MINDAU: EEG-CONDITIONED FACIAL ACTION UNIT EDITING VIA DUAL-STREAM MANIFOLD ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent brain decoding studies have made substantial progress in reconstructing externally perceived visual content from neural signals. However, using electroencephalography (EEG) recordings to guide facial expression editing remains largely unexplored and poses a distinct challenge: rather than recovering what a subject sees, it requires identifying facial-action related patterns from noisy EEG signals and grounding them in localized, identity-preserving expression edits. In this paper, we investigate EEG-conditioned facial image editing for fine-grained facial action unit (AU) control and propose **MindAU**, a unified framework for controlling facial AU edits from EEG signals. MindAU first learns noise-robust and AU-discriminative EEG representations through temporal masked reconstruction and AU classification supervision. It then bridges the modality gap via Dual-Stream Manifold Alignment, aligning EEG features with AU-level text semantics and identity-reduced visual displacement trajectories in the multimodal space of Qwen2.5-VL. Finally, MindAU incorporates EEG-aware Multimodal Rotary Positional Embeddings, landmark-guided reference masking, and AU-aware region supervision into a multimodal diffusion-based editor for high-fidelity identity-preserving editing. We also introduce **E-CAFE**, a curated benchmark for EEG-Conditioned Action-Unit Facial Editing with paired EEG-face editing samples and standardized evaluation protocols. Extensive experiments demonstrate the effectiveness of MindAU and suggest its potential as a step towards future assistive expression technologies for individuals with facial neuromuscular disorders.

1 INTRODUCTION

Recent brain decoding studies have shown remarkable progress in reconstructing externally perceived visual content from neural signals, ranging from static images Dai et al. (2025); Wang et al. (2024b); Kong et al. (2024) to dynamic videos Lu et al. (2024); Chen et al. (2023b); Liu et al. (2024). Most existing methods focus on recovering what a subject sees, typically by aligning neural representations with visual or semantic spaces through contrastive learning Sun et al. (2023); Scotti et al. (2023) or signal reconstruction Bai et al. (2024). However, an underexplored and potentially impactful direction is to move beyond passive perceptual reconstruction towards controllable facial behavior editing conditioned on brain activity, establishing a pathway from neural signals to interpretable facial actions. This is especially relevant when paired neural recordings and facial-expression observations are scarce: electroencephalography (EEG)-conditioned facial editing may enrich EEG–face supervision and facilitate cross-modal modeling between brain activity and facial dynamics. It may also support future internal-state modeling, such as affect- or pain-related analysis, and assistive communication interfaces where expressive facial behavior is driven by neural signals.

Facial expressions can be decomposed into facial action units (AUs), providing a structured and interpretable control space for linking neural signals to localized facial movements. Unlike reconstructing perceived stimuli, EEG-conditioned facial editing requires extracting subtle and noisy facial-action cues from EEG signals and grounding them in fine-grained, identity-preserving expression changes on a reference face, posing a distinct and challenging problem. To the best of our knowledge, no prior work has addressed EEG-conditioned facial image editing with AU-level control. To fill this gap, we propose MindAU, the first framework for EEG-conditioned facial image editing at the action-unit level, which also serves as an early step towards future assistive expression technologies

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.

for individuals with facial neuromuscular disorders Peitersen (2002); Jack & Schyns (2015); Hotton et al. (2020).

Despite its promise, EEG-driven facial expression editing is challenged by noisy signals, cross-modal alignment gaps, identity-preserving editing, and limited paired EEG-face data. To address these challenges, we propose MindAU, an EEG-conditioned generative framework for identity-preserving facial action unit editing. **First**, an AU-Aware EEG Encoder learns noise-robust and AU-discriminative representations via temporal masked reconstruction and AU classification supervision. **Second**, a Dual-Stream Manifold Alignment module maps EEG features to Qwen2.5-VL's multimodal space, aligning them with both AU-level text semantics and identity-reduced visual displacement trajectories. **Third**, we incorporate EEG-aware M-RoPE, landmark-guided reference masking, and AU-aware region supervision into a multimodal diffusion-based editor to enable EEG-driven local expression editing while preserving reference identity. **Finally**, we introduce **E-CAFE**, a curated benchmark built upon BU-EEG Li et al. (2020), containing 5,000 paired EEG–face editing samples and two standardized protocols for self-referenced and cross-identity evaluation.

The primary contributions are summarized as follows:

- We formulate EEG-conditioned facial expression editing as a fine-grained, identity-preserving facial action unit editing task. To the best of our knowledge, MindAU is the first framework specifically designed for EEG-conditioned action-unit-level facial image editing.
- We introduce an AU-Aware EEG Encoder with temporal masked reconstruction and AU classification supervision for noise-robust, AU-discriminative EEG representation learning.
- We propose Dual-Stream Manifold Alignment to bridge EEG representations with Qwen2.5-VL's multimodal space through AU-level text semantics and identity-reduced visual displacement trajectories.
- We adapt a multimodal diffusion-based editing backbone with EEG-aware M-RoPE, landmark-guided reference masking, and AU-aware region supervision to mitigate shortcut learning and preserve reference identity.
- We construct **E-CAFE**, a curated benchmark built upon BU-EEG with 5,000 paired EEG–face editing samples and two standardized evaluation protocols.

2 Related Work

2.1 Representation Learning for EEG Signals

Learning robust representations from noisy EEG signals is essential for brain decoding. Existing pre-training strategies mainly fall into two categories. Generative methods model EEG signal distributions through raw waveform reconstruction Bai et al. (2024); Barmpas et al. (2025) or time-frequency codebook learning Jiang et al. (2024a;b). Task-oriented approaches, often based on contrastive learning Song et al. (2023); Wang et al. (2024a); Camaret Ndir et al. (2025), map EEG representations into downstream spaces such as semantic classification Fu et al. (2025) or cross-modal retrieval Li et al. (2024). However, these methods are not explicitly grounded in facial action semantics. EEG-conditioned facial editing requires noise-robust and AU-discriminative features to control local facial movements from neural signals. We therefore introduce AU-aware EEG pre-training with temporal masked reconstruction and AU classification supervision.

2.2 Brain-Guided Image Generation and Editing

Brain-guided visual generation synthesizes images from neural signals such as fMRI and EEG. Early methods used GANs Kavasidis et al. (2017) to reconstruct perceived stimuli, whereas recent approaches adopt diffusion models Zeng et al. (2024); Bai et al. (2024); Fu et al. (2025). Most existing works focus on passive perceptual reconstruction from fMRI Chen et al. (2023a); Scotti et al. (2023) or EEG Wang et al. (2024a) recorded during stimulus viewing. In parallel, text-guided diffusion models Labs (2024); Esser et al. (2024) and facial editing models Team et al. (2025); Cai et al. (2025) have achieved high-fidelity semantic editing with strong identity preservation, but they rely on explicit text or image conditions and cannot be directly driven by noisy brain signals. Although several studies use brain signals as implicit guidance for image editing Davis et al. (2022); Yu et al. (2025b);

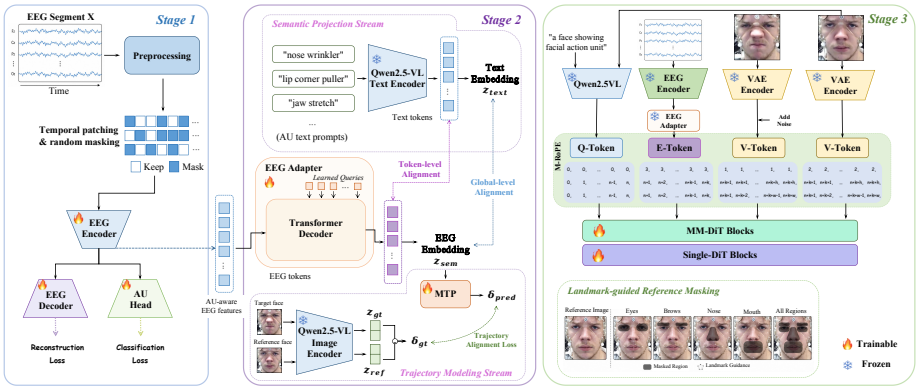


Figure 1: Overall architecture of MindAU. MindAU learns AU-discriminative EEG representations, aligns them with AU-level semantics and identity-reduced visual displacement trajectories, and conditions a multimodal diffusion editor for identity-preserving facial AU editing. MTP denotes the Manifold Trajectory Projector.

