UNI-NEUR2IMG: UNIFIED NEURAL SIGNAL-GUIDED IMAGE GENERATION, EDITING, AND STYLIZATION VIA DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Non-invasive Brain–Computer Interfaces (BCIs) provide a viable pathway for decoding visual perception, yet EEG-to-image translation remains constrained by low signal-to-noise ratio, temporal non-stationarity, and substantial trial- and subject-level variability. Treating EEG as a clean text-like prompt can expose the diffusion backbone to both signal and noise, motivating explicit control over neural conditioning. We propose **Uni-Neur2Img**, a unified Diffusion Transformer architecture and conditioning interface for generation, editing, and stylization, with separately fine-tuned task-specific weights. A dual-pathway **Neural Adapter** maps EEG to pooled and token conditioning spaces without discrete EEG-to-text decoding. The **Neural Signal Injection Module (NSIM)** combines adaptive LoRA, using EEG- and timestep-dependent gates and a rank schedule, with quality-aware cross-modal attention. A separate subject-adversarial objective regularizes EEG representations. A frozen T5/CLIP text teacher provides semantic supervision only in EEG+Text training and is omitted in EEG-only training and inference. We introduce **EEG-Style**, comprising 20,000 paired trials from five participants viewing 4,000 images across 1,000 styles, with a style-disjoint split. Experiments evaluate generation on EEGCVPR40, editing on LoongX, and stylization on EEG-Style. Results are reported as means over three random seeds. In pooled stylization experiments involving five participants, EEG-only conditioning outperforms text-only conditioning on five of seven metrics with the tested prompts, reducing Style Loss by 68.75% and Color Loss by 14.86%. Joint EEG+Text conditioning surpasses both individual modalities on six of seven metrics, with DINO as the exception. These findings support visually evoked EEG as style guidance and show complementary gains from EEG and text under the evaluated conditions.

1 INTRODUCTION

Reconstructing visual content from brain activity is a central goal of neural decoding (Takagi & Nishimoto, 2023; Du et al., 2023). Electroencephalography (EEG) offers non-invasive acquisition, millisecond temporal resolution, and accessibility relative to fMRI, motivating EEG-conditioned image generation (Bai et al., 2023; Li et al., 2024; Lan et al., 2023). However, low signal-to-noise ratio, temporal non-stationarity, and variability across trials and participants hinder reliable visual conditioning.

Early EEG-to-image methods explored generative adversarial models (Palazzo et al., 2017; Kavasidis et al., 2017), with subsequent work adopting diffusion-based reconstruction (Bai et al., 2023; Li et al., 2024). Diffusion Transformers (DiTs) (Peebles & Xie, 2023) offer expressive conditioning interfaces, but treating EEG as a clean prompt can propagate noise into the generative model. Reliable adaptation therefore requires explicit control over neural conditioning. For stylization, textual prompts may underspecify color, texture, and other appearance attributes. Evidence that visual attributes can be decoded from EEG (Hajonides et al., 2021; Bae & Luck, 2018) motivates studying whether EEG can guide stylization without text and whether joint conditioning improves appearance fidelity.

We propose **Uni-Neur2Img** (Fig. 1), a unified DiT architecture and conditioning interface for EEG-driven generation, editing, and stylization, with separately fine-tuned task-specific weights. A dual-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

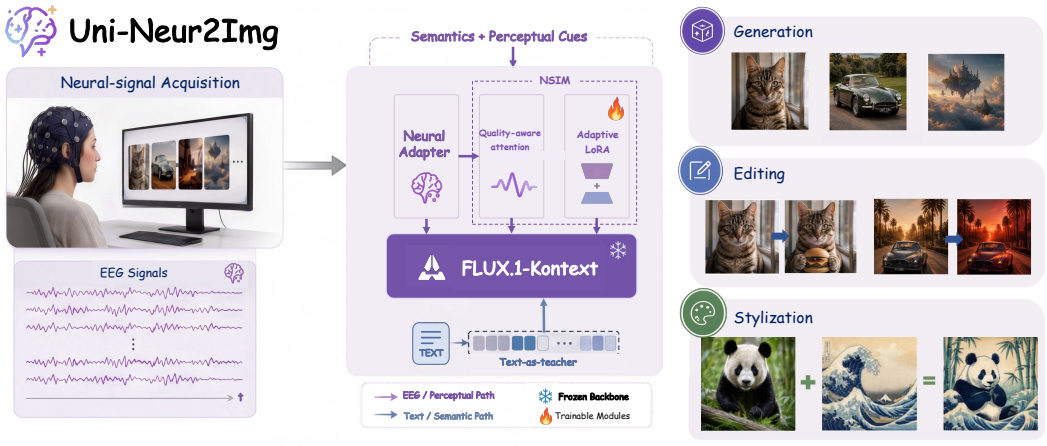

Figure 1: Uni-Neur2Img overview. Tasks are fine-tuned separately; the text teacher is used only with EEG+Text.

pathway **Neural Adapter** maps EEG to the pooled and token interfaces of FLUX.1-Kontext (Black Forest Labs et al., 2025) without discrete EEG-to-text decoding. The **Neural Signal Injection Module (NSIM)** combines an adaptive **LoRA pathway** with **quality-aware cross-modal attention**. Its condition-selective LoRA updates use EEG- and timestep-dependent gates and a timestep-dependent rank schedule; its attention bias depends on signal quality and timestep. Together, the Neural Adapter and NSIM align and inject EEG into the frozen backbone. A separate subject-adversarial training objective encourages subject-invariant representations. In the EEG+Text setting, a frozen T5/CLIP teacher provides semantic supervision; EEG-only training and inference omit this teacher while retaining the backbone's language-aligned conditioning space.

We investigate **image stylization directly conditioned on visually evoked EEG**, using responses to stylized images to guide a source image's appearance while preserving its content. To our knowledge, this task has not been studied previously. We collect **EEG-Style**, a new dataset of **20,000 paired EEG–image trials** from five participants viewing the same 4,000 images across 1,000 styles. Each style contains four content-diverse images. A shared split holds out 250 complete style groups across all participants, yielding 15,000 training and 5,000 test trials. This evaluates generalization to unseen styles without stimulus overlap across partitions.

Experiments on EEGCVPR40, LoongX, and EEG-Style evaluate all three tasks. With pooled training on the five EEG-Style participants, EEG-only conditioning achieves better mean performance than text-only conditioning on five of seven metrics under the tested prompts. Style Loss decreases by 68.75% and Color Loss by 14.86%, without textual prompts or a text teacher in the EEG-only setting. Joint EEG+Text conditioning surpasses both individual modalities on six of seven metrics, with DINO as the exception. These results support EEG-only style guidance and complementary benefits from joint conditioning under the evaluated conditions.

