Mind's Eye in 4D: 4D Scene Generation conditioned on fMRI Representations

Anonymous authors
Paper under double-blind review

Figure 1: **fMRI signals based BCI. (a)** Subject to respective visual stimuli, prior fMRI-to-image, fMRI-to-shape, and fMRI-to-video systems mainly produce static or single-view outputs. **(b)** By generating plausible dynamic 3D scenes conditioned on fMRI, our framework explores a step toward richer visual BCI outputs beyond previous methods.

Abstract

Brain-computer interfaces (BCIs) have enabled breakthroughs like translating fMRI signals into images or videos. However, human perception operates in a dynamic 3D world, processing information across both spatial and temporal dimensions. In this work, we introduce ***Brain-to-4D***, a novel fMRI-conditioned 4D scene generation that synthesizes dynamic 3D scenes conditioned on decoded fMRI signals. Since the subjects observe only 2D videos, the target is not to directly reconstruct physically exact 4D geometry from brain activity. Instead, our goal is to generate plausible 4D scenes whose high-level semantics are guided by fMRI representations, while unobserved multi-view geometry and high-frequency dynamics are completed by generative priors. To address the absence of paired fMRI–4D data, we propose ***Mind4D***, an innovative fMRI-conditioned 4D generation framework capable of learning asymmetric hierarchical representations from fMRI signals in a weakly supervised manner. Our approach captures both high-level and low-level representations, along with the decomposition of scene backgrounds and object foregrounds. By conditioning and integrating multiple generative priors for the foreground and background, Mind4D produces prior-guided semantic 4D visuals. Extensive experiments show that Mind4D generates multi-view consistent 4D visuals semantically aligned with brain activity.

1 Introduction

Brain-computer interfaces (BCIs) (Saha et al., 2021; Rashid et al., 2020; Wolpaw et al., 2002) are increasingly recognized for enabling direct communication through brain activity. Among non-invasive approaches, functional magnetic resonance imaging (fMRI) has been widely adopted for various BCI functions. With recent generative AI advances, state-of-the-art fMRI-to-visual methods now reconstruct images (Takagi & Nishimoto, 2023; Scotti et al., 2023), videos (Chen et al., 2024; Gong et al., 2025a), and 3D shapes (Gao et al., 2023) (see Figure 1**(a)**).

However, humans possess an innate ability to perceive and interpret scenes in both spatial and temporal dimensions, even during fleeting thoughts (Heft, 2010; Kiverstein & Rietveld, 2021; Wang & Spelke, 2002). Immersed in a dynamic three-dimensional environment, the human brain inherently encodes and integrates spatiotemporal information at a subconscious level (Zhang et al., 2020; Schonhaut et al., 2023; Howard, 2017;

Xu et al., 2021). Inspired by this process, we propose **Brain-to-4D**, an fMRI-conditioned 4D generation framework that creates 4D visuals conditioned on decoded fMRI signals (Figure 1**(b)**). Since the visual stimulus is a 2D video and paired fMRI–4D supervision is unavailable, Brain-to-4D does not directly decode exact physical 4D geometry from brain activity. Instead, it synthesizes plausible dynamic 3D scenes by using fMRI-derived semantic and spatiotemporal representations as conditions, while relying on generative priors to complete unobserved views, 3D structure, and high-frequency motion. This formulation studies how decoded brain representations can guide richer visual generation under incomplete supervision.

Building a brain-to-4D generation system is inherently challenging due to the instantaneous nature of brain responses. It prevents simultaneous multi-view capture. The high cost and temporal limitations of fMRI acquisition also make it difficult to collect continuous fMRI-visual pairs. To overcome these challenges, we propose **Mind4D**, a novel framework for fMRI-conditioned 4D generation. Rather than attempting to extract high-frequency kinematics directly from low-frequency fMRI signals, Mind4D decodes coarse semantic and spatiotemporal representations from brain activity, and leverages them to condition powerful video generative models. Furthermore, Mind4D designs a novel weakly supervised learning approach that circumvents the need for direct fMRI-4D pairings. Instead, it exploits existing fMRI-2D video pairs to extract pseudo-labels. To handle the inevitable error accumulation from these pseudo-labels, we propose an asymmetric hierarchical representation learning scheme with vector quantization (VQ). These representations capture both high-level semantic information and low-level visual details, while robustly decomposing the scene into distinct background and foreground components. By conditioning foreground and background generative priors with these purified signals, Mind4D synthesizes high-quality, semantically aligned 4D visuals with preserved multi-view consistency.

Our **contributions** are three-fold: **(i)** We introduce Brain-to-4D generation, a novel fMRI-conditioned setting that synthesizes plausible dynamic 3D scenes from decoded brain representations. **(ii)** We propose Mind4D, an innovative framework that integrates asymmetric hierarchical representation learning with generative priors for fMRI-conditioned, prior-guided 4D synthesis. **(iii)** We establish new benchmarks for brain-to-4D generation. Experiments demonstrate that Mind4D generates multi-view-consistent 4D visuals that are semantically aligned with brain activity. Additional exploratory visualizations of region-of-interest conditioning are provided in the supplementary material.

2 Related work

2.1 Neural decoding for BCIs

Existing neural decoding studies have focused on extracting essential representations of brain signals for visual content decoding and object recognition (Liu et al., 2025a; Saha et al., 2021; Rashid et al., 2020; Lawhern et al., 2018; Guger et al., 2024; Abdulkader et al., 2015). However, they often fail to produce detailed visuals directly from brain signals. These investigations have facilitated advances in reconstructing images (Beliy et al., 2019; Chen et al., 2023; Huo et al., 2024), videos (Chen et al., 2024; Lu et al., 2025; Liu et al., 2025b), and geometry (Gao et al., 2023; 2024) from fMRI using techniques such as self-supervised dual-network architectures (Beliy et al., 2019), latent-space diffusion (Scotti et al., 2023; 2024; Gong et al., 2025a) or hierarchical latent variable models aligned with cortical processing (Takagi & Nishimoto, 2023; Miliotou et al., 2023). Concurrent works explore cross-subject alignment (Wang et al., 2024; Gong et al., 2025b; Li et al., 2024), hierarchical latent decomposition (Li et al., 2025), cortical-inspired decoding (Wang et al., 2025), and multi-shot reconstruction with LLM priors (Jiang et al., 2024a). Despite this progress, existing fMRI-to-visual reconstructions mainly focus on single-view or static objects. They do not directly address the generation of dynamic 3D scenes that can be rendered from multiple viewpoints over time. Meanwhile, humans inherently encode spatial and temporal information subconsciously (Schonhaut et al., 2023; Howard, 2017; Xu et al., 2021; Rolls & Treves, 2011; Naselaris et al., 2011). Motivated by this, we propose Brain-to-4D, an fMRI-conditioned 4D scene generation, where decoded brain representations are used to guide plausible multi-view dynamic synthesis.

2.2 3D/4D generation

Recent advances in text/image-based 3D generation (Poole et al., 2023; Lin et al., 2023; Wang et al., 2023; Tang et al., 2024; Liu et al., 2023; Shi et al., 2023) are predominantly based on strong 3D representations, including NeRF (Mildenhall et al., 2020), DMTet (Shen et al., 2021) or Gaussian splatting (Kerbl et al., 2023), which leverage score distillation sampling (Poole et al., 2023) (SDS) and extensive 3D datasets (Deitke et al., 2023; Yu et al., 2023; Wu et al., 2023). With the emergence of 4D representations (Wu et al., 2024; Pumarola et al., 2020; Cao & Johnson, 2023; Yang et al., 2024; 2023), these techniques have also been extended to generate dynamic 3D scenes (Jiang et al., 2024b; Ren et al., 2023; Tang et al., 2024). Our approach takes a step further by leveraging diffusion models as generative priors, which will be guided through hierarchical representations to bridge the gap between fMRI and 4D generation. Different from text- or image-conditioned 4D generation, the conditioning signal in our setting is decoded from noisy and low-temporal-resolution brain activity. The main challenge is therefore not only 4D synthesis itself, but also how to learn fMRI representations that can reliably interface with foreground and background generative priors when paired fMRI-4D supervision is unavailable.

3 Preliminary

3.1 Deformable 3D Gaussian splatting

3D Gaussian splatting (3DGS) (Kerbl et al., 2023) represents a 3D scene with a set of Gaussians. Each Gaussian is characterized by position mean $\mu \in \mathbb{R}^3$, covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$, color $\mathbf{c} \in \mathbb{R}^3$, and opacity $\alpha \in \mathbb{R}$. The color of each pixel results from the 2D projection of these 3D Gaussians and the depth volumetric rendering. In the dynamic setting, we adopt the deformable 3DGS (Wu et al., 2024) as the 4DGS representation, which uses an additional network Φ to predict the deformation of $S = \{\mu, \Sigma, \alpha\}$ given timestamp τ: $\tilde{S} = \Phi(S, \tau)$, where $\tilde{S}$ denotes the deformed attributes of S. With these deformed attributes, we can render images at different timestamps.

3.2 Score distillation sampling

Score distillation sampling (SDS) provides a method to distill knowledge from a pre-trained diffusion model ϵ_ϕ. Specifically, when an image I is rendered from a scene representation (*e.g.* 3DGS) parameterized by θ, the SDS gradient is calculated as:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, I_t) = \mathbb{E}\left[w(t)\left(\epsilon_\phi(I_t; t, c) - \epsilon\right)\frac{\partial I_t}{\partial \theta}\right], \tag{1}$$

where $w(t)$ is defined as the weighting function in SDS as defined in (Poole et al., 2023). I_t is the perturbed image with noise ϵ at time step t, and c is the condition.

3.3 Vector quantization

Vector quantization (VQ) involves mapping continuous input embeddings to discrete codebook entries. Given an input embedding $z_{\text{h}} \in \mathbb{R}^D$, the quantized embedding z_{h}^{vq} is determined by selecting the closest codebook vector from a set of codebook entries $\{g_{\text{j}} \in \mathbb{R}^D\}_{j=1}^K$ based on $z_{\text{h}}^{vq} = g_k$, where $k = \text{argmin}_j \|z_{\text{h}} - g_j\|$.

4 Method

We present **Mind4D**, a pioneering fMRI-guided 4D generation framework, as depicted in Figure 2. The key idea of Mind4D is to extract asymmetric hierarchical representations from fMRI signals, which condition generative priors for 4D synthesis.

