exp(s(\mathbf{z}_{v,k}, \mathbf{v}_j))}\right].\end{aligned} \tag{19}$$

The first bracketed term constructs the mapping from fMRI embeddings to their constituent visual targets, while the second term enforces the reverse alignment. This symmetric formulation is identically applied to the textual modality to yield $\mathcal{L}_{\mathrm{CLIP}_t}$, substituting the visual embedding $\mathbf{v}$ with the corresponding CLIP text embedding $\mathbf{t}$ and utilizing the projected fMRI semantic representation $\mathbf{z}_{t,i}$. The overall encoding objective is thus given by:

$$\mathcal{L}_{\text{enc}} = \mathcal{L}_{\mathrm{CLIP}_v} + \lambda_t \mathcal{L}_{\mathrm{CLIP}_t}. \tag{20}$$

Table 3: Per-subject results of MINDSHAPE. These scores were averaged across participants for the values shown in Table 1.

Subject	Semantic-level		Geometry-level				
	Acc@2 ↑	Acc@10 ↑	IoU ↑	CD ↓	EMD ↓	F@0.02 ↑	F@0.05 ↑
Subject 1	0.921	0.717	8.606	4.293	5.365	0.167	0.390
Subject 2	0.909	0.654	8.502	4.811	5.299	0.168	0.375
Subject 3	0.907	0.660	8.388	4.639	5.511	0.166	0.368
Subject 4	0.925	0.712	9.009	4.098	5.313	0.179	0.391
Subject 5	0.878	0.598	8.450	4.363	5.481	0.164	0.370
Subject 6	0.866	0.568	9.036	4.200	5.294	0.176	0.400
Subject 7	0.877	0.562	8.390	4.584	5.478	0.167	0.379
Subject 8	0.921	0.672	7.705	4.890	5.365	0.150	0.353

C Subject-Specific Results

This appendix complements the aggregated numbers in the main paper with a closer look at how MINDSHAPE behaves on individual subjects. Quantitatively, Table 3 lists the per-subject metrics. Qualitatively, Figure 5 visualises six stimulus objects reconstructed across multiple subjects from their own fMRI responses. IoU is computed at 32^3 surface-voxel resolution; IoU, CD and EMD are reported after multiplying by 100; F-score is evaluated at $\tau \in \{0.02,\ 0.05\}$.

Inter-Subject Qualitative Analysis. Figure 5 provides a visual counterpart to the per-subject metrics in Table 3. Using the ground-truth row as reference, the reconstructions remain category-consistent across subjects and preserve comparable global proportions for several stimulus categories, echoing the compact metric range in the table, where F@0.02 remains within 0.150–0.179 and IoU within 7.7–9.0. The stable silhouettes and part layouts suggest that MINDSHAPE recovers a shared category-level shape prior from individual fMRI responses, while the final textured meshes retain subject-specific variation in color, material and fine surface appearance. The figure also exposes meaningful failure cases, where Subject 4's cabinet is decoded with a flatter credenza-like body and Subject 3's sedan receives an implausible multicolor texture. Subject 1's watercraft further illustrates that individual variability can destabilize shape decoding, yet the semantic condition helps correct the generation toward the intended category. Overall, MINDSHAPE makes a balanced effort to produce 3D assets that are both semantically correct and geometrically faithful.

D Inference Model

To clarify the 3D generative prior used in this paper, we briefly summarize flow matching. Unlike standard diffusion models that learn a stepwise denoising distribution, flow matching formulates generation as a continuous-time deterministic transport process. Given a real sample z_0 and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we construct the linear interpolation

$$z_t = (1 - t)z_0 + t\epsilon, \tag{21}$$

and learn a velocity field $v_\theta(z_t, t)$ to approximate the instantaneous change along the trajectory. At inference time, the model starts from the noisy endpoint z_1 and integrates backward along the ordinary differential equation

$$\frac{dz_t}{dt} = v_\theta(z_t, t) \tag{22}$$

to recover the clean sample z_0.

TRELLIS [51] applies this idea in a two-stage 3D generation pipeline. The first-stage structure flow generates a coarse structural latent in a compact 3D latent space and decodes it into a voxel structure. The second-stage appearance flow then completes local appearance features for active voxels, ultimately decoding them into textured meshes.

