

Follow-Your-Mind: Towards Inversion-Free Brain-Driven Visual Context Synthesis and Editing

Haodong Jing¹, Panqi Yang¹, Rongchao Zhang², Zhipeng Liu³,
Yajun Liu⁴, Xuehai Bai⁵, Yongqiang Ma^{1*}, and Nanning Zheng^{1*}

¹National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University;

²Peking University; ³Northeastern University; ⁴Shanghai Jiao Tong University;

⁵Independent Researcher

jinghd@stu.xjtu.edu.cn, nnzheng@xjtu.edu.cn

Abstract. Deciphering human visual intent and manipulating visual content via brain signals represents a significant milestone in neuro-generative AI. Recent studies have demonstrated impressive image reconstruction from brain activity, yet they often rely on slow iterative diffusion inversion processes, making it difficult to achieve real-time editing and fails to unify generation with precise context modification. Rectified Flow has recently emerged as a powerful solution for fast, deterministic generation. Inspired by this, we propose MinD-Flow, a pioneering inversion-free framework for unified brain-driven synthesis and editing. It incorporates a Neuro-aware Encoder to achieve unified encoding of heterogeneous brain signals. Combined with the Neuro-Perceiver Bridge (NPB), it aligns brain semantics with visual anchors, mapping the ‘Mind Space’ to the generative latent space. Furthermore, the Decoupled Flow Guidance (DFG) and Adaptive Soft Masking overcome the conflict between structural preservation and semantic changes, enabling precise local editing without noise inversion. We introduce Brain-Gen Benchmark to standardize the evaluation of Neuro-context/creation/editing tasks. Comprehensive experiments show SOTA performance, providing a versatile tool for BCI applications.

Keywords: Brain Decoding · Image Editing · Visual Generation · BCI

1 Introduction

“I dream my painting and I paint my dream.” — Vincent van Gogh

Humans perceive and understand the world through vision, combining it with intrinsic cognition to create countless exquisite works of art [79, 80]. This formidable creativity stems from the brain’s direct reasoning of visual intent, embodying its nature as a powerful “generative model” [11, 50]. Although computer-vision fields have made efforts to simulate this process for contextual image

* Corresponding authors: Nanning Zheng and Yongqiang Ma.

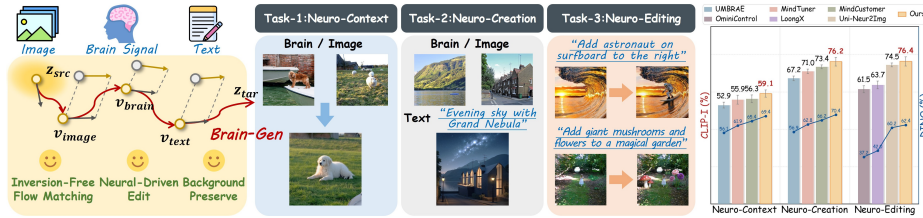


Fig. 1: The **MinD-Flow** based on inversion-free flows to address challenges of consistency and fine-grained control in brain-driven visual manipulation.

synthesis and editing [15, 16, 19, 21, 39, 40, 57, 83], interaction methods reliant on external information have consistently failed to capture the direct psychological intentions. Therefore, directly manipulating visual content using brain signals from cognitive processes represents not only a crucial topic in neuroscience but also a milestone for neuro-generative AI. Achieving this capability holds the promise of transforming BCI from passive “decoding tools” into active “creative partners” [46, 53], enabling truly human-centered visual manipulation.

Driven by this vision, the emergence of generative models and high-quality neural datasets has enabled numerous studies to demonstrate exceptional capabilities in reconstructing images from brain signals [7, 12, 18, 20, 25–27, 54, 64, 66, 71]. Some research has explored using brain signals as conditional prompts for “brain-image fusion” [77, 78] or “brain-driven editing” [4, 87], and constructing multimodal datasets, validating the potential of brain signals as input modalities for image manipulation. However, existing approaches, whether based on LDM [52] or advanced DiT [47] architectures, suffer from inherent fidelity degradation, preventing precise alignment with human intent. Recently, Rectified Flow [37] has emerged as a powerful solution by connecting data distributions via linear ordinary differential equation (ODE), offering an approach to eliminate the noise reversal cycles in diffusion models [72, 76]. Thus, we want to answer a core question: **Can Rectified Flow enable a novel brain-driven paradigm to achieve more general and accurate visual editing?**

Reviewing existing research on brain-driven visual manipulation, we identify two key challenges. First, the diffusion-based “inversion-generation” paradigm struggles to achieve precise local control. Existing methods typically rely on DDIM [58] or Null-Text Inversion [43] to map back to the noise space. This process is complex and the stochastic nature of diffusion trajectories makes it prone to introducing cumulative errors, preventing strict preservation of unedited regions [31]. Consequently, *we need to achieve deterministic, structure-preserving editing*. Second, the balance between encoding heterogeneous brain signals and resolving multimodal conflicts remains unresolved. Brain signals exhibit high spatio-temporal noise [42], it is difficult for MLP or vanilla state-space models [14] to extract robust representations. In multimodal synthesis involving “brain-image-text”, implicit brain signals are often overshadowed by explicit visual or textual

anchors, leading to modality dominance [68]. *We need to develop a better guidance mechanism to strike a balance between original structure and injecting signals.*

Based on the above observations, we propose **MinD-Flow**, the first framework for brain-driven visual context synthesis and editing based on Rectified Flow, as shown in Fig. 1. MinD-Flow comprises three key components. First, we propose the **Neuro-aware Hybrid Encoder**, which unifies heterogeneous brain signals by disentangling spatial topology from fMRI and temporal fluctuations from EEG into a cohesive semantic representation. Second, we introduce the **Neuro-Perceiver Bridge (NPB)** to align these neural semantics with visual anchors. It synthesizes multi-modal features into a compact representation via learnable queries and gated fusion, effectively mapping the “mind space” to the generative latent space. Additionally, we develop the **Inversion-Free Neuro-Flow Solver** featuring **Decoupled Flow Guidance (DFG)** and **Neuro-Adaptive Soft Masking**. This mechanism resolves the conflict between structural preservation and semantic modification, enabling precise local editing and high-fidelity generation in a single forward pass without costly noise inversion. To evaluate the capabilities of MinD-Flow, we extended the popular fMRI dataset NSD [3] and the novel brain-editing dataset L-Mind [87] to define multimodal visual manipulation tasks, which constitute a new benchmark for brain understanding.

Overall, our contributions can be summarized as follows:

- We introduce MinD-Flow, a framework for unified brain-driven visual synthesis and editing based on Rectified Flow. It incorporates Neuro-aware Hybrid Encoder to unify heterogeneous brain signals and employs Decoupled Flow Guidance with Neuro-Adaptive Soft Masking. It resolves the conflict between structural preservation and semantic changes, enabling precise local editing.
- We construct the Brain-Gen Benchmark, integrating the NSD [3] and L-Mind [87] datasets. Unlike standard benchmarks focused on reconstruction, it systematically assesses hierarchical capabilities across three tasks: Neuro-Context, Neuro-Creation, and Neuro-Editing, aiming to standardize the evaluation of brain-driven visual manipulation.
- MinD-Flow achieves SOTA performance, significantly superior in both fidelity and semantic consistency. It achieves precise neural editing with minimal function evaluations (NFE) thanks to inversion-free design, offering real-time performance suitable for practical BCI applications.