This approach is designed to tackle the challenge of partially aligned modalities between 2D video supervision and 4D scene targets, circumventing the need for paired fMRI-4D data. During weakly supervised pre-training in the first stage, we optimize encoders, decoders, and extractors to extract asymmetric hierarchical

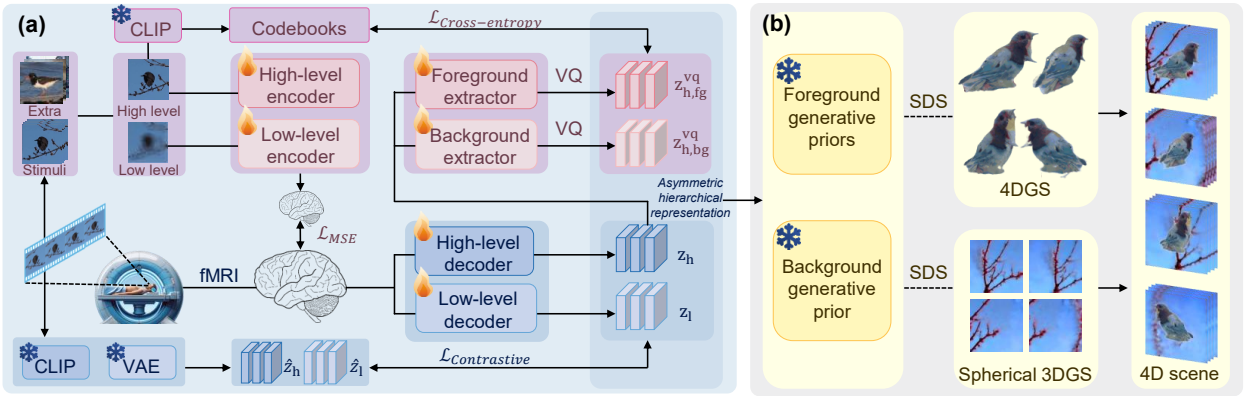

Figure 2: **Overview of Mind4D.** Without labeled fMRI-to-4D training data, we design a neural-guided framework that integrates asymmetric hierarchical representation learning with rich generative priors for fMRI-conditioned 4D synthesis. **(a)** In weakly-supervised pre-training in the first stage, we decode fMRI into both high-level semantic (z_{h}) and low-level spatiotemporal (z_{l}) representations, optimized in a contrastive manner. We design fMRI encoders that can translate extra images to fMRI signals to improve the generalization of hierarchical representations. Asymmetrically, we extract foreground ($z^{vq}_{\mathrm{h,fg}}$) and background ($z^{vq}_{\mathrm{h,bg}}$) representations from z_{h}, leveraging vector quantization (VQ) to ensure the stability of hierarchical representations. **(b)** In the second inference stage, we apply pre-trained model to extract hierarchical representation $\{z_{\mathrm{h}}, z_{\mathrm{l}}, z^{vq}_{\mathrm{h,fg}}, z^{vq}_{\mathrm{h,bg}}\}$ directly from fMRI, which conditions both foreground and background generative priors to optimize 4DGS for foreground and spherical 3DGS for background, which are then integrated for the target 4D scene composition.

representations from brain signals using fMRI-2D training set (Section 4.1). During inference, we apply pre-trained model to extract asymmetric hierarchical representations from solely brain signals in the testing set. These hierarchical representations serve as conditions for separate generative priors for the foreground and background, which are then integrated for 4D synthesis (Section 4.2). Since paired fMRI-4D data are unattainable, our generated novel views could only be interpreted as prior-guided plausible completions rather than directly perceived 3D geometry.

4.1 Hierarchical representation learning

4.1.1 fMRI decoding

Inspired by the human brain's ability to dynamically decode fleeting sensory input across spatial and temporal dimensions (Heft, 2010; Kiverstein & Rietveld, 2021; Wang & Spelke, 2002), we design fMRI decoding to decode hierarchical representations from fleeting brain signals. When perceiving scenes, humans capture both broad outlines and intricate details (Hegdé, 2008; Petras et al., 2019; Friston, 2008). To reflect this dual-resolution processing, our model adopts a two-branch architecture that explicitly separates low-level and high-level processing, allowing specialized pathways to handle global structure and fine-grained information, respectively.

Specifically, paired fMRI-2D data contain video stimuli $\hat{V}$ and N sequences of brain signals $\hat{X} = [\hat{X}_1, \hat{X}_2, ..., \hat{X}_N]$. In fMRI decoding, brain signals $\hat{X}$ are mapped into visual representations $z_{\mathrm{h}} = F_{\mathrm{d,h}}(\hat{X}), z_{\mathrm{l}} = F_{\mathrm{d,l}}(\hat{X})$ through the high-level decoder $F_{\mathrm{d,h}}$ and the low-level decoder $F_{\mathrm{d,l}}$. During training, the high-level representation z_{h} is supervised from CLIP (Radford et al., 2021) embedding of key frame $\hat{I}$ chosen from video $\hat{I} \in \hat{V}$ through soft contrastive regulation (Scotti et al., 2023) (SoftCLIP):

$$\mathcal{L}_{\mathrm{D,H}} = \mathcal{L}_{\mathrm{SoftCLIP}}(z_{\mathrm{h}}, \hat{z}_{\mathrm{h}}),\ \hat{z}_{\mathrm{h}} = \mathrm{CLIP}(\hat{I}). \tag{2}$$

On the other hand, the low-level representation z_{l} is supervised from SD VAE (Rombach et al., 2022) embedding of $\hat{V}$ through contrastive regulation, with

$$\mathcal{L}_{\mathrm{D,L}} = \mathcal{L}_{\mathrm{Contrastive}}(z_{\mathrm{l}}, \hat{z}_{\mathrm{l}}),\ \hat{z}_{\mathrm{l}} = \mathrm{VAE}(\hat{V}). \tag{3}$$

The high-level representation captures features of objects and scenes from key frames. While fMRI blood-oxygen-level-dependent (BOLD) signals inherently suffer from low temporal resolution (typically TR = 1 $\sim$ 2s) and physiological delay, neuroscience studies suggest that the visual cortex encodes time-integrated motion semantics rather than frame-by-frame kinematics (Hegdé, 2008). Therefore, instead of forcing the model to extract high-frequency frame-wise details directly from low-frequency fMRI, our low-level representation z_{l} focuses on capturing these abstract, temporally-aggregated motion semantics by mapping the entire rough video into VAE representations. The high-frequency dynamics are subsequently synthesized in a physically and semantically plausible manner by the rich temporal knowledge embedded in video generative priors.

4.1.2 fMRI encoding regularization

To mitigate overfitting on the limited fMRI-video paired dataset, we introduce an auxiliary visual-to-fMRI encoding branch as a regularizer. For the ground truth image $\hat{I} \in \hat{V}$ sampled from training videos, we apply a Gaussian blur to yield $\hat{I}' = \mathrm{Blur}(\hat{I})$, intentionally removing high-frequency textures to match the inherently low spatial resolution of BOLD fMRI signals. We employ both high-level encoder $F_{\mathrm{e,h}}$ and low-level encoder $F_{\mathrm{e,l}}$ to encode images into simulated brain signals: $X_{\mathrm{h}} = F_{\mathrm{e,h}}(\hat{I}), X_{\mathrm{l}} = F_{\mathrm{e,l}}(\hat{I}')$. These encoders are supervised through Mean Square Error (MSE):

$$\mathcal{L}_{\mathrm{E,H}} = \mathcal{L}_{\mathrm{MSE}}(X_{\mathrm{h}}, \hat{X}),\ \mathcal{L}_{\mathrm{E,L}} = \mathcal{L}_{\mathrm{MSE}}(X_{\mathrm{l}}, \hat{X}). \tag{4}$$

More importantly, for external images $\hat{I}_{extra}$ without corresponding fMRI pairs, the encoders generate surrogate brain signals $X_{extra} = F_{\mathrm{e,h}}(\hat{I}_{extra})$. These surrogate pairs are subsequently fed into decoders $F_{\mathrm{d,h}}, F_{\mathrm{d,l}}$, effectively simulating the continuous visual perception process and injecting diverse visual priors to enhance the generalization of the decoders.

4.1.3 Vector quantized semantic extraction

To effectively disentangle complex scenes, we propose an asymmetric representation learning module that extracts decoupled object and scene semantics directly from the high-level representation z_h. Specifically, we design two extractors $F_{\mathrm{fg}}, F_{\mathrm{bg}}$ to split the high-level representation $z_{\mathrm{h}} = F_{\mathrm{d,h}}(\hat{X})$ into representations that represent background and foreground information: $z_{\mathrm{h,fg}} = F_{\mathrm{fg}}(z_{\mathrm{h}}), z_{\mathrm{h,bg}} = F_{\mathrm{bg}}(z_{\mathrm{h}})$. The resulting $z_{\mathrm{h,fg}}, z_{\mathrm{h,bg}}$ is then vector quantized by:

$$\begin{aligned} z^{vq}_{\mathrm{h,fg}} &= g_{\mathrm{fg}}[k],\ k = \mathrm{argmin}_j \left\| z_{\mathrm{h,fg}} - g_{\mathrm{fg}}[j] \right\|, \\ z^{vq}_{\mathrm{h,bg}} &= g_{\mathrm{bg}}[k],\ k = \mathrm{argmin}_j \left\| z_{\mathrm{h,bg}} - g_{\mathrm{bg}}[j] \right\|, \end{aligned} \tag{5}$$

where $g_{\mathrm{fg}} \in \mathbb{R}^{K_{\mathrm{fg}} \times D_{\mathrm{fg}}}, g_{\mathrm{bg}} \in \mathbb{R}^{K_{\mathrm{bg}} \times D_{\mathrm{bg}}}$ are the codebooks with $K_{\mathrm{fg/bg}}$ entries and $D_{\mathrm{fg/bg}}$ dimension. Since paired fMRI-4D data is unavailable, our weakly supervised strategy leverages 2D video stimuli $\hat{V}$ as an intermediate bridge. We extract pseudo-labels for foreground and background from $\hat{V}$ to map the stimuli into the CLIP (Radford et al., 2021) space and initialize the codebooks. To mitigate the error accumulation caused by imperfect 2D pseudo-labels during 4D generation, we rely on the vector quantization (VQ) bottleneck. The codebooks g_{fg} and g_{bg} act as semantic anchors that filter out high-frequency noise and segmentation artifacts from the pseudo-labels, ensuring that the extracted representations $z^{vq}_{\mathrm{h,fg}}$ and $z^{vq}_{\mathrm{h,bg}}$ provide clean and robust guidance for the subsequent 4D optimization. The extractors $F_{\mathrm{fg}}, F_{\mathrm{bg}}$ are optimized using cross-entropy (CE) loss as equation (6):

$$\mathcal{L}_{\mathrm{VQ}} = \mathcal{L}_{\mathrm{CE}}(z_{\mathrm{h,fg}}, g_{\mathrm{fg}}) + \mathcal{L}_{\mathrm{CE}}(z_{\mathrm{h,bg}}, g_{\mathrm{bg}}). \tag{6}$$

One key advantage of vector quantization is its ability to bypass the curse of dimensionality. By constraining the size of the latent space $K \ll n$, we significantly improve the regularization of the model and avoid

overfitting (Peng et al., 2023) in high-dimensional feature spaces. Furthermore, our approach yields a tighter KL divergence between the quantized and empirical distributions compared to the non-quantized case, as indicated by Theorem 4.1. It shows that the quantized latent space representation z_{h}^{vq} (either $z_{\mathrm{h,fg}}^{vq}$ or $z_{\mathrm{h,bg}}^{vq}$) yields a much tighter approximation to the true distribution compared to the nonquantized z_{h}, which is crucial for robust latent representations.

Theorem 4.1 *Denote $p(z_{\mathrm{h}})$ as the distribution of the embeddings without vector quantization and $p(\hat{z}_{\mathrm{h}})$ as the smooth-approximated empirical distribution from samples, $p(z_{\mathrm{h}}^{vq})$ and $p(\hat{z}_{\mathrm{h}}^{vq})$ as their vector quantized counterparts. Then,*

$$KL(p(z_{\mathrm{h}}^{vq})||p(\hat{z}_{\mathrm{h}}^{vq})) \ll KL(p(z_{\mathrm{h}})||p(\hat{z}_{\mathrm{h}})). \tag{7}$$

Additionally, Theorem 4.2 shows our vector quantized approach also significantly reduces entropy. This ensures that the model is less likely to capture irrelevant data-specific noise, thereby enhancing generalization to unseen data.