Our main contributions are:

1) **Unified architecture:** A shared EEG conditioning interface for generation, editing, and stylization, with task-specific fine-tuning.
2) **Integrated neural conditioning:** A Neural Adapter for EEG alignment and NSIM for signal injection, combining adaptive LoRA with quality-aware cross-modal attention.
3) **New task and dataset:** EEG-guided image stylization and EEG-Style, comprising 20,000 paired trials across 1,000 styles with a style-disjoint split.
4) **EEG style guidance and modality complementarity:** In pooled EEG-Style evaluation with the tested prompts, EEG-only conditioning surpasses text-only conditioning in style and color fidelity, while joint conditioning improves six of seven metrics over both individual modalities.

2 Related Work

2.1 Visual decoding using neural signals

EEG visual decoding spans image reconstruction, retrieval, and video reconstruction (Lan et al., 2023; Liu et al., 2024; Xiang et al., 2026). Reconstruction methods commonly align EEG and visual representations with diffusion priors (Zhang et al., 2025; Li et al., 2024; Bai et al., 2023). Recent work extends this direction through frequency-oriented semantic alignment in D^2-FOSA (Yu et al., 2026), while AVDE explores autoregressive image generation (Dai et al., 2026). EEGMirror uses montage-agnostic self-supervised pretraining for EEG-to-video decoding (Liu et al., 2025). Visual decoding also uses fMRI (Miyawaki et al., 2008) and MEG (Kauppi et al., 2015); EEG offers millisecond temporal resolution with portable, relatively low-cost acquisition (Tangermann et al., 2012; Li et al., 2024). Decodable attributes such as hue (Hajonides et al., 2021), orientation (Bae & Luck, 2018), and spatial position (Foster et al., 2017) motivate investigating EEG as a source of style guidance.

NEED addresses cross-subject and cross-task reconstruction through subject adaptation (Huang et al., 2025b), while SATTC uses label-free calibration for cross-subject retrieval (Huang & Zhu, 2026). Block-design temporal confounds can survive image-disjoint splits (Li et al., 2021); EEG-Style therefore holds out style groups and acquisition blocks across participants. Our encoder and objectives build on state-space temporal modeling (Gu et al., 2022) and domain-adversarial learning (Ganin & Lempitsky, 2015).

2.2 Controllable image generation

DragGAN supports point-based GAN editing (Pan et al., 2023). Diffusion models inject conditions into the denoiser. On U-Net backbones (Ronneberger et al., 2015), lightweight external branches such as ControlNet (Zhang et al., 2023), T2I-Adapter (Mou et al., 2024), and IP-Adapter (Ye et al., 2023) accept spatial conditions or image prompts while keeping the base model frozen. Diffusion Transformers (Peebles & Xie, 2023; Chen et al., 2024) provide token-level conditioning interfaces and support LoRA-based in-context adaptation (Huang et al., 2024; Hu et al., 2022; Huang et al., 2025a). The FLUX (Black Forest Labs, 2024) family integrates rectified flow (Liu et al., 2023; Lipman et al., 2023) with this backbone, and OminiControl (Tan et al., 2025) reuses the DiT itself to encode conditional images, yielding a minimalist unified pathway that we adopt as the host for EEG-conditioned generation, editing, and stylization.

Brain-Supervised Image Editing uses EEG responses to learn GAN editing directions (Davis et al., 2022), while LoongX fuses multimodal neural signals for DiT-based image editing (Zhou et al., 2025). NeuroAdapter conditions latent diffusion on fMRI representations and analyzes attention during denoising (Feng et al., 2026). Our focus is an EEG conditioning interface that combines dual-pathway alignment with adaptive injection across generation, editing, and stylization, using task-specific weights.

2.3 Uncertainty-aware fusion and adaptive low-rank adaptation

QAN learns quality-weighted sample aggregation for image-set recognition (Liu et al., 2017), while Trusted Multi-View Classification fuses views at the evidence level (Han et al., 2021). We adapt quality-aware fusion to generative attention through a timestep- and quality-conditioned bias. MoLE (Wu et al., 2024) and X-LoRA (Buehler & Buehler, 2024) extend gated mixture-of-experts (Shazeer et al., 2017) to adapters, while T-LoRA schedules effective rank across diffusion timesteps (Soboleva et al., 2025). NSIM combines EEG- and timestep-dependent gates for condition-selective LoRA, a timestep-dependent rank schedule, and a quality-aware cross-modal attention bias. Its contribution is their integration with direct EEG alignment for three synthesis tasks, without guaranteeing noise removal or subject invariance.

3 Dataset

We evaluate EEG-conditioned image generation on EEGCVPR40 (Spampinato et al., 2017; Palazzo et al., 2021), image editing on L-Mind, the dataset introduced with LoongX (Zhou et al., 2025),

and image stylization on our newly collected *EEG-Style* dataset. These datasets differ in electrode montage, participant population, and acquisition paradigm; they evaluate the same architecture with task-specific fine-tuning. EEG-Style provides a testbed for the contribution of visually evoked EEG to color, texture, and stylistic reconstruction. The recorded responses are elicited by viewed images, so these experiments do not establish decoding of imagined images, dreams, emotions, or unconstrained aesthetic intent.

EEG-Style acquisition. EEG-Style contains 20,000 trials from five participants, each viewing the same 4,000 unique Midjourney-generated stylized target images organized into 1,000 style groups with four content-diverse images per group. Source–target pairs come from the same stylization collection sourced from the Midjourney website.[1] Each source is the original image corresponding to its stylized target. Participants view the final target image during EEG recording; the model receives the paired source image and the EEG evoked by that target, with text additionally supplied in the joint setting. The target serves as the supervised reference. This is paired stylization guided by target-evoked EEG, not transfer from an unrelated style exemplar.

The 64-channel device recorded 59 scalp EEG channels, one ECG channel, and four EOG channels at 1,000 Hz. Each stimulus was displayed for 1 s; the 0.5 s analysis segment contains 500 samples per channel (59×500). The four images sharing a style formed a block, with a 2 s rest between blocks and extended breaks after every 50 blocks. Preprocessing comprised notch filtering, ICA-based EOG/ECG artifact correction, and 1–80 Hz band-pass filtering. The protocol received institutional ethics approval and participants provided written informed consent.

Style-disjoint partition. We randomly assign 750 complete style groups to training and 250 to testing. All four target images in a group remain together, yielding 3,000 training and 1,000 test targets per participant, or 15,000 training and 5,000 test trials overall. The assignments are shared across participants, so no held-out target or style group enters training through another participant. This partition separates the corresponding acquisition blocks and evaluates held-out styles. The main multi-subject setting pools all five participants; LOSO instead tests on an unseen participant. Style-disjoint splitting alone does not rule out all session or stimulus confounds. Public benchmarks follow their respective evaluation protocols.