Chen et al. (2024); Yu et al. (2025a), they mainly rely on neural responses to external stimuli rather than grounding facial-action-related EEG cues in localized expression edits. Recent work has also explored EEG-based facial synthesis. Mind-to-Face Xiong et al. (2025) maps EEG signals to dense 3D facial geometry for subject-specific avatar synthesis, whereas our method targets reference-based facial image editing with transferable AU-level control and identity preservation.

3 METHODOLOGY

Given an EEG segment $\mathbf{X}$ and a reference face image $\mathbf{I}_{ref}$, MindAU generates an edited image $\hat{\mathbf{I}}$ that preserves the identity in $\mathbf{I}_{ref}$ while expressing the facial action implied by $\mathbf{X}$. It proceeds in three stages: (i) *Action-Unit-Aware EEG Pre-training* learns AU-discriminative EEG representations; (ii) *Dual-Stream Manifold Alignment* maps these representations to semantic embeddings and identity-reduced visual displacement trajectories; and (iii) *EEG-Conditioned Facial Editing*, which conditions an MM-DiT-based face editing backbone on the aligned EEG tokens and the reference face to produce identity-preserving expression edits. We also introduce **E-CAFE**, a benchmark for **E**EG-**C**onditioned **A**ction-unit **F**acial **E**diting.

3.1 AU-AWARE EEG ENCODER PRE-TRAINING

EEG preprocessing. For each trial, we remove DC offset, apply a fourth-order Butterworth band-pass filter between 0.1 and 50 Hz, and perform channel-wise Z-score normalization. This standard preprocessing removes slow drift, attenuates high-frequency noise, and normalizes channel-wise signal scales before temporal masked modeling.

Temporal Masked Modeling. To learn robust representations from noisy EEG signals, we adopt a temporal Masked Autoencoder (MAE) backbone He et al. (2022), following DreamDiffusion Bai et al. (2024). Let $\mathbf{X} \in \mathbb{R}^{C\times T}$ denote a preprocessed EEG segment with C channels and T time points. We partition $\mathbf{X}$ into temporal patches and randomly mask a subset to obtain $\tilde{\mathbf{X}}$. The encoder E_θ maps $\tilde{\mathbf{X}}$ to $\mathbf{z} = E_\theta(\tilde{\mathbf{X}})$, and the decoder D_ϕ reconstructs the masked patches from $\mathbf{z}$. This objective encourages contextual temporal modeling rather than reliance on local noisy fluctuations:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{M} \odot (\mathbf{X} - D_\phi(\mathbf{z}))\|_2^2, \tag{1}$$

where $\mathbf{M}$ is the temporal patch mask and $\odot$ denotes element-wise multiplication.

AU Semantic Supervision. Since signal reconstruction lacks explicit AU-level semantic grounding, we add an AU classification objective to learn AU-discriminative EEG features. Specifically, AU prediction is formulated as a multi-class classification task over predefined AU categories. The encoder feature $\mathbf{z}$ is fed into a classification head $f_{cls}(\cdot)$, and the loss is

$$\mathcal{L}_{cls} = \text{CE}(f_{cls}(\mathbf{z}), y), \tag{2}$$

where y is the ground-truth AU label and $\mathrm{CE}(\cdot,\cdot)$ denotes cross-entropy. The Stage-1 objective is

$$\mathcal{L}_{stage1} = \mathcal{L}_{rec} + \lambda_{cls}\mathcal{L}_{cls}, \tag{3}$$

where λ_{cls} balances reconstruction and AU supervision. This pre-training yields temporally contextualized and AU-aware EEG representations for subsequent cross-modal alignment.

3.2 Dual-Stream Manifold Alignment

Although the Stage-1 encoder captures AU-correlated EEG features, its latent space is not directly compatible with the semantic space used by pretrained multimodal generative models. We therefore introduce *Dual-Stream Manifold Alignment* to bridge this gap. The *Semantic Projection Stream* maps EEG features to the text-semantic space of Qwen2.5-VL, while the *Trajectory Modeling Stream* aligns them with identity-reduced visual displacement trajectories for fine-grained editing.

Semantic Projection Stream. We introduce an EEG Adapter to convert Stage-1 EEG features into semantic tokens compatible with Qwen2.5-VL Bai et al. (2025). Given the EEG feature sequence, $Q = 32$ learnable query tokens cross-attend to it through a Transformer Decoder. These queries are randomly initialized and optimized with the EEG Adapter. The resulting query features are projected into the embedding space of the frozen Qwen2.5-VL text encoder, forming the EEG semantic tokens. We use their pooled representation $\mathbf{z}_{sem}$ for global-level alignment and retain the full token sequence as E-tokens for token-level supervision and Stage-3 editing.

As semantic targets, we encode AU-specific natural language descriptions with the frozen Qwen2.5-VL text encoder and denote the pooled text embedding as $\mathbf{z}_{text}$. We first align $\mathbf{z}_{sem}$ and $\mathbf{z}_{text}$ using a cross-modal supervised contrastive loss:

$$\mathcal{L}_{align} = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{|P(i)|}\sum_{j\in P(i)}\log\frac{\exp(\mathrm{sim}(\mathbf{z}_{sem,i},\mathbf{z}_{text,j})/\tau)}{\sum_{k=1}^{B}\exp(\mathrm{sim}(\mathbf{z}_{sem,i},\mathbf{z}_{text,k})/\tau)}, \tag{4}$$

where $P(i)$ denotes the set of text samples sharing the same AU label as EEG sample i, $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity, and τ is the temperature.

To preserve token-level semantic structure, we further align EEG tokens with teacher text tokens. Let $\mathbf{E}_i = \{\mathbf{e}_{i,1},\ldots,\mathbf{e}_{i,Q}\}$ be the EEG tokens of sample i, and $\mathbf{T}_i = \{\mathbf{t}_{i,1},\ldots,\mathbf{t}_{i,L}\}$ be its valid teacher text tokens. For each EEG token, we keep its maximum similarity to any valid text token:

$$s_{i,q} = \max_{l:\,\mu_{i,l}=1}\mathrm{sim}(\mathbf{e}_{i,q},\mathbf{t}_{i,l}), \tag{5}$$

where $\mu_{i,l}$ is the text validity mask. The token-level alignment loss is

$$\mathcal{L}_{tok} = 1 - \frac{1}{BQ}\sum_{i=1}^{B}\sum_{q=1}^{Q}s_{i,q}, \tag{6}$$

where Q is the number of EEG tokens. This loss avoids enforcing one-to-one positional correspondence and instead allows each EEG token to align with its most relevant text token, preserving local semantic structure beyond the pooled embedding.

Trajectory Modeling Stream. Although the Semantic Projection Stream aligns EEG features with AU-level text semantics, text supervision is categorical and cannot capture within-AU variation, such as expression intensity. We therefore introduce a Trajectory Modeling Stream to supervise EEG features with identity-reduced visual displacement trajectories.

Given the visual embeddings of the reference image $\mathbf{z}_{ref}$ and ground-truth target image $\mathbf{z}_{gt}$ extracted by the frozen Qwen2.5-VL visual encoder Bai et al. (2025), we compute the raw visual displacement as

$$\boldsymbol{\delta}_{raw} = \mathbf{z}_{gt} - \mathbf{z}_{ref}.$$

Since this displacement may contain identity-related components, we reduce the component aligned with the reference embedding by projecting it onto the orthogonal complement of $\mathbf{z}_{ref}$:

$$\boldsymbol{\delta}_{gt} = \boldsymbol{\delta}_{raw} - \frac{\boldsymbol{\delta}_{raw}\cdot\mathbf{z}_{ref}}{\|\mathbf{z}_{ref}\|^2}\mathbf{z}_{ref}. \tag{7}$$

The resulting $\boldsymbol{\delta}_{gt}$ serves as an identity-reduced visual displacement target that emphasizes expression-specific deformation relative to the reference face.

We train the Manifold Trajectory Projector (MTP) to predict an EEG-driven trajectory $\boldsymbol{\delta}_{pred}$ from the pooled EEG representation. The trajectory loss combines directional consistency and magnitude matching:

$$\mathcal{L}_{traj} = \lambda_{cos}\big(1 - \cos(\boldsymbol{\delta}_{pred}, \boldsymbol{\delta}_{gt})\big) + \frac{\lambda_{L1}}{D}\|\boldsymbol{\delta}_{pred} - \boldsymbol{\delta}_{gt}\|_1, \tag{8}$$

where D is the feature dimension, and λ_{cos} and λ_{L1} balance the two terms.