2 Related Work

Neural-Driven Visual Decoding. Neural-driven visual decoding aims to decipher brain activity to reconstruct visual perception [36], early pioneering works primarily relied on linear regression [30] or bayesian inference [45] to retrieve or reconstruct simple geometric images, and methods based on CNN and GAN achieved end-to-end reconstruction of natural images [44, 51, 56]. Recently, LDM [52], leveraging its powerful generative capabilities improved generations texture details and semantic consistency. In fMRI decoding, works such as

MinD-Vis [7], the MindEye series [54, 55], UMBRAE [71], NeuroPictor [18], and others [8, 12, 25, 41, 64, 81], have successfully achieved high-fidelity image reconstruction by mapping brain signals to latent spaces of CLIP [48]. Meanwhile, EEG decoding research aims to address challenges in low SNR and spatial resolution using pre-trained models, as evidenced by emerging works such as ATM [33], NICE [60, 61], and CogCap [82], among others [28, 29, 35, 38, 67, 70]. Moreover, recent studies have explored novel architectures to address cross-subject variability and precise representation alignment, thereby mitigating inference latency and enhancing biological plausibility [5, 10, 17]. However, existing work remains far from practical BCI applications. We believe that shifting our perspective to consider how brain signals can be applied to downstream visual tasks will help establish a paradigm more aligned with real-world applications.

Visual Editing with Generative Model. Image editing [16] aims to modify specific regions based on user intent while faithfully preserving the background and global structure. With the ascendance of diffusion models, the technical paradigm has shifted from GAN-based latent manipulation to high-fidelity inpainting leveraging diffusion priors [2]. To balance efficiency and fidelity, training-free paradigms based on attention control have become mainstream [22, 23, 73–75, 85, 86]. Prompt-to-Prompt [15] demonstrated that cross-attention maps govern spatial layout, combined with feature injection strategies like Plug-and-Play [63] and Null-Text Inversion [43], significantly enhance structural consistency. Recent research has evolved towards greater efficiency and precision, works [69, 88] optimize attention mechanisms to achieve pixel-level background preservation. Furthermore, with the emergence of Flow Matching [34] and Rectified Flow, novel methods such as FlowEdit [31], FlowDC [24] aim to address the noise issues inherent in DDIM inversion, theoretically guaranteeing editing reversibility. Shifting the perspective to brain decoding, pioneering studies have explored brain-guided image editing [77, 78] and introduced domain-specific multimodal datasets [87]. However, to advance towards more generalizable neural visual editing, we propose that incorporating Rectified Flow paradigms, grounded in general visual decoding datasets [3], constitutes a highly promising avenue for exploration.

3 Methodology

The **MinD-Flow** framework is shown in Fig. 2, it consists of three components: 1) The Neuro-aware Hybrid Encoder unifies heterogeneous brain signals by disentangling spatial topology and temporal fluctuations. 2) The Neuro-Perceiver Bridge (NPB) synthesizes multi-modal features into a compact representation via learnable queries and gated fusion. 3) The Inversion-Free Neuro-Flow Solver enables deterministic generation and precise editing through Decoupled Flow Guidance (DFG), eliminating the need for inversion loops.

We aim to achieve unified generation across heterogeneous neural modalities from N subjects (each subject providing M samples). For fMRI, let $\mathcal{D}_{\text{fMRI}} = \{\mathbf{x}_{s,i}^f, \mathbf{y}_i\}_{s=1, i=1}^{N, M_f}$, where $\mathbf{x}_{s,i}^f \in \mathbb{R}^V$ represents the voxel-wise activity (ROIs)

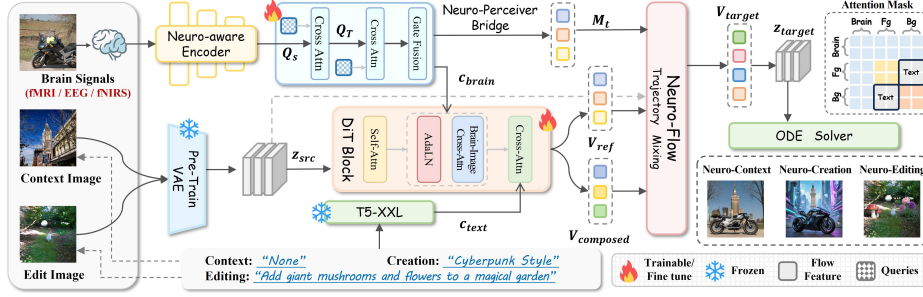


Fig. 2: Overview of MinD-Flow. Neuro-aware encoder processes heterogeneous brain signals using spatial and temporal streams, fusing them into a unified condition. Visual anchors and text prompts are integrated into the DiT block to guide generation. Finally, the solver employs trajectory mixing with an adaptive mask to blend reference and target flows, enabling precise editing without inversion.

for image y_i . For temporal signals, we define $\mathcal{D}_{\text{Temp}} = \{\mathbf{x}_{s,j}^\tau, \mathbf{c}_j, \mathbf{y}_j\}_{s=1,j=1}^{N,M_\tau}$, encompassing both EEG and fNIRS modalities. Here, $\mathbf{x}_{s,j}^\tau \in \mathbb{R}^{C \times T}$ denotes time-series signals (C channels, T timepoints) aligned with editing instruction $\mathbf{c}_j$ and image $\mathbf{y}_j$. Our goal is to learn a shared encoder $\mathcal{E}_{\text{brain}}$ that maps these disparate physical signals $\mathbf{x}$ into a unified semantic manifold aligned with the generative latent space.

3.1 Hybrid Encoding Stage

Neuro-aware Encoder. For heterogeneous brain signals, we treat it as a generalized “Neural Spectrogram” varying along spatial and temporal dimensions.

For the **Spatial Stream (fMRI)**, we employ a Masked Volumetric Autoencoder to capture topological semantics. Given the voxel volume $\mathbf{x}^f$, we apply a random masking operator $\mathcal{M}_\rho$ with ratio $\rho = 0.75$. The visible patches are projected into embeddings and processed by ViT [9]:

$$\mathbf{F}_{\text{spa}} = \text{ViT}_{\text{enc}}(\text{Patch}(\mathbf{x}^f \odot \mathbf{m}) + \mathbf{E}_{\text{pos}}), \quad \mathbf{m} \sim \mathcal{M}_\rho, \quad (1)$$

where $\mathbf{E}_{\text{pos}}$ are learnable positional embeddings. This forces the encoder to infer global context from sparse activation, yielding spatial tokens $\mathbf{F}_{\text{spa}} \in \mathbb{R}^{N_s \times D}$.

For the **Temporal Stream (EEG & fNIRS)**, we utilize a unified 1D-CNN augmented with RoPE [62] to model dynamic neural fluctuations. This stream processes both rapid electrical oscillations (EEG) and slower hemodynamic responses (fNIRS). Unlike absolute encodings, RoPE injects relative position information by rotating the query-key pairs, preserving the phase of neural oscillations and hemodynamic trends. Formally, for a token at time t , the embedding is transformed via a rotation matrix $\mathbf{R}_{\theta,t}$:

$$f_{q,k}(\mathbf{x}_t, t) = \mathbf{R}_{\theta,t}(W_{q,k}\mathbf{x}_t), \quad \mathbf{R}_{\theta,t} = \begin{pmatrix} \cos t\theta & -\sin t\theta \\ \sin t\theta & \cos t\theta \end{pmatrix}. \quad (2)$$

These rotated vectors used to compute self-attention scores, yields time-aware features $\mathbf{F}_{temp} \in \mathbb{R}^{T \times D}$ that explicitly encode the “when” of the editing intent.

Visual & Text Anchors. We leverage pre-trained foundation models as semantic anchors. First, for the visual target, we employ a pre-trained VAE to compress the ground-truth image $\mathbf{y}$ into a low-dimensional latent code $\mathbf{z}_{img} = \mathcal{E}(\mathbf{y})$. We refer to this latent as the visual condition $\mathbf{c}_{img}$. Second, for linguistic guidance in neuro-editing tasks, we utilize a frozen T5-XXL Encoder [49]. Text prompts are converted into sequence embeddings $\mathbf{c}_{text}$. By freezing these anchors, we ensure that the Neuro-aware Encoder learns to map noisy brain activity into a well-structured latent space.