4 METHODOLOGY

4.1 OVERVIEW AND PROBLEM FORMULATION

For an EEG trial $\mathbf{x} \in \mathbb{R}^{C\times T}$, a source image I_{src} when applicable, and an optional text prompt y, we encode the available inputs into conditioning signals $\mathbf{c}$ and model

$$p_\theta\big(I_{\mathrm{tgt}} \,\big|\, \mathbf{x}, I_{\mathrm{src}}, y\big) \;=\; p_\theta\big(I_{\mathrm{tgt}} \,\big|\, \mathbf{c}\big). \tag{1}$$

The conditioning encoders are deterministic. Generation uses EEG-only conditioning; editing and stylization additionally receive a source image. For EEG-Style, I_{src} is the original image paired with I_{tgt}, and $\mathbf{x}$ is recorded while the participant views I_{tgt} (Section 3). Auxiliary conditioning depends on the input setting.

As shown in Figure 2, the dual-pathway Neural Adapter maps EEG directly to the conditioning interfaces of frozen FLUX backbones (Black Forest Labs, 2024; Black Forest Labs et al., 2025), avoiding a discrete EEG-to-text decoding step. The Neural Signal Injection Module (NSIM) integrates an adaptive, condition-selective LoRA pathway with quality-aware cross-modal attention to regulate neural conditioning across denoising steps. The three tasks share the architecture and objective, with separately trained task-specific weights.

4.2 EEG ENCODER AND DUAL-PATHWAY NEURAL ADAPTER

The encoder first reweights EEG channels using a learnable channel prior. Differentiable frequency-band filtering and multi-scale temporal convolutions produce a band summary $\mathbf{f}_{\mathrm{band}}$, while a multi-branch encoder with structured state-space temporal modeling (Gu et al., 2022) extracts complementary temporal features. A projection head maps these features to $\mathbf{Z} \in \mathbb{R}^{16\times 64\times 64}$.

[1] https://www.midjourney.com/

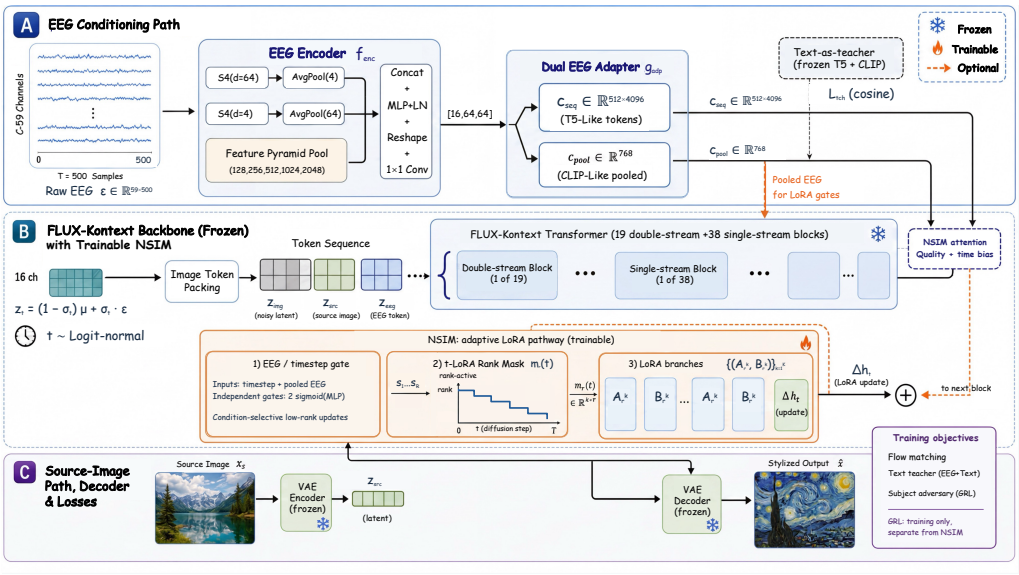

Figure 2: **Uni-Neur2Img** uses a shared EEG conditioning interface with task-specific weights on frozen FLUX backbones. The dual-pathway Neural Adapter aligns EEG; NSIM integrates adaptive LoRA and quality-aware cross-modal attention.

The Neural Adapter maps $\mathbf{Z}$ to the token-level and pooled prompt interfaces of FLUX. A 1×1 convolution followed by reshaping produces sequence embeddings $\mathbf{p}_{\text{seq}}$, and global average pooling followed by a linear projection produces a pooled embedding $\mathbf{p}_{\text{pool}}$. These EEG-derived embeddings occupy the interfaces normally supplied by T5 hidden states and the CLIP pooled output, respectively. The band summary enters both pathways through zero-initialized projections. Appendix A.1 gives the input dimensions and channel-weighting definition.

4.3 Neural Signal Injection Module (NSIM)

We denote the packed target and conditioning tokens by $\mathbf{H} = [\mathbf{s}_{\text{tgt}} \,\|\, \mathbf{s}_{\text{cond}} \,\|\, \mathbf{s}_{\text{eeg}}]$. Here $\mathbf{s}_{\text{tgt}}$ represents the noisy target-image latent, $\mathbf{s}_{\text{cond}}$ the task-specific auxiliary conditioning, and $\mathbf{s}_{\text{eeg}} = \text{Pack}(\mathbf{Z})$ the EEG representation. NSIM contains an adaptive LoRA pathway for condition-selective low-rank updates (Eq. 2) and quality-aware cross-modal attention (Eq. 3).

Adaptive LoRA pathway. We attach low-rank branches to the query, key, value, and output projections in each double- and single-stream block. For the query projection,

$$\tilde{\mathbf{q}} \;=\; \mathbf{W}_q\mathbf{H} \;+\; \sum_{i=1}^{n_{\text{LoRA}}} g_i(t, \mathbf{p}_{\text{pool}})\, \mathbf{B}_q^{(i)} \big[\mathbf{m}_R^{(i)}(t) \odot \mathbf{A}_q^{(i)}(\mathbf{m}^{(i)} \odot \mathbf{H})\big], \tag{2}$$

where $\mathbf{m}^{(i)}$ selects the conditioning positions assigned to branch i, and $(\mathbf{A}_q^{(i)}, \mathbf{B}_q^{(i)})$ are its low-rank factors. Given timestep t and the pooled EEG embedding, the router produces an independent sigmoid gate $g_i = 2\sigma(z_i) \in (0, 2)$ for each branch. The deterministic mask $\mathbf{m}_R^{(i)}(t)$ increases the active rank from $r_{\min}$ at the noise endpoint to R_i at the data endpoint (Soboleva et al., 2025). Together, the gates and rank masks vary adaptation strength and capacity across denoising steps. Appendix A.3 specifies the router, rank schedule, and initialization.