The overall Stage-2 objective is

$$\mathcal{L}_{stage2} = \mathcal{L}_{align} + \lambda_{tok}\mathcal{L}_{tok} + \lambda_{traj}\mathcal{L}_{traj}, \tag{9}$$

where λ_{tok} and λ_{traj} balance token-level semantic alignment and trajectory supervision. To provide coarse magnitude-aware supervision for Stage-2 alignment, we synthesize identity-diversified expression-transition pairs with Nano Banana 2 Google (2026). For each real sample, we use the reference-to-target facial change, including the target AU and observed expression strength, to guide the same intended transition on four generated identities. Rather than serving as exact trajectory targets, these pairs provide an approximate prior on EEG-associated deformation magnitude across identities, helping the EEG Adapter learn identity-invariant AU deformation and coarse intensity variation before fine-tuning on real EEG–face pairs.

3.3 EEG-Conditioned Facial Editing

With the aligned EEG representations, MindAU builds on a Multimodal Diffusion Transformer (MM-DiT)Esser et al. (2024) for identity-preserving facial editing. As shown in Figure 1 (Stage 3), a fixed text prompt and a reference image are encoded by the frozen Qwen2.5-VL to produce semantic Q-tokens, while the frozen EEG encoder and EEG Adapter map the EEG signal into 32 E-tokens. The ground-truth image is encoded by the VAE and perturbed with Gaussian noise to form the noisy latent V-token$_{noise}$, and the reference image is encoded as V-token$_{ref}$ to provide identity information.

To fuse these heterogeneous tokens in a shared attention space, we redesign Multimodal Rotary Positional Embedding (M-RoPE)Team et al. (2025). The positional encoding contains one modality dimension, which distinguishes text, EEG, ground-truth image, and reference image, and two spatial dimensions, which encode image geometry. Image tokens use their natural (h, w) coordinates, while each 1D text or EEG token uses its sequence index for both spatial dimensions. Following LongCat-Image Team et al. (2025), the backbone consists of double-stream blocks followed by single-stream blocks.

We train the model with Conditional Flow Matching (CFM). Let $\mathbf{z}_0$ denote the latent of the ground-truth image and $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote Gaussian noise. The interpolation path is

$$\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1, \tag{10}$$

with target velocity

$$\mathbf{v}_{target} = \frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}t} = \mathbf{z}_1 - \mathbf{z}_0. \tag{11}$$

Given the multimodal condition set

$$\mathcal{C} = \{\text{Q-token}, \text{E-token}, \text{V-token}_{ref}\}, \tag{12}$$

the model v_θ predicts the velocity from $\mathbf{z}_t$, yielding the flow-matching loss

$$\mathcal{L}_{FM} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1}\left[\|v_\theta(\mathbf{z}_t, t, \mathcal{C}) - (\mathbf{z}_1 - \mathbf{z}_0)\|_2^2\right]. \tag{13}$$

At inference, the model takes only the EEG segment, the reference image, and the fixed text prompt as conditions; the ground-truth image is used only for training.

Training on limited data can lead to a degenerate shortcut: as the reference image already accounts for most target-image content, the model may learn a reference-dominant near-identity mapping and under-utilize EEG conditions for expression control. To mitigate this issue, we introduce landmark-guided reference masking and AU-aware region supervision.

Landmark-Guided Reference Masking. To reduce expression leakage from the reference branch, we mask AU-relevant regions in the reference image before encoding it as a visual condition. The masks are derived from reference-image landmarks and expanded to cover surrounding AU-related facial regions, including eyes, brows, nose, and mouth. We use expanded ellipses for eyes, brows, and mouth, and a convex hull for the nose; all masks are further dilated to cover nearby expression-related texture changes. We use a progressive masking schedule, decreasing the fraction of training samples with masked references from 70% to 30%. This encourages early reliance on EEG conditions while gradually restoring full reference information for identity-preserving editing at inference.

AU-aware region supervision. In addition to the full-image flow-matching loss, we introduce two spatial losses computed from the same prediction residual. For the b-th sample, the channel-averaged spatial error map in latent space is

$$E_b(h,w) = \frac{1}{D}\sum_{c=1}^{D}\left(v_\theta(\mathbf{z}_{t,b}, t, \mathcal{C}_b)_{c,h,w} - \mathbf{v}_{target,b,c,h,w}\right)^2, \tag{14}$$

where D is the number of latent channels.

We first construct a facial-region mask by taking the union of the four AU-relevant regions:

$$M^{(b)}_{\text{region}} = \max\left(M^{(b)}_{\text{eye}}, M^{(b)}_{\text{brow}}, M^{(b)}_{\text{nose}}, M^{(b)}_{\text{mouth}}\right). \tag{15}$$

The region loss averages prediction error over this facial action-relevant area:

$$\mathcal{L}_{\text{region}} = \frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{h,w} M^{(b)}_{\text{region}}(h,w)\, E_b(h,w)}{\sum_{h,w} M^{(b)}_{\text{region}}(h,w) + \epsilon}, \tag{16}$$

where ϵ is a small constant for numerical stability.

We further define an AU-focused mask according to the target AU label:

$$M^{(b)}_{\text{focus}} = \begin{cases} M^{(b)}_{\text{brow}}, & a_b \in \{AU1, AU2, AU4\}, \\ M^{(b)}_{\text{eye}}, & a_b = AU5, \\ M^{(b)}_{\text{nose}}, & a_b = AU9, \\ M^{(b)}_{\text{mouth}}, & a_b \in \{AU12, AU15, AU17, AU25, AU27\}. \end{cases} \tag{17}$$

The AU-focused loss is

$$\mathcal{L}_{\text{focus}} = \frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{h,w} M^{(b)}_{\text{focus}}(h,w)\, E_b(h,w)}{\sum_{h,w} M^{(b)}_{\text{focus}}(h,w) + \epsilon}. \tag{18}$$

The final Stage-3 objective is

$$\mathcal{L}_{\text{stage3}} = \mathcal{L}_{FM} + \lambda_{\text{region}}\mathcal{L}_{\text{region}} + \lambda_{\text{focus}}\mathcal{L}_{\text{focus}}, \tag{19}$$

where λ_{region} and λ_{focus} balance the two auxiliary losses.

3.4 E-CAFE: A Benchmark for EEG-Conditioned Action Unit Editing

To support standardized evaluation of EEG-driven facial editing, we introduce **E-CAFE** (**E**EG-**C**onditioned **A**ction-unit **F**acial **E**diting), a curated benchmark built upon BU-EEG Li et al. (2020). E-CAFE provides paired EEG–face editing samples together with two evaluation protocols.

For subject-independent evaluation, we split BU-EEG by identity, reserving two subjects as a strictly held-out test set and using the remaining subjects for training. We first apply the Face Alignment Network (FAN) Bulat & Tzimiropoulos (2017) to detect facial landmarks and filter frames inconsistent with the target AU activation. For each retained target frame, we extract the corresponding preceding EEG segment of 48 time points and a reference face image. The resulting task is to transform the reference image into the target expression under the condition of the EEG segment. For testing, we further manually curate neutral-to-target sequences to reduce expression leakage from the reference image, ensuring that expression changes are primarily driven by EEG conditions.

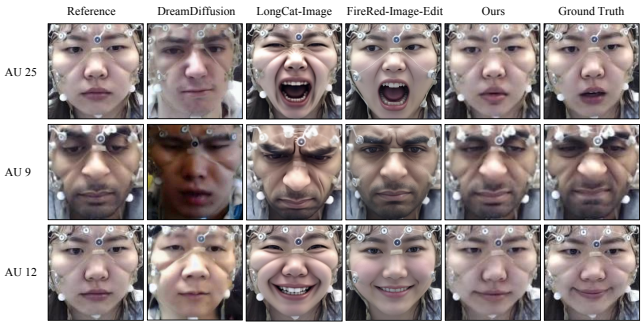


Figure 2: Visual comparison of our approach with other methods. The first, second, and third rows correspond to AU25 (Lips Part), AU9 (Nose Wrinkler), and AU12 (Lip Corner Puller).

