EEGRFusion: Uncertainty-Aware EEG-to-Image Evidence for Adjunct Bedside Assessment of CMD in Disorders of Consciousness

Chenyuan Hong[1], Yueqin Diao[1], Lei Wang[1], Ziyan Huang[1], Jiongning Zhao[1], Hanyi Yu[1], and Yanwu Xu[1,2⋆]

[1] South China University of Technology, Guangzhou, Guangdong, China
[2] Pazhou Lab, Guangzhou, Guangdong, China
ywxu@ieee.org

Abstract. Assessments for disorders of consciousness are frequently confounded by absent or inconsistent behavioral output, particularly in cognitive motor dissociation, motivating objective and interpretable EEG-based visual-semantic evidence as an adjunct to bedside evaluation. We present EEGRFusion, an EEG-to-image framework that projects EEG into a visual-semantic space and enables (i) zero-shot image retrieval and (ii) EEG-conditioned image generation, together providing clinician-reviewable cues of stimulus-evoked processing. For robust EEG representation, we introduce MAMD, a montage-aware temporal dual-path encoder with a Mamba backbone. For generation, to address the CLIP semantic bottleneck, EEG variability and related issues, we learn an EEG-to-CLIP conditional prior using Rectified Flow and decode it with a frozen diffusion backbone equipped with an IP-Control Fusion Adapter to mitigate semantic and structural drifts. Experiments on THINGS-EEG demonstrate improvements in retrieval accuracy and reconstruction quality over baselines. We further provide uncertainty heatmaps demonstrating multi-sample variability to facilitate risk-aware interpretation. The code is available at https://github.com/hcy1026/EEGRFusion.

Keywords: EEG-to-image generation · Disorders of consciousness · Cognitive motor dissociation · Robust EEG encoding · Control adapter.

1 Introduction

Disorders of consciousness (DoC) encompass clinical entities such as unresponsive wakefulness syndrome, minimally conscious state, and related syndromes, becoming a global clinical priority [7]. Although the Coma Recovery Scale–Revised [6] is the clinical diagnosis criterion, it depends on overt behavior that may be inconsistent in patients with **cognitive motor dissociation (CMD)**. Consequently, diagnostic errors persist. Previous clinical reviews report up to 30% misclassification between clinical entities and up to 25% of DoC

⋆ Corresponding author

patients may still exhibit covert neural signatures [16, 1]. Taken together, these observations underscore the urgent need for clinically interpretable neurophysiological evidence to aid behavioral assessment in patients with CMD.

Importantly, neural visual decoding and image reconstruction are supported by recent progress in contrastive learning and generative modeling, notably CLIP and diffusion, that substantially improve semantic fidelity and perceptual realism [22, 17, 19, 25]. Although recent studies have demonstrated high-quality image reconstruction from fMRI [17], fMRI has limited clinical scalability. In the bedside CMD setting, EEG-to-image is therefore an appealing, reviewable form. However, making EEG a reliable conditioning signal for a high-capacity image generator remains challenging [22, 17, 13, 2]. CLIP spaces provide a strong anchor for high-level semantics, whereas EEG contains abundant low-level perceptual traces, which can be suppressed and lead to category-level generations rather than instance-faithful reconstructions. Moreover, real-world EEG exhibits substantial trial-to-trial variability even within a subject. Consequently, mapping EEG to a single deterministic image embedding can be brittle under regression-style objectives, and the resulting instability may further manifest as structural drift and inconsistency during diffusion decoding.