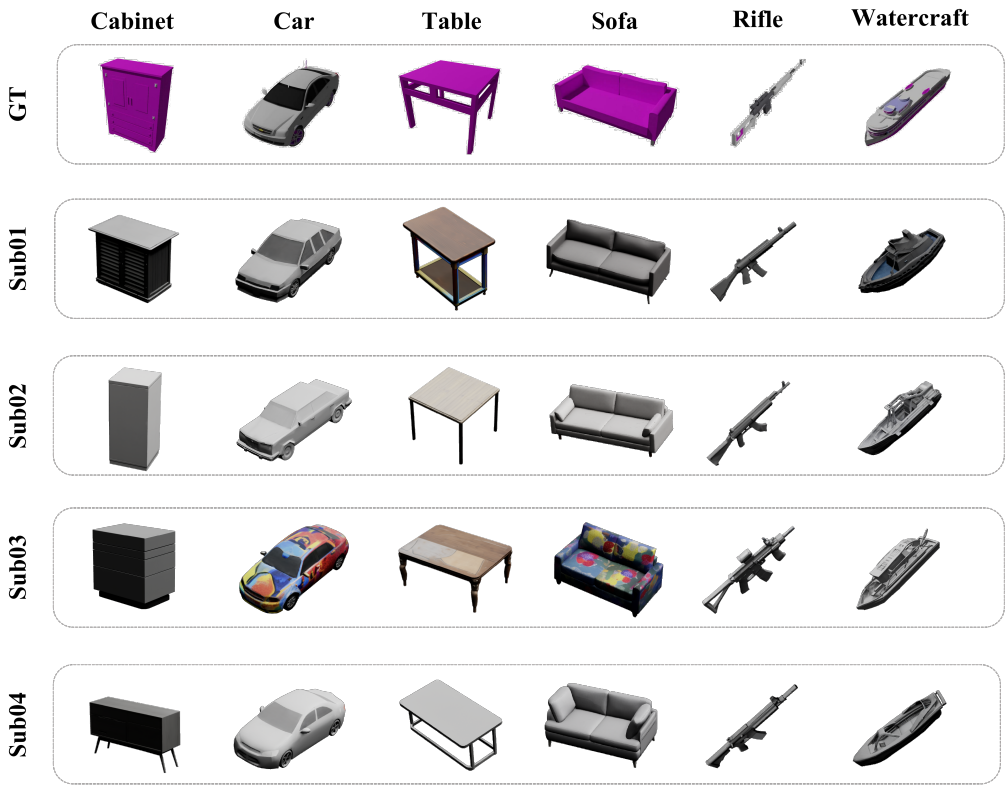

Figure 5: Inter-subject qualitative reconstructions on fMRI-Shape. The top row shows the ground-truth stimulus objects, and Rows 2–5 show the corresponding textured reconstructions from four different subjects, each generated from that subject's own fMRI response. Across subjects, the reconstructions preserve the target object categories and global layouts while exhibiting subject-specific variation in surface texture, colour and fine-grained appearance.

The inference pipeline injects explicit 3D geometric conditions directly into the test-time structure-flow inference process without retraining the generator. Concretely, given a spatial control signal x_c (e.g., superquadrics), the pipeline first maps it into the structure latent space using the TRELLIS structure encoder, obtaining $z_{c,0} = E(x_c)$. Instead of starting the denoising process from pure noise, it constructs an intermediate state at a chosen time step t_0 :

$$z_{t_0} = t_0 z_1 + (1 - t_0) z_{c,0}, \tag{23}$$

where $z_1 \sim \mathcal{N}(0, I)$. Starting from this partially noised control latent, the pretrained structure flow continues the reverse process from t_0 to 0. In this way, the control geometry is not introduced through an extra network branch or an auxiliary loss. Instead, it directly modifies the starting point of the generation trajectory. As a result, the final shape can be explicitly constrained in 3D space without breaking the pretrained model prior.

From this perspective, MindShape first decodes a set of superquadric (SQ) parameters and a fine-grained caption from fMRI, then converts the SQ set into an explicit spatial control signal and the caption into a semantic condition, and finally feeds both signals into a pretrained 3D generator. The SQ branch provides structural priors such as global shape, part layout, relative scale, and pose, while the text branch complements the generation with category, material, color, and appearance attributes. Because the SQ condition is injected into the structure-flow latent trajectory rather than used only as a post-hoc filter or a multi-view intermediate, it provides a more direct and interpretable constraint on the final 3D geometry. This is one reason why the proposed method can improve geometric stability and spatial faithfulness while preserving semantic consistency.