Quality-aware cross-modal attention. We compute a non-trainable scalar $c \in [0, 1]$ from spectral-power and peak-to-peak amplitude proxies (Appendix A.2). With local batch size one, the amplitude term is constant and only the spectral term varies across trials. This is a heuristic control signal, not a calibrated reliability probability. For modality blocks a and b, we replace a fixed attention mask with the additive bias

$$\mathbf{B}_{ab}(t, c) \;=\; \text{LSig}\big([\text{MLP}_{\text{mask}}([\gamma(t) \,\|\, \mathbf{W}_c c])]_{ab}\big) \;+\; \text{logmax}(c, \epsilon) \cdot \mathbf{1}\big[(a, b) \in \mathcal{E}_\times\big], \tag{3}$$

where $\mathrm{LSig} = \log \sigma$, $\gamma(t)$ embeds the timestep, $\epsilon = 10^{-3}$, and $\mathcal{E}_\times$ is the set of cross-modal edges. The bias is broadcast to token pairs and added to scaled dot-product attention logits. The learned term adapts interactions to timestep and trial quality, while the logarithmic term penalizes cross-modal edges more strongly as c decreases. This extends quality-aware fusion (Liu et al., 2017; Han et al., 2021) to generative attention. Initialization and limiting behavior are detailed in Appendix A.4.

4.4 Training objective

Text-teacher supervision. In the EEG+Text setting, the frozen FLUX text encoders provide semantic targets for both adapter pathways: $\bar{\mathbf{p}}_{\mathrm{seq}} = E_{\mathrm{T5}}(y)$ and $\bar{\mathbf{p}}_{\mathrm{pool}} = E_{\mathrm{CLIP}}(y)$. We align the EEG embeddings with these targets using

$$\mathcal{L}_{\mathrm{text}} \;=\; w_t\Big[\tfrac{1}{L_e}\sum_{l=1}^{L_e}\mathcal{L}_{\cos}\big([\mathbf{p}_{\mathrm{seq}}]_l, [\bar{\mathbf{p}}_{\mathrm{seq}}]_l\big) + \alpha_{\mathrm{p}}\,\mathcal{L}_{\cos}\big(\mathbf{p}_{\mathrm{pool}}, \bar{\mathbf{p}}_{\mathrm{pool}}\big)\Big], \tag{4}$$

where L_e is the sequence length, $\mathcal{L}_{\cos}$ denotes cosine distance, α_{p} weights pooled alignment, and w_t weights the teacher loss. The sequence and pooled terms supervise token-level and global semantic alignment, respectively. Teacher supervision is applied only during training, with both text encoders frozen. EEG-only training and inference omit text input and the text teacher, with $\mathcal{L}_{\mathrm{text}} = 0$. Thus, using the pretrained prompt interfaces does not itself require text input; language supervision is added only in the EEG+Text setting.

Flow-matching and total loss. We combine flow matching with the auxiliary losses applicable to each training setting:

$$\mathcal{L} \;=\; \mathbb{E}\big[w(t)\,\big\|\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{v}\big\|_2^2\big] \;+\; \lambda_{\mathrm{grl}}\mathcal{L}_{\mathrm{grl}} \;+\; \mathcal{L}_{\mathrm{text}}. \tag{5}$$

Here $\hat{\mathbf{v}}_\theta$ predicts the velocity target $\mathbf{v}$ from the noisy image latent $\mathbf{x}_t$, $w(t)$ weights the flow-matching loss, and $\mathbf{c} = \{\mathbf{p}_{\mathrm{pool}}, \mathbf{p}_{\mathrm{seq}}, \mathbf{s}_{\mathrm{cond}}, \mathbf{s}_{\mathrm{eeg}}, t, c\}$ collects the conditioning signals. The subject-adversarial loss $\mathcal{L}_{\mathrm{grl}}$ uses gradient reversal (Ganin & Lempitsky, 2015) on the pooled EEG embedding to discourage subject-discriminative features when multiple labeled subjects are available; the head is an auxiliary training component, separate from NSIM, and is omitted at inference. Appendix A.5 gives the subject-adversarial objective and its applicability to different training settings.

The DiT, VAE, and text encoders remain frozen. We optimize the EEG encoder, Neural Adapter, NSIM (LoRA branches, gates, and soft-mask network), and the auxiliary subject classifier when applicable.

5 Experiment

5.1 Implementation Details

We use pretrained FLUX.1-dev and FLUX.1-Kontext diffusion transformers, with separate task-specific fine-tuning for each subject setting. Training is conducted on $8\times$ A100 (80 GB) GPUs with an input resolution of 512×512 and a per-GPU batch size of 1 (effective batch size 8 without gradient accumulation). We use the AdamW optimizer with learning rate 1×10^{-4}. We use bf16 mixed precision and maximum NSIM LoRA rank 128. Gradients are clipped at a norm of 1.0. The learning rate follows a constant schedule with a 500-step warm-up.

Random seeds and reporting. We repeat each experiment conducted in this work, including reproduced baselines and ablations, with three random seeds and report the mean across runs. Reported uncertainty intervals are 95% confidence intervals across these runs, where available.

5.2 EEG-Guided Image Generation

Training data flow. We evaluate EEG-guided generation with EEG-only conditioning, without text input during training or inference.

Quantitative Comparison. Table 1 compares generation performance on EEGCVPR40 under single- and multi-subject settings. Uni-Neur2Img achieves the highest IS and ACC and the lowest FID and

Table 1: Image generation on EEGCVPR40. Percentages denote changes relative to GWIT; dashes indicate unreported results.

Single-subject				
Method	**IS ↑**	**FID ↓**	**ACC ↑**	**LPIPS ↓**
BrainVis	31.52	121.02	0.49	—
GWIT	33.32	80.47	0.91	0.81
Uni-Neur2Img	**35.19** ± 0.639 (+5.61%)	**74.70** ± 0.958 (-7.17%)	**0.93** ± 0.027 (+2.20%)	**0.75** ± 0.020 (-7.41%)

Multi-subject				
Method	**IS ↑**	**FID ↓**	**ACC ↑**	**LPIPS ↓**
EEGStyleGAN-ADA	10.82	174.13	—	—
GWIT	33.87	78.11	0.91	0.77
Uni-Neur2Img	**36.19** ± 0.699 (+6.85%)	**70.72** ± 0.912 (-9.46%)	**0.92** ± 0.025 (+1.10%)	**0.73** ± 0.018 (-5.2%)

Note. Baselines: BrainVis (Fu et al., 2025), GWIT (Lopez et al., 2025), and EEGStyleGAN-ADA (Singh et al., 2024).