Motivated by these practical constraints, we present **EEGRFusion** as an EEG-to-image framework. The goal is not to visualize raw EEG waveforms, but to decode visual evidence by retrieving the viewed image and reconstructing a stimulus-consistent image from EEG responses. We learn an EEG-to-CLIP conditional prior via **Rectified Flow**, allowing trial-to-trial variability to be represented as a conditional distribution rather than being forced into a single point estimate [15]. During diffusion decoding, we retain CLIP for high-level content, while introducing a lightweight **Control** module to inject low-level structural guidance and reduce layout drift [26, 20, 19, 25]. We report multi-sample **uncertainty heatmaps** given the uncertainty-bearing nature of EEG. Since EEG cannot directly serve as a reliable conditioning signal for a high-capacity generator, we build a stable EEG backbone by encoding EEG into robust representations via a **montage-aware temporal dual** tokenization and a long-sequence **Bi-Mamba** backbone [4, 29], improving robustness under low SNR, montage perturbations, and channel dropout. In future clinical visual stimulus paradigms, we interpret retrieval-based N-way identification and generation-based reconstructions as clinician-reviewable readouts of stimulus-evoked processing in DoC.

We summarize our contributions as follows:

1. We propose EEGRFusion, an EEG-to-image framework that produces clinician-reviewable, uncertainty-aware retrieval and generation outputs as potential adjunct evidence for future bedside assessment in DoC, particularly CMD.
2. We introduce MAMD, a montage-aware EEG encoder with temporal dual tokenization that improves robustness under bedside EEG conditions. We further use retrieval as a probe of embedding quality.
3. We develop RFFA, a Rectified-Flow-based prior and Fusion Adapter controlled diffusion generator that trains a Control Adapter reducing structural drift and enabling heatmaps for improved clinical reviewability.

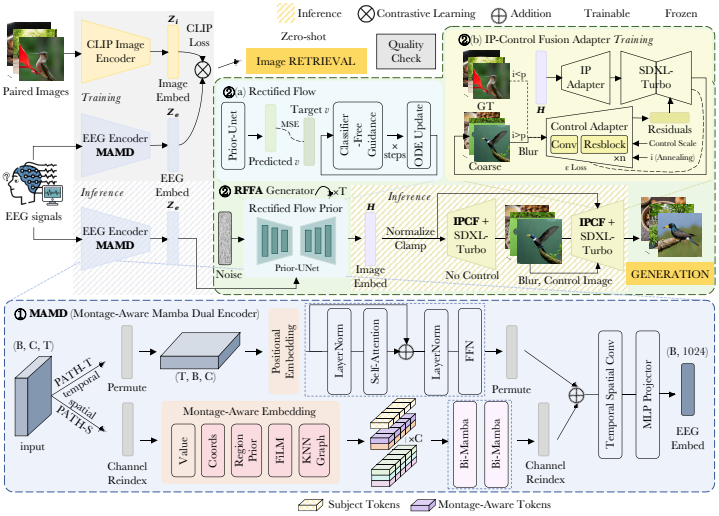

Fig. 1. Overview of EEGRFusion. Our method comprises three components: (i) a MAMD (Montage-Aware Mamba Dual) encoder; (ii) an RFFA (Rectified-Flow prior and Fusion Adapter) generator that learns a Rectified Flow prior; and (iii) an IP-Control Fusion adapter with SDXL-Turbo.

2 Methodology

Our model **EEGRFusion** maps EEG signals into an interpretable and generative visual-semantic space, enabling two complementary capabilities: (i) EEG-to-image retrieval and (ii) EEG-to-image generation. As illustrated in **Fig. 1**, EEGRFusion comprises three components: (i) a MAMD (Montage-Aware Mamba Dual) encoder aligning EEG and image embeddings; (ii) an RFFA (Rectified-Flow prior and Fusion Adapter) generator that learns a Rectified Flow prior to model EEG and image embeddings; and (iii) an IP-Control Fusion adapter with SDXL-Turbo for coarse-to-fine image generation refinement, improving structural stability and visual fidelity.

2.1 MAMD: Montage-Aware Mamba Dual Encoder

Electroencephalogram (EEG) differs fundamentally from typical modalities, since it is an inherently coupled **spatiotemporal** signal. In addition, practical EEG introduces low SNR, subject variability and montage perturbations. Motivated by these properties, we design MAMD as a dual-path encoder: **Temporal Path** models long-range temporal dependencies for global semantic dynamics, and **Spatial (Montage) Path** injects electrode geometry and regional priors to enforce topographic consistency and montage robustness.