Theorem 4.2 *Denote L as the CLIP (Radford et al., 2021) space boundary size, $H(z_{\mathrm{h}}^{vq})$ as the entropy of distribution of vector quantized embeddings, and $H(z_{\mathrm{h}})$ as the entropy of Riemann-Discrete approximated distribution without vector quantization. Then we have $H(z_{\mathrm{h}}) > H(z_{\mathrm{h}}^{vq})$,*

$$H(z_{\mathrm{h}}) - H(z_{\mathrm{h}}^{vq}) = O\left(\log\left(L^d/K\right)\right). \tag{8}$$

The proof can be found in the supplementary. The empirical validation and ablation in Section 5.6 illustrate the substantive role of the quantized latent space representation. As a result, the whole loss in our training stage is:

$$\mathcal{L} = \lambda_{\mathrm{E}}(\mathcal{L}_{\mathrm{E,H}} + \mathcal{L}_{\mathrm{E,L}}) + \lambda_{\mathrm{D}}(\mathcal{L}_{\mathrm{D,H}} + \mathcal{L}_{\mathrm{D,L}}) + \lambda_{\mathrm{VQ}}\mathcal{L}_{\mathrm{VQ}}. \tag{9}$$

Different from previous weakly supervised approaches (Zhou, 2018; Mahajan et al., 2018; Zheng et al., 2021) that assume uni-modality labels are available, our framework solves the partially aligned modalities between existing fMRI-2D pairs and fMRI-4D targets. This is achieved by extracting asymmetric hierarchical representations from the fMRI-2D supervision and serving these hierarchical representations as the condition for the 4D scene targets.

4.2 Generative priors for fMRI-based 4D synthesis

4.2.1 Hierarchical representation guidance

The core contribution of this work lies in the robust fMRI representation learning described in Section 4.1. To validate the richness of our extracted semantics $(z_{\mathrm{h}}, z_{\mathrm{l}}, z_{\mathrm{h,fg}}^{vq}, z_{\mathrm{h,bg}}^{vq})$, we deliberately employ an assembly of generative models as a downstream generation pipeline. By adopting these frozen, pre-trained models rather than designing a proprietary 4D generator from scratch, we demonstrate that our fMRI representations act as robust semantic conditions compatible with state-of-the-art generative priors.

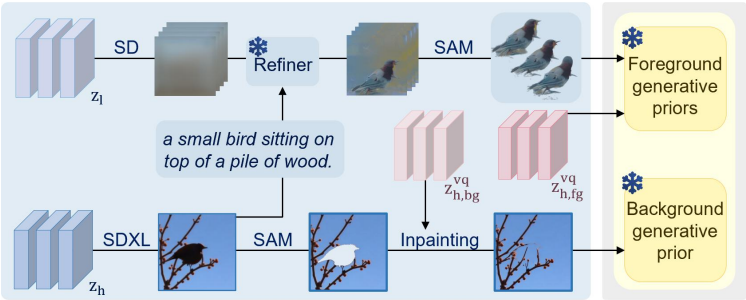

Figure 3: **Hierarchical representation guidance.** The hierarchical representations $z_{\mathrm{l}}, z_{\mathrm{h}}$ generate both coarse videos and fine key-frame visuals, which are then refined and segmented to get foreground visuals to condition the foreground generative priors with $z_{\mathrm{h,fg}}^{vq}$. The $z_{\mathrm{h,bg}}^{vq}$ semantically steers the inpainting of key-frame background, conditioning the background generative prior.

During inference, we extract representations $z_{\mathrm{h}}, z_{\mathrm{l}}, z_{\mathrm{h,fg}}^{vq}, z_{\mathrm{h,bg}}^{vq}$ from fMRI. In this downstream phase, the boundaries between brain decoding and model generation are clearly delineated: the fMRI representations provide the core macroscopic semantic control, while the powerful generative priors synthesize the missing high-frequency visual and temporal details to ensure

photorealism (Figure 3). The semantic representation z_{h} is used to generate key-frame visuals $I_{\mathrm{h}} = \mathrm{SDXL}(z_{\mathrm{h}})$ through SDXL (Podell et al., 2023), while the spatiotemporal representation z_{l} is applied to generate low-level videos $V_{\mathrm{l}} = \text{VAE-D}(z_{\mathrm{l}})$ frame by frame through the VAE decoder of Stable Diffusion (Rombach et al., 2022). The key-frame visuals I_{h}, low-level videos V_{l} and the caption generated from key-frame visuals $c_{\mathrm{h}} = \mathrm{BLIP}(I_{\mathrm{h}})$ are integrated together to generate refined video V_{r} through a pretrained refiner F_{r}:$V_{\mathrm{r}} = F_{\mathrm{r}}(V_l; c_{\mathrm{h}}, I_{\mathrm{h}})$. The refined video V_{r} is further segmented into foreground part through SAM (Kirillov et al., 2023). For the background part, we remove the objects in key-frame visuals I_{h} and inpaint it using $z^{vq}_{\mathrm{h,bg}}$ as semantic guidance.

4.2.2 Foreground and background generative priors

Modeling 4D scenes faces two challenges: (1) Foreground and background present intrinsically different characteristics (*e.g.*, dynamic vs. static); (2) Camera perspectives in 4D scenes often blur out nearby objects dynamically. To tackle these issues, we propose decomposed generation with both foreground and background generative priors.

The foreground generative priors are aimed at optimizing 4DGS (*i.e.*, deformable 3D Gaussians introduced in Section 3), driven by foreground video V_{r} and semantic representation $z^{vq}_{\mathrm{h,fg}}$. The foreground optimization, 3D Gaussians and its deformation are optimized by foreground generative priors using SDS loss as equation (1). Along with a L_2 loss between rendered image $\hat{I}_{\mathrm{ref}}$ under reference views with $I_{\mathrm{ref}} \in V_{\mathrm{r}}$, the total loss $\mathcal{L}_{\mathrm{fg}}$ for foreground modeling can be expressed by:

$$\mathcal{L}_{\mathrm{fg}} = \lambda_{\mathrm{img}}\mathcal{L}_{\mathrm{SDS,img}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{SDS,sem}} + \lambda_{\mathrm{ref}}\|\hat{I}_{\mathrm{ref}} - I_{\mathrm{ref}}\|_2^2, \tag{10}$$

where λ_* are balancing weights, with an image-based generative prior ("img") and a semantic-based generative prior ("sem") referring to (AI, 2023) and (Shi et al., 2023) guidance, respectively. To ensure the generated 4D scene possesses valid 3D geometric structures rather than degrading into a "flat" 2D video, the optimization is rigorously constrained by multi-view consistency. By rendering the 4DGS from varied camera viewpoints and applying SDS with 3D-aware zero-shot novel view synthesis priors (Shi et al., 2023), we enforce strict spatial coherence. Note that ground-truth depth or multi-view geometry is unavailable for the observed 2D stimuli. Consequently, we do not evaluate metric 3D accuracy. Instead, we evaluate whether the generated scenes maintain semantic consistency across views and temporal coherence. This is achieved through multi-view CLIP similarity, temporal CLIP consistency, and a user study on perceived 3D/4D plausibility.

The background generative prior optimizes spherical 3D Gaussians whose points are scattered on a spherical surface. This spherical 3DGS is designed to mimic a panoramic environment map, which supports rendering from 360° viewpoints. A scene-level 3D-aware generative prior (Sargent et al., 2023) extends the inpainted background image I_{h} into a consistent 360° environment. The background loss $\mathcal{L}_{\mathrm{bg}}$ is:

$$\mathcal{L}_{\mathrm{bg}} = \lambda_{\mathrm{bg}}\mathcal{L}_{\mathrm{SDS,bg}} + \lambda_{\mathrm{ref}}\|\hat{I}_{\mathrm{h}} - I_{\mathrm{h}}\|_2^2, \tag{11}$$

where λ_* refers to balancing weights and $\mathcal{L}_{\mathrm{SDS,bg}}$ refers to SDS with background generative prior.

4.2.3 Integration

Finally, we integrate the separately optimized foreground and background. To obtain the composite image I_c, we render both foreground image I_{fg} and background image I_{bg} with a foreground mask M, and then blend them by: $I_c = I_{\mathrm{fg}} \odot M + I_{\mathrm{bg}} \odot (1 - M)$. At this stage, to ensure temporal consistency in the final composition, we leverage an existing fMRI-to-video prior (Chen et al., 2024) purely as an optional regularizer. This produces a reference video $\{I_{\mathrm{rv}_k}\}_{k=1}^{T}$ by denoising a noise-perturbed composite video $\{I_{c_k}\}_{k=1}^{T}$. A simple L_2 loss between I_{c_k} and I_{rv_k} is applied to train both 4DGS and spherical 3DGS.

5 Experiments

5.1 Dataset

Our research extends the publicly available fMRI-video dataset (Wen et al., 2018). fMRI data are acquired using a 3T MRI scanner at a repetition time (TR) of 2 seconds, comprising 18 segments of 8-minute video

Table 1: **Quantitative comparison with fMRI-to-visual methods.** Previous native 2D methods lack renderable 3D/4D structure and therefore serve as reference-view ("Front") comparisons. Rather than targeting a new front-view reconstruction benchmark alone, our goal is to pioneer the fMRI-conditioned 4D generation setting, where the generated scene can be rendered from multiple viewpoints over time. The front-view results of Mind4D also reflect that the 4D pipeline preserves the semantic and perceptual information available in the observed stimulus. The best performance is highlighted in **bold**.

Method	View	Video-based		Frame-based			4D-based	
		VCS-2 ↑	VCS-50 ↑	ICS-2 ↑	ICS-50 ↑	SSIM ↑	CLIP-T ↑	CLIP-T-G ↑
MinD-Video (Chen et al., 2024)	Front	0.8545	0.1318	0.7962	0.2061	0.2466	0.9434	-
Mind-Animator (Lu et al., 2025)	Front	0.8727	0.2189	0.8365	0.2655	0.3942	0.8516	-
NeuroClips (Gong et al., 2025a)	Front	0.8827	**0.2291**	0.8368	0.3020	0.3432	0.9409	-
Neurons (Wang et al., 2025)	Front	0.8815	0.2193	0.8573	0.3251	0.3741	0.9502	-
Mind4D (Ours)	Front	**0.8918**	0.2209	**0.8742**	**0.4532**	**0.4262**	**0.9683**	-
	Side	0.8255	0.2009	0.8641	0.3115	0.4129	0.9717	-
	Back	0.8682	0.2336	0.8986	0.3736	0.4249	0.9761	-
	Mean	0.8618	0.2185	0.8790	0.3794	0.4213	0.9720	0.9380

clips, resulting in 4,320 training video-fMRI pairs, and 5 segments for 1,200 testing samples. For each video-fMRI pair, a single frame is randomly selected as the ground truth image for background supervision. Besides, we automatically extracted pseudo-labels for foreground objects (Krizhevsky et al., 2009) and background scenes (Bansal, 2019). Lacking 4D annotations, we employ semantic embeddings of these labels as a codebook to supervise our foreground and background extractors. Following standard practices in fMRI decoding (Chen et al., 2024), all models are trained and evaluated in a subject-specific (within-subject) setting. Since large-scale fMRI-video datasets remain scarce, our experiments should be viewed as a proof-of-concept validation on this acquisition protocol. Generalization across scanners, subjects, and stimulus distributions remains an important direction for future work.

5.2 Metrics

In line with Chen et al. (2024), we employ the Structural Similarity Index Measure (SSIM) for pixel-level accuracy and classification-based score for semantic accuracy with respect to ground truth visual stimuli. The classification score compares the top-1 accuracy between ground truth and rendered frames in selected $N = 2$ and $N = 50$ classes. Both image and video classifiers are used, designed as ICS-N and VCS-N, respectively. A major challenge in brain-to-4D generation is evaluating 3D plausibility without ground-truth 4D scenes. Note that simultaneous multi-view ground truth from the brain is impossible, and the original 3D geometry of the 2D stimulus is unknown. It is impractical to evaluate on the absolute accuracy of the physical 3D geometry. Consequently, our evaluation focuses on multi-view consistency and semantic plausibility rather than metric multi-view accuracy (*i.e.*, the exact 3D shape), which remains a fundamental limitation when learning from 2D stimuli. First, we adopt Multi-view CLIP Similarity (CLIP-T-G), which renders a 360° camera trajectory around the 4D scene and computes the average CLIP similarity across different spatial viewpoints to ensure multi-view semantic consistency. Second, following Yin et al. (2023); Pan et al. (2024b), we incorporate CLIP-T as a dynamic metric, which evaluates temporal smoothness by computing the CLIP similarity between adjacent frames in a rendered video. To further inspect 3D geometric consistency, we conduct a user study by showing videos to 20 participants and asking them to judge how well the content presents 3D/4D geometry. The success rate (SR, ratio of "yes" answers) is used as a perceptual indicator of 3D/4D geometric plausibility. For 4D benchmark, we render a 4D model from the front view (reference view), side views, and back view, with each view evaluated separately in 100 cases.