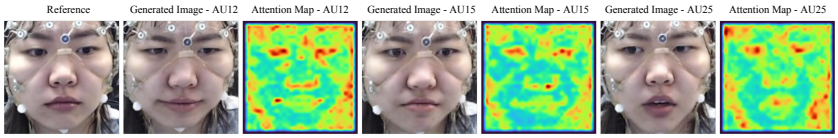


Figure 3: Visualization of EEG attention-sensitive spatial regions in generated images.

E-CAFE includes two complementary evaluation protocols. **(1) Self-Referenced Editing** uses the reference image and EEG signal from the same held-out identity (I_A, E_A). This setting evaluates reconstruction fidelity on unseen subjects by measuring whether the model can translate EEG signals into the corresponding facial dynamics while preserving identity. **(2) Cross-Referenced Editing** evaluates transferability and disentanglement. EEG signals from held-out subjects are used to drive 128 synthetic identities generated by Z-Image Cai et al. (2025), covering diverse ages and genders. This setting tests whether the learned EEG representations capture AU-related expression semantics independent of the source identity.

4 Experiments

4.1 Datasets

We utilize the BU-EEG dataset Li et al. (2020), which includes 28 participants performing 10 distinct AUs (AU1, AU2, AU4, AU5, AU9, AU12, AU15, AU17, AU25, AU27). It provides synchronized 128-channel EEG signals and facial videos sampled at 250 Hz and 24 fps, respectively.

Evaluation Metrics. We adopt a multi-dimensional protocol to comprehensively assess our framework. To quantify image quality and diversity, we compute the **Fréchet Inception Distance (FID)** Heusel et al. (2017) to measure the distributional discrepancy between synthesized and ground-truth images. To evaluate identity preservation, we calculate the **Cosine Similarity (CSIM)** between the ArcFace Deng et al. (2019) embeddings of the generated face and the reference identity. For expression fidelity, we report **AU Accuracy (AU ACC)** using OpenGraphAU Luo et al. (2022), initialized from the model pre-trained on BP4D Zhang et al. (2014) and further adapted on the E-CAFE training split. In addition, we report the **CLIP Score** Hessel et al. (2021) to measure the semantic consistency between generated images and the corresponding conditioning descriptions. More details about the evaluation protocol are provided in the Appendix.

4.2 Implementation Details

Our training pipeline consists of three stages: (1) AU-Aware EEG Encoder Pre-training, (2) Dual-Stream Manifold Alignment, and (3) EEG-Conditioned Image Editing. All experiments are conducted on NVIDIA RTX 6000 Ada GPUs. More details are provided in the Appendix.

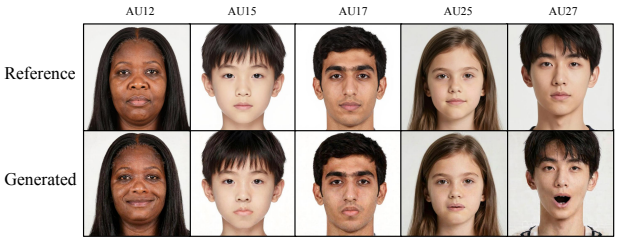


Figure 4: Qualitative results of Cross-Referenced Editing.

Table 1: Quantitative comparison with existing methods. Pipeline baselines first predict AU categories from EEG and then edit images with fixed AU descriptions. "GT Ref." denotes results on ground-truth target images.

Method	CSIM(%) ↑	FID ↓	AU ACC(%) ↑	CLIP Score ↑
DreamDiffusion Bai et al. (2024)	16.59	106.25	9.23	13.85
LongCat-Image Team et al. (2025)	31.65	59.40	20.88	**18.89**
FireRed-Image-Edit Team et al. (2026)	38.42	80.83	19.48	17.40
Ours	**75.33**	**43.35**	**27.71**	17.61
GT Ref.	–		74.29	19.44

4.3 Evaluation

Quantitative Comparison. Since no existing method is designed for EEG-conditioned facial editing, we compare MindAU with three representative baselines under the Self-Referenced Editing protocol. DreamDiffusion Bai et al. (2024) is fine-tuned as an EEG-to-image baseline, generating images directly from test EEG signals without a reference image. We further construct two natural pipeline baselines with LongCat-Image Team et al. (2025) and FireRed-Image-Edit Team et al. (2026), decomposing EEG-conditioned editing into EEG-to-AU recognition followed by text-guided reference-based editing. For a fair comparison, the EEG-to-AU classifier used in these pipelines shares the same Stage-1 EEG encoder and AU classification head as MindAU, achieving 32.93% standalone AU accuracy on the held-out test split. As shown in Table 1, MindAU achieves the best performance on the three most task-critical metrics. Compared with DreamDiffusion, MindAU benefits from the reference image as an explicit identity condition and better preserves identity while translating EEG signals into facial actions. Compared with the two text-guided pipelines, MindAU avoids compressing EEG signals into a discrete AU label before editing. Fixed AU text prompts provide only coarse control and often lead to exaggerated facial deformations or unrecoverable errors when the EEG-to-AU classifier predicts an incorrect label. In contrast, MindAU conditions the editor on continuous EEG-aligned representations, enabling more AU-faithful, intensity-aware, and identity-preserving edits. LongCat-Image achieves the highest CLIP Score, likely due to its direct use of fixed AU text prompts. However, CLIP Score mainly measures coarse text-image consistency and does not fully reflect EEG faithfulness or expression-intensity alignment. We therefore interpret CLIP Score together with CSIM, FID, and AU ACC.

Qualitative Comparison. Figure 2 provides qualitative comparisons. DreamDiffusion often fails to preserve reference identity because it generates images directly from EEG without using the reference image. Text-guided pipelines better retain the reference structure but tend to produce canonical or overly strong expressions, such as exaggerated smiles for AU12, due to fixed textual AU prompts. They are also sensitive to AU recognition errors, where an incorrect predicted AU constrains the downstream editor. MindAU directly conditions generation on continuous EEG-aligned tokens. It produces subtler mouth movements for AU25 and AU12 and a nose-related deformation closer to the target for AU9, while better preserving identity. Figure 3 further shows EEG-sensitive attention maps concentrated on AU-relevant facial regions, indicating that EEG conditions guide local facial generation. Figure 4 shows AU-specific edits on synthetic reference identities under Cross-Referenced Editing. More results are provided in the Appendix.

Table 2: Ablation studies under the Self-Referenced and Cross-Referenced Editing protocols. "Direct" and "SupCon" denote pairwise cosine EEG–text alignment and supervised contrastive global-level alignment, respectively.

Configuration						Self-Referenced				Cross-Referenced		
AU	Align	$\mathcal{L}_{tok}$	$\mathcal{L}_{traj}$	$\mathcal{L}_{reg}$	$\mathcal{L}_{foc}$	CSIM↑	FID↓	AU ACC↑	CLIP↑	CSIM↑	AU ACC↑	CLIP↑
	Direct					74.96	51.09	13.25	14.92	57.88	9.74	19.37
✓	Direct					71.81	45.85	15.66	16.24	54.70	11.14	20.27
✓	SupCon					74.25	43.45	23.89	16.31	51.59	11.88	20.40
✓	SupCon	✓				74.70	**41.87**	20.88	16.02	52.89	11.86	20.43
✓	SupCon		✓			74.11	42.98	24.10	16.14	53.24	11.95	20.41
✓	SupCon	✓	✓			73.50	43.59	24.71	16.26	53.48	12.01	20.52
✓	SupCon	✓	✓	✓		73.34	43.84	26.49	16.78	59.27	17.45	20.62
✓	SupCon	✓	✓	✓	✓	**75.33**	43.35	**27.71**	**17.61**	**66.21**	**20.70**	**21.09**

4.4 Ablation Studies

We conduct ablations under both Self-Referenced and Cross-Referenced Editing protocols. As shown in Table 2, AU classification supervision improves AU ACC and CLIP Score, confirming that explicit AU labels help the EEG encoder learn expression-discriminative representations. Replacing direct pairwise alignment with supervised contrastive alignment brings a large AU ACC gain, highlighting the importance of AU-level semantic structure.