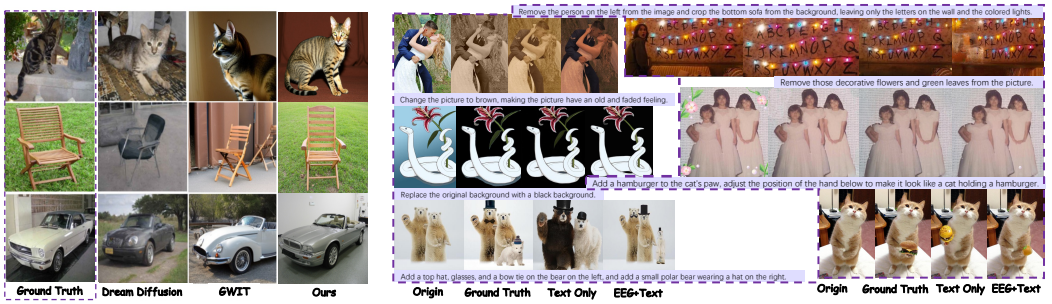

Figure 3: (a) EEG-Driven Image Generation on the CVPR40 Dataset: Qualitative Comparison of Uni-Neur2Img with DreamDiffusion, GWIT, and Original Images; (b) Qualitative Comparison of Uni-Neur2Img on the LoongX Dataset: Original, Ground Truth, Text-Only, and EEG+Text Results.

LPIPS among the reported baselines in both settings. Relative to GWIT, it improves IS by 5.61% and reduces FID by 7.17% in the single-subject setting; the corresponding gains are 6.85% and 9.46% in the multi-subject setting. Higher ACC and lower LPIPS indicate better category consistency and perceptual agreement with reference images under these metrics; IS and FID characterize distribution-level sample quality.

Qualitative Comparison. Figure 3(a) compares generated images with the reference stimuli. Uni-Neur2Img captures category-specific details, including the cat's striped fur, and the chair's slatted structure and green background. Compared with DreamDiffusion and GWIT, the displayed examples show closer correspondence to the reference appearance while retaining clear contours and coherent object structure, consistent with the quantitative results.

5.3 Neural-Guided Image Editing

Quantitative Comparison. Table 2 compares Uni-Neur2Img with OminiControl and LoongX across text-only, EEG-only, neural-signal-only, and Neural+Speech settings. Uni-Neur2Img consistently achieves lower L1 and L2 and higher CLIP-I and DINO than the corresponding baseline in all four settings. Joint neural-signal and speech conditioning yields the best values on these four metrics. Compared with LoongX under the same Neural+Speech setting, our method reduces L1 and L2 by 43.48% and 48.98%, respectively, while improving CLIP-I by 16.82% and DINO by 43.16%. These gains indicate improved reconstruction accuracy and image-level semantic consistency. The modality comparison in Table 7 further shows that EEG+Text improves four of five metrics over Text Only, reducing L1 and L2 by 19.76% and 20.14%, while Text Only retains higher CLIP-T.

Qualitative Analysis. Figure 3(b) illustrates color, structural, and global edits. Our model produces localized color changes, preserves subject identity and pose during structural edits, and adjusts global color tone while retaining fine detail. In the bear example, EEG+Text better preserves fur color than Text Only when adding accessories. Consistent with the quantitative gains over LoongX in Table 2, these examples support neural-guided image editing and motivate accessible, intuitive creative interfaces.

Table 2: Image editing on LoongX with reported 95% CIs. Percentages denote changes relative to the baseline within each input setting; bold indicates the best result within each setting.

Method	L1 ↓	L2 ↓	CLIP-I ↑	DINO ↑	CLIP-T ↑
OminiControl (Text Only)	0.2632 ± 0.006	0.1161 ± 0.010	0.6558 ± 0.010	0.4636 ± 0.017	0.2549 ± 0.008
Uni-Neur2Img (Text Only)	**0.1827** ± 0.005 (-30.59%)	**0.0690** ± 0.009 (-40.57%)	**0.7259** ± 0.010 (+10.69%)	**0.5673** ± 0.016 (+22.37%)	**0.2563** ± 0.006 (+0.55%)
LoongX (EEG Only)	0.2641 ± 0.006	0.1078 ± 0.009	0.5457 ± 0.009	0.2963 ± 0.015	0.2251 ± 0.008
Uni-Neur2Img (EEG Only)	**0.2122** ± 0.006 (-19.65%)	**0.0854** ± 0.009 (-20.78%)	**0.6475** ± 0.009 (+18.65%)	**0.4442** ± 0.015 (+49.92%)	**0.2280** ± 0.007 (+1.29%)
LoongX (Neural Signals)	0.2509 ± 0.006	0.1029 ± 0.009	0.6605 ± 0.009	0.4812 ± 0.015	**0.2436** ± 0.008
Uni-Neur2Img (Neural Signals)	**0.1979** ± 0.005 (-21.12%)	**0.0813** ± 0.008 (-20.99%)	**0.6878** ± 0.009 (+4.13%)	**0.5260** ± 0.014 (+9.31%)	0.2321 ± 0.007 (-4.72%)
LoongX (Neural+Speech)	0.2594 ± 0.006	0.1080 ± 0.009	0.6374 ± 0.009	0.4205 ± 0.014	**0.2588** ± 0.009
Uni-Neur2Img (Neural+Speech)	**0.1466** ± 0.005	**0.0551** ± 0.008	**0.7446** ± 0.008	**0.6020** ± 0.0014	0.2522 ± 0.006

Table 3: Single- and multi-subject stylization on EEG-Style. Percentages denote changes relative to EEG Only (single-subject) and Text Only (multi-subject). Bold indicates the best result within each block.

Method	Input	CLIP-I ↑	DINO ↑	PSNR ↑	SSIM ↑	PDist ↓	Style loss ↓	Color loss ↓
IP-Adapter	**Text Only**	0.7016 (CI: —)	0.3268 (CI: —)	7.4663 (CI: —)	0.2112 (CI: —)	0.8751 (CI: —)	0.0854 (CI: —)	0.3182 (CI: —)
Uni-Neur2Img (Single)	**EEG Only**	0.7487 ± 0.015	0.4089 ± 0.015	9.4166 ± 0.247	0.2573 ± 0.008	0.7332 ± 0.018	0.2681 ± 0.009	0.3092 ± 0.010
	Text Only	**0.8260** ± 0.017 (+10.32%)	**0.6179** ± 0.015 (+51.11%)	9.2378 ± 0.237 (-1.90%)	0.2733 ± 0.008 (+6.22%)	0.7089 ± 0.018 (-3.31%)	0.2935 ± 0.009	0.3498 ± 0.010
	Text+EEG	0.8182 ± 0.018 (+9.28%)	0.5741 ± 0.015 (+40.40%)	**11.0139** ± 0.264 (+16.96%)	**0.3168** ± 0.008 (+23.12%)	**0.6250** ± 0.019 (-14.76%)	**0.2514** ± 0.009 (-6.23%)	**0.2946** ± 0.010 (-4.72%)
Uni-Neur2Img (Multi)	**Text Only**	0.8260 ± 0.017	**0.6179** ± 0.015	9.2378 ± 0.237	0.2733 ± 0.008	0.7089 ± 0.018	0.2045 ± 0.009	0.3122 ± 0.010
	EEG Only	0.7975 ± 0.018	0.5284 ± 0.015	9.9694 ± 0.264 (+7.92%)	0.2886 ± 0.008 (+5.60%)	0.6660 ± 0.019 (-6.05%)	0.0639 ± 0.009 (-68.75%)	0.2658 ± 0.010 (-14.86%)
	Text+EEG	**0.8347** ± 0.018 (+1.05%)	0.6102 ± 0.015	**11.0748** ± 0.264 (+19.89%)	**0.3316** ± 0.008 (+21.33%)	**0.6069** ± 0.019 (-14.39%)	**0.0450** ± 0.009 (-78.00%)	**0.2365** ± 0.010 (-24.25%)

Note. Baseline: IP-Adapter (Ye et al., 2023). Its CIs are unavailable.