3.2 Neuro-Perceiver Bridge

Learnable Neuro-Queries. To unify the distinct representations from the two streams, we introduce a factorized querying strategy. We define two sets of learnable latent queries: Spatial Queries $\mathbf{Q}_S \in \mathbb{R}^{K \times D}$ and Temporal Queries $\mathbf{Q}_T \in \mathbb{R}^{K \times D}$. $\mathbf{Q}_S$ interacts with spatial features $\mathbf{F}_{spa}$ to extract static semantic context (e.g., object identity), while $\mathbf{Q}_T$ interacts with $\mathbf{F}_{temp}$ to capture dynamic intent (e.g., editing direction). This design acts as an *information bottleneck*: the fixed number of queries ($K \ll N_{voxels}$) forces the model to filter out physiological noise and retain only salient semantics. For a stream $i \in \{S, T\}$, the intermediate embedding $\mathbf{H}_i$ is computed via Cross-Attention:

$$\mathbf{H}_i = \text{Softmax} \left(\frac{\mathbf{Q}_i (\mathbf{F}_i \mathbf{W}_K^i)^T}{\sqrt{d}} \right) (\mathbf{F}_i \mathbf{W}_V^i), \quad (3)$$

where $\mathbf{W}_K^i, \mathbf{W}_V^i$ are projection matrices specific to each modality.

Adaptive Gated Fusion. To synthesize these streams into a unified condition, we propose a learnable gating mechanism. We fuse the spatial and temporal streams using a gate $\lambda = \sigma(\alpha) \in (0, 1)$ to produce the final brain condition $\mathbf{c}_{brain}$:

$$\mathbf{c}_{brain} = \sigma(\alpha) \cdot \mathbf{H}_S + (1 - \sigma(\alpha)) \cdot \mathbf{H}_T, \quad (4)$$

where $\sigma(\cdot)$ is the sigmoid function and α is a learnable scalar initialized to 0. The gate $\sigma(\alpha)$ allows the model to shift its focus flexibly, ensuring that $\mathbf{c}_{brain}$ is a compact, noise-filtered representation ready for the DiT generator.

3.3 Inversion-Free Neuro-Flow Solver

Flow Matching with Neuro-Conditions. Building upon the latent space defined in Sec. 3.1 and the principles of Rectified Flow, we define **Neuro-Flow**. Neuro-Flow models a deterministic Ordinary Differential Equation (ODE) that connects the noise distribution π_0 to the data distribution π_1 , ensuring stable generation. We define a time-dependent vector field v_θ conditioned on the multi-modal set $\mathcal{C} = \{\mathbf{c}_{brain}, \mathbf{c}_{img}, \mathbf{c}_{text}\}$:

$$d\mathbf{z}_t = v_\theta(\mathbf{z}_t, t, \mathcal{C})dt, \quad \mathbf{z}_0 \sim \mathcal{N}(0, \mathbf{I}). \quad (5)$$

For Context or Editing tasks, $\mathbf{c}_{img}$ serves as a visual anchor. To inject these conditions, we employ a dual-conditioning strategy:

- **Global Modulation:** The pooled brain embedding modulates global statistics via *adaLN*, setting the visual tone.
- **Fine-Grained Interaction:** The sequence of brain tokens $\mathbf{c}_{brain}$ interacts with image latents via *Cross-Attention*, guiding spatial synthesis.

This deterministic formulation is crucial, as it allows us to perform editing by modifying the flow trajectory directly, bypassing the error-prone inversion process required by stochastic samplers.

Decoupled Flow Guidance. To harmonize conflicting modalities, structural constraints from reference images ($\mathbf{c}_{img}$), explicit semantics from text ($\mathbf{c}_{text}$), and implicit intents from brain signals ($\mathbf{c}_{brain}$), we propose a **Compositional Flow** strategy for Decoupled Flow Guidance. As illustrated in Fig. 3, directly mixing conditions can lead to semantic conflicts. Instead, we decouple the guidance into separate velocity fields relative to a reference trajectory.

Let $v_{ref}(\mathbf{z}_t, t) = v_\theta(\mathbf{z}_t, t, \mathbf{c}_{img})$ denote the *Reference Flow* that reconstructs the source image. We define the editing forces as deviations from this reference:

$$\Delta v_{brain} = v_\theta(\mathbf{z}_t, t, \mathbf{c}_{brain}) - v_{ref}(\mathbf{z}_t, t), \Delta v_{text} = v_\theta(\mathbf{z}_t, t, \mathbf{c}_{text}) - v_{ref}(\mathbf{z}_t, t). \quad (6)$$

Instead of a single joint prediction, we define the Global Guided Velocity $v_{composed}$ as a weighted composition of these delta flows:

$$v_{composed}(\mathbf{z}_t, t) = v_{ref}(\mathbf{z}_t, t) + \gamma_b \cdot \Delta v_{brain} + \gamma_t \cdot \Delta v_{text}, \quad (7)$$

where γ_b and γ_t are guidance scales controlling the strength of the brain/text influence. This velocity field represents the ideal editing direction that incorporates multi-modal instructions while respecting the structural layout of the source.

Neuro-Adaptive Soft Masking. For the editing task, the goal is to manipulate specific regions driven by brain/text while faithfully preserving the background.

Traditional diffusion-based editing relies on inverting the image to noise, which is computationally expensive and often degrades high-frequency details. To address this, we propose an Inversion-Free Trajectory Mixing strategy.

While standard generation ODE defines $t = 0$ as the noise domain, a direct path connecting the source and target trajectories naturally originates at the source image. Thus, we initialize the ODE with the source data latent: $\mathbf{z}_{t=0}^{edit} = \mathbf{z}_{src}$. Conditioned on the source image, the model implicitly

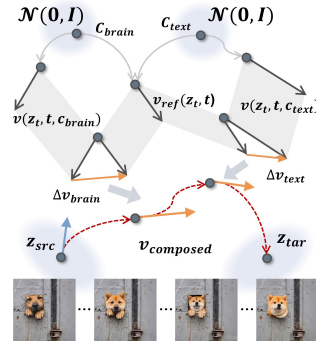


Fig. 3: Image Editing Based on DFG: By jointly deriving a stable Global Guided smoothing field from the observation fields of brain and text, it facilitates faithful attainment to $\mathbf{z}_{tar}$.

targets the source latent itself, predicting a velocity close to zero. Leveraging this property, we compute a **Neuro-Adaptive Mask** $\mathbf{M}_t$ to localize the edit by thresholding the cross-attention maps between the intermediate latent $\mathbf{z}_t$ and the condition $\mathbf{c}_{brain}/\mathbf{c}_{text}$. We define the final editing as a masked deviation from the source flow:

$$d\mathbf{z}_t^{edit} = \underbrace{(v_{ref}(\mathbf{z}_t, t))}_{\text{Source Flow}} + \mathbf{M}_t \odot \underbrace{(v_{composed}(\mathbf{z}_t, t) - v_{ref}(\mathbf{z}_t, t))}_{\text{Delta Flow}(\Delta v)} dt. \quad (8)$$

This formulation is mathematically elegant:

- **Background (Bg)** ($\mathbf{M}_t \approx 0$): The latent trajectory follows v_{ref} exactly, effectively *locking* the background features to the source image.
- **Foreground (Fg)** ($\mathbf{M}_t \approx 1$): The flow transitions to $v_{composed}$, allowing the brain-text composition to drive the generation.

The attention mask is illustrated in Fig. 2, by solving this ODE from $t = 0$ to 1, we achieve precise brain editing in a single forward pass, avoiding the errors inherent in inversion-based methods.