$\mathcal{L}_{tok}$ and $\mathcal{L}_{traj}$ provide complementary supervision. The token-level loss preserves fine-grained semantic structure and yields the best FID, but its flexible token matching does not enforce global AU discrimination or deformation consistency when used alone. The trajectory loss introduces deformation-aware supervision, improving AU fidelity by aligning EEG features with identity-reduced visual changes. Stage-3 spatial losses further improve editing accuracy by concentrating optimization on AU-relevant regions. The full model achieves the best overall trade-off across CSIM, AU ACC, and CLIP Score. Table 2 shows a similar trend under Cross-Referenced Editing. These component-wise results are complemented by a masking ablation (Table 5) and, in Appendix I, controls for EMG influence and reference-image dependence, a MagicFace comparison, and blinded perceptual ratings.

4.5 Limitations

MindAU is evaluated on 28 healthy participants performing individually elicited AUs. Our frequency- and channel-based analysis suggests that preprocessing reduces the influence of face-adjacent activity, but EEG and facial EMG cannot be fully disentangled in BU-EEG. Accordingly, the results do not establish transfer to imagined or attempted facial movements or to patients with facial paralysis. Although we evaluate held-out participants and different reference identities, the small cohort and controlled acquisition setting limit conclusions about broader populations and unconstrained conditions. BU-EEG also lacks synchronized annotations for simultaneous multi-AU actions and temporal expressions, so reliable multi-AU and dynamic editing remain outside the demonstrated scope.

5 Conclusion

MindAU introduces EEG-conditioned facial AU editing with identity-preserving image generation. Its AU-aware EEG encoder, Dual-Stream Manifold Alignment, and EEG-conditioned editor connect neural signals to localized facial changes. We also introduce E-CAFE, a benchmark with subject-independent and cross-identity evaluation protocols. On the main benchmark, MindAU outperforms EEG-to-image and text-guided pipeline baselines in AU accuracy and identity similarity. These findings support the feasibility of single-AU editing from EEG recorded during performed facial actions.

AI USE STATEMENT

We used Nano Banana 2 to generate identity-diversified expression-transition image pairs for Stage-2 pretraining and Z-Image to generate synthetic reference faces for cross-referenced evaluation, as described in the paper. We also used AI tools to assist with polishing some sentences and correcting spelling and grammar. The authors take full responsibility for the manuscript, including its claims, results, and references.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yunpeng Bai, Xintao Wang, Yan-Pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: High-quality eeg-to-image generation with temporal masked signal modeling and clip alignment. In *European Conference on Computer Vision*, pp. 472–488. Springer, 2024.

Konstantinos Barmpas, Na Lee, Alexandros Koliousis, Yannis Panagakis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Neurorvq: Multi-scale eeg tokenization for generative large brainwave models. *arXiv preprint arXiv:2510.13068*, 2025.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pp. 1021–1030, 2017.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Tidiane Camaret Ndir, Robin Tibor Schirrmeister, and Tonio Ball. Eeg-clip: Learning eeg representations from natural language descriptions. *Frontiers in Robotics and AI*, 12:1625731, 2025.

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

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

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36:24841–24858, 2023b.

Yuqin Dai, Zhouheng Yao, Chunfeng Song, Qihao Zheng, Weijian Mai, Kunyu Peng, Shuai Lu, Wanli Ouyang, Jian Yang, and Jiamin Wu. Mindaligner: Explicit brain functional alignment for cross-subject visual decoding from limited fmri data. *arXiv preprint arXiv:2502.05034*, 2025.

Keith M Davis, 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*, pp. 18480–18489, 2022.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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

Google. Gemini 3.1 Flash Image Preview (Nano Banana 2). https://ai.google.dev/gemini-api/docs/image-generation, 2026. Accessed: 2026-05-01.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

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, 2021.

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. *Advances in neural information processing systems*, 30, 2017.

Matthew Hotton, Esme Huggons, Claire Hamlet, Danielle Shore, David Johnson, Jonathan H Norris, Sarah Kilcoyne, and Louise Dalton. The psychosocial impact of facial palsy: a systematic review. *British Journal of Health Psychology*, 25(3):695–727, 2020.

Rachael E Jack and Philippe G Schyns. The human face as a dynamic tool for social communication. *Current Biology*, 25(14):R621–R634, 2015.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. *arXiv preprint arXiv:2409.00101*, 2024a.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024b.

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.

Xiangtao Kong, Kexin Huang, Ping Li, and Lei Zhang. Toward generalizing visual brain decoding to unseen subjects. *arXiv preprint arXiv:2410.14445*, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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.

Xiaotian Li, Xiang Zhang, Huiyuan Yang, Wenna Duan, Weiying Dai, and Lijun Yin. An eeg-based multi-modal emotion database with both posed and authentic facial actions for emotion analysis. In *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)*, pp. 336–343. IEEE, 2020.

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. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024.

Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, Xujin Li, and Huiguang He. Animate your thoughts: Decoupled reconstruction of dynamic natural vision from slow brain activity. *arXiv preprint arXiv:2405.03280*, 2024.

Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022.

Erik Peitersen. Bell's palsy: the spontaneous course of 2,500 peripheral facial nerve palsies of different etiologies. *Acta oto-laryngologica*, 122(7):4–30, 2002.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023.

Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. Contrast, attend and diffuse to decode high-resolution images from brain activities. *Advances in Neural Information Processing Systems*, 36:12332–12348, 2023.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

Ling Wang, Chen Wu, and Lin Wang. Braindreamer: Reasoning-coherent and controllable image generation from eeg brain signals via language guidance. *arXiv preprint arXiv:2409.14021*, 2024a.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11333–11342, 2024b.

Mengting Wei, Tuomas Varanka, Xingxun Jiang, Huai-Qian Khor, and Guoying Zhao. Magicface: High-fidelity facial expression editing with action-unit control. *arXiv preprint arXiv:2501.02260*, 2025.

Haolin Xiong, Tianwen Fu, Pratusha Bhuvana Prasad, Yunxuan Cai, Haiwei Chen, Wenbin Teng, Hanyuan Xiao, and Yajie Zhao. Mind-to-face: Neural-driven photorealistic avatar synthesis via eeg decoding. *arXiv preprint arXiv:2512.04313*, 2025.

Muzhou Yu, Shuyun Lin, Lei Ma, Bo Lei, and Kaisheng Ma. Mindcustomer: Multi-context image generation blended with brain signal. In *Forty-second International Conference on Machine Learning*, 2025a.

Muzhou Yu, Shuyun Lin, Hongwei Yan, and Kaisheng Ma. Mindpainter: Efficient brain-conditioned painting of natural images via cross-modal self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14468–14476, 2025b.

Bohan Zeng, Shanglin Li, Xuhui Liu, Sicheng Gao, Xiaolong Jiang, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Controllable mind visual diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6935–6943, 2024.

Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

A LANDMARK-GUIDED PROGRESSIVE MASKING STRATEGY

A.1 REFERENCE-IMAGE MASKING

In Stage 3, masking is applied only to the reference image, while the target image remains unmasked. This design aims to suppress direct appearance leakage from expression-relevant facial regions in the reference image, thereby encouraging the model to rely more heavily on EEG conditioning when reconstructing the target expression. During training, we adopt a linear masking-ratio schedule: at the beginning, 70% of the reference images are masked over key facial regions and 30% are left unchanged; the masking ratio is then linearly annealed until 3,000 training steps, after which 30% of the reference images are masked and 70% remain unaltered.

We use the standard 68-point facial landmark representation to define six local facial subregions that are closely related to facial expressions. Specifically, the landmark indices are grouped as follows: left eye, right eye, left eyebrow, right eyebrow, nose, and mouth (Table 3). These regions are used to construct expression-relevant masks based on the activated Action Units (AUs).

Table 3: Definition of local facial regions based on 68 facial landmarks. Landmark indices follow the standard 68-point annotation protocol.

Facial region	Landmark indices
Left eye	36–41
Right eye	42–47
Left eyebrow	17–21
Right eyebrow	22–26
Nose	27–35
Mouth	48–67

To determine which facial regions should be masked, we map each activated AU to its corresponding local facial region. For AUs related to eyebrow motion, including AU1, AU2, and AU4, both eyebrow regions are selected. AU5 is associated with the eye regions, AU9 with the nose region, and mouth-related AUs, including AU12, AU15, AU17, AU25, and AU27, are mapped to the mouth region.

A.2 GEOMETRIC CONSTRUCTION OF EXPRESSION-RELEVANT MASKS

Given the AU-selected facial regions, we construct geometric masks from the corresponding 68-point facial landmarks. The goal is to mask not only the sparse landmark locations, but also the surrounding expression-related appearance cues, such as wrinkles, local contours, and skin deformations.