5.4 EEG-Guided Image Stylization

Quantitative Comparison. Table 3 compares EEG and text conditioning using the tested prompts; the multi-subject setting pools the five recorded participants. In the single-subject setting, Text Only outperforms EEG Only on four metrics—CLIP-I, DINO, SSIM, and PDist—with gains of 10.32% in CLIP-I and 51.11% in DINO. EEG Only nevertheless achieves higher PSNR and lower Style Loss and Color Loss. In the multi-subject setting, EEG Only outperforms Text Only on five of seven metrics: PSNR, SSIM, PDist, Style Loss, and Color Loss. The latter two decrease by **68.75%** and **14.86%**, respectively, while Text Only retains higher CLIP-I and DINO.

In the multi-subject setting, joint Text+EEG improves all seven metrics over EEG Only and six over Text Only, except DINO. Style Loss reaches **0.0450**, a **29.58%** reduction from EEG Only and **78.00%** from Text Only. Joint conditioning also achieves the best PSNR, SSIM, PDist, Style Loss, and Color Loss in the single-subject setting. These three-seed means support EEG-only style guidance and complementary benefits from joint conditioning under the tested conditions.

Qualitative Comparison. Figure 4 illustrates differences in color fidelity. In the sunset example, EEG Only captures the target's pink sky and cool sea tones, whereas Text Only shifts toward yellow and orange. For the pig among clouds, EEG Only more closely matches the target's pink-and-white palette, while Text Only introduces golden hues. Text Only often preserves clearer structure; EEG Only more closely matches target colors under the tested prompts, consistent with its lower Color Loss in both subject settings.

5.5 Ablation Study

We ablate components on multi-subject EEG-Style with EEG+Text (Table 4). **Text teacher and Neural Adapter.** The teacher improves semantic similarity and Style Loss, supporting semantic supervision. The dual-pathway adapter improves all seven metrics, supporting joint sequence-level and pooled conditioning for faithful stylization.

NSIM. Quality-aware conditioning favors semantic similarity, rank scheduling improves style and color fidelity, and soft masking benefits structural similarity. Timestep-aware routing has mixed effects. These results suggest complementary roles, with no individual mechanism leading on all metrics. **Cross-task pretraining.** Gains in semantic and structural similarity accompany losses elsewhere, indicating selective transfer . **Cross-subject evaluation.** LOSO preserves comparable CLIP-I but worsens other metrics, suggesting that semantic alignment alone is insufficient for faithful stylization of unseen subjects.

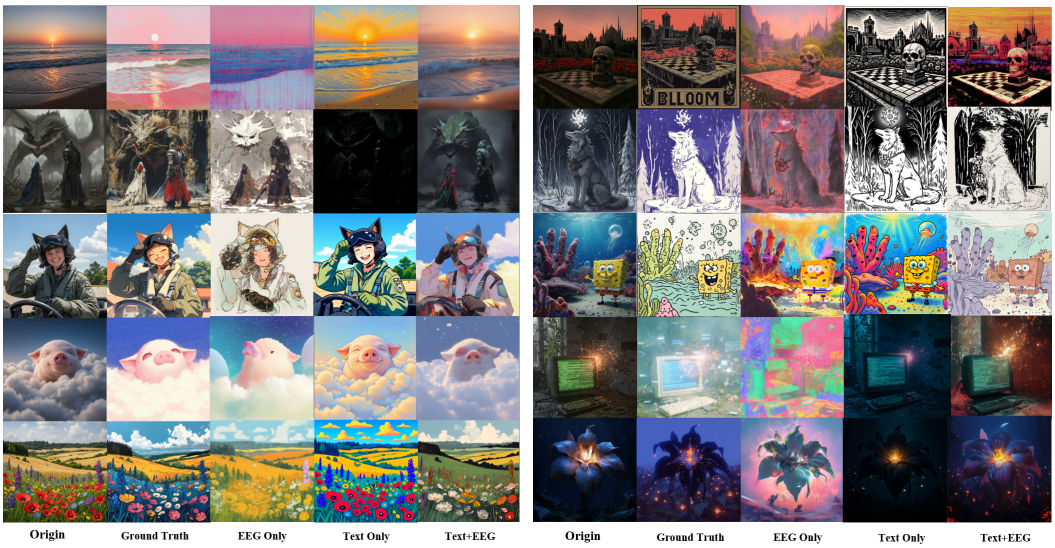

Figure 4: Qualitative comparison on EEG-Style. We show original images, ground truth stylizations, EEG Only, Text Only, and EEG+Text results across representative style editing cases.

Table 4: Ablations on multi-subject EEG-Style with EEG+Text. Unchecked dual-pathway entries use a single pathway. Values are means over three random seeds, with 95% CIs where available. Bold indicates block-wise best results; gray highlights the full model. Pretraining and LOSO are evaluated separately.