To maintain consistency in symbols, let a batch of EEG segments be denoted as $\mathbf{X} \in \mathbb{R}^{B\times C\times T}$, where B is the batch size, C is the number of channels, and T is

the number of temporal samples. A frozen CLIP image encoder E_{CLIP} produces image embeddings $\mathbf{Z}_i \in \mathbb{R}^{B\times d}$ from images, while a trainable EEG encoder MAMD, denoted as f_θ, produces EEG embeddings $\mathbf{Z}_e = f_\theta(\mathbf{X}) \in \mathbb{R}^{B\times d}$.

In the **Temporal Path**, we treat time as the token dimension and channels as the feature dimension. Specifically, omitting the batch dimension, we permute each EEG sample from $\mathbb{R}^{C\times T}$ to $\mathbb{R}^{T\times C}$, apply a linear projection and add positional encoding. The resulting sequence is processed by a stack of linear and self-attention blocks followed by a reverse permutation to produce a temporal representation $\mathbf{U}_T$. In the **Spatial (Montage) Path**, we first standardize channel indexing to an internal montage order and convert each channel segment into a token via value projection. Each token is then augmented with **montage-aware priors**: Fourier-feature encoding maps electrode coordinates into multi-frequency sinusoidal features and coarse region priors group electrodes into broad scalp regions. We further perform **FiLM** [18] for token modulation, and **kNN graph mixing** in electrode space to propagate local neighborhood information [24]. The resulting sequence is then processed by **Bi-Mamba** blocks, leveraging linear-time selective state-space modeling with bidirectional scanning [29] to produce a montage-aware representation $\mathbf{U}_M$, which is fused with $\mathbf{U}_T$ by element-wise addition. Finally, a lightweight temporal-spatial convolution and a projection head produce the EEG embedding $\mathbf{Z}_e$.

For cross-modal contrastive learning, given paired EEG and image embeddings $\{(Z_k^e, Z_k^i)\}_{k=1}^B$ in a batch, we compute cosine similarity in the shared latent space and adopt CLIP-style symmetric InfoNCE loss:

$$L_{\text{C}} = -\frac{1}{2B}\sum_{k=1}^{B}\left[\log\frac{\exp(\text{sim}(Z_k^e, Z_k^i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(Z_k^e, Z_j^i)/\tau)} + \log\frac{\exp(\text{sim}(Z_k^i, Z_k^e)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(Z_k^i, Z_j^e)/\tau)}\right] \tag{1}$$

where τ is the learnable temperature. Following Radford et al. [20], we tailor the objective to tasks: retrieval uses only the CLIP loss to align EEG-image pairs to improve representation discriminability, while generation adds a Mean Squared Error (MSE) term $L_{\text{M}} = \frac{1}{B}\sum_{k=1}^{B}\|Z_k^e - Z_k^i\|_2^2$ to enforce regression consistency:

$$Loss = \lambda \cdot L_{\text{C}} + (1-\lambda)\cdot L_{\text{M}} \tag{2}$$

where λ is a balancing hyperparameter that weights the CLIP and MSE losses.

2.2 RFFA–Rectified Flow Prior in the Embedding Space

With EEG-image alignment established, the generation problem can be reframed as sampling an image embedding $\mathbf{H} \in \mathbb{R}^d$ from a conditional distribution $p_\theta(\mathbf{Z}_i \mid \mathbf{Z}_e)$. Taking the idea from Mind's Eye [21] and ATMS [13], we learn a conditional generative prior in the CLIP embedding space [2]. Rectified Flow models generation as integrating an Ordinary Differential Equation (ODE) velocity field from noise to data, offering stable training and efficient sampling [15].