For the eyes, eyebrows, and mouth, we use expanded elliptical masks. Let $\mathcal{P}_r = \{(x_i, y_i)\}_{i=1}^{N_r}$ denote the landmark set of region r. We first compute the axis-aligned bounding box of the landmarks and define its center and half sizes as

$$c_x = \frac{x_{\min} + x_{\max}}{2}, \quad c_y = \frac{y_{\min} + y_{\max}}{2}, \tag{20}$$

$$h_w = \frac{x_{\max} - x_{\min}}{2}, \quad h_h = \frac{y_{\max} - y_{\min}}{2}. \tag{21}$$

The initial ellipse axes are obtained by region-specific scaling factors:

$$a_0 = s_x h_w, \quad b_0 = s_y h_h, \tag{22}$$

where a_0 and b_0 denote the horizontal and vertical semi-axes, respectively. We further allow a small region-specific center offset:

$$c_x \leftarrow c_x + o_x a_0, \quad c_y \leftarrow c_y + o_y b_0. \tag{23}$$

To ensure that all landmarks are covered, we first enlarge the axes according to the maximum landmark deviation:

$$a = \max\left(a_0,\ m_{\text{cover}} \max_i |x_i - c_x|\right), \tag{24}$$

$$b = \max\left(b_0,\ m_{\text{cover}} \max_i |y_i - c_y|\right). \tag{25}$$

We then verify landmark inclusion under the normalized ellipse equation:

$$\rho = \max_i \left[\left(\frac{x_i - c_x}{a}\right)^2 + \left(\frac{y_i - c_y}{b}\right)^2\right]. \tag{26}$$

If $\rho > 1$, the ellipse axes are further enlarged by a factor of $\sqrt{\rho}$, ensuring that all landmarks lie inside the ellipse:

$$a \leftarrow \sqrt{\rho}a, \quad b \leftarrow \sqrt{\rho}b. \tag{27}$$

Finally, we apply an additional post-expansion factor m_{post}:

$$a \leftarrow m_{\text{post}}a, \quad b \leftarrow m_{\text{post}}b. \tag{28}$$

The resulting elliptical mask for region r is defined as

$$\mathcal{M}_r = \left\{(u, v) \;\middle|\; \left(\frac{u - c_x}{a}\right)^2 + \left(\frac{v - c_y}{b}\right)^2 \le 1\right\}. \tag{29}$$

The region-specific parameters used in our implementation are listed in Table 4.

Table 4: Parameters for expanded elliptical masks. The same parameters are used for the left and right instances of eyes and eyebrows.

Region	s_x	s_y	o_x	o_y	m_{cover}	m_{post}
Eyes	1.55	1.85	0.00	0.00	1.10	1.18
Eyebrows	1.75	2.80	0.00	-0.18	1.12	1.18
Mouth	1.90	2.20	0.00	0.12	1.12	1.18

For the nose region, we use a convex-hull mask rather than an ellipse:

$$\mathcal{M}_{\text{nose}} = \text{ConvHull}(\mathcal{P}_{\text{nose}}). \tag{30}$$

This design is sufficient because the nose region is relatively compact and exhibits less non-rigid deformation compared with the eyes, eyebrows, and mouth.

Given the set of facial regions selected by activated AUs, the final expression-relevant mask is obtained as the union of all corresponding regional masks:

$$\mathcal{M} = \bigcup_{r \in \mathcal{R}_{\text{AU}}} \mathcal{M}_r, \tag{31}$$

where $\mathcal{R}_{\text{AU}}$ denotes the set of regions determined by the AU-to-region mapping.

A.3 Ablation Study

We conducted an ablation study on the use of the masking strategy. The results without the masking strategy are presented in Table 5. It can be observed that while the CSIM is very high and the FID is remarkably low, the accuracy is significantly poor. As shown in Figure 5, we found that without the masking strategy, the model suffers from severe "shortcut learning," where it simply replicates the reference image.

B Implementation Details

All experiments are conducted on 8 NVIDIA RTX 6000 Ada GPUs with 48GB memory. Stage 1 takes approximately 2 GPU-hours, Stage 2 takes approximately 8*36 GPU-hours, and Stage 3 takes approximately 8*48 GPU-hours.

Stage 1. We employ a Transformer-based asymmetric Masked Autoencoder (MAE)He et al. (2022) with a 6-layer encoder and a 4-layer decoder. The model is pre-trained on EEG signals with an input

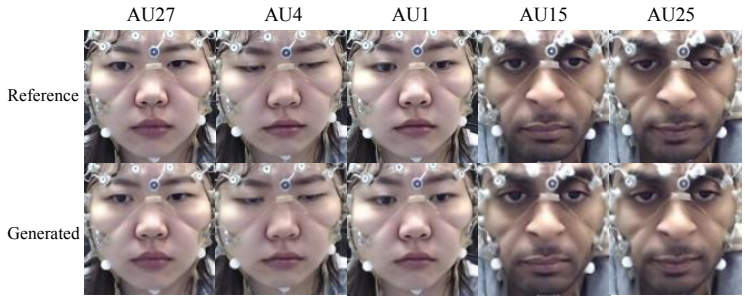


Figure 5: Qualitative comparison of the masking ablation under the self-reference setting.

Table 5: Ablation study on expression-relevant masking under the self-reference setting.

Method	CSIM(%) ↑	FID ↓	AU ACC(%) ↑	CLIP Score ↑
w/o Masking (Shortcut)	**90.45**	**35.12**	5.02	15.49
Ours (Full Model)	75.33	43.35	**27.71**	17.61

dimensionality of 128×48. To facilitate robust temporal representation learning, we apply a 50% masking ratio and optimize using AdamW for 80 epochs with a batch size of 32.

Stage 2. We align the EEG encoder with the Qwen2.5-VL Bai et al. (2025) model via a multi-task learning framework. Optimization is performed using AdamW for 5,000 steps with a global batch size of 64. To mitigate memory bottlenecks, we pre-extract and cache visual features for both reference and ground-truth images using the frozen Qwen2.5-VL vision encoder Bai et al. (2025). This strategy bypasses the need to load the heavy vision backbone during alignment, significantly accelerating training.

Stage 3. We initialize the model using pre-trained parameters from LongCat-Image Team et al. (2025). The model is trained for 5,000 steps at 512×512 resolution with a per-GPU batch size of 8.

We further construct two natural pipeline baselines with LongCat-Image Team et al. (2025) and FireRed-Image-Edit Team et al. (2026), decomposing EEG-conditioned editing into EEG-to-AU recognition followed by text-guided reference-based editing. In these pipelines, the downstream text-guided editors are kept frozen and used in a plug-and-play setting with the same reference image and fixed AU text prompts. These baselines evaluate a practical off-the-shelf alternative to EEG-conditioned editing, rather than a task-specific retraining of large text-guided editing models. For a fair comparison on the EEG decoding component, the EEG-to-AU classifier shares the same Stage-1 EEG encoder and AU classification head as MindAU, achieving 32.93% standalone AU accuracy on the held-out test split.

C E-CAFE Construction Details

In this section, we present a detailed visualization of E-CAFE. Fig. 6 illustrates a subset of the test dataset for Self-Referenced Editing. To improve the reliability of EEG–image supervision, we employ a two-stage filtering process: landmark-based automatic pre-filtering followed by manual verification. We first extract 68-point facial landmarks for each candidate target frame using an offline face-alignment detector and discard samples with missing frames, failed face detection, or invalid landmarks. We then compute simple AU-specific geometric scores from normalized landmarks, such as brow–eye distance, eye-opening height, nose–upper-lip deformation, mouth width, and inner-mouth opening, to remove weak or ambiguous AU activations.

The remaining samples are manually verified to remove residual landmark failures, ambiguous expressions, and reference frames with expression leakage. This filtering procedure improves the

quality of EEG–image supervision by retaining visually reliable and AU-consistent pairs for both EEG-to-expression alignment and final image editing. Fig. 7 displays a subset of face images generated by Z-Image Cai et al. (2025) used in Cross-Referenced Editing.

The 5,000 paired samples are balanced across the ten individually elicited AUs, with 500 samples per AU. For a target frame at time t, the 48-point EEG segment precedes that frame, and the reference face is selected at or nearest to the beginning of the corresponding EEG interval. Thus, each reference–EEG–target triplet represents a facial transition over the same interval.