Table 2: **Quantitative comparison with 3D generation methods.** We compare against Zero123 (Liu et al., 2023), MVDream (Shi et al., 2023), and ZeroNVS (Sargent et al., 2023). Since these baselines are not native fMRI-conditioned methods, they are adapted using decoded fMRI intermediates under the same test samples and comparable optimization budgets. The video-based and frame-based results are averaged across all views. The best scores are highlighted in **bold**.

Method	Video-based		Frame-based			4D-based		
	VCS-2 ↑	VCS-50 ↑	ICS-2 ↑	ICS-50 ↑	SSIM ↑	CLIP-T ↑	CLIP-T-G ↑	SR ↑
Zero123	0.8429	0.1787	0.7372	0.2942	0.3784	0.9531	0.9348	0.6455
MVDream	0.8374	0.1219	0.7144	0.1186	0.2919	0.9517	0.9338	0.5730
ZeroNVS	0.4608	0.0687	0.5668	0.0708	0.2462	0.9288	0.9307	0.2865
Mind4D (Ours)	**0.8618**	**0.2185**	**0.8790**	**0.3794**	**0.4213**	**0.9720**	**0.9380**	**0.8220**

5.3 Implementation details

Our designed high-level encoder $F_{\text{e,h}}$, decoder $F_{\text{d,h}}$, low-level encoder $F_{\text{e,l}}$, decoder $F_{\text{d,l}}$, foreground extractor F_{fg} and background extractor F_{bg} are all MLP structures. The external images $\hat{I}_{extra}$ are selected from COCO (Lin et al., 2014). We set $\lambda_{\text{D}} = 1.0$, $\lambda_{\text{E}} = 0.3, \lambda_{\text{VQ}} = 0.1$ for pre-training. Foreground generative prior uses pretrained models from (AI, 2023; Shi et al., 2023), while the background generative prior employs (Sargent et al., 2023). The fMRI-to-video generative prior exploits structures from (Chen et al., 2024). We set $\lambda_{\text{img}} = 1, \lambda_{\text{sem}} = 0.5, \lambda_{\text{ref}} = 10,000, \lambda_{\text{env}} = 1$. Regarding computational cost, the weakly-supervised pre-training takes ∼40 hours on a single NVIDIA A6000 GPU (requiring 48GB VRAM). The subsequent 4DGS SDS optimization takes ∼30 minutes per scene. Crucially, once generated, the 4D scenes support real-time interactive rendering at >15 FPS, fulfilling the requirement for continuous BCI exploration. See supplementary for further network architecture details.

5.4 4D generation results

We present our 4D generation results in Figure 4, Table 1. Note that side and back views are not observed in the original stimulus. Therefore, they should be interpreted as prior-guided completions rather than ground-truth 4D reconstructions. For comparison with fMRI-to-visual methods, we include results with MinD-Video (Chen et al., 2024), Mind-Animator (Lu et al., 2025), NeuroClips (Gong et al., 2025a), and Neurons (Wang et al., 2025). In the visual results in Figure 4, all previous methods are limited to 2D generation and are frequently inconsistent with visual stimuli. In comparison, our method preserves stronger consistency with the observed stimuli while extending fMRI-conditioned visual generation from single-view outputs to renderable dynamic 3D scenes.

Moreover, our method achieves a higher SSIM (Wang et al., 2004) score from the reference view (front view) as detailed in Table 1. Regarding semantic-level metrics, our method achieves strong performance from the reference front view. We emphasize that Mind4D is designed to evaluate whether fMRI-derived representations can guide 4D scene synthesis, rather than merely competing for a new SOTA on 2D front-view reconstruction. The higher metrics compared to native 2D methods (e.g., MinD-Video, NeuroClips) naturally arise from the integration of powerful downstream 4D generative priors. This indicates that our unified framework successfully preserves the robust perceptual capabilities of SOTA 2D front-end methods. Mind4D fundamentally extends them into a dedicated, neural-guided framework for 4D perception. For CLIP-T scores assessing the 4D effect, our results demonstrate both dynamic and spatial smoothness, outperforming previous methods that focus on single-view output.

We further compare Mind4D against 3D/4D generation baselines, including ZeroNVS (Sargent et al., 2023), Zero123 (Liu et al., 2023), and MVDream (Shi et al., 2023). Existing 3D/4D modeling or multi-view diffusion methods cannot be applied for fMRI-conditioned 4D generation since they do not take fMRI as input. Consequently, we adapt them through the same decoded intermediates used by our pipeline. This is achieved by streamlining two successive steps: (i) fMRI→video: We derive low-level video V_l and text c_h

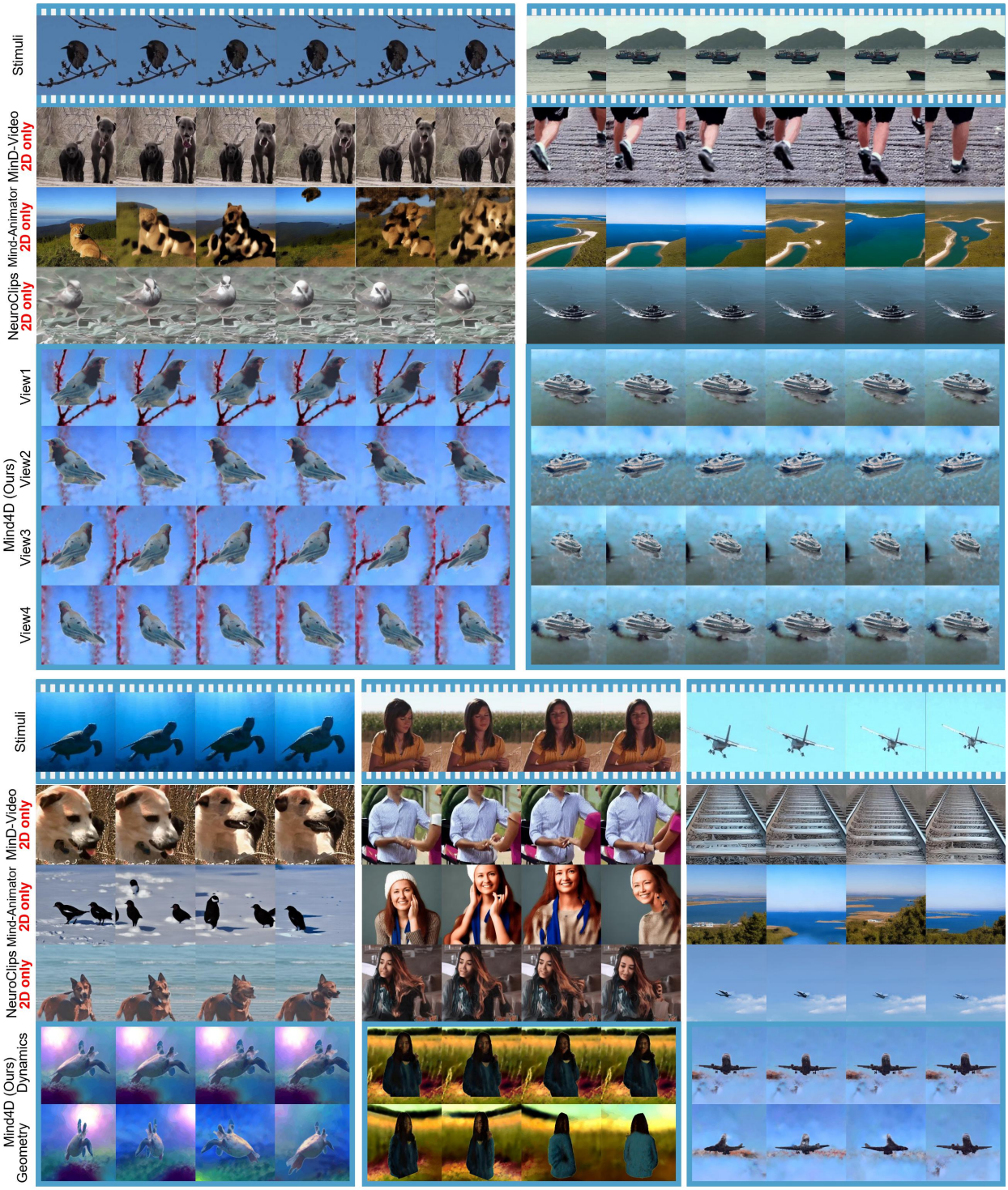

Figure 4: **Multi-view 4D scenarios of Mind4D**. Previous methods (MinD-Video (Chen et al., 2024), Mind-Animator (Lu et al., 2025) and NeuroClips (Gong et al., 2025a)) are limited in **2D** and the results usually do not align well with visual stimuli. In comparison, Mind4D produces dynamic 3D scenes that remain semantically consistent with the observed stimulus while supporting plausible novel-view rendering. Please refer to the supplementary for additional comparisons and visualizations.

Table 3: **Ablation on fMRI condition and representations.** The results show that properly matched fMRI signals are crucial for driving the downstream 4D priors. Our proposed hierarchical representation significantly outperforms single-stream alternatives in guiding 4D synthesis.

Condition	VCS-2 ↑	VCS-50 ↑	ICS-2 ↑	ICS-50 ↑	SSIM ↑
Noise	0.5082	0.0215	0.5036	0.0198	0.1425
Shuffling	0.5275	0.0244	0.5302	0.0227	0.1613
Single-stream	0.8805	0.2118	0.8491	0.3624	0.3817
Mind4D (Ours)	**0.8918**	**0.2209**	**0.8742**	**0.4532**	**0.4262**

using our fMRI decoding methods; (ii) Video→4D: We then use these outputs (V_l or c_h) as direct inputs to a diffusion model to generate the final Gaussians. Specifically, Zero123 receives the decoded reference image as image guidance, MVDream receives the BLIP caption generated from the decoded key frame as text guidance, and ZeroNVS receives the inpainted background/reference-view image for scene-level novel-view synthesis. All adapted baselines use the same test samples, the same fMRI decoder outputs, and the same optimization budget as Mind4D whenever applicable. This adaptation is intended to test whether standard 3D/4D priors can directly consume fMRI-derived visual cues, rather than to claim that these methods were originally designed for neural decoding. Therefore, the comparison measures the effectiveness of the proposed representation-to-4D interface and decomposition strategy, not only the standalone capacity of the underlying 3D/4D generators. As illustrated in Table 2, our Mind4D consistently outperforms all alternatives across visual consistency scores (VCS), identity consistency scores (ICS), and perceptual metrics (SSIM, CLIP-T, CLIP-T-G, and SR).

5.5 Role of fMRI and Hierarchical Representations

To delineate the boundaries between brain decoding and generative priors, we explicitly investigate the complementary roles of fMRI semantics and diffusion priors in 4D scene generation. Recognizing the inherently low temporal resolution of fMRI (e.g., TR = 2s), it is impractical to extract high-frequency kinematic details directly from brain signals. Instead, we assess whether the extracted fMRI representations serve as reliable macroscopic semantic and spatiotemporal anchors, guiding the generative priors that upsample the high-frequency dynamics.