During training, we define the linear interpolation and the target velocity:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad \mathbf{v}^*(\mathbf{x}_t, t) = \mathbf{x}_1 - \mathbf{x}_0 \tag{3}$$

where $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the noise endpoint, $\mathbf{x}_1 = \mathbf{Z}_i$ is the target CLIP image-embedding endpoint, $t \sim \mathcal{U}(0, 1)$ is a uniformly sampled time point, $\mathbf{x}_t$ is the linear interpolation at time t, and $\mathbf{v}^*()$ is the target velocity. A conditional Prior-UNet predicts $\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{Z}_e)$ and is trained by MSE loss against the target velocity. At inference, we integrate the learned ODE starting from $\mathbf{x}_{t=0}$ to finally obtain $\hat{\mathbf{h}} = \mathbf{x}_{t=1}$. To improve conditional fidelity, we use classifier-free guidance [11] to combine conditional and unconditional velocity predictions:

$$\mathbf{v}_{\text{cfg}}(\mathbf{x}_t, t, \mathbf{Z}_e) = (1 + w) \cdot \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{Z}_e) - w \cdot \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing) \tag{4}$$

where w is the guidance scale and $\varnothing$ denotes the unconditional prediction. Finally, we normalize and clamp $\hat{\mathbf{h}}$ to stabilize downstream conditioning.

2.3 RFFA–IP-Control Fusion Adapter for Image Refining

Even with a strong semantic embedding condition, structural drift can still occur due to the inherent variability of bedside EEG. To stabilize layout while preserving semantics, we introduce **IP-Control Fusion Adapter**, which combines an IP-Adapter-based semantic injection [25] with a lightweight ControlNet-style residual control branch [26] on top of the SDXL-Turbo diffusion backbone [19].

The Control Adapter takes a control image $\mathbf{I}_{\text{ctrl}}$ and produces multi-scale residual features $\mathbf{R}^{(l)}$ that are injected into the frozen UNet features $\mathbf{F}^{(l)}$:

$$\tilde{\mathbf{F}}^{(l)} = \mathbf{F}^{(l)} + \sigma_l(s)\, \mathbf{R}^{(l)} \tag{5}$$

where $\sigma_l(s)$ denotes the control scale at training step s. We employ an **annealing** schedule for σ so that the control influence is introduced gradually during training, which is warmed up over 2000 steps from 0.01 to 0.1, preventing early-stage instability and encouraging the model to first learn semantic consistency.

Training follows the standard DDPM diffusion noise-prediction objective [10]. We keep pretrained SDXL-Turbo and IP-Adapter frozen and optimize only the Control Adapter parameters by minimizing the ϵ-loss. Inference follows a **coarse-to-fine** procedure: first generate a coarse image using IP-Adapter only, then blur it to derive a control image, and finally generate a refined image with both semantic (IP) and structural (Control) conditions enabled.

A key detail is how the control image $\mathbf{I}_{\text{ctrl}}$ is defined during training. At inference, $\mathbf{I}_{\text{ctrl}}$ is derived from the model's coarse prediction. However, using such self-generated controls during early training provides little structure supervision. Therefore, our Control Adapter is trained with a **mixed GT-and-generative** control image. Specifically, we construct two control cues:

1. **Stable ground-truth-based control:** Apply a Gaussian-blur low-pass operator to the ground-truth image to obtain $\mathbf{I}_{\text{ctrl}}^{\text{gt}}$.
2. **Inference-mimicking generative control:** Generate a coarse prediction using the same diffusion backbone under IP-Adapter only with sampling steps of 2, then apply the same blur operator to obtain $\mathbf{I}_{\text{ctrl}}^{\text{coarse}}$.