E Implementation Details

Data and inputs. We follow the official fMRI-Shape preprocessing protocol and use the released 256×256 voxel-projection maps as the fMRI input to the encoder. Each example pairs a short fMRI response with the corresponding 3D object stimulus rendered as a 6-frame multi-view clip ($n = 6$). During training, the visual target is sampled per example uniformly from the six rendered views, which acts as a view-level augmentation. At inference and rendering we use a fixed view index for reproducibility; image-domain metrics are aggregated over the multi-view scores as described below.

Brain Representation Encoding. The fMRI encoder is a ViT (patch 16, depth 24, 16 heads, embedding dim 1024) initialized from the pretrained LEA backbone, followed by a lightweight dimensional adapter that projects to the CLIP embedding space. Stage 1 uses a contrastive objective with MixCo augmentation, where the visual branch is aligned to CLIP image features extracted from a randomly sampled rendered view, and the text branch is aligned to CLIP text features of category-level captions, with λ_t=0.25 and MixCo coefficient α=0.15. We optimize with AdamW under a OneCycle schedule (max lr $= 3\times10^{-4}$), batch size 64, for 150 epochs.

Geometry-Semantic Decoding. The Stage-1 backbone is frozen, and two task-specific heads are trained on top of the brain representation: a Superquadric decoder and a fine-grained text decoder. The SQ decoder is a Transformer set predictor with 16 learnable query slots, hidden dim 512, 8 heads, 3 layers, and dropout 0.1; each slot regresses an existence logit, the two shape exponents, 3D scale, 3D translation, and a 6D rotation. The set-prediction loss uses Hungarian matching with weights λ_{exist}=1.0, λ_{ϵ}=1.0, λ_{scale}=2.0, λ_{trans}=1.0, and λ_{rot}=5.0; the matching-cost weights $\alpha_{(\cdot)}$ are set equal to the corresponding $\lambda_{(\cdot)}$. The text decoder is a GPT-2 prefix tuner with 4 learned prefix tokens and a maximum generation length of 20, supervised by token-level cross-entropy; its input is detached from the shared backbone so that caption supervision does not back-propagate into the brain representation. Stage 2 is trained with batch size 24, max lr $= 5\times10^{-5}$ on a OneCycle schedule, for 100 epochs.

3D Reconstruction Inference. At test time, the fMRI signal is mapped into the brain representation, the SQ decoder produces a set of candidate primitives, and the text decoder generates a fine-grained caption. The decoded SQ set is rasterized into the spatial latent and the predicted caption is used as the textual prompt; the conditional 3D generator runs 25 denoising steps with a classifier-free-guidance scale of 7.5. The resulting textured mesh is jointly conditioned on the decoded geometric layout and the decoded semantics.

Computational resources. All training and inference experiments were completed on a single NVIDIA H200 GPU.

F Broader Impact

This work studies fMRI-conditioned 3D reconstruction in a controlled object-centric setting. In the long term, brain-inspired 3D asset reconstruction could contribute to immersive human–computer interaction, VR/AR content creation, neurorehabilitation interfaces, and assistive systems that help users externalize perceived or imagined spatial content. Such applications may lower the barrier for creating 3D assets and provide new tools for studying how the human brain represents object geometry, category semantics, and appearance.

At the same time, neural decoding raises important privacy and consent concerns. Although current fMRI-based reconstruction remains far from practical mind reading, requires controlled acquisition, and is limited by subject-specific variability, progress in this area should be accompanied by strict safeguards for brain data. We advocate using de-identified and consented datasets, transparent data governance, and application boundaries that prevent involuntary decoding or misuse of neural signals. The goal of this work is scientific understanding and assistive interaction, not surveillance or unauthorized inference of private mental content.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the scope of MINDSHAPE, including fMRI-to-3D reconstruction, SQ-guided geometry, and the reported experimental claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Sec. 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Model, training, inference, metrics, and compute details are specified in Sec. 3, Sec. 4, and Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper uses publicly available fMRI-Shape data and describes the reproduction protocol. Our code will be released after acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and testing settings are given in Sec. 4 and Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are averaged over eight subjects, with per-subject variability reported in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute setup is reported in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses a previously released fMRI dataset and does not collect new human-subject data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential benefits and privacy risks in Appendix F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release private neural data or a new high-risk public model.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets, baselines, pretrained models, and 3D generation components are cited in Sec. 2, Sec. 4, and Appendix E.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, or released asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not conduct new crowdsourcing or new human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper uses an existing released dataset and does not collect new human-subject data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The authors did not use LLMs as an important, original, or non-standard component of the research process; any use for writing or formatting assistance does not affect the core methodology or scientific claims.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.