Component	Configuration			Metrics						
	Dual pathway	NSIM	Text teacher	CLIP-I ↑	DINO ↑	PSNR ↑	SSIM ↑	PDist ↓	Style loss ↓	Color loss ↓
Overall components				0.7514 ± 0.013	0.5806 ± 0.015	9.704 ± 0.219	0.2615 ± 0.007	0.6859 ± 0.018	0.1409 ± 0.009	0.2782 ± 0.0010
			✓	0.7686 ± 0.012	0.5889 ± 0.016	9.7377 ± 0.272	0.2684 ± 0.009	0.6915 ± 0.018	0.1385 ± 0.009	0.2684 ± 0.009
		✓	✓	0.7843 ± 0.017	0.5968 ± 0.013	9.7906 ± 0.258	0.2669 ± 0.008	0.6830 ± 0.020	0.1256 ± 0.008	0.2492 ± 0.012
	✓	✓	✓	**0.8347** ± 0.018	**0.6102** ± 0.015	**11.0748** ± 0.264	**0.3316** ± 0.008	**0.6069** ± 0.019	**0.0450** ± 0.009	**0.2365** ± 0.010
NSIM	Regular (no NSIM)			0.7686 ± 0.015	0.5889 ± 0.016	9.7377 ± 0.249	0.2684 ± 0.009	0.6915 ± 0.018	0.1385 ± 0.010	0.2684 ± 0.010
	Quality-aware			**0.8217** ± 0.014	**0.6201** ± 0.015	9.7058 ± 0.252	0.2544 ± 0.008	**0.6687** ± 0.017	0.1494 ± 0.009	0.2477 ± 0.011
	Timestep-aware			0.7783 ± 0.016	0.5876 ± 0.014	9.8347 ± 0.263	0.2582 ± 0.008	0.6986 ± 0.018	0.1398 ± 0.010	0.2453 ± 0.012
	Rank schedule			0.7665 ± 0.015	0.5876 ± 0.014	**10.0500** ± 0.258	0.2774 ± 0.009	0.6939 ± 0.018	**0.0954** ± 0.008	**0.2115** ± 0.010
	Soft mask			0.7829 ± 0.016	0.6072 ± 0.015	9.7193 ± 0.237	**0.2780** ± 0.009	0.6739 ± 0.019	0.1347 ± 0.010	0.2566 ± 0.011
Neural Adapter	Single-pathway			0.7843 ± 0.017	0.5968 ± 0.013	9.7906 ± 0.258	0.2669 ± 0.008	0.6830 ± 0.020	0.1256 ± 0.008	0.2492 ± 0.012
	Dual-pathway (full)			**0.8347** ± 0.018	**0.6102** ± 0.015	**11.0748** ± 0.264	**0.3316** ± 0.008	**0.6069** ± 0.019	**0.0450** ± 0.009	**0.2365** ± 0.010
Pre-train	Generate + edit			0.8467 ± 0.017	0.6241 ± 0.014	11.0152 ± 0.0225	0.3523 ± 0.009	0.6139 ± 0.024	0.0570 ± 0.010	0.2379 ± 0.014
LOSO	Held-out subject			0.8382 ± 0.016	0.5938 ± 0.012	10.7440 ± 0.0268	0.2907 ± 0.010	0.6689 ± 0.025	0.0870 ± 0.010	0.2547 ± 0.014

6 Conclusion and Limitations

We presented Uni-Neur2Img for EEG-guided generation, editing, and stylization, alongside the EEG-Style benchmark. The framework combines neural feature alignment, adaptive LoRA, and quality-aware attention, achieving strong performance across three benchmarks with task-specific fine-tuning. In the pooled five-participant setting, EEG-only stylization improves style and color fidelity over the tested text-only baseline without textual prompts or a text teacher. Joint EEG and text conditioning improves most metrics over either modality alone, while text-only conditioning retains higher DINO similarity. Pooled EEG-only training also improves over single-subject training, but the effects of data scale and participant diversity cannot be isolated. These findings suggest that target-evoked EEG provides useful signals for stylization and can complement textual guidance.

Limitations. The small participant cohort and evaluation on synthetic styles under controlled conditions limit generalizability. Future work should broaden participant and visual diversity and examine natural viewing and interactive use.

Ethics Statement

The EEG-Style data collection protocol was approved by the institutional ethics committee, and experiments were conducted under the guidance of professionals with neuroscience and clinical EEG expertise. All participants provided written informed consent after receiving an explanation of the study procedures and were explicitly informed that their data would be used for research and scientific publications. Each participant received financial compensation for their time and participation.

EEG recordings may contain sensitive personal information. Any future data release should therefore remain within the scope of participant consent and ethics approval, remove direct identifiers, and include safeguards against re-identification and unauthorized use. Dataset documentation should describe collection procedures, intended uses, and limitations, while respecting the licenses and permissions associated with existing datasets and source images.

Our experiments concern offline image synthesis from visually evoked neural responses; they do not establish unrestricted thought decoding or support clinical inference. Potential misuse includes non-consensual profiling, surveillance, and deceptive image generation. The limited participant diversity and synthetic visual stimuli may also introduce demographic and stylistic biases, limiting generalization. Future applications should assess these risks and avoid unsupported claims about individuals or populations. The authors remain responsible for the integrity of the reported methods, results, and disclosures, including the use of language models described in Appendix H.

Reproducibility Statement

We describe the model architecture and training objectives in Section 4 and Appendix A. Section 3 specifies the source–target–EEG pairing, dataset sources, preprocessing, and partitioning; Appendix B adds acquisition details. Section 5.1 reports hardware, optimization settings, and repeated-run reporting. Appendix C.2 describes the metrics, and Appendix E reports extended ablations and cross-subject evaluations. Tables 3 and 4 report the means and available confidence intervals for the main stylization and ablation comparisons.

References

Gi-Yeul Bae and Steven J. Luck. Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials. *Journal of Neuroscience*, 38(2):409–422, 2018. doi: 10.1523/JNEUROSCI.2860-17.2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. DreamDiffusion: Generating high-quality images from brain EEG signals. arXiv preprint arXiv:2306.16934, 2023. URL https://arxiv.org/abs/2306.16934.

Black Forest Labs. FLUX.1-dev, 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-dev. Model card.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. URL https://arxiv.org/abs/2506.15742.

Eric L. Buehler and Markus J. Buehler. X-LoRA: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design. *APL Machine Learning*, 2(2):026119, 2024. doi: 10.1063/5.0203126.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Sicheng Dai, Hongwang Xiao, Shan Yu, and Qiwei Ye. Autoregressive visual decoding from EEG signals. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a4dfcbcfa1f0425cd18aafa35a68019a-Abstract-Conference.html.

Keith M. Davis, III, Carlos de la Torre-Ortiz, and Tuukka Ruotsalo. Brain-supervised image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18480–18489, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Davis_Brain-Supervised_Image_Editing_CVPR_2022_paper.html.

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

Ahmed Fares, Sheng-hua Zhong, and Jianmin Jiang. Brain-media: A dual conditioned and lateralization supported GAN (DCLS-GAN) towards visualization of image-evoked brain activities. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1764–1772, 2020. doi: 10.1145/3394171.3413858. URL https://doi.org/10.1145/3394171.3413858.

Pinyuan Feng, Hossein Adeli, Wenxuan Guo, Fan Cheng, Ethan Hwang, and Nikolaus Kriegeskorte. Towards interpretable visual decoding with attention to brain representations. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/46cb65d30bd4931f3e71df8239c92ff0-Abstract-Conference.html.

Joshua J. Foster, Emma M. Bsales, Russell J. Jaffe, and Edward Awh. Alpha-band activity reveals spontaneous representations of spatial position in visual working memory. *Current Biology*, 27 (20):3216–3223.e6, 2017. doi: 10.1016/j.cub.2017.09.031.

Honghao Fu, Hao Wang, Jing Jih Chin, and Zhiqi Shen. BrainVis: Exploring the bridge between brain and visual signals via image reconstruction. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189. PMLR, 2015. URL https://proceedings.mlr.press/v37/ganin15.html.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

Jasper E. Hajonides, Anna C. Nobre, Freek van Ede, and Mark G. Stokes. Decoding visual colour from scalp electroencephalography measurements. *NeuroImage*, 237:118030, 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2021.118030. URL https://www.sciencedirect.com/science/article/pii/S1053811921003074.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations (ICLR)*, 2021.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369. IEEE, 2010. doi: 10.1109/ICPR.2010.579.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context LoRA for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. URL https://arxiv.org/abs/2410.23775.