During training, we select the actual control image $\mathbf{I}_{\text{ctrl}}$ from these two sources:

$$p_{\text{coarse}}(s) = p_{\max} \cdot \min\left(1, \frac{s}{S_{\text{anneal}}}\right) \tag{6}$$

with probability $p_{\text{coarse}}(s)$, we use $\mathbf{I}_{\text{ctrl}}^{\text{coarse}}$, otherwise we use $\mathbf{I}_{\text{ctrl}}^{\text{gt}}$. In addition, we warm up the p_{coarse} to $p_{\max} = 0.8$ over the first $S_{\text{anneal}}(= 8000)$ steps.

3 Experiments

3.1 Datasets and Implementations

Dataset: We conduct retrieval and generation experiments on the THINGS-EEG [9], a large-scale EEG corpus acquired from 10 subjects during a visual recognition task. The experiment uses a Rapid Serial Visual Presentation paradigm with an orthogonal target-detection task to ensure sustained attention [9, 8]. For each subject, the dataset contains 16,540 training images, each repeated 4 times, and 200 test images, each repeated 80 times. THINGS-EEG serves as a reasonable healthy-subject benchmark for algorithmic validation under stimulus-locked EEG before clinical validation in DoC patients.

Implementation: We follow the experimental protocols and data splits from Li et al. [13]. All experiments are conducted on a single NVIDIA GeForce RTX 4090 with 24 GB. The proposed MAMD encoder is implemented in PyTorch and trained for 30 epochs with a batch size of 64, using the Adam optimizer with learning rate 1×10^{-4} and initial temperature 0.07. For the generation module, the Control Adapter is optimized using AdamW with learning rate 2×10^{-5} and weight decay 1×10^{-2}. During inference, we set guidance scale to 5.0 and use 16 sampling steps for generation to reduce latency for time-sensitive clinical use.

3.2 EEG Encoder Performance

Retrieval experiments are conducted within subject. We evaluate encoder performance through zero-shot retrieval-based N-way identification, reporting Top-1 accuracy for $N \in \{2, 4, 10, 50, 100, 200\}$, where each EEG embedding is matched against one ground-truth image and $N - 1$ distractors. N candidates are ranked by EEG-image embedding similarity. For the most challenging setting $N = 200$, we additionally report Top-5 accuracy. Beyond aggregate accuracy, we report the 2-way (i.e. $N = 2$) z-score to quantify separability relative to chance and 2-way-acc at 75% coverage as accuracy evaluated on the 75% most confident predictions, characterizing reliability under selective prediction [5], as bedside decisions can be preferentially made on high-confidence trials.

To ensure fair comparisons, we keep the training protocol and objective constant. Specifically, all methods are trained with the same batch size and an identical loss formulation. In **Table 1** and **Table 2**, we report results from the

Table 1. Retrieval-based identification performance for different encoders under the same experimental conditions. We report Top-1 within-subject accuracy for 2-, 4-, 10-, 50- and 100-way. All values are reported as mean ± std in percentage.

Encoders	2-way ↑	4-way ↑	10-way ↑	50-way ↑	100-way ↑
CTNet[28]	94.05 ± 1.68	80.20 ± 3.27	62.10 ± 3.59	31.35 ± 4.24	21.45 ± 5.15
MSVTNet[14]	93.95 ± 1.85	78.25 ± 3.72	58.40 ± 6.13	26.90 ± 5.79	18.35 ± 5.17
CBraMod[23]	89.45 ± 2.98	70.25 ± 5.35	50.80 ± 6.14	22.80 ± 5.14	14.85 ± 3.65
PBT[12]	92.60 ± 3.13	78.25 ± 6.12	59.70 ± 8.27	30.75 ± 7.16	22.00 ± 5.52
InceptionMI[27]	92.80 ± 2.40	77.85 ± 7.34	58.90 ± 7.90	29.65 ± 6.84	19.40 ± 6.58
EEGNex[3]	94.30 ± 1.31	81.50 ± 4.06	61.15 ± 5.90	31.65 ± 4.79	21.75 ± 4.03
ATMS[13]	96.75 ± 1.43	87.45 ± 3.63	72.20 ± 4.20	44.40 ± 6.33	34.25 ± 6.28
MAMD(Ours)	**97.25±1.19**	**87.70±3.07**	**74.25±4.96**	**46.55±6.46**	**35.05±5.04**

Table 2. We report 200-way Top-1 and Top-5 within-subject accuracy, together with reliability diagnostics. All values are reported as mean ± std in percentage (2-way z-score is raw).