To verify the necessity of the brain signal itself, we conduct critical ablations (Table 3). First, we replace the learned fMRI representation with dimension-matched Gaussian noise ("Noise"). Second, to rigorously demonstrate that the generation is driven by the matched fMRI semantics rather than generic dataset biases, we introduce an fMRI Shuffling control ("Shuffling"), where we randomly pair fMRI signals with mismatched test videos. As shown in Figure 5 and Table 3, removing meaningful fMRI conditioning ("Noise") or using mismatched signals ("Shuffling") leads to catastrophic structural distortions and severe semantic drift.

Furthermore, to demonstrate that our performance gains stem from the proposed hierarchical architecture rather than merely the downstream 4D priors, we construct a strong baseline: **Single-stream**. We replace our asymmetric hierarchical representations with single-stream features from a state-of-the-art 2D method (NeuroClips (Gong et al., 2025a)), while keeping the downstream 4D generation pipeline identical. As shown in Table 3, without our explicit foreground/background decoupling ($z^{vq}_{\text{h,fg}}, z^{vq}_{\text{h,bg}}$) to guide the SDS optimization, the single-stream baseline struggles with semantic ambiguity and yields inferior 4D consistency. This confirms that our proposed representation specifically is what enables effective 4D synthesis.

5.6 Ablation study

Beyond the architectural ablations in Table 3, we further ablate on fMRI encoding, vector quantization, decomposed generation of the foreground and background, and hierarchical representation. The ablation results show that all previous designs are crucial for fMRI extraction and 4D synthesis. See supplementary for details.

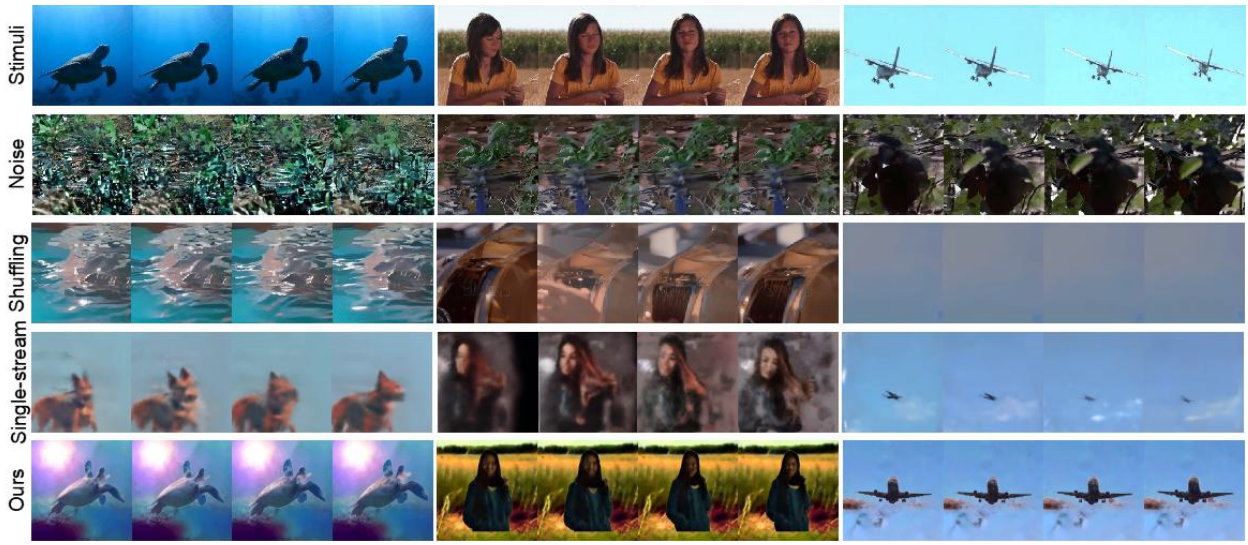

Figure 5: **Examples for conditional importance of fMRI signals.** Clearly, 4D scene generation is genuinely driven by brain signals rather than dominated by generative prior.

6 Conclusion

In this study, we introduce Mind4D, an fMRI-conditioned 4D scene generation framework that synthesizes plausible dynamic 3D scenes from decoded brain representations. The core idea is to learn hierarchical fMRI representations that can serve as conditions for foreground and background generative priors under the absence of paired fMRI-4D supervision. Experiments show that matched fMRI conditioning and hierarchical representations improve semantic alignment, temporal consistency, and multi-view plausibility compared with adapted baselines. We hope this work can open up and foster more advanced research toward richer visual generation from brain signals.

Broader Impact. While Mind4D produces vivid, dynamic 3D scenes, these outputs are prior-guided hypotheses conditioned on noisy fMRI responses, not literal "mind reading." The generated content heavily relies on the hallucination capabilities of generative models and should not be treated as an exact reconstruction of a person's thoughts, memories, or intentions. As brain decoding technologies advance, any potential real-world deployment must strictly prioritize user privacy and explicit informed consent.

References

Sarah N Abdulkader, Ayman Atia, and Mostafa-Sami M Mostafa. Brain computer interfacing: Applications and challenges. *Egyptian Informatics Journal*, 2015.

Stability AI. Stable zero123: Quality 3d object generation from single images. `https://stability.ai/news/stable-zero123-3d-generation`, 2023.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 2022.

Puneet Bansal. Intel image classification. `https://www.kaggle.com/datasets/puneet6060/intel-image-classification`, 2019.

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

Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *CVPR*, 2023.

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

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. In *NeurIPS*, 2024.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023.

Karl Friston. Hierarchical models in the brain. *PLoS computational biology*, 2008.

Jianxiong Gao, Yuqian Fu, Yun Wang, Xuelin Qian, Jianfeng Feng, and Yanwei Fu. Mind-3d: Reconstruct high-quality 3d objects in human brain. *arXiv preprint*, 2023.

Jianxiong Gao, Yuqian Fu, Yun Wang, Xuelin Qian, Jianfeng Feng, and Yanwei Fu. fmri-3d: A comprehensive dataset for enhancing fmri-based 3d reconstruction. *arXiv preprint*, 2024.

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. In *NeurIPS*, 2025a.

Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *AAAI*, 2025b.

Christoph Guger, Nuri Firat Ince, Milena Korostenskaja, and Brendan Z Allison. Brain-computer interface research: A state-of-the-art summary. *Brain-Computer Interface Research: A State-of-the-Art Summary 11*, 2024.

Harry Heft. Affordances and the perception of landscape. *Innovative approaches to researching landscape and health*, 2010.

Jay Hegdé. Time course of visual perception: coarse-to-fine processing and beyond. *Progress in neurobiology*, 2008.

Marc W Howard. Temporal and spatial context in the mind and brain. *Current opinion in behavioral sciences*, 2017.

Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *ECCV*, 2024.

Shuai Jiang, Zhu Meng, Delong Liu, Haiwen Li, Fei Su, and Zhicheng Zhao. Mindshot: Brain decoding framework using only one image. *arXiv preprint*, 2024a.

Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. In *ICLR*, 2024b.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023.

Insub Kim, Sang Wook Hong, Steven K Shevell, and Won Mok Shim. Neural representations of perceptual color experience in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Julian Kiverstein and Erik Rietveld. Scaling-up skilled intentionality to linguistic thought. *Synthese*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*, 2009.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 2018.

Chong Li, Xuelin Qian, Yun Wang, Jingyang Huo, Xiangyang Xue, Yanwei Fu, and Jianfeng Feng. Enhancing cross-subject fmri-to-video decoding with global-local functional alignment. In *ECCV*, 2024.

Chong Li, Jingyang Huo, Weikang Gong, Yanwei Fu, Xiangyang Xue, and Jianfeng Feng. Decofuse: Decomposing and fusing the" what"," where", and" how" for brain-inspired fmri-to-video decoding. *arXiv preprint*, 2025.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023.

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

Pengyu Liu, Guohua Dong, Dan Guo, Kun Li, Fengling Li, Xun Yang, Meng Wang, and Xiaomin Ying. A survey on fmri-based brain decoding for reconstructing multimodal stimuli. *arXiv preprint*, 2025a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023.

Xuanhao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. In *NeurIPS*, 2025b.

Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, and Huiguang He. Animate your thoughts: Decoupled reconstruction of dynamic natural vision from slow brain activity. *ICLR*, 2025.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Eleni Miliotou, Panagiotis Kyriakis, Jason D Hinman, Andrei Irimia, and Paul Bogdan. Generative decoding of visual stimuli. In *ICML*, 2023.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 2011.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Zijie Pan, Jiachen Lu, Xiatian Zhu, and Li Zhang. Enhancing high-resolution 3d generation through pixel-wise gradient clipping. In *ICLR*, 2024a.

Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Fast dynamic 3d object generation from a single-view video. *arXiv preprint*, 2024b.

Dehua Peng, Zhipeng Gui, and Huayi Wu. Interpreting the curse of dimensionality from distance concentration and manifold effect. *arXiv preprint*, 2023.

Kirsten Petras, Sanne Ten Oever, Christianne Jacobs, and Valerie Goffaux. Coarse-to-fine information integration in human vision. *NeuroImage*, 2019.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2020.

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

Mamunur Rashid, Norizam Sulaiman, Anwar PP Abdul Majeed, Rabiu Muazu Musa, Ahmad Fakhri Ab Nasir, Bifta Sama Bari, and Sabira Khatun. Current status, challenges, and possible solutions of eeg-based brain-computer interface: a comprehensive review. *Frontiers in neurorobotics*, 2020.

Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint*, 2023.

Edmund T Rolls and Alessandro Treves. The neuronal encoding of information in the brain. *Progress in neurobiology*, 2011.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Simanto Saha, Khondaker A Mamun, Khawza Ahmed, Raqibul Mostafa, Ganesh R Naik, Sam Darvishi, Ahsan H Khandoker, and Mathias Baumert. Progress in brain computer interface: Challenges and opportunities. *Frontiers in systems neuroscience*, 2021.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint*, 2023.

Daniel R Schonhaut, Zahra M Aghajan, Michael J Kahana, and Itzhak Fried. A neural code for time and space in the human brain. *Cell Reports*, 2023.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *NeurIPS*, 2023.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, 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. In *ICLR*, 2024.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *NeurIPS*, 2021.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint*, 2023.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, 2023.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024.

Frank Tong. Primary visual cortex and visual awareness. *Nature reviews neuroscience*, 2003.

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.

Ranxiao Frances Wang and Elizabeth S Spelke. Human spatial representation: Insights from animals. *Trends in cognitive sciences*, 2002.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *CVPR*, 2024.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. In *IEEE TIP*, 2004.

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 2018.

Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 2002.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *CVPR*, 2023.

Lichao Xu, Minpeng Xu, Tzyy-Ping Jung, and Dong Ming. Review of brain encoding and decoding mechanisms for eeg-based brain–computer interface. *Cognitive neurodynamics*, 2021.

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *ICLR*, 2023.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *CVPR*, 2024.

Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint*, 2023.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023.

Youhui Zhang, Peng Qu, Yu Ji, Weihao Zhang, Guangrong Gao, Guanrui Wang, Sen Song, Guoqi Li, Wenguang Chen, Weimin Zheng, et al. A system hierarchy for brain-inspired computing. *Nature*, 2020.

Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *ICCV*, 2021.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 2018.

A Limitations and future work

This work studies fMRI-conditioned 4D generation under limited supervision, and several limitations remain. We only examine the feasibility of Mind4D on (Wen et al., 2018), and future work should evaluate the framework on additional datasets, subjects, and acquisition settings. We acknowledge several limitations that provide promising avenues for future research:

Temporal resolution and kinematic fidelity. Due to the fundamentally low temporal resolution of fMRI (TR = 2s), the decoded signals primarily provide coarse semantic cues and spatiotemporal trajectories (e.g., "a bird flying"). High-frequency, specific kinematic details (e.g., the precise angle of a left turn) are largely synthesized by the video diffusion priors rather than deterministically decoded from the brain. Future works leveraging higher-resolution recording modalities (e.g., MEG or EEG) could bridge this gap to decode exact kinematic patterns.