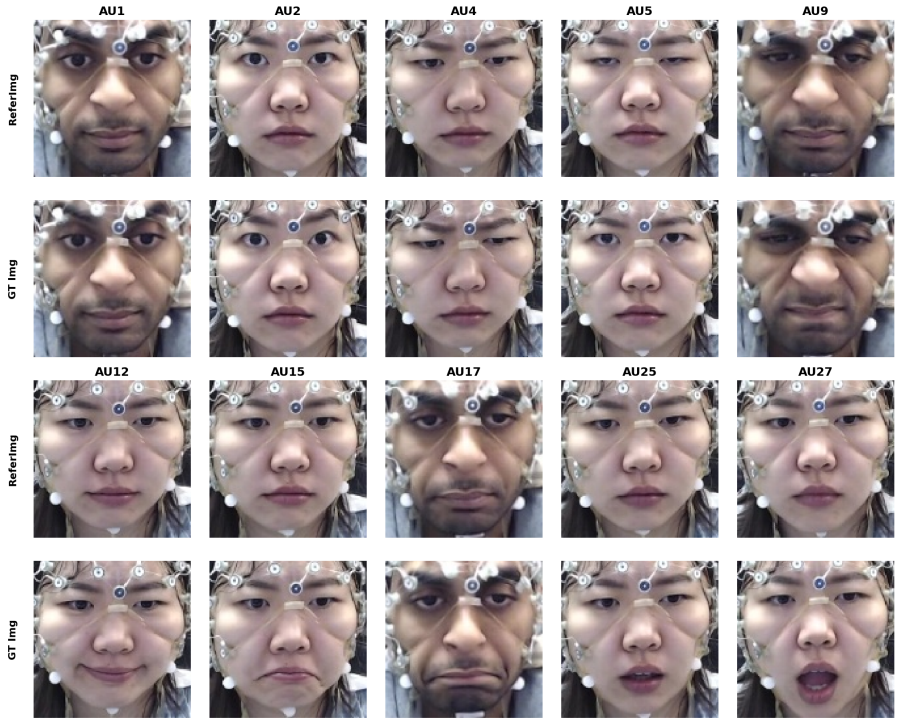


Figure 6: Visualization of samples from the Self-Referenced Editing test subset.



Figure 7: Visualization of reference faces from a subset of the Cross-Referenced Editing test dataset.

Table 6: Action unit (AU) labels and their corresponding text descriptions used in this work.

AU	Text Description
AU1	inner brows raised
AU2	outer brows raised
AU4	brows lowered and drawn together
AU5	upper eyelids raised
AU9	nose wrinkled
AU12	lip corners pulled up
AU15	lip corners pulled down
AU17	chin raised
AU25	lips parted
AU27	mouth wide open

Table 7: Ablation study on the effectiveness of generated pre-training data in the second stage.

Method	CSIM(%) ↑	FID ↓	AU ACC(%) ↑	CLIP Score ↑
Ours (w/o pretrain)	74.29	45.22	27.10	17.98
Ours (Full Model)	75.33	43.35	27.71	17.61

D AU Text Descriptions

Table 6 lists the AU-specific text descriptions used in this work. These descriptions serve as textual semantic targets for the Semantic Projection Stream in Stage 2, where they are encoded by the frozen Qwen2.5-VL text encoder for EEG–text alignment. The same descriptions are also used as fixed AU prompts for the text-guided pipeline baselines to ensure consistent semantic conditioning across methods.

E More Visualization

Fig. 8 and Fig. 9 present additional visualization results.

F Dual-Stream Manifold Alignment

As shown in Fig. 10, we visualize the features of EEG tokens extracted from the train set. The results indicate that our EEG Adapter effectively discriminates between distinct AU features while eliminating interference from different subjects. To investigate the impact of utilizing generated pre-training data in the second stage, we conducted an ablation study, as shown in Table 7. The results demonstrate that our full model outperforms the variant without pre-training (w/o pretrain) across key metrics, achieving a higher CSIM, a lower FID, and an improved AU ACC.

G Failure Case Analysis

Figure 11 shows representative failure cases of MindAU. Although the proposed landmark-guided reference masking strategy reduces reference-dominant shortcut learning, the model can still occasionally under-utilize EEG conditions when the target expression is subtle or the reference and target images are visually similar. Besides incorrect AU generation, we observe a small number of cases where the output nearly copies the reference image, producing insufficient expression change. This suggests that fully preventing reference shortcut behavior remains challenging under limited paired EEG–face supervision.



Figure 8: Qualitative results of Self-Referenced Editing on the test dataset.

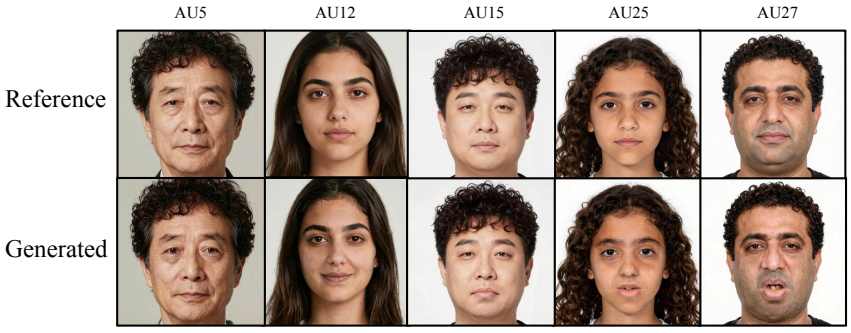


Figure 9: Qualitative results of Cross-Referenced Editing on the test dataset.

H Inference Details

H.1 Inference Pipeline

At inference time, we follow the same conditioning structure used during Stage-3 training. Given a reference face image and an input EEG segment, we first preprocess the reference image into two branches: a 512×512 branch for the VAE encoder and a 256×256 branch for the Qwen2.5-VL encoder. The text condition uses the same fixed prompt as training, namely *"a face showing facial action unit"*. This fixed text prompt is encoded together with the reference image by the Qwen2.5-VL encoder, and we retain the prompt-token hidden states corresponding to the editable text span.

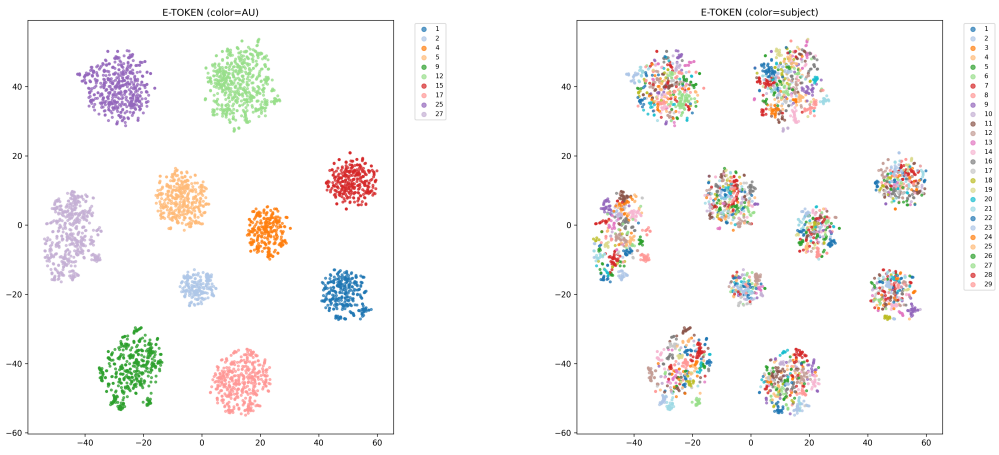


Figure 10: Visualization of E-token feature clustering extracted by the EEG adapter on the training set. Left: labeled by AU; Right: labeled by Subject.

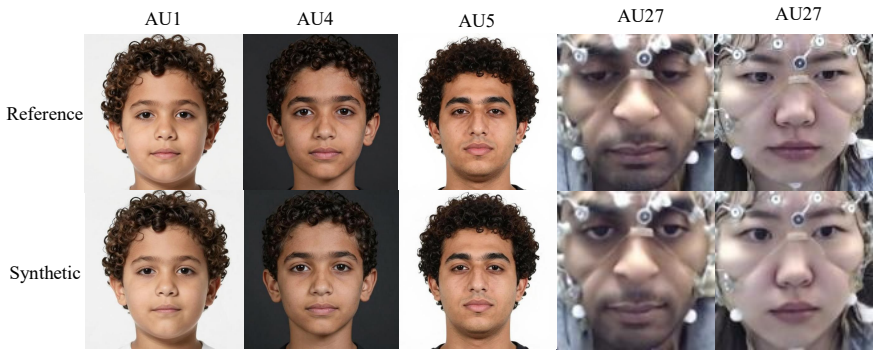


Figure 11: Representative failure cases of MindAU. Some outputs exhibit incorrect AU generation or insufficient expression change, where the generated image remains overly close to the reference image.