Encoders	200-way ↑	200-way-top5 ↑	2-way-z ↑	2-way-acc@cov75 ↑
CTNet[28]	14.45 ± 3.93	40.15 ± 5.90	12.45 ± 0.49	98.47 ± 0.99
MSVTNet[14]	11.70 ± 4.05	34.75 ± 6.24	12.43 ± 0.52	98.00 ± 1.07
CBraMod[23]	9.60 ± 2.96	28.55 ± 5.47	11.16 ± 0.84	95.20 ± 2.49
PBT[12]	14.30 ± 4.23	38.50 ± 8.48	12.05 ± 0.88	97.33 ± 2.53
InceptionMI[27]	13.55 ± 4.83	37.35 ± 9.31	12.11 ± 0.68	97.67 ± 1.96
EEGNex[3]	14.60 ± 3.67	39.50 ± 4.96	12.53 ± 0.37	98.87 ± 0.79
ATMS[13]	24.90 ± 5.51	55.55 ± 7.13	13.22 ± 0.43	99.40 ± 0.92
MAMD(Ours)	**26.20±5.03**	**56.25±6.68**	**13.28±0.34**	**99.73±0.44**

epoch with the **best 2-way** accuracy for each encoder, as this criterion is stable and clinically interpretable, reflecting robust binary discrimination commonly used in bedside EEG paradigms.

Our MAMD encoder demonstrates the **best overall performance**, distinguished by its consistently high accuracy, competitive stability and strong reliability. Collectively, beyond serving as an encoder evaluation, the retrieval task also provides a clinically relevant paradigm for assisting the diagnosis of CMD.

3.3 Image Generation Performance

Following prior studies, generation results are reported on Subject 8 for consistency and fair comparison and are evaluated at three levels: (i) pixel-level fidelity, measuring intensity agreement with the ground-truth target; (ii) structural similarity, capturing structure preservation and gradient-based distortions; and (iii) high-level semantics, assessed via two-way identification in the feature space of pretrained visual encoders. Results are summarized in **Table 3**.

We output an **uncertainty heatmap** computed from 10 stochastic generations as **Fig. 2** to explicitly reflect both object reliability and structural stability. Regions with higher uncertainty are highlighted as low-confidence evi-

Table 3. Reconstruction quality comparison in Subject 8 with pixel-level, structural and high-level metrics. (Control scale is set to 0.01.)

Methods	Pixel-level		Structural		High-level		
	PixCorr ↑	PSNR ↑	SSIM ↑	GMSD ↓	AlexNet5 ↑	CLIP ↑	SwAV ↓
EEGImage[13]	0.1788	10.6133	0.4596	0.3008	0.5750	0.5548	0.6518
DreamDiffusion[2]	0.0919	11.8308	0.4224	**0.2830**	0.4969	0.5096	0.6993
EEGRFusion	**0.2577**	**12.2634**	**0.4864**	0.2881	**0.8753**	**0.7878**	**0.6458**

Table 4. Ablation of key components in our generation pipeline. Comparison of reconstruction quality in Subject 8. (Control scale is set to 0.01.)