3.4 Training Objectives & Inference

Multi-Modal Flow Matching Objective. Given the limited scale of neural datasets, we only optimize the newly introduced conditioning modules. To enable the *Decoupled Guidance* (Eq. (7)) during inference, we employ stochastic Condition Dropout. We introduce a Bernoulli mask vector $\mathbf{b} \in \{0, 1\}^3$ that randomly replaces elements in the condition set $\mathcal{C}$ with a null token $\emptyset$. This strategy forces the model to learn both the joint distribution $p(\mathbf{z}_1|\mathcal{C})$ and the marginals (e.g., $p(\mathbf{z}_1|\mathbf{c}_{img})$), ensuring robustness against missing modalities. Furthermore, to accelerate semantic convergence, we impose a regularization term aligning the pooled brain representation with the target image’s CLIP embedding. The total objective is formulated as:

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{z}_0, \mathbf{z}_1, \mathbf{b}} [\|v_\theta(\mathbf{z}_t, t, \mathcal{C} \odot \mathbf{b}) - (\mathbf{z}_1 - \mathbf{z}_0)\|^2] + \lambda_{sem}(1 - \cos(\bar{\mathbf{c}}_{brain}, \mathbf{e}_{clip})). \quad (9)$$

Efficient Inference via Euler Solver. During inference, we solve the learned ODE from $t = 0$ to 1 using the Euler method. Unlike diffusion models requiring 50–100 steps, our rectified flow formulation yields nearly straight trajectories, allowing high-fidelity generation with as few as $N = 5$ function evaluations (NFE). For editing, we compute the source trajectory v_{ref} and target trajectory v_{target} in parallel, applying the neuro-adaptive mask $\mathbf{M}_t$ at each step as defined in Eq. (8). This efficiency makes MinD-Flow suitable for real-time applications.

4 Experiments

4.1 Brain-Gen Benchmark

We construct a comprehensive brain-guided generation benchmark, Brain-Gen Benchmark, to systematically assess the controllability and fidelity of brain-driven

Table 1: Neuro-Context. ‘FLUX.1-R/’ refers the blending results obtained by FLUX.1 Redux [13] using reconstructed images and context image. We highlight quantitative comparisons with baselines, and report results across different subjects.

Methods	Pixel-level		Perceptual-level			Semantic-level		
	L1↓	L2↓	LPIPS↓	PDist↓	FID↓	CLIP-I↑	DINO↑	CLIP-IQA↑
FLUX.1-R/MindEye [54]	0.1956	0.0782	0.4215	0.4389	44.83	0.5312	0.5734	0.8645
FLUX.1-R/NeuroPictor [18]	0.2108	0.0845	0.4567	0.4692	49.25	0.5248	0.5421	0.8592
FLUX.1-R/UMBRAE [71]	0.2045	0.0812	0.4398	0.4510	47.66	0.5289	0.5612	0.8610
FLUX.1-R/MindTuner [12]	0.1742	0.0695	0.3581	0.3924	39.12	0.5586	0.6189	0.8850
MindCustomer [77]	0.1583	0.0617	0.3294	0.3418	36.57	0.5630	0.6540	0.8920
MinD-Flow	0.1427	0.0543	0.3068	0.2905	32.18	0.5914	0.6937	0.9103
Subject-1	0.1398	0.0526	0.3012	0.2847	31.24	0.5982	0.7021	0.9134
Subject-2	0.1449	0.0557	0.3114	0.2946	32.89	0.5867	0.6884	0.9076
Subject-5	0.1434	0.0549	0.3087	0.2923	32.56	0.5893	0.6912	0.9092
Subject-7	0.1427	0.0540	0.3059	0.2904	32.03	0.5914	0.6931	0.9111

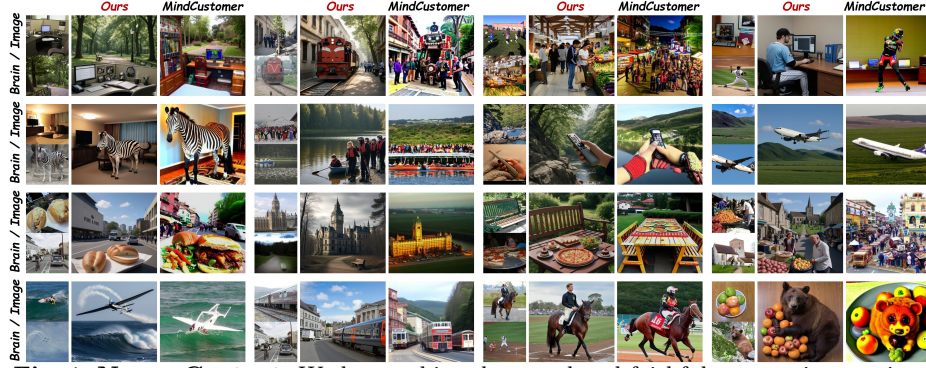


Fig. 4: Neuro-Context. We have achieved natural and faithful context integration.

visual manipulation. We integrate the **Natural Scenes Dataset (NSD)** [3], **L-Mind Dataset** (from LoongX) [87], which provides massive scales of high-quality brain-image-text pairs. We aim to explore the model’s capability to manipulate visual content driven by implicit neural intents through three hierarchical tasks:

- **Neuro-Context.** Adapted from [77], the Dual Contexts Generation task evaluates the fusion of visual cues and fMRI with ‘None’ text prompts. Utilizing the high-fidelity brain representations in NSD, leverage the underlying MS-COCO source to incorporate unpaired images. The objective is to synthesize a coherent image that preserves the structure of the reference image while injecting personalized semantic content derived from brain activity.
- **Neuro-Creation.** This task integrates three modalities: text prompts, reference images, and fMRI signals. The goal is to harmonize explicit semantic instructions with implicit neural contexts. We utilize NSD’s rich triplets, grounded in standard MS-COCO image-text pairs, to assess the model’s capacity for complex multi-modal blending and high-quality generation.
- **Neuro-Editing.** Following protocols in LoongX [87], this task focuses on semantic-aware local modifications driven solely by implicit neural instruc-

Table 2: Neuro-Creation. ‘FLUX.1-K/’ refers the blending results obtained by FLUX.1 Kontext [32] reconstructed images and text-image pairs.

Methods	Pixel-level		Perceptual-level			Semantic-level		
	L1↓	L2↓	LPIPS↓	PDist↓	FID↓	CLIP-I↑	DINO↑	CLIP-T↑
FLUX.1-K/MindEye [54]	0.1843	0.0712	0.3921	0.4215	39.84	0.6842	0.5813	0.2945
FLUX.1-K/NeuroPictor [18]	0.2019	0.0805	0.4317	0.4582	44.12	0.6519	0.5436	0.2912
FLUX.1-K/UMBRAE [71]	0.1904	0.0746	0.4105	0.4328	41.56	0.6724	0.5692	0.2934
FLUX.1-K/MindTuner [12]	0.1628	0.0634	0.3412	0.3745	35.29	0.7105	0.6281	0.3056
MindCustomer [77]	0.1512	0.0589	0.3156	0.3384	33.18	0.7342	0.6615	0.3124
MinD-Flow	0.1387	0.0512	0.2943	0.2916	29.74	0.7618	0.7042	0.3258
Subject-1	0.1354	0.0498	0.2887	0.2854	28.92	0.7684	0.7123	0.3289
Subject-2	0.1412	0.0524	0.2991	0.2967	30.45	0.7562	0.6984	0.3231
Subject-5	0.1396	0.0518	0.2956	0.2932	30.12	0.7601	0.7015	0.3247
Subject-7	0.1385	0.0509	0.2938	0.2911	29.48	0.7625	0.7046	0.3265

Fig. 5: Neuro-Creation. The text context provides additional semantic information for image-brain fusion, enabling effective guided and brain-driven multimodal generation.

tions. The model must decode editing intents from brain signals to modify specific regions of a reference image while strictly maintaining its original spatial layout and background consistency.

Metrics. We employ a multi-dimensional metrics: L1 and L2 for pixel-level fidelity; LPIPS [84] and PDist for perceptual similarity; CLIP-I/T [48], DINO [6], and CLIP-IQA [65] for semantic consistency and text alignment; and FID to assess overall distribution quality.