Multi-stage error propagation. Our pipeline integrates multiple pretrained generative components to demonstrate the compatibility of the proposed fMRI representations with existing 3D/4D generation modules. However, this multi-stage nature inherently introduces risks of error propagation. For instance, if the VAE decoding produces severe visual artifacts, or the Segment Anything Model (SAM) fails to accurately isolate the foreground object from the background due to complex occlusion, these errors will propagate into the final 4DGS and spherical 3DGS optimization stages. This can result in semantic ambiguity between the foreground and background, or floating artifacts in the 3D scene. A promising future direction is to develop an end-to-end, unified 4D generation model that directly ingests fMRI representations, thereby circumventing the complexities and intermediate failure points of the current multi-stage pipeline.

4D evaluation metrics. Due to the absence of 4D ground truth, current evaluations of geometry are limited to CLIP similarity and the user study. As this study represents the first work in brain-to-4D generation, our focus is on evaluating consistency across semantics, texture, and geometry. This is achieved through multiple widely adopted metrics: VCS-2, VCS-50, ICS-2, ICS-50, SSIM, CLIP-T and CLIP-T-G. Our propose SR in user study partly evaluate geometric quality to some extent. However, more objective geometry metrics on 3D/4D fidelity, such as depth consistency, mesh normal consistency, or multi-view photometric error, remain unexplored. These metrics objective geometry metrics could be further evaluated once 4D ground truth is available.

Generative prior biases and hallucination. The overall quality of the generated 4D content relies heavily on frozen generative priors (AI, 2023; Sargent et al., 2023; Shi et al., 2023; Gong et al., 2025a). It is crucial to distinguish between what is "decoded from the brain" and what is "hallucinated by the generative model." The fMRI representations primarily provide macroscopic semantic anchors ("what" and "where"), while the high-frequency 3D structures, textures, and unseen viewpoints are plausibly hallucinated by the priors. Consequently, when the visual stimuli are out-of-distribution (OOD) with respect to the pre-training datasets of these priors, the generated scenes may exhibit biases or hallucinate incorrect geometries (e.g., generating a generic bird shape instead of the specific species observed). Due to the fundamental absence of 4D ground truth from brain signals, reliance on such external knowledge is necessary in current research, but future work should aim to better disentangle neural decoding from prior-driven hallucination.

Dataset and generalization limitations. As a proof-of-concept for the Brain-to-4D task, our experiments are conducted on a single publicly available fMRI-video dataset (Wen et al., 2018). Following standard fMRI decoding practices, our learned encoders and decoders are subject-specific, meaning they must be trained individually for each participant. Adapting Mind4D to new subjects without extensive retraining would require incorporating cross-subject alignment techniques, which remains an open challenge. Furthermore, the stimulus distribution in the dataset is relatively narrow. We observed that highly unusual scenes, extreme lighting conditions, or rapid, complex actions occasionally lead to generation failures, as they challenge both the fMRI decoding capacity and the downstream generative priors. Expanding to larger, more diverse datasets is a critical next step. **Potential applications.** The current system is an offline exploratory generation framework and is not ready for direct real-world deployment. In the long term, fMRI-conditioned 4D generation may support richer visualization tools for neuroscience, assistive communication, neurorehabilitation, and creative interfaces. Such applications require explicit informed consent, privacy protection,

Figure 6: **Ablation on fMRI encoding.** Our design of fMRI encoders enhances the generalizability of hierarchical representations.

Table 4: **Ablation study on fMRI encoding.** Incorporating the fMRI encoder consistently improves performance across most evaluation metrics.

Encoding	VCS-2	VCS-50	ICS-2	ICS-50	SSIM	CLIP-T	CLIP-T-G	SR
w/o	0.7938	0.1113	0.7587	0.1167	0.3535	0.9707	**0.9382**	0.7957
w/	**0.8618**	**0.2185**	**0.8790**	**0.3794**	**0.4213**	**0.9720**	0.9380	**0.8300**

and careful communication that the generated scenes are plausible visual hypotheses rather than faithful readings of a person's thoughts or intentions. Potential long-term directions include: (1) **Neuroscience visualization.** Exploring how decoded visual representations interact with generative scene priors. (2) **Assistive communication.** Providing richer visual outputs for users who cannot communicate through conventional channels. (3) **Neurorehabilitation.** Constructing controlled dynamic environments for therapeutic or diagnostic research. (4) **Creative interfaces.** Supporting human-in-the-loop visual creation from neural or multimodal signals.

B More ablations

B.1 fMRI encoding

Figure 6 and Table 4 highlights the importance of our designed fMRI encoding process, which significantly improves overall quality. This is achieved by introducing surrogate fMRI from extra data, which can enhance the generalization of our hierarchical representations.

B.2 Impacts of hierarchical representations

We further investigate the impacts of hierarchical representations: semantic representation $z^{vq}_{\mathrm{h,fg}}$ and video V_r derived from representations $z_{\mathrm{h}}, z_{\mathrm{l}}$ on foreground generation, as shown in Figure 9. Since the reference frames are typically out of distribution of the training data (Deitke et al., 2023) used for generative prior models, the baseline "w/o sem" that relies solely on image-based generative prior (with no guidance from semantic representation $z^{vq}_{\mathrm{h,fg}}$ and semantic-based generative prior) fails to produce satisfactory 3D shapes. In addition, the results using only semantic-based generative prior ("w/o img") do not accurately reflect the brain-related images.

B.3 Vector quantized semantic extraction

Figure 7 highlights the crucial role of vector quantization (VQ) semantic representation extraction in our Mind4D framework. Without VQ, high-level representation $z_{h,fg}$ results in ineffective image generation, which has cosine similarity of only 0.073, caused by the high variation with fMRI and data scarcity. In

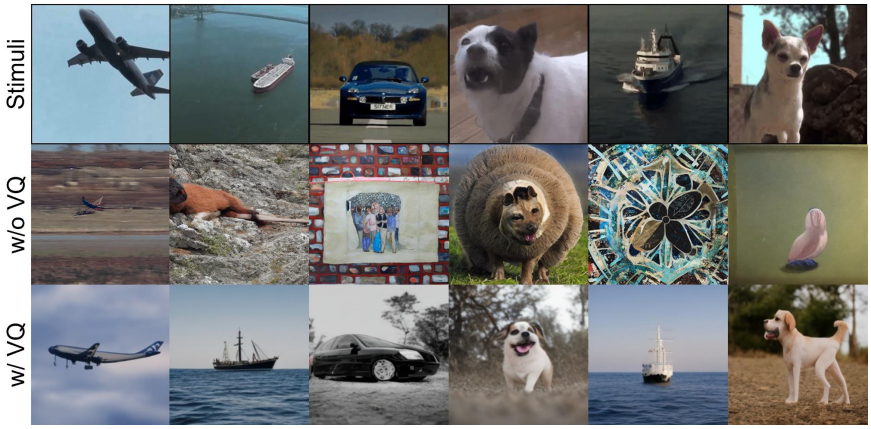

Figure 7: **Ablation on vector quantization (VQ).** VQ significantly improves the stability of hierarchical representations.

Table 5: **Quantitative analysis of the vector quantization (VQ) design.** Our VQ formulation substantially reduces both entropy and KL divergence in foreground and background regions, by several orders of magnitude.

VQ	Entropy (foreground)	Entropy (background)	KL (foreground)	KL (background)
w/o	3.13×10^5	2.58×10^5	5.62×10^2	5.11×10^2
w	**3.38**	**1.79**	1.66×10^{-2}	9.58×10^{-3}

comparison, our designed foreground and background extractors capture the semantic information, with an increased cosine similarity of 0.789, facilitating accurate reproduction of 4D scenes. We further conduct empirical tests on entropy and KL divergence as an empirical validation. As shown in Table 5, our VQ design significantly reduces entropy and KL divergence by four to five orders of magnitude, which is consistent with our theoretical results.

B.4 Ablation on decomposed generation strategy

In Figure 8, we perform an ablation study on the decomposed generation of the foreground and background. Our method adopts both decoupled scene representations (*i.e.*, 4DGS and spherical 3DGS) and decoupled optimization strategy. For ablation, we study the coupled optimization strategy, which employs composite videos as a reference while simultaneously optimizing scene representations. We also investigate the coupled scene representation, which relies solely on 4DGS to jointly model foreground and background. As shown in Figure 8, we find that the coupling of the foreground and background poses the challenge to the optimization of 4D scene, while the decoupled design achieves the best geometry and avoids the ambiguity between the foreground and background.

C Implementation details

C.1 Hierarchical representation learning

The high-level decoder $F_{d,h}$ first employs an MLP to map the fMRI data into a 1024-dimensional vector. This is followed by four MLPs with residual connections to further extract fMRI features. The output is then transformed into 257×768-dimensional shared feature representations, which employs diffusion prior for the high-level representation Z_h. The low-level decoder $F_{d,l}$ utilizes shallow MLP modules to extend fMRI data to match the dimensions of video frame lengths. Subsequently, a Temporal Upsampling module (Gong et al., 2025a) is applied sequentially to generate the embedding z_l, which can then be fed into the Stable

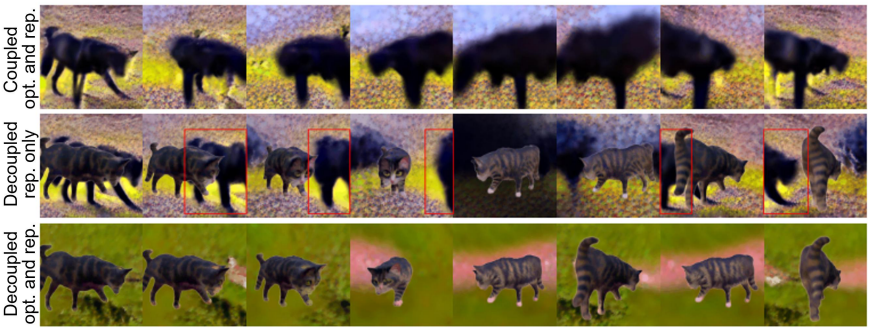

Figure 8: **Ablation on decoupling-coupling.** The "rep." denotes scene representation and the "opt." denotes scene optimization. The coupling of representations leads to bad geometry and the coupling of optimization leads to ambiguity.

Figure 9: **Ablation on the hierarchical representations for foreground generative priors.** Without either semantic (sem) representation $z^{vq}_{\mathrm{h,fg}}$ or guided image (img) generated from $z_{\mathrm{h}}, z_{\mathrm{l}}$ for 3D appearance guidance, the rendering quality decreases significantly.

Diffusion VAE decoder. Both high-level and low-level encoding process begins with inputting image patches into a 12-layer vision transformer with a token size of 768, pre-trained using the DINOv2 (Oquab et al., 2024) to extract features. In the transformer architecture, we choose the query features, as they are known to exhibit more object-centric feature representations to extract pertinent information from the encoder's output, which is used to predict fMRI.

Both the foreground and background extractors use two-layer MLPs to map z_h into $z_{\mathrm{h,fg}}, z_{\mathrm{h,bg}}$. The codebook dimensions for foreground modeling are set to $D_{\mathrm{fg}} = 77 \times 1024$, aligned with Shi et al. (2023), while the background modeling follows Takagi & Nishimoto (2023) with dimensions of $D_{\mathrm{bg}} = 77 \times 768$. Given the practical challenges in acquiring sufficient 4D stimuli for end-to-end optimization, these codebooks are crafted around specific categories of foreground objects (Krizhevsky et al., 2009) and background scenes (Bansal, 2019). The foreground object categories refer to the category of CIFAR-100 (Krizhevsky et al., 2009), while the background scene categories exploit the categories of Intel Image Classification (Bansal, 2019). The number of codewords of foreground and background is set in correspondence with the number of categories in these protocols.