Methods	MAMD	RF	CA	Pixel-level	Structural		High-level	
				PixCorr ↑	SSIM ↑	GMSD ↓	AlexNet5 ↑	SwAV ↓
EEGImage[13]	✗	✗	✗	0.1788	0.4596	0.3008	0.5750	0.6518
	✗	✓	✗	0.2206	0.4742	0.2918	0.8882	0.6369
	✗	✓	✓	0.2356	0.4762	0.2898	0.8703	0.6463
EEGRFusion	✓	✓	✗	0.2412	0.4846	0.2909	**0.8930**	**0.6336**
	✓	✓	✓	**0.2577**	**0.4864**	**0.2881**	0.8753	0.6458

dence, encouraging additional clinical verification and reducing over-reliance on the automated output. Notably, with structural drift controlled, the heatmap is most informative in the common mixed regime, where some samples remain stimulus-consistent while others drift off-target.

Ablation study To isolate the contribution of each major component in our pipeline, as **Table 4** shows, we conduct controlled ablations and report a compact subset of the assessment metrics. Specifically, we vary (i) the generation backbone (standard diffusion/Rectified Flow), (ii) the presence of Control Adapter, and (iii) the EEG encoder. Control Adapter strengthens structural adherence while slightly suppressing semantic expressiveness, yet it still improves the perceived visual quality and stability of the generated images.

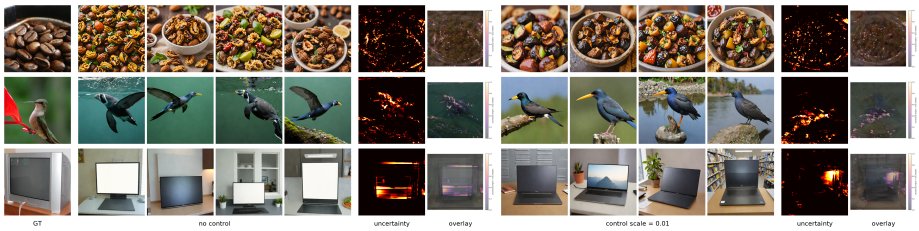

Fig. 2. Example reconstructions for Subject 8. We visualize the ground-truth image, generated reconstructions, uncertainty maps computed from 10 stochastic generations, and overlay maps under (i) no control and (ii) control scale = 0.01, showing that control improves reconstruction quality and reduces structural inconsistency.

4 Discussion and Conclusion

This paper proposes EEGRFusion, which integrates a montage-aware MAMD encoder, a Rectified-Flow embedding prior, and an IP-Control Fusion Adapter to deliver more stable reconstructions and uncertainty-aware evidence. EEGRFusion addresses key challenges in EEG-conditioned generation, including the CLIP semantic bottleneck that may suppress fine-grained perceptual cues as well as EEG's trial-to-trial variability and low SNR, thereby improving fidelity, stability, and clinical reviewability and moving EEG-to-image modeling closer to practical assistive assessment for DoC. Constrained by the absence of paired DoC/CMD patient EEG-image data, this work should therefore be viewed as algorithmic validation on a healthy-subject benchmark rather than clinical deployment. Future research will perform comprehensive evaluation on clinical datasets, while further mitigating the semantic trade-off introduced by structural control.

Acknowledgments. This study was supported by Major Science and Technology Special Projects for Cancer, Cardiovascular, Respiratory and Metabolic Diseases (No. 2025ZD0547202); the Guangzhou Science and Technology Department (No. 2024D03J0013).

Disclosure of Interests. The authors declare no competing interests.

References

1. Aubinet, C., Claassen, J., Edlow, B.L., Fischer, D., Gosseries, O., Koch, C., Kondziella, D., Massimini, M., Young, M.J.: Covert consciousness: What's in a name? Brain **148**(12), 4248–4252 (2025)
2. Bai, Y., Wang, X., Cao, Y.P., Ge, Y., Yuan, C., Shan, Y.: DreamDiffusion: High-quality EEG-to-image generation with temporal masked signal modeling and CLIP alignment. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol. 15089, pp. 472–488. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72751-1_27
3. Chen, X., Teng, X., Chen, H., Pan, Y., Geyer, P.: Toward reliable signals decoding for electroencephalogram: A benchmark study to EEGNeX. Biomedical Signal Processing and Control **87**, 105475 (2024)
4. Dao, T., Gu, A.: Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In: Proceedings of the 41st International Conference on Machine Learning. pp. 10041–10071. PMLR (2024)
5. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 4879–4888. Curran Associates Inc., Red Hook, NY, USA (2017)
6. Giacino, J.T., Kalmar, K., Whyte, J.: The JFK coma recovery scale–revised: Measurement characteristics and diagnostic utility. Archives of Physical Medicine and Rehabilitation **85**(12), 2020–2029 (2004)
7. Giacino, J.T., Katz, D.I., Schiff, N.D., et al.: Practice guideline update recommendations summary: Disorders of consciousness: Report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology; the American Congress of Rehabilitation Medicine; and the National