4.2 Implementation Details

The NPB utilizes $K = 64$ learnable queries per modality to compress multi-modal signals. The generative backbone is based on the MM-DiT from FLUX.1-dev [1]. MinD-Flow is trained on 8 NVIDIA A800 GPUs with batch size of 256 for 50k iterations. We use the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with a learning rate of $1e-4$ to $2e-5$ via a cosine annealing schedule following a linear warmup of 2k steps. The λ_{sem} is set to 0.1. To enable decoupled guidance, we apply stochastic condition dropout with a probability of 0.1 for each modality. During inference, we use the Euler solver with 5 function evaluations (NFE). The guidance scales

are set to $\gamma_b = 3.5$ and $\gamma_t = 4.0$. For Neuro-Editing, the adaptive mask $\mathbf{M}_t$ is obtained by thresholding the cross-attention maps at the 60th percentile.

4.3 Neuro-Context Results

We present quantitative results on the Neuro-Context in Tab. 1. Given the lack of additional baselines, we constructed a baseline by integrating the results of brain decoding methods [12, 18, 54, 71] using FLUX.1, termed ‘FLUX.1/Method’, and we followed the same approach for the Neuro-Creation. Compared to MindCustomer, L1 and L2 decrease by 9.9% and 12.0%, indicating that the inversion-free solver effectively eliminates reconstruction errors typically introduced by diffusion inversion processes. Furthermore, MinD-Flow exhibits outstanding perceptual and semantic alignment, indicating that our generated images are more photorealistic and structurally consistent with the reference. The consistent improvements across 4 subjects further validate the robustness of the encoding strategy against cross-subject variability. Qualitative comparisons in Fig. 4 show that baseline methods often struggle to balance structural preservation and semantic modification, leading to blurred backgrounds or identity loss. MinD-Flow strictly preserves the spatial layout of the reference while accurately rendering the neural intent. This precise control is attributed to DFG, which harmonizes the conflict between visual cues and brain signals.

4.4 Neuro-Creation Results

Evidenced in Tab. 2, MinD-Flow demonstrates superior capability in this complex multi-modal synthesis. Quantitatively, it reduces L1 and L2 by 8.3% and 13.1%, respectively. Crucially, in terms of semantic alignment, our method achieves 4.3% increase in CLIP-T and 3.8% rise in CLIP-I. These gains confirm that the NPB effectively integrates text prompts as high-level semantic guides alongside the fine-grained visual context decoded from fMRI, ensuring that neither modality overshadows the other. Fig. 5 further illustrates this balance. Unlike previous methods that often suffer from semantic conflict, MinD-Flow generates coherent composites. By leveraging Decoupled Flow Guidance, our model ensures that text instructions refine specific attributes (*e.g.*, modifying materials, understanding spatial relationships, and transforming styles) while the brain signals dictate the visual context, resulting in a faithful and unified synthesis across diverse scenes.

4.5 Neuro-Editing Results

As shown in Tab. 3, in Neuro-Editing, MinD-Flow establishes a new SOTA across all input modalities. In the most challenging setting using Pure Neural Signals, our method demonstrates a remarkable advantage over LoongX and the Flow-based method Uni-Neur2Img. Specifically, MinD-Flow (EEG+fNIRS) reduces the L1 by 24.5% and LPIPS by 19.4%, boosting CLIP-I by 2.7%. This substantial margin confirms that the Neuro-aware Encoder effectively disentangles brain dynamics

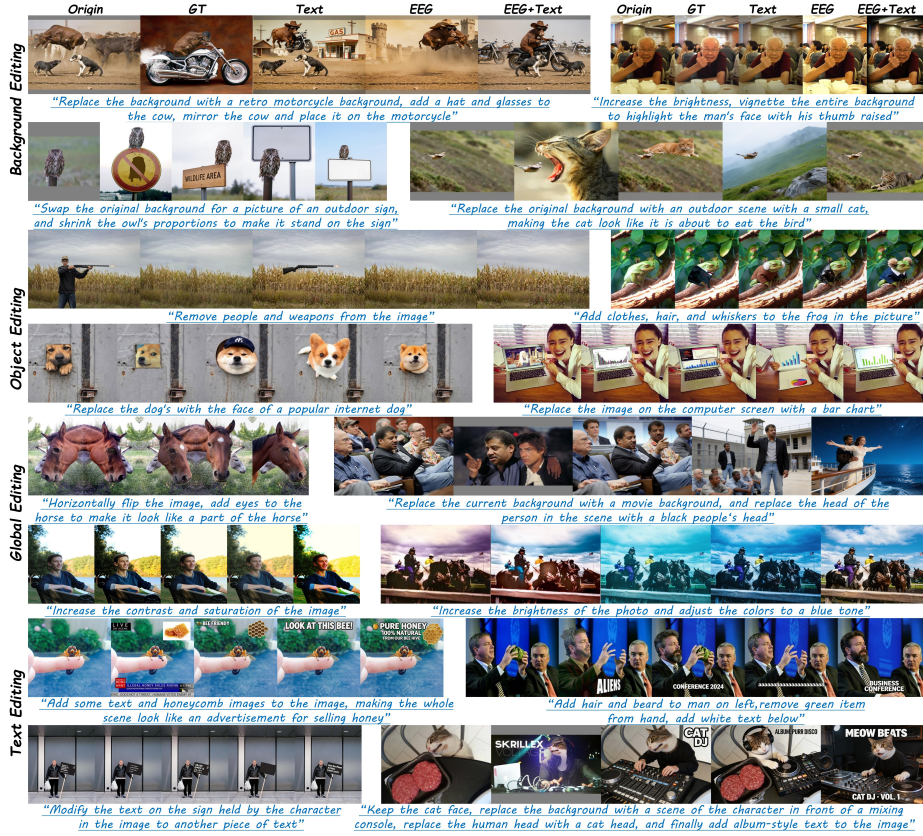


Fig. 6: Neuro-Editing. We present a qualitative comparison of four editing tasks under different modality inputs: (a) background, (b) object, (c) global, and (d) text editing tasks. We achieve consistent rational and faithful editing.

to decode precise editing instructions without text. Furthermore, the integration of text yields the highest overall performance. MinD-Flow (EEG+fNIRS+Text) achieves the best scores (L1: 0.1356, FID: 29.13), significantly outperforming LoongX (Neural+Speech) with a 47.7% reduction in pixel-level error. Fig. 6 validates this versatility across 4 editing scenarios: background, object, global, and text editing. Unlike diffusion-based inversion methods that often introduce artifacts or distort non-edited regions, our Inversion-Free Neuro-Flow Solver, empowered by Neuro-Adaptive Soft Masking, ensures that the background remains pixel-perfect while the foreground is modified faithfully according to the intent.

4.6 Ablation & Discussion

Ablation of Different Modules. We conducted ablation study in Tab. 4 and Fig. 7 involving 4 dimensions: encoding, fusion strategy, flow guidance, and losses.

Table 3: Neuro-Editing. We report quantitative comparisons between our method and baseline across five input conditions: pure text, pure EEG signals, all neural signals (EEG/fNIRS), text-enhanced EEG signals, and text-enhanced neural signals.