We set $\lambda_{\mathrm{D}} = 1.0$, $\lambda_{\mathrm{E}} = 0.3$ and $\lambda_{\mathrm{VQ}} = 0.1$ for training. The whole high-level encoding process is trained with 150 epochs, while the whole low-level encoding process is trained with 50 epochs. After each 10 epochs, we apply trained encoders to encode extra images from the training set of the Natural Scenes Dataset (NSD) (Allen et al., 2022) to update surrogate fMRI. To encourage temporal alignment between video and

fMRI signals and allow fair comparison with prior work, we downsampled the videos from 30 FPS to 3 FPS. Hence, we adopted a fixed delay for simplicity and consistency, as adopted in (Gong et al., 2025a). Training all fMRI encoders, decoders, and extractors is a one-time process that takes approximately two days on one NVIDIA A6000 GPU. Once completed, the parameters are fixed for subsequent 4D generation from any fMRI.

C.2 Generation

In the generation stage, we implement our pipeline based on the DreamGaussian4D (Ren et al., 2023), a framework focusing on efficient 4D generation. Training involves 500 steps for static foreground and background, 1,000 steps for dynamic foreground, and 50 steps for joint refinement. The Gaussians are initialized with 5,000 random points for foreground inside a sphere of and 200,000 random points for background around a sphere of radius 5. Densification is performed every 50 steps. For balancing weights, we set $\lambda_{\text{img}} = 1, \lambda_{\text{sem}} = 0.5, \lambda_{\text{ref}} = 10,000, \lambda_{\text{env}} = 1$. All hyperparameters are selected based on common experimental tuning, and are set to the same value for all individuals.

For diffusion guidance, we use pretrained models from Stable Zero123 (AI, 2023) and MVDream (Shi et al., 2023) object-level 3D-aware diffusion, adopt ZeroNVS (Sargent et al., 2023) as 2D prior in scene-level 3D-aware diffusion, and apply MinD-Video (Chen et al., 2024) for fMRI-to-video diffusion prior. Considering the unstable training of Gaussians in the generative manner, we follow Pan et al. (2024a) to manually clip the gradient of rendered image pixel-wisely. **Computational Cost.** We clarify the computational cost across different stages. (1) *Weakly-supervised pre-training*: Training all fMRI encoders and extractors is a one-time process per subject, taking approximately 40 hours on a single NVIDIA A6000 GPU. (2) *4D Scene Generation (SDS Optimization)*: The generation pipeline, including the SDS optimization for both foreground 4DGS and background spherical 3DGS, takes about 30 minutes per scene on an A6000 GPU. Considering the complexity of 4D generation, this offline scene-building time is highly competitive. (3) *Real-time Interaction*: Once the 4D scene is generated from the fMRI signal, the lightweight 4DGS representations are saved. Subsequent continuous exploration and rendering of the 4D scene achieve over 15 frames per second (FPS), fully supporting a real-time user experience.

D Further results on 4D generation

Figure 12 exhibits comparison of our method and other fMRI-based video reconstructions. Note that all previous works are only limited in 2D videos. In comparison, Mind4D proposed and realized the 4D generation with both spatial and temporal dimensions from brain signals. Additionally, Figure 11 shows the overrall 4D effects where dynamic images rendered from different viewpoints at different timestamps. Figure 10 shows more samples with subjects 1-3.

E Exploratory visualizations of ROI-conditioned generation

We apply Mind4D to two exploratory visualizations: analyzing region-of-interest (ROI) interpretability and simulating localized signal disruptions (Figure 13). Our design focuses on four specific groups within the visual cortex: primary (V1), associative (V2, V3, V4), dynamic (MT, MST, LIP), and synthesis (TPOJ) visual cortex. For each group, we examine its correlation with the generated 4D priors by masking out other region groups and conditioning the generation on each ROI group separately. To explore the visual effects of neural signal variations, we introduce controlled perturbations to the signals of each group and analyze how the resulting synthesized 4D scenes degrade, providing a novel visualization tool for functional impact. It is important to note that these are exploratory visual simulations of generative prior degradation, not actual neuroscientific facts or clinical diagnostics.

E.1 ROI interpretation via generated scenes

We explore the function of specific Region-of-Interest (ROI) groups (Figure 13(b)). When conditioned solely on the V1 region, the generation captures basic spatial layouts and edges, aligning with V1's role

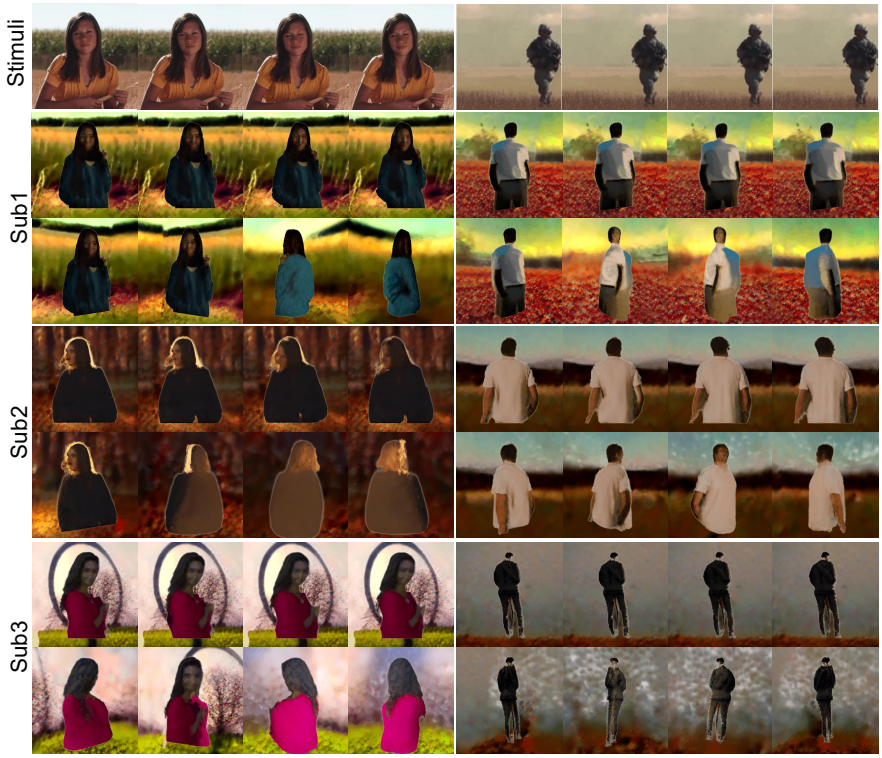

Figure 10: Samples from different subjects.

in processing low-level visual features. Conditioning on associative regions (V2, V3, V4) yields certain texture details but struggles with coherent global structures, reflecting their intermediate processing role. Meanwhile, conditioning on spatiomotion regions (MT, MST, LIP) primarily influences the overall motion and flow of the generated scene, contributing less to static complex patterns. The TPOJ region facilitates a more cohesive visual generation, consistent with its role in information integration. These observations provide a novel, albeit indirect, way to visualize ROI functionality in visual perception (Tong, 2003; Kim et al., 2020).

E.2 Visualizing localized signal perturbations

We demonstrate how our framework visually simulates the effects of localized signal disruptions (Figure 13(c)). By introducing controlled perturbations to specific fMRI regions, we observe distinct degradation patterns in the synthesized 4D scenes. Perturbing the primary (V1) visual regions leads to severe structural collapse. Perturbations in the integration (TPOJ) region result in disjointed scene perception, indicating its importance in maintaining global coherence. In contrast, perturbing spatiomotion regions (MT, MST, LIP) primarily affects temporal consistency and motion direction. These visualizations offer a way to inspect the role of localized neural signals in guiding generative priors.

F Proof of Theorem 4.1

In sparse sampling where the dimensionality of the decoded latent space $d = \dim(z_{\mathrm{h}})$ significantly exceeds the number of training samples n, that is $d \gg n$, the probability distribution $p(z_{\mathrm{h}})$ is not adequately represented. The empirical approximation $\hat{p}_\delta(z_{\mathrm{h}})$ constructed from a limited number of samples fails to capture substantial portions of the probability mass inherent to $p(z_{\mathrm{h}})$.

To avoid the degenerate case where the empirical density is exactly zero outside the observed samples (which would yield an infinite KL divergence), we consider a smoothed empirical density $\hat{p}_\delta(z_{\mathrm{h}})$ (e.g., kernel density

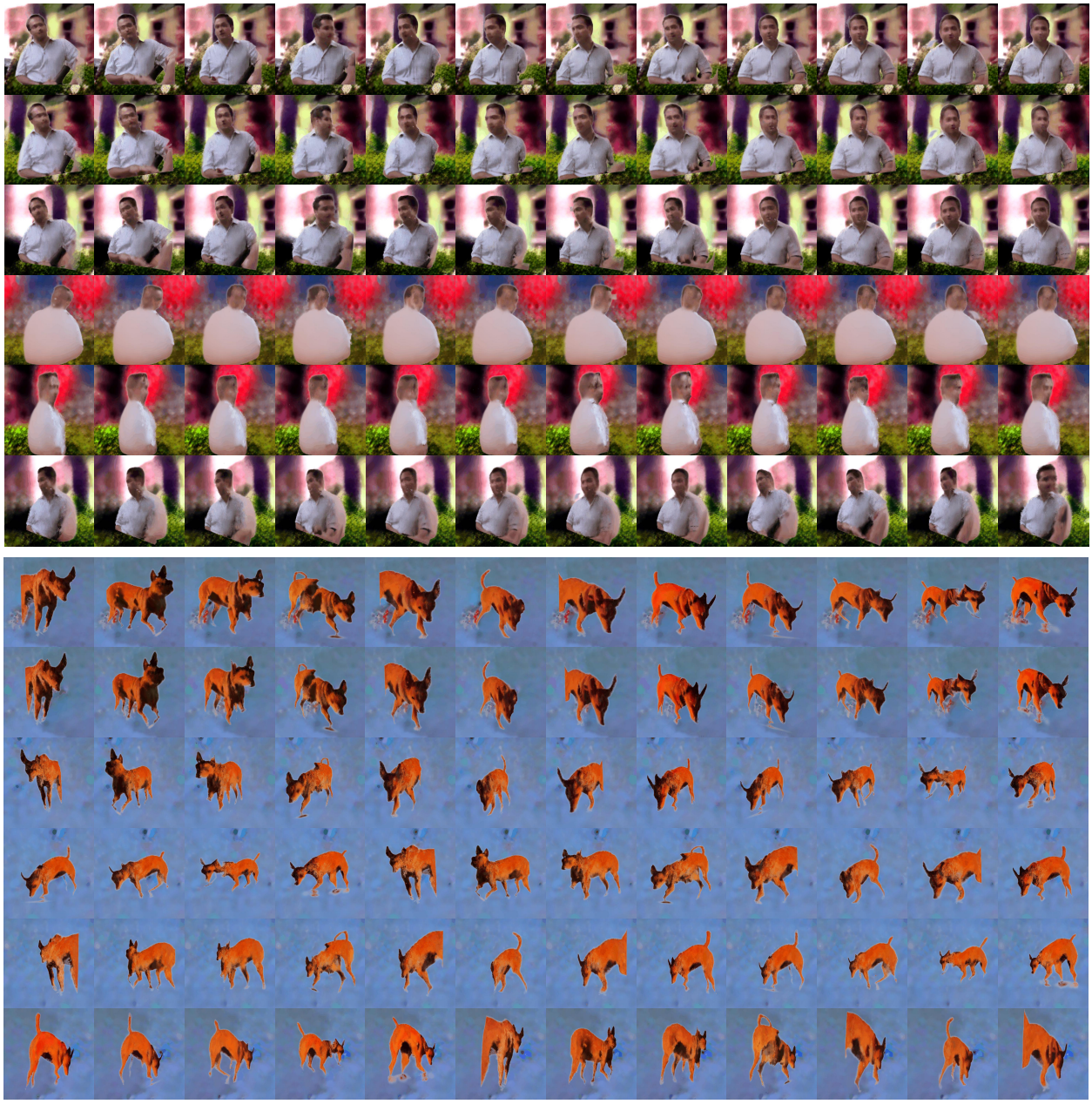

Figure 11: **4D results of two cases.** For each case, we show 6 viewpoints and 12 consecutive frames.

estimate with a small floor $\delta > 0$). In particular, for a neighborhood radius $r > 0$, for points that satisfy $\|z_{\mathrm{h}} - t_i\| > r$ for all $i \in \{1, \ldots, n\}$ (where t_i are the decoded training samples), we assume $\hat{p}_\delta(z_{\mathrm{h}}) \leq \delta$.