The EEG signal is processed by the Stage-1 EEG Encoder and Stage-2 EEG adapter. The adapter maps the EEG input into a sequence of 32 E-tokens, each with hidden dimension 3584. These E-tokens are appended to the text prompt embeddings, yielding the final multimodal conditioning sequence. We keep distinct modality identifiers for text and EEG tokens, so that the diffusion transformer can distinguish linguistic and EEG conditions while attending to both within a unified context space.

The reference image is encoded by the VAE into latent space, while the target latent is initialized from Gaussian noise. We use the same FlowMatch Euler scheduler as in training-time validation, with 30 denoising steps for train–test consistency. After denoising, the predicted latent is decoded by the VAE and resized back to the original reference-image resolution for saving.

H.2 EEG-Free Inference

To isolate the contribution of EEG conditioning, we further introduce an EEG-free inference setting. In this setting, the overall inference pipeline remains unchanged, including the reference image, fixed text prompt, LoRA checkpoint, and sampling configuration. The only modification is applied to the EEG branch: instead of feeding the EEG-derived tokens, we replace the appended E-token block

Table 8: Comparison between EEG-free inference and EEG-conditioned model under self-reference protocol.

Method	CSIM(%) ↑	FID ↓	AU ACC(%) ↑	CLIP Score ↑
EEG-free	25.12	136.2	7.22	14.43
Ours (Full Model)	75.33	43.35	27.71	17.61

Table 9: Stage-1 AU classification accuracy under frequency and channel controls.

Channels	Preprocessing	AU ACC (%)
All	Original broadband	47.93
Without face-adjacent channels	Original broadband	29.70
All	0.1–50 Hz	32.93
Without face-adjacent channels	0.1–50 Hz	28.95

with an all-zero token matrix of the same shape:

$$\mathbf{E}_{\text{free}} = \mathbf{0} \in \mathbb{R}^{N_e \times d},$$

where $N_e = 32$ denotes the number of EEG tokens and $d = 3584$ is the token embedding dimension.

As shown in Table 8, removing EEG conditioning leads to a substantial performance degradation across all evaluation metrics. Specifically, the EEG-free setting obtains only 25.12% CSIM and 7.22% AU ACC, while the full model achieves 75.33% and 27.71%, respectively. Meanwhile, the FID increases sharply from 43.35 to 136.2, indicating a clear decline in image quality and distribution fidelity.

I Additional Evaluation and Controls

I.1 EMG-Related Sensitivity Analysis

To examine possible facial-muscle contributions to AU decoding, we compare Stage-1 classification with and without face-adjacent EEG channels, using either the original broadband signal or the 0.1–50 Hz preprocessing used by MindAU. As shown in Table 9, removing face-adjacent channels reduces accuracy from 47.93% to 29.70% before filtering, compared with 32.93% to 28.95% after filtering. This suggests that band-pass preprocessing reduces the contribution of activity concentrated near the face. However, removing these channels can also remove neural information, and EMG can overlap with the retained frequency range. This diagnostic does not isolate a purely neural signal or establish transfer to attempted or imagined actions.

I.2 Fixed-Reference EEG Control

We select one neutral face from the AU1 test subset as a global reference and apply EEG segments from different held-out participants and AU categories to this same image. Thus, the visual input and inference pipeline remain fixed while the EEG input changes. A stricter AU-wise setting excludes reference-subject EEG for the associated AU categories while retaining that subject's EEG for other categories.

Table 10: Editing with a single fixed neutral reference face.

Setting	CSIM (%) ↑	FID ↓	AU ACC (%) ↑	CLIP ↑
Original self-reference	75.33	43.35	27.71	17.61
Fixed reference, cross-subject EEG	76.43	42.72	27.53	17.57
Fixed reference, AU-wise exclusion	78.32	41.97	26.91	17.45

Table 10 compares these settings with the original self-reference protocol. With one fixed face, AU accuracy remains close to the original result (27.53% versus 27.71%), while CSIM, FID, and CLIP change only modestly. The stricter AU-wise exclusion retains 26.91% AU accuracy. This is a targeted diagnostic with one global reference image; it does not replace a comprehensive multi-reference evaluation or a paired-versus-shuffled EEG control.

I.3 Additional MagicFace Baseline

We also evaluate a classify-then-edit pipeline using MagicFace (Wei et al., 2025) with the same reference images and predicted AU conditions. MagicFace does not support AU27, so AU accuracy is compared on the nine AU categories supported by both methods: 17.29% for MagicFace and 29.06% for MindAU. MagicFace often makes only small changes to the reference face, which can favor identity and image-distribution metrics without producing the requested AU deformation. The nine-AU result is kept separate from the ten-AU comparison in Table 1.

I.4 Blinded Perceptual Evaluation

We conduct a blinded ranking study on 30 held-out cases with ten lay participants and three independent domain experts. Each rater sees the reference image, target AU description, ground-truth target, and anonymized outputs from DreamDiffusion, LongCat-Image, FireRed-Image-Edit, and MindAU. Output order is independently randomized. Raters rank the four outputs from 1 (best) to 4 (worst) for AU and intensity match, identity preservation, realism, and absence of artifacts. Table 11 reports both mean rank and Top-1 preference rate.

Table 11: Blinded perceptual study: mean rank / Top-1 preference rate (%). Lower mean rank and higher Top-1 rate are better.

Raters	Method	AU/intensity	Identity	Realism	Artifact-free
Lay	DreamDiffusion	3.76 / 0.0	4.00 / 0.0	2.73 / 16.7	2.74 / 15.7
Lay	LongCat-Image	2.77 / 2.7	2.54 / 9.0	3.06 / 12.0	3.13 / 12.0
Lay	FireRed-Image-Edit	2.31 / 7.3	2.23 / 13.7	2.30 / 27.3	2.36 / 23.3
Lay	MindAU	1.16 / 90.0	1.23 / 77.3	1.91 / 44.0	1.76 / 49.0
Experts	DreamDiffusion	3.82 / 0.0	4.00 / 0.0	2.71 / 14.4	2.68 / 18.9
Experts	LongCat-Image	2.73 / 2.2	2.57 / 5.6	3.04 / 12.2	3.19 / 10.0
Experts	FireRed-Image-Edit	2.36 / 4.4	2.28 / 10.0	2.24 / 33.3	2.36 / 22.2
Experts	MindAU	1.09 / 93.3	1.16 / 84.4	2.00 / 40.0	1.78 / 48.9

I.5 Exploratory Multi-AU Composition

As an inference-only exploration, we sum learned embeddings from three separately recorded single-AU EEG segments and condition the frozen editor on the combined embedding. In two of ten tested combinations, the output qualitatively displays cues of all three intended AUs. Because the EEG segments were not recorded during simultaneous AU activation and most combinations were unsuccessful, this result does not establish reliable multi-AU EEG decoding or editing.

J Asset Licenses and Terms of Use

Table 12 lists the third-party assets and their licenses or access terms. We use them for research, do not redistribute third-party data or weights unless permitted, and follow provider access agreements when no standard public license is available.

Table 12: Existing assets used in this work and their licenses or terms of use.

Asset	License or terms of use
BU-EEG	Original BU-EEG data-use terms / access agreement.
Qwen2.5-VL-7B-Instruct	Apache-2.0 for the 7B checkpoint. If a different Qwen2.5-VL variant is used, we follow the license attached to that specific checkpoint.
LongCat-Image-Edit	Apache-2.0.
FireRed-Image-Edit-1.0	Apache-2.0 for code and weights.
Z-Image	Apache-2.0.
Gemini / Nano Banana 2	Google Gemini API Additional Terms of Service.
Face Alignment Network	BSD-3-Clause.
OpenGraphAU	MIT for the official repository code.
ArcFace / InsightFace	MIT for code; released pretrained models may be restricted to non-commercial research use depending on the checkpoint.