	Methods	Pixel-level		Perceptual-level			Semantic-level		
		L1↓	L2↓	LPIPS↓	PDist↓	FID↓	CLIP-I↑	DINO↑	CLIP-T↑
Text	OminiControl (Text)	0.2632	0.1161	0.5821	0.5342	55.37	0.6558	0.4636	0.2549
	OminiControl (Speech)	0.2714	0.1209	0.6134	0.6219	62.84	0.6146	0.3717	0.2501
	Uni-Neur2Img (Text)	0.1827	0.0690	0.4123	0.4287	38.92	0.7259	0.5673	0.2563
	MinD-Flow (Text)	0.1743	0.0612	0.3984	0.4107	36.58	0.7314	0.5782	0.2603
	vs. Uni-Neur2Img (Text)	↓4.6%	↓11.3%	↓3.4%	↓4.2%	↓6.0%	↑0.8%	↑1.9%	↑1.6%
Pure Neural	LoongX (EEG)	0.2641	0.1078	0.6382	0.6947	78.14	0.5457	0.2963	0.2251
	LoongX (All Neural)	0.2509	0.1029	0.5479	0.5166	49.23	0.6605	0.4812	0.2436
	Uni-Neur2Img (EEG)	0.2122	0.0854	0.4836	0.5491	52.68	0.6475	0.4442	0.2280
	MinD-Flow (EEG)	0.1987	0.0765	0.4621	0.5239	48.13	0.6592	0.4576	0.2348
	vs. Uni-Neur2Img (EEG)	↓6.4%	↓10.4%	↓4.4%	↓4.6%	↓8.6%	↑1.8%	↑3.0%	↑3.0%
	MinD-Flow (EEG+fNIRS)	0.1894	0.0718	0.4416	0.4952	44.27	0.6783	0.4965	0.2491
	vs. LoongX (All Neural)	↓24.5%	↓30.2%	↓19.4%	↓4.1%	↓10.1%	↑2.7%	↑3.2%	↑2.3%
Multimodal	LoongX (Neural+Speech)	0.2594	0.1080	0.5712	0.5733	58.41	0.6374	0.4205	0.2588
	Uni-Neur2Img (EEG+Text)	0.1466	0.0551	0.3458	0.3874	32.19	0.7446	0.6020	0.2522
	MinD-Flow (EEG+Text)	0.1402	0.0514	0.3317	0.3729	30.84	0.7512	0.6108	0.2587
	vs. Uni-Neur2Img (EEG+Text)	↓4.4%	↓6.7%	↓4.1%	↓3.7%	↓4.2%	↑0.9%	↑1.5%	↑2.6%
	MinD-Flow (EEG+fNIRS+Text)	0.1356	0.0489	0.3195	0.3604	29.13	0.7638	0.6241	0.2645
	vs. LoongX (Neur+Speech)	↓47.7%	↓54.7%	↓44.1%	↓37.1%	↓50.1%	↑19.8%	↑48.4%	↑2.2%

We compared RoPE with Absolute Position Embedding, yields a consistent performance gain (Creation CLIP-I: 0.6723 vs. 0.6415). This suggests that RoPE better captures the relative dependencies within long-sequence EEG. Furthermore, simple concatenation fails to effectively balance different brain signals, leading to suboptimal semantic alignment. While cross-attention improves interaction, our Adaptive Gated Fusion achieves the best results. By dynamically weighing the contribution of each modality, effectively synthesizes a high-fidelity condition. Comparing Classifier-Free Guidance (CFG) with DFG, we observe a significant improvement in structural (Editing LPIPS: 0.3195 vs. 0.3685) and semantic consistency, results in Fig. 7 further corroborate this. In contrast, the decoupled strategy separates the editing force from the reference flow, allowing precise local manipulation driven by brain/text while strictly maintaining the global layout. Finally, incorporating the $\mathcal{L}_{sem}$ further boosts the CLIP-I score, forcing map noisy signals into the semantic manifold of the CLIP space.

Effect of Inversion-Free. We compare against three inversion-based baselines on editing in Tab. 5: DDIM Inversion, Null-Text Inversion, and standard Flow ODE Inversion. Consequently, our method only uses 5 steps can outperform the strongest baseline (Null-Text, 50 steps) in both pixel-level fidelity and semantic consistency (CLIP-I: 0.7638 vs. 0.7453), this demonstrates that our Inversion-Free method is efficient while maintaining high-fidelity editing. MinD-Flow achieves superior performance without requiring iterative noise mapping, by modeling the editing process as a direct trajectory mixing problem rather than a reconstruction-via-inversion task, we eliminate the quantization errors inherent in inversion loops.

Pareto Frontiers Analysis. To rigorously evaluate the trade-off between structural and semantic, we plot the Pareto frontiers in Fig. 8 (A). MinD-Flow significantly dominates both standard diffusion and the naive baseline. This

Table 4: Ablation of Different Modules. We analyze the impact of each component. ‘Abs.’ refers to Absolute Position Embeddings. ‘CFG’ refers to Classifier-Free Guidance.

Different Ablation Configurations				Neuro-Creation				Neuro-Editing			
Neuro Enc.	Fusion	Guidance	Loss	LPIPS↓	CLIP-I↑	DINO↑	CLIP-T↑	LPIPS↓	CLIP-I↑	DINO↑	CLIP-T↑
Abs.	Concat	CFG	MSE	0.4102	0.6415	0.5328	0.2814	0.4215	0.6482	0.4815	0.2312
Abs.	Concat	CFG	FM	0.3654	0.6723	0.5819	0.2905	0.3842	0.6814	0.5217	0.2405
Abs.	Gated	Decoupled	FM+Sem	0.3105	0.7214	0.6512	0.3189	0.3489	0.7315	0.5912	0.2581
RoPE	Concat	Decoupled	FM+Sem	0.3082	0.7356	0.6715	0.3201	0.3452	0.7402	0.6015	0.2602
RoPE	Cross-Attn	Decoupled	FM+Sem	0.3015	0.7489	0.6892	0.3224	0.3415	0.7510	0.6125	0.2618
RoPE	Gated	CFG	FM+Sem	0.3342	0.7105	0.6421	0.2956	0.3685	0.7015	0.5518	0.2415
RoPE	Gated	Decoupled	FM	0.2998	0.7512	0.6915	0.3185	0.3392	0.7556	0.6152	0.2598
RoPE	Gated	Decoupled	FM+Sem	0.2943	0.7618	0.7042	0.3258	0.3195	0.7638	0.6241	0.2645

**Fig. 7:** Ablation of different modules and generation policies in MinD-Flow.

demonstrates that DFG efficiently injects neural semantics while minimizing structural distortion, a capability that inversion-based methods struggle to achieve. Furthermore, the ‘No Mask’ approach yields high semantic scores but suffers from severe structure loss (high LPIPS), whereas the ‘Hard Mask’ preserves structure but results in poor semantic blending. In contrast, Neuro-Adaptive Soft Masking achieves an *optimal balance*, ensuring seamless integration of brain-driven modifications while strictly respecting the original visual context.

Neural Analysis. We further investigate the biological plausibility in Fig. 8 (B). The ROI-wise importance map reveals early visual areas exhibit strong correlation with structural features, whereas LOC/PPA dominate semantic understanding and scene layout, and the RSC brain region contributes more prominently to the alignment encoding of context. The EEG frequency analysis unveils spectral preferences for different editing. Global editing is primarily driven by low-frequency (θ, α), and fine-grained Object editing relies on high-frequency (β, γ). This finding suggests that visual context updates align with rhythmic cognitive states (*e.g.*, precise object manipulation demands active, high-frequency neural processing).

User Study. To evaluate subjective quality, we conducted a user study with more than 30 participants. Participants evaluated pairwise results based on four criteria: visual quality, background preservation, text alignment, and overall satisfaction. As shown in Tab. 6, MinD-Flow achieves a commanding lead, securing a 74.8% and 73.5% overall preference against MindCustomer and LoongX, and dominates

Table 5: Ablation of Different Generation Policies. We compare MinD-Flow against DDIM Inversion, the Null-Text Inversion, and standard Flow Inversion.

Method	Steps	L1↓	LPIPS↓	CLIP-I↑	DINO↑	CLIP-T↑
DDIM Inv. [58]	10	0.2147	0.4413	0.6158	0.4824	0.2249
	50	0.1853	0.4027	0.6514	0.5138	0.2352
Null-Text Inv. [43]	10	0.1917	0.4108	0.6342	0.5019	0.2316
	50	0.1394	0.3287	0.7453	0.6082	0.2558
Flow ODE Inv. [59]	10	0.1683	0.3846	0.6917	0.5528	0.2414
	50	0.1529	0.3562	0.7154	0.5847	0.2483
Ours	5	0.1356	0.3195	0.7638	0.6241	0.2645

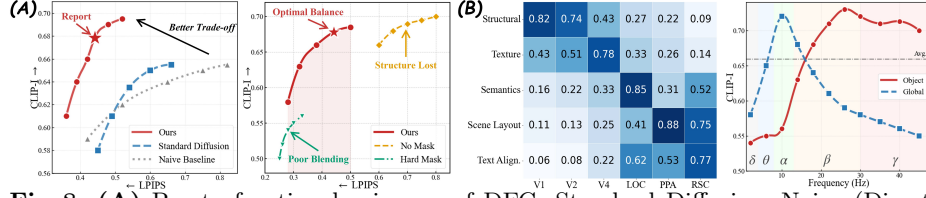


Fig. 8: (A) Pareto frontier dominance of DFG, Standard Diffusion, Naive (Direct Injection), and the different masking methods. We consistently achieve the optimal tradeoff and performance. **(B)** ROI-wise importance maps for Context/Creation tasks, along with the impact of different frequency EEG on Object/Global editing performance.

in background preservation. This confirms that our Inversion-Free Neuro-Flow Solver effectively eliminates the structural distortion common in inversion method, delivering precise editing that aligns strictly with human preferences.