Institute on Disability, Independent Living, and Rehabilitation Research. Neurology **91**(10), 450–460 (2018)
8. Gifford, A.T., Dwivedi, K., Roig, G., Cichy, R.M.: A large and rich EEG dataset for modeling human visual object recognition. NeuroImage **264**, 119754 (2022)
9. Hebart, M.N., Contier, O., Teichmann, L., et al.: THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. eLife **12**, e82580 (2023)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. pp. 6840–6851. Curran Associates Inc., Red Hook, NY, USA (2020)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
12. Klein, T., Minakowski, P., Sager, S.: Flexible patched brain transformer model for EEG decoding. Scientific Reports **15**, 10935 (2025)
13. Li, D., Wei, C., Li, S., Zou, J., Liu, Q.: Visual decoding and reconstruction via eeg embeddings with guided diffusion. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems. pp. 102822–102864. Curran Associates Inc., Red Hook, NY, USA (2024)
14. Liu, K., Yang, T., Yu, Z., Yi, W., Yu, H., Wang, G., Wu, W.: MSVTNet: Multiscale vision transformer neural network for EEG-based motor imagery decoding. IEEE Journal of Biomedical and Health Informatics **28**(12), 7126–7137 (2024)
15. 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)
16. Murtaugh, B., Shapiro Rosenbaum, A.: Clinical application of recommendations for neurobehavioral assessment in disorders of consciousness: An interdisciplinary approach. Frontiers in Human Neuroscience **17**, 1129466 (2023)
17. Ozcelik, F., VanRullen, R.: Natural scene reconstruction from fMRI signals using generative latent diffusion. Scientific Reports **13**, 15666 (2023)
18. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1), 3942–3951 (2018)
19. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
21. Scotti, P.S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., Dempster, A.J., Verlinde, N., Yundler, E., Weisberg, D., Norman, K.A., Abraham, T.M.: Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. arXiv preprint arXiv:2305.18274 (2023)
22. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14453–14463. IEEE, Vancouver, BC, Canada (2023)
23. Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., Pan, G.: CBraMod: A criss-cross brain foundation model for EEG decoding. In: Proceedings of the 13th International Conference on Learning Representations (ICLR) (2025)

24. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds. ACM Transactions on Graphics **38**(5), 146:1–146:12 (2019)
25. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
26. Zavadski, D., Feiden, J.F., Rother, C.: ControlNet-XS: Rethinking the control of text-to-image diffusion models as feedback-control systems. In: Leonardis, A., Ricci, E., Roth, S., et al. (eds.) Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol. 15146, pp. 343–362. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73223-2_20
27. Zhang, C., Kim, Y.K., Eskandarian, A.: EEG-inception: An accurate and robust end-to-end neural network for EEG-based motor imagery classification. Journal of Neural Engineering **18**(4), 046014 (2021)
28. Zhao, W., Jiang, X., Zhang, B., Xiao, S., Weng, S.: CTNet: A convolutional transformer network for EEG-based motor imagery classification. Scientific Reports **14**, 20237 (2024)
29. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision Mamba: Efficient visual representation learning with bidirectional state space model. In: Proceedings of the 41st International Conference on Machine Learning. pp. 62429–62442. PMLR (2024)