Denote R_{in} as the union of all proximal areas around the training samples:

$$R_{\mathrm{in}} = \bigcup_{i=1}^{n} U_i, \quad \text{where } U_i = \{u \in A : \|u - t_i\| \leq r\}, \tag{12}$$

and let R_{out} represent the complement region in the latent space A, which is far from the training samples:

$$R_{\text{out}} = A \setminus R_{\text{in}}. \tag{13}$$

Then the KL divergence without Vector Quantization satisfies:

$$\begin{aligned}
&KL(p(z_{\text{h}}) \,||\, \hat{p}_\delta(z_{\text{h}})) \\
&= \int p(z_{\text{h}}) \log \frac{p(z_{\text{h}})}{\hat{p}_\delta(z_{\text{h}})} dz_{\text{h}} \\
&= \int p(z_{\text{h}}) \log p(z_{\text{h}}) dz_{\text{h}} - \int_{R_{\text{in}}} p(z_{\text{h}}) \log \hat{p}_\delta(z_{\text{h}}) dz_{\text{h}} \\
&\quad - \int_{R_{\text{out}}} p(z_{\text{h}}) \log \hat{p}_\delta(z_{\text{h}}) dz_{\text{h}} \\
&\geq \int p(z_{\text{h}}) \log p(z_{\text{h}}) dz_{\text{h}} - \int_{R_{\text{in}}} p(z_{\text{h}}) \log \hat{p}_\delta(z_{\text{h}}) dz_{\text{h}} \\
&\quad - \int_{R_{\text{out}}} p(z_{\text{h}}) dz_{\text{h}} \cdot \log(\delta) \\
&= O\left(\log \frac{1}{\delta}\right),
\end{aligned} \tag{14}$$

which is relatively large when $\delta \to 0$ (as long as $p(R_{\text{out}})$ is not negligible).

In an ideal scenario where the dataset is sufficiently large and evenly distributed, the region R_{out} diminishes, effectively becoming negligible. Consequently, we could expect that:

$$KL(p(z_{\text{h}}) \,||\, \hat{p}_\delta(z_{\text{h}})) = O(1), \tag{15}$$

as $R_{\text{out}} \to 0$. Conversely, in our setting where fMRI samples are sparse ($n \ll d$), a substantial region of R_{out} persists, indicating a significant divergence in the decoded latent space.

After vector quantization, the number of samples n greatly exceeds the number of quantization bins K. Let $p_k = \mathbb{P}(z_{\text{h}}^{vq} = k)$ and $\hat{p}_k$ be the corresponding empirical frequency (optionally with a small additive smoothing to avoid $\hat{p}_k = 0$). When $n \gg K$, standard concentration implies $\hat{p}_k$ is close to p_k and bounded away from zero for bins with non-negligible mass, yielding a bounded KL divergence:

$$KL(p(z_{\text{h}}^{vq}) \,||\, \hat{p}(z_{\text{h}}^{vq})) \;=\; \sum_{k=1}^{K} p_k \log \frac{p_k}{\hat{p}_k} \;=\; O(1). \tag{16}$$

As a result,

$$KL(p(z_{\text{h}}^{vq}) \,||\, \hat{p}(z_{\text{h}}^{vq})) \ll KL(p(z_{\text{h}}) \,||\, \hat{p}_\delta(z_{\text{h}})) \,. \tag{17}$$

G Proof of Theorem 4.2

Assume that the latent variable $z_{\text{h}} \in \mathbb{R}^d$ lies in a bounded region $A \subset \mathbb{R}^d$, e.g., a hyperrectangle $A = \prod_{i=1}^{d} [a_i, b_i]$ in a pretrained CLIP space (Radford et al., 2021). Consider a uniform grid partition of A into M subrectangles (cells) $\{S_j\}_{j=1}^{M}$ of equal volume ΔV. Let $f(z_{\text{h}})$ be the density of z_{h} on A and define the induced discrete probabilities $p_j = \int_{S_j} f(z_{\text{h}})\, dz_{\text{h}}$.

We define the Riemann-sum discretized entropy (up to an additive constant) as

$$H_\Delta(z_{\text{h}}) \;=\; -\sum_{j=1}^{M} p_j \log \frac{p_j}{\Delta V}. \tag{18}$$

Vector quantization induces a many-to-one mapping from cells to K codebook bins. Let B_k be the set of cell indices assigned to bin k and let $q_k = \sum_{j \in B_k} p_j$. The entropy of the quantized discrete variable is

$$H(z_{\mathrm{h}}^{vq}) \;=\; -\sum_{k=1}^{K} q_k \log q_k. \tag{19}$$

Let J be the cell index random variable with $\mathbb{P}(J = j) = p_j$ and let K be the induced codebook index with $\mathbb{P}(K = k) = q_k$. Since K is a deterministic function of J, we have the standard decomposition

$$H(J) \;=\; H(K) + H(J \,|\, K), \tag{20}$$

which implies $H(J) - H(K) = H(J \,|\, K) \geq 0$, i.e., the quantization (binning) cannot increase entropy. Moreover, for each k, the conditional entropy satisfies $H(J \,|\, K = k) \leq \log |B_k|$, hence

$$0 \leq H(J) - H(K) = H(J \,|\, K) \leq \log J_{\max}, \tag{21}$$

where $J_{\max} = \max_k |B_k|$ is the maximum number of cells mapped to one bin.

Finally, note that $H_\Delta(z_{\mathrm{h}}) = H(J) + \log \frac{1}{\Delta V}$. Therefore,

$$H_\Delta(z_{\mathrm{h}}) - H(z_{\mathrm{h}}^{vq}) \;=\; \log \frac{1}{\Delta V} + \big(H(J) - H(K)\big) \;=\; \log \frac{1}{\Delta V} + O(\log J_{\max}). \tag{22}$$

For a uniform grid in a bounded region, $M = \mathrm{Vol}(A)/\Delta V = O(L^d/\Delta V)$ for some characteristic scale L, and under typical balanced assignments one has $J_{\max} = O(M/K)$, yielding

$$H_\Delta(z_{\mathrm{h}}) - H(z_{\mathrm{h}}^{vq}) \;=\; \log \frac{1}{\Delta V} + O\!\left(\log \frac{M}{K}\right) \;=\; \log \frac{1}{\Delta V} + O\!\left(\log \frac{L^d}{K \Delta V}\right). \tag{23}$$

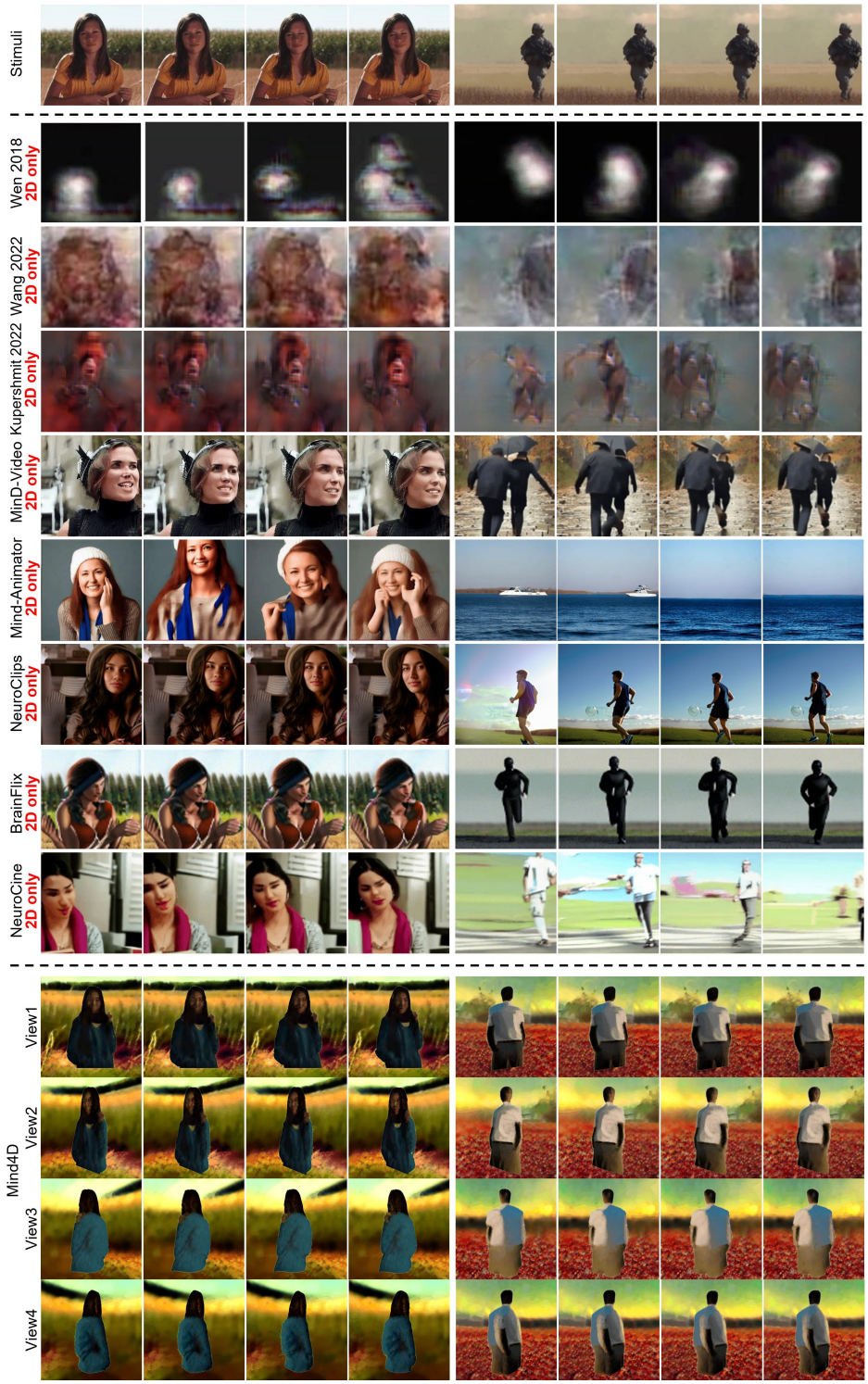

Figure 12: **Comparisons with more fMRI-based video generation.** All previous methods are limited in **2D** when there is only 2D supervision. In comparison, Mind4D pinoeers the **Brain-to-4D** function through a novel weakly supervised framework.

Figure 13: **Exploratory visualization of ROI-conditioned generation.** Our proposed Mind4D enables separate encoding of distinct visual cortex groups, facilitating a novel visualization of how different regions guide 4D generative priors. For each region shown in (a), we first evaluate ROI interpretability by independently applying Mind4D to each ROI and generating corresponding 4D synthesized representations, as illustrated in (b). We introduce controlled perturbations to individual regions and leverage Mind4D to visually explore the effects of these localized disruptions in the synthesized 4D scenes, as shown in (c). Note that these are exploratory visualizations of how generative priors react to perturbed signals, rather than clinical diagnostics.