5 Conclusion

We introduce **MinD-Flow**, an innovative inversion-free framework for brain-driven visual context synthesis and editing based on Rectified Flow. Its core idea is a deterministic flow matching paradigm, with the Neuro-aware Hybrid Encoder for unifying heterogeneous brain signals and the Decoupled Flow Guidance combined with Neuro-Adaptive Soft Masking for precise local control. This enables precise local editing driven by neural intents to be executed in a single forward pass, faithfully preserving the original visual structure. Extensive experiments on the proposed Brain-Gen Benchmark demonstrate outstanding performance, paving the way for practical next-generation BCI systems.

Acknowledgments

This work was supported by the New Generation AI-National Science and Technology Major Project No. 2025ZD0124201, the National Natural Science Foundation of China NO. 32541016, 62436005, the Brain Networks and Brain-Inspired Intelligence Science Breakthrough Pilot Project No. JYB2025XDXM504.

Table 6: User Study. Participants were asked to select their preferred results. We report the win rate of MinD-Flow, ‘Method*’ indicates the method using FLUX.1 fusion.

Ours vs.	Quality↑	BackG.↑	Text↑	Overall↑
UMBRAE*	73.5%	96.2%	82.1%	87.4%
MindTuner*	65.9%	92.4%	77.8%	83.6%
MindCustomer	59.3%	89.1%	76.4%	74.8%
LoongX	75.6%	68.9%	71.2%	73.5%
Uni-Neur2Img	60.0%	53.7%	58.6%	61.9%
Random	50.0%	50.0%	50.0%	50.0%

References

1. Black forest labs. flux.1 development model, 2024. 2, 3, 9
2. Alimisis, P., Mademlis, I., Radoglou-Grammatikis, P., Sarigiannidis, P., Papadopoulos, G.T.: Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions. *Artificial Intelligence Review* **58**(4), 112 (2025)
3. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al.: A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience* **25**(1), 116–126 (2022)
4. Bai, X., Yu, R., Xiu, J., Zhou, P., Xia, J., Ji, P.: Uni-neur2img: Unified neural signal-guided image generation, editing, and stylization via diffusion transformers. *arXiv preprint arXiv:2512.18635* (2025)
5. Beliy, R., Zalcher, A., Kogman, J., Wasserman, N., Irani, M.: Brain-it: Image reconstruction from fmri via brain-interaction transformer. *arXiv preprint arXiv:2510.25976* (2025)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9650–9660 (2021)
7. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22710–22720 (2023)
8. Dai, Y., Yao, Z., Song, C., Zheng, Q., Mai, W., Peng, K., Lu, S., Ouyang, W., Yang, J., Wu, J.: Mindaligner: Explicit brain functional alignment for cross-subject visual decoding from limited fmri data. *arXiv preprint arXiv:2502.05034* (2025)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
10. Feng, P., Adeli, H., Guo, W., Cheng, F., Hwang, E., Kriegeskorte, N.: Towards interpretable visual decoding with attention to brain representations. *arXiv preprint arXiv:2509.23566* (2025)
11. Friston, K.: The free-energy principle: a unified brain theory? *Nature reviews neuroscience* **11**(2), 127–138 (2010)
12. Gong, Z., Zhang, Q., Bao, G., Zhu, L., Xu, R., Liu, K., Hu, L., Miao, D.: Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 14247–14255 (2025)
13. Greenberg, O.: Demystifying flux architecture. *arXiv preprint arXiv:2507.09595* (2025)
14. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021)
15. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022)
16. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)

17. Huang, Z., Zhang, E., Cai, Y., Qiu, W., Yang, C., Chen, E., Zhang, X., Ying, R., Zhou, D., Yan, Y.: Seeing through the brain: New insights from decoding visual stimuli with fmri. arXiv preprint arXiv:2510.16196 (2025)
18. Huo, J., Wang, Y., Wang, Y., Qian, X., Li, C., Fu, Y., Feng, J.: Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In: European Conference on Computer Vision. pp. 56–73. Springer (2024)
19. Jia, J., Liu, Y., Sun, Y., Chen, H., Qin, F., Wang, C., Peng, Y.: Geodesic prototype matching via diffusion maps for interpretable fine-grained recognition. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1786–1790. IEEE (2026)
20. Jia, J., Liu, Y., Yang, C., Sun, Y., Qin, F., Wang, C., Peng, Y.: Brain-hgcn: A hyperbolic graph convolutional network for brain functional network analysis. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6391–6395. IEEE (2026)
21. Jia, J., Zheng, H., Wu, Y., Chen, H., Wang, H., Bu, J., Wu, L.: Scout: Selective coupling via optimal unbalanced transport for interpretable text classification. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6413–6431 (2026)
22. Jiang, D., Chen, H., Jing, H., Ma, Y., Zheng, N.: Mrsp: Learn multi-representations of single primitive for compositional zero-shot learning. In: European Conference on Computer Vision. pp. 1–17. Springer (2024)
23. Jiang, D., Jing, H., Ma, Y., Zheng, N.: Beyond image classification: A video benchmark and dual-branch hybrid discrimination framework for compositional zero-shot learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9860–9869 (2025)
24. Jiang, Y., Wang, Z., Wang, Y., Yu, J., Zhuang, Y., Xiao, J., Chen, L.: Flowdc: Flow-based decoupling-decay for complex image editing. arXiv preprint arXiv:2512.11395 (2025)
25. Jing, H., Jiang, D., Ma, Y., Hua, H., Huang, B., Zheng, N.: Beyond brain decoding: Visual-semantic reconstructions to mental creation extension based on fmri. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19258–19268 (2025)
26. Jing, H., Jiang, D., Wang, J., Jia, J., Li, Y., Ma, Y., Zheng, N.: 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. pp. 19033–19044 (2026)
27. Jing, H., Ma, Y., Yang, P., Hua, H., Zheng, N.: Pinpointing visual content: Disentangled features in multimodal model for eeg representation learning and decoding. Knowledge-Based Systems p. 114212 (2025)
28. Jing, H., Ma, Y., Yang, P., Li, H., Huang, S., Chen, B., Zheng, N.: Damind: Zero-shot visual cross-domain alignment and representation for eeg decoding. IEEE Transactions on Image Processing (2026)
29. Jing, H., Yang, P., Jiang, D., Liu, Z., Zheng, N., Ma, Y.: Evoke: Efficient and high-fidelity eeg-to-video reconstruction via decoupling implicit neural representation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40 (2026)
30. Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L.: Identifying natural images from human brain activity. *Nature* **452**(7185), 352–355 (2008)
31. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)

32. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontekst: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
33. Li, D., Wei, C., Li, S., Zou, J., Qin, H., Liu, Q.: Visual decoding and reconstruction via eeg embeddings with guided diffusion. arXiv preprint arXiv:2403.07721 (2024)
34. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
35. Liu, A., Jing, H., Liu, Y., Ma, Y., Zheng, N.: Hidden states in llms improve eeg representation learning and visual decoding. In: ECAI. vol. 392, pp. 2130–7 (2024)
36. Liu, P., Dong, G., Guo, D., Li, K., Li, F., Yang, X., Wang, M., Ying, X.: A survey on fmri-based brain decoding for reconstructing multimodal stimuli. arXiv preprint arXiv:2503.15978 (2025)
37. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
38. Liu, X.H., Liu, Y.K., Wang, Y., Ren, K., Shi, H., Wang, Z., Li, D., Lu, B.L., Zheng, W.L.: Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems* **37**, 72245–72273 (2024)
39. Liu, Z., Duan, P., Chu, Q., Kuhlmann, L., Zhang, C., Yue, W., Tang, X., Zhang, B.: An attributed multiplex network enabled gnn-based stock predictor with observable and non-observable information. *Expert Systems with Applications* p. 129018 (2025)
40. Liu, Z., Duan, P., Xue, X., Zhang, C., Yue, W., Zhang, B.: A dynamic hypergraph attention network: Capturing market-wide spatiotemporal dependencies for stock selection. *Applied Soft Computing* **169**, 112524 (2025)
41. Ma, Y., Zhang, W., Du, M., Jing, H., Zheng, N.: Hierarchical bayesian causality network to extract high-level semantic information in visual cortex. *International Journal of Neural Systems* **34**(01), 2450002 (2024)
42. Mišić, B., Mills, T., Taylor, M.J., McIntosh, A.R.: Brain noise is task dependent and region specific. *Journal of Neurophysiology* **104**(5), 2667–2676 (2010)
43. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6038–6047 (2023)
44. Mozafari, M., Reddy, L., VanRullen, R.: Reconstructing natural scenes from fmri patterns using bigbigan. In: *2020 International joint conference on neural networks (IJCNN)*. pp. 1–8. IEEE (2020)
45. Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L.: Bayesian reconstruction of natural images from human brain activity. *Neuron* **63**(6), 902–915 (2009)
46. Nicolas-Alonso, L.F., Gomez-Gil, J.: Brain computer interfaces, a review. *sensors* **12**(2), 1211–1279 (2012)
47. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4195–4205 (2023)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
49. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**(140), 1–67 (2020)

50. Rao, R.P., Ballard, D.H.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* **2**(1), 79–87 (1999)
51. Ren, Z., Li, J., Xue, X., Li, X., Yang, F., Jiao, Z., Gao, X.: Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage* **228**, 117602 (2021)
52. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10684–10695 (2022)
53. Saha, S., Mamun, K.A., Ahmed, K., Mostafa, R., Naik, G.R., Darvishi, S., Khandoker, A.H., Baumert, M.: Progress in brain computer interface: Challenges and opportunities. *Frontiers in systems neuroscience* **15**, 578875 (2021)
54. Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., et al.: Reconstructing the mind’s eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems* **36**, 24705–24728 (2023)
55. Scotti, P.S., Tripathy, M., Villanueva, C.K.T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K.A., et al.: Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207* (2024)
56. Shen, G., Horikawa, T., Majima, K., Kamitani, Y.: Deep image reconstruction from human brain activity. *PLoS computational biology* **15**(1), e1006633 (2019)
57. Shuai, X., Ding, H., Ma, X., Tu, R., Jiang, Y.G., Tao, D.: A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555* (2024)
58. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020)
59. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020)
60. Song, Y., Liu, B., Li, X., Shi, N., Wang, Y., Gao, X.: Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234* (2023)
61. Song, Y., Wang, Y., He, H., Gao, X.: Recognizing natural images from eeg with language-guided contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems* (2025)
62. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024)
63. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 1921–1930 (2023)
64. Wang, H., Lu, J., Li, H., Li, X.: Zebra: Towards zero-shot cross-subject generalization for universal brain visual decoding. *arXiv preprint arXiv:2510.27128* (2025)
65. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 37, pp. 2555–2563 (2023)
66. Wei, Y., Ma, C., Gao, J., Wang, C., Zhang, S., Gong, B., Tan, S., Yuan, H., Shan, H.: 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*. pp. 28211–28223 (2026)

67. Wu, H., Li, Q., Zhang, C., He, Z., Ying, X.: Bridging the vision-brain gap with an uncertainty-aware blur prior. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 2246–2257 (2025)
68. Wu, H., Tang, M., Zheng, X., Jiang, H.: When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552* (2025)
69. Wu, J., Li, Z., Qin, H., Liu, X., Kong, L., Zhang, Y., Yang, X.: Flashedit: Decoupling speed, structure, and semantics for precise image editing. *arXiv preprint arXiv:2509.22244* (2025)
70. Wu, L., Li, J., Ren, Z., Zhang, K., Gao, X.: Shrinking the teacher: An adaptive teaching paradigm for asymmetric eeg-vision alignment. *arXiv preprint arXiv:2511.11422* (2025)
71. Xia, W., de Charette, R., Oztireli, C., Xue, J.H.: Umbrae: Unified multimodal brain decoding. In: *European Conference on Computer Vision*. pp. 242–259. Springer (2024)
72. Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., Jiang, Y.G.: A survey on video diffusion models. *ACM Computing Surveys* **57**(2), 1–42 (2024)
73. Yang, P., Jing, H., Chao, J., Xiang, T., Lin, L., Hu, Y., Luo, Y., Ma, Y.: Muse: Resolving manifold misalignment in visual tokenization via topological orthogonality. *arXiv preprint arXiv:2605.05646* (2026)
74. Yang, P., Jing, H., Zheng, N., Ma, Y.: Unihoi: Unified human-object interaction understanding via unified token space. *arXiv preprint arXiv:2511.15046* (2025)
75. Yang, P., Jing, H., Zheng, N., Ma, Y.: Unibvr: Balancing visual and reasoning abilities in unified 3d scene understanding. *Neurocomputing* p. 132599 (2026)
76. Yang, Y., Jin, M., Wen, H., Zhang, C., Liang, Y., Ma, L., Wang, Y., Liu, C., Yang, B., Xu, Z., et al.: A survey on diffusion models for time series and spatio-temporal data. *ACM Computing Surveys* **58**(8), 1–39 (2026)
77. Yu, M., Lin, S., Ma, L., Lei, B., Ma, K.: Mindcustomer: Multi-context image generation blended with brain signal. In: *International Conference on Machine Learning*. pp. 73197–73220. PMLR (2025)
78. Yu, M., Lin, S., Yan, H., Ma, K.: Mindpainter: Efficient brain-conditioned painting of natural images via cross-modal self-supervised learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 14468–14476 (2025)
79. Zeki, S.: Art and the brain. *Journal of Consciousness Studies* **6**(6-7), 76–96 (1999)
80. Zeki, S.: Art and the brain. In: *The Brain*, pp. 71–103. Routledge (2017)
81. Zeng, B., Li, S., Liu, X., Gao, S., Jiang, X., Tang, X., Hu, Y., Liu, J., Zhang, B.: Controllable mind visual diffusion model. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 38, pp. 6935–6943 (2024)
82. Zhang, K., He, L., Jiang, X., Lu, W., Wang, D., Gao, X.: Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 14486–14493 (2025)
83. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 3836–3847 (2023)
84. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)
85. Zhang, X., Jing, H., Chen, H., Ma, Y., Zheng, N.: Refiner: Fine-grained cross-modal concepts refinement for compositional zero-shot learning. In: *ICASSP 2025*

- 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
- 86. Zhang, X., Jing, H., Ma, Y., Zheng, N.: Decoupling primitive with experts: Dynamic feature alignment for compositional zero-shot learning. In: The Fourteenth International Conference on Learning Representations
 - 87. Zhou, P., Xia, J., Peng, X., Zhao, W., Ye, Z., Li, Z., Yang, S., Pan, J., Chen, Y., Wang, Z., et al.: Neural-driven image editing. arXiv preprint arXiv:2507.05397 (2025)
 - 88. Zhu, T., Zhang, S., Shao, J., Tang, Y.: Kv-edit: Training-free image editing for precise background preservation. arXiv preprint arXiv:2502.17363 (2025)