Qunjie Huang and Weina Zhu. SATTC: Structure-aware label-free test-time calibration for cross-subject EEG-to-image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16887–16896, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_SATTC_Structure-Aware_Label-Free_Test-Time_Calibration_for_Cross-Subject_EEG-to-Image_Retrieval_CVPR_2026_paper.html.

Shijie Huang, Yiren Song, Yuxuan Zhang, Hailong Guo, Xueyin Wang, Mike Zheng Shou, and Jiaming Liu. PhotoDoodle: Learning artistic image editing from few-shot pairwise data. *arXiv preprint arXiv:2502.14397*, 2025a. URL https://arxiv.org/abs/2502.14397.

Shuai Huang, Huan Luo, Haodong Jing, Qixian Zhang, Litao Chang, Yating Feng, Xiao Lin, Chendong Qin, Han Chen, Shuwen Jia, Siyi Sun, and Yongxiong Wang. NEED: Cross-subject and cross-task generalization for video and image reconstruction from EEG signals. In *Advances in Neural Information Processing Systems*, volume 38, pp. 191829–191868, 2025b. doi: 10.52202/085713-5761. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fca83589e85cb061631b7ebc5db5d6bd-Abstract-Conference.html.

Bernd Jähne. *Digital Image Processing*. Springer, 6th edition, 2005. doi: 10.1007/3-540-27563-0.

Jukka-Pekka Kauppi, Melih Kandemir, Veli-Matti Saarinen, Lotta Hirvenkari, Lauri Parkkonen, Arto Klami, Riitta Hari, and Samuel Kaski. Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals. *NeuroImage*, 112:288–298, 2015. doi: 10.1016/j.neuroimage.2014.12.079.

Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2Image: Converting brain signals into images. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1809–1817, 2017. doi: 10.1145/3123266.3127907.

Sanchita Khare, Rajiv Nayan Choubey, Loveleen Amar, and Venkanna Udutalapalli. NeuroVision: perceived image regeneration using cProGAN. *Neural Computing and Applications*, 34(8):5979–5991, 2022. doi: 10.1007/s00521-021-06774-1.

Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. Seeing through the brain: image reconstruction of visual perception from human brain signals. *arXiv preprint arXiv:2308.02510*, 2023. URL https://arxiv.org/abs/2308.02510.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024. URL https://arxiv.org/abs/2403.07721.

Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey M. Siskind. The perils and pitfalls of block design for EEG classification experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):316–333, 2021. doi: 10.1109/TPAMI.2020.2973153.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Xuan-Hao 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 *Advances in Neural Information Processing Systems*, volume 37, pp. 72245–72273, 2024. doi: 10.52202/079017-2306. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/84bad835faaf48f24d990072bb5b80ee-Abstract-Conference.html.

Xuan-Hao Liu, Bao-Liang Lu, and Wei-Long Zheng. EEGMirror: Leveraging EEG data in the wild via montage-agnostic self-supervision for EEG to video decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18273–18283, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_EEGMirror_Leveraging_EEG_Data_in_the_Wild_via_Montage-Agnostic_Self-Supervision_ICCV_2025_paper.html.

Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5790–5799, 2017.

Eleonora Lopez, Luigi Sigillo, Federica Colonnese, Massimo Panella, and Danilo Comminiello. Guess what I think: Streamlined EEG-to-image generation with latent diffusion models. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10890059.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C. Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008. doi: 10.1016/j.neuron.2008.11.004.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024. doi: 10.1609/aaai.v38i5.28226.

Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, and Mubarak Shah. Generative adversarial networks conditioned by brain signals. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3410–3418, 2017.

Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, 2021. doi: 10.1109/TPAMI.2020.2995909.

Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your GAN: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023. doi: 10.1145/3588432.3591500. URL https://arxiv.org/abs/2305.10973.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Bujar Raufi and Luca Longo. An evaluation of the EEG alpha-to-theta and theta-to-alpha band ratios as indexes of mental workload. *Frontiers in Neuroinformatics*, 16:861967, 2022. doi: 10.3389/fninf.2022.861967.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Prajwal Singh, Dwip Dalal, Gautam Vashishtha, Krishna Miyapuram, and Shanmuganathan Raman. Learning robust deep visual representations from EEG brain recordings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7553–7562, 2024.

Vera Soboleva, Aibek Alanov, Andrey Kuznetsov, and Konstantin Sobolev. T-LoRA: Single image diffusion model customization without overfitting. *arXiv preprint arXiv:2507.05964*, 2025. URL https://arxiv.org/abs/2507.05964.

Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4503–4511, 2017. doi: 10.1109/CVPR.2017.479.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14453–14463, 2023.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14940–14950, 2025.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *International Conference on Learning Representations (ICLR)*, 2024.

Xin Xiang, Wenhui Zhou, Haonan Zhu, Yunrui Li, Guojun Dai, and Lili Lin. EEG-driven natural image reconstruction with regional semantic awareness. *Pattern Recognition*, 172:112589, 2026. doi: 10.1016/j.patcog.2025.112589. URL https://www.sciencedirect.com/science/article/pii/S003132032501252X.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. URL https://arxiv.org/abs/2308.06721.

Chenglong Yu, Shuai Shen, Xiangsheng Li, and Yang Li. D^2-FOSA: Dual-diffusion guided EEG-to-image reconstruction with frequency-oriented semantic alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26698–26710, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yu_D2-FOSA_Dual-Diffusion_Guided_EEG-to-Image_Reconstruction_with_Frequency-Oriented_Semantic_Alignment_CVPR_2026_paper.html.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. CognitionCapturer: Decoding visual stimuli from human EEG signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025. doi: 10.1609/aaai.v39i13.33587. URL https://ojs.aaai.org/index.php/AAAI/article/view/33587.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Xiao Zheng, Wanzhong Chen, Mingyang Li, Tao Zhang, Yang You, and Yun Jiang. Decoding human brain activity with deep learning. *Biomedical Signal Processing and Control*, 56:101730, 2020. doi: 10.1016/j.bspc.2019.101730.

Pengfei Zhou, Jie Xia, Xiaopeng Peng, Wangbo Zhao, Zilong Ye, Zekai Li, Suorong Yang, Jiadong Pan, Yuanxiang Chen, Ziqiao Wang, Kai Wang, Qian Zheng, Xiaojun Chang, Gang Pan, Shurong Dong, Kaipeng Zhang, and Yang You. Neural-driven image editing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/9138f908b3cfe807ee3288dbfd1156d3-Abstract-Conference.html.